Computational Methods for Single-Cell RNA Sequencing

ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Computational Methods for Single-Cell RNA Sequencing

Brian Hie,[1,*] Joshua Peters,[2,3,*] Sarah K. Nyquist,[1,3,4,*] Alex K. Shalek,[3,5] Bonnie Berger,[1,6] and Bryan D. Bryson[2,3]

[1]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; email: bab@mit.edu

[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; email: bryand@mit.edu

[3]Ragon Institute of MGH, MIT, and Harvard, Cambridge, Massachusetts 02139, USA

[4]Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[5]Department of Chemistry, Institute for Medical Engineering & Science (IMES), and Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[6]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Annu. Rev. Biomed. Data Sci. 2020. 3:339–64

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-012220-100601

*These authors contributed equally to this article

Keywords

computational methods, single-cell RNA sequencing, data integration, gene regulatory networks, dimensionality reduction, clustering

Abstract

Single-cell RNA sequencing (scRNA-seq) has provided a high-dimensional catalog of millions of cells across species and diseases. These data have spurred the development of hundreds of computational tools to derive novel biological insights. Here, we outline the components of scRNA-seq analytical pipelines and the computational methods that underlie these steps. We describe available methods, highlight well-executed benchmarking studies, and identify opportunities for additional benchmarking studies and computational methods. As the biochemical approaches for single-cell omics advance, we propose coupled development of robust analytical pipelines suited for the challenges that new data present and principled selection of analytical methods that are suited for the biological questions to be addressed.

1. INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) technologies generate datasets that describe the state of individual cells with unprecedented resolution. Advances in technologies for scRNA-seq have resulted in increasingly complex datasets with millions of cells and thousands of features captured per cell (1). The number of single-cell transcriptomes reported has exploded since the introduction of the technology and has been further fueled by global efforts like the Human Cell Atlas (2–5). However, unlike their bulk counterparts, scRNA-seq data experience unique challenges derived from sparsity, heterogeneity, and scale. These data also provide opportunities to derive novel insights about the behavior of subpopulations of cells. As a consequence of these unique challenges and opportunities, new computational methods for scRNA-seq have become necessary for advancing the field of single-cell omics.

A range of technologies are available for users to generate single-cell transcriptomes, and these technologies have been benchmarked by several groups (6–8). In parallel with the explosive growth of biochemical methods for single-cell omics, there has been complementary growth in the number of computational tools available for the analysis of these data (9). The wealth of computational methods available now places an increasing onus on researchers to choose the right tools for the job.

Computational methods for scRNA-seq analysis span a range of functions, from alignment to quantification, batch correction, dimensionality reduction, and clustering (**Figure 1**). Ultimately, scRNA-seq analysis requires principled method selection and execution informed by the characteristics of the data and the biological hypotheses (see the sidebar titled Experimental Planning for Effective Computational Analysis). Here, we review an array of computational methods available for scRNA-seq analysis, highlight benchmarking studies of existing computational methods, and identify opportunities for the future.

2. PREPROCESSING

The typical starting point for scRNA-seq data analysis is a sparse gene expression matrix describing gene abundances per cell. These data are based on the number of reads aligned to each annotated gene or transcript. Experiments that leverage unique molecular identiers (UMIs) are not normalized to gene length, whereas full-length protocols normalize to length by converting to units of transcripts per million. Principled generation of the expression matrix is critical because it is the foundation for most downstream analyses.

2.1. Alignment and Quantification

High-throughput sequencing data are processed to generate an expression matrix. First, reads must be appropriated to their respective cellular barcode in a process known as demultiplexing. The reads are then aligned to a reference genome then quantified per gene or transcript. Commonly, splice-aware aligners like STAR (10), HISAT2 (11), and Tophat2 (12) are paired with expression quantification methods like RSEM (13), htseq-count (14), or featureCounts (15). Newer tools have combined alignment and quantification, such as the Rsubread package (16) and Cell Ranger pipeline (17). Several parameters can be tuned in these algorithms, primarily focused on ambiguous or multimapped reads. Van den Berge et al. (18) have reviewed the logic of and considerations for method selection for alignment, and we refer the reader there. Recent benchmarking publications from the Human Cell Atlas used zUMIs (19) and scumi (7) to standardize their benchmarking efforts across most scRNA-seq protocols, even those without UMIs. Several notable improvements of common tools have focused on efficiency and scalability,

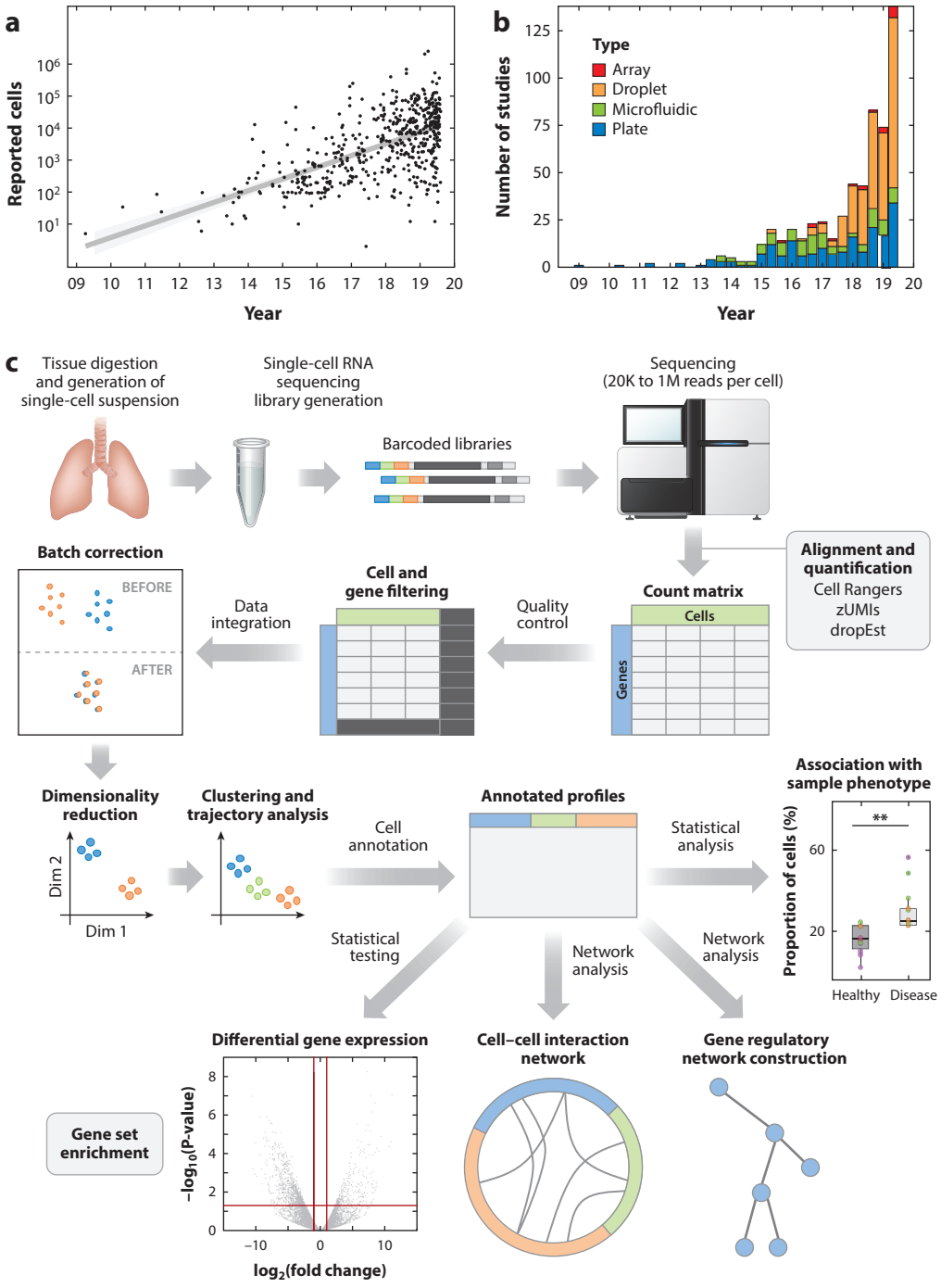

Figure 1

Single-cell RNA sequencing has seen enormous growth since its first demonstration in 2009, as evidenced by the number of sequenced cellular transcriptomes reported per study. (*a*) Number of cells reported per study plotted over time. (*b*) Quarterly summary of technologies used per study published. Panel *a*,*b* data from Reference 4. (*c*) Common single-cell RNA sequencing analysis pipeline annotated with popular approaches and tools.

EXPERIMENTAL PLANNING FOR EFFECTIVE COMPUTATIONAL ANALYSIS

For successful analyses, there are several factors to consider prior to performing an experiment. Addressing these issues beforehand can improve the subsequent computational analyses.

Choosing How Many Cells and How Many Reads per Cell

Choosing the number of cells and the depth of sequencing remains a challenging question. While Zhang et al. (25) have suggested that one read per gene per cell is sufficient to capture a gene's distribution, this may shift based on the biological genes and cell types of interest. For instance, more reads may be worthwhile for a time course experiment, whereas sampling more cells may be preferred for cell type discovery. Recently, Svensson et al. (26) suggested that there is substantial benefit from increased sequencing depth until 15,000 reads per cell, after which cell number or sequencing depth provides similar marginal benefit.

Minimizing Batch Effects

Unbalanced experimental design can lead to batch effects confounding computational analyses (27–29). Balancing experimental conditions across individual technical runs is critical. Several methods have been introduced to identify and control for batch effects, most notably surface labeling (30, 31). These reagents allow samples to be pooled for improved experimental design. Employing these strategies can improve analyses, as they can enhance doublet detection and reduce batch effects.

Identifying Potential Cell Types Present

Although one of the most attractive applications of scRNA-seq is high-dimensional cytometry, prior knowledge of what cell types are present and their proportions can prove useful. This information can be used to inform filtering metrics for barcodes and cells, normalization, and rare cell identification. If rare cell types exist and are of interest, these cells may need to be enriched prior to the experiment (32).

Processing Artifacts

Recently, numerous groups have highlighted the impact of dissociation on cellular profiles (33, 34). O'Flanagan et al. (34) identified a stress response, including *FOS*, *JUN*, and heat shock proteins, mediated by collagenase dissociation for certain subsets of cells. Considering the effects of sample processing can help identify expression patterns that may confound clustering and arise in differential expression analysis.

notably STARsolo (10), bustools (20, 21), and Alevin (22). scRNA-seq experiments of thousands of cells experience additional demultiplexing challenges including sequencing errors and hopping of barcodes/indices, which can prevent the collapse of reads to their appropriate UMI, cell, and library barcode. In large-scale, complex experiments, these issues require attention, and several methods are available to correct for UMI and cellular barcode errors (23) and index hopping (24).

2.2. Quality Filtering

After generation of the expression matrix, barcodes may remain that represent unwanted transcriptomes like doublets, dying cells, or contamination. Singlet identification and ambient correction can be performed to address these concerns. It can be informative to begin with lenient parameters initially and revisit these steps after developing a sense of the biological structure of the data. We refer to transcriptional profiles that confidently map to unique barcodes as cells, although we recognize that confident identification of single cells is a nontrivial challenge.

2.2.1. Singlet identification. In droplet-based scRNA-seq methods, there are risks that barcodes may not represent bona fide cells or that more than one cell may be paired with a single barcode. These empty droplets and multiple-cell barcodes, called doublets, confuse downstream analyses by introducing artificial populations of cells. While empty droplet identification methods like EmptyDrops (35) are typically included in pipelines, doublet identification is not. For droplet- and array-based methods, computational doublet identification methods include DoubletFinder (36) and Scrublet (37), which simulate doublets by pairing single cells to classify true doublets through nearest neighbor approaches. Identification of doublets can be visualized in low-dimensional space to assess the validity of the results. Several methods such as scSplit (38) and souporcell (39) have been developed to offer genotype-free demultiplexing and doublet detection. Unlike demuxlet, souporcell calls genomic variants using realigned reads and clusters these to identify doublets, demultiplex samples, and account for ambient RNA (39). Identifying potential doublets is an important step in analyses to avoid potentially confounding profiles in later analyses.

2.2.2. Ambient correction. Ambient RNA captured during scRNA-seq experiments also represents a technical artifact that can confound downstream analysis. Dissociation of tissue calls for the assessment of potential ambient RNA contamination originating from abundant RNAs from lysed or dying cells prior to emulsification. As a result, cells may also contain free-floating RNA, thus confounding cellular profiles. Vieira Braga et al. (40) utilized SoupX (41) to identify ambient RNA profiles from empty droplets based on the number of UMIs per barcode. SoupX also offers a method to correct the expression of single cells that requires prior knowledge about cluster-specific gene expression to estimate ambient contamination, which may be challenging in novel applications. Smillie et al. (42) grouped cells based on broad cell types and then compared expression of established markers within each group to all other types. By fitting a linear model to this relationship and analyzing the residuals, they identified contaminating genes to be excluded in differential expression (DE) results. Given the importance of ambient correction as a quality control consideration, comparatively few applicable methods exist, and the problem would benefit from further methodological development.

2.2.3. Low-quality cells and genes. It is important to identify low-quality cells and genes that are present within individual samples in order to ensure effective downstream analyses. Identifying low-quality cells can be challenging and is highly dependent on the biological properties of the samples being analyzed. It is worth noting that low transcript capture may not always reflect technical artifacts and merits careful consideration. Beyond cells with low-quality sequencing metrics, the percentage of mitochondrial reads is most commonly used to exclude cells. Illicic et al. (43) used the C1 platform to visually inspect and classify cells before sequencing. They identified mitochondrial reads as a major predictor of cell quality, reasoning that increased reads indicate broken or lysed cells prior to mRNA capture. Other gene families such as heat shock proteins and ribosomal proteins, as well as technical factors such as library size and mapping rates, can also be used to perform more comprehensive quality control. Genes that are expressed in a low percentage of cells can also be removed, although this number of cells should also be significantly less than the expected proportion of rare cell types. It is often desirable to have a less stringent initial filtering followed by more stringent and specific artifact identification during downstream analyses like clustering and DE in order to prevent removing bona fide cell profiles.

2.3. Data Sketching and Summarization

Given the growing size of scRNA-seq data (2), some methods have been developed to reduce the number of data points to consider, thereby minimizing the amount of computational resources

required for large-scale analysis. One approach, data sketching, is to obtain a small, representative subset of cells. The simplest sketching strategy is to randomly select a subset of cells, where each cell has an equal probability of being chosen. Instead of uniform sampling, an efficient density-aware sampling method called geometric sketching (44) preferentially selects a heterogeneous subset of cells, therefore preserving rare cell types within the sketch. Another approach to reduce analytic complexity is to summarize a dataset by first clustering the cells with a scalable algorithm (Section 5.6) and by then performing further analysis only on the statistics computed for each cluster of cells, such as average gene expression (45, 46) or gene coexpression (47). Summarizing data based on clusters of cells has also been used to accelerate algorithms for trajectory analysis (48). Although summarization methods will lose some amount of information, these approaches enable the rapid analysis of millions or even billions of cells.

2.4. Imputation

scRNA-seq data are highly sparse (i.e., many gene expression values are measured as zero) and most nonzero measurements are low abundance (for example, one or two UMI counts), especially for experiments performed using less efficient scRNA-seq technologies or lower sequencing depth. Small technical artifacts such as experimental contamination or computational alignment errors could result in data with a low signal-to-noise ratio (SNR), thus reducing the ability of downstream analysis to distinguish subtle biological patterns. One approach adopted by some is to improve the SNR by imputing scRNA-seq signal based on strong, high-abundance signal, either by amplifying correlated low-abundance signal or even by inferring a positive expression for values measured as zero by the original experiment (49–53). A separate line of work in the scRNA-seq quality control literature, however, has focused on profiling imputation methods and their underlying assumptions, particularly since imputation methods assume different noise models and the noise patterns themselves can vary greatly among experiments (54, 55). In particular, Andrews & Hemberg (55) found that all of the imputation methods in their benchmark introduced false positive signal into DE and marker gene analysis. It should be noted that analyses with low tolerance for false positives can still operate in the high-abundance regime of unimputed data and obtain informative results. Moreover, increasing the SNR of scRNA-seq experiments when profiling very subtle biological processes will benefit most from collaborative efforts among data analysts and technologists, not only with regard to computational signal processing but also in improvements to the underlying biochemical approaches themselves.

3. ACCOUNTING FOR TECHNICAL AND BIOLOGICAL VARIATION

Several sources of variation (biological and technical) are present in the count matrix. Normalization, scaling, variable gene identification, and latent variable modeling are employed to minimize technical variation, while retaining biological variation (**Figure 2**). The field has not reached a consensus on which methods and transformations are most appropriate for scRNA-seq analysis.

3.1. Normalization

Normalization seeks to enable consistent comparison of gene measurements across cells, including technical variation due to the number of reads sequenced or the number of transcripts identified per cell (57–59). Normalization methods are dependent on the specific experimental characteristics of a given study. In software packages like Scanpy, Cell Ranger, and Seurat, gene counts for a specific cell are normalized by the total number of counts per cell (known as the size factor), scaled by a factor (commonly around 10^4 to 10^6), added to some pseudocount, and then log-transformed

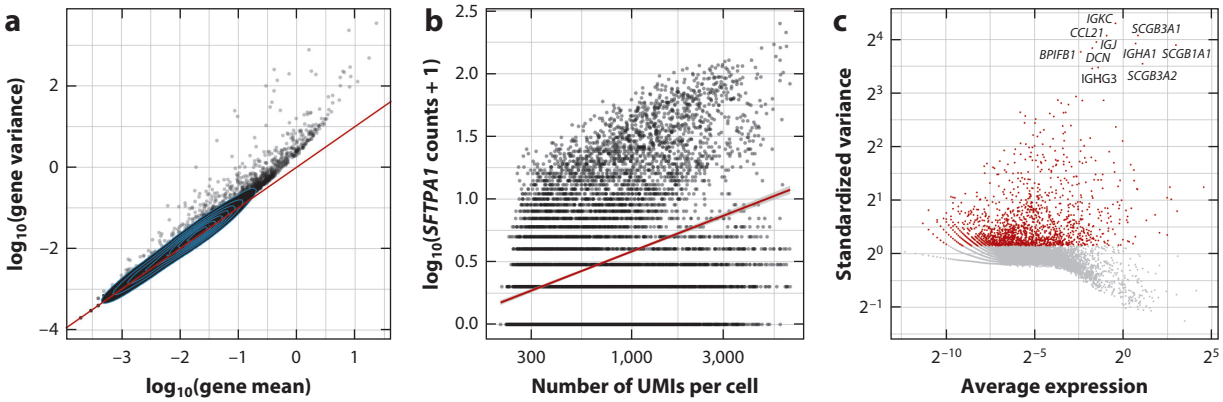

Figure 2

Inspecting the properties of cellular profiles is critical. (*a*) A plot of $\log_{10}$(variance) versus $\log_{10}$(mean) highlights the overdispersion compared to the Poisson model (*red*). (*b*) A plot of log-transformed *SFTPA1* counts versus unique molecular identifiers (UMIs) per cell shows a dependence on expression depth that needs to be accounted for. (*c*) Common visualization and variable gene selection method (56) for downstream analyses. Data in panels *a–c* are from lung Drop-seq data from Vieira Braga et al. (40).

(56, 60). Normalizing each cell by its size factor can be impacted by outlier gene expression and ignores biologically relevant differences in cell size and total mRNA between cell types and states. Potential artifacts can be introduced by this transformation when size factors are diverse and gene counts are low (61). Although these methods are popular and perform well (62, 63), additional methods have proposed various modeling assumptions concerning the sparsity and the underlying distribution of gene expression within cells. SCnorm (58), scran (61), Linnorm (64), Census (65) and DESeq2 (66) have all proven successful in certain cases, but the performance of these methods can vary between datasets (62, 67). Many methods apply a log transformation to the count matrix to reduce the overall influence of the naturally higher variability of high-abundance genes. Recent publications highlight issues with this approach; in particular, Townes et al. (68) highlighted that zero inflation is introduced as data are log-transformed, and as Lun (69) also showed, this can introduce artifacts that resemble heterogeneous structure. Lun suggested filtering out low-quality cells and increasing the pseudocount added before log transformation (69).

Recently, SCTransform offers a new approach that models the counts per gene as a result of sequencing depth using a regularized negative binomial model (59). The difference between each gene's predicted expression and measured expression thereby represents the biological expression of the gene within that cell with technical variation removed. These statistical parameters, such as the expression depth dependence, can be visualized to highlight confounding dependence pre- and postnormalization. As technologies advance to improve transcript capture, comprehensive benchmarking across normalization methods and technologies is needed to determine best practices.

3.2. Variable Gene Identification

Identifying highly variable genes is primarily used to focus downstream dimensionality reduction and clustering on a subset of genes that are responsible for much of the biological variation within the data. Brennecke et al. (70) introduced the identification of variable genes using the squared coefficient of variation (CV^2) across cells. A curve is fit to the normalized read counts against the CV^2 and per-gene CV^2 is tested against a threshold of basal biological-derived variance. Klein

et al. (71) employed a modification of this model that utilizes a different distribution for their calculated measure of gene variance above basal noise. Seurat v3 (56) and scran (61) implement similar strategies, where a loess curve is fitted to the mean of each gene and plotted against its variance. In contrast, each method utilizes different ranking procedures: scran tests if a corrected variance for each gene is greater than zero, and Seurat v3 ranks genes based on the variance of standardized gene values per cell. The number of variable genes selected can range from 500 to 5,000 depending on cutoffs, although 2,000–5,000 variable genes can be selected for complex data that are properly normalized. Yip et al. (72) evaluated several variable gene selection methods based on their clustering performance, which favored scran by highlighting broad disagreement between methods. Given that this benchmarking study also included normalization and only a few small datasets, more benchmarking of feature selection methods is warranted.

3.3. Integration

While typical unsupervised scRNA-seq methods are designed to find biological structure in the underlying data, not all of this variation is interesting. For example, multiple technical replicates of the same biological sample may contain noise patterns specific to experimental batch that are then reflected in downstream analysis, such as clusters that separate cells according to both cell type and batch. In addition to batch effect correction, researchers may desire to transform scRNA-seq data in a way that eliminates other sources of undesired variation, a process termed integration. A common application of integration is to preserve cell type heterogeneity but remove differences due to technical effects, like experimental batch or sequencing technology, or biological effects, like donor variation or evolutionary differences separating biological species. Importantly, integrative methods require specifying what variation to preserve and what variation to remove, which may vary based on the desired downstream application.

Methods applied to bulk RNA-seq data can also be applied in the scRNA-seq setting. These include linear systems that model and remove known covariates (73) such as the percentage of mitochondrial reads, the average expression of curated cell cycle genes (74), and the number of UMIs per cell. Nonlinear integration methods developed for bulk RNA-seq try to model the noise distribution associated with known covariates across all of the given observed data (75). Bulk models, however, often cannot model more complex noise distributions associated with scRNA-seq data, which contain a more heterogeneous collection of cells (76, 77) and do not scale well to the large numbers of transcriptomes observed in single-cell experiments.

Many methods have been developed specifically for scRNA-seq dataset integration, in which intradataset variation is mostly preserved and interdataset variation is prone to be removed. The most common class of methods for integration are locality based, where locality is typically computed as Euclidean distance in gene expression space, or a similar distance metric. Many locality-based methods transform data to minimize the distance between the nearest neighbors of cells in one dataset within another dataset. Haghverdi et al. (76) introduced the mnnCorrect method based on a robust nearest neighbor alignment algorithm known as mutual nearest neighbors (MNN) or best buddies matching (78), in which two cells are aligned across two datasets only if each was a nearest neighbor of the other. The Scanorama method (79) built upon this MNN-based strategy to enable practical integration of many large-scale experiments by focusing on computational efficiency and minimizing overcorrection of interdataset variation. Subsequent methods including BBKNN (batch-balanced *k*-nearest neighbors) (80), Conos (clustering on network of samples) (81), and Seurat v3 (56) have further refined nearest neighbor–based approaches with emphases on scalability and minimizing overcorrection. Although not based on nearest neighbors matching, the Harmony method (82) is another locality-based integration method that uses a scalable soft-clustering approach to iteratively remove distances representing undesired variation among cells.

Other approaches to integration rely on learning similar structure shared across datasets. The Seurat v2 method (77) uses canonical correlation analysis to transform datasets into an embedding space that maximizes the similarity of interdataset structure. The LIGER (linked inference of genomic experimental relationships) method (83) employs a linear model that decomposes scRNA-seq data into a signal that is dataset specific and a signal that is shared across all datasets, while the scVI (single-cell variational inference) method (84) accomplishes a conceptually similar decomposition but instead uses a latent variable model instantiated by nonlinear neural networks. Rather than transforming data before applying downstream clustering methods, other methods first cluster datasets separately and then apply locality- or correlation-based approaches to identify similar clusters across datasets (56, 85). The Coscape method leverages locality in correlation space to achieve integrative analysis that does not assume all interdataset variation should be removed, enabling comparison of datasets separated by meaningful biological variation. For example, Coscape is able to reconstruct developmental trajectories across temporally diverse studies (47).

Different approaches can quantify the extent of undercorrection, in which unwanted variation persists after integration, and overcorrection, in which desired variation is removed or obscured by integration. These strategies first require specifying a label for each cell, where cells with the same label should be more similar than cells with different labels. The silhouette coefficient (86), used in several scRNA-seq integration studies (62, 76, 79), assigns a score to each cell that increases if the cell is close to cells with the same label (based on a distance metric, such as Euclidean distance in integrated embedding space) and decreases if the cell is close to cells with different labels. Other approaches to quantifying integration quality have been based on statistical hypothesis testing (63), information theoretic measurements of diversity (80, 82), and the composition of local nearest neighborhoods (56, 63). Visualizing integrated datasets, which is reviewed in Section 4.2 below, can also provide intuition into the transformed data.

Integrative efforts have moved beyond the realm of single-cell transcriptomics to include additional genomic, epigenomic, proteomic, and spatial modalities, which has has been reviewed at length by Stuart & Satija (87). While many integrative challenges are similar to those in the scRNA-seq setting, a particularly important challenge unique to multimodal integration is weighing how much information to consider from different data types, especially when measurements of the same feature disagree across different modalities.

4. DIMENSIONALITY REDUCTION AND REPRESENTATION LEARNING

scRNA-seq data represent a set of cells wherein each cell is featurized by gene abundances, but additional transformations of these features such as dimensionality reduction are useful in many instances as well.

While computational pipelines for gene quantification can often identify transcripts belonging to the tens of thousands of possible genes within an organism, only a subset of these genes will be both present in the assayed sample and measured with nontrivial variability. Most genes will have zero abundance in most or all cells and can be removed with minimal impact to the results of downstream analysis; however, removing lowly expressed genes does improve the efficiency of computational analyses, which often have a runtime or memory dependence on the number of genes considered. Identifying such genes to remove is often as simple as setting a filter cutoff based on the percentage of cells with nonzero expression or based on a summary of the total expression of that gene (e.g, mean expression across all cells). Other methods for gene filtering are based on statistical measures of expression variability and are discussed in greater detail in Section 3.2.

In addition to removing lowly expressed genes, a key concept underlying further dimensionality reduction of scRNA-seq data is that cells do not have completely random gene expression values and are not evenly distributed across all data dimensions. Not only are many genes lowly expressed, but many genes have highly correlated expression values across cells, resulting in large amounts of information redundancy. As a result, the effective dimension of the data is often much lower than the total number of genes profiled, a property inherent in many types of biological data (88–90). The scRNA-seq literature sometimes refers to this property as single cells occupying a "low-dimensional manifold."

4.1. Linear Decomposition Methods

Many dimensionality reduction techniques aim to learn a compact set of features by combining information across multiple genes into each individual feature. Most feature summarization methods are based on linear decomposition models like principal component analysis (PCA; finds orthogonal features of maximum variation), independent component analysis (finds statistically independent features that best reconstruct the original data), or nonnegative matrix factorization (finds features, often interpreted as gene modules, that combine expression across multiple correlated genes); these methods are reviewed in the context of genomic data by Stein-O'Brien et al. (91). Instead of a cell being featurized by hundreds or thousands of genes, a cell is instead featurized by a much smaller number of components.

Notably, the analyst must choose the number of decomposition components. This parameter selection is typically done by comparing the information contained in the original dataset to the information contained in the lower-dimensional dataset, as quantified by the respective objective function for each dimensionality reduction method. In nearly all biological data, the amount of new information captured within the lower-dimensional dataset will decrease as the number of new components increases (89, 91). To select the number of components, researchers typically choose a cutoff after which the marginal information gain from using additional components is minimal, for example, by selecting the number of principal components that captures 95% of the total dataset variance, or by visually inspecting the value of the objective function as it decreases with the number of components, as in Reference 83. In general, feature summarization using linear decomposition methods results in a much more efficient representation that captures much of the biological variation among cells and reduces the contribution of noisy outlier signal.

4.2. Visualization

An important and ubiquitous dimensionality reduction problem is visualization, which entails learning a two- or three-dimensional representation of each cell that captures some aspect of the dataset structure in a more human-intuitive feature space. Visualizations of scRNA-seq data most often take the form a scatter plot in which each point corresponds to a single cell, which in many instances results in beautiful, pointillistic displays. Linear decomposition algorithms (Section 4.1) can be used to learn visualization embeddings; for example, PCA is commonly used to visualize cells along the two axes corresponding to maximum variability across cells. Nonlinear algorithms can potentially incorporate richer structural information to prevent dense overcrowding within a visualization but may also introduce unrepresentative distortion; deeper interpretation of the data should occur in the more informative, higher-dimensional feature space. A common nonlinear visualization technique is *t*-distributed stochastic neighbor embedding (t-SNE) (92), which learns a low-dimensional embedding in which the distribution of pairwise distances among cells forms a reasonably good information theoretic approximation of the distribution of pairwise distances in the original, high-dimensional space. t-SNE provides tuning parameters that enable

users to vary the amount of density distortion, for which a helpful tutorial has been provided by Wattenberg et al. (93). Additional algorithmic extensions to t-SNE (94–96) have mostly focused on reducing the runtime and efficiency of the original t-SNE algorithm, which scales quadratically with the number of cells. A different set of nonlinear visualization approaches are based on the *k*-nearest neighbor (KNN) graph in which edges connect a cell to its *k*-nearest neighbors. A KNN graph can be visualized according to a force-directed layout (97, 98), inspired by physical simulations of attraction and repulsion, in which unconnected cells are more likely to be separated by greater distances. The KNN graph can also be converted into a visualizable embedding using the uniform manifold approximation and projection (UMAP) algorithm (99, 100). UMAP provides a set of heuristics that, like force-directed embeddings, enable visualization of the KNN graphical topology, but it is controlled by a set of parameters that enable greater density distortion even in the presence of KNN graph edges. Other visualization techniques have recently been deveoped that are designed specifically for single-cell data, in particular for settings in which analysts desire greater preservation of global structures (101–103).

4.3. Nonlinear Representation Learning

Representation learning beyond dimensionality reduction and data visualization is an exciting and actively developing field. As discussed in Section 3.3, it is possible to learn embeddings that allow for data comparison and integration across diverse scRNA-seq studies. Nonlinear models, and particularly deep neural network–based models, have also been used to learn embeddings specifically applied to other scRNA-seq analytic domains, including imputation and cell type assignment (49, 84, 104). Another notable application involves learning an embedding for a cell or a cell type that enables useful latent space arithmetic. Nonlinear representation learning techniques in other fields have demonstrated that it is possible to encode the data into an embedding space in which vector arithmetic approximates complex transformations in the decoded data space. For example, word co-occurrence–based embeddings like word2vec (105) and GloVe (106) can capture simple semantic relationships such that `embedding('king') − embedding('man') + embedding('woman')` is close to `embedding('queen')`, and similar arithmetic in a latent space of human faces learned by a variational autoencoder (VAE) (107, 108) has enabled interpolation between a smiling and a neutral face. In the context of scRNA-seq data, Lotfollahi et al. used a VAE-based model to learn nonlinear embeddings in which vector arithmetic approximates complex cellular changes including drug and infection response (109). As single-cell biological datasets grow larger and more abundant, algorithms that leverage large amounts of training data to learn complex biological models will become useful in an increasing variety of tasks.

5. CLUSTERING

One of the goals of scRNA-seq analysis is to describe the heterogeneity of the cells in a sample. For example, we expect to see both shared and different expression patterns representing cell types and cell states. Clustering is used to separate cells into these groups to allow for downstream comparison between groups.

An optimal clustering separates cells of different states into different groups and places cells of the same type in the same group. **Figure 3** demonstrates several ways that this process can go wrong in scRNA-seq data: undersclustering, overclustering, and cluster splitting. **Figure 3*a*,*b*** highlights the true clustering of a hypothetical scRNAseq dataset. Cells are underclustered if cells of different types are assigned to the same cluster, masking variation in the data (**Figure 3*c***). Cells are overclustered if multiple clusters represent the same cell type (**Figure 3*e***). Cluster splitting

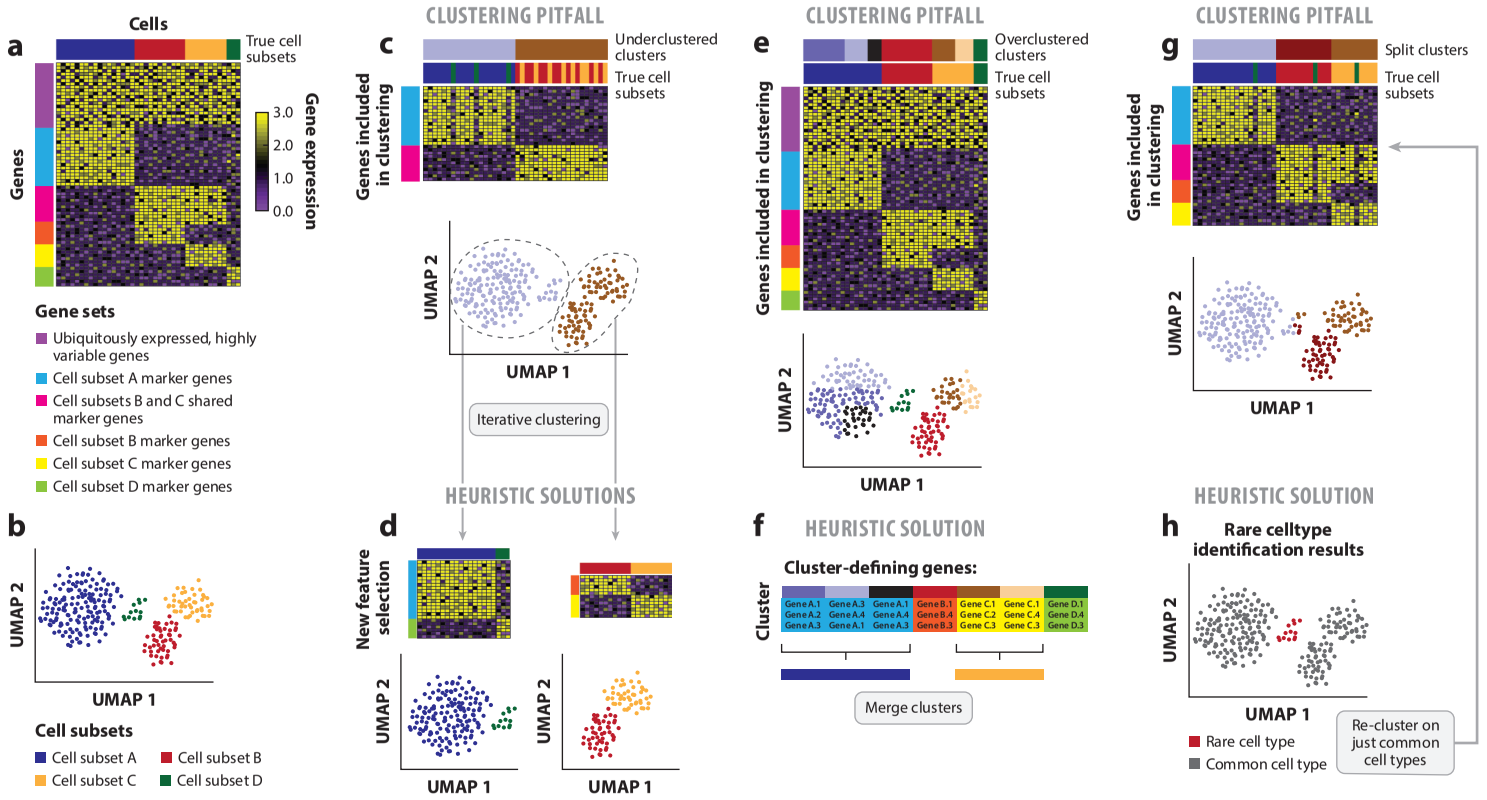

Figure 3

(*a*) Heatmap of gene expression by cell, including marker genes for each cell type and uninformative genes. (*b*) Cells colored by their true cell subset labels in dimensionality-reduced space. (*c*) Underclustering can occur when marker genes that distinguish between similar cell subsets (e.g., B and C) or are unique to one cell type (e.g., D) are not included as features. (*d*) Iterative clustering can split clusters from panel *c* when variable genes are reselected for the cells in these clusters. (*e*) Overclustering can occur when uninformative but variable genes are included, resulting in clusters with no distinguishing marker genes. (*f*) Overclustered cells in panel *d* can be merged based on shared marker genes. (*g*) Rare cell types are split between clusters when their marker genes are excluded from clustering. (*h*) Rare cell type discovery methods identify rare cells before clustering.

occurs when some cells of the same type are scattered between different clusters dominated by other cell types (**Figure 3*g***), a common pitfall in discovering rare cell subsets. Because of the wide range of cell type proportions in a dataset, it is possible for a single clustering to experience a combination of these issues, with some cells overclustered, some underclustered, and some cell types split between several clusters.

scRNA-seq data face several challenges that make out-of-the-box clustering methods from other fields less applicable. Clustering methods use the similarity between cells defined over a set of genes to assign cells to clusters, so gene selection has a large influence on clustering results (110). For example, inclusion of genes whose variation across cells is not informative to the biologically meaningful clustering could lead to overclustering (**Figure 3*e***). Omission of all genes that distinguish cell types leads to underclustering if these cells are similar to another cell type (**Figure 3*c***) or to cluster splitting if the stochastic expression of genes defining other clusters makes them appear more similar to several clusters (**Figure 3*g***).

Clustering methods seek to solve this optimization problem by varying clustering algorithms, feature selection algorithms, and cell–cell similarity metrics. Countless approaches have been proposed that are the result of the combinatorics of the above components (110–112), and method comparisons emphasize that the choice of the optimal clustering methods depends on the data (55, 111, 113, 114).

Clustering techniques can be split into several general categories: hierarchical clustering, network community detection, iterative clustering, model-based clustering, spectral clustering, and consensus clustering (111). While some methods use a combination of these techniques, in this section we review the ways that these types of techniques avoid overclustering, underclustering, and cluster splitting. As new clustering techniques emerge and scRNA-seq datasets become larger, additional comparisons of clustering techniques will be necessary.

5.1. Network Community Detection and Hierarchical Clustering

In network community detection techniques, a graph (i.e., a network) of cells connected by a measure of their pairwise cell–cell similarity is clustered by identifying densely connected regions of nodes (60, 115, 116); these methods are therefore sometimes termed graph-based clustering. Hierarchical clustering comprises either splitting or joining cells iteratively into groups based on similarity metrics (117, 118).

These clustering approaches require the choice of a cell-to-cell similarity metric that, along with feature selection, can have a large impact on clustering results. This is reviewed by Kim et al. (119). While parameters of these clustering methods can usually be tuned to toggle the number of clusters and deal with over- and underclustering, if the genes that define the difference between two cell types are excluded, these cell types will not be split and the results will be underclustered. It is difficult to find an optimal set of genes and parameters that cluster all cells to the optimal resolution, so clusters are sometimes merged together after overclustering (**Figure 3*f***). Some of the popular clustering techniques are network community detection methods, specifically the Louvain and Leiden methods (116), which are available in the Seurat and scanpy toolkits (60, 115).

5.2. Iterative Clustering and Rare Cell Detection

Since hierarchical clustering and community detection are prone to under- or overclustering, it is common to take an iterative divide-and-conquer approach, which begins with a coarse separation of cells and then reclusters each group independently, as reviewed in Reference 111 (**Figure 3*d***). This iterative reclustering is able to deal with cell types of different resolutions by performing feature selection again for each individual cluster before further clustering these reduced sets

(120, 121). Iterative clustering avoids overclustering by halting iterations along one branch when all meaningful clusters are found and avoids underclustering by continuing to repeat feature selection and clustering until cells are clustered sufficiently. Several clustering methods are available to automate this scheme (122–124).

Iterative clustering does not avoid splitting rare cell types into different clusters (**Figure 3*g***). If a cell type is split between clusters at an earlier iteration, even if the genes that discern this cell type are considered at a later stage of clustering, the low numbers of these rare cells in each subcluster make them difficult to identify. Some algorithms contain specific optimizations for the identification of rare cell subsets in the full dataset (**Figure 3*h***) (125–129). These methods can be incorporated into the iterative clustering approach to identify rare cells before clustering larger cell types (**Figure 3*h***).

5.3. Model-Based Methods

Model-based clustering methods avoid dependence on cell–cell similarity metrics by directly modeling gene expression as counts and inferring cluster membership from these models. There is some controversy about the validity of normalizing scRNA-seq count data to continuous values prior to the selection of variable genes. By avoiding this assumption, model based methods claim to discover more accurate clusters.

Techniques such as BISCUIT (Bayesian inference for single-cell clustering and imputing) (130), BAMM-SC (Bayesian mixture model for single-cell sequencing) (131), and Para-DPMM (parallelized split merge sampling on Dirichlet process mixture model) (132) include inference of other features beyond cluster assignments such as marker genes and cell trajectories, compensate for batch correction (131) and technical variation (130), and estimate the uncertainty of cluster assignments, allowing for soft cluster memberships (130–132). Model-based methods have not been included in independent benchmarking analyses, so it remains unclear if they show practical improvements over other techniques.

5.4. Consensus Clustering

Since clustering techniques often give different divisions of clusters, it can be useful to use a consensus of several techniques (133). These approaches avoid over- and underclustering and rare cell splitting by using information from multiple resolutions and by using hyperparameter selections (134) of clusters to define the final clusters. This type of approach could be extended to use the consensus of multiple gene selection choices to further minimize the splitting of rare cell types.

5.5. Reference-Assisted Clustering

While unsupervised clustering has been the main focus of scRNA-seq cell type identification pipelines, recent techniques have proposed taking advantage of biological knowledge to annotate scRNA-seq cell types through supervised classification. These methods label cells based on their similarities to transcriptional profiles of cells in the literature (114). This reduces the burden of manually annotating cell types from lists of marker genes, allows researchers to consider cell types they would not ordinarily consider, and removes the feature selection step. Since reference-assisted clustering labels focuses on identifying those cell types that are present in existing databases, the methods are unlikely to label new cell types. Additionally, since the databases are usually derived from different types of technologies, comparisons of scRNA-seq data to the existing data can be hindered by differences in data types.

5.6. Scalability of Clustering Methods

Clustering methods will need to process a rapidly growing number of cells as the throughput of scRNA-seq methods increases. Calculation of a cell–cell similarity metric between all cells, a common subroutine within many clustering algorithms, is not computationally scalable to large cell numbers. Some approaches improve the asymptotic efficiency of hierarchical or graph-based clustering by running the clustering algorithm only on the KNN graph (60), which has a complexity that scales with the number of cells multiplied by a factor of k, which is typically small (rather than the graph containing all cell–cell similarities, which scales quadratically with the number of cells). Other approaches improve efficiency by clustering on downsampled representative subsets of cells (44, 45, 128) (Section 2.3).

Besides the highly parallelized Para-DPMM (132), model-based methods have difficulty scaling to large numbers of cells. Some recent approaches using neural networks also model gene expression as count data and emphasize their scalability to large numbers of cells (135–137). Petegrosso et al. (111) compared the runtime and clustering performance of several methods on a dataset of 10^5 cells, but comparisons with newer methods are needed.

5.7. Cellular Trajectories

Another approach to interrogate the heterogeneity of a scRNA-seq dataset is to arrange cells along a trajectory. Originally described for cellular differentiation, trajectory inference methods are often termed pseudotime approaches, although these orderings do not necessarily refer literally to time (138). These methods aim to arrange cells along an axis of variation that depends on the choice of both cells and genes. While often applied to cells undergoing differentiation in which a progenitor cell and a more differentiated cell are found in the data, cell orderings can be meaningful across any number of continua, such as spatial gradients and response to external stimuli (139).

Most trajectory inference methods start by computing cell–cell distances. They then order the cells in a low-dimensional manifold computed from these distances under the assumption that adjacent cells in a trajectory will be closer on this manifold. Like clustering, gene selection influences the ability of these methods to discover cell orderings and biases the types of orderings they infer.

An extensive benchmarking analysis compared 45 tools for trajectory inference and developed a package to efficiently run multiple tools and compare their outputs (140). This analysis emphasized that method choice depends on the type of trajectory in the data, and running multiple methods may yield vastly different results. While currently available methods are successful at ordering cells in trajectories that follow one path, the problem of describing more complex trajectory structures is still an area of development.

Another recent objective in scRNA-seq analysis is to derive insights pertaining to the directionality of these trajectories. Because a single sample is only a snapshot of cells, it is difficult to infer with confidence which of many dynamic processes could have led to a cellular trajectory without some additional information (141). Studies with data from multiple, closely spaced time points where the trajectory of interest proceeds along the time course can take advantage of recently developed inference methods that use time points to inform the trajectory, such as TASIC (temporal assignment of single cells) (142) or Waddington-OT (optimal transport) (143). Another innovative method to attribute dynamics to cellular trajectories is RNA velocity (144). This approach uses the ratio of spliced and unspliced RNA transcripts in each cell to compute a vector describing a cell's current state and future direction. Methods like this show great promise for enhanced analysis in studies where multiple samples are not available but dynamic processes are expected.

6. DIFFERENTIAL EXPRESSION AND GENE SET ENRICHMENT

DE is a common procedure in RNA sequencing, and single-cell analysis presents unique challenges due to the nature of the data (69). The goal of DE analysis is to identify robust, true expression differences between two or more groups of interest. The underlying data, the selection of the comparison groups, and the statistical model or test used all influence the results of differential gene expression analyses. Many users initially opt for false discovery rate–adjusted Wilcoxon rank sum tests due to their simplicity, speed, and performance (145, 146). Beyond this nonparametric approach, a plethora of methods have been developed using various modeling approaches to account for confounding technical factors and improve accuracy. Generalized linear models are popular [e.g., Monocle 3 (147), MAST (model-based analysis of single-cell transcriptomics) (148), DESeq2 (66)] owing to both their flexibility in distribution choice and their ability to account for other factors that may influence DE results. Unfortunately, *p*-value inflation is common in DE analysis, particularly between clusters (149). This continues to pose a challenge for accurate detection of differentially expressed genes. Robust DE analysis is critical, considering the increasing complexity of test groups, such as multilevel experimental designs and clinical samples.

Several benchmarks offer some idea about the state of tools for DE analysis (145, 150–153). Early benchmarks from Dal Molin et al. (152) and Jaakkola et al. (151) focused on a mixture of bulk (e.g., DESeq) and single-cell methods [e.g., MAST, SCDE (single-cell DE analysis)] benchmarked against non-UMI datasets. Both made somewhat conflicting statements regarding the suitability of bulk DE methods in single-cell analyses, highlighting the need for proper attention in selecting and applying these tools. In a benchmarking study, 36 tools were tested against each other using a group of prefiltered and unfiltered full-length and UMI-based datasets (145). Several methods performed poorly; however, commonly used methods were sufficient in controlling type-1 error. In fact, the *t*-test and Wilcoxon rank sum test are ranked highly, along with bulk-based limma variants and single-cell-focused MAST variants. Vieth et al. (154) performed a more expansive benchmark, which included both normalization and DE. In contrast to the results in the above-mentioned benchmarking study (145), MAST performed poorly compared to the other methods. Unsurprisingly, normalization was a significant factor in DE analysis. For instance, if appropriate normalization was applied, many methods generated similar discrimination for true positives in both 10x Genomics and Smart-seq2 data (154). Similarly, Hafemeister & Satija demonstrated that a simple *t*-test between two equivalent cell types, where UMI counts are downsampled in one group, identifies many false positive differentially expressed genes (59).

Several groups have proposed new approaches to address shortcomings in DE analysis. Zhang et al. (155) introduced a method to control for the selection bias that arises as clusters are defined that generates artificially low *p*-values and hence false discoveries. They introduced a framework to correct for the introduction of induced separation with the sample groups. Ntranos et al. (156) revisited logistic regression for single-cell data, reasoning that the current scale of sampling enables appropriate fitting. As an example, Stuart et al. (56) utilized this approach to account for the donor source of each cell. Crowell et al. (157) offered a new simulation framework, muscat, for looking at differential states between conditions. They demonstrated definitively that, although extensively used, the bulk methods edgeR (158) and limma (73) are effective at comparing aggregated sets of cells between two conditions. These three examples represent expanding areas of research to control for artificially induced differences, utilize the wealth of information in single-cell data, and perform accurate testing across sample stratifications.

Trajectory analysis has expanded the potential sample groups, covariates, and null hypotheses that can be included in differential testing. Van den Berge et al. (159) introduced TradeSeq, a toolbox of trajectory-based methods that build and expand upon previously published methods. A generalized additive model as a function of the inferred pseudotime can often be used to

perform a range of differential tests, both within and between lineages. In their paper, they intuitively presented the potential comparisons and outcomes with their negative binomial model (or zero-inflated negative binomial). Cao et al. (147) presented an additional method for both general cluster-based and trajectory-based differential analysis termed graph autocorrelation analysis based on Moran's *I* metric.

After producing tables of differentially expressed genes between groups of interest, gene set enrichment facilitates quick interpretation of biological themes. Bulk methods like GSEA (gene set enrichment analysis) can be applied to these gene sets (160). Strict filtering should be employed before enrichment to ensure accurate results, considering the additional sources of false positives. AUCell offers a per-cell enrichment of specific gene sets by calculating the area under the curve across ranked gene expression versus the number of genes within the gene set (161). This produces bimodal distributions of gene set scores across all cells, where a proportion of cells have notable enrichment. Given a particular gene set, Tirosh et al. (74) introduced a scoring method that averages the expression level of all genes within the gene set, correcting for basal expression of random control genes in similar expression bins. This method provides a convenient way to visualize and additionally validate module expression.

Regardless of the tool chosen, *p*-values must be corrected for multiple hypotheses (162). Additionally, properly accounting for covariates requires additional attention, highlighted by Luecken & Theis (149). Before biochemical validation, users should additionally address the count distribution of the gene of interest across various technical factors. Ensuring the tool is appropriately suited for the statistical characteristics of the data is critical to prevent the generation of inappropriate conclusions.

7. NETWORK RECONSTRUCTION

Once cells are organized into clusters and cell types, network reconstruction techniques can be used to form hypotheses about the underlying behaviors of the cells in these clusters beyond their marker genes (163). Bulk RNA-seq studies have been widely successful in identifying gene regulatory networks (GRNs) using measurements of populations of cells under different conditions, but network inference methods developed for bulk RNA-seq data are not adept at inferring the types of networks empowered by scRNA-seq data (164). In single-cell experiments, we can harness the coexpression of genes in single cells instead of full samples to understand the networks underlying these states (165). This structure enables different types of networks to be considered from scRNA-seq data: GRNs, gene–gene coexpression networks, and cell–cell interaction networks.

7.1. Gene Regulatory Networks

GRNs are representations of the factors in a cell that control the transcription of genes into mRNA. In order to describe a causal relationship between one gene product and the transcription of a gene, researchers must show that perturbing the upstream factor alters the behavior of the downstream gene. Since scRNA-seq data represent a snapshot of the heterogeneity of cells in a sample, the lack of stimulation or time variation makes this formal inference of causality impossible from a single experiment (166, 167). An approximation of causality in differentiation networks is possible when temporal data are available, such as from a time series experiment (168, 169) or trajectory inference analysis (170, 171). These approaches assume that a change in expression of one gene at an earlier time point could indicate that gene as the cause of a change in expression of a different gene at a later time. Recently, the BEELINE framework has been developed to benchmark these scRNA-seq GRN inference approaches using a unified model (172), and it has shown promising results for GRN inference from scRNA-seq data.

7.2. Coexpression Networks

While causality or directionality of GRNs is difficult to conclude, scRNA-seq empowers the discovery of gene–gene coexpression networks. These are weighted, undirected networks with genes connected by a measure of the frequency of their coexpression in single cells. scRNA-seq analysis methods harness these networks to discover gene modules (173), predict upstream transcription factors that drive expression (161), provide guilt-by-association indications of gene function, cluster cells based on coexpressed genes (174, 175), and construct interstudy trajectories (47). Innovations in constructing gene–gene coexpression networks, such as optimization of the choice of coexpression metric (176), will improve these applications, but more work is needed to reach a consensus on how to build these networks from scRNA-seq data.

7.3. Cell–Cell Interaction Networks

Computational methods have been developed to summarize the possible receptor–ligand interactions to understand cell–cell communication from scRNA-seq data (163).

There are obstacles that have not been addressed despite recent progress in these analyses. Current techniques rely on curated lists of receptor–ligand pairs, which require effort to create and are subject to biases toward more heavily researched areas. Some repositories have been created for this purpose (177). Still, mRNA expression of receptors and ligands is an imperfect measure of protein expression, as receptor and ligand transcripts are not always detectable in scRNA-seq data and could be influenced by posttranscriptional regulation. Arneson et al. (178) proposed an alternative approach to score cells based on expression of genes involved in ligand production or receptor signaling. Cell–cell interaction networks show promise and have been applied in some studies with great success. The applicability of these methods to general studies remains unclear given the lack of benchmarking or ground truth comparisons available for this analysis.

8. CONCLUSIONS AND OUTLOOK

It is an incredibly exciting time in the field of single-cell omics. scRNA-seq has transformed our understanding of biological systems by providing precision approaches to characterize the transcriptomes of individual cells. At the same time, these methods have presented new challenges and opportunities for computational methods to robustly derive novel biological insights. Ultimately, it is important to remember that analysis of single-cell omics data must be conducted with deep consideration of the biological questions to be answered in the analysis. At times, keeping up with developments and improvements in the field can be incredibly challenging. As additional layers of biological information (RNA velocities, protein expression, chromatin accessibility, etc.) are integrated into single-cell workflows, new computational methods will have to accommodate the subtleties of these data. Complex experimental designs, especially clinical studies, large-scale atlases, and multimodal integration will continue to call for principled implementation of these computational tools. As new tools are developed, we recommend that researchers building computational tools consider the usability and postpublication development of their packages. In some cases, packages are difficult to integrate with common workflows in the field, thus hindering their adoption. Lastly, we believe that benchmarking studies for computational methods with well-designed biological experiments will have tremendous value going forward. Tools like Dynverse (140) epitomize our vision for the future of benchmarking of computational methods: publicly available resources, user-friendly interfaces, and cross-platform compatibility. While it is difficult to predict the next decade of technological advancements in the field of single-cell omics, we envision a future where computational methods advance in parallel with the biochemical tools,

with comprehensive benchmarking, well-documented methods, and new or unexpected biological insights.

SUMMARY POINTS

1. Single-cell RNA sequencing (scRNA-seq) analysis is influenced by numerous experimental and computational decisions. The choice of computational tools is critical.
2. The field has produced and benchmarked a range of tools for common analyses, although more comprehensive benchmarks are needed.
3. Analysis requires iterative, principled approaches that are biologically aware.

FUTURE ISSUES

1. Building robust and interpretable models are needed for analyzing scRNA-seq data beyond cell and gene clustering.
2. An explosion of computational methods creates challenges for standardization of analytical methods and ongoing maintenance of existing methods.

DISCLOSURE STATEMENT

A.K.S. has received compensation from and is a member of the scientific advisory board of Cellarity, Cogen Immune Medicines, and Honeycomb Biotechnologies. The authors are not aware of any other affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to thank members of the Berger, Blainey, Bryson, and Shalek labs for helpful discussions. We also would like to recognize and thank the authors of numerous tools we were unable to include within this review. B.H. was partially supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program.

LITERATURE CITED

1. Tanay A, Regev A. 2017. Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541:331–38
2. Svensson V, Vento-Tormo R, Teichmann SA. 2018. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13:599–604
3. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, et al. 2017. The Human Cell Atlas. *eLife* 6:e27041
4. Svensson V, da Veiga Beltrame E. 2019. A curated database reveals trends in single cell transcriptomics. bioRxiv 742304. **https://doi.org/10.1101/742304**
5. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6:377–82
6. Chen X, Teichmann SA, Meyer KB. 2018. From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annu. Rev. Biomed. Data Sci.* 1:29–51

7. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, et al. 2019. Systematic comparative analysis of single cell RNA-sequencing methods. bioRxiv 632216. **https://doi.org/10.1101/632216**
8. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, MacCarthy DJ, et al. 2019. Benchmarking single-cell RNA sequencing protocols for Cell Atlas Projects. bioRxiv 630087. **https://doi.org/10.1101/630087**
9. Zappia L, Phipson B, Oshlack A. 2018. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Comput. Biol.* 14:e1006245
10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner: supplementary data. *Bioinformatics* 29:15–21
11. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37:907–15
12. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36
13. Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform.* 12:323
14. Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–69
15. Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–30
16. Liao Y, Smyth GK, Shi W. 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* 47:e47
17. 10x Genomics. 2019. *What is Cell Ranger?* Tech. Support Memo., 10x Genomics, Pleasanton, CA. **https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger**
18. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, et al. 2019. RNA sequencing data: Hitchhiker's guide to expression analysis. *Annu. Rev. Biomed. Data Sci.* 2:139–73
19. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. 2018. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* 7(6):giy059
20. Melsted P, Ntranos V, Pachter L. 2019. The barcode, UMI, set format and BUStools. *Bioinformatics* 35(21):4472–73
21. Melsted P, Booeshaghi AS, Gao F, Beltrame E, Lu L, et al. 2019. Modular and efficient pre-processing of single-cell RNA-seq. bioRxiv 673285. **https://doi.org/10.1101/673285**
22. Srivastava A, Malik L, Smith T, Sudbery I, Patro R. 2019. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.* 20:65
23. Petukhov V, Guo J, Baryawno N, Severe N, Scadden DT, et al. 2018. dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.* 19:78
24. Farouni R, Najafabadi HS. 2019. Statistical modeling, estimation, and remediation of sample index hopping in multiplexed droplet-based single-cell RNA-seq data. bioRxiv 617225. **https://doi.org/10.1101/617225**
25. Zhang MJ, Ntranos V, Tse D. 2018. One read per cell per gene is optimal for single-cell RNA-seq. bioRxiv 389296. **https://doi.org/10.1101/389296**
26. Svensson V, Beltrame EdV, Pachter L. 2019. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. bioRxiv 762773. **https://doi.org/10.1101/762773**
27. Baran-Gale J, Chandra T, Kirschner K. 2018. Experimental design for single-cell RNA sequencing. *Brief. Funct. Genom.* 17:233–39
28. Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, et al. 2017. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 7:39921
29. Lafzi A, Moutinho C, Picelli S, Heyn H. 2018. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* 13:2742–57
30. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, et al. 2018. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* 19:224
31. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, et al. 2019. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* 16:619–26

32. Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, et al. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356:eaah4573
33. van den Brink SC, Sage F, Vértesy Ã, Spanjaard B, Peterson-Maduro J, et al. 2017. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* 14:935–36
34. O'Flanagan CH, Campbell KR, Zhang AW, Kabeer F, Lim JLP, et al. 2019. Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* 20:210
35. Lun AT, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. 2019. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20:63
36. McGinnis CS, Murrow LM, Gartner ZJ. 2019. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8:329–37
37. Wolock SL, Lopez R, Klein AM. 2019. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* 8:281–91
38. Xu J, Falconer C, Nguyen Q, Crawford J, McKinnon BD, et al. 2019. Genotype-free demultiplexing of pooled single-cell RNA-Seq. *Genome Biol.* 20:290
39. Heaton H, Talman AM, Knights A, Imaz M, Durbin R, et al. 2019. souporcell: robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes. bioRxiv 699637. **https://doi.org/10.1101/699637**
40. Vieira Braga FA, Kar G, Berg M, Carpaij OA, Polanski K, et al. 2019. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* 25:1153–63
41. Young MD, Behjati S. 2018. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. bioRxiv 303727. **https://doi.org/10.1101/303727**
42. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, et al. 2019. intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 178:714–30
43. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, et al. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17:29
44. Hie B, Cho H, DeMeo B, Bryson B, Berger B. 2019. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Syst.* 8:483–93
45. Iacono G, Mereu E, Guillaumet-Adkins A, Corominas R, Cuscó I, et al. 2018. bigSCale: an analytical framework for big-scale single-cell data. *Genome Res.* 28:878–90
46. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, et al. 2018. MetaCell: analysis of single cell RNA-seq data using *K*-nn graph partitions. *Genome Biol.* 20:206
47. Hie B, Cho H, Bryson B, Berger B. 2019. Coexpression uncovers a unified single-cell transcriptomic landscape. bioRxiv 719088. **https://doi.org/10.1101/719088**
48. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, et al. 2019. PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* 29:59
49. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. 2019. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* 10:390
50. Huang M, Wang J, Torre E, Dueck H, Shaffer S, et al. 2018. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15:539–42
51. Li WV, Li JJ. 2018. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9:997
52. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, et al. 2018. Recovering gene interactions from single-cell data using data diffusion. *Cell* 174:716–29
53. Linderman GC, Zhao J, Kluger Y. 2018. Zero-preserving imputation of sc RNA-seq data using low-rank approximation. bioRxiv 397588. **https://doi.org/10.1101/397588**
54. Zhang L, Zhang S. 2018. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* In press
55. Andrews TS, Hemberg M. 2018. False signals induced by single-cell imputation. *F1000Research* 7:1740
56. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, et al. 2019. Comprehensive integration of single-cell data. *Cell* 177:1888–902

57. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14:565–71
58. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, et al. 2017. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14:584–86
59. Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20:296
60. Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19:15
61. Lun AT, Bach K, Marioni JC. 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75
62. Tian L, Dong X, Freytag S, Lê Cao KA, Su S, et al. 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* 16:479–87
63. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. 2019. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16:43–49
64. Yip SH, Wang P, Kocher JPA, Sham PC, Wang J. 2017. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* 45:e179
65. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. 2017. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14:309–15
66. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550
67. Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, et al. 2019. Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell Syst.* 8:315–28
68. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. 2019. Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. *Genome Biol.* 20:295
69. Lun A. 2018. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. bioRxiv 404962. **https://doi.org/10.1101/404962**
70. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, et al. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10:1093–98
71. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161:1187–201
72. Yip SH, Sham PC, Wang J. 2018. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief. Bioinform.* 20:1583–89
73. Smyth GK. 2005. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, ed. R Gentleman, VJ Carey, W Huber, RA Irizarry, S Dudoit, pp. 397–420. New York: Springer
74. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, et al. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352:189–96
75. Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118–27
76. Haghverdi L, Lun A, Morgan M, Marioni J. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36:421–27
77. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36:411–20
78. Dekel T, Oron S, Rubinstein M, Avidan S, Freeman WT. 2015. Best-buddies similarity for robust template matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2021–29. Los Alamitos, CA: IEEE Comput. Soc.
79. Hie B, Bryson B, Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* 37:685–91
80. Polański K, Park JE, Young MD, Miao Z, Meyer KB, Teichmann SA. 2019. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*. In press
81. Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, et al. 2019. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* 16:695–98

82. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, et al. 2018. Fast, sensitive, and accurate integration of single cell data with Harmony. *Nat. Methods* 16:1289–96
83. Welch J, Kozareva V, Ferrara A, Vanderburg C, Martin C, Macosko E. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177:1873–87
84. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15:1053–58
85. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. 2018. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* 9:884
86. Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20:53–65
87. Stuart T, Satija R. 2019. Integrative single-cell analysis. *Nat. Rev. Genet.* 20:257–72
88. Loh PR, Baym M, Berger B. 2012. Compressive genomics. *Nat. Biotechnol.* 30:627–30
89. Yu YW, Daniels NM, Danko DC, Berger B. 2015. Entropy-scaling search of massive biological data. *Cell Syst.* 1:130–40
90. Cleary B, Cong L, Cheung A, Lander ES, Regev A. 2017. Efficient generation of transcriptomic profiles by random composite measurements. *Cell* 171:1424–36
91. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, et al. 2018. Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet.* 34:790–805
92. van der Maaten LJP, Hinton GE. 2008. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* 9:2579–605
93. Wattenberg M, Viégas F, Johnson I. 2016. How to use t-SNE effectively. *Distill.* **http://doi.org/10.23915/distill.00002**
94. Cho H, Berger B, Peng J. 2018. Generalizable and scalable visualization of single-cell data using neural networks. *Cell Syst.* 7:185–91
95. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. 2019. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* 16:243–45
96. van der Maaten L. 2014. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* 15:3221–45
97. Jacomy M, Venturini T, Heymann S, Bastian M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE* 9(6):e98679
98. Weinreb C, Wolock S, Klein AM. 2018. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* 34:1246–48
99. McInnes L, Healy J. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML]
100. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IW, et al. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37:38–44
101. Chen Z, An S, Bai X, Gong F, Ma L, Wan L. 2019. DensityPath: an algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell RNA sequencing data. *Bioinformatics* 35:2593–601
102. An S, Ma L, Wan L. 2019. TSEE: an elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell RNA sequencing data. *BMC Genom.* 20:224
103. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, et al. 2019. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* 37:1482–92
104. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. 2019. Harmonization and annotation of single-cell transcriptomics data with deep generative models. bioRxiv 532895. **https://doi.org/10.1101/532895**
105. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. **https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality**
106. Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–43. Stroudsburg, PA: Assoc. Comput. Linguist.

107. Kingma DP, Welling M. 2014. *Auto-encoding variational Bayes*. Paper presented at International Conference on Learning Representations (ICLR 2014), Banff, Can., Apr. 14–16
108. Rezende DJ, Mohamed S, Wierstra D. 2014. Stochastic backpropagation and approximate inference in deep generative models. *Proc. Mach. Learn. Res.* 32(2):1278–86
109. Lotfollahi M, Wolf FA, Theis FJ. 2019. scGen predicts single-cell perturbation responses. *Nat. Methods* 16:715–21
110. Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20:273–82
111. Petegrosso R, Li Z, Kuang R. 2019. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinform.* 2019:bbz063
112. Zeng T, Dai H. 2019. Single-cell RNA sequencing-based computational analysis to describe disease heterogeneity. *Front. Genet.* 10:629
113. Duò A, Robinson MD, Soneson C. 2018. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* 7:1141
114. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, et al. 2019. A comparison of automatic cell identification methods for single-cell RNA-sequencing data. *Genome Biol.* 20:294
115. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, et al. 2018. Comprehensive integration of single cell data. *Cell* 177(7):1888–902.e21
116. Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 9:5233
117. žurauskienė J, Yau C. 2016. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* 17:140
118. Lin P, Troup M, Ho JWK. 2017. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59
119. Kim T, Chen IR, Lin Y, Wang AYY, Yang JYH, Yang P. 2018. Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.* 20(6):2316–26
120. Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, et al. 2017. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* 20:484–96
121. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, et al. 2016. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19:335–46
122. Ding H, Wang W, Califano A. 2018. iterClust: a statistical framework for iterative clustering analysis. *Bioinformatics* 34:2865–66
123. Hu MW, Kim DW, Liu S, Zack DJ, Blackshaw S, Qian J. 2019. PanoView: an iterative clustering for single-cell RNA sequencing data. *PLOS Comput. Biol.* 15(8):e1007040
124. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, et al. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347:1138–42
125. Jindal A, Gupta P, Jayadeva, Sengupta D. 2018. Discovery of rare cells from voluminous single cell expression data. *Nat. Commun.* 9:4719
126. Tsoucas D, Yuan GC. 2018. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol.* 19:58
127. Wegmann R, Neri M, Schuierer S, Bilican B, Hartkopf H, et al. 2019. CellSIUS provides sensitive and specific detection of rare cell populations from complex single cell RNA-seq data. *Genome Biol.* 20:142
128. Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. 2018. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res.* 46:e36
129. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, et al. 2016. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19:266–77
130. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, et al. 2018. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* 174:1293–308
131. Sun Z, Chen L, Xin H, Jiang Y, Huang Q, et al. 2019. A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nat. Commun.* 10:1649
132. Duan T, Pinto JP, Xie X. 2019. Parallel clustering of single cell transcriptomic data with split-merge sampling on Dirichlet process mixtures. *Bioinformatics* 35:953–61

133. Yang Y, Huh R, Culpepper HW, Lin Y, Love MI, Li Y. 2019. SAFE-clustering: single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data. *Bioinformatics* 35:1269–77
134. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, et al. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14:483–86
135. Tian T, Wan J, Song Q, Wei Z. 2019. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* 1:191–98
136. Rashid S, Shah S, Bar-Joseph Z, Pandya R. 2019. Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics* 2019:btz095
137. Srinivasan S, Johnson NT, Korkin D. 2019. A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. bioRxiv 511626. **https://doi.org/10.1101/511626**
138. Tritschler S, Büttner M, Fischer DS, Lange M, Bergen V, et al. 2019. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* 146:dev170506
139. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, et al. 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* 4438:eaax4438
140. Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37(5):547–54
141. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. 2018. Fundamental limits on dynamic inference from single-cell snapshots. *PNAS* 115:E2467–76
142. Rashid S, Kotton DN, Bar-Joseph Z. 2017. TASIC: determining branching models from time series single cell data. *Bioinformatics* 33:2504–12
143. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, et al. 2019. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 176:928–43
144. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, et al. 2018. RNA velocity of single cells. *Nature* 560:494–98
145. Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15:255–61
146. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, et al. 2019. Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* 17:137–45
147. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, et al. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566:496–502
148. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278
149. Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15:e8746
150. Wang T, Li B, Nelson CE, Nabavi S. 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* 20:40
151. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. 2017. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* 18:735–43
152. Dal Molin A, Baruzzo G, Di Camillo B. 2017. Single-cell RNA-sequencing: assessment of differential expression analysis methods. *Front. Genet.* 8:62
153. Miao Z, Zhang X. 2016. Differential expression analyses for single-cell RNA-seq: old questions on new data. *Quant. Biol.* 4:243–60
154. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. 2019. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* 10:4667
155. Zhang JM, Kamath GM, Tse DN. 2019. Valid postclustering differential analysis for single-cell RNA-seq. *Cell* 9(4):383–92.e6
156. Ntranos V, Yi L, Melsted P, Pachter L. 2019. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* 16:163–66
157. Crowell HL, Soneson C, Germain PL, Calini D, Collin L, et al. 2019. On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. bioRxiv 713412. **https://doi.org/10.1101/713412**

158. Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–40
159. Van den Berge K, de Bézieux HR, Street K, Saelens W, Cannoodt R, et al. 2019. Trajectory-based differential expression analysis for single-cell sequencing data. bioRxiv 623397. **https://doi.org/10.1101/623397**
160. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102:15545–50
161. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14:1083–86
162. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300
163. Blencowe M, Arneson D, Ding J, Chen YW, Saleem Z, Yang X. 2019. Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerg. Top. Life Sci.* 3(4):379–98
164. Chen S, Mar JC. 2018. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinform.* 19:232
165. Todorov H, Cannoodt R, Saelens W, Saeys Y. 2019. Network inference from single-cell transcriptomic data. In *Gene Regulatory Networks: Methods and Protocols*, ed. G Sanguinetti, VA Huynh-Thu, pp. 235–49. New York: Humana Press
166. Iacono G, Massoni-Badosa R, Heyn H. 2019. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol.* 20:110
167. Fiers MWEJ, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z, Aerts S. 2018. Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genom.* 17:246–54
168. Papili Gao N, Ud-Dean SMM, Gandrillon O, Gunawan R. 2018. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 34:258–66
169. Sanchez-Castillo M, Blanco D, Tienda-Luna IM, Carrion MC, Huang Y. 2018. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* 34:964–70
170. Ocone A, Haghverdi L, Mueller NS, Theis FJ. 2015. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* 31:i89–96
171. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, et al. 2017. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* 33(15):2314–21
172. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. 2019. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17:147–54
173. Liu H, Li P, Zhu M, Wang X, Lu J, Yu T. 2016. Nonlinear network reconstruction from gene expression data using marginal dependencies measured by DCOL. *PLOS ONE* 11:e0158247
174. Peng H, Zeng X, Zhou Y, Zhang D, Nussinov R, Cheng F. 2019. A component overlapping attribute clustering (COAC) algorithm for single-cell RNA sequencing data analysis and potential pathobiological implications. *PLOS Comput. Biol.* 15:e1006772
175. Mohammadi S, Ravindra V, Gleich DF, Grama A. 2018. A geometric approach to characterize the functional identity of single cells. *Nat. Commun.* 9:1516
176. Skinnider MA, Squair JW, Foster LJ. 2019. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* 16:381–86
177. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. 2019. CellPhoneDB v2.0: inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. bioRxiv 680926. **https://doi.org/10.1101/680926**
178. Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, et al. 2019. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.* 10:963