JOURNAL OF Bio-X Research

REVIEW ARTICLE
http://dx.doi.org/10.1097/JBR.0000000000000053

OPEN

Dropout imputation and batch effect correction for single-cell RNA sequencing data

Gang Li[a], Yuchen Yang[b], Eric Van Buren[c], Yun Li[b,c,d,*]

Abstract

Single-cell RNA sequencing (scRNA-seq) allows researchers to examine the transcriptome at the single-cell level and has been increasingly employed as technologies continue to advance. Due to technical and biological reasons unique to scRNA-seq data, denoising and batch effect correction are almost indispensable to ensure valid and powerful data analysis. However, various aspects of scRNA-seq data pose grand challenges for such essential tasks pertaining to data pre-processing, normalization or harmonization. In this review, we first discuss properties of scRNA-seq data that contribute to the challenges for denoising and batch effect correction from a computational perspective. We then focus on reviewing several state-of-the-art methods for dropout imputation and batch effect correction, comparing their strengths and weaknesses. Finally, we benchmarked three widely used correction tools using two hematopoietic scRNA-seq datasets to show their performance in a real data application.

Keywords: batch effect, deep learning, dropout, imputation, single-cell RNA sequencing

Introduction

In living organisms, cells are the fundamental composition and functional units.[1] Identification and characterization of cell types and their biological functions from a mass of heterogeneous cells are of great interest and importance in understanding the molecular mechanisms underlying growth, development and disease.[2–4] In recent years, RNA sequencing (RNA-seq) has been widely used to study the transcriptome as well as to help construct gene expression networks underlying the complex processes of cellular proliferation, differentiation, and reprograming.[5–7] However, for most genes, expression levels are found to vary dramatically both across different cell types and even across single cells of the same type. Possible reasons underlying such variation in gene expression profiles include different cellular functions, developmental stages, cell cycle phase, and adjacent microenvironments,[8–10] among others. When RNA-seq data are generated from bulk tissue or many cells in aggregate (commonly referred to as bulk RNA-seq in the literature), they measure the average expression across many cells of potentially different types and/or across states, and thus may mask biologically varying functional capacities across cell types or across single cells.[11] In contrast to bulk RNA-seq, single-cell RNA-seq (scRNA-seq) provides researchers refined resolution to investigate cellular heterogeneity in gene expression profiles, as well as to discover novel cell types and to infer cell fates, presenting enormous potential from basic science studies of cell biology, all the way to facilitating transformative clinical applications.[7,12–17]

scRNA-seq based studies are becoming increasingly common as technological improvements allow sequencing of an increasing number of cells measured with ever-improving accuracy. As scRNA-seq technologies continue to advance and mature, together with decreasing sequencing costs, analyses integrating multiple scRNA-seq datasets have become common practice. Such analyses enable joint investigation of gene expression profiles across multiple scRNA-seq datasets collected across different conditions or time points, and/or from different laboratories or experimental assays. Integrating data in an unbiased and valid manner enables researchers to better understand transcriptional dynamics during a certain biological process (eg, across different development stages) at the single-cell level by leveraging as much data as possible. Such integration approaches also encourage and enable re-using published datasets, maximizing the value of scRNA-seq datasets already generated and substantially reducing the costs associated with generation of new data. However, integration cannot be carried out by simply pooling multiple datasets naively together. To ensure drawing valid scientific conclusions, before integrating multiple datasets, it is critical to first take careful steps to denoise naturally sparse scRNA-seq data and to properly adjust for batch effects across datasets. In this review, we discuss computational strategies commonly employed for denoising and batch effect correction of scRNA-seq data. Specifically, the review is organized as follows. We first describe the commonly used scRNA-seq technologies, focusing on aspects that later lead to challenges in computational analysis. We then highlight several state-of-art methods for dropout imputation and batch effect correction, with an emphasis on their advantages and drawbacks in practical applications. Finally, we benchmark some of these methods using two real scRNA-seq datasets studying hematopoietic cells.

[a] Department of Statistics and Operations Research, [b] Department of Genetics, [c] Department of Biostatistics, [d] Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[*] Corresponding author: Yun Li, Department of Genetics, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, Campus Box 7264, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: yunli@med.unc.edu.

Journal of Bio-X Research (2019) 2:169–177

Received: 11 October 2019; Accepted: 2 November 2019

Published online 17 December 2019

http://dx.doi.org/10.1097/JBR.0000000000000053

Database search strategy

The articles used in this review were largely based on the authors' knowledge. But they could be retrieved by searching the terms "single-cell RNA sequencing", "batch effect correction" and "dropout imputation" via Google Scholar and PubMed. The results from such search can be further screened by title and abstract to focus on methodological work for batch effect correction and dropout imputation for single-cell RNA sequencing data. In addition, we recommend conducting electronic searches for papers that cited batch effect correction methods, such as "Seurat", "MNN", and "LIGER", and dropout imputation methods, such as "MAGIC", "scImpute", "VIPER", "DCA" and "SAVER". Results from these additional searches are further filtered again to focus on methodological work or review articles.

Experimental technologies for single-cell RNA sequencing

Since the first scRNA-seq experiment was published in 2009,[18] we have been witnessing rapid development of scRNA-seq technologies. Technological advancements manifest in many different ways, particularly in the number of cells that can be profiled simultaneously and dramatically reduced costs per single cell.[19] Different scRNA-seq technologies are distinct in various aspects. We highlight below two key aspects: strategy for single-cell isolation, and transcript coverage.[20] Regarding the first aspect, there are two popular strategies for single cell partitioning: plate-based platforms and microdroplet-based microfluidics. Plate-based methods, as adopted by technologies such as SMART-seq,[21] MARS-Seq,[22] and Fluidigm C1,[23] first lyse single cells, then isolate and place them into individual wells on a single plate by flow-activated cell sorting. Microdroplet-based microfluidic platforms, such as those employed by Drop-seq,[24] inDrop,[25] and Gencode (10X Genomics),[26] load cells and gel beads into the channels of a microfluidic chip to generate gel beads each containing transcripts from one single cell (identified by a cell barcode unique within each gel bead, or droplet). Plate-based platforms suffer from rather limited number of single cells that can be profiled at the same time (maximally 800 single cells per run for Fluidigm C1 system, for instance). Comparatively, microfluidic technologies enable simultaneously capturing tens of thousands of cells with a reduced reagent cost. These larger sample sizes provide researchers a more powerful and comprehensive way to investigate the transcriptional dynamics underlying various biological processes—one example is the identification of rare cell types from a mass of highly heterogeneous cells during cell type differentiation or reprogramming.[19,27] Because of these advantages, microfluidic platforms technologies have become more widely adopted commercial scRNA-seq platforms.[19,28] Furthermore, scRNA-seq is an area of fast growth, and a lot of new sequencing technologies have been recently developed. For instance, a new method called SPLiT-seq uses the combinatorial indexing solution.[29] Compared to previous methods, SPLit-seq has 2 advantages:

1) it uses the single cells themselves as partitioning compartments instead of separating cells into custom microwells or microfluidics; and
2) SPLiT-seq enables simultaneous sequencing of tens of thousands of cells from multiple biological samples in one single experiment. It can largely reduce the batch effects introduced during the processes of library preparation and sequencing.[29]

The other key technological aspect is the captured regions of transcripts enabled by library preparation.[19] Some of technologies, such as Smart-seq,[21] Quartz-Seq[30] and MATQ-seq,[31] offer full-length transcript coverage. Other methods (relatively few), for example, STRT-seq, only capture the 5'-end of transcripts.[32] The more commonly used platforms, including MARS-seq,[22] Drop-seq,[24] 10X Genomics[26] and SPLiT-seq,[29] only capture the 3'-end of transcripts. The methods producing full-length transcripts typically also offer a higher capture rate of transcripts and thus enable researchers to investigate alternative-splicing events and allele-specific expression at single-cell level. However, the arguable disadvantage of full-length transcript libraries is that they require higher sequencing depth for accurate quantification of transcripts in full length. Comparatively, technologies only capturing 3'-end require relatively shallow sequencing depth, and therefore have a substantially reduce sequencing cost.

Here we only highlight a few widely-used scRNA-seq experimental technologies, for a more comprehensive review, please refer to Hwang et al.[19]

Computational challenges for single-cell RNA sequencing data

Compared to bulk RNA-seq, scRNA-seq can simultaneously investigate transcriptional profiles of tens of thousands of single cells, which enables researches to investigate cellular heterogeneity in gene expression as well as transcriptional kinetics, to identify new cell types/states and to recover cell trajectory during proliferation and differentiation processes. However, multiple characteristics specific to scRNA-seq data present large challenges to computational analyses.

First, scRNA-seq data often exhibit high level of heterogeneity in the composition of cell populations, owing to both natural variability in cell type proportions across individuals and systematic biases created during single-cell capture, library preparation and sequencing. Such heterogeneity makes many tasks difficult. For example, the detection and classification of rare cell types can be challenging because of their low abundance. Although there are several tools, such as GiniClust[33] and RaceID,[13] designed for identifying rare cell types from heterogeneous scRNA-seq datasets, these methods suffer from impaired performance in common cell type clustering.[28] Furthermore, heterogeneity poses additional challenges when integrating multiple scRNA-seq datasets. Such challenges cannot be readily addressed by standard methods used by bulk RNA-seq. One reason is that these bulk RNA-seq methods assume a uniform cell population compositions across different samples, which may mask the underlying biological structure and be unrealistic.[34]

Second, the high dimensionality of scRNA-seq data poses additional challenges for data analysis. Even with increasingly larger numbers of single cells evaluated in a single experiment, the number of genes/features can often substantially exceed the number of single cells. Under such a scenario, directly applying classic statistical methods would result in problematic parameter estimation. In the context of batch effect correction, for instance, high dimensionality may lead to inaccurate estimates of the distances between pairs of single cells, thus resulting in sub-optimally, if not wrongly, corrected expression profiles across batches. These sub-optimally adjusted expression matrices could subsequently lead to many problems in downstream analysis, including cell type misclassification and false positives and/or false negatives in the identification of differentially expressed

genes. Moreover, the high dimensionality of scRNA-seq data can easily render computational costs prohibitive, especially in datasets involving a large number of single cells or from a complicated study design. For these reasons, dimension reduction is a standard and important step in the pre-processing of scRNA-seq data. Various methods have been utilized to reduce the dimensionality of scRNA-seq data. For example, t-distributed stochastic neighbor embedding is one commonly used dimension reduction method in scRNA-seq analysis.[28,35,36] There are many alternative dimension reduction methods, such as canonical component analysis employed by Seurat,[37] latent factor analysis used ZIFA,[38] integrative non-negative matrix factorization (iNMF)[39] adopted by Link Inference of Genomic Experimental Relationships (LIGER),[40] and variational autoencoder exploited by single-cell variational inference.[41] Assuming that the true biological signals in the original data can be represented in some lower-dimensional manifold, these dimension reduction techniques can enhance both statistical and computational efficiency when analyzing scRNA-seq data. Unfortunately, choosing among the many options for dimension reduction and deciding on the appropriate number of dimensions to represent an arbitrary high-dimensional dataset remains more art than science.

Lastly, scRNA-seq data frequently contain excessive zero counts. In some datasets, particularly those involving a large number of single cells and without aggressive quality filtering (eg, on minimum number of single cells where a gene is expressed), more than 80% of counts are observed as zero.[28] It is difficult to separate two sources of zeros: "true," meaning that there is truly no corresponding transcript expressed in the corresponding single cell; or "false," meaning that true non-zero number of transcripts is measured as zero count in data due to low capture rate, insufficient sequencing depth, or other technological factors such that the observed zero does not reflect the underlying true expression level.[28,42] Despite improvements in scRNA-seq technologies, these false zeros, which are frequently referred to as dropout events, still exist and can introduce substantial noise and bias to many downstream analyses, including cell type classification, cell trajectory construction, differential expression analysis, and integration across multiple experiments. Simply ignoring dropout events underestimates gene expression level in a cell and can lead to misleading results. Therefore, many methods have been developed for dropout imputation in scRNA-seq data (see more details in the following section).

Computational tools for dropout imputation in single-cell RNA sequencing data

As mentioned, dropout events can bias downstream analyses if ignored. Therefore, imputing dropout events and denoising scRNA-seq data are crucial preprocessing steps to ensure valid and powerful analysis, as well as to make data more comparable across different batches. In recent years, a diverse collection of scRNA-seq denoising algorithms have been proposed. Here we review several state-of-the-art dropout imputation methods (Table 1).

Markov Affinity-based Graph Imputation of Cells (MAGIC) is an algorithm to perform denoising and dropout imputation in scRNA datasets using diffusion geometry.[43] The lower-dimensional manifold assumed to be underlying scRNA datasets is learned though a diffusion operator coupled with signal processing principles. Subsequently, a transformed cell-to-cell similarity matrix is built on the low dimensional space. The rationale behind MAGIC is that mapping cellular phenotypes (reflected by single-cell transcriptome profiles) onto the low dimensional manifold can effectively recover the true expression levels hidden under dropout events. By borrowing information across similar cells, selected via data diffusion, MAGIC fills in the missing transcripts or dropout events. MAGIC demonstrates that recovers gene-gene relationship by filling. MAGIC is one of the first paper to impute dropout event. One major caveat is that MAGIC projects data to low dimensional space, which makes it inevitably lose some natural biological variability across cells[44] and abolishes a key feature of single-cell sequencing data.[45]

scImpute is a statistical method which robustly imputes scRNA-seq data.[46] scImpute aims to impute only the zero counts that are most likely to be dropout events. This can limit biases introduced to the true biological zero counts and thus improve power and validity in downstream analyses, including clustering and differential gene analyses. Both MAGIC and scImpute pool data across similar cells for each gene, which might lead to over-smoothing and wipe out genuine biological variability in gene expression. Also, neither MAGIC nor scImpute provides a measure of uncertainty quantification for the estimated values.

Single-cell Analysis Via Expression Recovery (SAVER) is an expression recovery method for unique molecule index (UMI)-based scRNA-seq data.[44] SAVER assumes that the count in each gene of each cell follows a Poisson-gamma mixture (negative binomial) model and uses the posterior mean as the recovered or imputed gene expression value. Once the parameters are estimated through an empirical Bayes-like approach with Poisson Lasso regression, SAVER can recover the true gene expression levels and subsequently better reflect biological gene-to-gene correlations and distribution-level features. However, because SAVER relies on Markov Chain Monte Carlo algorithms to infer parameters, it is computationally intensive and might not be suitable for large datasets.

Table 1

Tools for dropout imputation in single-cell scRNA sequencing data.

Name (reference)	Year	Method type	Strengths	Limitations
MAGIC[43]	2018	Diffusion geometry	One of the first methods	Could over-correct, leading to loss of natural biological variability, no measure of uncertainty quantification
scImpute[46]	2018	Zero-inflated mixture	Robust, focusing on imputing likely dropout events rather than all observed zeros	Might over-smooth, no measure of uncertainty quantification
SAVER[44]	2018	Poisson-gamma mixture	Has uncertainty quantification	Computationally intensive
VIPER[45]	2018	Non-negative regression	Free of tuning, computationally efficient	No measure of uncertainty quantification
DCA[47]	2019	Deep count autoencoder	Directly models counts in a deep autoencoder framework	Requiring tuning, thus computationally intensive

DCA = deep count autoencoder, MAGIC = Markov affinity-based graph imputation of cells, SAVER = single-cell analysis via expression recovery, VIPER = variability-preserving imputation for expression recovery.

Variability-preserving ImPutation for Expression Recovery (VIPER) focuses squarely on imputing zero values via a non-negative regression model.[45] In VIPER's framework, for each cell of interest, a most predicative sparse set of local neighborhood cells is identified to impute zero counts in that cell under study. VIPER is free of tuning parameters and computationally efficient as compared to Bayesian methods with many hyper-parameters that need to be tuned. More importantly, VIPER claims that it can better preserve gene expression variability after imputation, demonstrated in real data applications (Fig. 5 in the original paper).[45] However, VIPER has similar drawbacks to MAGIC and scImpute: it does not provide any uncertainty quantification either.

Deep Count Autoencoder (DCA) adopts a neural network framework for denoising scRNA datasets.[47] The main idea behind DCA is to infer parameters associated with a negative binomial or zero-inflated negative Binomial distribution assumed for the input count data, using a deep neural network framework. DCA employs such negative binomial distributions both with and without zero-inflation to model expression count data, allowing overdispersion and dropout events, as well as taking into account the sparsity of scRNA datasets. Many techniques, including standard regularization methods and neural network "dropout" (distinct from dropout in scRNA-seq data, "dropout" here is a neural network terminology, and refers to the special regularization technique where certain units [both hidden and visible] are dropped out in a neural network), have been employed to avoid overfitting. By utilizing the deep count autoencoder neural network architecture and gradient-based optimization algorithms, DCA is computationally efficient. Specifically, computational costs scale only linearly with the number of cells. However, tuning parameters in DCA can be computationally intensive.

In summary, many powerful imputation tools have been developed and tailored specifically for scRNA datasets. These methods have demonstrated their utility in both simulated and real datasets. By properly imputing dropout counts in scRNA-seq, we can recover the true underlying biological structure and therefore enhance validity and/or statistical power in downstream analyses. However, each imputation method has its own strengths and weaknesses. Suitable imputation methods must be carefully selected and tuned based on nature of the data under study to maximally reduce false positive findings, to improve statistical power to detect true positive signals, and to enhance reproducibility. For instance, Andrews and Hemberg[48] reported low reproducibility of cell-type-specific markers identified after dropout imputation via several different methods, suggesting potentially false signals introduced by imputation.

Computational tools for batch effects correction of single-cell RNA sequencing data

Large-scale scRNA-seq studies including tens of thousands to even millions of cells (eg, the Human Cell Atlas),[4,49] have become increasingly feasible with rapid improvement in single-cell capture and library preparation technologies as well as decreasing sequencing costs. Large datasets typically involve interrogation of cells across multiple batches, from different laboratories, across varying time points, and/or via different experimental protocols and techniques. Batch effects, also understood as systematic differences between cells from different batches, present large challenges to integrative analyses across multiple experiments. When not properly corrected, batch effects may lead to false positives as well as false negatives for many analyses including identification of novel cell type(s) and detection of differentially expressed genes (Fig. 1).[28,42]

Traditional batch effect correction methods, for example limma[50] and ComBat,[51] are mainly based on linear regression, where batch effects are modeled either as known variables and regressed out from the raw joint data matrix. These methods have proven to be valuable in correcting batch effects for bulk RNA-seq data.[50–52] However, these methods, proposed largely for bulk data, are not designed to address issues unique to scRNA-seq data, including the aforementioned over-dispersion, excessive zero counts, and more pronounced heterogeneity across single-cell samples. Applying these correction methods designed for bulk RNA-seq to scRNA-seq datasets can result in improperly corrected data which can subsequently result in misleading findings and/or failure to reveal true underlying variation. Recently, many batch effect correction methods have been developed for scRNA-seq data to address the aforementioned challenges.[53,54] Instead of providing a complete list of all the correction methods proposed in recent literature, here we select three state-of-the-art methods to review.

One integrative algorithm for jointly analyzing multiple scRNA-seq datasets is LIGER.[40] The main goal of LIGER is to infer cell types across datasets by simultaneously characterizing shared and specific features between different modalities or batches. iNMF is applied to reduce the high dimensionality of each cell into one shared and one batch-specific set of factors that represent the common and unique biological features, respectively. In the resulting factor space, LIGER constructs a shard factor neighborhood graph where cells are connected to the nearest neighborhoods having similar patterns based on the maximum factor loadings, and then performs a joint clustering using the shared factor neighborhood graph constructed. Compared to alternative methods that only focus on similarities among datasets,[34,51,55,56] LIGER also takes the batch-specific features into consideration, which can maximally recover the latent differentiations among different batches.[40] For example, benchmarking results of two datasets of human peripheral blood mononuclear cells showed that the integrative analysis by LIGER can well preserve the underlying cell-type structures after correction.[40] However, LIGER is designed for inferring cell types simultaneously from multiple scRNA-seq datasets. It does not provide batch-corrected expression profiles for the single cells. Because of that, it is incredibly challenging, if not impossible, to use other single-cell analysis packages in a valid manner for other downstream analyses such as cell trajectory analysis,[6,57–61] or analysis of differentially expressed genes.[62–65]

Realizing the importance of generating batch-corrected gene expression profiles for scRNA-seq data, a few groups have developed strategies to correct gene expression values across different batches. For example, MNN adopts the information of mutual nearest neighbors for correcting technical biases between expression profiles of different batches.[34] MNN first globally scales the data into a cosine space, then identifies nearest neighbors across batches, and finally matches pairs of mutual nearest neighbors, which are assumed to be cells from the same cell type or state. MNN calculates a correction vector specific for each pair of mutual nearest neighbors to represent the technical variation between batches, and then derives a correction vector for each cell by calculating a weighted average of all correction vectors with its mutual nearest neighbors. Finally, the gene expression profiles of all the cells in all the batches are corrected according to these cell-specific batch-effect correction vectors.

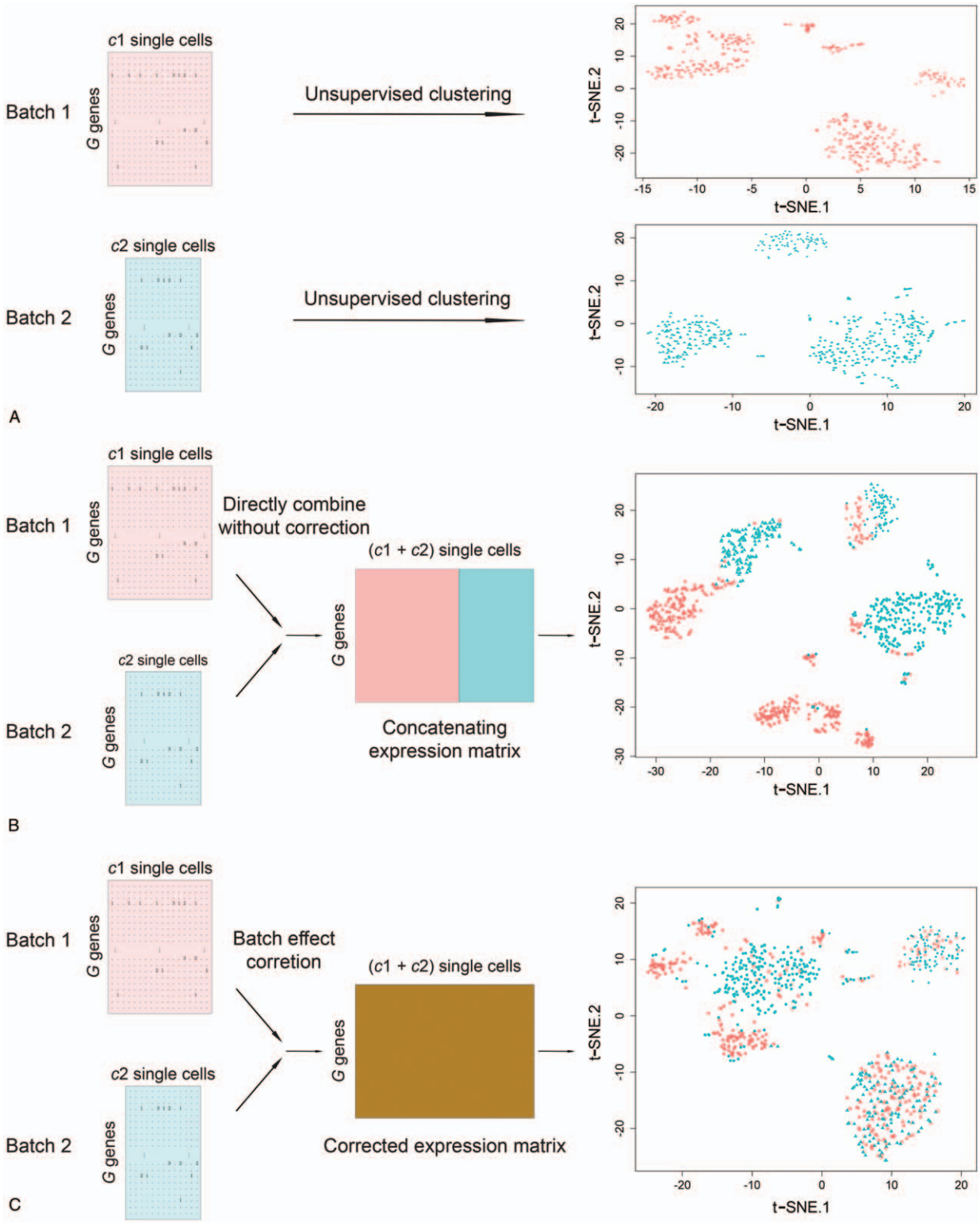

Figure 1. Framework of batch effect correction. (A) Separate analyses of multiple batches. (B) Joint analysis of simply concatenated matrix. (C) Joint analysis of batch corrected expression matrix. t-SNE=t-distributed stochastic neighbor embedding.

Compared to the alternative methods for bulk RNA-seq which assume identical cell composition in different batches, MNN only requires at least one cell population is shared across batches, and MNN has demonstrated superior performance over standard methods developed for bulk RNA-seq data, such as limma and Combat.[34,54] However, MNN makes a rather strong assumption that true biological differences are orthogonal to those due to batch effect. At the minimum, this assumption requires that the variation from batch effect is much smaller than that from biological effect. However, the orthogonality assumption might

Table 2

Major characteristics of the benchmarking mouse hematopoietic datasets.

Batch ID	Technical platform	CMP	MEP	GMP	# of cells evaluated*	# of cells in total	Reference
Batch 1	SMART-seq2	328	362	123	813	1920	Paul et al[69]
Batch 2	MARS-seq	481	1095	1154	2730	2730	Nestorowa et al[68]

* Number of cells used for the benchmarking analysis in this paper. CMP = common myeloid progenitor, GMP = granulocyte-monocyte progenitor, MEP = megakaryocyte-erythrocyte progenitor.

not hold in real data, particularly given that different batches may differ in many aspects, including samples used, single-cell capture method, or library preparation approach. Under non-orthogonal scenarios, MNN will not be optimal using its global (ignoring cell type information) nearest neighbor search strategy, leading to undesired correction results.[66]

Another popular batch effect correction method is Seurat.[54,55,64] In the latest version (v3), Seurat aligns cells from the same cell population or state across different modalities or batches.[37] This approach is conceptually similar to that adopted by MNN. In contrast to MNN, however, Seurat v3 first performs cross-batch dimension reduction using diagonalized canonical correlation analysis, and then detects "anchors" between datasets in the low dimensional space. Here, an anchor refers a pair of cells from the same biological state from different batches (hereafter we use anchor and anchor pair interchangeably). To ensure the accuracy of anchor detection, Seurat v3 assigns each anchor pair a score according to the level of shared mutual neighbors between the two corresponding cells in the anchor pair. In addition to the anchor score calculated above, Seurat v3 also computes a cell similarity score (as the name indicates, this score is specific to each cell) by quantifying the distance between the cell under study and its nearest "neighbors" ("neighbors" here refer to cells that form anchor pairs with the cell under investigation). Correction vectors are then computed for each cell by averaging across the anchor-level correction vectors with its nearest neighbors weighted by anchor score and cell similarity score. The corrected expression matrix is then inferred by subtracting the weighted correction vectors from the original expression profiles. Because of canonical correlation analysis's advantages in identifying shared biological markers and conserved gene correlation patterns,[55,67] this type of batch effect adjustment is robust across a full spectrum of scales of batch effects with respect to the true biological effect, particularly under scenarios with extensive technical variation.[37]

In summary, these batch effect methods allow researchers to carry out integrative analyses of multiple scRNA-seq datasets, which can boost statistical power as well as enhance robustness and validity in downstream analyses that could include defining cell types and profiling gene regulations across different cellular conditions and processes.

Real data application

We benchmarked three widely-used batch effect correction methods, MNN, Seurat v3, and LIGER, on two hematopoietic scRNA-seq datasets produced using two different sequencing platforms, MARs-seq and SMART-seq2.[68,69]

The first batch produced by MARS-seq consists of 1920 cells of six major cell types, and the second batch generated by SMART-seq2 contains 2730 of 3 cell types. Between the two batches, three types of cells are shared: common myeloid progenitor cells, granulocyte-monocyte progenitor cells and megakaryocyte-erythrocyte progenitor cells. For simplicity, only single cells belonging to these three shared cell types were extracted for our benchmarking analyses (Table 2).

Batch effect correction was carried out following the instruction of the 3 methods. Corrected results of all the 3 cell types were visualized by t-distributed stochastic neighbor embedding (Fig. 2A–D).[35,36] To quantify the mixing of single cells across batches using each of the 3 batch correction methods, we fitted two-way multivariate analysis of variance (MANOVA) models in the merged dataset after batch effect correction using each method. We calculated 2 *F* statistics from MANOVA: one for batch; and the other for cell type (here, three different cell types). These 2 *F* statistics represent differences between batches or cell types, respectively. Therefore, for the merged dataset, smaller batch *F* values are more desirable, since they indicate better mixing across batches. Whereas larger cell-type *F* values are preferred because they mean more differentiations retained between cell types. In the mouse hematopoietic datasets, all the three correction methods can substantially mitigate the discrepancy between the two datasets (Fig. 2E): LIGER outperforms Seurat v3 and MNN, which perform similarly (the lowest batch *F* value from LIGER in Fig. 2E). Moreover, LIGER also seems to reveal more differentiation between different cell types than the other two methods (largest cell-type *F* value from LIGER in Fig. 2F). Interestingly, MNN correction led to an even smaller cell-type *F* value, suggesting that MNN might lose some biological structures after correction.

In summary, in our real data application, LIGER seems to outperform the other two methods. However, as different datasets have varying properties, it may be impossible to find one batch effect correction that if always preferred for all datasets. We recommend that researchers take caution when choosing desired correction method(s) suitable for their data.

Future perspectives

As discussed, proper analysis of scRNA-seq data was very complex given various levels of biases, uncertainties, as well as sheer high dimensionality. Myriads of computational methods have been developed in recent years to facilitate scRNA-seq data analysis. We have focused in this review on methods for data pre-processing, specifically on dropout imputation and batch effect correction. Excellent reviews exist for various downstream analyses, including cell type clustering,[28] differential expression,[70] pseudo-time or cell trajectory estimation.[58]

We focus on preprocessing methods because we view them indispensable to ensure validity and enhance power for any downstream analysis. With the help of better batch effect correction and imputation tools, one can impute dropout events and integrate multiple scRNA-seq datasets more accurately and robustly. These pre-processing procedures can help researchers acquire larger, better harmonized datasets with more *bona fide* biological patterns recovered or retained, therefore boosting statistical power in downstream analysis.

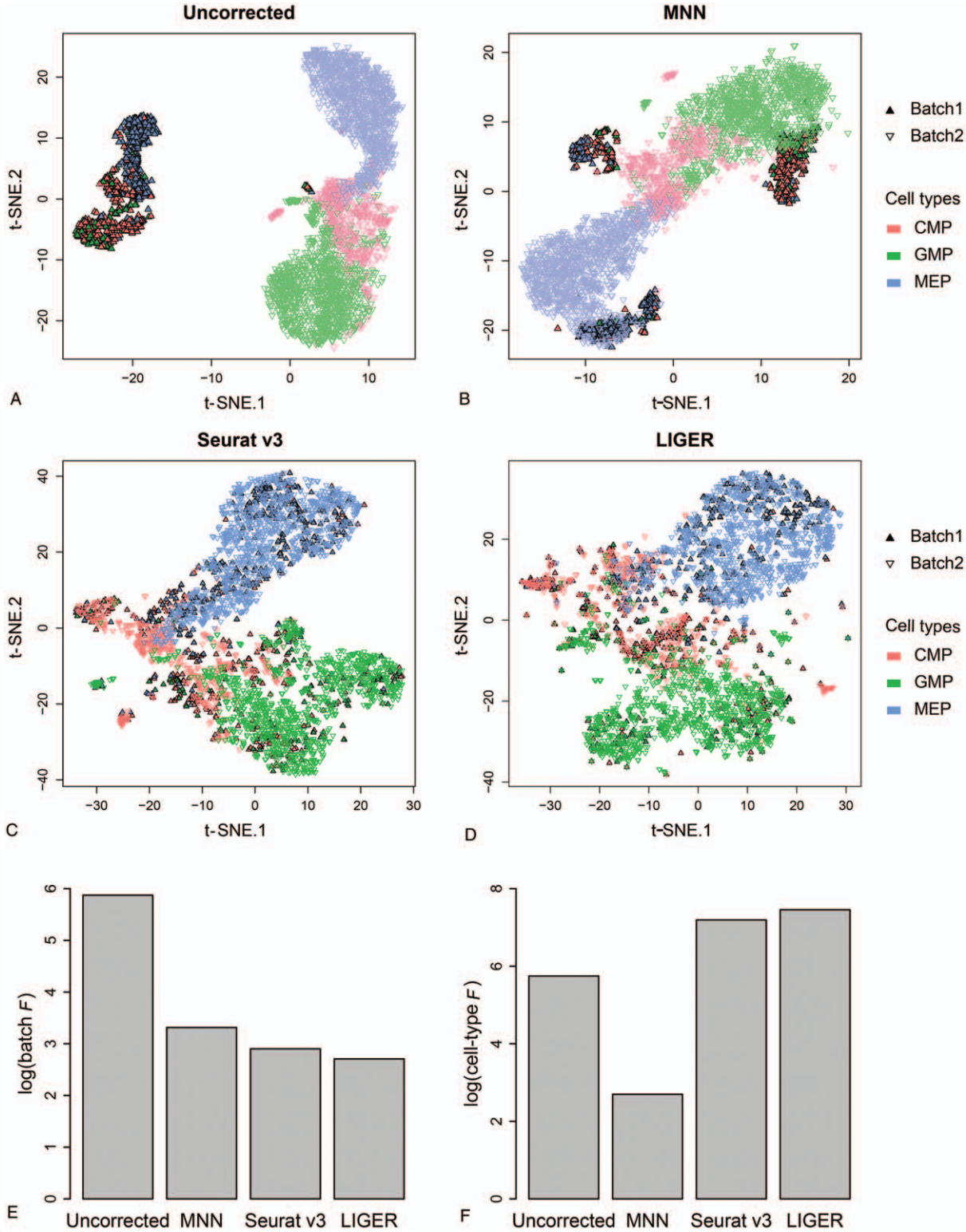

Figure 2. t-SNE visualization of uncorrected and corrected results. (A) t-SNE plots for 2 mouse hematopoietic datasets before correction. Solid and inverted triangle represent the first and second batch, respectively; and different cell types are shown in different colors. (B–D) t-SNE plots after correction by MNN, Seurat v3 and LIGER. (E) Logarithms of batch *F*-statistics. (F) Logarithms of cell-type *F*-statistics. CMP = common myeloid progenitor, GMP = granulocyte-monocyte progenitor, LIGER = link inference of genomic experimental relationships, MEP = megakaryocyte-erythrocyte progenitor, t-SNE = t-distributed stochastic neighbor embedding.

Each aforementioned tool, either for dropout imputation or for batch effect correction, has its strengths and limitations. For instance, a considerable proportion, if not the majority, of dropout imputation methods do not provide uncertainty quantification. This is because such uncertainty estimation incurs additional computational costs, which can easily become too prohibitive for scRNA-seq datasets that are getting ever increasingly large. Furthermore, when performing batch effect

correction across different scRNA-seq datasets, it is distinctly possible to overcorrect and lose true underlying cell type structures as a result. The old saying of "there is no free lunch" applies. Therefore, investigators are recommended to choose methods based on the nature of their data as well as which properties are most desired. From an optimistic perspective, we are hopeful that future methods could integrate more advantages of those methods in certain unified framework with affordable computational cost.

Although deep learning methods are often regarded as black box algorithms, they have demonstrated impressive values in many fields, and have already contributed to the analysis of scRNA-seq data. For example, the DCA method[47] was proposed to better denoise and impute dropout for scRNA-seq datasets and the single cell variational inference[41] method was developed to aggregate information across different batches to achieve higher accuracy for downstream analyses such as clustering and differential expression. With the development of transfer learning and other sophisticated neural networks, we anticipate accurate and efficient dropout imputation and batch effects correction methods with deep learning approaches.

With the continuous and rapid advancement of scRNA-seq technologies, we anticipate having more complete (eg, better capture rate, full-length transcripts, enrichment of rare cell types), more accurate (eg, with UMI to mitigate biases introduced by PCR), and more additional information (eg, cell type, cell location information), for the analyses of scRNA-seq data. In the foreseeable future, if these additional pieces of information are routinely available in scRNA-seq datasets, we expect increasingly more supervised or semi-supervised methods for dropout imputation and batch effect correction to be introduced. In addition, the integration of scRNA-seq data with other omics data (for instance DNA methylation, open chromatin status [ATAC-seq], or chromatin interactome [Hi-C]), both at bulk and single cell resolution, would allow more comprehensive borrowing information from different technologies and/or different omics assays, and thus hopefully provides a more complete understanding of the genome and the regulatory landscape.

Acknowledgements

We would like to express deep gratitude to Dr. Jeremy M. Simon (Department of Genetics, University of North Carolina at Chapel Hill, USA) for his useful suggestions on this research work. We also thank the Data Science Core of the Intellectual and Developmental Disabilities Research Center (IDDRC).

Author contributions

GL and YY wrote the manuscript. GL and YY analyzed the real data and provided visualization. EVB and YL revised and edited the manuscript. All authors approved the final version of the manuscript.

Financial support

None.

Conflicts of interest

The authors declare that they have no conflict of interest.

References

[1] Harris H. The birth of the cell. New Haven and London: Yale University Press; 2000.
[2] Waddington C. The strategy of the genes: a discussion of some aspects of theoretical biology. London: Allen & Unwin; 1957.
[3] Littman DR, Rudensky AY. Th17 and regulatory T cells in mediating and restraining inflammation. Cell 2010;140:845–858.
[4] Regev A, Teichmann SA, Lander ES, et al. The Human Cell Atlas. Elife 2017;6:e27041.
[5] Darmanis S, Sloan SA, Zhang Y, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A 2015;112:7285–7290.
[6] Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–386.
[7] Treutlein B, Brownfield DG, Wu AR, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–375.
[8] Buganim Y, Faddah DA, Cheng AW, et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell 2012;150:1209–1222.
[9] Shalek AK, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 2013;498:236–240.
[10] Tang F, Barbacioru C, Bao S, et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 2010;6:468–478.
[11] Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods 2011;8:S6–11.
[12] Arsenio J, Kakaradov B, Metz PJ, et al. Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. Nat Immunol 2014;15:365–372.
[13] Grün D, Lyubimova A, Kester L, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 2015;525:251–255.
[14] Jia C, Hu Y, Kelly D, et al. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res 2017;45:10978–10988.
[15] Kalisky T, Quake SR. Single-cell genomics. Nat Methods 2011;8:311–314.
[16] Mahata B, Zhang X, Kolodziejczyk AA, et al. Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep 2014;7:1130–1142.
[17] Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 2018;18:35–45.
[18] Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–382.
[19] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50:96.
[20] Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. Front Genet 2019;10:317.
[21] Picelli S, Faridani OR, Björklund ÅK, et al. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 2014;9:171–181.
[22] Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 2014;343:776–779.
[23] Xin Y, Kim J, Ni M, et al. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proc Natl Acad Sci U S A 2016;113:3293–3298.
[24] Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161:1202–1214.
[25] Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;161:1187–1201.
[26] Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049.
[27] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14:618–630.
[28] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273–282.
[29] Rosenberg AB, Roco CM, Muscat RA, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 2018;360:176–182.

[30] Sasagawa Y, Nikaido I, Hayashi T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol 2013;14:R31.
[31] Sheng K, Cao W, Niu Y, et al. Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat Methods 2017;14:267–270.
[32] Islam S, Kjallquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 2011;21:1160–1167.
[33] Jiang L, Chen H, Pinello L, et al. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 2016;17:144.
[34] Haghverdi L, Lun ATL, Morgan MD, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36:421–427.
[35] Maaten LVD, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–2605.
[36] van der Maaten L. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 2014;15:3221–3245.
[37] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data Cell 2019;177: 1888-1902.e21.
[38] Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 2015;16:241.
[39] Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 2016;32:1–8.
[40] Welch JD, Kozareva V, Ferreira A, et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 2019;177:1873-1887.e17.
[41] Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods 2018;15:1053–1058.
[42] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16:133–145.
[43] van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;174:716-729.e27.
[44] Huang M, Wang J, Torre E, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods 2018;15:539–542.
[45] Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol 2018;19:196.
[46] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 2018;9:997.
[47] Eraslan G, Simon LM, Mircea M, et al. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun 2019;10:390.
[48] Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Res 2018;7:1740.
[49] Rozenblatt-Rosen O, Stubbington MJT, Regev A, et al. The Human Cell Atlas: from vision to reality. Nature 2017;550:451–453.
[50] Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor. New York, NY: Springer; 2005.
[51] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8: 118–127.
[52] Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res 2014;42:e161.
[53] Adey AC. Integration of single-cell genomics datasets. Cell 2019;177: 1677–1679.
[54] Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet 2019;20:257–272.
[55] Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–420.
[56] Risso D, Ngai J, Speed TP, et al. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 2014;32: 896–902.
[57] Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res 2016;44:e117.
[58] Saelens W, Cannoodt R, Todorov H, et al. A comparison of single-cell trajectory inference methods. Nat Biotechnol 2019;37:547–554.
[59] Welch JD, Hartemink AJ, Prins JF. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol 2016;17:106.
[60] Haghverdi L, Büttner M, Wolf FA, et al. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods 2016;13:845–848.
[61] Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017;14:979–982.
[62] Fan J, Salathia N, Liu R, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods 2016;13:241–244.
[63] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11:740–742.
[64] Van Buren E, Hu M, Weng C, et al. TWO-SIGMA: a novel TWO-component SInGle cell Model-based Association method for single-cell RNA-seq data. bioRxiv 2019;709238.
[65] Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015;16:278.
[66] Yang Y, Li G, Qian H, et al. SMNN: Batch Effect Correction for Single-cell RNA-seq data via supervised mutual nearest neighbor detection. bioRxiv 2019;672261.
[67] Meng C, Zeleznik OA, Thallinger GG, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform 2016;17:628–641.
[68] Nestorowa S, Hamey FK, Pijuan Sala B, et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 2016;128:e20–e31.
[69] Paul F, Arkin Y, Giladi A, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell 2015;163:1663–1677.
[70] Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 2018;15:255–261.