

MINI REVIEW

Emerging deep learning methods for single-cell RNA-seq data analysis

Jie Zheng*, Ke Wang

School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

* Correspondence: zhengjie@shanghaitech.edu.cn

Received June 24, 2019; Revised September 27, 2019; Accepted October 14, 2019

Deep learning is making major breakthrough in several areas of bioinformatics. Anticipating that this will occur soon for the single-cell RNA-seq data analysis, we review newly published deep learning methods that help tackle computational challenges. Autoencoders are found to be the dominant approach. However, methods based on deep generative models such as generative adversarial networks (GANs) are also emerging in this area.

Keywords: single-cell; RNA-seq; deep learning; autoencoder

Author summary: Single-cell RNA sequencing (scRNA-seq) and deep learning are revolutionizing the fields of biomedicine and artificial intelligence respectively. Due to features of scRNA-seq data (*e.g.*, large sample sizes, high dimensionality), deep learning looks a promising technique for the data analysis. This is the first review at the intersection of deep learning and scRNA-seq technologies. After listing computational challenges, we describe key ideas of representative deep learning methods and compare their strengths and limitations. Autoencoders have been used the most, although generative models are also emerging. We anticipate an explosive development of new methods in this young area.

INTRODUCTION

Deep learning is revolutionizing many data-rich research fields, such as biology and medicine [1]. Almost parallel to the booming of deep learning in the past decade, single-cell technology has become a major driving force behind rapid advances in biomedicine. With unprecedented quantity and resolution, single-cell data offer exciting opportunities to uncover many secrets about life, such as cancer drug resistance, gene regulation in embryonic development, mechanisms of stem cell differentiation and reprogramming [2–4]. With an increasing speed of data generation (jumping from hundreds of cells to millions of cells in the past few years), single-cell biology calls for powerful bioinformatics methods. Against this backdrop of big data, it is not surprising that single-cell bioinformatics would have intersection with deep learning. In this review, we are going to focus on the single-cell RNA-seq (scRNA-seq) data, because it is currently one of the most prevalent types of data in molecular biology, and by directly profiling the gene expression levels of cells it

links up dynamics at the molecular level and the cellular level.

Despite intense development of bioinformatics techniques for scRNA-seq data analysis in the past 5 years, some challenges are still not fully addressed, *e.g.*, dropout events, batch effect, noise, high dimensionality, and scalability. On the other hand, deep learning researchers need understand these computational challenges in order to use deep learning to help with the data analysis and knowledge discovery. By summarizing the challenges and state-of-the-art methods of applying deep learning to scRNA-seq data analysis, this review may contribute to building a bridge between single-cell bioinformatics and deep learning.

To use existing deep learning methods or develop new ones, we need obtain datasets of sufficient quantity and quality. Specialized scRNA-seq databases containing pre-processed data have been developed. For instance, PanglaoDB is an scRNA-seq database collected from mouse and human cells [5]. It contains pre-processed and pre-computed analyses from more than 1,054 single-cell

experiments covering most major single-cell platforms and protocols, based on more than 4 million cells from a wide range of tissues and organs. Besides, there have been many scRNA-seq datasets released along with journal publications (e.g., [6–8]), some of which have been used in the papers reviewed below.

CHALLENGES OF scRNA-SEQ DATA ANALYSIS

To develop useful algorithms and software tools for the analysis of scRNA-seq data, it is crucial to recognize and understand the computational challenges posed by the analysis tasks. Although many approaches have been proposed to address the challenges, more and better techniques are still needed. Moreover, with fast advances in single-cell technologies, new challenges of data analysis would appear. The following is a list of challenges that have been recently addressed by deep learning methods. Due to space limitation, state-of-the-art methods that are not deep learning will not be reviewed here. For details, readers can refer to a few reviews focused on the challenges in scRNA-seq data analysis [9–11].

Batch effect

High-throughput single-cell RNA-seq data are often collected in multiple batches, with different conditions, platforms or by different laboratories. It is inevitable that the difference among batches leads to different gene expression values, which might be confused with the biological variations due to inherent inter-cell heterogeneity. Such a technical bias is called *batch effect*. If not corrected, batch effect would result in spurious structures in the data and misleading conclusions in downstream analysis. It is worth noting that batch effect is not unique to single-cell data analysis. Recently published methods for batch effect removal in scRNA-seq data include canonical correlation analysis (CCA) [12] and mutual nearest neighbors (MNN) [13].

Dropout events

For some lowly expressed genes, their small numbers of RNA molecules and the stochastic nature of transcriptional processes could lead to spurious zero entries in the expression matrices of scRNA-seq data. This is called *dropout* [14]. To correct for the zero-inflation bias caused by dropout, numerous statistical methods (especially imputation) have been recently proposed [15,16]. Accidentally, in deep learning, the term “dropout” refers to a regularization technique that can address overfitting [17,18]. When the dropout technique is applied in deep

learning, every neuron has a chance of being temporarily ignored in the model training, thereby preventing co-adaptation among neighboring neurons and increasing the robustness and generalizability of the whole neural network.

Technical noise

In addition to batch effect and dropout events, some other technical factors could also cause biases in the scRNA-seq data, especially for lowly expressed genes, such as cDNA amplification bias, cell cycle effects, insufficient sequencing depth, etc., and such biases are called *technical noise*. On the other hand, single-cell data contain intrinsic biological variability which can reveal valuable insights about the mechanisms of gene regulation at single-cell level. It is therefore crucial and challenging to separate the technical noise from the biological noise [19]. To tackle the challenge, powerful biological techniques, such as ERCC (external RNA control consortium) spike-in [20] and unique molecular identifier (UMI) [21], have been developed and widely used. Nonetheless, computational techniques are still needed to further correct for the technical noise. The past few years have seen an increasing number of denoising methods, e.g., normalization [22,23], as well as statistical models for both biological noise and technical noise [24].

Curse of dimensionality

A key step in scRNA-seq data analysis is dimensionality reduction. Typically, an scRNA-seq dataset contains the expression profiles of a large number of genes, where each gene corresponds to a dimension, and the expression profile of each cell corresponds to a data point in the high-dimensional cell state space. In some data analysis steps (e.g., clustering), the distance between data points plays crucial roles. In a high-dimensional space, however, as the data points become sparse, the distance measures (e.g., Euclidean distance, Mahalanobis distance and Manhattan distance) lose their effectiveness, rendering the concept of nearest neighbor unclear and data analysis problems difficult. This is called the *curse of dimensionality* [25]. Moreover, it could lead to the issue of overfitting, especially when the number of data points is relatively small. One way to alleviate the issues caused by high dimensionality is to increase the data, but in most cases that would be infeasible because the amount of data required would increase exponentially with the dimensionality. Thus, an alternative and practical solution is dimensionality reduction [26].

Many approaches of dimensionality reduction have been employed or developed for scRNA-seq data, e.g., PCA (principal component analysis) [27], t-SNE [28,29],

diffusion map [13], GPLVM [30,31], SIMLR [32], and UMAP [33], etc. Recently, some dimensionality reduction methods are devised to take into account characteristics that are unique to scRNA-seq data, such as ZIFA which explicitly tackles dropout events [34]. Tested on simulated and real data, these methods have been demonstrated to be effective in extracting salient factors from the high-dimensional scRNA-seq datasets, and help improve performance in various downstream analysis, *e.g.*, clustering, visualization, cell type discovery, developmental trajectory reconstruction, pseudotime inference, and gene regulatory network inference. However, the existing methods for dimensionality reduction still have some limitations, such as the lack of robustness to random sampling, unable to capture global structures while focusing on local structures of the data, sensitivity to parameters, and high computational cost, etc.

Scalability

While dimensionality reduction mainly deals with the large number of genes in scRNA-seq data, the other key parameter of data size is the number of cells. Both parameters of data size pose the challenge of scalability. Since the birth of droplet-based scRNA-seq technique (*i.e.*, Drop-seq) [35], the amount of cells profiled in each experiment has reached tens of thousands and often millions. High-throughput single-cell projects, such as the Human Cell Atlas (HCA) project [4], are generating data of many cells, which calls for more efficient and scalable algorithms for modeling and data analysis. For instance, some methods for dimensionality reduction and clustering require multiplication of two $N \times N$ matrices, where N is the number of cells. Besides algorithmic innovations, some parallel and high-performance computing techniques, *e.g.*, graphical processing units (GPUs), are also frequently used in different areas of bioinformatics [36].

AUTOENCODERS

In the past 5 years, numerous statistical and machine learning methods have been proposed to address the above issues of scRNA-seq data analysis. However, deep learning techniques have been involved in tackling these issues only since around 2017. Among these techniques, autoencoder has been the most popular so far.

Autoencoders are artificial neural networks for unsupervised learning. Sometimes it is also counted as self-supervised learning as the target output is directly from the input data. The main idea of autoencoder is to learn efficient representations of data by forcing the neural network to reconstruct the input dataset itself as accurately as possible under some constraints [37]. While directly learning the identity function would be

easy but useless, imposing constraints on the internal hidden layers of the neural network (*e.g.*, lower dimensions) can force the model to ignore irrelevant information and capture most essential patterns in the data. A typical autoencoder comprises an encoder (which converts the input data to an internal representation) and a decoder (which generates output from the internal representation). The loss function is usually the reconstruction error based on some measure of distance between the input and output data (*e.g.*, Euclidean distance or Kullback–Leiler divergence). Autoencoders are often used for dimensionality reduction. Actually, principal component analysis (PCA) can be considered a special case of autoencoder when the cost function is the Mean Squared Error and only linear activation functions are used. Other applications of autoencoders include pretraining for supervised learning, feature extraction, information retrieval, etc.

Dependent on specific applications, there are variants of autoencoders, such as stacked (or deep) autoencoders (where multiple hidden layers pile up) [38], denoising autoencoders (where noise is added to the input in order to push the autoencoder to recognize the informative patterns in the data) [39], and variational autoencoders (where a sampling layer lies between the encoder and decoder such that the output data consist of instances sampled from the same distribution as the input data) [40].

DEEP LEARNING METHODS FOR scRNA-SEQ DATA ANALYSIS

In the following, we briefly review several deep learning methods that have been applied to scRNA-seq data analysis, published between 2017 and 2019. Although the number of papers is still small in this direction, an explosion of publications is likely to come soon.

Shaham *et al.* proposed one of the first methods that use neural networks to remove batch effect [41]. They used the residual neural networks (ResNets), which can be very deep by avoiding exploding or vanishing gradients and increasing performance with depth. ResNets can easily learn mapping functions that are close to the identity function, and thus are suitable for calibrating a source sample to match the target sample in the batch effect removal. Moreover, the maximum mean discrepancy (MMD), a measure for distance between two probability distributions, was used to encode the loss function in the ResNets. The authors have applied the MMD-ResNet method to remove batch effects in both mass cytometry and scRNA-seq data.

Batch effect removal is usually done before cell clustering in scRNA-seq data analysis. However, a recently proposed algorithm called “DESC” (deep embedding algorithm for single-cell clustering) combines

clustering and batch effect removal in an iterative framework [42], considering the fact that different cell types have different degrees of vulnerability to batch effect. Here, a stacked autoencoder is used for dimensionality reduction, as a step of pretraining and parameter initialization for the iterative clustering. The trained encoder layers represent the mapping from the original data to the lower-dimensional representation.

To address the issue of dropout, several imputation methods have been developed, *e.g.*, MAGIC [43], scImpute [44] and drImpute [45]. Inspired by the recent success of autoencoders for sparse matrix imputation in collaborative filtering for recommendation systems, Talwar *et al.* proposed an autoencoder-based method called “AutoImpute” to handle the dropout in scRNA-seq data [46]. The authors used overcomplete autoencoders, which aim to regenerate the imputed expression matrix by focusing on the non-zero entries in the input sparse matrix.

Eraslan *et al.* proposed a method called “DCA” (deep count autoencoder) to solve the problems of denoising and imputation together [47]. The main idea is to define the loss function in terms of noise models, *e.g.*, negative binomial (NB) and zero-inflation negative binomial (ZINB). When the ZINB noise model is used, the loss function is the likelihood of the ZINB distribution. The output layer of DCA includes three neuron nodes for each gene, representing the mean of the NB distribution (as the denoised data), and two parameters of the ZINB distribution (*i.e.*, dispersion and dropout probability). Compared with other imputation and denoising methods for scRNA-seq data, DCA has two advantages. One is the ability to capture the nonlinear dependencies among genes, and the other is the scalability to millions of cells thanks to the efficiency of autoencoder and the support for GPU usage.

Several deep learning methods for the dimensionality reduction of scRNA-seq data have been recently proposed. Lin *et al.* compared 4 neural network architectures to learn the representation of scRNA-seq data, and used denoising autoencoder (DAE) to do unsupervised pre-training [48]. Their main goal, however, was to do supervised learning for disentangling cell types and database queries to infer cell types or states. They also found that the neural networks incorporating prior knowledge of protein-protein and protein-DNA interactions can perform better. Moreover, analysis of the learned models can yield biological insights, demonstrating some degree of interpretability of the neural networks. Ding *et al.* used the variational autoencoder (VAE) to infer the approximate posterior distributions of low-dimensional latent variables, thereby learning a parametric mapping from a high-dimensional space to a low-dimensional embedding [49]. Compared with popular existing meth-

ods such as t-SNE, their method named “scvis” can capture the global structure in the data, is more robust to noise, and has better interpretability thanks to the probabilistic nature of the latent variable model. However, according to Becht *et al.* [33], the running time of scvis is expensive especially for dimensionality reduction. Wang and Gu proposed a method called “VASC”, which uses the deep variational autoencoder for dimensionality reduction and visualization of scRNA-seq data [50]. The architecture of VASC includes the encoder network, the decoder network, and the zero-inflated layer which simulates the dropout events. Compared with existing methods such as PCA, t-SNE and ZIFA, VASC can capture nonlinear patterns in the data and has broader data compatibility. Moreover, VASC could recover cell developmental processes based on dimensionality reduction. However, it is still insufficient for the recovery of cell differentiation trajectories. Integrating the gene ontology (GO) with deep neural networks, Peng *et al.* proposed both an unsupervised method called “GOAE” (Gene Ontology AutoEncoder) and a supervised method called “GONN” (Gene Ontology Neural Network) for dimensionality reduction and clustering [51]. Their experimental results show that, by incorporating the prior knowledge from GO, both clustering performance and interpretability of the neural networks can be improved. However, the model is sensitive to the threshold in dealing with the GO terms.

Most of the above methods are each focused on one or two tasks of data processing and analysis. It would be desirable, however, to integrate different tools into one joint framework. Lopez *et al.* developed an integrative software tool called “scVI” (single-cell variational inference) which can carry out tasks including batch correction, library-size bias correction, dropout correction, imputation, dimensionality reduction, clustering, visualization [52]. The rationale is that different analysis tasks can reuse a common low-dimensional representation of scRNA-seq data to increase consistency and flexibility. scVI is a probabilistic approach based on the hierarchical Bayesian model. It uses variational autoencoders for dimensionality reduction and neural networks for multiple tasks such as estimating the dropout probability. Note that the architecture of the scVI algorithm is built on a highly modular deep learning framework, and thus better results could be obtained by testing other combinations of modules (*e.g.*, nonlinearities, regularization). Another hierarchical Bayesian model with a deep autoencoder for data denoising, named SAVER-X, was published recently [53]. In addition, SAVER-X employs transfer learning to automate the process of cross-study information sharing and data integration. As such, the authors were able to do cross-species scRNA-seq data analysis from the mouse

cells to human cells so that the issue of human cell shortage could be addressed. More recently, a deep recurrent learning method called “scScope” was proposed to conduct batch effect removal, dropout imputation, and cell subpopulation identification and so on [54]. At the core of scScope is an autoencoder, with a layer for batch effect correction and a layer for imputation. The imputation layer returns to the beginning of the encoder, forming a recurrent network structure. If there is only one iteration, the framework is a standard autoencoder. Moreover, like other deep learning methods, scScope offers support for parallel training using GPUs, thereby promising to have the scalability for millions of cells.

Most of the deep learning methods reviewed above are summarized in Table 1.

DISCUSSIONS

The methods reviewed here have mostly used autoencoders, a popular class of unsupervised machine learning methods, probably because labeled data are often unavailable for scRNA-seq datasets, which renders supervised learning inapplicable.

Despite the exciting development of deep learning methods for scRNA-seq data analysis, there are still some limitations which could point to interesting directions of future work. First, deep learning is mostly used for data preprocessing (*e.g.*, dimensionality reduction, denoising, and imputation), serving to enhance rather than directly carry out downstream analysis tasks (*e.g.*, lineage

identification, gene regulatory network inference). Secondly, in most of the cases, it is still unclear whether deep learning methods are significantly better than traditional statistical or machine learning methods. An interesting future work would be to compare deep learning methods with other methods for specific tasks, to gain insights about when and why deep learning would perform better. Thirdly, the reported performance of deep learning methods may be sensitive to the values of hyperparameters [55], and the robustness and generalizability of deep learning methods should be assessed by testing on some third-party data. Of course, this issue is not specific to scRNA-seq data analysis. Fourthly, the integration of scRNA-seq data with other types of single-cell data (*e.g.*, single-cell Hi-C data, scATAC-seq data, single-cell proteomics and cell imaging data) could be facilitated by some deep learning methods in the future [56–58]. Last but not least, more interpretability is needed for scientific knowledge discovery. Although some of the reviewed papers have included biological interpretation of the hidden layers of autoencoders, a desirable aspect of interpretability is the automatic construction of a model that is able to simulate the single-cell dynamics of transcription. For that, generative adversarial network (GAN) models might be promising [59]. Although not as popular as some other deep learning methods in genomics, GAN has recently been applied to single-cell genomics [60]. For example, Ghahramani *et al.* used a GAN model to generate synthetic scRNA-seq data by simulation and to achieve dimensionality reduction [61].

Table 1 Different deep learning methods for scRNA-seq data analysis

Names	References	Years	Methods	Goals
Lin’s method	Lin <i>et al.</i> [48]	2017	PCA-based dimensionality reduction with denoising autoencoders	Dimensionality reduction, cell grouping, inference of cell type or state
AutoImpute	Talwar <i>et al.</i> [46]	2018	Autoencoder-based sparse gene expression matrix imputation	Deal with dropout events
scVI	Lopez <i>et al.</i> [52]	2018	Hierarchical Bayesian model and variational autoencoder	Tackle batch correction, library-size bias, dropout, imputation, and visualization, etc.
VASC	Wang & Gu [50]	2018	Variational autoencoder	Model the dropout events and find the nonlinear hierarchical feature representations of the original data
scvis	Ding <i>et al.</i> [49]	2018	Variational autoencoder	Model and visualize structures in scRNA-seq data
scScope	Deng <i>et al.</i> [54]	2019	Autoencoder with recurrent structure	Conduct batch effect removal, dropout imputation, cell subpopulation identification
DCA	Eraslan <i>et al.</i> [47]	2019	Deep count autoencoder with a ZINB loss function	Remove technical variation to improve downstream analyses
SAVER-X	Wang <i>et al.</i> [53]	2019	Bayesian hierarchical model and deep autoencoder with transfer learning	Leverage existing data to improve the quality of new scRNA-seq datasets

Their model has a higher interpretability of parameters compared with other models. A recent method for data integration, called the Manifold-Aligning GAN (MAGAN), was proposed by Amodio *et al.* [62]. MAGAN aligns two manifolds to maintain pointwise correspondence in order to integrate scRNA-seq data and proteomic data (e.g., mass cytometry).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Jie Zheng and Ke Wang declare that they have no conflict of interests.

This article is a review article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P. M., Zietz, M., Hoffman, M. M., *et al.* (2018) Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface*, 15, 20170387
- Tang, F., Lao, K. and Surani, M. A. (2011) Development and applications of single-cell transcriptome analysis. *Nat. Methods*, 8, S6–S11
- Berg, J. (2018) Exploring organisms cell by cell. *Science*, 362, 1333
- Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al.* (2017) The Human Cell Atlas. *eLife*, 6, e270416
- Franzén, O., Gan, L.-M. and Björkegren, J.L. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019, baz046
- Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, 20, 1131–1139
- Deng, Q., Ramsköld, D., Reinius, B. and Sandberg, R. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343, 193–196
- Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Illicic, T., Henriksson, J., Natarajan, K. N., Tuck, A. C., Gao, X., Bühler, M., Liu, P., *et al.* (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17, 471–485
- Kiselev, V. Y., Andrews, T. S. and Hemberg, M. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, 20, 273–282
- Stegle, O., Teichmann, S. A. and Marioni, J. C. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16, 133–145
- Poirion, O. B., Zhu, X., Ching, T. and Garmire, L. (2016) Single-cell transcriptomics bioinformatics and computational challenges. *Front. Genet.*, 7, 163
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36, 411–420
- Haghverdi, L., Büttner, F. and Theis, F. J. (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31, 2989–2998
- Kharchenko, P. V., Silberstein, L. and Scadden, D. T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11, 740–742
- Zhang, L. and Zhang, S. (2018) Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*
- Miao, Z., Deng, K., Wang, X. and Zhang, X. (2018) DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34, 3223–3224
- Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.R. (2012) Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580v1*
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15, 1929–1958
- Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10, 1093–1095
- Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R. and Oliver, B. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, 21, 1543–1551
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11, 163–166
- Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A. and Wang, W. (2015) Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, 31, 2225–2227
- Ding, B., Zheng, L. and Wang, W. (2017) Assessment of single cell RNA-Seq normalization methods. *G3 (Bethesda)*, 7, 2039–2045
- Kim, J. K., Kolodziejczyk, A. A., Illicic, T., Teichmann, S. A. and Marioni, J. C. (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.*, 6, 8687
- Bellman, R. and Corporation, R. (1957) *Dynamic programming*. Princeton: Princeton University Press
- Van Der Maaten, L., Postma, E. and Van den Herik, J. (2009) Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.*, 10, 13
- Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublot, J. T., Yosef, N., *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510, 363–369
- van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9, 2579–2605
- Amir, A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G.

- P. and Pe'er, D. (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, 31, 545–552
30. Lawrence, N. D. (2004) Gaussian process latent variable models for visualisation of high dimensional data. *Adv. in Neural Inf. Proc. Sys.*, 16, 329–336
31. Buettner, F. and Theis, F. J. (2012) A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, 28, i626–i632
32. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. and Batzoglou, S. (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, 14, 414–416
33. Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F. and Newell, E. W. (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, 37, 38–44
34. Pierson, E. and Yau, C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16, 241
35. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161, 1202–1214
36. Nobile, M. S., Cazzaniga, P., Tangherloni, A. and Besozzi, D. (2017) Graphics processing units in bioinformatics, computational biology and systems biology. *Brief. Bioinformatics*, 18, 870–885
37. Bourlard, H. and Kamp, Y. (1988) Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybern.*, 59, 291–294
38. Hinton, G. E. and Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. *Science*, 313, 504–507
39. Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.A. (2008) Extracting and composing robust features with denoising auto-encoders. In *Proceedings of the 25th International Conference on Machine learning*, pp. 1096–1103. ACM: Helsinki, Finland
40. Kingma, D.P. and Welling, M. (2013) Auto-encoding variational bayes. *arXiv:1312.6114v10*
41. Shaham, U., Stanton, K. P., Zhao, J., Li, H., Raddassi, K., Montgomery, R. and Kluger, Y. (2017) Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33, 2539–2546
42. Li, X., Lyu, Y., Park, J., Zhang, J., Stambolian, D., Susztak, K., Hu, G., Li, M. (2019) Deep learning enables accurate clustering and batch effect removal in single-cell RNA-seq analysis. *bioRxiv*, 530378
43. van Dijk, D., Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K. R., Chaffer, C. L., *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174, 716–729 e27
44. Li, W. V. and Li, J. J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, 9, 997
45. Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N. and Garry, D. J. (2018) DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19, 220
46. Talwar, D., Mongia, A., Sengupta, D. and Majumdar, A. (2018) AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.*, 8, 16329
47. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. and Theis, F. J. (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, 10, 390
48. Lin, C., Jain, S., Kim, H. and Bar-Joseph, Z. (2017) Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.*, 45, e156
49. Ding, J., Condon, A. and Shah, S. P. (2018) Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.*, 9, 2002
50. Wang, D. and Gu, J. (2018) VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genom. Proteom. Bioinf.*, 16, 320–331
51. Peng, J., Wang, X. and Shang, X. (2019) Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. *BMC Bioinformatics*, 20, 284
52. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. and Yosef, N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15, 1053–1058
53. Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C. and Zhang, N. R. (2019) Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods*, 16, 875–878
54. Deng, Y., Bao, F., Dai, Q., Wu, L. F. and Altschuler, S. J. (2019) Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat. Methods*, 16, 311–314
55. Hu, Q. and Greene, C. S. (2019) Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pac. Symp. Biocomput.*, 24, 362–373
56. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., MauckIII, W. M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R. (2019) Comprehensive integration of single-cell data. *Cell*, 177, 1888–1902 e21
57. Bhardwaj, V., Heyne, S., Sikora, K., Rabbani, L., Rauer, M., Kilpert, F., Richter, A. S., Ryan, D. P. and Manke, T. (2019) snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics*, btz436
58. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. and Smibert, P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, 14, 865–868
59. Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D.S., Krebs, C.F. and Bonn, S. (2018) Realistic *in silico* generation and augmentation of single cell RNA-seq data using Generative Adversarial Neural Networks. *bioRxiv*, 390153
60. Eraslan, G., Avsec, Ž., Gagneur, J. and Theis, F. J. (2019) Deep learning: new computational modelling techniques for genomics.

- Nat. Rev. Genet., 20, 389–403
61. Ghahramani, A., Watt, F. M. and Luscombe, N. M. (2018) Generative adversarial networks uncover epidermal regulators and predict single cell perturbations. bioRxiv, 262501
62. Amodio, M. and Krishnaswamy, S. (2018) MAGAN: Aligning biological manifolds. arXiv,1803.00385