Received: 3 December 2020 | Revised: 5 April 2021 | Accepted: 7 April 2021

DOI: 10.1002/widm.1414

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

Big data analytics in single-cell transcriptomics: Five grand opportunities

Namrata Bhattacharya[1,2] | **Colleen C. Nelson**[2,3] | **Gaurav Ahuja**[4] | **Debarka Sengupta**[1,2,4,5]

[1]Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi, India

[2]Australian Prostate Cancer Research Centre—Queensland, Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology (QUT), Brisbane, Queensland, Australia

[3]Princess Alexandra Hospital, Translational Research Institute, Woolloongabba, Queensland, Australia

[4]Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi, India

[5]Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi, India

Correspondence
Debarka Sengupta, Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi 110020, India.
Email: debarka@iiitd.ac.in

Gaurav Ahuja, Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi 110020, India.
Email: gaurav.ahuja@iiitd.ac.in

Colleen C. Nelson, Australian Prostate Cancer Research Centre—Queensland, Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology (QUT), Brisbane, QLD 4000, Australia.
Email: colleen.nelson@qut.edu.au

Edited by Sushmita Mitra, Associate Editor, and Witold Pedrycz, Editor-in-Chief

Abstract

Single-cell omics technologies provide biologists with a new dimension for systematically dissecting the underlying complexities within biological systems. These powerful technologies have triggered a wave of rapid development and deployment of new computational tools capable of teasing out critical insights by analysis of large volumes of omics data at single-cell resolution. Some of the key advancements include identifying molecular signatures imparting cellular identities, their evolutionary relationships, identifying novel and rare cell-types, and establishing a direct link between cellular genotypes and phenotypes. With the sharp increase in the throughput of single-cell platforms, the demand for efficient computational algorithms has become prominent. As such, devising novel computational strategies is critical to ensure optimal use of this wealth of molecular data for gaining newer insights into cellular biology. Here we discuss some of the grand opportunities of computational breakthroughs which would accelerate single-cell research. These are: predicting cellular identity, single-cell guided in silico drug screening for precision medicine, transfer learning methods to handle sparsity and heterogeneity of expression data, establishing genotype–phenotype relationships at single-cell resolution, and developing computational platforms for handling big data.

This article is categorized under:

Algorithmic Development > Biological Data Mining

Fundamental Concepts of Data and Knowledge > Big Data Mining

Technologies > Machine Learning

KEYWORDS

big data, CRISPRi/a, drug screening, personalized medicine, single-cell RNA sequencing

1 | INTRODUCTION

Since the advent of single-cell RNA sequencing technologies in 2009, large volumes of single-cell data have been accumulated at an unprecedented rate. Recently implemented multi-omics measurements of single-cell genome, transcriptome, and proteome are opening new avenues to reveal cellular heterogeneity and cell type lineage relationships. Multi-omic single-cell technologies have the potential to reveal the nuances of the association between dynamic molecular profiles and how these co-determine higher-order cellular phenotypes (Colomé-Tatché & Theis, 2018; X. Tang, Huang, Lei, Luo, & Zhu, 2019). Because of its potential to advance our understanding of functional heterogeneity among seemingly similar cells in the context of the associated microenvironment, in the past decade, single-cell transcriptomics has been adopted in both basic science and clinical studies (Aldridge & Teichmann, 2020; Eberwine, Sul, Bartfai, & Kim, 2014). Moreover, it also enables the detection of rare cells, albeit ectopically expressed transcripts such as olfactory receptors, whose expression in the tissue level transcriptomics was mostly overshadowed (Gupta, Lalit, Biswas, & Maulik, 2020; Kalra, Mittal, Bajoria, et al., 2020; Kalra, Mittal, Gupta, et al., 2020). The exponential growth of profiled cell numbers, thus far, has closely mirrored the "Moore's Law" (Aldridge & Teichmann, 2020). Consequently, medical science breakthroughs in the future will be dependent on the ability to process and interpret large volumes of single-cell omics data, which will continue to offer the challenge of dimensionality as the sequencing cost declines. As a result, the field of genomic data science is now presented with the next big task concerning efficient storage, processing, and analysis of sequence data deluge stemming from single-cell studies (Angerer et al., 2017).

Single-cell transcriptomics enables high throughput profiling of thousands of single-cells in a single experiment. In 2025, the data produced from single-cell sequencing experiments are expected to exceed 1 zetta-bases = 106 peta-bases per year (Stephens et al., 2015). This has led to increased use of the catchphrase "big data," which refers to the exponential growth in the volume, variety, and velocity of data (Altaf-Ul-Amin, Afendi, Kiboi, & Kanaya, 2014; Pal, Mondal, Das, Khatua, & Ghosh, 2020). Most of the computational methods developed for ultra-large datasets are either computationally expensive or over-simplistic (Sinha, Kumar, Kumar, Bandyopadhyay, & Sengupta, 2018). Big data technologies incorporate parallelization, visualization, and distribution, which disentangle the hidden associations within large datasets (Yu & Lin, 2016). Big data technologies are a potent source for biologists and bioinformaticians to bridge the gap for designing scalable methods for handling high-dimensional data without compromising accuracy (Angerer et al., 2017; Sinha et al., 2018). Besides the scalability issues, a key challenge involved in handling single-cell data is to tackle high levels of noise and technical bias that may be generated due to small amounts of starting nucleic acids (Colomé-Tatché & Theis, 2018).

Given the high anticipated value of single-cell transcriptomics and the exponential growth of single-cell RNA sequencing (scRNA-seq) data in the next 5–10 years, here we highlight five grand opportunities in single-cell genomics and provide new dimensions towards handling these ultra-large datasets (Figure 1). We begin with the technical challenges and opportunities associated with cell-type annotation in the human body. Then, we discuss the new computational options for drug discovery processes for precision medicine. We then focus on transfer learning methods for single-cell data analysis that allows efficient representation of one single-cell gene expression data by accruing the learning from another independent data. Furthermore, we address scRNA-seq methods based on pooled genetic perturbation screens that use Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology to mute individual genes in single-cells. Finally, we discuss the adaptation of single-cell investigations to big data infrastructure and the platforms.

2 | DATA ANALYTICS BEST-PRACTICES IN SINGLE-CELL TRANSCRIPTOMICS: A SURVEY

The scRNA-seq datasets comprising vast numbers of cells are becoming commonly available across the globe, creating a data revolution for the field of genomics. To date, the primary focus of single-cell research has concentrated on quantifying tissue heterogeneity, discovering novel cell-types and states, discerning developmental trajectory, and assessing response to therapy (Lafzi, Moutinho, Picelli, & Heyn, 2018; Lähnemann et al., 2020). These are achieved by standard analytics techniques such as dimension reduction, differential expression analysis, clustering, and classification. However, the majority of these techniques struggle to scale with the increase in data volume. This section reviews the commonly used single-cell analysis techniques such as preprocessing, clustering, dimensionality reduction, cell-type classification, and differential expression analysis (Figure 2a).

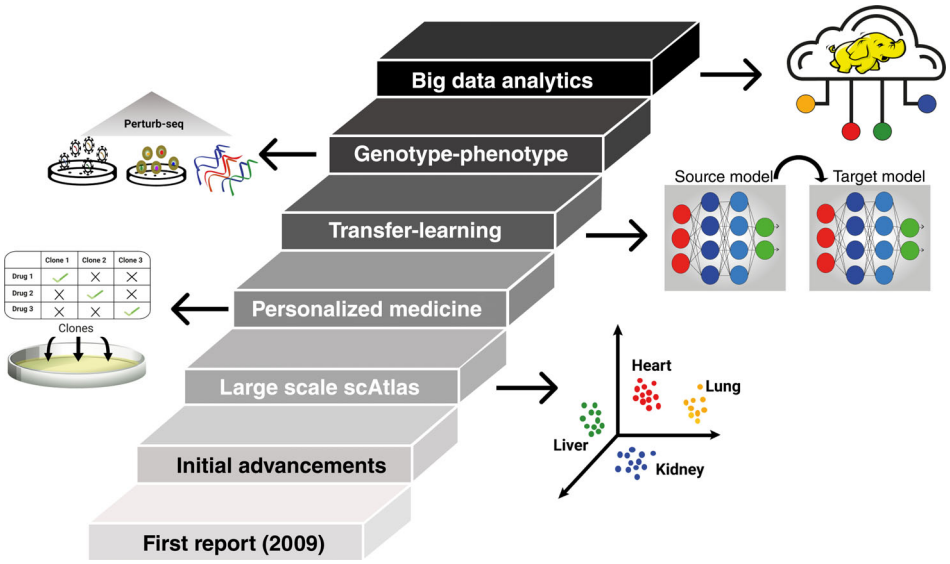

FIGURE 1 Schematic representation of the emergent opportunities for big data analytics in the field of single-cell transcriptomics

2.1 | Standard data preprocessing steps

The progress in scRNA-seq technology has prompted great opportunities for the quantitative characterization of cell-types (C. Wang, Gao, & Liu, 2020). However, due to nominal amounts of starting material found in an individual cell, the RNA must be amplified (Phipson, Zappia, & Oshlack, 2017). This process introduces errors and biases, resulting in unequal coverage, noise, and inaccurate quantification. Hence a significant challenge lies in filtering the data to include only high-quality cells and expressed genes followed by normalization of the expression profile to account for over-dispersed count value. For unbiased filtering of cells, Ilicic et al. (2016) developed a classification approach capable of processing raw data and removing most low-quality cells. A combination of biological and technical features capable of distinguishing low from high quality cells and microscopy annotation are used to train a SVM model that is capable of predicting low quality cells in the datasets. Optimal gene filtering for single-cell data (OGFSC) is a gene-filtering method that filters out genes based on constructing a thresholding curve (Hao, Cao, Huang, Zou, & Han, 2019). The OGFSC algorithm models the dependency between means and variances of gene expression levels by a multiple linear model method. Eigen analysis is performed to identify the thresholding curve to filter out corrupted genes from the rest genes.

Highly variable genes discovery identifies the genes that contribute to the cell-to-cell variation in a seemingly homogenous cell population (Yip, Sham, & Wang, 2019). Bayesian analysis of single-cell sequencing data (BASiCS) is a Bayesian hierarchical model-based detection criterion for highly or lowly variable genes within a cell population (Vallejos, Marioni, & Richardson, 2015). BASiCS uses a joint model of biological and spike-in genes based on a Poisson structure to quantify unexplained technical noise. Normalization across cells is applied to ignore the sources of cell-to-cell expression variability attributable to technical covariates such as the number of molecules detected in each cell. Hafemeister and Satija (2019) presented a statistical modeling framework based on a generalized linear model (GLM) for the normalization and variance stabilization of molecular count data. SCnorm uses quantile regression for between-sample normalization that addresses the artifacts which bias the downstream analysis (Bacher et al., 2017). For every gene, SCnorm estimates the dependency of transcript expression on sequencing depth and group genes with similar dependency. Then, normalized estimates of expression are computed using intragroup adjustment of sequencing depth with a scaling factor. To allow scaling to a larger dataset in lesser time, Linnorm is a linear model and normality-based normalizing transformation algorithm with time complexity $O(n * \log(n))$ (Yip, Wang, Kocher, Sham, & Wang, 2017). Linnorm performs a prior logarithmic transformation on the expression data for linear regression analysis between the expression of each sample and the expression mean across samples. This allows expression levels to be adjusted both linearly and exponentially.

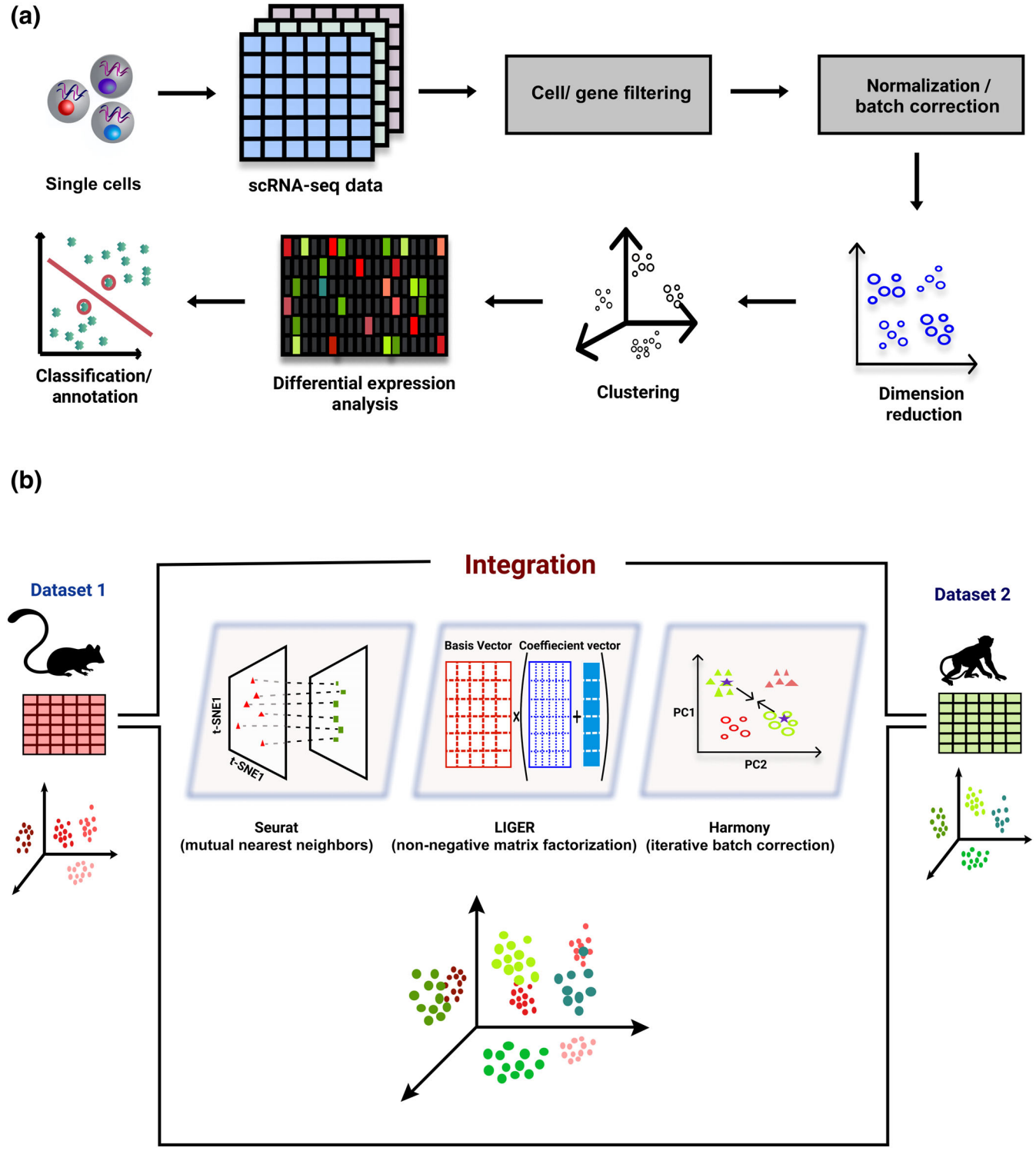

FIGURE 2 (a) Standard steps involved in the downstream analysis of scRNA-seq data. (b) Schematic diagram illustrating some of the widely adopted computational algorithms practiced for integrative analysis of multi-modal scRNA-seq data across species, conditions, and technology platforms

Batch effects arise from integration of scRNA-seq datasets, processed in different laboratories, using different chemistry and conditions. Batch correction methods are absolutely critical to enable integrative analysis of single-cell data (Haghverdi, Lun, Morgan, & Marioni, 2018; Tran et al., 2020). Haghverdi et al. (2018) proposed a batch correction strategy based on the detection of mutual nearest neighbors (MNNs) between batches in a high-dimensional expression

space. MNNs identify cells between different experimental batches that have mutually similar expression profiles, thus belonging to the same cell type or state. A prominent feature of the MNN correction method is that it uses a Gaussian kernel to adjust for local variations in the batch effects. A novel deep learning approach has been proposed by Shaham et al. (2017) for the systematic removal of batch effects. For batch correction, it works by minimizing the maximum mean discrepancy between the multivariate distributions of two replicates using a residual neural network, measured in separate batches. Harmony is a robust, scalable, and flexible multi-dataset integration algorithm that employs soft clustering (Korsunsky et al., 2019). It projects cells into a common reduced dimensional embedding where cells are grouped by cell type rather than data-specific conditions. Scanorama is another single-cell batch correction algorithm that approximates nearest neighbors using locality sensitive hashing (LSH) (Hie, Bryson, & Berger, 2019). It effectively integrates heterogeneous collections of scRNA-seq data by identifying and merging the shared cell types across all pairs of datasets.

2.2 | Discerning tissue heterogeneity with clustering

Defining cell-types through clustering based on transcriptome similarity is the most important step involved in the analysis of scRNA-seq data (Kiselev, Andrews, & Hemberg, 2019). Clustering enables finding differentially expressed genes between distinct cell lineages and conducting co-expression analysis across the individual clusters (Kharchenko, Silberstein, & Scadden, 2014; McKenzie et al., 2018; van Dam, Võsa, van der Graaf, Franke, & de Magalhães, 2018). Significant research is taking place to design scalable clustering methods for large volumes of data. DropClust leverages LSH to represent scRNA-seq profiles as binary codes without compromising clustering accuracy (Sinha et al., 2018; Sinha, Sinha, Saha, Bandyopadhyay, & Sengupta, 2019). It allows ultra-fast, memory-friendly clustering, and clean low-dimensional visualization of the cells. Single-cell Analysis via Iterative Clustering identifies the set of signature genes that helps identify optimal single-cell clusters (L. Yang, Liu, Lu, Riggs, & Wu, 2017). It employs an iterative k-means clustering for performing an exhaustive search for the best signature genes within the search space. The parameters are defined by several combinations of centers k and p-values obtained using analysis of variance tests. Single-cell consensus clustering uses the cluster-based similarity partitioning algorithm (Strehl & Ghosh, 2002) to combine multiple distinct clustering outcomes into a consensus matrix (Kiselev et al., 2017). The consensus matrix indicates how often each pair of cells appears in the same cluster. Complete-linkage hierarchical clustering of the consensus matrix gives the final result. Spearman subsampling-clustering-classification (SSCC) is a clustering framework for large-scale scRNA-seq data (Ren, Zheng, & Zhang, 2019). SSCC is a machine learning-based technique, including random projection and feature construction, that can cluster single-cells efficiently in $O(n)$ time. Clustering through Imputation and Dimensionality Reduction uses imputation to reduce the repercussions of dropouts in a scRNA-seq dataset (Lin, Troup, & Ho, 2017). It integrates dimension reduction and hierarchical clustering on estimated pairwise cell distances and identifies the clusters based on the Calinski–Harabasz index.

2.3 | Dimensionality reduction

Single-cell transcriptomics studies feature thousands to hundreds of thousands of single-cells (Lähnemann et al., 2020). It is necessary to project the high-dimensional data to a lower-dimensional space for two main reasons: (a) to ensure that the clusters are not resulted due to overfitting of expression data, and (b) to visualize the relative positioning of various cell-types in a 2-D cellular map (Tadist, Najah, Nikolov, Mrabti, & Zahi, 2019). Dimensionality reduction algorithms have significant applications in feature selection, denoising, imputation, removal of batch effects, rare cell-type detection, searching of similar cells in a large database of gene expression profiles, and so on (Tsuyuzaki, Sato, Sato, & Nikaido, 2020).

Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) are some of the popularly used dimensionality reduction algorithms (Hotelling, 1933; McInnes, Healy, & Melville, 2018; van der Maaten & Hinton, 2008). PCA is an unsupervised linear algorithm that projects data points in a low-dimensional space while retaining information. t-SNE is an unsupervised nonlinear algorithm. One of the most significant differences between PCA and t-SNE is that the latter also preserves local similarity among the data points. The most crucial t-SNE parameter, perplexity, governs the adequate number of neighbors and controls the Gaussian kernel's width to compute similarities between

points (Kobak & Berens, 2019). However, both PCA and t-SNE face limitations for large-scale data owing to long computation time and the inability to represent vast datasets meaningfully. UMAP is a relatively new dimensional reduction technique introduced by McInnes et al. in 2018. It is a graph-based algorithm similar to t-SNE that uses both supervised and unsupervised dimension reduction. It generates a high-dimensional graph representation of the data, subsequently optimizing a low-dimensional graph to be structurally identical.

ZIFA uses an updated probabilistic PCA approach that includes a zero-inflated model to account for dropout cases in large scRNA-seq datasets (Pierson & Yau, 2015). It is based on the empirical observation that lowly expressed genes are more likely to be affected by dropout than highly expressed genes. Badaoui, Amar, Hassou, Zoglat, and Okou (2017) proposed an approach based on feature extraction and selection. The feature extraction is based on correlation and rank analysis leading to a reduction in the number of variables. The feature selection chooses nonredundant and informative features based on a variant of linear discriminant analysis (Cai & Liu, 2011). It allows retrieval of the variables that contain complete information for the proper classification of cells. Townes et al. outlined an analysis framework with UMI counts based on a multinomial model. For non-normal distributions, it involves a generalized principal component analysis (GLM-PCA) model and a feature selection model based on deviance (Townes, Hicks, Aryee, & Irizarry, 2019).

2.4 | Single-cell classification

Rapid accumulation of profiled single-cells from diverse healthy and diseased tissues now presents the opportunity for automatic characterization of cell-types by referring to annotated single-cell data. Classifying the cell-types is essential to investigate their functional and pathological roles.

scPred is a generalized method that accurately classifies single-cells using a combination of unbiased feature selection and a probability-based classification method (Alquicira-Hernandez, Sathe, Ji, Nguyen, & Powell, 2019). scPred recognizes the transcriptional signatures that describe a cell-type. This information can be used to classify an individual cell based on its transcriptional profile, thus allowing characterization of heterogeneous cell populations. The gene expression matrix is eigen decomposed via singular value decomposition to obtain orthonormal linear combinations of the gene expression values. Wilcoxon signed-rank test filters principal components (PCs) with a low contribution to the variance. The resultant cell-PC matrix is used to train the support vector machine (SVM) model (Cortes & Vapnik, 1995). CaSTLe is a transfer learning-based cell-type classifier. It is based on the XGBoost (Chen & Guestrin, 2016) classification model built on the robust selection and transformation features (Lieberman, Rokach, & Shay, 2018). CaSTLe is well suited for classifying large datasets, having the additional advantage of being parallelizable. scReClassify uses PCA for dimension reduction and a semi-supervised learning approach to correctly identify and reclassify misclassified cells to their correct cell-types (T. Kim et al., 2019). Given the expression matrix and initial cell-type annotation for a scRNAseq dataset, scReClassify uses PCA followed by post hoc classification using AdaSampling with either SVM or random forest (Breiman, 2001).

2.5 | Statistical analysis of differential expression

Differential gene expression analysis between different cell populations provides insights into the intrinsic and extrinsic cellular processes in organisms, for example, uncovering driver genes in cancer (Myers, von Lersner, Robbins, & Sang, 2015; T. Wang, Li, Nelson, & Nabavi, 2019). Differential gene expression analysis on scRNA-seq cell-type clusters reveals the transcriptomic basis of tissue heterogeneity. Single-cell differential expression (SCDE) models each cell's measurement as a mixture of two probabilistic processes: one is zero-inflated negative binomial distribution, and another is the dropout component (Kharchenko et al., 2014). SCDE employs a Bayesian model to measure the likelihood of a gene being differentially expressed in one population versus another. Model-based analysis of single-cell transcriptomics (MAST) performs supervised analysis based on GLMs (Finak et al., 2015). MAST models discrete expression rates and continuous positive mean expression values to account for the bimodal distribution of single-cell transcriptome. The expression level conditioned on a cell expressing the gene is modeled as Gaussian, and the gene expression rate is modeled using logistic regression. ROSeq is another robust algorithm that models read counts for a gene as a function of ranks using a discrete generalized beta

distribution (Gupta, Mohanty, et al., 2020). ROSeq uses Wald's test to probe into differential expression across cell subtypes.

However, analyzing large numbers of cells poses the challenge of extended processing time and limited computing resources. Notably, bigSCale is a scalable pipeline for differential expression analysis, cell clustering, and marker identification that addresses the challenges associated with large datasets (Iacono et al., 2018). bigSCale uses a directed convolution strategy to handle the heterogeneity and sparsity. Nonparametric differential expression for single-cells (NODES) introduces a nonparametric method for detecting differential expression of >1,000 cells with increased accuracy and speed (Sengupta, Rayan, Lim, Lim, & Prabhakar, 2016). NODES put forth a pseudo-counted quantile normalization technique that reduces technical variability and improves the quality of downstream analyses.

2.6 | Missing value imputation

A significant challenge in scRNA-seq data is the "dropout" phenomenon caused by low capturing, high amplification bias, and low sequencing efficiency that affects all cells to varying degrees. This leads to a large percentage of genes expressed with zeros or low read counts. This dropout problem makes the expression matrix highly sparse. As a result, the downstream analysis such as clustering, classification, differential expression analysis, and pseudo-time analysis is hampered.

Several "imputation" methods have been developed to tackle the dropouts (Mukherjee, Zhang, Fan, Seelig, & Kannan, 2018; Peng, Zhu, Yin, & Tan, 2019). To impute dropouts, matrix completion based imputation for single-cell RNA-seq data (McImpute) uses a technique based on low-rank matrix completion of sparse single-cell expression data (Mongia, Sengupta, & Majumdar, 2019). McImpute applies soft thresholding iteratively on singular values without making any assumption about the expression data distribution. Deep learning-based pipeline, deepMc, imputes missing values in gene expression data using a deep matrix factorization-based method (Mongia, Sengupta, & Majumdar, 2020). SAVER is a model-based method that models UMI counts using a negative binomial distribution (Huang et al., 2018). SAVER leverages gene-to-gene relationships to retrieve the original expression value of each gene and then uses an empirical Bayes-like approach with a Poisson LASSO regression leveraging gene expression of other genes to estimate the dispersion parameter. Markov affinity-based graph imputation of cells (MAGIC) denoises the cell count matrix by performing data diffusion across similar cells to fill in missing transcripts (van Dijk et al., 2018). MAGIC leverages the large sample sizes in scRNA-seq to share information and uses a low-pass filtering technique identical to those used to clarify blurry and grainy images. MAGIC uses diffusion along an affinity-based graph structure to construct a weighted affinity matrix for the imputation of cells. AutoImpute is a deep learning-based imputation method for sparse gene expression matrix that uses over-complete autoencoders (Talwar, Mongia, Sengupta, & Majumdar, 2018). AutoImpute learns the inherent distribution of the input gene expression and imputes the missing value in the recovered matrix with minimal disruption to the biologically silent genes.

3 | GRAND OPPORTUNITIES

With the rapidly expanding commercial availability of various next-generation sequencing technologies over the past 10 years, even modest-sized laboratories and institutions worldwide can effectively produce petabytes of data in weeks. Present-day average next-generation sequencers can generate 1,000 Gb per month. With the advent of scRNA-seq technologies in 2009, there has been significant growth in scRNA-seq data (Hwang, Lee, & Bang, 2018; F. Tang et al., 2009). With the increasing availability of scRNA-seq platforms since then, we have reached a point where biomedical researchers can leverage large amounts of publicly available scRNA-seq data to augment their studies. This section discusses the five overarching opportunities in the single-cell studies, their status, and the expansion scope.

3.1 | Querying cell atlases

The Human Cell Atlas (HCA) project is a collaborative effort that targets defining all cell types in the human body in terms of their distinct molecular profiles (Regev et al., 2017). HCA associates these data with cellular descriptions to precisely investigate physiological states, regulatory circuitry, formative directions, and interaction of cells (Panina,

Karagiannis, Kurtz, Stacey, & Fujibuchi, 2020). This represents a cell as a superposition of "basis vectors," each of which determines a different but potentially dependent aspect of cellular function and organization (Wagner, Regev, & Yosef, 2016). Tabula Muris project and Mouse Cell Atlas are a similar effort to create a comprehensive catalogue of tissues in the mouse (Guo, Chen, & Zhou, n.d.; Tabula Muris Consortium et al., 2018). The gathered data have great potential for advancement in medicine and biomedical research intended for clinical applications (Panina et al., 2020). There are two major opportunities for cell annotation: one is to design a scalable algorithm for cell searching. Another is to create a central repository for storing cell metadata in a data lake from multiple sources in a homogeneous format.

Cell-type identification is possible by clustering the cells based on the proximity of their expression profiles and assigning labels to each group by cell population annotation. MARS is a meta-learning approach that annotates novel cell-types (Brbić et al., 2020). The method uses a deep learning method to transfer knowledge of cell embeddings across heterogeneous experiments. It annotates cells by probabilistically defining a cell-type based on nearest landmarks in the embedding space. However, the annotation step is challenging and time-consuming. The caveats become more pronounced with the increasing number of samples, hence intercepting quick and reproducible annotations (Abdelaal et al., 2019; Peyvandipour, Shafi, Saberian, & Draghici, 2020). Adding to this challenge are the massive amounts of data generated in the HCA project. For example, a recent work has generated an unprecedented amount of scRNA-seq data for 530,000 cells (Panina et al., 2020; Rosa, Pires, Zimmermannova, & Pereira, 2020). New fast and efficient algorithms need to be designed to overcome these limitations of characterizing unannotated single-cells.

CellAtlasSearch is a massively parallel and light-weight novel search engine for high-dimensional expression data based on LSH (Srivastava, Iyer, Kumar, & Sengupta, 2018). It uses graphical processing unit (GPU) computing and big data mining techniques for scalability. It calculates relative cell proximity by generating bit-vectors/hash-codes for each of the CellAtlasSearch database's reference expression profiles. GPU friendly version of LSH facilitates improved speed and accuracy of cell-type representation and querying. CellFishing.jl is a similar technique for searching atlas-scale datasets for similar cells based on their expression patterns and detecting genes based on LSH (Sato, Tsuyuzaki, Shimizu, & Nikaido, 2019). Like CellAtlasSearch, it reduces the computational time and space required for searching cells.

A large-scale effort is underway to create comprehensive and well-annotated reference atlases with millions of cells covering the full range of cell states. A single-cell is characterized by its mutation status, expression data, cell subtype, and several other metadata. Integrating multiple datasets collected under different technical and biological conditions is required to create these reference datasets (Kang, Nathan, Millard, Rumker, & Moody, 2020). The single-cell cohorts built using data generated from multiple platforms will increase the efficiency of cell-type identification. The emerging diversity of scRNA-seq datasets represents diverse biological/clinical conditions and is often generated using different technology platforms/chemistry, thereby posing a challenge for integrative analysis (Figure 2b). With the rapid growth in single-cell research, scRNA sequencing platforms are becoming more prevalent and diverse. The scRNA-seq library preparation includes protocols such as CEL-seq2 (Hashimshony et al., 2016), Drop-seq (Macosko et al., 2015), MARS-seq (Jaitin et al., 2014), Smart-seq, and Smart-seq2 (Picelli et al., 2013, 2014). There lies an opportunity for creating a central repository for integrating the different studies and storing the big data in a homogenous format. Databases such as PanglaoDB (Franzén, Gan, & Björkegren, 2019) and workflows such as recount (Collado-Torres, Nellore, Kammers, & Ellis, 2016), Symphony (Kang et al., 2020), LIGER (Welch et al., 2019), Seurat (Butler, Hoffman, Smibert, Papalexi, & Satija, 2018), and Harmony integrate data generated from multiple studies and present them through a unified framework to make a computational method more effective.

3.2 | Drug recommendation

Drug Discovery is a multifactorial process of developing drugs for the effective treatment of a given disease. In the era of big data in drug development, high-throughput screening techniques have driven an explosion of the data generated in biomedical research and healthcare systems. Drug discovery methods require methods to handle the enormous volume of structured and unstructured biomedical data (Jainul Fathima & Murugaboopathi, 2018). Effective data analysis and interpretation lead to a better understanding of diseases and the development of more personalized diagnostics and therapeutics (Merelli, Pérez-Sánchez, Gesing, & D'Agostino, 2014).

Personalized medicine has accelerated the pace of change in predictive science, where optimized medication can be tailored to individual patients. It provides better disease diagnostics, prognostication, prediction of treatment response, and drug discovery for several diseases, including cancer, with greater precision. Initiatives for precision medicine for

disease treatment must consider the changes in the molecular system. In the upcoming years, clinical-context-specific large-scale and complex datasets for precision medicine will be generated. This will require developing flexible frameworks to discover principal molecules for the future of drug development. The electronic patient health record is a big data source containing patient-associated clinical, demographic, genetic, and treatment information (Leff & Yang, 2015). Big data paradigms will drive new ways to store and access data for annotating and integrating it and making it available to find therapeutic targets and drugs for specific diseases. With protected patient privacy, cloud computing can ideally serve the large-scale analysis of patient data (Qian, Zhu, & Hoshida, 2019). It can narrow data integration gaps by addressing the massive computational resources required for big data-driven genomics research (Langmead & Nellore, 2018).

Single-cell analysis techniques studying cell development and differentiation have revealed cell-to-cell heterogeneity related to disease prognosis and treatment response at the level of individual cell resolution. The complexity of genetic, functional, and compositional heterogeneity of healthy and diseased tissues hinders the design of an accurate disease model. This can confound the interpretation of biomarker levels and patient responses to specific therapies. However, with the exponential growth of scRNA-seq data, it is required to design proper bioinformatics and machine learning methods to mine the rules associated with them (J. Yang, Liao, Zhang, & Xu, 2020). The burgeoning big single-cell RNA-seq data and development of complex patient-derived cancer models will help predict the response of multiple drugs in parallel and combinations for individual samples or tumor clones. It will be useful in designing tailored therapies (Adam et al., 2020; Anchang et al., 2018).

3.3 | Transfer learning single-cell representation

Transfer learning is a concept that data from a different domain can be used to train a predictive model. For instance, a model trained on widely available datasets, such as natural images, can be transferred to a target model that will perform similar tasks in a different domain, such as medical imaging, where training data are scarce (Koumakis, 2020).

Current single-cell gene-expression datasets pose an imbalance between a large number of genes, in the order of tens of thousands, and a small number of samples, in the order of a few hundred. Furthermore, technical noise obscures precise distinctions between cell states, preventing accurate quantification of genes with low expression. This causes scRNA-seq data to be noisy and sparse. As the scale of scRNA-seq studies continues to grow, more noisy and sparse data are being generated over time. Repeating denoising and clustering every time new data is generated is time-consuming. Transfer learning is an emerging approach for single-cell data used to denoise newly generated cells while utilizing information learned from analysis of previous data. SAVER-X is a transfer learning-based method that remarkably improves data quality of the sparse and noisy data. SAVER-X model discovers gene-to-gene relationships across data obtained from different labs with varying conditions (J. Wang, Agarwal, et al., 2019). SAVER-X couples a deep autoencoder with a Bayesian model to denoise new target datasets.

Transfer learning is gaining popularity in cell-type classification of single-cell data. CaSTLe is a parallel and efficient transfer learning-based method for classifying cells using previously labeled scRNA-seq datasets (Lieberman et al., 2018). The method includes a selection and transformation of feature, succeeded by the XGBoost algorithm. CaSTLe being parallelizable, is well suited for large datasets. ItClust is a cell type identification method that uses transfer learning to leverage cell-type-specific gene expression information learned from well-labeled source data, to help cluster and classify cell-types on newly generated unlabeled target data. However, due to computational memory and speed constraints, many recent tools have limitations in scaling to large datasets (Deng, Bao, Dai, Wu, & Altschuler, n.d.). Hence, new approaches are required to extract informative representations from these high-dimensional scRNA-seq profiles.

3.4 | Inferring gene regulation from perturbation experiments

To examine functional CRISPR screening in a single-cell granularity, single-cell CRISPR screening techniques combine pooled CRISPR screening with scRNA-seq (Adamson et al., 2016; Datlinger et al., 2017; Dixit et al., 2016). It can yield a wealth of data on the biological networks that underpin cellular response (Holding, Cook, & Markowetz, 2020). The available information makes large-scale studies about the disease progression, gene functions, and therapeutic response at a single-cell resolution more affordable. Single-cell CRISPR screens allow comprehensive readouts of cellular

phenotypes assessing both the CRISPR-mediated gene edits and gene knockdowns and the resulting perturbed gene expression profile (J.-X. Tang et al., 2018).

While a large and comprehensive amount of data can be a benefit, it has presented a significant computational challenge due to sparsity and noise (Liu & Trapnell, 2016). This has created an utter emergency in developing new and robust computational methods for better handling the resulting data. Perturb-seq is a high-throughput method that uses a reverse genetics approach (Dixit et al., 2016). They developed the multi-input, multi-output single-cell analysis (MIMOSCA) computational framework, which uses a regularized linear model to estimate the impact of perturbations on gene expression in a massively parallel manner. MIMOSCA coalesces multiplexed CRISPR-mediated gene inactivation with scRNA-seq to evaluate comprehensive gene expression phenotypes for each perturbation. It uses an elastic net regularization to fit the coefficient matrix, reducing the number of hypotheses tested and to address the correlated covariates and noisy data. Datlinger et al. (2017) designed CROP-seq for CRISPR droplet sequencing combining pooled CRISPR screening. Adamson et al. (2016) bridged the gap of functional genomics efforts facing tradeoffs between the number of perturbations examined and the complexity of phenotypes measured by Perturb-seq. An integrated pipeline MUSIC performs model-based discerning of single-cell CRISPR screening data (Duan et al., 2019). The pipeline can quantify and prioritize the perturbation effect of individual genes on cell phenotypes. And also handles the challenges due to data sparsity, and noise existing in single-cell CRISPR screening data analysis. MUSIC applies a data imputation step to improve the data quality. This is followed by a topic models based computational framework to handle single-cell CRISPR screening data.

Perturbation experiments constitute the central means to study cellular networks. Each experiment is a combinatorial perturbation of multiple genes. Gene knock-out using CRISPR-Cas allows experimental interventions on cellular networks and measurement of their effects. The main goal is to figure out which cell states are present in a heterogeneous sample, how cells switch between states, and which states are relevant to the biological process under investigation. Mapping the gene regulatory networks underlying cell states helps understand cellular heterogeneity (Fiers et al., 2018; Tiuryn & Szczurek, 2019). Differential equation-based local dynamic Bayesian network (DELDBN) is a new gene regulatory network inference hybrid algorithm that combines ordinary differential equation models and dynamic Bayesian network analysis (Li, Li, Krishnan, & Liu, 2011). To improve the speed and scalability, the DELDBN implemented a DBN learning algorithm using a local causality discovery algorithm by identifying a Markov blanket (Margaritis & Thrun, 2000) of the transcription rate. There are huge opportunities to develop more scalable algorithms since reverse-engineering large-sized network gene regulatory networks remain a challenge. As the network graph's size increases, it becomes difficult to process the graph in commodity hardware due to space constraints and execution time.

3.5 | Upgrading data storage and analytics platform

Petabytes of omics data are generated from clinical settings all over the world. These vast data need to be processed very rapidly for quick clinical decisions (Bhattacharya, Mondal, & Khatua, 2019). The rapid increase of single-cell RNA seq data demands efficient infrastructure for big data storage and analysis.

Hadoop is a powerful Java-based open-source tool that uses MapReduce architecture to store, design, and analyze very large datasets, such as Google File System (Ghemawat, Gobioff, & Leung, 2003). The fundamental part of Apache Hadoop is the Hadoop Distributed File System (HDFS) and MapReduce programming model. MapReduce is a programming model consisting of a mapper and reducer applied for distributed storage and processing of large datasets. Apache Spark is an open-source Hadoop replacement that uses resilient distributed datasets. Spark supports in-memory data processing, which makes it 100 times faster than Hadoop in multi-pass analytics. Spark provides native bindings for the Java, Scala, Python, and R programming languages. It is a distributed data processing platform that includes higher-level libraries supporting SQL queries, streaming data, machine learning, and graph processing (Dash, Shakyawar, Sharma, & Kaushik, 2019). Cloud computing is a platform that allows users to rent computers and storage from large data centers. It is well known for its elasticity, reproducibility, and confidentiality features. Several commercial cloud service platforms are currently available, including Amazon Web Services, Google Cloud Platform, Microsoft Azure, and so on (Qian et al., 2019). Large data files can be uploaded and posted to the cloud using cloud services. These data can be interfaced with by the server, allowing users to seamlessly integrate them with other data for use by downstream analysis tools.

Yu and Lin (2016) proposed a workflow for inter-institutional scRNA-seq data integration using Hadoop, a big data framework for scRNA-seq sequencing datasets. As part of this workflow, the inter-institutional scRNA-seq data are stored and managed in the Hadoop layer using HDFS. Single-cell expression profiles are normalized across different studies in the normalization layer to control the cross-assay technical variation. The differential expression, co-expression, and bi-clustering analysis are implemented on normalized data to identify the pattern in gene expression profiles in the analysis layer. The verification layer analyses the biological significance of the input gene set. A cloud-based framework Falco accelerates read alignment and transcript assembly of full-length scRNA-seq data (A. Yang, Kishore, Phipps, & Ho, 2019). It enables parallelization of existing RNA-seq processing pipelines using big data platforms such as Apache Hadoop and Apache Spark, allowing for efficient processing of scRNA-seq data.

Researchers use GPUs as the solution for high-performance computing. GPUs use general purpose computing on GPUs methodology to process big data. The two main properties of GPU are data parallelization and high throughput (Akizur & Muniyandi, 2018). Taylor-Weiner et al. (2019) showed that high efficiency of commonly used methods in genomics methods such as SignatureAnalyzer (J. Kim et al., 2016) and FastQTL (Ongen, Buil, Brown, Dermitzakis, & Delaneau, 2016) is achieved at low cost using GPUs. It leverages general-purpose libraries for computing, such as PyTorch and TensorFlow. They demonstrated runtime reductions of over 200-fold and cost reductions of 5–10-fold when compared to CPUs. For unrivaled speed in data processing and querying, CellAtlasSearch employs a GPU-friendly version of LSH (Srivastava et al., 2018).

Notably, the big data framework and high-performance computing offer the opportunity to identify significant correlations in new dimensionalities. The algorithmic complexity of large datasets can cope with within an acceptable run time using these platforms.

4 | CONCLUSION

Big data and scRNA-seq are the two fastest-growing fields of research. With the introduction of new technologies, high throughput scRNA-seq data are amassing at an unprecedented rate, necessitating the use of big data algorithms and technology platforms. Big data not only provides a framework for storing, processing, transforming, and visualizing data from various sources, but it also enables significant changes in cross-study data sharing. One of the main goals of big data scRNA-seq is to integrate heterogeneous, sparse, and large-scale datasets to construct models that can predict complex molecular signatures of the phenotypes or clinical outcomes.

Overall, this review focuses on various aspects of single-cell data analysis algorithms, including clustering, classification, dimension reduction, and differential analysis. We have also focused on the five opportunities in scRNA-seq data analysis and how the integration with big data technologies will lift it to a new height. While machine learning and statistical methods have been in use for genomic data analysis for decades now, we predict that big data algorithms will be indispensable to fully exploit the wealth lies in scRNA-seq data.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Namrata Bhattacharya: Conceptualization; visualization; writing-original draft. **Colleen C. Nelson:** Conceptualization; funding acquisition; project administration; supervision; writing-review & editing. **Gaurav Ahuja:** Conceptualization; funding acquisition; project administration; supervision; writing-review & editing. **Debarka Sengupta:** Conceptualization; funding acquisition; project administration; supervision; visualization; writing-original draft; writing-review & editing.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Namrata Bhattacharya https://orcid.org/0000-0002-5666-2551
Colleen C. Nelson https://orcid.org/0000-0001-6410-4843
Debarka Sengupta https://orcid.org/0000-0002-6353-5411

RELATED WIREs ARTICLES

Formatting biological big data for modern machine learning in drug discover (DOI: 10.1002/wcms.1408)

REFERENCES

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, *20*(1), 194. https://doi.org/10.1186/s13059-019-1795-z.

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: Challenges and recent progress. *NPJ Precision Oncology*, *4*, 19. https://doi.org/10.1038/s41698-020-0122-1.

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., ... Weissman, J. S. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, *167*(7), 1867–1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

Akizur, R. M., & Muniyandi, M. R. C. (2018). Review of GPU implementation to process of RNA sequence on cancer. *Informatics in Medicine Unlocked*, *10*, 17–26. https://doi.org/10.1016/j.imu.2017.10.008.

Aldridge, S., & Teichmann, S. A. (2020). Single cell transcriptomics comes of age. *Nature Communications*, *11*(1), 4307. https://doi.org/10.1038/s41467-020-18158-5.

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., & Powell, J. E. (2019). scPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, *20*(1), 264. https://doi.org/10.1186/s13059-019-1862-5.

Altaf-Ul-Amin, M., Afendi, F. M., Kiboi, S. K., & Kanaya, S. (2014). Systems biology in the context of big data and networks. *BioMed Research International*, *2014*, 428570. https://doi.org/10.1155/2014/428570.

Anchang, B., Davis, K. L., Fienberg, H. G., Williamson, B. D., Bendall, S. C., Karacosta, L. G., ... Plevritis, S. K. (2018). DRUG-NEM: Optimizing drug combinations using single-cell perturbation response to account for intratumoral heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(18), E4294–E4303. https://doi.org/10.1073/pnas.1711365115.

Angerer, P., Simon, L., Tritschler, S., Alexander Wolf, F., Fischer, D., & Theis, F. J. (2017). Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, *4*, 85–91. https://doi.org/10.1016/j.coisb.2017.07.004

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., ... Kendziorski, C. (2017). SCnorm: Robust normalization of single-cell RNA-seq data. *Nature Methods*, *14*(6), 584–586. https://doi.org/10.1038/nmeth.4263.

Badaoui, F., Amar, A., Hassou, L. A., Zoglat, A., & Okou, C. G. (2017). Dimensionality reduction and class prediction algorithm with application to microarray big data. *Journal of Big Data*, *4*(1), 1–11. https://doi.org/10.1186/s40537-017-0093-4

Bhattacharya, N., Mondal, S., & Khatua, S. (2019). A MapReduce-based association rule mining using Hadoop cluster—An application of disease analysis. In *Innovations in computer science and engineering* (pp. 533–541). Singapore: Springer. https://doi.org/10.1007/978-981-13-7082-3_61.

Brbić, M., Zitnik, M., Wang, S., Pisco, A. O., Altman, R. B., Darmanis, S., & Leskovec, J. (2020). MARS: Discovering novel cell types across heterogeneous single-cell experiments. *Nature Methods*, *17*, 1200–1206. https://doi.org/10.1038/s41592-020-00979-3

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/a:1010933404324

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096.

Cai, T., & Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, *106*(496), 1566–1577. https://doi.org/10.1198/jasa.2011.tm11199.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Collado-Torres, L., Nellore, A., Kammers, K., & Ellis, S. E. (2016). recount: A large-scale resource of analysis-ready RNA-seq expression data. *bioRxiv*. https://doi.org/10.1101/068478v1.abstract

Colomé-Tatché, M., & Theis, F. J. (2018). Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, *7*, 54–59. https://doi.org/10.1016/j.coisb.2018.01.003.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, *6*(1), 54. https://doi.org/10.1186/s40537-019-0217-0.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., ... Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, *14*(3), 297–301. https://doi.org/10.1038/nmeth.4177.

Deng, Y., Bao, F., Dai, Q., Wu, L. F., & Altschuler, S. J. (n.d.). Massive single-cell RNA-seq analysis and imputation via deep learning. https://doi.org/10.1101/315556

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., ... Regev, A. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, *167*(7), 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., ... Liu, Q. (2019). Model-based understanding of single-cell CRISPR screening. *Nature Communications*, *10*(1), 2233. https://doi.org/10.1038/s41467-019-10216-x.

Eberwine, J., Sul, J.-Y., Bartfai, T., & Kim, J. (2014). The promise of single-cell sequencing. *Nature Methods*, *11*(1), 25–27. https://doi.org/10.1038/nmeth.2769.

Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., & Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, *17*(4), 246–254. https://doi.org/10.1093/bfgp/elx046.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., ... Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.

Franzén, O., Gan, L.-M., & Björkegren, J. L. M. (2019). PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database: The Journal of Biological Databases and Curation*, *2019*, baz046. https://doi.org/10.1093/database/baz046

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. In *Proceedings of the nineteenth ACM symposium on operating systems principles* (pp. 29–43).

Guo, G., Chen, M., & Zhou, Y. (n.d.). *MCA – Mouse Cell Atlas*. Retrieved from http://bis.zju.edu.cn/MCA/

Gupta, K., Lalit, M., Biswas, A., & Maulik, U. (2020). ROSeq: Modeling expression ranks for noise-tolerant differential expression analysis of scRNA-Seq data. *bioRxiv*. https://doi.org/10.1101/374025v2.abstract

Gupta, K., Mohanty, S. K., Mittal, A., Kalra, S., Kumar, S., Mishra, T., ... Ahuja, G. (2020). The cellular basis of the loss of smell in 2019-nCoV-infected individuals. *Briefings in Bioinformatics*, *22*, 873–881. https://doi.org/10.1093/bib/bbaa168

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1), 1–15. https://doi.org/10.1186/s13059-019-1874-1.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427. https://doi.org/10.1038/nbt.4091.

Hao, J., Cao, W., Huang, J., Zou, X., & Han, Z.-G. (2019). Optimal gene filtering for single-cell data (OGFSC)-a gene filtering algorithm for single-cell RNA-seq data. *Bioinformatics*, *35*(15), 2602–2609. https://doi.org/10.1093/bioinformatics/bty1016.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., ... Yanai, I. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, *17*, 77. https://doi.org/10.1186/s13059-016-0938-8.

Hie, B., Bryson, B., & Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, *37*(6), 685–691. https://doi.org/10.1038/s41587-019-0113-3.

Holding, A. N., Cook, H. V., & Markowetz, F. (2020). Data generation and network reconstruction strategies for single cell transcriptomic profiles of CRISPR-mediated gene perturbations. *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms*, *1863*(6), 194441. https://doi.org/10.1016/j.bbagrm.2019.194441.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., ... Zhang, N. R. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nature Methods*, *15*(7), 539–542. https://doi.org/10.1038/s41592-018-0033-z.

Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, *50*(8), 96. https://doi.org/10.1038/s12276-018-0071-8.

Iacono, G., Mereu, E., Guillaumet-Adkins, A., Corominas, R., Cuscó, I., Rodríguez-Esteban, G., ... Heyn, H. (2018). bigSCale: An analytical framework for big-scale single-cell data. *Genome Research*, *28*(6), 878–890. https://doi.org/10.1101/gr.230771.117.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., & Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, *17*, 1–15. https://doi.org/10.1186/s13059-016-0888-1

Jainul Fathima, A., & Murugaboopathi, G. (2018). A novel customized big data analytics framework for drug discovery. *Journal of Cyber Security and Mobility*, *7*(1), 145–160.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., ... Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651.

Kalra, S., Mittal, A., Bajoria, M., Mishra, T., Maryam, S., Sengupta, D., & Ahuja, G. (2020). Challenges and possible solutions for decoding extranasal olfactory receptors. *The FEBS Journal*. *288*(8). https://doi.org/10.1111/febs.15606

Kalra, S., Mittal, A., Gupta, K., Singhal, V., Gupta, A., Mishra, T., ... Ahuja, G. (2020). Analysis of single-cell transcriptomes links enrichment of olfactory receptors with cancer cell differentiation status and prognosis. *Communications Biology*, *3*(1), 506.

Kang, J. B., Nathan, A., Millard, N., Rumker, L., & Moody, D. B. (2020). Efficient and precise single-cell reference atlas mapping with Symphony. *bioRxiv*. https://doi.org/10.1101/2020.11.18.389189v1.abstract.

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740–742. https://doi.org/10.1038/nmeth.2967.

Kim, J., Mouw, K. W., Polak, P., Braunstein, L. Z., Kamburov, A., Kwiatkowski, D. J., ... Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, *48*(6), 600–606. https://doi.org/10.1038/ng.3557

Kim, T., Lo, K., Geddes, T. A., Kim, H. J., Yang, J. Y. H., & Yang, P. (2019). scReClassify: Post hoc cell type classification of single-cell rNA-seq data. *BMC Genomics*, *20*(Suppl. 9), 913. https://doi.org/10.1186/s12864-019-6305-x.

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews. Genetics*, *20*(5), 273–282. https://doi.org/10.1038/s41576-018-0088-9.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., ... Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, *14*(5), 483–486. https://doi.org/10.1038/nmeth.4236.

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1), 5416. https://doi.org/10.1038/s41467-019-13056-x.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., ... Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

Koumakis, L. (2020). Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, *18*, 1466–1473. https://doi.org/10.1016/j.csbj.2020.06.017.

Lafzi, A., Moutinho, C., Picelli, S., & Heyn, H. (2018). Tutorial: Guidelines for the experimental design of single-cell RNA sequencing studies. *Nature Protocols*, *13*(12), 2742–2757. https://doi.org/10.1038/s41596-018-0073-y.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., ... Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6.

Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews. Genetics*, *19*(4), 208–219. https://doi.org/10.1038/nrg.2017.113.

Leff, D. R., & Yang, G.-Z. (2015). Big data for precision medicine. *Proceedings of the Estonian Academy of Sciences: Engineering*, *1*(3), 277–279. https://doi.org/10.15302/j-eng-2015075.

Li, Z., Li, P., Krishnan, A., & Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics*, *27*(19), 2686–2691. https://doi.org/10.1093/bioinformatics/btr454.

Lieberman, Y., Rokach, L., & Shay, T. (2018). CaSTLe – Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One*, *13*(10), e0205499. https://doi.org/10.1371/journal.pone.0205499.

Lin, P., Troup, M., & Ho, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, *18*(1), 59. https://doi.org/10.1186/s13059-017-1188-0.

Liu, S., & Trapnell, C. (2016). Single-cell transcriptome sequencing: Recent advances and remaining challenges. *F1000Research*, *5*, F1000. https://doi.org/10.12688/f1000research.7223.1

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., ... McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214.

Margaritis, D., & Thrun, S. (2000). Bayesian network induction via local neighborhoods. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in neural information processing systems* (Vol. 12, pp. 505–511). MIT Press.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. In *arXiv [stat. ML]*. arXiv. http://arxiv.org/abs/1802.03426

McKenzie, A. T., Wang, M., Hauberg, M. E., Fullard, J. F., Kozlenkov, A., Keenan, A., ... Zhang, B. (2018). Brain cell type specific gene expression and co-expression network architectures. *Scientific Reports*, *8*(1), 8868. https://doi.org/10.1038/s41598-018-27293-5.

Merelli, I., Pérez-Sánchez, H., Gesing, S., & D'Agostino, D. (2014). Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *BioMed Research International*, *2014*, 134023. https://doi.org/10.1155/2014/134023.

Mongia, A., Sengupta, D., & Majumdar, A. (2019). McImpute: Matrix completion based imputation for single cell RNA-seq data. *Frontiers in Genetics*, *10*, 9. https://doi.org/10.3389/fgene.2019.00009.

Mongia, A., Sengupta, D., & Majumdar, A. (2020). deepMc: Deep matrix completion for imputation of single-cell RNA-seq data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *27*(7), 1011–1019. https://doi.org/10.1089/cmb.2019.0278.

Mukherjee, S., Zhang, Y., Fan, J., Seelig, G., & Kannan, S. (2018). Scalable preprocessing for sparse scRNA-seq data exploiting prior knowledge. *Bioinformatics*, *34*(13), i124–i132. https://doi.org/10.1093/bioinformatics/bty293.

Myers, J. S., von Lersner, A. K., Robbins, C. J., & Sang, Q.-X. A. (2015). Differentially expressed genes and signature pathways of human prostate cancer. *PLoS One*, *10*(12), e0145322. https://doi.org/10.1371/journal.pone.0145322.

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, *32*(10), 1479–1485. https://doi.org/10.1093/bioinformatics/btv722

Pal, S., Mondal, S., Das, G., Khatua, S., & Ghosh, Z. (2020). Big data in biology: The hope and present-day challenges in it. *Gene Reports*, *21*, 100869. https://doi.org/10.1016/j.genrep.2020.100869.

Panina, Y., Karagiannis, P., Kurtz, A., Stacey, G. N., & Fujibuchi, W. (2020). Human cell atlas and cell-type authentication for regenerative medicine. *Experimental & Molecular Medicine*, *52*(9), 1443–1451. https://doi.org/10.1038/s12276-020-0421-1.

Peng, T., Zhu, Q., Yin, P., & Tan, K. (2019). SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biology*, *20*(1), 88. https://doi.org/10.1186/s13059-019-1681-8.

Peyvandipour, A., Shafi, A., Saberian, N., & Draghici, S. (2020). Identification of cell types from single cell data using stable clustering. *Scientific Reports*, *10*(1), 12349. https://doi.org/10.1038/s41598-020-66848-3.

Phipson, B., Zappia, L., & Oshlack, A. (2017). Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research*, *6*, 595. https://doi.org/10.12688/f1000research.11290.1

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1098. https://doi.org/10.1038/nmeth.2639.

Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006.

Pierson, E., & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, *16*, 241. https://doi.org/10.1186/s13059-015-0805-z.

Qian, T., Zhu, S., & Hoshida, Y. (2019). Use of big data in drug development for precision medicine: An update. *Expert Review of Precision Medicine and Drug Development*, *4*(3), 189–200.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., ... Human Cell Atlas Meeting Participants. (2017). The Human Cell Atlas. *eLife*, *6*, e27041. https://doi.org/10.7554/eLife.27041

Ren, X., Zheng, L., & Zhang, Z. (2019). SSCC: A novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data. *Genomics, Proteomics & Bioinformatics*, *17*(2), 201–210. https://doi.org/10.1016/j.gpb.2018.10.003.

Rosa, F., Pires, C., Zimmermannova, O., & Pereira, C.-F. (2020). Direct reprogramming of mouse embryonic fibroblasts to conventional type 1 dendritic cells by enforced expression of transcription factors. *Bio-Protocol*, *10*(10), e3619. https://doi.org/10.21769/bioprotoc.3619

Sato, K., Tsuyuzaki, K., Shimizu, K., & Nikaido, I. (2019). CellFishing.jl: An ultrafast and scalable cell search method for single-cell RNA sequencing. *Genome Biology*, *20*(1), 31. https://doi.org/10.1186/s13059-019-1639-x

Sengupta, D., Rayan, N. A., Lim, M., Lim, B., & Prabhakar, S. (2016). Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv*. https://doi.org/10.1101/049734v1.abstract

Shaham, U., Stanton, K. P., Zhao, J., Li, H., Raddassi, K., Montgomery, R., & Kluger, Y. (2017). Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, *33*(16), 2539–2546. https://doi.org/10.1093/bioinformatics/btx196.

Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S., & Sengupta, D. (2018). dropClust: Efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Research*, *46*(6), e36–e36. https://doi.org/10.1093/nar/gky007

Sinha, D., Sinha, P., Saha, R., Bandyopadhyay, S., & Sengupta, D. (2019). Improved dropClust R package with integrative analysis support for scRNA-seq data. *In Bioinformatics*, btz823. https://doi.org/10.1093/bioinformatics/btz823

Srivastava, D., Iyer, A., Kumar, V., & Sengupta, D. (2018). CellAtlasSearch: A scalable search engine for single cells. *Nucleic Acids Research*, *46*(W1), W141–W147. https://doi.org/10.1093/nar/gky421

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7), e1002195.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research: JMLR*, *3*(December), 583–617.

Tabula Muris Consortium, & Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, & Principal investigators. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727), 367–372.

Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F., & Zahi, A. (2019). Feature selection methods and genomic big data: A systematic review. *Journal of Big Data*, *6*(1), 79. https://doi.org/10.1186/s40537-019-0241-0

Talwar, D., Mongia, A., Sengupta, D., & Majumdar, A. (2018). AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Scientific Reports*, *8*(1), 16329.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., ... Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382.

Tang, J.-X., Chen, D., Deng, S.-L., Li, J., Li, Y., Fu, Z., ... Liu, Y.-X. (2018). CRISPR/Cas9-mediated genome editing induces gene knockdown by altering the pre-mRNA splicing in mice. *BMC Biotechnology*, *18*(1), 61.

Tang, X., Huang, Y., Lei, J., Luo, H., & Zhu, X. (2019). The single-cell sequencing: New developments and medical applications. *Cell & Bioscience*, *9*, 53.

Taylor-Weiner, A., Aguet, F., Haradhvala, N. J., Gosai, S., Anand, S., Kim, J., ... Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. *Genome Biology*, *20*(1), 228. https://doi.org/10.1186/s13059-019-1836-7.

Tiuryn, J., & Szczurek, E. (2019). Learning signaling networks from combinatorial perturbations by exploiting siRNA off-target effects. *Bioinformatics*, *35*(14), i605–i614. https://doi.org/10.1093/bioinformatics/btz334.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, *20*(1), 295. https://doi.org/10.1186/s13059-019-1861-6.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1), 12. https://doi.org/10.1186/s13059-019-1850-9.

Tsuyuzaki, K., Sato, H., Sato, K., & Nikaido, I. (2020). Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology*, *21*(1), 9. https://doi.org/10.1186/s13059-019-1900-3.

Vallejos, C. A., Marioni, J. C., & Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, *11*(6), e1004333. https://doi.org/10.1371/journal.pcbi.1004333

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics*, *19*(4), 575–592. https://doi.org/10.1093/bib/bbw139.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research: JMLR*, *9*(86), 2579–2605.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., ... Pe'er, D. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, *174*(3), 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, *34*(11), 1145–1160. https://doi.org/10.1038/nbt.3711.

Wang, C., Gao, X., & Liu, J. (2020). Impact of data preprocessing on cell-type clustering based on single-cell RNA-seq data. *BMC Bioinformatics*, *21*(1), 1–13. https://doi.org/10.1186/s12859-020-03797-8.

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., & Zhang, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, *16*(9), 875–878. https://doi.org/10.1038/s41592-019-0537-1.

Wang, T., Li, B., Nelson, C. E., & Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, *20*(1), 40. https://doi.org/10.1186/s12859-019-2599-6.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, *177*(7), 1873–1887.e17. https://doi.org/10.1016/j.cell.2019.05.006

Yang, A., Kishore, A., Phipps, B., & Ho, J. W. K. (2019). Cloud accelerated alignment and assembly of full-length single-cell RNA-seq data using Falco. *BMC Genomics*, *20*(Suppl. 10), 927. https://doi.org/10.1186/s12864-019-6341-6.

Yang, J., Liao, B., Zhang, T., & Xu, Y. (2020). Editorial: Bioinformatics analysis of single cell sequencing data and applications in precision medicine. *Frontiers in Genetics*, *10*, 1358. https://doi.org/10.3389/fgene.2019.01358.

Yang, L., Liu, J., Lu, Q., Riggs, A. D., & Wu, X. (2017). SAIC: An iterative clustering approach for analysis of single cell RNA-seq data. *BMC Genomics*, *18*(Suppl. 6), 689. https://doi.org/10.1186/s12864-017-4019-5.

Yip, S. H., Sham, P. C., & Wang, J. (2019). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings in Bioinformatics*, *20*(4), 1583–1589. https://doi.org/10.1093/bib/bby011.

Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C., & Wang, J. (2017). Linnorm: Improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Research*, *45*(22), e179. https://doi.org/10.1093/nar/gkx828.

Yu, P., & Lin, W. (2016). Single-cell transcriptome study as big data. *Genomics, Proteomics & Bioinformatics*, *14*, 21–30. https://doi.org/10.1016/j.gpb.2016.01.005

How to cite this article: Bhattacharya N, Nelson CC, Ahuja G, Sengupta D. Big data analytics in single-cell transcriptomics: Five grand opportunities. *WIREs Data Mining Knowl Discov*. 2021;e1414. https://doi.org/10.1002/widm.1414