 doi:10.20944/preprints202209.0327.v1

Bioinformatic Methods for Identifying Differentially Abundant Subpopulations in scRNA-seq Data

Giona Kleinberg[1,*], Ashwini Shinde[1], Sai Batchu[2], Michael Joseph Diaz[3], Kevin Thomas Root[3], Brandon Lucke-Wold[4]

[1]Northeastern University, Boston, MA, United States

[2]Montville, NJ, United States

[3]University of Florida, College of Medicine, Gainesville, FL, United States

[4]Department of Neurosurgery, University of Florida, Gainesville, FL, United States

*Corresponding author

Brandon Lucke-Wold MD, PhD, MCTS

Department of Neurosurgery

University of Florida

Brandon.Lucke-Wold@neurosurgery.ufl.edu

Abstract

Single-cell RNA sequencing data facilitates investigation of cell heterogeneity and subpopulations as well as differentially abundant states however modern single-cell RNA sequencing datasets are growing in size and complexity requiring advances in the bioinformatic methods that analyze them. Many methods exist for each step of analysis including read alignment, normalization, quality control, batch effect correction, imputation and dimensionality reduction. With so many options to choose from at each step of the analysis, benchmarking and a synthesis of the literature on the methods available is necessary to inform biological

researchers on the most optimal workflow for their data. Here, recent key methods of analysis are highlighted with a focus on methods that facilitate identification of cell subpopulations and differentially abundant cell states. With a constantly expanding toolset for each step in single-cell RNA sequencing dataset analysis, biological researchers should stay informed to utilize the most applicable methods for their own analyses.

Keywords

scRNA-seq, bioinformatics, subpopulations, analysis methods, single-cell RNA sequencing,

Introduction

Single-cell RNA sequencing (scRNA-seq) technology allows the investigation of the state of individual cells at high resolution. Single cell transcriptomes and other scRNA-seq datasets are rapidly increasing in quantity and complexity due to advances in scRNA-seq technology with modern datasets having thousands of features and millions of cells.[1] Existing methods of analysis of scRNA-seq data are abundantly available and have been benchmarked periodically.[2–4] Methods of analysis are developed quickly and frequently to complement the growth of single-cell transcriptome generation which despite frequent benchmarking, results in many modern bioinformatic methods not being discussed in current reviews hindering their immediacy and impact.

Selection of the ideal method of scRNA-seq analysis is informed by the characteristics and specifics of the dataset being analyzed as well as the hypotheses being tested. Due to the large quantity and complexity of bioinformatic methods available, this selection is not always straightforward, especially for biological researchers without a computational background.

 doi:10.20944/preprints202209.0327.v1

The large quantity of bioinformatic methods for scRNA-seq data analysis requires a review to allow researchers to more rapidly find the best method of analysis for a given dataset. Here, such a review is provided highlighting the standout modern methods available, comparing their usability and unique features, and identifying areas of analysis that could benefit from further method development. A specific focus is given to methods of identifying differentially abundant cell populations due to the great utility such analysis can provide when investigating cell heterogeneity.

Materials and Methods

The initial literature search was conducted using the PubMed Advanced Search Builder with MeSH keyword searches in order to identify modern methods of analysis. To investigate the current bioinformatic methods of scRNA-seq analysis available, the search term “scRNA-seq” AND “data analysis” was used which gave 185 results of which 20 were selected. Next, to place an emphasis on methods identifying differentially abundant cell populations, the search term "scRNA-seq" AND "subpopulations" was used which yielded 226 results of which 41 were selected.

Inclusion criteria involved full-text articles primarily consisting of bioinformatic methods, technique applications, meta-analyses, and reviews published within the last 10 years. All articles selected were written in the English language and published in the United States. To increase the accuracy of the information synthesized and to maintain integrity in the review of the literature, non-peer-reviewed articles were excluded (as determined by authors GK and AS). Any discrepancies in eligibility determinations were discussed openly and resolved by authors MD and SB.

Results

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 September 2022 doi:10.20944/preprints202209.0327.v1

Expression and Read Alignment

Before cell subpopulations can be identified in an scRNA-seq analysis, several steps must be completed. The most common first steps in bioinformatic pipelines for scRNA-seq data analysis are expression and read alignment, quality control, and normalization. Methods for completing these steps are well established and benchmarked although there are always new techniques that increase efficiency, accuracy, and resolution.[5, 6]

Quality of scRNA-seq data is variable and quantified by the mapping ratio of reads. Mapping tools that can provide this quantification are plentiful as scRNA-seq and bulk RNA-seq usually sequence transcripts as fastq formatted reads.[7, 8] Since there is no difference in read alignment, bulk RNA-seq mapping tools can also be applied to scRNA-seq data.[9] Some tools, such as Scissor or Digitaldlsorter integrate bulk expression data with scRNA-seq data to link phenotypes to cell subpopulations.[10, 11] The integration of bulk RNA-seq and scRNA-seq analysis methods provides a comprehensive set of tools for read alignment in this way. Popular alignment methods such as TopHat2, HISAT, and STAR are well benchmarked and used widely.[12–14] Compared to each other, HISAT is the fastest, followed by STAR and then TopHat2 although STAR uses much more memory since it is suffix-array based.[9]

For expression quantification, the standard methods used for genome-guided assembly include Cufflinks, Stringtie, and RSEM with Pertea et al. stating that StringTie performs the greatest.[15–17] Methods that do not require a reference genome (De novo) are typically lower accuracy and therefore are primarily applied only to organisms that do not have a reference genome available.[18]

Quality Control and Normalization

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 September 2022 doi:10.20944/preprints202209.0327.v1

Quality control is very important as even in the most sensitive protocols, dropout events (transcripts that cannot be detected) are common generating low-quality data. Dead, mixed, or broken cells often lead to misinterpretation of data in downstream analysis when they are not identified and accounted for early on which, especially in clinical applications, could have disastrous consequences.[19–21] Some commonly used methods of quality control include Scater and SinQC.[22–24] Data must also be normalized before downstream analysis in order to adjust for technical biases and noise which are inherent to scRNA-seq data due to the challenging protocol and low starting materials of the scRNA-seq procedure.[25] Popular approaches for normalization of bulk RNA-seq data including DESeq2 and trimmed mean of M values (TMM) can be applied to scRNA-seq datasets as well.[26, 27] Methods designed specifically for scRNA-seq data have also been designed due to the large amount of zero-expression values and variation present in scRNA-seq data. Some of the more popular methods include SCnorm and SAMstrt but there are many more approaches available.[28, 29] Computational workflows for quality control and normalization are also available online using tools such as Bioconductor which is a relatively accessible option for low level scRNA-seq data analysis.[30]

Imputation

Failed RNA amplification during scRNA-seq can cause missing values and dropouts within a dataset.[31] Imputation methods such as CMF-Impute allow for accurate correction of these dropout entries for scRNA-seq expression matrices.[32] Another cell subpopulation based bounded low-rank (PBLR) method for scRNA-seq data imputation was proposed by Zhang et Al. which takes into account the high heterogeneity of scRNA-seq data and the effect of gene expression levels on dropout events. PBLR can effectively recover information hidden by dropouts as well as identify cell sub-populations.[33]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 September 2022 doi:10.20944/preprints202209.0327.v1

Batch Effect

Due to the increasing use due to decreasing cost of scRNA-seq, the quantity of scRNA-seq datasets available is getting larger allowing many studies to investigate the transcriptomes of large amounts of cells. As a result, batch effect is an increasing issue due to technical variation in scRNA-seq procedures, environment, and time.[34] Systematic error is introduced as a result of this effect which can cause misinformed conclusions after downstream analysis. Many methods have been proposed to handle such an effect.

The k-nearest-neighbor batch effect test (kBet) was developed specifically for detecting batch effects in scRNA-seq data. The knn batch effect test uses repeated χ^2-tests to compare the batch label composition of random local cell neighborhoods with global composition and output a rejection rate of the hypothesis (both batch compositions are the same).[35] Unlike methods of quality control and normalization, many batch correction methods using linear regression were designed with bulk RNA-seq in mind and assume that in each batch the cell population composition is identical, which is not usually the case in scRNA-seq data. Using the difference in expression of mutual nearest neighbors (MNNs) between batches can provide an estimate of the batch effect that can be honed by averaging across many MNN pairs.[36] Other tools such as Seurat allow for the alignment of different scRNA-seq datasets through integration based on common sources of variation within the datasets.[34]

Dimensionality reduction

Deep learning methods are frequently used however often struggle to cluster the high dimensional and increasing size of modern scRNA-seq datasets.[37–39] The most widely used linear dimensionality reduction algorithm is the PCA (Principal Component Analysis). Significant principal components can be used for nonlinear dimensionality reduction and to dependably

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 September 2022 doi:10.20944/preprints202209.0327.v1

indicate sources of heterogeneity in a dataset.[40] Another popular yet nonlinear dimensionality reduction method is the Uniform Manifold Approximation and Projection (UMAP) method. UMAP preserves the local data structure as well as the global structure, surpassing the alternative t-Distributed Stochastic Neighbor Embedding (t-SNE) method with the ability to preserve the global structure of the data and a shorter run time for very large-scale scRNA datasets.[40] Recent tools such as DivBiclust, PanoView, and scziDesk have been developed in order to make such analysis possible on larger datasets through biclustering or iterative clustering that can scale with dataset size reducing the need for dimensionality reduction.[41–44] An even newer unsupervised clustering approach called scGAC also seeks to analyze high dimensional and sparse datasets. The method utilizes latent relationship information across cells to graphically obtain cell clusters.[45]

Discussion

Identifying cell subpopulations

A primary application of scRNA-seq analysis is the identification of differentially abundant cell subpopulations as different populations typically highlight unique cell types.[38, 46–48] Identifying such cell subpopulations is an important step in understanding cell heterogeneity.[49–52] The standard in analysis pipelines is to apply this identification step at least after quality control and normalization in order to not introduce misleading artifacts.[9] There is a large quantity of tools that seek to accomplish this task and while some methods such as scPopCorn and SCCLRR claim outperformance of other tools during benchmark tests, not all workflows are identical and biological datasets often require different features for analyses.[53, 54]

For example, modern tools can vary in accessibility to such subpopulation analysis. The web-based application CHARTS allows for the investigation of cell subpopulations within tumors across public scRNA-seq cancer data sets.[55] This is a stark contrast to the many methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 September 2022 doi:10.20944/preprints202209.0327.v1

existing code workflows and packages that may require extensive programming knowledge to use effectively.

More developmental but less accessible tools such as f-scLVM and the MscNMF framework allow for the inference of interpretable factors that lead to such heterogeneity within cell subpopulations.[56, 57] Knowing more about factors that underpin cell heterogeneity can highlight new features to be analyzed by modern bioinformatic methods. For example, Poirion et al. propose the use of single nucleotide variations as alternative features for subpopulation identification.[58]

When identifying or investigating rare subpopulations that could contain only a few cells, many techniques are unable to characterize the different cell subpopulations accurately.[59, 60] One viable option is the MicroCellClust method which allows for rare subpopulation identification of exceedingly specific expression profiles that are highly expressed in some cells even if they are lowly expressed in peripheral cells.[61]

Other emerging techniques utilizing spatial-omics are furthering the utility of this analysis by also providing information on neighboring interactions of each identified cell state or subpopulation.[62, 63] Techniques such as those proposed by Moncada et al. facilitate this investigation of spatial patterns of gene expression using microarray-based spatial transcriptomics.[64] Others such as the method proposed by Ye et al. use co-expression network analysis to detect interactive gene groups using gene-to-gene interactions.[65] Dynamic methods such as these are important for datasets that may have nonlinear trends. Another example is the scRCMF method which specializes in characterizing subpopulations of cells that may transition from state to state.[21]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 September 2022 doi:10.20944/preprints202209.0327.v1

Differential expression analysis

Further analysis of identified cell populations can provide even more information on biological differences between conditions or cell types.[66, 67] A common method is differential expression analysis which facilitates identification of genes expressed differentially between identified distinct cell subpopulations.[25] Many tools are available for this analysis such as the muscat R package which was developed through a simulation based survey of the many state analyses and methods commonly used for differential state analysis.[68] Since many methods of detecting these differentially expressed subpopulations relies on initial clustering of cells of all cells which can miss more localized differences, new methods of analysis have been developed such as DA-seq which operates more locally to identify differentially abundant cell populations without being restricted to clustering.[69] Differentially expressed testing methods like the Bayesian approach and MAST can handle the presence of dropout elements in scRNA-seq data. Another and more efficient test for large scRNA-seq datasets is the Wilcoxon rank-sum test. After identifying the gene signatures of clusters, analyses like Gene Ontology Enrichment Analysis (GOEA) and Gene Set Enrichment Analysis (GSEA) can be implemented to identify the active biological processes in each cell's cluster.[40]

More research needs to be done on developing methods that scale with dataset size. As modern scRNA-seq methods increase in complexity, the datasets that are generated also increase in complexity as well as size. By focusing development on methods with performance independent or minimally diminished by dataset scale, workflows and pipelines using such methods will stay relevant for longer. Furthermore, more research should be conducted to employ more features when training algorithms to analyze scRNA-seq data using artificial intelligence. Through identifying additional features such as spatial-omics and cell-cell interactions, artificial intelligence models will be able to become more specialized and accurate.

 doi:10.20944/preprints202209.0327.v1

Lastly, a greater emphasis on usability should be undertaken. As the number of scRNA-seq data sets, especially those publicly available, increases, data analysis techniques requiring significant programming knowledge serve less utility as biological researchers without a computational background will be unable to use or troubleshoot them effectively. Methods designed with usability in mind in the form of a user interface or thorough documentation are more accessible and can therefore be used by a larger number of researchers. By increasing the ability of a model to be applied to as many biological investigations as possible, the utility of such a model can be improved. Overall, modern scRNA-seq analysis methods for identifying differentially abundant subpopulations are as widely varied and plentiful as they are complex. Those looking to apply such bioinformatic methods should spend time to consider the options available in order to find the best method suited to their biological hypothesis.

Statements and declarations

Author contributions

All authors contributed to the study conception and design. G Kleinberg, A Shinde, M Diaz, and S Batchu conducted the literature review. The first draft of the manuscript was written by G Kleinberg and A Shinde. B Lucke-Wold supervised. All authors commented on previous versions of the manuscript. All authors read, edited, and approved the final manuscript.

Funding

This material is based upon work supported by a Northeastern University Undergraduate Research and Fellowships PEAK Experiences Award.

Disclosure statement

The authors declare that they have no conflicts of interest.

References

[1] Tanay, A.; Regev, A. Scaling Single-Cell Genomics from Phenomenology to Mechanism. *Nature*, **2017**, *541* (7637), 331–338. https://doi.org/10.1038/nature21350.

[2] Chen, X.; Teichmann, S. A.; Meyer, K. B. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. *Annu. Rev. Biomed. Data Sci.*, **2018**, *1* (1), 29–51. https://doi.org/10.1146/annurev-biodatasci-080917-013452.

[3] Ding, J.; Adiconis, X.; Simmons, S. K.; Kowalczyk, M. S.; Hession, C. C.; Marjanovic, N. D.; Hughes, T. K.; Wadsworth, M. H.; Burks, T.; Nguyen, L. T.; et al. Systematic Comparative Analysis of Single Cell RNA-Sequencing Methods. bioRxiv May 23, 2019, p 632216. https://doi.org/10.1101/632216.

[4] Mereu, E.; Lafzi, A.; Moutinho, C.; Ziegenhain, C.; MacCarthy, D. J.; Alvarez, A.; Batlle, E.; Sagar; Grün, D.; Lau, J. K.; et al. Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects. bioRxiv May 13, 2019, p 630087. https://doi.org/10.1101/630087.

[5] Lieberman, B.; Kusi, M.; Hung, C.-N.; Chou, C.-W.; He, N.; Ho, Y.-Y.; Taverna, J. A.; Huang, T. H. M.; Chen, C.-L. Toward Uncharted Territory of Cellular Heterogeneity: Advances and Applications of Single-Cell RNA-Seq. *J. Transl. Genet. Genomics*, **2021**, *5*, 1–21. https://doi.org/10.20517/jtgg.2020.51.

[6] Poirion, O. B.; Zhu, X.; Ching, T.; Garmire, L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. *Front. Genet.*, **2016**, *7*, 163. https://doi.org/10.3389/fgene.2016.00163.

[7] Li, H.; Homer, N. A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing. *Brief. Bioinform.*, **2010**, *11* (5), 473–483. https://doi.org/10.1093/bib/bbq015.

[8] Chen, G.; Wang, C.; Shi, T. Overview of Available Methods for Diverse RNA-Seq Data Analyses. *Sci. China Life Sci.*, **2011**, *54* (12), 1121–1128. https://doi.org/10.1007/s11427-011-4255-x.

[9] Chen, G.; Ning, B.; Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.*, **2019**, *10*.

[10] Sun, D.; Guan, X.; Moran, A. E.; Wu, L.-Y.; Qian, D. Z.; Schedin, P.; Dai, M.-S.; Danilov, A. V.; Alumkal, J. J.; Adey, A. C.; et al. Identifying Phenotype-Associated Subpopulations by Integrating Bulk and Single-Cell Sequencing Data. *Nat. Biotechnol.*, **2021**. https://doi.org/10.1038/s41587-021-01091-3.

[11] Torroja, C.; Sanchez-Cabo, F. Digitaldlsorter: Deep-Learning on ScRNA-Seq to Deconvolute Gene Expression Data. *Front. Genet.*, **2019**, *10*, 978. https://doi.org/10.3389/fgene.2019.00978.

[12] Kim, D.; Langmead, B.; Salzberg, S. L. HISAT: A Fast Spliced Aligner with Low Memory Requirements. *Nat. Methods*, **2015**, *12* (4), 357–360. https://doi.org/10.1038/nmeth.3317.

[13] Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.; Salzberg, S. L. TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions. *Genome Biol.*, **2013**, *14* (4), R36. https://doi.org/10.1186/gb-2013-14-4-r36.

[14] Dobin, A.; Gingeras, T. R. Mapping RNA-Seq Reads with STAR. *Curr. Protoc. Bioinforma.*, **2015**, *51* (1), 11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51.

[15] Trapnell, C.; Williams, B. A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M. J.; Salzberg, S. L.; Wold, B. J.; Pachter, L. Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. *Nat. Biotechnol.*, **2010**, *28* (5), 511–515. https://doi.org/10.1038/nbt.1621.

[16] Li, B.; Dewey, C. N. RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome. *BMC Bioinformatics*, **2011**, *12* (1), 323. https://doi.org/10.1186/1471-2105-12-323.

[17] Pertea, M.; Pertea, G. M.; Antonescu, C. M.; Chang, T.-C.; Mendell, J. T.; Salzberg, S. L. StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. *Nat. Biotechnol.*, **2015**, *33* (3), 290–295. https://doi.org/10.1038/nbt.3122.
[18] Garber, M.; Grabherr, M. G.; Guttman, M.; Trapnell, C. Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq. *Nat. Methods*, **2011**, *8* (6), 469–477. https://doi.org/10.1038/nmeth.1613.
[19] Ilicic, T.; Kim, J. K.; Kolodziejczyk, A. A.; Bagger, F. O.; McCarthy, D. J.; Marioni, J. C.; Teichmann, S. A. Classification of Low Quality Cells from Single-Cell RNA-Seq Data. *Genome Biol.*, **2016**, *17* (1), 29. https://doi.org/10.1186/s13059-016-0888-1.
[20] Zhu, Y.; Huang, Y.; Tan, Y.; Zhao, W.; Tian, Q. Single-Cell RNA Sequencing in Hematological Diseases. *Proteomics*, **2020**, *20* (13), e1900228. https://doi.org/10.1002/pmic.201900228.
[21] Zheng, K.; Lin, L.; Jiang, W.; Chen, L.; Zhang, X.; Zhang, Q.; Ren, Y.; Hao, J. Single-Cell RNA-Seq Reveals the Transcriptional Landscape in Ischemic Stroke. *J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab.*, **2022**, *42* (1), 56–73. https://doi.org/10.1177/0271678X211026770.
[22] Stegle, O.; Teichmann, S. A.; Marioni, J. C. Computational and Analytical Challenges in Single-Cell Transcriptomics. *Nat. Rev. Genet.*, **2015**, *16* (3), 133–145. https://doi.org/10.1038/nrg3833.
[23] Jiang, P.; Thomson, J. A.; Stewart, R. Quality Control of Single-Cell RNA-Seq by SinQC. *Bioinformatics*, **2016**, *32* (16), 2514–2516. https://doi.org/10.1093/bioinformatics/btw176.
[24] McCarthy, D. J.; Campbell, K. R.; Lun, A. T. L.; Wills, Q. F. Scater: Pre-Processing, Quality Control, Normalization and Visualization of Single-Cell RNA-Seq Data in R. *Bioinformatics*, **2017**, *33* (8), 1179–1186. https://doi.org/10.1093/bioinformatics/btw777.
[25] Vallejos, C. A.; Richardson, S.; Marioni, J. C. Beyond Comparisons of Means: Understanding Changes in Gene Expression at the Single-Cell Level. *Genome Biol.*, **2016**, *17* (1), 70. https://doi.org/10.1186/s13059-016-0930-3.
[26] Love, M. I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.*, **2014**, *15* (12), 550. https://doi.org/10.1186/s13059-014-0550-8.
[27] Robinson, M. D.; Oshlack, A. A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. *Genome Biol.*, **2010**, *11* (3), R25. https://doi.org/10.1186/gb-2010-11-3-r25.
[28] Bacher, R.; Chu, L.-F.; Leng, N.; Gasch, A. P.; Thomson, J. A.; Stewart, R. M.; Newton, M.; Kendziorski, C. SCnorm: Robust Normalization of Single-Cell RNA-Seq Data. *Nat. Methods*, **2017**, *14* (6), 584–586. https://doi.org/10.1038/nmeth.4263.
[29] Katayama, S.; Töhönen, V.; Linnarsson, S.; Kere, J. SAMstrt: Statistical Test for Differential Expression in Single-Cell Transcriptome with Spike-in Normalization. *Bioinformatics*, **2013**, *29* (22), 2943–2945. https://doi.org/10.1093/bioinformatics/btt511.
[30] Lun, A. T. L.; McCarthy, D. J.; Marioni, J. C. A Step-by-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data with Bioconductor. *F1000Research*, **2016**, *5*, 2122. https://doi.org/10.12688/f1000research.9501.2.
[31] Soneson, C.; Robinson, M. D. Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis. *Nat. Methods*, **2018**, *15* (4), 255–261. https://doi.org/10.1038/nmeth.4612.
[32] Xu, J.; Cai, L.; Liao, B.; Zhu, W.; Yang, J. CMF-Impute: An Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinforma. Oxf. Engl.*, **2020**, *36* (10), 3139–3147. https://doi.org/10.1093/bioinformatics/btaa109.
[33] Zhang, L.; Zhang, S. Imputing Single-Cell RNA-Seq Data by Considering Cell Heterogeneity and Prior Expression of Dropouts. *J. Mol. Cell Biol.*, **2020**, *13* (1), 29–40. https://doi.org/10.1093/jmcb/mjaa052.

[34] Butler, A.; Hoffman, P.; Smibert, P.; Papalexi, E.; Satija, R. Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nat. Biotechnol.*, **2018**, *36* (5), 411–420. https://doi.org/10.1038/nbt.4096.
[35] Lütge, A.; Zyprych-Walczak, J.; Brykczynska Kunzmann, U.; Crowell, H. L.; Calini, D.; Malhotra, D.; Soneson, C.; Robinson, M. D. CellMixS: Quantifying and Visualizing Batch Effects in Single-Cell RNA-Seq Data. *Life Sci. Alliance*, **2021**, *4* (6), e202001004. https://doi.org/10.26508/lsa.202001004.
[36] Haghverdi, L.; Lun, A. T. L.; Morgan, M. D.; Marioni, J. C. Batch Effects in Single-Cell RNA Sequencing Data Are Corrected by Matching Mutual Nearest Neighbours. *Nat. Biotechnol.*, **2018**, *36* (5), 421–427. https://doi.org/10.1038/nbt.4091.
[37] Duò, A.; Robinson, M. D.; Soneson, C. A Systematic Performance Evaluation of Clustering Methods for Single-Cell RNA-Seq Data. *F1000Research*, **2018**, *7*, 1141. https://doi.org/10.12688/f1000research.15666.3.
[38] Chovatiya, G.; Ghuwalewala, S.; Walter, L. D.; Cosgrove, B. D.; Tumbar, T. High-Resolution Single-Cell Transcriptomics Reveals Heterogeneity of Self-Renewing Hair Follicle Stem Cells. *Exp. Dermatol.*, **2021**, *30* (4), 457–471. https://doi.org/10.1111/exd.14262.
[39] Mircea, M.; Hochane, M.; Fan, X.; Chuva de Sousa Lopes, S. M.; Garlaschelli, D.; Semrau, S. Phiclust: A Clusterability Measure for Single-Cell Transcriptomics Reveals Phenotypic Subpopulations. *Genome Biol.*, **2022**, *23* (1), 18. https://doi.org/10.1186/s13059-021-02590-x.
[40] Slovin, S.; Carissimo, A.; Panariello, F.; Grimaldi, A.; Bouché, V.; Gambardella, G.; Cacchiarelli, D. Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. *Methods Mol. Biol. Clifton NJ*, **2021**, *2284*, 343–365. https://doi.org/10.1007/978-1-0716-1307-8_19.
[41] Fang, Q.; Su, D.; Ng, W.; Feng, J. An Effective Biclustering-Based Framework for Identifying Cell Subpopulations From ScRNA-Seq Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2021**, *18* (6), 2249–2260. https://doi.org/10.1109/TCBB.2020.2979717.
[42] Chen, L.; Wang, W.; Zhai, Y.; Deng, M. Deep Soft K-Means Clustering with Self-Training for Single-Cell RNA Sequence Data. *NAR Genomics Bioinforma.*, **2020**, *2* (2), lqaa039. https://doi.org/10.1093/nargab/lqaa039.
[43] Shi, F.; Huang, H. Identifying Cell Subpopulations and Their Genetic Drivers from Single-Cell RNA-Seq Data Using a Biclustering Approach. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **2017**, *24* (7), 663–674. https://doi.org/10.1089/cmb.2017.0049.
[44] Hu, M.-W.; Kim, D. W.; Liu, S.; Zack, D. J.; Blackshaw, S.; Qian, J. PanoView: An Iterative Clustering Method for Single-Cell RNA Sequencing Data. *PLoS Comput. Biol.*, **2019**, *15* (8), e1007040. https://doi.org/10.1371/journal.pcbi.1007040.
[45] Cheng, Y.; Ma, X. ScGAC: A Graph Attentional Architecture for Clustering Single-Cell RNA-Seq Data. *Bioinforma. Oxf. Engl.*, **2022**, btac099. https://doi.org/10.1093/bioinformatics/btac099.
[46] Lee, J.; Geng, S.; Li, S.; Li, L. Single Cell RNA-Seq and Machine Learning Reveal Novel Subpopulations in Low-Grade Inflammatory Monocytes With Unique Regulatory Circuits. *Front. Immunol.*, **2021**, *12*, 627036. https://doi.org/10.3389/fimmu.2021.627036.
[47] Obradovic, A.; Chowdhury, N.; Haake, S. M.; Ager, C.; Wang, V.; Vlahos, L.; Guo, X. V.; Aggen, D. H.; Rathmell, W. K.; Jonasch, E.; et al. Single-Cell Protein Activity Analysis Identifies Recurrence-Associated Renal Tumor Macrophages. *Cell*, **2021**, *184* (11), 2988-3005.e16. https://doi.org/10.1016/j.cell.2021.04.038.
[48] Peng, J.; Sun, B.-F.; Chen, C.-Y.; Zhou, J.-Y.; Chen, Y.-S.; Chen, H.; Liu, L.; Huang, D.; Jiang, J.; Cui, G.-S.; et al. Single-Cell RNA-Seq Highlights Intra-Tumoral Heterogeneity and Malignant Progression in Pancreatic Ductal Adenocarcinoma. *Cell Res.*, **2019**, *29* (9), 725–738. https://doi.org/10.1038/s41422-019-0195-y.

[49] Crinier, A.; Milpied, P.; Escalière, B.; Piperoglou, C.; Galluso, J.; Balsamo, A.; Spinelli, L.; Cervera-Marzal, I.; Ebbo, M.; Girard-Madoux, M.; et al. High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. *Immunity*, **2018**, *49* (5), 971-986.e5. https://doi.org/10.1016/j.immuni.2018.09.009.

[50] Hundertmark, J.; Berger, H.; Tacke, F. Single Cell RNA Sequencing in NASH. *Methods Mol. Biol. Clifton NJ*, **2022**, *2455*, 181–202. https://doi.org/10.1007/978-1-0716-2128-8_15.

[51] Ji, A. L.; Rubin, A. J.; Thrane, K.; Jiang, S.; Reynolds, D. L.; Meyers, R. M.; Guo, M. G.; George, B. M.; Mollbrink, A.; Bergenstråhle, J.; et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell*, **2020**, *182* (2), 497-514.e22. https://doi.org/10.1016/j.cell.2020.05.039.

[52] Karaayvaz, M.; Cristea, S.; Gillespie, S. M.; Patel, A. P.; Mylvaganam, R.; Luo, C. C.; Specht, M. C.; Bernstein, B. E.; Michor, F.; Ellisen, L. W. Unravelling Subclonal Heterogeneity and Aggressive Disease States in TNBC through Single-Cell RNA-Seq. *Nat. Commun.*, **2018**, *9* (1), 3588. https://doi.org/10.1038/s41467-018-06052-0.

[53] Zhang, W.; Li, Y.; Zou, X. SCCLRR: A Robust Computational Method for Accurate Clustering Single Cell RNA-Seq Data. *IEEE J. Biomed. Health Inform.*, **2021**, *25* (1), 247–256. https://doi.org/10.1109/JBHI.2020.2991172.

[54] Wang, Y.; Hoinka, J.; Przytycka, T. M. Subpopulation Detection and Their Comparative Analysis across Single-Cell Experiments with ScPopCorn. *Cell Syst.*, **2019**, *8* (6), 506-513.e5. https://doi.org/10.1016/j.cels.2019.05.007.

[55] Bernstein, M. N.; Ni, Z.; Collins, M.; Burkard, M. E.; Kendziorski, C.; Stewart, R. CHARTS: A Web Application for Characterizing and Comparing Tumor Subpopulations in Publicly Available Single-Cell RNA-Seq Data Sets. *BMC Bioinformatics*, **2021**, *22* (1), 83. https://doi.org/10.1186/s12859-021-04021-x.

[56] Buettner, F.; Pratanwanich, N.; McCarthy, D. J.; Marioni, J. C.; Stegle, O. F-ScLVM: Scalable and Versatile Factor Analysis for Single-Cell RNA-Seq. *Genome Biol.*, **2017**, *18* (1), 212. https://doi.org/10.1186/s13059-017-1334-8.

[57] Wang, C.-Y.; Gao, Y.-L.; Kong, X.-Z.; Liu, J.-X.; Zheng, C.-H. Unsupervised Cluster Analysis and Gene Marker Extraction of ScRNA-Seq Data Based On Non-Negative Matrix Factorization. *IEEE J. Biomed. Health Inform.*, **2022**, *26* (1), 458–467. https://doi.org/10.1109/JBHI.2021.3091506.

[58] Poirion, O.; Zhu, X.; Ching, T.; Garmire, L. X. Using Single Nucleotide Variations in Single-Cell RNA-Seq to Identify Subpopulations and Genotype-Phenotype Linkage. *Nat. Commun.*, **2018**, *9* (1), 4892. https://doi.org/10.1038/s41467-018-07170-5.

[59] Roth, R.; Kim, S.; Kim, J.; Rhee, S. Single-Cell and Spatial Transcriptomics Approaches of Cardiovascular Development and Disease. *BMB Rep.*, **2020**, *53* (8), 393–399.

[60] Ryu, K. H.; Huang, L.; Kang, H. M.; Schiefelbein, J. Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. *Plant Physiol.*, **2019**, *179* (4), 1444–1456. https://doi.org/10.1104/pp.18.01482.

[61] Gerniers, A.; Bricard, O.; Dupont, P. MicroCellClust: Mining Rare and Highly Specific Subpopulations from Single-Cell Expression Data. *Bioinforma. Oxf. Engl.*, **2021**, btab239. https://doi.org/10.1093/bioinformatics/btab239.

[62] Bingham, G. C.; Lee, F.; Naba, A.; Barker, T. H. Spatial-Omics: Novel Approaches to Probe Cell Heterogeneity and Extracellular Matrix Biology. *Matrix Biol. J. Int. Soc. Matrix Biol.*, **2020**, *91–92*, 152–166. https://doi.org/10.1016/j.matbio.2020.04.004.

[63] Longo, S. K.; Guo, M. G.; Ji, A. L.; Khavari, P. A. Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics. *Nat. Rev. Genet.*, **2021**, *22* (10), 627–644. https://doi.org/10.1038/s41576-021-00370-8.

[64] Moncada, R.; Barkley, D.; Wagner, F.; Chiodin, M.; Devlin, J. C.; Baron, M.; Hajdu, C. H.; Simeone, D. M.; Yanai, I. Integrating Microarray-Based Spatial Transcriptomics and

Single-Cell RNA-Seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas. *Nat. Biotechnol.*, **2020**, *38* (3), 333–342. https://doi.org/10.1038/s41587-019-0392-8.

[65] Ye, X.; Zhang, W.; Futamura, Y.; Sakurai, T. Detecting Interactive Gene Groups for Single-Cell RNA-Seq Data Based on Co-Expression Network Analysis and Subgraph Learning. *Cells*, **2020**, *9* (9), E1938. https://doi.org/10.3390/cells9091938.

[66] McDavid, A.; Finak, G.; Chattopadyay, P. K.; Dominguez, M.; Lamoreaux, L.; Ma, S. S.; Roederer, M.; Gottardo, R. Data Exploration, Quality Control and Testing in Single-Cell QPCR-Based Gene Expression Experiments. *Bioinformatics*, **2013**, *29* (4), 461–467. https://doi.org/10.1093/bioinformatics/bts714.

[67] Wu, K.; Lin, K.; Li, X.; Yuan, X.; Xu, P.; Ni, P.; Xu, D. Redefining Tumor-Associated Macrophage Subpopulations and Functions in the Tumor Microenvironment. *Front. Immunol.*, **2020**, *11*, 1731. https://doi.org/10.3389/fimmu.2020.01731.

[68] Crowell, H. L.; Soneson, C.; Germain, P.-L.; Calini, D.; Collin, L.; Raposo, C.; Malhotra, D.; Robinson, M. D. Muscat Detects Subpopulation-Specific State Transitions from Multi-Sample Multi-Condition Single-Cell Transcriptomics Data. *Nat. Commun.*, **2020**, *11* (1), 6077. https://doi.org/10.1038/s41467-020-19894-4.

[69] Zhao, J.; Jaffe, A.; Li, H.; Lindenbaum, O.; Sefik, E.; Jackson, R.; Cheng, X.; Flavell, R. A.; Kluger, Y. Detection of Differentially Abundant Cell Subpopulations in ScRNA-Seq Data. *Proc. Natl. Acad. Sci. U. S. A.*, **2021**, *118* (22), e2100293118. https://doi.org/10.1073/pnas.2100293118.