Principle of single-cell RNA-seq data analysis for case-control studies

Abstract

Main

Case-control studies are commonly adopted in biomedicine research to discover the risk factors associated with diseases. It's especially suitable for pioneering studies in diseases lacking the revealing clear mechanisms to design randomized trials and for rare diseases or for biomedical labs which are not feasible to recruit enough cases to do cohort or longitudinal studies[1].

The advent of omics-based strategies drives the risk factors to be investigated to a more precise and personalized molecular level. Transcriptome, for example, reflects the tissue or cellular functional states linking to diseases and helps evaluate which functionally expressed genes are specifically associated with a phenotype in the key tissues. These strategies are based on the accuracy and resolution of data generation and learning, i.e., new sequencing technology and tailored data analysis approaches.

The past decade has seen a revolution in single-cell or single-nuclei sequencing (scRNA-seq/ snRNA-seq) since the first bona fide study in 2009[2]. They provide a relative abundance of different transcript species in cells or nuclei. The readouts cover, from the transcript abundance that characterizes cell identity to functional transcripts relevant to cell states, and the transcripts that contribute to individual specificity or are induced from treatment, pathologies, stimuli, or even from experimental procedures. Since cells are residents of a vast "landscape" of possible states[3], the powerful high-covered sampling on many cells in a static time, given the ergodic theory[4], could reflect the distribution of cellular heterogeneity in transcriptome within a short time, i.e., the order of the sampled cells in pseudo-time[5]. This is better resolved given the transcription[6] or post-transcription kinetics[7] that shapes the cellular dynamic landscape in terms of the transcriptome.

Despite amazing progress in sequencing technology to produce omics data from single cells or nuclei to atlas the common characteristics of cell types in an organism, the case-control study leveraging the barcode resolution remains challenging, with less than a hundred published research articles from 2018 (**Fig. 1a**). There are 111 datasets available on 80 diseases from human cell atlas[8]. If we extend the scope to include also the treatment studies, there are 155 NIH Bioprojects registered with scRNA or snRNA-seq performed on patient samples or patient-derived cell systems (**Fig. 1b**) (url for this collection: https://www.ncbi.nlm.nih.gov/sites/myncbi/hong.jiang.8/collections/63241581/public/). The study is costly in both the library preparation and sequencing steps (**Table 1**). Therefore, the study requires meticulous consideration given a limited budget in the number of samples, number of cells per sample, and number of reads per cell to meet the needed power according to the specific aim of the study[9]. Moreover, the medical insights from the high-cost study are limited by the correctness of data analysis in decomposing the variance due to individual or batch differences and the depth of analysis to utilize the rich sampling of the nuclei or cells. Numerous computational analysis strategies were established to overcome the difficulties from study design, and noise removal to probing biological questions such as the differential expression (DE) or eQTLs.

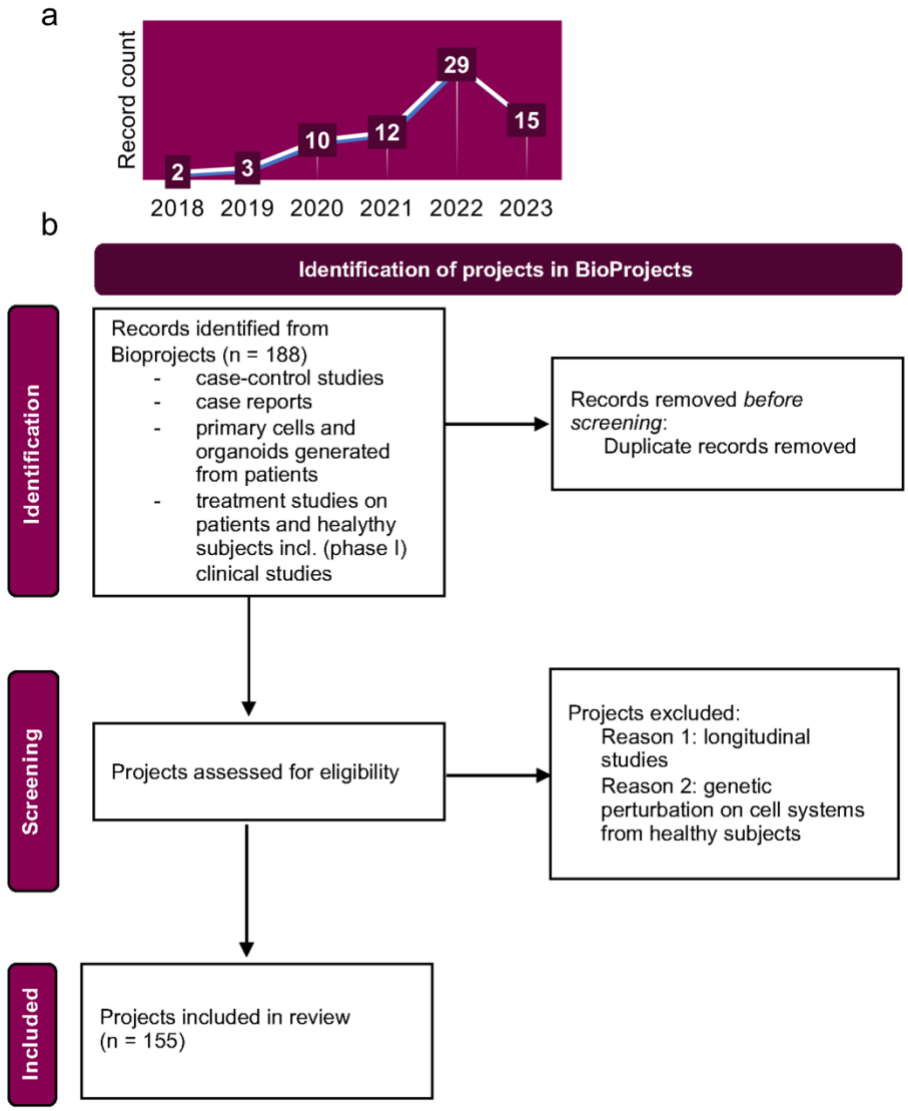

Fig.1 a) The number of publications with the topic in snRNA-seq or scRNA-seq case-control or human disease studies. Record counts are derived from Clarivate Web of Science. © Copyright Clarivate 2023. All rights reserved. **b)** Identification of projects in Bioprojects from NIH. The diagram is modified based on PRISMA diagram[10].

Technology	Library Preparation costs per cell	Sequencing costs/ M reads
10X Genomics	0.05€ - 0.12€	3.42€
Drop-Seq	0.09€	3.42€
Smart-Seq2	13€	3.42€

Table 1 Experimental cost per technology. Library preparation cost estimation (per cell) and sequencing cost estimation (per 1 million reads) for three of the most common single-cell RNA-seq technologies in Euro (€). For 10X Genomics, the cost depends on the number of cells per lane, an overloading of each lane with 20,000 cells generates costs of 0.05€ per cell, a loading with 8,000 cells per lane costs of 0.12€ per cell.

The key factors defining promising analysis strategies are also in frequent updates. Firstly, computational frameworks such as Seurat[11] and Scanpy[12] adopt structured sparse matrices to store large-scale sc- or snRNA-seq data with the annotation to features and metadata to observations. These data are integrated and deposited onto large databases for various species and diseases (research areas) such as the human cell atlas[8], PlantscRNAdb[13], and CancerSCEM[14]. The datasets are easily

accessible via repertories such as scfind[15] and R package scRNAseq. In the analysis processes, the high dimensional data can be represented as manifolds or networks. Then a whole bunch of manifold learning and network analysis tools could be applied. The statistical models are used to do the hypothesis testing to identify the associated features, such as generalized linear model (GLM) fit negative binomial model and Wald test or likelihood ratio test in DESeq[16]; two-part GLM for dropouts and non-zeros (hurdle distribution) in MAST[17]; Bayesian frameworks such as SCDE[18] and lvm-DE[19]; IDEAS using the permutation test[20] and MARBLES using the Markov model[21]. To find lower dimensional and meaningful patterns from the raw matrices, matrices decomposition methods including principal component analysis (the implements were benchmarked[22]), non-negative matrix factorization[23], Independent Component Analysis (reviewed[24]) and laten embedding multivariate gression[25] are applied. The machine learning and deep learning methods are not absent. In theory, any supervised questions could be modeled via some kind of deep network model. In practice, generative models simulate droplets or backgrounds[26]. Transfer learning to make use of the large databases. Not to say that cell type annotation is inherently a classification question that complies well with deep learning classifiers.

Although the xx and xx are comprehensively reviewed elsewhere, here we focus on understanding the data structure from the sequencing platforms as well as the study design of the expansion case-control studies. Based on these, we detailed how emerging analysis approaches were applied to solve the bottlenecks embedded in the study design and data structures. Although we are only at the door of new computational analysis tools to advance biomedical research, we suggest that the synergistic combination of bioinformatics and experimental testing and validation can markedly facilitate the important risk factors to be identified.

Understand the data from scRNA-seq in case-control studies

Understand the medical questions and study design.

The study is designed to answer certain medical questions and all the analytic means should match the questions. The understanding of the study design supports further clarification on how each element of the count matrix is generated (from a generative view). The incorporation of biomedical priors will optimize the analysis. As said, one biggest feature of biological data is the nature of signal and noise, which can only be comprehended with biomedical knowledge.

Classic, however, when …,. Case-control studies can sample more than one case group and/or more than one control group when there are unbeknownst conditions that are differently deviated from investigated cases or there are multiple disease subtypes. For example, in a single-cell RNA sequencing study of a COVID-19 patient with psoriasis treated with ustekinumab, both healthy and patients with untreated psoriasis lesions were used as controls to protect against the possibility that the identified cell types affected, and transcriptional changes were associated with psoriasis[27]. (A subtype example: Single-cell RNA sequencing identifies macrophage transcriptional heterogeneities in granulomatous diseases) For different cases, the controls should be chosen accordingly, i.e., no one-fit-all control groups. (Give examples and cite) When a study collects data on more than a single case and single control group, modeling the joint relationship between all groups and confounders and exposures of interest offers the possibility of comparing the cell types affected and genes differentially expressed from different confounders or exposures.

Address cells as samples in following paragraphs

Power analysis

As in general statistics, the power analysis describes the relation among (biological) effect size (in DEG analysis is usually conferred as fold change), sample size, statistical power (True positive rate, not committing false negative errors), and significance (alpha as the threshold of committing false positive errors, in DEG analysis is usually conferred as False Discovery rate) is also essential in planning single-cell RNA-seq studies to get ideal power in different hypothesis testing such as to detect differentially expressed genes. In hindsight, although not recommended, as some case-control studies are confined by the cohort size at hand, sample size is not a plannable factor to achieve the desired power, the power analysis could be utilized to interpret and justify the testing results. By

simulating the gene count data given a certain effect size and sample size and ground-truth "true positive rate", at a significance level, from the statistical testing on the simulated data, one could compare the testing results to the provided ground truth to get the power of the postulated study design.

Why single nuclei

The confounded composition of cell types from sequencing data has long been a bottleneck for the detection of precise cellular and molecular targets associated with phenotypes and diseases. Tissues recorded in the human protein atlas are composed on average of over 10 cell types[28]. In case-control studies, it's rather important to distinguish disease-affected cells and healthy functional cells or even cells such as immune cells that ingeniously constrain the disease progression. Quantifying features barcoded with the source cells or nuclei can help with this separation. Compared to single-cell sequencing, the advantages of single-nuclei sequencing are that it does not require the preservation of cellular integrity during sample preparation, especially dissociation. Therefore, it's especially good for case-control studies that utilize the samples from biobank. Anyway, if one would collect fresh samples, why wouldn't he design a randomized trial?

Barcode and UMI-based data structure

The intrisinc difference compared to bulks equnecing data is the biological samples contains cells (observations) that sampled from different populations, while in bulk-seq the data from biological samples in a group are supposed to be sampled from a population.

While the scRNA-seq revolution starts with single-cell cDNA amplification and sequencing on SOLiD platform and there are currently a variety of single-cell isolation, library construction, and sequencing technologies available and under developing[2], the focus in this review is only on the workflow of 10x Chromium for 3' whole transcriptome gene expression profiling. For reviews on sequencing technologies, one can refer to.

10x Chromium adopts the microfluidics technology for single-cell isolation[29] and barcode primer bead in the droplets into the microfluidics system to generate the sequencing libraries. To control for the amplification duplicates, UMI is incorporated into each read in the initial library. The library is compatible with NGS short-read sequencing on Illumina sequencers. Then the readout of the NGS sequencer is a file in FASTQ format recording the nucleotide sequence of the reads in the library and its quality score[30]. If the reads are sequenced by Illumina sequencer, usually one can find a unique instrument name from Illumina software in the read name. The reads are then mapped to the corresponding reference genome. Aided by an annotated file, the reads mapped confidentially to transcriptome and uniquely to a gene are identified and the UMIs on them are to be counted to get the UMI count matric for downstream analysis. The usage of the mappers and barcode correction were summarized in the review.

Curse of dimensionality and non-biological zeros in scRNA-seq data

The sample space of scRNA-seq data without practical assumptions is a hypercube in n-dimensions as in $(\mathbb{R}^+)^n$, with each mRNA or gene species a dimension, leading to $n > 10{,}000$. In general, this high dimensionality will lead to curses in many regards that affects the fundamental analysis of scRNA-seq data. The essence is sourced from two facets, one is in the "vastness" of high-dimensional sample space, as the volume of the hypercube increases exponentially with the dimension $V_{hypercube} \sim l^n$. Another is the accumulation of the noises. These will lead to problems illustrated in Box 1.

Box 1

- The sample scarcity: Usually it's impossible in the medical field to collect an exponentially increased number of samples. It's futile or impossible to expect real-world samples to cover the sample space.
- Useless distance metrics in high-dimensional Euclidian space (**Fig. 2**), which lead to
  - loss of closeness: The closeness, if we defined as the samples within a distance of r with an upper limit, geometrically circumcise a subspace in $(\mathbb{R}^+)^n$ as a hypersphere with a radius r.

Mathematically the volume of such a space is the integral of area defined within $\sqrt[2]{\Sigma_{i=1}^{n} x_i^2} \leq r$ (Eq1), which gives $V_n(R) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} R^n \sim \frac{1}{\sqrt{n\pi}} \left(\frac{2\pi e}{n}\right)^{\frac{n}{2}} R^n$, $\lim_{n\to\infty} V_n(R) = 0$. Colloquially, as in Eq1, $\lim_{n\to\infty} x_i = 0$, which means the points on the periphery of the hypersphere are approaching the origin. This means no points are in the hypersphere, i.e., "close" to each other.

- little difference in distances between samples: Colloquially, there are too many pairs of many combinations of features (samples) at the same distance. Geometrically most samples are located close to the edges of the hypercube.

This affects the fundamental measurement of the sample difference in case-control studies and clustering based on distances.

- Inconsistent variance metric by eigenvalues and eigenvectors:
  - The accumulation of noises and the possibility that the noises on one dimension jeopardize the whole data structure.
- Hughes phenomenon in machine learning[31]: with a fixed number of training samples, the expected predictive power of a classifier or regressor deteriorates beyond a certain number of dimensions. This is a phenomenon sourced from the sample scarcity.

The accumulation of small noises in high-dimensional statistics will cause problems in sparsity in sampling (loss of closeness), making it hard to collect enough samples which encompass sufficient feature space; useless distance metrics (because the distance to origin is large?, one dimension contribute a small part) explanation:: too many possibilities (observation space) that has the same distance, i.e., too many combination of feature values that have the same distance, then we need much more samples to sample enough different distant space. With the same amount of observations, the clustering will fail since most of the data are in relatively similar distances. which renders many basic jobs such as PCA (variance does not converge to true variances), nearest neighbor search not trivial, and the important clustering to capture the observations (barcodes) coming from the biologically similar cell types, and hampers the very fundamental of comparing cases to controls is to measure the distance between samples (inconsistency of statistics); Specifically for PCA, which is a fundamental step to select features for most downstream analysis, suffers from the large variance of noises especially when the detection rate is low; to extract the features contributing to whether sample are case or control, even with the simplest linear model, the possible parameter values grow exponentially as the number of features grows and the false space increases as each of the features bearing noises and certain features will affect the global models; As more machine learning methods are used to analyze scRNA-seq data, the observations needed to cover the high dimensional space to train a model are one of the bottlenecks and notoriously known as "Hughes phenomenon".

For the first problem, dozens of methods were developed to impute data to overcome the sparse sampling. MAGIC, RECODE.

The straightforward question is scRNA-seq case-control study is a data mining problem to find the features or combinations of features that are associated with the interesting observations of cases.

Solution: Data representation as manifolds

The real-world scRNA-seq sample space is rarely on n-dimension, And the patterns or information in the data that constrain the data into a lower-dimensional subspace are with significance to specialize scRNA-seq data among the high dimensional data in other fields such as image processing and to be learned as biological insights.

Dropout: Mention but do not emphasis since UMI-based is not very prominent in this, the drop-out rates vary among the cells, depending on the quality of a particular library, cell type, or RNA-seq protocol

Firstly, the RNA-seq count data across samples fit certain statistical distributions. The process of sequencing a cDNA library to a depth of N can be likened to repeatedly sampling N times, with each RNA species drawn with a probability (p) proportional to the abundance of that RNA species in the library. These events, involving the successful sampling of cDNA sequences from a particular RNA species, are independent integer events. When N is sufficiently large, this distribution converges towards a Poisson distribution. However, it is often more practical to assume that the variance and

mean are independent random variables. Consequently, the negative binomial distribution (NB) is frequently employed to model scRNA-seq data. This distribution accounts for the probability of not successfully sampling a certain RNA species N-n times before achieving n successful samplings (where n is the entry in the count matrix). Evaluating the goodness-of-fit of these distributions reveals the applicability of these distributions. In 12 out of 18 examined 10X Genomics datasets, NB distribution outperformed xx, xx, and xx[32]. However, in certain UMI-based datasets, a Poisson distribution may provide a good fit[33]. It is noteworthy that single-cell data often exhibit zero inflation[34], whereas UMI-based quantification typically lacks this feature[35]. In some cases, the zero-inflated negative binomial model has been found to improve fitness by only 1.6%[32]. Nevertheless, it is prudent for researchers to meticulously examine the distribution of gene expression values in their specific datasets, taking into account the presence of dropout events, before selecting a distribution to model their data and construct background controls. Additionally, when choosing an appropriate normalization method, one must consider the raw distribution, ensuring that the chosen method aligns with the assumptions necessary for downstream analysis.

Second, the information between these genes is redundant, in that the gene expression events and data counts are not usually independent (e.g., historically we know for decades that certain gene expressions have high correlation[36] or mutual information[37]). While it's unrealistic to expect real co-linearity and the data lie on a lower dimensional plane (i.e., remove one dimension), the single-cell RNA-seq data falls in the manifold hypothesis, in which these high-dimensional spaces are generated from low-dimensional manifolds. The manifold still needs the full dimensions to characterize but locally can be characterized in lower dimensions. Driven by this hypothesis, we could alleviate the curse of dimensionality by reducing dimensions while preserving global or local structures. To preserve the global geometric structure, methods such as principal component analysis, isometric mapping, diffusion maps, and PHATE were often used. To represent the data on subspaces while preserving the local structure, methods such as locally linear embedding, Laplacian eigenmaps, and stochastic neighbor embedding were often used. The manifolds already being applied are summarized in Table 2. The challenges are to preserve the information to be indistinguishable from the original information while removing all redundancies and noises, so that's where autoencoders and GANs can come in. To measure the properties of these manifolds, metrics such as scaler curvaton, metric tensor, geodesics, and homology were utilized. Established on these manifolds, methods to learn biological insights were developed, such as batch correction and network comparing algorithms.

Solution: Data Imputation

General signals and noises

The intrinsic signatures of snRNA-seq data compared to tensors generated from other fields such as imaging processing in the signal and noise definition and data distribution in terms of specific questions being asked. The noises are not always Gaussian with low amplitudes. Technically, the general noises underlying most questions are from amplification bias, cell cycle effects, library size differences, and low RNA capture rate.

Strategies streaming the scRNA-seq for case-control studies (~3500 words, one fig)

Data selection and denoising

A typical scRNA-seq experiment generates over xx of data per sample. However, the high throughput droplet-based scRNA-seq technologies are particularly sparse with high dropout rates for features (unique molecules, i.e., genes) and empty observations composed of ambient RNAs or doublets. The data are in general confounded with other technical noises from amplification bias, cell cycle effects, library size differences, and low RNA capture rate.

To distinguish the background noisy observations, the general idea is to nominate or simulate noisy observations and call the observations significantly deviate from them genuine cells or nuclei. The general idea is to nominate background observations from the data or simulate background observations learned from the data. Then the observations significantly deviate from them and are called genuine cells/nuclei. In DropletUtils, the ambient RNA pool is estimated with the gene count

across barcodes with total counts less than 100. In Cell Ranger, the expected number of recovered cells is estimated by the OrdMag algorithm, and then the barcodes with counts less than the lowest in the expected recovered cells were nominated as background observations. In both tools, the significant deviation from the estimated background observations, i.e., the genuine cells, were tested using a Dirichlet-multinomial model of UMI count sampling and using a knee point filter to ensure that barcodes with large total counts are always retained. Among the deep learning models, deep generative models are adopted to generate the background-contaminated counts. The denoising autoencoders with the loss function match the biological "true" data distribution such as output denoising ZINB distributed features were applied to retain the essential biological features but drop the noises.

Compare gene expression between groups and associate gene expression to disease phenotypes.

The intrinsic difference here is that in bulk sequencing, the samples in a group are independent samples from a population distribution while the scRNA-seq data from each sample are from multiple populations (cell types). The natural way then is to calculate the cluster-based pseudo-bulk expression matrices, then in each cluster, the data structure becomes the same as the bulk RNA-seq data. After standard normalization, the differential expression test using regression models as known in edgeR[38], DESeq2[16] , limma[39] can be applied (**Fig. 2**). Other than the gene expression values, the factors that are divergent between groups and influence the estimated variance and coefficients of gene expressions are often used as covariates. This usually includes age, sex, day of the experiment, percentages of cells in the cell-cluster or cell-type in the total of cells and, GEM batch (defined by nuclear isolation, GEM generation, and barcoding performed for all samples in one batch in the same 10X GEM generation run, v2 or v3 chemistry). Another strategy is to compare the expression in individual cells, using statistical tests such as t-test, Wilcoxon rank-sum test, logistic regression[40], negative binomial and Poisson generalized linear models, likelihood ratio test[41], and the two-part hurdle model implemented by MAST[17]. However, in a benchmark study used the differential expression (DE) results from eighteen bulk RNA-seq datasets matching cell populations as gold standard, the pseudo bulk methods outperformed such single-cell methods measured by the concordance between DE results in bulk versus scRNA-seq datasets[42].

Deciphering the mechanism by associate gene expression variations to genetic or epigenetic variations

Outlook of -omics in deciphering disease mechanisms in case-control studies

Medical and physiological concerns

Spatial

Disease circuits

Acknowledge

Reference

1. Levin, K.A. Study design I. *Evid Based Dent* **6**, 78-79 (2005).
2. Tang, F.C., *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-U386 (2009).
3. Waddington, C. The Strategy of the Genes. (George Allen & Unwin, 1957).
4. Wergeland, H. Simple Proof of the Ergodic Theorem. *Acta Chem Scand* **12**, 1117-1123 (1958).

5. Trapnell, C., *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-U251 (2014).
6. La Manno, G., *et al.* RNA velocity of single cells. *Nature* **560**, 494-+ (2018).
7. Xu, Z., Sziraki, A., Lee, J., Zhou, W. & Cao, J. PerturbSci-Kinetics: Dissecting key regulators of transcriptome kinetics through scalable single-cell RNA profiling of pooled CRISPR screens. *bioRxiv* (2023).
8. Regev, A., *et al.* The Human Cell Atlas. *Elife* **6**(2017).
9. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat Protoc* **13**, 2742-2757 (2018).
10. Page, M.J., *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews (74, pg 790, 2021). *Rev Esp Cardiol* **75**, 192-192 (2022).
11. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-U206 (2015).
12. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**(2018).
13. Chen, H.Y., *et al.* PlantscRNAdb: A database for plant single-cell RNA analysis. *Mol Plant* **14**, 855-857 (2021).
14. Zeng, J.Y., *et al.* CancerSCEM: a database of single-cell expression map across various human cancers. *Nucleic Acids Res* **50**, D1147-D1155 (2022).
15. Lee, J.T.H., Patikas, N., Kiselev, V.Y. & Hemberg, M. Fast searches of large collections of single-cell data using scfind. *Nat Methods* **18**, 262-+ (2021).
16. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
17. Finak, G., *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015).
18. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**, 740-U184 (2014).
19. Boyeau, P., *et al.* An empirical Bayes method for differential expression analysis of single cells with deep generative models. *P Natl Acad Sci USA* **120**(2023).
20. Zhang, M.Q., *et al.* IDEAS: individual level differential expression analysis for single-cell RNA-seq data. *Genome Biology* **23**(2022).
21. Zhu, B.Q., Li, H.Y., Zhang, L., Chandra, S.S. & Zhao, H.Y. A Markov random field model-based approach for differentially expressed gene detection from single-cell RNA-seq data. *Brief Bioinform* **23**(2022).
22. Tsuyuzaki, K., Sato, H., Sato, K. & Nikaido, I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology* **21**(2020).
23. Elyanow, R., Dumitrascu, B., Engelhardt, B.E. & Raphael, B.J. netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res* **30**, 195-204 (2020).
24. Sompairac, N., *et al.* Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *Int J Mol Sci* **20**(2019).
25. Ahlmann-Eltze, C. & Huber, W. Analysis of multi-condition single-cell data with latent embedding multivariate regression. *bioRxiv*, 2023.2003.2006.531268 (2023).
26. Fleming, S.J., *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat Methods* **20**, 1323-1335 (2023).

27. Rindler, K., *et al.* Single-cell RNA sequencing analysis of a COVID-19-associated maculopapular rash in a patient with psoriasis treated with ustekinumab. *J Dermatol* **50**, 1052-1057 (2023).
28. Alam, O. A single-cell-type transcriptomics map of human tissues. *Nat Genet* **53**, 1275-1275 (2021).
29. Marcus, J.S., Anderson, W.F. & Quake, S.R. Microfluidic single-cell mRNA isolation and analysis. *Anal Chem* **78**, 3084-3089 (2006).
30. Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. & Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767-1771 (2010).
31. Trunk, G.V. A problem of dimensionality: a simple example. *IEEE Trans Pattern Anal Mach Intell* **1**, 306-307 (1979).
32. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486-3488 (2017).
33. Ziegenhain, C., *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631-+ (2017).
34. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications* **11**(2020).
35. Svensson, V. Reply to: UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nat Biotechnol* **39**, 160 (2021).
36. Butte, A.J. & Kohane, I.S. Unsupervised knowledge discovery in medical databases using relevance networks. *J Am Med Inform Assn*, 711-715 (1999).
37. Butte, A.J. & Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 418-429 (2000).
38. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
39. Ritchie, M.E., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
40. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat Methods* **16**, 163-166 (2019).
41. McDavid, A., *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461-467 (2013).
42. Squair, J.W., *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).