ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

Q1Q4 Methods and challenges in the analysis of single-cell RNA-sequencing data

Q3 Pablo G. Camara

Abstract

The recent advent of highly parallelizable single-cell RNA-sequencing technologies has opened a new window into the study of cell differentiation, commitment, and diversity. Rapid advances in the development of these technologies are being accompanied by the design of computational methods tailored to address the challenges presented by the analysis of single-cell RNA-sequencing data. This review provides a concise overview of some of the steps, algorithms, and approaches that are currently used in the analysis of single-cell RNA-sequencing data, with an emphasis on recent developments.

Addresses

Department of Genetics, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Corresponding author: Camara, Pablo G (pcamara@upenn.edu)

Current Opinion in Systems Biology 2017, ■:1–7

This review comes from a themed issue on **Genomics and epigenomics**

Edited by **Raul Rabadan** and **Peter Sims**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.coisb.2017.12.007

2452-3100/© 2017 Published by Elsevier Ltd.

Introduction

Highly parallelizable single-cell RNA-sequencing (scRNA-seq) [1–4] has emerged in the past decade as a new tool for simultaneously accessing the transcriptome of thousands of individual cells, and is becoming a standard element in the toolbox of molecular biologists. The new-born field of single-cell transcriptomics (or single-cell genomics, in general) is evolving rapidly [5]. Advances in the field have affected not only the cost of single-cell library preparation, which has been largely reduced by the development of microfluidic approaches [6–10], but also the capability to perform measurements in methanol-fixed [11,12] or fresh-frozen samples [13–15], or to concurrently quantify mRNA and protein levels of individual cells [16,17]. This massive technological progress has been accompanied by the development of numerous algorithms for analyzing scRNA-seq data. These analytic tools are designed to overcome key challenges that originate from the substantial technical variability and sparseness inherent in scRNA-seq data. Despite these challenges, the field is now entering a mature stage and standard analytic pipelines are beginning to emerge, some of which are regularly updated and maintained (e.g. Refs. [7,18]).

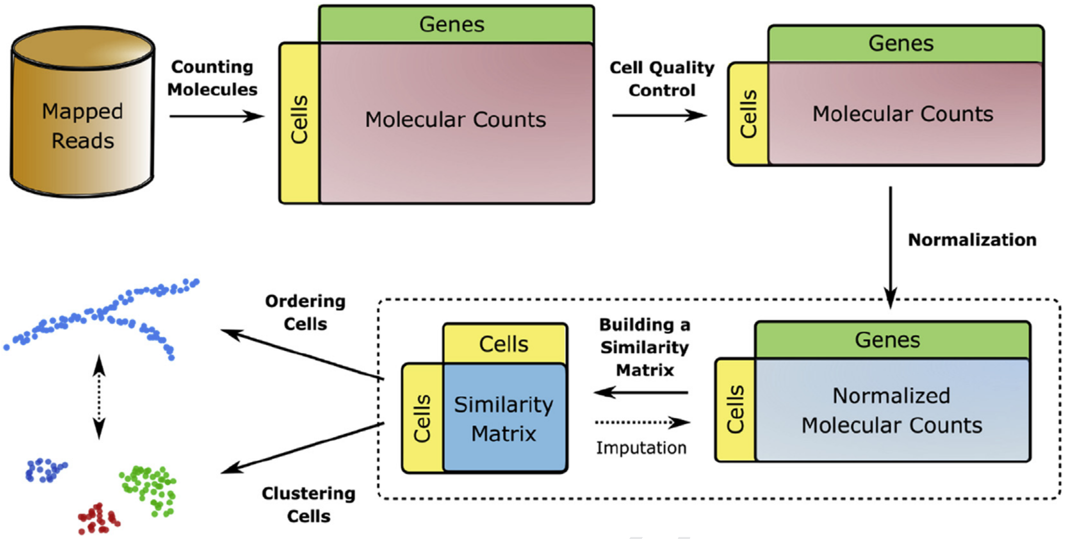
This review provides a concise overview of some of the analytic approaches and techniques that are used in scRNA-seq, with an emphasis on the developments that have occurred over the past two years. The goal here is not to comprehensively review all existing algorithms (at least 139 tools specifically designed for analyzing scRNA-seq data are currently available (http://www.scrna-tools.org/)), but to present the main steps in a typical, generic pipeline used for scRNA-seq data analysis (summarized in Figure 1), the unique features and difficulties associated with each step, and some of the current algorithms used to address those difficulties. We do not discuss the design of scRNA-seq experiments, the existing methods for single-cell isolation and amplification, or the mapping and filtering of raw reads. For a discussion of these topics, or for a broader view of the topics addressed here, we refer the reader to other review articles [19–23].

Counting molecules

In scRNA-seq experiments, the aim is to obtain estimates of the number of mRNA molecules of each gene in each sequenced cell. The number of mapped reads is a random variable that depends in an unknown manner on the number of mRNA molecules present in the cell. This dependence is heavily affected by the extent of cDNA amplification, the sequencing depth, and the efficiency of the capture and reverse-transcription reactions. Thus, differences among cells in these quantities can largely obscure the relationship between the number of reads and the molecular counts of a gene. The use of external spike-in RNA of known concentration [24] and/or unique molecular identifiers (UMIs) [25] allows researchers to control experimentally for differences in cDNA amplification, which enhances the accuracy of the estimates of the number of captured and reverse-transcribed mRNA molecules of each gene.

The use of UMIs substantially reduces the amount of technical variability in scRNA-seq experiments. In a recent study [26], the amount of technical variability was estimated to be reduced by approximately 50% when expression levels were measured using UMIs instead of reads. Several available tools systematically and accurately estimate molecular counts in UMI-based scRNA-seq experiments [27–29]. These tools account

Figure 1



The structure of a generic pipeline for the analysis of scRNA-seq data. Schematic summarizing the basic steps in a generic scRNA-seq data analysis pipeline starting with the mapped reads. Each step corresponds to a different section of this review.

for sequence errors introduced in the UMIs during library preparation or sequencing, which can lead to overestimation of the molecular counts. Whereas the simplest tools collapse UMI sequences assigned to the same gene at a low edit distance from each other [27,28], more advanced methods, like dropEst [29], use a Bayesian approach to estimate the probability of a UMI being erroneous based on several measured quantities. dropEst also implements a bootstrap algorithm to account for the probability of UMI collisions (a same UMI being assigned to two separate mRNA molecules of the same gene). Although the number of possible UMI sequences in an scRNA-seq experiment is generally large, the probability of UMI collisions can be unexpectedly large due to highly skewed UMI frequency distributions [29].

In experiments where UMIs are not available, such as those conducted using non-barcoded scRNA-seq protocols, the addition of external spike-in RNA of known concentration is an alternative approach to adjust for differences in the extent of cDNA amplification [24,30]. However, when neither spike-in RNA controls nor UMIs are experimentally available, inferring molecular counts becomes particularly challenging. The algorithm Census [31] assumes that all detectably expressed genes measured below the mode of the log-transformed read distribution in a cell should be present at the level of ~1 cDNA copy. This assumption is particularly justified in scRNA-seq protocols that produce full-length cDNAs [1–3]. Although this approach cannot control for nonlinear cDNA amplification biases during library preparation, the use of Census counts in experiments where UMIs and spike-in RNAs are unavailable leads to marked improvements with respect to read counts [31].

Cell quality control

Low-quality cells must be removed to avoid artifacts in the analysis of scRNA-seq data. The systematic identification of low-quality cells has been recently studied by Ilicic et al. [32]. By using microscopy data from a Fluidigm C1 platform, from which broken cells, doublets, and empty wells can be visually identified, Ilicic et al. trained a support vector machine to classify cells into low- and high-quality cells based on the scRNA-seq data. Among a set of 20 biological and technical features derived from the data, the percentage of mapped reads, the expression of mitochondrially encoded genes (whose RNA is retained even when leakage of cytoplasmic RNA occurs, as in the case of broken cells), and the correlation with the mean expression profile were found to be powerful predictors of cell quality [32]. However, the application of this approach is limited by the unavailability of microscopy data to train the classifier in most scRNA-seq experiments, and application to datasets involving cell types other than the one used in the training set leads to poor predictions. To overcome these limitations, scPipe [28] implements an unsupervised outlier-based method for classifying low-quality cells by using some of the aforementioned biological and technical features as input. Alternatively, training sets containing errors can be constructed from the features mentioned above [29]. In all cases, extra care must be exercised when using these methods in the case of heterogeneous cellular populations, because genuine biological differences can produce substantial changes in the output of these classifiers.

Normalization

Identifying suitable normalization schemes for scRNA-seq data is a complex problem, and a part of the recent literature on scRNA-seq has centered around this topic [21,33]. The capture efficiency of scRNA-seq experiments (i.e., the fraction of mRNA molecules in the cells that is reverse transcribed, amplified, and sequenced) is typically in the 10%—30% range, depending on the experimental protocol and cell type. Such low efficiencies lead to sparse, zero-inflated data. Furthermore, the readout of each gene is affected by factors as diverse as the total amount of mRNA that was originally present in the cell (and which depends on the cell type, cell cycle stage, etc.), stochastic expression effects such as transcriptional bursting [34,35], and differences in the capture, reverse transcription, and sequencing depth across cells. These effects (and presumably others that remain undocumented) should be considered when comparing the expression profiles of any two cells. Unlike in the case of population-level RNA-seq experiments, where replicates can be readily established, obtaining genuine replicates in scRNA-seq experiments is challenging, because each cell is unique and typically can be measured only once. Therefore, to account for the aforementioned effects, assumptions must be made regarding the expression profiles of the cells. For instance, one common assumption is that most genes are not differentially expressed among cells, and thus for most genes, the observed variability is expected to be derived from the effects mentioned above and not from genuine differential expression. These requirements make it particularly hard to identify adequate normalization schemes in experiments that involve highly heterogeneous cell populations, such as in longitudinal cell-differentiation experiments, where these assumptions are easily violated.

Normalizing the molecular counts and comparing the expression profile of cells are therefore deeply interrelated problems that, optimally, should be addressed jointly. Although we describe a few algorithms developed for this purpose in the next section, it is a widespread approach (motivated by its scalability to large datasets) to address the problem of normalization and that of comparing expression profiles (or building a similarity matrix) separately [33]. To this end, scaling factors are used to regress out some of the undesirable dependences discussed above. In the simplest approach, referred to as library-size normalization, the molecular counts of each gene are normalized by the total number of counts in the cell. This approach compensates for differences in cell size and sequencing depth. However, problems arise when disparities in the expression of a set of genes lead to differences in the total amount of mRNA in the cell [36]. In such scenarios, library-size normalization leads to spurious effects, and genes that are expressed at the same level might appear to be differentially expressed when comparing the expression profiles of two cells. Although relatively more advanced normalization schemes exist that account for this effect, such as DESeq [36] and TMM [37], none of these approaches consider differences in the capture and reverse-transcription efficiency among cells or other systematic effects present in scRNA-seq data.

Several normalization schemes specifically tailored to scRNA-seq data have been reported recently [38—40]. To lessen the effect of missing data in the molecular counts of individual cells, scran constructs synthetic population-level RNA-seq datasets by pooling molecular counts across sampled sets of cells [39]. These synthetic datasets are then used to compute global, population-level scaling factors, which are deconvolved to obtain the individual factors of each cell. In a case study conducted using simulated data to compare various normalization schemes, the performance of scran was substantially superior to that of other schemes that were not specifically designed for scRNA-seq [33]. However, a limitation of all between-sample normalization schemes, where a single scaling factor is applied to each cell, is that they cannot account for gene-specific biases. Accordingly, scRNA-seq experiments were recently reported to exhibit systematic variation in the relationship between gene-specific expression and sequencing depth, which cannot be accounted for using a single scaling factor [38]. To alleviate this problem, the algorithm SCnorm regresses out differences in the sequencing depth while accounting for potential differences across groups of genes [38]. Therefore, the use of a combination of tools such as scran and SCnorm appears to be a favorable procedure for normalizing the data in several scRNA-seq experiments.

Building a similarity matrix

Normalization schemes based on scaling factors adjust for the technical variability that originates from differences in library size and sequencing depth. Nevertheless, the effect of missing data and zero-inflation must be also considered when comparing the expression profiles of cells, since zeroes are not affected by scaling factors. The problem of building a similarity matrix is to a large extent an imputation and/or feature-selection problem in the gene space. Therefore, the simplest approach to building a similarity matrix involves computing correlation or Euclidean pairwise distances by using only highly expressed and variable genes. Because the fraction of missing information in the case of highly expressed genes is smaller than that in the case of other genes [26], highly expressed genes can more accurately account for genuine biological differences among cells. The number of genes used can be determined by modeling the probability of dropout events as a logistic function of the transcript abundance [41]. In more complex but related approaches, principal component analysis is performed in gene space and a similarity matrix is defined based on considering sets of principal components instead of genes [42]. The number of principal components can be determined from random matrix theory arguments [43]. Alternatively, a similarity matrix can also be built using sets of annotated genes (e.g., gene ontology categories) for which the amount of variance explained by the first principal component is significantly higher than expected [44]. All these methods rely on feature selection in the gene space.

In a second type of approach, kernel-smoothing methods are used to lessen the effects of dropouts and impute some of the missing data. For instance, the algorithm MAGIC assumes that the underlying structure of the denoised dataset should be that of a manifold, and uses a diffusion-based approach to smooth the dataset into such a structure [45]. Although expression spaces more complex than manifolds can be envisioned (e.g., stratified spaces where the intrinsic dimension changes across the space), MAGIC has shown excellent results in recovering gene—gene interactions in several biological settings, which suggests that the manifold assumption is a favorable approximation in several scenarios [45]. Kernel-based approaches can also use additional input from the user to constrain the similarity matrix and impute missing data. For instance, SIMLR uses an optimization framework that combines multiple kernels in such a manner that the resulting similarity matrix features an approximate block-diagonal structure with rank equal to the expected number of cell types (which is a parameter of the algorithm) [46]. Because imputing missing data and learning a similarity matrix are highly interrelated problems, all kernel-based algorithms must make assumptions based on certain preconceived notions of similarity before data imputation.

A third strategy employed for learning a similarity matrix involves using statistical zero-inflation or hurdle models. In these approaches, a certain degree of homogeneity across the cells must be assumed to fit the underlying model. The algorithm ZIFA implements a latent model based on factor analysis augmented to accommodate zero-inflation [47]. Similarly, SCDE uses a two-component mixture model to estimate the sequencing depth, drop-out rate, and amplification noise in each cell [48]. MAST implements a hurdle model in which the probability of detection of each gene is modeled using logistic regression [49]. In cases where exogenous spike-in RNA is available, BASiCS can be also used to jointly estimate cell-specific scaling factors and technical variability [40]. Fitting these models to the data is a computationally intensive task. Therefore, the use of these normalization approaches is typically constrained to small and relatively homogeneous datasets.

Clustering cells

Once a similarity matrix that is robust against technical variability has been constructed, standard clustering methods can be applied to group cells according to their expression profile and identify cell types. However, because of the high degree of technical variability present in scRNA-seq data, the approaches described in the preceding section for building a similarity matrix are, in practice, not optimal; thus, several clustering algorithms specifically tailored to scRNA-seq data have been developed [7,42,50—55]. Some of the most recent strategies involve combining clustering outcomes across the space of parameters [53], leveraging the use of reference population-level transcriptomes [54], or using inferred gene regulatory networks by linking *cis*-regulatory sequences to single-cell transcript abundances [55].

With standard clustering methods, a general drawback is that their capacity to identify rare cell types is limited. Because populations of rare cells typically contribute little to the total gene variability of the dataset, in the clustering process, the weights of the genes that are specific to rare cell populations are small. Two algorithms have been specifically designed to identify rare cell populations by using scRNA-seq data: RaceID leverages the use of a multistep k-medoids clustering, where outlier cells are identified and clustered apart [56,57]. Similarly, GiniClust builds upon the Gini index, originally developed to study social inequality [58], to identify genes that are specifically expressed by groups of rare cells [59].

Once cell types have been identified, differential-expression and gene-set enrichment analyses can be performed to understand the biological basis of differences among cell types. Accounting for missing data is particularly when performing these analyses, and several algorithms have been specifically developed for the purpose [44,48,49,60].

Ordering cells

The transcriptional profiles of cells frequently do not naturally form clusters but mostly continuous structures. For instance, in developmental progression, multipotent cells differentiate into one or more cell types in a continuous or semi-continuous fashion, often evolving throughout several transient states. Similarly, in immune and inflammatory responses, immune cells exhibit substantial plasticity and acquire a broad continuum of activation states, ranging from proinflammatory to immunosuppressive states. In these and other scenarios, the use of clustering methods is not an optimal approach for dissecting the continuous transcriptional structure of the system under study.

Several computational approaches attempt to order cells according to their transcriptional profile [18,61—68].

Most of these methods have been designed for the analysis of developmental trajectories, and include techniques for pseudo-time inference and branch-point identification. These approaches have yielded highly accurate results in multiple studies on cell differentiation and commitment. However, with these methods, tree-like differentiation structures featuring a unique origin (a single stem cell type) are frequently assumed and thus cannot capture some of the most complex relationships of dynamic cellular systems, such as the presence of independent lineages, convergent lineages, or the coupling of cell cycle. Although additional research is required to overcome these limitations, several available frameworks appear promising because they present the ability to construct low-dimensional representations that preserve the local relationships of the expression space in an unbiassed manner. Diffusion maps [69] provide a natural low-dimensional embedding for the multiscale geometric organization of high-dimensional data. Destiny [61], Diffusion Pseudotime [62], and Wishbone [63] build upon diffusion maps to respectively identify combinations of genes associated with the dynamics of a cellular process, infer a pseudo-temporal ordering, and identify branch points in developmental trajectories. Similarly, scTDA [67] builds upon low-dimensional topological representations [70] that preserve local relationships of the expression space to identify and characterize differentially expressed genes without the requirement of predefining cellular populations. These genes are then used to dissect the transcriptional events that underlie the dynamic cellular process without assuming any specific structure. A most recent algorithm, WADDINGTON-OT [68], adapts the mathematics of optimal transport to the analysis of scRNA-seq data of dynamic cellular systems, permitting the recovery of complex transcriptional relationships in an unbiassed fashion. All these frameworks can be readily used for unbiased dissection of the transcriptional events underlying complex dynamic systems.

Conclusions

Here, we have reviewed some of the challenges associated with the main steps in the analysis of scRNA-seq data and some of the recent computational developments to address these challenges. In the past few years, powerful algorithms for analyzing scRNA-seq data have emerged. However, scRNA-seq is a rapidly evolving field and will certainly continue to undergo further technological development in the coming years, which will pose new analytic challenges. As single-cell technologies continue to develop, computational tools for the integration of distinct forms of single-cell data will be increasingly required. In this regard, algorithms for integrating scRNA-seq data with imaging data [71,72] and T-cell-receptor sequence data [73], and for integrating multiple scRNA-seq datasets [74], are already becoming available, and we expect these tools to play a central role in this field in the future. Algorithms for the inference of large copy number variants using scRNA-seq [75–78] will progressively lead to new perspectives in the study of intra-tumor heterogeneity. The development of suitable simulation platforms [79], benchmarking schemes, and gold-standard datasets will also foster the calibration of existing tools for analyzing scRNA-seq data. These and other developments will undoubtedly affect ongoing large-scale scRNA-seq efforts, such as the Human Cell Atlas [80], and contribute toward extending the boundaries of cell biology.

Acknowledgments

I am grateful to L. Aparicio, A. Iavarone, E. Kandror, A. Lasorella, A. Levine, B. Li, T. Maniatis, P. van Nieuwenhuizen, R. Rabadán, A. Rizvi, T. Roberts, M. Shen, M. Shibata, and P. Sims for collaboration in related projects to the topic of this review, and to Hao Wu for comments on an early version of the manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, *et al.*: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Method* 2009, **6**:377–382.

2. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R: **Smart-seq2 for sensitive full-length transcriptome profiling in single cells**. *Nat Method* 2013, **10**: 1096–1098.

3. Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, *et al.*: **Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells**. *Nat Biotechnol* 2012, **30**: 777–782.

4. Hashimshony T, Wagner F, Sher N, Yanai I: **CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification**. *Cell Rep* 2012, **2**:666–673.

5. Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ: **Single cells make big data: new challenges and opportunities in transcriptomics**. *Current Opinion in Systems Biology* 2017, **4**: 85–91.

6. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW: **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells**. *Cell* 2015, **161**:1187–1201.

7. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, *et al.*: **Highly parallel Genome-wide expression profiling of individual cells using nanoliter droplets**. *Cell* 2015, **161**: 1202–1214.

8. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, *et al.*: **Massively parallel digital transcriptional profiling of single cells**. *Nat Commun* 2017, **8**:14049.

9. Bose S, Wan Z, Carr A, Rizvi AH, Vieira G, Pe'er D, Sims PA: **Scalable microfluidics for single-cell RNA printing and sequencing**. *Genome Biol* 2015, **16**:120.

10. Yuan J, Sims PA: **An automated microwell platform for large-scale single cell RNA-seq**. *Sci Rep* 2016, **6**:33883.

11. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, *et al.*: **Comprehensive single-cell transcriptional profiling of a multicellular organism**. *Science* 2017, **357**:661–667.

12. Alles J, Karaiskos N, Praktiknjo SD, Grosswendt S, Wahle P, Ruffault PL, Ayoub S, Schreyer L, Boltengagen A, Birchmeier C, *et al.*: **Cell fixation and preservation for droplet-based single-cell transcriptomics**. *BMC Biol* 2017, **15**:44.

13. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, Wildberg A, Gao D, Fung HL, Chen S, *et al.*: **Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain**. *Science* 2016, **352**:1586–1590.

14. Hu P, Fabyanic E, Zhou Z, Wu H: *sNucDrop-Seq: dissecting cell-type composition and neuronal activity state in mammalian brains by massively parallel single-nucleus RNA-Seq*. bioRxiv. 2017: 154476.

15. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, Choudhury SR, Aguet F, Gelfand E, Ardlie K, *et al.*: **Massively parallel single-nucleus RNA-seq with DroNc-seq**. *Nat Method* 2017. Q2

16. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA: **Multiplexed quantification of proteins and transcripts in single cells**. *Nat Biotechnol* 2017.

17. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P: **Simultaneous epitope and transcriptome measurement in single cells**. *Nat Method* 2017, **14**:865–868.

18. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL: **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells**. *Nat Biotechnol* 2014, **32**: 381–386.

19. Saliba AE, Westermann AJ, Gorski SA, Vogel J: **Single-cell RNA-seq: advances and future challenges**. *Nucleic Acids Res* 2014, **42**:8845–8860.

20. Grun D, van Oudenaarden A: **Design and analysis of single-cell sequencing experiments**. *Cell* 2015, **163**:799–810.

21. Stegle O, Teichmann SA, Marioni JC: **Computational and analytical challenges in single-cell transcriptomics**. *Nat Rev Genet* 2015, **16**:133–145.

22. Bacher R, Kendziorski C: **Design and computational analysis of single-cell RNA-sequencing experiments**. *Genome Biol* 2016, **17**:63.

23. Wagner A, Regev A, Yosef N: **Revealing the vectors of cellular identity with single-cell genomics**. *Nat Biotechnol* 2016, **34**: 1145–1160.

24. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, *et al.*: **Accounting for technical noise in single-cell RNA-seq experiments**. *Nat Method* 2013, **10**:1093–1095.

25. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson S: **Quantitative single-cell RNA-seq with unique molecular identifiers**. *Nat Method* 2014, **11**: 163–166.

26. Grun D, Kester L, van Oudenaarden A: **Validation of noise models for single-cell transcriptomics**. *Nat Method* 2014, **11**: 637–640.

27. Smith T, Heger A, Sudbery I: **UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy**. *Genome Res* 2017, **27**:491–499.

28. Tian L, Su S, Amann-Zalcenstein D, Biben C, Naik SH, Ritchie ME: *scPipe: a flexible data preprocessing pipeline for single-cell RNA-sequencing data*. bioRxiv. 2017:175927.

29. • Petukhov V, Guo J, Baryawno N, Severe N, Scadden D, Samsonova MG, Kharchenko PV: *Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments*. bioRxiv. 2017:171496.

The authors provide a comprehensive treatment of molecular counts in UMI-based scRNA-seq experiements, identifying in detail the different sources of biases and implementing corrections for those biases.

30. Katayama S, Tohonen V, Linnarsson S, Kere J: **SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization**. *Bioinformatics* 2013, **29**: 2943–2945.

31. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C: **Single-cell mRNA quantification and differential analysis with Census**. *Nat Method* 2017, **14**:309–315.

32. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA: **Classification of low quality cells from single-cell RNA-seq data**. *Genome Biol* 2016, **17**:29.

33. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC: **Normalizing single-cell RNA sequencing data: challenges and opportunities**. *Nat Method* 2017, **14**:565–571.

34. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S: **Stochastic mRNA synthesis in mammalian cells**. *PLoS Biol* 2006, **4**:e309.

35. Kaern M, Elston TC, Blake WJ, Collins JJ: **Stochasticity in gene expression: from theories to phenotypes**. *Nat Rev Genet* 2005, **6**:451–464.

36. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**:R106.

37. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biol* 2010, **11**:R25.

38. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C: **SCnorm: robust normalization of single-cell RNA-seq data**. *Nat Method* 2017, **14**:584–586.

39. • Lun AT, Bach K, Marioni JC: **Pooling across cells to normalize single-cell RNA sequencing data with many zero counts**. *Genome Biol* 2016, **17**:75.

This paper introduces an approach to the normalization of scRNA-seq data based on the devolution of pool-based size factors. Similar to population-based RNA-seq normalization schemes, this approach is highly scalable. However, contrary to those approaches, it is rather insensitive to zero-inflation.

40. Vallejos CA, Marioni JC, Richardson S: **BASiCS: bayesian analysis of single-cell sequencing data**. *PLoS Comput Biol* 2015, **11**:e1004333.

41. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, *et al.*: **Single-cell RNA-seq reveals dynamic paracrine control of cellular variation**. *Nature* 2014, **510**:363–369.

42. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, *et al.*: **Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics**. *Cell* 2016, **166**:1308–1323. e1330.

43. Marchenko VA, Pastur LA: **Distribution of eigenvalues for some sets of random matrices**. *Matematicheskii Sbornik* 1967, **114**:507–536.

44. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan JB, Zhang K, Chun J, *et al.*: **Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis**. *Nat Method* 2016, **13**:241–244.

45. • van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D: *MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data*. BioRxiv. 2017:111591.

An imputation method for scRNA-seq data based on diffusion maps is introduced in this paper, revealing gene–gene interactions which otherwise would be hard to characterize from scRNA-seq data.

46. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S: **Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning**. *Nat Method* 2017, **14**:414–416.

47. Pierson E, Yau C: **ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis**. *Genome Biol* 2015, **16**:241.

48. Kharchenko PV, Silberstein L, Scadden DT: **Bayesian approach to single-cell differential expression analysis**. *Nat Method* 2014, **11**:740–742.

49. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, *et al.*: **MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data**. *Genome Biol* 2015, **16**: 278.

50. Zurauskiene J, Yau C: **pcaReduce: hierarchical clustering of single cell transcriptional profiles**. *BMC Bioinf* 2016, **17**:140.

51. Xu C, Su Z: **Identification of cell types from single-cell transcriptomes using a novel clustering method**. *Bioinformatics* 2015, **31**:1974–1980.

52. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, *et al.*: **Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq**. *Science* 2015, **347**:1138–1142.

53. •• Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, *et al.*: **SC3: consensus clustering of single-cell RNA-seq data**. *Nat Method* 2017, **14**:483–486.

The authors present a consensus clustering algorithm based on the similarity partitioning algorithm. This approach achieves high accuracy and robustness against technical variability in scRNA-seq experiments.

54. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, *et al.*: **Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors**. *Nat Genet* 2017, **49**:708–718.

55. Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, *et al.*: **SCENIC: single-cell regulatory network inference and clustering**. *Nat Method* 2017.

56. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A: **Single-cell messenger RNA sequencing reveals rare intestinal cell types**. *Nature* 2015, **525**:251–255.

57. Grun D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, *et al.*: **De novo prediction of stem cell identity using single-cell transcriptome data**. *Cell Stem Cell* 2016, **19**: 266–277.

58. Gini C. In *Variabilità e mutabilità. Reprinted in memorie di metodologica statistica*. Edited by Pizetti E, Salvemini T, Rome: Libreria Eredi Virgilio Veschi; 1912.

59. Jiang L, Chen H, Pinello L, Yuan GC: **GiniClust: detecting rare cell types from single-cell gene expression data with Gini index**. *Genome Biol* 2016, **17**:144.

this paper, an algorithm that adapts the Gini index to detect rare cell types using scRNA-seq data is presented. The identification of rare cell types using clustering algorithms is particularly challenging, and this is one of the few algorithms specifically designed to that end.

60. Vallejos CA, Richardson S, Marioni JC: **Beyond comparisons of means: understanding changes in gene expression at the single-cell level**. *Genome Biol* 2016, **17**:70.

61. • Angerer P, Haghverdi L, Buttner M, Theis FJ, Marr C, Buettner F: **destiny: diffusion maps for large-scale single-cell data in R**. *Bioinformatics* 2016, **32**:1241–1243.

Diffusion maps are becoming a widespread approach to reduce the dimensionality of scRNA-seq data while preserving local relationships. This is one of the first papers to adopt the use of diffussion maps for the analysis of scRNA-seq experiments.

62. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ: **Diffusion pseudotime robustly reconstructs lineage branching**. *Nat Method* 2016, **13**:845–848.

63. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D: **Wishbone identifies bifurcating developmental trajectories from single-cell data**. *Nat Biotechnol* 2016, **34**:637–645.

64. Welch JD, Hartemink AJ, Prins JF: **SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data**. *Genome Biol* 2016, **17**:106.

65. • Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H, Trapnell C: *Reversed graph embedding resolves complex single-cell developmental trajectories*. 2017:110668. bioRxiv.

66. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC: **Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape**. *Proc Natl Acad Sci USA* 2014, **111**:E5643–E5650.

67. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R: **Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development**. *Nat Biotechnol* 2017.

Topological methods constitute an approach to reduce the dimensionality of scRNA-seq data while preserving local relationships. This paper established the foundations for the application of topological methods to scRNA-seq data analysis.

68. • Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, Liu S, Lin S, Berube P, Lee L, *et al.*: *Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming*. bioRxiv. 2017:191056.

Inferring the cellular dynamics of a differentiating system using scRNA-seq snapshops is particularly challenging and there are fundamental limits to it. This paper adapts the mathematics of optimal-transport to formulate a framework for dynamical inference using longitudinal scRNA-seq data under controlled assumptions.

69. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW: **Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods**. *Proc Natl Acad Sci USA* 2005, **102**:7432–7437.

70. Singh G, Mémoli F, Carlsson GE: **Topological methods for the analysis of high dimensional data sets and 3D object recognition**. In *SPBG*. Citeseer; 2007:91–100.

71. Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC: **High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin**. *Nat Biotechnol* 2015, **33**:503–509.

72. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of single-cell gene expression data**. *Nat Biotechnol* 2015, **33**:495–502.

73. Stubbington MJT, Lonnberg T, Proserpio V, Clare S, Speak AO, Dougan G, Teichmann SA: **T cell fate and clonality inference from single-cell transcriptomes**. *Nat Method* 2016, **13**:329–332.

74. Butler A, Satija R: *Integrated analysis of single cell transcriptomic data across conditions, technologies, and species*. bioRxiv. 2017. 164889.

75. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, *et al.*: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma**. *Science* 2014, **344**:1396–1401.

76. Muller S, Liu SJ, Di Lullo E, Malatesta M, Pollen AA, Nowakowski TJ, Kohanbash G, Aghi M, Kriegstein AR, Lim DA, *et al.*: **Single-cell sequencing maps gene expression to mutational phylogenies in PDGF- and EGF-driven gliomas**. *Mol Syst Biol* 2016, **12**:889.

77. Tirosh I, Izar B, Prakadan SM, Wadsworth 2nd MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, *et al.*: **Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq**. *Science* 2016, **352**:189–196.

78. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, *et al.*: **Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma**. *Nature* 2016, **539**:309–313.

79. Zappia L, Phipson B, Oshlack A: **Splatter: simulation of single-cell RNA sequencing data**. *Genome Biol* 2017, **18**:174.

80. Regev A, Teichmann S, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M: *The human cell Atlas*. bioRxiv. 2017. 121202.