Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Single cells make big data: New challenges and opportunities in transcriptomics

Philipp Angerer[1], Lukas Simon[1], Sophie Tritschler[1], F. Alexander Wolf[1], David Fischer[1] and Fabian J. Theis[1,2]

Abstract

Recent technological advances have enabled unprecedented insight into transcriptomics at the level of single cells. Single cell transcriptomics enables the measurement of transcriptomic information of thousands of single cells in a single experiment. The volume and complexity of resulting data make it a paradigm of big data. Consequently, the field is presented with new scientific and, in particular, analytical challenges where currently no scalable solutions exist. At the same time, exciting opportunities arise from increased resolution of single-cell RNA sequencing data and improved statistical power of ever growing datasets. Big single cell RNA sequencing data promises valuable insights into cellular heterogeneity which may significantly improve our understanding of biology and human disease. This review focuses on single cell transcriptomics and highlights the inherent opportunities and challenges in the context of big data analytics.

Addresses

[1] Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany
[2] Department of Mathematics, Technical University of Munich, Garching, Germany

Corresponding author: Theis, Fabian J. Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany. (fabian.theis@helmholtz-muenchen.de)

Current Opinion in Systems Biology 2017, **4**:85–91

This review comes from a themed issue on **Big data acquisition and analysis (2017)**

Edited by **Pascal Falter-Braun and Michael A. Calderwood**

For a complete overview see the Issue and the Editorial

Available online 15 July 2017

http://dx.doi.org/10.1016/j.coisb.2017.07.004

Keywords

Single-cell RNA-seq, Big data, Single-cell transcriptomics, Machine learning.

Intro

The term "big data" was first coined in 1997 and initially denoted the problem of data not fitting into memory and therefore being too big to be processed by conventional means [1]. The definition was expanded to the four "V"s volume — the amount of data, velocity — the required processing speed, veracity — trustworthiness and availability, and variety — necessary model complexity [2]. The traditional scientific big data field is astronomy because of the huge *volume* of image data produced by telescopes with a high daily *velocity* [3]. Big data has also reached biology, mainly driven through the advent of next generation sequencing technology. For biologists, assessing *veracity* through statistical means is nothing new.

Recent technological advances now allow the profiling of single cells at a *variety* of omic layers (genomes, epigenomes, transcriptomes and proteomes) at an unprecedented level of resolution [4]. Single cell transcriptomics (SCT) entails the profiling of all messenger RNAs present in a single cell and constitutes the most widely-used sc profiling technology [4]. Unlike bulk RNA-seq profiling where sequencing libraries are generated from thousands of cells, scRNA-seq technologies isolate single cells and generate cell-specific sequencing libraries (e.g. Fluidigm [5]) mark RNA content with a cell-specific molecular barcode [6–9]. Both approaches generate gene expression estimates at the single cell level [10]. SCT enables, for the first time, the measurement of the transcriptomic information of thousands, and up to millions of single cells, in a single experiment [7]. The complexity of SCT data coupled with the massive volume inherent to next generation sequencing data makes it a paradigm of big data.

SCT profiles an increasing number of cells, while the underlying amount of raw data per experiment does not nearly grow as fast. Thereby computational needs for data preprocessing and storage stay relatively constant, while the need for development of analytics dealing with a large number of cells is critical. As one of the most popular single-cell technologies with the largest scalability, this review will focus on single-cell transcriptomics and highlight its challenges and opportunities with a particular focus on analytics.

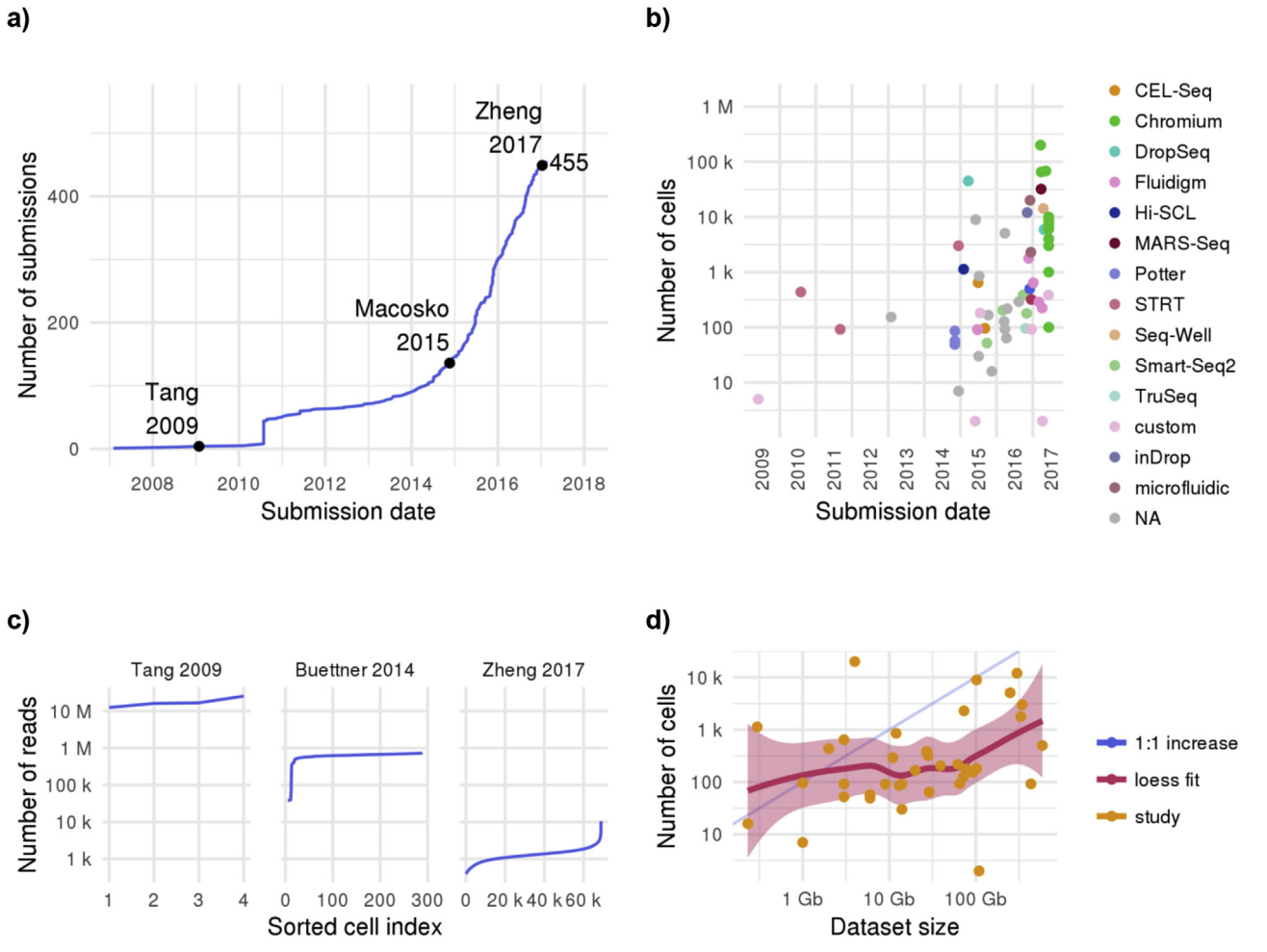
The growth of single-cell transcriptomics

Transcriptome analysis focuses on the profiling of the complete set of RNA molecules in a given experiment and is mainly used to study gene expression. The three dominating technologies in the field, quantitative reverse transcription polymerase chain reaction (qRT-PCR), microarrays and RNA sequencing (RNA-seq), have now all been extended to single cell applications

[11–13]. Up to 500 single cell expression studies are now already available in databases (Figure 1**a**) with increasingly high numbers of cells (Figure 1**b**).

QRT-PCR measures PCR product accumulation by hybridisation with a fluorescent marker and provides up until this date the most sensitive and reproducible quantification of gene expression [14,15]. However, it is limited by low throughput, high costs per cell and high labour effort. While the number of cells measured can be increased using a highly parallelized microfluidic approach, the number of genes is harder to scale up [12]. Multiplexing is only possible up to 4 fluorescent dyes [16] and for each gene one assay and specific primers that need extensive prior testing are needed [17]. More fundamentally this means that qRT-PCR is a hypothesis-driven approach that requires selection of target genes based on prior knowledge, which potentially leads to a biased analysis. One of the first studies to employ single-cell qRT-PCR investigated intercellular expression heterogeneity of globin genes in response to hemin treatment in 46 single K562 cells. Performing this study at single cell resolution allowed the researchers to conclude that globin mRNA levels are initially widely distributed and that the hemin treatment response is high, with the exception of a subset of cells with low initial globin expression [18]. More recent studies have dramatically increased the number of cells profiled in a single experiment. For example, we profiled almost four thousand single cells using single-cell qRT-PCR to study early blood development. Clustering analysis grouped blood progenitor cells from different

Figure 1



a) Increasingly many single cell studies are being published. A query for single cell transcriptome data reveals a steep increase in submissions of single cell data to GEO. See Supplementary Table 1 for the queries used. **b)** Read counts for a few exemplary studies are inversely correlated to cell counts. **c)** Improvement of scRNA-seq techniques allow increasingly larger cell numbers per study. Numbers were obtained from GEO study and sample descriptions, count matrices and papers. **d)** Even though the number of cells per experiment is increasing as per **c**, the data size (compressed format) is increasing less strongly. The file size distribution of the compressed raw reads per dataset peaks between 20 Gb and 120 Gb. Sizes were obtained for all GEO single cell datasets, cell counts as in **a** via its associated SRA sequence data.

developmental stages. The fact that not only cells from the same stage form a cluster implies that their development is asynchronous. We also reconstructed a regulatory network that reflects known pathways and predicts new candidates for regulatory motifs in those pathways [19]. The large number of cells was essential to identify the developmental lineages, which applied to an even larger degree in a subsequent analysis in which we developed a novel algorithm to study diffusion pseudotime [20]. Pseudotime is a similarity-based ordering of transcriptomic states (cells) along the trajectory of a continuous developmental process. It approximates progression of a stereotypic cell in real time and, hence, reveals gene expression dynamics and factors triggering state transitions. The concept of pseudotime requires a resolution of single cell heterogeneities, which precludes insights from bulk RNA-seq data. Using our algorithm we were reconstructed the lineage branching events and identified driver genes.

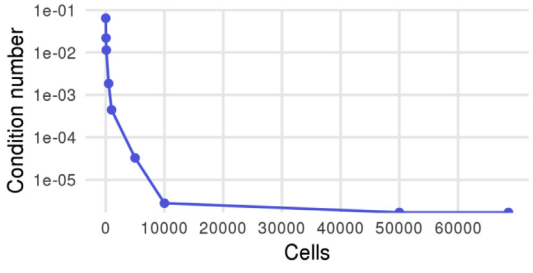
Another popular transcriptome quantification method, microarrays, has also been extended to single cell applications [11]. In comparison to qRT-PCR, microarrays enable transcriptome-wide analyses by using predesigned RNA probes to capture all known transcripts. Ramos and colleagues studied gene expression in 12 stem cells using microarray technology, an early assessment of expression variability in stem cells [21]. However, several limitations hindered microarrays to become a method of choice in single cell transcriptomics. A complete set of probes covering also non-coding regions and splice variants is very expensive. In addition, they suffer from limited sensitivity and dynamic range and require large amounts of starting material, which is problematic in single cells [22,23]. The advent of next-generation sequencing technology has brought the most recent and unbiased transcriptome measurement approach, RNA-seq [24]. A major advantage of RNA-seq in comparison to qRT-PCR and microarrays is the fact that it enables the unbiased profiling of the entire transcriptome. The extension of RNA-seq to single cells was first driven by well-based approaches including Fluidigm C1 [5], SMART-seq [25] and the more scalable MARS-Seq [9]. More recent droplet-based approaches such as Drop-seq [6], inDrop [8] and 10× Genomics' Chromium [7] have substantially increased the number of cells that can be profiled in a single experiment. High numbers of cells can be reached by automating the isolation of lysis and reaction for tens of thousands of cells per run using gel beads or water droplets in oil as inexpensive and scalable medium. This way, the largest data set, as of early 2017, contains more than 1.3 million transcriptomes of mouse brain cells. The community is currently eagerly awaiting results from this new order of magnitude in data size and the analytic possibilities it enables, such as detection of rare subpopulations. One of the limiting factors of next generation sequencing technology is the sequencing costs. Single-cell RNA-seq (scRNA-seq) experiments profile many cells, but often with reduced sequencing depth when comparing the total reads per sample between a single cell and bulk RNA-seq sample (Figure 1**c**). Therefore, the computational demands of data handling and pre-processing do not increase for single-cell relative to bulk experiments. The possibility of determining the depth needed means this is not necessarily a tradeoff [26].

While the data volume of scRNA-seq experiments remains comparable to bulk experiments, the number of samples in and the sparseness of the resulting gene expression tables have increased dramatically. Bulk genomic data has been suffering from the so called 'small n large p' problem, i.e. having many parameters (for example genes) with much lower number of samples. Regularization techniques such as LASSO are commonly used to deal with this 'ill-conditionedness' by decreasing the number of parameters in the model. scRNA-seq for the first time allows to address this problem with data containing more cells than genes ($p \simeq 5k{-}20k$, $n > 10k{-}100k$). As the number of cells increases, the expression matrices become better conditioned (Figure 2). As a result, we expect the application of more standard data analyses techniques without the need for strong regularization techniques designed to counteract ill-conditionedness.

Beyond single-cell transcriptomics

High-throughput sequencing and other technological advances have revolutionised the fields of single-cell genomics and epigenomics. Single cell genome sequencing has its major applications in studying the diversity in microbial ecosystems, evolution and in evaluating the role of genetic mosaicism, especially in cancer [27,28]. Epigenetic mechanisms regulate gene expression within genotypically identical cells and thereby enable the differentiation and maintenance of diverse functional and stable cellular states [29,30].

Figure 2



With increasing number of cells, sample matrices become well-conditioned and novel data analyses become possible. Here we plot condition numbers of differently sized gene expression matrices randomly sampled from the 10× genomics 68k PBMC data. A low condition number indicates well-conditionedness and therefore for example that a matrix inversion becomes possible, which is necessary for many transformations common in analysis.

The first 200 full genomes of single human cells were sequenced in 2011 to model tumour evolution [28] and decreasing costs lead to a steady increase of increasingly voluminous data [31]. Epigenetic profiling has been successfully adapted to single cells despite the inability to amplify epigenetic modifications (review by Refs. [29,30]). In single cell genomics and epigenomics the trade-off between coverage/costs and cell number is more impactful than in transcriptomics [30,32].

A cell's proteome ties genotype to phenotype by defining its response to the various internal and external stimuli. The most advanced sub-field in single cell proteomics is snapshot data, where the number of proteins and cells assayed in parallel could be significantly scaled up in the past few years with novel tools developed on the principles of immunohistochemistry such as flow cytometry [33,34], mass cytometry [35] and microchip-based platforms [36,37]. The upper limit to the number of marker proteins that can be measured simultaneously still sits at a hundred.

Imaging of living cells in vitro and in vivo has resulted in movies featuring high resolutions in both space and time. Using fluorescent light sheet microscopy (SPIM), a calcium imaging approach allowed to capture a zebrafish larval brain with $\sim 10^5$ neurons at a frequency of 0.8 Hz [38], which can result in 1 TB of data per hour and specimen [39]. Imaging flow cytometry allows to record 30 megapixels per second, which can amount to 10,000−100,000 multispectral images per sample [40,41]. Single cell time-lapse movies can amount up to dozens of terabytes of data [42]. While multi-omics single cell data has not yet scaled for large cell numbers, approaches such as scTrio-seq provide larger amounts of per-cell data than single-omics data and have potential for large cell numbers. Single cell triple omics sequencing [43,44].

Technical challenges in SCT data analytics

While single cell data volume only lies in the terabyte range and can therefore be stored locally, its processing and analysis requires parallelisation. Transcriptome data processing like alignment and counting can be performed independently per cell and is therefore a "embarrassingly parallel" − or loosely-coupled − application [45], except for potential batch effects. Combined with the fact that raw sizes are not growing considerably (Figure 1**d**), the main challenge lies in enabling common and novel downstream analysis for ever-growing cell numbers. Often, this analysis − just as processing of other data types − would require to relate all cells to each other and therefore to load the entire data matrix into memory for a traditional computational analysis. Parallelisation accordingly requires a more tightly-coupled and therefore more challenging parallelism [46]. For example more sophisticated interfaces such as the Message Passing Interface (MPI) are better suited best for those kinds of tasks compared to a batch job submission system. MPI and similar interfaces allow efficient and frequent communication between machines connected by a fast network. This leaves a cluster still more efficient than a (heterogeneous) cloud-based solution [47]. Apache's Hadoop is the base for most existing big data RNA-seq applications − single cell or not [3,48], but Spark and other upcoming frameworks are similarly promising.

Scientific challenges

Variety in and of data is a classic biological problem pertaining also to big data [48]. While there are clear opportunities in bigger volumes of data, there are technical, statistical and interpretative challenges rising alongside. The exploration of the cellular heterogeneity of tissues is the focus of many recent scRNA-seq studies [49] and of the human cell atlas project (humancellatlas.org). In most cases, cell populations cannot be experimentally purified from bulk data, whereas single cell data allows for in silico dissection of a mixture of cells into different molecular (e.g. transcriptomic) states. Common approaches target inference of visually interpretable cell clusters or cell/gene networks. This typically involves in-memory implementations of dimension reduction, differential analysis, clustering, network inference, or pseudotime ordering [26,50]. As an example, heterogeneity of tissues is often visually analysed based on a dimension reductions such as Principal Component Analysis (PCA) or Gaussian Process Latent Variable Model (GPLVM) [51]. A popular method are t-Stochastic Neighborhood Embedding (t-SNE) plots, as t-SNE was capable of reproducing prior cell biological knowledge in many cases, for example separating the 8 types of 44,808 retinal cells into distinct clusters [6,7]. One has to be aware that this type of visualisation is an unsupervised approach to finding structure in a data set. As the number of cells in studies increases, the number of putative transcriptomic states will increase. In the case of t-SNE the default result is an increased count of distinct clusters, potentially influencing the interpretation by the scientist performing the analysis. The same applies to other problems such as network synthesis [19,52]. It will therefore be of increasing importance to put appropriate model selection procedures into place to guard against overfitting. Model selection procedures for the number of clusters used to model a sample can be as simple as Akaike or Bayesian information criterion. Model selection is a tradeoff of complex models which are liable to fit noise ("overfitting") against simplistic models which are not capable to reconstruct the full underlying signal of interest. A common scenario in single-cell transcriptomics is the search for rare or new cell types in large data sets: Model complexity is the number of cell types used to explain the data. In scRNA-seq,

likelihood-based model selection is still difficult because multiple effects impede modelling of the data generating process: Drop-out measurements and unclear steady-state and non-steady-state transcript distributions. Advances in statistical modelling of scRNA-seq data will therefore be closely linked to advances in model selection of explorative heterogeneity analysis.

A second immediate analytic strategy used in most transcriptomic studies is the dissection of processes into contributions of different genes. A widely used approach is differential expression analysis. Most commonly used differential expression algorithms are likelihood-based [53–55] and include model fitting steps: The scalability of these methods ("algorithmic complexity") to large data sets becomes increasingly important as algorithms differ in the maximal data set size they can cope within an acceptable run time.

Beyond the computational analysis, challenges arise when it comes to biological interpretation. With the ultimate goal to relate molecular measurements to cellular function in health and disease, the wealth of information contained in SCT data needs to be transformed into relevant biological knowledge. One has to keep in mind that computational approaches based on big data can generate hypotheses and only complement but not substitute traditional experimental validation. In particular, only from transcriptional profiles obtained from scRNA-seq data it is difficult to make conclusive statements on the functional state of distinct cell subtypes and gene expression just hints at underlying molecular mechanisms. To date all solely big data-driven and unbiased single cell analyses remain largely descriptive and lack mechanistic insights. Extracting the important and relevant information that should guide follow-up experiments such as (genetic) perturbation or lineage-tracing experiments requires, besides a computing infrastructure and robust statistical or machine-learning methods, a clear research question and biological domain expertise.

Vice versa, big data are a powerful source for biologists to improve experimental design and focus their research [56]. Many big data sets are "under-analysed" and could be reused to address other biological problems [57]. This requires a certain rethinking in traditional biology which tends to create evermore new data. Now, with the large body of accumulated data, it is likely that insights will come faster and experiments will be more efficient, if existing data is integrated and used to carefully interpret, validate and falsify hypotheses and generate new predictions.

Opportunities and summary

What can we hope to gain from higher cell numbers? High sample numbers are a necessity if we deal with complex models. The success of deep learning models, which feature millions of parameters, in other areas of science is in large part due the availability of a high amount of data [58]. While the analysis of single cell biological data has been dominated by relatively simple models — simple meaning few parameters — SCT will enable training complex machine learning methods to give unprecedented insight into biology at the single cell level. Furthermore, in experiments that study biological processes, the sampling frequency is often a limiting factor. Quite generally, one is not only interested in what happens at the beginning and the end of a process, for example, a healthy and a diseased state, but also in what happens "in between". As the sampling conditions in experiments can hardly be controlled in many settings, high sample numbers are the only guaranteed way to "fill in gaps" in the data. This also applies to imputation, which is a necessity in the generally sparse and noisy single cell transcriptomics data [59,60]. When considering a prototypical problem in which one is interested in classifying cell types based on molecular data, that is, a discrete case, one can hope that high cell numbers reveal rare subtypes and increase our confidence in statistical hypothesis tests [6]. As an example, in a tissue sample a small number — relative to the whole sample — of precancerous cells is less likely to be seen as inconsequential outliers when the absolute sizes of sample and subpopulation are larger [61]. The aforementioned confidence is crucial when studying complex hypothesis, for example, in the context of combinatorial gene regulation.

It's becoming increasingly clear that our notion of "cell type" serves merely our own understanding of the interplay of cells. What exists instead is a landscape of cell tendencies and specialisations that are as diverse as they are plastic [62]. Hypothesis free isolations of these "cell tags" from big experiments will help to see beyond the rigid cell classifications today and allow us identify previously unseen similarities and distinctions between cells.

In summary, SCT has entered the regime of big data. With the development of novel and robust tools for big data analytics, we can harness the increased statistical and machine learning power. We hope to grasp the chances in discovering weak signals (e.g. discovery of rare subpopulations), differentially regulated subnetworks or lineage transitions between cell types. With the availability of increasingly large-scale data sets, we can also embrace robust method comparisons and multi-dataset evaluations. By embracing scale as additional factor in SCT data analytics, we will be able to harness SCT's potential for other areas of biological and medical research.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

S. T. and D. F. are supported by a DFG Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM).
F. A. W. acknowledges support by the Helmholtz Postdoc Programme, Initiative and Networking Fund of the Helmholtz Association.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.coisb.2017.07.004.

References

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

- • of special interest
- •• of outstanding interest

1. Cox M, Ellsworth D: *Application-controlled demand paging for out-of-core visualization*. Proceedings. Visualization '97 (Cat. No. 97CB36155). 1997, http://dx.doi.org/10.1109/visual.1997.663888.

2. Demchenko Y, Grosso P, de Laat C, Membrey P: *Addressing big data issues in scientific data infrastructure*. 2013 International Conference on Collaboration Technologies and Systems (CTS). 2013, http://dx.doi.org/10.1109/cts.2013.6567203.

3. O'Driscoll A, Daugelaite J, Sleator RD: **"Big data", Hadoop and cloud computing in genomics**. *J Biomed Inf* 2013, **46**:774–781.

4. Linnarsson S, Teichmann SA: **Single-cell genomics: coming of age**. *Genome Biol* 2016, **17**:97.

5. Xin Y, Kim J, Ni M, Wei Y, Okamoto H, Lee J, Adler C, Cavino K, Murphy AJ, Yancopoulos GD, Lin HC, Gromada J: **Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells**. *Proc Natl Acad Sci U S A* 2016, **113**: 3293–3298.

6. •• Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA: **Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets**. *Cell* 2015, **161**:1202–1214.

Published back-to-back, the two papers Klein et al., 2015 and Macosko et al., 2015 introduce microfluidic devices that automate the creation and population of droplets that will contain the necessary reactions for massively scalable single-cell transcriptome analysis.

7. •• Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH: **Massively parallel digital transcriptional profiling of single cells**. *Nat Commun* 2017, **8**:14049.

A commercialisation and extension of the aforementioned microfluidic groundwork, 10× Genomics enables every lab to leverage the cheap and scalable technology, regardless of the availability of hardware to manufacture microfluidic devices.

8. •• 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.

Published back-to-back, the two papers Klein et al., 2015 and Macosko et al., 2015 introduce microfluidic devices that automate the creation and population of droplets that will contain the necessary reactions for massively scalable single-cell transcriptome analysis.

9. • Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I: **Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types**. *Science* 2014, **343**: 776–779.

MARS-Seq is the first protocol to introduce scalability for scRNA-Seq. Other than in well-less microfluidic devices, a robotic approach is used to place cells into wells.

10. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based technologies will revolutionize whole-organism science**. *Nat Rev Genet* 2013, **14**:618–630.

11. Esumi S, Wu S-X, Yanagawa Y, Obata K, Sugimoto Y, Tamamaki N: **Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors**. *Neurosci Res* 2008, **60**:439–451.

12. White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, Hansen CL: **High-throughput microfluidic single-cell RT-qPCR**. *Proc Natl Acad Sci U S A* 2011, **108**:13999–14004.

13. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods* 2009, **6**:377–382.

14. Heid CA, Stevens J, Livak KJ, Williams PM: **Real time quantitative PCR**. *Genome Res* 1996, **6**:986–994.

15. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA: **The technology and biology of single-cell RNA sequencing**. *Mol Cell* 2015, **58**:610–620.

16. Zhong Q, Bhattacharya S, Kotsopoulos S, Olson J, Taly V, Griffiths AD, Link DR, Larson JW: **Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR**. *Lab Chip* 2011, **11**:2167–2174.

17. Citri A, Pang ZP, Südhof TC, Wernig M, Malenka RC: **Comprehensive qPCR profiling of gene expression in single neuronal cells**. *Nat Protoc* 2011, **7**:118–127.

18. Smith RD, Malley JD, Schechter AN: **Quantitative analysis of globin gene induction in single human erythroleukemic cells**. *Nucleic Acids Res* 2000, **28**:4998–5004.

19. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa S-I, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B: **Decoding the regulatory network of early blood development from single-cell gene expression measurements**. *Nat Biotechnol* 2015, **33**:269–276.

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

21. Ramos CA, Bowman TA, Boles NC, Merchant AA, Zheng Y, Parra I, Fuqua SAW, Shaw CA, Goodell MA: **Evidence for diversity in transcriptional profiles of single hematopoietic stem cells**. *PLoS Genet* 2006, **2**. e159.

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

23. Nygaard V, Hovig E: **Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling**. *Nucleic Acids Res* 2006, **34**:996–1014.

24. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies**. *Nat Rev Genet* 2010, **11**: 843–854.

25. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R: **Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells**. *Nat Biotechnol* 2012, **30**:777–782.

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

27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**:455–477.

28. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L,

Krasnitz A, Richard McCombie W, Hicks J, Wigler M: **Tumour evolution inferred by single-cell sequencing**. *Nature* 2011, **472**:90–94.

29. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W: **Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity**. *Genome Biol* 2016, **17**:72.

30. Schwartzman O, Tanay A: **Single-cell epigenomics: techniques and emerging applications**. *Nat Rev Genet* 2015, **16**:716–726.

31. Wang Y, Navin NE: **Advances and applications of single-cell sequencing technologies**. *Mol Cell* 2015, **58**:598–609.

32. Gawad C, Koh W, Quake SR: **Single-cell genome sequencing: current state of the science**. *Nat Rev Genet* 2016, **17**:175–188.

33. O.D. Perez, P.O. Krutzik, G.P. Nolan, Flow cytometric analysis of kinase signaling cascades, in: Flow Cytometry Protocols, n.d.: pp. 067–094.

34. Perfetto SP, Chattopadhyay PK, Roederer M: **Seventeen-colour flow cytometry: unravelling the immune system**. *Nat Rev Immunol* 2004, **4**:648–655.

35. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, Tanner SD: **Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry**. *Anal Chem* 2009, **81**:6813–6822.

36. Love JC, Ronan JL, Grotenbreg GM, van der Veen AG, Ploegh HL: **A microengraving method for rapid selection of single cells producing antigen-specific antibodies**. *Nat Biotechnol* 2006, **24**:703–707.

37. Ma C, Fan R, Ahmad H, Shi Q, Comin-Anduix B, Chodon T, Koya RC, Liu C-C, Kwong GA, Radu CG, Ribas A, Heath JR: **A clinical microchip for evaluation of single immune cells reveals high functional heterogeneity in phenotypically similar T cells**. *Nat Med* 2011, **17**:738–743.

38. Ahrens MB, Orger MB, Robson DN, Li JM, Keller PJ: **Whole-brain functional imaging at cellular resolution using light-sheet microscopy**. *Nat Methods* 2013, **10**:413–420.

39. Perez CC, Lauri A, Symvoulidis P, Cappetta M, Erdmann A, Westmeyer GG: **Calcium neuroimaging in behaving zebrafish larvae using a turn-key light field camera**. *J Biomed Opt* 2015, **20**:096009.

40. George TC, Basiji DA, Hall BE, Lynch DH, Ortyn WE, Perry DJ, Seo MJ, Zimmerman CA, Morrissey PJ: **Distinguishing modes of cell death using the ImageStream multispectral imaging flow cytometer**. *Cytometry A* 2004, **59**:237–245.

41. Henery S, George T, Hall B, Basiji D, Ortyn W, Morrissey P: **Quantitative image based apoptotic index measurement using multispectral imaging flow cytometry: a comparison with standard photometric methods**. *Apoptosis* 2008, **13**:1054–1063.

42. Schroeder T: **Long-term single-cell imaging of mammalian stem cells**. *Nat Methods* 2011, **8**:S30–S35.

43. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, Peng J: **Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas**. *Cell Res* 2016, **26**:304–319.

44. Macaulay IC, Ponting CP, Voet T: **Single-cell multiomics: multiple measurements from single cells**. *Trends Genet* 2017, **33**:155–168.

45. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP: **Computational solutions to large-scale data management and analysis**. *Nat Rev Genet* 2010, **11**:647–657.

46. Tokunaga T, Hirose O, Kawaguchi S, Toyoshima Y, Teramoto T, Ikebata H, Kuge S, Ishihara T, Iino Y, Yoshida R: **Automated detection and tracking of many cells by using 4D live-cell imaging data**. *Bioinformatics* 2014, **30**:i43–51.

47. Zhai Y, Liu M, Zhai J, Ma X, Chen W: *Cloud versus in-house cluster*. State of the Practice Reports on - SC '11. 2011, http://dx.doi.org/10.1145/2063348.2063363.

48. Yu P, Lin W: **Single-cell transcriptome study as big data**. *Genomics Proteomics Bioinforma* 2016, **14**:21–30.

49. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, Melton DA, Yanai I: **A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure**. *Cell Syst* 2016, **3**:346–360. e4.

50. • Marr C, Zhou JX, Huang S: **Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots**. *Curr Opin Biotechnol* 2016, **39**:207–214.
This review outlines the common challenges and practices in single-cell transcriptomics and epigenomics, calling for going towards a theory-based analysis.

51. • Buettner F, Theis FJ: **A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst**. *Bioinformatics* 2012, **28**:i626–i632.
Structured GP-LVMs were one of the first nonlinear dimension reduction methods for single-cell gene expression data, resolving more subpopulations than the popular PCA.

52. AlJadda K, Korayem M, Ortiz C, Grainger T, Miller JA, York WS: *PGMHD: a scalable probabilistic graphical model for massive hierarchical data problems*. 2014 IEEE International Conference on Big Data (Big Data). 2014, http://dx.doi.org/10.1109/bigdata.2014.7004213.

53. Ho Sui SJ, Begley K, Reilly D, Chapman B, McGovern R, Rocca-Sera P, Maguire E, Altschuler GM, Hansen TAA, Sompallae R, Krivtsov A, Shivdasani RA, Armstrong SA, Culhane AC, Correll M, Sansone S-A, Hofmann O, Hide W: **The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons**. *Nucleic Acids Res* 2012, **40**:D984–D991.

54. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, Juliana McElrath M, Prlic M, Linsley PS, Gottardo R: **MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data**. *Genome Biol* 2015, **16**, http://dx.doi.org/10.1186/s13059-015-0844-5.

55. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**:550.

56. Dolinski K, Troyanskaya OG: **Implications of big data for cell biology**. *Mol Biol Cell* 2015, **26**:2575–2578.

57. Marx V: **Biology: the big challenges of big data**. *Nature* 2013, **498**:255–260.

58. Buggenthin F, Marr C, Schwarzfischer M, Hoppe PS, Hilsenbeck O, Schroeder T, Theis FJ: **An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy**. *BMC Bioinforma* 2013, **14**:297.

59. Buettner F, Pratanwanich N, Marioni JC, Stegle O: *Scalable latent-factor models applied to single-cell RNA-seq data separate biological drivers from confounding effects*. 2016, http://dx.doi.org/10.1101/087775.

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

61. Grün 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.

62. Clevers H, Rafelski S, Elowitz M, Klein A, Shendure J, Trapnell C, Lein E, Lundberg E, Uhlen M, Martinez-Arias A, Sanes JR, Blainey P, Eberwine J, Kim J, Love JC: **What is your conceptual definition of "Cell Type" in the context of a mature organism?** *Cell Syst* 2017, **4**:255–259.