# Single-Cell Transcriptional Analysis

Angela R. Wu,[1] Jianbin Wang,[2] Aaron M. Streets,[3] and Yanyi Huang[4]

[1]Division of Life Science and Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China; email: angelawu@ust.hk

[2]School of Life Sciences and Center for Life Sciences, Tsinghua University, Beijing 100084, China; email: jianbinwang@mail.tsinghua.edu.cn

[3]Department of Bioengineering, University of California, Berkeley, California 94720; email: astreets@berkeley.edu

[4]Biodynamic Optical Imaging Center (BIOPIC), Beijing Advanced Innovation Center for Genomics (ICG), College of Engineering, and Center for Life Sciences, Peking University, Beijing 100871, China; email: yanyi@pku.edu.cn

Annu. Rev. Anal. Chem. 2017. 10:439–62

First published as a Review in Advance on March 16, 2017

The *Annual Review of Analytical Chemistry* is online at anchem.annualreviews.org

https://doi.org/10.1146/annurev-anchem-061516-045228



## Keywords

RNA-seq, gene expression, heterogeneity, microfluidics, next-generation sequencing

## Abstract

Despite being a relatively recent technological development, single-cell transcriptional analysis through high-throughput sequencing has already been used in hundreds of fruitful studies to make exciting new biological discoveries that would otherwise be challenging or even impossible. Consequently, this has fueled a virtuous cycle of even greater interest in the field and compelled development of further improved technical methodologies and approaches. Thanks to the combined efforts of the research community, including the fields of biochemistry and molecular biology, technology and instrumentation, data science, computational biology, and bioinformatics, the single-cell RNA-sequencing field is advancing at a pace that is both astounding and unprecedented. In this review, we provide a broad introduction to this revolutionary technology by presenting the state-of-the-art in sample preparation methodologies, technology platforms, and computational analysis methods, while highlighting the key considerations for designing, executing, and interpreting a study using single-cell RNA sequencing.

## 1. INTRODUCTION

Cells are the fundamental unit of life: In our body, millions of cells work in harmony and coordination to carry out basic physiological functions required for life. Yet many of the pivotal points in critical biological processes and biological systems, such as the genesis and evolution of cancers and the development and fate commitment of early-stage embryos, are determined by the biology of a single cell. The ability to interrogate single-cell states undoubtedly provides great insight on critical biological processes that are important to both health and disease, and for this reason, single-cell technologies have emerged as an indispensable tool for the modern life scientist (1). Although the cell's genome is the foundational genetic blueprint from which cellular actions originate, transcriptional and translational activity can vary dramatically according to cellular identity and in response to environmental changes. Proteins are the effector molecules within cells that directly bring about changes in cellular identity, dynamics, and function. But despite ongoing efforts, quantitative measurement of proteins, particularly at the single-cell level, remains challenging due to the inability to amplify proteins and efficiently read their sequences. Thus, RNA transcripts, functioning directly upstream of proteins, have become a useful albeit imperfect proxy for appraising protein abundance. Furthermore, as we expand our knowledge of the transcriptional landscapes in a multitude of cell types, RNA transcriptomes are increasingly useful as quantitative descriptions of cellular identity, independent of their relationship with protein abundance.

Genome-wide transcriptional profiling paints a global picture of gene expression, generating multidimensional data sets that ultimately provide a unique cellular signature encoded by the correlations between levels of gene expression (2–4). Despite the appearance of being uniform, cellular heterogeneity exists in every cell population, rendering conventional ensemble-level transcriptional analyses woefully inaccurate in their reflection of the diversity and heterogeneity of a biological system. Furthermore, ensemble-level interrogations inevitably miss the discovery of minority cell subpopulations due to their technical inability to recognize weak biological signals from noisy background (5–8). Transcriptional analysis at the single-cell level is becoming the ultimate technology of choice for precisely and accurately reconstructing the composition of a functional biological unit at the highest resolution and for elucidating the networking and dynamic interplay between individual cells.

In the past decade, there has been a paradigm shift in the choice of transcriptional analysis technologies. Researchers in many fields now favor the use of next-generation sequencing (NGS) approaches over previously popular multiplex reverse-transcription polymerase chain reactions (RT-PCRs) or hybridization-based microarray assays, mainly due to the significantly improved accuracy and precision of NGS. Compared to microarray and real-time PCR, transcriptional sequencing (RNA-seq) has several advantages. First, RNA-seq scans the full length of each transcript instead of examining restricted loci as in microarray or real-time PCR. Second, RNA-seq does not require probe or primer design; therefore, no a priori knowledge of the genome or transcriptome sequence is needed. Routine RNA-seq analysis does use a reference genome for mapping sequenced reads, but in the event there is no reference, de novo transcript assembly is also feasible. Third, RNA-seq can easily cover the whole transcriptome, whereas real-time PCR is targeted and limited in scope. Fourth, a typical next-generation sequencer requires small amounts of library material, typically on the order of nanograms, whereas microarrays demand micrograms of material. RNA-seq is now also more affordable than microarray. Given these advantages, RNA-seq has become the routine method for transcriptional analysis.

For the aforementioned reasons, the applications of single-cell transcriptional analysis in basic science research are broad and growing rapidly. By providing higher resolution, single-cell RNA-seq can directly interrogate rare cells, such as stem cells or low-abundance progenitors, within a

population dominated by other cell types. Researchers have already applied this tool to identify novel cell types within previously well-studied tissue and organ systems (9–11). In another application, by profiling individual cells and assessing the gradual change between them, researchers are able to elucidate transitional states and relationships between cells, such as during the course of differentiation or reprogramming (12, 13). Although single-cell RNA-seq is currently still too costly for routine use, there are some efforts to translate this technology into the clinical setting in the context of personalized noninvasive cancer diagnostics, where the targeted circulating tumor cells are too rare to obtain in bulk (14, 15).

Since 2009, the pairing of RNA-seq technology with single-cell transcriptional analysis has rapidly brought great successes in numerous application areas (16, 17). Every single cell contains only picograms of RNA; therefore, it is essential to amplify this RNA to produce enough material for the construction of NGS libraries. Many protocols, each with its own advantages and limitations, have been developed, and it is anticipated that many more will be developed in coming years to continue improving their performance. In this review, we cover both experimental and computational technologies of single-cell RNA-seq. We elaborate on the strategies of experimental design and considerations for protocol selection, enumerate the various bioinformatic processing methods, and discuss the challenges of this evolving field. We focus on the technological development and data analysis, rather than detail the discoveries made by these technologies in fundamental biology or medicine, and we aim to offer objective assessments on prevailing methods that have been reported. Because this field is growing rapidly, it is entirely possible that by the time this review is published, there will be more recent techniques available. Nevertheless, we hope that this review not only provides a showcase of available technologies but will also inspire new inventions to overcome existing challenges in the field.

## 2. WHAT CAN WE LEARN FROM SINGLE-CELL RNA-SEQ?

To plan an effective single-cell RNA-seq experiment, one must first determine what types of information are needed to address the research question being investigated. Different experimental protocols for capturing and processing RNA transcripts from single cells provide different facets of information about the transcriptional state of the cell, due to the limitations of each approach inherent in its molecular biology. For example, although data from most single-cell RNA-seq workflows allow classification of cells into subtypes based on their global transcriptional signature, some workflows provide full-length transcript sequences that allow for discovery of novel transcript isoforms, whereas others may only provide information at the gene level. In this section, we discuss the types of biological insight that can be gained from single-cell RNA-seq and some of the advantages and disadvantages of popular workflows in that context.

### 2.1. Gene Counting

The first wave of single-cell RNA-seq studies focused on gene expression dynamics during early embryonic development. Tang and colleagues (18) traced the transition of cells from the inner cell mass of mouse blastocysts to mouse pluripotent embryonic stem cells using single-cell RNA-seq analysis. They detected significant molecular transitions and major changes in transcript variants over the course of this cell type transition in development. A few years later, Yan et al. (19) analyzed the same process in human early embryos. Studies led by Durruthy-Durruthy (20) and Petropoulos (21) further expanded the scope of these investigations and explored the lineage establishment process. These results helped researchers to understand the regulatory roadmap from single-cell zygote to blastocytes. Aside from its application in early embryonic development, single-cell

RNA-seq demonstrated its strength in illuminating various other dynamic processes: Researchers detected substantial variation between dendritic cells upon identical immune stimulation (22). Tan et al. (23) tested the "one-neuron-one-receptor" rule by sequencing the transcriptomes of over 100 single olfactory sensory neurons and found that, contrary to prior belief, a subset of these cells expressed multiple olfactory receptors. The aforementioned studies shed light on a new way to investigate developmental processes, which benefit greatly from studies at the single-cell level. Additionally, in response to rising interest in the newly discovered biological functions of noncoding RNA, previously thought to be junk RNA, the scope of RNA-seq was expanded beyond just messenger RNA (mRNA) to allow profiling of microRNA and long noncoding RNA (lncRNA) as well. One study utilized strand-specific single-cell RNA-seq to identify cell type–specific lncRNA (24).

## 2.2. RNA Splicing

Beyond simple counting to ascertain gene level abundance, the intricate biology of the cell is influenced and regulated by an additional dimension of transcriptional variation: the production of transcript isoforms by alternative splicing. Scientists once speculated that the human genome has more than 50,000 genes (25, 26). With the completion of the draft human genome sequence, the gene number settled to approximately 20,000–30,000, depending on how a gene is defined (27, 28). This is on par with the number of genes found in other species that are much less biologically complex than humans, such as mouse or even some plants. Roughly the same number of genes can generate different magnitudes of complexity by alternative gene splicing mechanisms, which create transcriptional variation via combinatorial joining of different exons in the gene and via the use of alternative transcription start and stop sites (29). RNA-seq can scan the full-length transcript, so it is easy to detect alternative splicing by searching for exon–exon junctions or by de novo transcript assembly (30–32). Another specific splicing isoform is the circular RNA (circRNA), which has been recently discovered in many species and is expected to harbor important biological functions (33, 34). Because circRNAs do not contain poly(A) tails at the 3′-ends, they cannot be captured by most single-cell RNA-seq approaches that have been optimized for poly(A)-containing eukaryotic mRNAs. Recently developed poly(A)-independent capture methods have enabled amplification and sequencing of low-abundance circRNAs from single cells and have also been applied to investigate early-stage development of mouse and human embryos (35, 36).

## 2.3. Cell Typing

Researchers have long been familiar with the idea that different cell types have different gene expression signatures. Although many studies were fruitful by utilizing the gene markers of known cell types to further select for and study them, other studies sought to characterize less-understood cell types or identify new types of cells, precisely by targeting those cells without known gene markers. Aided by the rapid decrease in cost of single-cell whole-transcriptome amplification and NGS, many studies now routinely report data from hundreds or even thousands of single cells (37). One major advantage of this kind of large-scale single-cell RNA-seq project, compared with analyzing fewer cells, is the implementation of unsupervised gene expression classification with high statistical power: Now, cells with similar gene expression profiles can be grouped together without having prior knowledge of any cell type markers. Novel cell types can thus be discovered with appropriate analysis methods, as first demonstrated by Treutlein et al. (9) in their identification of alveolar epithelial bipotent progenitor cells. Since then, this so-called reverse tissue engineering approach has helped researchers to examine the cellular heterogeneity in the brain (10, 38), placenta (39),

lymphoid system (40), and tumors (41), as well as to identify novel and rare cell types in these cellular populations (42). It is not necessary to have isoform-level information to classify cells into their subpopulations, as demonstrated by Macosko and colleagues (43) and many others (44–46); in contrast, if full-length transcripts are assessed, it is possible to uncover novel cell type–specific biology based on isoform variants, as demonstrated by the study of neurexin splicing in neurons by Treutlein et al. (47). As a powerful extension from simply identifying subpopulations of cells, researchers have also created computational algorithms that can place these subpopulations on a biological or pseudotemporal trajectory. For example, the lineage of cell types in the hematopoietic system of zebrafish was shown to consist of a continuous spectrum of differentiation using single-cell RNA-seq (48). This was first accomplished by grouping cells by subtype and then placing them along a pseudotemporal axis on the basis of the similarities of transcriptional signatures of each cell cluster compared to those of other clusters.

## 3. EXPERIMENTAL DESIGN AND PROCEDURES

After determining the desired outputs from the single-cell RNA-seq for the specific research question at hand, whether it is gene-level abundance measurements or full-length transcripts that can provide transcript-level detail, the overall experimental design and specific protocol should be considered. The overall experimental design of single-cell experiments poses a unique set of questions, and an ill-planned experiment could introduce confounding artifacts that are problematic and irresolvable in the downstream interpretation of biological significance. Selection of appropriate biochemistry for each step of the workflow may be constrained by characteristics of the samples being studied, for example, when limited numbers of cells are available. It will also depend on the types of analyses that will be performed on the resulting data and which biases are acceptable in a trade-off with other factors. In this section, we discuss important elements in overall experimental planning, as well as the strengths and limitations of specific protocols based on their molecular biology mechanisms. We also review recent developments in technology and instrumentation that can aid the single-cell RNA-seq experimental workflow.

### 3.1. Experimental Planning

Although it is generally desirable to sequence more cells with deeper coverage, the scale of any specific project is inevitably limited by time and budget. Striking a balance between the number of cells profiled and the sequencing depth achieved for each cell within given budget constraints is critical to the success of a project. In addition, it is also essential to include appropriate controls. In this section, we describe the considerations for these aspects of experimental planning for studies using single-cell RNA-seq.

**3.1.1. Cell number versus sequencing depth.** How many cells do we need to analyze, and how many sequencing reads do we need for each single cell [i.e., the sequencing depth (49)]? The simplest way to objectively illustrate the complexity, diversity, heterogeneity, and spatial organization of a group of cells at the single-cell level is to sequence as many cells as possible, as deeply as possible. Although more cells and deeper sequencing provide more information, they also mean greater experimental costs and computational challenges in data analysis. Unless there are infinite resources at one's disposal, there will always be a trade-off between cell number and sequencing depth.

In one strategy, the number of cells to be profiled can be reduced by performing a cursory screen on selected cellular phenotypes or some other cell biomarkers, which can effectively enrich

specific groups of cells for further RNA-seq analysis at moderate or high sequencing depth per cell (50). This strategy is feasible and realistic when studying a system with some prior knowledge but also runs the risk of passing up important novel discoveries. In another scenario, with a large number of cells being analyzed, each cell's sequencing depth must be carefully considered so as to remain within budgetary and computational constraints, while still obtaining sufficient information for meaningful analysis and interpretation. A general guideline for the sequencing depth is that the depth should not be too shallow to avoid missing the transcripts with relatively lower abundance or with stochastic expression patterns; moreover, it should not be too deep, as the number of detectable genes reaches saturation at a certain depth (51, 52). The optimal depth varies depending on cell type and experimental protocols. Based on one previous study (53), the ability to confidently capture most detectable genes in a mammalian single-cell RNA-seq sample, 1–2 million reads should be sequenced for each cell. Other studies suggest that as few as 200,000 reads may be sufficient, but such shallow depths of sequencing may result in underrepresentation of certain low-abundance transcripts as well as other biases caused by pooling multiple samples during sequencing. It is also recommended that during the first single-cell RNA-seq experiment for a particular cell type, one should err on the side of caution by sequencing to a greater depth, as the number of unique transcripts per cell is difficult to predict and can vary greatly based on cell type, cell size, and culture conditions. Cancer cell lines, for example, can express on average over 10,000 genes per cell; in comparison, mouse embryonic lung cells express around 3,000 genes per cell (9).

**3.1.2. Experimental design.** A carefully designed experiment will not only save cost and time, but it also provides data with significantly higher precision and accuracy for highly confident and novel discoveries. Unique to single-cell studies and unlike most other experiments, single-cell RNA-seq has no real biological replicates. Every single cell may be unique, and simply averaging the results of multiple cells will not improve the signal-to-noise ratio, but this is rather likely to further dedifferentiate the signal from the background. However, the absence of biological replicates does not mean there is no way to identify true differentially expressed genes between two individual cells. One important issue is the intersample normalization for quantitative comparison. The use of exogenous RNA molecules as a spike-in, such as the set designed by the External RNA Controls Consortium (ERCC) (54–56), has become a routine process to help identify poor quality samples and reduce the experimental variation between samples. Recent studies also point out that batch effects cause major distortions between experiments operated by different persons, carried out on different instruments, or conducted with different reagents (57, 58). Careful experimental planning, including taking precautions such as distributing different cells into different batches of experiments, helps eliminate or reduce the batch effect. For example, the workflow to perform RNA-seq on 50 individual cells of type A and 50 individual cells of type B is estimated to take two whole days, because the laboratory's throughput is approximately 50 cells per day. One plan is to perform RNA-seq on all 50 type A cells on day 1 and all 50 type B cells on day 2. An alternative plan is to process 25 of each cell type on day 1 and then do another 25 of each type on day 2. Although the latter strategy is more experimentally tedious, it is preferred, because in the first plan, any batch-related effect on experiments conducted on different days will confound the interpretation of differences between cell types.

## 3.2. Single-Cell RNA-Seq Protocols

Although NGS requires far less sample input compared to microarray, nanograms or more of DNA are still mandatory for most library preparation methods. Even the oocyte, which arguably contains

the highest amount of RNA among all cell types, is orders of magnitude away from meeting this requirement. Therefore, reverse transcription and amplification are crucial in single-cell RNA-seq experiments. A typical single-cell RNA-seq protocol consists of three major modular steps: double-stranded complementary DNA (cDNA) synthesis, cDNA amplification, and sequencing library preparation, as shown in **Figure 1**.

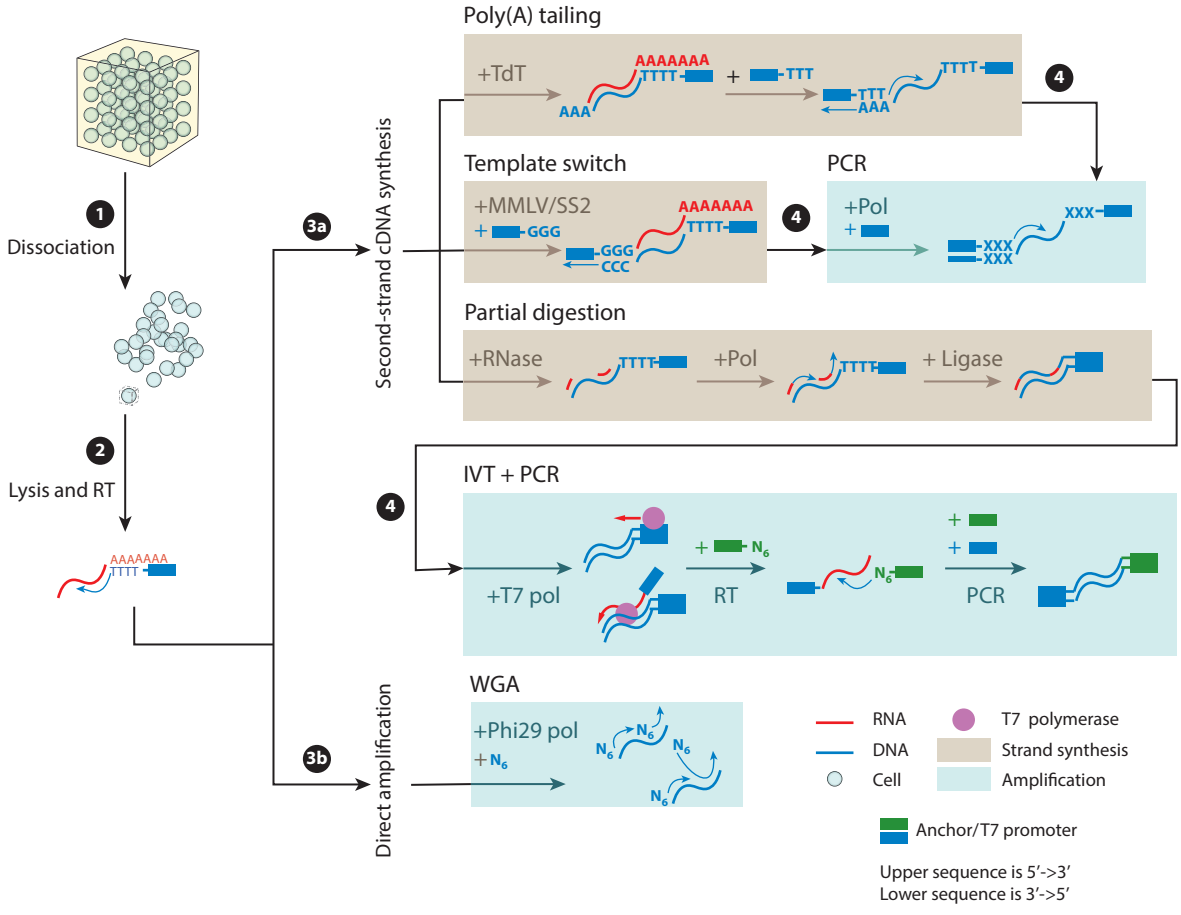


**Figure 1**

Single-cell RNA-seq experimental pipeline. This schematic depicts the essential reaction steps necessary to generate a cDNA library for next-generation sequencing. In step ❶, samples are dissociated into a single-cell suspension. In step ❷, each single cell is lysed, and mRNA molecules are reverse transcribed into first-strand cDNA with anchored poly(T) primers. Certain protocols produce second-strand cDNA (step ❸), which requires annealing of new primers. The poly(A) tailing method uses TdT to add poly(A) tails to the 3′-end of first-strand cDNA, whereas the template switch method utilizes the cytosine tailing activity of MMLV/SS2 to recruit template switching oligos. The second-strand cDNA with anchors on both ends can be further amplified by PCR in step ❹. RNA templates that are partially digested by RNase can also serve as primers for second-strand cDNA synthesis. The double-stranded cDNA can be amplified by IVT and further converted to DNA by reverse transcription and PCR. WGA-based amplification requires no second-strand cDNA synthesis; first-strand cDNA is directly amplified by Phi29 pol and random N6 primers in step ❸. Abbreviations: cDNA, complementary DNA; IVT, in vitro transcription; MMLV, Moloney murine leukemia virus; PCR, polymerase chain reaction; Phi29, bacteriophage ϕ29; pol, DNA polymerase; RNA-seq, RNA transcriptional sequencing; RT, reverse transcription; SS2, SuperScript™ II reverse transcriptase; TdT, terminal deoxynucleotidyl transferase; WGA, whole-genome amplification.

**3.2.1. Double-stranded cDNA synthesis.** Typical mRNAs have their signature poly(A) tails. Almost all published methods, except for specially designed projects (35), utilize oligo(dT) primers for transcript capture and first-strand cDNA synthesis. With specially designed 5′ sequences in the oligo(dT) primer, researchers can incorporate anchor sequences and even tagging barcodes into the newly synthesized cDNA. The real challenge comes from generating the second-strand cDNA. Several strategies have been introduced to the field, and each has its own strength and weaknesses.

***3.2.1.1. Second-strand cDNA synthesis.*** The RNase H activity in retroviral reverse transcriptase competes with its own polymerase activity against the RNA template, resulting in decreased formation of the RNA template-primer complex. This issue may be partially tolerated when there is a large amount of template RNA, but at the single-cell level, it could render the experiment a failure by degrading the limited amounts of template RNA. The removal of RNase H activity improves the reverse transcriptase's processivity and is crucial for the amplification of long mRNA. Therefore, Moloney murine leukemia virus (MMLV) and its successors used in first-strand cDNA synthesis all had their RNase H activity abolished by genetic engineering. The DNA–RNA duplex products offer the perfect template for a classic second-strand synthesis method (59). Small amounts of *Escherichia coli* RNase H partially digest the RNA within the duplex and produce short RNA primers. In the presence of *E. coli* DNA polymerase I, these RNA primers promote DNA synthesis at multiple sites along the cDNA template. DNA ligase seals the nicks between newly synthesized fragments, and T4 DNA polymerase further polishes the ends to produce blunt full-length double-stranded cDNA. This short RNA priming method can be easily combined with first-strand synthesis and produces full-length cDNA with minimal bias. It is widely used in bulk RNA-seq experiments; however, application on single cells requires further amplification of cDNA. Special amplification strategies are necessary due to the lack of universal primer sequences.

***3.2.1.2. Poly(A) tailing and related approaches.*** The first single-cell RNA-seq study was demonstrated by Tang et al. in 2009 (16). Now known as the Tang2009 protocol, this amplification method uses an oligo(dT) primer with an anchor sequence (UP1) to capture a single cell's mRNA transcripts. The anchor sequence provides a universal priming site for future amplification. Terminal deoxynucleotidyl transferase then adds poly(A) tails to the 3′-ends of first-strand cDNAs, so that a second oligo(dT) primer with a different anchor sequence (UP2) can initiate the second-strand cDNA synthesis. The resulting double-stranded cDNA can then be amplified by PCR with UP1 and UP2 as a pair of primers. The Tang2009 protocol introduced some important modifications to the original homomeric tailing strategy, which realized the transition of single-cell transcriptional analysis from microarray to RNA-seq. Those modifications include longer reverse transcription and PCR extension duration, as well as primer modification to eliminate end bias. The one-tube reaction style, which requires highly compatible buffers in all steps, was also faithfully followed to avoid loss of material during transition and purification and subsequently became a major consideration in many other protocols. Later studies have further modified the Tang2009 protocol to expand its application. Sasagawa et al. (60) added T7 promoter during first-strand synthesis, so that the amplified cDNA from PCR can be used for both sequencing and RNA microarray (Quartz-seq). SUPeR-seq supplemented random oligomers to the first-strand synthesis reaction to also simultaneously amplify nonpoly(A)-tailed linear transcripts and circular RNAs (35).

***3.2.1.3. Template-switching approaches.*** The homomeric tailing strategy offers stable performance and has helped produce some of the most elegant single-cell RNA-seq data to date.

However, the strategy requires many reagent addition steps, making it labor intensive and error prone. The intrinsic template-switching property of MMLV reverse transcriptase naturally tails the first-strand cDNA with three to six cytosines upon the completion of synthesis. A helper oligo tailed by three Gs can further direct the incorporation of a unique sequence at the 3′-end of first-strand cDNAs. The template-switching strategy was first implemented as single-cell tagged reverse transcription (STRT) for RNA-seq (STRT-seq) (61). The Sandberg group (14) later adapted the same strategy and introduced a protocol called Smart-Seq. Although the template-switching mechanism simplified the experimental operation, its mRNA-to-cDNA conversion efficiency is suboptimal, mainly due to the low thermal stability between the short stretch of G-C pairs. Both STRT-seq and Smart-Seq used three guanosines in their template-switching oligos (TSOs), which do not bind strongly enough to the newly synthesized cDNA tail under reverse transcription conditions. To solve this problem, the more recent Smart-Seq2 protocol exchanged a guanosine for a locked nucleic acid (LNA) guanosine at the TSO 3′-end (62), leading to much stronger binding affinity between the TSO and cDNA tails due to increased thermal stability of LNA:DNA base pairs (1–8°C per LNA monomer). Other major modifications implemented in Smart-Seq2 include the addition of the methyl group donor betaine and increased $Mg^{2+}$ concentration in the reaction buffer. Because template switching happens mainly at the end of cDNA, Smart-Seq2 demonstrated satisfactory full-length transcript amplification. The distribution bias within the transcripts is also greatly reduced. Given the ease of implementation and reduced end bias, Smart-Seq2 has become the most widely used single-cell whole-transcriptome amplification method. It is now available as both open access and commercially available kits.

**3.2.2. cDNA amplification.** Minute amounts of cDNA have to be amplified after reverse transcription. Typical methods include PCR, in vitro transcription (IVT), and rolling circle amplification (RCA).

***3.2.2.1. PCR amplification.*** The homomeric tailing strategy and template-switching mechanism enable the incorporation of universal anchor sequences at both ends of cDNA. PCR is the routine method to amplify the cDNA in many mature protocols (16, 43, 62). It is well known, however, that PCR efficiency is sensitive to G-C content and amplicon length. Excess rounds of PCR inevitably introduce bias to the gene expression profile (63).

***3.2.2.2. Linear amplification.*** IVT employs linear amplification, which leads to less bias (37, 44, 64, 65). It requires the promoter sequence at only one end of cDNA, which can be incorporated during first-strand synthesis. Second-strand synthesis can be initiated by short RNA primers, as mentioned above. CEL-Seq (64) and its successor CEL-Seq2 (65), as well as MARS-seq (37), employ this strategy to minimize bias during amplification. However, an additional round of reverse transcription is required after IVT and reintroduces 3′-bias.

***3.2.2.3. Whole-genome amplification.*** Full-length double-stranded cDNA can also be amplified by whole-genome amplification methods. Pan et al. (66) demonstrated two amplification methods for cDNA made with a short RNA priming strategy. The first method circularized the cDNA and performed hyperbranch RCA, whereas the second method used semirandomly primed PCR to amplify the overlapping segments along cDNA. These two methods can be versatile because they require no anchor sequence within cDNA. However, the first method inevitably suffers from low circularization efficiency as well as the amplification bias of Phi29 DNA polymerase, whereas the second method introduces bias by random priming. Recent development of emulsion multiple displacement amplification (67, 68) may alleviate the bias problem.

**3.2.3. Library preparation.** Currently, all next-generation sequencers require some form of library generation from the sample before the samples can be sequenced. Based on the selected experimental design, either the entire full-length cDNA or only parts of the cDNA molecule are made ready for downstream sequencing. In this section, we discuss the advantages and disadvantages of sequencing the full-length cDNA compared with sequencing only the cDNA ends.

***3.2.3.1. Full-length cDNA.*** Amplified full-length cDNA can be processed in the same way as genomic DNA for NGS library preparation. Although most of the studies use Illumina platforms, similar protocols can also generate libraries for other sequencers. The classic library construction method first breaks the DNA into smaller fragments using ultrasound or hydroshearing, followed by end repair to produce blunt and phosphorylated DNA molecules. Sequencing adaptors can be ligated to the ends by A-T or blunt-end ligation (16). For a large number of single cells, however, shearing individual samples becomes impractical. A Tn5 transposase-based method combines the fragmentation and adaptor ligation into a single step (69). Most single-cell RNA-seq studies now employ such a method to achieve high throughput with high reproducibility.

***3.2.3.2. End counting.*** Despite the reduction of sequencing cost, full-transcriptome analysis of over 1,000 single cells still costs a fortune. As most of the single-cell RNA-seq studies focus on transcript counting, it is a reasonable compromise to sequence only the transcript ends. The end-counting strategy also circumvents the 5′-3′ coverage bias. Special combinations of primers can selectively amplify the end regions of cDNA. CEL-Seq focuses on 3′-ends, whereas STRT-Seq examines 5′-ends of the transcripts.

## 3.3. Technological Development to Facilitate Single-Cell RNA-Seq

Even after having selected one of the various aforementioned experiment protocols, researchers still have many different choices, some of which are shown in **Figure 2**, to decide how to implement the protocols. From traditional manual pipetting to high-tech devices, each technology has its milieu.

**3.3.1. Traditional manual pipetting and mouth pipetting.** In some ways, our eyes and hands are our most responsive and reliable tools. Micropipetting ensures the lowest doublet contamination because the whole process is under direct visual examination. A typical setup consists of an inverted microscope and a mouth pipette. Although pneumatic pipette controllers are available from many brands, some skilled researchers still prefer using their mouths. Aside from visual examination, micropipetting has the flexibility to allow extensive cell washes: Picked single cells can be first injected into large buffer drops to wash off extracellular contamination. This washing step is essential to ensure data quality because cell rupture is inevitable during tissue dissociation, and excess nucleic acids together with other cellular contents are released from neighboring cells. Single cells are especially sensitive to even minute amounts of such contamination. Up to three rounds of washing are typically implemented during experiments. Micropipette picking does suffer from obvious limitations, such as low throughput and a lack of sorting capability. Well-trained personnel can pick up to 100 single cells in an hour. Certain cell surface stains can be incorporated to eliminate dead cells, but sophisticated surface marker interrogation is not possible. Therefore, early-stage embryos are the ideal candidate application, as these cells are present in low numbers but compose highly pure populations (16, 18, 70). After the cells are dispensed into individual PCR wells, subsequent experiments are typically carried out with manual pipetting.

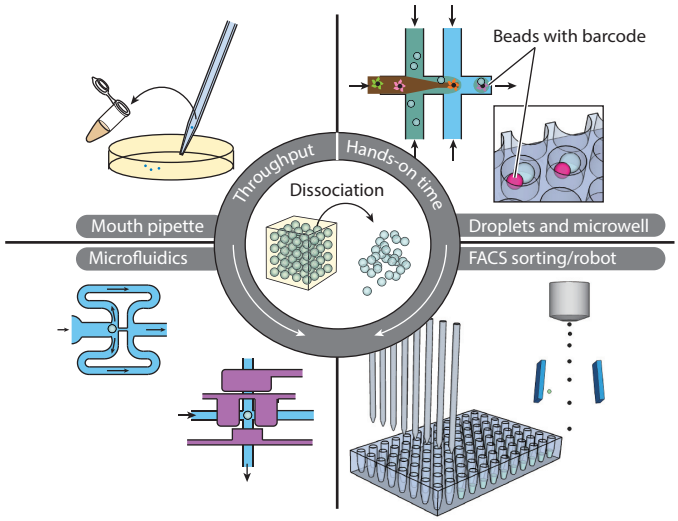


**Figure 2**

Techniques for single-cell capture and segregation. These four commonly used technologies for sorting single cells have trade-offs in experimental labor and throughput. Hands-on time increases in the clockwise direction starting from the top right quadrant, and throughput increases in the counterclockwise direction beginning from the top left quadrant. Sorting single cells with a mouth pipette, for example, is the most labor intensive and has the least throughput. Abbreviation: FACS, fluorescence-activated cell sorting.

**3.3.2. Lab automation: cytometers and liquid-handling robots.** When the cell population has high levels of heterogeneity, or when higher throughput is desired, automation comes into play. With the capacity to scan thousands of cells per second with multiple spectroscopic channels, and with single-cell dispensing capability, fluorescence-activated cell sorting (FACS) is the most widely used technology in single-cell genomics. High throughput does come with compromises to other aspects. FACS typically requires hundreds of thousands of cells as input, which is difficult for precious samples such as embryos and biopsies. Doublet contamination is a significant problem for single-cell analysis using flow cytometers, even with extensive scattering examination. High pressure and shear force within the flow stream may also damage the cells and adversely alter the transcription profile. To address some of these issues, FACS can be used in conjunction with manual picking: When a small number of high-purity single cells are needed, cells of interest can be selected based on surface marker staining, followed by manual washing and dispensing. Further processing of a large number of single cells also requires automation. Liquid-handling robots, especially those with submicroliter accuracy, are standard instruments for large genome centers. A growing desire for smaller reaction volumes has sparked the wave of replacements of liquid-handling robots; piezo-controlled printing inkjet robots, which can dispense nanoliters of reagents, are a promising new development in the field. Auxiliary equipment such as humidity controllers may also be necessary for working with submicroliter volumes to avoid premature and unwanted liquid evaporation. To the best of our knowledge, there is currently no complete commercially available solution for such demands.

**3.3.3. Microfluidic reactors.** To control even smaller volumes of reagents, researchers turned to microfluidics. Originally developed for analytical chemistry, microfluidic chips can manipulate nanoliters or even picoliters of reagents. Large-scale microfluidic integration has enabled high-throughput parallel genomic experiments, such as cell isolation, nucleic acid capture, reverse

transcription, and DNA amplification. The major advantage of microfluidics comes with its small reaction volume, typically three orders of magnitude smaller than bench-top reactions. With the same amount of start material in a single cell, the template concentration is effectively three orders of magnitude higher, which means a higher reaction rate and yield in the microfluidic device. As such, more transcripts can be captured, and more single-stranded cDNA can be converted into double-stranded DNA than in bench-top reactions. Smaller reaction volumes naturally produce smaller amounts of products. Fortunately for microfluidics, downstream applications such as library construction do not need more than nanograms of DNA. Reagent cost is inherently much lower for microfluidics than for bench-top reactions. Direct comparisons between microfluidic and bench-top single-cell whole-transcriptome amplifications with the same chemistry have shown that more genes can be reproducibly detected in the microfluidic group (71). At least in the single-cell genomics field, microfluidics turns small into big. Microfluidics also reduces the amount of hands-on work, which is the largest source of irreproducibility. Accordingly, microfluidic sample preparation produces data with higher correlation among technical replicates than those produced using bench-top reactions. Microfluidics is certainly not perfect; this method has a rather long learning curve for most life scientists. Currently, the ability to design, fabricate, and operate sophisticated microfluidic devices is still confined to mostly engineering laboratories and a very small proportion of biology laboratories. Commercially available microfluidic devices are much more user friendly, but the high prices of such devices and auxiliary instruments counterbalance all potential savings from reduced reagents. Given that microfluidics does not offer much higher throughput than FACS, it is easy to understand why the adoption of microfluidics has been limited in comparison to FACS.

**3.3.4. The quest for ultrahigh throughput.** With increasing interest in studying highly heterogeneous cell populations, such as those from blood, brain, and tumors, there is an unprecedented desire to perform RNA-seq on a very large number of single cells. Recent studies routinely report data from more than 1,000 individual cells, which is within the capacity of the aforementioned technologies but very labor intensive and expensive. In the latest high-throughput microfluidic single-cell preparation device design, 800 cells are labeled with 40 barcodes so that cDNA from every 40 cells can be pooled prior to library construction. It has become clear to researchers that, to further improve the sample processing throughput and to reduce cost, sample pooling is necessary and should be implemented as early as possible in the experimental protocol. Jaitin et al. (37) barcoded each well in a 384-well plate so that every 192 barcoded single cells could be pooled before second-strand synthesis. With this barcoding and pooling approach, the team managed to sequence over 4,000 single cells or approximately 10 such 384-well plates. Despite being highly efficient, barcoding single wells requires extensive, laborious pipetting, which was achieved by a liquid-handling robot in the original publication (37). Fan et al. (72) developed CytoSeq, which spreads cells and barcoded beads into microwells to further simplify the barcoding process. The microwells were carefully designed to accommodate only one bead and one cell. Transcripts from each cell were annealed to separate beads before being pooled together for downstream reverse transcription and amplification. The technology can easily be scaled up to process tens of thousands of single cells with more microwells. An automated version was reported recently with improved efficiency and reduced cross-contamination risk (46). To be completely free from physical wells, two teams simultaneously developed water-in-oil droplet-based methods, termed Drop-Seq and inDrop RNA-seq. Both methods combined single cells with single barcoded beads using microfluidic droplet generators. Over 10,000 single cells were profiled in each study.

Despite the excitement generated over this technology, the barrier to entry for microfluidics remains high for many life scientists, and the requirements to fabricate the devices and assemble

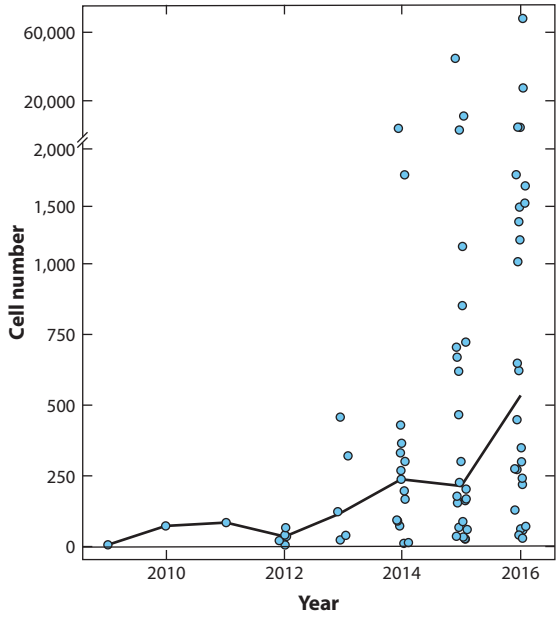


**Figure 3**

Number of single cells analyzed in published studies. Each dot represents a published study utilizing single-cell RNA-seq. The solid line connects the median number of cells per study in each year.

the system keep many researchers away. To alleviate these challenges, the Drop-Seq team established an online forum to further facilitate the adoption of their platform system. Another recent study has demonstrated the incorporation of droplet-based single-cell barcoding with a commercially available platform (45). Simpler operation and guaranteed reproducibility from commercial kits will likely promote the acceptance of droplet-based single-cell RNA-seq among life science researchers. Driven by the rapid development of new methods, the number of single-cell RNA-seq–based studies has significantly increased since 2009. **Figure 3** plots the numbers of single-cell analyses in some representative studies. The increasing numbers of single cells in a study are dramatic. In contrast, most studies were still based on fewer than 1,000 cells. This could be due to the limit of precious samples, such as circulating tumor cells and embryos, and the time gap between the introduction and adoption of ultrahigh-throughput methods.

## 4. TECHNICAL PERFORMANCE OF SINGLE-CELL RNA-SEQ

For any quantitative measurement, it is crucial to thoroughly evaluate technical performance so that the confidence with which a measurement is made can be stated alongside the measurement itself. This is especially important in single-cell RNA-seq experiments that provide such rich and complex data sets that the fundamental limitations of the technology can sometimes be overlooked. Three of the most essential technical specifications in all measurements, which are equally crucial to the interpretation of RNA sequencing data, are sensitivity, accuracy, and precision. Although these metrics are typically well understood, they are worth revisiting in the context of single-cell RNA-seq, as they are critical to the understanding, analysis, and interpretation of single-cell RNA-seq results.

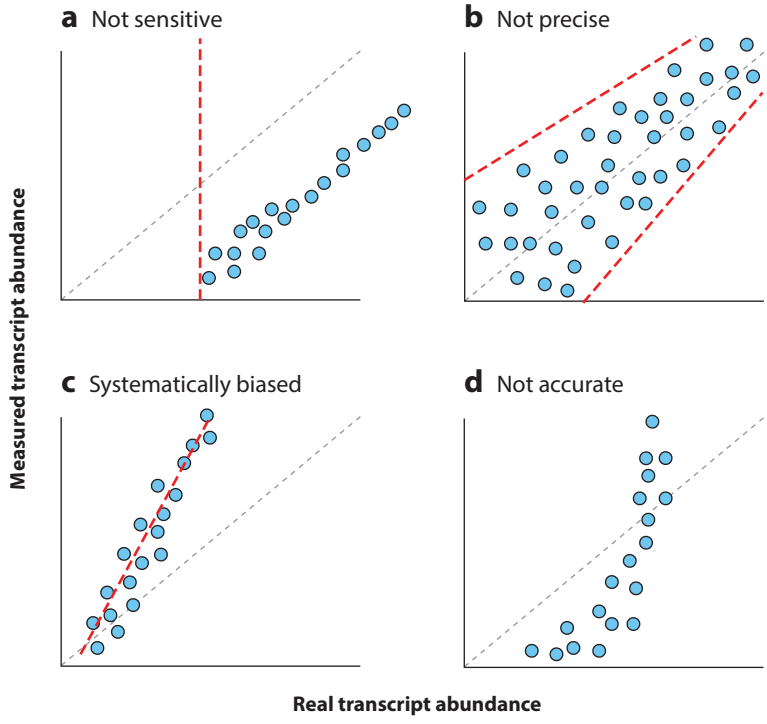


**Figure 4**

A schematic representation of the basic metrics for technical performance in single-cell RNA-seq. (*a*) Sensitivity is determined by the transcript detection limit, represented by the vertical red dashed line. The low sensitivity depicted here can indeed be a form of systematic bias, but it can be remedied with a multiplicative factor. (*b*) Low precision yields a large coefficient of variation in the estimation of mean expression levels. The two red dashed lines represent confidence intervals of the measurement. (*c*) Systematic bias can lead to an inaccurate measurement. The systematic bias depicted here can be compensated with a linear transformation of the red dashed line. (*d*) Unreported or unknown bias can lead to severely inaccurate estimates of gene expression level. Abbreviation: RNA-seq, RNA transcriptional sequencing.

## 4.1. Sensitivity

Sensitivity in single-cell RNA-seq generally refers to the transcript abundance detection limit or the lowest number of RNA transcripts that can be reproducibly detected (**Figure 4*a***). At the single-molecule level, the detection rate of one RNA transcript can be considered as the quantum efficiency of the measurement (71). This metric has been shown to rely heavily on sequencing depth up to about one million reads per cell (71, 73). Beyond a sufficient sequencing depth, however, sensitivity is an intrinsic limitation that is highly dependent on the protocol itself, as mentioned in the previous section. One common and useful interpretation of sensitivity is the transcript capture efficiency or the percentage of RNA transcripts that is counted by the measurement. This is a practical assessment of sensitivity because it asks: What portion of the cell's RNA are we measuring? Grün et al. (53) used single-molecule fluorescence in situ hybridization (FISH) as a reference and reported a capture efficiency of approximately 10% for the CEL-Seq protocol. We have measured as high as 40% capture efficiency of known spike-in controls (71). The total number of reproducibly detected genes is a metric that is often used as a convenient surrogate for sensitivity because it is straightforward to measure. However, this metric strongly depends on cell type.

### 4.2. Accuracy

Accuracy in single-cell RNA-seq refers to the extent to which the final expression measurement recapitulates the actual transcript abundance distribution. Accuracy can be limited by sources intrinsic to the protocol that introduce systematic bias, such as exponential PCR amplification or sequence-dependent transcript capture efficiency. In certain cases, if systematic bias distorts the abundance measurement in a linear or measureable fashion, it can be calibrated and corrected for (**Figure 4*c***). In other cases, accuracy is limited by unreported or nonlinear bias (**Figure 4*d***).

### 4.3. Precision

Precision is inversely proportional to technical noise in the RNA-seq measurement. Technical noise is often defined as the coefficient of variation or the mean-normalized standard deviation of replicate expression measurements of the same sample (**Figure 4*b***). Precision is a critical metric for single-cell analysis because the goal is typically to measure real biological variation between individual cells. Nonetheless, precision is arguably the largest limitation of current single-cell RNA-seq protocols. We have taken advantage of microfluidic platforms to quantitatively compare and benchmark various single-cell RNA-seq protocols (71, 74). Microfluidic chambers provide a convenient platform to minimize experimental variation and perform technical replicate experiments to assess precision. Additionally, RNA transcript capture, reverse transcription, and amplification show improved detection efficiency in nanoliter reactors (71, 74).

## 5. QUANTIFICATION OF SINGLE-CELL RNA-SEQ

Although it is crucial to have a quantitative understanding of these fundamental performance metrics to properly interpret single-cell RNA-seq data, it is almost impossible to characterize these metrics in an absolute sense. As single-cell RNA-seq has increasingly become the gold standard for transcriptome-wide gene expression measurement, there are increasingly fewer competing technologies to benchmark small conditional RNA-seq against. Furthermore, every cell is a unique sample, so it is challenging to perform replicate experiments to genuinely assess reproducibility. Thus, the choice of measurement unit is another crucial decision for analysis of single-cell RNA-seq data.

### 5.1. Units of Expression

Choosing a robust measurement unit is important when comparing expression levels of thousands of genes between large numbers of single cells. It is also necessary to understand the advantages and limitations of the chosen measurement unit when assessing performance. In any NGS experiment, the fundamental measurement unit is the read. A read is a sequence word of length determined by the sequencing platform and represents a fragment of cDNA that was reverse transcribed from an RNA transcript and amplified. The number of reads that map to a given gene is the most basic measurement for the expression level of that gene and is often referred to as the raw read count. However, because genes have variable lengths, and because every cell is not always sequenced with the same depth, it is usually necessary to report a normalized unit of gene expression. A common normalized unit is reads per kilobase of transcript per million mapped reads (RPKM). RPKM is essentially a raw read count normalized to the transcript length and the sequencing depth. Fragments per kilobase of transcript per million mapped fragments (FPKM) (75) is a similar unit calculated by the Cufflinks tool that accounts for paired-end reads, in which case a

single fragment produces two reads. RPKM and FPKM units are slightly biased because reads are randomly distributed among cells during sequencing (76). The unit transcripts per million (TPM) is a widely used alternative to RPKM and FPKM because it avoids this bias. One drawback of these three units is that they are relative measurements of cDNA abundance in sequencing libraries. That is to say, RPKM, FPKM, and TPM do not measure the absolute transcript abundance; instead, they estimate the relative proportions of amplified cDNA per sample, which is a distorted representation of transcript distribution because of many rounds of PCR amplification.

## 5.2. Spike-In Controls

To assess technical performance of single-cell RNA-seq, it is necessary to have a positive control to compare with endogenous RNA transcript abundance measurements. ERCC RNA standards are a pool of synthetic transcripts with known sequence and abundance (56). When spiked into a single-cell RNA-seq experiment, they can provide this positive control. ERCC spike-in controls are a useful way to construct plots such as those depicted in **Figure 4**, and it has become routine to use ERCC spike-in controls to evaluate sensitivity, accuracy, and precision in single-cell RNA-seq experiments. Recently, Svensson and colleagues (73) conducted a meta-analysis of published ERCC spike-in controls used with a variety of single-cell RNA-seq protocols to compare accuracy and sensitivity across platforms. Interestingly, they found that ERCC quantification underestimates sensitivity of some single-cell RNA-seq protocols. Spike-in controls can also be used to perform regression-based normalization on raw read counts to produce estimates for absolute transcript abundance. However, the capture and reverse transcription of synthetic, cell-free ERCC transcripts does not accurately recapitulate the processing of endogenous RNA transcripts. Therefore, regression-based normalization with ERCC controls can potentially introduce unwanted bias.

A powerful technique to remove amplification bias and achieve absolute transcript counting is the incorporation of a random sequence of nucleotides to the reverse transcription primers (77). This barcode, often referred to as a unique molecular identifier (UMI), is reverse transcribed with the mRNA and incorporated into the cDNA. cDNA amplicons that originate from the same template will all contain this unique sequence. In this way, every read that contains an identical UMI can be collectively counted as a single transcript, and absolute gene expression can be estimated as the number of UMIs associated with a given gene. UMIs are an effective way of benchmarking performance because they are an absolute measurement (78). However, incorporation of barcodes can reduce capture efficiency (53) or lead to faster saturation of sequencing depth.

## 5.3. Normalization

In addition to variation in transcript length and sequencing depth, there are a variety of other experimental sources of systematic bias in single-cell RNA-seq. Fortunately, there is a large community of computational biologists, mathematicians, and computer scientists who have contributed to a growing body of literature and computational tools designed to address normalization and interpretation of single-cell RNA-seq data. Normalization is a crucial step in single-cell RNA-seq data analysis, particularly if data sets are being compared across experimental batches (79, 80). Systematic variation can arise from every step of the RNA-seq processing pipeline, and unwanted variation can even come from confounding biological factors, such as cell cycle or state. The goal of normalization is to remove systematic variation between cells and between experimental batches through statistical analysis of endogenous transcript expression or spike-in standards.

Risso et al. (81) showed that, whereas ERCC controls might not be sufficient for regression-based normalization, they can be used as an internal standard to model and remove unwanted

variation. They provide a tool called RUV (remove unwanted variation) that has been incorporated into a comprehensive normalization package in R called SCONE devised by Nir Yosef's group. Buettner et al. (82) used a latent variable model in conjunction with ERCC controls to remove biological variation that arises from the cell cycle. BASiCS is a normalization algorithm from Vallejos et al. (83) that uses a Bayesian hierarchical model to quantify variability, again with the aid of ERCC controls.

Limited precision and sensitivity in single-cell RNA-seq can lead to frequent dropout events, in which a transcript is not captured and is thus assigned a false zero value. High dropout rates associated with variability in sequencing depth or capture efficiency can lead to significant systematic bias in mean expression measurements. Lun et al. (84) address this issue by first pooling single-cell measurements and aggregating expression levels to remove false zero counts. The benefit of their normalization approach is that is does not rely on spike-in controls.

## 5.4. Noise

Unlike systematic bias that directly affects the accuracy of expression level estimation, technical noise in single-cell RNA-seq refers to the precision of the measurement. The analysis of hundreds of single-cell transcriptomes revealed a great deal of heterogeneity between individual cells. However, due to the low quantity of starting material, single-cell RNA-seq data are inherently more noisy than those of their bulk counterparts (71, 85, 86). To correctly interpret the heterogeneity present in single-cell transcriptomic data sets, it is therefore important to distinguish the three components of variability within them: (*a*) technical noise, arising from the random sampling of cellular transcripts during conversion from RNA to cDNA, from differences in initial transcript concentration, and from fluctuations in sample-to-sample sequencing efficiency; (*b*) biological noise, which is the variability of transcript abundance between cells of the same biological subtype, arising from stochasticity of gene expression such as transcriptional bursting; and (*c*) true biological variation, which marks the difference between distinct cellular subtypes and is the main component of variation that interests researchers. Several approaches to quantify and deconvolve noise from single-cell RNA-seq data have been described. The use of UMIs can help eliminate amplification biases (78), but this only accounts for part of the technical noise; variation due to inefficiencies in reverse transcription or sampling are not accounted for using this method. In another approach, the technical noise component is addressed by assessing the sample-to-sample variability of known quantities of RNA spike-in added to each single-cell experiment (71, 85). Compared to the expected quantity of each spiked-in transcript, the sample-to-sample variation observed for each transcript can be used to construct a model of the technical noise. The technical noise was found to be largely dependent on the mean level of gene expression: Highly expressed genes typically display low levels of technical noise, whereas transcripts with low abundance usually show high levels of technical noise (71, 85). Most recent studies employing single-cell RNA-seq analyzed hundreds of transcriptomes, providing sufficient statistical power to allow modeling of technical noise using the set of 92 ERCC spike-ins (11, 71, 85). If the data set is small and includes only a few cells, however, a pooled total RNA sample is recommended as the spike-in to provide a more robust noise model, with the caveat that total RNA spike-ins will take up a significant proportion of total sequenced reads (85). Once the technical noise is deconvolved, the biological noise can be inferred, typically by fitting to a negative binomial distribution that models transcriptional bursting of gene transcription (86, 87). To determine the interesting biological variation between cellular subpopulations, further analysis of the variation is required. For example, we can identify genes that are highly variable as defined by their $p$-value in the context of the overall biological noise of the experiment (85) or directly infer the distribution of biological noise using

a deconvolution of negative binomial distributions (53). Alternatively, Bayesian methods can be applied to model the measurement of each cell as a mixture of amplified transcripts and dropout or nontranscribed genes, where the mixing ratio of the two components is dependent on the gene expression over all cells being analyzed from the cell population in question (86).

## 6. INTERPRETATION AND VISUALIZATION OF UNDERLYING BIOLOGY

Many methods are used to glean biological insight from single-cell transcriptomic data, and these computational tools and packages are readily available (88–93). The challenge more often lies in the appropriate selection and application of tools based on the data being analyzed and the biological question at hand. Below, we discuss several considerations in the selection of data analysis tools and approaches.

### 6.1. Dimensionality Reduction

Analysis of single-cell RNA-seq data sets is computationally challenging because each cell is represented by over 30,000 attributes (i.e., the expression level of each of the 30,000 genes), and often each data point must be compared to every other data point in every other cell. For this reason, dimensionality reduction is a critical first step in the analysis of these data sets. Dimensionality reduction can also help to reveal the most striking variations and differences within the data and allow downstream visualization and intuitive interpretation.

**6.1.1. Principal component analysis and independent component analysis.** Principal component analysis (PCA) can be used for dimensional reduction by identifying axes, or principal components, that contain the greatest amount of variance in the data. This essentially reveals the sets of attributes, in this case genes, which contribute the most variance between cells. The process can be very useful in highlighting important features in the higher dimension that could be difficult to visualize directly (9, 82). Applications of PCA should be performed with care because certain implementations or algorithms assume that variables in the data are linearly related, whereas this is often not the case. Therefore, data with nonlinear variables that have been embedded using PCA cannot be interpreted using the classical intuition. In these situations, independent component analysis is a potential solution (12) in addition to t-distributed stochastic neighbor embedding (t-SNE). PCA is often followed by clustering analysis to further identify sets of genes that define cellular subgroups and the biological function and significance of each group.

**6.1.2. T-distributed stochastic neighbor embedding.** T-SNE is a machine learning algorithm for dimensional reduction that is superior to PCA in many aspects, including the ability to maintain similarities in local neighborhoods of data that may be important to the overall structure and dynamic of the data (94). Similar to PCA, it is an unsupervised method; unlike PCA, it does not require data variables to be linearly related. Pe'er and colleagues (95) created viSNE, which efficiently applies the t-SNE algorithm to biological high-dimensional data. It was used initially to study single-cell mass cytometry data from leukemia, then later used to analyze single-cell RNA-seq data (10, 43, 44, 83). T-SNE analysis can also be followed by other analyses to elucidate biological function.

### 6.2. Unsupervised Clustering Analysis

Clustering methods are among the first to be applied to large single-cell RNA-seq data sets and are also the most frequently used, as they are an established approach for bulk RNA-seq

analysis and are familiar to many researchers (9, 12, 91). The main goal of clustering is to separate individuals into subsets based on their similarity or distance between the data points. This allows heterogeneity within the population to be identified along with the genes that are responsible for these differences.

**6.2.1. Hierarchical clustering.** Advantages of clustering analysis are that there are usually no assumptions regarding the underlying distribution of the data, and they do not rely on existing cell markers. However, one should also be careful in the interpretation of the output. For example, given a data matrix, it is always possible to generate clusters even if no biologically meaningful grouping exists. Furthermore, distances between clusters on a dendrogram should be interpreted with care, taking note of the distance metric used. Thus, further statistical or functional validations are essential to confirm the biological significance of groupings.

**6.2.2. K-means clustering.** In k-means clustering, cells are placed into the cluster nearest to them based on their Euclidean distance from the cluster center (96). The number of clusters must be specified, which may pose a limitation if it is not known a priori how many cellular subtypes are expected. Furthermore, because k-means typically uses Euclidean distance, it is less useful if some other metric is preferred for measuring similarity between cells. In single-cell RNA-seq applications, it can be used to infer clusters for the construction of a network or tree indicating the progression of cell types along a trajectory, such as pseudotime or differentiation (12, 13, 75, 90).

## 6.3. Inference from Coexpression Gene Networks

The use of existing knowledge about biological networks and genes that work coordinately can help to enhance the discovery and interpretation of single-cell RNA-seq data sets. Buettner and colleagues (82) found that cell cycle differences are a potentially confounding factor of variation of gene expression within a biologically similar group of cells. Using previously annotated cell cycle genes, they removed variations correlating with cell cycle variation, which then allowed for higher confidence in the identification of other biological variation of interest.

Fan and colleagues (86) developed an analysis tool called PAGODA. In addition to looking for sources of variation in gene expression, this tool also identifies known pathways or novel gene sets that show coordinated variability over what is expected within the population of cells profiled. Arguably, this method is potentially more robust than using clustering methods alone, as it looks for patterns of variability in gene sets that have been annotated with biological significance. The use of gene sets also allows for identification of overlapping aspects of variation within the population, which is also not possible with hierarchical clustering.

## 6.4. Construction of Pseudotemporal Trajectories

Another fast-evolving approach is pseudotemporal trajectory construction. In many biological systems, cells do not necessarily occupy discrete states; instead, they exist in a continuum of states. This is the case, for example, in stem cell differentiation. These transitional states can be defined by the continuously evolving gene expression profile of single cells. In such circumstances, clustering does not capture the dynamic features of the transition. Tools such as Monocle (12) aim to arrange single cells in a pseudotimeline defined by their differentiation stage through ordering neighboring cells by the geometric distance of their gene expression profile. Construction of single-cell trajectories through the differentiation landscape reveals new markers for intermediate

states, shows branched pathways, and informs regulatory dynamics of differentiation. In addition to Monocle, this rapidly growing set of tools includes Wanderlust (97), Wishbone (92), Waterfall (90), and SCUBA (98). This genre of single-cell analysis tools may eventually drastically advance the field of single-cell transcriptomics.

## 7. CONCLUSIONS

Just a few years ago, when single-cell RNA-seq methodologies were first introduced, many scientists were skeptical about their feasibility in routine biological research due to a lack of understanding of their sensitivity, accuracy, precision, and practical limitations. Recently, the International Human Cell Atlas initiative was established, with the mission of creating a comprehensive reference map (atlas) of all human cells. The atlas aims to catalog every cell type in the body and characterize their location, function, and how they change over time. Undoubtedly, before the existence of single-cell RNA-seq technology, this initiative would have been impossible; now, as confidence in the technology is established, it is feasible. Just as high-throughput sequencing ignited the genomics era by enabling the sequencing of the first human genome, the advent of single-cell RNA-seq technology may prove to be the watershed moment in our understanding of human cellular diversity in health and disease.

In this review, we have provided a primer on the most popular methodologies to readers interested in single-cell RNA-seq technology and who may be interested in applying these approaches to their own research. We have discussed the main methodologies and techniques used in the field today and presented our view of the advantages and disadvantages of each. Despite the strong capabilities of the technologies described, there are still improvements that can be made. First, the suboptimal efficiency of capturing and sampling cellular transcripts is a bottleneck in single-cell RNA-seq sensitivity. Second, the instrumentation and devices available for assay automation and to improve throughput currently are inaccessible to many scientists due to their high cost or high technical learning curve. Third, the cost of sequencing, although greatly reduced from before, still forces the average user to choose between cell number and sequencing depth. Fourth, the computational modeling of confounding variables and technical variability in single-cell data analysis still has room for further improvement in accuracy. Another issue that was not discussed in this review is sample processing from primary tissues or cells into single-cell suspensions, which currently takes a long time and uses procedures that could affect cellular transcriptional profiles. There is great need to develop sample processing methods for frozen or fixed specimens or technological modalities for in situ or even in vivo profiling of transcriptomes from single cells. We hope that these discussions will help newcomers to the field with their experimental planning and decision-making and ultimately inspire them and others to make exciting discoveries and future developments in the field.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors thank Dr. Yusi Fu for creating **Figures 1**, **2**, and **3**. Work was supported by the Ministry of Science and Technology of China (2015AA0200601 to Y.H. and 2016YFC0900103 to J.W.), the National Natural Science Foundation of China (21525521 and 21327808 to Y.H.

and 21675098 to J.W.), and the Hong Kong University Grants Council and Hong Kong Research Grants Council (A.R.W.).

## LITERATURE CITED


1. Fritzsch FSO, Dusny C, Frick O, Schmid A. 2012. Single-cell analysis in biotechnology, systems biology, and biocatalysis. *Annu. Rev. Chem. Biomol. Eng.* 3:129–55
2. Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14(9):618–30
3. Sandberg R. 2014. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11(1):22–24
4. Junker JP, van Oudenaarden A. 2014. Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell* 157(1):8–11
5. Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16(3):133–45
6. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. 2015. The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58(4):610–20
7. Wang Y, Navin NE. 2015. Advances and applications of single-cell sequencing technologies. *Mol. Cell* 58(4):598–609
8. Leng N, Chu L-F, Barry C, Li Y, Choi J, et al. 2015. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* 12(10):947–50
9. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, et al. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509(7500):371–75
10. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, et al. 2015. A survey of human brain transcriptome diversity at the single cell level. *PNAS* 112(23):7285–90
11. Grün D, van Oudenaarden A. 2015. Design and analysis of single-cell sequencing experiments. *Cell* 163(4):799–810
12. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, et al. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32(4):381–86
13. Treutlein B, Lee QY, Camp JG, Mall M, Koh W, et al. 2016. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* 534(7607):391–95
14. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30(8):777–78
15. Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, et al. 2015. RNA-seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* 349(6254):1351–56
16. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6(5):377–82
17. Blainey PC, Quake SR. 2013. Dissecting genomic diversity, one cell at a time. *Nat. Methods* 11(1):19–21
18. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, et al. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* 6(5):468–78
19. Yan L, Yang M, Guo H, Yang L, Wu J, et al. 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20(9):1131–39
20. Durruthy-Durruthy J, Wossidlo M, Pai S, Takahashi Y, Kang G, et al. 2016. Spatiotemporal reconstruction of the human blastocyst by single-cell gene-expression analysis informs induction of naive pluripotency. *Dev. Cell* 38(1):100–15
21. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, et al. 2016. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165(4):1012–26
22. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, et al. 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510(7505):363–69
23. Tan L, Li Q, Xie XS. 2015. Olfactory sensory neurons transiently express multiple olfactory receptors during development. *Mol. Syst. Biol.* 11(12):844–44

24. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, et al. 2016. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.* 17(1):67
25. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252(5013):1651–56
26. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, et al. 1996. A gene map of the human genome. *Science* 274(5287):540–46
27. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291(5507):1304–51
28. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921
29. Modrek B, Lee C. 2002. A genomic view of alternative splicing. *Nat. Genet.* 30(1):13–19
30. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5(7):621–28
31. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10(1):57–63
32. Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–11
33. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, et al. 2014. CircRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* 56(1):55–66
34. Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L. 2014. Complementary sequence-mediated exon circularization. *Cell* 159(1):134–47
35. Fan X, Zhang X, Wu X, Guo H, Hu Y, et al. 2015. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* 16:148
36. Dang Y, Yan L, Hu B, Fan X, Ren Y, et al. 2016. Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biol.* 17(1):1–15
37. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, et al. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343(6172):776–79
38. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, et al. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226):1138–42
39. Nelson AC, Mould AW, Bikoff EK, Robertson EJ. 2016. Single-cell RNA-seq reveals cell type-specific transcriptional signatures at the maternal-foetal interface during pregnancy. *Nat. Commun.* 7:11414
40. Björklund ÅK, Forkel M, Picelli S, Konya V, Theorell J, et al. 2016. The heterogeneity of human CD127$^{+}$ innate lymphoid cells revealed by single-cell RNA sequencing. *Nat. Immunol.* 17(4):451–60
41. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, et al. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344(6190):1396–401
42. Zhou F, Li X, Wang W, Zhu P, Zhou J, et al. 2016. Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* 533(7604):487–92
43. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161(5):1202–14
44. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161(5):1187–201
45. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049
46. Yuan J, Sims PA. 2016. An automated microwell platform for large-scale single cell RNA-Seq. *Sci. Rep.* 6:33883
47. Treutlein B, Gokce O, Quake SR, Südhof TC. 2014. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *PNAS* 111(13):E1291–99
48. Macaulay IC, Svensson V, Labalette C, Ferreira L, Hamey F, et al. 2016. Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.* 14(4):966–77
49. Streets AM, Huang Y. 2014. How deep is enough in single-cell RNA-seq? *Nat. Biotechnol.* 32(10):1005–6
50. Drissen R, Buza-Vidas N, Woll P, Thongjuea S, Gambardella A, et al. 2016. Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat. Immunol.* 17(6):666–76

51. Liu Y, Zhou J, White KP. 2014. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* 30(3):301–4
52. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15(2):121–32
53. Grün D, Kester L, van Oudenaarden A. 2014. Validation of noise models for single-cell transcriptomics. *Nat. Methods* 11(6):637–40
54. Munro SA, Lund SP, Pine PS, Binder H, Clevert D-A, et al. 2014. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.* 5:5125
55. Lee H, Pine PS, McDaniel J, Salit M, Oliver B. 2016. External RNA controls consortium beta version update. *J. Genom.* 4:19–22
56. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, et al. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 21(9):1543–51
57. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22(6):839–51
58. Peixoto L, Risso D, Poplawski SG, Wimmer ME, Speed TP, et al. 2015. How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res.* 43(16):7664–74
59. Gubler U. 1987. Second-strand cDNA synthesis: mRNA fragments as primers. *Methods Enzymol.* 152:330–35
60. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, et al. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14(4):R31
61. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, et al. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21(7):1160–67
62. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg GOS, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10:1096–98
63. Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *PNAS* 109(4):1347–52
64. Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2(3):666–73
65. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, et al. 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17(1):77
66. Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, et al. 2012. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *PNAS* 110(2):594–99
67. Fu Y, Li C, Lu S, Zhou W, Tang F, et al. 2015. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *PNAS* 112(38):11923–28
68. Fu Y, Chen H, Liu L, Huang Y. 2016. Single cell total RNA sequencing through isothermal amplification in picoliter-droplet emulsion. *Anal. Chem.* 88(22):10795–99
69. Picelli S, Björklund ÅK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* 24(12):2033–40
70. Tang F, Lao K, Surani MA. 2011. Development and applications of single-cell transcriptome analysis. *Nat. Methods* 8(4):S6–11
71. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, et al. 2014. Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11(1):41–46
72. Fan HC, Fu GK, Fodor SPA. 2015. Combinatorial labeling of single cells for gene expression cytometry. *Science* 347(6222):1258367
73. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, et al. 2017. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14:381–87
74. Streets AM, Zhang X, Cao C, Pang Y, Wu X, et al. 2014. Microfluidic single-cell whole-transcriptome sequencing. *PNAS* 111(19):7048–53

75. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28(5):511–15
76. Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 131(4):281–85
77. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, et al. 2014. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *PNAS* 111(5):1891–96
78. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, et al. 2013. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11:163–66
79. Hicks SC, Teng M, Irizarry RA. 2015. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. bioRxiv 025528. **http://dx.doi.org/10.1101/025528**
80. Tung P-Y, Blischak JD, Hsiao C, Knowles DA, Burnett JE, et al. 2017. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 8:39921
81. Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32(9):896–902
82. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, et al. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33(2):155–60
83. Vallejos CA, Marioni JC, Richardson S. 2015. BASiCS: Bayesian analysis of single-cell sequencing data. *PLOS Comput. Biol.* 11(6):e1004333
84. Lun ATL, Bach K, Marioni JC. 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol*. 17(1):75
85. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, et al. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10(11):1093–95
86. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, et al. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13(3):241–44
87. Deng Q, Ramsköld D, Reinius B, Sandberg R. 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343(6167):193–96
88. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. 2015. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLOS Comput. Biol.* 11(11):e1004575
89. Juliá M, Telenti A, Rausell A. 2015. *Sincell*: An R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. *Bioinformatics* 31(20):3380–82
90. Shin J, Berg DA, Zhu Y, Shin JY, Song J, et al. 2015. Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell* 17(3):360–72
91. Žurauskienė J, Yau C. 2016. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform*. 17(1):140
92. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, et al. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34(6):637–45
93. Pierson E, Yau C. 2015. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol*. 16(1):241
94. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579–605
95. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, et al. 2013. ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* 31(6):545–52
96. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, et al. 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525(7568):251–55
97. Bendall SC, Davis KL, Amir ED, Tadmor MD, Simonds EF, et al. 2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157(3):714–25
98. Marco E, Karp RL, Guo G, Robson P, Hart AH, et al. 2014. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *PNAS* 111(52):E5643–50

Annual Review of Analytical Chemistry

Volume 10, 2017

# Contents