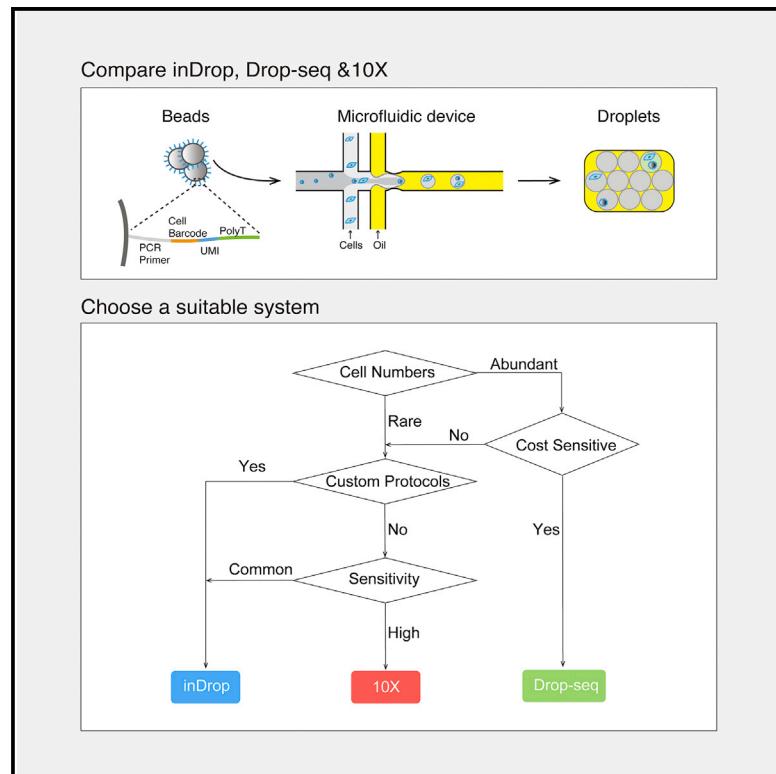


Molecular Cell

Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems

Graphical Abstract



Authors

Xiannian Zhang, Tianqi Li, Feng Liu, ..., Zeyao Li, Yanyi Huang, Jianbin Wang

Correspondence

yanyi@pku.edu.cn (Y.H.), jianbinwang@tsinghua.edu.cn (J.W.)

In Brief

Zhang et al. compare three prevalent droplet-based high-throughput scRNA-seq systems using unified sample and bioinformatics pipeline. They provide detailed analyses on system designs and performance, which would guide both future experimental design and system improvement.

Highlights

- Comprehensive diagrams for comparing the features of the three systems
- An open source versatile pipeline for all systems
- Systematic comparison on sensitivity, precision, bias, and costs
- Demonstration of Smart-seq2 protocols with inDrop platform



Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems

Xiannian Zhang,^{1,6} Tianqi Li,^{2,6} Feng Liu,^{3,6} Yaqi Chen,^{4,6} Jiacheng Yao,² Zeyao Li,⁵ Yanyi Huang,^{1,*} and Jianbin Wang^{2,7,*}

¹Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, College of Engineering, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

²School of Life Sciences, and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China

³National Research Center for Translational Medicine (Shanghai), State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

⁴Peking University Shenzhen Graduate School, Shenzhen 518055, China

⁵School of Life Sciences, and Peking-Tsinghua-NIBS Joint Graduate Program, Tsinghua University, Beijing 100084, China

⁶These authors contributed equally

⁷Lead Contact

*Correspondence: yanyi@pku.edu.cn (Y.H.), jianbinwang@tsinghua.edu.cn (J.W.)

<https://doi.org/10.1016/j.molcel.2018.10.020>

SUMMARY

Since its establishment in 2009, single-cell RNA sequencing (RNA-seq) has been a major driver behind progress in biomedical research. In developmental biology and stem cell studies, the ability to profile single cells confers particular benefits. Although most studies still focus on individual tissues or organs, the recent development of ultra-high-throughput single-cell RNA-seq has demonstrated potential power in characterizing more complex systems or even the entire body. However, although multiple ultra-high-throughput single-cell RNA-seq systems have attracted attention, no systematic comparison of these systems has been performed. Here, with the same cell line and bioinformatics pipeline, we developed directly comparable datasets for each of three widely used droplet-based ultra-high-throughput single-cell RNA-seq systems, inDrop, Drop-seq, and 10X Genomics Chromium. Although each system is capable of profiling single-cell transcriptomes, their detailed comparison revealed the distinguishing features and suitable applications for each system.

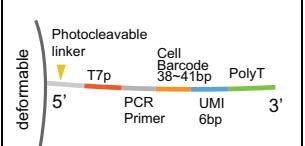
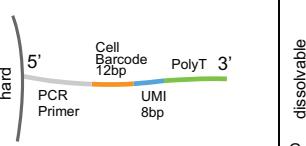
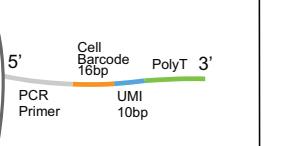
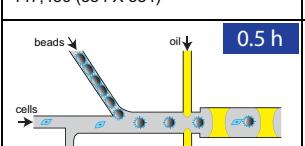
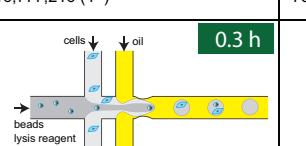
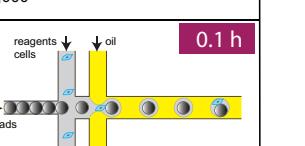
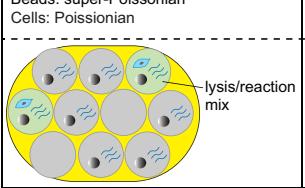
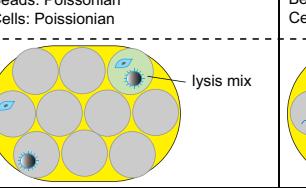
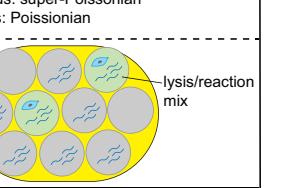
INTRODUCTION

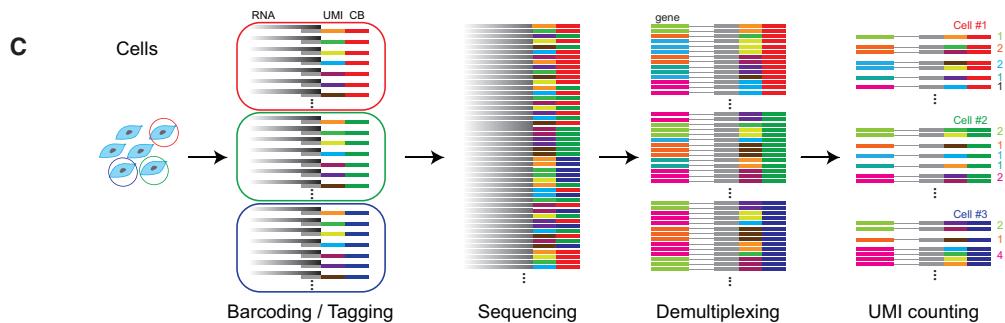
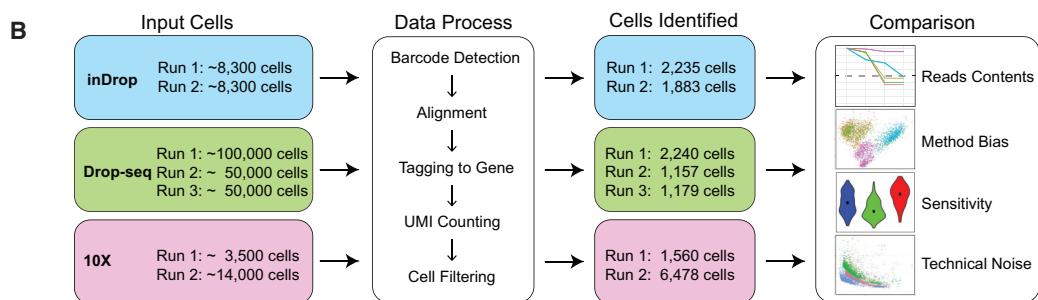
Single-cell RNA sequencing (scRNA-seq), which was first established in 2009 (Tang et al., 2009), has become one of the most powerful approaches for revealing biological heterogeneity. The ability to manipulate picograms of RNA in single cells has enabled the performance of studies with unprecedented temporal and spatial resolution. Based on the substantial data of the whole transcriptome, scRNA-seq has provided comprehensive information on landscapes of gene expression and their regulatory interactions at the finest resolution, enabling accurate and precise depiction of cell types and states (Grün and van Oude-

naarden, 2015; Tanay and Regev, 2017; Wu et al., 2017). In the last decade, the sensitivity and precision of mRNA quantification through scRNA-seq have been greatly improved (Hashimshony et al., 2016; Picelli et al., 2014), leading to revolutionary discoveries in many fields, such as cell type identification in various tissues or organs (Jaitin et al., 2014; Lake et al., 2016; Papalexi and Satija, 2018; Treutlein et al., 2014; Villani et al., 2017); tracing cell lineage and fate commitment in embryonic development and cell differentiation (Olsson et al., 2016; Semrau et al., 2017; Tirosh et al., 2016; Yan et al., 2013); drawing inferences on transcriptional dynamics and regulatory networks (Deng et al., 2014; Dixit et al., 2016); and identifying the development, evolution, and heterogeneity of tumors (Patel et al., 2014; Treutlein et al., 2014; Venteicher et al., 2017).

The experimental throughput is always a major concern in the design of scRNA-seq experiments. In some biological systems, such as early-stage embryos, only dozens of cells are required to achieve critical findings (Yan et al., 2013). However, owing to tissue complexity, the dynamicity of the cell cycle, or other intrinsic variations (Buetchner et al., 2015), as well as technical noise (Brennecke et al., 2013), RNA-seq data from a small number of cells are typically inadequate to reflect the state of biological samples comprehensively (Tanay and Regev, 2017). The sensitivity of transcriptome detection is known to become rapidly saturated with increasing sequencing depth (Svensson et al., 2017). The shallow sequencing of massively sampled single cells can effectively reduce random variation and define cell types through clustering analysis, providing a more robust approach (Pollen et al., 2014; Streets and Huang, 2014; Svensson et al., 2018). For large-scale scRNA-seq studies, a major technical hurdle is the cost of preparing a large number of cDNA libraries. Laboratory automation can overcome the laboriousness of this approach, but the reagents are still expensive (Jaitin et al., 2014). A few recently reported microfluidic approaches have demonstrated various advantages in scRNA-seq (Prakadan et al., 2017). For example, small-volume reactors may improve reaction efficiency and reduce technical noise when coupled with appropriate chemistry (Streets et al., 2014; Wu et al., 2014). Moreover, lab-on-a-chip operations have also made single-cell isolation much easier than manual cell picking



	inDrop	Drop-seq	10X
Barcoded Primer Bead			
Cell Barcode Capacity	147,456 (384 X 384)	16,777,216 (4 ¹²)	734,000
Droplet Generation	 Beads: super-Poissonian Cells: Poissonian	 Beads: Poissonian Cells: Poissonian	 Beads: super-Poissonian Cells: Poissonian
Emulsion			
Reaction in Droplets	2.5 h <ul style="list-style-type: none">• cell lysis• primer release by UV• mRNA capture• reverse transcription	0.3 h <ul style="list-style-type: none">• cell lysis• mRNA capture on beads	1 h <ul style="list-style-type: none">• cell lysis• primer release by bead dissolving• reverse transcription and template switch
Reaction after Demulsification	28 h <ul style="list-style-type: none">• 2nd strand synthesis• in vitro transcription• RNA fragmentation• RT-PCR	9 h <ul style="list-style-type: none">• RT and template switch• PCR• Tn5 tagmentation• PCR	7 h <ul style="list-style-type: none">• PCR• cDNA fragmentation and ligation• PCR



(legend on next page)

(Shalek et al., 2014). Microwell-based scRNA-seq methods (Fan et al., 2015; Han et al., 2018) have also exhibited advantages in terms of low cost and high throughput. However, owing to the lack of commercially available instruments or detailed protocols, microwell-based scRNA-seq has not been widely adopted.

Droplet microfluidics can achieve rapid compartmentation and encapsulation at a frequency of up to dozens of thousands of droplets per second and be easily scaled to produce millions of droplets, each having a nanoliter volume to accommodate single-cell reactions (Agresti et al., 2010). The microfluidic pipeline layout is very simple, consisting mainly of microchannels introducing or collecting reagents and samples (Duncombe et al., 2015). This droplet strategy greatly increases the reaction throughput and dramatically reduces the cost. Currently, there are three prevalent droplet-based systems for high-throughput scRNA-seq, namely, inDrop (Briggs et al., 2018; Klein et al., 2015; Wagner et al., 2018; Zilionis et al., 2017), Drop-seq (Farrell et al., 2018; Macosko et al., 2015), and 10X Genomics Chromium (10X) (Zheng et al., 2017). All of these have been demonstrated to be robust and practical in generating cDNA libraries for thousands of cells in a single run at acceptable cost. All three methods use similar designs to generate droplets, use on-bead primers with barcodes to differentiate individual cells, and apply unique molecular identifier (UMI) for bias correction (Kivioja et al., 2011). Despite these similarities, they involve different approaches for bead manufacturing, barcode design, and cDNA amplification and thus have different experimental protocols. Given these differences in system specifications and potentially in the results of transcriptome analysis (Ziegenhain et al., 2017), there is a need for a systematic and unbiased comparison among these methods.

Here, we compare the performance of these three approaches using the same sample with a unified data processing pipeline. We generated two to three replicates for each method using the lymphoblastoid cell line GM12891. The mean sequencing depth was around 50,000 reads per cell barcode. We also developed a versatile and rapid data processing workflow and applied it for all datasets. Cell capture efficiency, effective read proportion, cell barcode error rate, and transcript detection sensitivity were analyzed and compared. The results reveal strengths and weaknesses in each system and provide guidance for the selection of the most appropriate system in future research.

RESULTS

System Overview

Among the three systems, inDrop and Drop-seq have been extensively described in the literature, whereas 10X is a commercial platform whose design details have not been fully disclosed.

We here attempt to dissect these systems to the best of our ability based on the information that we could collect. In all three systems, the cell barcodes are embedded in microbead-tethered primers (Figure 1A). The DNA sequences of on-bead primers share a common structure, containing a PCR handle, cell barcode, UMI, and poly-T. The primer on the inDrop beads also has a photo-cleavable moiety and a T7 promoter. However, the beads are fabricated with different materials. The beads used in 10X and inDrop systems are made of hydrogel, and Drop-seq uses brittle resin. Normally, beads and cells are introduced at low concentration to reduce the chance of forming doublets; that is, two cells or two beads are encapsulated in a single droplet. Therefore, for Drop-seq that uses small hard beads, encapsulation of one bead and one cell in the same droplet follows a double Poisson distribution. The hydrogel beads are soft and deformable, closely packed in the microfluidic channel, and their encapsulation can be synchronized to achieve a super-Poissonian distribution (Figure 1A; Abate et al., 2009). Although 100% single-bead occupancy is very difficult due to inevitable variation in bead size, the cell capture efficiency can reach markedly higher levels in 10X and inDrop approaches. 10X is reported to have ~80% bead occupancy and a cell capture rate of ~50% (Zheng et al., 2017).

The material of the beads may also influence the quantity and density of DNA primers. The use of a hydrogel for 10X and inDrop allows the immobilization of primers throughout the beads, whereas the smaller Drop-seq beads can only carry primers on the surface. After encapsulation, the entire beads from 10X are dissolved to release all of the primers into the solution phase to boost the efficiency of mRNA capture. inDrop also mobilizes the primers by UV-irradiation-induced cleavage. In contrast, Drop-seq uses surface-tethered primers to capture the mRNA molecules, which could reduce the capture efficiency compared with that for 10X and inDrop.

Reverse transcription is carried out within droplets for 10X and inDrop before demulsification. Instead, Drop-seq only captures the transcripts without cDNA conversion. Reverse transcription in droplets can confer more uniform results due to the isolation of many local reactions and the reduction of reaction competition. It is also known that the performance of a reaction in a limited volume, such as a droplet, enhances the specificity of cDNA conversion and relative yield (Streets et al., 2014). The three systems adopt different strategies for cDNA amplification. inDrop employs CEL-seq (Hashimshony et al., 2012), whereas 10X and Drop-seq follow a template-switching protocol (Macosko et al., 2015; Zheng et al., 2017) similar to the popular Smart-seq chemistry (Ramsköld et al., 2012). The *in vitro* transcription step in inDrop extends the library preparation time beyond 24 hr, although both Drop-seq and 10X pipelines can be completed within a day.

Figure 1. Overview of the Three Platforms, Experimental Design, and Data Analysis Pipeline

- (A) Schematic and comparison of experimental features of the three systems. They differ in terms of barcode design, library size, emulsion, and downstream reactions.
 - (B) Experimental scheme summary. Two or three replicates were performed for each platform, and the same data processing pipeline was used for downstream analysis. The numbers of input and recovered cells are labeled.
 - (C) Overview of the data processing pipeline workflow. The sequencing reads that result from barcoding and tagging in reverse transcription are first demultiplexed by their cell barcodes and then the UMIs mapped to each gene are aggregated and counted.
- See also Figure S1.

Experimental Design and Data Processing

We used GM12891, a human lymphoblastoid cell line, for our comparative study. Biological replicates were set up for all three systems, with various cell input numbers on different days and in different batches (Figure 1B). We adjusted the sequencing depth to obtain comparable numbers of reads per cell barcode across the three systems (see STAR Methods).

Each system has its own data processing pipeline. However, none of them can directly handle data generated by other systems due to differences in read structures. Each analysis pipeline has to cope with system-dependent data characteristics, for example, the tolerance of cell barcode errors. Besides, the analysis pipelines use different strategies in some key processes, such as gene tagging. All of these differences may introduce bias in gene quantification, which is not ideal when attempting to perform a fair comparison among the systems. To solve this problem, we developed a versatile pipeline that accepts data from all of these systems and generates matrices of UMI counts (Figure 1C). We applied this pipeline to our data and conducted comparisons on sensitivity, precision, and bias in an objective way.

The pipeline package is freely available online (<https://github.com/beiseq/baseqDrops>) for download. It was designed to accept paired-end sequencing data with one end (read 1) containing the cell barcode and UMI and the other end (read 2) containing the transcript sequence. The pipeline first identifies cell barcodes in read 1's raw data. After removing cell barcodes with read counts that are too low (miscellaneous barcodes), the pipeline corrects cell barcode errors (see STAR Methods). These errors may have been introduced during on-bead primer synthesis and also during PCR or sequencing steps. Reads from the same cell barcodes are aggregated, and invalid cell barcodes are removed after filtering by read counts. For 10X and inDrop, in which barcodes are not completely random, the pipeline further filters the cell barcodes based on manufacturers' whitelists.

Read 2 sequences are mapped to the human reference genome (hg38) using STAR (Dobin et al., 2013) and then tagged to the corresponding genes. We also processed the datasets with each protocol's official pipeline. We then compared the obtained results with those from our versatile pipeline. The expression levels of the majority of genes and the UMI counts in each barcode were found to be highly consistent among the different data processing methods (see STAR Methods; Figures S2A and S2B). To confirm the accuracy of transforming aligned reads to the corresponding genes, we performed simulation by generating around 2 million reads based on the cell line's gene expression profile (see STAR Methods). More than 99% of the reads (2,229,156 out of 2,251,529) were tagged to the correct gene (see STAR Methods; Figure S2C). The remaining 1% of ambiguous reads were mainly derived from genes with paralogs or overlapping genes, such as RPL41/AC090498.1 or IGHA1/IGHA2 (Table S2). After read-to-gene assignment, the reads for each gene in each cell were grouped and their UMIs were aggregated and counted by allowing a 1-bp mismatch, thus generating a gene expression UMI matrix.

The processing speed of this pipeline was optimized by reducing the read and write payload, which is a common bottle-

neck. For example, ~50% of reads from inDrop data have an invalid sequence structure. By removing these reads, we can increase the data processing efficiency. Furthermore, the reads are split into multiple (typically $N = 16$) files, based on the cell barcode prefix, which enables parallel processing.

Quality of Primers on Beads

The barcode library size determines the maximum capacity for a single experimental run using droplet-based scRNA-seq. A small cell barcode library might result in barcode collision and artificial doublets. In the information accompanying the three systems, theoretical cell barcode library sizes of 1.47×10^5 (inDrop), 1.6×10^7 (Drop-seq), and 7.34×10^5 (10X) are claimed. However, the effective barcode library size may be smaller than the designed value. We estimated the proportion of effective barcodes by analyzing the barcode collisions between multiple runs from each system (see STAR Methods). The likelihood analysis demonstrated the relative probability of observing such a number of collisions at different effective barcode fractions (Figure 2A). For inDrop, our results suggest an effective barcode proportion of around 30%, although 100% effectiveness is also possible with smaller possibility. The analysis is less powerful for larger libraries, but we can still try to determine the lower bound of effective proportion for Drop-seq (~10%) and 10X (~40%). The likelihood of an effective barcode proportion smaller than the lower bound is relatively low. Thus, by rough estimation, the effective barcode size is $\sim 5 \times 10^4$ for inDrop and at least 1×10^6 for Drop-seq and 3×10^5 for 10X (see STAR Methods).

One-barcode-one-bead is the key requirement for all three systems. However, owing to the imperfection in the chemistry of DNA synthesis, asynchronous base addition is inevitable. Inconsistency in the sequences of cell barcodes could thus arise within the same bead. Such presence of errors in cell barcodes would result in inflation of the number of detected single cells, which requires careful correction. We aggregated the cell barcodes within 1 Hamming distance. For each valid cell barcode, the proportion of the corrected reads (which contains errors in raw barcode sequences) to the total reads after correction is calculated as the cell barcode error rate (Figure 2B), which reflects the general quality of on-bead DNA primers. 10X beads showed few mismatches in cell barcodes, indicating good quality control in bead fabrication. In contrast, more than half of the cell barcodes contained obvious mismatches in the other two systems. Specifically, about 10% of Drop-seq beads contained a one-base deletion in cell barcodes, which also required extra care during data analysis (see STAR Methods).

We further analyzed the base composition of UMI, which could reflect its synthesis and usage bias (Figure 2C; Table S1). All systems showed bias or preference for poly-T due to its affinity to the poly-A tail of mRNA. We also found the enrichment of poly-C in inDrop and of poly-G in Drop-seq and 10X. Such patterns, predominantly due to DNA synthesis bias, may cause system-dependent skewness of the RNA-seq results.

The primary filtering criterion for valid cell barcodes is based on the total number of raw reads, which largely reflects the abundance of cellular mRNAs. A cell barcode with more reads is more likely to originate from a real cell. The cell barcodes were sorted and visualized by their read counts, and we observed different

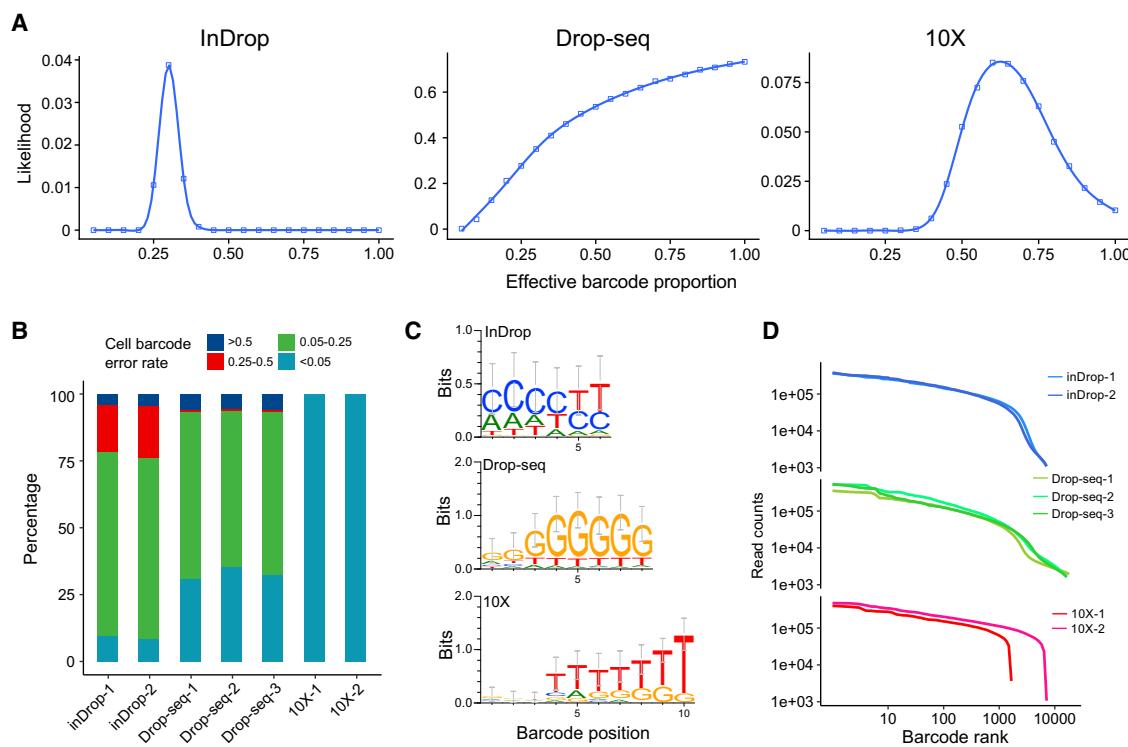


Figure 2. On-Bead Primer Library Size and Quality Assessment

(A) Estimation of effective cell barcode library size for each system. The likelihood of different effective barcode proportion is shown. The likelihood analysis is based on the observed barcode collisions between different samples from the same system (see STAR Methods).

(B) Distribution of cell barcode error rate. The error rate was measured as the proportion of corrected reads (1-bp mismatch) relative to the total reads.

(C) The motif of the top 50 frequently used UMIs for each system.

(D) The primary estimation of the valid cell barcode numbers according to the read counts. Cell barcodes in the same sample are ordered by their read counts. The top N cell barcodes are selected according to input cell numbers and experimental capture efficiency.

See also Table S1.

features in the three systems (Figure 2D). For 10X, a sharp cliff indicated the distinct difference in read counts between barcodes from healthy cells and others. For inDrop, there was a similar but subtler cliff. For Drop-seq, however, there was no obvious cliff on the read-count curve for a clear cutoff. This might have originated from the wide size distribution of beads used by Drop-seq. We noticed that the size of beads used in inDrop or 10X was more uniform than that in Drop-seq (Figure S1), probably due to the difficulties in size control when fabricating resin beads.

Data Processing Steps and Results

It is challenging to accurately determine the cell number, represented by cell barcodes, in each sample. This is due to the large dispersion in cellular mRNA molecular counts and their capture efficiency. We attempted multiple strategies to estimate the valid cell numbers (see STAR Methods; Figure S3). Many of these methods rely on certain assumptions about the reads or UMIs distribution or cell composition, which might not apply for all protocols or situations. We implemented a strategy that started from a given number of cells determined experimentally, followed by strict quality control filtering (UMIs $\geq 1,000$ and nearest correlation ≥ 0.6).

This strategy has been implemented by multiple groups in recently reported high-throughput scRNA-seq studies. For each run, the number of recovered cells could be roughly estimated by considering the number of input cells, cell capture ratio, and downstream reaction success ratio, in accordance with system-specific protocols. Then, the estimated cells were further filtered to satisfy the quality control criteria (see STAR Methods).

The reads split into each valid cell barcode are first aligned to the human genome to analyze the distribution of reads throughout the genome (Figure 3A). Drop-seq and 10X have about 65% of the reads mapped to UTR (mainly 3' UTR) and exon regions, although this proportion in inDrop is only about 45%. After the tagging of reads that map to gene bodies, the numbers of detectable genes can be obtained (Figure 3B). The gene number drop off in 10X-1 as well as Drop-seq-2 and Drop-seq-3 is due to small numbers of input cells. The number of genes declines in accordance with the number of reads within a cell, except for several outliers in Drop-seq data. We use those detected genes to demonstrate the bias of read distribution along the gene body (Figure 3C). The reads were mainly derived from the 3' end of the mRNA for all three systems, consistent with

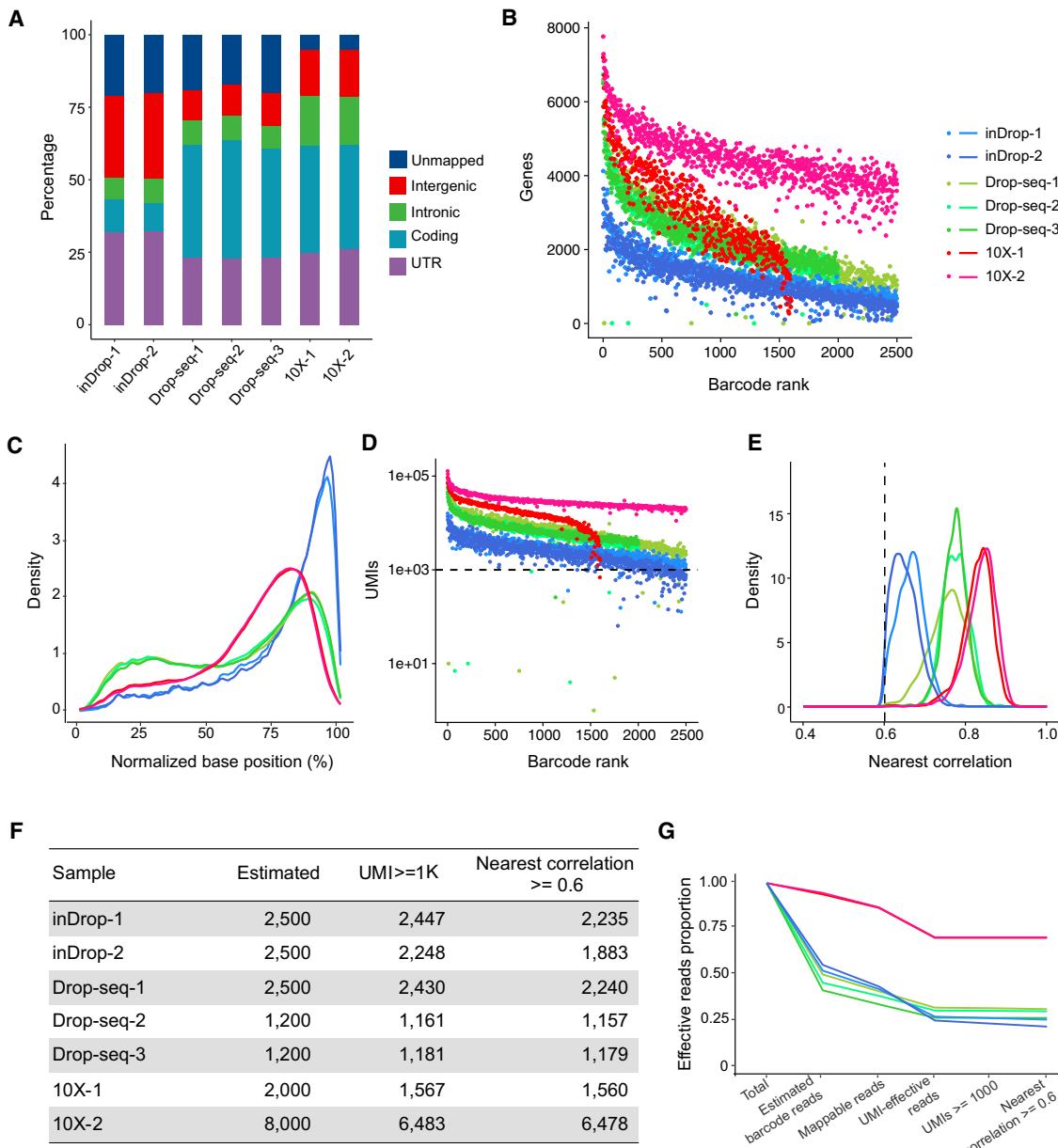


Figure 3. Data Processing Steps and Results

- (A) Read composition after mapping to the genome. Percentages of reads mapped to different genomic regions and unmapped reads are shown.
- (B) The number of genes detected with cell barcode ranked by read counts.
- (C) Normalized read distribution across the gene body from the 5' to the 3' end.
- (D) The number of UMIs with cell barcode ranked by read counts.
- (E) The distribution of cells' nearest correlation with all other cells from the same sample (see STAR Methods); a threshold of 0.6 is applied for quality control.
- (F) The number of valid cell barcodes retained after each step of quality control filtering.
- (G) The proportion of effective reads after each step of quality control process (see STAR Methods).

See also Figures S2 and S3.

their library construction strategies. Drop-seq data showed a bimodal distribution, most likely due to the same PCR anchor sequences being used at both ends of cDNA molecules.

We performed cell barcode filtering based on the total count of UMIs (transcripts) in each experimental run (Figure 3D). With a

total UMI cutoff of 1,000, most of the cell barcodes passed the filter, which indicates that the estimated cell number is sound. To further remove possible artifacts caused by barcode errors, we checked the similarity of expression profiles between similar cell barcodes. If the expression profile of a cell barcode was

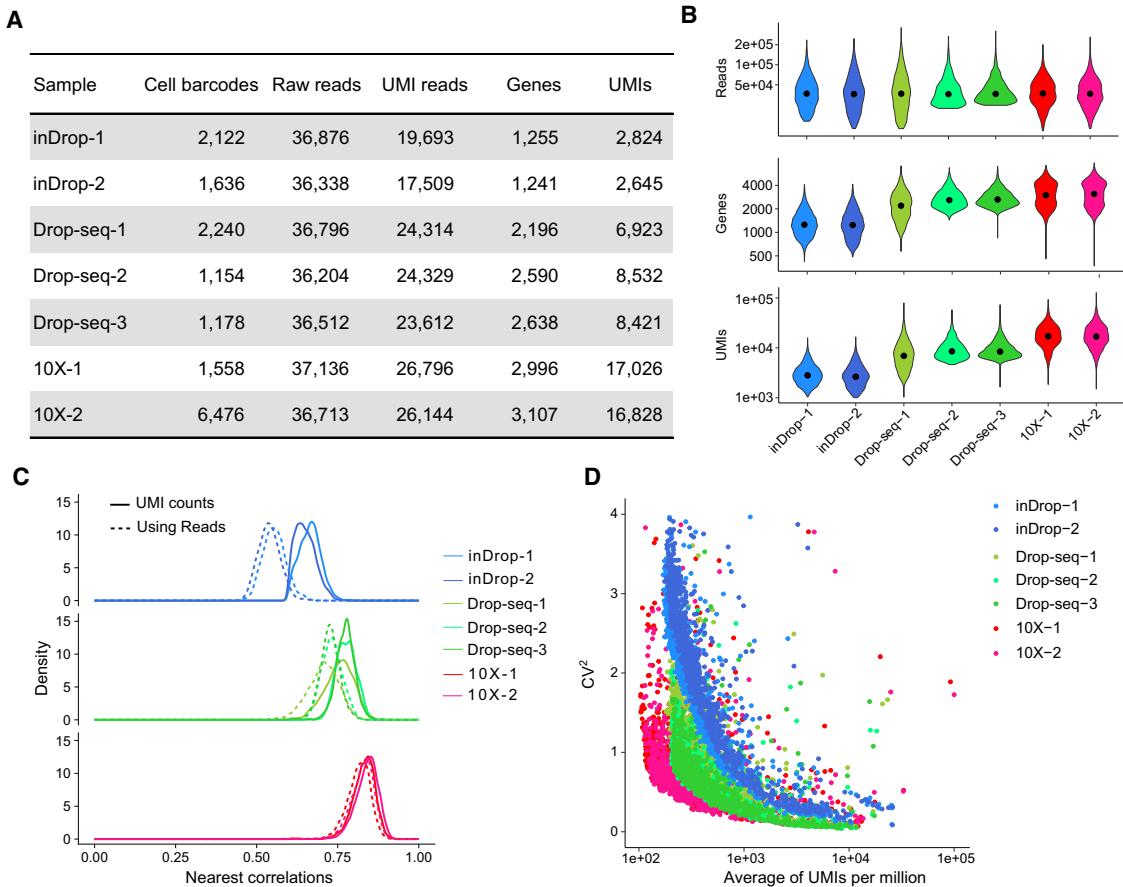


Figure 4. Demonstration of the Sensitivity and Technical Noise of Each Platform

(A) Summary of cell barcode numbers, read counts, and molecular detection performance. The data are down sampled to obtain a uniform level of raw reads across all samples (see [STAR Methods](#)).
(B) The distribution of raw reads, UMIs, and genes detected.
(C) Technical noise measured by the nearest correlation between one cell barcode and every other cell barcode within the same sample. Gene quantifications through UMI counts (solid line) and read counts (dashed line) are both adopted.
(D) The CV-mean (CV squared) plot of each system. The technical noise is measured at the gene level.
See also [Figures S4](#) and [S5](#) and [Tables S3](#) and [S5](#).

markedly different from its closest cell barcode neighbor (Spearman's correlation ≤ 0.6 ; see [STAR Methods](#)), we discarded the barcode ([Figure 3E](#); see [STAR Methods](#)).

Through all of these steps, we obtained various numbers of cells in each experiment ([Figure 3F](#)). The proportion of effective reads (reads from valid barcodes) was ~75% for 10X, ~25% for inDrop, and ~30% for Drop-seq ([Figure 3G](#)). The proportion of such reads should be maximized to reduce wastage of sequencing capacity.

Sensitivity of UMI and Gene Detection

The sensitivity of gene detection is a fundamental indicator of the performance of scRNA-seq. It reflects the overall efficiency of a method for capturing a single mRNA molecule for reverse transcription, second-strand synthesis, and pre-amplification. It further influences and determines the precision and accuracy of gene expression quantification. With the same cell line as an

input sample, the sensitivity can be depicted simply with the recovered UMIs and gene counts ([Figure 4A](#)). The UMI and gene numbers gradually become saturated for cell barcodes with increasing read counts ([Figures S4A](#) and [S4B](#)). We found that the log-transformed UMI count is highly correlated (Spearman's correlation $r > 0.9$) with the number of detected genes ([Figure S4C](#)). This shows that sequencing depth may influence the numbers of UMIs and genes detected. For a fair comparison among the three different systems, we normalized the dataset to achieve a uniform raw read level (~36K/cell) before gene expression analysis (see [STAR Methods](#)). The technical replicates from the same system showed high consistency and reproducibility. 10X had the highest sensitivity, capturing over 17,000 transcripts from ~3,000 genes on average. This performance was consistent, regardless of the number of input cells. Drop-seq detected ~8,000 transcripts from ~2,500 genes. Meanwhile, the inDrop system had lower sensitivity, detecting

~2,700 UMIs from ~1,250 genes. The read distribution is more skewed in inDrop and Drop-seq, for which the majority of cell barcodes have relatively low read counts (Figure 4B).

Technical Noise and Precision

Technical noise reflects the variation conferred by experimental randomness, including transcript dropout in reverse transcription and the bias associated with PCR amplification. Precision can be assessed by the concordance of the transcriptome among technical replicates. A major purpose of performing single-cell RNA-seq is to cluster cells into different subgroups based on their gene expression profiles, typically for discovering and characterizing new cell types or states. Clustering is based on the similarities or distances of gene expression patterns among cells. Large technical noise or variation will distort the actual distances and obscure subtle biological differences between cells, thus lowering the resolution of cell grouping. Many efforts have been made to reduce the technical noise, such as the use of UMI to eliminate the quantification error caused by amplification bias.

Although we here use an apparently homogeneous cell line, there is still intrinsic biological noise or heterogeneity (Prakadan et al., 2017). In our dataset, the total variation consists of technical and biological components, which are difficult to separate. Here, we assume that biological noise is consistent among samples and that technical noise dominates the variation in the datasets. The noise levels of housekeeping genes (which show a minimal level of biological noise) and other genes have similar distributions, which indicates the low level of biological noise compared with technical noise (Figure S5; see STAR Methods). Thus, the overall total variation should reflect the technical noise level.

The overall total variation is characterized as the nearest Spearman's correlation between a specific cell barcode and every other cell barcode in the entire dataset (see STAR Methods). Many clustering or classification strategies, such as k-means and hierarchical clustering, are based on the nearest correlation between the cells. To identify minor cell types, the nearest correlation among these minor cells should be high to enable their separation from other cells. To validate the effect of UMIs in reducing the PCR amplification noise of gene counting, we performed the analysis using both UMI counts and raw read counts for the quantification of gene expression. The results (Figure 4C) show that 10X and Drop-seq have lower technical noise levels than inDrop. For all three systems, gene expression profiles characterized by UMI have reduced noise compared with those using raw counts, confirming the effectiveness of UMI in noise reduction. It is noteworthy that such noise is more severe in inDrop data, probably due to the use of random primers during library construction. For 10X, however, the usage of UMI does not dramatically reduce noise. This is probably due to relatively even amplification during 10X sample preparation. In addition, most UMIs were sequenced only two to three times, suggesting a less saturated sequencing depth. For deeper sequencing, the use of UMI can probably reduce the noise further.

The technical variation at the gene level can be measured by the coefficient of variation (CV) of normalized UMI (UMIs per million) counts across all cells (Figure 4D; see STAR Methods). This provides a view of the technical noise on the whole gene

expression profile. All systems show reduced variation for genes with higher expression levels. Generally, 10X has the lowest technical noise, followed by Drop-seq and then inDrop. Interestingly, many of the most highly expressed genes are quite noisy, especially in the 10X data. We examined these genes (normalized UMI \geq 2,000; CV \geq 0.5) and found that most of them were the cell line's most highly expressed genes or mitochondrial genes (Table S3). High noise in these genes was probably introduced by the stochastic manner of bursts by which transcription occurs (Sanchez and Golding, 2013).

Saturation of Sensitivity and Precision at Low Sequencing Depth

The ability to detect transcripts present at a low level could be enhanced by performing deeper sequencing. However, there is a trade-off between costs and sensitivity, especially for high-throughput experiments. Empirically, it has been shown that each cell gets 10,000–100,000 reads in high-throughput scRNA-seq experiments, whereas for conventional scRNA-seq data, the corresponding value is usually \sim 1 million reads per cell (Baran-Gale et al., 2018). A previous study based on a mathematical model suggested that shallow sequencing (1% of conventional depth) can also be informative regarding cell status (Heimberg et al., 2016). We randomly subsampled sequencing data and analyzed the corresponding changes in sensitivity and precision (Figures 5A, 5B, and S6). The fitted saturation curves of UMI and gene counts help to determine a suitable sequencing depth for most applications.

All of the systems show diminishing returns at higher depths. For more sensitive methods, it is possible to detect the same level of UMIs with fewer reads. All three methods can reach a threshold of 1,000 UMIs with fewer than 10K reads. 10X can detect 10,000 UMIs with about 20K reads as a median, although for Drop-seq, the value is 50K. These both exceed the capacity of inDrop. We also evaluated how many reads per cell are needed to reach 80% of the total saturated UMIs for Drop-seq (\sim 80K) and inDrop (\sim 60K; Figure S6A). In contrast, 10X requires \sim 200K reads/cell to accomplish this due to the higher sensitivity. Detection sensitivity of gene numbers saturated faster. To reach the 80% saturation level, \sim 30K reads/cell are needed for inDrop or Drop-seq, while \sim 80K reads/cell are needed for 10X (Figure S6B).

Other than sensitivity, precision also determines a system's resolution for making biological discoveries. Here, the precision is measured as the nearest correlation between one cell and the others, which also indicates the level of technical noise. We investigated how the precision level was affected by the sequencing depth and found that the precision index rapidly saturated with increasing read depth (\geq 20,000 effective reads) for all three systems (Figure 5C).

These results help us to establish appropriate empirical guidelines for experimental design. For the most commonly performed tasks, such as cell typing, a median number of 20,000 reads/cell should be sufficient. However, it should be noted that these results are from a cell line with abundant mRNA. The desired sequencing depth should be considered based on both the sensitivity of protocols and the input RNA content. For cells with lower transcription activities, such as primary cells, a lower level of sequencing depth could be sufficient for each protocol.

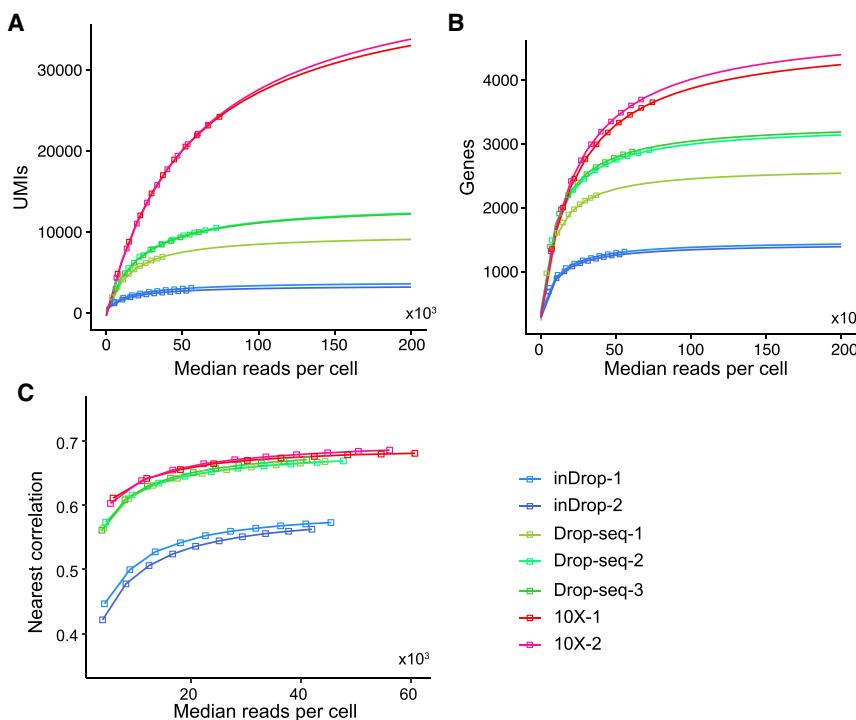


Figure 5. Transcriptome Analysis Sensitivity and Noise Level at Different Sequencing Depths by Subsampling Analysis

(A and B) Median numbers of UMIs (A) and genes (B) detected for each sample with increasing effective read counts.

(C) Transcriptome analysis noise level saturates quickly with sequencing depth. The noise was measured as the nearest correlation (see STAR Methods).

See also Figure S6.

sets together from the same method. However, different protocols have obvious bias related to gene length and GC content. Thus, combining these datasets directly will introduce extra divergence.

DISCUSSION

We have compared the three most widely used droplet-based high-throughput single-cell RNA-seq systems, inDrop, Drop-seq, and 10X, using the same cell sample and a unified data processing pipeline to

reduce bias in experimental design and data analyses. Technical replicates were included to identify possible batch-dependent artifacts. For each system, we sequenced thousands of single cells. Through quantitative analysis of a few key parameters using our unified data processing pipeline, we have clarified the characteristics of each system. Generally, after filtering out artifacts and errors, all three systems produced quality data for single-cell expression profiling. The cell typing analysis indicated obscure batch effects but noticeable clustering bias in association with the system of choice. This indicates that cell typing analysis using datasets from a mixture of systems is technically challenging and should be avoided.

In this study, we chose a lymphoblastoid cell line for the analysis because cell line quality is highly controllable. At least for technical evaluation, we wished to reduce the variation of sample quality on the obtained results as much as possible. However, direct comparisons using primary cells, especially those with low mRNA contents, would be more informative. To expand the scope of our study, we further processed HEK293 cells with 10X system and included some datasets produced by the original developers of the three systems (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017). As summarized in Table S5, 10X demonstrates higher sensitivity, detecting roughly twice as many of UMIs as inDrop and Drop-seq do from various kinds of cell. The results from the inDrop developers are better than ours. We attribute this discrepancy to batch-to-batch variation in bead quality. As we showed above, inDrop cell barcode error rate is much higher than those of Drop-seq and 10X (Figure 2B). Being labeled with defective barcodes would deem the transcripts undetectable since the very beginning. More than half of inDrop sequencing data were wasted due to a failure of matching with the cell barcodes in

Bias in Gene Quantification

To comprehensively compare the transcriptomes depicted by different systems, we conducted dimension reduction with principal-component analysis (PCA) and *t*-distributed stochastic neighbor embedding (tSNE) analyses (Figure 6A). Almost all of the cells were robustly separated and clustered according to their system of origin. Although there is biological and technical variation within cells from the same run, which results in great diversity in sequencing reads, and in gene and UMI counts, the bias between different systems still exceeds the level of these variations. As the replicates are processed in different batches and days, the batch effect is also obscure. Within the same system, different batches of data show a very homogeneous distribution (Figure S7).

The separation of cells by system indicates that there is system-specific quantification bias at the gene level. Potential biases in the mRNA enrichment at the gene level could be related to three major factors: expression abundance (normalized to UMIs per million); gene length; and GC content. We hence selected the top 100 marker genes (see STAR Methods) from each method and analyzed the distribution of these factors (Figures 6B–6D). These genes showed consistent expression intensity among biological replicates. We found that, compared with the other systems, 10X slightly favored shorter genes and genes with higher GC content, whereas Drop-seq better detected genes with lower GC content. This observation echoes a previous report describing that Drop-seq overestimates transcription of genes with low GC ratio or long sequence (Macosko et al., 2015).

In summary, all of the methods appear to be very consistent and homogeneous among technical replicates from different batches. This indicates the validity of combining different data-

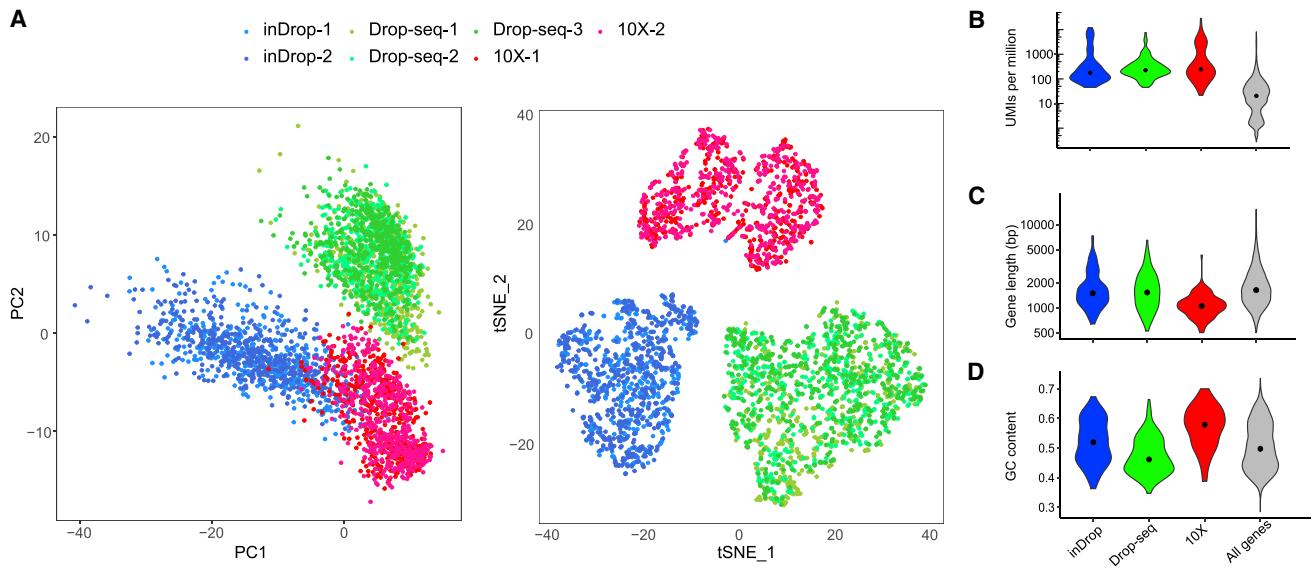


Figure 6. Transcriptome Analysis Bias in the Three Systems

(A and B) Visualization of cell barcodes of all three systems clustered by PCA (A) and tSNE (B).

(B–D) Demonstration of transcriptome analysis bias in aspect of gene expression level (B), gene length (C), and GC content (D). The top 100 marker genes from each system were used for demonstration. The distributions of all genes (in gray color) are also shown for comparison. See also Figure S7.

our data. Feedback from other inDrop users showed that the equivalent proportions from different batches of beads range from 25% to 65% (unpublished data). We also tested the impact of mRNA content on system performance. When using half of HEK293 cDNA for downstream library preparation, we detected roughly half UMI as in normal HEK293 (Table S5). All these abovementioned results suggest that our findings based on the lymphoblastoid cell line can be generalized to other cell types.

For all three systems, the beads are specifically provided by the particular manufacturer and would probably be difficult to produce in small laboratories. Thus, the quality of the beads, such as their size dispersity, is particularly important to define the robustness and uniformity of reverse transcription and further reactions. Moreover, the fidelity and purity of the barcode sequences on each bead are also key factors affecting the bioinformatics pipeline, for which artifacts and errors should be minimized.

Our comparison shows that 10X generally has higher molecular sensitivity and precision and less technical noise. As a more maturely commercialized system, the 10X protocol should have been extensively optimized, which is partially reflected in the barcode design and quality control of bead manufacture. However, high-performance optimization also comes with a high price tag. Specifically, the instrument costs more than \$50,000 and the per-cell cost is around \$0.50, even without considering the sequencing cost or instrument depreciation (Table S6).

With small compromises in sensitivity and precision, Drop-seq exhibits a significant advantage in experimental cost compared with 10X, which is typically the major concern when a large number of single cells are needed. As an open-source system

(except for the beads), Drop-seq has gained popularity since its introduction in 2015. As of the time of writing, the Drop-seq protocol has been downloaded nearly 60,000 times. Building up the whole system costs less than \$30,000. The experimental cost of Drop-seq is about \$0.10 per cell (Table S6). Drop-seq is thus a reasonable choice for individual labs, given its balanced performance and economical nature.

To a certain extent, inDrop can be considered an open-source version of 10X. Both of them use hydrogel beads for super-Poissonian loading. Their on-bead primers are both releasable to facilitate the capture of transcripts. The instrument cost is comparable to that of 10X, and the per-cell cost is about half that of 10X (Table S6). We attribute the lower performance of inDrop to its excessive cDNA amplification (Hashimshony et al., 2016), as well as the fact that the protocol has yet to be completely optimized. As an open-source system, inDrop can adopt other chemistries and be easily modified for different types of RNA-seq protocols. In a preliminary experiment, we tested the implementation of Smart-seq2, the most widely used scRNA-seq protocol, on the inDrop system. The cDNA profile closely resembled conventional Smart-seq2 products (Figure 7A). We further tested different conditions for reverse transcription and cDNA amplification. Similar to the results generated by the official protocol, a significant proportion (~40%) of reads in the new data could not be assigned to genuine cell barcodes. Our briefly optimized protocol generated results for UMI and gene detection comparable to those with the official protocol (Figures 7B–7D; Table S4). Although the sensitivity of transcription detection was still lower than in the other two systems, our preliminary results demonstrated the flexibility of inDrop and that the system could be desirable for nonstandard approaches or technical development.

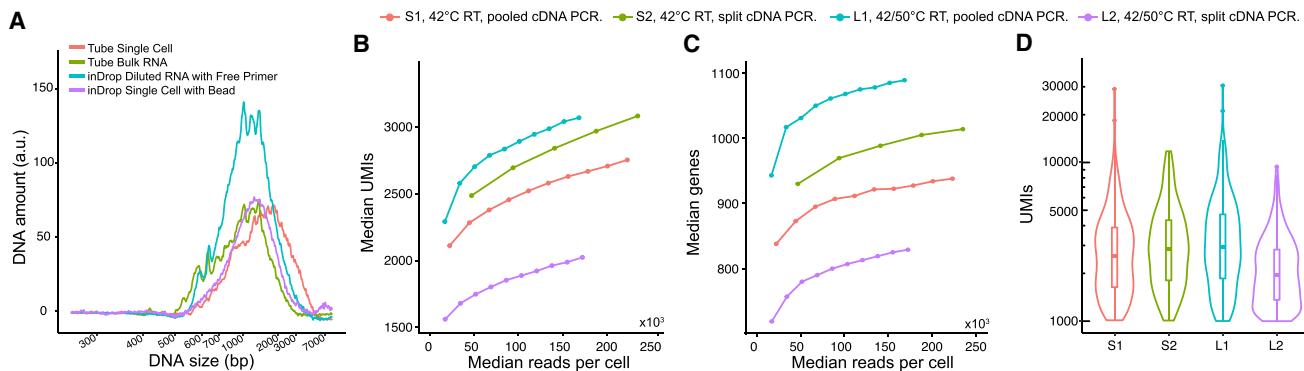


Figure 7. Adopting the Smart-seq2 Protocol in the inDrop Platform

(A) Comparison of cDNA fragment size between Smart-seq2 performed in tube and inDrop platform.
 (B and C) Four kinds of reaction with different reaction temperatures and PCR amplification strategies were performed (S1, S2, L1, and L2, see STAR Methods). Their median detected UMI (B) and gene (C) counts at various sequencing depths are shown.
 (D) The UMI distributions for four conditions at uniform sequencing depth (100K reads). The L1 condition has better sensitivity.
 See also Table S4.

With all of the system-specific features mentioned above, we proposed guidance to facilitate the choice of a suitable droplet-based scRNA-seq system for ultra-high-throughput single-cell studies. Although most projects work with relatively large cell numbers, precious samples, such as human embryos, require efficient cell capture. A super-Poissonian distribution of cell capture could be essential for such samples. The requirements regarding the experimental cost and efficiency of transcript detection depend on the specific scenario. Generally, all three systems offer satisfactory transcript detection efficiency, and higher efficiency is associated with higher experimental cost. By rule of thumb, 10X is currently a safe choice for most applications. When the sample is abundant, Drop-seq could be more cost efficient. In contrast, when the detection of low-abundance transcripts is optional, or a custom protocol is desired, inDrop becomes a better choice.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
 - Cell lines
- **METHOD DETAILS**
 - Cell lines and cell preparation
 - Drop-seq experiment
 - inDrop experiment
 - 10X system experiment
 - Calculation of cell capture efficiency
 - Cell barcodes and UMI assignment
 - Analysis of effective cell barcode library size
 - Barcode correction
 - Determination of valid cell barcodes
 - Alignment and tagging reads to genes
 - UMI corrections

- Comparison with protocols' official analysis pipeline
- Distribution of reads on the gene structures
- Quality control and filtering of cells
- Proportion of effective reads
- Technical noise analysis
- Technical noise of housekeeping genes
- Sequencing depth subsampling and normalization
- Smart-seq2 protocol in inDrop system
- Comparison of public datasets of the three systems
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
 - Sequencing depth fitting and prediction
 - PCA and tSNE analysis
- **DATA AND SOFTWARE AVAILABILITY**

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and six tables and can be found with this article online at <https://doi.org/10.1016/j.molcel.2018.10.020>.

ACKNOWLEDGMENTS

We thank Fan Wu for testing the data process pipeline. This work was supported by Ministry of Science and Technology of China (2016YFC0900100 to J.W. and 2018YFA0108100 to Y.H.), the National Natural Science Foundation of China (21675098 to J.W. and 21327808 and 21525521 to Y.H.), and Beijing Advanced Innovation Center for Genomics.

AUTHOR CONTRIBUTIONS

Y.H. and J.W. conceived the project. T.L., F.L., and J.Y. performed experiments. X.Z., Z.L., Y.C., and J.Y. analyzed data. All authors participated in manuscript preparation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 2, 2018

Revised: August 25, 2018

Accepted: October 12, 2018

Published: November 21, 2018

REFERENCES

- Abate, A.R., Chen, C.H., Agresti, J.J., and Weitz, D.A. (2009). Beating Poisson encapsulation statistics using close-packed ordering. *Lab Chip* 9, 2628–2631.
- Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.C., Marquez, M., Kilianov, A.M., Griffiths, A.D., and Weitz, D.A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc. Natl. Acad. Sci. USA* 107, 4004–4009.
- Baran-Gale, J., Chandra, T., and Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics* 17, 233–239.
- Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1095.
- Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., and Klein, A.M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360, eaar5780.
- Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160.
- Butler, A., Hoffman, P., Smibert, P., Papalex, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420.
- Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196.
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
- Duncombe, T.A., Tentori, A.M., and Herr, A.E. (2015). Microfluidics: reframing biological enquiry. *Nat. Rev. Mol. Cell Biol.* 16, 554–567.
- Fan, H.C., Fu, G.K., and Fodor, S.P. (2015). Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* 347, 1258367.
- Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* 360, eaar3131.
- Grün, D., and van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell* 163, 799–810.
- Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by Microwell-seq. *Cell* 172, 1091–1107.e17.
- Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* 2, 666–673.
- Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17, 77.
- Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* 2, 239–250.
- Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779.
- Kivioja, T., Vähäautio, A., Karlsson, K., Bonke, M., Enge, M., Linnares, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74.
- Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201.
- Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S., et al. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 352, 1586–1590.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214.
- Olsson, A., Venkatasubramanian, M., Chaudhri, V.K., Aronow, B.J., Salomonis, N., Singh, H., and Grimes, H.L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* 537, 698–702.
- Papalex, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18, 35–45.
- Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401.
- Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181.
- Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058.
- Prakadan, S.M., Shalek, A.K., and Weitz, D.A. (2017). Scaling by shrinking: empowering single-cell ‘omics’ with microfluidic devices. *Nat. Rev. Genet.* 18, 345–361.
- Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782.
- Sanchez, A., and Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. *Science* 342, 1188–1193.
- Semrau, S., Goldmann, J.E., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., and van Oudenaarden, A. (2017). Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.* 8, 1096.
- Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510, 363–369.
- Streets, A.M., and Huang, Y. (2014). How deep is enough in single-cell RNA-seq? *Nat. Biotechnol.* 32, 1005–1006.
- Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., and Huang, Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* 111, 7048–7053.
- Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387.
- Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604.
- Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenome-logy to mechanism. *Nature* 541, 331–338.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382.

- Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C., Mount, C., Filbin, M.G., et al. (2016). Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma. *Nature* 539, 309–313.
- Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375.
- Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 355, eaai8478.
- Villani, A.C., Satija, R., Reynolds, G., Sarkisova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356, eaah4573.
- Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981–987.
- Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., and Quake, S.R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11, 41–46.
- Wu, A.R., Wang, J., Streets, A.M., and Huang, Y. (2017). Single-cell transcriptional analysis. *Annu. Rev. Anal. Chem. (Palo Alto, Calif.)* 10, 439–462.
- Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139.
- Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049.
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65, 631–643.e4.
- Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A.M., and Mazutis, L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* 12, 44–73.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
Drop-seq beads	Chemgenes Barcoded Bead SeqB	Cat#MACOSKO-2011-10
INDROP SINGLE-CELL RNA SEQ KIT	1cellbio, Inc	N/A
Chromium Single Cell 3' Library & Gel Bead Kit v2	10X Genomics	Cat#PN-120237
Deposited Data		
Raw data from Drop-seq original paper	NCBI GEO	GEO: GSE63472
Raw data from inDrop original paper	NCBI GEO	GEO: GSE65525.
Raw data from 10X original paper	10Xgenomics	https://support.10xgenomics.com/single-cell-gene-expression/datasets
Raw and processed dataset	This paper	GEO: GSE111912
Experimental Models: Cell Lines		
GM12891 cell line	Coriell Institute	N/A
Software and Algorithms		
R	R Core Team	http://www.R-project.org
Seurat	Rahul Satija	https://satijalab.org/seurat/
STAR	Alexander Dobin	https://github.com/alexdobin/STAR
baseqDrops	This paper	https://github.com/beiseq/baseqDrops
Samtools	Li and Durbin, 2009	http://samtools.sourceforge.net/
Cellranger	10XGenomics	https://github.com/10XGenomics/cellranger

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jianbin Wang (jianbinwang@tsinghua.edu.cn).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

The GM12891 cell line (Male) was purchased from the Coriell Institute. The complete growth medium was made using RPMI-1640 medium with L-glutamine (cat. No.11875-085; Life Technologies), 10% fetal bovine serum (cat. No. 16000-044; Life Technologies), and 1% penicillin and streptomycin. The cell line was incubated with 5% carbon dioxide at 37°C in a culture flask.

METHOD DETAILS

Cell lines and cell preparation

The GM12891 cell line was purchased from the Coriell Institute. The complete growth medium was made using RPMI-1640 medium with L-glutamine (cat. No.11875-085; Life Technologies), 10% fetal bovine serum (cat. No. 16000-044; Life Technologies), and 1% penicillin and streptomycin. The cell line was incubated with 5% carbon dioxide at 37°C in a culture flask. The cell concentration was maintained between 5x10⁵ and 1 × 10⁶ cells/mL. Before the experiment, the general cell condition was confirmed under a microscope. Regular circular cells and some cell aggregates indicate a good cell state. Cells were collected by centrifugation at 150 g for 5 min and subsequently counted with a hemocytometer.

Drop-seq experiment

Cells were washed with PBS-BSA three times and filtered with a 40-µm cell drainer. We then counted the cells and adjusted the concentration to 100 cells/µL. All subsequent steps were carried out as detailed online by Macosko et al. (<http://mccarrolllab.org/dropseq/>). Briefly, we loaded droplet-making oil, cells in PBS-BSA, and barcoded beads (cat. No. MACOSKO-2011-10; Chemgenes Barcoded Bead SeqB) in lysis buffer into the droplet, generating a microfluidic device. Cells were lysed in the droplets to release mRNA. Beads captured mRNAs in the droplets. After demulsification, beads were pooled together. We conducted reverse

transcription and ExoI digestion. Two thousand beads (or 100 STAMPs) were aliquoted into one PCR tube for PCR amplification. PCR products were pooled together and purified using AMPure XP Beads. We further constructed libraries and performed sequencing on an Illumina HiSeq 4000 with custom Drop-seq read 1 primer.

inDrop experiment

We purchased the inDrop instrument and hydrogel beads from 1CellBio. We performed all steps in accordance with the manufacturer's protocol: Silanization Protocol v4.0 (<https://1cell-bio.com/wp-content/uploads/2017/10/Silanization-Protocol-v4.pdf>), inDrop Single-Cell Reverse Transcription Protocol v2.1 (<https://1cell-bio.com/wp-content/uploads/2017/11/inDrop-Single-Cell-Encapsulation-and-Reverse-Transcription-Protocol-Version-2.1.pdf>), and inDrop Library Preparation Protocol v1.2 (https://1cell-bio.com/wp-content/uploads/2017/03/InDrop_LibraryPrep_Protocol_v1.2.pdf). In short, we silanized the microfluidic chip and pre-processed the hydrogel beads before each experiment. The droplet-making oil, cell resuspension, and RT/lysis buffer were loaded into the chip for the generation of droplets. The emulsion was collected in a tube on ice and irradiated by UV light to release primers. Reverse transcription proceeded in the droplets. After demulsification, hydrogel beads were filtered. The RT product was digested by ExoI/HinfI and purified using AMPure XP beads. Second-strand cDNA was synthesized using NEB second-strand synthesis kit. After IVT, the RNA product was fragmented and reverse-transcribed via random primers. The product was purified by AMPure XP Beads and quantified by qPCR. We further constructed libraries and performed sequencing on an Illumina HiSeq 4000.

10X system experiment

We performed all steps following the 10X protocol. We used the Chromium Single Cell 3' Library & Gel Bead Kit v2 (10X Genomics). In short, all samples and reagents were prepared and loaded into the chip. Then, we ran the Chromium Controller for droplet generation. Reverse transcription was conducted in the droplets. We recovered cDNA through demulsification and bead purification. Pre-amplified cDNA was further subjected to library preparation. Libraries were sequenced on an Illumina Hiseq 4000.

Calculation of cell capture efficiency

The number of cells loaded into each system was counted as described above. The number of captured cells was determined using the number of cell barcodes passing the quality threshold (total UMIs \geq 1,000, nearest correlation \geq 0.6).

Cell barcodes and UMI assignment

For 10X, we obtained the cell barcode and UMI from each read pair by extracting the first 16 bp and the following 8 bp from read 1 sequences. In a similar way, the barcodes of Drop-seq can be accessed. inDrop's barcode design is more complicated as the full cell barcode contains two parts (named CB1 and CB2), which are separated by a 22-bp spacer sequence called W1. The length of CB1 ranges from 8 to 12 bp. We first located the W1 sequence by tolerating up to two mismatches. Then, we could determine the length of CB1 and the whole cell barcode sequence. We aggregated the retrieved cell barcodes and exported the count data for downstream analysis. For 10X and inDrop, the barcodes were further filtered by requiring their presence in the barcode whitelist in the associated protocols.

Analysis of effective cell barcode library size

The effective barcode library was estimated by comparing and counting the common barcodes between datasets from multiple experiments. It is possible to estimate the barcode library size from a single experiment in theory, by using a Poisson distribution model. However, it requires a precise number of input cells which is practically impossible to obtain. We therefore employed pairwise analysis. We hypothesized that the effective or real barcode library is a subset of the designed whole library and is consistent across different experiments. In each sample, the detected cell barcodes are randomly sampled from the effective library (without replacement); by chance, there should be some barcodes that recur in more than one sample, which are named barcode collisions. The number of collisions between two samples with given numbers of detected cells is mainly determined by and reflects the actual barcode library size.

The ideal number of collisions between samples under claimed library size are estimated by running simulations for multiple times (10000). For each simulation, we sampled (without replacement) given number of cell barcodes according to cell number in two samples from a pool of barcodes of the claimed library size, respectively. The mean number of recurred barcodes between two samples in all simulations are used as estimated collision number. For inDrop, the estimated collision number is 28, but observation revealed 92 of them. For Drop-seq, which has three samples, the barcodes of Drop-seq-3 and Drop-seq-2 are combined. The observed value is zero, while the simulation reported an mean of 0.3. For 10X, by simulation, the mean number of collisions is 13.7 upon sampling 1560 and 6478 cell barcodes from a library of 734,000 cell barcodes. The observed collision number is 22.

We further calculated the likelihood of detecting such number of collisions under a series of effective library sizes (ranging from 5% to 100% of the claimed size, at intervals of 5%). For each assumed library size, 10,000 random samplings were conducted using the number of cell barcodes in two experiments under such a library size. The proportion of observations of all simulation results that reported the same number of collisions as in our experiment is referred to as the likelihood of the corresponding library size. The likelihood analysis helps to indicate the probability of the actual library size based on our experimental observations.

Barcode correction

Ideally, the cell barcode sequences on the same beads should be the same. However, owing to errors in DNA synthesis, there are mismatches or even deletions in the cell barcode sequences. These errors will cause the sequencing reads from the same cell to be spliced into other barcodes, which would inflate the cell number. We adopt a relatively simple method to correct all kinds of errors. Specifically, the raw barcodes are sorted by abundance, and cell barcodes within 1-bp mismatch are aggregated into the consensus barcode with higher read counts. After the aggregation, each cell barcode consists of the original reads and corrected (but containing barcode errors) reads. The proportion of the corrected reads relative to the total reads for each barcode is calculated and defined as the cell barcode error rate. The error rate of cell barcodes with more than 20K reads is used for the statistical analysis.

Determination of valid cell barcodes

The number of cell barcodes after correction greatly exceeds the number of input cells, which is named as “barcode inflation.” Most of these cell barcodes are derived from droplets with beads but no cells. We assume that a valid cell barcode from a high-quality cell should have sufficient mRNA molecules and higher read counts. There are multiple ways to infer the valid cell barcodes with the read count data. We adopt multiple strategies to determine the cell numbers (Figure S3). We first obtain a rough estimate of the captured cell numbers by considering the cell inputs, cell capture ratio, and other factors (referred to as Estimated). The numbers of cells with more than 20K and 40K reads are also determined (referred to as Reads \geq 20/40K). We also followed the method used in the Cell Ranger pipeline, which set a threshold for UMIs of cells. The hypothesis is that the top 1% of cells contains about 10 times as many UMIs as a typical cell. The resulting estimated cell number is referred to as the UMI (1/10). We also observed a distinct knee point on a plot of log-transformed barcode rank and barcode reads (Figure 2D). We generated the derivation of the plot of log-transformed barcode rank and barcode reads (Figure S3A) for 10X samples, and observed a sharp cliff around the estimated cell numbers. However, this phenomenon is subtler or absent for inDrop and Drop-seq samples, respectively.

The methods discussed above are all based on certain hypotheses or assumptions that may not apply for all methods and cell compositions. Here, we determine the actual number of cells in each sample by quality control and filtering of the number of cells by rough estimation based on experimental information. The valid cells are required to have at least 1,000 UMIs and the nearest correlation should exceed 0.6 (see Methods below).

Alignment and tagging reads to genes

The reads are aligned to a reference genome (GRCh38) using STAR, a high-performance aligner for RNA-seq data (Dobin et al., 2013). Most of the reads can be uniquely mapped and the reads can be tagged according to the annotation. A read is required to have at least 50% of its length mapped and overlapping with an exon region. For multiple aligned reads, the reads should be from the same gene to be counted.

UMI corrections

The UMIs could also be influenced by sequencing errors. In our process, the UMIs in the same gene from the same cell barcode are sorted by their counts. UMIs within 1 Hamming distance are aggregated. We observed that, for some highly expressed genes with hundreds to thousands of UMIs, the amount of time required for UMI correction may increase exponentially. We disabled the correction for those genes with UMI counts larger than 100 by default.

Comparison with protocols' official analysis pipeline

The official analysis pipeline was downloaded and performed under their instructions, respectively. We combined the UMI table for around 100 cell barcodes for the official pipeline and our versatile pipeline (Figure S2A). The gene expression levels were compared between the two analyses. We also calculated and compared the total numbers of UMIs for each cell barcode with the two pipelines (Figure S2B).

Distribution of reads on the gene structures

Picard Tools (<http://broadinstitute.github.io/picard/>) and its RNASEqMetrics command were used to analyze the distribution of mapped reads on the genome and gene bodies. The gene structure annotation for GRCh38 was downloaded in refFlat format from the UCSC genome browser. The following regions were counted: coding, UTRs, intronic, and intergenic regions.

Quality control and filtering of cells

The UMI counts reflect the molecular diversity of the transcriptome. We evaluated the base contents and numbers of UMIs for all cells. The qualified cell barcodes should contain more than 1,000 UMIs. We further generated a Spearman's correlation matrix for all valid cell barcodes using their UMI counts of the top 500 genes. Then, we obtained the nearest neighbor for each cell barcode and calculated the maximum pairwise correlation. These values revealed outliers, so cell barcodes with nearest correlation < 0.6 were removed.

Proportion of effective reads

The estimated barcode reads counts reads from all estimated barcodes before filtering, and the reads mapped to genome are called as mappable reads. The UMI-effective reads counts reads that are mapped to gene body regions which contribute to molecular counting. The remaining reads are further filtered by two quality control filtering.

Technical noise analysis

The nearest correlation analysis was performed by calculating Spearman's correlation coefficients for all valid pairwise barcodes. To evaluate the technical noise at the gene level, we randomly selected 500 cells from each sample with their top 1000 highly expressed genes. To adjust for the differences in molecular number and capture efficiency, the UMIs were normalized by multiplication by a factor that made the sum of normalized UMI counts equal 1,000,000 for each cell (UMIs per million).

Technical noise of housekeeping genes

The list of housekeeping genes (HK genes) was downloaded from <https://www.tau.ac.il/~elieis/HKG/>. The housekeeping genes and nonhousekeeping genes (non-HK genes) were labeled with different colors to depict their distribution regarding the overall noise level. The HK genes were associated with the lowest amount of biological noise and their overall noise level could be approximately treated as the technical noise. The similarity of the distributions between HK and non-HK genes showed that the biological noise level was notably below the technical noise level.

Sequencing depth subsampling and normalization

To ensure a fair comparison of the performances of the three protocols without concern for their different sequencing depths, we conducted subsampling analysis. We subsampled the raw sequencing reads for raw sequencing depths from 10% to 90% at intervals of 5%. The resulting data were then subjected to the same processing pipeline as described above. We obtained the corresponding cell barcodes, median raw sequencing depth, UMIs, and genes at each sequencing depth. As Drop-Seq-1 was sequenced at a relatively low sequencing depth (~37K reads), we selected the subsampling ratio for each sample to make the median sequencing depth (raw reads) approach 37,000.

Smart-seq2 protocol in inDrop system

We tested the Smart-seq2-based protocol on the inDrop platform. In the droplet generation step, SuperScript III was replaced by SuperScript II (10 μ L/100 μ L RT inlet) and 2 μ L of 100 μ M Template Switch Oligo (TSO, AAGCAGTGGTATCAACGCAGAGTA CATrGrGrG) was also added to 100 μ L of RT inlet. At the reverse transcription step, two strategies were tested. One involved exposure to 42°C for 1.5 h (42°C RT) and the other involved 42°C for 1.5 h followed by 50°C for 2 min and 42°C for 2 min for 10 cycles (42/50°C RT). After reverse transcription, demulsification was performed the same as with the standard inDrop protocol. The aqueous phase was purified using 0.6 \times AMPure beads. At the cDNA amplification step, two strategies were tested. One involved all of the cDNA being pooled together for amplification (Pool PCR) and the other involved splitting cDNA into tubes with about 200 recovered cells per tube and pooling together after amplification (Split PCR).

Comparison of public datasets of the three systems

We collected multiple datasets by these three systems preferential from the original published papers. For datasets without the provision of raw sequencing read files, we used the processed UMI expression tables downloaded from Gene Expression Omnibus (GEO). The raw sequencing files were subsampled to obtain a relatively uniform sequencing depth (about 30K reads per cell). We also included a HEK293 sample by 10X prepared in our own laboratory. The cDNAs after demulsification was split into two halves for downstream reactions. So for each barcode, it contains about half the amount of a normal HEK293 cell. We obtain the raw files (Fastq/SRA) of the following datasets: mouse ES cells by inDrop ([Klein et al., 2015](#)), brain cells by 10X, pan T cells by 10X, HEK293 by Drop-seq ([Macosko et al., 2015](#)) and 10X. These datasets were reanalyzed using our own developed pipeline to obtain digital expression matrices. To identify valid cell barcodes, all datasets were filtered using standard of UMIs > 1000. The unavailable values are labeled by "NA."

The same kind of cells makes a more fair comparison among the systems. Datasets of zebrafish from similar stage exist for all systems. inDrop and 10X can be compared by the pan T cell or T regulatory cells. Drop-seq and 10X can be compared using their HEK293 datasets.

QUANTIFICATION AND STATISTICAL ANALYSIS

Sequencing depth fitting and prediction

When samples are sequenced at different depths, it interferes with the comparison between samples and methods. Thus, we randomly subsampled the raw reads to different sequencing depths (from 10% to 90%, at intervals of 10%) and measured their performance. The numbers of detected UMIs and genes, and the precision level index for each depth were calculated as described above. The fitting and prediction of the saturation level of genes and UMIs were performed using the following equation.

$$y = a + \frac{b}{x+c}$$

PCA and tSNE analysis

We use the Seurat package (<https://satijalab.org/seurat/>) (Butler et al., 2018) for PCA and tSNE analysis. We randomly selected 500 cells from each sample for efficiency. The pipeline proceeds through the cell and gene filtering, data normalization, and identification of the most variable genes. The pipeline also reports the marker genes for each sample which are higher expressed compared to other samples. We generated the top 100 marker genes for each sample and these genes are aggregated according to their systems. The length and GC content for each gene are calculated by averaging the value from all of the corresponding transcripts.

DATA AND SOFTWARE AVAILABILITY

The accession number for the data reported in this paper is GEO: GSE111912. The data processing pipeline (baseqDrops) is available at github (<https://github.com/beiseq/baseqDrops>).

Supplemental Information

Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems

Xiannian Zhang, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, Zeyao Li, Yanyi Huang, and Jianbin Wang

Supplemental figures and tables

Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems

Xiannian Zhang,^{1,#} Tianqi Li,^{2,#} Feng Liu,^{3,#} Yaqi Chen,^{4,#} Jiacheng Yao,² Zeyao Li,⁵ Yanyi Huang,^{1,*} and Jianbin Wang^{2,*}

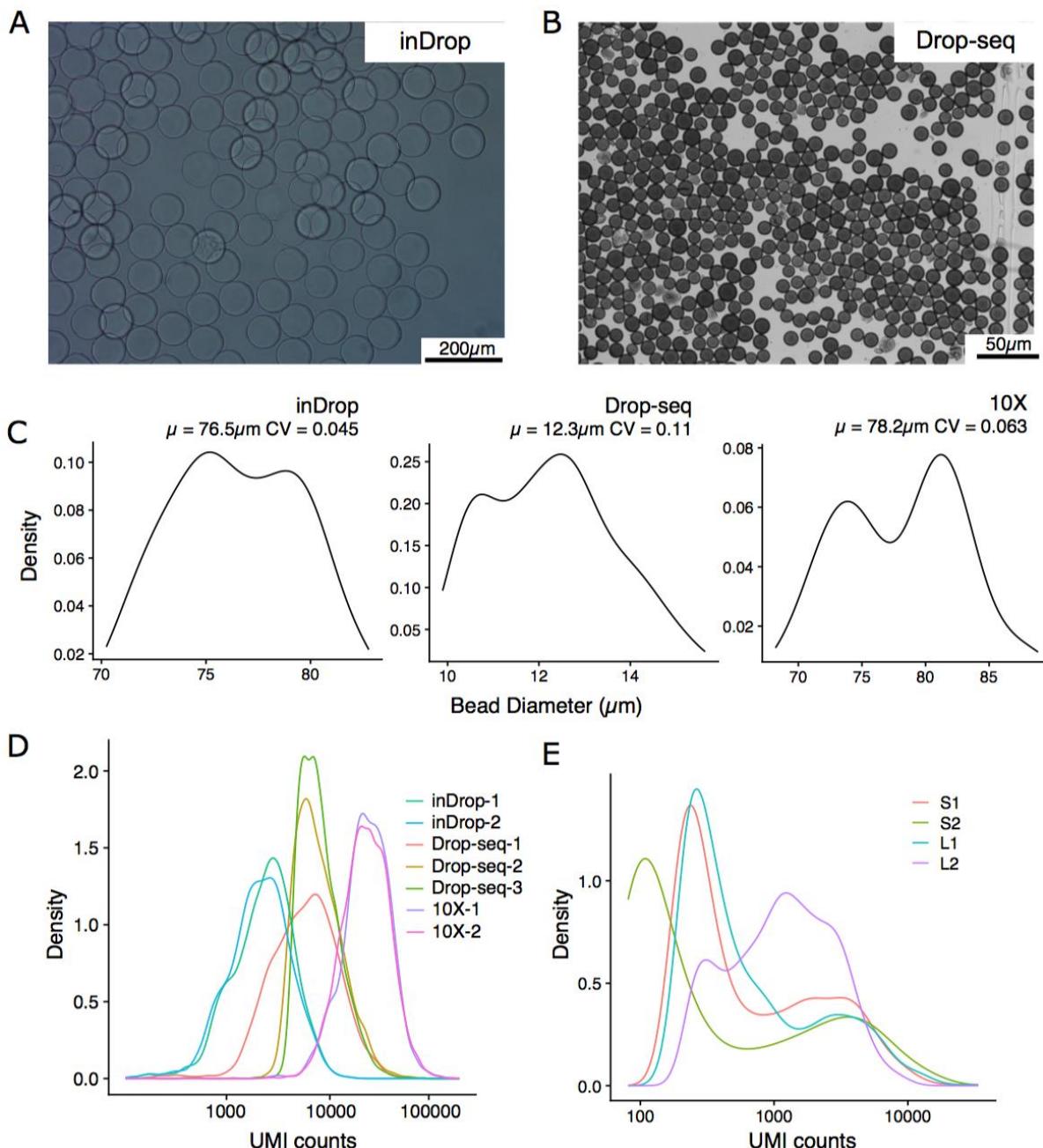


Figure S1 (Related to Figures 1 and 7). The size distribution of Drop-seq and inDrop beads. Bright-field microscopic images of Drop-seq beads (A) and inDrop beads (B) under different magnifications. (C) The bead diameter was measured through image processing and the density distribution of the bead diameter was plotted for inDrop, Drop-seq, and 10X. The mean and coefficient of variation (CV) of bead diameter are indicated. (D) The UMI distribution of all samples followed a unimodal distribution. (E) The UMI distribution of the four Smart-seq2 inDrop samples. The sample names and their corresponding conditions are shown in Figure 7.

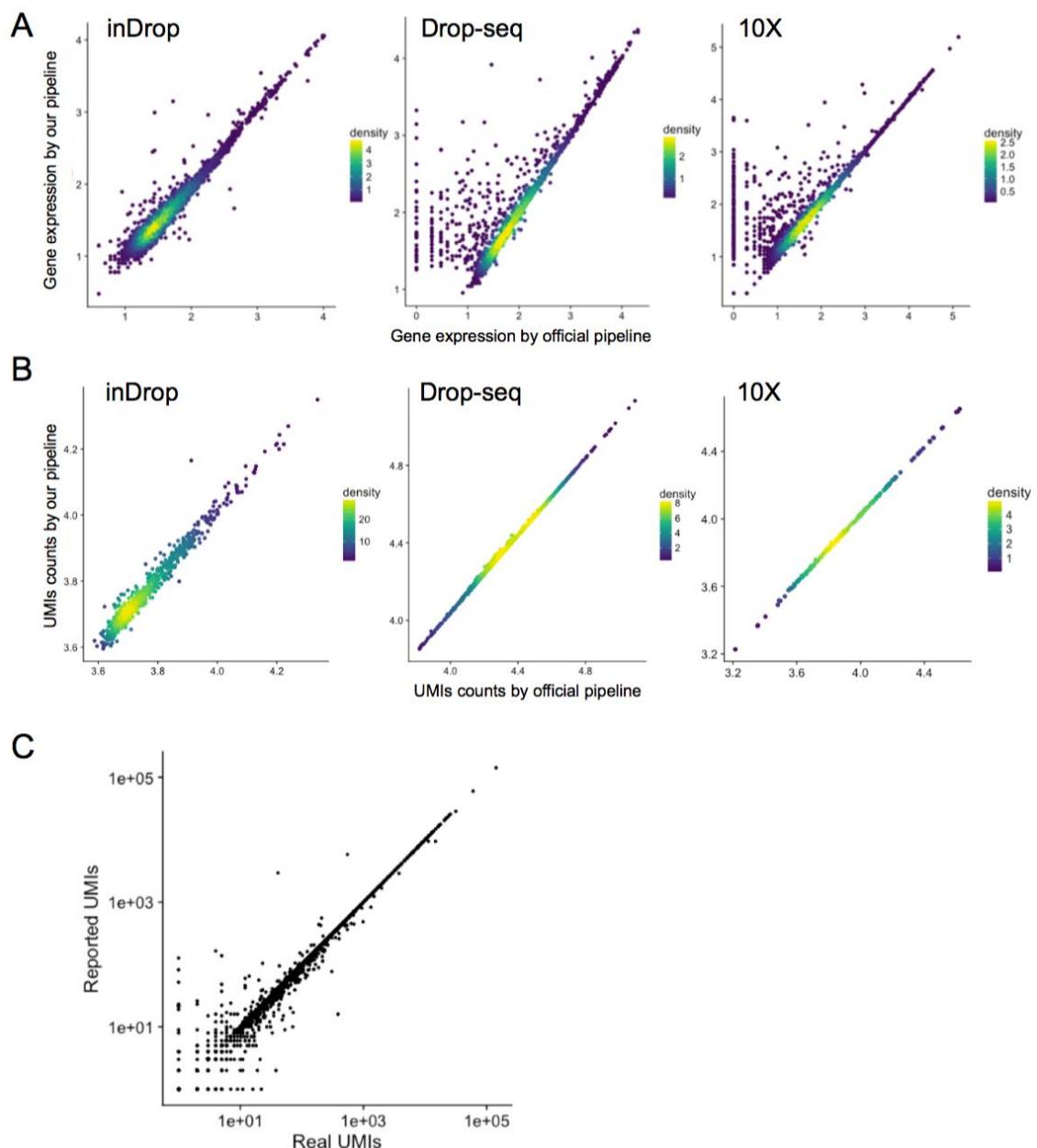


Figure S2 (Related to Figure 3). Validity of the data processing pipeline. (A) We compared the aggregated gene expression level (log10-transformed) of approximately 100 cells between the official analysis pipeline and our pipeline. (B) The total numbers of UMIs (log10-transformed) in each cell quantified by the official pipeline and our pipeline show high consistency. (C) Comparing the actual number of UMIs and that estimated by our pipeline using simulation data.

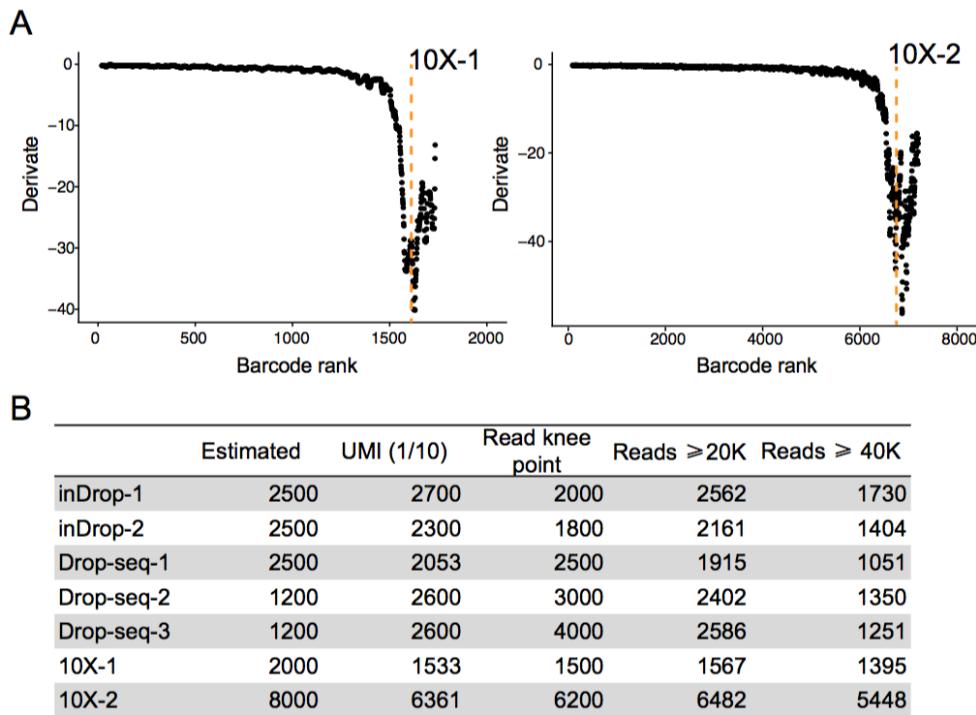


Figure S3 (Related to Figure 3). Determining the valid cell numbers. (A) The derivation of the plot of barcode rank vs. read count; a cliff can be seen around the estimated cell numbers. This corresponds to a knee point in the plot of barcode rank vs. read count. The cell numbers estimated by this method (referred to as the read knee point method) are labeled by a dashed line (B) The numbers of cells estimated by different methods (see Methods). Here, the term “estimated” indicates that the number is obtained by considering all experimental evidence. UMI (1/10) means that the lower level of UMI is set as 1/10th of the level of the top 1% of barcodes. Reads $\geq 20K$ or 40K will require the corresponding minimal number of reads.

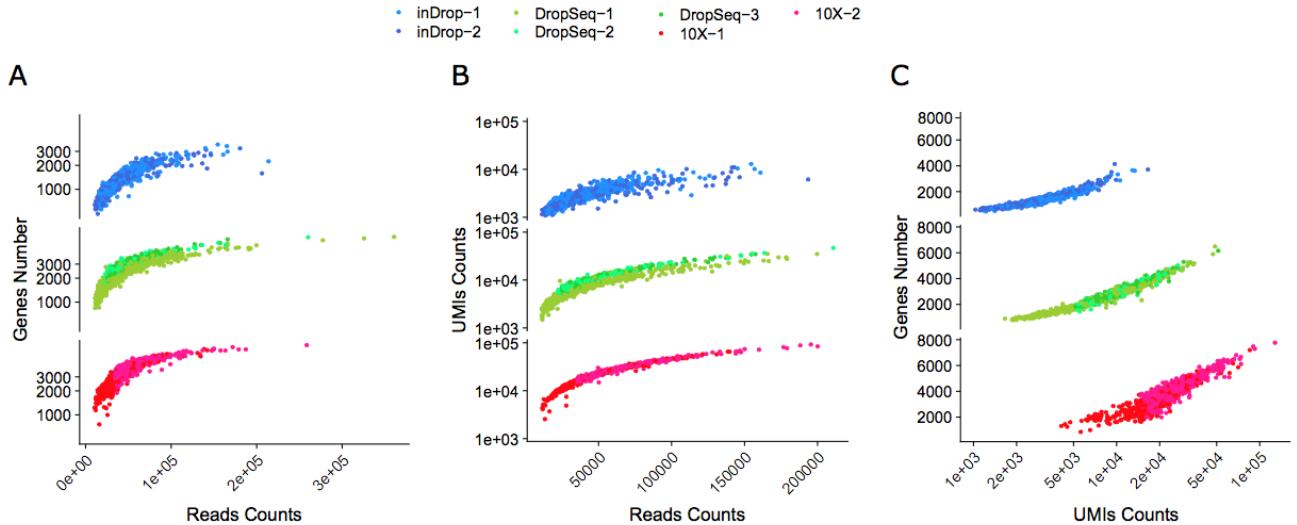


Figure S4 (Related to Figure 4). Pair-wise relationship between the measurements of gene expression sensitivity and precision in all cells from the same sample. (A) Gene number vs. read count. (B) UMI count vs. read count. (C) Gene number vs. UMI count.

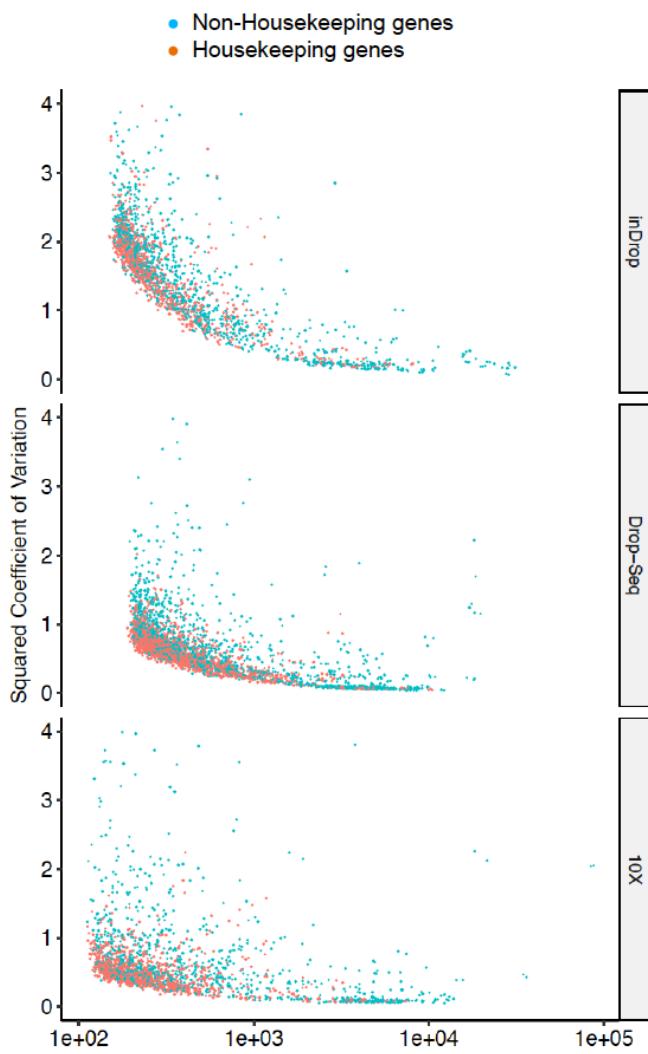


Figure S5 (Related to Figure 4). Technical and biological noise level of each method.

The housekeeping and nonhousekeeping genes are labeled with different colors. The similarity of their distributions indicates that the level of technical noise overwhelms the biological noise.

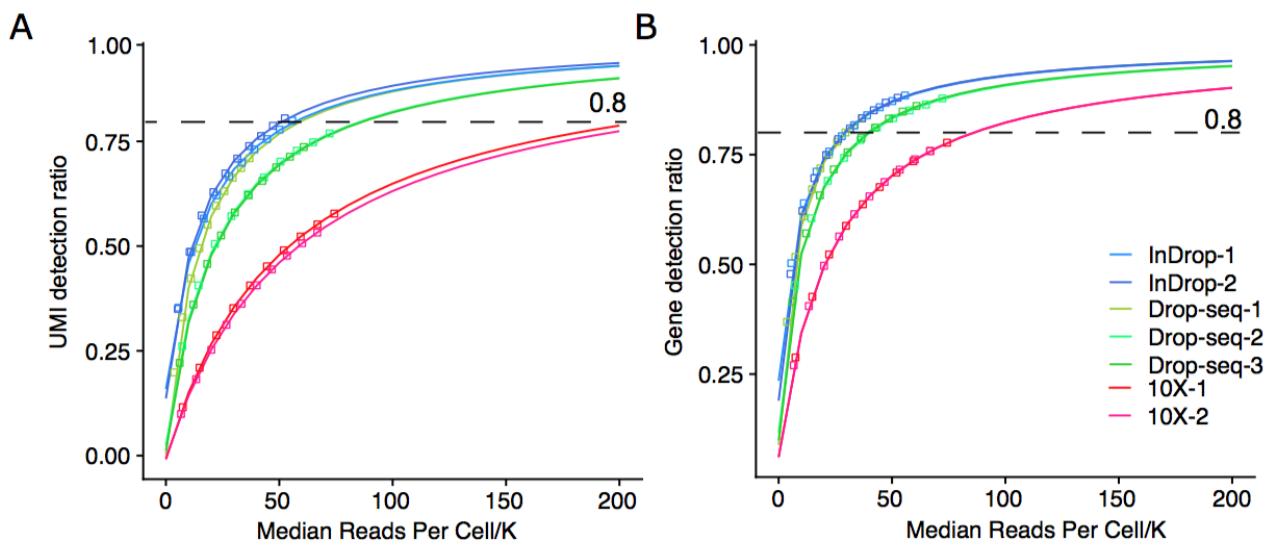


Figure S6 (Related to Figure 5). Sequencing depth saturation analysis. UMIs (A) and genes (B) detection ratio at each subsampled sequencing depth are plotted. The subsampling depth is measured as the median number of reads of all valid cells in a sample. The curve is fitted, and the saturation level is predicted. The detection ratio is normalized to the saturation level of UMI or gene numbers in each sample (see Methods).

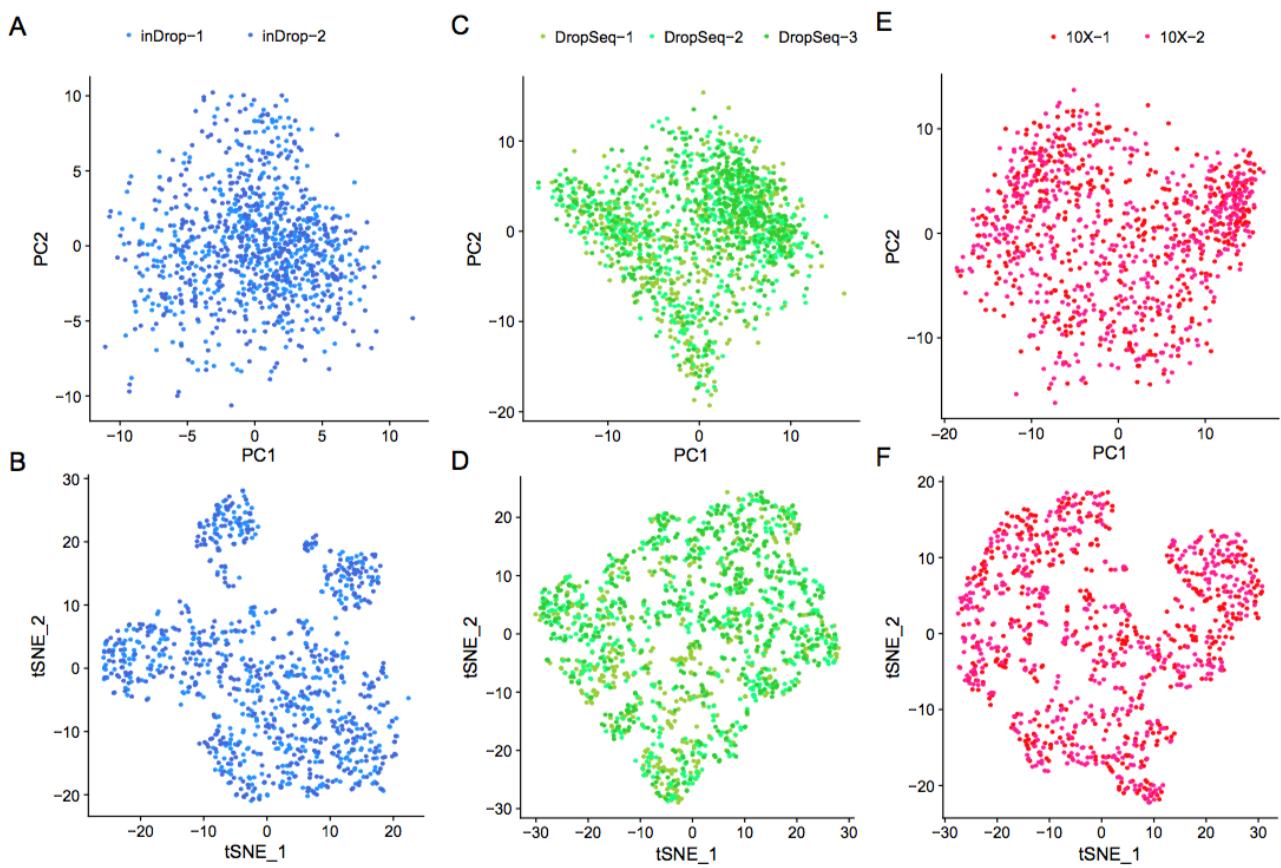


Figure S7 (Related to Figure 6). Clustering of cells from samples with the same system using PCA and tSNE analyses. The systems are inDrop (A, B), Drop-seq (C, D), and 10X (E, F). The dimension reduction methods are PCA (A, C, E) and tSNE (B, D, F).

Table S1 (Related to Figure 2). The top 20 detected UMI sequences in each system.

inDrop		Drop-seq		10X	
UMI Sequence	Depth	UMI Sequence	Depth	UMI Sequence	Depth
CCCCCC	20	TTTTTTTT	36.6	TTTTTTTTTT	44.8
CCCCCT	9.7	GGGGGGGG	17.9	GTTTTTTTTT	14.7
CCCCCA	7.9	CTTTTTTT	13.1	ATTTTTTTTT	14.7
TTTTTT	7.1	AGGGGGGG	11.7	CTTTTTTTTT	12
GCCCCC	6.8	GAGGGGGG	11.3	GGTTTTTTTT	9.2
ACCCCC	6.8	GGGGGGGT	10.7	TATTTTTTTT	7.2
CCCCCG	6.5	CGGGGGGG	10.5	AATTTTTTTT	7.1
CCCCTT	6.1	ATTTTTTT	10.4	GCTTTTTTTT	7
ATTTTT	6.1	GGAGGGGG	10.2	GATTTTTTTT	6.9
CCGCC	6	TGGGGGGG	10.1	TGTTTTTTTT	6.8
AATTTT	5.8	GGGGGAGG	10	AGTTTTTTTT	6.6
CCGCC	5.7	GTTTTTTT	10	CATTTTTTTT	6.4
CCCACC	5.6	GGGGGGGA	10	TCTTTTTTTT	6.2
CGCCCC	5.5	GGGAGGGG	9.9	ACTTTTTTTT	5.9
CCACCC	5.4	GGGGAGGG	9.3	CGTTTTTTTT	5.8
AAATT	5.4	GTGGGGGG	8.7	CCTTTTTTTT	5.4
CACCCC	5.3	GGGGGGAG	8.7	GGATTTTTTT	5.2
CCCCAC	5.3	TCTTTTTT	8.7	GTATTTTTTT	4.9
CCCTCC	5.2	CCTTTTTT	8.6	AGGTTTTTTT	4.7
AAAAAA	5.1	AAGGGGGG	8.6	GGCTTTTTTT	4.7

The normalized depth is calculated as demonstrated in Method.

Table S2 (related to STAR methods). The mislabeled genes in read-to-gene tagging simulation.

Real gene	Mislabeled as
RPL41	AC090498.1
IGHA1	IGHA2
RPS15A	RP11-1035H13.3
IGLC3	IGLC2
MIF	AP000350.10
RPL36A	RPL36A-HNRNPH2
IFITM1	IFITM2
ARPC1B	ARPC1A
C20orf24	TGIF2-C20orf24
TUBA1B	TUBA1C
RPL36	HSD11B1L
PSMA1	NA
FKBP1A	FKBP1C
NDUFA11	AC024592.12
AC090498.1	RPL41
PA2G4	RP11-603J24.9
HNRNPA1	HNRNPA1L2
HNRNPA1L2	HNRNPA1
VAMP2	RP11-599B13.6
NDUFB8	RP11-411B6.6
FKBP11	AC073610.5
NDUFB8	NA
TMSB4X	NA
TUBA1B	TUBA1A

Mislabeled genes in read-to-gene tagging simulation (see Methods). Most of these genes overlap in the genome or have similar sequences.

Table S3 (related to Figure 4). The highly variable genes from each system. The genes are ranked by the coefficient of variance (CV). The criteria of filtering are shown in Methods.

10X			Drop-seq			inDrop		
Gene	CV	Mean_Norm_UMI	Gene	CV	Mean_Norm_UMI	Gene	CV	Mean_Norm_UMI
IGLC3	2.551	2264.403	HIST1H4C	1.352	3083.957	IGLC2	2.236	5811.328
JCHAIN	1.950	4792.352	IGLC2	1.328	17186.632	IGHM	2.120	7665.682
IGHA2	1.401	2186.472	IGHM	1.123	17755.828	MALAT1	1.435	3064.135
IGHA1	1.394	25045.873	HSPA5	0.992	2918.335	MTRNR2L12	1.082	6777.936
IGLC2	1.364	105380.787	MALAT1	0.963	10510.735	MT-ATP8	0.917	2413.889
FTH1	0.853	4607.734	FTH1	0.891	3429.481	FTH1	0.907	3365.871
ISG15	0.852	2200.911	FTL	0.821	3181.098	HLA-DRB1	0.892	2431.433
ACTB	0.850	7805.019	HSP90B1	0.746	4117.068	PTMA	0.815	3285.668
FTL	0.811	6396.760	MT-ATP8	0.713	2108.570	SRGN	0.798	6019.020
ACTG1	0.765	2816.229	ACTB	0.692	4465.837	HLA-DRA	0.756	3967.253
TMSB4X	0.707	33627.828	PDIA3	0.685	2405.884	MT-ND5	0.756	5430.640
PTMA	0.705	4846.445	MT-ND5	0.659	2831.306	IFITM2	0.688	2359.603
GAPDH	0.696	5946.673	HSP90AA1	0.563	2780.779	PFN1	0.685	2810.210
CD74	0.694	4239.958	MT-ATP6	0.556	5223.957	CD74	0.664	4899.420
PFN1	0.653	2963.908	HSPD1	0.545	2797.749	ACTB	0.659	16324.214
MT-ATP6	0.635	2122.471	MT-CYB	0.521	2809.355	HLA-C	0.652	2820.471
TMSB10	0.605	4284.120	MT-ND2	0.516	4022.501	RPS13	0.649	3211.120
MALAT1	0.603	14898.154	MT-ND3	0.514	2345.705	HLA-B	0.647	2862.303
YBX1	0.588	2171.211	HNRNPA2B1	0.510	2623.687	NPM1	0.646	2572.229
RPS24	0.566	2671.524	MT-ND4	0.510	3831.651	HLA-A	0.637	2451.213
HLA-B	0.553	2257.010	MT-CO2	0.506	5319.697	GAPDH	0.631	2739.710
MT-CO3	0.544	7057.905	MT-ND1	0.504	4194.428	MT-CYB	0.609	10615.063
MT-ND2	0.542	3116.763	NCL	0.503	3348.467	PPIA	0.602	2400.896
MT-CYB	0.542	3246.955				TMSB4X	0.591	15151.985
SUB1	0.530	2086.202				RPL39	0.573	2512.036
						RPS20	0.562	2539.395
						RPS10	0.560	2131.890
						UBA52	0.560	2410.621
						EIF1	0.557	4003.951
						MT-ATP6	0.557	15550.839
						ACTG1	0.556	5873.674
						MT-ND3	0.554	17580.354
						RPL19	0.522	2308.035
						RPLP1	0.520	4997.513
						RPL14	0.512	2590.718
						MT-ND2	0.512	17652.350

Table S4 (Related to Figure 7). Statistics of performance on gene and UMI detection for four experiments of Smart-seq2 in the inDrop system with various conditions.

	Raw Reads	Valid Reads(with barcode in whitelist)	Claimed Cells	Claimed Cell Reads	Claimed Cell UMI Reads	Valid CB counts	Median reads	Median UMI-effective reads	Median gene counts	Median UMI counts
L1	203213268	123,672,238	1300	100,595,865	61,688,856	366	101543	62509	1075	2939
L2	257489533	165,149,928	900	124,648,424	80,839,124	492	108251	70966	822	1949
S1	293336617	182,604,144	1,300	152,093,346	100,414,242	440	113533	76759	925	2560
S2	221840332	144,839,268	650	123,550,924	80,182,371	176	101817	66333	996	2841

L1 is RT+Pool PCR. L2 is 42/50 °C RT+Split PCR. S1 is 42 °C and RT+Pool PCR. S2 is 42 °C RT+Split PCR. The RT reaction temperature for L1 and L2 is 42/50 °C, while that for S1 and S2 is 42 °C. In S1 and L1, the cDNA is pooled together for amplification. In S2 and L2, the cDNA is split for PCR.

Table S5. (Related to Figure 4) The summarization of some published datasets from the three systems. The tissues or cell lines, cell numbers, sequencing depth and detected UMIs are demonstrated.

Protocol	Cell Type	Source	UMI >= 1000 CBs	Median UMIs	Total reads per cell	Mapped reads per cell
inDrop	HEK293	Ours*	2,500	6,373	39,189	35,034
inDrop	Zebrafish 14hpf	3	4,001	7,259	53,379	37,831
inDrop	Mouse ESC	1	703	5,667	40,088	34,119
inDrop	Human Treg	2	1,726	1,647	32,774	20,217
Drop-seq	HEK293	4	587	11,472	32,243	28,771
Drop-seq	Zebrafish 12hpf	5	4,413	2,497	NA	11,945
Drop-seq	Mouser Retina	4	3,207	4,398	NA	NA
10X	HEK293	8	504	21,400	32,361	31,260
10X	HEK293*	Ours**	2,288	9,015*	16,321*	15,522
10X	Zebrafish 12hpf	5	3,000	10,577	NA	NA
10X	Mouse Brain	6	931	8,545	56,718	53,372
10X	Human T Cells	7	3,465	3,731	31,788	29,709

Reference:

1. Klein A M, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells[J]. Cell, 2015, 161(5): 1187-1201.
2. Zemmour D, Zilionis R, Kiner E, et al. Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR[J]. Nature immunology, 2018, 19(3): 291.
3. Wagner D E, Weinreb C, Collins Z M, et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo[J]. Science, 2018, 360(6392): 981-987.
4. Macosko E Z, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets[J]. Cell, 2015, 161(5): 1202-1214.
5. Farrell J A, Wang Y, Riesenfeld S J, et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis[J]. Science, 2018, 360(6392): eaar3131.
6. https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons_900
7. https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/t_3k
8. https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_1k

Notation:

*A different batch of beads.

**We used half of the cDNA for downstream library preparation and sequencing (see Methods).

Table S6. (Related to STAR Methods) The estimation of RNA-seq library preparation costs per cell for all three platforms.

	inDrop	Drop-seq	10X
Instrument price	\$50,000	\$30,000	\$50,000
Instrument cost per cell ¹	\$0.07	\$0.04	\$0.07
Running cost per cell	\$0.10	\$0.10	\$0.50
Sequencing cost per cell ²	\$0.30	\$0.30	\$0.30
Overall cost per cell	\$0.47	\$0.44	\$0.87

Notation:

1: We assume each platform is used to prepare 1000 cells per day and each instrument can provide service for 2 years.

2: Assuming each cells get 50000 reads on average (150bp pair-end).