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] Zeyao Li,[5] Yanyi Huang,[1,*] and Jianbin Wang[2,*]

[1] Beijing Advanced Innovation Center for Genomics (ICG), Biodynamics Optical Imaging Center (BIOPIC), School of Life Sciences, College of Engineering, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China. [2] School of Life Sciences, and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China. [3] National Research Center for Translational Medicine (Shanghai), State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China. [4] Peking University Shenzhen Graduate School, Shenzhen 518055, China. [5] School of Life Sciences, and Peking-Tsinghua-NIBS Joint Graduate Program, Tsinghua University, Beijing 100084, China..

[#] These authors contributed equally to this work.

* Corresponding authors. E-mail: yanyi@pku.edu.cn (Y.H.) or jianbinwang@tsinghua.edu.cn (J.W.)

Summary

Since its debut in 2009, single-cell RNA-seq has been a major propeller behind biomedical research progress. Developmental biology and stem cell studies especially benefit from the ability to profile single cells. While most studies still focus on individual tissues or organs, recent development of ultra-high-throughput single-cell RNA-seq has demonstrated potential power to depict more complexed system or even the entire body. Though multiple ultra-high-throughput single-cell RNA-seq systems have acquired attention, systematic comparison of these systems is yet available. Here we focus on three prevalent droplet-based ultra-high-throughput single-cell RNA-seq systems, inDrop, Drop-seq, and 10X Genomics Chromium. While each system is capable of profiling single-cell transcriptome, detailed comparison revealed distinguishing features and suitable application scenario for each system.

Introduction

Single-cell RNA-seq (scRNA-seq), firstly introduced in 2009 (Tang et al., 2009), has become one of the most powerful approaches to unveil the heterogeneous nature of biology. The ability to manipulate picograms of RNA material in single cells enabled studies with unprecedented temporal and spatial resolution. Based on the large data of whole-transcriptome, scRNA-seq has provided comprehensive landscapes of gene expression and their regulatory interactions at the finest resolution, enabling accurate and precise depiction of cell types and cell states (Grun and van Oudenaarden, 2015; Tanay and Regev, 2017; Wu et al., 2017). In the past decade, the sensitivity and precision of mRNA quantification through scRNA-seq has been greatly improved (Hashimshony et al., 2016; Picelli et al., 2014), bringing revolutionary discoveries to many fields such as cell type deconstruction 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), inferring transcription dynamics and regulatory networks (Deng et al., 2014; Dixit et al., 2016) and identifying the development,

evolution and heterogeneity in tumors (Patel et al., 2014; Treutlein et al., 2014; Venteicher et al., 2017) .

The experimental throughput is always a major concern in scRNA-seq experimental design. In some particular biological systems, such as the early stage embryos, only dozens of cells is sufficient to achieve critical findings (Yan et al., 2013). However, due to tissue complexity, dynamic cell cycle or other intrinsic variations (Buettner et al., 2015), and technical noise (Brennecke et al., 2013), RNA-seq data from a small number of cells is typically not adequate to reflect the full image of biological samples (Tanay and Regev, 2017). The transcriptome detection sensitivity is known to saturate quickly with increasing sequencing depth (Svensson et al., 2017). Shallow sequencing of massively sampled single cells can effectively reduce random variation and define cell types through clustering analysis, 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 to prepare a large number of cDNA libraries. Laboratory automation can solve the labor problem but the reagent cost is still high (Jaitin et al., 2014). A few recently published microfluidic approaches have demonstrated various advantages in scRNA-seq (Prakadan et al., 2017). For example, small volume reactors may facilitate reaction efficiency and reduce technical noise when coupled with proper chemistry (Streets et al., 2014; Wu et al., 2014). Lab-on-chip operations also make single cell isolation much easier than manual picking (Shalek et al., 2014). Microwell-based scRNA-seq systems (Fan et al., 2015; Han et al., 2018) proved their advantages in low cost and high throughput. However, due to the lack of commercialized instrument or detailed protocols, microwell-based scRNA-seq has not been widely adopted.

Droplet microfluidics can achieve fast compartmentation and encapsulation at the frequency up to dozens of thousands per second and be easily scaled to produce millions of droplets, each having nanoliter volume to accommodate single-cell reactions (Agresti et al., 2010). The microfluidic pipeline layout is very simple, consisting mainly micro-channels introducing/collecting reagents and samples (Duncombe et al., 2015). This droplet strategy greatly elevates the reaction throughput and dramatically reduces cost. Currently, there are three prevalent droplet-base 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 them have been demonstrated to be robust and practical in generating cDNA libraries for thousands of cells in a single run with acceptable cost. All three methods use similar design 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 take different approaches for bead manufacturing, barcode design, cDNA amplification, and thus have different experimental protocols. Given these differences in system specifications and potentially in transcriptome analysis (Ziegenhain et al., 2017), a systematic and unbiased comparison among these methods is highly desired.

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 lymphoblastoid cell line GM12891. The average sequencing depth was around 50,000-60,000 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 ratio, barcode detection error and transcript detection sensitivity were analyzed and compared. The results reveal strength and weakness in each system and provide guidance for system selection in future research.

Results

System overview

In all three systems, the cell barcodes are embedded in the micro-bead-tethered primers (Figure 1A). The DNA sequences of on-bead primers share a common structure, containing 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 while 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 Poisson distribution (Figure 1A). As a result, the cell capture efficiency can be much higher in 10X and inDrop approaches. 10X is reported to have ~80% bead occupancy, and ~50% cell capture rate (Zheng et al., 2017).

The material of beads may also influence the quantity and density of DNA primers. 10X and inDrop's adoption of hydrogel 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 the primers into the solution phase to boost mRNA capture efficiency. InDrop also mobilizes the primers by UV irradiation induced cleavage. In contrast, Drop-seq uses surface-tethered primers to capture the mRNA molecules, and the capture efficiency could be problematic compared with 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 bring more uniform results due to the limited reagents encapsulated. It is also known that reaction in limited volume like droplets enhanced the yield of cDNA conversion (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 (Ramskold et al., 2012). The *in vitro* transcription step in inDrop extends the library preparation time beyond 24 hours, while both Drop-seq and 10X pipelines can be finished within the same day.

Experimental design and data processing

We used GM12891, a human lymphoblastoid cell line for our comparative study. Biological replicates were setup for all three systems, inDrop, Drop-seq and 10X, with various cell inputs in different days and batches (Figure 1B). We adjusted sequencing depth to obtain a comparable level of read-per-cell-barcode across the three systems.

Each system has its own data processing pipeline. However, none of them can directly handle data generated by other systems due to difference 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 these differences may introduce bias in gene quantification, which is not ideal for a fair comparison between systems. To solve the problem, we develop a versatile pipeline which accepts data from all these systems and generates matrices of UMI counts (Figure 1C). We applied this pipeline to our data and conducted the comparisons on sensitivity, precision and bias in an objective way.

The script of pipeline is available online (https://github.com/basedata10/DropRNA) for free 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 raw data. After removing cell barcodes with too low read counts, the pipeline corrects cell barcode errors (see Methods for details). These errors may have been introduced during on-bead primer synthesis, and also during PCR or sequencing steps. Reads with the same cell barcodes are aggregated, and invalid cell barcodes are removed after filtering by read counts. For 10X and inDrop where 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. To validate the accuracy of transforming aligned reads to the correspond genes, we performed simulation by generating around 2.2 million reads based on the cell line's gene expression profile (ref, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM424360). More than 99% of the reads (2,229,156 out of 2,251,529) were tagged to the correct gene. The remaining 1% ambiguous reads were mainly derived from genes with paralogs or overlapped genes, such as RPL41/ AC090498.1 or IGHA1/IGHA2 (Table S2)

After read-to-gene assignment, the analysis pipeline further removed duplicate reads based on UMI information and produces a gene expression matrix.

The processing speed of this pipeline has been optimized by reducing the read/write payload which is a common bottleneck. For example, ~50% of reads from inDrop data have invalid sequence structure. By removing these reads we can increase the data processing efficiency correspondingly. Furthermore, storage of the intermediate results such as tagging reads with gene site is generally avoided. The reads are split into multiple (typically 16 or 64) files, based on cell barcode prefix, to enable parallel processing.

Quality of primers on beads

The complexity of cell barcodes determines the maximum capacity for a single experimental run using droplet-based scRNA-seq. Low cell barcode complexity might result in barcode collision and produce artificial doublets. The three systems claim theoretical cell barcode complexity of 1.47×10^5 (inDrop), 1.6×10^7 (Drop-seq) and 7.34×10^5 (10X), respectively. However, the effective complexity may be lower than the designed value. We first estimated the effective cell barcode complexity by counting barcode collision events between different runs from each system (Figure 2A, see Methods). For 10X, the effective cell barcode complexity was close to the claimed value. However, only about 30% (~ 5×10^4) of inDrop's designed cell barcodes were observed. Drop-seq used arbitrary random sequences for cell barcode, and by rough estimation, the effective cell barcodes were about 1×10^6.

One-barcode-one-bead is the key requirement for all three systems. However, due to imperfection of the DNA synthesis chemistry, asynchronous base addition is inevitable. Cell barcodes sequence inconsistency could happen within the same bead. Error-containing cell barcodes result in the inflation of detected single cell number, and hence need to be corrected carefully. We aggregated the cell barcodes within 1 Hamming distance. For each valid cell barcode, the ratio between read counts before and after correction is calculated as barcode quality ratio (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 fabrications. Meanwhile, more than half of cell barcodes contained obvious mismatches in the other two systems. Specifically, about 10% of Drop-seq beads contained 1-base deletion in cell barcodes, which also required extra caution during data analysis (see 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 preferences 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 poly-G in Drop-seq and 10X. Such patterns, majorly due to DNA synthesis bias, may cause system-dependent skewness of the RNA-seq results.

It is challenging to determine the accurate cell number, represented by cell barcodes, in each sample. This is due to the great dispersion on cellular mRNA molecular counts and their capture efficiency. 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 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. We could determine the cell number by detecting the inflection point of the distribution curve. For Drop-seq, however, there was no obvious cliff on the read-count curve for a clear cut-off. This might originate from the wide size distribution of beads used by Drop-seq. We noticed that the size of beads used in inDrop or 10X are more uniform than that in Drop-seq (Figure S1), likely due to the difficulties in size control when fabricating resin beads. We used the unified threshold of 40,000 reads to remove invalid cell barcodes for Drop-seq as well as the other two systems. Such threshold level is also close to the obvious inflection points for 10X and inDrop. The numbers of cell barcodes and percentage of valid reads retained after read count filtering are shown in Figure 2E and 2F.

Data processing steps and results

The reads split into each valid cell barcodes are first aligned to the human genome to analyze the distribution of reads throughout the genome regions (Figure 3A). Drop-seq has more than 75% of the reads mapped to UTR (mainly 3'UTR) and exonic regions, while this ratio in inDrop is only about 55%. The effective read number in each experimental run has hence declined (Figure 2F). After the tagging of reads that mapped to gene bodies, the numbers of detectable genes can be obtained (Figure 3B). The number of genes declines

in accordance with the number of reads within a cell excepts for several outliers in Drop-seq data. We use those detected genes to demonstrate the bias of reads 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 their library construction strategies. Drop-seq data showed a bimodal distribution, most likely due to the same PCR anchor sequences at both ends of cDNA molecules.

We performed another round of cell barcode filtering based on the total count of UMIs (transcripts) in each experimental run (Figure 2E, 2F and 3D). With a total UMI cut-off of 1,000, about 2-3% of valid cell barcodes in Drop-seq data failed to meet this requirement, while almost all data of the other two systems passed this filter (filtering rate 0% for 10X and 0.3% for inDrop).

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 so different from its closest cell barcode neighbor (Spearman correlation <= 0.6, see Methods), we discarded the cell barcode. For 10X, nearly all the cell barcodes with enough read counts passed this nearest correlation filtering. InDrop (~95%) and Drop-seq (~97%) cell barcodes also passed with high percentage (Figure 2E, 2F and 3E). At last, the percentage of effective reads is ~70% for 10X, and ~30% for inDrop and Drop-seq. The ratio of such reads should be maximized to reduce waste of sequencing capacity (Figure 2F).

We noticed that many of the discarded cell barcodes in Drop-seq results were present in all three experimental runs (see Methods). This is likely because Drop-seq uses random 12-mers rather than predesigned sequences as cell barcodes. Thus, some highly abundant sequences, for example the adaptors or poly-A containing sequences, could be mistaken as cell barcodes. A cell barcode whitelist can help filter out these artificial cell barcodes.

Sensitivity of detecting UMIs and genes

The sensitivity of gene detection is the fundamental indicator of scRNA-seq performance. It reflects the overall efficiency of a method 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 input sample, the sensitivity can be depicted simply with the recovered UMI (Figure 4A) and gene counts (Figure 4B). The technical replicates from the same system showed high consistency and reproducibility. 10X had the highest sensitivity, capturing over 26,000 transcripts from ~4,000 genes on average. This performance is consistent regardless of input cell numbers. Drop-seq detected ~10,000 transcripts from ~2,600 genes. The inDrop system had relatively lower sensitivity, detecting ~4,200 UMIs from ~1,400 genes. The read distribution is more skewed in inDrop and Drop-seq, where the majority of cell barcodes have relatively low read counts (Figure 4B). The UMI and gene numbers get gradually saturated for cell barcodes with increasing reads counts (Figure S2A-B). Interestingly, we found a that log-transformed UMI counts is highly correlation (Spearman correlation r>0.9) with the gene numbers detected (Figure S2C).

Technical Noise and Precision

Technical noise is a reflection of the variation brought by experimental randomness, including transcript dropout in reverse transcription and the bias of PCR amplification. Precision can be assessed by the concordance of transcriptome among technical replicates. The main 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 real distances and obscure subtle biological differences between cells, thus lowering the resolution of cell grouping. Many efforts have been taken to reduce the technical noise such as the use of UMI to eliminate the quantification error caused by amplification bias.

Although we are using a seemingly homogeneous cell line, there still exists intrinsic biological noise or heterogeneity (Prakadan et al., 2017). In our dataset, the total variation is contributed by technical and biological parts, which are hard to separate. Here we assume that biological noise is consistent among samples and the technical noise dominates the variations in the datasets. Then the overall total variation can reflect the technical noise level.

The overall total variation is first characterized as the median Spearman correlation (name as Median Correlations) between a specific cell barcode and every other cell barcode in the entire datasets (see Methods). In general, it depicts the similarity preserved among biological replicates after the introduction of technical noise. To validate the effect of UMIs in reduction of PCR amplification noise of gene counting, we performed the analysis using UMI counts and raw read counts, respectively, for gene expression quantification. The result (Figure 4C) shows 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 than those using raw counts, thus verifying 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.

We further performed analysis using the highest Spearman correlation (name as Nearest Correlations) (Figure 4D). This metric shows the distance between a cell and its nearest neighbor in the sense of transcription profile. It can serve as a particular indicator of a system's capacity of identifying minor cell-types with a few cells. Again, consideration of UMI information can effectively reduce noise.

The technical variation at the gene level can be measured by the coefficient of variation (CV) of normalized UMI counts across all cells (Figure 4E). 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, then inDrop. Interestingly, many highest expressed genes are quite noisy, especially in the 10X data. We examined these genes (normalized UMI >= 2,000, CV>=0.5) and found most of them being the cell line's most abundant genes or mitochondrial genes (Table S3). High noise in these genes is probably introduced by stochastically burst manner of transcription (Sanchez and Golding, 2013).

Saturation of sensitivity and precision at low sequencing depth

The ability to detect low abundant transcripts could be enhanced with deeper sequencing. However, there is a trade-off between costs and sensitivity, especially for high-throughput

experiments. It is an empirical practice that each cell gets 10,000-100,000 reads in high-throughput scRNA-seq experiments, whereas conventional scRNA-seq data usually ~1 million reads per cell (Baran-Gale et al., 2017). We randomly subsampled sequencing data and analyzed the corresponding changes in sensitivity and precision (Figure 5A-B, Figure S3). The fitted saturation curves of UMI and gene counts help determine a suitable sequencing depth for most applications.

All the systems show a diminishing return at a higher depth. For inDrop and Drop-seq, 80% of the total saturated UMIs can be detected with about 60,000 read/cell. Meanwhile, 10X need ~180,000 read/cell to accomplish the same goal due to the higher sensitivity. Detection sensitivity of gene numbers saturated faster. To reach the 80% saturation level, ~30,000 read/cell is needed for inDrop or Drop-seq, and ~70,000 for 10X.

Other than sensitivity, precision also determines a system's resolution for making biological discoveries. Here, the precision is measured as the Median Correlations between one cell and the others, which also indicates the level of technical noise. We checked how the precision level was affected by the sequencing depth and found that the precision index quickly saturated with read depth (>=20,000 effective reads) for all three systems (Figure 5C). These results help us estimate an empirical guideline for experimental design. For the prevailing tasks such as cell typing, a median number of 20,000 read/cell should be sufficient.

Bias in gene quantification

To comprehensively compare the transcriptome depicted by different systems, we conducted dimension reduction with principle component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE) analyses (Figure 6A). Almost all the cells were robustly separated and clustered according to their system origin. This provides an evidence that factors such as sequencing depth, gene number and UMI counts are not the major factors for cell typing analysis. As the replicates are processed in different batches and days, batch effect is also obscure. Within the same system, different batches of data show very homogeneous distribution (Figure S4).

The separation of cells by systems indicates that there exists system-specific quantification bias at gene level. Potential biases in the mRNA enrichment at gene level could be related to three major factors: expression abundance, gene length and GC content. We hence selected the top 100 marker genes (see Methods) from each method and analyzed the distribution on these factors (Figure 6B-D). These genes show consistent expression intensity among biological replicates. We found that, compared to the others, 10X slightly favored shorter genes and the genes with higher GC contents, whereas Drop-seq better detected genes with lower GC contents. This observation echoes with previous report that Drop-seq overestimates the low GC ratio or longer transcripts (Macosko et al., 2015).

In summary, all the methods seem to be very consistent and homogeneous among technical replicates from different batches. This enables the validity of combining different dataset together from the same method. However, different protocols have obvious bias related to gene length and GC contents. Thus, it will introduce extra divergence from methods by combining these datasets together directly.

Discussion

We have compared three most popular droplet-based high-throughput single-cell RNA-seq systems, inDrop, Drop-seq, and 10X, using identical cell sample and a unified data processing pipeline to reduce bias in experimental design and data analyses. Technical replicates are included to verify the possible batch-dependent artifacts. For each system, we have sequenced thousands of single cells. Through quantitatively analyzing a few key parameters using our unified data processing pipeline, we have learned 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 indicate obscure batch effects, but noticeable clustering bias in association with systems of choice. Such observation indicates that cell typing analysis using datasets from mixture of systems is technically challenging and should be avoided.

For all three systems, the beads are specifically provided by the corresponding manufacturers and unlikely to be easily home-made. Thus, the quality of the

beads, such as size dispersity, is especially 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 is also a key factor affecting the bioinformatics pipeline, which has to rule out the corresponding artifacts and errors.

Our comparison shows that 10X generally has higher molecular sensitivity and precision, and less technical noise. As a more maturely commercialized system, 10X protocol should have been extensively optimized, which is partially reflected by their barcode design and quality control of bead manufacture. However, high performance optimization also comes with high price tags. The per cell cost for 10X is around $1, without consideration of sequencing cost or instrument depreciation.

With small compromises in sensitivity and precision, Drop-seq presents significant advantage in experimental cost, which commonly becomes the major concern when a large number of single cells are needed. The experimental cost of Drop-seq is about a few cents per cell. As an open-source system (except for the beads), Drop-seq has gained popularity since its debut in 2015. As of the time we prepared this manuscript, Drop-seq protocol has been downloaded nearly 60,000 times. Drop-seq can be a reasonable choice for individual labs for its balanced performance and budget.

To certain extent, inDrop can be considered an open-source version of 10X. Both of them use hydrogel beads for super Poisson loading. Their on-bead primers are both releasable to facilitate the capture of transcripts. We attribute the relative lower performance of inDrop to its excessive cDNA amplification (Hashimshony et al., 2016), as well as the unfinished optimization of the protocol. As an open-source system, inDrop can adopt other chemistries, and be easily modified for different types of RNA-seq protocols. In a preliminary attempt, we tested the implementation of Smart-seq2, the most widely used scRNA-seq protocol, on inDrop system. The cDNA profile resembled closely conventional Smart-seq2

products (Figure 7A). In general, inDrop could be a desirable system for non-standard approaches or technical development.

With all these system specific features mentioned above, we propose a flowchart (Figure 7B) to help with the choice of suitable droplet-based scRNA-seq system for ultra-high-throughput single-cell studies. While most projects work with relatively large cell numbers, very precious samples such as human embryos require efficient cell capture. Super-Poisson distribution of cell capture could be essential for such precious samples. The requirement for experimental cost and transcript detection efficiency could be subjective. Generally, all three systems offer satisfiable transcript detection efficiency, and higher efficiency comes with higher experimental cost. By rule of thumb, currently 10X is a safe choice for most applications. When the sample is abundant, Drop-seq could be more cost-efficient. When detection of low abundant transcripts is optional, or custom protocol is desired, inDrop becomes a better choice.

Acknowledgments

We thank F.W. for testing the data process pipeline. This work was supported by National Natural Science Foundation of China (21327808, 21525521 to Yanyi Huang and 21675098 to Jianbin Wang), Ministry of Science and Technology of China (2016YFC0900100 to Jianbin Wang), and Beijing Advance Innovation Center for Genomics.

Author Contributions

Y.H. and J.W. designed the experiments. T.L. and F.L. performed the experiments. X.Z., Z.L. and Y.C. analyzed the data. X.Z., J.W. and Y.H. wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.C., Marquez, M., Klibanov, A.M., Griffiths, A.D., and Weitz, D.A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proc Natl Acad Sci U S A *107*, 4004-4009.

Baran-Gale, J., Chandra, T., and Kirschner, K. (2017). Experimental design for single-cell RNA sequencing. Brief Funct Genomics.

Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C.*, et al.* (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.

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., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol.

Deng, Q., Ramskold, 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 e1817.

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.

Grun, 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 e1017.

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.

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.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A.*, et al.* (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776-779.

Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, 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.

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.

Papalexi, 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., Bjorklund, 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*.

Ramskold, 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.*, et al.* (2014). Microfluidic single-cell whole-transcriptome sequencing. Proc Natl Acad Sci U S A *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 phenomenology 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 oligodendroglioma. 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*.

Villani, A.C., Satija, R., Reynolds, G., Sarkizova, 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*.

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.

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.*, et al.* (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 e634.

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.

Figure titles and legends

Figure 1: System and experiment overview. (A) Overview and comparison of experimental features of the three systems. (B) Experimental scheme summary. (C) Overview of the data processing pipeline work flow. The same data processing pipeline is applied for all three systems.

Figure 2. On-bead primer quality and data filtering. (A) Estimation of real cell barcode complexity for each system (see Methods). The observed barcode collisions between different samples of the same system determines the most possible real cell barcode library size. (B) Cell barcode error rate distribution. For each cell barcode, its error rate was measured as the total percentage of reads with 1 Hamming distance in the total corrected 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. A uniform cutoff of 40,000 reads per cell barcodes is applied to all samples. (E) The number of valid cell barcodes after each step of quality control filtering. (F) The percentage of reads remained after each process or quality control step.

Figure 3. Data processing steps and results. (A) Percentage of reads mapped to different genomic regions are calculated after alignment to genome. (B) The number of genes detected with cell barcode ranked by read counts. (C) Normalized reads distribution across gene body from 5’ to 3’ end. (D) The number of UMIs with cell barcode ranked by their read counts. (E) The distribution of Spearman correlation between one cell barcode and every other cell barcode within the same sample.

Figure 4: Transcription analysis sensitivity and noise. (A) Summary of cell barcode numbers, median read counts, and median gene and molecular detection performance for each sample. (B) The distribution of counts of reads, UMIs and genes. (C and D) Transcription analysis noise measured as the median (C) and highest (D) Spearman correlation between one cell barcode and every other cell barcode within the same sample. Gene quantification through UMI counts (solid

line) and read counts (dashed line) are both adopted. (E) The CV-mean (CV squared) plot of each system.

Figure 5: Dependence of transcription analysis sensitivity and noise on sequencing depth. Median numbers of UMIs (A) and genes (B) detected for each sample with increasing effective read counts. The data points are calculated and fitted to get their respective saturation value (See Methods). (C) Transcription analysis noise level saturates quickly with sequencing depth. The noise was measured as the median Spearman correlation between one cell barcode and every other cell barcode within the same sample.

Figure 6: Transcription analysis bias in the three systems. (A-B) Visualization of cell barcodes of all three systems clustered by PCA and tSNE. (B-D) Demonstration of transcription analysis bias in gene expression level (B), gene length (C) and GC content (D). The top 100 marker genes from each system were used for demonstration.

Figure 7: The flexibility of inDrop and guidance for choosing the suitable system. (A) The size analysis of cDNA from Smart-seq2 protocol in inDrop system. (B) Flowchart to help making a decision on the suitable systems according to demand.

Methods

Cell lines and cell preparation

GM12891 cell line was bought from Coriell Institute. The complete growth medium was made by RPMI-1640 Medium with L-glutamine (Life Technologies cat. No.11875-085), 10% fetal bovine serum (Life Technologies cat. No.16000-044) and 1% penicillin and streptomycin. The cell line was incubated with 5% carbon dioxide at 37 ℃ in culture flask. Cell concentration is maintained between $5x10^5$ and 10^6 cells/mL. Before experiment, general cell condition is confirmed under microscope. Regular circle shape cells and some cell aggregates indicate a good cell state. Cells were collected by centrifugation at 150 g for 5 min and later counted with hemocytometer.

Drop-seq experiment

Cells were washed with PBS-BSA for three times and filtered with a 40-micron cell drainer. We then counted the cells and adjusted concentration to 100 cells/μL. All subsequent steps were carried out as detailed online by Macosko et al (http://mccarrolllab.com/dropseq/). Briefly, we loaded droplet-making oil, cells in PBS-BSA and barcoded beads (Chemgenes Barcoded Bead SeqB, cat. No. MACOSKO-2011-10) in lysis buffer into the droplet generating microfluidic device. Cells were lysed in droplets to released mRNA. Beads captured mRNAs in droplets. After demulsification, beads were pooled together. We conducted reverse transcription and ExoI digestion. Every 2,000 beads (or 100 STAMPs) were aliquoted into one PCR tube for PCR amplification. PCR product was pooled together and purified by AMPure XP Beads. We further constructed libraries and sequenced 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 following 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 droplets generation. Emulsion was collected in tube on ice and irradiated by a UV light to release primers. Reverse transcription was proceeded in droplets. After demulsification, hydrogel beads were filtered. The RT product was digested by ExoI/HinfI and purified by AMPure XP beads. Second strand cDNA was synthesized using NEB second-strand synthesis kit. After IVT, 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 sequenced on an Illumina HiSeq 4000.

10X system experiment

We performed all steps following 10X protocol. 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 droplets. We recovered cDNA through demulsification and beads purification. Pre-amplified cDNA further went through library preparation. Libraries were sequenced on an Illumina HiSeq 4000.

Cell capture efficiency calculation

The number of cells loaded into each system is counted as described above. The number of captured cells was determined using the number of Cell barcodes passing the quality threshold (40k raw reads, total UMIs>=1,000, nearest correlation ≥ 0.6).

Cell barcodes and UMI assignment

For 10X, we obtain the cell barcode and UMI from each read pair by extracting the first 16bp and following 8bp from read 1. Similarity, the barcode of Drop-seq can be accessed. The inDrop's design of barcode is more complicated as the full cell barcode contains two parts (name CB1, CB2) which are separated by a 22bp

spacer sequence called W1. The length of CB1 ranges from 8-12bp. We first located the W1 sequence by tolerating up to 2 mismatches. Then the length of CB1 and the whole cell barcode sequence can be determined. The retrieved cell barcodes are aggregated and the counts data are exported for downstream analysis.

Cell barcode complexity analysis

We count the number of cell barcodes that appeared in multiple samples (name as collision cell barcodes, or collision_CB). To estimate the collision_CB counts from pure statistical possibility, we perform a simulation on the number of collision_CB assuming the real library is the size as the method claimed.

For 10X, by simulation, the mean number of collision_CB is 10 by sampling 1395 and 5447 cell barcodes from 734,000 cell barcodes library. The observed collision_CB number is 13, which is close (~77%) to the theoretical value.

For inDrop, the simulation reported a mean collision_CB counts of 17. However, we detected 61 collisions between the 2 inDrop samples. This indicates that the real number of cell barcodes on the beads is roughly about 28% (17/61) of what is designed.

For Drop-seq, the simulation reports a mean collision_CB counts of 0.08 using the latter two of the samples. Surprisingly, we observe three such collisions between Drop-seq_2 and Drop-seq_3 (CB as ACGTTATCAGTT, TATGTTTGGTTT and TAACTCTCGCAT). The average collision_CB count is 1 in pairwise for the Drop-seq samples and by roughly estimation, the effective Cell barcodes size is about 1M.

Barcode correction

For ideal situation, the cell barcode sequences on the same beads should be the same. However, due 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 be spliced into other barcodes and the cell number

would be inflated. We adopt a relatively simple method to correct all kinds of errors. We rank the raw barcodes by their abundance, and cell barcodes within 1 bp mismatch are aggregated into the consensus barcode with higher read counts. After the aggregated, each cell barcode gets the aggregated counts considering errors. The ratio between the original and corrected counts for each cell barcode is calculated as mismatch ratio.

Determination of the valid cell barcodes

The number of cell barcodes after correction greatly exceed the number of input cells. Most of these cell barcodes are derived from droplets with beads but no cell. We assume that a valid cell barcode from a high-quality cell should have enough mRNA molecules and higher read counts. There are multiple ways for inferring the valid cell barcodes with the read counts data. Here we ranked the cell barcodes by read counts. For an ideal sample, there should be an inflection point between the valid and invalid cell barcodes. If there is no obvious inflection point, a minimum read count of 20,000 is required.

Alignment and tagging reads to genes

The reads are aligned to 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 read can be tagged according to the annotation. A read is required to have at least 50% of its length mapped and overlapped with the exon region. For the multiple aligned reads, the reads should be from the same gene to be counted.

UMI corrections

The UMIs also could 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 abundant genes with hundreds to thousands of UMIs, the time consumption may increase exponentially. We disabled the correction for those genes with UMI counts larger than 100 by default.

Reads Distribution on the gene structures

The Picard Tools (http://broadinstitute.github.io/picard/) and its RNASeqMetrics command is used to analysis the distribution of mapped reads on genome and gene bodies. The gene structures annotation for GRCh38 is downloaded as refFlat format from UCSC genome browser. The following regions are counted: coding, UTRs, intronic and intergenic regions.

Quality control and filtering of cells

The UMI counts reflect the molecular diversity of the transcriptome. We check the base contents and numbers of UMIs for all cells. The qualified cell barcodes should contain more than 1,000 UMIs. We further generate the Spearman correlation matrix for all valid Cell barcodes using their UMI counts of top 500 genes. Then we get the nearest neighbor for each cell barcode and calculated the maximum pair-wise correlation. These values reveal outliers, and cell barcodes with <0.6 nearest correlation are removed.

Technical noise

The median and highest correlation analysis are calculated as spearman correlation coefficients for all pair-wise valid barcodes. Nearest correlation is the Spearman correlation efficiency between one single cell and its nearest neighbors. Median correlations is the median Spearman correlation efficiency between one single cell barcode and every other cell barcode in the entire datasets.

For evaluating the technical noise, we randomly picked 500 cells from each sample. We picked the highest expressed 1,000 genes. To compensate the different molecular detection sensitivity, we normalize the UMIs into UMI per Millions.

Sequencing depth subsampling and fitting

We randomly subsample the total sequencing data with step size of 10%. The number of detected UMIs, genes and the precision level index for each depth are

calculated as described above. The fitting and prediction for the salutation level of genes and UMIs is performed using the model below.

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

PCA and tSNE analysis

We use the Seruat package (http://satijalab.org/seurat/) (Butler et al., 2018) for PCA and tSNE analysis. We randomly picked 500 cells from each sample. The pipeline goes through cell and gene filtering, data normalization and finding the most variable genes. The pipeline also reports the marker genes for each sample. We generate the top 100 marker gene for each sample which are highly expressed in one sample compared to others. The marker genes from the same protocols are combined. The length and GC content for each gene is calculated by averaging the value from all the corresponding transcripts.

Data availability

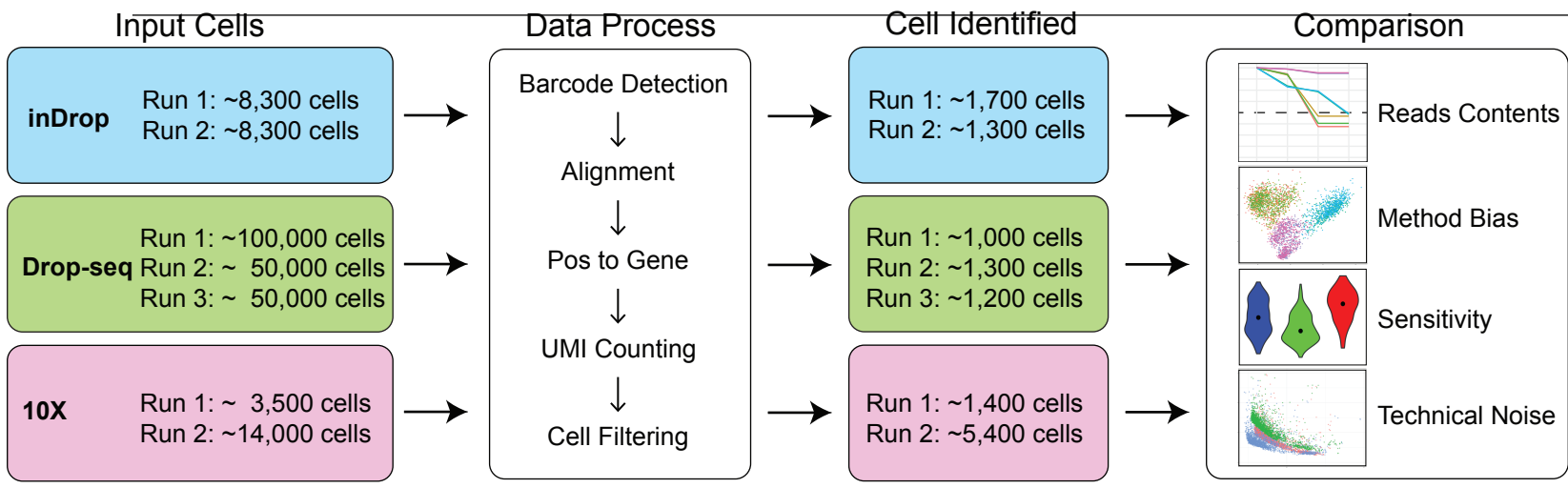
The data can be accessed in GSE111912 in GEO.

Figure 1

A

	inDrop	Drop-seq	10X
Barcoded Primer Bead	deformable; Photocleavable linker; T7p; Cell Barcode 38~41bp; PolyT; 5'; PCR Primer; UMI 6bp; 3'; Hydrogel	hard; 5'; Cell Barcode 12bp; PolyT 3'; PCR Primer; UMI 8bp	dissolvable; 5'; Cell Barcode 16bp; PolyT 3'; PCR Primer; 10bp UMI; Gel
Cell Barcode Capacity	147,456 (384 X 384)	16,777,216 (4^{12})	734,000
Droplet Generation	0.5 h; beads; oil; cells; RT/lysis reagents; oil; Beads: super-Poissonian; Cells: Poissionian	0.3 h; cells; oil; beads; lysis reagent; cells; oil; Beads: Poissonian; Cells: Poissionian	0.1 h; reagents; cells; oil; beads; cells; reagents; oil; Beads: super-Poissonian; Cells: Poissionian
Emulsion	lysis/reaction mix	lysis mix	lysis/reaction mix
Reaction in Droplets	2.5 h; • cell lysis • primer release by UV • mRNA capture • reverse transcription	0.3 h; • cell lysis • mRNA capture on beads	1 h; • cell lysis • primer release by bead dissolving • reverse transcription and template switch
Reaction after Demulsification	28 h; • 2nd strand synthesis • in vitro transcription • RNA fragmentation • RT-PCR	9 h; • RT and template switch • PCR • Tn5 tagmentation • PCR	7 h; • PCR • cDNA fragmentation and ligation • PCR

B



C

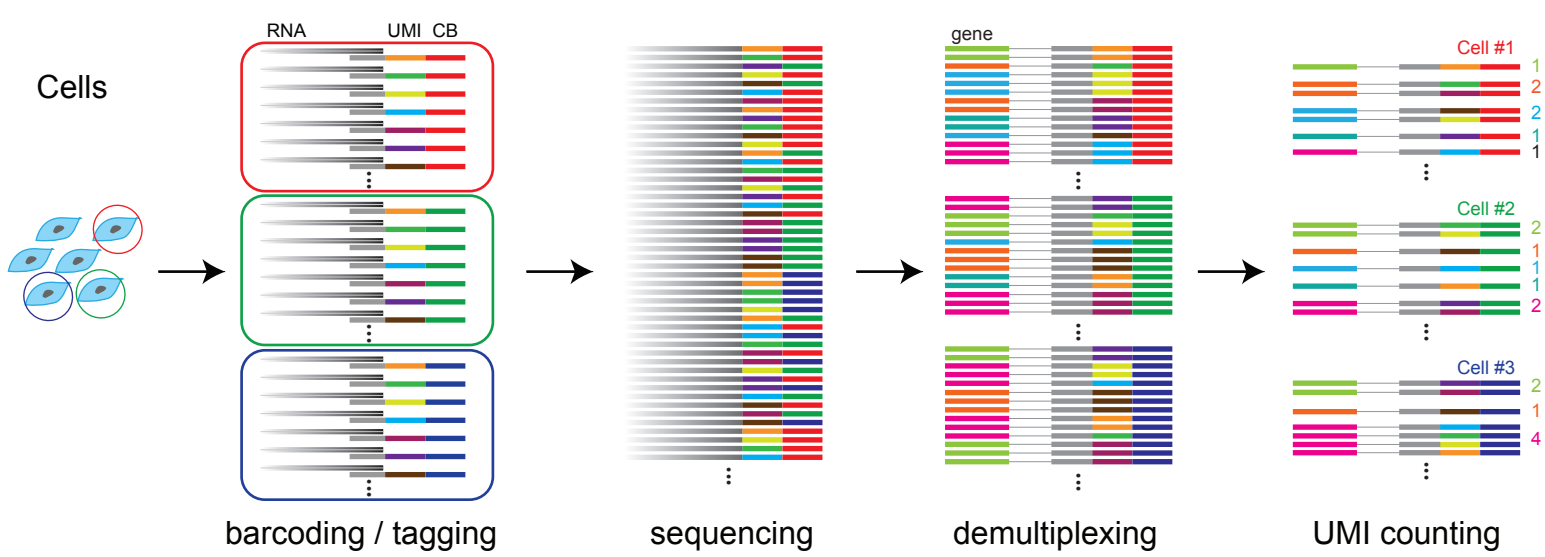
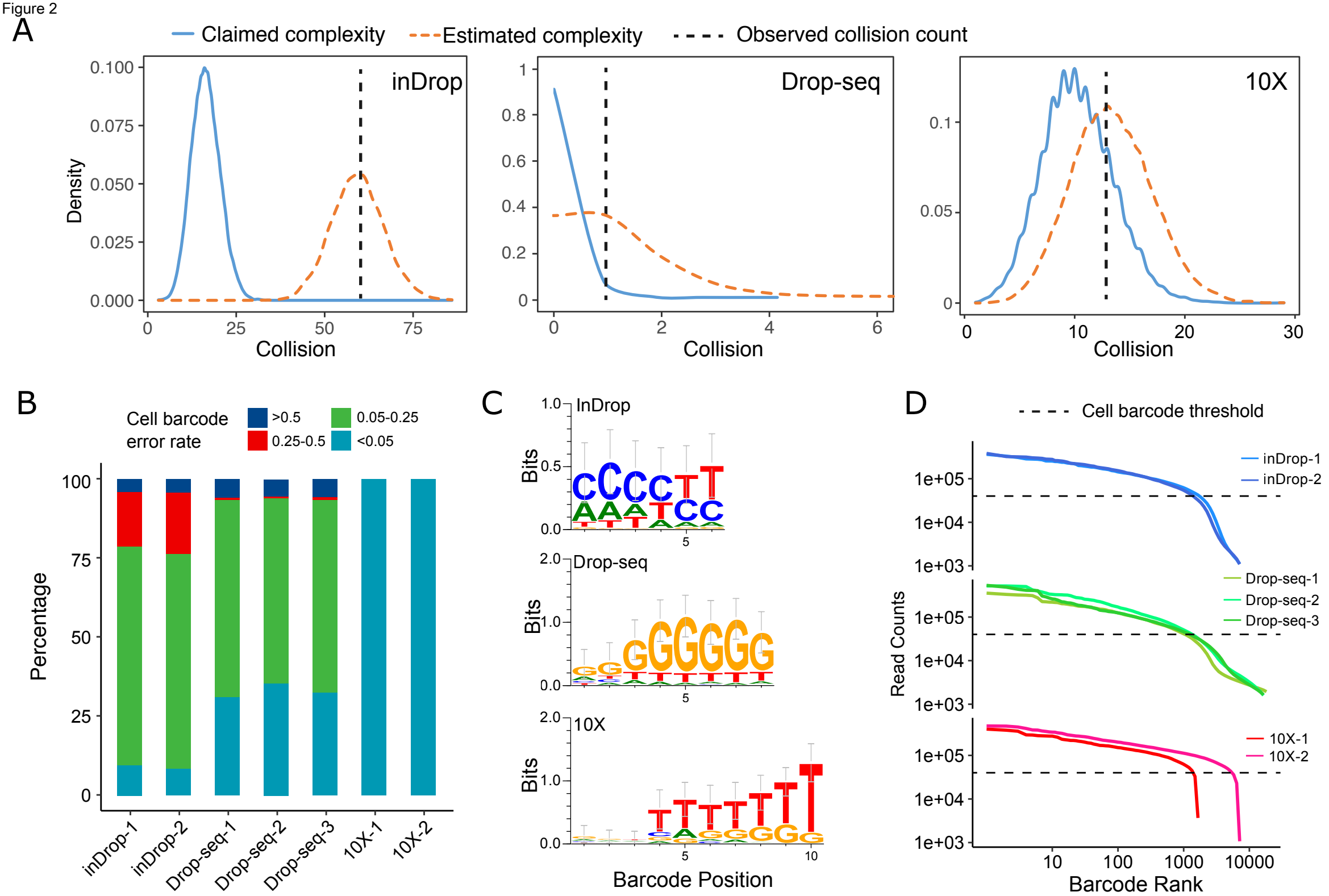


Figure 2



E

Sample	Reads ≥ 40k	UMIs ≥ 1k	Nearest corr ≥ 0.6
inDrop-1	1,730	1,724	1,665
inDrop-2	1,404	1,403	1,328
Drop-seq-1	1,051	1,032	1,029
Drop-seq-2	1,350	1,307	1,301
Drop-seq-3	1,251	1,231	1,228
10X-1	1,395	1,395	1,395
10X-2	5,448	5,448	5,447

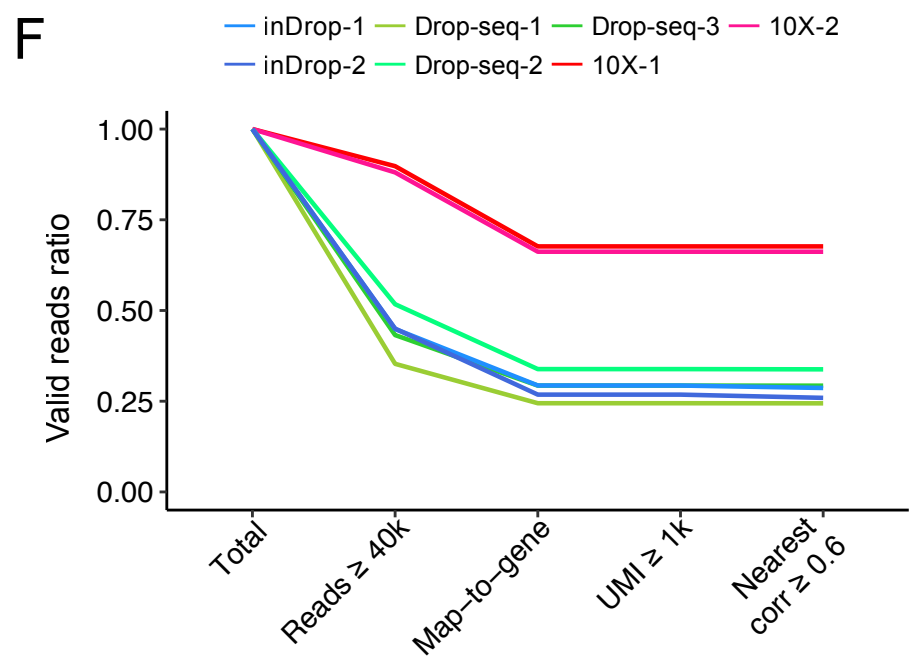


Figure 3

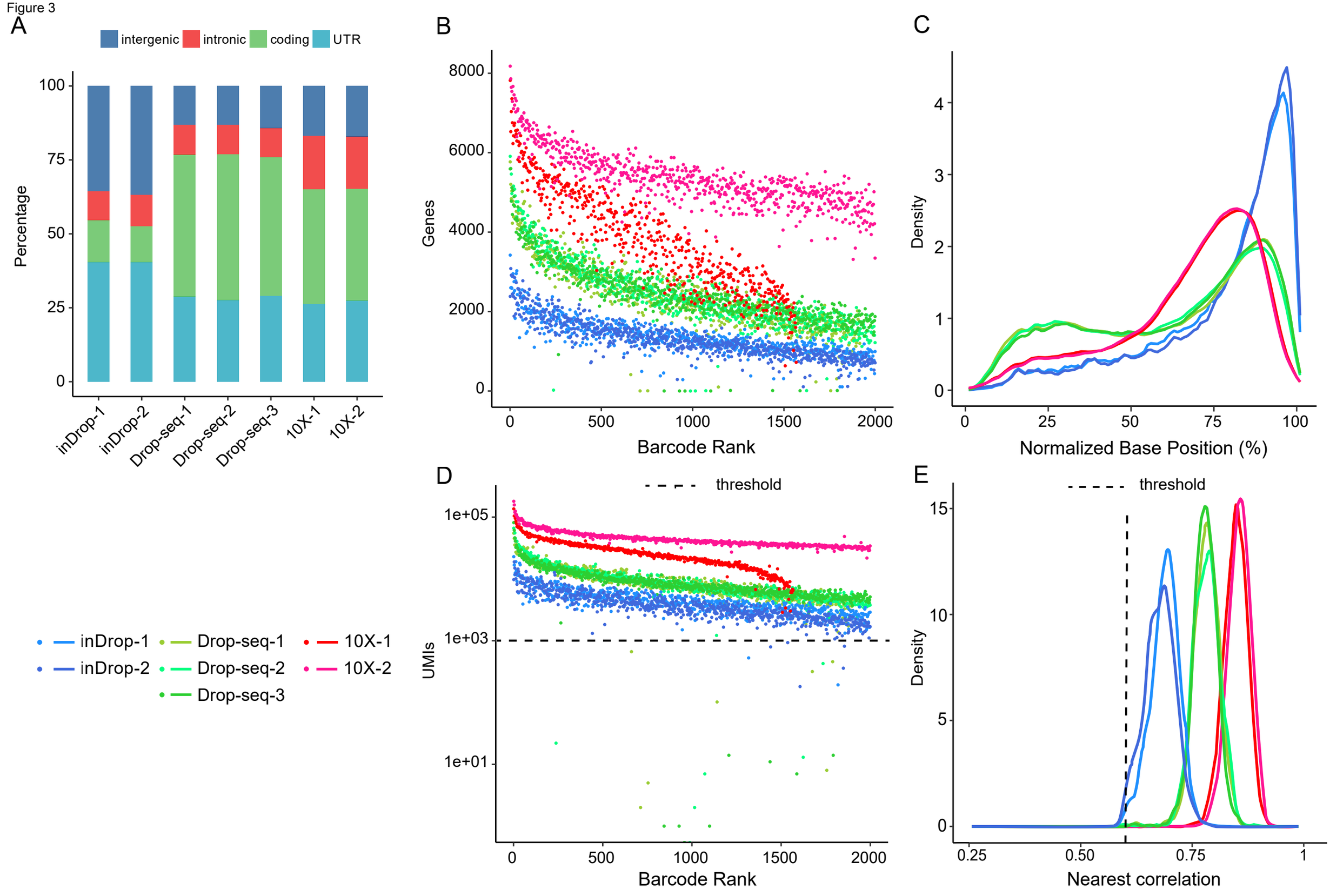
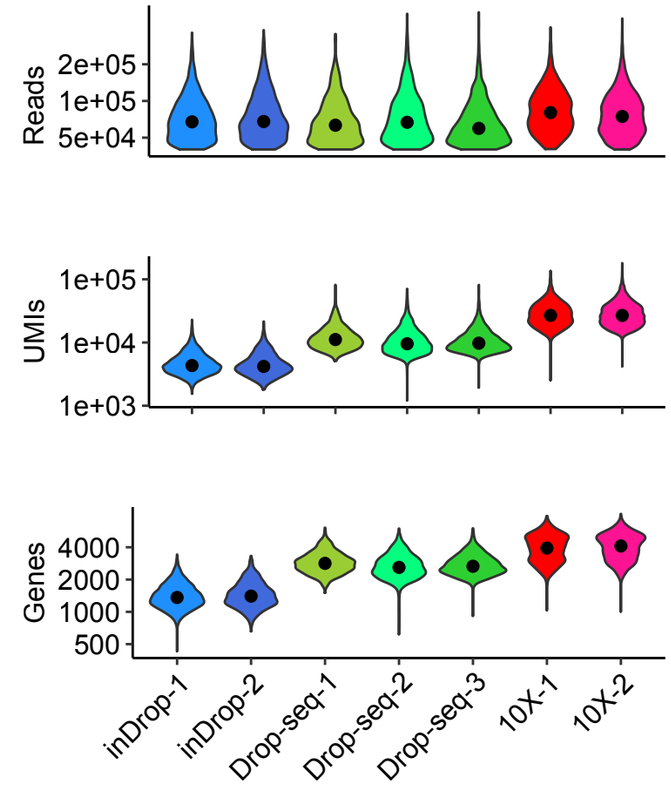

A
intergenic
intronic
coding
UTR
Percentage
100
75
50
25
0
inDrop-1
inDrop-2
Drop-seq-1
Drop-seq-2
Drop-seq-3
10X-1
10X-2
B
Genes
8000
6000
4000
2000
0
0
500
1000
1500
2000
Barcode Rank
C
Density
4
3
2
1
0
0
25
50
75
100
Normalized Base Position (%)
D
threshold
UMIs
1e+05
1e+03
1e+01
0
500
1000
1500
2000
Barcode Rank
E
threshold
Density
15
10
5
0
0.25
0.50
0.75
1
Nearest correlation
inDrop-1
inDrop-2
Drop-seq-1
Drop-seq-2
Drop-seq-3
10X-1
10X-2

Figure 4

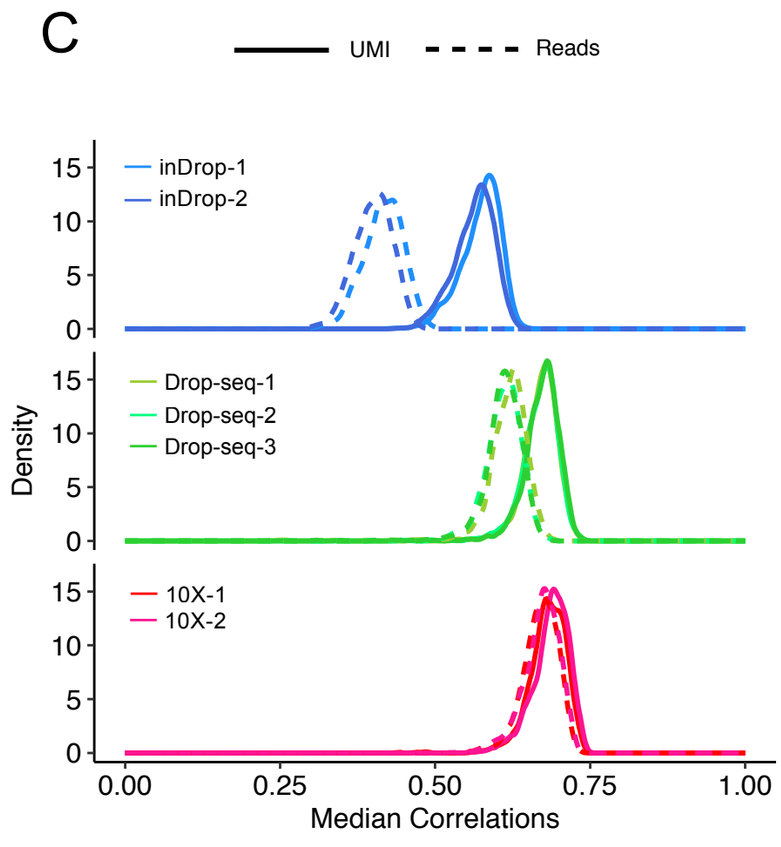
A

Sample	CBs	Reads	Genes	UMIs
inDrop-1	1,665	67,400	1,361	4,322
inDrop-2	1,328	67,810	1,400	4,183
Drop-seq-1	1,029	63,264	2,830	11,146
Drop-seq-2	1,301	66,844	2,595	9,553
Drop-seq-3	1,228	59,590	2,659	9,763
10X-1	1,395	80,215	3,930	26,838
10X-2	5,447	74,643	4,111	26,814

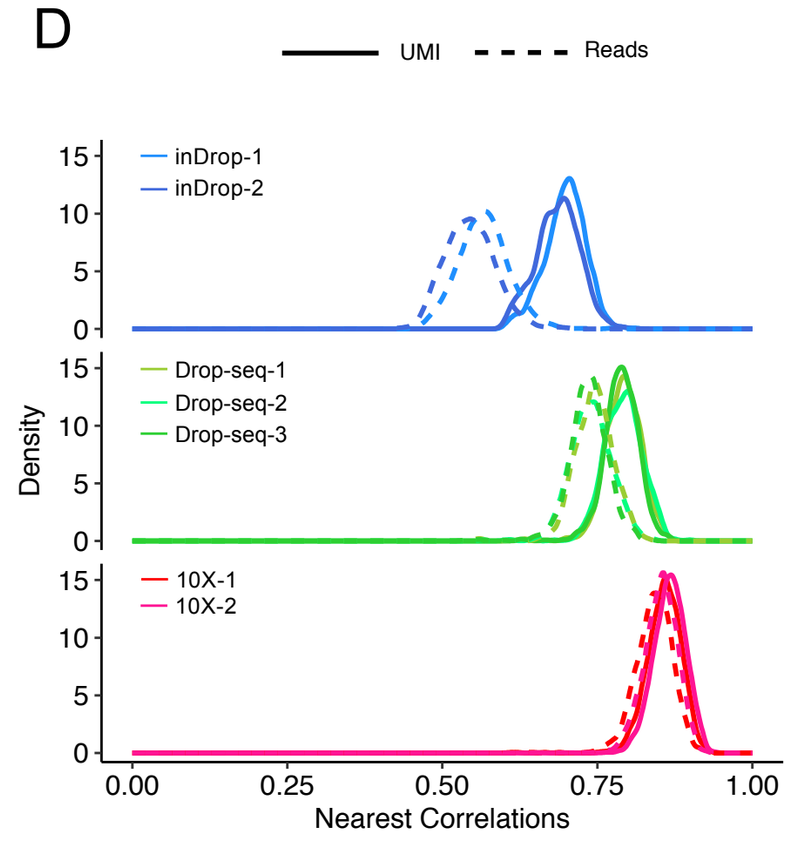
B



C



D



E

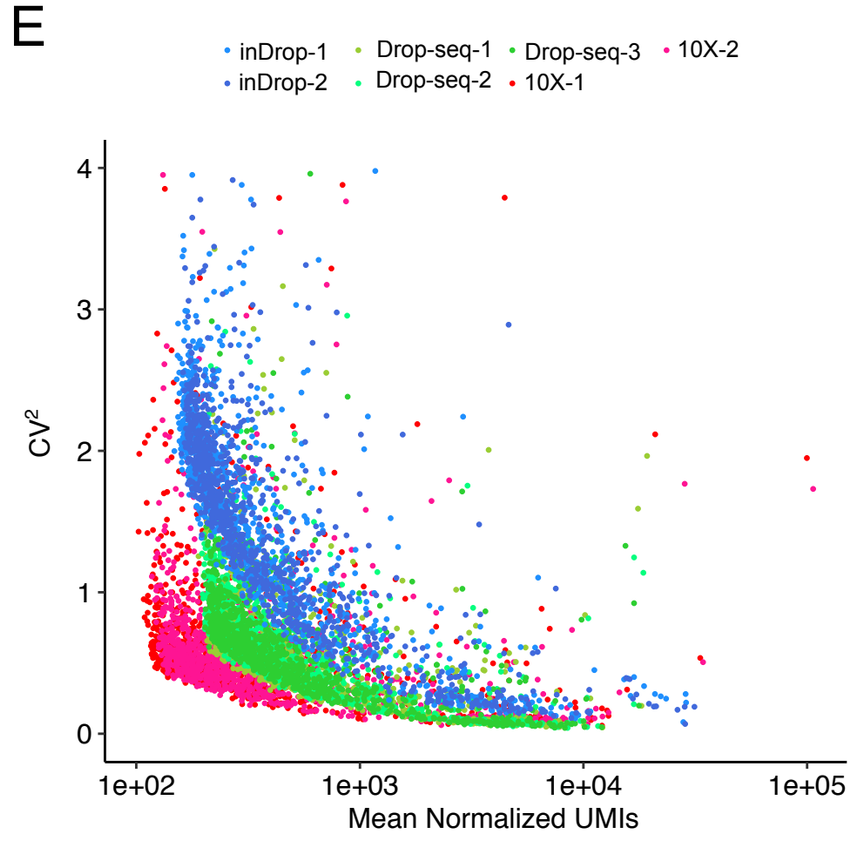


Figure 5

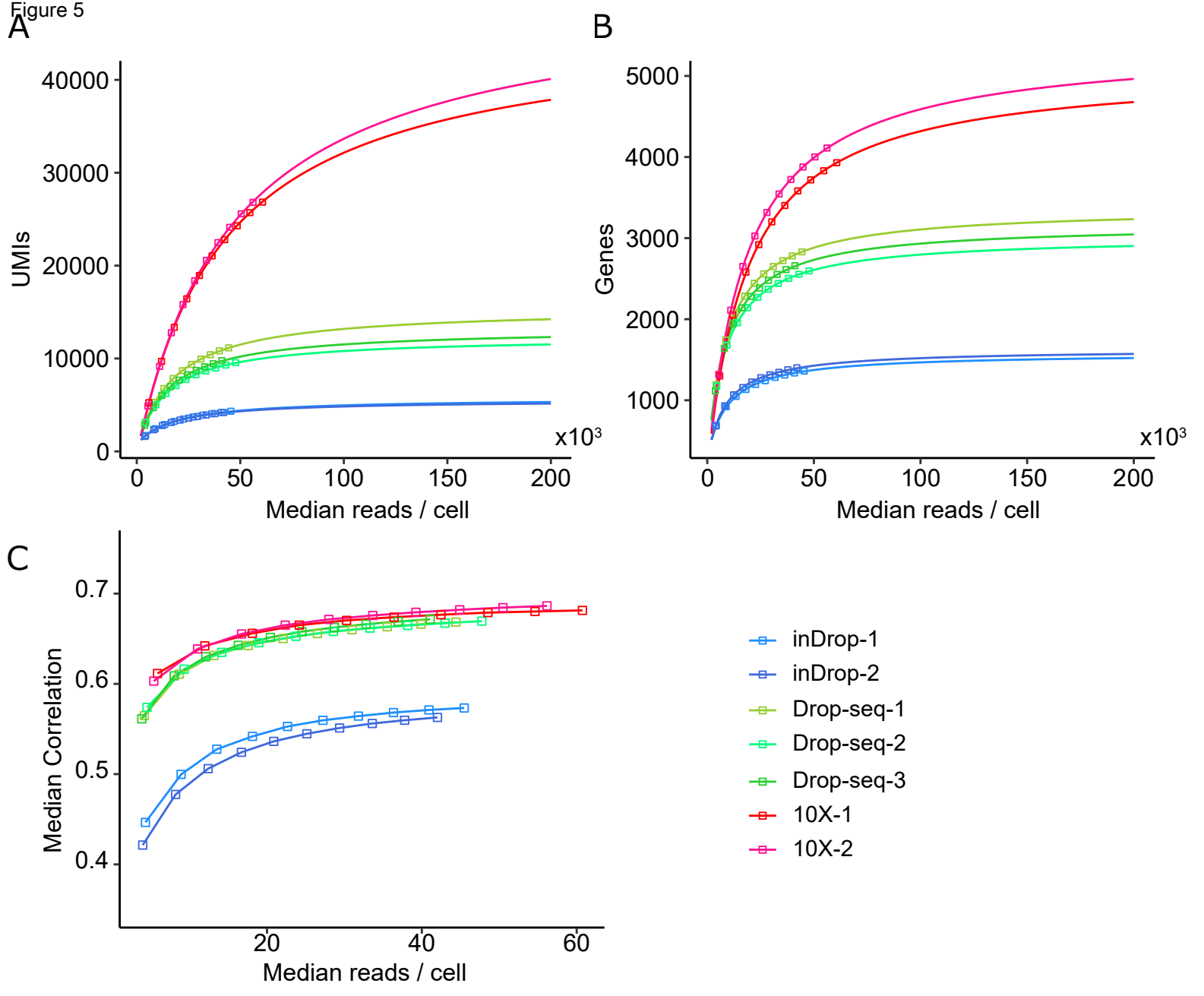

A
UMIs
40000
30000
20000
10000
0
x10³
0
50
100
150
200
Median reads / cell
B
Genes
5000
4000
3000
2000
1000
x10³
0
50
100
150
200
Median reads / cell
C
Median Correlation
0.7
0.6
0.5
0.4
20
40
60
Median reads / cell
inDrop-1
inDrop-2
Drop-seq-1
Drop-seq-2
Drop-seq-3
10X-1
10X-2

Figure 6

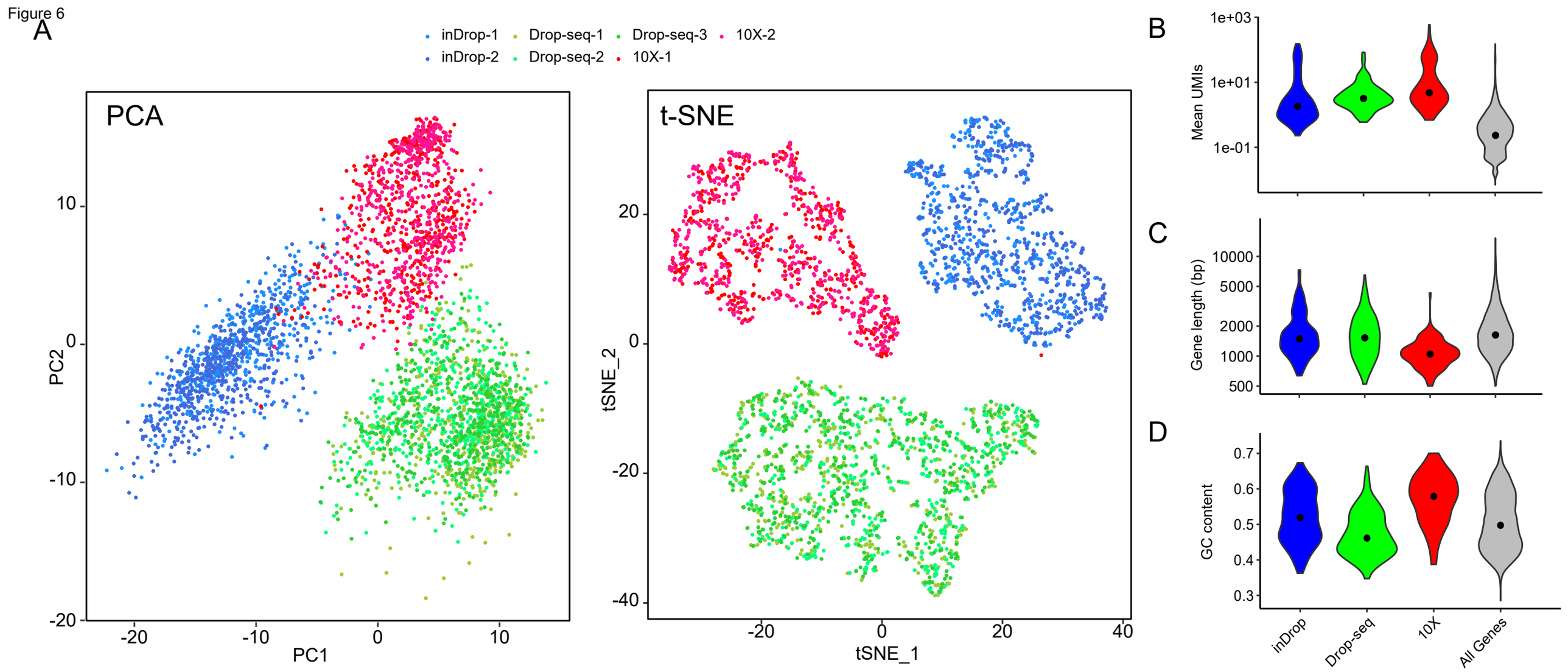
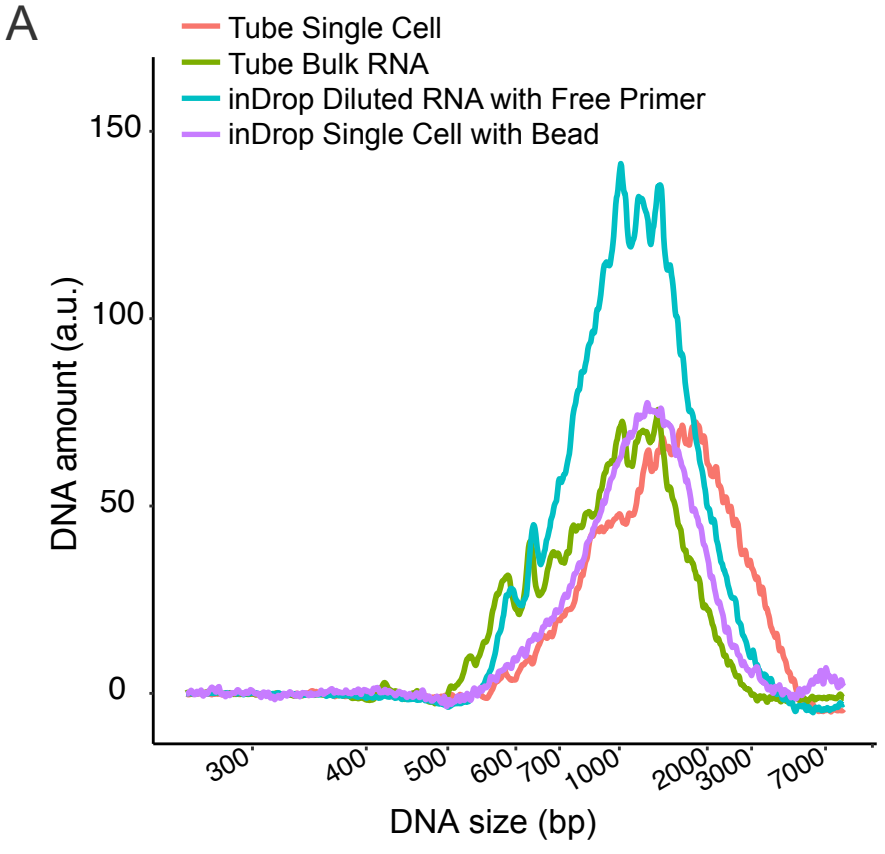
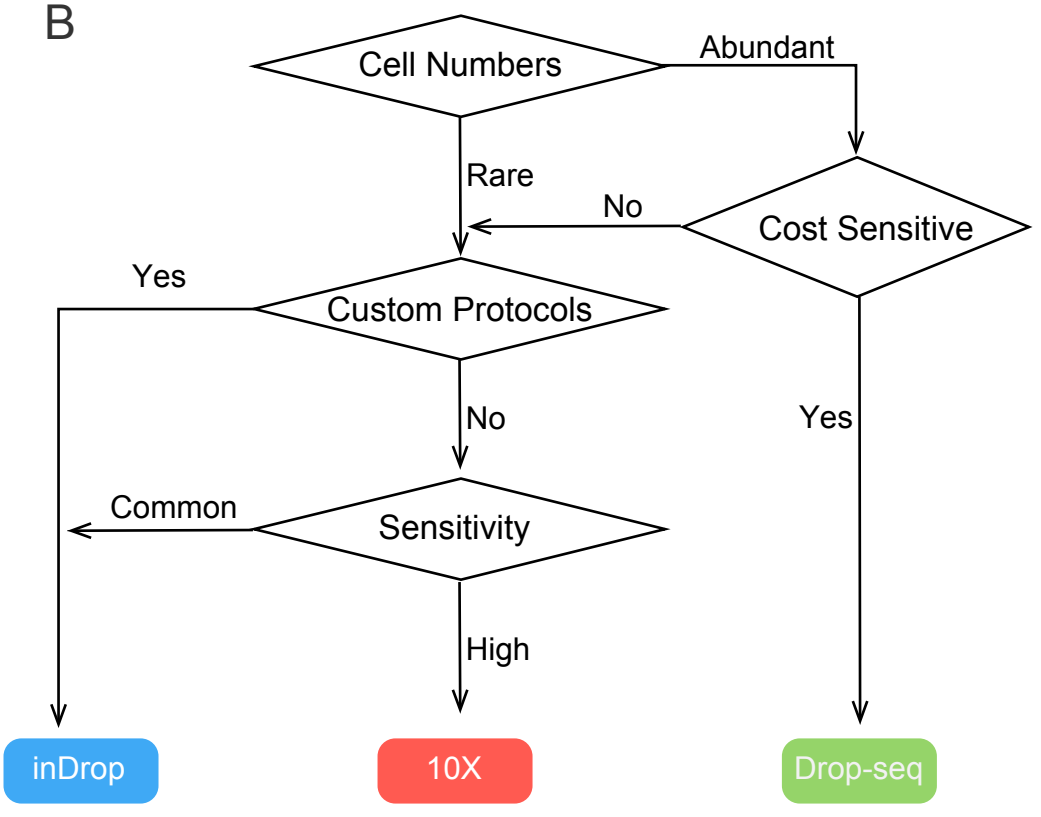

A
inDrop-1
inDrop-2
Drop-seq-1
Drop-seq-2
Drop-seq-3
10X-1
10X-2
PCA
PC1
PC2
t-SNE
tSNE_1
tSNE_2
B
Mean UMIs
1e+03
1e+01
1e-01
C
Gene length (bp)
10000
5000
2000
1000
500
D
GC content
0.7
0.6
0.5
0.4
0.3
inDrop
Drop-seq
10X
All Genes

Figure 7


A
Tube Single Cell
Tube Bulk RNA
inDrop Diluted RNA with Free Primer
inDrop Single Cell with Bead
DNA amount (a.u.)
150
100
50
0
300
400
500
600
700
1000
2000
3000
7000
DNA size (bp)


B
Cell Numbers
Abundant
Rare
No
Cost Sensitive
Yes
Custom Protocols
No
Yes
Common
Sensitivity
High
inDrop
10X
Drop-seq