A statistical simulator scDesign for rational scRNA-seq experimental design

Wei Vivian Li [1], Jingyi Jessica Li [1,2,*]

Abstract

Single-cell RNA-sequencing (scRNA-seq) has revolutionized biological sciences by revealing genome-wide gene expression levels within an individual cell. However, a critical challenge faced by researchers is how to optimize the choices of sequencing platforms, sequencing depths, and cell numbers in designing scRNA-seq experiments, so as to balance the exploration of the depth and breadth of transcriptome information. Here we present a flexible and robust simulator, scDesign, the first statistical framework for researchers to quantitatively assess practical scRNA-seq experimental design in the context of differential gene expression analysis. In addition to experimental design, scDesign also assists computational method development by generating high-quality synthetic scRNA-seq datasets under customized experimental settings. In an evaluation based on 17 cell types and six different protocols, scDesign outperformed four state-of-the-art scRNA-seq simulation methods and led to rational experimental design.

1 Introduction

The emergence and rapid development of single-cell RNA sequencing (scRNA-seq) technologies offer unprecedented opportunities for investigating transcriptional mechanisms underlying biological and medical phenomena at the individual-cell resolution [1, 2, 3]. While bulk RNA sequencing has been widely used to capture the average transcriptome information in a batch of cells [4], scRNA-seq allows the investigation of transcriptome variation across from thousands to millions of cells. The scRNA-seq technologies have enabled researchers to investigate fundamental biomedical questions such as cellular composition of various tissues and cell types [5, 6], cell differentiation trajectories [7, 8], and spatial and temporal dynamics of single cells [9, 10]. Important discoveries have been made from scRNA-seq data and advanced our understanding of diseases such as neurological disorders [11, 12] and tumorigenesis [13, 14].

Since the first scRNA-seq study was published in 2009 [15], more than twenty scRNA-seq experimental protocols have been developed [16, 17, 18, 19, 20, 21]. An effective scRNA-seq

[1] Department of Statistics, University of California, Los Angeles, CA 90095-1554
[2] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[*] To whom correspondence should be addressed. Email: jli@stat.ucla.edu

experimental design requires careful consideration of the target research question as well as the experimental budget, and a typical design in practice consists of two steps. First, researchers need to select a proper protocol among the available ones, and the primary consideration is the choice between a tag-based protocol that allows the integration of unique molecular identifiers (UMIs) [22] and a full-length protocol that captures full-length transcripts and allows the addition of the External RNA Control Consortium (ERCC) spike-ins [21, 23]. The tag-based protocols (e.g., Drop-seq [18] and inDrop [17]) are designed to obtain a broad but shallow view of the transcriptomes across many cells, while the full-length protocols (e.g., Smart-seq2 [16] and Fluidigm C1 [24]) provide a deeper and more accurate account of the gene expression in fewer cells. Thus, the choice between the two types of protocols depends on the research question. For example, a study about gene expression dynamics during stem cell differentiation requires accurate gene expression measurements, so it should opt for a full-length protocol. In contrast, in a study aiming to identify a previously unknown cell phase during the differentiation, it is necessary to sequence a large number of cells to capture the possibly transient phase. Hence, choosing a tag-based protocol is reasonable. In the second step, to optimize an experiment with a selected protocol and a fixed budget, researchers again need to choose between exploring the depth or breadth of transcriptome information, which sums up to determining the appropriate number of cells to sequence [25, 26, 27, 28].

However, in contrast to the classical experimental design [29] guided by certain theoretical optimality (e.g., the maximum power of a statistical test), the scRNA-seq experimental design is impeded by various sources of data noises, making a reasonable theoretical analysis tremendously difficult [30, 31]. Especially, scRNA-seq data are characterized by excess zeros resulted from *dropout* events, in which a gene is expressed in a cell but its mRNA transcripts are undetected. As a result, many commonly used statistical assumptions are not directly applicable to modeling scRNA-seq data. For example, Baran-Gale et al. proposed using a negative binomial model to estimate the number of cells to sequence, so that the resulting experiment is expected to capture at least a specified number of cells from the rarest cell type [25]. However, the estimation accuracy depends on the idealized negative binomial model assumption, which real scRNA-seq data usually do not closely follow. In contrast to model-based design approaches [25, 32, 33], multiple scRNA-seq studies used descriptive statistics to provide qualitative guidance instead of well-defined optimization criteria for experimental design [34, 35, 21, 26]. However, because the various descriptive statistics were proposed from different perspectives, their resulting experimental designs are difficult to unify to guide practices. For example, one study reported that the sensitivity of most protocols saturates at approximately one million reads per cell [34], while another study found that the saturation occurs at around 4.5 million reads per cell [35]. Hence, the first study suggested to sequence more cells than the second study did. The reason for this discrepancy is that the two studies defined the sensitivity in different ways: the first study used the gene detection rate (i.e., the percentage of genes detected as expressed), while the second study used the minimum number of input RNA molecules required for confidently detecting a spike-in

control [36].

In this paper, we propose a statistical simulator scDesign for optimizing scRNA-seq experimental design from the perspective of detecting differentially expressed (DE) genes between two biological conditions (determined before an experiment) or two cell states (inferred after an experiment), a major scRNA-seq data analysis task. Given a pre-defined significance level (e.g., a false discovery rate or a p-value threshold), the power of an scRNA-seq experiment for detecting DE genes is jointly determined by the sensitivity of detecting gene expression, the accuracy of measuring gene expression, and the number of cells sequenced for each cell state [35, 34]. For each protocol and a specified total sequencing depth (i.e., the total number of reads in an scRNA-seq experiment), the cell-wise sequencing depth (i.e., the expected number of reads per cell) decreases as the cell number increases [2]. However, existing power analysis methods for scRNA-seq experiments unrealistically assume a fixed cell-wise sequencing depth, which does not change as the cell number varies [34, 37]. Therefore, the practical scRNA-seq experimental design calls a new approach that accounts for various characteristics and constraints of a real scRNA-seq experiment.

ScDesign is a simulation-based experimental design framework that has several unique advantages. First, scDesign is protocol- and data-adaptive. It learns scRNA-seq data characteristics from rapidly accumulating public scRNA-seq data generated under diverse settings. For example, 622 series of scRNA-seq datasets are currently available in the Gene Expression Omnibus (GEO) database [38]. There are also newly developed scRNA-seq databases such as SCPortalen (70 studies with $67,146$ cells) [39], scRNASeqDB (36 studies with $8,910$ cells) [40], and the Single Cell Portal (43 studies with $496,366$ cells). Second, scDesign generates synthetic data that well mimic real scRNA-seq data under the same experimental settings, providing a basis for using its synthetic data to guide practical scRNA-seq experimental design. Third, scDesign is flexible in accommodating user-specific analysis needs. Its synthetic data have the same format as real scRNA-seq data, and users can use scDesign to evaluate the performance of downstream analysis, such as gene differential expression and cell clustering, under various experimental settings at no experimental cost. Assisted by the evaluation results, users will be able to design an scRNA-seq experiment based on the setting leading to the best performance according to their specified criteria.

2 Results

2.1 The statistical framework of scDesign

We develop scDesign based on a realistic statistical generative framework that utilizes both existing real scRNA-seq data and reasonable assumptions mimicking various experimental processes. In contrast to the existing simulation methods for scRNA-seq data [41, 42, 43, 37], scDesign has a unique advantage in its use of a mixture model to account for dropout events. This is

motivated by the successful applications of our previously developed imputation method, scImpute, for recovering dropout gene expression values in scRNA-seq data [44] (see Methods). This mixture model allows scDesign to overcome the dropout hurdle in learning the key gene expression characteristics from real scRNA-seq data, so that scDesign generates synthetic data highly similar to real data in multiple aspects. Depending on whether the task is to design an scRNA-seq experiment to sequence one or two batches of cells, scDesign has the corresponding one-state mode (Figure 1a, Methods) as well as the two-state mode (Figure 1b, Methods). In the one-state mode, scDesign leverages the information in a real scRNA-seq dataset from one biological condition (e.g., treatment or control) or one cell state (e.g., T cells) to generate a single scRNA-seq dataset given an experimental setting, i.e., a pre-specified total sequencing depth and a cell number. Specifically, scDesign first estimates five parameters from the real scRNA-seq dataset, including two cell-wise and three gene-wise parameters, which jointly define the key characteristics of scRNA-seq data. Second, scDesign simulates ideal gene expression levels for new cells of the same biological condition or cell state based on the estimated gene expression mean and variance parameters. Third, scDesign introduces dropout values based on the estimated gene-wise and cell-wise dropout parameters to mimic the actual dropout events in an scRNA-seq experiment. Fourth, scDesign outputs a synthetic gene expression matrix with entries as read counts. In the two-state mode, scDesign leverages the information in two real scRNA-seq datasets from different biological conditions or cell states to generate two scRNA-seq datasets given an experimental setting. In other words, the simulation by scDesign mimics an experiment where two groups of cells from two biological conditions or cell states are sequenced together. Similar to the one-state mode, scDesign independently simulates ideal gene expression levels for new cells of the two cell states, introduces dropout values based on the estimated dropout parameters of each state, and generate observed gene read counts by accounting for the fact that RNA molecules from the two batches of cells compete for the total sequencing depth. Finally, scDesign outputs two gene expression count matrices, one for each condition or state. It is worth noting that the scDesign framework is directly generalizable to more than two biological conditions or cell states.

2.2 scDesign captures key characteristics of scRNA-seq data

We first demonstrate that scDesign accurately captures six key characteristics of real scRNA-seq data, so it serves as a reliable data simulator to assist scRNA-seq experimental design and to benchmark computational methods. To assess the simulation performance of scDesign as compared with four other simulation methods, splat, powsimR, Lun, and scDD, we compared the simulated data generated by each method with the real data from various protocols and settings. Both splat and powsimR are tools specifically designed for simulating single-cell RNA-seq data [43, 37]; Lun denotes the simulation design introduced by Lun et al. [41]; scDD denotes the simulation method designed to evaluate the differential expression method scDD [42]. We considered six experimental protocols, Smart-seq2 [16], Drop-seq [18], 10x Genomics [19], Fluidigm C1 (SMARTer) [24], inDrop [17], and Seq-Well [20], and we collected three real scRNA-seq gene

read count matrices of distinct cell types from each protocol (Table S1). In summary, we used 18 real count matrices of 17 cell types from two species (human and mouse) to evaluate the five simulation methods.

We applied scDesign and the other four simulation methods to each real count matrix to estimate gene expression parameters and simulate a new count matrix with the same matrix dimensions (see Methods). We note that scDesign is the only method that considers the total sequencing depth, i.e., the total read count in the real count matrix. We compared each pair of real and simulated count matrices in terms of six summary statistics, including four gene-wise statistics (the count mean, the count variance, the count coefficient of variation (cv), and the gene-wise zero fraction) and two cell-wise statistics (the library size and the cell-wise zero fraction) (see Methods). If a simulation method is able to mimic real scRNA-seq experiments, each of the six statistics should have similar empirical distributions in the simulated and the corresponding real data. Based on this evaluation criterion, our results show that scDesign well mimics real scRNA-seq experiments based on all the six experimental protocols, even though those protocols generate data with distinct properties. For example, data from Smart-seq2 and Fluidigm C1 have relatively larger library sizes and smaller count cvs (Figures 2, S1), while data from the other four protocols have smaller library sizes, larger count cvs, and larger gene-wise and cell-wise zero fractions (Figures 3, S2, S3, S4). The simulated data by scDesign successfully capture these characteristics. In detail, we measured the similarity between each summary statistics' empirical distributions in real and the corresponding simulated data by each simulation method, using the Kolmogorov-Smirnov (KS) distance, whose value is between 0 and 1 and a smaller value indicates greater similarity (see Methods). Comparing the KS distances of the five methods, we found that scDesign performs the best for four protocols: Smart-seq2, Fluidigm C1, Seq-Well, and inDrop (Figures 2, S1, 3, S4); scDesign and powsimR are the best two methods for 10x Genomics (Figure S2); scDesign, splat, and powsimR have comparably good performance for Drop-seq (Figure S3). In summary, scDesign is ranked the best in 83 comparisons and the second best in 24 comparisons, among the total of 108 comparisons (six statistics for each of the 18 datasets). The demonstrated advantage of scDesign is rooted in its ability to incorporate both parametric and non-parametric methods to simulate scRNA-seq gene count data. By constructing a mixture model adapted from scImpute [44], scDesign explicitly models the gene-wise parameters from the real data. When generating cell-wise parameters for the simulated new cells, scDesign uses different sampling techniques for each parameter to capture its distribution characteristic. In terms of the method stability, scDesign and Lun are the only two methods that successfully estimated parameters and simulated data for all the 18 datasets, while the other three methods had errors for a number of datasets: scDD encountered errors for seven datasets, while splat and powsimR each had errors with one dataset.

2.3 scDesign guides rational scRNA-seq experimental design

Given a fixed sequencing depth in designing an scRNA-seq experiment, scDesign assists users to predict the optimal numbers of cells for sequencing. In the context of gene differential expression analysis of two biological conditions or cell states, the cell number is optimal if its resulting scRNA-seq data lead to the most accurate detection of DE genes, where the accuracy depends on a user-specified criterion, e.g., a statistical test's power given a significance level. We consider two scenarios: (1) cells from the two biological conditions or cell states are prepared as two separate libraries and sequenced independently; (2) cells from the two biological conditions or cell states are prepared in the same library and sequenced together. For simplicity, we will refer to "biological conditions" as "cell states" in the following text. Scenario (1) includes many studies that investigated cells collected at two differentiating time points [45], cells of the same tissue type from patients and healthy subjects [46], or cells of the same type but exposed to different experimental treatments [47]. The experimental design under scenario (1) aims to select the optimal cell numbers simultaneously for two libraries, so that the subsequent DE analysis becomes the most accurate given a user-specified criterion. On the other hand, scenario (2) includes many scRNA-seq studies that sequenced an *in vivo* tissue sample, e.g., the peripheral blood mononuclear cell sample [19], which is composed of a mixture of cell subtypes [18]. In scenario (2), DE analysis is performed on a pair of known or putative cell subtypes within the sequenced sample. We consider the experimental design to optimize the DE analysis between two pre-selected cell subtypes under scenario (2).

In scenario (1), the constraints are the total sequencing depths of the two cell states, and scDesign aims to determine the optimal cell number for each cell state, among a set of candidate cell numbers. scDesign simulates a new count matrix of each state based on a real count matrix of the same state, for each pre-specified sequencing depth and cell number (see Methods). Once obtaining the simulated count matrices corresponding to various candidate cell numbers, scDesign assesses the accuracy of DE gene identification using five metrics: precision, recall, true negative rate, F1 score (the harmonic mean of precision and recall), and F2 score (the harmonic mean of true negative rate and recall) (Table S2; Methods). We applied scDesign to optimize the designs of 14 example experiments (Table S3). In every experiment, we set the sequencing depth to 100 million reads, a typical depth used in real scRNA-seq experiments. We approximated real experimental scenarios by assuming that the libraries of the two cell states have the same number of cells. We considereded eight candidate cell numbers per cell state: 64, 128, 256, 512, 1024, 2048, 4096, and 8192. The DE genes between two cell states were identified using the two-sample t test (see Methods).

Our results suggest that given a criterion in the DE analysis, the optimal cell number is jointly determined by multiple technical factors, including the experimental protocol and the unwanted variation introduced by sequencing, as well as biological factors, such as the intra- and inter- state cellular heterogeneity (Table S3). Two factors are notable. First, when cells of the same two states are sequenced, the optimal cell number varies with protocols. For example, between two subtypes

of glial cells: astrocytes and oligodendrocytes, 512 cells per state is the optimal cell number that maximizes the recall in DE analysis when Fluidigm C1 is used, but the number becomes 4096 per state when inDrop is used (Figure 4). If users choose the F1 score as the criterion, the optimal cell number per state is 128 and 1024 for Fluidigm C1 and inDrop, respectively. Interestingly, Fluidigm C1 and inDrop require vastly different cell numbers to reach the same level of accuracy in DE analysis, and inDrop generally needs more cells than Fluidigm C1. This result is reasonable, since inDrop is a tag-based protocol that is advantageous in capturing more cells but disadvantageous in measuring each cell accurately, compared with the full-length protocol Fluidigm C1. Second, under the same protocol, the optimal cell number depends on the transcriptome similarity of the two cell states. For instance, with Smart-seq2, 512 cells need to be sequenced per state to maximize the recall in identifying DE genes between two dendrocyte subtypes, but only 256 cells per state are needed when dendrocytes are compared with monocytes (Figure 5). If the goal is to maximize the F2 score, the optimal cell number for comparing the two dendrocyte subtypes remains 512 per state, but the number reduces to 128 for comparing dendrocytes with monocytes. It is worth noting that the optimal cell number for both comparisons becomes 64, the smallest candidate cell number, when the criterion is the precision or the true negative rate (Table S3). The reason is that only the genes with strong DE signals are detectable with a small sample size (cell number) in any statistical testing. Hence, with a reasonable lower bound on the cell number, the DE genes detected at a smaller cell number have a higher precision. Unlike the precision, the largest recall in DE analysis is mostly achieved at a medium to large cell number. In all the experimental designs we evaluated, the recall rate of DE genes first increases with the cell number and then decreases after reaching a peak (Figures 4 and 5). These results demonstrate the trade-off between the cell number and the cell-wise library size (i.e., cell-wise gene expression capture rate) in scRNA-seq experiments. A combination of a small cell number and a large cell-wise library size ensures the identification of the DE genes with strong DE signals (i.e., achieving a high precision rate), but the small cell number may prohibit the detection of the DE genes with small to medium DE signals (i.e., sacrificing the recall rate). On the other hand, a combination of a reasonably large cell number and a small cell-wise library size increases the recall rate in detecting DE genes but compromises the precision rate due to high dropout rates (Figure S5). We also performed the DE analysis by replacing the two-sample t test with an scRNA-seq DE method MAST [48] (Table S4). The optimal cell number remains 64 per state in all comparisons, when the criterion is the precision. The optimal cell numbers defined by the recall have small differences from the t test results (Table S3), but the scale and trend remain largely consistent.

In scenario (2), the constraint is the total sequencing depth of one experiment with at least two cell states, and the goal is to determine the optimal total cell number for that experiment given a criterion in DE analysis. scDesign simulates a new count matrix of each cell state based on a real count matrix from the same state, with pre-specified total sequencing depth, total cell number, and cell proportions of the two cell states of interest (see Methods). We applied scDesign to evaluate the designs of 12 example experiments (Table S5). In every experiment, we set the

sequencing depth to 100 million reads, and we considered six total cell numbers: 512, 1024, 2048, 4096, 8192, and $16,384$. We estimated the cell proportions of the two cell states of interest from the corresponding real data (Table S5). In practical applications of scDesign, the cell state proportions can be inferred from public data or literature [5, 49, 18, 20].

In contrast to scenario (1), the optimal total cell number in scenario (2) depends on an additional factor: the cell state proportions, aside from the technical and biological factors we have discussed. The two cell states of interest may present in various proportions depending on biological conditions and experimental protocols, and larger cell state proportions in general reduce the demand of a larger total cell number. For example, the estimated cell state proportions of astrocytes and oligodendrocytes in a human brain sample are 19.2% and 14.9%, respectively [50], and 1024 cells are needed to maximize the recall with Fluidigm C1 (Figure 6). In a mouse visual cortex sample, however, the estimated proportions of the same two cell types are 8.8% and 13.1%, respectively, and $16,384$ cells are required to achieve the highest recall with inDrop (Figure 6). Given an experimental protocol, the optimal total cell number depends on both the two cell state proportions and the magnitude of gene expression differences between the two cell states. For example, the proportions of CD4 cells, CD8 cells, and B cells in a human peripheral blood mononuclear sample are 17.2%, 10.2%, and 7.3%, respectively [20]. Two important facts about this experiment are: first, the proportion of CD8 cells is higher than the proportion of B cells; second, the magnitude of gene expression differences is larger between CD4 and B cells than between CD4 and CD8 cells. With Seq-Well as the experimental protocol, the DE analysis of CD4 vs. B cells only needs $4,096$ and $8,192$ cells to achieve the highest F1 and F2 scores, respectively. On the other hand, the DE analysis of CD4 vs. CD8 requires $16,384$ cells to maximize either the F1 score or the F2 score (Figure 7). To further assess the effects of cell state proportions on DE analysis, we synthesized CD4 and B cells with multiple hypothetical cell proportions: $10\%, 20\%, 30\%,$ and 40% (Figure S6), among which the mixture of 40% B cells and $20-30\%$ CD4 cells led to the minimum cell number required to maximize the recall and precision. It is worth noting that we did not allow the proportions of B cells and CD4 cells add up to 100%, because in real experiments that sequence *in vivo* tissue samples, it is almost impossible to only sequence the two cell states of interest. Determining the optimal cell state proportions given a total cell number is especially useful when the cell states of interest can be enriched by fluorescence-activated cell sorting [47] or flow cytometry [51] before the sequencing step [31].

2.4 scDesign assists scRNA-seq method development

In addition to assisting single-cell experimental design, scDesign can also simulate scRNA-seq data to evaluate and benchmark various computational methods for differential gene expression analysis, single cell clustering analysis, gene expression dimension reduction, etc. Due to excess zeros resulting from dropout events and the fact that each gene's expression level in each cell is only measured once, the ground truth of individual genes' expression levels in individual cells cannot be accurately estimated from scRNA-seq data. Also, cellular identities

of individual cells are difficult to pre-determine in most experiments, and they often need to be inferred from sequencing data afterwards. Lacking the aforementioned ground truth encumbers the development of computational methods to decipher information from scRNA-seq data. Direct evaluation of computational methods relies on experimental validation, which is often unavailable for computationalists, and indirect biological interpretation from downstream analysis is used instead as a not-so-ideal substitute. Empowered by its ability to generate synthetic scRNA-seq data that well mimic real scRNA-seq data and have ground truth information, scDesign provides a flexible framework to benchmark computational methods for various scRNA-seq data analysis tasks.

We first demonstrated the application of scDesign to evaluating and comparing DE methods. We considered a baseline DE method, i.e., the two-sample t test, and four DE methods (MAST [48], SCDE [52], DESeq2 [53], and edgeR [54]) specifically designed for scRNA-seq data. Here both DESeq2 and edgeR denote their single-cell-adapted versions, where gene expression values are weighted by the weights estimated from a zero inflated negative Binomial model before the statistical testing step [55]. We evaluated scDesign using real scRNA-seq data of six cell types: dendrocytes (Smart-seq2, 63.6% zero count), oligodendrocytes (Fluidigm C1, 62.9% zero count), interneurons (inDrop, 75.3% zero counts), retinal ganglions (Drop-Seq, 78.3% zero counts), enterocytes (10x Genomics, 82.0% zero counts), and natural killer cells (Seq-Well, 88.0% zero counts) (Table S1). Based on the real data of each cell type, we simulated a pair of count matrices, with one matrix containing the original gene expression levels and the other including up-regulated and down-regulated genes (each type of DE genes have a pre-specified percentage). In the first setting, we set the percentage to 5% and sampled the fold changes of those DE genes' expression values uniformly from the interval $[2, 5]$ (see Methods). Then we evaluated the performance of the five DE methods by comparing the areas under their precision-recall curves (Figure 8). With Smart-seq2 and Fluidigm C1, MAST and SCDE were the only two methods that achieved better accuracy than the two-sample t test, but overall the three methods had comparable precision and recall. With inDrop and 10x Genomics, edgeR became the best DE method, followed by MAST and SCDE. With Drop-seq and Seq-Well, the most accurate method was SCDE, and the baseline two-sample t test had poor performance. These simulation results suggest that scRNA-seq data from the 10x Genomics, inDrop, Drop-seq, and Seq-Well protocols need more specialized statistical modeling in the DE analysis, compared with Smart-seq2 and Fluidigm C1. In the second setting, we set the percentage of up-regulated and down-regulated genes in each comparison to 10% and sampled the fold changes of these DE genes uniformly from the interval $[4, 5]$ (see Methods). Due to the increased magnitude of fold changes, the DE methods overall demonstrated improved accuracy (Figure S7), but the relative accuracy of the five DE methods was consistent with that under the first setting.

We next demonstrated the application of scDesign to comparing dimension reduction methods. We considered four dimension reduction methods: principal component analysis (PCA), t-distributed stochastic neighbor embedding (tSNE) [56], independent component analysis (ICA)

[57], and ZINB-WaVE [58]. We evaluated scDesign using the same real scRNA-seq data of the six cell types (each with a different protocol) used in our last demonstration for comparing DE methods. Based on the real data of each cell type, we simulated four synthetic count matrices, representing four cell states following a differentiation path. We first simulated the cell state at the starting point of differentiation based on the real data, and then we simulated each of the three subsequent cell states with a pre-specified percentage of up-regulated and down-regulated genes from its previous state. In the first setting, we set the percentage to 5% and sampled the fold changes of those DE genes' expression values uniformly from $[2, 5]$ (see Methods), a range sufficient for all the four dimension reduction methods to distinguish the four cell states in the first two dimensions, for all the six scRNA-seq protocols (Figure 9). Among the four dimension reduction methods, PCA, tSNE, and ICA had the tendency to divide cells from the same state into two disjoint clusters with the Drop-seq and 10x data, while ZINB-WaVE resulted in four clear clusters of the four cell states with all the six protocols. In the second setting, we set the percentage of up-regulated and down-regulated genes to 3% (Figure S8) and sampled the fold changes of those DE genes' expression values uniformly from $[1.5, 2]$ (see Methods). Since the differentiation effect was reduced from the first setting, tSNE did not separate the four cell states well with the 10x data, and ICA failed to distinguish the four states with the Fluidigm C1 and 10x data. The above results demonstrate the capacity of scDesign in helping developers evaluate competing computational methods for the same purpose (e.g., DE analysis or dimension reduction), and in assisting users to select the appropriate method for analyzing scRNA-seq data from a specific protocol.

3 Discussion

The scRNA-seq technologies have become an essential tool for studying various biological and biomedical problems, but one unresolved challenge is how to balance the trade-off between exploring the depth or breadth of transcriptome information in experimental design. We introduce scDesign, the first statistical and computational simulator that enables rational and practical scRNA-seq experimental design. By integrating statistical assumptions and real scRNA-seq datasets from public repositories into its generative framework, scDesign is able to mimic the real experimental processes and simulate synthetic scRNA-seq datasets that well capture gene expression characteristics in real data. In addition, scDesign is a flexible and reproducible simulator that is capable of modeling protocol-specific scRNA-seq data generated under multiple biological and experimental conditions. We conducted a comprehensive comparison of scDesign and four other scRNA-seq simulation methods (splat, powsimR, Lun, and scDD) based on datasets from 17 different cell types and six experimental protocols. The comparison suggests that scDesign generates synthetic data with the largest resemblance to real scRNA-seq data regardless of cell types and protocols.

Using its simulated data, scDesign performs power analysis on differential gene expression analysis to provide a quantitative and objective standard for designing future experiments. In the

context of differential gene expression analysis between two cell states, scDesign suggests an optimal cell number given a fixed sequencing depth, in the trade-off between a deeper sequencing of a smaller number of cells or a shallower sequencing of a larger number of cells. Specifically, we demonstrated the use of scDesign in two scenarios, where cells from the two states are sequenced as two separate libraries or as one pooled library. We evaluated the experimental designs for 14 and 12 scRNA-seq studies under the two scenarios, respectively. Our results for the first time demonstrate how the optimal experimental design depends on the scRNA-seq protocol and the intra and inter cell state transcriptome heterogeneity. In addition, our results revealed a general phenomenon that a deeper sequencing of a smaller number of cells leads to a higher precision in DE analysis. In contrast to the precision, maximizing the recall of DE analysis requires finding a balance between the cell-wise sequencing depth and the cell number, because our results show that the recall first increases and then decreases as we increase the cell number with the total sequencing depth fixed. scDesign enables researchers to design effective scRNA-seq experiments without pre-experimental costs in an objective manner, for example, guided by the expected power in downstream DE analysis.

Aside from enhancing future experimental design, another main contribution of scDesign is to assist computational method development for scRNA-seq. Since large-scale benchmark data are not yet available in the field, computationalists typically rely on scRNA-seq datasets from public repositories to test and evaluate new methods and algorithms. However, quality control and normalization of real data are themselves ongoing research questions, making the evaluation results in many method papers not comparable nor reproducible [59, 34]. To tackle this challenge, scDesign allows users to generate synthetic scRNA-seq datasets with user-specified experimental protocols, sequencing depths, cell states, cell numbers, as well as pre-specified differentially expressed genes. Given that scDesign generates synthetic data with known truth and well mimicking real data, users can leverage its synthetic data to comprehensively evaluate computational and statistical methods in a flexible, reproducible, and comparable way. For example, we compared five DE methods (the two-sample t test, MAST, SCDE, DESeq2, and edgeR) and four dimension reduction methods (PCA, tSNE, ICA, and ZINB-WaVE) using synthetic data generated by scDesign. Those comparison results provide useful guidance for researchers to select the most appropriate computational method to analyze real data.

We expect scDesign to assist scRNA-seq experimental design for a vast array of currently available experimental protocols. scDesign incorporates real scRNA-seq data that are publically available into its statistical framework to make flexible decisions based on the protocol and cell states used in the target study. To extend scDesign's ability to evaluate experimental designs for cell states whose scRNA-seq data are not yet publicly available, a future direction is to incorporate bulk RNA-seq data of the same type as a surrogate and estimate gene expression parameters from the bulk data. Otherwise, pilot experiments need to be conducted to collect data for experimental design [60]. Another future extension of scDesign is to find the optimal experimental design in the context of other types of downstream analyses besides the differential

gene expression analysis, such as the detection of novel cell sub-types or the recovery of temporal transcriptome trajectories [28]. We expect scDesign to be an effective bioinformatic tool that assists rational scRNA-seq experiment design based on specific research goals and benchmarks competing scRNA-seq computational methods.

4 Methods

scDesign for scRNA-seq data simulation

In this section, we describe how scDesign generates simulated RNA-seq data given existing real scRNA-seq data from a certain cell state. These simulated count matrices capture the characteristics of real count matrices, and they thus can be used to assist the development of computational methods and evaluate the performance of those methods under user-specified settings.

Simulating a single count matrix

Given a real single-cell count matrix with I genes and J_0 cells, the goal of this subsection is to generate a new count matrix with I genes and J cells, under the constraint that the new matrix has a total of S reads (Figure 1a). Both J and S are user-specified parameters. This resembles the real scenario where both the cell number and the total read number (i.e., the total sequencing depth) need to be pre-determined before an scRNA-seq experiment.

1. Estimate parameters from real scRNA-seq data.

   Denote the real single-cell count matrix by $\boldsymbol{X}^{\text{real}}$, whose I rows and J_0 columns represent the genes and cells, respectively. About the two cell-wise parameters, for each cell j we estimated its library size as

   $$\hat{s}_{0j} = \sum_{i=1}^{I} X_{ij}^{\text{real}},\ j = 1, \ldots, J_0\,,$$

   and its cell-wise dropout rate as

   $$\hat{q}_{0j} = \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}\{X_{ij}^{\text{real}} = 0\},\ j = 1, \ldots, J_0\,.$$

   Then we fit the cell library sizes $\hat{s}_{01}, \ldots, \hat{s}_{0J_0}$ using a Normal distribution, and the estimated mean and standard deviation are denoted as $\hat{\mu}_s$ and $\hat{\sigma}_s$, respectively.

   To estimate the three gene-wise parameters, we first normalized the read counts by their corresponding library sizes (so that the normalized cell library sizes became 10^6) and then

performed a logarithmic transformation on the normalized values. The transformed matrix is denoted as $\boldsymbol{X}^{\text{log}}$, where

$$X_{ij}^{\text{log}} = \log_{10}\left(\frac{10^6}{\hat{s}_{0j}} X_{ij}^{\text{real}} + 1.01\right) .$$

Using the Gamma-Normal mixture model described in the scImpute method [44], for each gene i we estimated its gene-wise dropout rate and mean and standard deviation of its expression. The scImpute method models the expression levels of gene i as independently and identically distributed (i.i.d.) random variable $X_{i1}^{\text{log}}, \ldots, X_{iJ_0}^{\text{log}}$ following the density function

$$f_i(x) = \lambda_{0i}\,\text{Gamma}\left(x; \alpha_{0i}, \beta_{0i}\right) + (1 - \lambda_{0i})\,\text{Normal}\left(x; \mu_{0i}, \sigma_{0i}^2\right) \ , \ x \in \mathbb{R} \, ,$$

where λ_{0i} is gene i's dropout rate, α_{0i} and β_{0i} are the shape and rate parameters of the Gamma distribution, and μ_{0i} and σ_{0i} are the mean and standard deviation of the Normal distribution. The Gamma component describes the distribution of gene expression levels when dropout occurs, while the Normal component represents the distribution of actual gene expression levels. The parameters in this model can be estimated by the Expectation-Maximization (EM) algorithm and the resulting dropout rate, mean, and standard deviation estimates are denoted as $\hat{\lambda}_{0i}$, $\hat{\mu}_{0i}$, and $\hat{\sigma}_{0i}$, respectively. We then used a Gamma distribution to fit the estimated gene mean expression levels $\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}$ and denoted the estimated shape and scale parameters as $\hat{k}_0$ and $\hat{\theta}_0$.

To summarize, we estimated cell-wise and gene-wise parameters from the real count matrix. The estimated cell-wise parameters included the cell library size $\hat{s}_{0j}$ and the cell-wise dropout rate $\hat{q}_{0j}$ for each cell j, $j = 1, \ldots, J_0$; the estimated gene-wise parameters included the mean expression $\hat{\mu}_{0i}$, the standard deviation $\hat{\sigma}_{0i}$, and the gene-wise dropout rate $\hat{\lambda}_{0i}$ for each gene i, $i = 1, \ldots, I$.

2. Simulate ideal gene expression values.

   In this step, we simulated the ideal expression values independently for each gene without considering varying cell library sizes and the dropout issue. For each gene i ($i = 1, \ldots, I$), we first simulated its mean expression from the Gamma distribution: $\mu_i \sim$ Gamma$(\hat{k}_0, \hat{\theta}_0)$. Then we simulated the standard deviation of gene i by stratified sampling from the binned observations, which we processed from the real count matrix. Specifically, we divided the estimated gene mean expression values $\{\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}\}$ into B intervals, and we used $\hat{\mu}_{0(k)}$ to denote the k-th order statistic of $\{\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}\}$. Then, the first interval was $\left(-\infty, \hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)} - \hat{\mu}_{0(1)}}{B}\right]$, the b-th interval ($1 < b < B$) was $\left(\hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)} - \hat{\mu}_{0(1)}}{B}(b-1), \hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)} - \hat{\mu}_{0(1)}}{B} b\right]$, and the B-th interval was $\left(\hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)} - \hat{\mu}_{0(1)}}{B}(B-1), +\infty\right]$. We defined $\hat{z}_{0i} = b$ if $\hat{\mu}_{0i}$ belonged to the b-th bin, and similarly we defined $z_i = b$ if μ_i belonged to the b-th bin. We simulated the standard deviation σ_i of gene i by randomly sampling one value from the stratified gene standard deviations estimated from the real data (in step 1): $\sigma_i \sim$ Uniform$(\{\hat{\sigma}_{0i'}$:

$\hat{z}_{0i'} = z_i, i' = 1, \ldots, I\})$. Finally, we generated the ideal expression matrix $\boldsymbol{X}^{\text{ideal}}$, where $X_{ij}^{\text{ideal}} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_i, \sigma_i^2)$, $j = 1, \ldots, J$.

3. Introduce dropout events.

   In this step, we introduced dropout events into the synthetic count matrix, while accounting for the variability of both gene-wise and cell-wise dropout rates. The cell-wise dropout rate in a synthetic cell j was simulated as $q_j \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\{\hat{q}_{01}, \ldots, \hat{q}_{0J_0}\})$, $j = 1, \ldots, J$. For each gene i ($i = 1, \ldots, I$), we simulated its gene-wise dropout rate λ_i by sampling one value from the stratified dropout rates estimated from the real data: $\lambda_i \sim \text{Uniform}(\{\hat{\lambda}_{0i'} : \hat{z}_{0i'} = z_i, i' = 1, \ldots, I\})$. Then, we simulated the number of dropout events of gene i: $n_i \sim \text{Binomial}(J, \lambda_i)$. In other words, gene i was affected by the dropout events in n_i cells. These n_i cells were sampled without replacement from the cell population $\{1, 2, \ldots, J\}$, with cell j being selected with probability $\frac{q_j}{\sum_{j=1}^{J} q_j}$. We denoted the sampling results by I_{ij}, with $I_{ij} = 1$ indicating that gene i is a dropout in cell j and $I_{ij} = 0$ indicating that gene i is successfully amplified in cell j, $j = 1, \ldots, J$. We performed the above simulation steps independently for gene i, $i = 1, \ldots, I$.

   Then we obtained the synthetic count matrix with dropout events $\boldsymbol{X}^{\text{drop}}$, where

$$X_{ij}^{\text{drop}} = \left[10^{X_{ij}^{\text{ideal}} \mathbb{I}\{I_{ij}=0\}} - 1.01\right],$$

   where $[x]$ means rounding x to its nearest integer. Please note that $\boldsymbol{X}^{\text{drop}}$ is on the count scale.

4. Simulate the final count matrix.

   We first simulated the library size of each synthetic cell j: $s_j \overset{\text{i.i.d.}}{\sim} \text{Normal}(\hat{\mu}_s, \hat{\sigma}_s^2)$, $j = 1, \ldots, J$, and then we calculated the expected proportion of each entry in the count matrix

$$P_{ij} = \frac{s_j X_{ij}^{\text{drop}}}{\sum_{i=1}^{I} \sum_{j=1}^{J} s_j X_{ij}^{\text{drop}}}.$$

   Finally, we obtained the final synthetic count matrix $\boldsymbol{X}^{\text{syn}}$, which is constrained by the sequencing depth S, by simulating its counts from the multinomial distribution:

$$\left(X_{11}^{\text{syn}}, \ldots, X_{1J}^{\text{syn}}, \ldots, X_{I1}^{\text{syn}}, \ldots, X_{IJ}^{\text{syn}}\right) \sim \text{Multinomial}\left(S, (P_{11}, \ldots, P_{1J}, \ldots, P_{I1}, \ldots, P_{IJ})\right).$$

Simulating multiple count matrices following a differentiation path

Given a real dataset with I genes and J_0 cells, the goal of this section is to generate G ($G \geq 2$) new count matrices, each of which has I genes, J synthetic cells, and a total of S reads. The synthetic data should represent G cell states following a specified differentiation path with known DE genes, such that these data serve as a good basis for benchmarking single-cell data analysis

and method development. When generating the G synthetic count matrices, we assume that the G cell states follow a differentiation path, with a p_{up} proportion of up-regulated genes and a p_{down} proportion of down-regulated genes from state g to state $g+1$ ($g = 1, \ldots, G-1$).

1. Estimate parameters from real scRNA-seq data.

   As described in **Simulating a single count matrix**, from the real count matrix $\boldsymbol{X}^{\text{real}}_{I \times J_0}$, we obtained the following parameter estimates: (1) the mean $\hat{\mu}_s$ and the standard deviation $\hat{\sigma}_s$ of the Normal distribution used to model the cell library sizes; (2) the cell-wise dropout rates $\hat{q}_{01}, \ldots, \hat{q}_{0J_0}$; (3) the gene-wise dropout rate $\hat{\lambda}_{0i}$, mean $\hat{\mu}_{0i}$, and standard deviation $\hat{\sigma}_{0i}$ of gene i, $i = 1, \ldots, I$. A Gamma distribution was used to fit the estimated gene mean expression $\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}$ and the estimated shape and scale parameters are denoted as $\hat{k}_0$ and $\hat{\theta}_0$. The above parameter estimates were used to simulate the expression parameters of state 1, while the parameters of state $g+1$ depended on the parameters of its previous state g.

2. Simulate gene mean expression values of the G states.

   In this step, we simulated the log-scale mean gene expression values under each cell state, without considering dropout events. We assumed that from state g to state $g+1$, the proportions of up-regulated and down-regulated genes were p_{up} and p_{down}, respectively. The fold changes of gene mean expression levels were independently and uniformly distributed within $[f_l, f_u]$.

   We used μ_i^g to denote the mean expression of gene i in cell state g. For cell state 1, we simulated μ_i^1 from the Gamma distribution: $\mu_i^1 \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\hat{k}_0, \hat{\theta}_0)$, $i = 1, \ldots, I$. Then given $\mu_1^g, \ldots, \mu_I^g$, we simulated $\mu_1^{g+1}, \ldots, \mu_I^{g+1}$ ($g = 1, \ldots, G-1$) as follows.

   (a) We simulated the number of up-regulated genes n^g_{up}, and the number of down-regulated genes n^g_{down} from a Multinomial distribution:

   $$\left(n^g_{\text{up}}, n^g_{\text{down}}, I - n^g_{\text{up}} - n^g_{\text{down}}\right) \sim \text{Multinomial}\left(I, \left(p_{\text{up}}, p_{\text{down}}, 1 - p_{\text{up}} - p_{\text{down}}\right)\right) .$$

   (b) We randomly drew the $n^g_{\text{up}} + n^g_{\text{down}}$ DE genes from the gene population $\{1, \ldots, I\}$ without replacement and denoted

   $$d_i^g = \begin{cases} 1, & \text{if gene } i \text{ is up-regulated} \\ -1, & \text{if gene } i \text{ is down-regulated} \\ 0, & \text{otherwise} \end{cases} .$$

(c) We simulated μ_I^{g+1}, the mean expression of gene i in state $g+1$, as follows:

$$\mu_i^{g+1} = \begin{cases} \mu_i^g + \log_{10} f_i^g & \text{if } d_i^g = 1 \\ \mu_i^g - \log_{10} f_i^g & \text{if } d_i^g = -1 \\ \mu_i^g, & \text{otherwise} \end{cases},$$

where $f_i^g \overset{\text{i.i.d}}{\sim} \text{Uniform}[f_l, f_u])$.

3. Simulate the count matrices.

   With the mean gene expression $\mu_1^g, \ldots, \mu_I^g$, we simulated the count matrix $\boldsymbol{X}^{\text{syn},g}$ under each state g independently following steps 2-4 in **Simulating a single count matrix**. Please note that we estimated and simulated other cell-wise and gene-wise parameters also by following **Simulating a single count matrix**. We kept the estimated parameters the same for all the cell states, and we simulated the cell-wise parameters of synthetic cells independently across all the states.

scDesign for scRNA-seq experimental design

scDesign aims to determine the best number of cells to sequence given a fixed sequencing depth (i.e., the total number of RNA-seq reads in an experiment), such that the resulting RNA-seq data are optimized for differential gene expression analysis. In this section, we denote the two real count matrices as $\boldsymbol{X}^{\text{real1}}$, with I rows representing genes and J_{01} columns representing cells, and $\boldsymbol{X}^{\text{real2}}$, with I rows representing genes and J_{02} columns representing cells. Without loss of generality, we assume that the two matrices, which represent two cell states, have the same genes listed in the same order. We introduce how to simulate a synthetic count matrix for each state with scDesign, and the procedure can be repeated with varying cell numbers to obtain synthetic data for power analysis.

Scenario (1)

Given $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$, the goal of scDesign in scenario (1) is to generate a synthetic count matrix with I genes and J_1 cells for state 1, and a synthetic count matrix with I genes and J_2 cells for state 2. We assume that the cells of the two states are sequenced independently. Cell states 1 and 2 have sequencing depths of S_1 and S_2, respectively. For each state g $(g = 1, 2)$, we followed **Simulating a single count matrix** to simulate a count matrix $\boldsymbol{X}_{I \times J_g}^{\text{syn},g}$. The only difference is in step 2, where we directly set $\mu_i^g = \hat{\mu}_{0i}^g$ and $\sigma_i^g = \hat{\sigma}_{0i}^g$, $i = 1, \ldots, I$, instead of simulating new parameters. This requirement is to ensure that the rows in the two simulated matrices still represent the same set of real genes, and the power analysis based on the simulated data is biologically meaningful.

Scenario (2)

Now we consider the case where the cells of two cell states are jointly sequenced. Suppose that the two cell states are mixed in one biological sample, and the experimental setting is that J cells in the sample are to be sequenced to generate S RNA-seq reads in total. We assume that the two cell states present in fractions of p_1 and p_2 in the sample, respectively. That is, $0 < p_1 < 1$, $0 < p_2 < 1$, and $p_1 + p_2 \leq 1$. When $p_1 + p_2 < 1$, there are more than two cell states present in the same sample. The goal of scDesign in scenario (2) is to simulate count matrices for the two selected cell states, based on a real count matrix of each state (Figure 1b).

1. Determine cell numbers.

   We denote the numbers of cells from state 1, state 2, and the remaining states as J_1, J_2, and J_r, respectively. We sampled these numbers from a Multinomial distribution:

$$(J_1, J_2, J_r) \sim \text{Multinomial}\left(J, (p_1, p_2, 1 - p_1 - p_2)\right) .$$

2. Simulate count matrices with dropout events.

   Following step 1-3 in **Simulating a single count matrix**, we simulated two count matrices $\boldsymbol{X}^{\text{drop1}}_{I\times J_1}$ and $\boldsymbol{X}^{\text{drop2}}_{I\times J_2}$ for cell states 1 and 2, respectively. The only difference was in step 2, where we directly set $\mu_i^g = \hat{\mu}_{0i}^g$ and $\sigma_i^g = \hat{\sigma}_{0i}^g$, $i = 1, \ldots, I$, to ensure that the rows in the synthetic count matrices represented the same set of real genes.

3. Simulate the final count matrices.

   We first simulated the library sizes of the cells in the two states:

$$s_j^1 \sim \text{Normal}\left(\hat{\mu}_s^1, (\hat{\sigma}_s^1)^2\right) ,\ j = 1, \ldots, J_1 ;$$
$$s_j^2 \sim \text{Normal}\left(\hat{\mu}_s^2, (\hat{\sigma}_s^2)^2\right) ,\ j = 1, \ldots, J_2 ,$$

   where $\hat{\mu}_s^1$ and $\hat{\sigma}_s^1$ are estimated from $\boldsymbol{X}^{\text{real1}}$, and $\hat{\mu}_s^2$ and $\hat{\sigma}_s^2$ are estimated from $\boldsymbol{X}^{\text{real2}}$, as described in **Simulating a single count matrix**. Then we combined the two count matrices to obtain the expected proportion matrix $\boldsymbol{P}_{I\times(J_1+J_2)}$:

$$P_{ij} = \frac{s_j^1 X_{ij}^{\text{drop1}}}{\sum_{i=1}^{I}\sum_{j'=1}^{J_1} s_{j'}^1 X_{ij'}^{\text{drop1}} + \sum_{i=1}^{I}\sum_{j''=1}^{J_2} s_{j''}^2 X_{ij''}^{\text{drop2}}},\ j = 1, \ldots, J_1 ;$$
$$P_{i(j+J_1)} = \frac{s_j^2 X_{ij}^{\text{drop2}}}{\sum_{i=1}^{I}\sum_{j'=1}^{J_1} s_{j'}^1 X_{ij'}^{\text{drop1}} + \sum_{i=1}^{I}\sum_{j''=1}^{J_2} s_{j''}^2 X_{ij''}^{\text{drop2}}},\ j = 1, \ldots, J_2 .$$

   In the expected proportion matrix $\boldsymbol{P}$, the first J_1 columns and the last J_2 columns give the expected proportions of genes in cell states 1 and 2, respectively. Since the total number of reads is S, we assume the total number of reads from the two states together is $[S(J_1 +$

$J_2)/J]$, where $[x]$ denotes the nearest integer to x. Then we simulated the final count matrix $\boldsymbol{X}^{\text{syn}}_{I\times(J_1+J_2)}$ constrained by the sequencing depth from a Multinomial distribution:

$$(\boldsymbol{X}^{\text{syn}}_{11},\ldots,\boldsymbol{X}^{\text{syn}}_{1(J_1+J_2)},\ldots,\boldsymbol{X}^{\text{syn}}_{I1},\ldots,\boldsymbol{X}^{\text{syn}}_{I(J_1+J_2)}) \sim \\ \text{Multinomial}\left(\left[\frac{S(J_1+J_2)}{J}\right],(P_{11},\ldots,P_{1(J_1+J_2)},\ldots,P_{I1},\ldots,P_{I(J_1+J_2)})\right).$$

The final count matrix of cell state 1 is $\boldsymbol{X}^{\text{syn},1}_{I\times J_1}$, where $X^{\text{syn},1}_{ij}=X^{\text{syn}}_{ij}$. The final count matrix of cell state 2 is $\boldsymbol{X}^{\text{syn},2}_{I\times J_2}$, where $X^{\text{syn},2}_{ij}=X^{\text{syn}}_{i(j+J_1)}$.

Power analysis of DE detection with scDesign

We introduce two experimental designs in **scDesign for scRNA-seq experimental design**. If the two cell states are sequenced separately, the design needs specification of the sequencing depth S and the cell numbers J_1 in state 1 and J_2 in state 2. If the two cell states are sequenced together, the design needs specification of the sequencing depth S and the total cell number J. The goal of power analysis is to determine the best choice of cell number(s) to optimize the downstream DE analysis between two cell states, given a fixed S.

Given $\boldsymbol{X}^{\text{real1}}_{I\times J_{01}}$ and $\boldsymbol{X}^{\text{real2}}_{I\times J_{02}}$ from two different cell states, for each gene i we estimated its mean expression values in the mixture model adapted from scImpute (see **Simulating a single count matrix**) as $\hat{\mu}^g_{0i}$ and $\hat{\sigma}^g_{0i}$ for state g $(g=1,2)$. Then we calculated an effect score of gene i to denote its differential expression strength:

$$h_i=\frac{|\hat{\mu}^1_{0i}-\hat{\mu}^2_{0i}|}{\hat{\sigma}^1_{0i}+\hat{\sigma}^2_{0i}}.$$

The top N genes with the largest h_i's are used as the true DE genes to be compared with the detected DE genes from the simulated data, and this gene set is denoted as A^0. We set $N=1000$ in our analysis.

Given an experimental design, we simulate B count matrices $\{\boldsymbol{X}^{\text{syn},11},\ldots,\boldsymbol{X}^{\text{syn},1B}\}$ for cell state 1, and B count matrices $\{\boldsymbol{X}^{\text{syn},21},\ldots,\boldsymbol{X}^{\text{syn},2B}\}$ for cell state 2. By performing DE analysis on $\boldsymbol{X}^{\text{syn},1b}$ and $\boldsymbol{X}^{\text{syn},2b}$, we identified a DE gene set A^b. Denoting the gene population set as Ω, we calculated five accuracy metrics: precision (a^b_1), recall (a^b_2), true negative rate (a^b_3), F1 (a^b_4), and F2 (a^b_5):

$$a^b_1=\frac{|A^0\cap A^b|}{|A^b|},\ a^b_2=\frac{|A^0\cap A^b|}{|A^0|},\ a^b_3=1-\frac{|A^b\setminus A^0|}{|\Omega\setminus A^0|}$$
$$a^b_4=2\frac{a^b_1\times a^b_2}{a^b_1+a^b_2},\ a^b_5=2\frac{a^b_2\times a^b_3}{a^b_2+a^b_3}.$$

Then we averaged each of the five metrics calculated over the B sets of data as $a_i=\frac{1}{B}\sum_{b=1}^{B}a^b_i$, $i=1,\ldots,5$. Finally, we repeated the above steps for each candidate cell number and selected the cell number that maximizes the user-specified metric among the five metrics.

In our analysis, we set $N = 1000$ and $B = 100$. The DE method used in the simulation is the two sample t test, which is applied to the non-zero gene expression values. In real data applications, users are suggested to use the DE method of their choice for the experimental design.

Comparison of different simulation methods

The splat, Lun, and scDD simulation methods were implemented using the R package `splatter` version 1.3.3.9010. The powsimR method was implemented using the R package `powsimR` version 1.1.0. The scDesign method was implemented using the R package `scDesign` version 0.0.1.

We denote a $\log 10$-transformed count matrix as $\boldsymbol{X}_{I\times J}$, with rows representing genes and columns representing cells. For gene i ($i = 1, \ldots, I$), we define its count mean $T_i^1 = \frac{1}{J}\sum_{j=1}^{J} X_{ij}$, count variance $T_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(X_{ij} - T_i^1)^2$, coefficient of variance (cv) $T_i^3 = \frac{\sqrt{T_i^2}}{T_i^1}$, and gene-wise zero proportion $T_i^4 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{I}\{X_{ij} = 0\}$. For each cell j ($j = 1, \ldots, J$), we calculated its library size $T_j^5 = \sum_{i=1}^{I} X_{ij}$ and cell-wise zero proportion $T_j^6 = \frac{1}{I}\sum_{i=1}^{I}\mathbb{I}\{X_{ij} = 0\}$. For each real $\log 10$-transformed matrix, we calculated the values of the six statistics

$$\left\{\left(T_i^{1,\text{real}}, T_i^{2,\text{real}}, T_i^{3,\text{real}}, T_i^{4,\text{real}}, T_j^{5,\text{real}}, T_j^{6,\text{real}}\right) : i = 1, \ldots, I, j = 1, \ldots, J\right\}$$

and denote the resulting empirical distribution of the k-th statistic as F^k, $k = 1, , \ldots, 6$. For each synthetic $\log 10$-transformed matrix, we also calculated the values of the six statistics

$$\left\{\left(T_i^{1,\text{syn}}, T_i^{2,\text{syn}}, T_i^{3,\text{syn}}, T_i^{4,\text{syn}}, T_j^{5,\text{syn}}, T_j^{6,\text{syn}}\right) : i = 1, \ldots, I, j = 1, \ldots, J\right\}$$

and denote the resulting empirical distribution of the k-th statistic as G^k, $k = 1, , \ldots, 6$. Finally, to evaluate the quality of the synthetic data, we calculated the Kolmogorov-Smirnov (KS) distance between F^k and G^k is calculated as

$$D(F^k, G^k) = \sup_x |F^k(x) - G^k(x)|, \; x \in \mathbb{R}\,.$$

5 Software availability

The R package `scDesign` is freely available at https://github.com/Vivianstats/scDesign.

6 Acknowledgement

This work was supported by the following grants: UCLA Dissertation Year Fellowship (to W.V.L), and National Science Foundation DMS-1613338, NIH/NIGMS R01GM120507, PhRMA Foundation Research Starter Grant in Informatics, Johnson & Johnson WiSTEM2D Award, and Sloan Research Fellowship (to J.J.L).

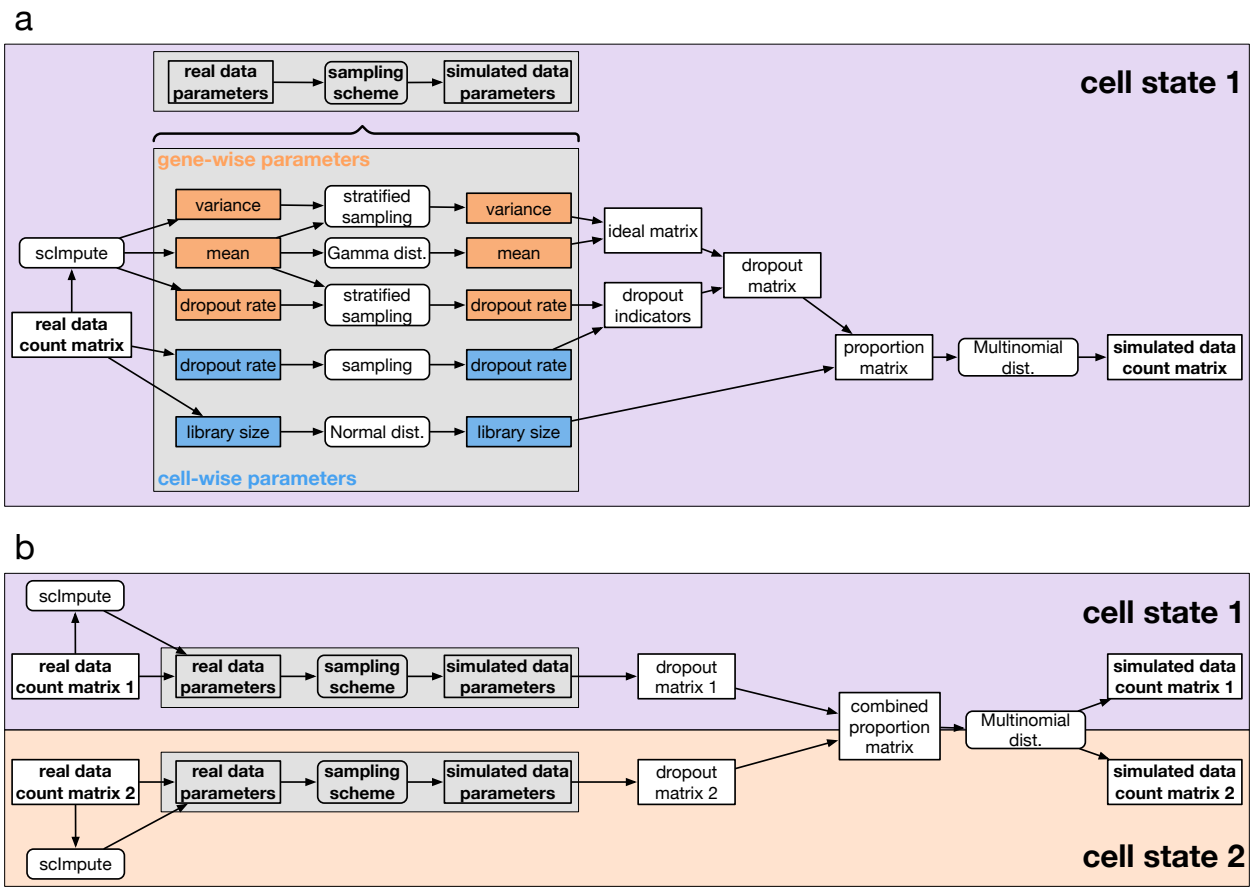


Figure 1: The statistical framework of scDesign. **a**: The simulation process of a count matrix of a single cell state (one-state mode). **b**: The joint simulation process of two count matrices of two different cell states (two-state mode).

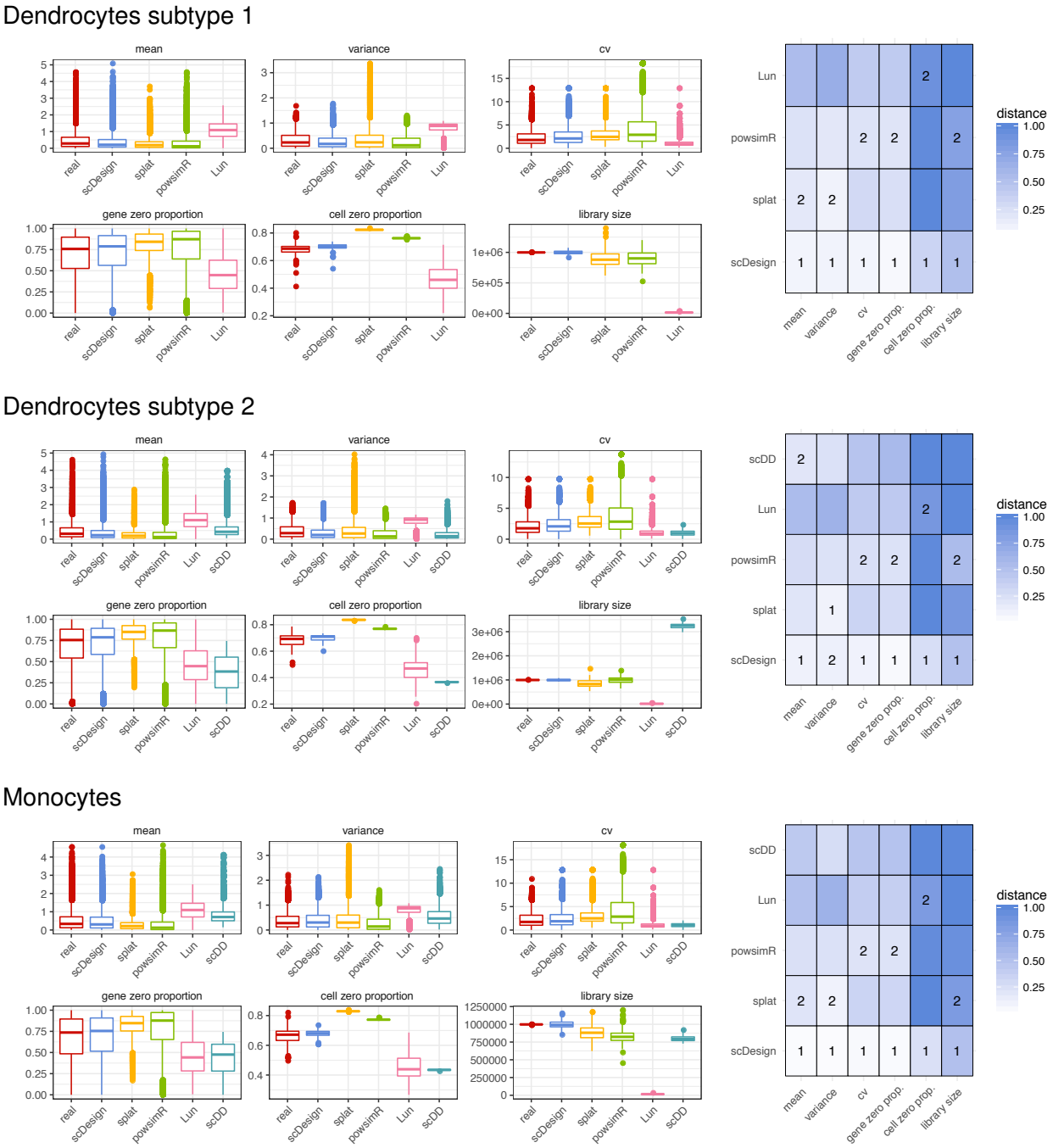


Figure 2: Comparison of scDesign and the other four simulation methods based on the Smart-seq2 protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. Note that scDD failed to simulate data for the dendrocytes subtype1 dataset.

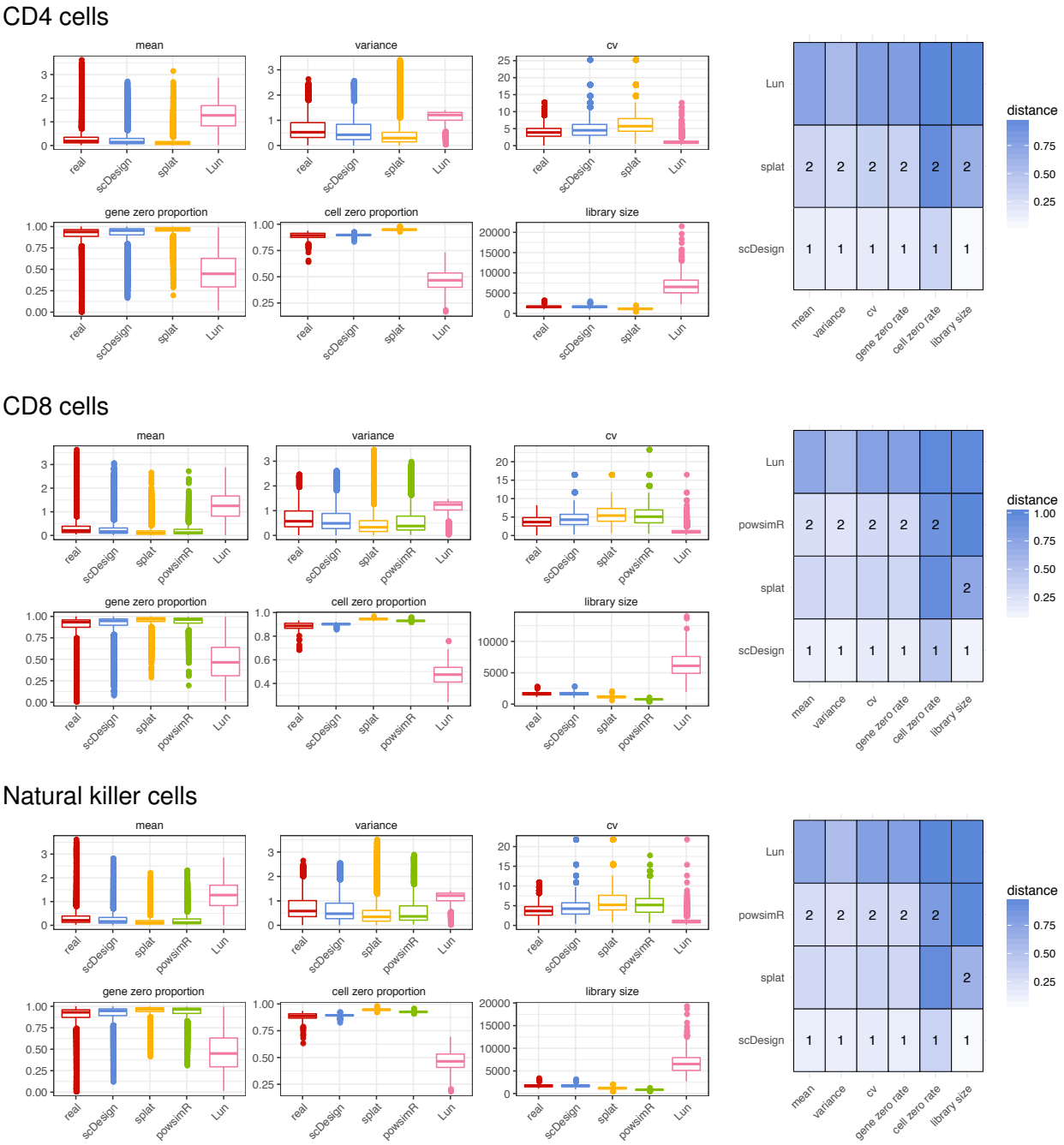
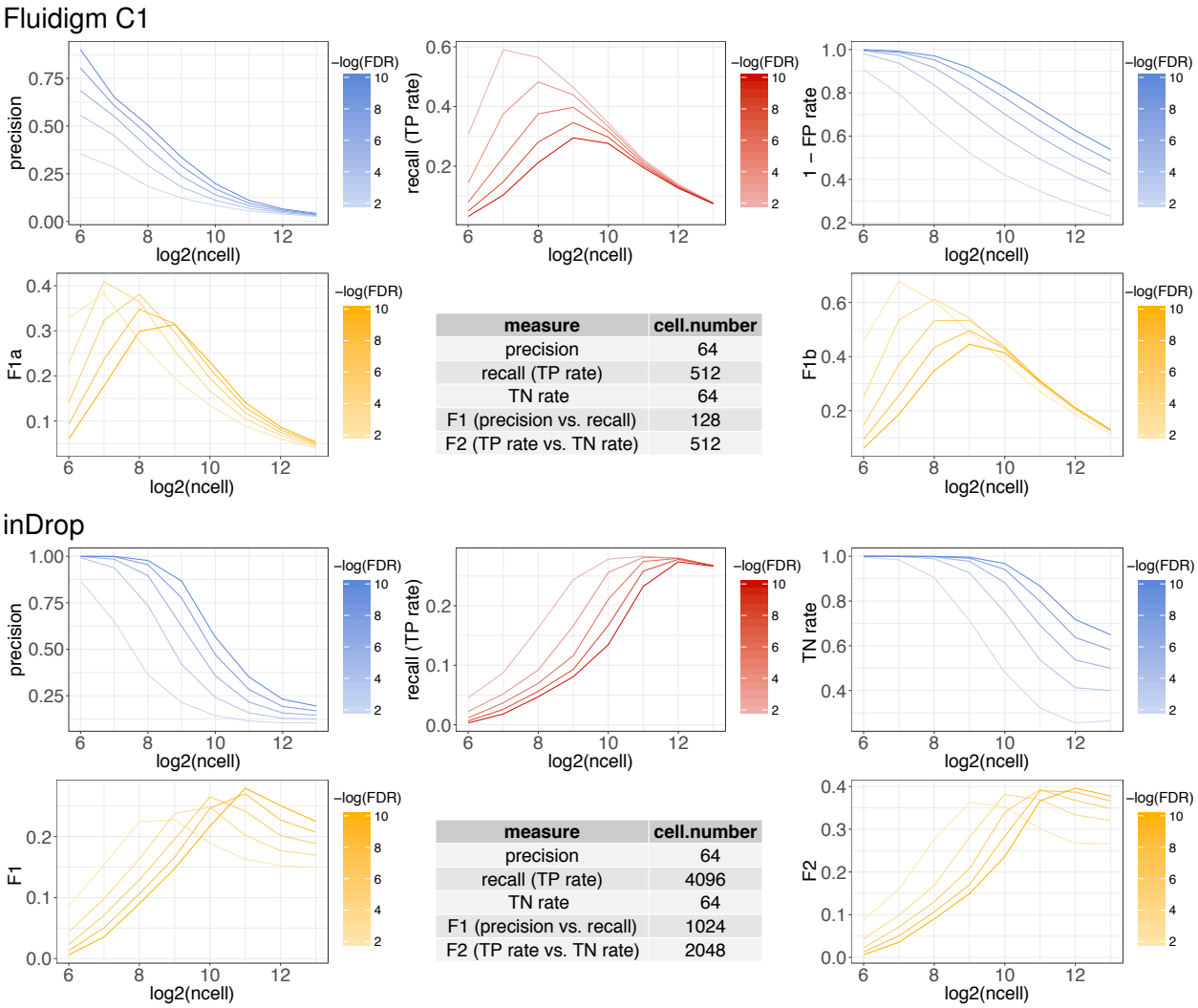


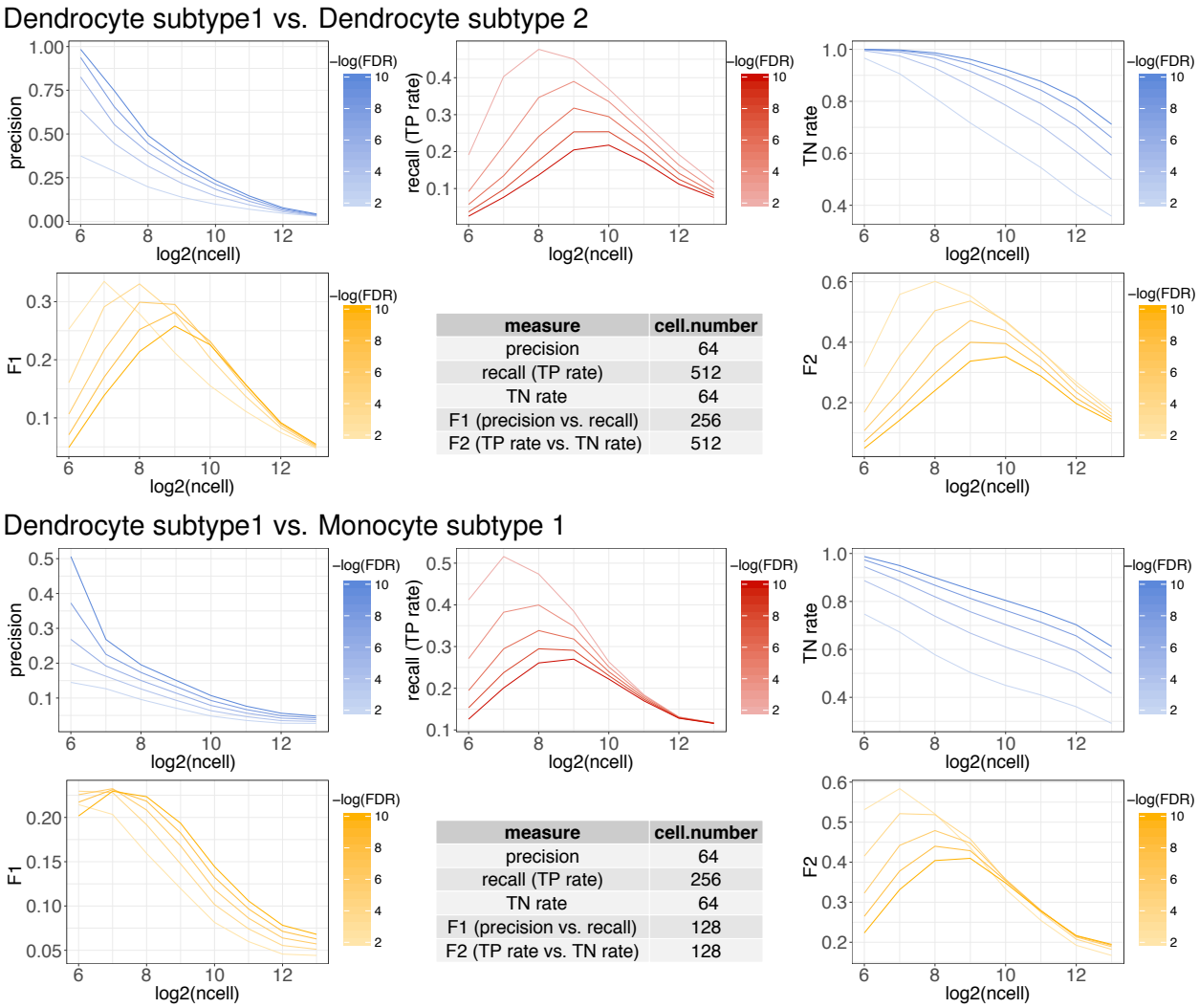
Figure 3: Comparison of scDesign and the other four simulation methods based on the Seqwell protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. Note that scDD failed to simulate data for the three datasets, and powsimR failed to simulate data for the CD4 dataset.



measure	cell.number
precision	64
recall (TP rate)	512
TN rate	64
F1 (precision vs. recall)	128
F2 (TP rate vs. TN rate)	512

measure	cell.number
precision	64
recall (TP rate)	4096
TN rate	64
F1 (precision vs. recall)	1024
F2 (TP rate vs. TN rate)	2048

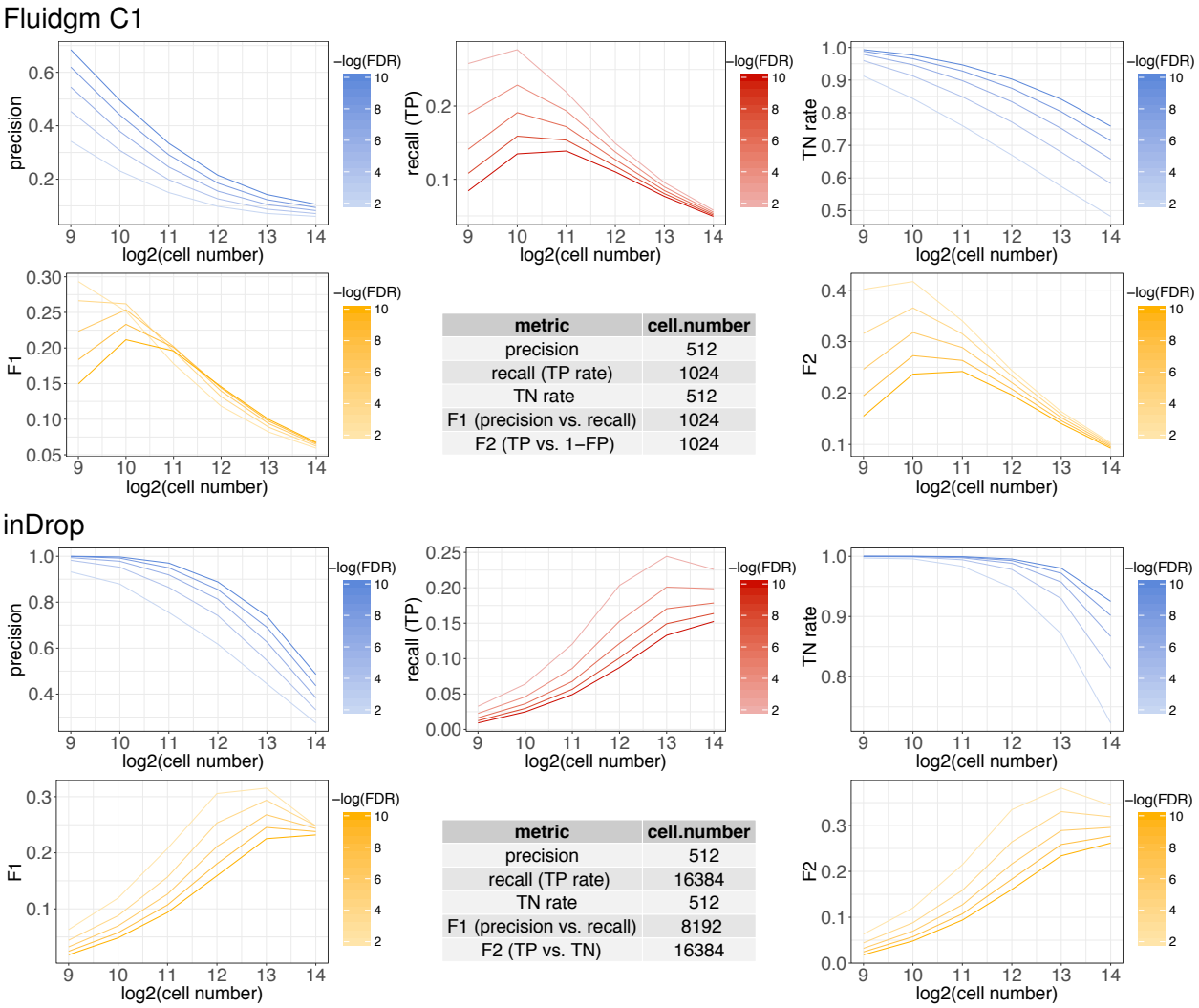
Figure 4: Power analysis for DE studies comparing astrocytes and oligodendrocytes (scenario 1). The thresholds on the false discovery rates (FDRs) (to identify DE genes) are denoted in the color legends.



measure	cell.number
precision	64
recall (TP rate)	512
TN rate	64
F1 (precision vs. recall)	256
F2 (TP rate vs. TN rate)	512

measure	cell.number
precision	64
recall (TP rate)	256
TN rate	64
F1 (precision vs. recall)	128
F2 (TP rate vs. TN rate)	128

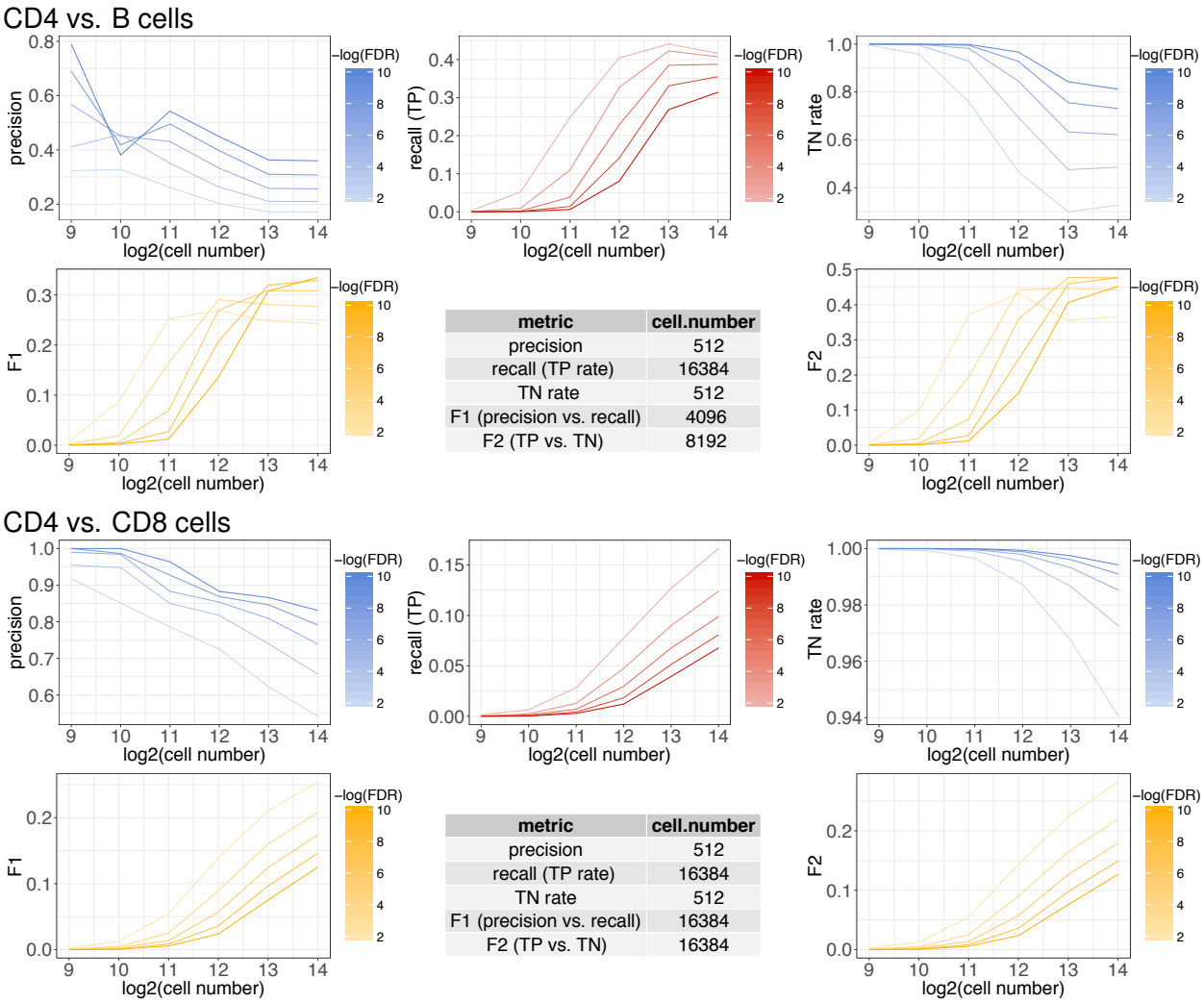
Figure 5: Power analysis for DE studies comparing dendrocytes and monocytes (scenario 1). The thresholds on the false discovery rates (FDRs) (to identify DE genes) are denoted in the color legends.



metric	cell.number
precision	512
recall (TP rate)	1024
TN rate	512
F1 (precision vs. recall)	1024
F2 (TP vs. 1−FP)	1024

metric	cell.number
precision	512
recall (TP rate)	16384
TN rate	512
F1 (precision vs. recall)	8192
F2 (TP vs. TN)	16384

Figure 6: Power analysis for DE studies comparing astrocytes and oligodendrocytes (scenario 2). The thresholds on the false discovery rates (FDRs) (to identify DE genes) are denoted in the color legends.



metric	cell.number
precision	512
recall (TP rate)	16384
TN rate	512
F1 (precision vs. recall)	4096
F2 (TP vs. TN)	8192

metric	cell.number
precision	512
recall (TP rate)	16384
TN rate	512
F1 (precision vs. recall)	16384
F2 (TP vs. TN)	16384

Figure 7: Power analysis for DE studies comparing immune cells with the Seqwell protocol (scenario 2). The thresholds on the false discovery rates (FDRs) (to identify DE genes) are denoted in the color legends.

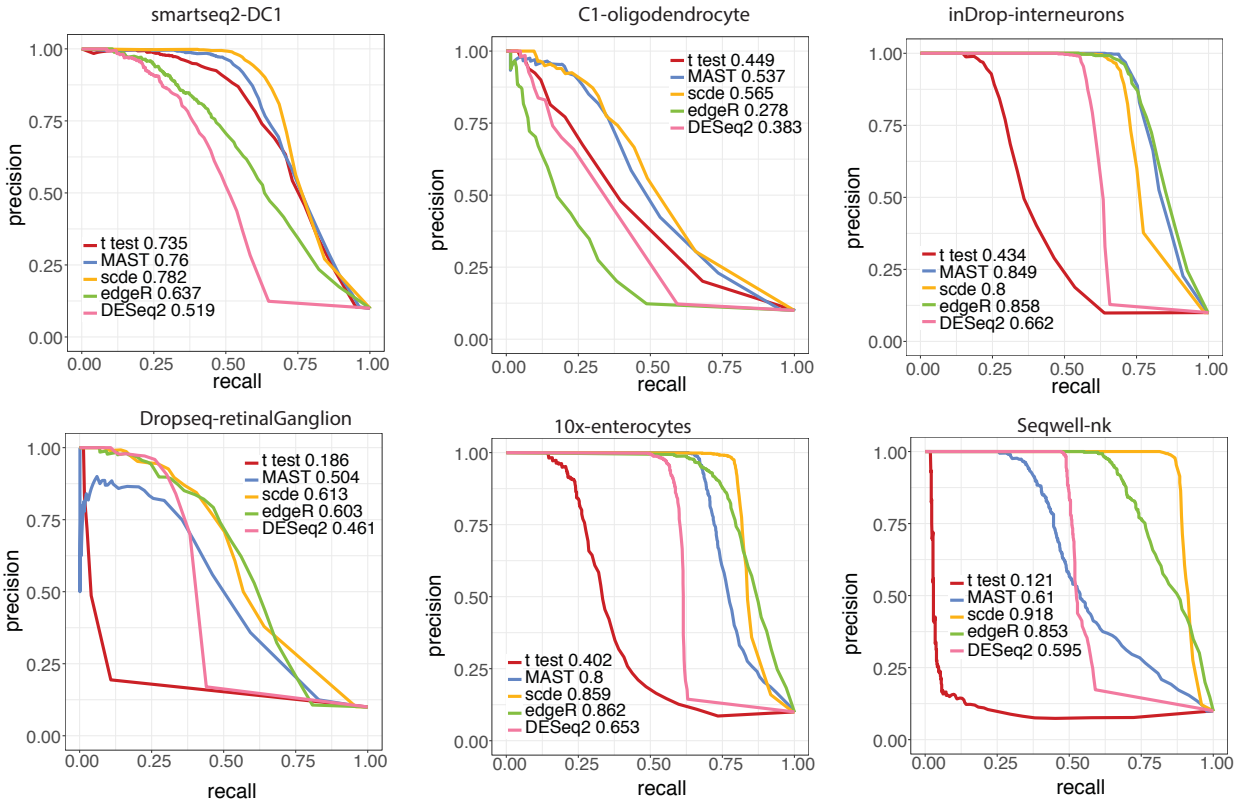


Figure 8: Comparison of scRNA-seq DE methods in the first setting. The precision-recall curves of the five DE methods are drawn for the six scRNA-seq protocols, respectively. Corresponding area under the curve (AUC) are given in the plots.

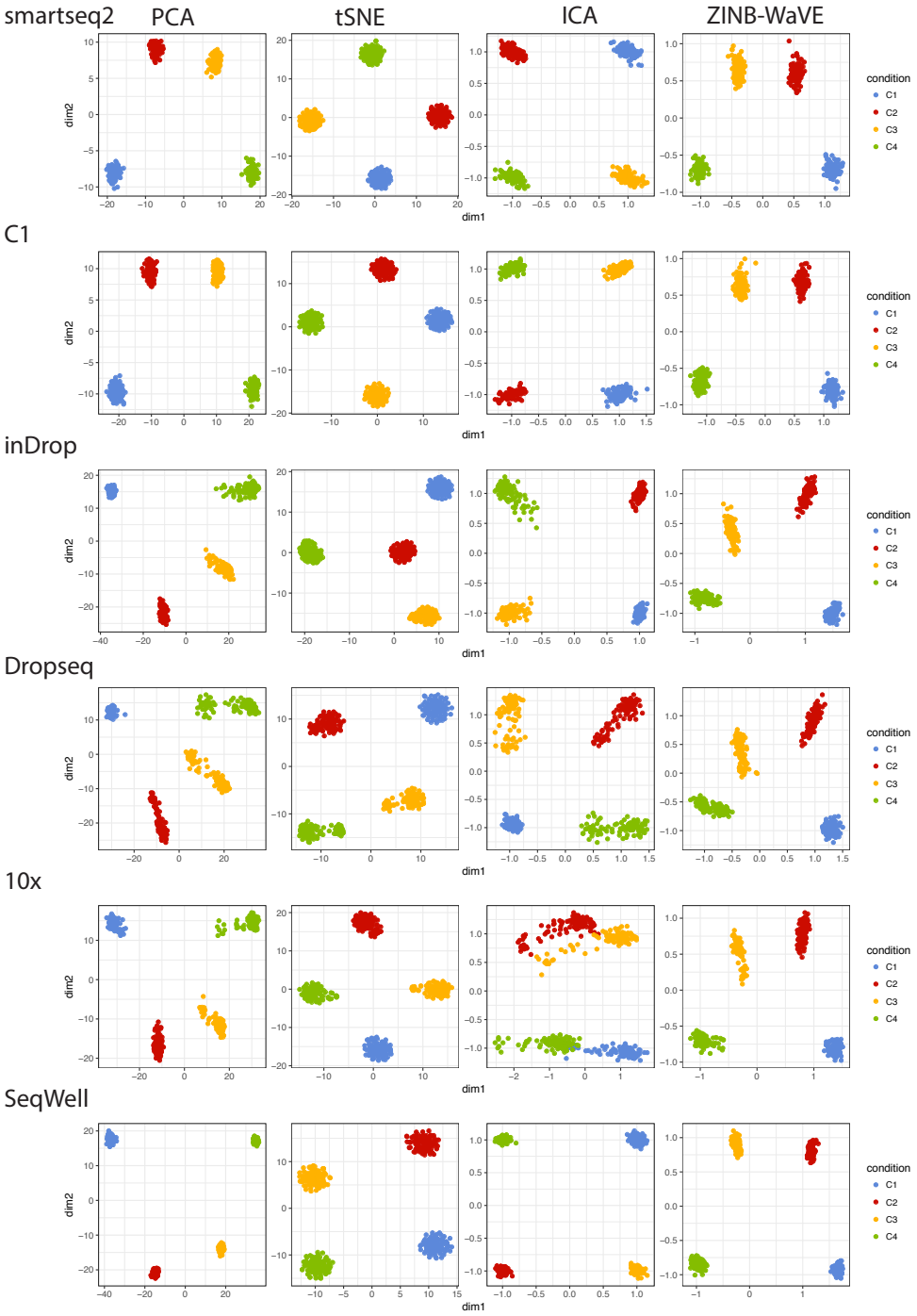


Figure 9: Comparison of scRNA-seq dimension reduction methods in the first setting. The first two dimensions resulting from the four dimension reduction methods are given for the six scRNA-seq protocols, respectively.

References

[1] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145, 2016.

[2] Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9(1):75, 2017.

[3] S Steven Potter. Single-cell rna sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology*, page 1, 2018.

[4] Wei Vivian Li and Jingyi Jessica Li. Modeling and analysis of rna-seq data: a review from a statistical perspective. *arXiv preprint arXiv:1804.06050*, 2018.

[5] Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4):593–607, 2016.

[6] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251, 2015.

[7] Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E Sun, et al. Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing. *Nature*, 500(7464):593, 2013.

[8] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015.

[9] Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C Marioni. High-throughput spatial mapping of single-cell rna-seq data to tissue of origin. *Nature biotechnology*, 33(5):503, 2015.

[10] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236, 2013.

[11] Paul W Hook, Sarah A McClymont, Gabrielle H Cannon, William D Law, A Jennifer Morton, Loyal A Goff, and Andrew S McCallion. Single-cell rna-seq of mouse dopaminergic neurons

informs candidate gene selection for sporadic parkinson disease. *The American Journal of Human Genetics*, 102(3):427–446, 2018.

[12] Nathan G Skene, Julien Bryois, Trygve E Bakken, Gerome Breen, James J Crowley, Héléna A Gaspar, Paola Giusti-Rodriguez, Rebecca D Hodge, Jeremy A Miller, Ana B Muñoz-Manchado, et al. Genetic identification of brain cell types underlying schizophrenia. *Nature genetics*, page 1, 2018.

[13] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, page 1254257, 2014.

[14] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

[15] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377, 2009.

[16] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096, 2013.

[17] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[18] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[19] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8, 2017.

[20] Todd M Gierahn, Marc H Wadsworth II, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nature methods*, 14(4):395, 2017.

[21] Dominic Grün and Alexander van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810, 2015.

[22] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72, 2012.

[23] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1):63, 2016.

[24] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053, 2014.

[25] Jeanette Baran-Gale, Tamir Chandra, and Kristina Kirschner. Experimental design for single-cell rna sequencing. *Briefings in functional genomics*, 2017.

[26] Simone Rizzetto, Auda A Eltahla, Peijie Lin, Rowena Bull, Andrew R Lloyd, Joshua WK Ho, Vanessa Venturi, and Fabio Luciani. Impact of sequencing depth and read length on single cell rna sequencing data of t cells. *Scientific Reports*, 7(1):12781, 2017.

[27] Vilas Menon. Clustering single cells: a review of approaches on high-and low-depth single-cell rna-seq data. *Briefings in functional genomics*, 2018.

[28] Bianca Dumitrascu, Karen Feng, and Barbara E Engelhardt. Gt-ts: Experimental design for maximizing cell type discovery in single-cell data. *bioRxiv*, page 386540, 2018.

[29] Gerry P Quinn and Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002.

[30] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241, 2015.

[31] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[32] How many cells do we need to sample so that we see at least n cells of each type? https://satijalab.org/howmanycells. Accessed: 2018-08-08.

[33] Bianca Dumitrascu, Karen Feng Feng, and Barbara E Engelhardt. Gt-ts: experimental design for maximizing cell type discovery in single-cell data. *bioRxiv*, page 386540, 2018.

[34] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

[35] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14(4):381, 2017.

[36] Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for rna-seq experiments. *Genome research*, 21(9):1543–1551, 2011.

[37] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: power analysis for bulk and single cell rna-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

[38] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[39] Imad Abugessaisa, Shuhei Noguchi, Michael Böttcher, Akira Hasegawa, Tsukasa Kouno, Sachi Kato, Yuhki Tada, Hiroki Ura, Kuniya Abe, Jay W Shin, et al. Scportalen: human and mouse single-cell centric database. *Nucleic acids research*, 46(D1):D781–D787, 2017.

[40] Yuan Cao, Junjie Zhu, Peilin Jia, and Zhongming Zhao. scrnaseqdb: A database for rna-seq based gene expression profiles in human single cells. *Genes*, 8(12):368, 2017.

[41] Aaron TL Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):75, 2016.

[42] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17(1):222, 2016.

[43] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[44] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.

[45] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.

[46] Max Schelker, Sonia Feau, Jinyan Du, Nav Ranu, Edda Klipp, Gavin MacBeath, Birgit Schoeberl, and Andreas Raue. Estimation of immune cell content in tumour tissue using single-cell rna-seq data. *Nature communications*, 8(1):2032, 2017.

[47] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[48] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):278, 2015.

[49] Ziyi Chen, Anfei Huang, Jiya Sun, Taijiao Jiang, F Xiao-Feng Qin, and Aiping Wu. Inference of immune cell composition on the expression profiles of mouse tissue. *Scientific reports*, 7:40508, 2017.

[50] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

[51] A Yu Yen-Rei, Emily G OKoren, Danielle F Hotten, Matthew J Kan, David Kopin, Erik R Nelson, Loretta Que, and Michael D Gunn. A protocol for the comprehensive flow cytometric analysis of immune cells in normal and inflamed murine non-lymphoid tissues. *PloS one*, 11(3):e0150606, 2016.

[52] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[53] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[54] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[55] Koen Van den Berge, Charlotte Soneson, Michael I Love, Mark D Robinson, and Lieven Clement. zinger: unlocking rna-seq tools for zero-inflation and single cell applications. *bioRxiv*, page 157982, 2017.

[56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[57] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[58] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):284, 2018.

[59] Davis J McCarthy, Kieran R Campbell, Aaron TL Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, 2017.

[60] Aniruddha Chatterjee, Antonio Ahn, Euan J Rodger, Peter A Stockwell, and Michael R Eccles. A guide for designing and analyzing rna-seq data. In *Gene Expression Analysis*, pages 35–80. Springer, 2018.

Supplementary Materials

Table S1: Details of real scRNA-seq data. Cell type information is from the original work that published the corresponding data.

data	accession	species	cell type	protocol
1	GSE94820	human	dendrocyte1 (165) dendrocyte2 (94) monocyte2 (163)	Smart-Seq2
2	GSM1626793	mouse	bipolar (919) cones (241) rods (3746) retinal ganglion (70)	Drop-seq
3	GSE92332	mouse	goblet (510) stem (1267) tuft (166)	10x
4	GSE67835	human	astrocyte (49) neuron (122) oligodendrocyte (38)	Fluidigm C1
5	GSE102827	mouse	excitatory (1040) neuron (116) oligodendrocyte (286) astrocyte (189)	inDrop
6	GSM2486333	human	natural killer (471) CD4 (634) CD8 (269) B cell (376)	Seq-Well

Table S2: The five metrics used to evaluate accuracy of DE analysis.

metric	abbreviation	meaning
recall (true positive rate)	recall (TP)	ability of a method to identify all true DE genes
precision	precision	ability of a method to identify only true DE genes
1-false positive rate	1-FP	
harmonic mean of recall and precision	F1	combined ability of a method to identify all and only true DE genes
harmonic mean of TP and 1-FP	F2	

Table S3: The optimal cell numbers in experimental design (scenario 1) using t test. Five DE accuracy measures are calculated in every comparison: precision, recall, TN (true negative rate), F1, and F2. For each measure, the smallest cell number that leads to the highest accuracy is recorded as the optimal number.

protocol	cell type 1	cell type 2	precision	recall	TN	F1	F2
Smart-Seq2	dendrocyte1	monocyte1	64	256	64	128	128
Smart-Seq2	dendrocyte1	dendrocyte2	64	512	64	256	512
Drop-seq	cone	retinal ganglion	64	1024	64	512	512
Drop-seq	cone	rod	64	2048	64	1024	512
10x	tuft	goblet	64	2048	64	1024	4096
10x	tuft	stem	64	4096	64	2048	4096
C1	neuron	astrocyte	64	512	64	128	512
C1	neuron	oligodendrocyte	64	512	64	128	512
C1	astrocyte	oligodendrocyte	64	512	64	128	512
inDrop	astrocyte	oligodendrocyte	64	4096	64	1024	2048
inDrop	excitatory	interneuron	64	4096	64	2048	4096
inDrop	excitatory	oligodendrocyte	64	1024	64	128	512
Seq-Well	CD4	B cell	64	2048	64	512	512
Seq-Well	CD4	CD8	64	8192	64	8192	8192

Table S4: The optimal cell numbers in experimental design (scenario 1) using MAST. Five DE accuracy measures are calculated in every comparison: precision, recall, TN (true negative rate), F1, and F2. For each measure, the smallest cell number that leads to the highest accuracy is recorded as the optimal number.

protocol	cell type 1	cell type 2	precision	recall	TN	F1	F2
Smart-Seq2	dendrocyte1	monocyte1	64	128	64	64	128
Smart-Seq2	dendrocyte1	dendrocyte2	64	256	64	128	256
Drop-seq	cone	retinal ganglion	64	1024	64	512	1024
Drop-seq	cone	rod	64	4096	64	4096	4096
10x	tuft	goblet	64	1024	64	512	1024
10x	tuft	stem	64	2048	64	512	2048
C1	neuron	astrocyte	64	512	64	128	512
C1	neuron	oligodendrocyte	64	512	64	256	512
C1	astrocyte	oligodendrocyte	64	256	8192	256	256
inDrop	astrocyte	oligodendrocyte	64	2048	64	1024	2048
inDrop	excitatory	interneuron	64	1024	64	256	1024
inDrop	excitatory	oligodendrocyte	64	2048	64	1024	2048
Seq-Well	CD4	B cell	64	2048	64	1024	2048
Seq-Well	CD4	CD8	64	4096	64	4096	4096

Table S5: The optimal cell numbers in experimental design (scenario 2) using t tests. Five DE accuracy measures are calculated in every comparison: precision, recall, TN (true negative rate), F1, and F2. For each measure, the smallest cell number that leads to the highest accuracy is recorded as the optimal number. The proportions of each cell type in the cell population are listed under the cell labels.

protocol	cell type 1	cell type 2	precision	recall	TN	F1	F2
Smart-Seq2	dendrocyte1 16.1%	monocyte1 15.8%	512	1024	512	512	1024
Smart-Seq2	dendrocyte1 16.1%	dendrocyte2 9.1%	512	2048	512	1024	2048
Drop-seq	cone 0.42%	retinal ganglion 0.1%	8192	16384	16384	16384	16384
Drop-seq	cone 0.42%	rod 65.6%	2048	16384	2048	16384	16384
10x	tuft 2.3%	goblet 7.1%	512	16384	512	16384	16384
10x	tuft 2.3%	Stem 17.6%	1024	16384	1024	16384	16384
C1	neuron 47.8%	astrocyte 19.2%	512	512	512	512	512
C1	astrocyte 19.2%	oligodendrocyte 14.9%	512	1024	512	1024	1024
inDrop	astrocyte 8.8%	oligodendrocyte 13.1%	512	16384	512	8192	16384
inDrop	excitatory 47.8%	interneuron 5.3%	512	16384	512	8192	16384
Seq-Well	CD4 17.2%	B cell 7.3%	512	16384	512	4096	8192
Seq-Well	CD4 17.2%	CD8 10.2%	512	16384	512	16384	16384

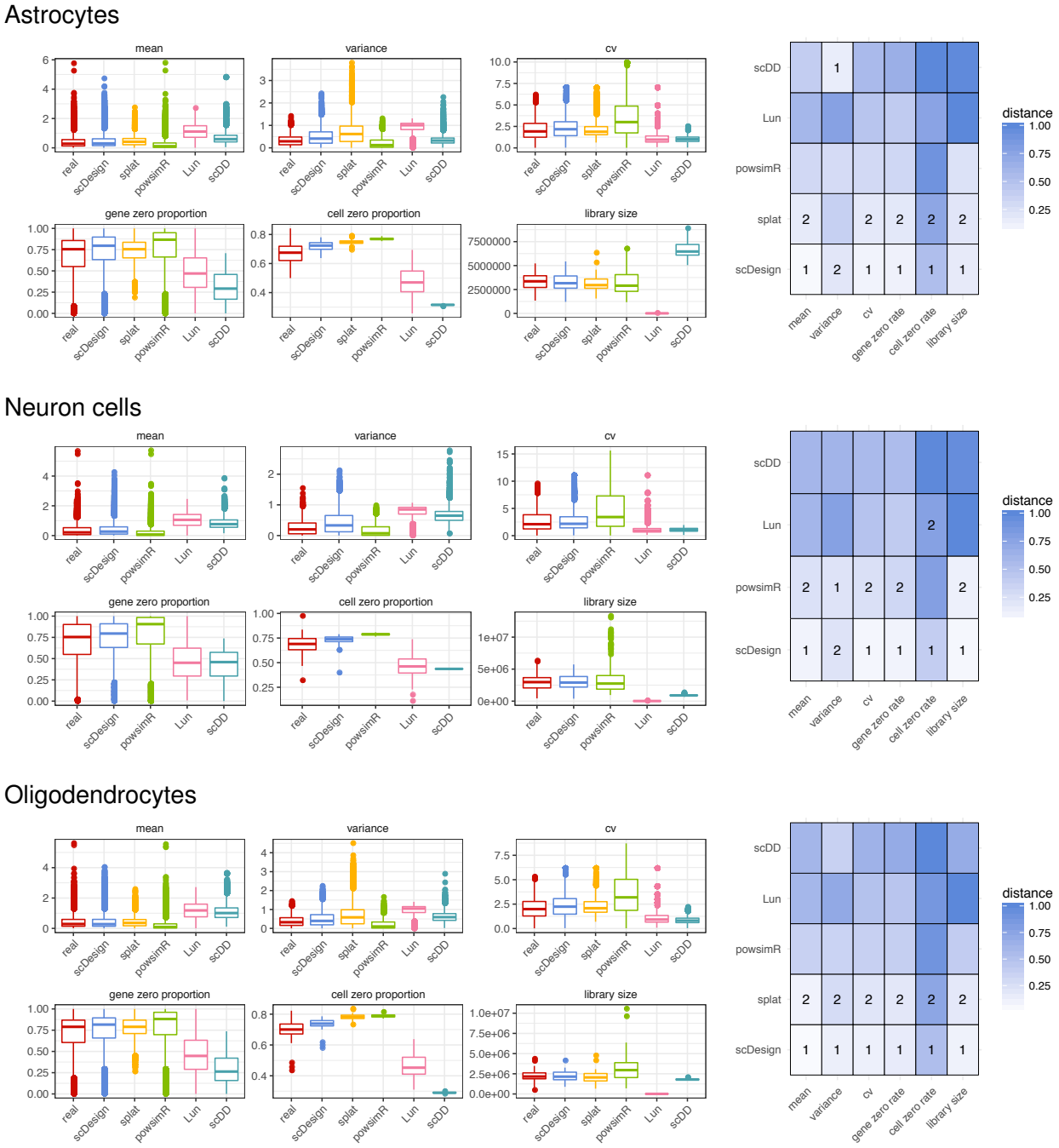


Figure S1: Comparison of scDesign and the other four simulation methods based on the Fluidigm C1 protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. Note that splat failed to simulate data for the neuron dataset.

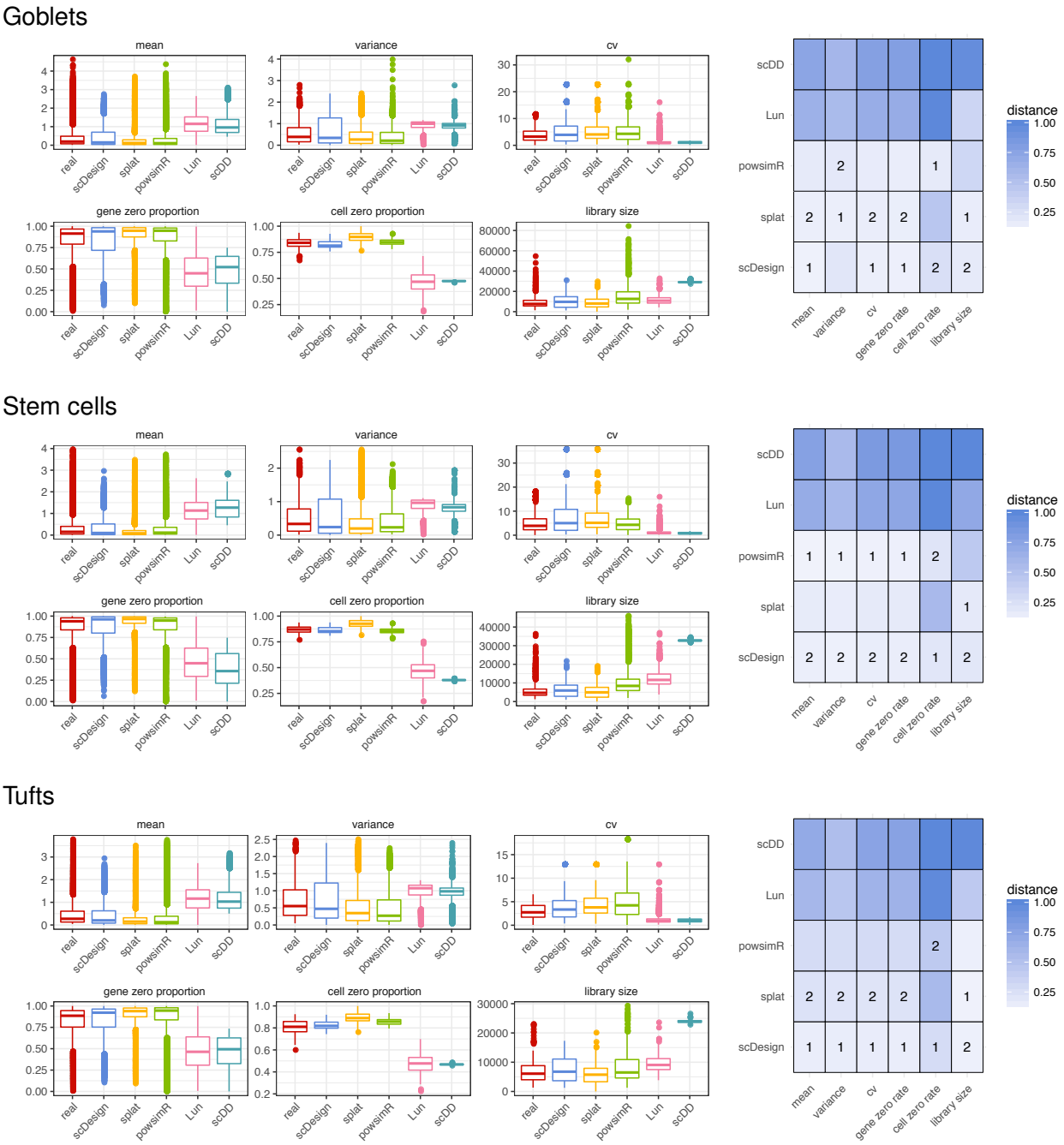


Figure S2: Comparison of scDesign and the other four simulation methods based on the 10x Genomics protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps.

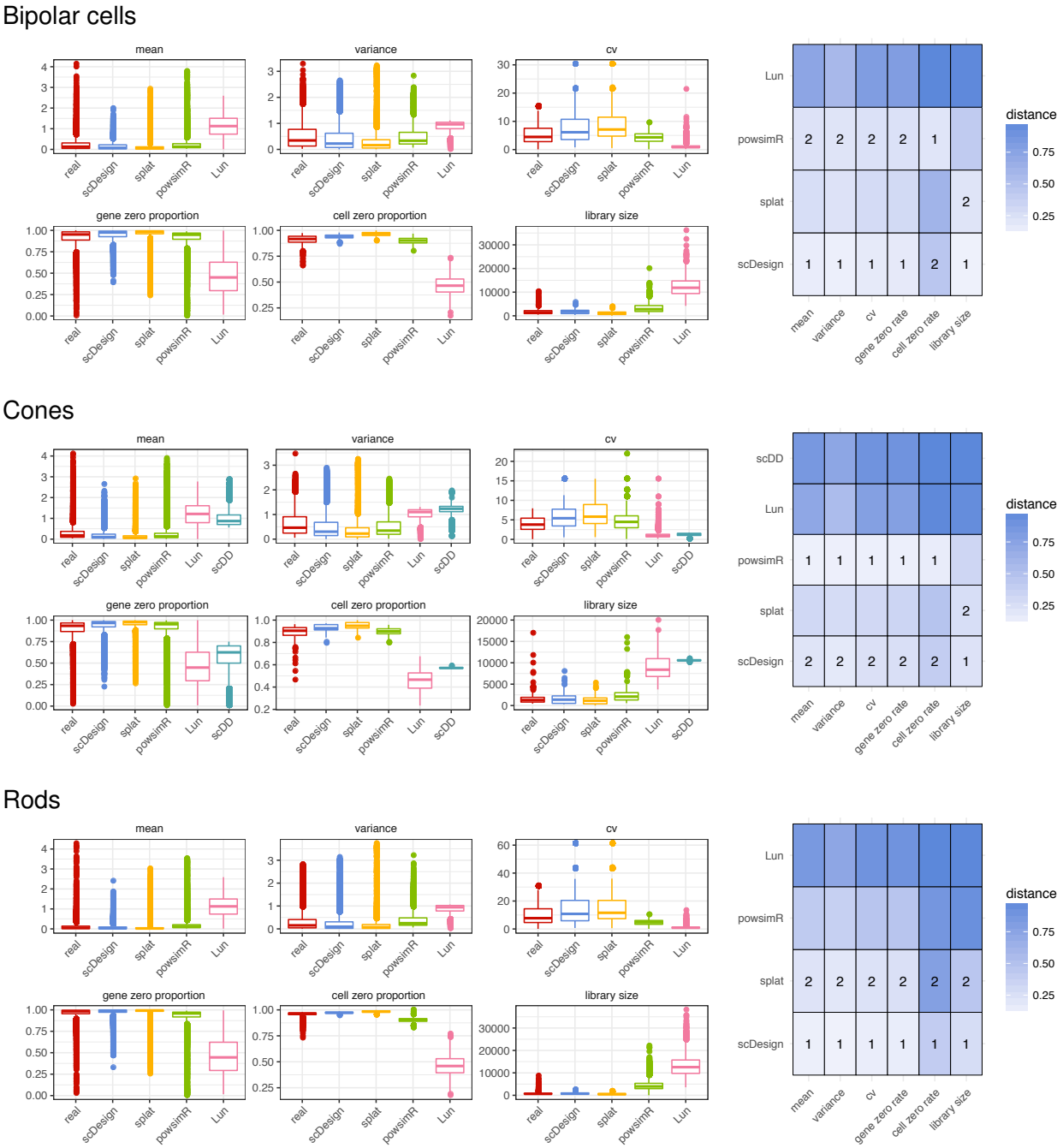


Figure S3: Comparison of scDesign and the other four simulation methods based on the Dropseq protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. Note that scDD failed to simulate data for the bipolar and rod datasets.

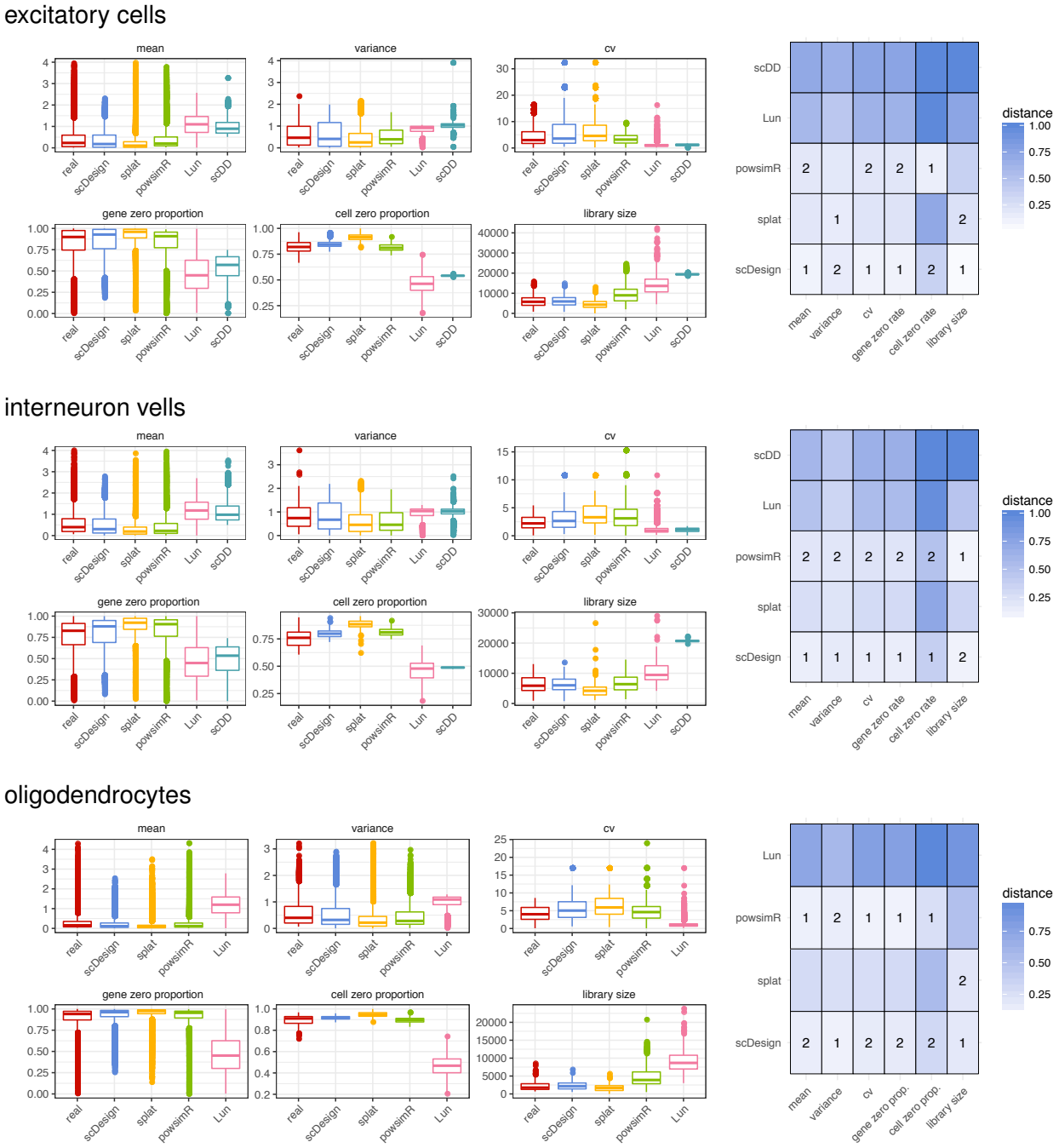


Figure S4: Comparison of scDesign and the other four simulation methods based on the inDrop protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. Note that scDD failed to simulate data for the oligodendrocyte dataset.

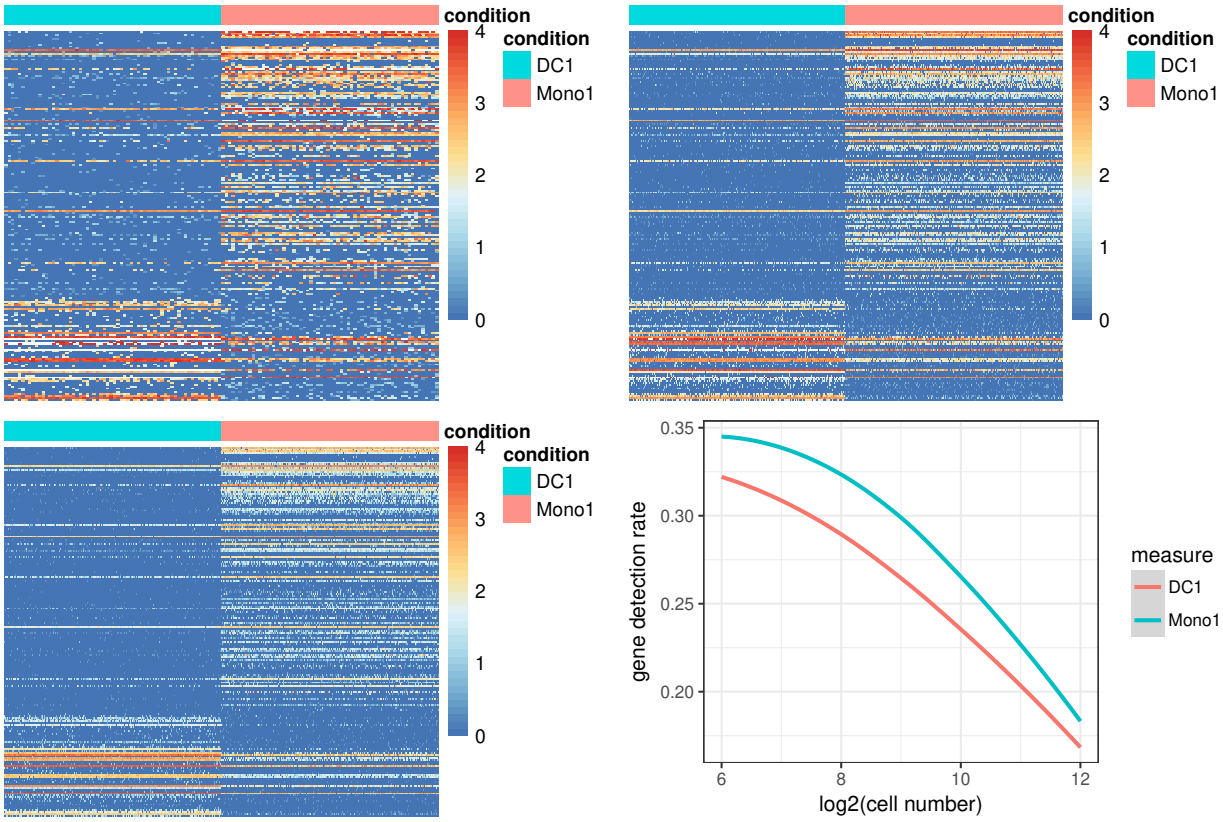


Figure S5: Trade-off between cell number and library size given a fixed total sequencing depth. The three heatmaps give the gene expression profiles of the top 200 DE genes when the simulated cell number of both dendrocytes (DC1) and monocytes (Mono1) is 64, 256, and 1024, respectively. The line plots give the median gene detection rates (i.e., the proportion of genes with non-zero read counts in a cell) in the two cell types as simulated cell number increases.

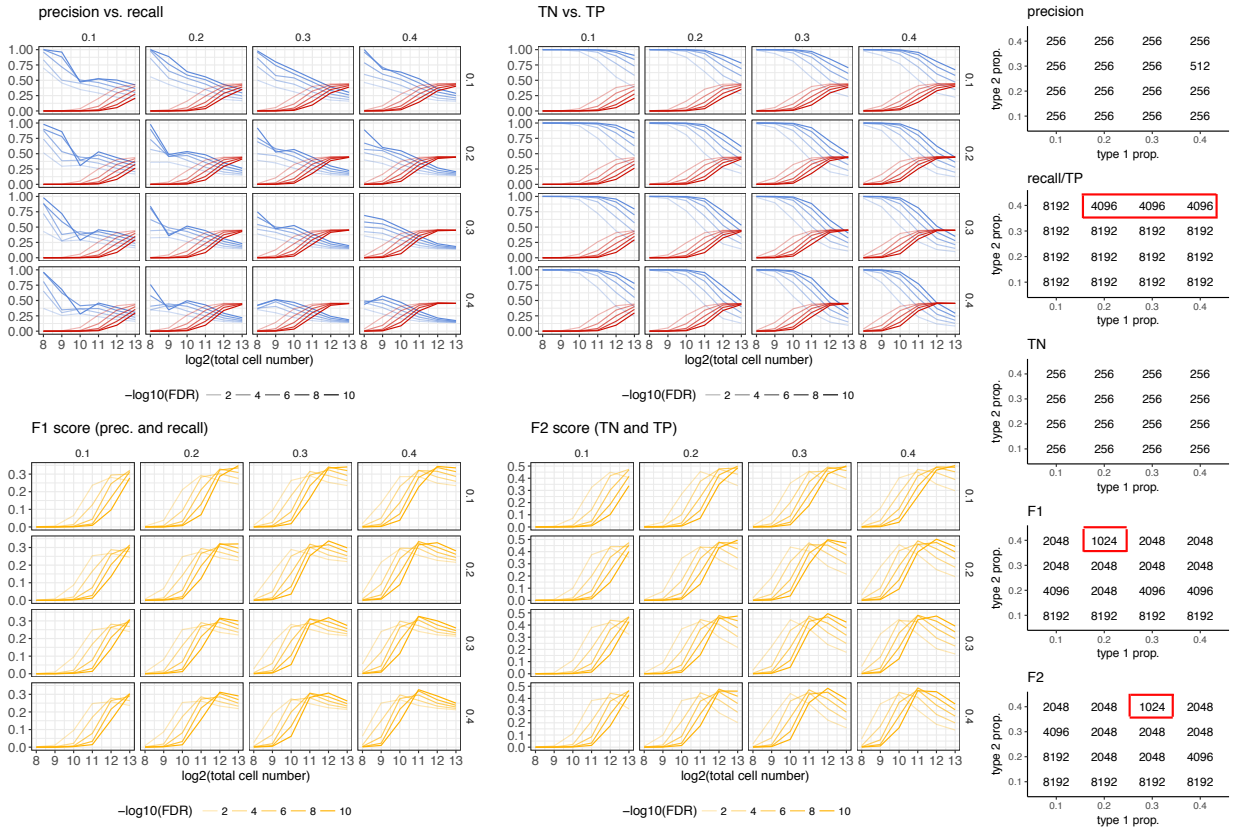


Figure S6: Power analysis for DE studies comparing CD4 and B cells with the Seqwell protocol (scenario 2). The thresholds on the false discovery rates (FDRs) (to identify DE genes) are denoted in the color legends. Different proportions ($0.1, 0.2, 0.3,$ and 0.4) of CD4 and B cells are considered in the experimental design. For the three metrics of recall, F1, and F2, the smallest total cell numbers leading to the best DE accuracy are marked in red boxes.

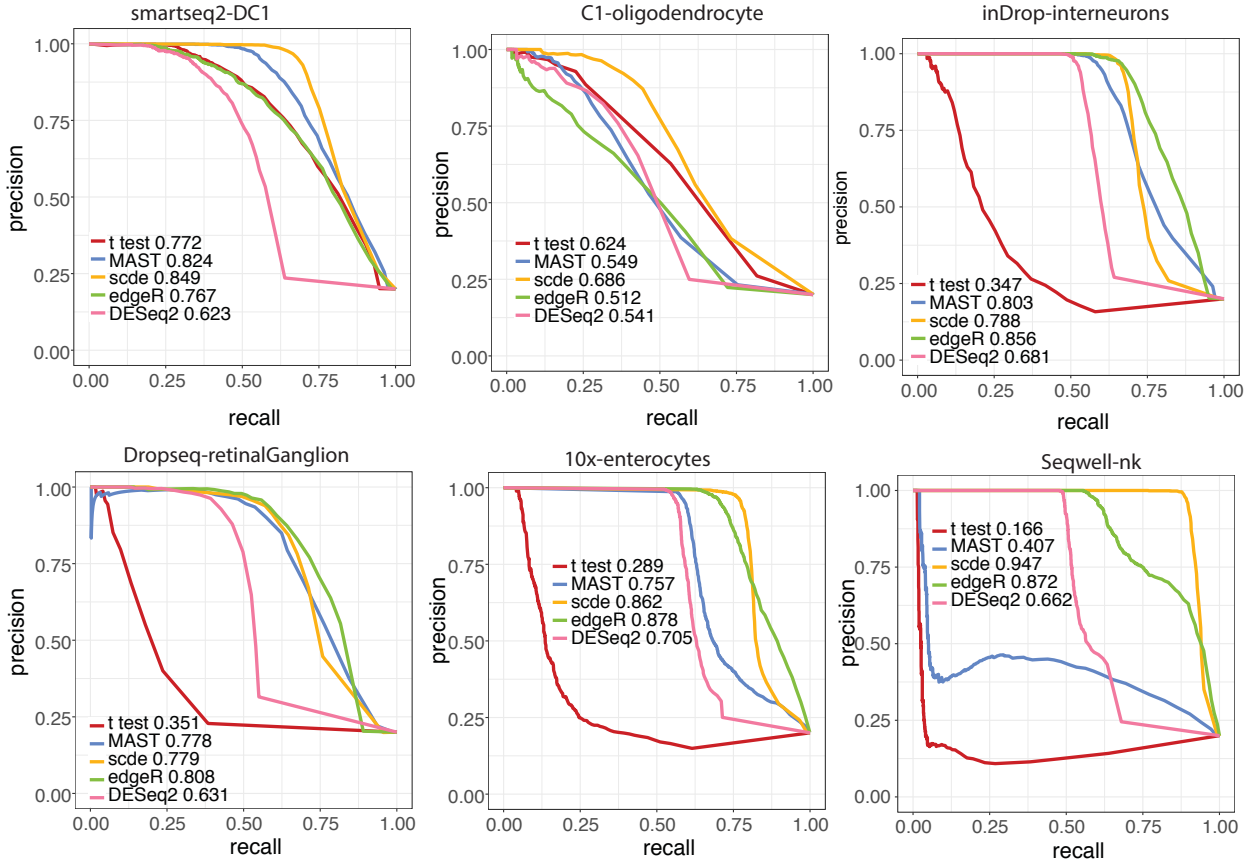


Figure S7: Comparison of scRNA-seq DE methods in the second setting. The precision-recall curves of the five DE methods are drawn for the six scRNA-seq protocols, respectively. Corresponding area under the curve (AUC) are given in the plots.

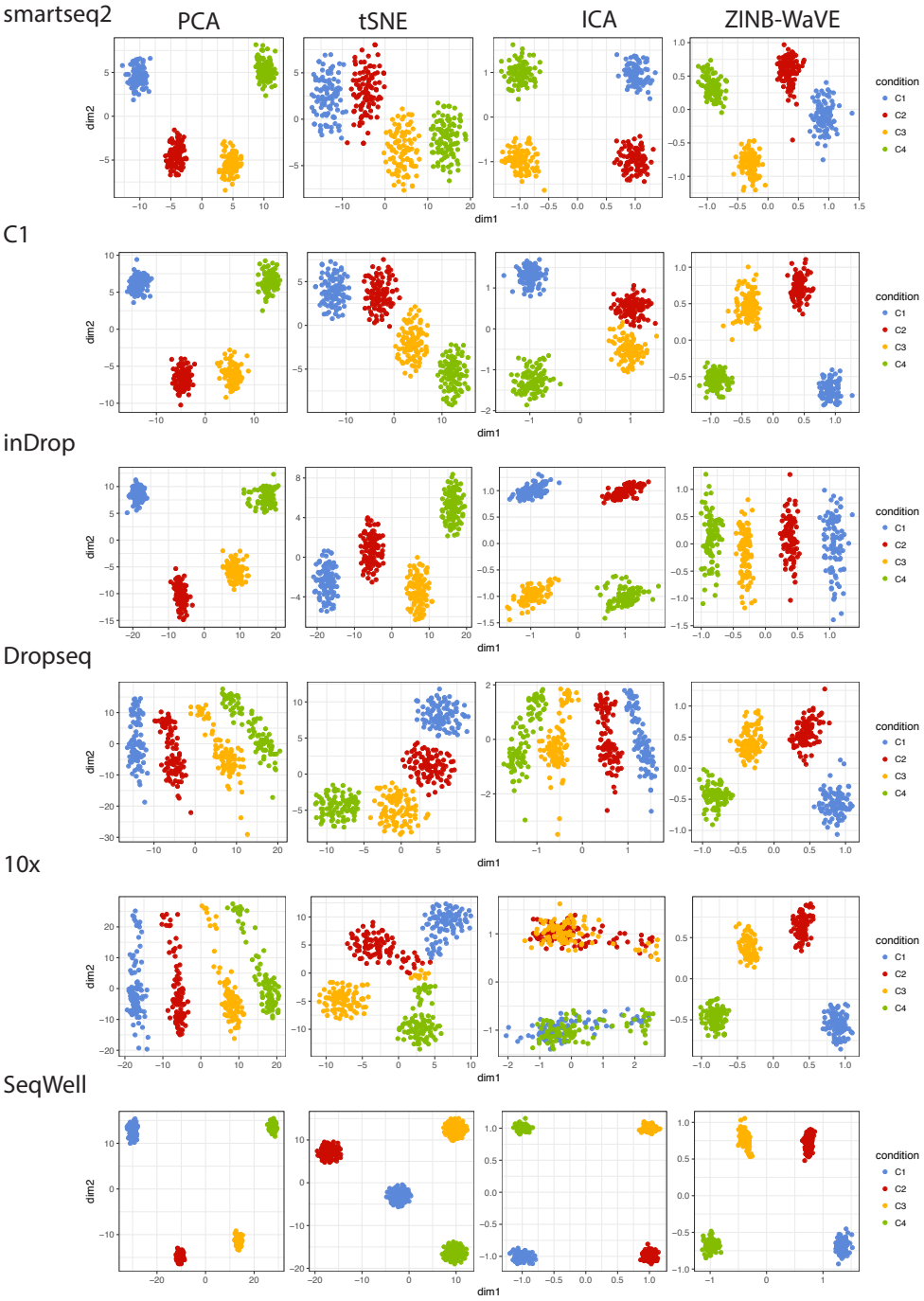


Figure S8: Comparison of scRNA-seq dimension reduction methods in the second setting. The first two dimensions resulting from the four dimension reduction methods are given for the six scRNA-seq protocols, respectively.