Probabilistic index models for testing differential expression in single cell RNA sequencing data

Alemu Takele Assefa[1], *Jo Vandesompele*[2,3,4] *and Olivier Thas*[1,3,5,6,]
[1]*Data analysis and mathematical modeling, Ghent University, Belgium,* [2]*Biomolecular Medicine, Ghent University, Belgium,* [3]*Cancer Research Institute Ghent, Ghent University, Belgium,* [4]*Center for Medical Genetics, Ghent University, Belgium,* [5]*National Institute for Applied Statistics Research, University of Wollongong, Australia, and* [6]*I-BioStat, Hasselt University, Belgium*

Abstract

Single-cell RNA sequencing (scRNA-seq) technologies profile gene expression patterns in individual cells. It is often of interest to test for differential expression (DE) between conditions, e.g. treatment and control or between cell types. Simulation studies have shown that non-parametric tests, such as the Wilcoxon-rank sum test, can robustly detect significant DE, with better performance than many parametric tools specifically developed for scRNA-seq data analysis. However, these classical rank tests cannot be used for complex experimental designs involving multiple groups, multiple factors and confounding variables. Further, rank based tests do not provide an interpretable measure of the effect size. We propose a semi-parametric approach based on probabilistic index models (PIM) that form a flexible class of models that generalize classical rank tests. Our method does not rely on strong distributional assumptions and it allows accounting for confounding factors. Moreover, our method allows for the estimation of the effect size in terms of a probabilistic index. Real data analysis demonstrated that PIM is capable of identifying biologically meaningful DE. Our simulation studies also show that tests for DE succeed well in controlling the false discovery rate at its nominal level, while maintaining good sensitivity as compared to competing methods.

Keywords: differential expression, single-cell RNA-sequencing, probabilistic index models, semi-parametric.

INTRODUCTION

Single cell RNA sequencing (scRNA-seq) technologies have become powerful tools to study the dynamics of gene expression at unprecedented high resolution [12, 31]. The data produced by such scRNA-seq technologies represent measures of transcript/gene abundance in individual cells. In the field of biology and medicine, it is actively used for various applications including delineating

population diversity, tracing cell lineages, classifying cell types, and genomic profiling of rare cells [31]. For example, analysis of scRNA-seq data is of paramount interest in studying diseases, such as cancer, to characterize intra-tumoral heterogeneity and to classify cancer cell sub-populations [31]. Of note, scRNA-seq data are known to be associated with particular issues that potentially challenge ordinary data analysis pipelines [17, 6]. Such issues include high abundance of zero counts, over-dispersion, and complex expression distributions (multimodal distributions) [3, 11, 24]. These issues are the result of either intrinsic biological variation or technical artifacts involved in various steps of the data processing pipeline [12, 11, 26]. Yet, the high level of noise can to some extent be compensated by the large number of cells processed in a scRNA-seq study [24].

One of the most common gene level analyses is testing for differential expression (DE). DE analysis aims at identifying a set of genes with different distributions of expression across populations of cells. Yet, the high magnitude of noise in scRNA-seq data challenge DE analysis in terms of controlling the type I error rate and sensitivity [23]. Various statistical methods for testing DE have been introduced. Many of them start from the assumption that the distribution of the data can be described by a particular parametric probability distribution. Among these, (zero inflated) negative binomial [29, 5] and two-component mixture distribution are common [1]. Parametric methods provide a flexible and straightforward modelling framework focusing on the difference in the mean expression across groups. Nevertheless, it is well known that the performance of parametric methods depends on the validity of the distributional assumptions, as well as on the precision of the parameter estimators. Simulation studies have demonstrated that non-parametric tests, such as the Wilcoxon rank sum test and SAMSeq [13], are equally useful for robustly testing DE with competitive performance in terms of sensitivity and false discovery rate (FDR) control [23, 34, 2]. Unfortunately, the rank tests available for scRNA-seq have a limited scope. First, they cannot be used for experiments involving comparison of multiple populations of cells or more than one factor of interest [23]. For example, Xin et al. [32] demonstrated cell type-specific DE between type-2 diabetics and controls while adjusting for subject specific characteristics such as gender and ethnicity. This design involves two main factors (disease status and cell type), their interaction effect and two additional nuisance factors, in which case simple rank-based tests cannot be used. Second, classical rank tests focus on testing for DE and they do not provide an estimate of an interpretable effect size measure. In this paper, we propose a new and flexible method that broadens the scope of rank-based tools and alleviates their limitation while retaining their robustness for testing DE in scRNA-seq data.

We propose a semi-parametric method based on Probabilistic Index Models (PIM) [25, 8], for testing DE in scRNA-seq data. PIMs entail a large class of semi-parametric models that can

generate many of the traditional rank tests, such as the Wilcoxon-Mann-Whitney test and the Kruskal-Wallis test [25, 8]. These models can be seen as the rank-equivalent of the generalized linear models (GLM). PIM methods do not rely on strong distributional assumptions and they inherit the robustness of rank methods. Moreover, PIMs come with parameters that possess an informative interpretation as effect sizes. PIMs can be used for complex experimental designs involving multiple groups of cells, and multiple discrete or continuous factors. PIMs form a regression framework and enable accounting for confounding factors that drive unwanted variations such as library size difference, batch, and cell cycle stage. Moreover, PIMs can be applied on the raw count data without the need for a normalization preprocessing step, and it can also be integrated with other data preprocessing methods, such as normalization, imputation, and removal of unwanted variation.

METHODS

Probabilistic index models

Let $(Y_i, X_i), i = 1, 2, \ldots, n$, be a set of n sample observations with Y the outcome variable and X the corresponding p-dimensional covariate. In a GLM the conditional mean $\mathrm{E}(Y|X)$ is modelled as a function of covariates through an appropriate link function. In contrast, a PIM models the conditional probability

$$\mathrm{P}(Y_i \preccurlyeq Y_j|X_i, X_j) = \mathrm{P}(Y_i < Y_j|X_i, X_j) + \frac{1}{2}\mathrm{P}(Y_i = Y_j|X_i, X_j), \tag{1}$$

where $Y_i \mid X_i$ and $Y_j \mid X_j$ have conditional distribution functions $F(\cdot; X_i)$ and $F(\cdot; X_j)$, respectively. The probability in (1) is known as the *probabilistic index (PI)* [25]. For a function $m(.)$ with range [0,1] and link function $g(.)$, PIM is defined as

$$\mathrm{P}(Y_i \preccurlyeq Y_j|X_i, X_j) = m(X_i, X_j; \beta) = g^{-1}(Z_{ij}^T\beta), \tag{2}$$

where β is the regression parameter and Z_{ij} is a function of the pair of covariates X_i and X_j. The latter is very often $Z_{ij} = X_j - X_i$. Similar to a GLM, the systematic component $m(X_i, X_j; \beta)$ is restricted to $g^{-1}(Z_{ij}^T\beta)$ with an appropriate link function $g(.)$. Throughout this paper, we use the *logit* link function (i.e. $g(x) = \log \frac{x}{1-x}$). To obtain the parameter estimates, say $\hat{\beta}$, the estimation method uses *pseudo observations* $I_{ij} = \mathrm{I}(Y_i < Y_j) + 0.5\mathrm{I}(Y_i = Y_j)$, where $\mathrm{I}(.)$ is the 0/1 indicator function. The pseudo-observations I_{ij} take values in $\{0,\ 0.5,\ 1\}$, and, like rank-based methods, they depend only on the ordering of the outcomes instead of the actual observations. It is also

showed that the estimator $\hat{\beta}$ is asymptotically normally distributed with a consistent estimator of its variance. We refer readers to [25], [8] and the supplementary file for further details about PIMs.

PIMs for testing differential gene expression

We propose PIMs for testing DE across experimental/biological conditions in scRNA-seq data. To illustrate how PIMs can be implemented for testing DE in scRNA-seq data, we first focus on a simple scenario with two groups of cells. Let Y_{gi} denote the normalized gene expression of gene $g = 1, 2, \ldots, G$ in cell $i = 1, 2, \ldots, n$. Later, we extend the model so that it can work directly on the raw counts. Let covariate $X = A$ be the treatment indicator, such that $A_i = 1$ if cell i is in one group and $A_i = 0$ if the cell is in the other group. We specify a PIM using the logit link function as

$$\text{logit}\left\{\mathrm{P}(Y_{gi} \preccurlyeq Y_{gj}|A_i, A_j)\right\} = \beta_g(A_j - A_i). \tag{3}$$

where the indices i and j refer cell i and j with associated grouping factor A_i and A_j, respectively. The parameter $\beta_g \in \mathbb{R}$ represents the effect of the factor A on the PI of the outcome. This effect is expressed as $P(Y_{gi} \preccurlyeq Y_{gj}|A_i = 0, A_j = 1) = \exp(\beta_g)/(1 + \exp(\beta_g))$. Once its estimate ($\hat{\beta}_g$) is obtained, the PI can be estimated as

$$\hat{\mathrm{P}}(Y_{gi} \preccurlyeq Y_{gj}|A_i, A_j) = \frac{e^{\hat{\beta}_g(A_j - A_i)}}{1 + e^{\hat{\beta}_g(A_j - A_i)}} \in [0, 1].$$

If there is a strong evidence that gene g is DE, then the estimated PI (PI= $\hat{\mathrm{P}}(Y_{gi} \preccurlyeq Y_{gj}|A_i = 0, A_j = 1)$) becomes close to 1 (if the gene expression is higher in the $A = 1$ group) or 0 (if lower in the $A = 0$ group). Under the null hypothesis (no DE), the estimated PI is expected to be 0.5, indicating that there is a 50% chance that the expression of gene g in a randomly selected cell from the $A = 0$ group is lower than that of a randomly selected cell from the $A = 1$ group (and vice versa). In other words, the PI indicates to what extent the distributions of the gene expression in the two groups are well distinguished. Thus, in this simple setting the PI is equivalent to the area under the receiver-operating-characteristics curve (AUC) in a classification problem. This is graphically illustrated in Figure 1.

Statistical hypothesis testing can be performed with the Wald-type test of [25]. For example, for the model in (3), the null hypothesis of no DE can be formulated as $H_0 : \beta_g = 0$. This is equivalent to testing $\mathrm{P}(Y_{gi} \preccurlyeq Y_{gj}|A_i = 0, A_j = 1) = 0.5$. Applying this test to every candidate gene results in a vector of raw p-values, to which the usual FDR control procedures, such as the Benjamin and

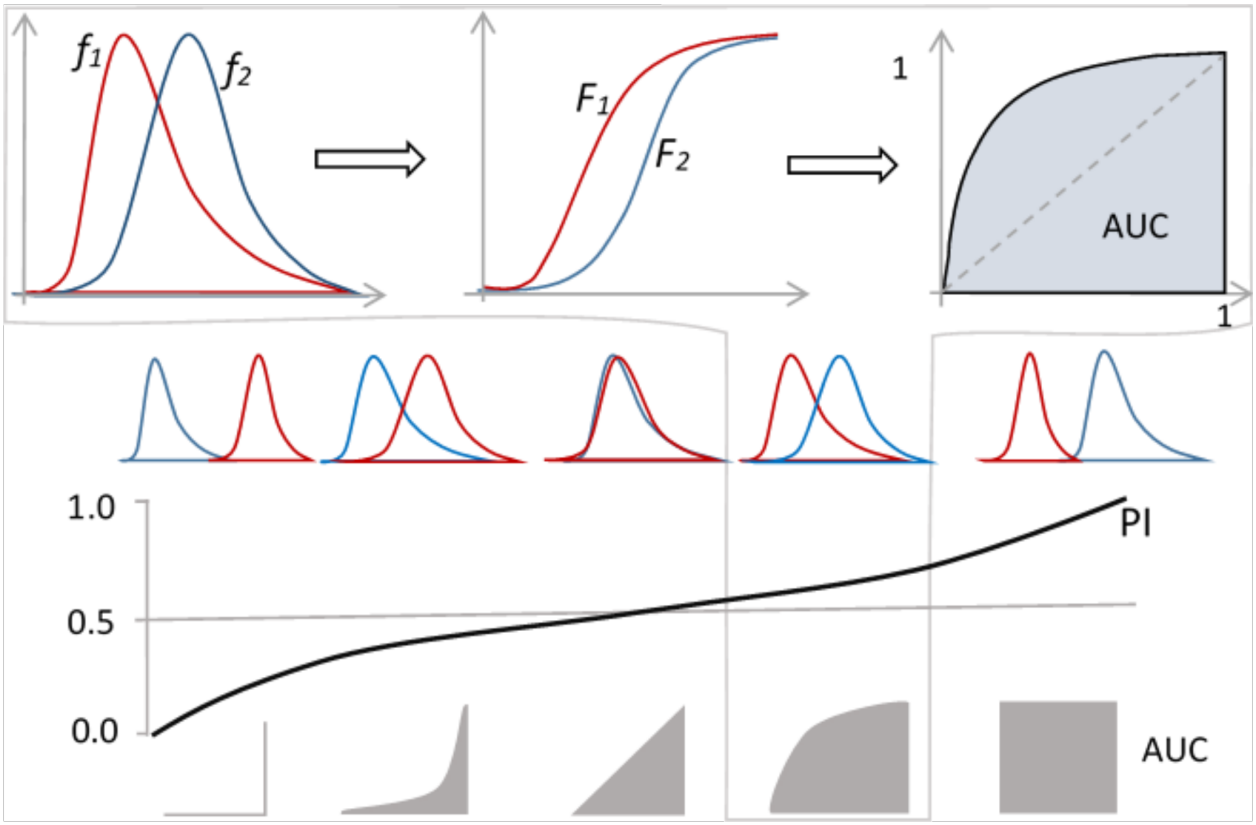

Figure 1: **Interpretation of the Probabilistic Index (PI).** The PI indicates to what extent the gene expression level distributions of two groups are different (f_1 versus f_2). In other words, the PI is equivalent to the area under the receiver-operating-characteristics curve (AUC) in a classification problem. Let Y_{i1} and Y_{i2} be the gene expression levels in group 1 and 2 with F_1 and F_2 their corresponding distribution functions. Assume we want to classify cells into two groups based on the expression level of a given gene (high or low expression at a particular threshold t). Thus, the PI =P($Y_{gi} \preccurlyeq Y_{gj}$) is equivalent to AUC for classifying a cell into group 2 (if expression$> t$) and group 1 (if expression$< t$) for all possible values of t. If the gene expression level is higher in group 2 than in group 1, then PI and AUC will be close to 1. Conversely, if the gene expression is higher in group 1 than in group 2, then PI and AUC will be close to 0. If there is no difference in expression, then both PI and AUC will be approximately 0.5

Hochberg [4] approach, can be applied. In the subsequent sections we will demonstrate that the p-values satisfy the assumptions required by such procedures.

Normalization and removal of unwanted variation

In the previous section, we assumed that the gene expression data were normalized per cell, using e.g. simple methods, such as transcripts per kilobase million (TPM) or counts per millions of reads (CPM) [20], or other (bulk RNA-seq) normalization methods that involve the calculation of a global scaling factor to normalize gene expression levels in each sample by a single constant. Importantly, typical scRNA-seq data are often confounded by several sources of unwanted variation among cells, such as library size, batch, cell size, cell cycle, and cell quality level that make cells not directly comparable [6]. Consequently, normalization of scRNA-seq data at best requires accounting for these nuisance factors. Hence, the meaning of this scaling factor is broader than in bulk RNA-seq data in which normalization is generally needed to remove library size differences [6, 28]. Moreover, the effect of such factors (e.g. cell cycle and batch) can be gene dependent [11]. Cole et al [6] and Risso et al [19] demonstrated that regression-based normalization is effective in addressing the technical factors in scRNA-seq data. PIMs allow for the incorporation of such known sources of variation. In this way, we can test for DE, while controlling for various technical factors at a gene level. In particular, let $\mathbf{W}$ be an $n \times q$ matrix of q cell level technical factors, then the PIM in (3) can be expressed as

$$\text{logit}\left\{\text{P}(Y_{gi} \preccurlyeq Y_{gj}|A_i, A_j, \mathbf{w}_i, \mathbf{w}_j)\right\} = \beta_g(A_j - A_i) + (\mathbf{w}_j - \mathbf{w}_i)^T\boldsymbol{\gamma}_g, \tag{4}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ are q-dimensional vectors of technical factors for cells i and j, respectively, and $\boldsymbol{\gamma}_g$ is a q-dimensional parameter vector for the technical factors. In this model, Y_{gi} is the observed expression of gene g in cell i. In particular, the adjusted effect of the group factor A on the PI becomes

$$\text{expit}(\beta_g) = P(Y_{gi} \preccurlyeq Y_{gj}|A_i = 0, A_j = 1, \mathbf{w}_j = \mathbf{w}_i),$$

where $\text{expit}(x) = \exp(x)/(1 + \exp(x))$. Thus, testing for DE between $A = 0$ and $A = 1$ is adjusted for the effect of the known nuisance factors $\mathbf{W}$.

In general, PIMs are flexible to be used for testing DE in a scenario with a multi-class factor or multiple discrete or continuous factors. For multiple factor studies, PIMs also allow for testing for interaction effects in which case one can test for the global effect or selected contrasts.

Single cell RNA-seq dataset

To benchmark PIM and demonstrate its application, we used two scRNA-seq datasets, from two different experimental protocols (see Table 1). The first dataset (A) is a cellular perturbation

experiment on the C1 instrument [30] (GEO accession GSE119984). This total RNA seq dataset contains 83 NGP neuroblastoma cells of which 31 were treated with 8μM of nutlin-3 and the other 52 cells were treated with vehicle (controls). The second scRNA-seq dataset (B) contains the same neuroblastoma cells under nutlin-3 treatment and control, except that the data are generated using the Chromium (10X Genomics) instrument (3' end squencing). We applied PIM and other competitor tools for testing DE between the treatment and control groups of cells in these two datasets.

Table 1: Summary of the scRNA-seq datasets

data	protocol	gene expression measure type	number of cells	number of genes	library size /10^5 (min-max)
A	SMARTer/C1	read counts	83	22,078	3.21 - 12.36
B	Chromium	UMI	7149	18,639	0.04 - 0.21

UMI=uique molecular identifier

Simulation methods

We implemented three simulation methods, each starting from other underlying assumptions.

Mock comparison: This simulation involves random assignment of cells from a single condition/population to two mock groups. Since we do not expect DE between mock groups, it enables the evaluation of controlling the false positive rate (FPR). In particular, cells within the control group (vehicle) in the two scRNA-seq datasets are randomly split into two mock groups between which no DE is expected. Any rejection of the null hypothesis at the 5% nominal per-comparison error rate (PCER) level is considered as false positive. From running 100 independent simulations, the actual FPR for each gene is approximated as the fraction of tests with un-adjusted p-value less than 5%.

Splat simulation – gamma-Poisson hierarchical model [33]: This simulation algorithm makes use of a gamma-Poisson hierarchical model in which the mean expression level for each gene is sampled from a gamma distribution and the count for each cell is subsequently sampled from a Poisson distribution. Splat also adds outlier mean expressions and it makes use of a mean-variance trend. It uses the logistic function for the observed relationship between the mean expression of a gene and the proportion of zero counts to add excess zeros representing technical noise. It is implemented using the *splatter* R Bioconductor package (version 1.6.1)[33]. The gamma-Poisson mixture distribution is a particular parametrization of the negative binomial (NB) model, and hence

the Splat simulation is equivalent to simulation from a NB model. Splat uses model parameters estimated from a real single cell RNA-seq data set. To add a set of DE genes, the mean expressions of randomly selected genes are multiplied by a factor, known as the fold-change.

SPsimSeq simulation [2]: Since the (zero inflated) NB distributional assumption for scRNA-seq data is debatable [22, 5], the results of the parametric simulations should be interpreted with care. SPsimSeq is a semi-parametric simulation method for bulk and scRNA-seq data. It makes use of no specific parametric distribution for the gene expressions. Instead, it generates data from a more realistic distribution, which can be estimated from a real source dataset by the semi-parametric method of [14] and [9]. In a first step, the log-CPM values from a given real data set are used for semi-parametrically estimating gene-wise distributions. It models the probability of zero counts as a function of the mean expression of the gene and the library size (read depth) of the cell (both in log scale). Zero counts are then added to the simulated data such that the observed relationship (zero probability to mean expression and library size) is maintained. It simulates DE by separately estimating the distributions of the gene expression from the different populations (for example treatment groups) in the source data, and subsequently sampling a new dataset from each group. It is implemented using the *SPsimSeq* R package (version 1.0.0)[2].

Competitor differential expression tools

We compared the PIM method with four other tools for testing DE. Three of the methods were specifically developed for analysis of scRNA-seq data. The fourth method is a non-parametric approach developed for micro-array and bulk RNA-seq data, but which shows good performance for testing DE in scRNA-seq data [23]. We briefly discuss each method in the subsequent paragraphs.

MAST [1] uses a hurdle model in which the zero counts are modelled by logistic regression, and the log TPM (or CPM) of the non-zero counts are modelled by normal linear regression. Both models include the cellular detection rate (fraction of detected genes per cell) as a covariate. MAST is implemented using *MAST* R biocondictor package (version 1.8.1)[1].

edgeR with Zinger: edgeR [18] is a very common tool for testing DE in bulk RNA-seq data. It fits NB regression models with moderated estimation of the gene-specific over-dispersion parameters. Zinger [29] calculates weights that can be used by edgeR to account for zero inflation. In particular, Zinger uses zero-inflated negative binomial models for calculating weights that indicate the probability that the observed zero expression belongs to the NB component. edgeR is implemented using *edgeR* [20] R bioconductor package (version 3.22.5) and Zinger is using *Zinger* [29] R package (version 0.1.0).

DESeq2 with Zinger: Like edgeR, DESeq2 [15] relies on the NB distribution for testing DE in bulk RNA-seq data. It uses an empirical Bayes moderation for estimating the over-dispersion parameters. The weights calculated by Zinger are also used with DESeq2. It is implemented using *DESeq2* R bioconductor package (version 1.20.0) [15].

SAMSeq[13] is a non-parametric method which was originally developed for testing DE in bulk RNA-seq data. It uses the Wilcoxon statistic in conjunction with a permutation plug-in method to estimate the false discovery rate to infer DE. It is implemented using the *samr* R software package (version 3.0) [27].

Real data analysis

We applied the 5 DE tools for testing DE between nutlin-3 treated and control cells from the two datasets (Table 1). The datasets first went through pre-processing steps using the *scatter* [17] R bioconductor package (version 1.6.3), which involves filtering cell with low quality metrics and genes with insufficient expression. The cell cycle stage is estimated using the *cyclone* function in *scran* [16] R bioconductor package (version 1.10.1). In particular, we compared the number of detected DE genes at the 5% nominal FDR among the 5 tools. We also looked at the distribution of the $\log_2$(CPM+1) as well as the distribution of zero counts within the set of DE genes detected by each tool. We also explored the cross data agreement of all tools in detecting DE genes. In addition to DE testing, we performed gene set enrichment analysis (GSEA) using the *fgsea*[21] R bioconductor package (version 1.8.0). In particular, we used a set TP53 pathway genes obtained from a study by Fischer [10]. Genes were ranked based their test statistic as in [29].

Software implementation for PIM

An R software package for DE analysis using PIM is accessible from a GitHub repository with name PIMseq (`https://github.com/CenterForStatistics-UGent/PIMseq`). Additionally, R codes used for all analyses reported in this manuscript are available in a GitHub repository (`https://github.com/CenterForStatistics-UGent/PIMseq-paper`).

RESULTS

Performance evaluation through simulation study

The ability to control the FDR without compromising the sensitivity to detect truly DE genes is an important requirement. This capability is evaluated in simulation studies in which gene expression data are generated with a built-in truth while retaining the characteristics of real data. Using three simulation procedures (Splat[33], SPsimSeq [2], and mock comparison), we evaluated the performance of PIM compared to edgeR (with Zinger), DESeq2 (with Zinger), MAST and SAMSeq. The simulation methods vary with respect to the data generating model (see Methods section). In general, two sets of simulations were conducted, each starting from other source data (see Table 1). Each simulation set involves the three simulation methods. The performance of the DE methods is evaluated by calculating the actual FDR and true discovery rate (TPR) over the simulation runs.

We compared the simulated datasets with the real datasets with respect to various characteristics proposed by [23, 33], such as the distribution of mean and variance expressions, mean-variance trend, distribution of zero fraction per gene, and variability among cells (supplementary file: Figure S1-S2). Although the Splat[33] simulation is flexible in simulating various scenarios, the simulated data showed the least resemblance with the real datasets. On the other hand, the SPsimSeq[2] procedure generated realistic scRNA-seq data that better capture the characteristics of the real dataset (supplementary file: Figure S1-S2).

Results from the simulation studies starting from source data A show that PIM succeeds in controlling the actual FDR below the nominal level (nominal FDR ranging from 0 to 0.4) for both Splat and SPsimSeq simulation procedures (Figure 2). edgeR and DESeq2 loose control of the actual FDR for the SPsimSeq simulations, which do not make use of parametric distributional assumption for the gene expression data. The actual TPR for PIM, SAMSeq and MAST are comparable and considerably higher than that of edgeR and DESeq2 for the SPsimSeq simulation. With the Splat simulation, edgeR and DESeq2 show slightly better TPR (up to 10% increase) than the other tools. This is expected because Splat generates data from the negative binomial distribution, which is also the working model for both edgeR and DESeq2. In contrast, PIM and SAMSeq show nearly the same performance for the two simulations, whereas MAST showed 20% reduction of sensitivity with the Splat simulation (Figure 2). The results starting from dataset B showed that all tools perform relatively poor, even though the number of cells is double of that of dataset A (Figure 3). This may be explained by the fact that scRNA-seq data with UMI-counts contain a large fraction

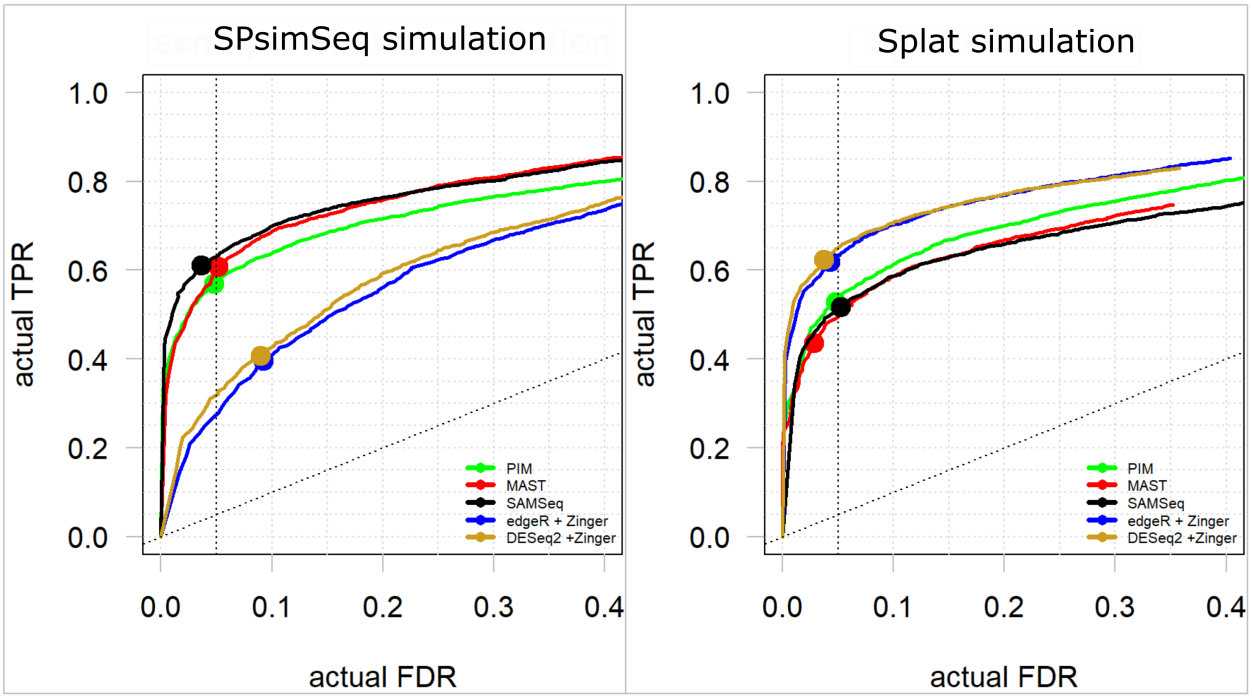

Figure 2: **Results from the simulation study starting from source dataset A.** Dataset A is neuroblastoma scRNA-seq data generated with SMARTer/C1 protocol. Each simulated dataset includes 2500 genes among which 10% DE, and 2 experimental groups with each 50 cells. The performance measures (actual FDR and TPR) are averaged over a total of 30 independent simulation runs. The curves represent actual FDR and TPR evaluated at nominal FDR levels ranging from 0 to 0.5, and the solid dots show the performances at the 5% nominal FDR level.

of zero counts, low expression magnitude and large variability among the library sizes [5]. With the SPsimSeq simulations, PIM shows better performance than all other tools, whereas with the Splat simulations PIM shows comparable performance to edgeR and DESeq2 in terms of the actual FDR and TPR (Figure 3). In contrast, SAMSeq and MAST lost control of the FDR and showed less than 25% sensitivity with both simulation procedures.

In addition, we evaluated the capability of PIM in controlling the type-I error rate (also known as false positive rate) using mock comparisons. Since SAMSeq does not return un-adjusted p-values, it is excluded from this comparison. Results from the mock comparison are presented in two ways. First, the observed average false positive proportions (FPP) across a number of mock data is nearly equal to the nominal 5% per-comparison error rate (PCER) for all tools (supplementary file: Figure S3). As the subsequent results show, the parametric tools are relatively conservative, leading to

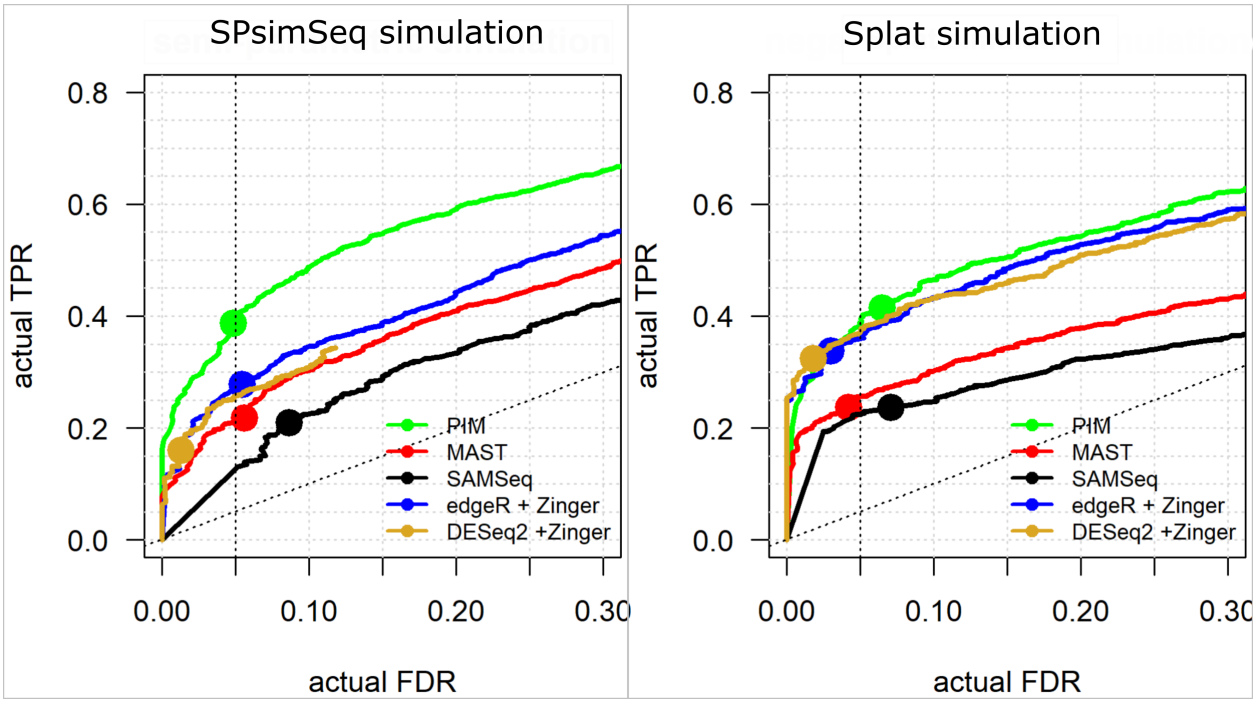

Figure 3: **Results from the simulation study starting from source dataset B.** Dataset B is neuroblastoma scRNA-seq data generated with Chromium protocol (UMI counts). Each simulated dataset includes 2500 genes among which 10% DE, and 2 experimental groups with each 100 cells. The performance measures (actual FDR and TPR) are averaged over a total of 30 independent simulation runs. The curves represents actual FDR and TPR evaluated at nominal FDR levels ranging from 0 to 0.5, and the solid dots show the performances at the 5% nominal FDR level.

smaller numbers of false positives. Second, since mock comparison simulates gene expression data under the null hypothesis (no DE), the distribution of raw p-values is expected to be uniform between 0 and 1. The distributions of raw p-values for PIM and all other considered DE tools are uniform except for DESeq2 (and slightly for MAST), which show a substantial deviation with higher probability mass near p-value=1 (also known as a conservative distribution; see supplementary file: Figure S4). This result is in line with the smallest observed average FPP for DESeq2 compared to all other tools (supplementary file: Figure S3). For uniform p-value distributions, as for PIM, it is anticipated that FDR control procedures, such as BenjaminiHochberg (BH) [4], can be used for FDR control.

DE analysis of nutlin-3 treatment of neuroblastoma cells

We applied PIM and the competitor tools for testing DE in two nutlin-3 treated neuroblastoma scRNA-seq datasets (see the data in Table 1). Nutlin-3 is a selective MDM2 antagonist, releasing TP53 from its targeted degradation, resulting in cell cycle arrest, followed by apoptotic cell death. The objective is to identify genes with DE between nutlin-3 treated and vehicle (control) cells. We focus on (i) comparing the number of genes detected as DE (at 5% nominal FDR) between PIM and other tools, (ii) exploring the set of DE genes in terms of the distribution of gene expression and fold change estimates (PI for PIM), (iii) TP53 gene set enrichment, and (iv) cross-data agreement.

For dataset A, we fitted the PIM with the treatment group and the log library size in the model. In this way, we test for DE by accounting for the library size differences. For dataset B, we extended the PIM model by adding the cell cycle stage (a discrete factor with three levels for the three cell cycle stages: G1, G2/M, and S) in the model. From dataset A, PIM identified a set of 355 genes, 225 down-regulated and 130 up-regulated genes in the nutlin-3 treated group (Figure 4-5). Compared to the other tools, PIM ranks second following SAMSeq, which detected over 430 DE genes. The parametric tools MAST, edgeR and DESeq2 identified relatively small numbers of DE genes (Figure 4). PIM and SAMSeq commonly identified 101 DE genes followed by 70 DE genes shared among all considered tools (Figure 4-panel a).

SAMSeq and PIM identified the highest number of DE genes. In particular, these tools identified almost all the DE genes identified by the parametric tools (Figure 4-panel a). SAMSeq uniquely called almost 200 genes DE. However, these genes have high variability and small log-fold-change estimates (< 0.5) compared to that of the other tools (supplementary file: Figure S6). Moreover, from exploring the distribution of the expression (log-CPM), specifically for the set of DE genes, we learn that SAMSeq and MAST tend to pick genes with relatively higher difference in the fraction of zero counts between the two treatment groups (Figure 4-panel b, supplementary file: Figure S7). In contrast, the set of DE genes identified by edgeR and DESeq2 have the smallest difference in the distribution of zero counts, presumably because Zinger down-weights the contribution of zero counts. Consequently, edgeR and DESeq2 uniquely identify those with difference only in the distribution of the non-zero expressions (Figure 4-panel b, supplementary file: Figure S7). In contrast, PIM detected DE genes with small to moderate difference in zero count distribution across the groups (Figure 4-panel b, supplementary file: Figure S7).

For dataset B, all tools but SAMSeq identify many DE genes (Figure 4-panel d). The two datasets are similar in terms of experimental design, but they resulted from different protocols

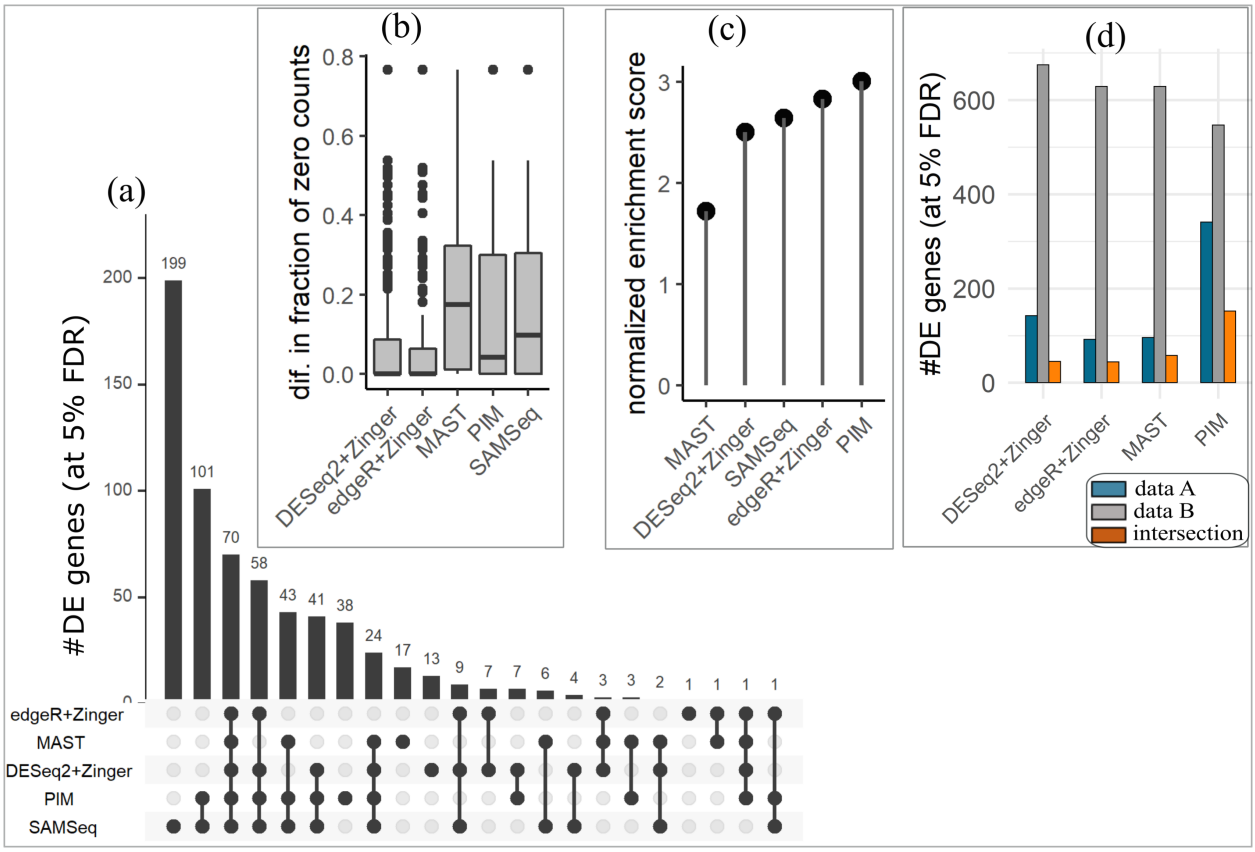

Figure 4: **Results from DE analysis of dataset A and B.** (a) UpSet plot to compare the set of DE genes detected by the 5 tools at the 5% FDR level. The vertical bars show the number of commonly detected genes between different combination of the tools. (b) The distribution of the difference (absolute) in the fraction of zero counts between nutlin-3 treated and control groups, particularly for the set of genes detected as DE by each tool. (c) The normalized enrichment score for TP53 gene set. (d) The number of DE genes detected by each tool in the two data sets (datasets A and B) and the numbers of genes detected in both datasets by each tool (cross data agreement).

(SMARTer/C1 and Chromium) and contain different gene expression units (read counts and UMI counts). In both datasets, we test for DE between the nutlin-3 and control group. This further allows us to evaluate the cross-data agreement of PIM compared to the competitor tools. SAMSeq failed for dataset B and was excluded from the cross-data agreement comparison. PIM identified 152 common DE genes in the two datasets, whereas MAST, edgeR, and DESeq2 identified 58, 44, and 45 genes, respectively, in both datasets (Figure 4-panel d). This result suggests that PIM has a better consistency across protocols and gene expression units.

The model parameters of PIM have an informative interpretation of the effect size in terms of PI. We now demonstrate that the estimated PI is a useful metric in addition to the fold-change used in parametric DE analysis. Fold changes express the biological signal in terms of the difference in the mean expression between conditions, with unbounded scale. Similarly, the PI quantifies the degree of difference in the distribution of expression at a scale ranging from 0 to 1 (see Figure 1). In Figure 5-panel a, we ranked genes according to their estimated PI in such a way that genes with strong evidence of DE appear on the left edge (if down-regulated in treated group) or at the right edge (if up-regulated in treated group). Note that, the estimated PIs are linearly correlated to the log-fold-change (Figure 5-panel b), suggesting that the PI estimates can also capture the biological signal. Note that we do not suggest that the fold-change should not be used. With parametric methods, the fold-change is still the effect size that correspond to the p-value. In a typical gene expression study, the majority of the genes are not DE, with estimated log-fold-change close to 0. This also holds for PIM, i.e. the estimated PI for the majority of the genes is around 0.5 (Figure 5-panel a nd c).

We also ran GSEA to find out if the DE genes detected by PIM have a biological meaning. Nutlin-3 liberates TP53 from MDM2-mediated inhibition and hence activates the TP53 transcriptional response pathway [10]. Therefore, we used a set of 116 TP53 pathway genes obtained from a meta-analysis consensus list by Fischer [10]. We compared the normalized enrichment score (NES) among the 5 DE tools including PIM (Figure 4-panel c). All the tools scored above 1.7 NES with p-value < 0.001. PIM attained the highest NES (NES=2.99) followed by edgeR (NES=2.85), SAMSeq (NES=2.64), DESeq2 (NES=2.53) and MAST (NES=1.72). In general, PIM significantly enriched the TP53 pathway for neuroblastoma cells treated with nutlin-3.

DISCUSSION

Advances in sequencing technology enable profiling of gene expression levels at individual cell level. ScRNA seq experiments are used for various applications in the field of biology and medicine, such as identifying genes whose expression differ under different conditions (also known as DE). Importantly, scRNA-seq data are characterized by high levels of noise resulting from various technical sources, such as limited RNA capture efficiency, amplification bias and differences in RNA content [12, 3]. Furthermore, the diverse biological characteristics of individual cells and the stochastic nature of gene expression challenge computational tools for testing DE.

Two broad classes of methods exist for testing DE, differing in terms of their underlying as-

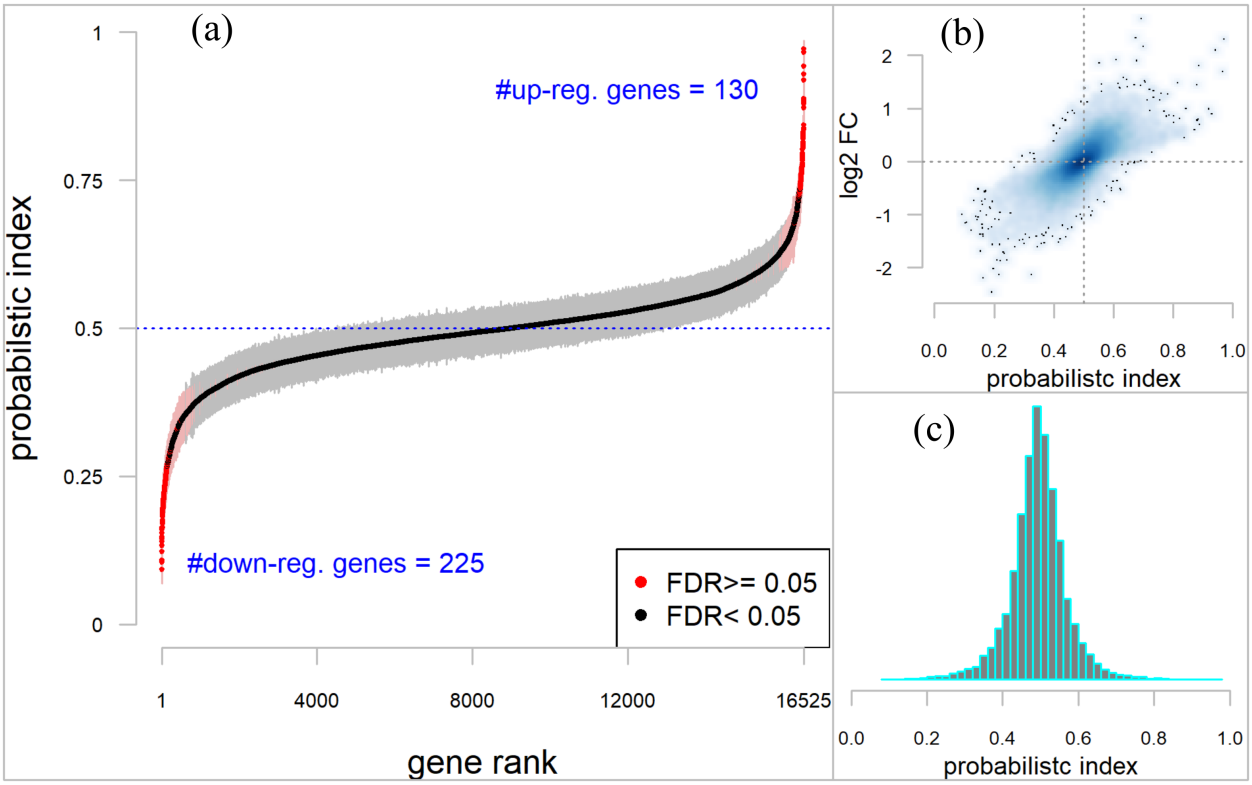

Figure 5: **PIM results for the DE analysis of dataset A.** (a) Gene ranking based on their estimated PI. The gray region indicates the 95% confidence interval estimate of the PI for each gene. (b) The distribution of PI across genes. (c) The relationship between estimated PI and a $\log_2$-fold-change ($\log_2$FC) estimates. The $\log_2$FC were obtained using DESeq2.

sumption for the gene expression data, i.e. parametric vs. non-parametric. Parametric tools such as edgeR and DESeq2 have been widely used for bulk and single cell RNA-seq studies. When used in conjunction with Zinger [29], they account for both the over-dispersion and zero-inflation issues in scRNA-seq data. MAST fits hurdle-Gaussian models to deal with zero inflation assuming that all zeroes result from technical factors. The performance of these tools seemed adequate using simulation studies that make use of their underlying distribution [23]. In contrast, non-parametric methods require minimal distributional assumptions and are considered useful when the underlying data generating distribution is unknown, as is the case for scRNA-seq data. Simple non-parametric methods like the Wilcoxon rank-sum test showed competitive or better performance relative to many of the parametric tools specifically developed for scRNA-seq data [23, 34]. Unfortunately, the usual non-parametric tools have a limited utility for scRNA-seq studies with complex experimental design. Moreover, they do not provide an informative estimate of the effect size, which is a

vital component of DE analysis. In this paper we introduce PIMs that generalize non-parametric tools and provides an informative effect size.

Using three simulation studies, we have demonstrated that PIMs succeed in controlling the FDR with reasonable sensitivity to detect true DE genes. The three simulation studies differ with respect to the underlying data generating assumption: fully parametric, semi-parametric, and non-parametric (mock comparison). Unlike the parametric tools, PIMs performs well under both simulation mechanisms. In contrast, edgeR and DESeq2 showed slightly better performance when the data were generated from negative binomial distributions. MAST and SAMSeq are greatly affected by the high fraction of zeroes in simulation studies involving UMI counts. In addition to simulation studies, we applied PIMs and the other tools for testing DE in two real and novel scRNA-seq datasets. These datasets represent two different types of experiments with different platforms, number of cells and units of gene expression, i.e. raw read counts versus UMI deduplicated reads. The results demonstrate that PIMs detect large numbers of DE genes with better cross-data agreement. Further, the PI ranked gene list is significantly positively enriched with genes from a pathway known to be activated in the cellular model system, demonstrating that PIM allows to attain biological insights from scRNA-seq experiments.

The standard parameter estimation methods for PIM involve comparison of all outcome pairs, increasing the computation time quadratically with the sample size. To improve PIMs scalability and computational efficiency, we made use of the Cox-proportional hazard model [7] translation of PIMs (see supplementary file for the detail). Finally, while we demonstrated the utility of PIMs for testing DE in scRNA-seq data, PIMs could also be used for bulk RNA-seq and microarray datasets with a sufficient number of replicates. In principle, PIMs can also be integrated with other data pre-processing pipelines such as normalization, batch correction, and imputation, making it a useful and general tool for single cell RNA seq differential gene expression analysis.

SUPPLEMENTARY DATA

Additional supporting results and discussion are available in a *Supplementary File*.

FUNDING

This study was sponsored by a Gent University Special Research Fund Concerted Research Actions (GOA grant number BOF16-GOA-023). OT and ATA are also acknowledge Research Fund Flanders

(FWO), Grant/Award Number G020214N.

Conflict of interest statement.

None declared.

References

[1] M. Aartsen, M. Ackermann, J. Adams, J. Aguilar, M. Ahlers, M. Ahrens, D. Altmann, T. Anderson, C. Arguelles, T. Arlen, et al. Search for prompt neutrino emission from gamma-ray bursts with icecube. *The Astrophysical Journal Letters*, 805(1):L5, 2015.

[2] A. T. Assefa, J. Vandesompele, and O. Thas. Spsimseq: semi-parametric simulation of bulk and single cell rna sequencing data. *bioRxiv*, 2019.

[3] R. Bacher and C. Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biology*, 17(1):63, Apr 2016.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[5] W. Chen, Y. Li, J. Easton, D. Finkelstein, G. Wu, and X. Chen. Umi-count modeling and differential expression analysis for single-cell rna sequencing. *Genome biology*, 19(1):70, 2018.

[6] M. B. Cole, D. Risso, A. Wagner, D. DeTomaso, J. Ngai, E. Purdom, S. Dudoit, and N. Yosef. Performance assessment and selection of normalization procedures for single-cell rna-seq. *bioRxiv*, 2017.

[7] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[8] J. De Neve and O. Thas. A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511):1276–1283, 2015.

[9] B. Efron, R. Tibshirani, et al. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

[10] M. Fischer. Census and evaluation of p53 target genes. *Oncogene*, 36(28):3943, 2017.

[11] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell rna- sequencing experiments. *bioRxiv*, 2017.

[12] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[13] J. Li and R. Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5):519–536, 2013.

[14] J. Lindsey. Construction and comparison of statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 418–425, 1974.

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

[16] A. T. Lun, D. J. McCarthy, and J. C. Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5, 2016.

[17] D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, 2017.

[18] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

[19] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

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

[21] A. Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*, page 060012, 2016.

[22] J. D. Silverman, K. Roche, S. Mukherjee, and L. A. David. Naught all zeros in sequence count data are the same. *bioRxiv*, 2018.

[23] C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[24] O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133, 2015.

[25] O. Thas, J. D. Neve, L. Clement, and J.-P. Ottoy. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):623–671, 2012.

[26] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

[27] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[28] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.

[29] K. Van den Berge, C. Soneson, M. I. Love, M. D. Robinson, and L. Clement. zinger: unlocking rna-seq tools for zero-inflation and single cell applications. *bioRxiv*, page 157982, 2017.

[30] K. Verboom, C. Everaert, N. Bolduc, K. J. Livak, N. Yigit, D. Rombaut, J. Anckaert, S. Lee, M. T. Ven, J. Kjems, F. Speleman, P. Mestdagh, and J. Vandesompele. SMARTer single cell total RNA sequencing. *Nucleic Acids Research*, 06 2019.

[31] Y. Wang and N. E. Navin. Advances and applications of single-cell sequencing technologies. *Molecular cell*, 58(4):598–609, 2015.

[32] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell metabolism*, 24(4):608–615, 2016.

[33] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[34] X. Z. Zhun Miao. Differential expression analyses for single-cell rna-seq: old questions on new data. *Quantitative Biology*, 4(4):243, 2016.