Published online 21 January 2022

Nucleic Acids Research, 2022, Vol. 50, No. 8 ***e48***
https://doi.org/10.1093/nar/gkac019

Benchmarking of a Bayesian single cell RNAseq differential gene expression test for dose–response study designs

Rance Nault [1,2,†], Satabdi Saha [3,†], Sudin Bhattacharya [4], Jack Dodson [1], Samiran Sinha[5], Tapabrata Maiti [3,*] and Tim Zacharewski [1,2,*]

[1]Department of Biochemistry & Molecular Biology, Michigan State University, East Lansing, MI, USA, [2]Institute for Integrative Toxicology, Michigan State University, East Lansing, MI 48824, USA, [3]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA, [4]Biomedical Engineering Department, Pharmacology & Toxicology, Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA and [5]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

Received September 27, 2021; Revised December 15, 2021; Editorial Decision January 04, 2022; Accepted January 07, 2022

ABSTRACT

The application of single-cell RNA sequencing (scRNAseq) for the evaluation of chemicals, drugs, and food contaminants presents the opportunity to consider cellular heterogeneity in pharmacological and toxicological responses. Current differential gene expression analysis (DGEA) methods focus primarily on two group comparisons, not multi-group dose–response study designs used in safety assessments. To benchmark DGEA methods for dose–response scRNAseq experiments, we proposed a multiplicity corrected Bayesian testing approach and compare it against 8 other methods including two frequentist fit-for-purpose tests using simulated and experimental data. Our Bayesian test method outperformed all other tests for a broad range of accuracy metrics including control of false positive error rates. Most notable, the fit-for-purpose and standard multiple group DGEA methods were superior to the two group scRNAseq methods for dose–response study designs. Collectively, our benchmarking of DGEA methods demonstrates the importance in considering study design when determining the most appropriate test methods.

INTRODUCTION

Single-cell transcriptomics enables researchers to investigate homeostasis, development and disease at unprecedented cellular resolution (1–5). As with any new innovative technology, diverse tools soon follow to address specific applications and unique challenges. Currently, there are dozens of differential gene expression analysis (DGEA) approaches for single-cell RNAseq (scRNAseq) data; developed based on differences in assumptions, statistical methodologies, and study designs (6–11). A recent comparison of 36 approaches demonstrated acceptable performance for common bulk RNAseq tools such as edgeR and limma-trend, and MAST for snRNAseq, as well as common statistical tests such as the Wilcoxon rank sum (WRS) and the *t*-test (9). However, most methods have been developed primarily for two group comparisons whereas study designs typical of pharmacology and toxicology experiments such as as dose–responses consist of multiple groups. The use of two sample tests for multiple group study designs elevate the type I error rate warranting further investigation of these methods for multiple group dose–response study designs (12).

Dose–response studies are used to derive the efficacy and/or safety margins such as effective dose and the point of departure (POD). Significant efforts by the toxicology and regulatory communities have suggested that acute (<14 days) and sub-acute (14–28 days) transcriptomic studies as viable alternative to the current standard 2-year rodent bioassay that significantly reduces the time and resources needed to assess risk (13–15). Gene expression profiling at single-cell resolution could further support such evaluations by identifying cell-specific dose-dependent responses indicative of an adverse event. The U.S. National Toxicology Program (NTP) recently reported a robust DGEA approach is essential to deriving biologically relevant PODs (15). However, concerns regarding the inclusion of false positives that produce less conservative POD estimates potentially leads to incorrect classification of mode-of-action

*To whom correspondence should be addressed. Tel: +1 517 353 3233; Fax: +1 517 353 9334; Email: maiti@stt.msu.edu
Correspondence may also be addressed to Tim Zacharewski. Tel: +1 517 355 1607; Fax: +1 517 353 9334; Email: tzachare@msu.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

(MoA), thus highlighting the importance of controlling type I error rates (16,17).

Unlike microarray and bulk RNAseq datasets, single-cell RNAseq (scRNAseq) data exhibits excess zero values due to the low per cell RNA input, biases in capture and amplification, transcriptional bursts, and other technical factors (18). This no expression (zero values), due to a conflation of both biological and technical factors, results in an excessive number of zeroes in an otherwise continuous measure (19). Therefore, traditional tests of differential gene expression, based on the assumption of a normal distribution, fail to correctly model the bimodality of single cell gene expression (20). Consequently, scRNAseq test methods usually consider the gene expression distribution as a mixture of a unexpressed (zero) and a positively (non-zero) expressed population (19–21). For example, the Seurat Bimod approach tests for differential gene expression using a likelihood ratio test designed for the said mixture population. MAST extends the Seurat Bimod test to a two-part generalized linear model structure capable of incorporating covariates (19,20). Given the improved performance of MAST (9,19–20), we hypothesized that multiple group tests developed assuming the same distributional framework would be most favorable for dose–response study designs. Furthermore, a Bayesian approach which considers prior knowledge is anticipated to minimize type I error rates (22,23).

The aim of the presented study is to evaluate the performance of existing and novel DGEA test methods on dose–response scRNAseq datasets. To reduce the rate of false positives we propose a novel, multiplicity corrected, Bayesian multiple group test (scBT) designed exclusively for DGEA of dose–response scRNAseq data. Two other fit-for-purpose frequentist multiple group tests are also examined: (i) a multiple group extension of the Seurat Bimod test and (ii) a simple extension of test (i) to a generalized linear model framework. Existing and proposed methods are benchmarked on simulated and real experimental dose–response datasets. Using simulated datasets we were able to investigate the influence of various parameters such as number of cells, and illustrate how using different test methods can aid in gaining biological insight on the role of individual cell types on the pathophysiological consequences of exposure.

MATERIALS AND METHODS

Animal handling and treatment

Male C57BL/6 mice aged postnatal day (PND) 25 were obtained from Charles Rivers Laboratories (Kingston, NY) were housed and treated as previously described (24). Mice were housed in Innovive cages (San Diego, CA) with ALPHA-dri bedding (Shepherd Specialty Papers, IL) at 23°C, 30–40% relative humidity, and a 12:12 h light:dark cycle. Aquavive water (Innovive) and Harlan Teklad 22/5 Rodent Diet 8940 (Madison, WI) was provided *ad libitum*. On PND 29, randomly assigned mice were gavaged at Zeitgeber time (ZT) 0 with 0.1 mL sesame oil vehicle (Sigma-Aldrich, St. Louis, MO), 0.01, 0.03, 0.1, 0.3, 1, 3, 10 or 30 μg/kg TCDD every 4 days for 28 days (7 total administered doses). At day 28 mice were euthanized by CO_2 asphyxiation and livers were immediately flash frozen in liquid nitrogen and stored at −80°C. All animal procedures were approved by the Michigan State University (MSU) Institutional Animal Care and Use Committee (IACUC) and reporting of *in vivo* experiments follow the Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) (25) and Minimum Information about Animal Toxicology Experiments (MIATE) guidelines (https://fairsharing.org/FAIRsharing.wYScsE).

Real scRNAseq and snRNAseq datasets

Hepatic single-nuclei RNA-sequencing (snRNAseq) was performed as previously described using the 10× Genomics Chromium Single Cell 3′ v3.1 kit (10X Genomics, Pleasanton, CA) (26). Briefly, nuclei were isolated using EZ Lysis Buffer (Sigma-Aldrich), homogenized by disposable Dounce homogenizer, washed, filtered using a 70-μm cell strainer. The nuclei pellet was resuspended in buffer containing DAPI (10 μg/ml), filtered using a 40-μm strainer, and immediately sorted using a BD FACSAria IIu (BD Biosciences, San Jose, CA) at the MSU Pharmacology and Toxicology Flow Cytometry Core (facs.iq.msu.edu/). Sequencing (150-bp paired end) was performed at a depth of 50 000 reads/cell using a NovaSeq6000 at Novogene (Beijing, China). CellRanger v3.0.2 (10× Genomics) was used to align reads to mouse gene models (mm10, release 93) including introns and exons to consider both pre-mRNA and mature mRNA gene models. Seurat was used to integrate and log-normalize expression data (27). The data is available on the Gene Expression Omnibus (GEO) at accession ID GSE184506 and R package versions are listed in Supplementary Table S1. Additional real datasets were publicly available. Hepatic whole-cell generated using the 10X Genomics platform was obtained from GEO (GSE129516) (5). Hepatic single-nuclei processed as the dose–response data for control and high dose TCDD treatment (0 and 30 μg/kg) was obtained from GEO (GSE148339). Peripheral blood mononuclear cell (PBMC) data also generated using the 10× Genomics platform and Seurat was obtained from the SeuratData R package (27).

Gene set enrichment analysis of experimental data was performed using the *fgsea v1.14* R package on gene lists sorted by significance values (e.g. *P*-value). Gene sets from BIOCARTA, KEGG, PANTHER and WIKIPATHWAYS were obtained from the Gene Set Knowledgebase (GSKB; http://ge-lab.org/gskb/) and filtered for gene sets containing 15–250 genes. Gene sets were agglomerated based on overlap of gene membership and only those showing ≥50% overlap were considered similar for subsequent network analyses. Visualization and calculation of measures of centrality were determined using *igraph v1.2.7*. Gene sets were considered enriched when adjusted *P*-value ≤0.05.

Dose–response data simulation

To simulate dose–response scRNAseq datasets we developed a wrapper for the Splatter R package (28). Splatter simulates counts using parameters estimated from real data to set the mean expressions, variance and outlier probability. Other parameters such as the number of cells, genes, probability of being differentially expressed, mean fold-change of DE genes (location) and standard deviation of

Table 1. Dose-response models for simulation of scRNAseq data

Model	Formula
Hill	$\mu(\text{dose}) = \gamma + (\text{v*dose}^n)/(k^n + \text{dose}^n)$
Exponential 2 or 3	$\mu(\text{dose}) = a\text{*exp}(\text{sign}(b\text{*dose}^d))$
Exponential 3 or 4	$\mu(\text{dose}) = a\text{*}(c-(c-1)\text{*exp}(\text{dose}^d))$
Power	$\mu(\text{dose}) = \gamma + \beta\text{dose}^\delta$

fold-change of DE genes (scale) were manually assigned to best reflect real data. The wrapper (SplattDR) leverages the group simulation feature of Splatter by applying a multiplicative factor estimated using dose–response models in Table 1 based on the US EPA Benchmark Dose Software (29). SplattDR R package is available at (https://github.com/zacharewskilab/splattdr) and R session information is listed in Supplementary Table S1.

Single cell RNA-seq hurdle model

We model the log-normalized gene expression matrix using a hurdle distribution wherein the rate of gene expression is assumed to follow a Bernoulli distribution and conditional on a cell expressing the gene, the log-normalized expression level is assumed to follow a Gaussian distribution (19). We denote $Y_{i,j}$ to be the log-normalized expression value of gene j in cell i, for $i = 1, \ldots n$ and $j = 1, \ldots p$. To characterize the bimodal properties of single cell data, for a given cell, a gene is defined to be either positively expressed or undetected. Define $R_{ij} = I[Y_{ij} > 0]$ to be the indicator variable denoting the presence or absence of expression for gene j in cell i. Following (19), the log-normalized gene expression values are modeled as follows:

$$\begin{aligned} Y_{i,j}\,|\,R_{i,j} &\sim \text{Normal}(\mu_j, \sigma_j^2), \\ Y_{i,j} &= 0 \text{ with probability 1 when } R_{i,j} = 0, \\ R_{i,j} &\sim \text{Bernoulli}(\omega_j), \end{aligned} \tag{1}$$

where μ_j and σ_j^2 denote the mean and variance of the gene expression level, conditional on the gene being expressed and ω_j denotes the rate of gene expression of gene j across all cells. Since the binary variable R_{ij} denotes the absence/presence of gene expression Y_{ij}, the Bernoulli distribution provides a natural representation as it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes-no question.

Hypothesis formulation. We now assume that our data has been collected under K conditions (doses), and denote the data by $D_{k,o} \equiv \{(Y_{k,i,j}, R_{k,i,j}), i = 1, \ldots, n_k\}$. The underlying populations for the sample data $D_{k,o}$ for $k = 1, 2, \ldots, K$, dose groups are assumed to be identified by the parameters $(\mu_{k,j}, \sigma_j^2, \omega_{k,j})$. The aim of this study is to test for difference in gene expression patterns between the different dose groups. Traditionally one would perform an ANOVA test to detect changes in mean across groups for samples with continuous measurements. However, to account for the bimodality in single cell gene expression distribution, the test should detect for changes in μ_j and ω_j simultaneously, as both could drive differential gene expression. Therefore we define,

$$\begin{aligned} H_0 :\ & \mu_{1,j} = \mu_{2,j} = \ldots \mu_{K,j} = \mu_j \text{ and} \\ & \omega_{1,j} = \omega_{2,j} = \ldots \omega_{K,j} = \omega_j. \end{aligned} \tag{2}$$

versus the alternative

$$\begin{aligned} H_a :\ & \mu_{k,j} \text{ is different for at least one } k \text{ and} \\ & \omega_{k,j} \text{ is different for at least one } k,\ k = 1, \ldots K \end{aligned}$$

Single cell Bayesian hurdle model analysis (scBT). Given the single cell RNA-seq hurdle model structure, we assume that *a priori*, given σ_j^2, $\mu_{k,j} \sim \text{Normal}(m_{k,0}, \tau_{k,\mu}\sigma_j^2)$, $\sigma_j^2 \sim IG(a_\sigma, b_\sigma)$, $\omega_{k,j} \sim \text{Beta}(a_{k,\omega}, b_{k,\omega})$, where IG is the inverse gamma distribution with shape a_σ and scale b_σ and $m_{k,0}$, $\tau_{k,\mu}$, a_σ, b_σ, $a_{k,\omega}$, $b_{k,\omega}$ are the hyperparameters. Given the large number of gene-wise model fits arising from a single cell experiment, there is a pressing need to allow for a parallel structure whereby the same model is fitted to each gene. The prior distributions on the parameters describe how the unknown coefficients $\mu_{k,j}$ $\omega_{k,j}$ and σ_j^2 vary across the genes and the dose groups while allowing for information borrowing between the genes. Now, based on the model assumptions, we propose a Bayesian test for simultaneously testing the differences in mean gene expression and dropout proportions as formulated in Equation (2). Under the null hypothesis the marginal likelihood is written as

$$\begin{aligned} \mathcal{L}_{H_0,j} = \int\int\int \prod_{k=1}^{K} \Bigg\{ \prod_{i=1}^{n_k} \Bigg[\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(Y_{k,i,j}-\mu_j)^2}{2\sigma_j^2} \right\} \\ \times \omega_j \Bigg]^{R_{k,i,j}} (1-\omega_j)^{1-R_{k,i,j}} \Bigg\} \times \pi(\mu_j|\sigma_j^2)\pi(\sigma_j^2)\pi(\omega_j) \\ \times d\mu_j d\sigma_j^2 d\omega_j. \end{aligned} \tag{3}$$

Under the alternative hypothesis we compute the marginal likelihood without any restriction on the K means $\mu_{k,j}$ and the dropout parameter $\omega_{k,j}$; $k = 1, 2, \ldots K$. Particularly, we assume that $\mu_{k,j} \sim \text{Normal}(m_{k,0}, \tau_{k,\mu}\sigma_j^2)$, and $\sigma_j^2 \sim IG(a_\sigma, b_\sigma)$, $\omega_{k,j} \sim \text{Beta}(a_{k,\omega}, b_{k,\omega})$; $k = 1, 2, \ldots K$. Now, the marginal likelihood under the alternative hypothesis is given by

$$\begin{aligned} \mathcal{L}_{H_a,j} = \int \cdots \int \Bigg\{ \prod_{k=1}^{K} \prod_{i=1}^{n_k} \Bigg[\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(Y_{k,i,j}-\mu_{k,j})^2}{2\sigma_j^2} \right\} \\ \times \omega_{k,j} \Bigg]^{R_{k,i,j}} (1-\omega_{k,j})^{1-R_{k,i,j}} \Bigg\} \times \prod_{k=1}^{K} \Big\{ \pi(\mu_{k,j})\pi(\omega_{k,j}) \Big\} \\ \times \pi(\sigma_j^2) \prod_{k=1}^{K} \Big\{ d\mu_{k,j} d\omega_{k,j} \Big\} d\sigma_j^2. \end{aligned} \tag{4}$$

The Bayes factor is then defined as

$$BF_{01,j} = \frac{\mathcal{L}_{H_0,j}}{\mathcal{L}_{H_a,j}} \times \frac{\pi(H_a)}{\pi(H_0)}, \tag{5}$$

where $\pi(H_a)$ and $\pi(H_0)$ are the prior probabilities for the alternative and null model, respectively. The hyperparameters are obtained by maximising the marginal likelihood un-

der the null and the alternative hypothesis. Detailed derivations of the likelihood function and the Bayes Factor are provided in Supplementary Material. Using the test of hypothesis described in Equation (2), scBT conducts a test of DE for each gene independently. To control for multiplicity we adopt the FDR correction approach discussed in (30). The rejection threshold is estimated in terms of the posterior probabilities of the null hypothesis, $p(H_{0,j}|D_j)$. For a target FDR α, the procedure rejects all hypotheses with $p(H_{0,j}|D_j) < \zeta$, where $p(H_{0,j}|D_j) = 1/(1 + 1/BF_{01,j})$ and ζ is the largest value such that $C(\zeta)/J(\zeta) \leq \alpha$ where, $J(\zeta) = \{j: p(H_{0,j}|D_j) \leq \zeta\}$ and $C(\zeta) = \sum_{j \in J(\zeta)} p(H_{o,j}|D_j)$.

Multiple group likelihood ratio test (LRT). To carry out a direct performance comparison with scBT, we extend the Seurat Bimod (19) for multiple groups. Assuming that all the K groups have the same variance σ_j^2 and omitting the index j for clarity, the likelihood ratio test can be defined as;

$$\Lambda(Y, R) = \frac{sup_{\theta \in H_0} L(\theta|Y, R)}{sup_{\theta \in H_a} L(\theta|Y, R)},$$

where the likelihood can be written as:

$$L(\theta|Y, R) = \prod_k \omega_k^{e_k}(1-\omega_k)^{n_k - e_k} \prod_{i \in C_k} f(Y_{i,k}|\mu_k, \sigma^2).$$

with Y and R representing the gene observation vector and the gene indicator vector across K dose groups, $\theta = \{\mu_k, \sigma^2, \pi_k, k = 1, \ldots, K\}$ is the vector of unknown parameters, C_k is the set of cells expressing the gene in group k (*i.e.* $C_k = \{i: R_{i,k} = 1\}$), $e_k = \sum_i R_{i,k}$ is the number of cells expressing the gene in group k and f is the density function of the normal distribution with parameters μ_k and σ^2. Following (19), it can be shown that the combined LRT is the product of a binomial and a normal LRT statistic, both of which can easily be derived using classical statistical theory. Applying classical asymptotic results about LRTs , $-2\log \Lambda(Y, R)$ converges to a χ^2 distribution with $(2K - 2)$ degrees of freedom under H_0. Detailed derivations of the test statistic can be found in the Supplementary Material.

Linear model-based likelihood ratio test (LRT linear). The generalized linear model approach MAST was identified as one of the top performing tests for pairwise differential expression testing (9,20). Deriving from their approach, the LRT multiple test is extended to a generalized linear model framework, where the mean and the dropout proportions are modelled as a linear function of the dose groups (assumed to be a continuous covariate). Using the same distributional assumptions defined in Equation (1) we fit a logistic regression model for the discrete variable R and a Gaussian linear model for the continuous variable Y conditional on $(R = 1)$ independently, as follows: $E(Y_{ij}|R_{ij} = 1) = m_{0,j} + m_{1,j}{*}d_i$ and $logit\{P(R_{ij} = 1)\} = \psi_{0,j} + \psi_{1,j}{*}d_i$, where d represents the continuous dose groups. Under this modelling approach, the null hypothesis described in Equation (2) can be rewritten as H_0: $E(Y_{ij}|R_{ij} = 1) = m_{0j}$ and $logit\{P(R_{i,j} = 1)\} = \psi_{0j}$. The regression models are fit using the *lm* and *brglm* functions in the *stats* and *brglm* R packages. The likelihood ratio test statistic is computed using the same statistical theory discussed for the LRT multiple test and it asymptotically follows a χ^2 distribution under H_0.

Benchmarking method selection

Our fit-for-purpose tests were benchmarked to existing differential expression testing methods or their multiple group equivalent based on previously reported performance, ability to consider multiple groups, or whether they served as foundation for the scBT and multiple group LRT (LRT multiple) tests developed here. Seurat Bimod served as foundation for the scBT and LRT multiple tests as previously outlined, and MAST was identified as one of the top performing test for two group comparisons (9). Similarly, limma-trend performed well for two sample comparisons and can consider multiple groups. The WRS test was identified as providing excellent balance between its ability to identify DE genes and speed, and is the default test for the Seurat R package for scRNAseq analysis. It was also reported that the *t*-test performed well and therefore we included the ANOVA and Kruskal-Wallis (KW) tests, a parametric and non-parametric alternative of the t-test for multiple group comparisons. All tests were run without correction for batch effects or other nuisance covariates. Multiplicity for each test was controlled using FDR correction. All tests, including scBT, LRT multiple and LRT Linear are available in our scBT R package (https://github.com/satabdisaha1288/scBT). R session information is listed Supplementary Table S1. A flow diagram outlines our benchmarking approach (Figure 1).

Seurat bimod. Seurat bimod test (19) is a pairwise differential gene expression testing approach developed assuming the single cell RNA-seq hurdle model framework. The test is formulated as H_0: the mean and the dropout parameters of the gene vector under two dose groups are equal versus H_a: the mean and the dropout parameters differ over the two groups. The LRT based test statistic $-2\log \Lambda(y, r)$ converges to a χ^2 distribution with 2 degrees of freedom under H_0. The computations are carried out using the R Package *Seurat*.

MAST. MAST (20) proposes a two-part generalized linear model for differential expression analysis of scRNAseq data. The first part models the rate of gene expression using logistic regression $logit(\omega_{ij}) = X_i\beta_j^{\omega}$ and the second part uses a linear model to express the positive gene-expression Y_{ij}, conditional on R_{ij} as $\mu_{ij} = X_i\beta_j^{\mu}$; where $\beta_j^{\omega} and \beta_j^{\mu}$ are the coefficients of the covariates used in the logistic and linear regression model respectively. A test with an asympotic χ^2 null distribution is employed for identifying DEGs and multiplicity is controlled using FDR correction (31). Despite the fact that LRT-linear and MAST have the same hurdle regression framework, the estimation process for the two methods has some significant differences. First, to achieve shrinkage of the continuous variance, MAST assumes a gamma prior distribution on the precision (inverse of variance) parameter and estimates its posterior maximum likelihood estimator (MLE) and uses that in place of the regular MLE of the precision parameter. Second, it fits a Bayesian logistic regression model for the discrete component by assuming Cauchy distribution priors centered at zero for the regression parameters. This is done to deal with cases of 'linear separation' where the parameter estimates diverge to

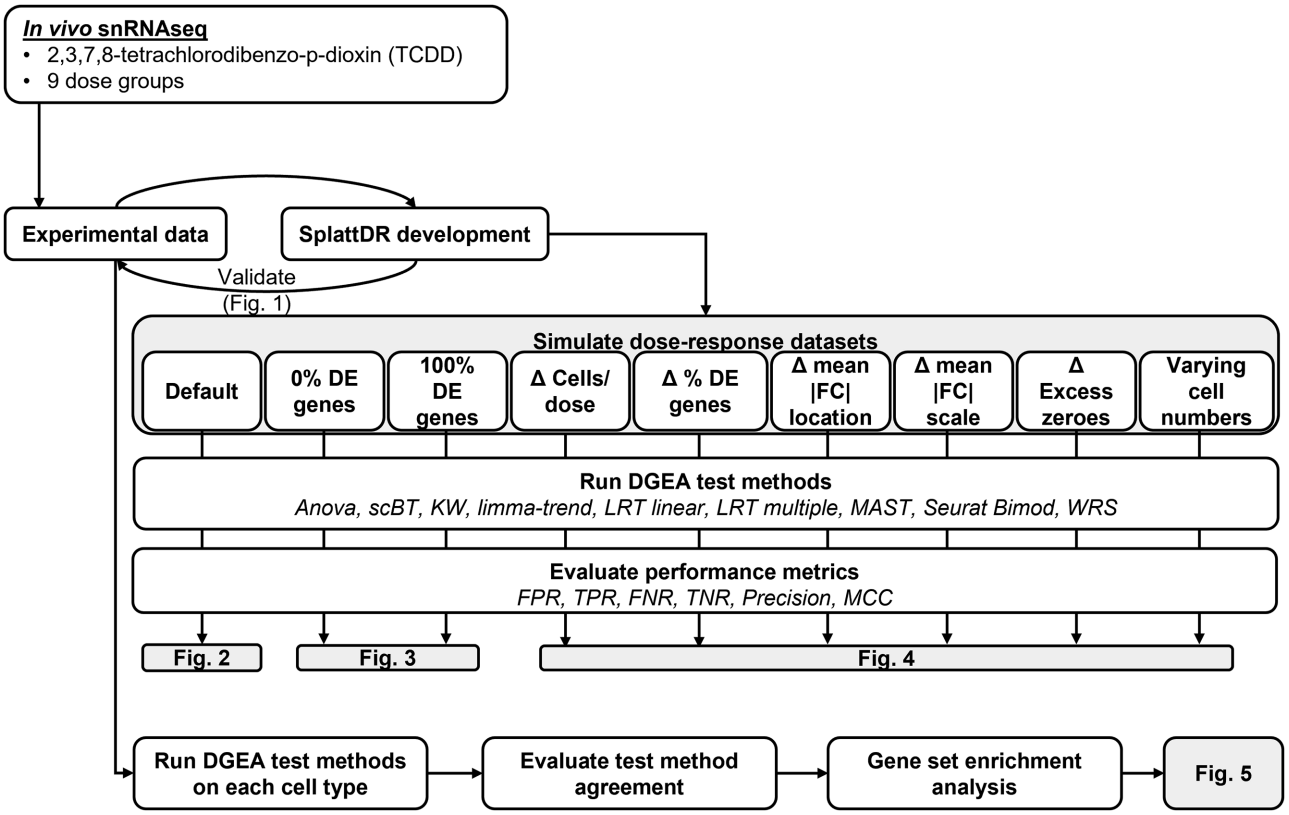

Figure 1. Flow diagram of the simulation, benchmarking, and experimental data evaluation strategy presented in the manuscript. Briefly, SplattDR was developed to simulate dose–response scRNAseq data and validated based on experimental dose–response data. Simulated datasets were generated varying diverse parameters 10 times and then used to assess the performance of each test method. Each test method was also assessed using experimental data from the hepatic snRNAseq dose response dataset obtained from male mice gavage every 4 days for 28 days with 0.01, 0.03, 0.1, 0.3, 1, 3, 10 or 30 μg/kg TCDD. Related figures for each analysis from the main body are noted.

$\pm\infty$ and the Fisher information matrix becomes singular. And finally, it considers the cellular detection rate defined as $CDR_i = \frac{1}{p}\sum_{j=1}^{P} R_{ij}$ to be a covariate in both the logistic and linear regression models. LRT linear on the other hand simply fits the non-Bayesian linear and the logistic regression models without considering variance shrinkage or adjustment for additional covariates.

limma-trend. Limma-trend (32) proposes a linear model based differential expression approach for modelling RNA-seq experiments of arbitrary complexity. Their framework models the mean gene expression as a function of several continuous and categorical covariates. A separate linear model is fitted for each gene, but the gene-wise models are linked by global parameters using the parametric empirical Bayes approaches (33). The global variance estimated by the empirical Bayes procedure also incorporates a mean variance trend, allowing better modelling of low abundance genes. Finally, test of differential gene expression is carried out by testing the significance of one or more coefficients of the fitted linear model.

Wilcoxon rank sum (WRS) test. WRS (34) test is a non-parametric test commonly used for pairwise DGE testing. The test is formulated as H_0: the distributions of the gene vector under two dose groups are equal versus H_a: the distributions are not equal.The test involves the calculation of the U statistic, which for large samples is approximately normally distributed. Since this is a pairwise test, a union is taken over all the genes found to be DE in each of the pairwise tests. The computations are carried out using the *wilcox.test* function in R package *stats* and multiplicity is controlled using FDR correction.

ANOVA. Analysis of variance (ANOVA) (35) is very commonly used for testing the differences among means in multiple groups. For a fixed gene j, it is assumed that the observed gene vector $y_{k,i,j}$ for cell i is grouped by dose. Assuming that $Y_{k,i,j} \sim Normal(\mu_{k,j}, \sigma_j^2)$, ANOVA aims to test the null hypothesis H_0: $\mu_{1,j} = \mu_{2,j} = \ldots \mu_{K,j} = \mu_j$ versus H_a: $\mu_{k,j}$, $i = 1, \ldots, n$; $j = 1, \ldots, p$; $k = 1, \ldots, K$ is different for at least one k. The test statistic is computed using the *aov* function in R package *stats* and it follows a F-distribution with $(K-1)$ and $(n-K)$ degrees of freedom. Multiplicity is controlled by applying FDR correction on the obtained P-values.

Kruskal–Wallis (KW) test. KW (36) test extends the WRS test for multiple groups. It is also a non-parametric extension of the ANOVA test. The test is formulated as; H_0: the distributions of the gene vector under K dose groups are equal versus H_a: the distributions are not equal. The com-

putation of the KW test statistic is carried out using the *kruskal.test* function in R-package *stats* and it asymptotically follows a χ^2 distribution with $K-1$ degrees of freedom. Multiplicity is controlled by applying FDR correction on the obtained *P*-values.

Benchmarking and sensitivity analyses

Benchmarking of DE test methods was performed on simulated datasets based on initial parameters derived from real dose–response snRNAseq data. The probability of differential expression was set to 10% with a 50% probability of being down-regulated, equally distributed among the dose–response models in Table 1. Batch parameters were used to include sample variation associated with data obtained from 3 individuals in each dose group. A total of 5000 genes were simulated for 4500 cells (500 per dose group) using the same doses as the real dataset. Sensitivity analyses varied each of the following parameters according to values in Supplementary Tables S2 and S3: cell abundance equally distributed among dose groups, varying cell numbers in each dose group, percent DE genes, proportion of downregulated DE genes, fold-change location or scale, and dropout rate. Each simulation was replicated 10 times using a different initial seed. Seeds and parameters listed in Supplementary Table S2 can be used to reproduce datasets under each condition. Method concordance was determined as area under the concordance curve (AUCC) for the top 100- or 500-ranked genes in simulated and real datasets, respectively, as previously described (37).

RESULTS

Dose-response single-cell data simulations

For benchmarking of DGEA methods, a ground truth is required. Existing simulation tools such as PowSimR, SymSim, SPsimSeq and Splatter are commonly used for power analyses, evaluating DE analysis methods, and testing cell clustering strategies (28,38–40). Tools such as SymSim and Splatter are also capable of simulating cell trajectories and model differentiation processes. Trajectories which exhibit non-linear changes over time or across different developmental stages are not unlike dose–response effects which change over a continuum of doses. However, dose-responsive changes commonly follow defined trajectories such as Hill, exponential, power, and linear models (29). To simulate dose–response scRNAseq data we developed a wrapper for the Splatter scRNAseq data simulation tool named SplattDR. SplattDR modified the Splatter grouped data simulation strategy by adjusting counts from means defined by one of the dose–response functions outlined in the Materials and Methods.

To demonstrate the modeling capability of SplattDR, 10 000 gene expression responses were simulated with a 10% probability of being differentially expressed, equally distributed across the dose–response models. Parameters used in Splatter were initially estimated from our experimental single nuclei RNAseq (snRNAseq) dose–response dataset. The simulated data compared to the experimental data showed the relationship between the mean expression, percentage of zeroes, and mean variance were consistent (Figures 2A, B). Estimation of the normalized root mean square deviation (NRMSD) from a curve fit to the experimental data indicated excellent concordance. This strong concordance was also maintained within distinct dose groups (Supplementary Figures S1 and S2). The distribution of log(fold-changes) between vehicle (dose 0) and the highest simulated dose (dose 9; 30 μg/kg) showed a more even distribution within a similar range compared to experimental data which was skewed towards induction (Figure 2C). However, the gene induction skew was captured by modulating the parameters affecting the probability of differential expression and the proportion of differentially repressed genes (Supplementary Figure S3). Principal components analysis (PCA) of the simulated data clearly showed the dose-dependent characteristics of scRNAseq data with distinct clusters increasing in separation with increasing dose (Figure 2D) which was also resolved by PCA within the experimental data (Supplementary Figure S4).

To our knowledge, no other published *in-vivo* dose–response scRNAseq datasets are available limiting the number of datasets to estimate initial parameters for simulation to date. To investigate whether existing datasets generated using a different study design (e.g. whole cells or different tissue source) could be used to derive initial parameters, we also simulated 10 000 genes starting with parameters estimated from (i) a two-dose liver snRNAseq (GSE148339), (ii) whole cell liver scRNAseq (GSE129516) and (iii) peripheral blood mononuclear cells (PBMC; GSE108313) datasets. When compared to a model fit for experimental data to determine the relation between mean expression and percent zeroes or mean variance, the NRMSD for data simulated from these datasets were between 1 and 10% with data simulated from whole cell data differing the most from the model fit (Figure 2E). We then explored whether parameters estimated from distinct cell types could replicate the characteristics of that same cell type (Figure 2F). Not surprisingly, using initial parameters derived from individual cell types in the experimental dose–response data had lower NRMSD than those derived from the whole cell dataset. Notably, when data derived from a lower abundant cell subtype was used to estimate starting parameters, the dose–response characteristics for that cell subtype was also poorly modeled (Figures 2E, F, S1–2).

Performance accuracy of DE test methods

We evaluated the performance of several differential gene expression analysis methods on simulated datasets consisting of nine dose groups of 500 cells each (4500 total) and 5000 genes with a 10% probability of being differentially expressed (500 differentially expressed genes). Selection criteria for test inclusion are outlined in the Materials and Methods section and included 9 test methods; ANOVA (35), single-cell Bayes hurdle model test (scBT), Kruskall–Wallis (KW) (36), limma-trend (32,33), likelihood-ratio test (LRT) linear and multiple, MAST (20), Seurat bimod (19) and WRS (34). With ground truth from simulated data, the sensitivity, specificity, and precision for each test method was computed. Area under the receiver-operating characteristic curve (AUROC) was used to measure test performance for correctly classified differentially expressed genes.

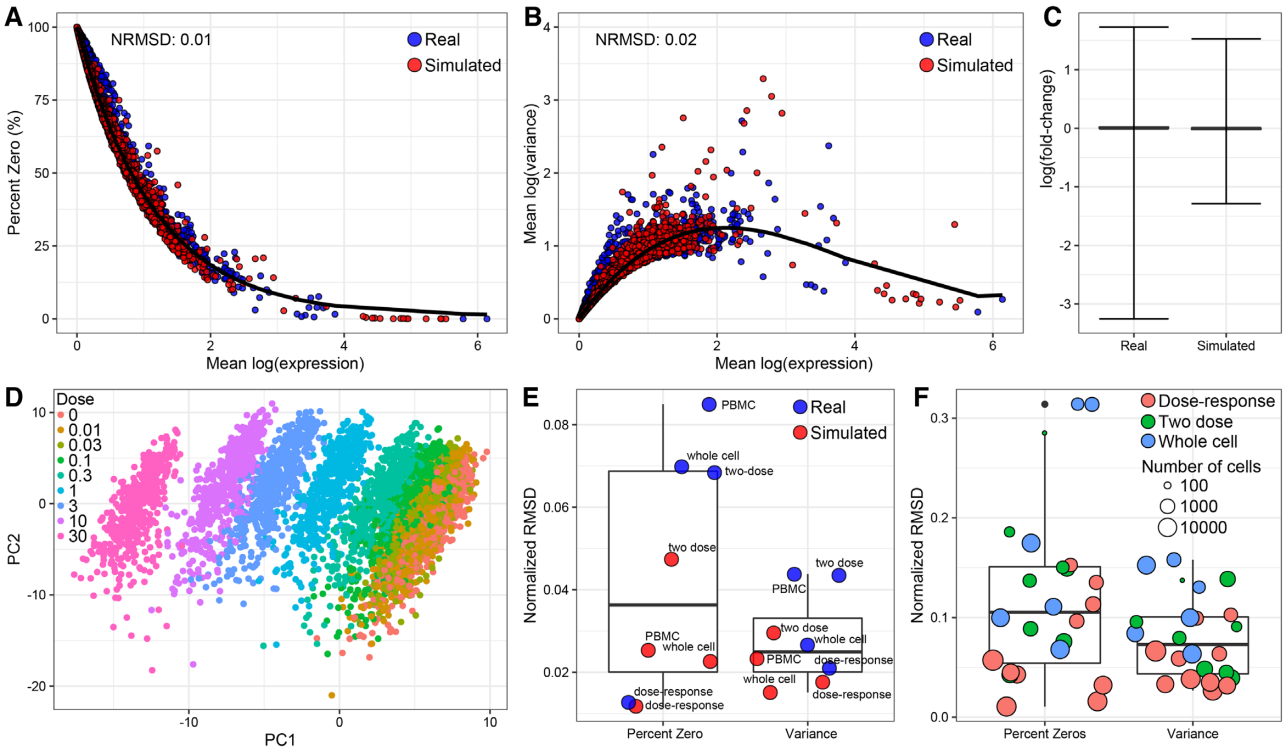

Figure 2. Comparison of simulated and real dose–response data. (**A**) Relationship between gene-wise mean expression and percent zeroes for simulated and real dose–response data. Simulation data consisted of 10 000 genes and nine dose groups based on parameters derived from experimental dose–response snRNAseq data. Black line represents a fitted model to the experimental data from which the normalized root mean square deviation (NRMSD) of simulated data was determined. (**B**) Relationship between gene-wise mean expression and variance for simulated and experimental data. NMRSD was calculated for simulated data from the fitted model represented as a black line. (**C**) Distribution of log(fold-changes) in experimental and simulated data showing the median and minimum and maximum values. (**D**) Principal components analysis of simulated data colored according to simulated dose groups. (**E**) NMRSD estimated relative to fitted model in A,B for simulated data generated from initial parameters derived from published hepatic scRNAseq (two dose; GSE148339), hepatic whole cell (whole cell; GSE129516), and peripheral blood mononuclear cell (PBMC; GSE108313) datasets. (**F**) NMRSD estimated relative to model fitted to cell-type specific experimental dose–response data when simulated from initial parameters estimated from that same cell type. Box and whisker plots show median NMRSD, 25th and 75th percentiles, and minimum and maximum values.

In unfiltered data, AUROC scores showed similar performance for most tests except scBT which had the largest AUROC among all test methods (Figure 3A). To account for the inherent class imbalance between differentially expressed and non-differentially expressed classes the area under the precision-recall curves (AUPRC) was also calculated. Similar to AUROCs, AUPRCs identified scBT as the best performing test (Figure 3C). In most standard differential expression testing pipelines genes expressed at low levels are removed to minimize false detection rates. Following filtering of genes expressed in $\geq$5% of cells in any dose group, scBT was consistently ranked as the best test based on AUROC and AUPRC scores. The performance of LRT linear test also improved, with comparable AUROC and AUPRC scores relative to scBT, suggesting LRT linear is poorly suited for genes expressed at low levels (Figures 3B–D).

AUROC and AUPRC reflect the performance of each test method with varying significance (i.e. *P*-value) thresholds. In the standard pipeline a fixed threshold is used, typically a *P*-value $\geq$0.05 after adjustment for multiple hypothesis testing (i.e. Bonferroni correction). For each method except scBT, the performance at an adjusted *P*-value $\geq$0.05 significance criteria was evaluated. In scBT analysis, a gene was considered differentially expressed when the estimated posterior probabilities of the null hypothesis, $p(H_{0,j}|D_j)$, was less than ζ, where the ζ value was chosen to achieve a target FDR of 0.05. scBT significantly outperformed all other tests in precision rates irrespective of low expression filtering (Figures 3E, S5). However, scBT was less effective in identifying true positives (Figures 3F, S5). Applying the filtering criteria improved the recall rates, but the precision rates remain largely unchanged (Figure 3E, F). Test method classification performance scores were estimated as the Matthews correlation coefficient (MCC) which is well suited for unbalanced data (41). We see that the scBT and LRT linear tests performed best for this metric on both unfiltered and filtered data (Figure 3G).

Type I error control and power

To investigate test performance in controlling type I errors (false positives), DGEA methods on simulated datasets were examined with 0% DE genes (i.e. negative control). Using the threshold for the computed posterior null probabilities, scBT identified only one false positive gene in 2 of 10 simulations (Figure 4A). ANOVA, scBT, KW, limma-trend and LRT linear had false positive rates (FPRs) below 3%

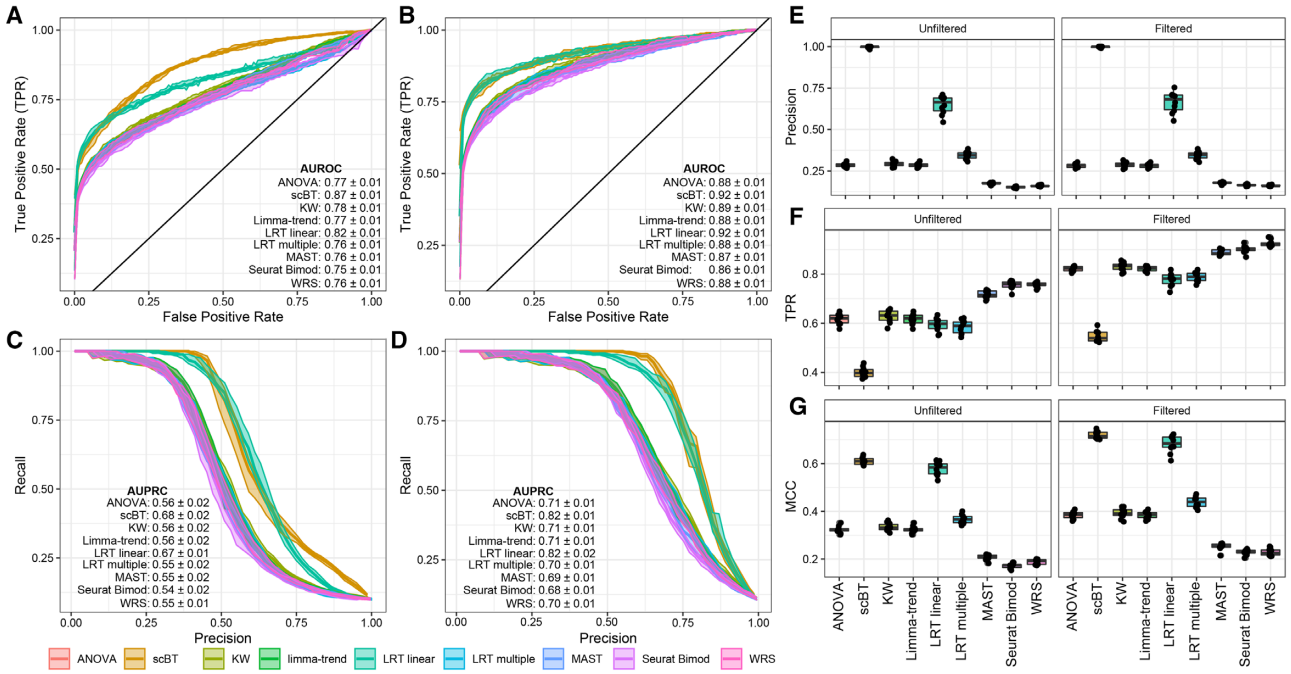

Figure 3. Classification performed of DE analysis tests. (**A**) ROCs estimated from simulated dose–response scRNAseq data for nine DE test methods including all genes expressed in at least one cell (unfiltered). (**B**) ROCs for nine DE test methods after filtering simulated dose–response scRNAseq data for genes expressed in only ≥5% of cells (low levels) in at least one dose group. (**C**) Precision-recall curves (PRCs) for nine DE test methods on unfiltered simulated dose–response scRNAseq data. (**D**) PRCs for nine DE test methods on filtered simulated dose–response scRNAseq data. Lines represent the mean values and shaded region reflects the standard deviation for 10 independent simulations. (**E**) Precision of DE test methods. (**F**) FPR of DE test methods. (**G**) MCC for test methods. (E–G) Box and whisker plots median values, 25th and 75th percentiles, and minimum and maximum values for 10 independent simulations. Points reflects values for each independent simulation. Panels display comparisons of unfiltered and filtered datasets.

indicating better performance compared to two group tests. After filtering for genes with low expression levels, scBT still correctly identified all the non-differentially expressed genes and was the best performing test. These are the same tests that had a better FPR control in initial simulations (Figure 3). To explore whether mean expression or percentage of zeroes influenced type I error rates, a logistic regression model was fit to negative control data. We predicted the probability for each gene to be identified as differentially expressed in the negative control data. While the curve for scBT is missing since few false positives were identified, the predicted FPR for all the other tests except LRT linear were also high for highly expressed genes with few zeroes (Figures 4B, C). Next, a positive control dataset with 100% differentially expressed genes was simulated to evaluate test performance for detecting true positives. All tests except scBT exhibited a false negative rate (FNR) ≥40% (Figure 4D). The best performing tests for FNR also had high FPR. Logistic model regression fitting for false negative classification of genes shows that the false negative rates were highest when the mean expression was either too high or too low for all tests (Figures 4E, F).

Parameter Sensitivity Analysis

Experimental scRNAseq datasets will vary between cell types, cell composition, and responses depending on the target tissue, treatment, number of cells sequenced, and more. For example, some distinct cell types are very abundant (e.g. hepatocytes), with others present at lower levels (e.g. portal fibroblasts) in hepatic scRNAseq datasets. Moreover, treatments such as exposure to a xenobiotic, can elicit dose-dependent changes in relative proportions of cell types such as the infiltration of immune cells (26). We investigated the impact by changing cell abundance from 25 to 2000 cells per dose group and observed an increase in the false positive rate (FPR) when increasing the number of cells (Supplementary Figure S6). The scBT and LRT linear tests were less sensitive to an increase in the FPR as cell abundance increased while the total positive rates (TPR + FPR) increased with cell abundance for all methods. Although all tests exhibited comparable performance at low cell numbers (≥500), as cell numbers increased scBT outperformed all other tests in both precision and MCC score (Figures 5A, S6). Comparison of AUROCs and AUPRCs across cell numbers showed that ANOVA, KW, limma-trend, and LRT linear tests performed best for a small number of cells, but the increase in AUROC was steeper for scBT (Supplementary Figures S7 and S8).

It was also evident from the experimental snRNAseq dataset that the number of cells per dose group was not fixed. We evaluated the performance of the test methods when the number of cells dose-dependently increased or decreased, and when the number of cells per dose group were taken from experimental data. Notably, while scBT had the best MCC for increasing number of cells per dose, LRT linear performed better than scBT when the number of cells decreased before and after filtering for genes expressed at low levels (Figure 5B). The shift in MCC between increasing and decreasing cell numbers for scBT appears to be driven

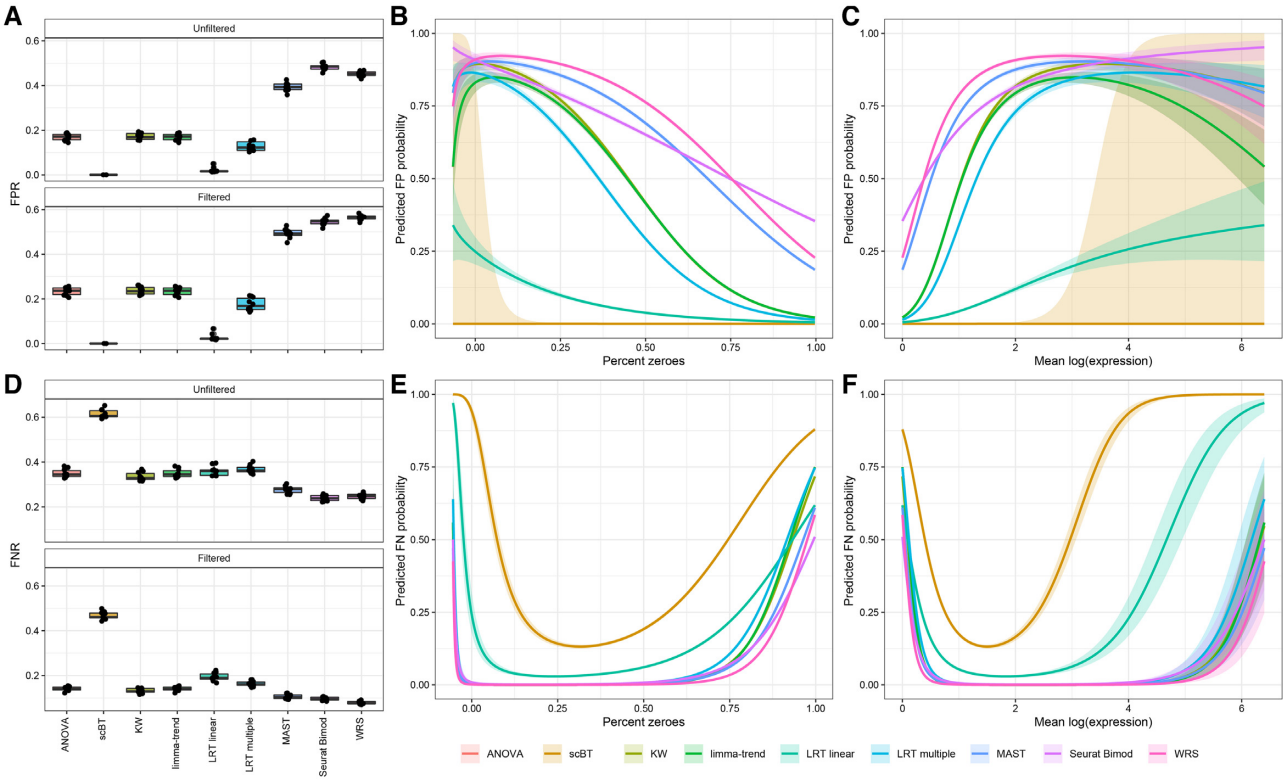

Figure 4. Evaluation of Type I and II error control. (**A**) False positive rate (FPR) of 9 differential expression test methods estimated from negative control (0% DE genes) simulated dose–response scRNAseq data including all genes expressed in at least 1 cell (unfiltered) and genes expressed in only ≥ 5% of cells in at least one dose group (filtered). (**B**, **C**) Logistic regression models were fitted to negative control data to predict the probability of false positive identification using percent zeroes and mean expression as covariates. Lines represent the predicted probability of false positive classification with the shaded region representing the 95% confidence interval. (**D**) False negative rate (FNR) of nine differential expression test methods estimated from positive control (100% DE genes) simulated dose–response scRNAseq data including unfiltered and filtered datasets. (**E**, **F**) Logistic regression models were fit to positive control data. Lines represent predicted probability of false negative classification with shaded region representing the 95% confidence interval.

by a concomitant decrease in TPR and increase in FNR (Supplementary Figures S9 and S10).

Unique chemical, drug, environmental contaminant, and natural product classes elicit distinct differential gene expression profiles defined by the MoA as well as by their metabolism, potency (sensitivity) and efficacy (maximal response). Differences between compound classes are reflected in the gene expression profile in (a) the proportion of differentially expressed genes, (b) the number of induced/repressed genes, (c) the mean fold-change for differentially expressed genes and (d) the distribution of fold-change for differentially expressed genes. These four parameters were modulated in simulated data to determine the effect of the percentage of differentially expressed genes (Supplementary Figures S11–S16), the fold-change distribution (*aka* scale; Supplementary Figures S17–S19), and the mean fold-change (*aka* location; Supplementary Figures S20–S22) on test performance. Among these scenarios, changing the proportion of repressed genes had little to no impact on test method performance (Figures 5C–F, S14).

Increasing the proportion of differentially expressed genes led to an improvement in MCC except for scBT and LRT linear, though these tests maintained the top MCC scores as well as AUROC and AUPRC (Figures 5C, S11–S13). As the magnitude of the effect increased, LRT linear performed best at the low end while scBT exhibited the greatest improvement in MCC (Figure 5D). Conversely, while the MCC decreased for most tests when modulating the fold-change scale of differentially expressed genes, scBT improved and was more stable (Figures 5E, S17–S19). As the proportion of unexpressed genes increased, the FPR increased with precision decreasing for all tests (Supplementary Figure S23). However, scBT was least affected, and maintained the highest MCC among all tests (Figure 5F). AUROC and AUPRC values also indicated that scBT consistently outperformed other test methods (Supplementary Figures S24–S25).

Test method agreement

To assess agreement between tests, the area under the concordance curve (AUCC) for each pair of tests for the top 100 genes ranked by adjusted *P*-value was calculated as previously described (9,37). All methods showed excellent concordance (AUCC ≥ 0.77) with LRT linear showing the poorest consistency compared to all other tests while the limma-trend and ANOVA tests showed perfect agreement with an AUCC of 1 (Supplementary Figure S5). Pairwise differential gene expression comparisons between Seurat Bimod, MAST and WRS had AUCC >0.95 AUCCs while the multiple group tests ANOVA, LRT multiple, KW, and scBT clustered together with AUCC ranging between 0.9

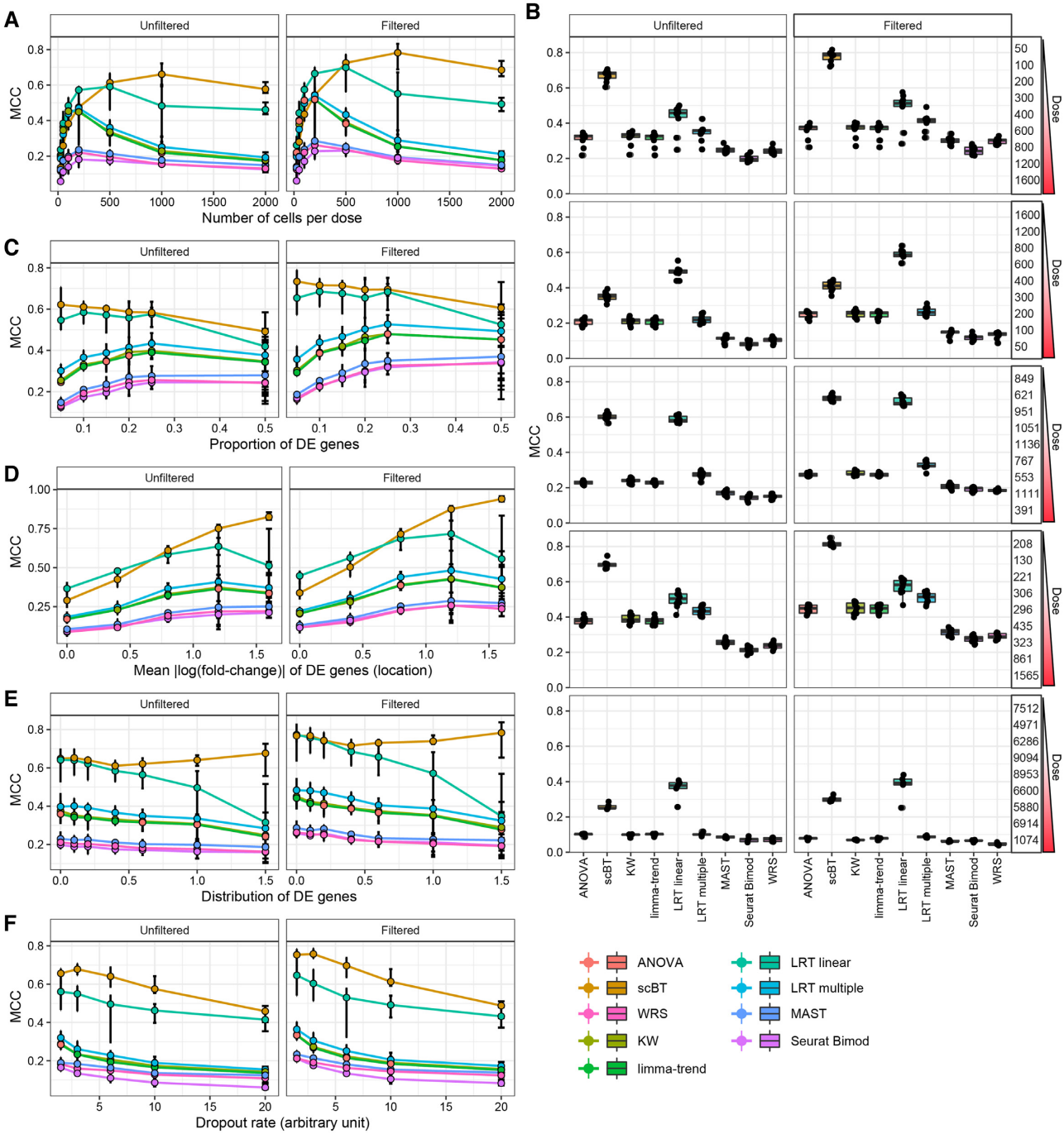

Figure 5. Matthews correlation coefficient (MCC) from sensitivity analyses of differential expression test methods. (**A**) MCC for nine DGEA test methods determined from simulated dose response data with varying number of cells per dose group. Simulations consisted of 5,000 genes with a probability of differential expression of 10% and 9 dose groups. (**B**) MCC for simulated data varying the cells numbers by dose group. The number of cells in each of the nine doses groups is shown on the right. (**C**) MCC for varying proportion of differentially expressed genes. (**D**) MCC when varying the mean fold-change (location) of repressed differentially expressed genes. (**E**) MCC for varying distribution of fold-change (scale) of differentially expressed genes. (**F**) MCC for varying dropout rates calculated as in Supplementary Table S3. Points represent median and error bars represent minimum to maximum values. Boxplots represent median, 25th to 75th percentile, and minimum to maximum values. Each analysis consisted of 10 replicate datasets including all genes expressed in at least one cell (unfiltered) and genes expressed in ≥5% of cells in at least one dose group (filtered).

and 1. In the absence of nuisance covariates, MAST and Seurat Bimod provided similar results, as expected given their similar mixture normal model structure. Likewise for ANOVA and limma-trend, both of which rely on normality assumptions for testing differential gene expression.

Real dose–response dataset DE analysis

Without ground truth for experimental data, the performance of the differential expression test methods was examined by first evaluating the agreement for each identified cell type (Figures 6, S26). Genes in the experimental dataset were considered differentially expressed when expressed in ≥5% of cells in at least one dose group and had a |fold-change| ≥ 1.5. In hepatocytes, the most abundant cell type, fewer than 5 genes were not detected in all test methods, with the majority missed by the WRS test (Figure 6A). Upon closer examination, those genes were not expressed in control hepatocytes. Not surprisingly, for all cell types, the largest intersection was between all tests indicating strong agreement within all test methods. Only a few tests identified a subset of unique genes as differentially expressed, which accounted for a very small fraction. For example, LRT linear identified 12 unique differentially expressed genes in portal fibroblasts, one of the least abundant cell types (Figure 6B). LRT linear was the best performing test for low cell numbers indicating that the 12 unique differentially expressed genes may in fact be true positives. Consistent with simulations of varying cell numbers (Figure 5A), 24 genes were not identified as differentially expressed by the scBT method for stellate cells which exhibit a dose-dependent decrease in numbers (Figures 6C, D). Although scBT outperformed other tests in most scenarios, it under performed in this scenario. Nevertheless, when ranking genes by significance level (i.e. *P*-values), AUCC were high for all pairwise comparisons.

To explore the biological insight gained by using the test methods, gene set enrichment analysis was performed by ranking genes following significance values (adjusted *P*-value or Bayes factor) on gene sets from BIOCARTA, KEGG, PANTHER and WIKIPATHWAYS. Gene sets were grouped based on their similarity in gene membership into a network for which centrality measures can be estimated. An examination of portal fibroblasts, which exhibited the most disagreement among test methods (Figure 6B), showed that multiple group test methods, particularly scBT had improved centrality metrics (centrality – number of edges; closeness – steps required to access other nodes; and betweenness – number of paths that go through a node) (Figure 6E). Visualization of significantly enriched terms identified enriched functions associated with growth factor and immune cell signaling in addition to expected terms such as xenobiotic metabolism and nuclear receptors involved in lipid metabolism (Figure 6F). Alternatively, WRS which did not find as many connected groups of functions, was largely limited to those identified by scBT except for the hormone signaling and tryptophan metabolism clusters (Supplementary Figure S27). While there is no ground truth from real data, greater agreement between similar gene sets from disparate sources (Supplementary Figure S28–S30) suggests that multiple group tests such as scBT provide more reliable findings. However, all the test methods produce comparable gene set enrichment results as expected since the most robust changes were identified by all the test methods.

DISCUSSION

The goal of this study was to compare the performance of newly developed DGEA test methods for dose–response experiments to existing analysis methods. Using simulated data to generate ground truth, we evaluated the performance of nine differential expression testing methods which were broadly classified as either fit-for-purpose, multiple group, or two group tests. Criteria for test method selection was based on previous benchmarking efforts for two group study designs identifying MAST, limma-trend, WRS, and t-test as the best performers (9,42). ANOVA and KW tests were also included for evaluating multiple group comparisons, and Seurat Bimod, for having the same modelling framework as scBT, LRT multiple and LRT linear tests. The test methods were ranked from best to worse (1–9) based on type I error rate, type II error rate, MCC, AUROC and AUPRC (Figure 7, Supplementary Table S4).

While several scRNAseq tools have been developed (28,38–40), none are developed to simulate dose–response models commonly identified in toxicological and pharmacological datasets (29,43). Our SplattDR wrapper for the Splatter package (28) was able to show that simulated data can effectively emulate key experimental scRNAseq data characteristics when simulation parameters were estimated from various Unique Molecular Identifier (UMI)-based datasets. In agreement with a previous report, technical and biological factors, such as cell type, does appear to influence gene dropout rates (18). We primarily focused on 10× Genomics UMI data given the unavailability of real experimental dose–response data generated using other platforms.

Overall, test method performance was consistent with their intended application. For example, fit-for-purpose tests scBT and LRT linear consistently ranked higher followed by multiple groups tests such as KW and LRT multiple. scBT exhibited the best overall performance with excellent FPR control and top ranked MCC while LRT linear struck a balance between type I and type II error rates. The scBT results are not surprising as Bayes factor-based tests have proven to be conservative and consequently more appropriate when false positives are of concern (22,23). In the context of investigating chemical or drug MoAs, false positives have the potential to lead to wasted effort and resources in attempts to validation and support findings (44). Moreover, when assessing a large number of genes, a 5% FP rate (*P*-value ≥ 0.05) can result in hundreds of FPs that skew MoA classifications (17).

A single test method was not expected to outperform all other tests under all conditions as previously demonstrated when comparing pairwise testing (6,9,42). Therefore, we assessed the strengths and limitations of each test method by varying parameters likely to change within and across various experimental datasets. The number and relative abundance of cell types is known to be affected by disease or treatment, and the distribution of differential expression influenced by the chemical, drug, or food contam-

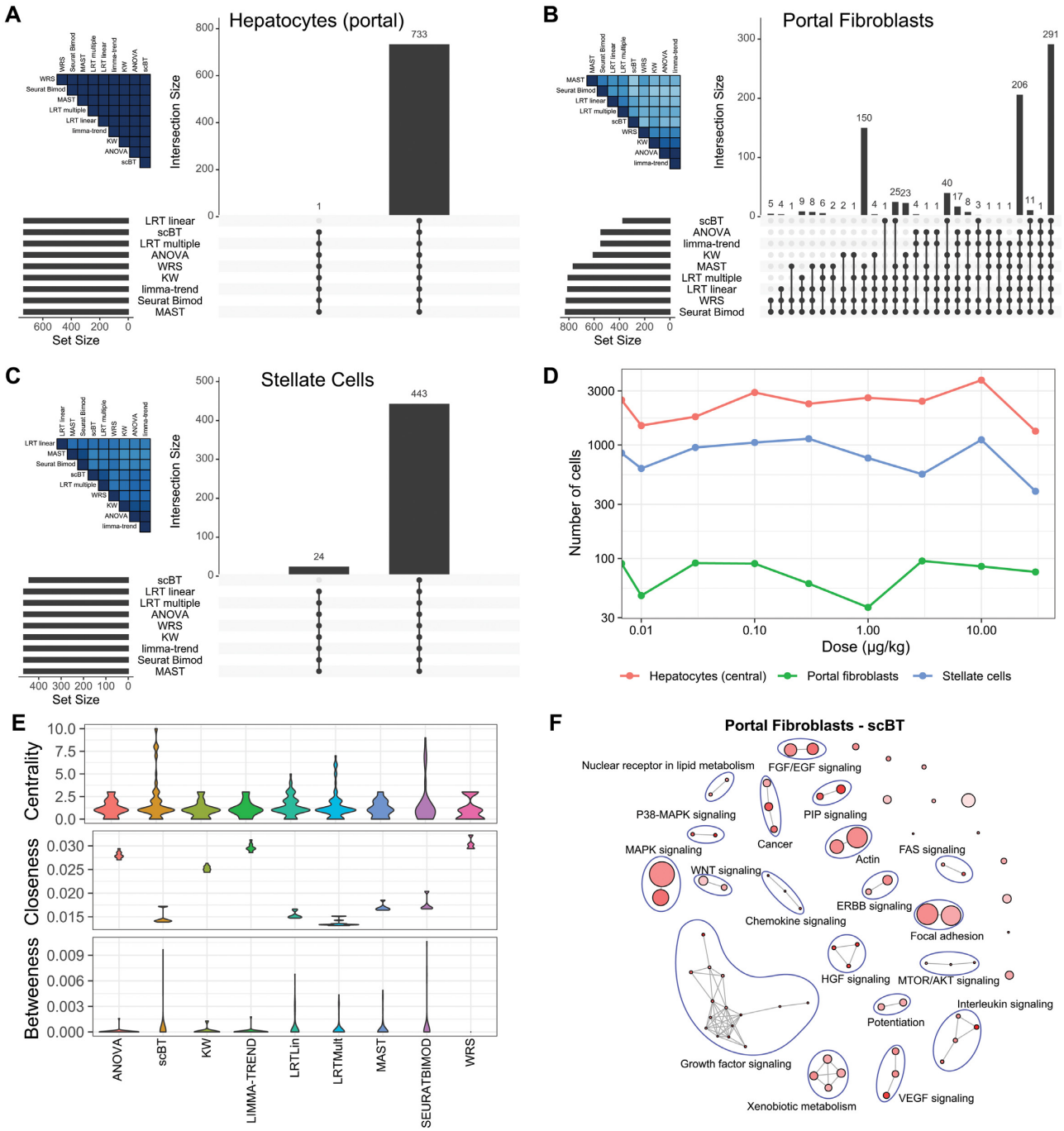

Figure 6. Agreement of differential expression test methods on experimental dose–response data. (**A**) Upset plot showing the intersection size of genes identified as differentially expressed by nine different test methods in hepatocytes from the portal region of the liver lobule. (**B**) Intersect of differentially expressed genes in portal fibroblasts. (**C**) Intersect size in hepatic stellate cells. Vertical bars represent the intersect size for test methods denoted by a black dot. Horizontal bars show the total number of differentially expressed genes identified within each test (set sizes). Only intersects for which genes were identified are shown. Genes were considered differentially expressed when (i) expressed in >5% of cells within any given dose group and (ii) exhibit a |fold-change| ≥ 1.5. A heatmap in the upper left corner of each panel shows the pairwise AUCC comparisons for the 500 lowest *P*-values. (**D**) Relative proportion of cell types identified in each dose group of the real dataset for the cell types in (A–C). Experimental snRNAseq data was obtained from male mice gavaged with sesame oil vehicle (vehicle control) or 0.01–30 μg/kg TCDD every 4 days for 28 days. (**E**) Graph metrics for gene set enrichment analysis of portal fibroblasts grouped by similarity in gene membership. Violin plots show distribution of node-wise values for each test method. (**F**) Network visualization of significantly enriched (adjusted *P*-value ≤ 0.05) gene sets using the Bayes factor ranked genes of portal fibroblasts. Groups of ≥2 nodes were manually annotated following commonality in the gene set names. Each node represents a gene set with the size of the node representing the number of genes in a gene set, and edges connect nodes with ≥50% overlap.

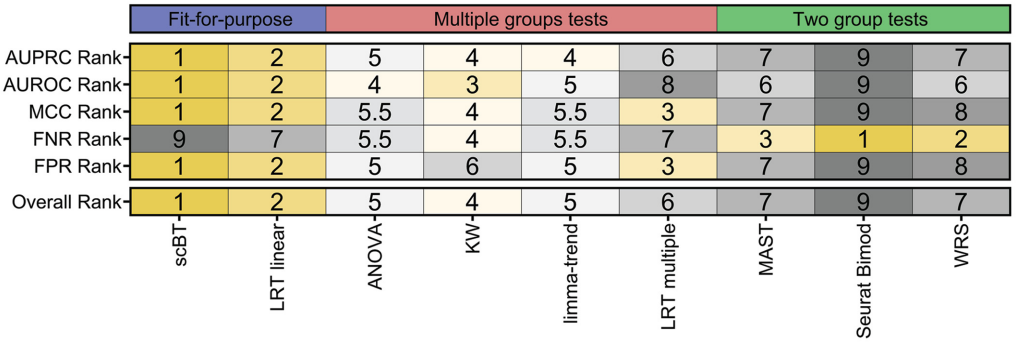

Figure 7. Median ranking of differential expression test methods across all simulations. The median rank of each test method was calculated for AUPRC, AUROC, MCC, FNR and FPR. Tests were grouped according to intended application including fit-for-purpose tests developed for the analysis of dose–response datasets, multiple group tests, and two group tests. The overall rank represents the median value for the five key metrics presented here.

inant being evaluated (5,26). scBT consistently ranked at the top under most scenarios, particularly when the mean and standard deviation of the fold-change for differentially expressed genes varied. However, scBT under performed in MCC when the number of cells decrease in a dose-dependent manner which would be expected in treatments which alter cell population sizes (e.g. inflammation). Under these circumstances LRT linear outperformed all other tests with scBT performing similar to the other test methods as evident when 24 differentially expressed genes were not identified by scBT within experimental data for stellate cells which experienced a dose-dependent decrease in relative abundance following TCDD treatment. Although excluding genes expressed at low levels generally improved the performance of all test methods, the comparative performance of test methods did not significantly change in most cases. We did not have access to experimental scRNAseq dose–response data, however, we expect that the scBT would perform equally well as with experimental snRNAseq data as the elevated number of zeroes are common to both types of data. Major differences between these types of data are (i) biases in gene detection and (ii) overall counts (26). Given the higher overall counts in scRNAseq data, test method such as scBT may even perform better.

DGEA provides biological information regarding the effects of exposure to chemicals, drugs, and food contaminants. As expected, gene set enrichment analyses did not dramatically differ in the enriched pathways which are driven by the most robust responses such as xenobiotic metabolism. However, when integrating gene sets from disparate sources we found gene sets that partially overlap in gene membership were consistently identified by multiple group test methods. For example, several gene sets related to growth factors and cell proliferation were identified by scBT but not WRS. Portal fibroblasts are implicated in proliferation of cholangiocytes and the secretion of growth factors during development. Enrichment of these terms suggests a functional role consistent with the induction of bile duct proliferation by TCDD (45,46). In contrast, WRS identified enrichment associated with tryptophan as well as oxytocin/thyrotropin-releasing-hormone pathways which has not been linked to the effects of TCDD on portal fibroblasts. Although ground truth for the complete experimental dataset is not available, the use of test methods such as scBT reduce experimental noise to identify leads warranting further analysis.

CONCLUSION

Collectively, our findings suggest that scBT and LRT linear fit-for-purpose tests are better suited for the differential expression analysis of dose–response studies and when false positives are of greater concern than false negatives. Moreover, consistent with previous benchmarking efforts, we show that common non-parametric tests such as KW outperform test methods developed for scRNAseq data when the study involves comparisons between multiple groups. Ultimately, each test method performs optimally under diverse scenarios. While the importance of controlling type I error rates is acknowledged, a balance must be struck with type II error rates. The tradeoff should be determined based on the individual research question being investigated. It may even be reasonable to apply different test methods to distinct cell types based on dropout rates, cell abundance, and changes in relative cell proportions given the strengths and weaknesses of each test method.

DATA AVAILABILITY

New dose-response single-nuclei RNA sequencing data has been deposited in the Gene Expression Omnibus (GEO) under the accession ID GSE184506. Publicly available single-nuclei and single-cell transcriptomic datasets were obtained from GEO under the accession IDs GSE108313 (PBMC), GSE129516 (hepatic whole cell), and GSE148339 (two dose hepatic single-nuclei). Simulated data was produced using our SplattDR R package can be reproduced using parameters from Table S2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Human Genome Research Institute [R21 HG010789 to T.R.Z., S.B.]; National Institutes of Environmental Health Sciences Superfund Research Program

[P42 ES004911 to T.R.Z.]; T.R. Zacharewski and S. Bhattacharya are partially supported by AgBioResearch at Michigan State University; T. Maiti and S. Saha are partially supported by NSF [DMS 1945824]. Funding for open access charge: National Human Genome Research Institute [R21 HG010789].
Conflict of interest statement. None declared.

REFERENCES

1. Halpern,K.B., Shenhav,R., Matcovitch-Natan,O., Toth,B., Lemze,D., Golan,M., Massasa,E.E., Baydatch,S., Landen,S., Moor,A.E. *et al.* (2017) Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, **542**, 352–356.
2. Trapnell,C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.
3. Mu,T., Xu,L., Zhong,Y., Liu,X., Zhao,Z., Huang,C., Lan,X., Lufei,C., Zhou,Y., Su,Y. *et al.* (2020) Embryonic liver developmental trajectory revealed by single-cell RNA sequencing in the Foxa2(eGFP) mouse. *Commun Biol.*, **3**, 642.
4. Guan,D., Xiong,Y., Trinh,T.M., Xiao,Y., Hu,W., Jiang,C., Dierickx,P., Jang,C., Rabinowitz,J.D. and Lazar,M.A. (2020) The hepatocyte clock and feeding control chronophysiology of multiple liver cell types. *Science*, **369**, 1388–1394.
5. Xiong,X., Kuang,H., Ansari,S., Liu,T., Gong,J., Wang,S., Zhao,X.Y., Ji,Y., Li,C., Guo,L. *et al.* (2019) Landscape of intercellular crosstalk in healthy and NASH liver revealed by single-cell secretome gene analysis. *Mol. Cell*, **75**, 644–660.
6. Vieth,B., Parekh,S., Ziegenhain,C., Enard,W. and Hellmann,I. (2019) A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.*, **10**, 4667.
7. Miao,Z., Deng,K., Wang,X. and Zhang,X. (2018) DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, **34**, 3223–3224.
8. Korthauer,K.D., Chu,L.F., Newton,M.A., Li,Y., Thomson,J., Stewart,R. and Kendziorski,C. (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**, 222.
9. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
10. Mou,T., Deng,W., Gu,F., Pawitan,Y. and Vu,T.N. (2019) Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. *Front. Genet.*, **10**, 1331.
11. Jaakkola,M.K., Seyednasrollah,F., Mehmood,A. and Elo,L.L. (2017) Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinformat.*, **18**, 735–743.
12. Kim,T.K. (2017) Understanding one-way ANOVA using conceptual figures. *Korean J. Anesthesiol.*, **70**, 22–26.
13. Farmahin,R., Gannon,A.M., Gagne,R., Rowan-Carroll,A., Kuo,B., Williams,A., Curran,I. and Yauk,C.L. (2019) Hepatic transcriptional dose–response analysis of male and female Fischer rats exposed to hexabromocyclododecane. *Food Chem. Toxicol.*, **133**, 110262.
14. Moffat,I., Chepelev,N., Labib,S., Bourdon-Lacombe,J., Kuo,B., Buick,J.K., Lemieux,F., Williams,A., Halappanavar,S., Malik,A. *et al.* (2015) Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo[a]pyrene in drinking water. *Cr. Rev. Toxicol.*, **45**, 1–43.
15. National Toxicology Program (2018) NTP research report on national toxicology program approach to genomic dose–response modeling. Research Triangle Park (NC): National Toxicology Program. *Natl. Toxicol. Program Res. Rep. Ser.*, http://dx.doi.org/10.22427/NTP-RR-5.
16. Webster,A.F., Chepelev,N., Gagne,R., Kuo,B., Recio,L., Williams,A. and Yauk,C.L. (2015) Impact of genomics platform and statistical filtering on transcriptional benchmark doses (BMD) and multiple approaches for selection of chemical point of departure (PoD). *PLoS One*, **10**, e0136764.
17. Gant,T.W. and Zhang,S.D. (2005) In pursuit of effective toxicogenomics. *Mutat. Res.*, **575**, 4–16.
18. Choi,K., Chen,Y., Skelly,D.A. and Churchill,G.A. (2020) Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.*, **21**, 183.
19. McDavid,A., Finak,G., Chattopadyay,P.K., Dominguez,M., Lamoreaux,L., Ma,S.S., Roederer,M. and Gottardo,R. (2013) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, **29**, 461–467.
20. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
21. Das,S. and Rai,S.N. (2021) SwarnSeq: an improved statistical approach for differential expression analysis of single-cell RNA-seq data. *Genomics*, **113**, 1308–1324.
22. Jeon,M. and De Boeck,P. (2017) Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychol Methods*, **22**, 340–360.
23. Li,Y., Liu,X.-B. and Yu,J. (2015) A Bayesian chi-squared test for hypothesis testing. *J. Econometrics*, **189**, 54–69.
24. Fader,K.A., Nault,R., Kirby,M.P., Markous,G., Matthews,J. and Zacharewski,T.R. (2017) Convergence of hepcidin deficiency, systemic iron overloading, heme accumulation, and REV-ERBalpha/beta activation in aryl hydrocarbon receptor-elicited hepatotoxicity. *Toxicol. Appl. Pharmacol.*, **321**, 1–17.
25. Percie du Sert,N., Hurst,V., Ahluwalia,A., Alam,S., Avey,M.T., Baker,M., Browne,W.J., Clark,A., Cuthill,I.C., Dirnagl,U. *et al.* (2020) The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *BMC Vet. Res.*, **16**, 242.
26. Nault,R., Fader,K.A., Bhattacharya,S. and Zacharewski,T.R. (2020) Single-nuclei RNA sequencing assessment of the hepatic effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin. *Cell. Mol. Gastroenterol. Hepatol.*, **11**, 147–159.
27. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
28. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
29. Davis,J.A., Gift,J.S. and Zhao,Q.J. (2011) Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1. *Toxicol. Appl. Pharmacol.*, **254**, 181–191.
30. Newton,M.A., Noueiry,A., Sarkar,D. and Ahlquist,P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.
31. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
32. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
33. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
34. Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics Bull.*, **1**, 80–83.
35. Fisher,R.A. (1921) On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3–32.
36. Kruskal,W.H. and Wallis,W.A. (1952) Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.*, **47**, 583–621.
37. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G., Geoghegan,J., Germino,G. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.
38. Vieth,B., Ziegenhain,C., Parekh,S., Enard,W. and Hellmann,I. (2017) powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, **33**, 3486–3488.
39. Zhang,X., Xu,C. and Yosef,N. (2019) Simulating multiple faceted variability in single cell RNA sequencing. *Nat. Commun.*, **10**, 2611.
40. Assefa,A.T., Vandesompele,J. and Thas,O. (2020) SPsimSeq: semi-parametric simulation of bulk and single-cell RNA-sequencing data. *Bioinformatics*, **36**, 3276–3278.
41. Chicco,D. and Jurman,G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.*, **21**, 6.

42. Dal Molin,A., Baruzzo,G. and Di Camillo,B. (2017) Single-Cell RNA-sequencing: assessment of differential expression analysis methods. *Front. Genet.*, **8**, 62.
43. Phillips,J.R., Svoboda,D.L., Tandon,A., Patel,S., Sedykh,A., Mav,D., Kuo,B., Yauk,C.L., Yang,L., Thomas,R.S. *et al.* (2019) BMDExpress 2: enhanced transcriptomic dose–response analysis workflow. *Bioinformatics*, **35**, 1780–1782.
44. Soufan,O., Ewald,J., Viau,C., Crump,D., Hecker,M., Basu,N. and Xia,J. (2019) T1000: a reduced gene set prioritized for toxicogenomic studies. *PeerJ*, **7**, e7975.
45. Fader,K.A., Nault,R., Zhang,C., Kumagai,K., Harkema,J.R. and Zacharewski,T.R. (2017) 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)-elicited effects on bile acid homeostasis: Alterations in biosynthesis, enterohepatic circulation, and microbial metabolism. *Sci. Rep.-UK*, **7**, 5921.
46. Wells,R.G. (2014) The portal fibroblast: not just a poor man's stellate cell. *Gastroenterology*, **147**, 41–47.