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

Samiran Sinha
Department of Statistics, Texas A&M University

WNAR 2023, Anchorage, Alaska

June 19, 2023

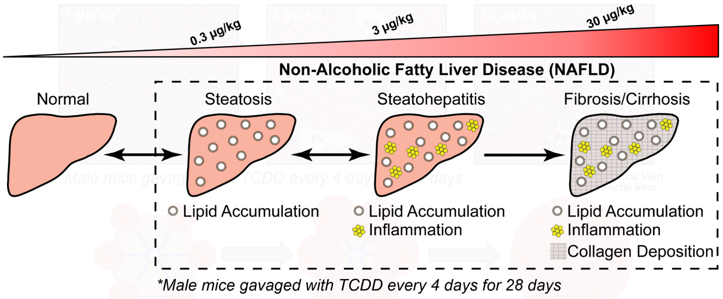
Collaborators: Rance Nault, Satabdi Saha, Sudin Bhattacharya, Tapabrata Maiti, Tim Zacharewski

Motivation

- 2,3,7,8-Tetrachlorodibenzo-p-Dioxin (TCDD)- a highly toxic and persistent organic pollutant
    - A by-products of a wide range of manufacturing processes including smelting, chlorine bleaching of paper pulp and the manufacturing of some herbicides and pesticides
    - Found in soils, sediments and food, especially dairy products, meat, fish and shellfish
- This contaminant can be found at the superfund sites

Objective

Fader et al. [1] used bulk RNAseq data on liver and on the intestine as well as flow cytometry (this is the only cell-specific level data) and histology (involving female mice)

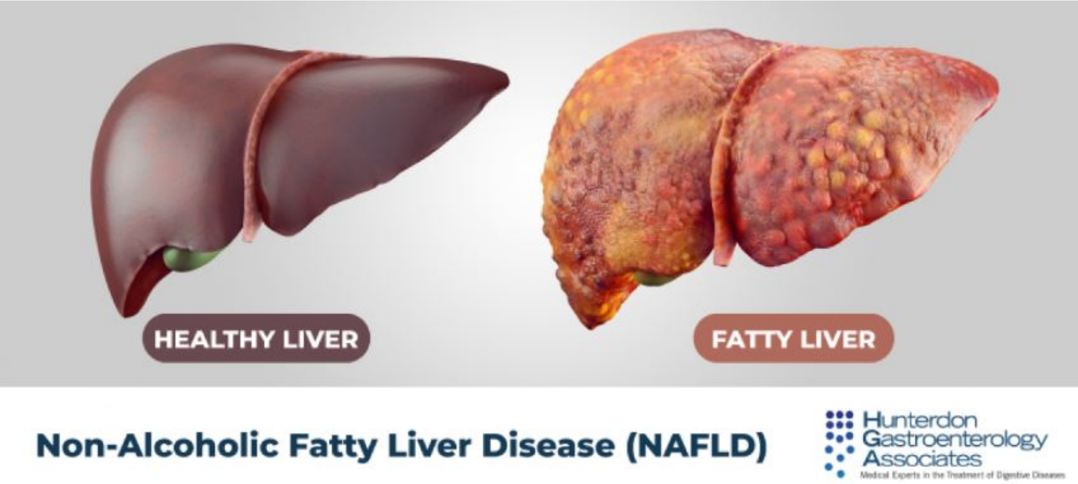


Fader et al. (2015). PMID26377647

Goal: Investigate dose-dependent TCDD-elicited hepatic cell-specific gene expression associated with the development of NAFLD among male mice using scRNAseq data

[1] Fader et al. (2015). 2,3,7,8-Tetrachlorodibenzo-p-Dioxin Alters Lipid Metabolism and Depletes Immune Cell Populations in the Jejunum of C57BL/6 Mice. Toxicol Sci. 2015 Dec;148(2):567-80.

Normal versus fatty liver


HEALTHY LIVER
FATTY LIVER
Non-Alcoholic Fatty Liver Disease (NAFLD)
Hunterdon Gastroenterology Associates
Medical Experts in the Treatment of Digestive Diseases

Experimental Design

- Randomly assigned male C57BL/6 mice to one one of the eight levels, 0 with 0.1 mL sesame oil vehicle (0-level or control) 0.01, 0.03, 0.1, 0.3, 1, 3, 10 or 30 µg/kg TCDD every 4 days for 28 days (How many mice in each dose level? 3, so the total was $3 \times 8 = 24$)
- Hepatic single-nuclei RNA-sequencing (snRNAseq) was performed using the 10× Genomics Chromium Single Cell $3'$ v3.1 kit
- These are mRNA measurements

National Toxicology Program's approach to genomic dose response modeling

- Design dose response (DR) experiment with a sufficient number of doses
- Design appropriate statistical test of hypothesis for deriving genes that exhibit minimum effect to treatment
- Fit parametric DR models derived from the Environmental Protection Agency (EPA) software to identify a biological potency estimate
- Group genes into predefined sets defined by gene ontologies and compute composite POD of the gene set
- Provide biological explanations for the selected set of genes and POD estimates

POD: Point of departure, the threshold dose level at which the gene expression starts to change from the control group

Challenges

- Traditional tests for pre-filtering: ANOVA
- scRNA-seq data is highly heterogeneous (across different cell types) and has a large number of zeroes
- Violates standard Gaussian assumptions
- No recommended method for differential gene expression analysis (DGEA) in multiple group single cell experimental studies
- Our contribution: Bayesian multiple group test (scBT)
  - Designed exclusively for dose–response scRNAseq data
  - FDR control

Distributional Assumptions

- $Y_{i,j,k}$: expression value of cell i, gene j, dose-level k, for $i = 1, \ldots n$ and $j = 1, \ldots p$, $k = 1, \ldots, K$
- $R_{i,j,k} = I[Y_{i,j,k} > 0]$: indicator denoting the presence of an expression
- Adopt the Hurdle model [2]

$$\begin{aligned} [Y_{i,j,k} | R_{i,j,k} = 1] &\sim \text{Normal}(\mu_{j,k}, \sigma_j^2), \\ \text{pr}(Y_{i,j,k} = 0 | R_{i,j,k} = 0) &= 1, \\ R_{i,j,k} &\sim \text{Bernoulli}(\omega_{j,k}), \end{aligned} \tag{1}$$

- $\mu_{j,k}$: the mean expression of the jth gene, level k, when it is expressed
- $\omega_{j,k}$: the rate of gene expression of gene j and dose-level k

[2] McDavid et al. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics, 29, 461–467.

Test of Hypothesis

- Need to account for the bimodality (due to the mixing of zeros and the positive valued numeric variable) in single cell gene expression distribution
- Design a test capable of detecting changes in the means and the inflation parameters, simultaneously, across the dose levels

$$H_{0,j} : \mu_{j,1} = \cdots = \mu_{j,K} = \mu_j \text{ and } \omega_{j,1} = \cdots = \omega_{j,K} = \omega_j$$

versus the alternative

$$H_{a,j} : H_{0,j} \text{ does not hold}$$

Bayesian Test- priors

For a given gene, say j, we use the following priors to calculate the marginal likelihoods

$H_{0,j}$	$H_{a,j}$
$\mu_{j,1} = \cdots = \mu_{j,K} = \mu_j$	
$\mu_j \sim N(m_0, \tau_\mu \sigma_j^2)$	$\mu_{j,k} \sim \text{Normal}(m_{k,0}, \tau_{k,\mu} \sigma_j^2)$
$\sigma_j^2 \sim IG(a_\sigma, b_\sigma)$	$\sigma_j^2 \sim IG(a_\sigma, b_\sigma)$
$\omega_{j,1} = \cdots = \omega_{j,K} = \omega_j$	
$\omega_j \sim \text{Beta}(a_\omega, b_\omega)$	$\omega_{k,j} \sim \text{Beta}(a_{k,\omega}, b_{k,\omega})$

Hyperparameters are obtained by maximising the marginal likelihood under the null and the alternative hypothesis

Bayesian Test of Hypothesis (scBT)

- Bayes factor

$$BF_{01,j} = \frac{\mathcal{L}_{H_0,j}}{\mathcal{L}_{H_a,j}} \times \frac{\pi(H_{a,j})}{\pi(H_{0,j})}$$

- $\pi(H_{a,j})$ and $\pi(H_{0,j})$: prior probabilities for alternative and null model
  Are $\pi(H_{a,j}) = \pi(H_{0,j})$? Yes.
- $\mathcal{L}_{H_0,j}$ and $\mathcal{L}_{H_a,j}$: marginal likelihood under the null and the alternative hypothesis.

Marginal Likelihood under $H_{0,j}$

$$\mathcal{L}_{H_{0,j}} = \frac{1}{(2\pi)^{(\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k})/2}} \times \frac{1}{\sqrt{1+\tau_\mu \sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k}}}$$
$$\times \frac{1}{\Gamma(a_\sigma) b_\sigma^{a_\sigma}} \times \frac{\Gamma(a_\sigma + (\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k})/2)}{(1/b_\sigma + \mathcal{A}_{tot}/2)^{a_\sigma + (\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k})/2}}$$
$$\times \frac{\text{Beta}(a_\omega + (\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k}), b_\omega + \sum_{k=1}^{K} n_k - (\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k}))}{\text{Beta}(a_\omega, b_\omega)}$$

where

$$\mathcal{A}_{tot} = \left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{k,i,j} Y_{i,j,k}^2 + \frac{m_0^2}{\tau_\mu}\right\} - \left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k} + \frac{1}{\tau_\mu}\right\}^{-1} \left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k} Y_{i,j,k} + \frac{m_0}{\tau_\mu}\right\}$$

Marginal Likelihood under $H_{a,j}$

$$\mathcal{L}_{H_{a,j}} = \frac{1}{(2\pi)^{(\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{k,i,j})/2}} \times \frac{1}{\prod_{k=1}^{K}\sqrt{1+\tau_{k,\mu}\sum_{i=1}^{n_k} R_{i,j,k}}}$$
$$\times \frac{1}{\Gamma(a_\sigma) b_\sigma^{a_\sigma}} \times \frac{\Gamma(a_\sigma + \sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k}/2)}{(1/b_\sigma + \sum_{k=1}^{K} \mathcal{A}_k/2)^{a_\sigma + \sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{i,j,k}/2}}$$
$$\times \prod_{k=1}^{K} \frac{\mathrm{Beta}(a_{k,\omega} + \sum_{i=1}^{n_k} R_{i,j,k}, b_{k,\omega} + n_k - \sum_{i=1}^{n_k} R_{i,j,k})}{\mathrm{Beta}(a_{k,\omega}, b_{k,\omega})}$$

where

$$\mathcal{A}_k = \left\{\sum_{i=1}^{n_k} R_{i,j,k} Y_{i,j,k}^2 + \frac{m_{k,0}^2}{\tau_{k,\mu}}\right\} - \left\{\sum_{i=1}^{n_k} R_{i,j,k} + \frac{1}{\tau_{k,\mu}}\right\}^{-1} \left\{\sum_{i=1}^{n_k} R_{i,j,k} Y_{i,j,k} + \frac{m_{k,0}}{\tau_{k,\mu}}\right\}^2$$

Multiplicity control

- D_j: Data for the jth gene
- Calculate the posterior probability of the null hypothesis

$$p(H_{0,j}|D_j) = \left(1 + \frac{1}{BF_{01,j}}\right)^{-1}$$

- For a target FDR α, we reject $H_{0,j}$ when $p(H_{0,j}|D_j) < \zeta$, [3]
  - ζ is the largest value such that $\frac{C(\zeta)}{J(\zeta)} \leq \alpha$
  - $J(\zeta) = \{j : p(H_{0,j}|D_j) \leq \zeta\}$ and $C(\zeta) = \sum_{j \in J(\zeta)} p(H_{0,j}|D_j)$
  - $\frac{C(\zeta)}{J(\zeta)}$: the average posterior probability of null hypothesis of the statistically significant genes

[3] Newton et al. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics, 5, 155–176.

Competing approaches

- Likelihood ratio test (LRT-multiple)
- LRT with a linear model for the mean and the inflation parameters (LRT-linear)
  - $\mu_{j,k} = \beta_{0,j} + \beta_{j,1} d_k$ (d_k: kth dose, jth gene)
  - $\text{logit}(\omega_{j,k}) = \gamma_{0,j} + \gamma_{j,1} d_k$
  - For the jth gene, test $H_{0,j} : \beta_{j,1} = 0, \gamma_{j,1} = 0$ versus $H_{a,j} : \beta_{j,1} \neq 0$ or $\gamma_{j,1} \neq 0$

Competing approaches (existing)

- limma-trend[4]: linear regression with dose as the explanatory variable
- MAST[5]: Model-based Analysis of Single-cell Transcriptomics
- Seurat Bimod[6]: (a pairwise test assuming the single cell RNA-seq hurdle model framework)
- WRS: Wilcoxon-Rank Sum test (it is pairwise test)
- ANOVA
- KW: Kruskal-Wallis test (nonparametric extension of one-way ANOVA)

For all these methods, we used the Benjamini-Hochberg adjusted p-value

[4] Law et al. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol., 15, R29.

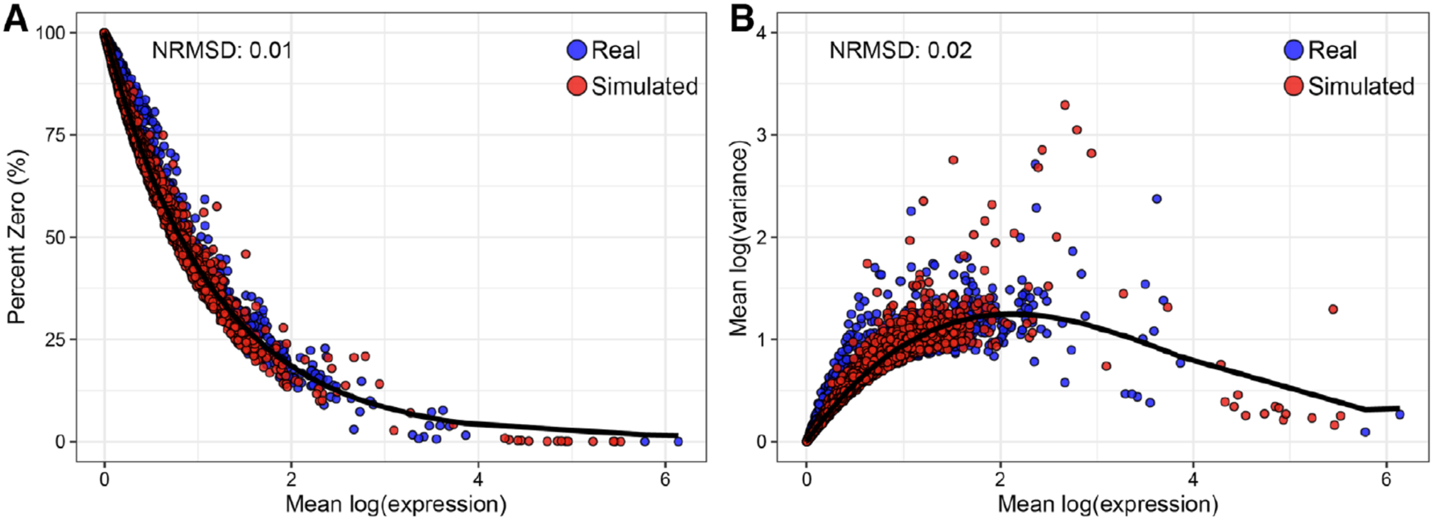
[5] Finak et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol., 16, 278.

[6] McDavid et al. (2013) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics, 29, 461–467.

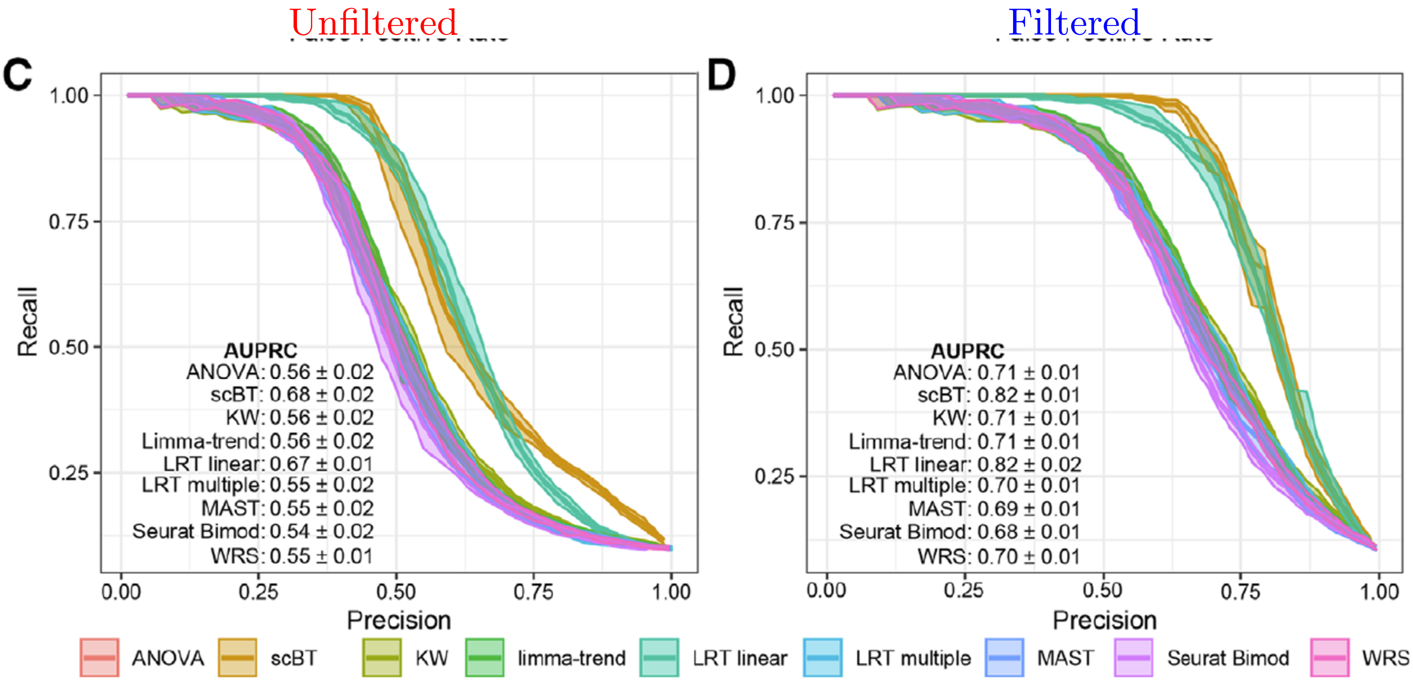
Simulations

- Data was simulated consisting of nine dose groups of 500 cells each (4500 total) and 5000 genes with a 10% probability of being DE
- Simulation of genes based on initial parameters derived from real DR data was reproduced 10 times
- Performance of scBT was benchmarked against 8 other DE analysis tests
- To investigate test performance in controlling type I errors, DGEA methods on simulated datasets were examined with 0 %DE genes (i.e. negative control)

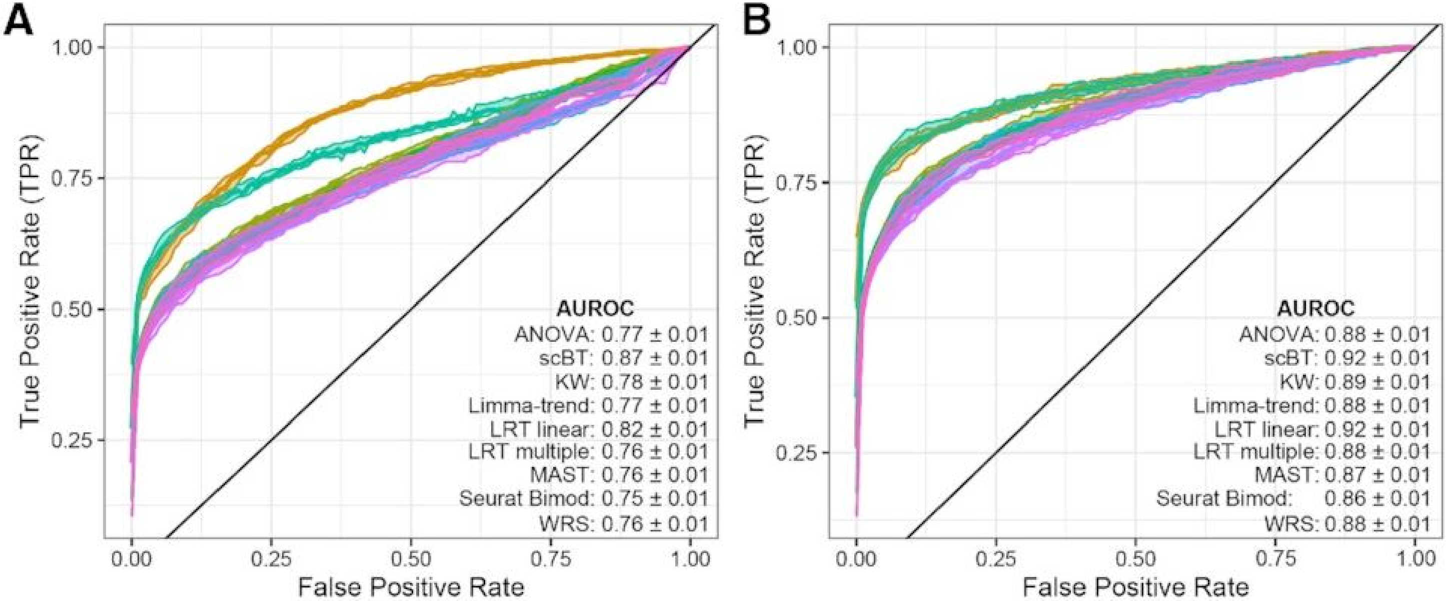
Simulations



Simulation Results (PRC)

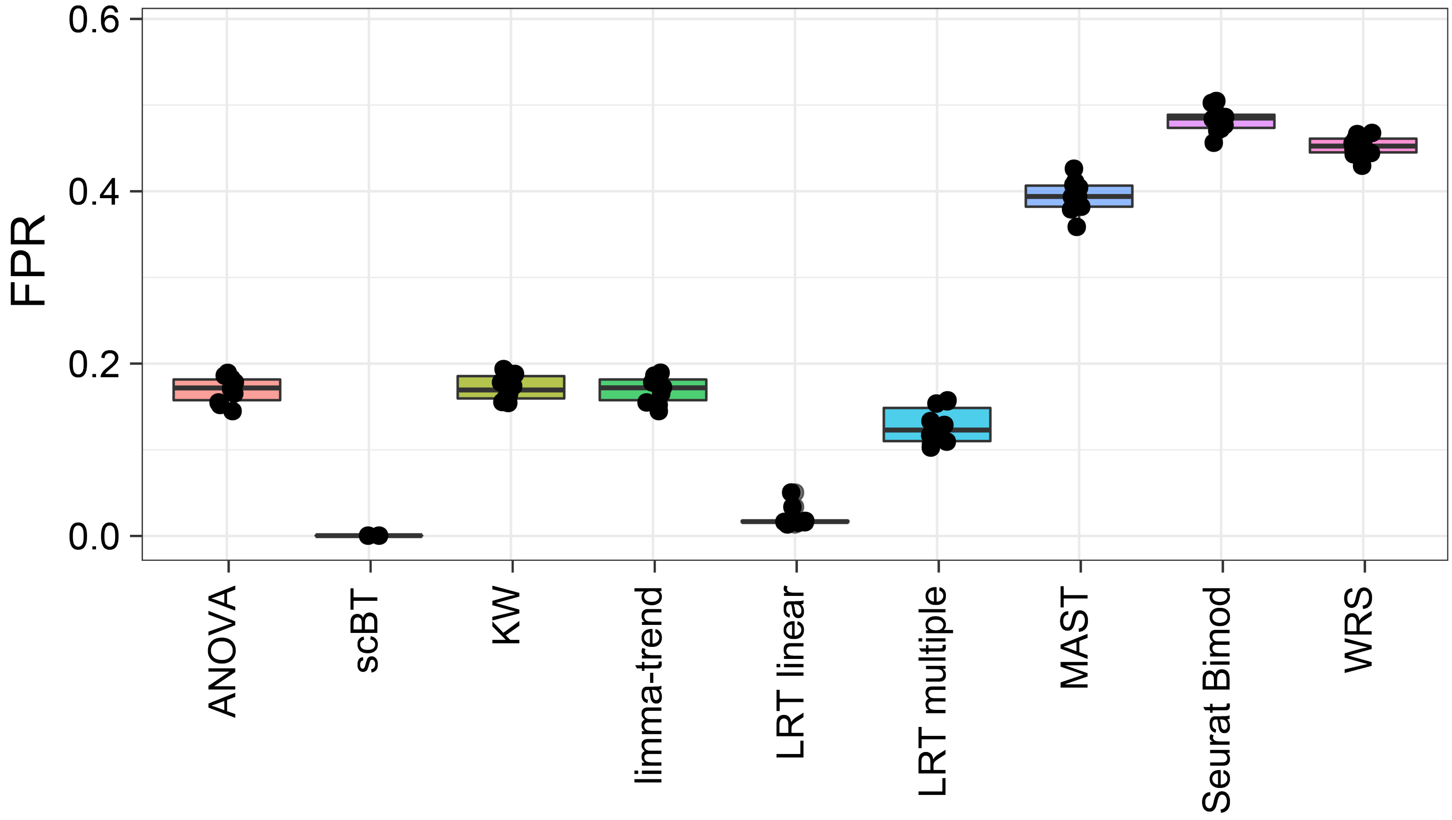


Simulation Results (ROC)


Unfiltered

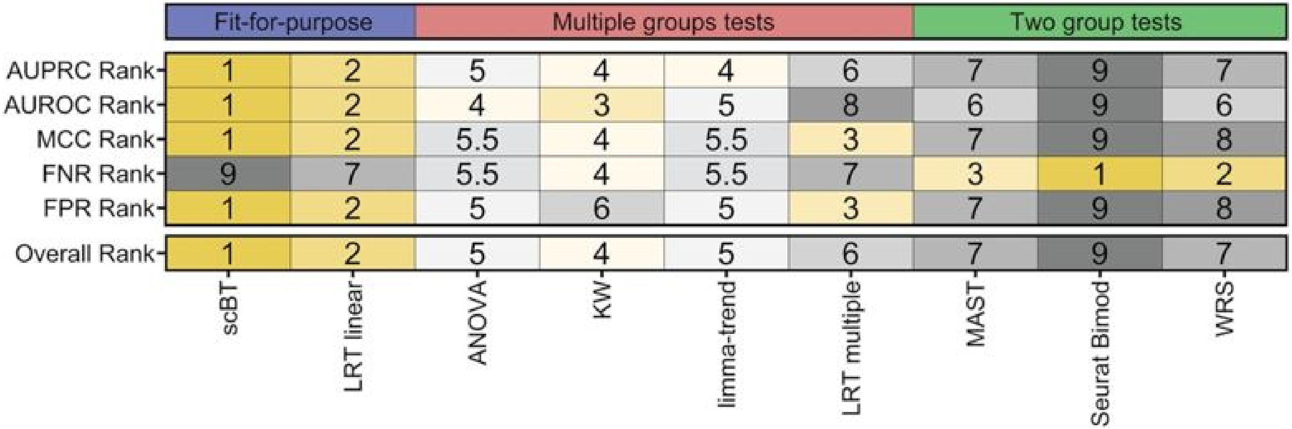
Filtered

Simulation Results



Simulation Results (performance ranking)

The methods were compared based on the 1) Mathew's correlation coefficients (MCC), 2) false positive rates (FPRs), 3) false negative rates (FNRs), 4) the area under the ROC curve (AUROC) and 5) the area under the precision recall curve (AUPRC)



	Fit-for-purpose		Multiple groups tests				Two group tests		
	scBT	LRT linear	ANOVA	KW	limma-trend	LRT multiple	MAST	Seurat Bimod	WRS
AUPRC Rank	1	2	5	4	4	6	7	9	7
AUROC Rank	1	2	4	3	5	8	6	9	6
MCC Rank	1	2	5.5	4	5.5	3	7	9	8
FNR Rank	9	7	5.5	4	5.5	7	3	1	2
FPR Rank	1	2	5	6	5	3	7	9	8
Overall Rank	1	2	5	4	5	6	7	9	7

Real Dose Response Dataset

- Total nuclei: 131613
- Average nuclei per dose group: 14624
- Seurat was used to integrate and log-normalize expression data
- Average of 1,665 genes were detected across all nuclei
- Applied the scBT method, used the FDR controlled selection criteria
- Filtering method: Genes in the experimental dataset were considered differentially expressed when expressed in $\geq$5% of cells in at least one dose group and had a $|$fold-change$| \geq 1.5$ in at least one of the seven treatment groups

Real Data Analysis

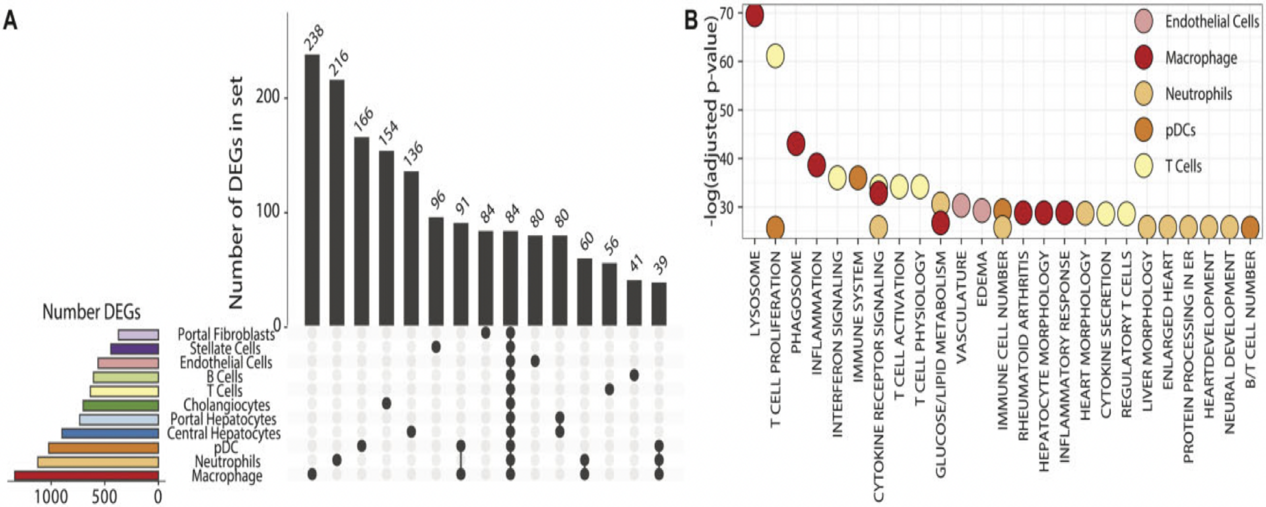


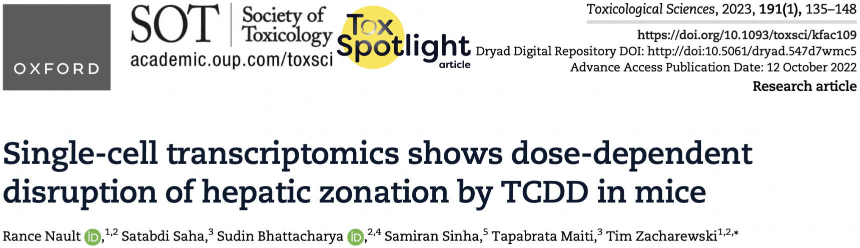
Figure 2. Set and functional analysis of hepatic differentially expressed genes (DEGs) from male mice gavaged with TCDD every 4 days for 28 days. A, UpSet plot of the 15 largest gene sets, in rank order, based on set analysis that identified both unique and common DEGs (Bayes Factor adjusted FDR ≥ 0.05 and |fold-change| ≥ 1.5) among all identified cell types. Set sizes represent the number of genes identified in only the cells as indicated by black circles. B, Functional analysis of DEGs for each cell type using gene lists from KEGG and MPO from the Gene Set Knowledgebase (GSKB; http://ge-lab.org/gskb/). Gene sets with ≥ 60% overlap were combined and manually annotated. The top 30 enriched functions (adjusted *p*-value) across all cell types are shown. A complete list is available in Supplementary Table 2.

Quick Summary

- Developed a multiplicity corrected Bayesian multiple group test (scBT), designed exclusively for DGEA of dose–response scRNAseq data
- In the context of investigating chemical or drug MoAs, false positives have the potential to lead to wasted effort and resources
- Simulations: scBT has excellent FPR control and top ranked AUPRC, but scBT has low power
- Real datasets: scBT detected biologically relevant genes in NAFLD development and progression
- The real data are deposited at the Gene Expression Omnibus

Contribution

- Simulator : R package SplattDR is available at https://github.com/zacharewskil
- Proposed test approaches : R package scBT is available at https://github.com/satabdisaha1



OXFORD

SOT | Society of Toxicology
academic.oup.com/toxsci

Tox Spotlight article

Toxicological Sciences, 2023, **191(1)**, 135–148

https://doi.org/10.1093/toxsci/kfac109

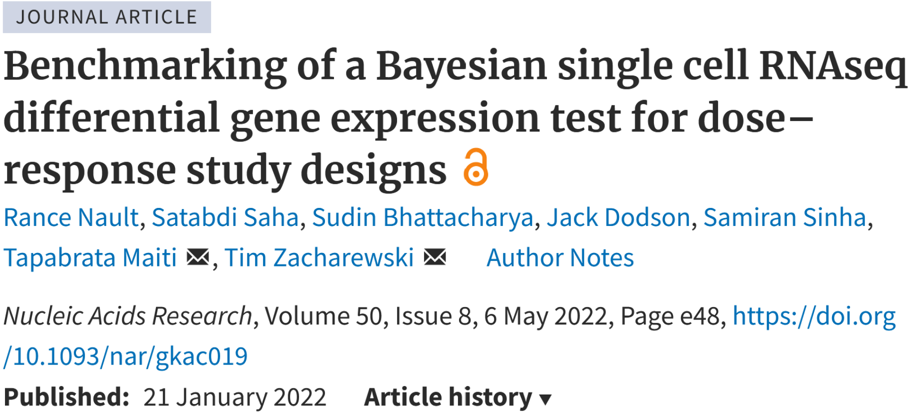
Dryad Digital Repository DOI: http://doi:10.5061/dryad.547d7wmc5

Advance Access Publication Date: 12 October 2022

Research article

Single-cell transcriptomics shows dose-dependent disruption of hepatic zonation by TCDD in mice

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



JOURNAL ARTICLE

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

Rance Nault, Satabdi Saha, Sudin Bhattacharya, Jack Dodson, Samiran Sinha, Tapabrata Maiti ✉, Tim Zacharewski ✉ Author Notes

Nucleic Acids Research, Volume 50, Issue 8, 6 May 2022, Page e48, https://doi.org/10.1093/nar/gkac019

Published: 21 January 2022 **Article history ▾**