Probabilistic index models for testing differential gene expression in single-cell RNA-seq data

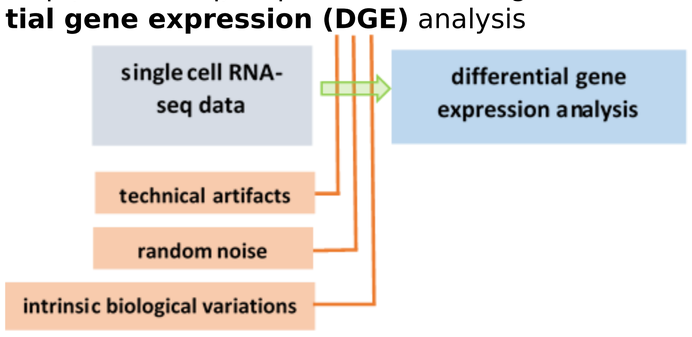
Alemu Takele Assefa[*1], **Jo Vandesompele**[2,3,4], **Olivier Thas**[1,3,5,6]

[1]Department of Data Analysis and Mathematical Modeling, Ghent University, Belgium; [2]Department of 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; [6]I-BioStat, Hasselt University, Belgium

* AlemuTakele.Assefa@UGent.BE

Introduction

The various sources of bias in single cell RNA-seq (scRNA-seq) experiment challenge **differential gene expression (DGE)** analysis



Parametric tools (e.g. generalized linear models with (zero-inflated) negative binomial family)

- flexible and easily interpretable modeling

Non-parametric tools (such as the Wilcoxon rank sum test, SAMSeq)

- simple, robust, and showed competitive performance for testing DGE in scRNA-seq data
- limitations: cannot be used for complex experimental designs (e.g. multi-factor), and do not provide an estimate of the effect size

Therefore, we propose a semi-parametric approach based on **Probabilistic Index Models** (PIM; [1, 2]), which form a class of models that generalizes classical rank tests.

PIM

- does not rely on strong distributional assumptions, and hence is robust
- is a regression framework, so that it can be used for complex experimental designs involving many factors of interest, e.g. treatment, sequencing depth, batch effect, . . . ,
- testing for DGE is augmented with an estimate of the effect size in terms of **probabilistic index (PI)**, which is straightforward for interpretation,
- unified approach, i.e. testing for DGE, normalization and reduction of unwanted variation can be done at the same time
- can be integrated with data pre-processing pipelines, e.g. normalization and data imputation

Probabilistic Index Models

For a particular gene, PIM is defined as

$$logit\{\text{Prob}(Y_i \preceq Y_j | X, U, Q)\} = \beta_{\mathbf{x}} X + \beta_{\mathbf{u}} U + \beta_{\mathbf{q}} Q$$

where Y_i and Y_j are the observed counts in cell i and j, respectively. X, U, and Q are vectors of factors of interest (treatment or cell type), normalization factors, and other sources of unwanted variations with their corresponding coefficient β_x, β_u, and β_q, respectively.

Testing for DGE is equivalent to testing the hypothesis

$$H_0 : \beta_x = 0 \quad \equiv \quad \text{Prob}(Y_i \preceq Y_j) = 0.5$$

with an estimate of effect size given by

$$\text{PI} = \frac{e^{\beta_x}}{1 + e^{\beta_x}} \in [0, 1]$$

Example

Data: Neuroblastoma cell line NGP scRNA-seq data, containing two group of cells: nutlin-3 treated (n_1=31) and vehicle (control) (n_2=52). It includes 15,439 genes expressed at least in 5 cells.
→ Objective: testing DGE between treated and control cells using PIM

The following model is fitted

$$logit\{\text{Prob}(Y_i \preceq Y_j)\} = \beta_1(X_j - X_i) + \beta_2(U_j - U_i)$$

where X_i is treatment indicator of cell i, such that X_i=1 for treated and 0 for control, and U_i and U_j are the log-library sizes in cell i and j.

Therefore, testing for DGE between treatment and control cells adjusting for library size differences (normalization) is equivalent to testing the hypothesis $H_A : \beta_1 \neq 0$.

Wald test is used to obtain p-values, and the Benjamini-Hochberg method is applied to control the false discovery rate (FDR).

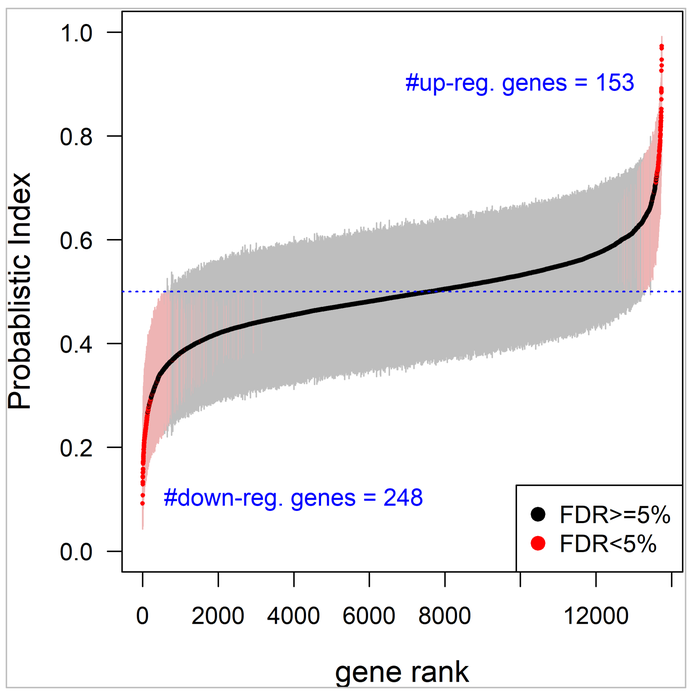


Figure 1: Gene ranking based on PI (with 95% confidence interval). The red color represents DE genes (FDR <5%), whereas non-DE genes are colored black.

A total of 401 (2.6%) genes detected as DE at 5% FDR. In addition to the adjusted p-values (adjusted to control FDR), the estimated PI can be used to declare DE.

In Figure 1, we ranked genes according to their estimated PI. Gene set enrichment analysis revealed that such ranking significantly enriched the TP53 pathway for NGP cells treated with nutlin-3 (normalized enrichment score 3.52, p-value < 0.001, Figure 2).

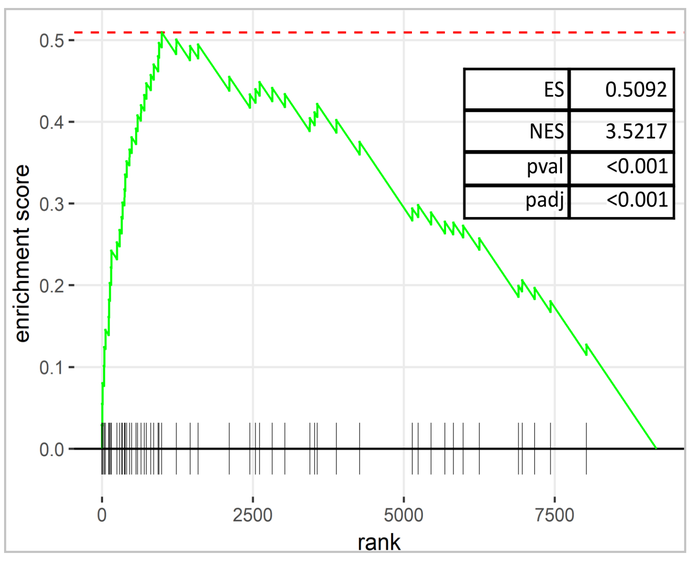


Figure 2: Gene set enrichment result for the TP53 pathway.

Also, in Figure 3, we showed visualization of DE genes that is equivalent to the volcano and MA plot using PI instead of log-fold change.

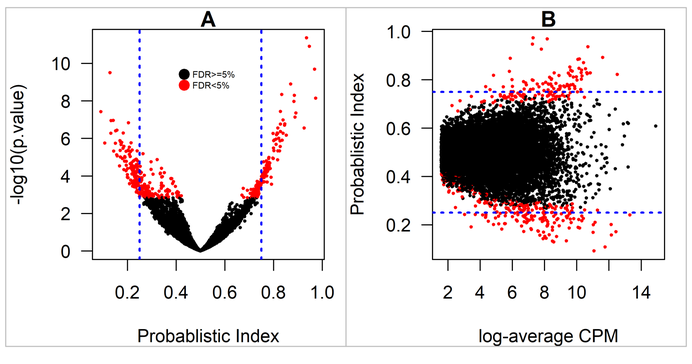


Figure 3: Alternative visualization of volcano plot (A) and MA plot (B). The blue dotted lines indicate PI = 0.25 and 0.75.

Performance evaluation

We have compared the performance of PIM with the state-of-the-art tools using simulation studies. In particular, a parametric simulation procedure was implemented to generate data from a mixture of Poisson-Gamma distributions (using `splatter` R package). The actual false discovery rate (FDR) and true positive rate (TPR) were used as measures of performance.

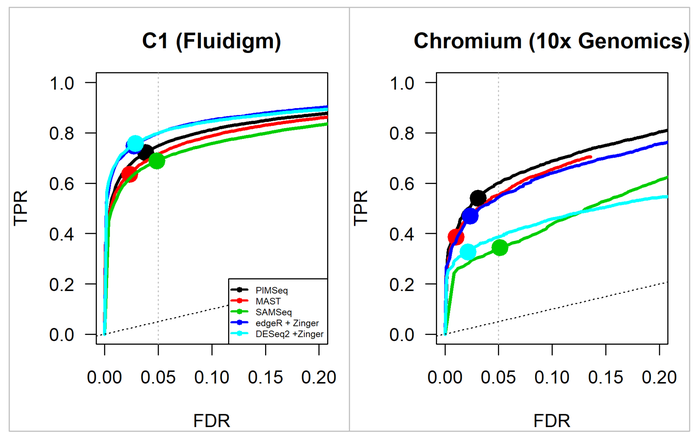


Figure 4: Performance of DGE tools for two C1 and Chromium protocols. The solid dots on each curve indicate the actual FDR and TPR at 5% nominal FDR.

Figure 4 shows that PIM has competitive performance to the parametric tools even when the data comes from negative binomial distribution. In some simulation settings (Chromium data), PIM performs best.

Breaking zero ties

By modelling the drop-out process in scRNA-seq experiment, we break the zero-ties, which further increases the efficiency of our method while preserving robustness (see Figure 5).

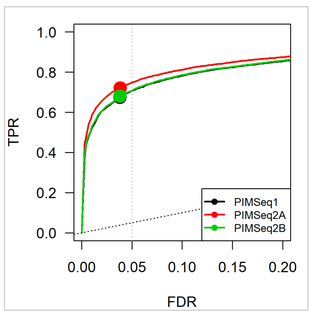


Figure 5: Performance of PIM with zero tie breaking strategy. PIMSeq1=the standard PIM, whereas PIMSeq2A and PIMSeq2B are PIM with zero-tie breaking methods when the true count is assumed to follow Poisson and negative binomial distributions, respectively.

Conclusions

- PIM does not rely on strong distributional assumptions and it is a robust approach for testing DGE in scRNA-seq
- It is adaptable to a wide range of experimental designs, and accounts for library size and other sources of variability
- It offers intuitively interpretable measure of the effect size, that augments DGE analysis
- The method can be considered as a generalization of the non-parametric methods (e.g. SAM) for testing DGE
- Our simulation studies demonstrate that PIM succeeds well in controlling the FDR at its nominal level, while showing good sensitivity as compared to competitor methods
- By modelling the drop-out process, we break the zero ties to further improve its performance while preserving robustness

References

[1] O. Thas, J. D. Neve, L. Clement, and J.-P. Ottoy, "Probabilistic index models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 4, pp. 623–671, 2012.

[2] J. De Neve and O. Thas, "A regression framework for rank tests based on the probabilistic index model," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 1276–1283, 2015.