Probabilistic Index Models for testing differential gene expression in single cell RNA-seq data

Alemu Takele Assefa[†1], Jo Vandesompele[‡1], Olivier Thas[§,1,2,3]

[1]Ghent University, Belgium; [2]University of Wollongong, Australia; [3]Hasselt University, Belgium

[†]AlemuTakele.Assefa@UGent.be, [‡]Jo.Vandesompele@UGent.be, [§]Olivier.Thas@UGent.be

March 26, 2019

Single cell RNA sequencing (scRNA-seq) data

- profiles gene expression patterns in individual cells
- data typically presented in a matrix

$$
\begin{array}{l|cccc}
 & \text{cell } 1 & \text{cell } 1 & \dots & \text{cell } n \\
\text{gene } 1 & y_{11} & y_{12} & \cdots & y_{1n} \\
\text{gene } 2 & y_{21} & y_{22} & \cdots & y_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{gene } G & y_{G1} & y_{G2} & \cdots & y_{Gn} \\
\hline
 & [\,N_1 & N_2 & \dots & N_n\,]
\end{array}
$$

from statistical point of view

- opportunity: high number of cells
- challenge: high noise level from various sources
  - technical noise because of low input material
  - intrinsic biological variability

⇒ scRNA-seq data

- sparse data
- complex distribution of gene expression

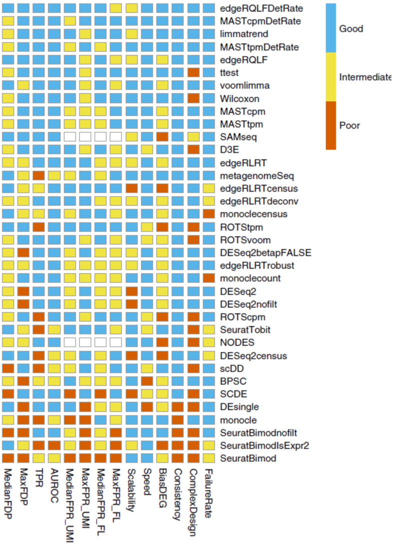
Differential gene expression (DGE) in scRNA-seq

DGE in scRNA-seq
⇒ identifies a set of genes with different distribution of expression across groups of cells

- parametric methods are often used for testing DGE
  e.g. NB or ZINB models
  + are flexible and easy for interpretation
  + account for various sources of variation
  + adaptable to many experimental design

- parametric assumptions do not always hold
  ⇒ tools relying on such assumption may thus under-perform

Benchmarking result by Soneson et al. Nature methods (2018)



- methods for bulk RNA-seq also work
- simple methods, such as t-test, WMW show good performance

---

non-parametric tools for testing DGE in scRNA-seq data

- showed better performance than many of the parametric tools but
- have limited scope
- no interpretable measure of fold-change (effect size)

Therefore, we suggest Probabilistic Index Models (PIM)[1] to widen the scope of non-parametric tools while

- being robust
- can be used for simple and complex experimental designs
- provide interpretable measure of the effect size

[1] Thas et al. JRSS-B (2012)

PIM

In PIM, we model the conditional probability

$$\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i, X_j) = \mathrm{P}(Y_{gi} < Y_{gj} | X_i, X_j) + \frac{1}{2}\mathrm{P}(Y_{gi} = Y_{gj} | X_i, X_j)$$

where Y_{gi} and Y_{gj} are the gene expression of gene g in cell i and j with their corresponding covariate information X_i and X_j, resp.

$\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i, X_j)$ is called the Probabilistic Index (PI)

PIM

- using a function $m(.)$ with range [0, 1], we model the PI as a function of X,

$$\boxed{\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i, X_j) = m(X_i, X_j; \beta_g)}$$

$m(X_i, X_j; \beta_g)$ satisfies some particular restrictions, see Thas et al. (2012)

- the parameter β_g represents the effect of X on the PI

- with an appropriate link function $g(.)$, such as logit,

$$\boxed{m(X_i, X_j; \beta_g) = g^{-1}(Z_{ij}^T \beta_g)}$$

where $Z_{ij} = X_j - X_i$ – one possible choice

Example

Let $(Y_{gi}, X_i), i = 1, \ldots, n$ are n i.i.d. r.v., where Y_{gi} is the normalized gene expression of gene g in cells i and X_i is a treatment group indicator of cell i ($X_i = 1$ for treatment and 0 for control).

Therefore, with a logit link function, we define PIM as

$$\boxed{\operatorname{logit}\left\{\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i, X_j)\right\} = \beta_g (X_j - X_i)}$$

- if $\beta_g = 0$, $\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i = 0, X_j = 1) = 0.5$
  $\Rightarrow$ probability that expression of gene g in a randomly selected cell from the control group is smaller than that of a randomly selected cell from the treatment group is 50% (and vice versa)
- $\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i = 0, X_j = 1) = \frac{e^{\beta_g}}{1+e^{\beta_g}} \in [0, 1]$

Example ... cont'd

- parameter estimation equation (score function)

$$\boxed{\sum_{(i,j)\in I_n} \mathrm{A}(\mathrm{Z}_{ij};\beta)\left\{I_{ij} - g^{-1}(\mathrm{Z}_{ij}^T\beta)\right\} = 0}$$

where $I_{ij} = \mathrm{I}(Y_i < Y_j) + 0.5\mathrm{I}(Y_i = Y_j) \in (0,\ 0.5,\ 1)$ – pseudo observations

- testing for no treatment effect, $H_0 : \beta_g = 0$,
  $\Rightarrow$ using Wald test of Thas et al (2012)[2]

- treatment effect size $\Leftrightarrow$ PI

$$\hat{\mathrm{P}}(Y_{gi} \preceq Y_{gj} | X_i = 0, X_j = 1) = expit\{\hat{\beta}_g\} \in [0, 1]$$

- Testing DGE for $G >> 1$ genes results in a vector of p-values
  $\Rightarrow$ Benjamini-Hochberg procedure to control false discovery rate (FDR)

[2] Thas et al. JRSS-B (2012)

Example: testing for DGE using PIMs

- Data:
  - Neuroblastoma cell line scRNA-seq data (SMARTer/C1)
  - two groups of cells: nutlin-3 treated (n_1=31) and control (n_2=52)
  - all cells came from a single biological sample and processed in a single batch
  - ≈12,000 genes, each with expression in at least 5 cells
- Objective: testing for DGE between nutlin-3 treated and control group of cells (X) adjusting for library size (N)
- PIM specification

$$\text{logit}\{\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i, X_j, N_i, N_j)\} = \underbrace{\beta_g^X (X_j - X_i)}_{\text{treatment effect}} + \underbrace{\beta_g^N (\log N_j - \log N_i)}_{\text{adjust for library size}}$$

Example: testing for DGE using PIMs ... cont'd

- PIM specification

$$\text{logit}\{\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i, X_j, N_i, N_j)\} = \underbrace{\beta_g^X (X_j - X_i)}_{\text{treatment effect}} + \underbrace{\beta_g^N (\log N_j - \log N_i)}_{\text{adjust for library size}}$$

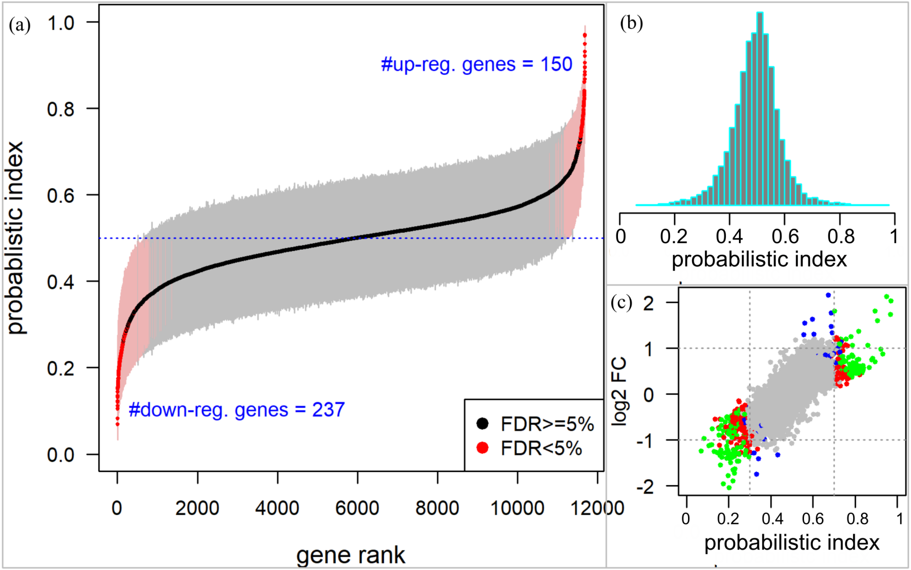
- the effect of nutlin-3 treatment for gene g given $N_i = N_j = n$,

$$\text{logit}\{\mathrm{P}(Y_{gi} \preceq Y_{gj} | X_i = 0, X_j = 1, N_i = n, N_j = n)\} = \beta_g^X$$

- ranking genes based on their estimated marginal PI of nutlin-3, i.e.

left edge	middle	right edge
PI → 0	PI ≈ 0.5	PI → 1
down regulated	no DGE	up regulated

Example: testing for DGE using PIMs ... results



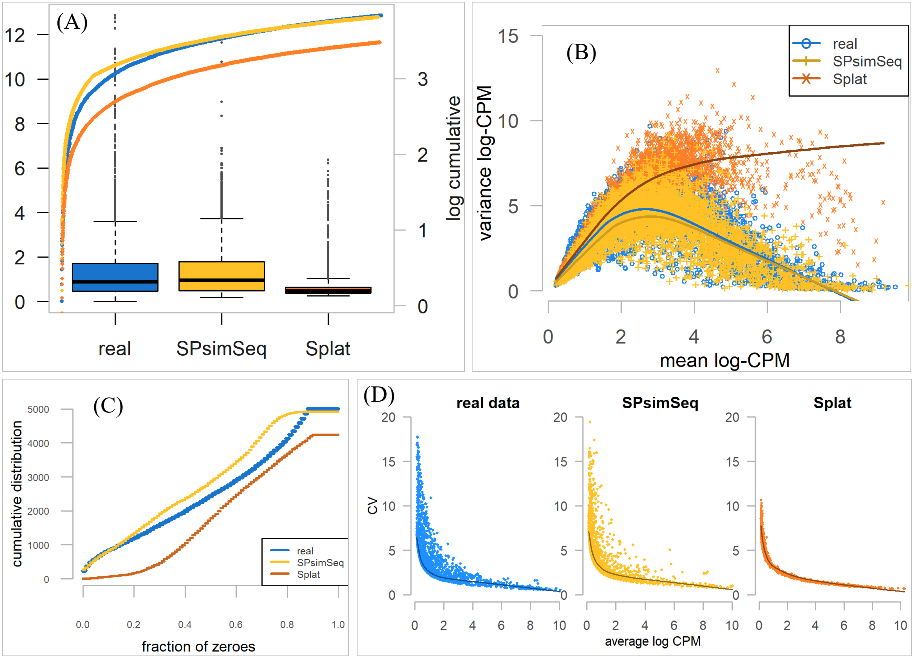
Performance evaluation ... simulation methods

Two sets of simulation methods

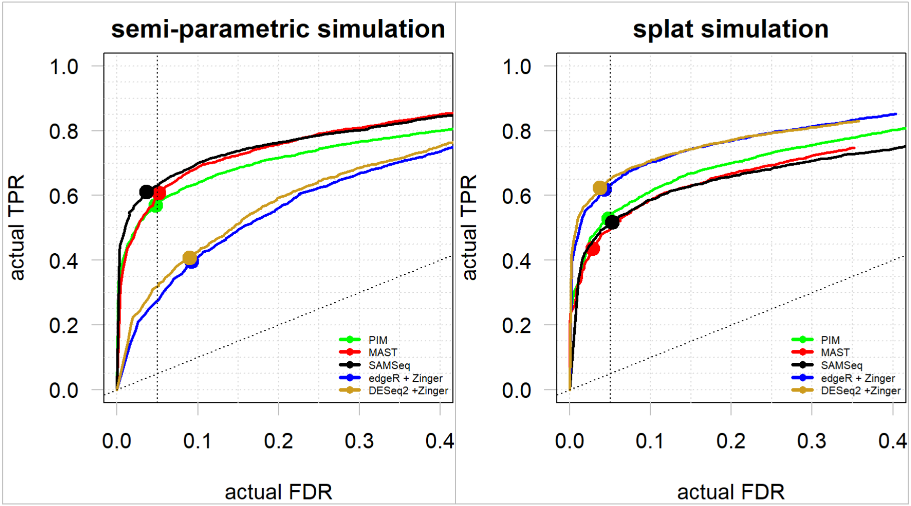
1. Splat simulation[3]: gamma-Poisson hierarchical model
   ⇒ Negative Binomial
   ⇒ fast and several scenario can be simulated

2. semi-paramatric simulation
   ⇒ sampling new data from the actual distribution of a real scRNA-seq data
   ⇒ involves two steps: construct density, and sample from the constructed density
   ⇒ generates realistic data

[3] Zappia et al Genome Biology (2017)

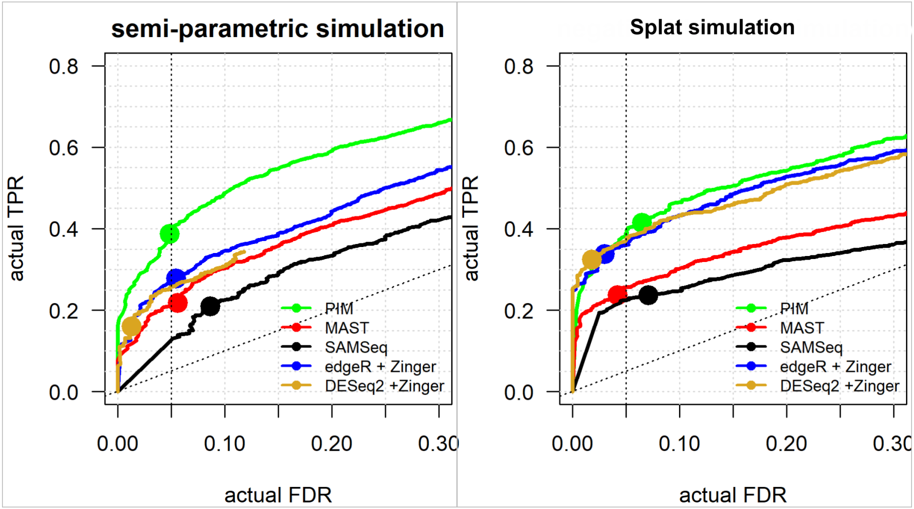

(A)
12
10
8
6
4
2
0
3
2
1
0
log cumulative
real
SPsimSeq
Splat
(B)
15
10
5
0
variance log-CPM
real
SPsimSeq
Splat
0
2
4
6
8
mean log-CPM
(C)
5000
4000
3000
2000
1000
0
cumulative distribution
real
SPsimSeq
Splat
0.0
0.2
0.4
0.6
0.8
1.0
fraction of zeroes
(D)
real data
SPsimSeq
Splat
20
15
10
5
0
CV
0
2
4
6
8
10
average log CPM

Performance evaluation ... simulation results



sim. design: 5000 genes, 2 group o f cells ($n_1 = n_2 = 50$), 10% DE genes, source data generated using SMARTer/C1 protocol, gene expression data in terms of read-counts.

Performance evaluation ... simulation results



sim. design: 5000 genes, 2 group o f cells ($n_1 = n_2 = 100$), 10% DE genes, source data generated using Chromium (10x Genomics) protocol, gene expression data in terms of UMI-counts.

Summary

PIM for testing DGE

- requires minimal distributional assumption
  ⇒ robust
- generalization of non-parametric methods
  ⇒ can be used for simple and complex experimental designs
  ⇒ PIM is more flexible than SAMSeq[4]
- interpretable effect size in terms of PI
  ⇒ meaningful gene ranking based on PI (in combination with p-values ot its standard error)
- valid under the presence of tied observations
- can be used for different measures of gene expression, such as read-counts and UMI-counts

[4] Li et al, Statistical methods in medical research (2013)