Unlocking RNA-seq tools for zero inflation and single cell applications using observation weights

Koen Van den Berge, Ghent University

Statistical Genomics, 2018-2019

The team

Koen Van den Berge*

Fanny Perraudeau*

Davide Risso

Jean-Philippe Vert

Charlotte Soneson

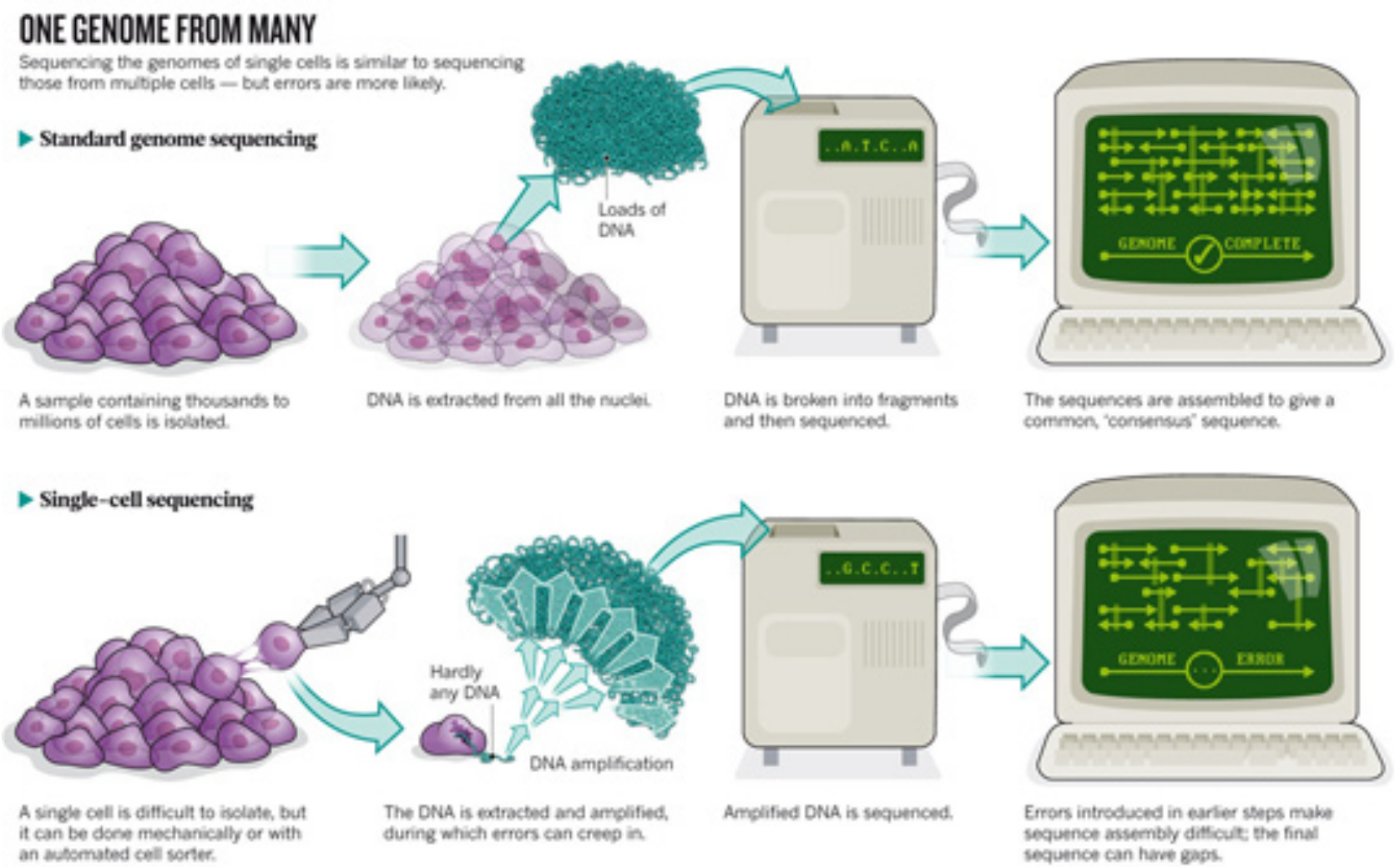
Michael Love

Mark Robinson

Sandrine Dudoit

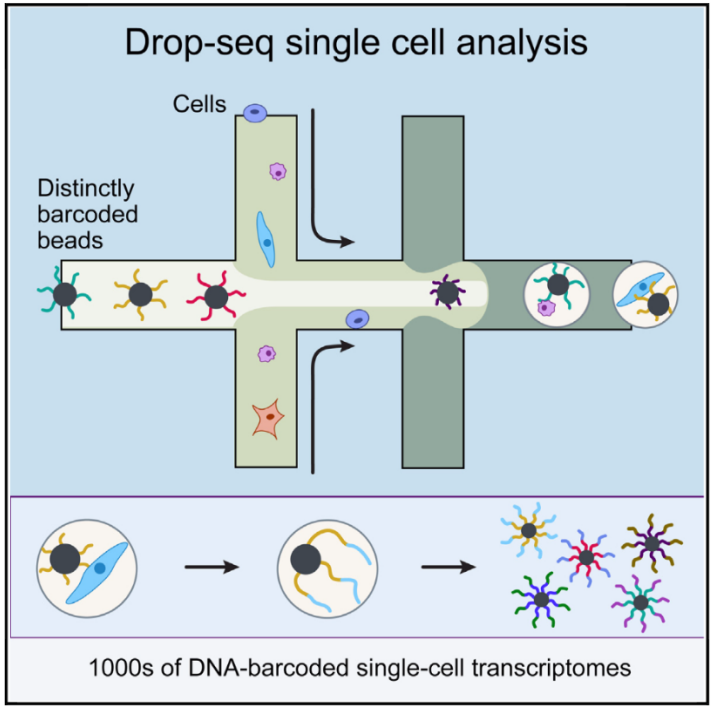
Lieven Clement

single-cell RNA-sequencing (scRNA-seq) is noisier than bulk RNA-seq

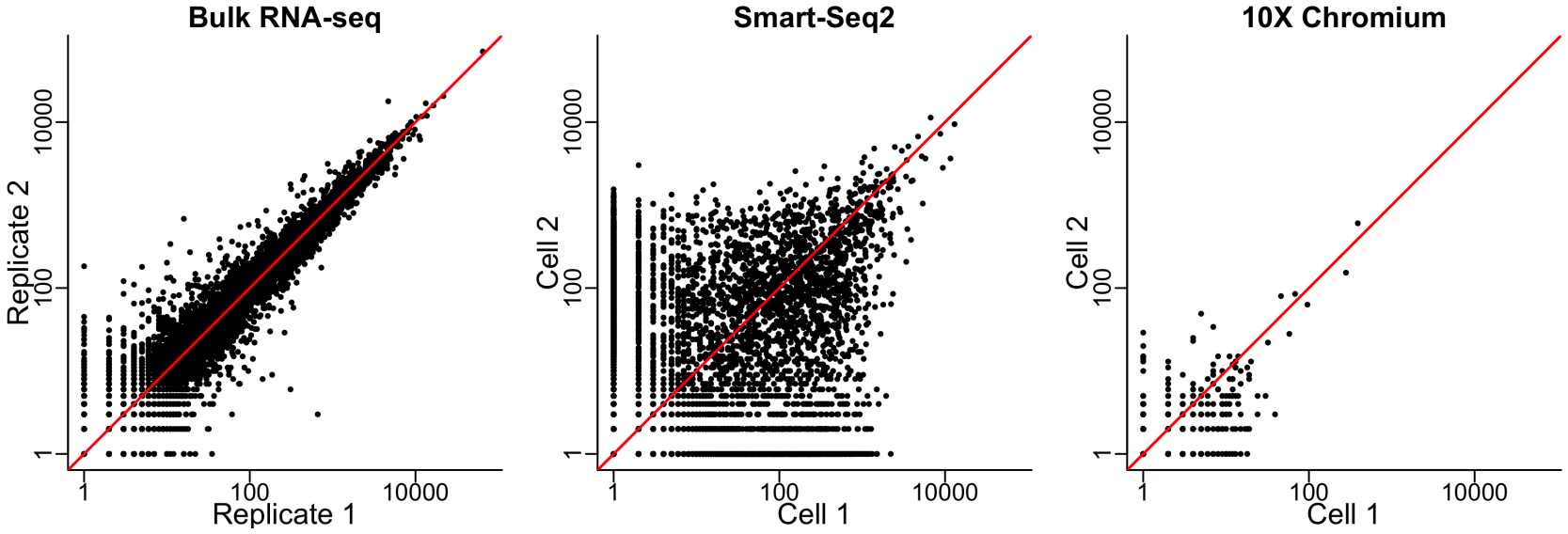


Single-cell RNA-seq protocols

- Full-length protocols (e.g., SMART-Seq2)
  - Cells must be isolated (manually, FACS, ...).
  - Library prep is typically plate-based; one well contains one cell.
- Droplet-based protocols (e.g., 10X, drop-seq)
  - Cells do not need to be isolated!
  - Cell-containing medium is mixed with bead-containing oil droplets.



single-cell RNA-sequencing (scRNA-seq) is noisier than bulk RNA-seq



data from [Pickrell et al. 2010, Fletcher et al. 2017, Zheng et al. 2017]

Bulk RNA-seq differential expression (DE) analysis

Popular methods (edgeR, DESeq2) adopt negative binomial (NB) models

$$
\begin{aligned}
y_{gi} &\sim NB(\mu_{gi}, \phi_g) \\
\log(\mu_{gi}) &= \eta_{gi} \\
\eta_{gi} &= \mathbf{X_i}\boldsymbol{\beta_g} + \log(O_i)
\end{aligned}
$$

with y_{gi} the expression count of gene g in sample i.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 1 2010, pages 139–140
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

Love *et al. Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8

Genome Biology

METHOD Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

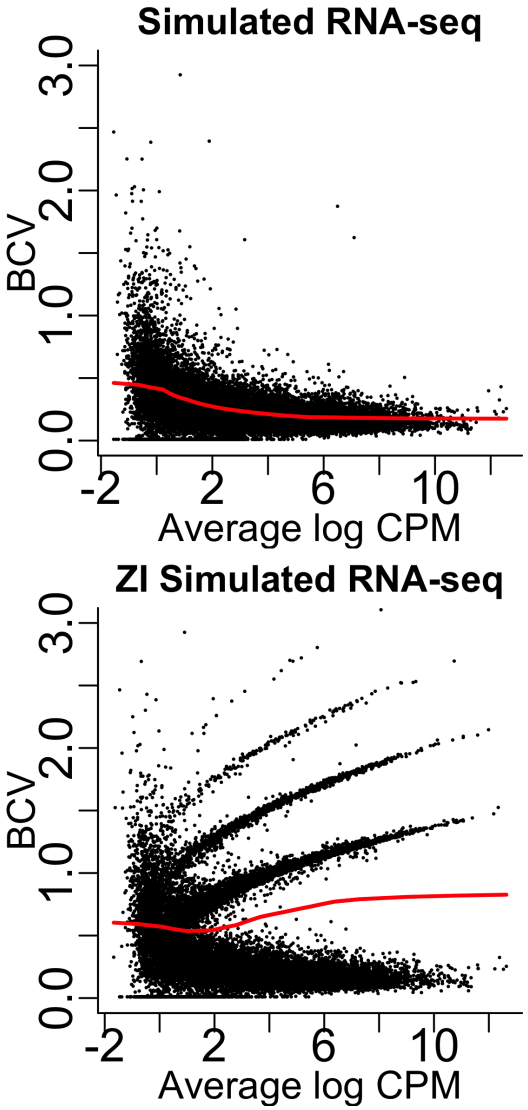
Bulk RNA-seq DE not worse than bespoke scRNA-seq tools

Jaakkoola *et al.* (2016), Bioinformatics:
"Our evaluations did not reveal systematic benefits of the currently available single-cell-specific methods."

Soneson & Robinson (2018), Nat. Meth.:
"We found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq."

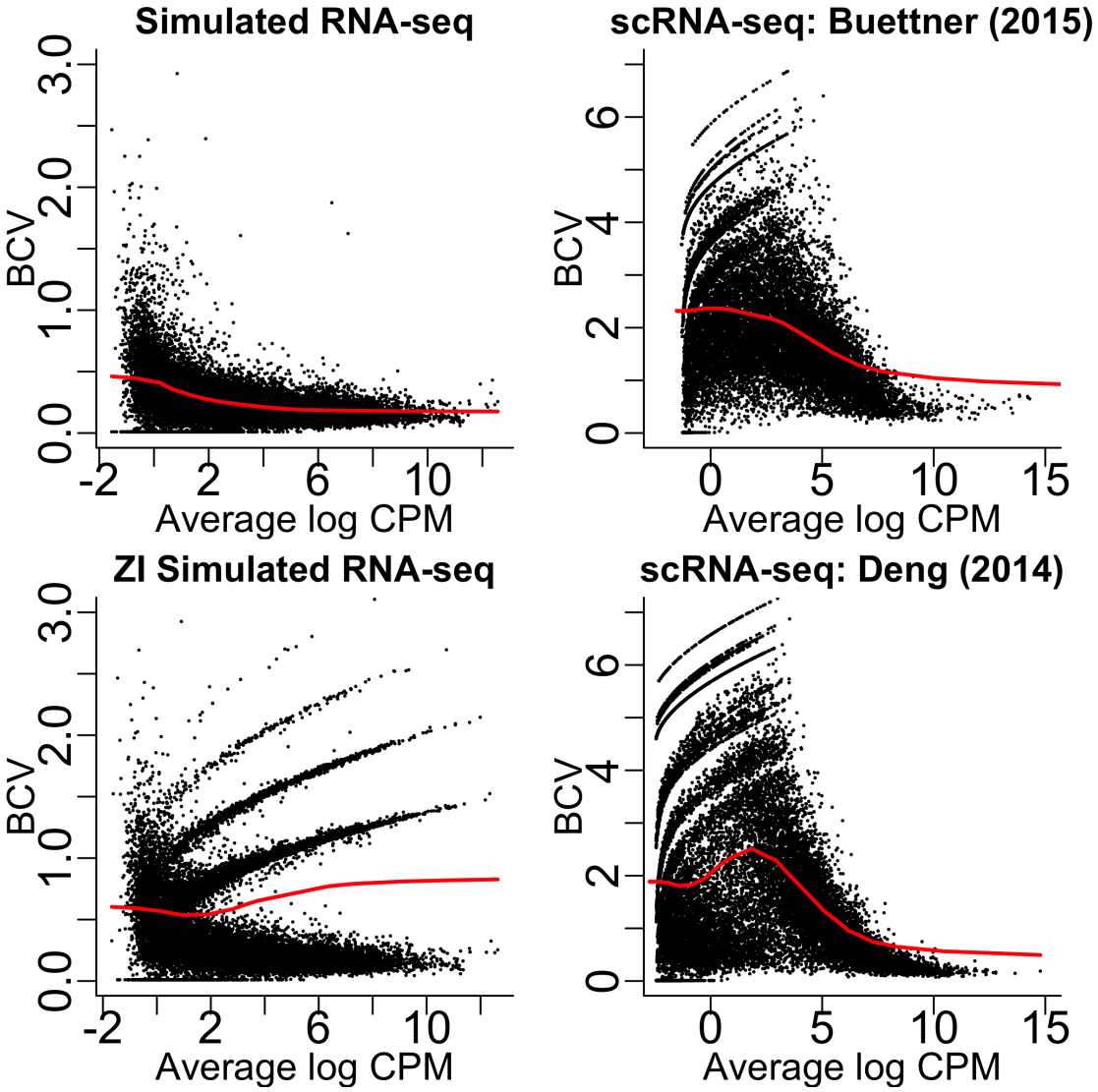
Bulk RNA-seq methods still break down due to ZI

Simulated (ZI-)bulk RNA-seq data using [Zhou *et al.* 2014] framework



Bulk RNA-seq methods still break down due to ZI

Simulated (ZI-)bulk RNA-seq data using [Zhou *et al.* 2014] framework



Observation weights unlock bulk RNA-seq tools towards zero inflation

Excess zeros observed $\rightarrow$ zero inflation
We propose to model counts with a zero inflated negative binomial (ZINB) distribution

$$f_{ZINB}(y_{gi}; \mu_{gi}, \phi_g, \pi_{gi}) = \pi_{gi}\delta + (1 - \pi_{gi}) f_{NB}(y_{gi}; \mu_{gi}, \phi_g). \quad (1)$$

Observation weights unlock bulk RNA-seq tools towards zero inflation

Excess zeros observed $\rightarrow$ zero inflation
We propose to model counts with a zero inflated negative binomial (ZINB) distribution

$$f_{ZINB}(y_{gi}; \mu_{gi}, \phi_g, \pi_{gi}) = \pi_{gi}\delta + (1 - \pi_{gi}) f_{NB}(y_{gi}; \mu_{gi}, \phi_g). \quad (1)$$

A ZINB model corresponds to a weighted NB where **observation weights** are posterior probabilities

$$w_{gi} = \frac{(1 - \pi_{gi}) f_{NB}(y_{gi}; \mu_{gi}, \phi_g)}{f_{ZINB}(y_{gi}; \mu_{gi}, \phi_g, \pi_{gi})} \quad (2)$$

Observation weights unlock bulk RNA-seq tools towards zero inflation

Excess zeros observed $\rightarrow$ zero inflation
We propose to model counts with a zero inflated negative binomial (ZINB) distribution

$$f_{ZINB}(y_{gi}; \mu_{gi}, \phi_g, \pi_{gi}) = \pi_{gi}\delta + (1 - \pi_{gi}) f_{NB}(y_{gi}; \mu_{gi}, \phi_g). \quad (1)$$

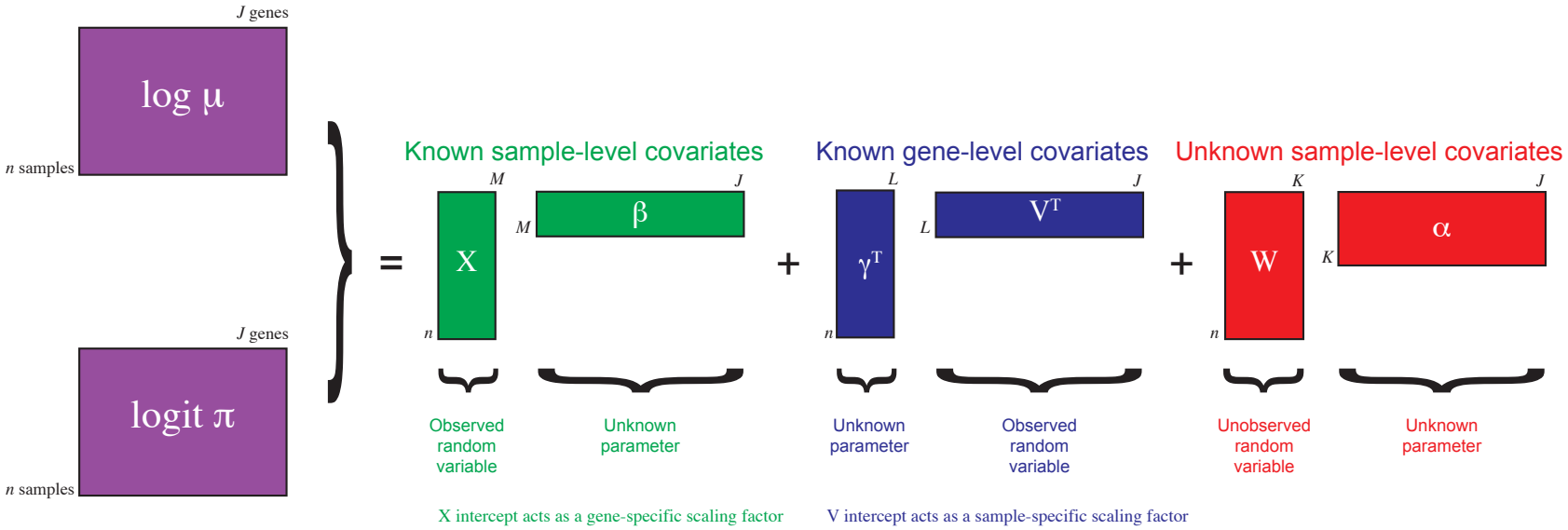
A ZINB model corresponds to a weighted NB where **observation weights** are posterior probabilities

$$w_{gi} = \frac{(1 - \pi_{gi}) f_{NB}(y_{gi}; \mu_{gi}, \phi_g)}{f_{ZINB}(y_{gi}; \mu_{gi}, \phi_g, \pi_{gi})} \quad (2)$$

Weights are used to unlock RNA-seq NB models (`edgeR`, `DESeq2`) for zero inflation [Van den Berge*, Perraudeau* *et al.*, 2018].

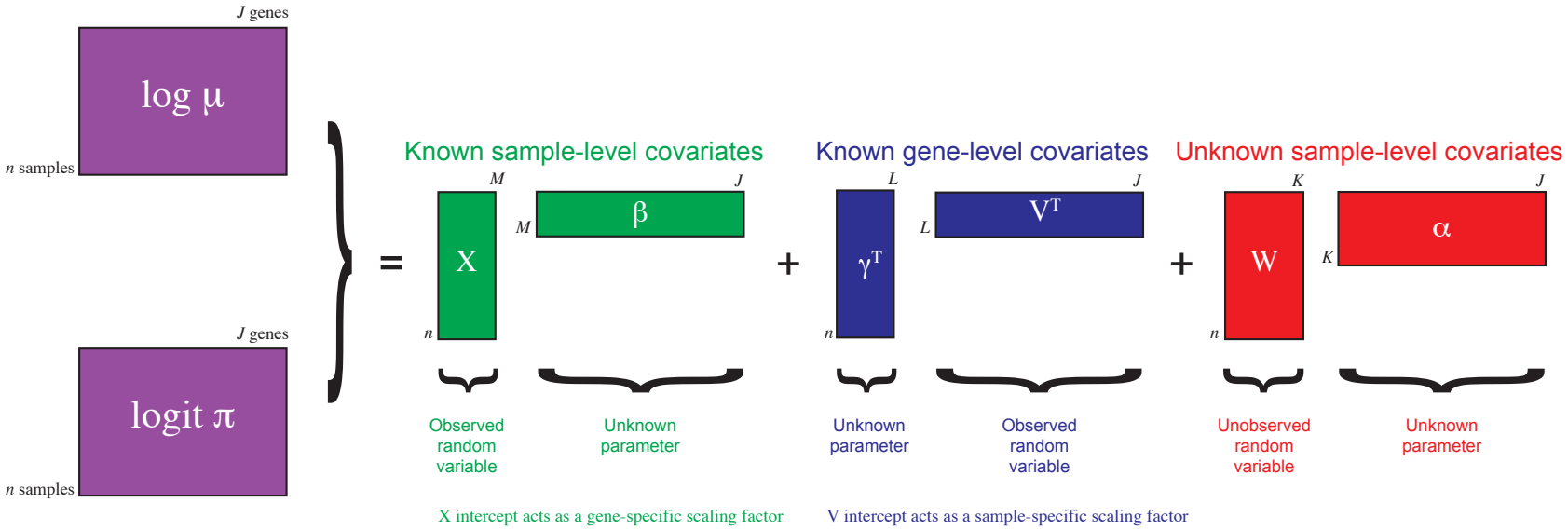
zinbwave can be used to fit ZINB models in scRNA-seq

Estimation of the ZINB parameters using penalized likelihood implemented in the ZINB-WaVE model [Risso et al. 2018]
Bioconductor: `http://bioconductor.org/packages/zinbwave/`



zinbwave can be used to fit ZINB models in scRNA-seq

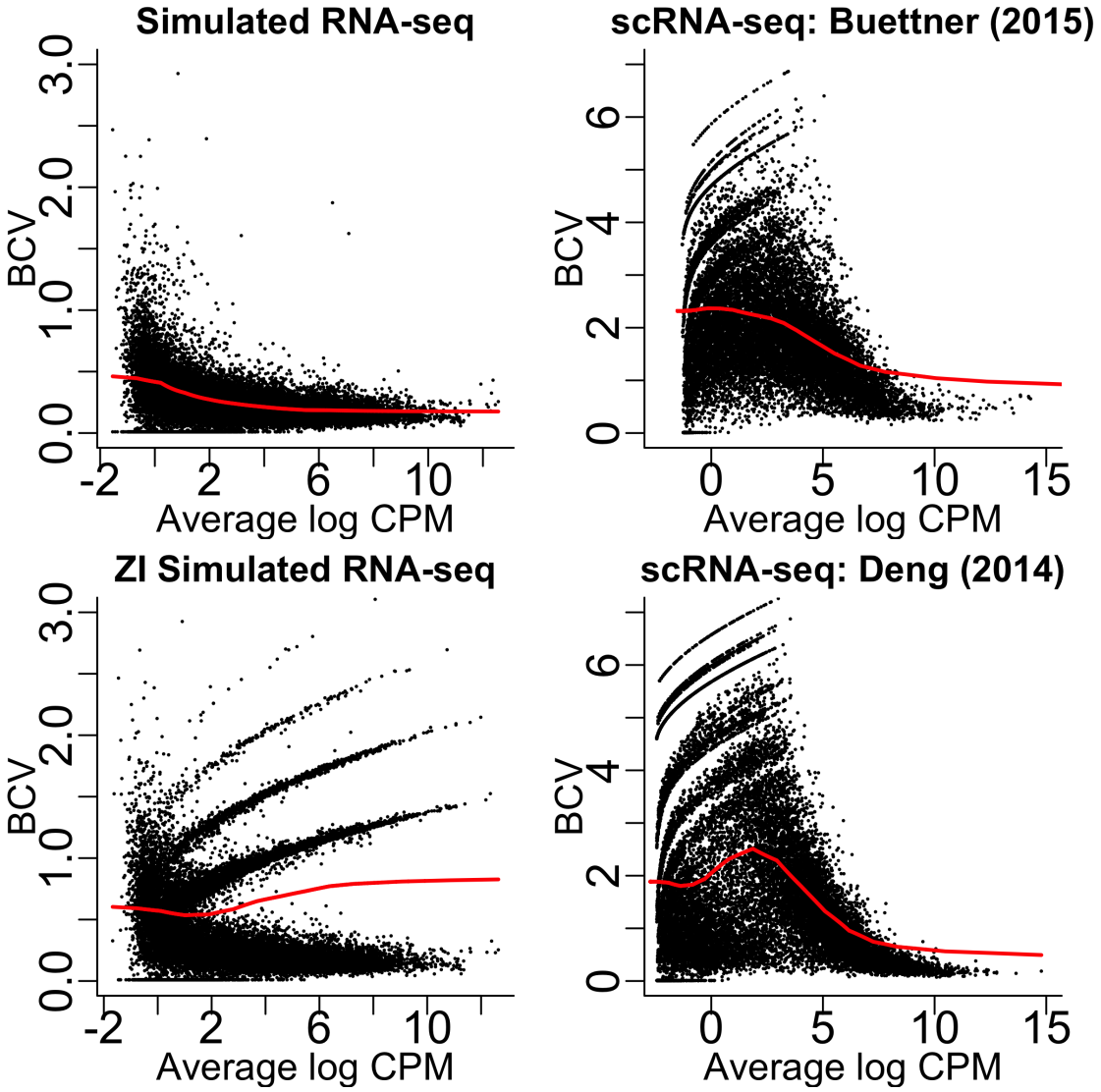
Estimation of the ZINB parameters using penalized likelihood implemented in the ZINB-WaVE model [Risso et al. 2018]
Bioconductor: `http://bioconductor.org/packages/zinbwave/`



Alternatively: EM-algorithm (see last couple of slides)

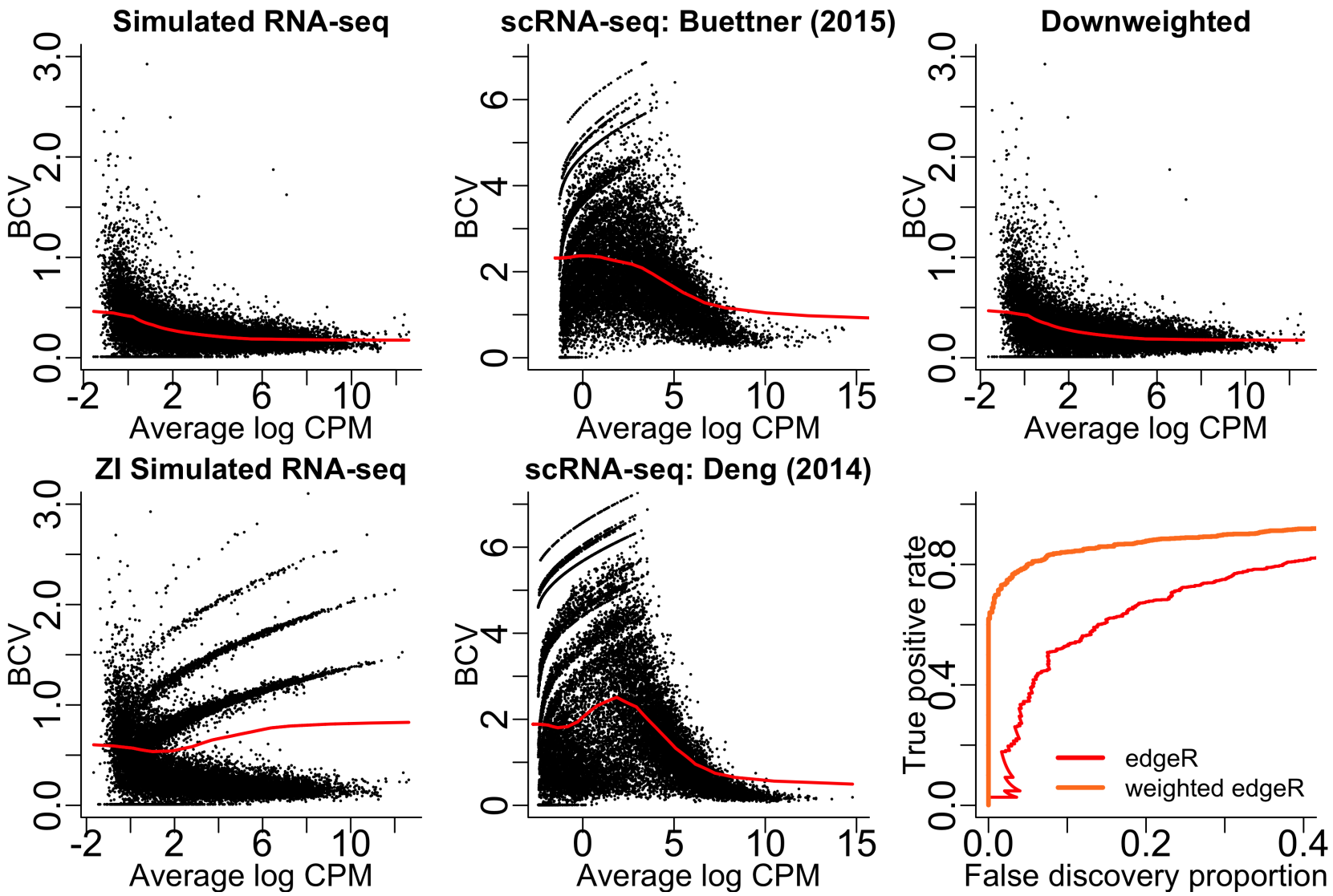
Downweighting excess zeros recovers mean-variance trend, resulting in high power

Simulated (ZI-)bulk RNA-seq data using [Zhou *et al.* 2014] framework



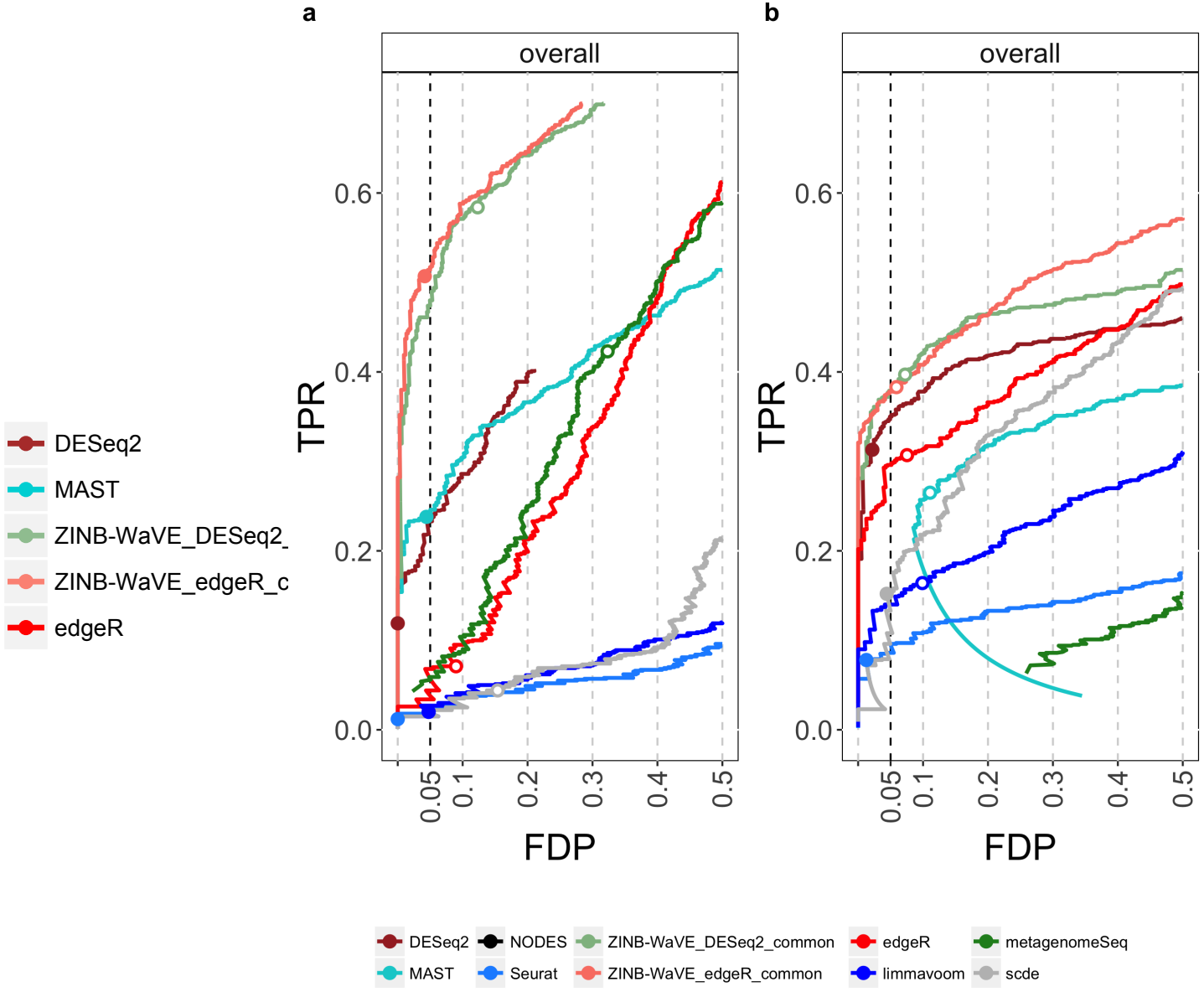
Downweighting excess zeros recovers mean-variance trend, resulting in high power

Simulated (ZI-)bulk RNA-seq data using [Zhou *et al.* 2014] framework



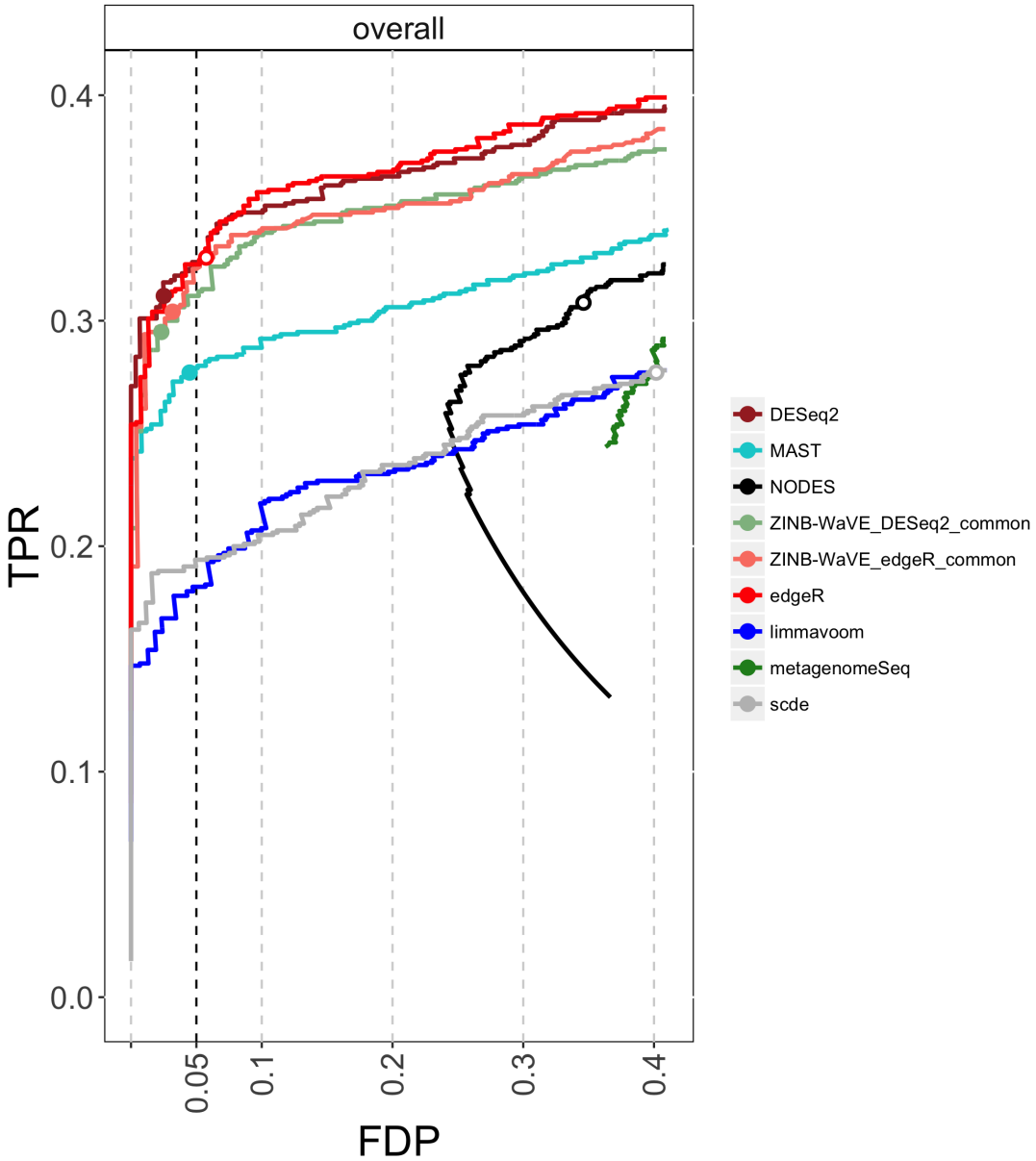
High power, good FDR control in scRNA-seq simulations

Full-length protocols

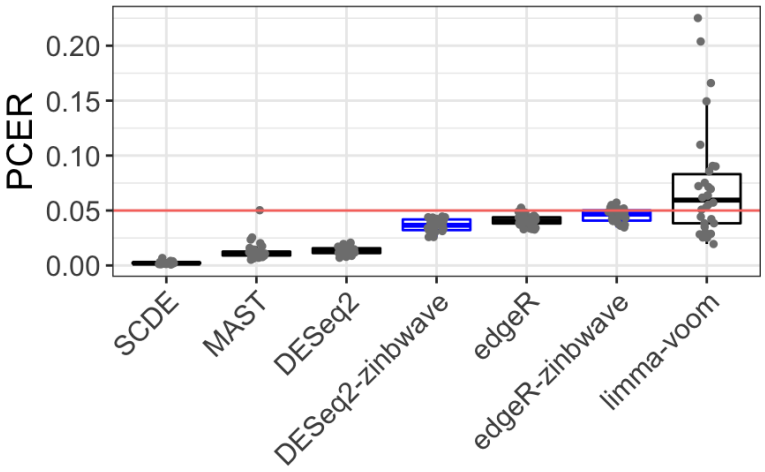


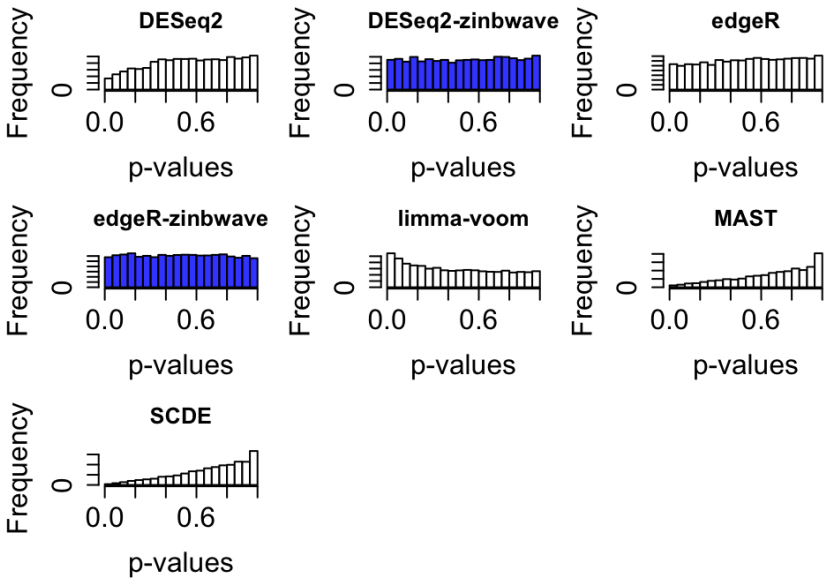
High power, good FDR control in scRNA-seq simulations

Droplet-based protocols, e.g. 10X Genomics, Drop-seq



Mock comparisons on real data show good FPR control

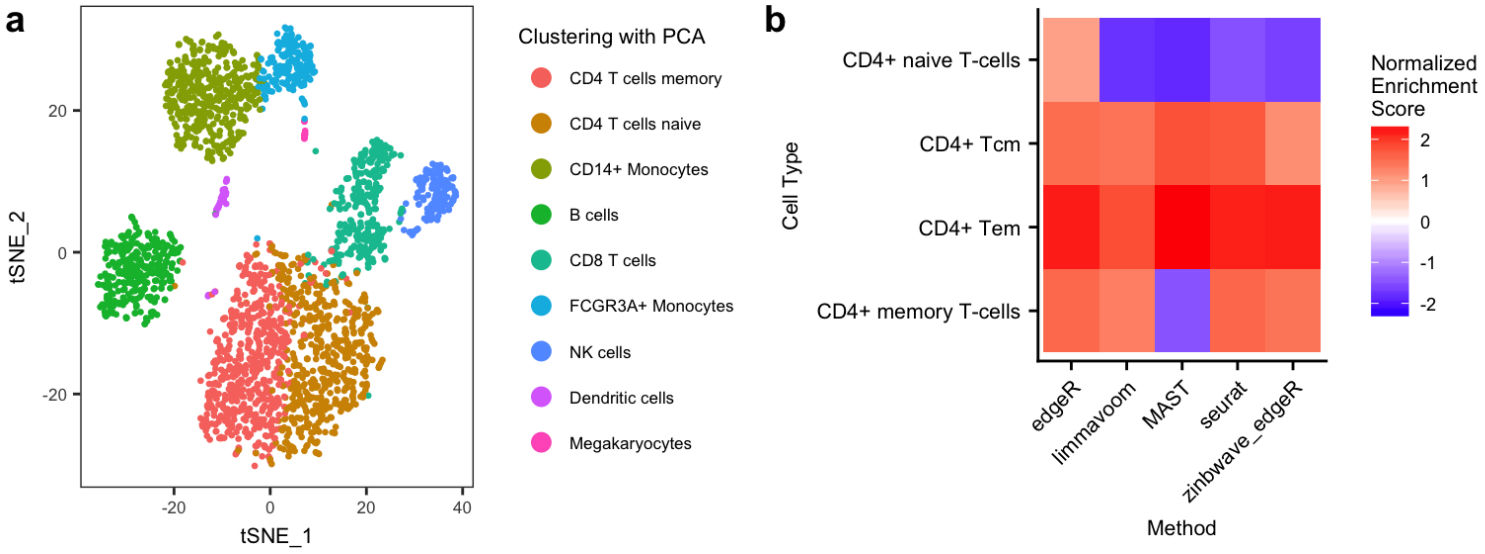




Non-UMI dataset on 622 neuronal cells from [Usoskin et al. 2015].
45 vs. 45 mock comparisons.

Downweighting leads to biologically meaningful results

10X Genomics PBMC dataset, preprocessed using tutorial from `Seurat`.



Method is implemented in `zinbwave` Bioc package

- `computeObservationalWeights` for weights calculation
- `edgeR:glmWeightedF` for ZI-adjusted inference
- `DESeq2:nbinomWaldTest` and `nbinomLRT` for ZI-adjusted inference
- Tutorial available in `zinbwave` vignette

METHOD Open Access

Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications

Koen Van den Berge[1,2†], Fanny Perraudeau[3†], Charlotte Soneson[4,5], Michael I. Love[6], Davide Risso[7], Jean-Philippe Vert[8,9,10,11], Mark D. Robinson[4,5], Sandrine Dudoit[3,12†] and Lieven Clement[1,2†*]

What follows are some slides on the EM algorithm used in the `zingeR` method.

zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications

Koen Van den Berge, Charlotte Soneson, Michael I. Love, Mark D. Robinson, Lieven Clement

doi: https://doi.org/10.1101/157982

This article is a preprint and has not been peer-reviewed [what does this mean?].

A zero-inflated negative binomial model

Distribution of counts y for gene g over samples i

$$y_{gi} \sim \pi_i \delta + (1 - \pi_i) f_{NB}(\mu_{gi}, \phi_g)$$

i.e. mixture distribution between point-mass at zero and negative binomial

Log-likelihood

$$l(y_{gi}) = \sum_i log \left\{ \pi_i \delta + (1 - \pi_i) f_{NB}(\mu_{gi}, \phi_g) \right\}$$

does not factorize
$\rightarrow$ very difficult to maximize!

Fitting a mixture distribution with EM

Estimate mixture using EM-algorithm: introduce latent variable $Z_{gi} \sim B(\pi_{gi})$ to assign zeros to the zero-inflation or count component. The joint density becomes

$$f(y_{gi}, z_{gi}) = f(y_{gi}|z_{gi}) f(z_{gi}) = [\pi_i \delta]^{z_{gi}} \left[(1 - \pi_i) f_{NB}(\mu_{gi}, \phi_g)\right]^{(1-z_{gi})}$$

Fitting a mixture distribution with EM

Estimate mixture using EM-algorithm: introduce latent variable $Z_{gi} \sim B(\pi_{gi})$ to assign zeros to the zero-inflation or count component. The joint density becomes

$$f(y_{gi}, z_{gi}) = f(y_{gi}|z_{gi})f(z_{gi}) = [\pi_i \delta]^{z_{gi}} \, [(1-\pi_i) f_{NB}(\mu_{gi}, \phi_g)]^{(1-z_{gi})}$$

Maximization of expected log-likelihood given the data:

$$Q = E(l(y_{gi}, z_{gi})|y_{gi})$$

$$= E(z_{gi}|y_{gi}) log \pi_i + E(z_{gi}|y_{gi}) log \delta + [1 - E(z_{gi}|y_{gi})] log(1-\pi_i) + [1 - E(z_{gi}|y_{gi})] log[f_{NB}(\mu_{gi}, \phi_g)]$$

1. **E-step:** Calculate expected likelihood
2. **M-step:** Maximize expected likelihood

EM-algorithm

E-step

- Calculate posterior probability that a zero belongs to zero-inflation component

$$E(z_{gi}|y_{gi}) = \frac{\hat{\pi}_i I(y_{gi} = 0)}{\hat{\pi}_i I(y_{gi} = 0) + (1 - \hat{\pi}_i) f_{NB}(y_{gi}; \hat{\mu}_{gi}, \hat{\phi}_g)}$$

EM-algorithm

E-step

- Calculate posterior probability that a zero belongs to zero-inflation component

$$E(z_{gi}|y_{gi}) = \frac{\hat{\pi}_i I(y_{gi} = 0)}{\hat{\pi}_i I(y_{gi} = 0) + (1 - \hat{\pi}_i) f_{NB}(y_{gi}; \hat{\mu}_{gi}, \hat{\phi}_g)}$$

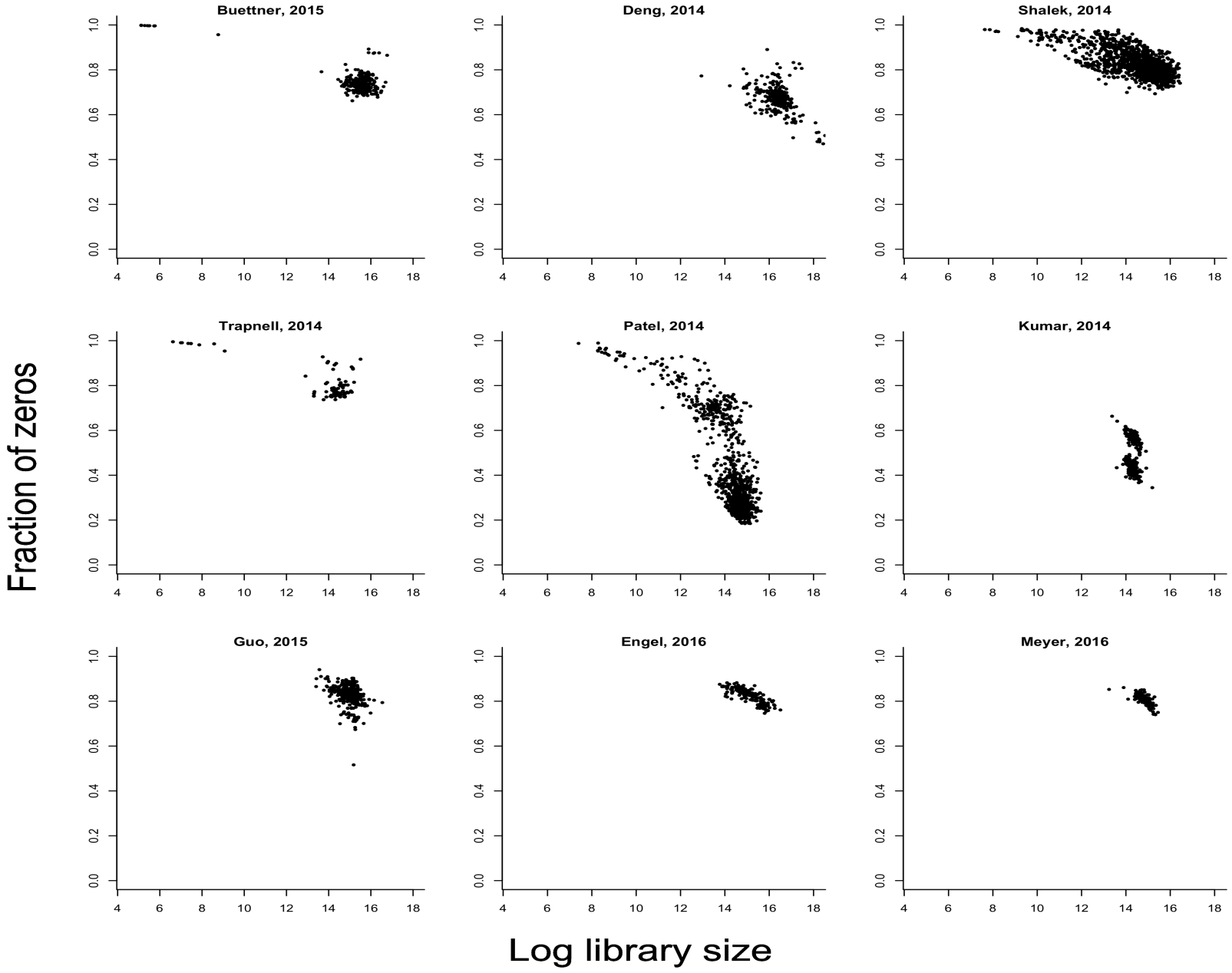
M-step

- Estimate NB component parameters μ_{gi} and ϕ_g using edgeR
  - Incorporate observation-level weights $w_{gi} = 1 - E_y(z_{gi})$ for counts y_{gi}
  - Because maximizing ZINB likelihood for NB model parameters is equivalent to maximizing a weighted NB likelihood.
- Estimate π_i using logistic regression model

$$log\left\{\frac{\pi_i}{1 - \pi_i}\right\} = \beta_0 + \beta_1 N_i$$

with N_i log library size of sample i

Why we use logistic regression with library size in the EM



Case study: Islam *et al.* (2014)

- Single-cell RNA-seq (scRNA-seq) allowed the study of 'sparse' cell populations.
- One of the first datasets we worked with was from Islam *et al.* (2014). It demonstrates scRNA-seq for 85 cells consisting of two cell populations in mouse: embryonic stem cells and fibroblasts.
- This paper was one of the first scRNA-seq studies and motivated our method development.
- Link to paper: `https://www.ncbi.nlm.nih.gov/pubmed/21543516`