Some challenges with single-cell gene expression data

Jean-Philippe Vert

MINES ParisTech

institut Curie
Together, let's beat cancer.

ENS

PSL

Gene expresion

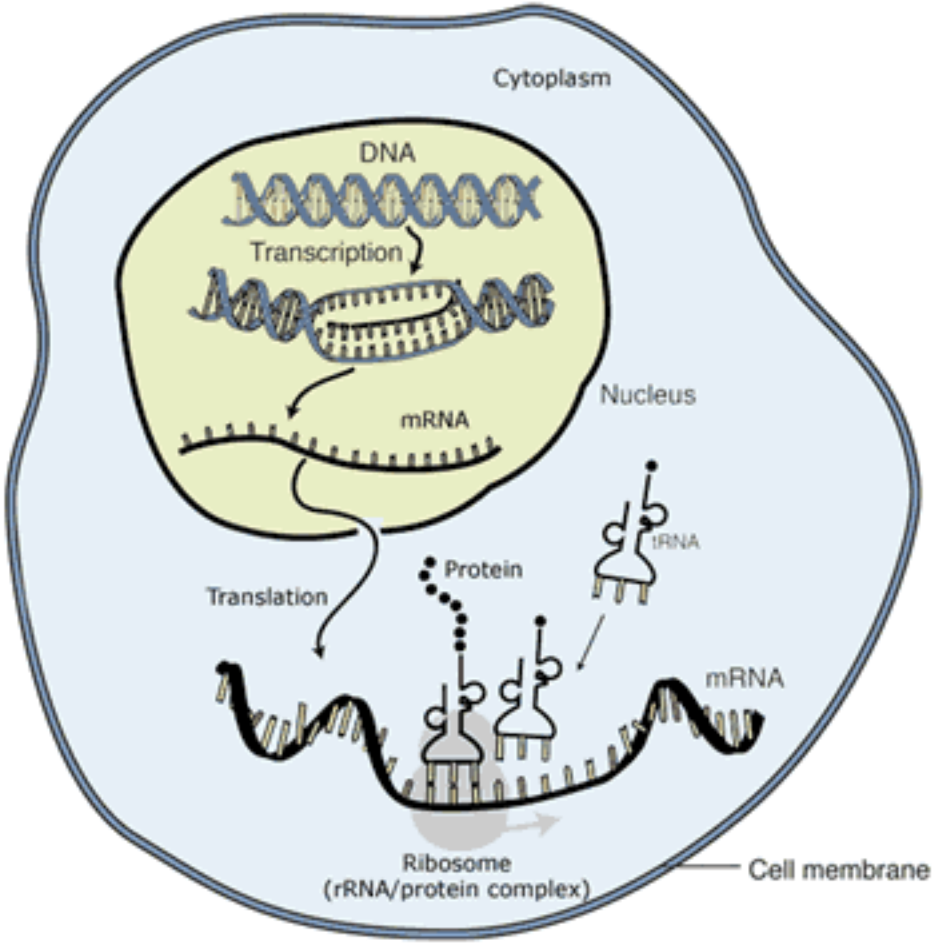
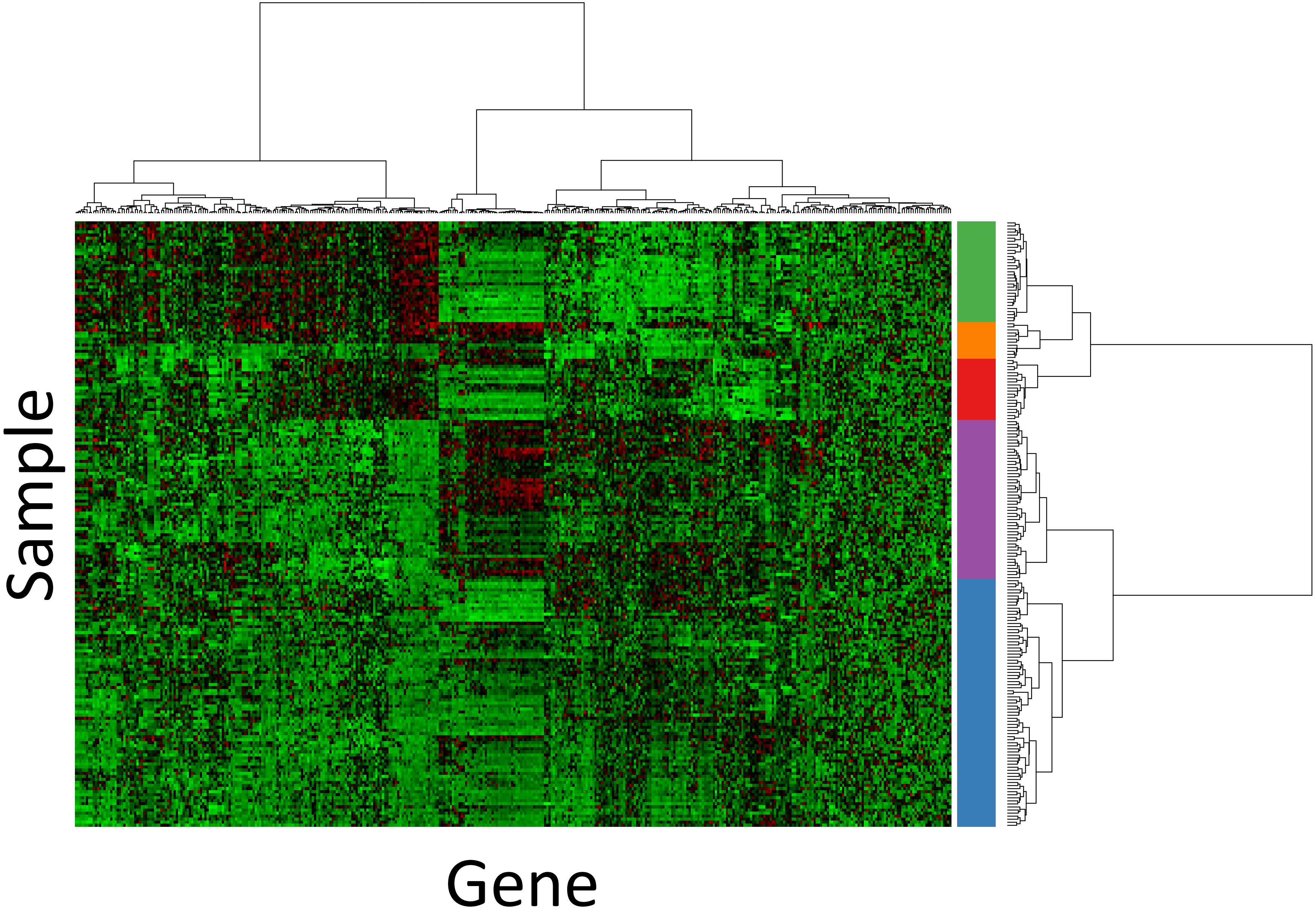


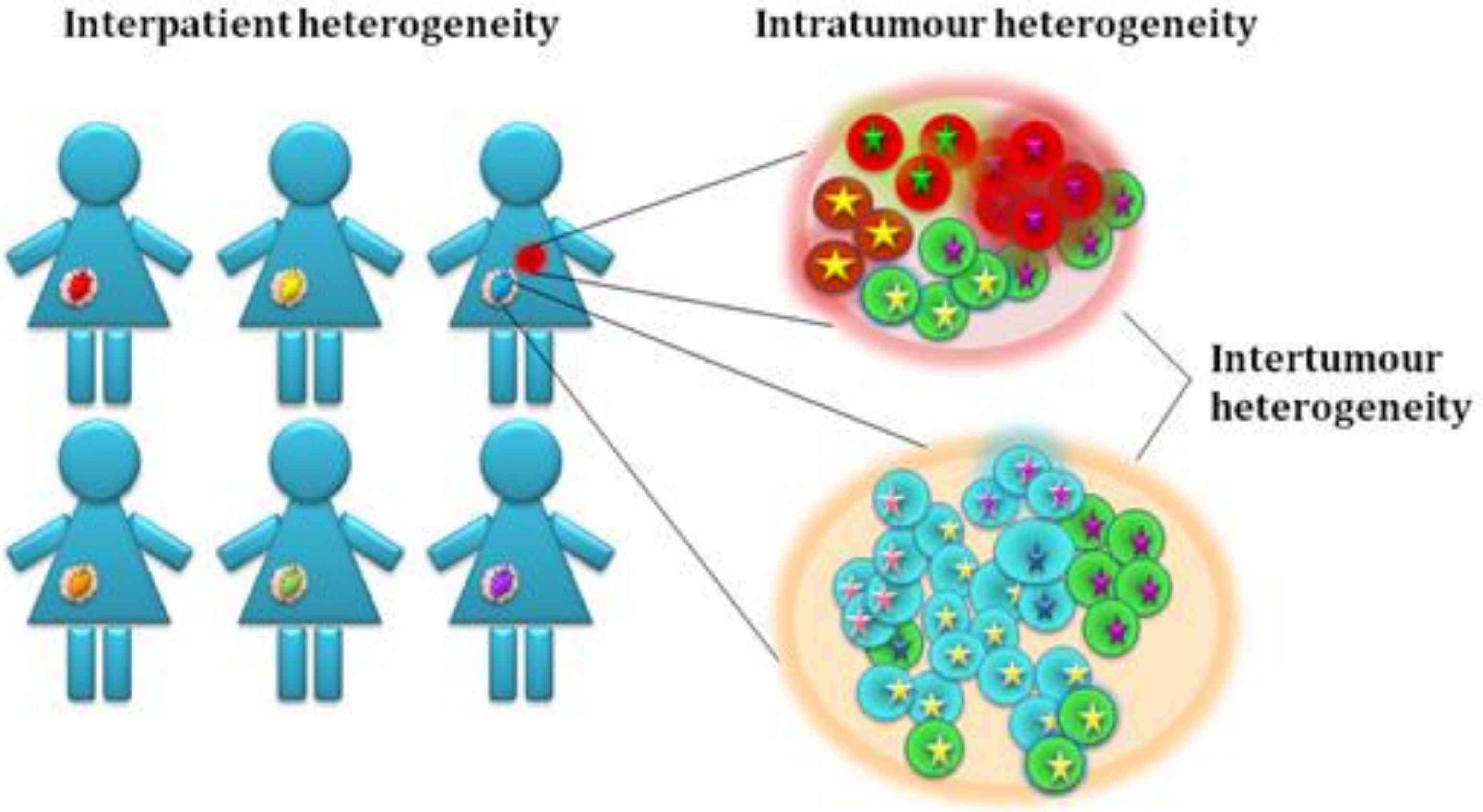
Image adapted from: National Human Genome Research Institute.

~20k genes
in human genome

« Bulk » gene expression



Each sample can be a complex mixture of different cells

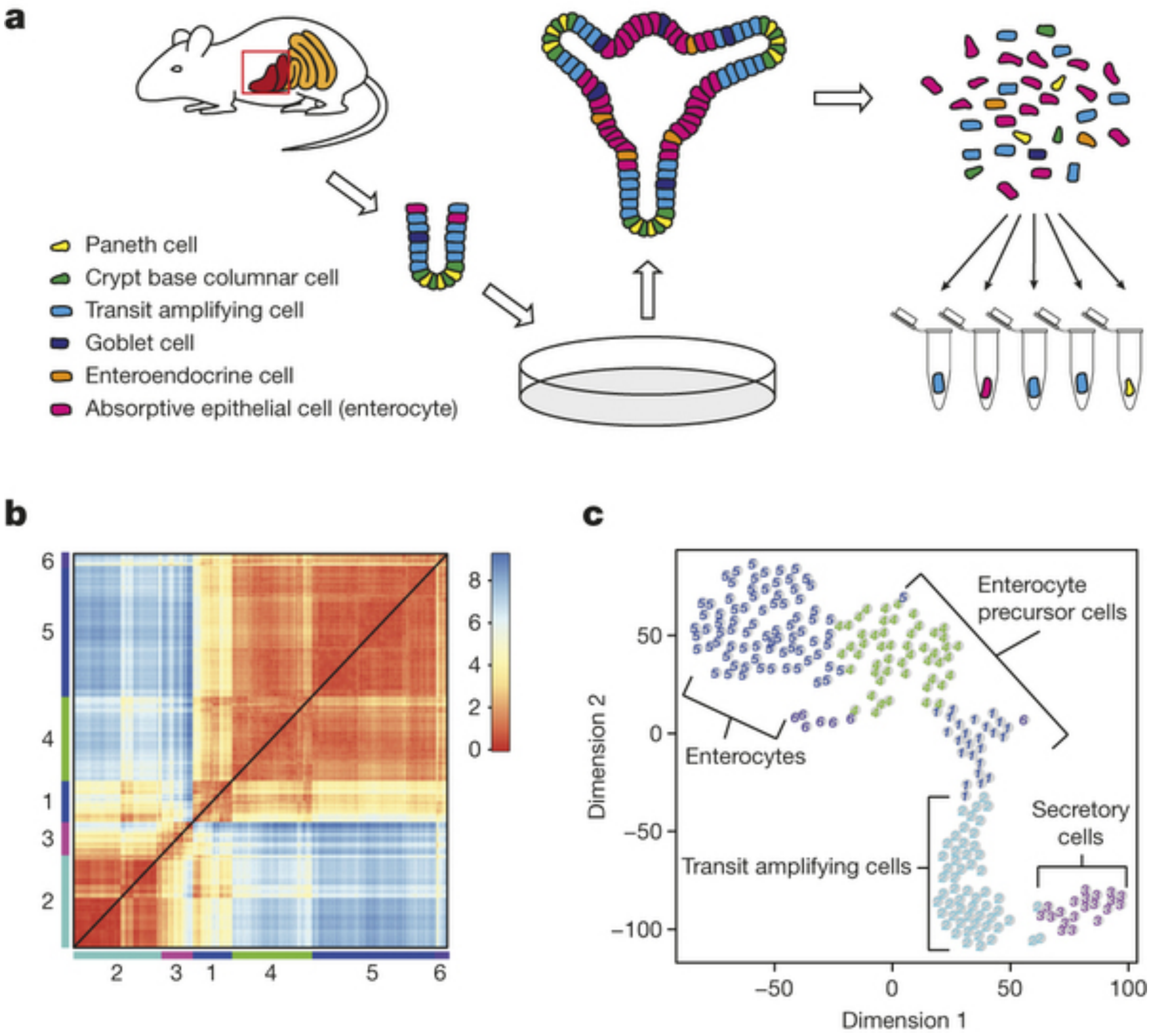

Interpatient heterogeneity
Intratumour heterogeneity
Intertumour heterogeneity

From Oslo University Hospital web page

« Bulk » vs « single-cell »

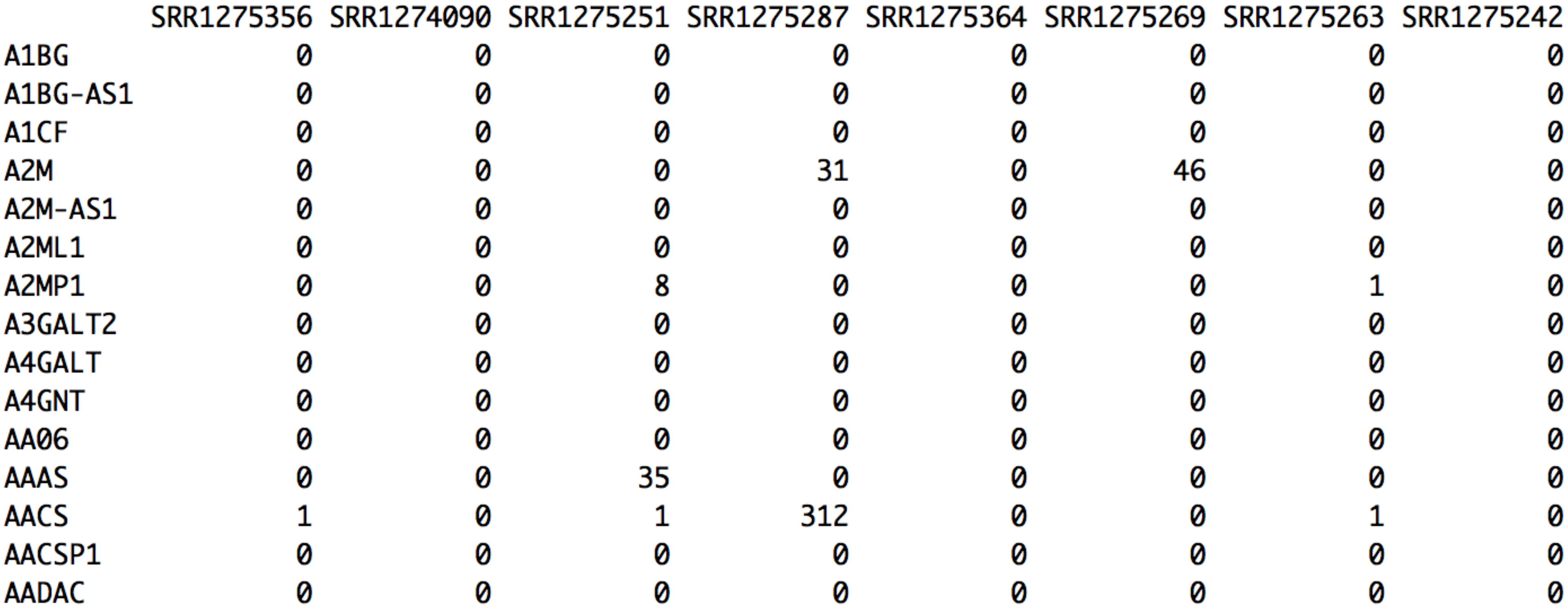
Inspired from slides of A. Regev

Single-cell RNA-seq



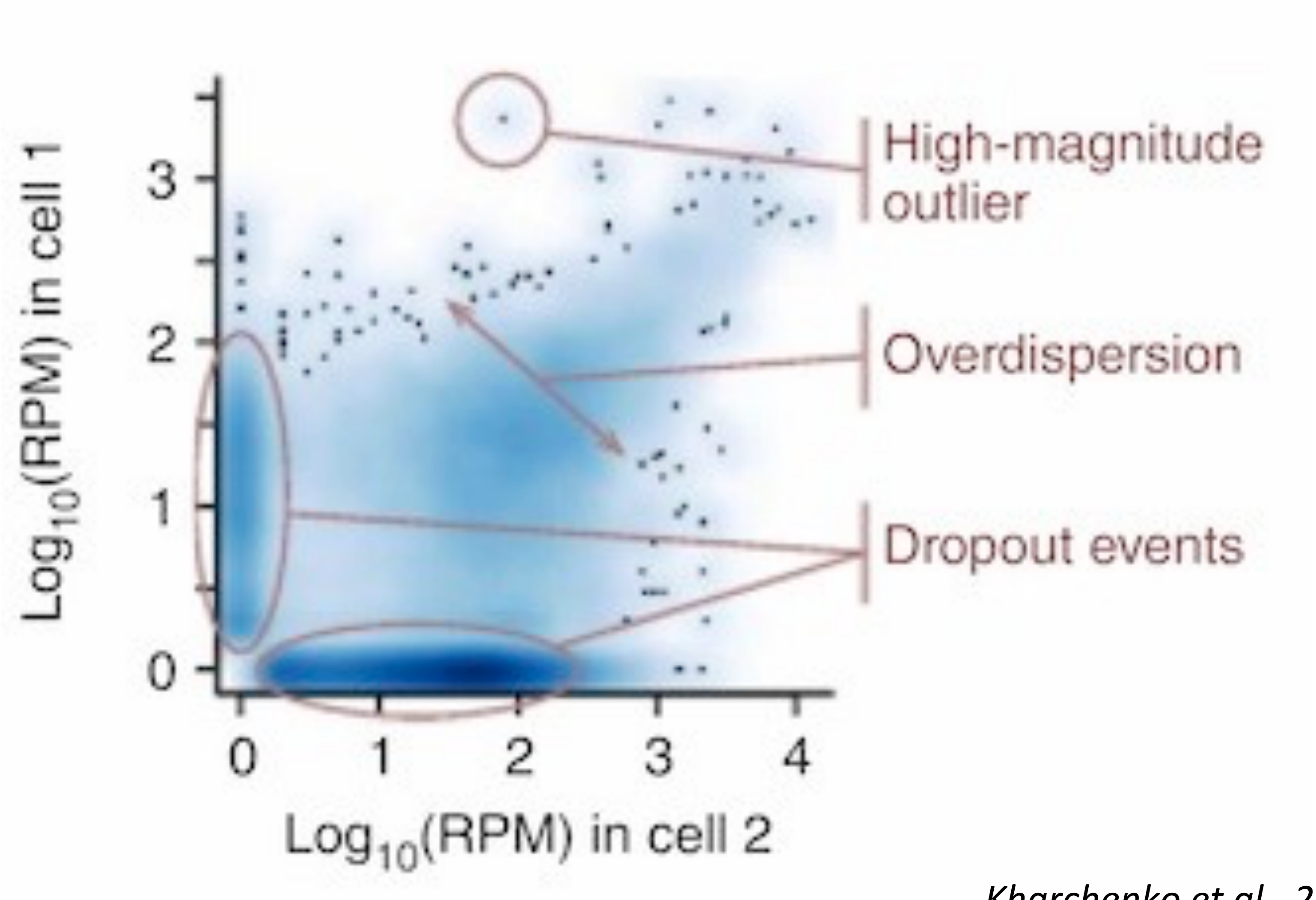
(Grün et al 2015)

The data



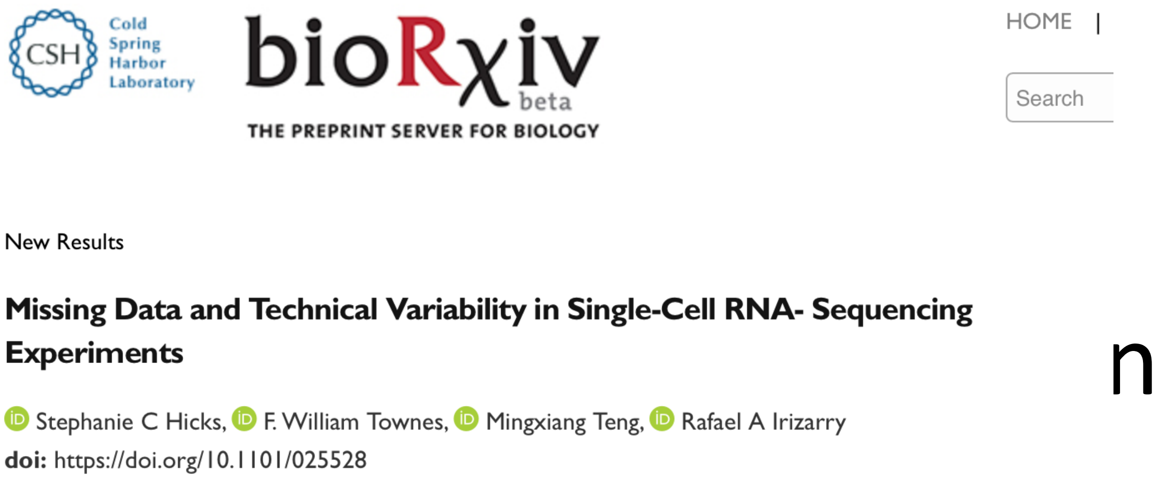
```
         SRR1275356 SRR1274090 SRR1275251 SRR1275287 SRR1275364 SRR1275269 SRR1275263 SRR1275242
A1BG              0          0          0          0          0          0          0          0
A1BG-AS1          0          0          0          0          0          0          0          0
A1CF              0          0          0          0          0          0          0          0
A2M               0          0          0         31          0         46          0          0
A2M-AS1           0          0          0          0          0          0          0          0
A2ML1             0          0          0          0          0          0          0          0
A2MP1             0          0          8          0          0          0          1          0
A3GALT2           0          0          0          0          0          0          0          0
A4GALT            0          0          0          0          0          0          0          0
A4GNT             0          0          0          0          0          0          0          0
AA06              0          0          0          0          0          0          0          0
AAAS              0          0         35          0          0          0          0          0
AACS              1          0          1        312          0          0          1          0
AACSP1            0          0          0          0          0          0          0          0
AADAC             0          0          0          0          0          0          0          0
```

Dropout, overdispersion…



Kharchenko et al., 2014

Batch effects, normalization…



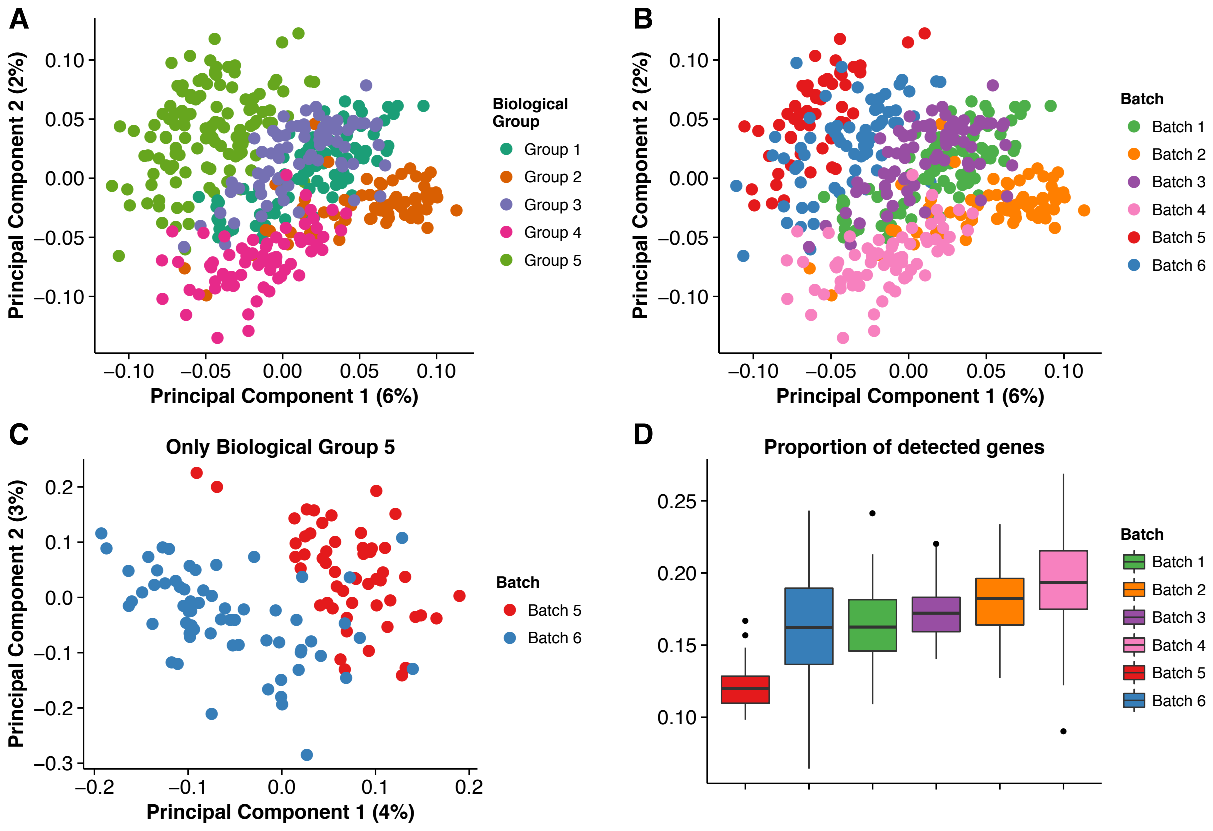
Cold Spring Harbor Laboratory
bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY
HOME |
Search

New Results

Missing Data and Technical Variability in Single-Cell RNA- Sequencing Experiments

Stephanie C Hicks, F. William Townes, Mingxiang Teng, Rafael A Irizarry

doi: https://doi.org/10.1101/025528

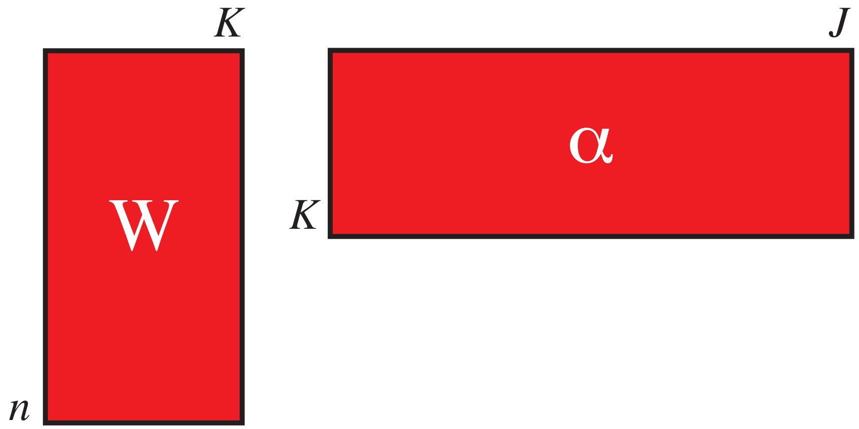


Some challenges

- Normalize for total count per cell?
- Remove unwanted variations? (batches, cell cycle, GC content, …)
- Distances between transcription profiles?
- Clustering / Visualization?
- Differential expression?
- Supervised classification?
- …

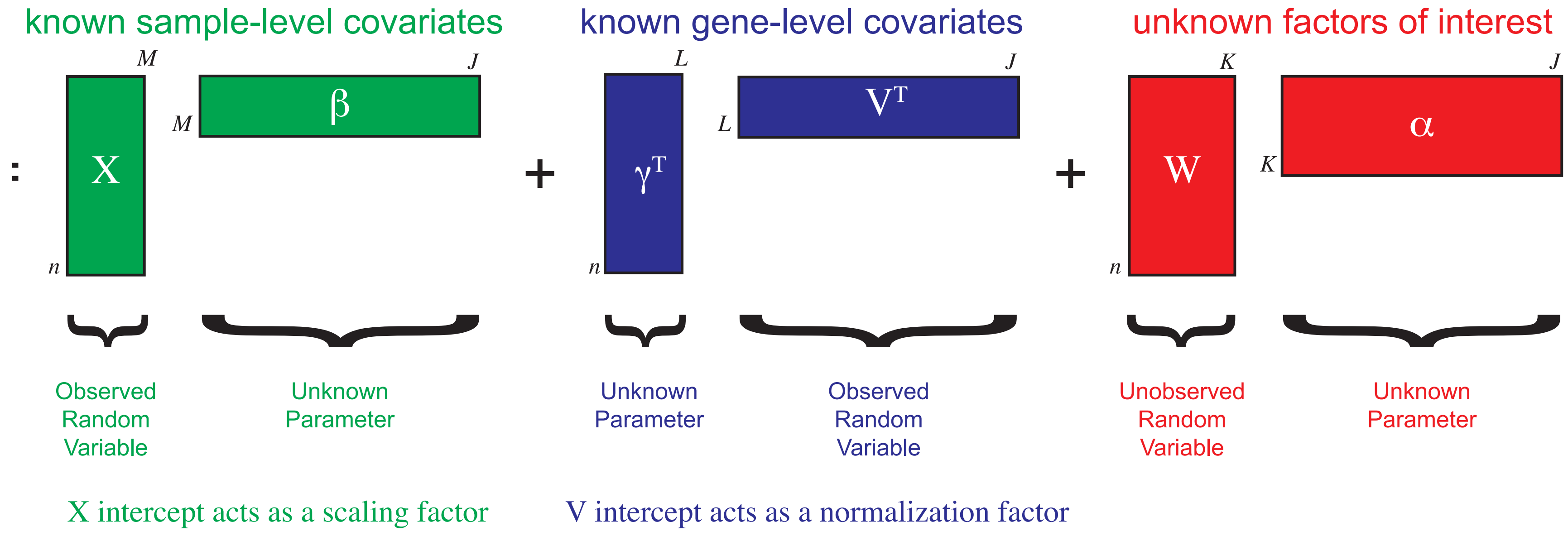
Dimension reduction (PCA/SVD)

$$E[Y] = W\alpha$$



Including known covariates (RUV)

$$E[Y] = X\beta + V\gamma + W\alpha$$

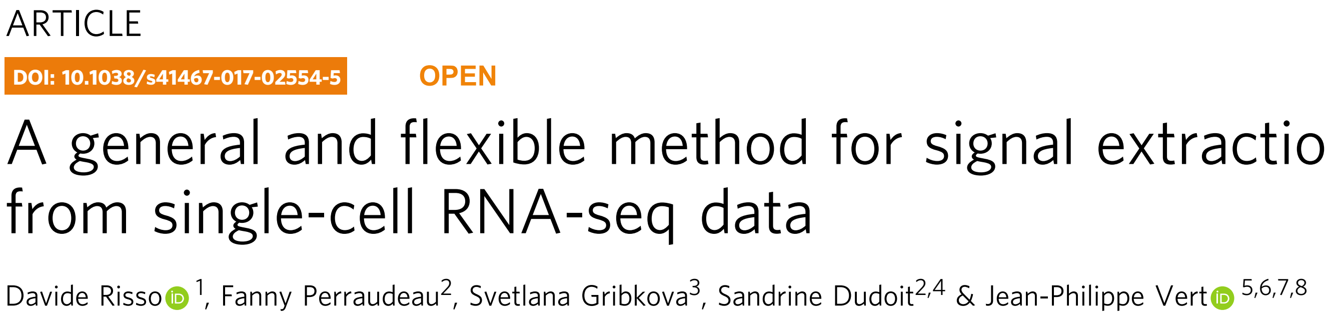

known sample-level covariates
known gene-level covariates
unknown factors of interest
M
J
M
n
X
β
:
+
L
J
L
n
γ^T
V^T
+
K
J
K
n
W
α
Observed Random Variable
Unknown Parameter
Unknown Parameter
Observed Random Variable
Unobserved Random Variable
Unknown Parameter
X intercept acts as a scaling factor
V intercept acts as a normalization factor

Jacob et al. (2013), Gagnon-Bartsch et al. (2013), Risso et al. (2014)

How to adapt PCA/SVD/RUV to scRNA-seq data?

$$E[Y] = X\beta + V\gamma + W\alpha$$

- discrete, non-Gaussian data
- dropouts



ARTICLE

DOI: 10.1038/s41467-017-02554-5

OPEN

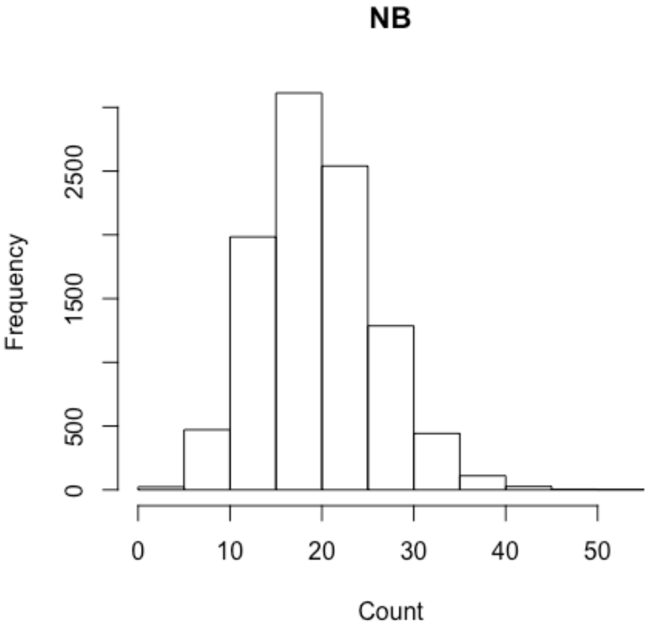
A general and flexible method for signal extraction from single-cell RNA-seq data

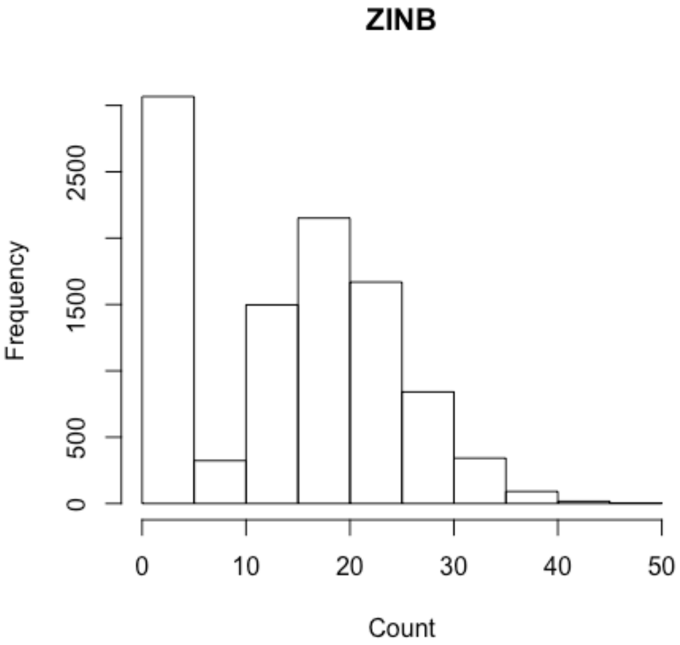
Davide Risso[1], Fanny Perraudeau[2], Svetlana Gribkova[3], Sandrine Dudoit[2,4] & Jean-Philippe Vert[5,6,7,8]

ZINB distribution to model a count

« Zero-Inflated Negative Binomial »

$$f_{NB}(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{y}, \quad \forall y \in \mathbb{N}.$$





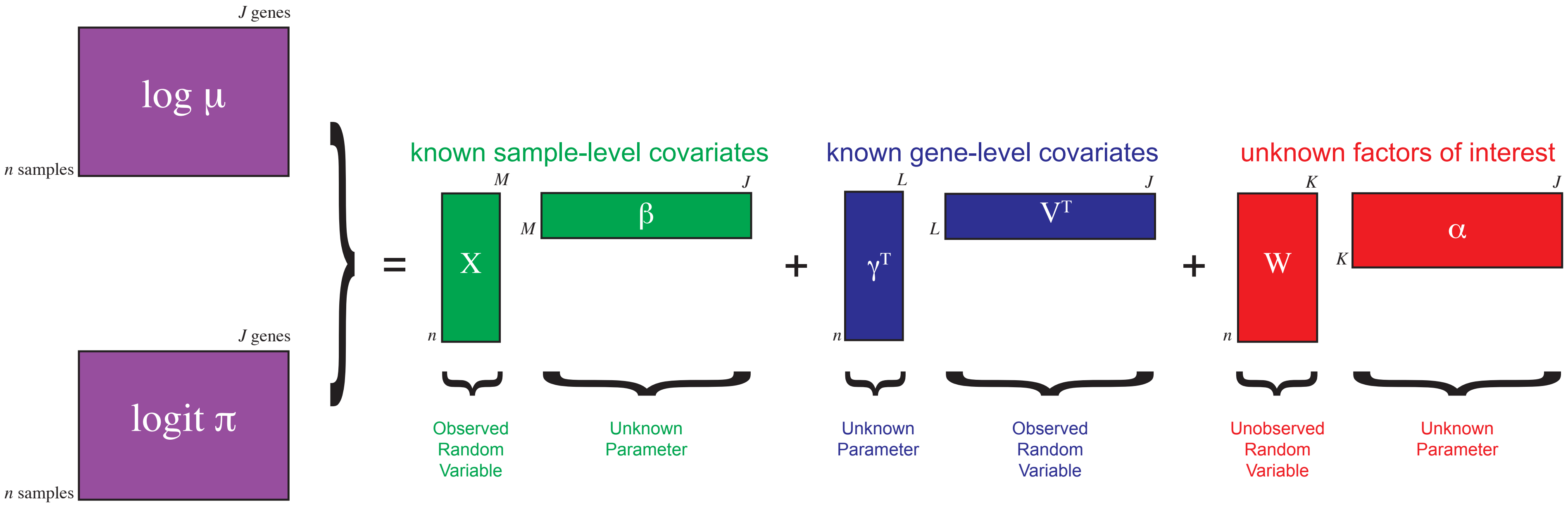
$$f_{ZINB}(y;\mu,\theta,\pi) = \pi\delta_0(y) + (1-\pi)f_{NB}(y;\mu,\theta), \quad \forall y \in \mathbb{N},$$

ZINB-WaVE model

$$\ln(\mu_{ij}) = \left(X\beta_\mu + (V\gamma_\mu)^\top + W\alpha_\mu + O_\mu\right)_{ij}$$

$$\text{logit}(\pi_{ij}) = \left(X\beta_\pi + (V\gamma_\pi)^\top + W\alpha_\pi + O_\pi\right)_{ij}$$

$$\ln(\theta_{ij}) = \zeta_j \,,$$



X intercept acts as a scaling factor

V intercept acts as a normalization factor

Usage

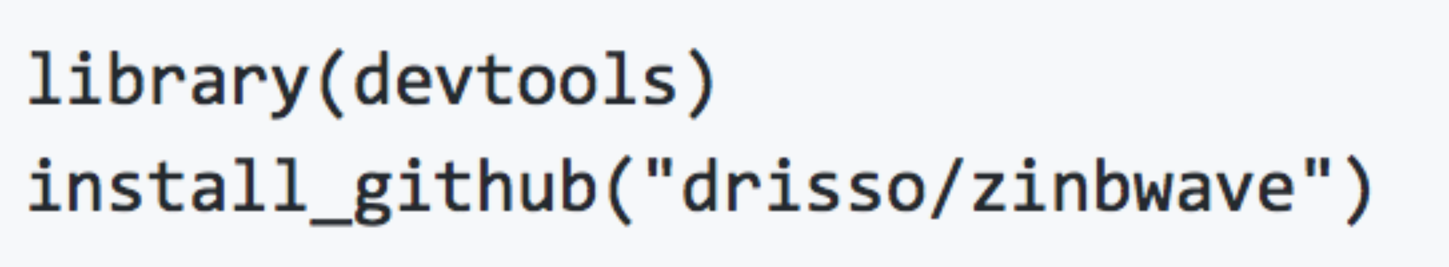
- X:
  - (1,…,1) for gene-specific offset
  - Batch effects, quality control
  - Experimental design
- V
  - (1,…,1) for cell-specific offset (size factor)
  - GC content, ...
- W,alpha: cell cycle, clusters, ... (like PCA)

Fitting the model

$$\max_{\beta,\gamma,W,\alpha,\zeta} \left\{ \ell(\beta,\gamma,W,\alpha,\zeta) - \mathrm{Pen}(\beta,\gamma,W,\alpha,\zeta) \right\}$$

$$\ell(\beta,\gamma,W,\alpha,\zeta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \ln f_{ZINB}(Y_{ij};\mu_{ij},\theta_{ij},\pi_{ij})$$

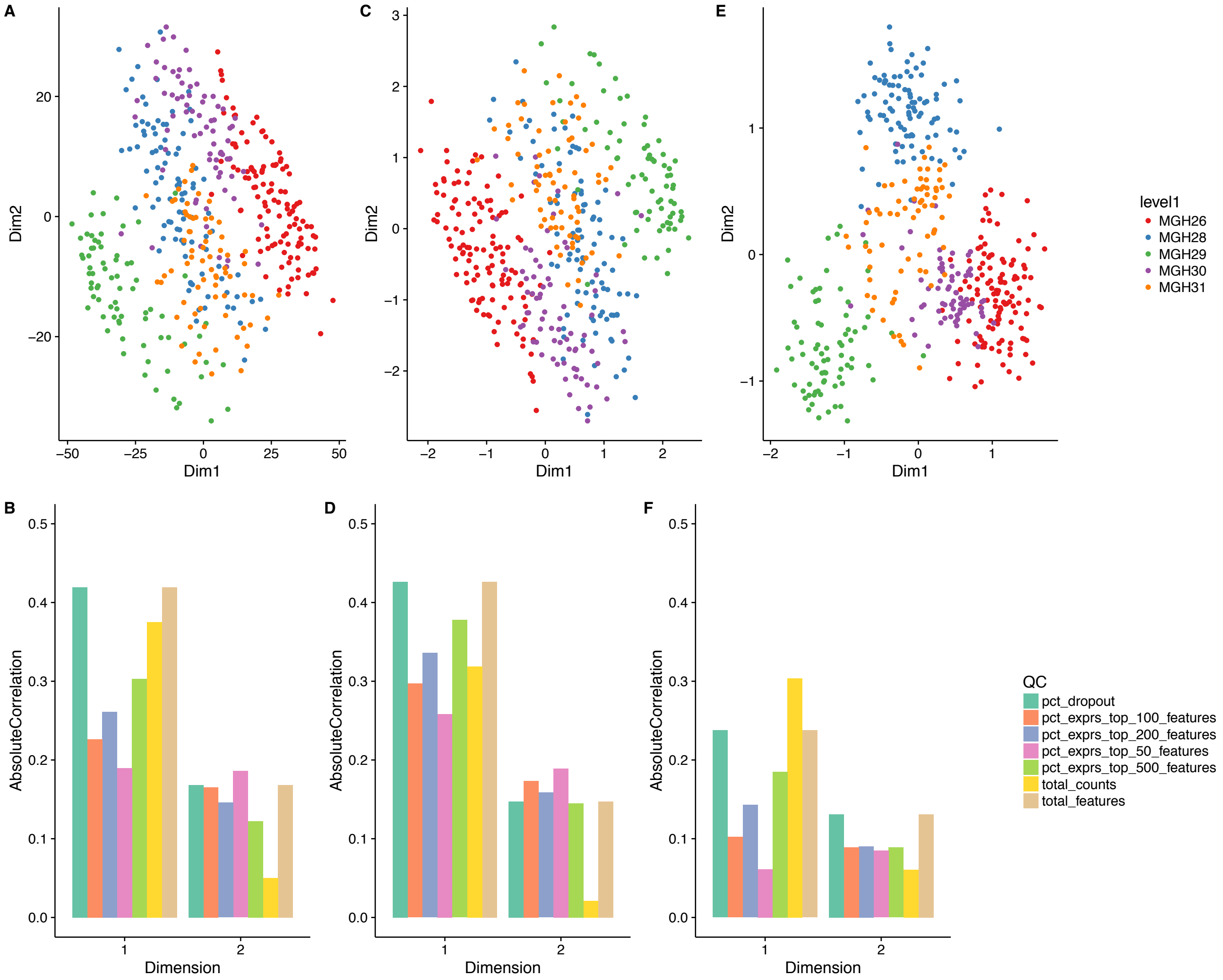
$$\mathrm{Pen}(\beta,\gamma,W,\alpha,\zeta) = \frac{\epsilon_\beta}{2}\|\beta^0\|^2 + \frac{\epsilon_\gamma}{2}\|\gamma^0\|^2 + \frac{\epsilon_W}{2}\|W\|^2 + \frac{\epsilon_\alpha}{2}\|\alpha\|^2 + \frac{\epsilon_\zeta}{2}\mathrm{Var}(\zeta)$$



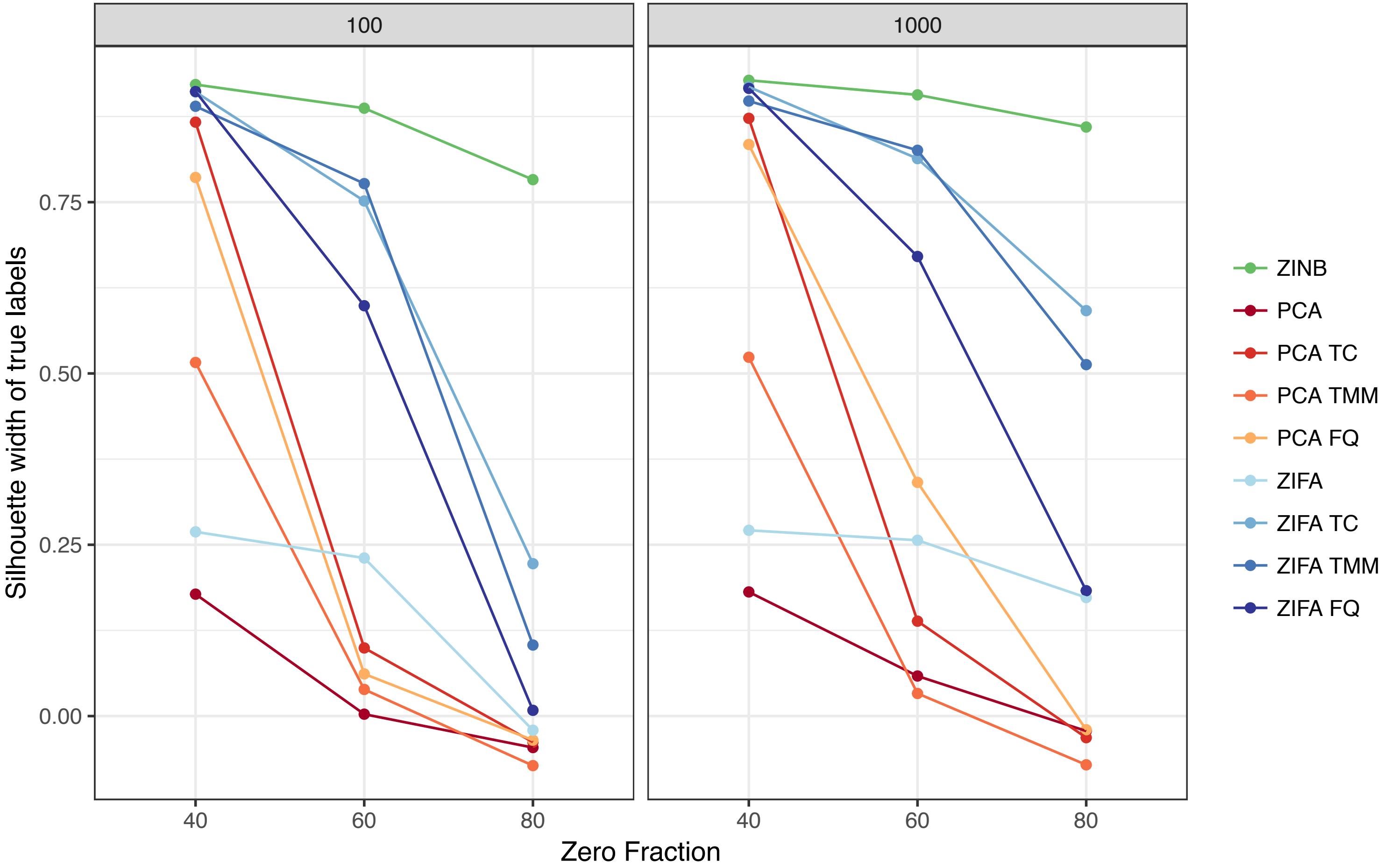
```
library(devtools)
install_github("drisso/zinbwave")
```

Glioblastoma data:
keeps less unwanted signal

PCA ZIFA ZINB-WaVE



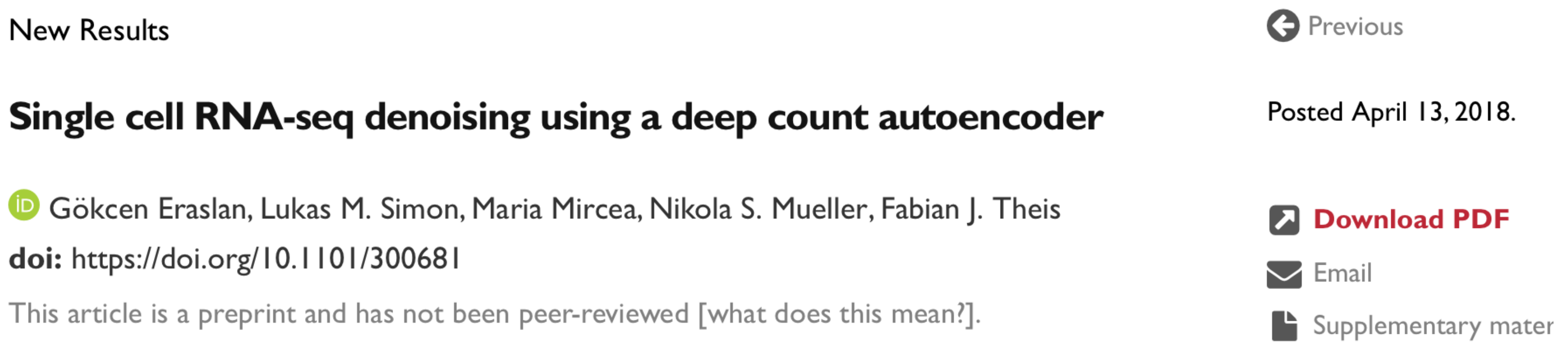
Simulation: robust cluster recovery



Simulation with the Lun & Marioni (2016) model

More recent work…

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY



New Results

Previous

Single cell RNA-seq denoising using a deep count autoencoder

Posted April 13, 2018.

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, Fabian J. Theis

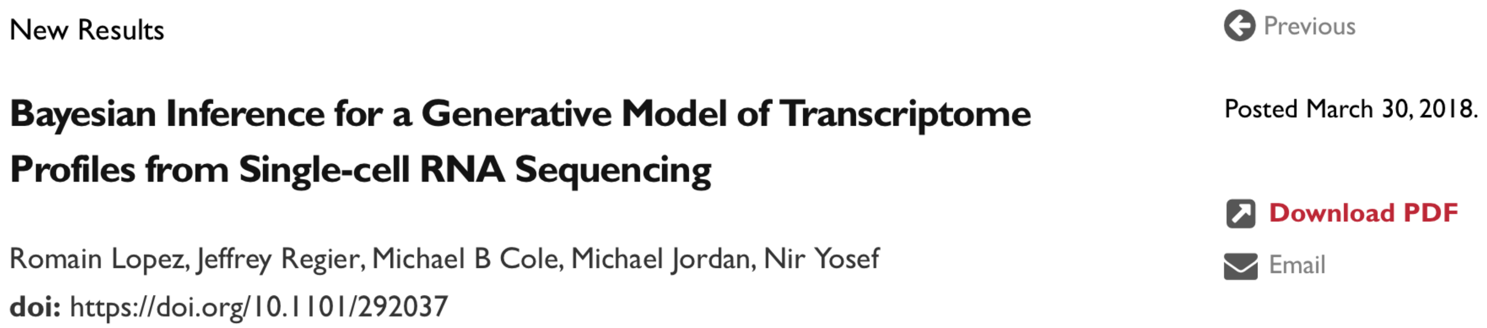
doi: https://doi.org/10.1101/300681

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

Download PDF

Email

Supplementary mater



New Results

Previous

Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing

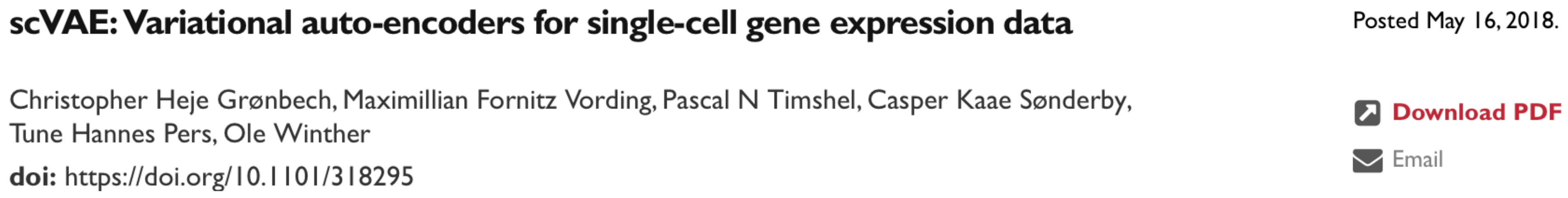
Posted March 30, 2018.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael Jordan, Nir Yosef

doi: https://doi.org/10.1101/292037

Download PDF

Email



scVAE: Variational auto-encoders for single-cell gene expression data

Posted May 16, 2018.

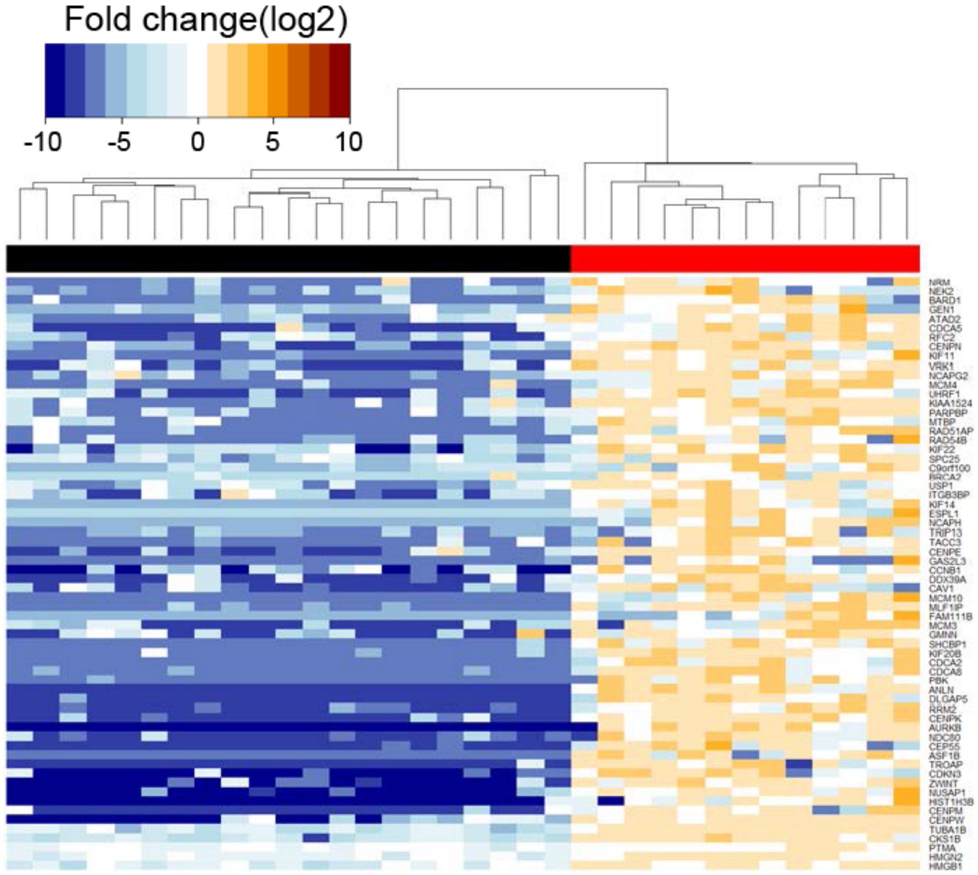
Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune Hannes Pers, Ole Winther

doi: https://doi.org/10.1101/318295

Download PDF

Email

Differential Expression (DE)

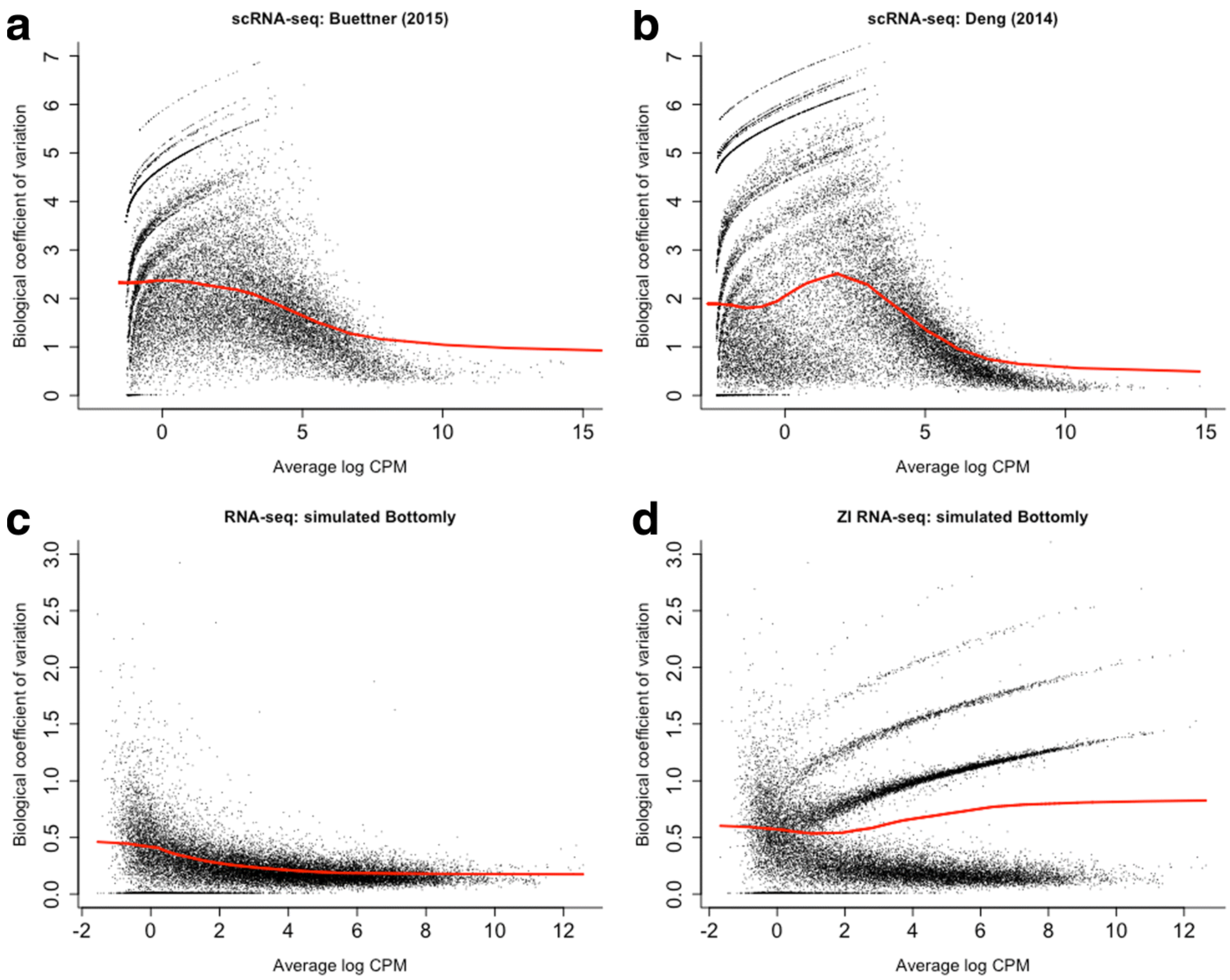


Dedicated tools for « bulk » RNA-seq

- DESeq2
- EdgeR
- …

Need to estimate mean & variance per gene

Zero inflation perturbs mean-variance relationship



Which 0's are dropout?

$$f_{ZINB}(y_{ij};\mu_{ij},\theta_j,\pi_{ij}) = \pi_{ij}\delta + (1-\pi_{ij})f_{NB}(y_{ij};\mu_{ij},\theta_j)$$

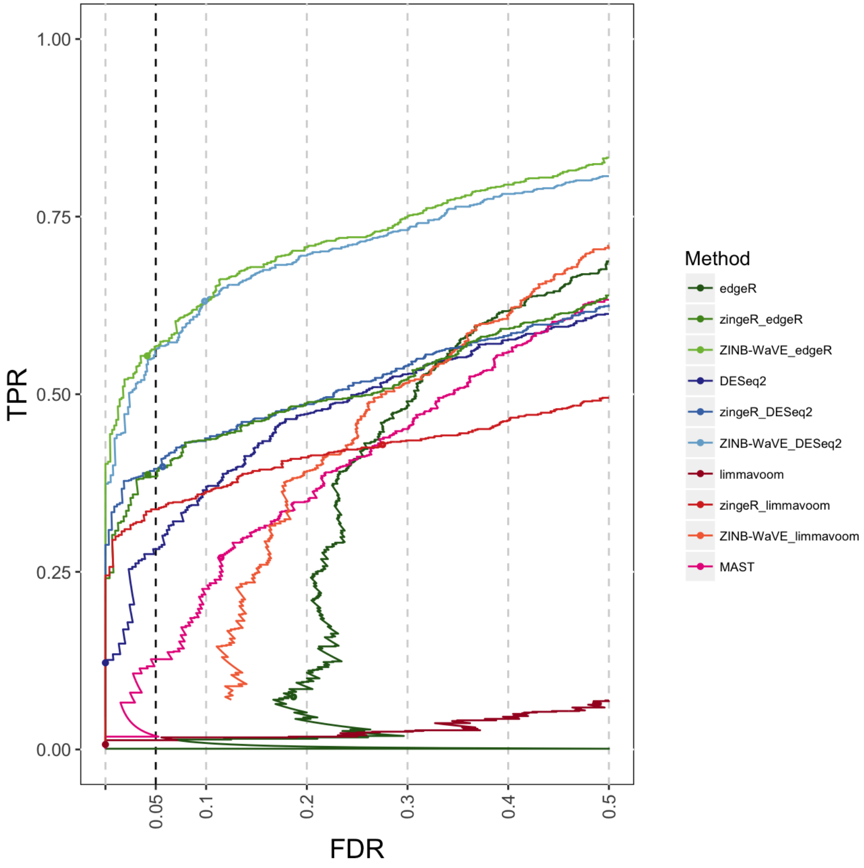
$$W_{ij} = \frac{(1-\pi_{ij})f_{NB}(y_{ij};\mu_{ij},\theta_j)}{f_{ZINB}(y_{ij};\mu_{ij},\theta_j,\pi_{ij})}$$

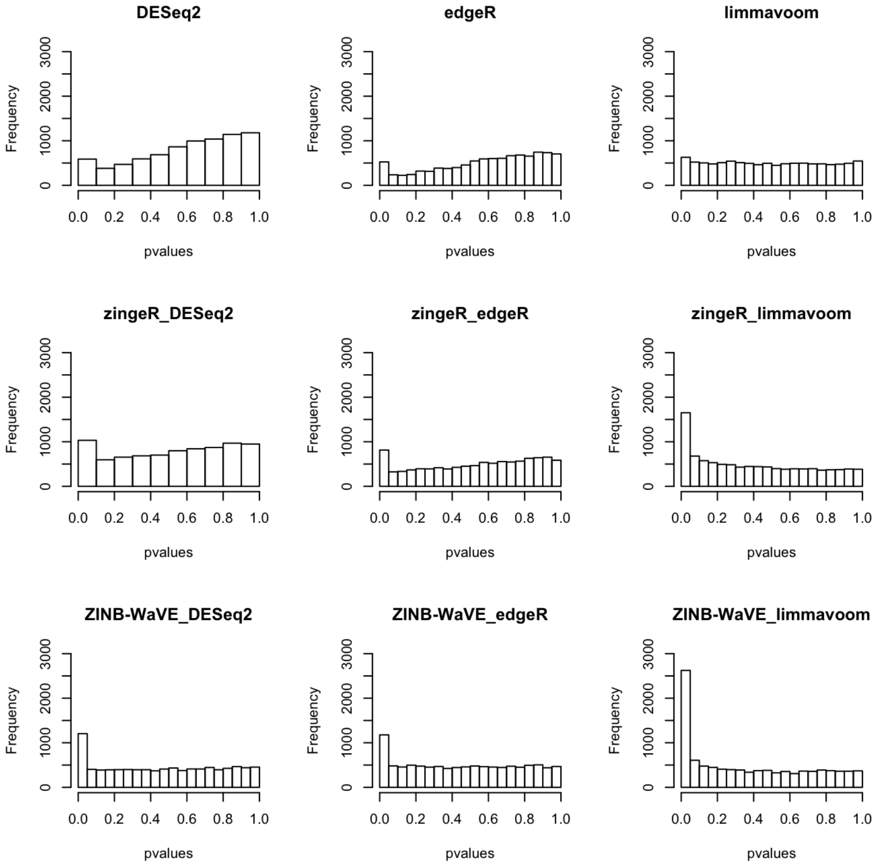
- *Posterior probability that Y_ij is not a dropout*
- *Can be used as an observation weight in methods for « bulk » RNA-seq*

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†*]





Supervised classification

- Given a set of labeled scRNA-seq profiles (e.g., cell types), how to learn a **sparse** classifier?
- Popular solution for bulk data: lasso / elastic net regression

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(w, x_i, y_i) + \lambda \Omega(w) \right\}$$

$$\Omega_{\text{enet}}(w) = \alpha \left\| w \right\|_2^2 + (1 - \alpha) \| w \|_1$$

From ridge to dropout regularization

- Ridge regularization is related to additive Gaussian noise in the data
- We should instead be robust to **dropout noise** in the data, suggesting to use instead **dropout regularization** *(altitude training)*

$$\min_{w \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} L(w, \delta_i \odot \frac{x_{i,}}{p}, y_i) \right)$$

Droplasso = Dropout + Lasso

B. Khalfaoui

$$\min_{w\in\mathbb{R}^d}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\delta_i\sim B(p)^d}L\left(w,\delta_i\odot\frac{x_{i,}}{p},y_i\right)+\lambda\left\|w\right\|_1\right)$$

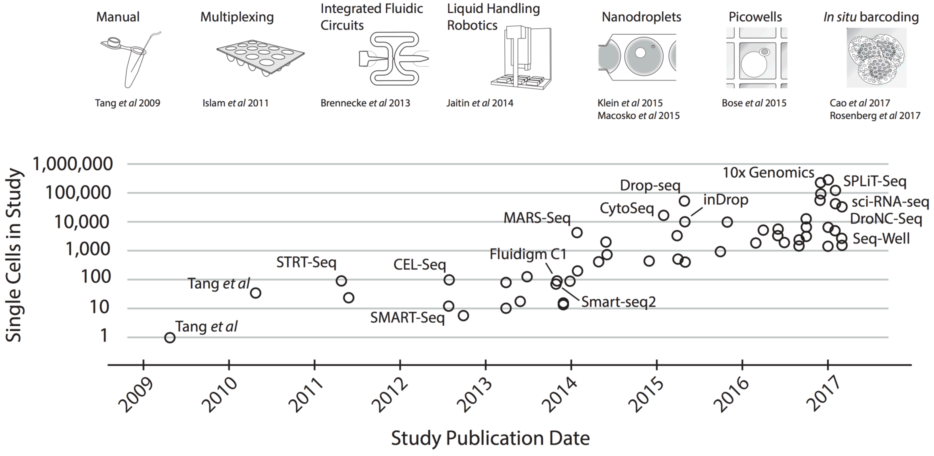
Dataset	Problem	Lasso	Elastic net	Dropout	Droplasso
EMTAB2805	G1 vs G2M	0.72	0.93	0.80	**0.95**
GSE74596	NKT0 vs NKT17	0.84	0.92	0.94	**0.97**
GSE63818	Primordial germ cells vs somatic	0.93	0.97	0.98	**0.99**
GSE48968	1h vs 4h LPS stimulation	0.95	0.96	0.96	**0.99**
GSE81861	Tumour vs normal	0.80	0.85	0.84	**0.90**

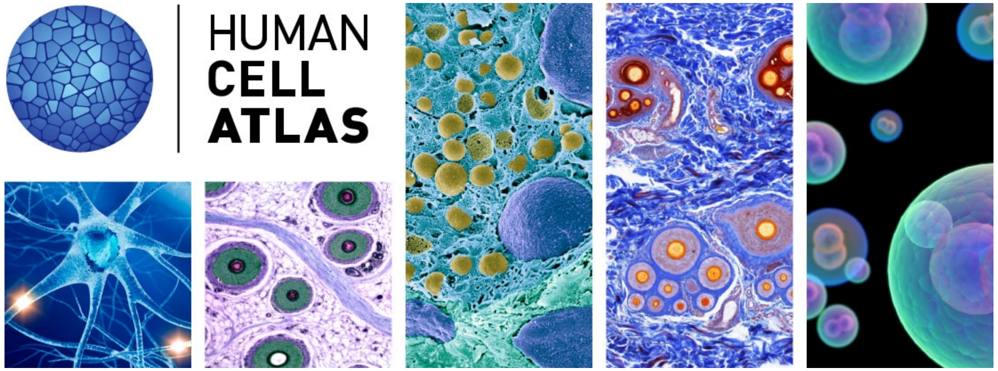
Preliminary results

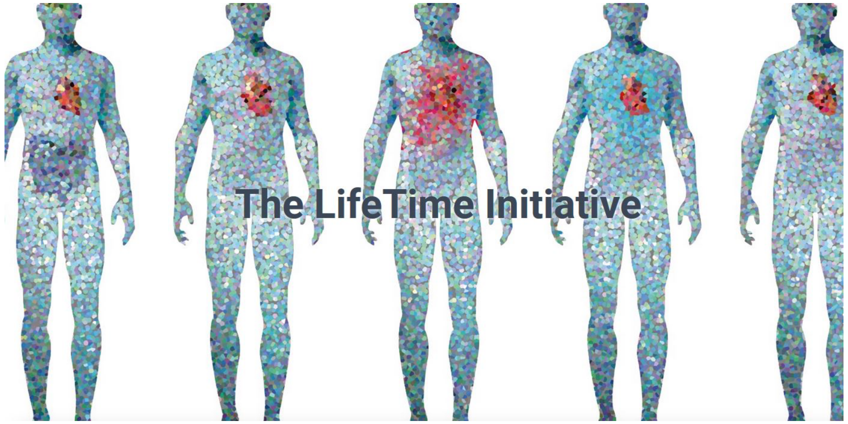
Much more ahead!

Single-Cell Multiomics: Multiple Measurements from Single Cells

Iain C. Macaulay,[1,*] Chris P. Ponting,[2,3,*] and Thierry Voet[2,4,*]







Thanks!

Fanny Perraudeau

Koen Van den Berge

Davide Risso

Beyrem Khalfaoui

Svetlana Gribkova

Charlotte Soneson

Michael Love

Mark Robinson

Lieven Clement

Sandrine Dudoit