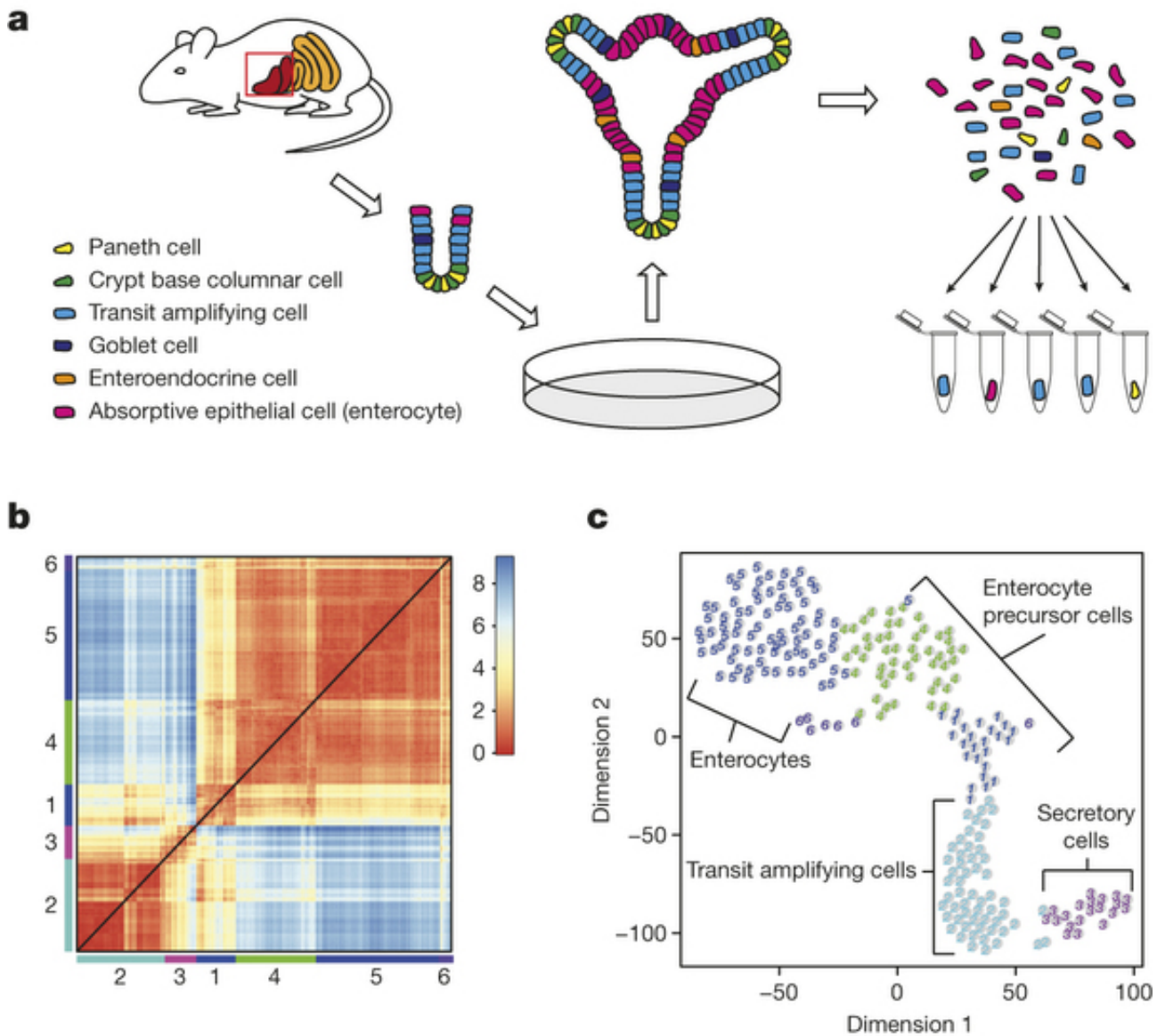
ZINB-WaVE: a general and flexible method for signal extraction from single-cell RNA-seq data

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

[1]Division of Biostatistics and Epidemiology, Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA.
[2]Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot, Paris, France
[3]Division of Biostatistics, School of Public Health, University of California, Berkeley, USA
[4]Department of Statistics, University of California, Berkeley, USA
[5]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Paris, France
[6]Institut Curie, Paris, France
[7]INSERM U900, Paris, France
[8]Ecole Normale Supérieure, Department of Mathematics and Applications, Paris, France

Single-cell RNA-seq

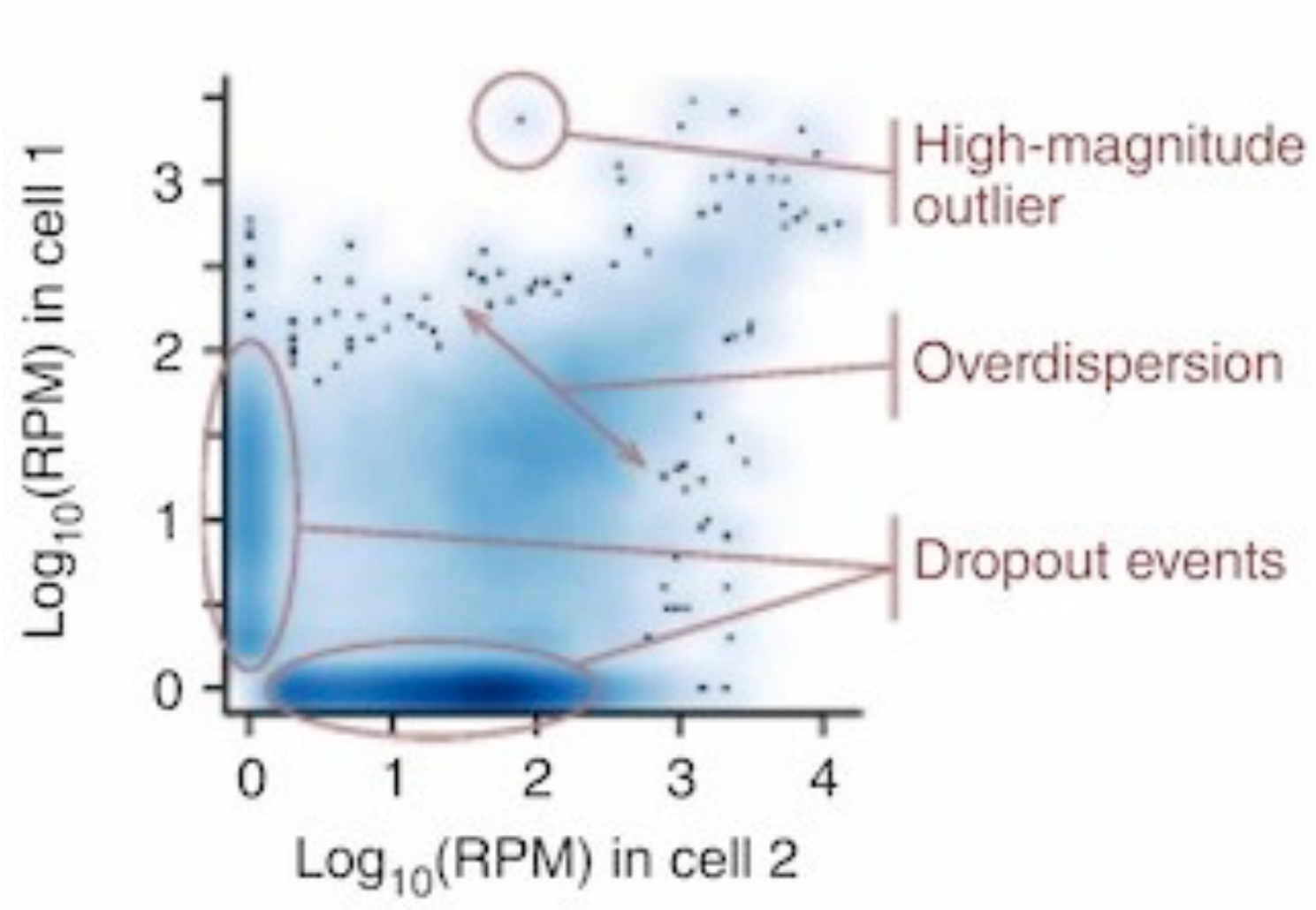


(Grün et al 2015)

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

Challenges

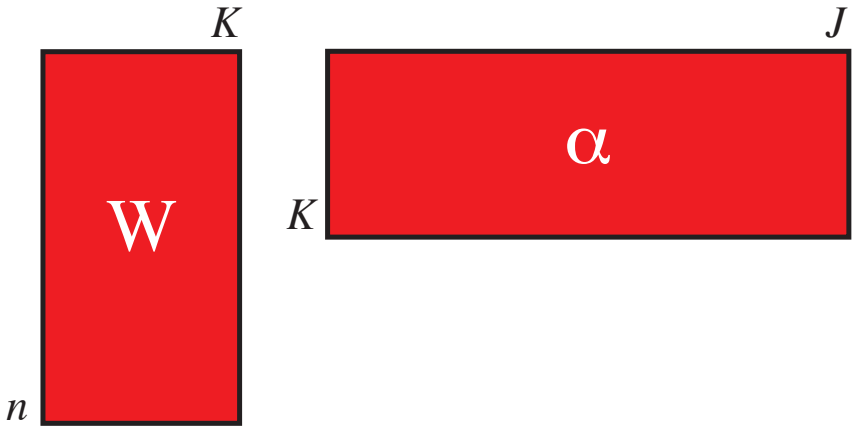
- Normalize for sequencing depth?
- Remove unwanted variations? (batches, cell cycle, GC content, …)
- Distances between transcription profiles?
- Clustering / Visualization?
- Differential expression?

Standard approach

- Massage the matrix
  - Y_ij = log(count_ij + 1) * size factor
  - Sometimes full quantile normalization
- Dimension reduction
  - PCA on Y
  - Keep around 50 dimensions
- Nonlinear embedding (t-SNE), clustering, …

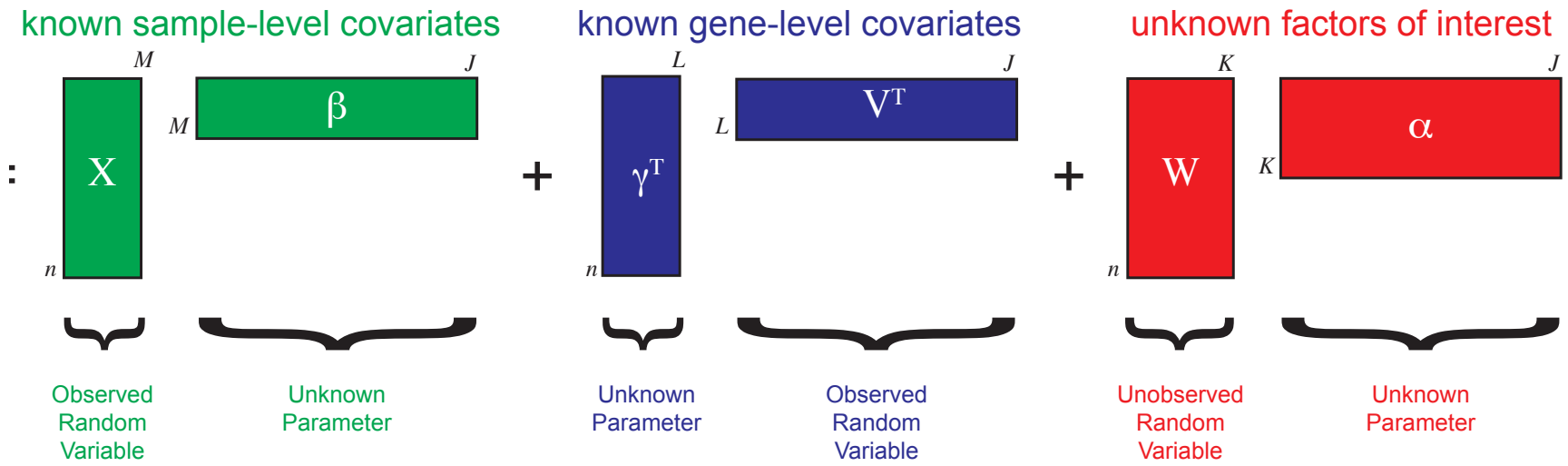
Dimension reduction (PCA/SVD)

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



Including known covariates (RUV)

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



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

Cold Spring Harbor Laboratory
bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

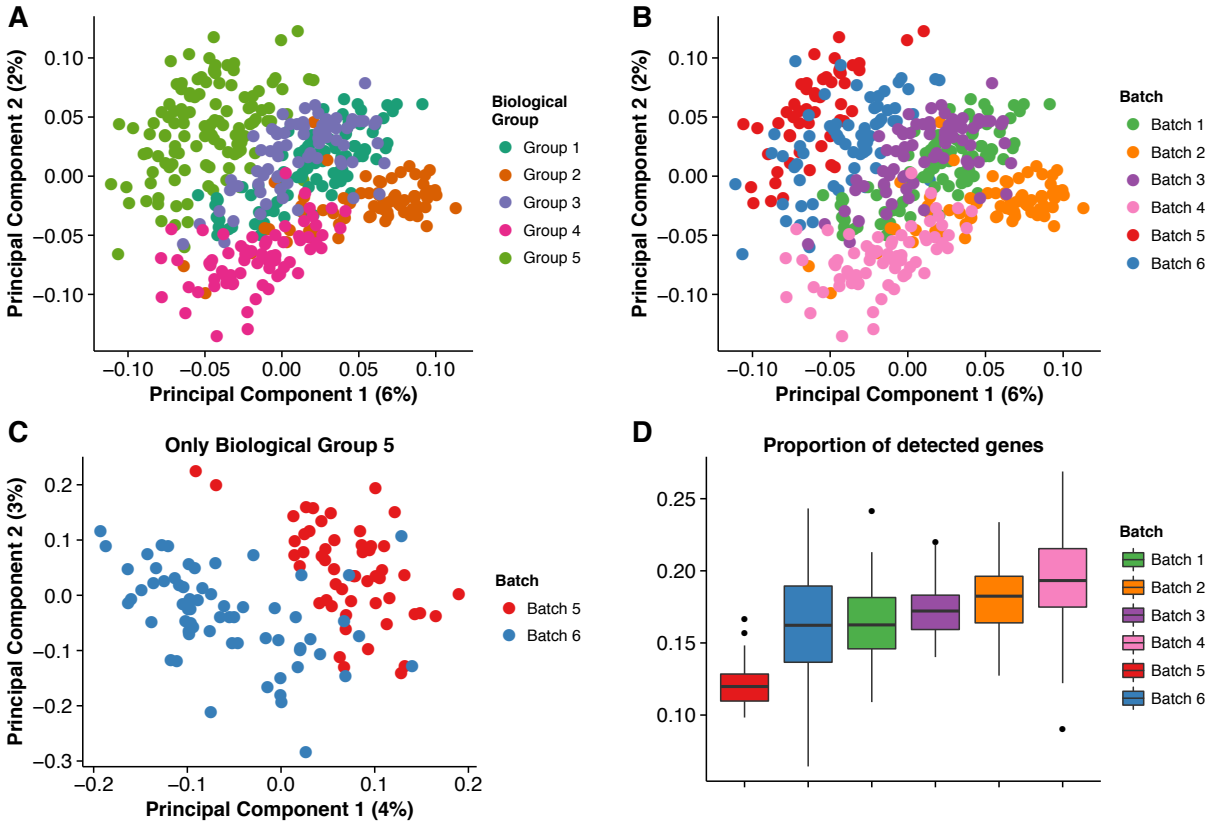
Some worrying results

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



Pierson and Yau *Genome Biology* (2015) 16:241
DOI 10.1186/s13059-015-0805-z

Genome Biology

SOFTWARE Open Access

ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis

Emma Pierson[1] and Christopher Yau[1,2*]

Interesting model

$$Z = W\alpha + \epsilon$$

$$Y_{ij} = \begin{cases} 0 & \text{with probability } \exp(-\lambda Z_{ij}^2) \\ Z_{ij} & \text{otherwise} \end{cases}$$

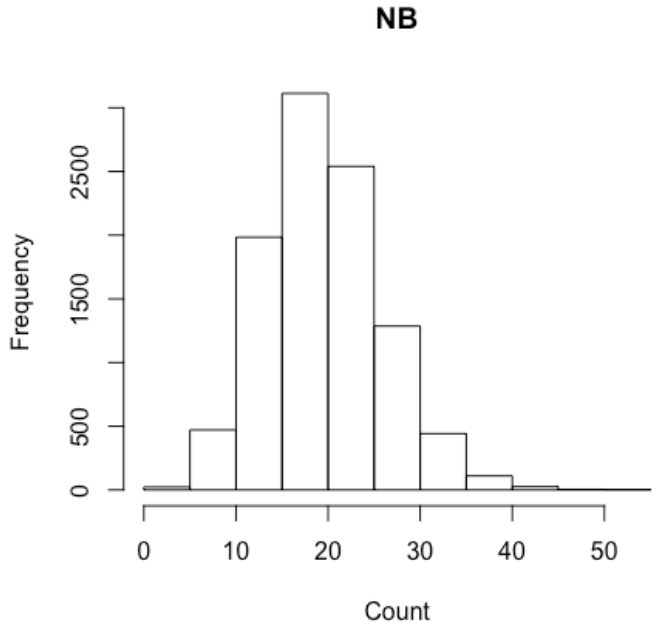
But:

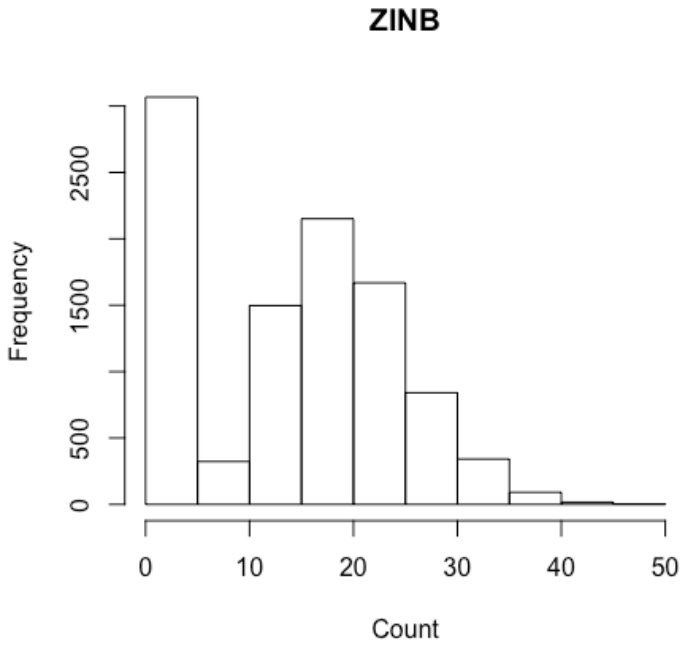
- *Models continuous data (log(count+1))*
- *Dropout probability as a fixed function of expression level*

ZINB distribution to model a count

« Zero-Inflated Negative Binomial »

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





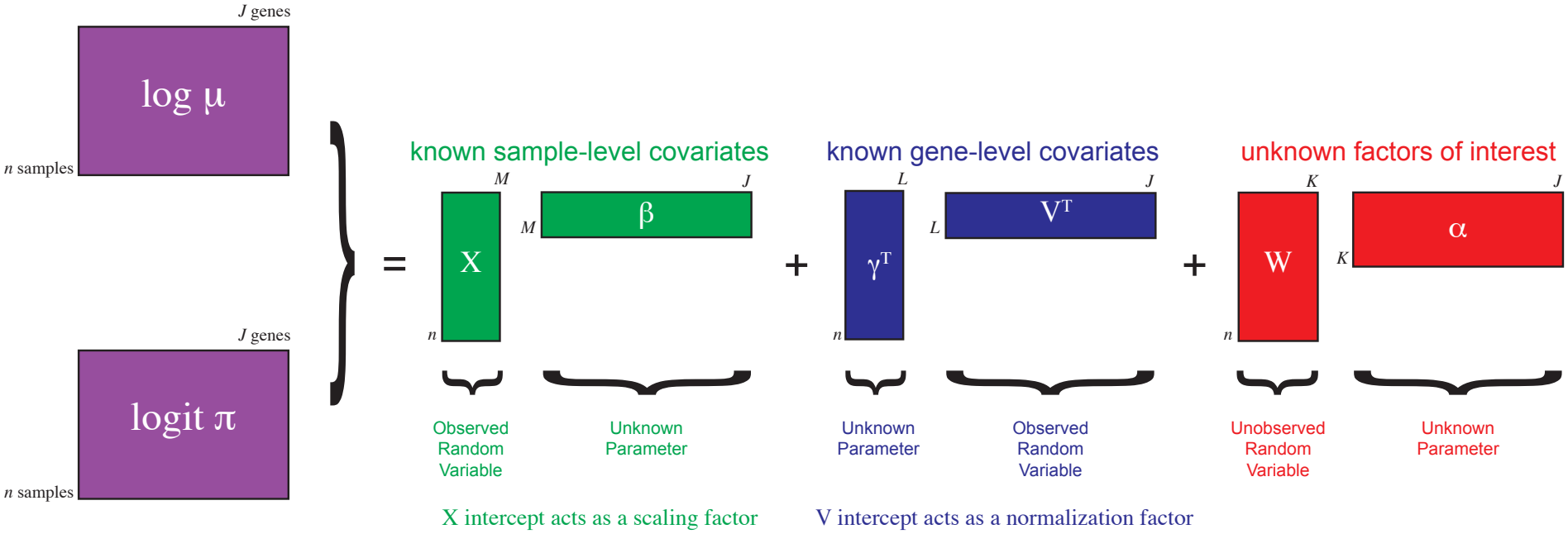
$$f_{ZINB}(y;\mu,\theta,\pi) = \pi\delta_0(y) + (1-\pi)f_{NB}(y;\mu,\theta), \quad \forall y \in \mathbb{N},$$

ZINB-WaVE model

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

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

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



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

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

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

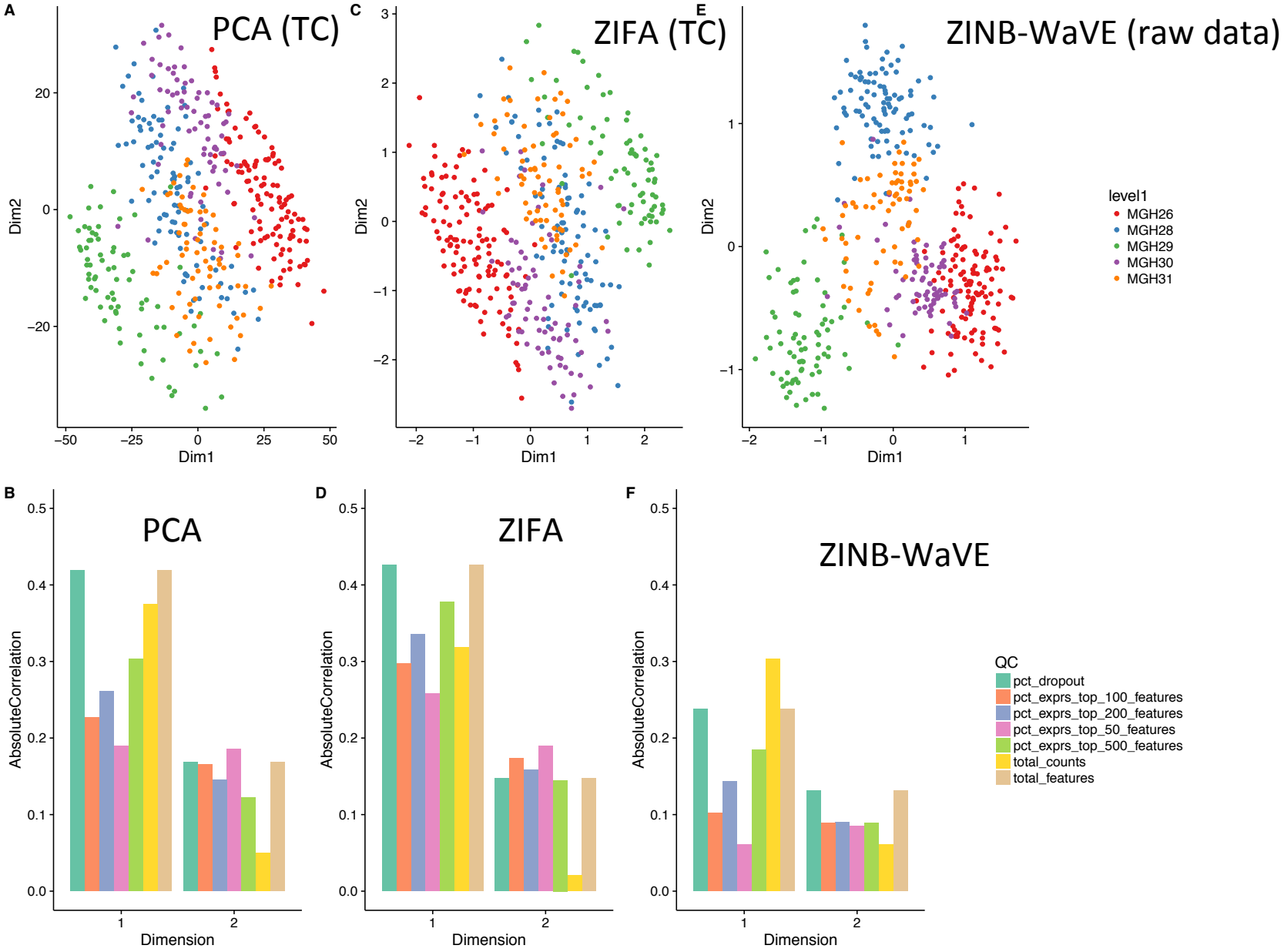
$$\text{Pen}(\beta,\gamma,W,\alpha,\zeta) = \frac{\epsilon_\beta}{2}\|\beta^0\|^2 + \frac{\epsilon_\gamma}{2}\|\gamma^0\|^2 + \frac{\epsilon_W}{2}\|W\|^2 + \frac{\epsilon_\alpha}{2}\|\alpha\|^2 + \frac{\epsilon_\zeta}{2}\text{Var}(\zeta)$$

Fitting the model

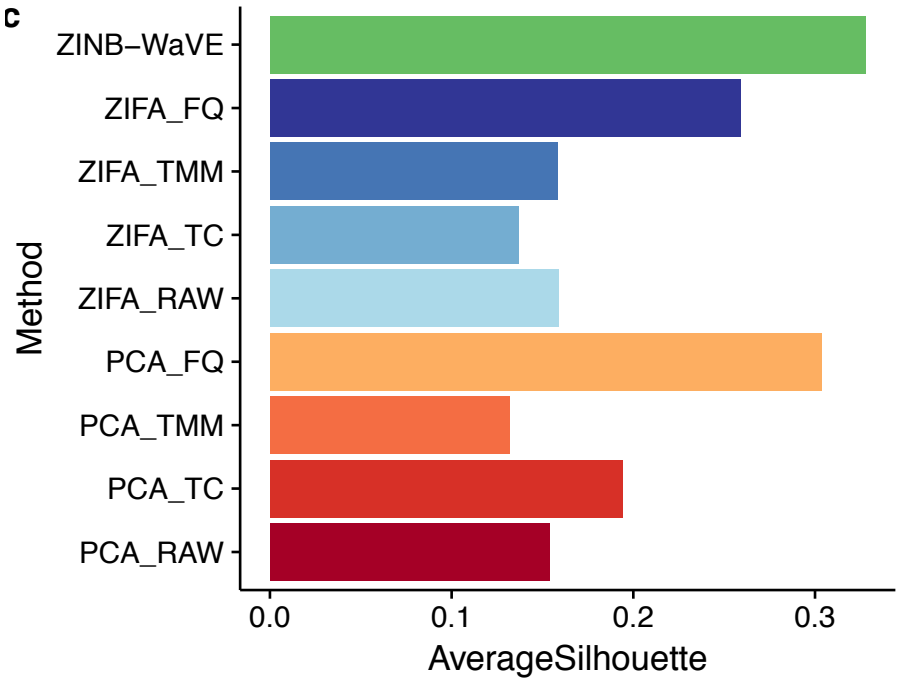
$$\max_{\beta,\gamma,W,\alpha,\zeta} \{\ell(\beta,\gamma,W,\alpha,\zeta) - \mathrm{Pen}(\beta,\gamma,W,\alpha,\zeta)\}$$

- Initialization
  - Uncouple mu and pi
- Iterate until convergence optimization of:
  - Dispersion (zeta)
  - Left factors (gamma, W)
  - Right factors (beta, alpha)
  - Orthogonalization (W, alpha)

Glioblastoma data

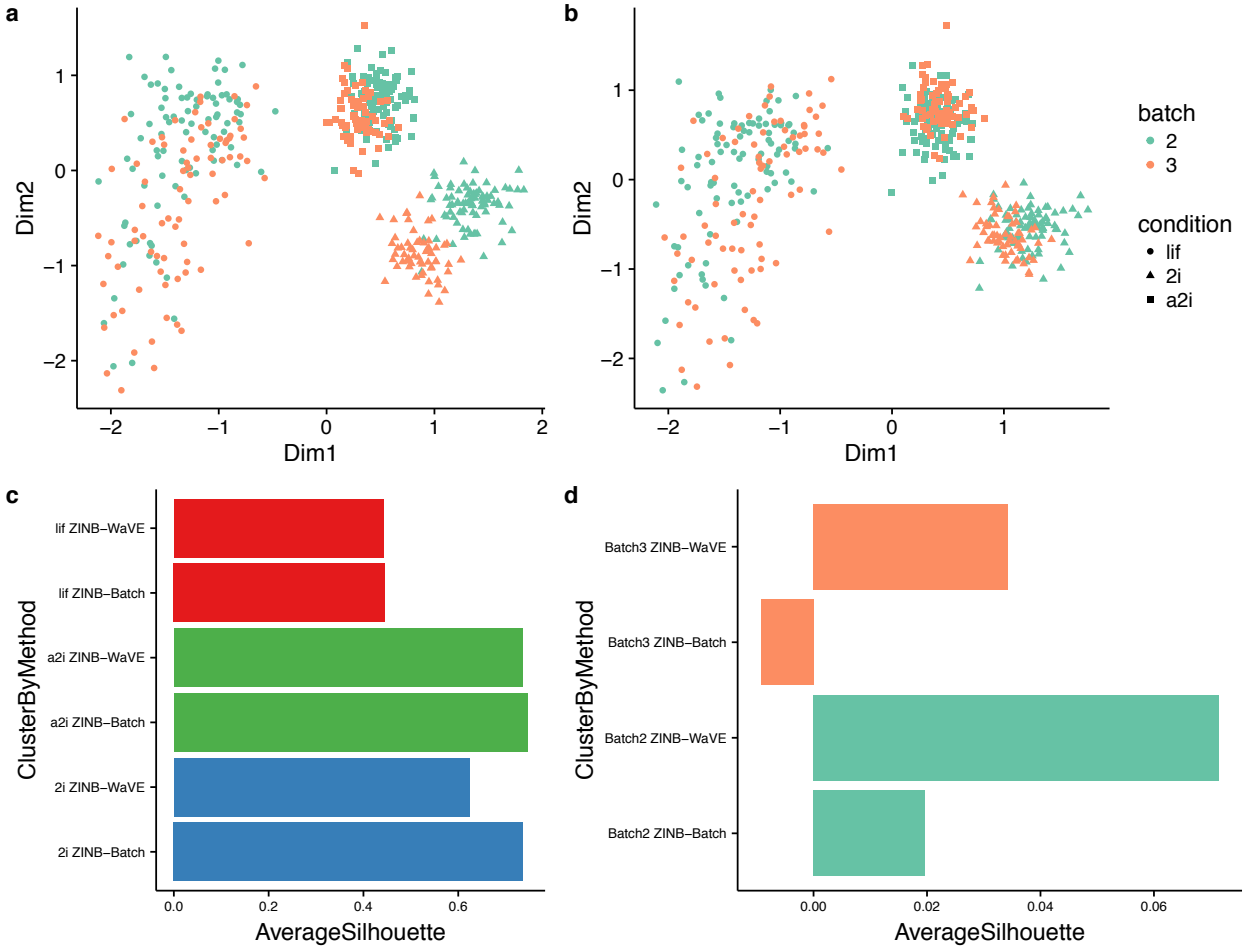


Glioblastoma data



- Less correlated with technical effects
- Better clusters cells by patient

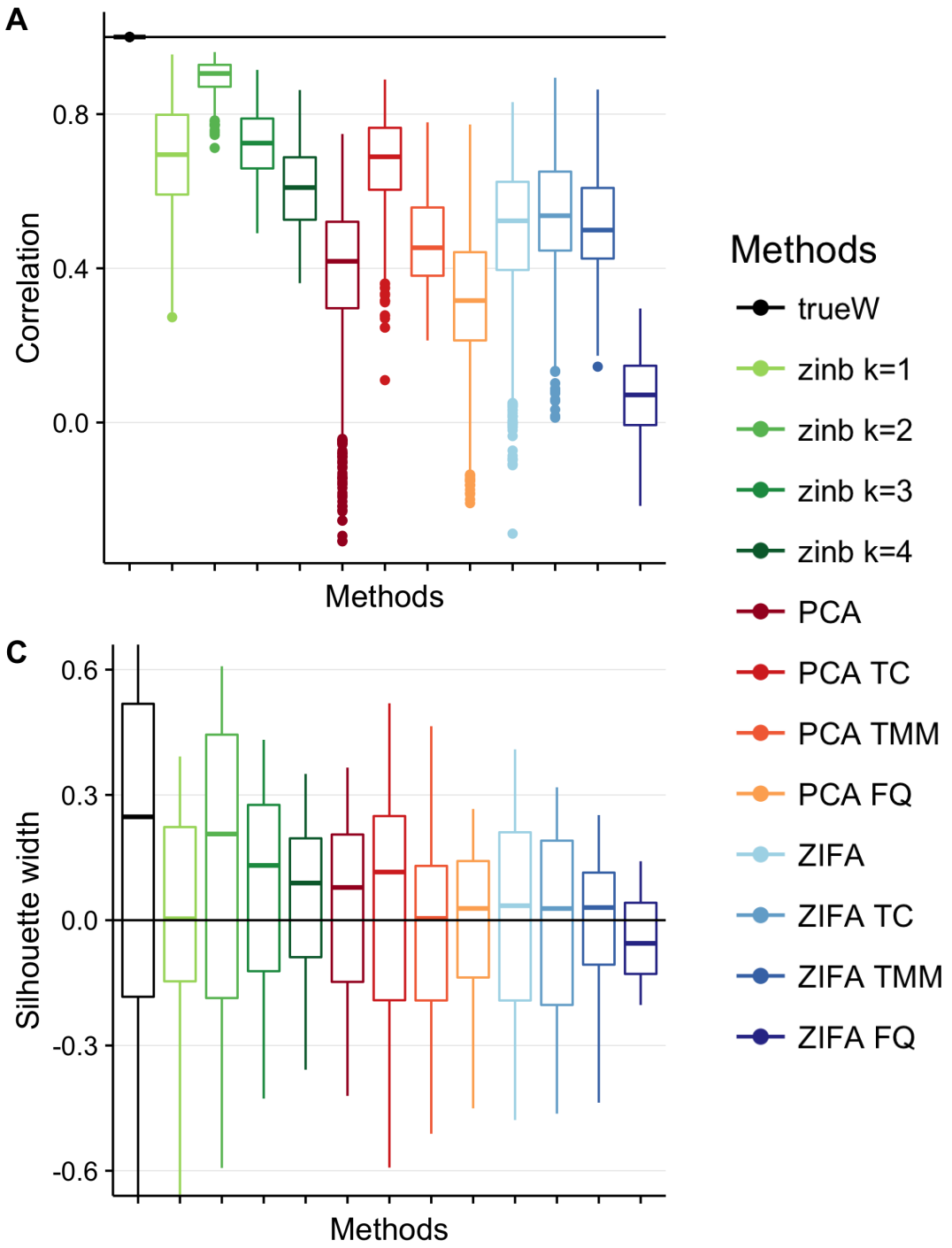
mESC data: decreasing batch effect



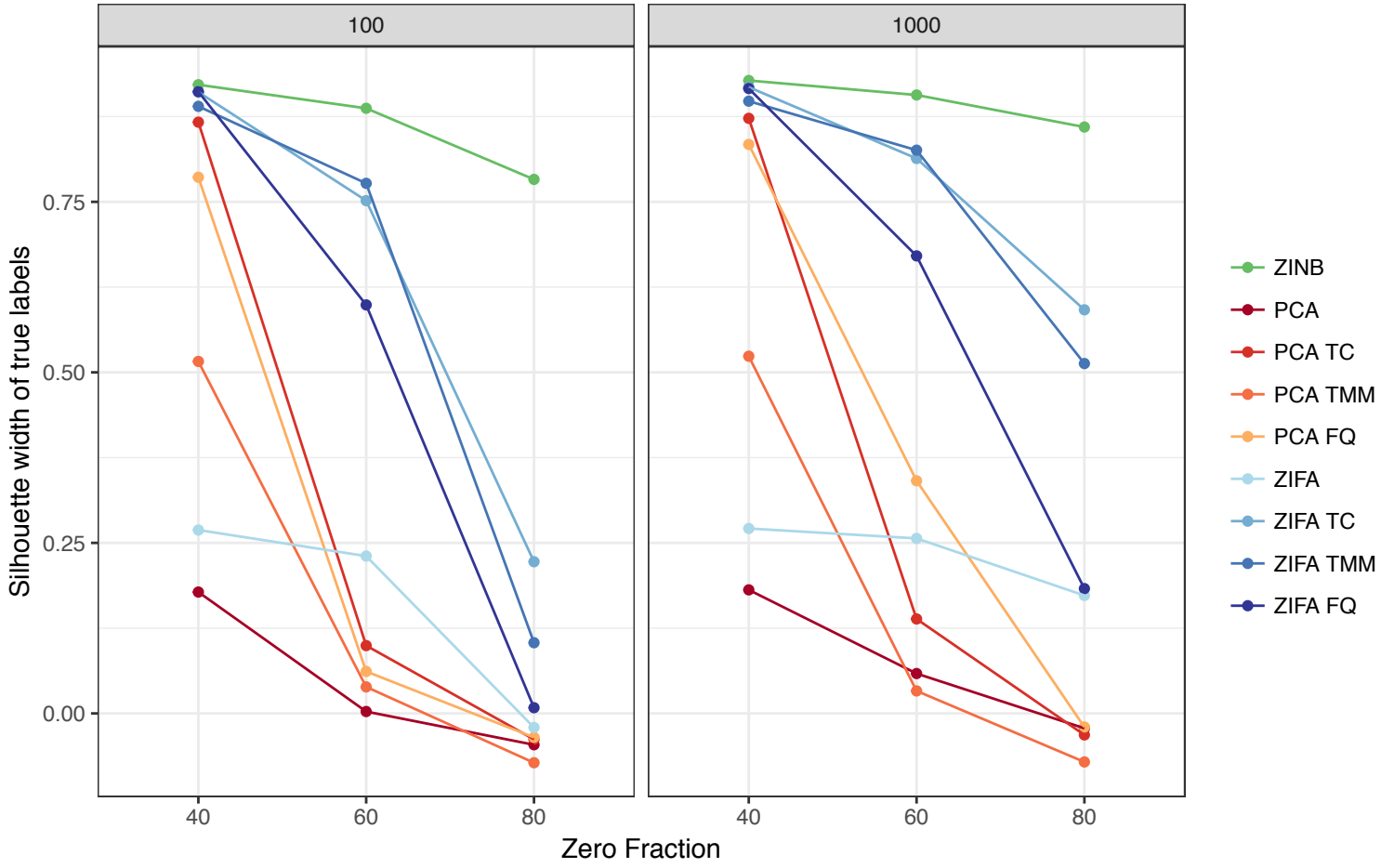
3 cultures media, 2 batches. Add batch as covariate

Simulations: W estimation

- Simulate clusters of single cells (from real data) with cell- and gene-level offsets
- Following the ZINB model with K=2 latent factors
- Check how well W is recovered, and the clustering is recovered

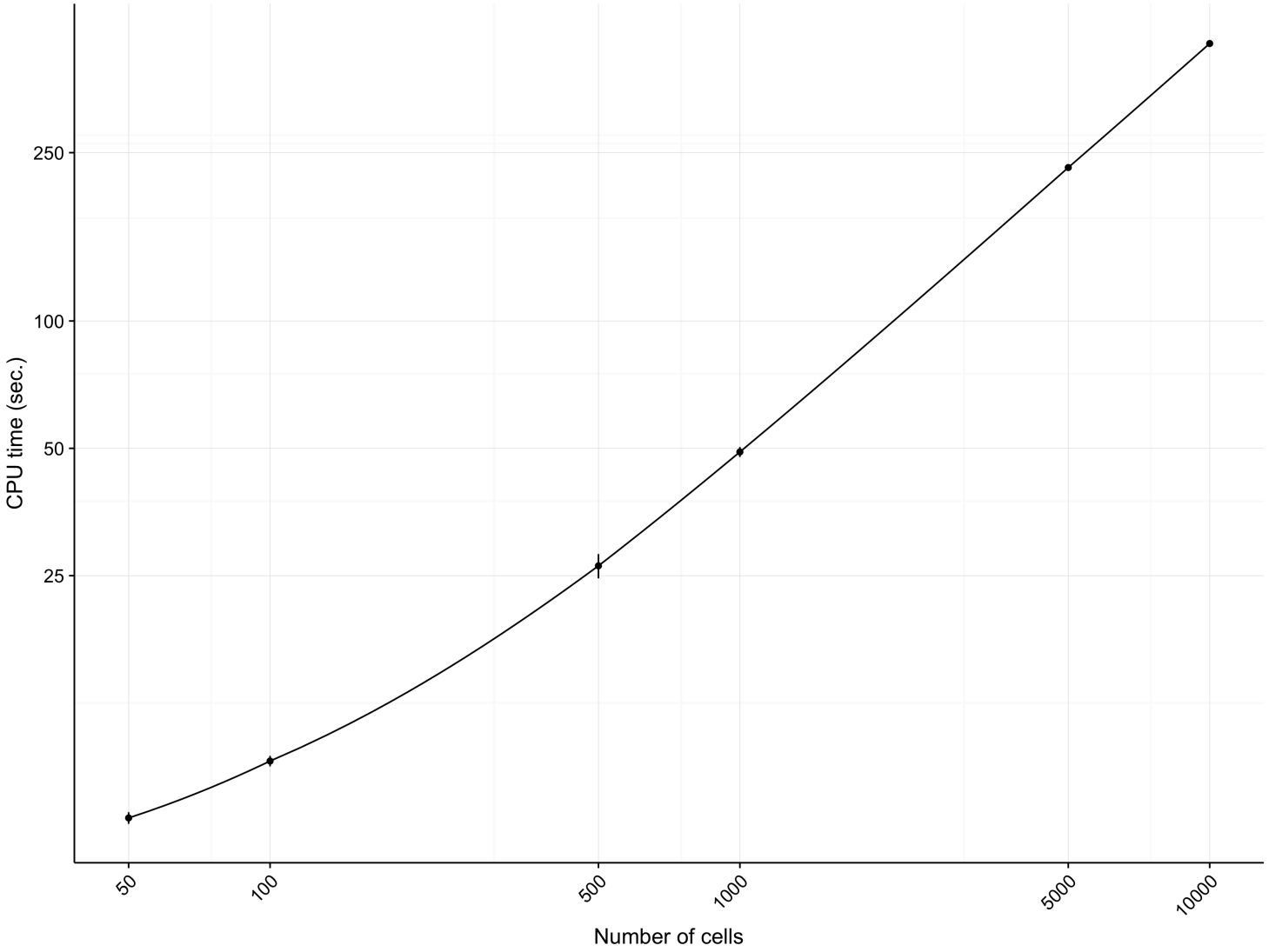


Simulation: cluster recovery



Simulation with the Lun & Marioni (2016) model

CPU time



On a recent iMac, 16GB of RAM, using 7 cores

Try it!

- https://github.com/drisso/zinbwave
- http://biorxiv.org/content/early/2017/04/06/125112

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

Conclusion

- A model:
  - Using ZINB distribution to model zero-inflated counts
  - With linear structure to include gene- or cell-specific covariates
  - And low-dimensional signal inferred automatically
- Fitting the model works on simulations
- On real data, better captures clustering than PCA or ZIFA
- Less correlated with batch / unwanted variations