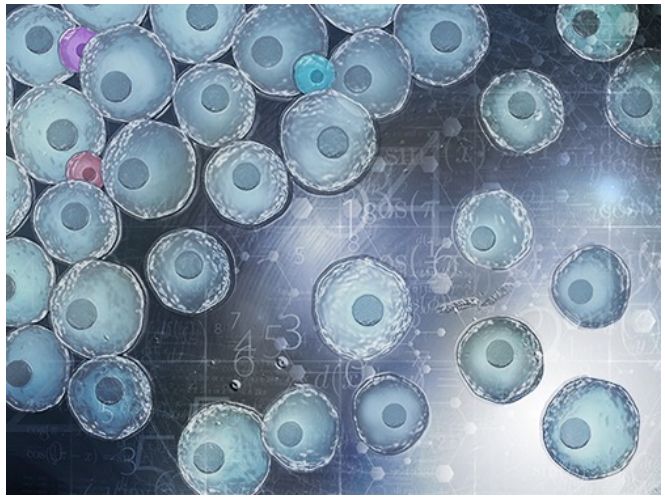
Latent variable models to disentangle sources of heterogeneity in single cell RNA-seq data

Florian Buettner

HelmholtzZentrum münchen
German Research Center for Environmental Health

EMBL-EBI

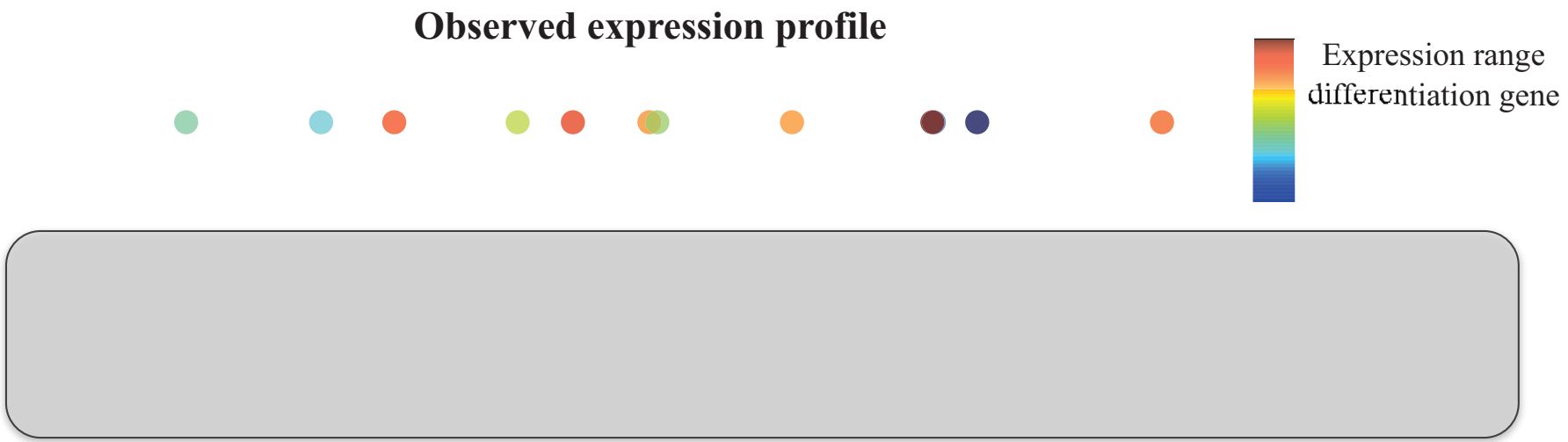
Single-cell RNA-seq



Multiple sources of variation between cells

- Variation of interest: differentiation, intra-tumoural heterogeneity
- Confounding variation: cell cycle/size, apoptosis
- Technical noise: drop-out effects, batch effects

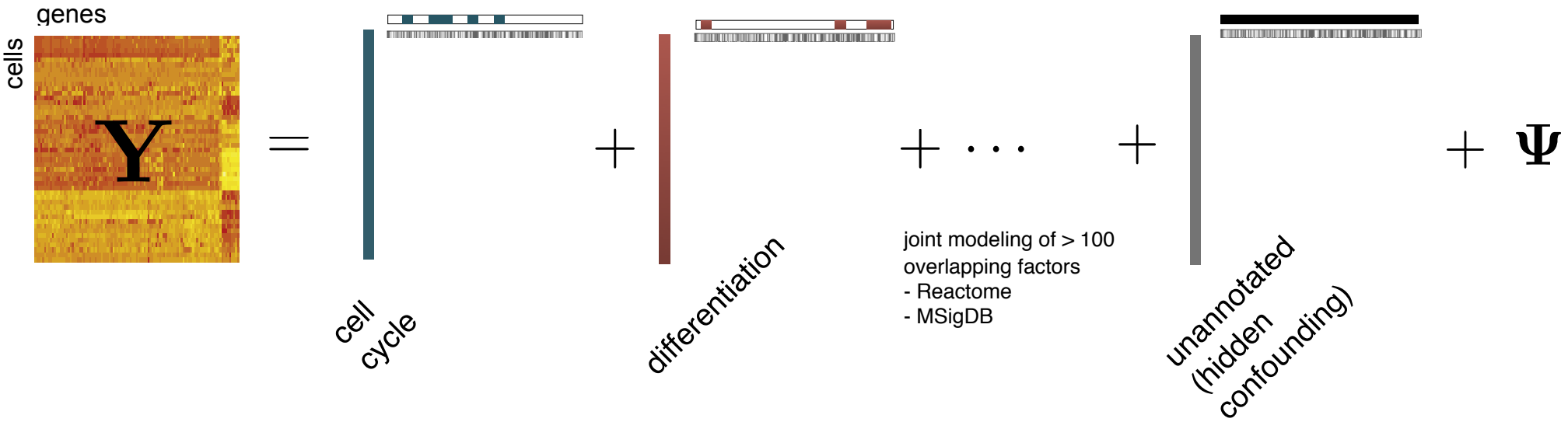
Expression variability is caused by the interplay of multiple latent processes



A versatile factor model to understand single-cell transcriptome variability

- Sparse factor analysis model untangling sources of variation for scRNA-seq in interpretable components
  - Use prior gene sets to infer interpretable factors
  - Explicit modelling of unannotated factors (confounding)
  - Refine gene set annotations
  - Accounting for scRNA-seq noise
  - Computational efficiency & scalability to larger datasets

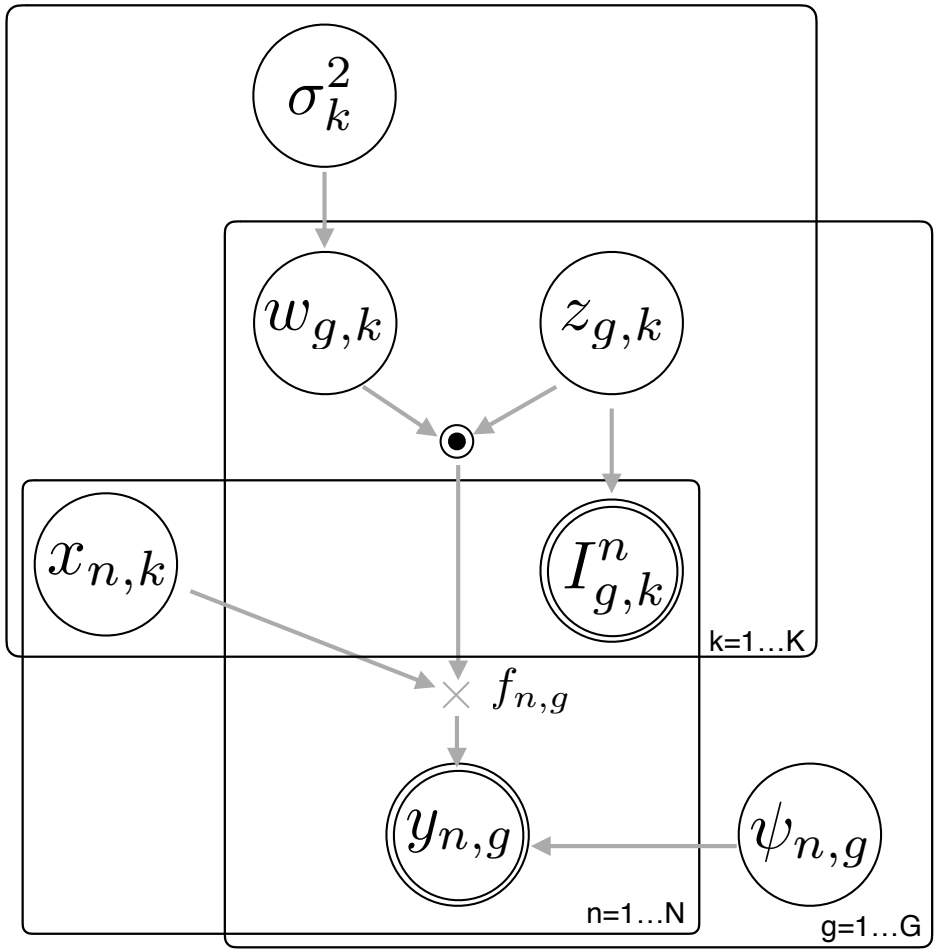
A factorial model for single-cell RNA-seq



$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{x}_k \left(\mathbf{w}_k \cdot \mathbf{z}_k\right)^T + \mathbf{\Psi}$$

Leek & Storey, 2007
Stegle et al., 2010
Gao et al., 2013

A factorial model for single-cell RNA-seq



$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{x}_k (\mathbf{w}_k \cdot \mathbf{z}_k)^T + \mathbf{\Psi}$$

- Modelling factor annotations

$$p(I_{g,k}^n = 1 \,|\, z_{g,k} = 1) =$$
$$\text{Bernoulli}(I_{g,k} = 1 \,|\, \text{TPR})$$

- ARD relevance prior on factors

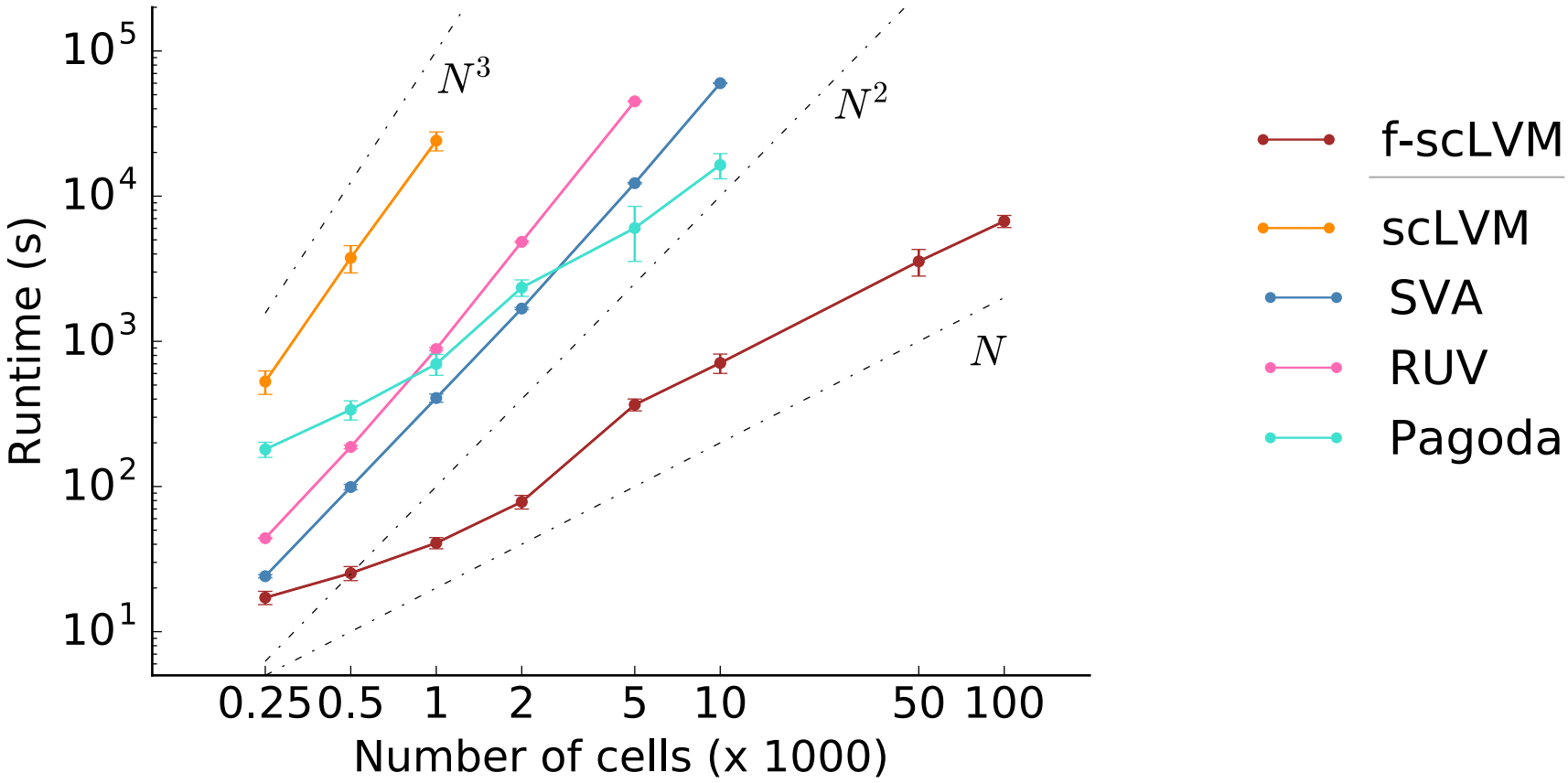
$$p(w_{g,k} \,|\, \sigma_k^2) = \mathcal{N}(w_{g,k} \,|\, 0, \sigma_k^2)$$
$$p(\sigma_k^2) = \text{Gamma}(\sigma_k^2 \,|\, a, b)$$

- A hurdle noise model to account for dropout events (e.g. Finak et al. 2015)

$$P(y_{ng}|f_{ng}) = \begin{cases} \frac{1}{1+\exp(f_{ng})} & \text{if } y_{ng} = 0 \\ \mathcal{N}(f_{ng}\ ,\ \sigma_g^2) & \text{otherwise} \end{cases}$$

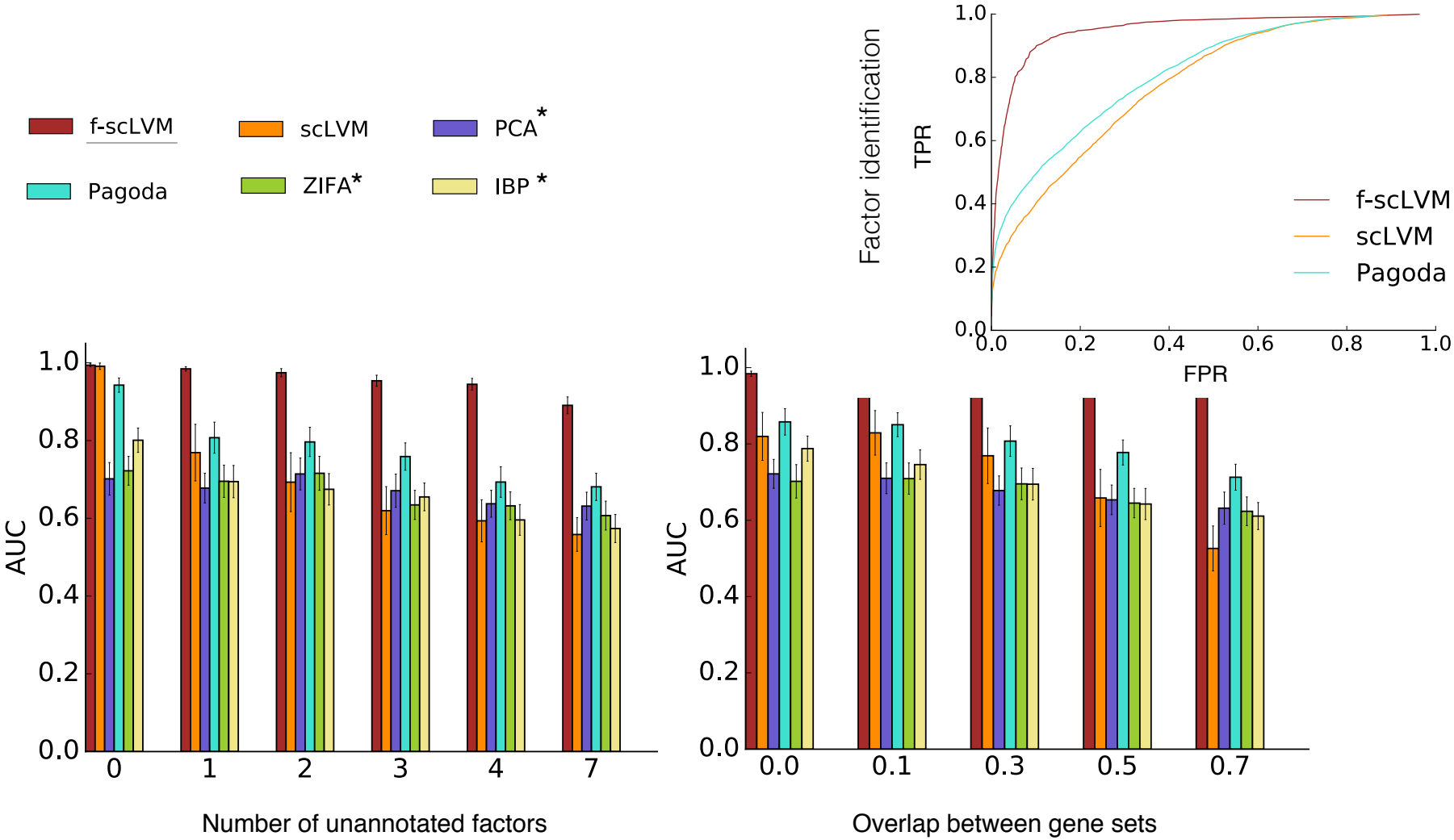
Scalability inference

- Factorised variational Bayes, retaining the coupling of the sparsity indicator and and factor weights $q(w_{g,k}, z_{g,k})$



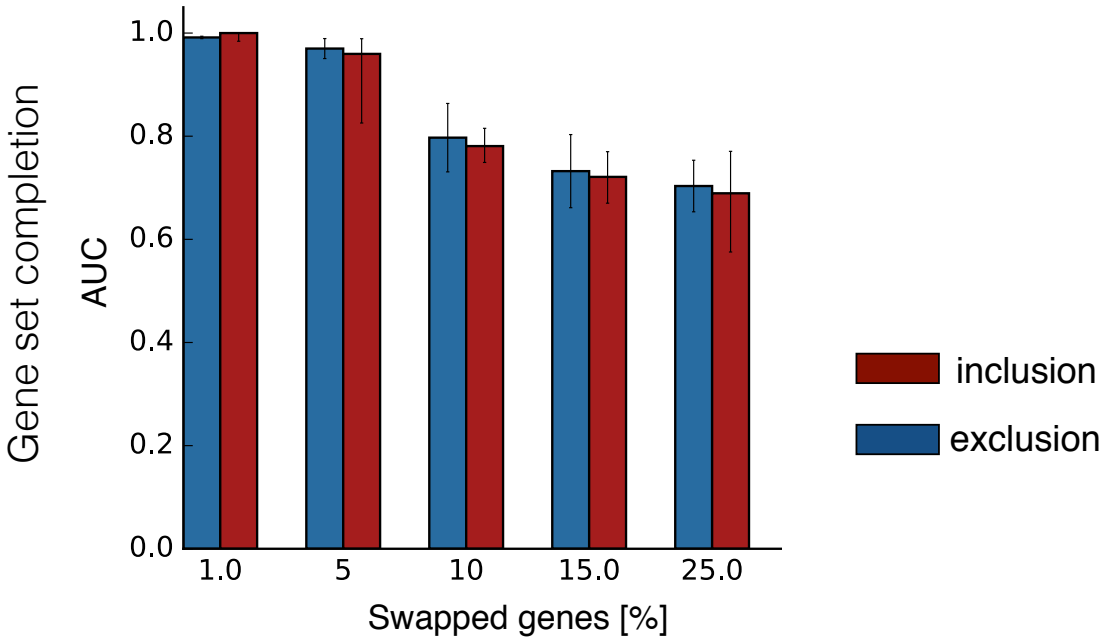
Compute time using 8 cores, fitting 100 factors.

Model validation I: determining expression drivers



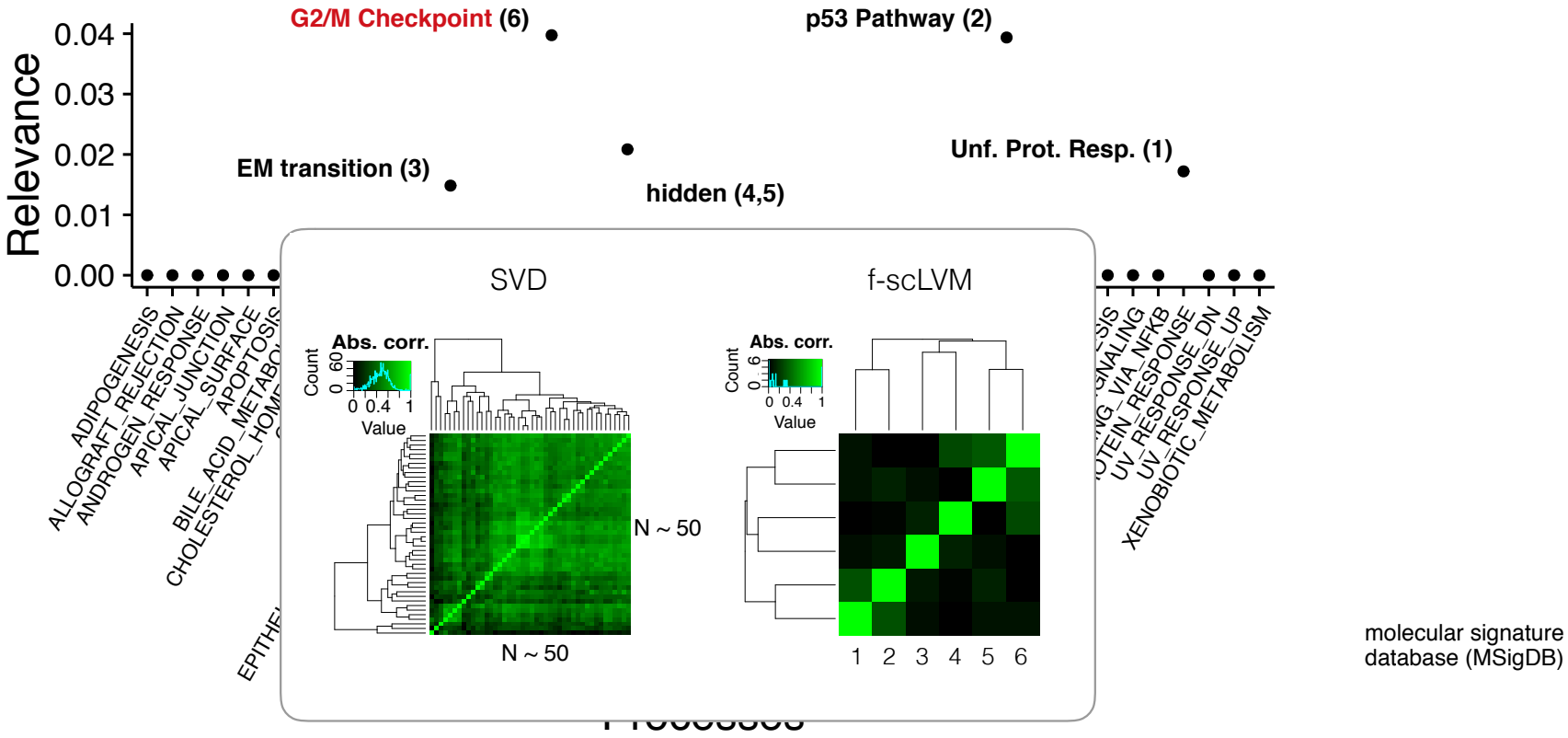
Simulated based on randomly selected Reactome terms.
Classical (sparse) factor analysis models require post-hoc annotation of factors; here using a competitive enrichment test.

Model validation II: refining factor annotations



Use case I: Identifying drivers of expression variability

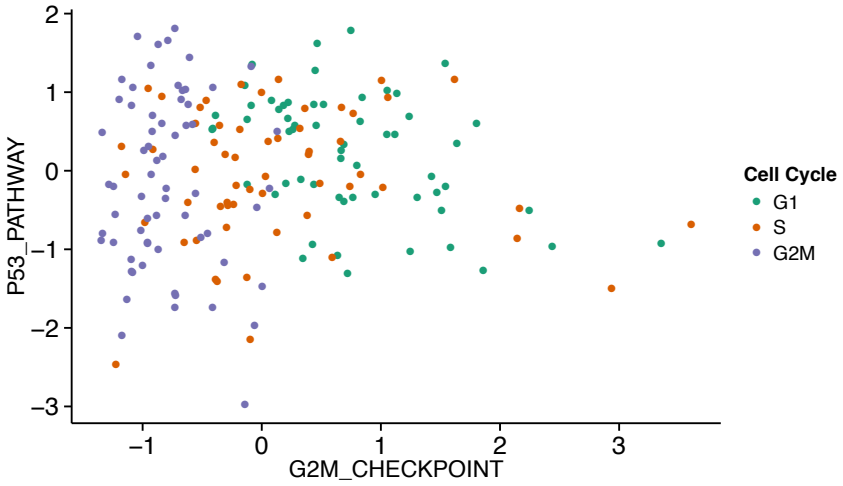
- To test our model, we used single-cell RNA-Seq data generated from ~300 ES cells collected at different stages of the cell cycle

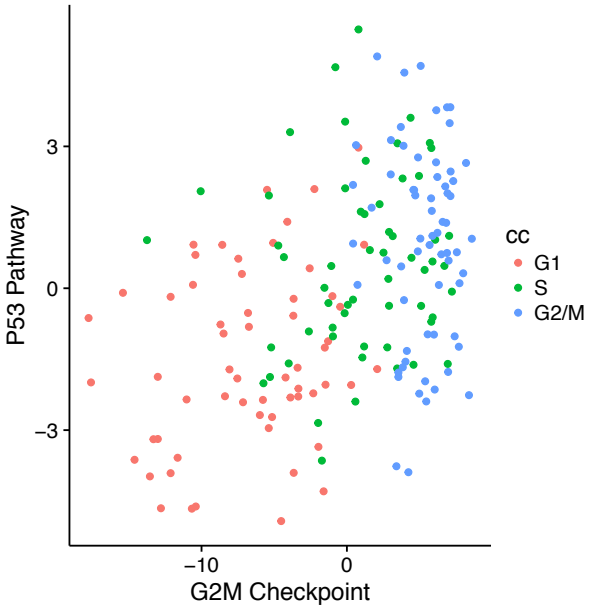


- factorial scLVM automatically discovers the presence of cell cycle

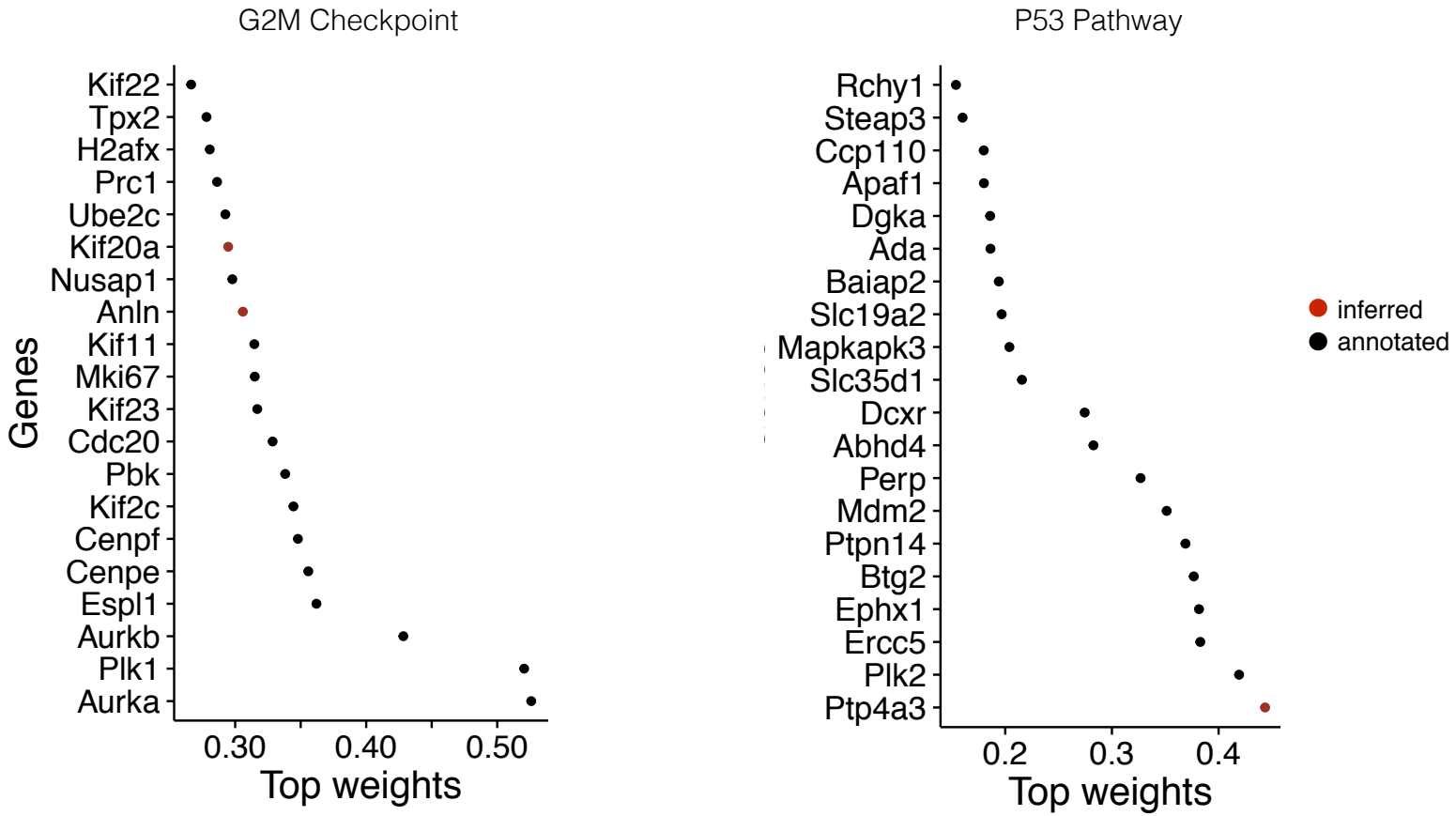
Use case I: Identifying drivers of expression variability

- To test our model, we used single-cell RNA-Seq data generated from ~300 ES cells collected at different stages of the cell cycle



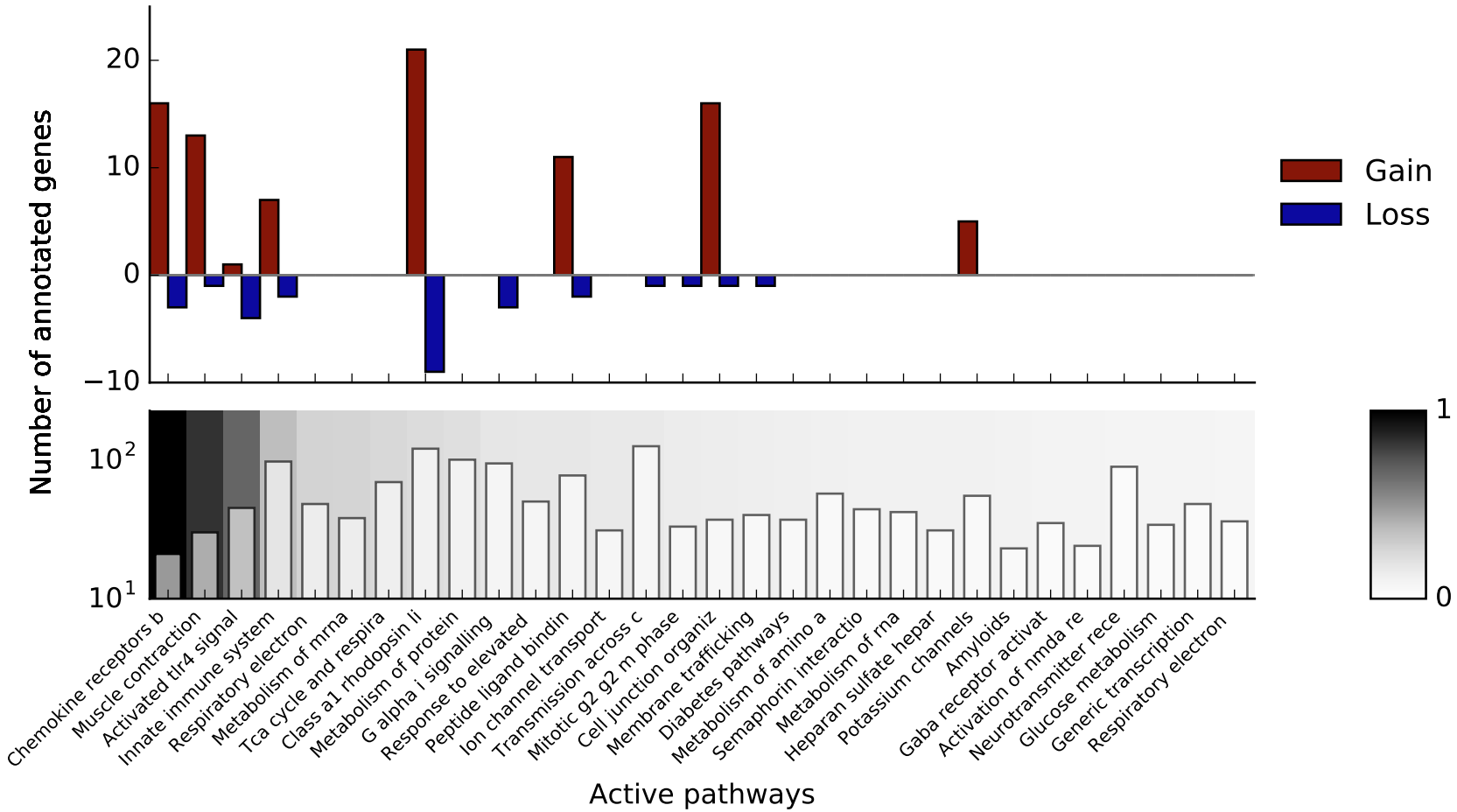


Use case II: Gene set completion

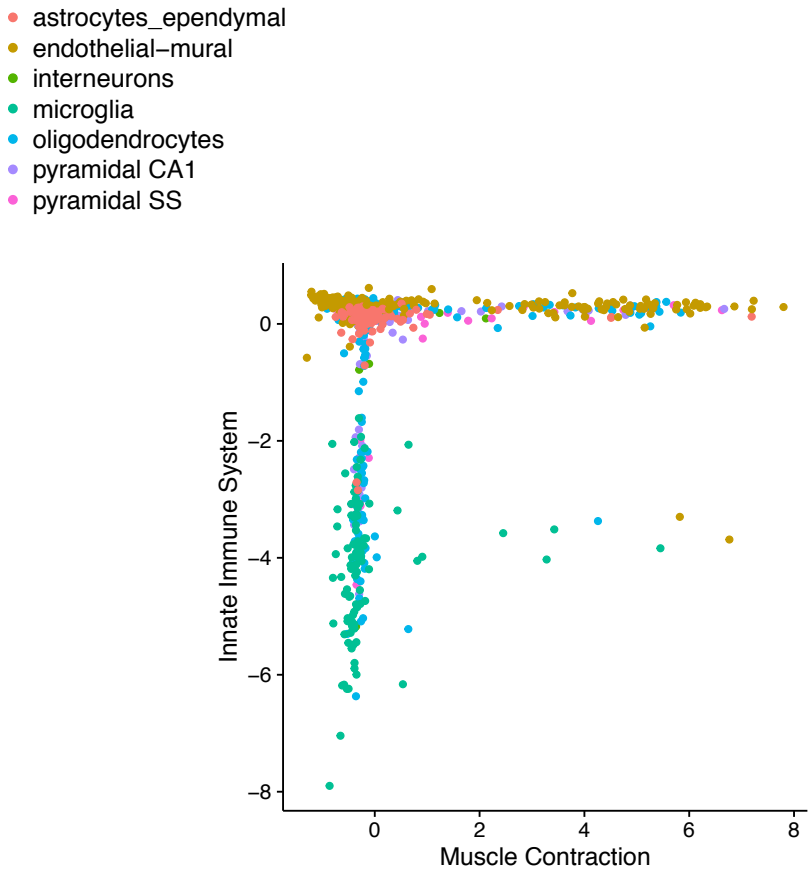


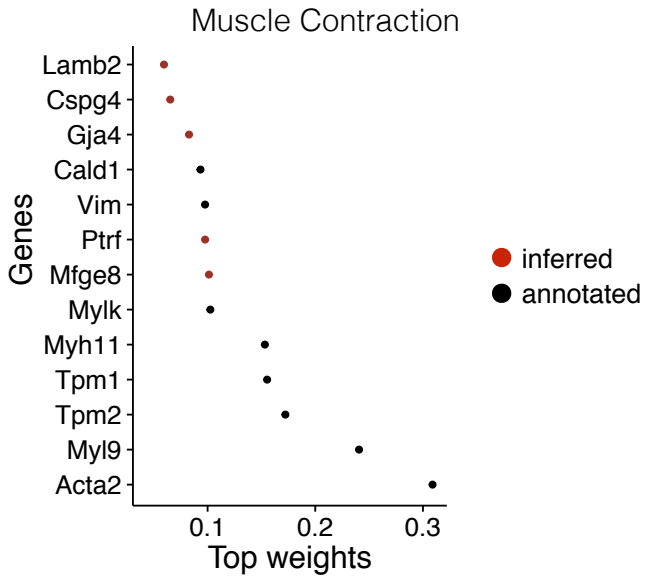
Use case II: Gene set completion

- To further illustrate the ability of our model to complete gene sets, we applied it to 3,005 neurons, classified in 7 subtypes (Zeisel et al. 2015)



Gene set completion - neurons

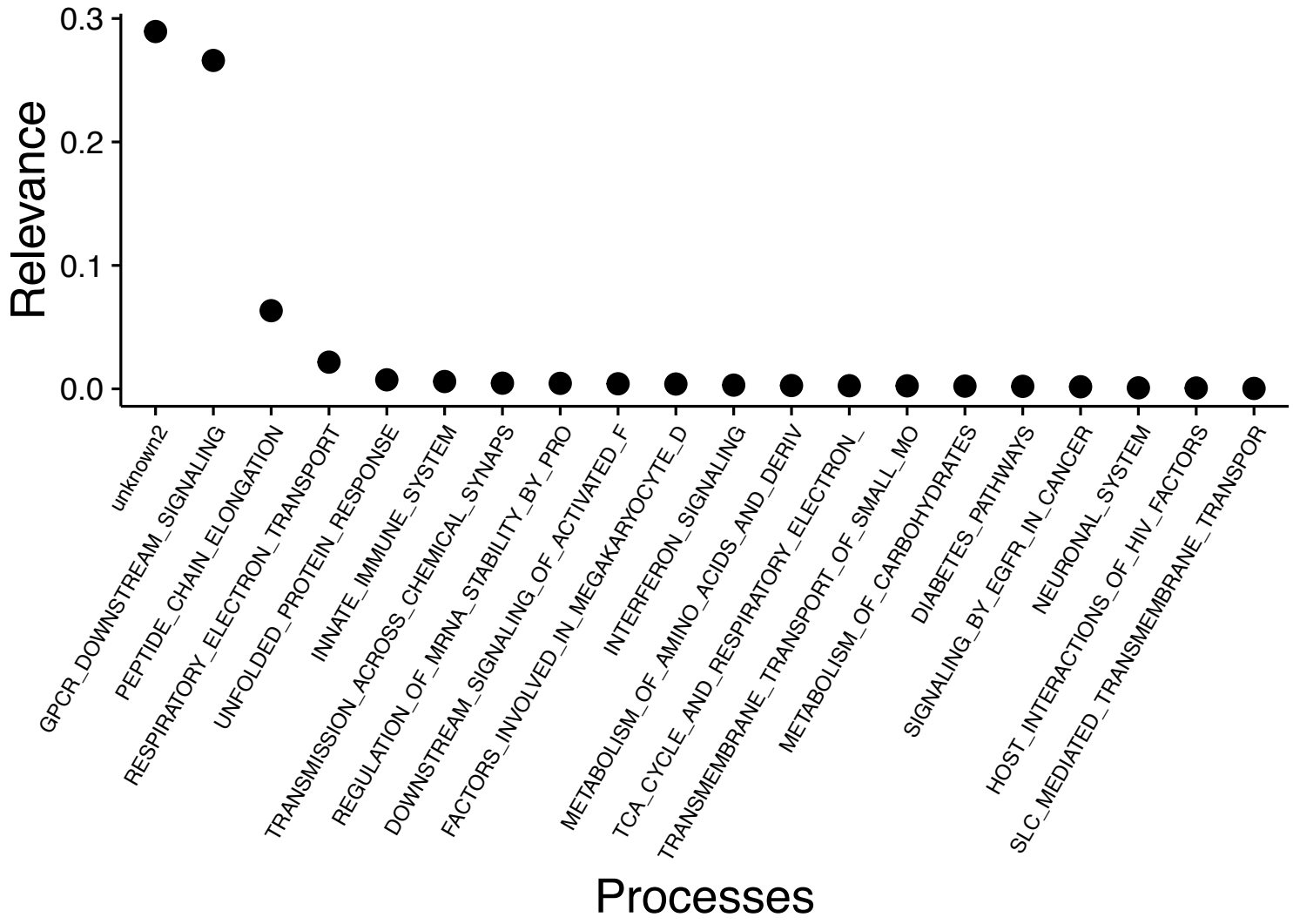




Known markers of vascular smooth muscle cells

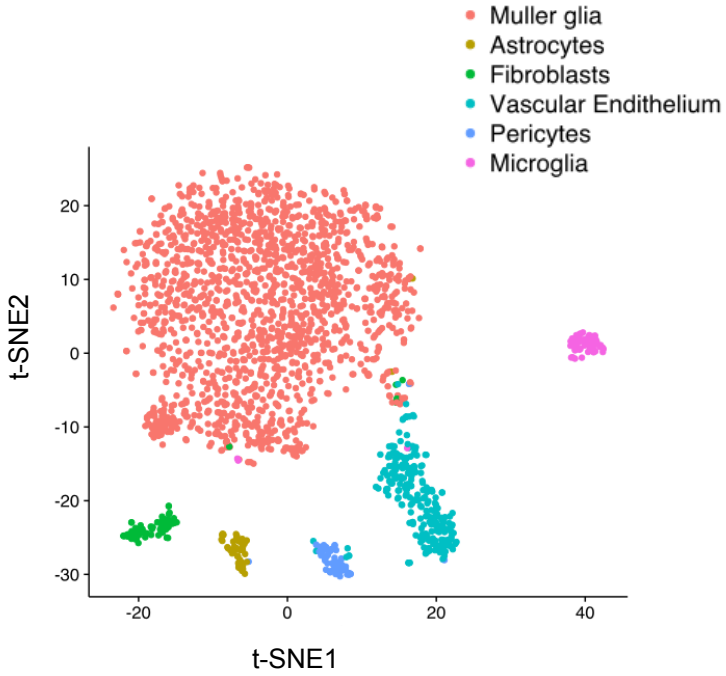
Use case III: Removing unwanted variation

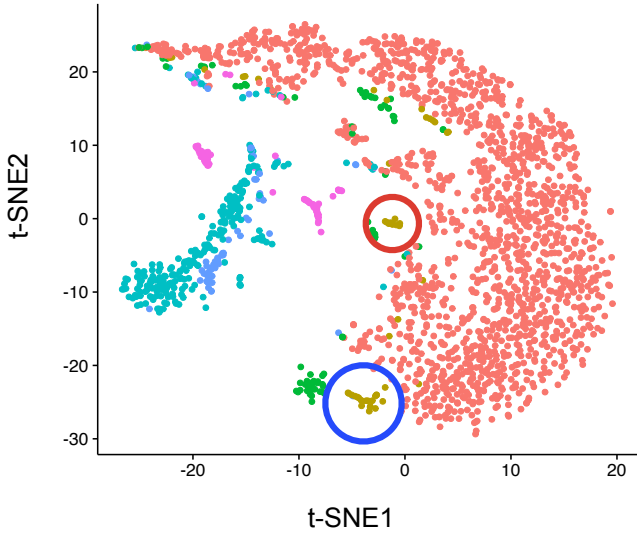
- Application to 45,000 retina cells profiled using Drop-seq

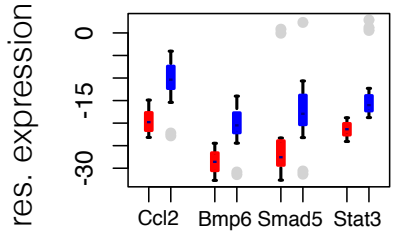


Use case III: Removing unwanted variation

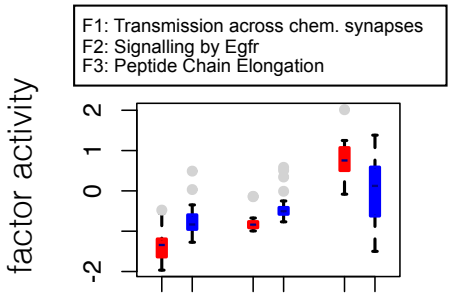
- Application to 45,000 retina cells profiled using Drop-seq





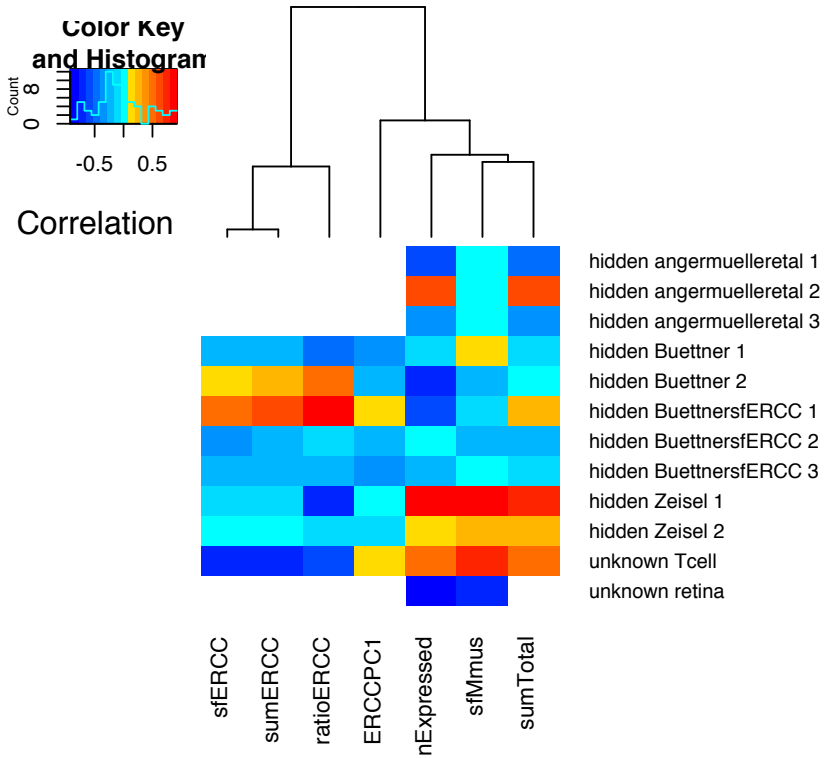


Differentially expressed genes



Differentially expressed factors

What are the inferred unannotated factors ?



Summary

- Sparse factor analysis models using informative gene sets allow decomposing the source of variation in scRNA-seq
  - identify both biological and technical sources of variation
  - complete gene sets
  - remove unwanted variation
- Fast variational approximations allow scaling these methods to very large datasets sizes with 100s of thousands of cells.

Acknowledgements

- Ploy Pratanwanich, John Marioni, Oli Stegle (EBI, Cambridge)

MRC Medical Research Council

- Software: github.com/PMBio/f-scLVM
- Contact: fbuettner.phys@gmail.com