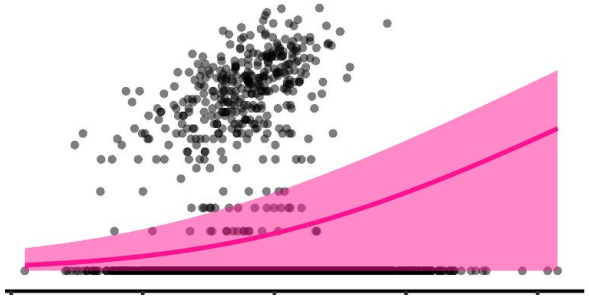
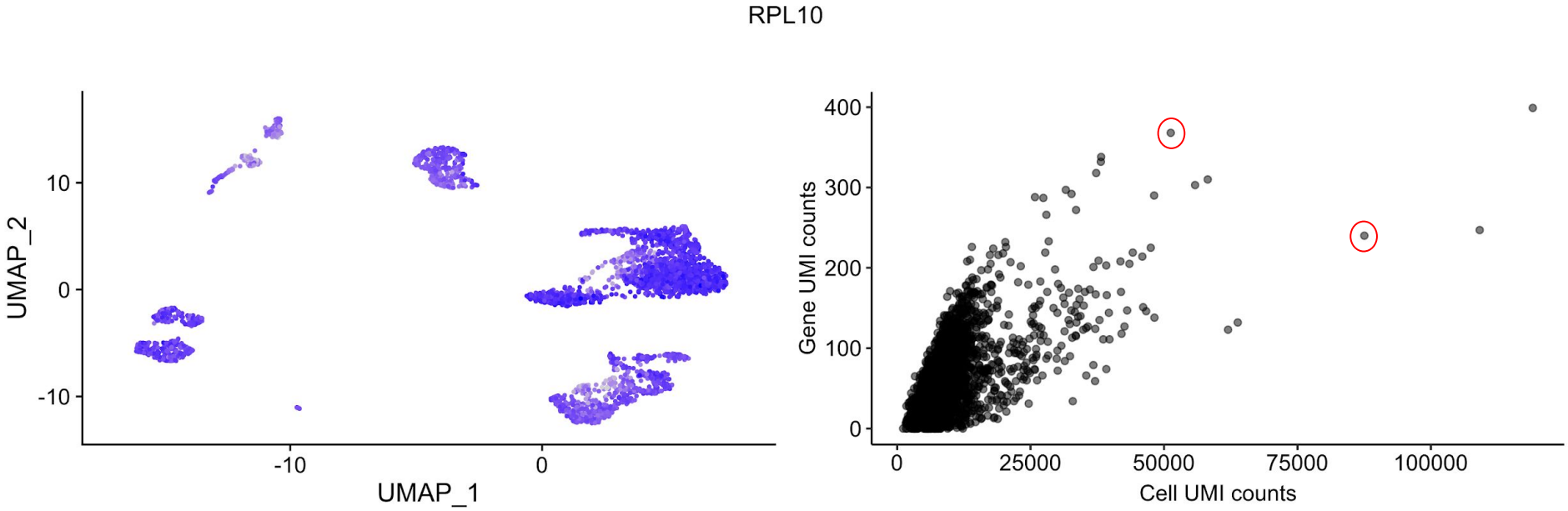
Count models for normalization of scRNA-seq data



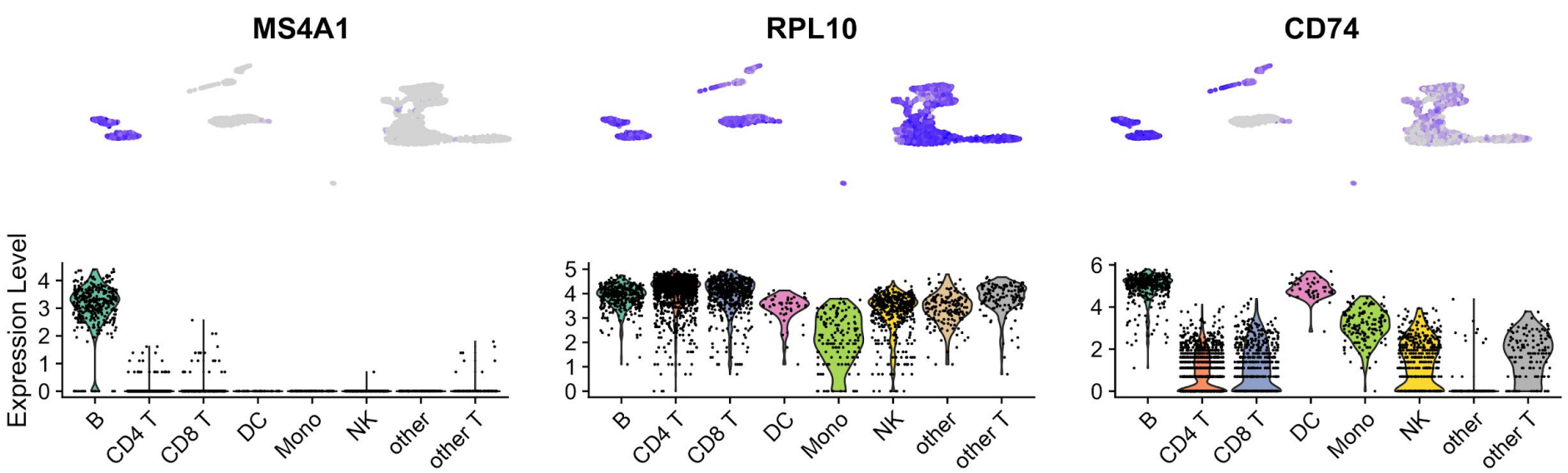
Saket Choudhary
Postdoc, Satija Lab
New York Genome Center

Many genes exhibit technical variation

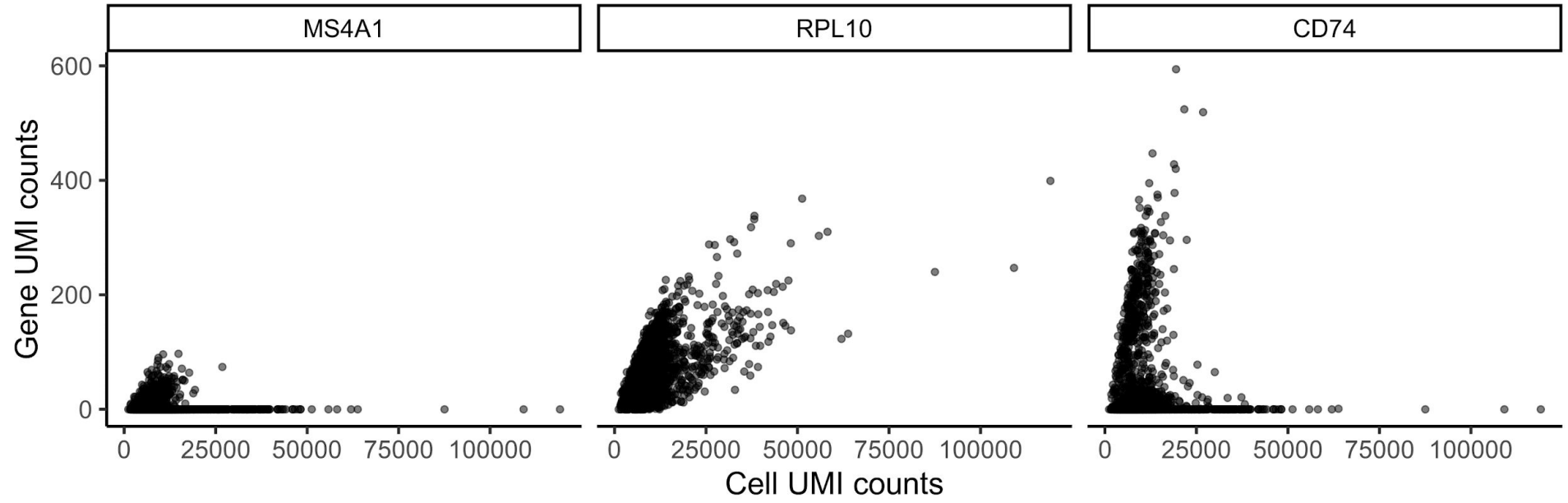


Data: Human PBMC Smart-seq3, 3k cells from Hagemann-Jensen et al., NBT 2020

Some genes exhibit both technical and biological variation



Some genes exhibit both technical and biological variation



Goal: Remove technical variation while retaining biological variation

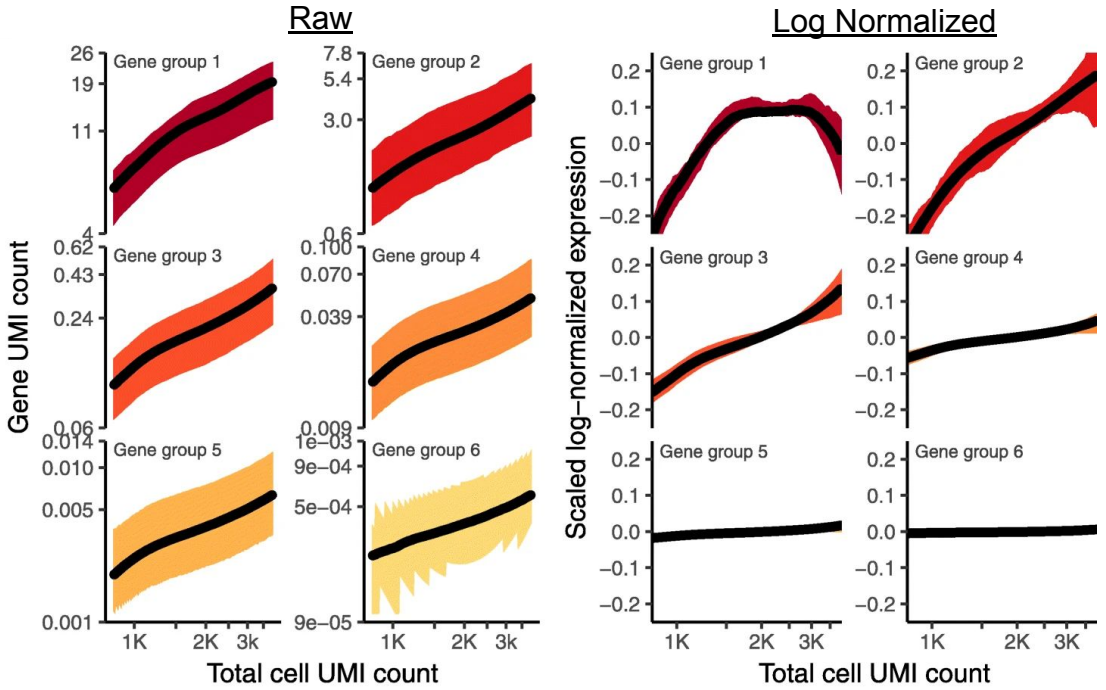
Standard Log Normalization

Step 1 (Size factor normalization):

- Goal: Correct for different sequencing depth
- Solution: Rescale cells to have same total UMI
- Dampens biological difference between cells

Step 2 (Log transformation):

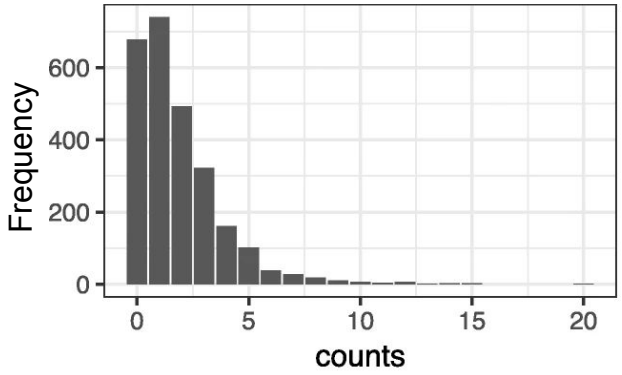
- Goal: Reduce impact of expression outliers
- Solution: Add pseudocount (+1) to avoid negative values
- Dampens both technical and biological variance



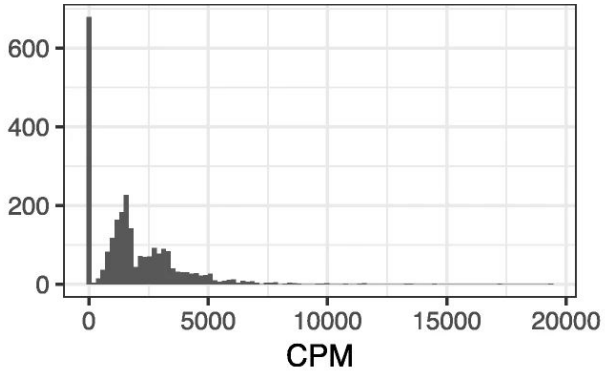
Hafemeister & Satija, Genome Biol. 2019

Differentially normalizes only low/medium abundance genes

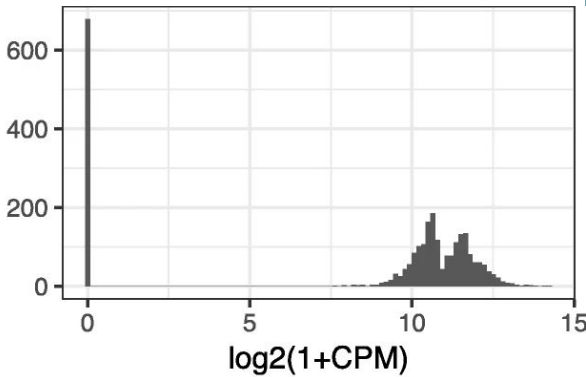
Standard Log Normalization



UMI counts



Step 1 LogNorm

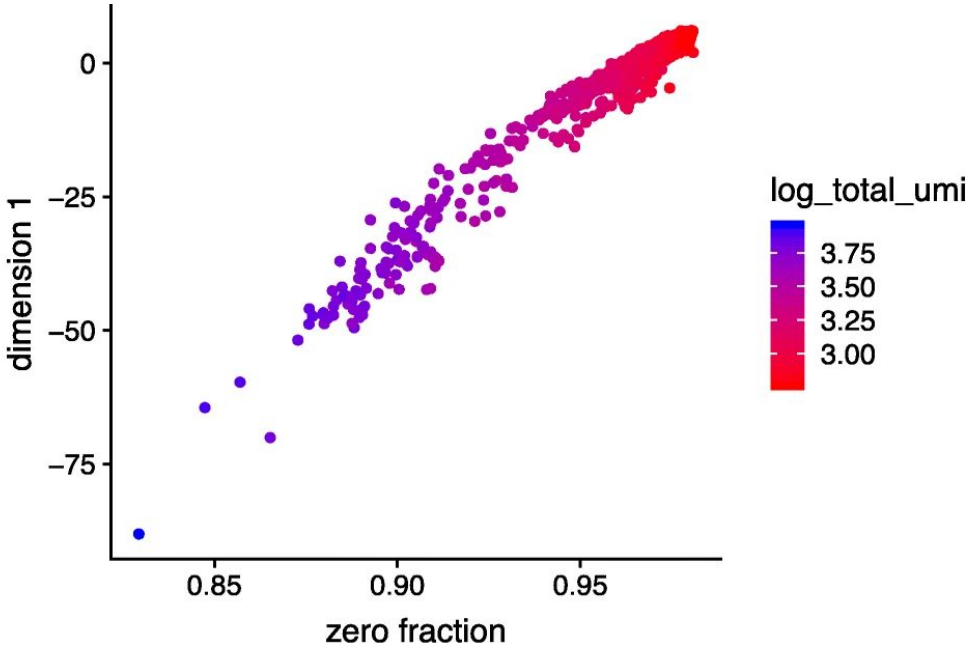


Step 2 LogNorm

Townes et al., Genome Biol. 2019

Distorts difference between zero and non-zero counts

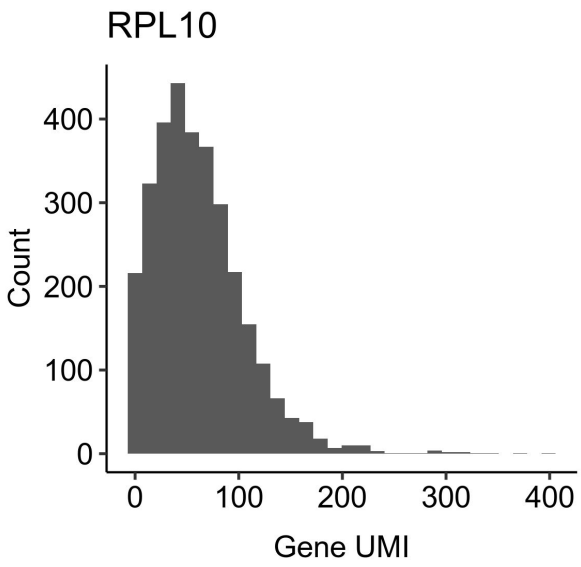
Standard Log Normalization

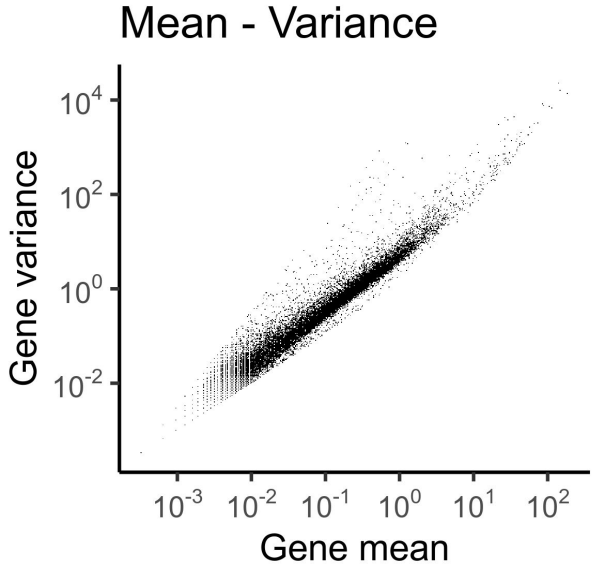


Townes et al., Genome Biol. 2019

A lot of variance even post normalization is explained by difference in sequencing depth

Count models for normalizing UMI data





- UMI counts are not normally distributed
- Variance depends on mean

Goal: Remove the effect of technical variation using an appropriate error model

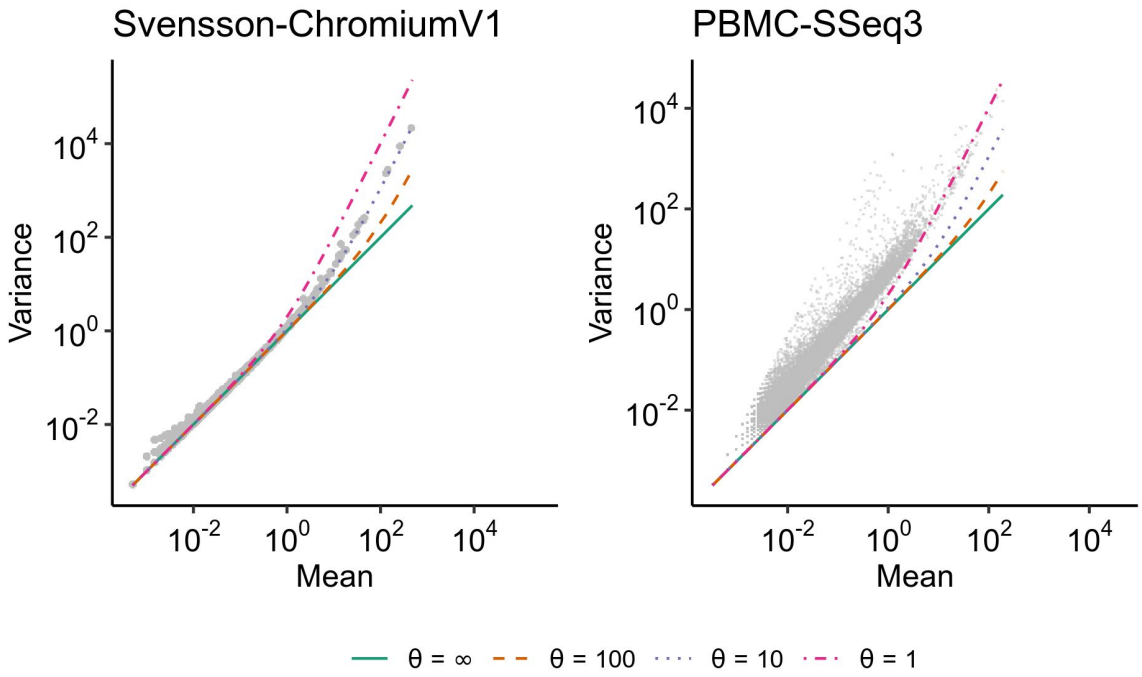
Negative Binomial distribution for modeling UMIs

Poisson:

Mean = μ
Variance = μ

Negative Binomial:

Mean = μ
Variance = $\mu + \mu^2/\theta$
Inverse-dispersion parameter = θ



Modeling technical noise

Goal

Learn a model that describes the **technical variance** in the data

Approach

- An “ideal” model should fit *perfectly* to genes showing only **technical variance**
- Biologically **‘interesting’** genes should **deviate** from the model
- Use deviations (residuals) from the model for downstream analysis

Generalized Linear Models: SCTransform

x_{gc} = Observed UMI count of gene g in cell c,

μ_{gc} = Expected UMI count of gene g in cell c,

n_c = Observed total UMI count of cell c,

$$x_{gc} \sim \text{NB}(\mu_{gc}, \boldsymbol{\theta_{gc}}),$$

$$\log \mu_{gc} = \boldsymbol{\beta_{g0}} + \boldsymbol{\beta_{g1}} n_c.$$

inverse-dispersion

Should fit the "uninteresting" genes *perfectly*

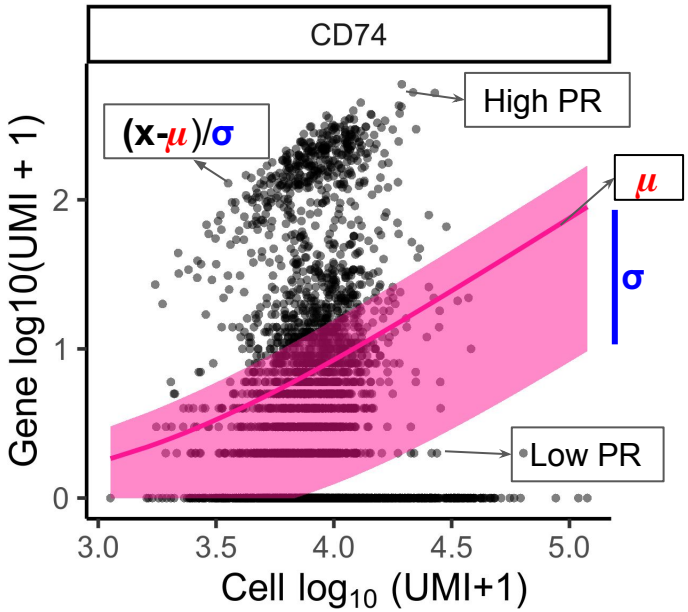
Intercept

Slope

$$\sigma_{gc} = \sqrt{\mu_{gc} + \frac{\mu_{gc}^2}{\theta_{gc}}},$$

Pearson residual

$$z_{gc} = \frac{x_{gc} - \mu_{gc}}{\sigma_{gc}}.$$



Idea: Learn a model of technical noise from the data

Generalized Linear Models: GLM-PCA

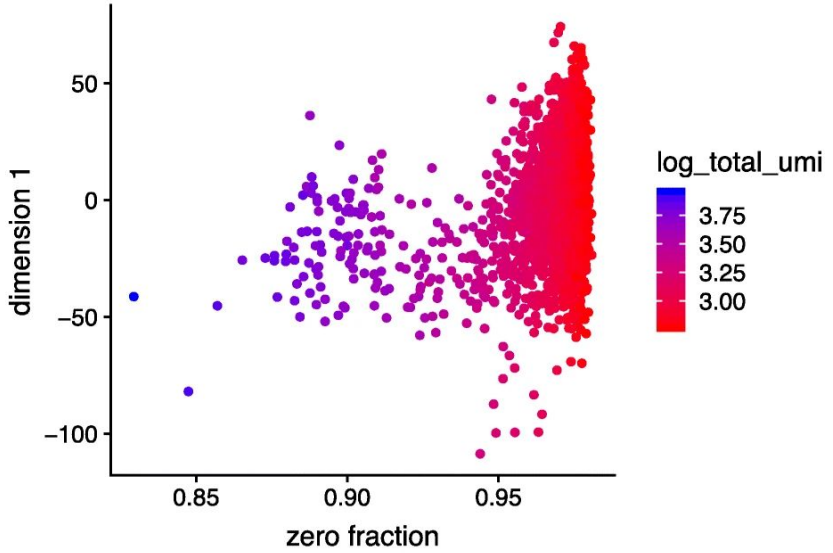
$$x_{gc} \sim \text{NB}(\mu_{gc}, \theta_{gc}) \text{ or Poisson}(\mu_{gc}),$$

$$\log \mu_{gc} = \log n_c + \sum_{l=0}^{k} U_{cl} V_{lg},$$

Loadings

PCs

$$D = \sum_{g,c} x_{gc} \log \frac{x_{gc}}{\mu_{gc}} - (x_{gc} - \mu_{gc})$$



Townes et al., Genome Biol. 2019

Idea: GLM residuals are asymptotically normal

How should we determine inverse-dispersion (θ)?

One θ for all datasets

New Results

Comment on this paper

Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data

Jan Lause, Philipp Berens, Dmitry Kobak

doi: https://doi.org/10.1101/2020.12.01.405886

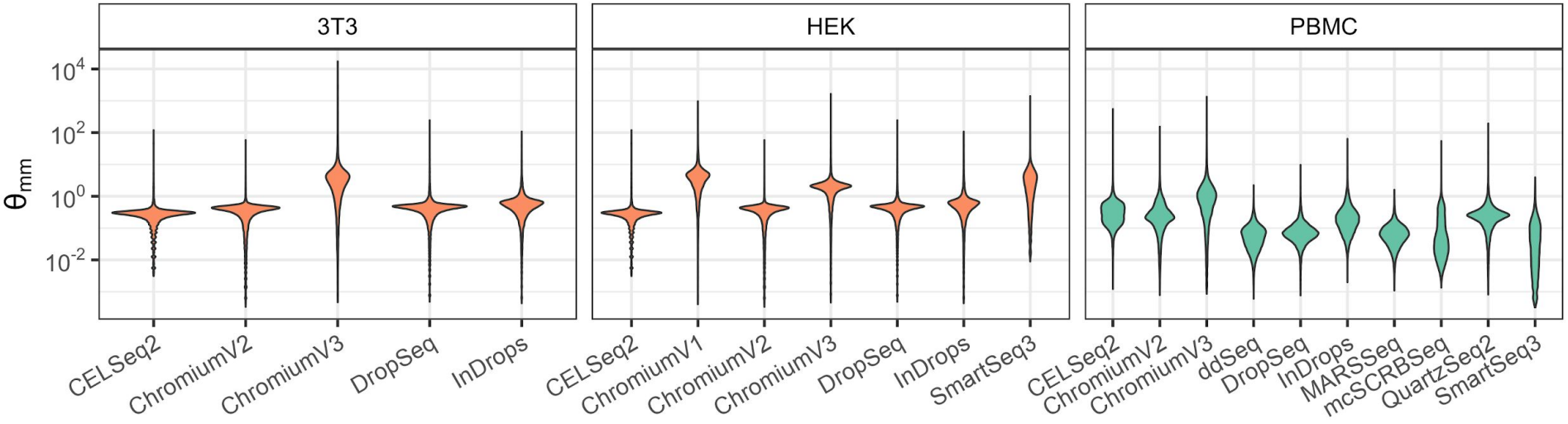
$$\beta_{g0} = \log \frac{1}{N} \sum_c x_{c,g} - \log \frac{1}{N} \sum_c n_c$$

$$\beta_{g1} = 1$$

$$\theta_{gc} = \{10, 100\}$$

Question: Can we fix dispersion estimates across all datasets?

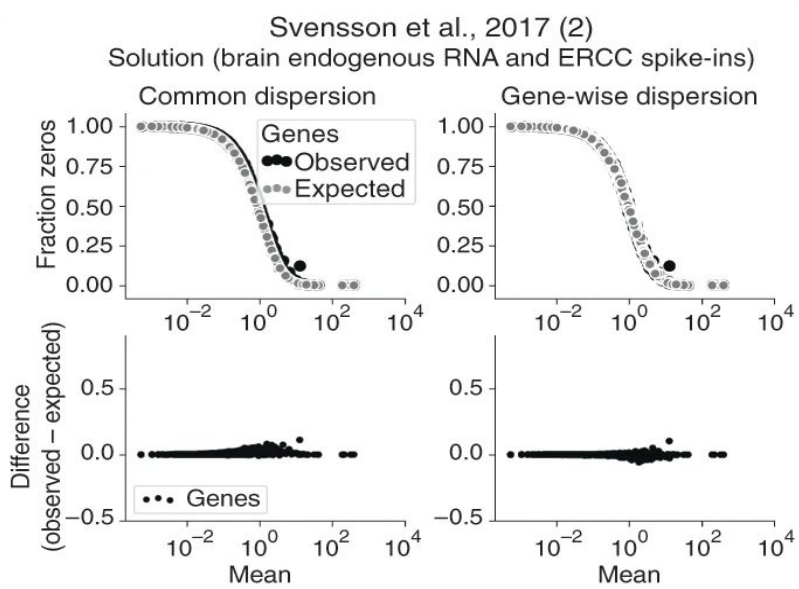
Dispersion varies across technologies

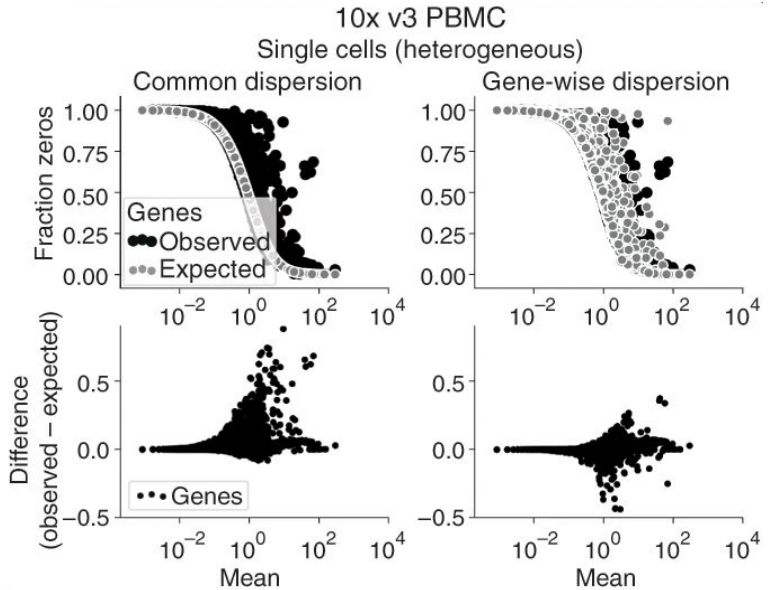


Data from: Mereu et al., NBT 2020; Ding et al., NBT 2020; Hagemann-Jensen et al., NBT 2020

Question: Can we estimate dispersion per gene per dataset?

Estimating dispersion per gene



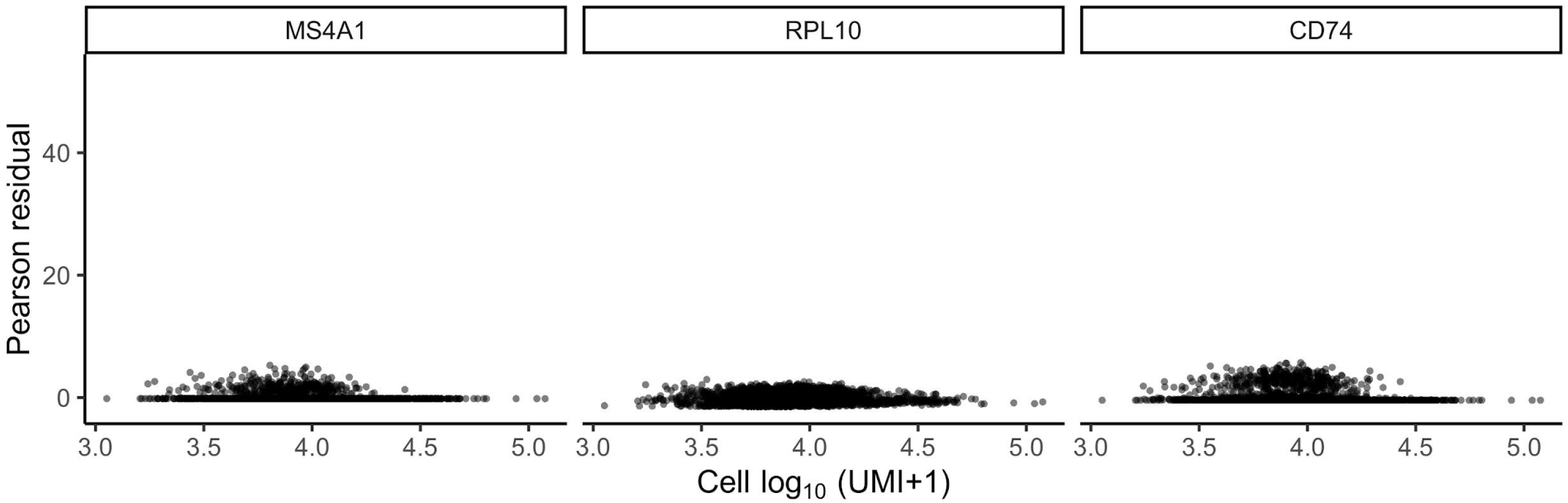


% deviation	4.7%	0.49%	9.65%	3.82%

Svensson, NBT 2020

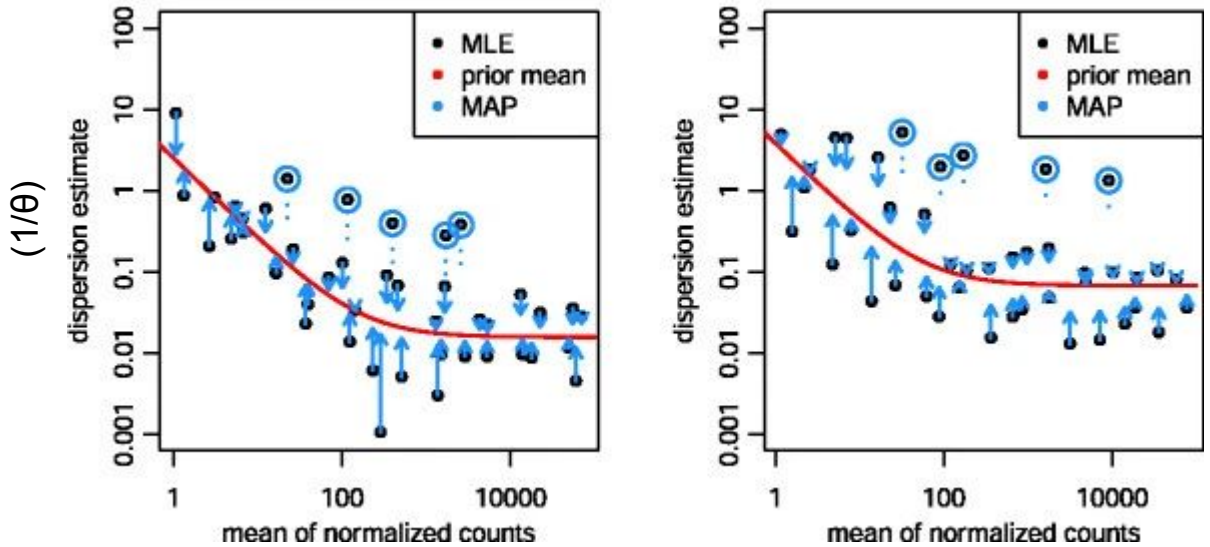
Genewise dispersion estimates lead to better fit

Learning per gene estimate dampens both biological and technical variance



Question: Is there a middle ground between learning per gene estimates and fixing dispersion estimates?

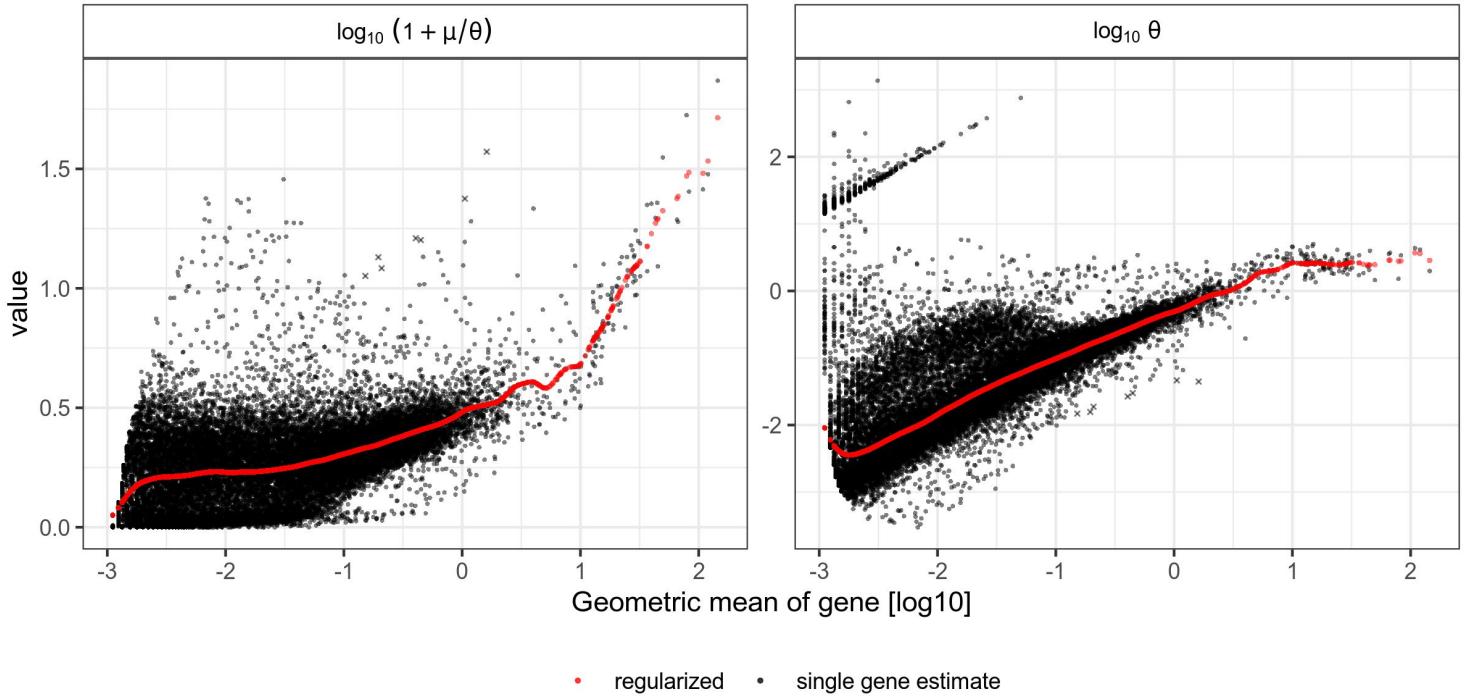
Sharing information across genes



Love et al., Genome Biol. 2014

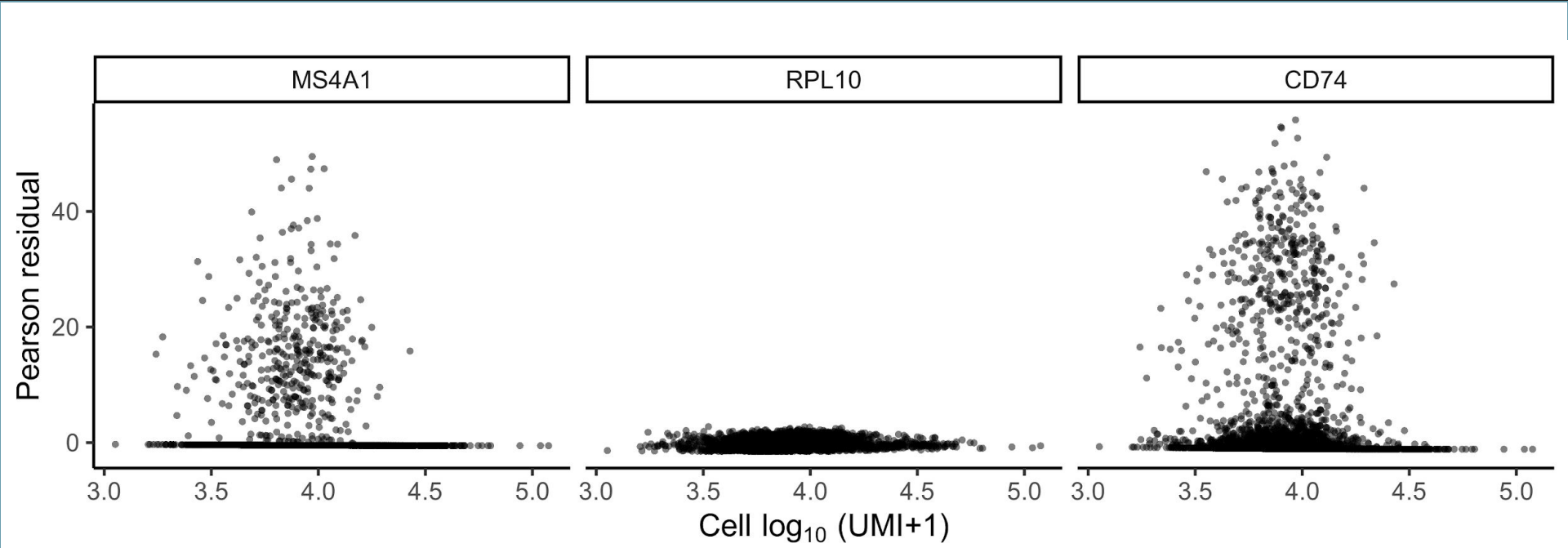
Genes with <u>similar mean</u> have <u>similar dispersion</u>

Sharing information across genes



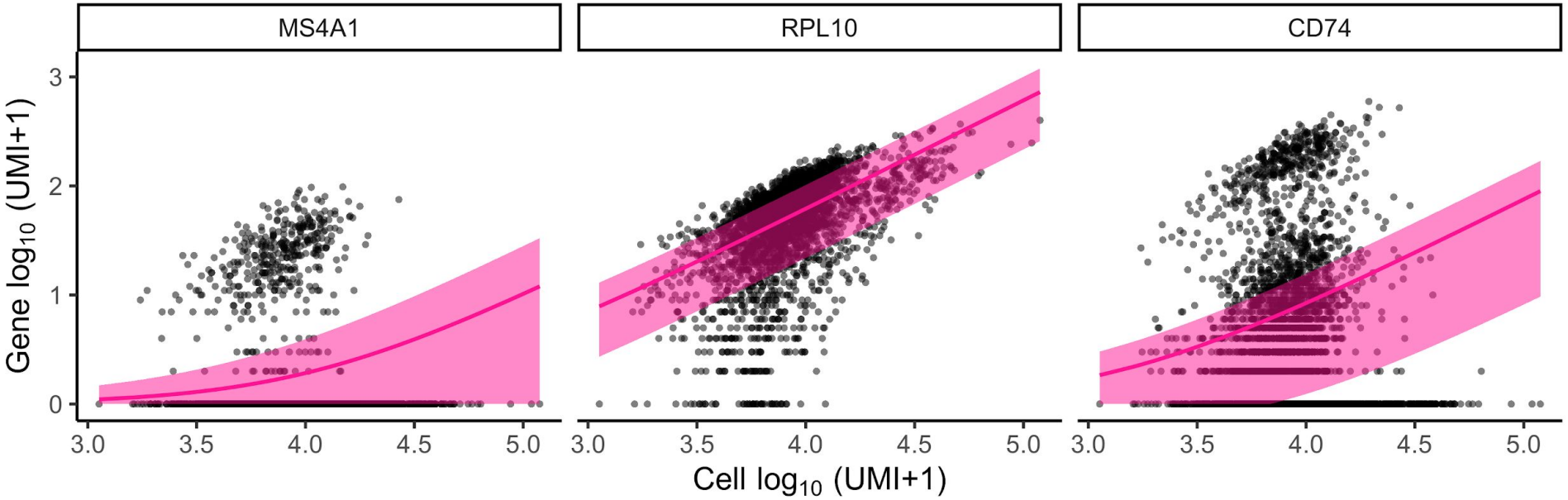
Model captures expected variation for a given gene mean by regularizing dispersion

SCTransform learns a model of technical noise from each dataset



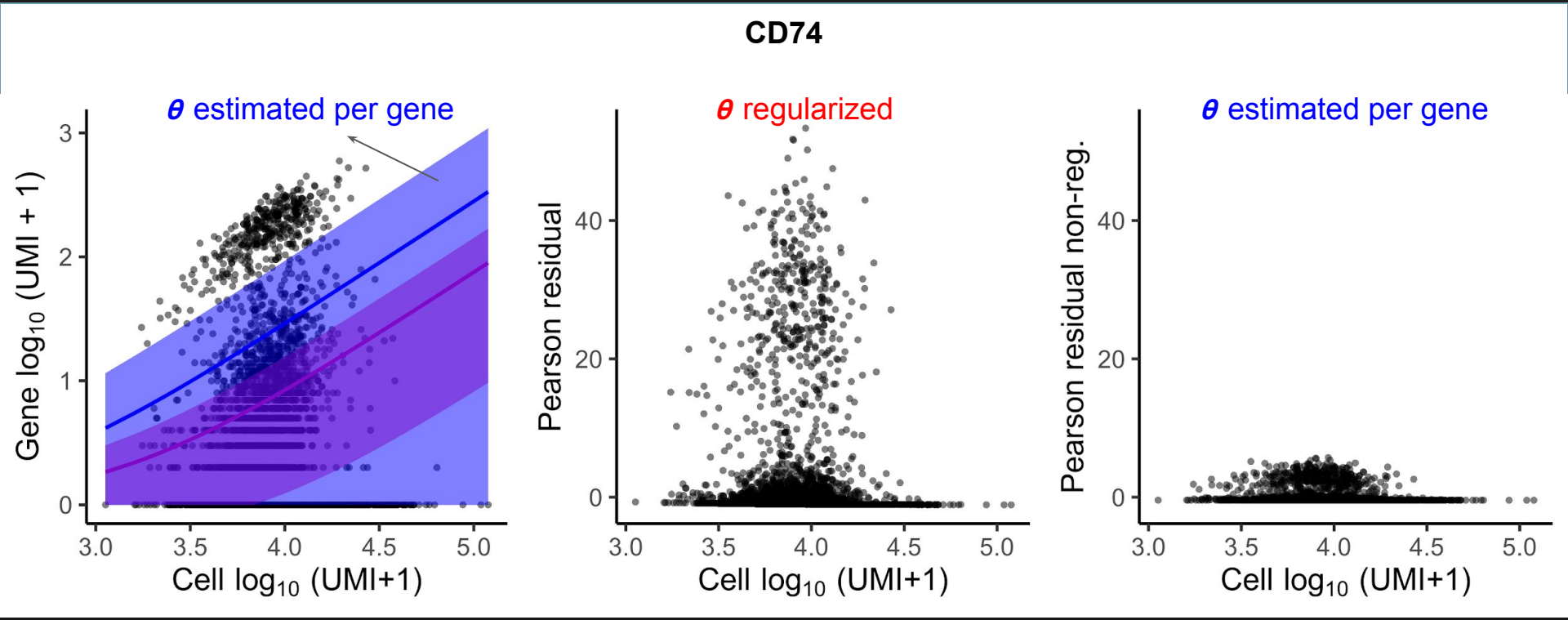
Variability that cannot be explained by the fit is <u>biological</u>

SCTransform learns a model of technical noise from each dataset



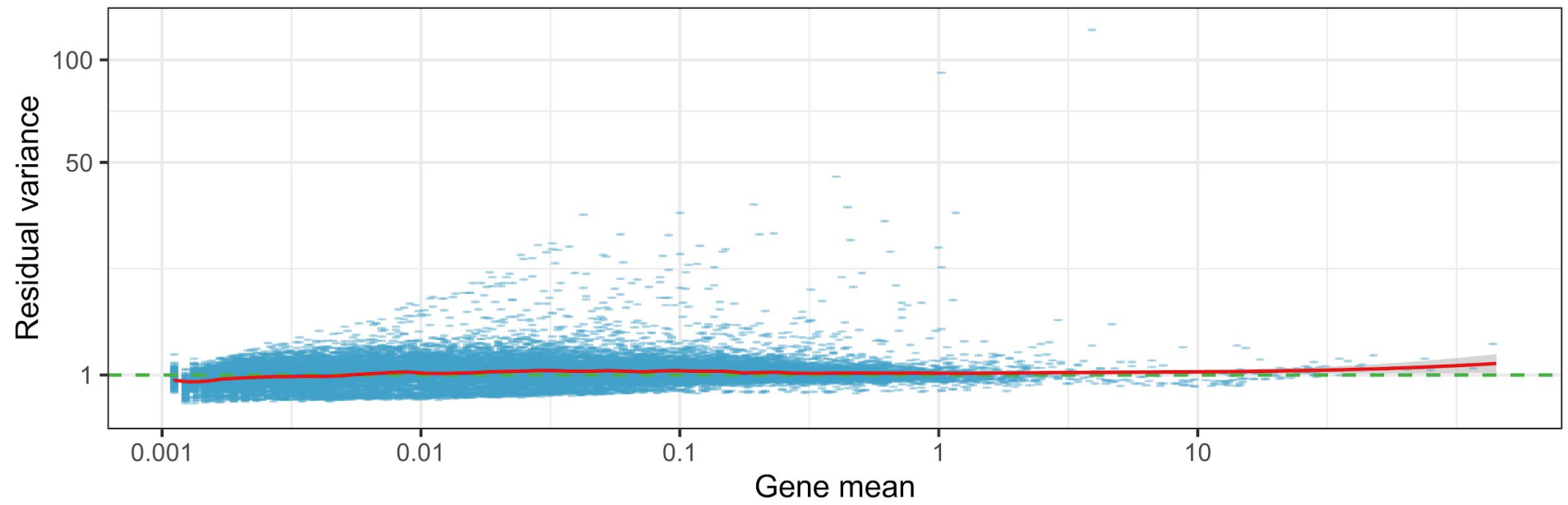
Biologically interesting genes show deviation from the model while uninteresting ones fit *perfectly*

SCTransform learns a model of technical noise from each dataset



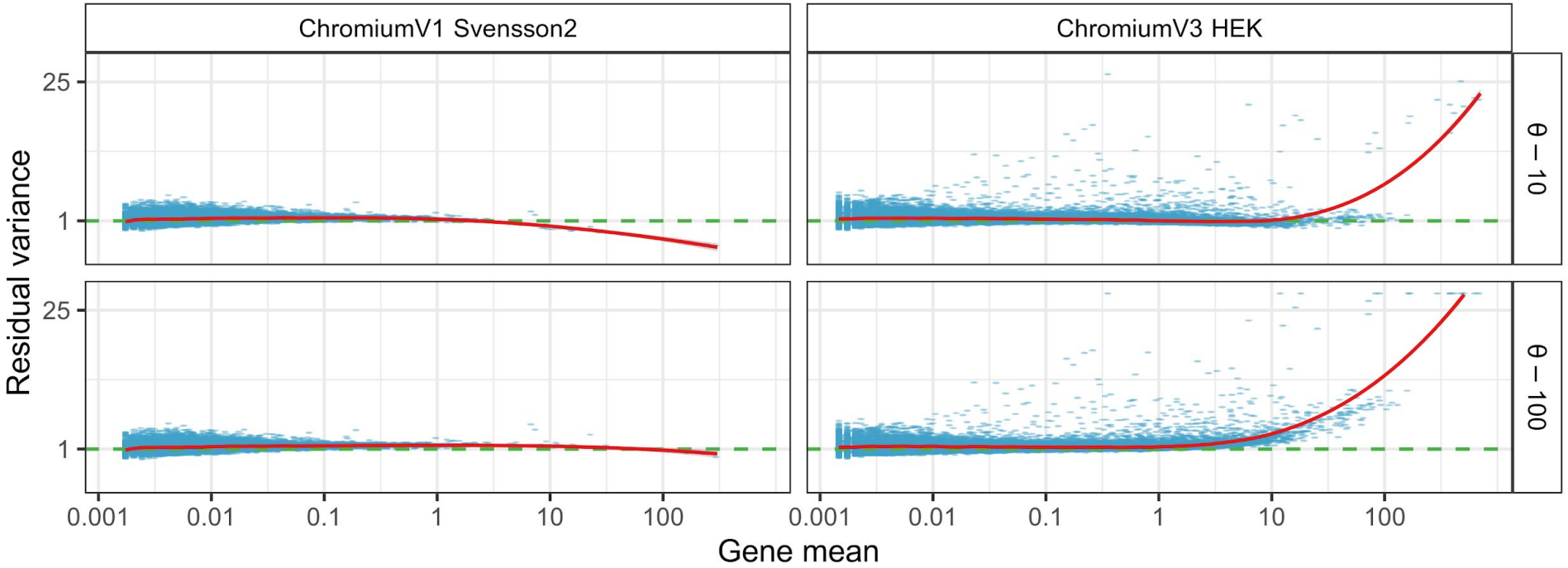
Per gene estimates dampen biological variability

SCTransform achieves variance stabilization



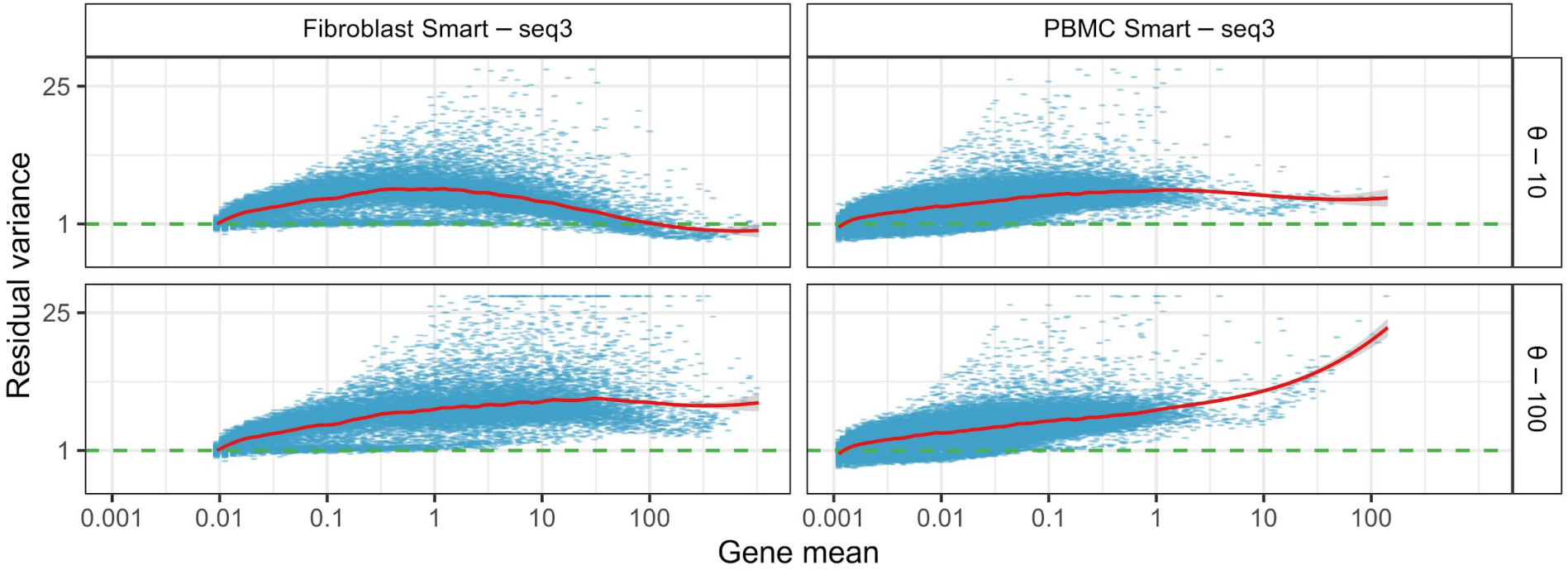
Residual variance is <u>uncorrelated</u> with gene mean

Variance stabilization: Fixed dispersion



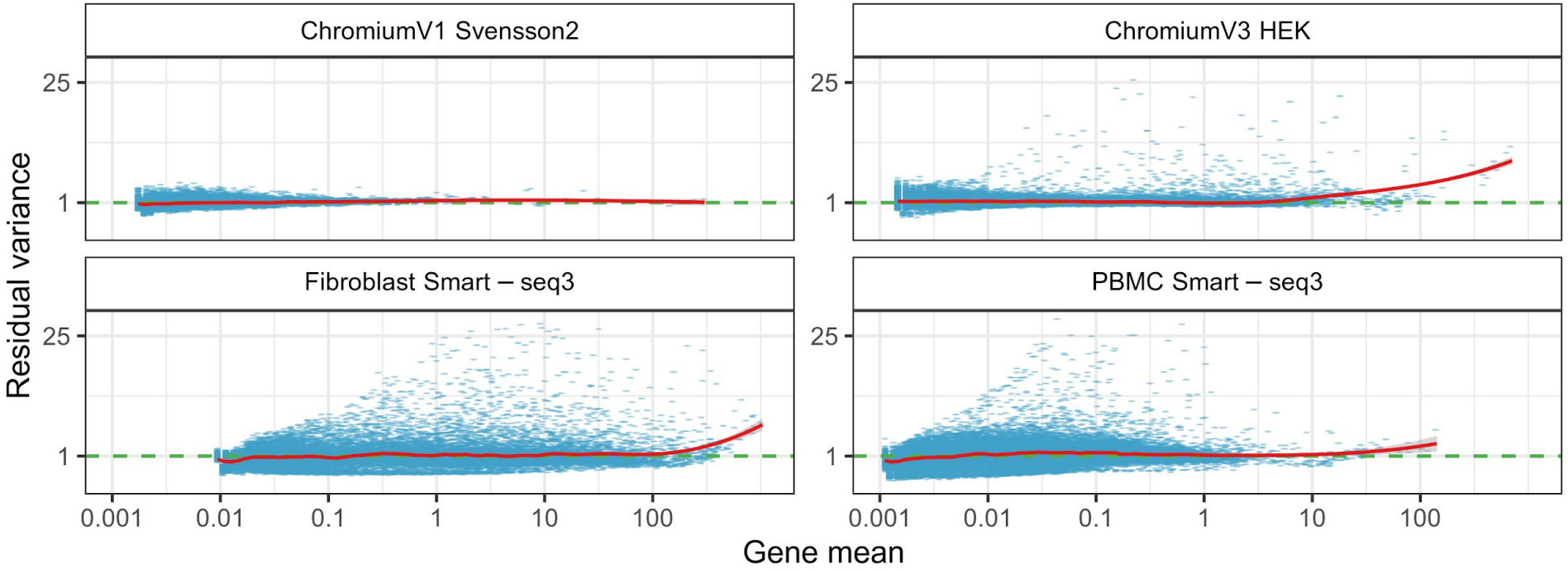
Residual variance is expected to be uncorrelated with gene mean

Variance stabilization: Fixed dispersion



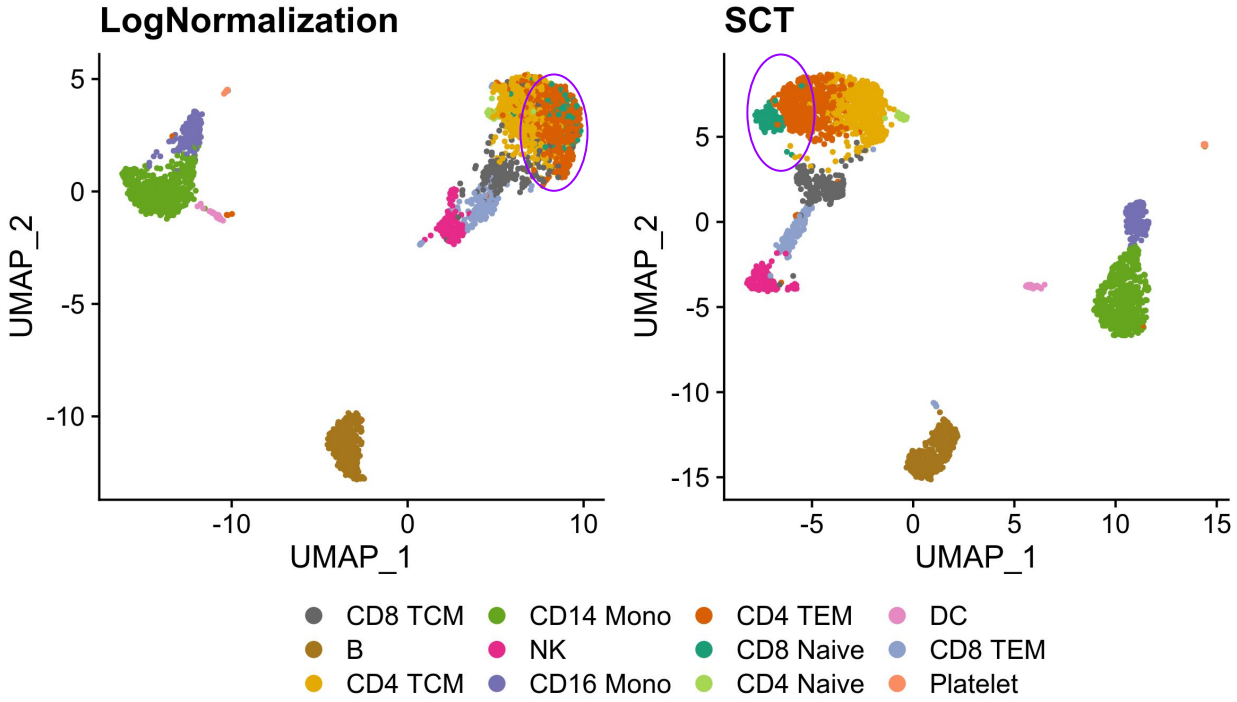
Residual variance is expected to be uncorrelated with gene mean

Variance stabilization: Regularized dispersion



Residual variance is <u>uncorrelated</u> with gene mean

SCTransform reveals more biological distinction



Data: PBMC3k (Chromium), 10X Genomics

Summary

- Standard Log Normalization introduces systematic errors
- GLM approach improves single-cell normalization
- Negative binomial dispersion (θ) can **vary substantially** across datasets and technologies
- We recommend learning and regularizing θ **separately** for each dataset

Other count models

[1] A. K. Sarkar and M. Stephens, “Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis,” *BioRxiv*, 2020 [Online].

[2] K. Choi, Y. Chen, D. A. Skelly, and G. A. Churchill, “Publisher Correction: Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics,” *Genome Biol.*, vol. 21, no. 1, p. 270, Nov. 2020, doi: 10.1186/s13059-020-02182-1. [Online].

[3] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, “Deep generative modeling for single-cell transcriptomics,” *Nat. Methods*, vol. 15, no. 12, pp. 1053–1058, Dec. 2018, doi: 10.1038/s41592-018-0229-2. [Online].

[4] T. H. Kim, X. Zhou, and M. Chen, “Demystifying ‘drop-outs’ in single-cell UMI data,” *Genome Biol.*, vol. 21, no. 1, p. 196, Aug. 2020, doi: 10.1186/s13059-020-02096-y. [Online].

[5] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, “Author Correction: Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model,” *Genome Biol.*, vol. 21, no. 1, p. 179, Jul. 2020, doi: 10.1186/s13059-020-02109-w. [Online].

[6] J. Lause, P. Berens, and D. Kobak, “Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data,” *bioRxiv*, 2020 [Online].

[7] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J. P. Vert, “ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data,” *BioRxiv*, 2017 [Online].

[8] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, “Single-cell RNA-seq denoising using a deep count autoencoder,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019 [Online].

[9] S. Sun, Y. Chen, Y. Liu, and X. Shang, “A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data,” *BMC Syst. Biol.*, vol. 13, no. Suppl 2, p. 28, Apr. 2019. [Online].

[10] M. Huang *et al.*, “SAVER: gene expression recovery for single-cell RNA sequencing,” *Nat. Methods*, vol. 15, no. 7, pp. 539–542, Jul. 2018, doi: 10.1038/s41592-018-0033-z. [Online].

[11] W. Tang *et al.*, “bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data,” *Bioinformatics*, vol. 36, no. 4, pp. 1174–1181, Feb. 2020, doi: 10.1093/bioinformatics/btz726. [Online].

Acknowledgements

Rahul Satija
Christoph Hafemeister
Andrew Butler
Avi Srivastava
Bill Mauck
Bingjie Zhang
Efi Papalexi
Jaison Jain
John Blair
Harm Wessels
Kristof Torkenczy
Paul Hoffman
Sandra Iborra
Shaista Madad
Tim Stuart
Yuhan Hao

NYU

NEW YORK GENOME CENTER