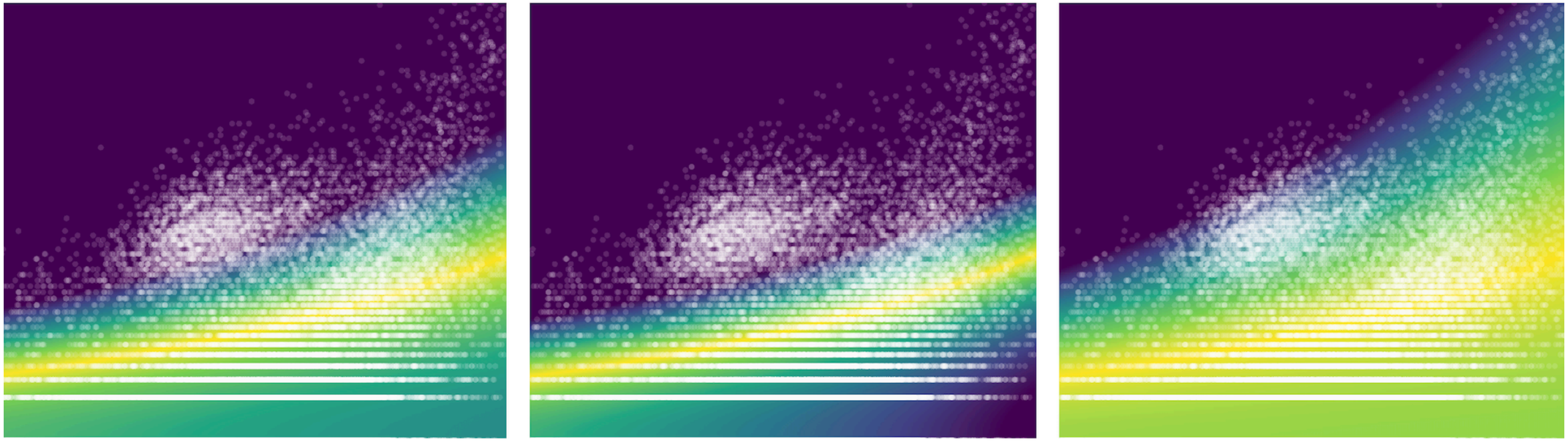
scRNA-seq data normalization using regularized negative binomial regression



Christoph Hafemeister
Rahul Satija's lab
New York Genome Center

normjam
Normalization Workshop
November 19, 2019

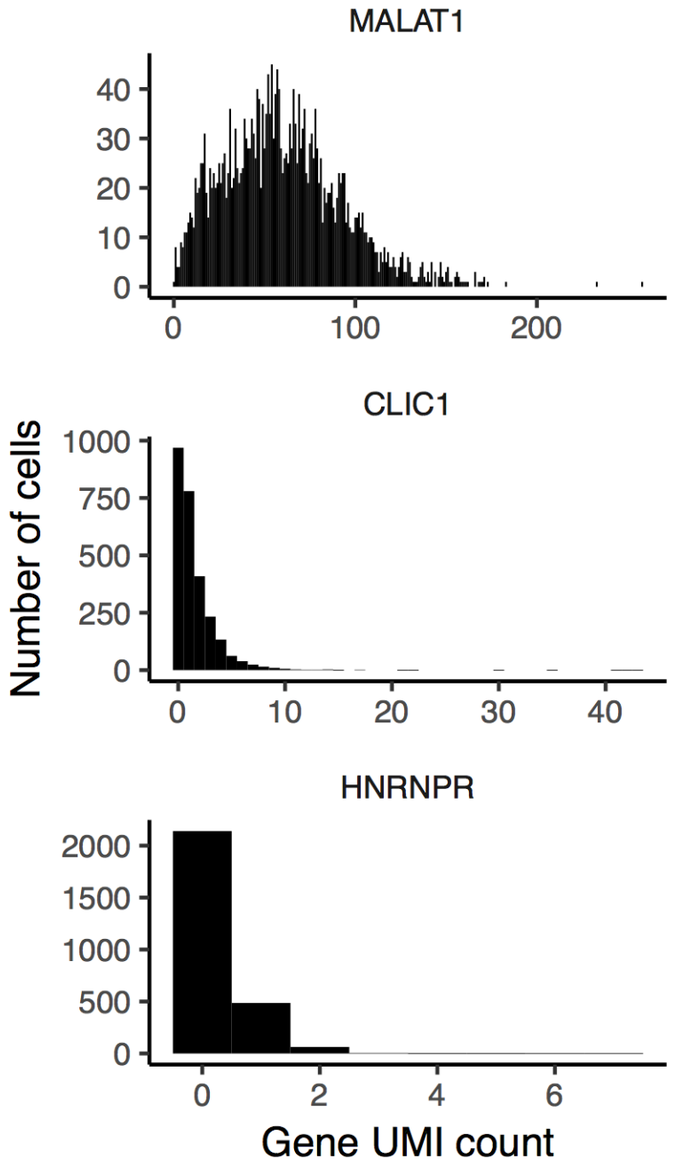
Normalization in the context of this talk

Adjust observed UMI counts across cells to a common scale while controlling for nuisance factors

Focus on sequencing depth differences between cells as technical nuisance factor and control for the variability it contributes to the observed counts

Important because downstream analysis steps should be driven by biological variability instead of technical factors

Example data: 2700 PBMCs from 10x



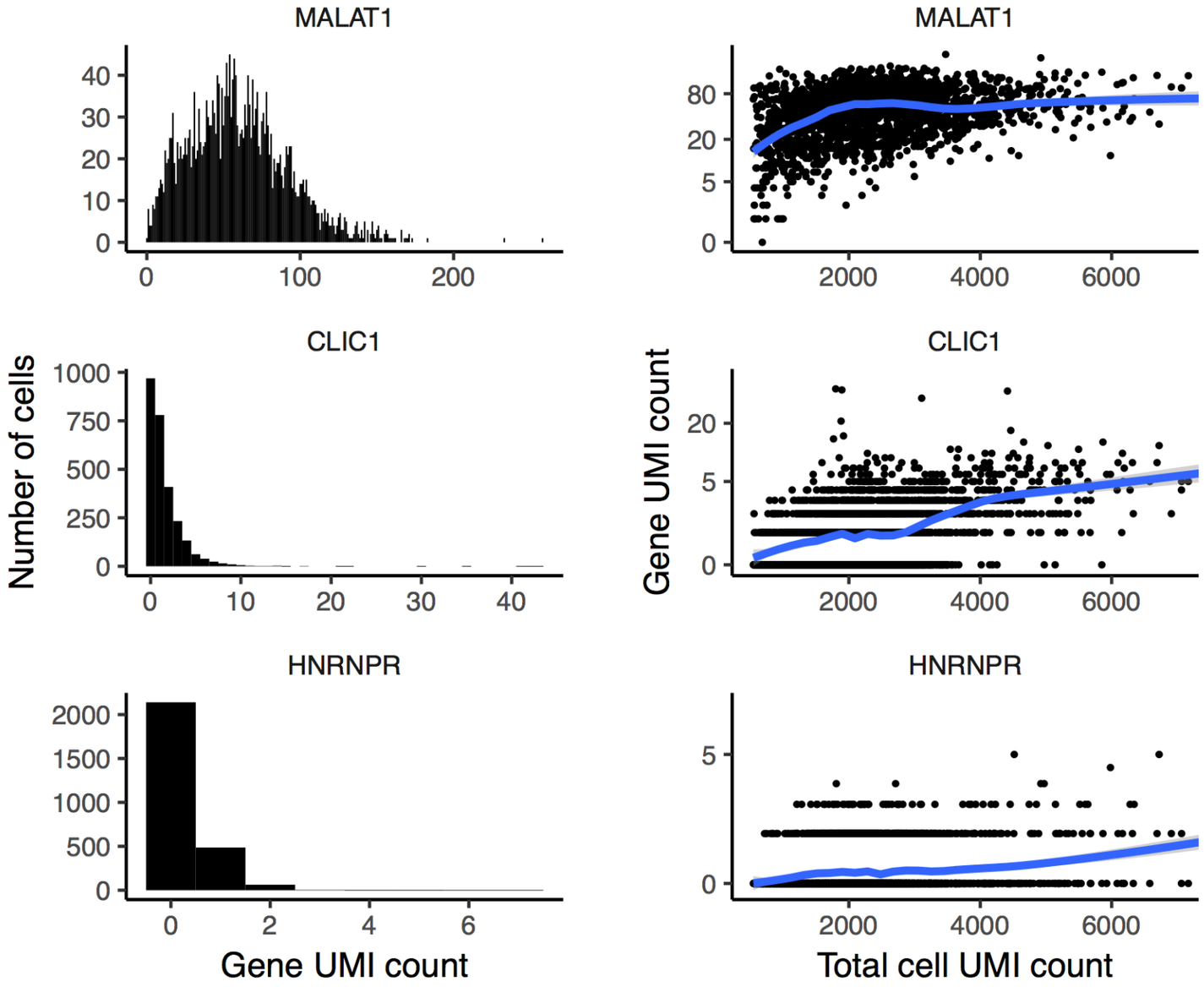
Rank based on mean count

1

200

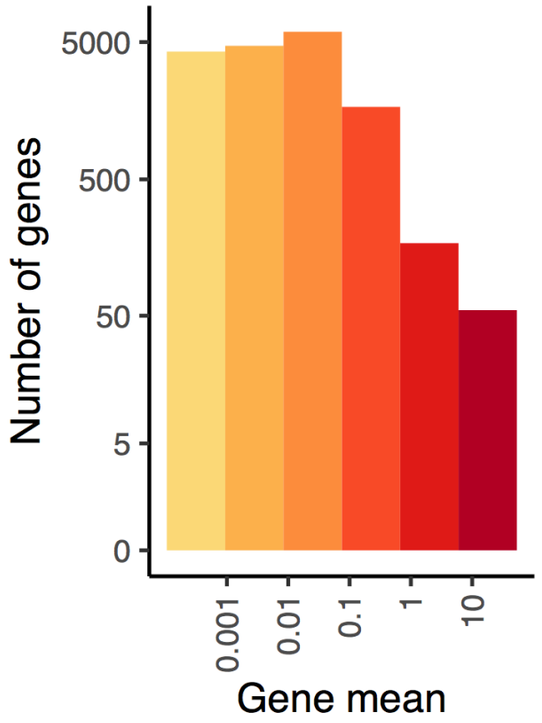
1000

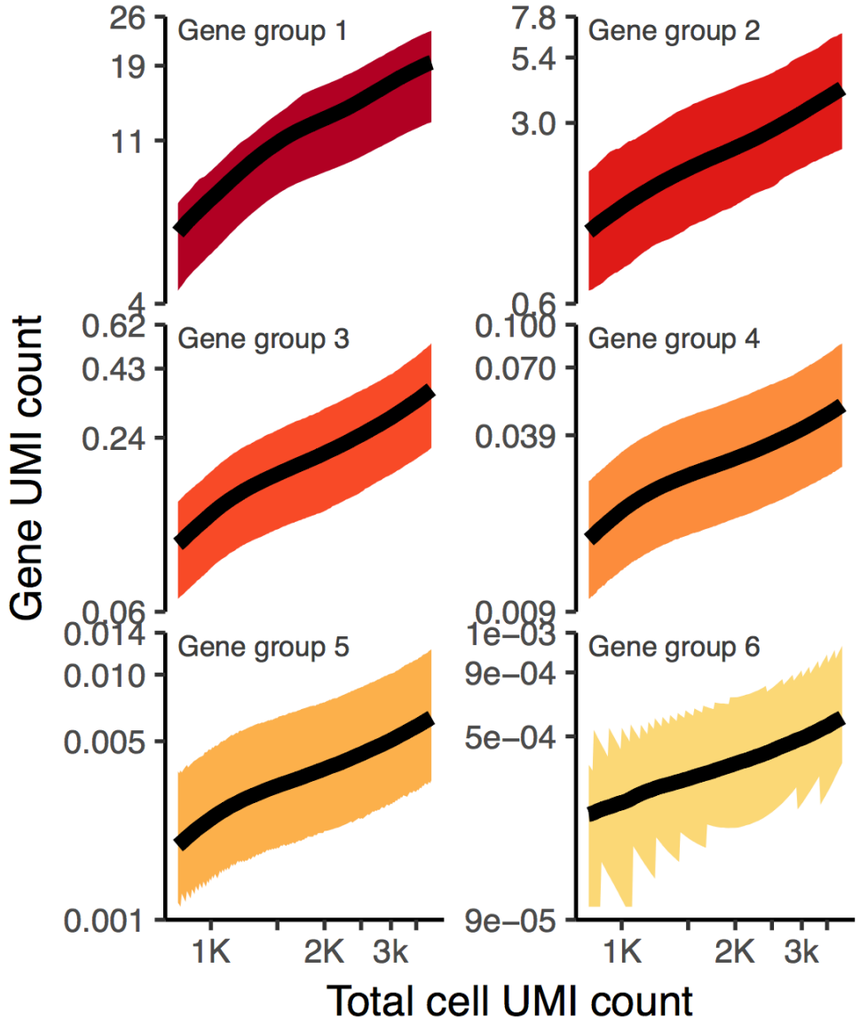
Sequencing depth of cell matters



Sequencing depth effect in 33k PBMC

Summarized loess fits for 16k genes





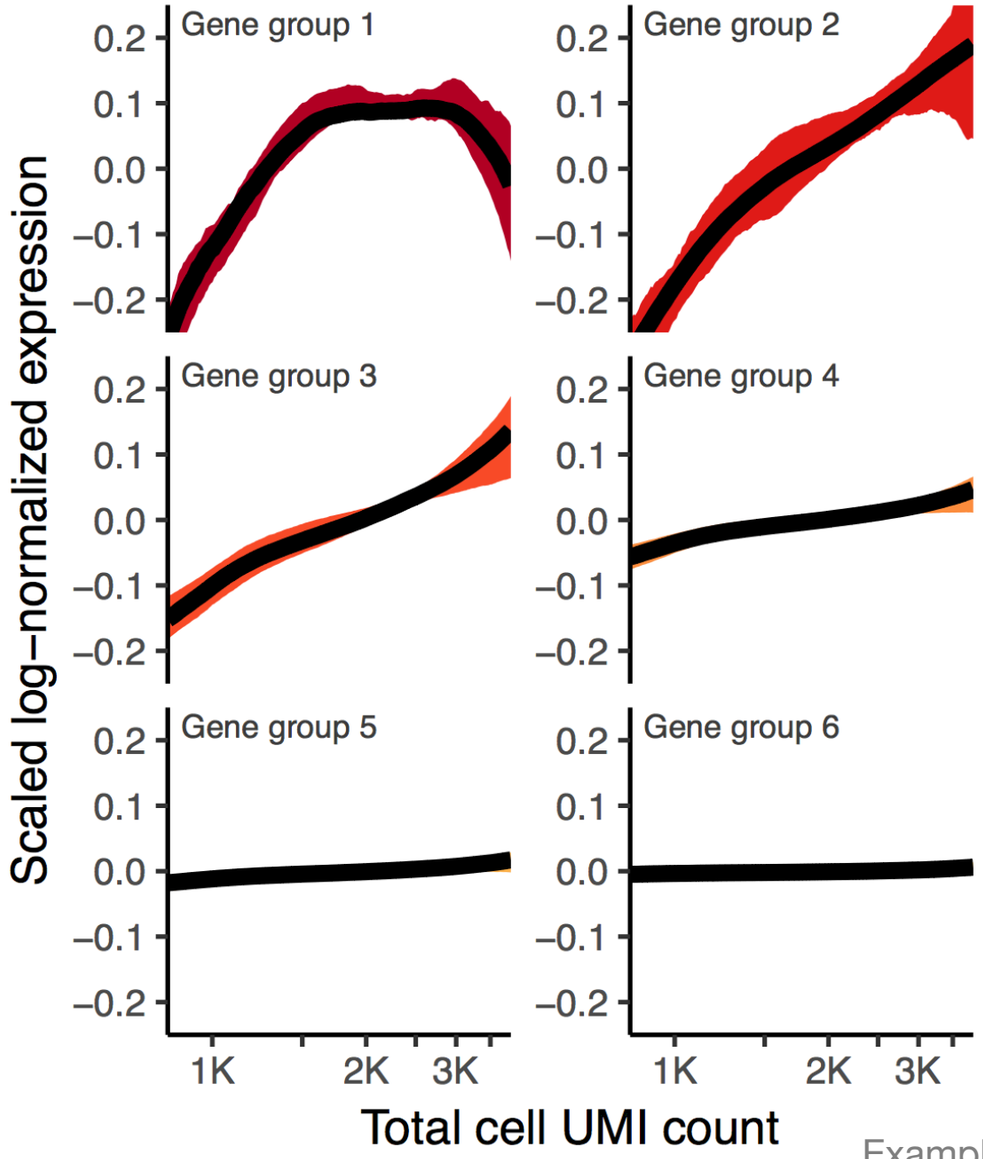
Example data: 33k PBMC

Log-normalization is not sufficient

Log-transformed counts per 10k

$$x_{ij} = \log\left(\frac{c_{ij}}{\sum_k c_{kj}} 10{,}000 + 1\right)$$

c_{ij} = UMI counts of gene i in cell j



Example data: 33k PBMC

Model the sequencing depth part of the signal

Can we learn a statistical model that describes technical variance in the data?

A good model should fit “boring” genes well

Biologically interesting genes should deviate from model

Deviation from model should drive the downstream analysis, e.g. clustering, marker discovery, cell alignment

Generalized linear model

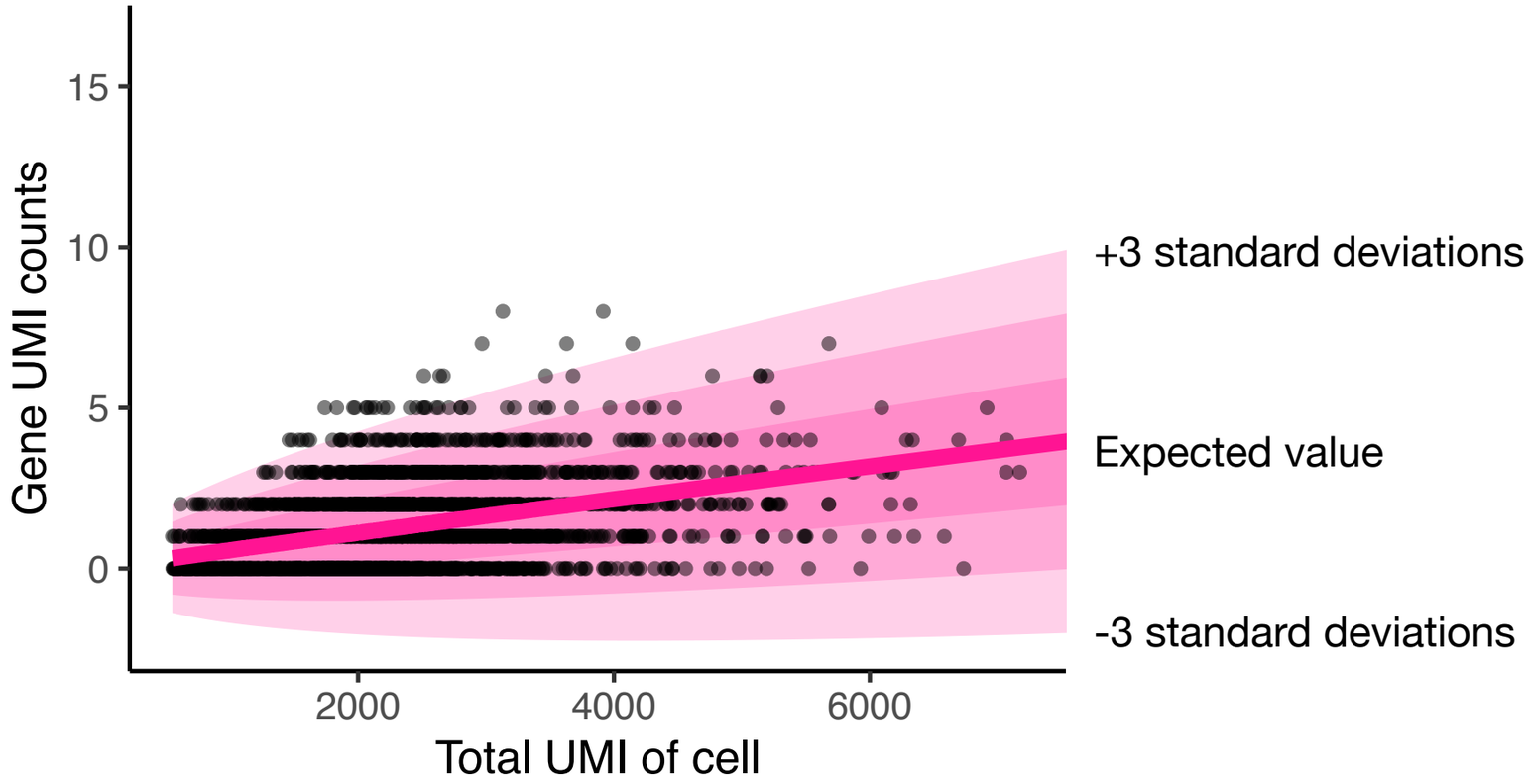
For a given gene, model log of expected value as function of cell attributes

$$\log(y) = \beta_0 + \beta_1 x$$

Several attributes and attribute combinations possible

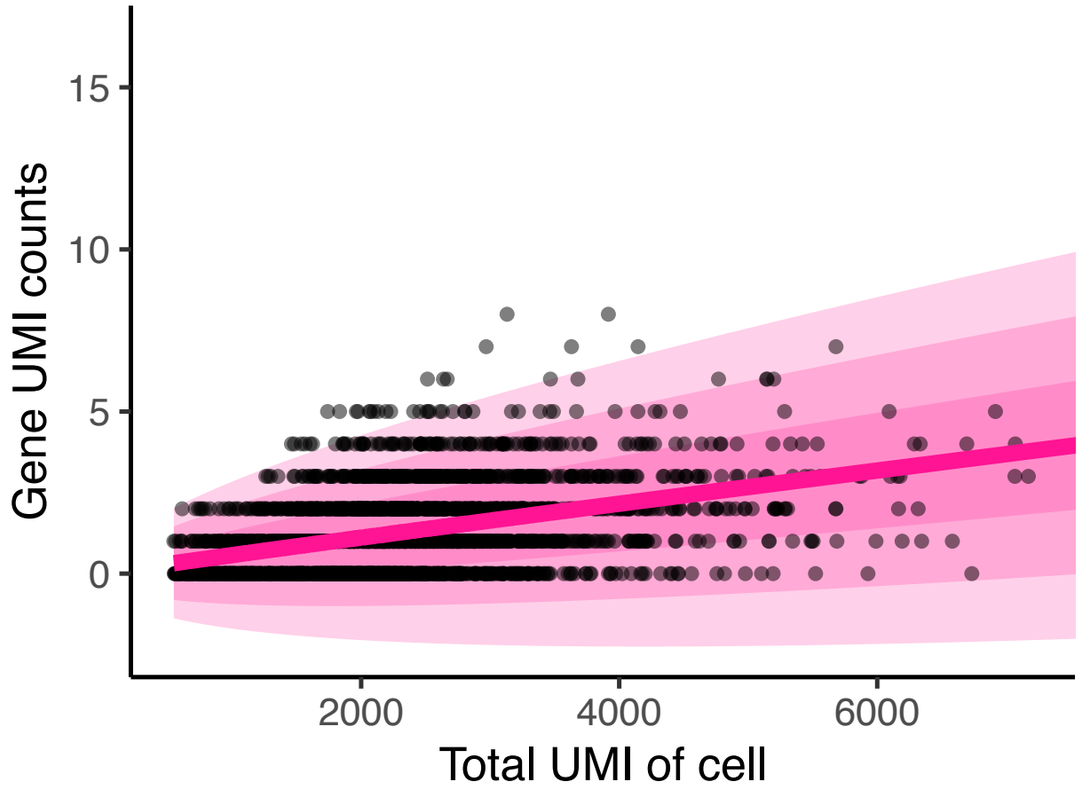
$$\log(E(\text{Gene UMI count})) = \text{Intercept} + \beta_1 \log(\text{Cell UMI count})$$

Simplest model for count data: Poisson



Example data: 2700 PBMC
Example gene: NBEAL1

Simplest model for count data: Poisson

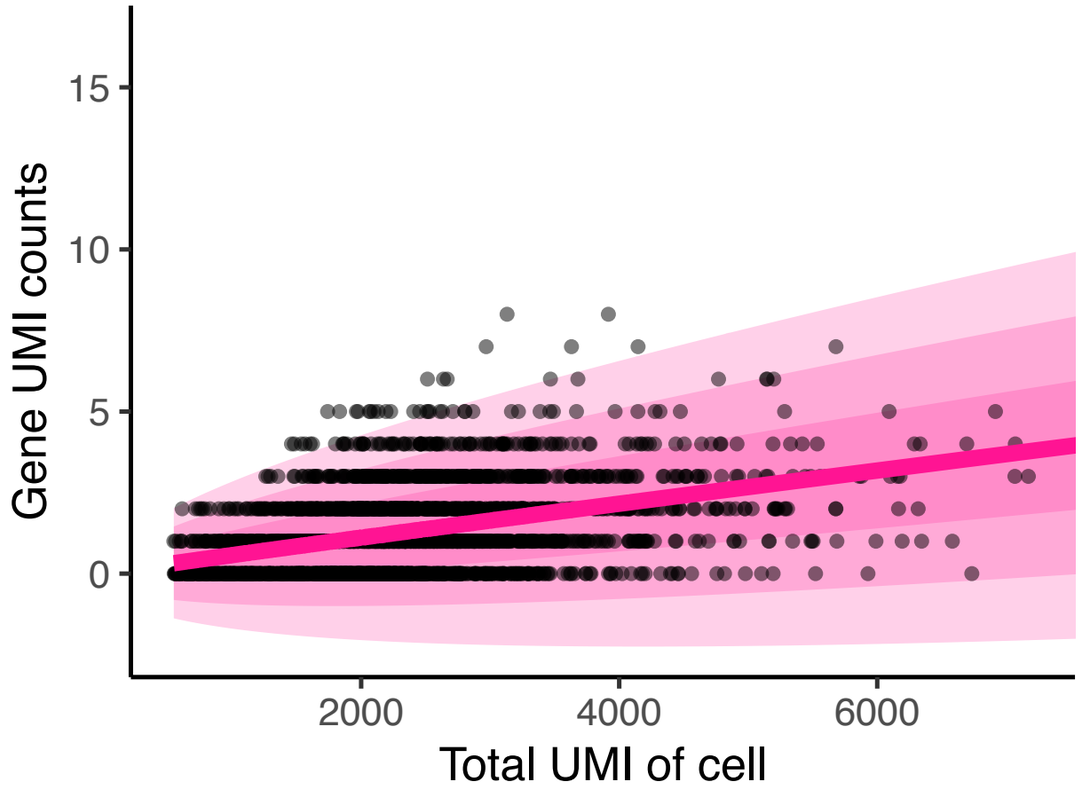


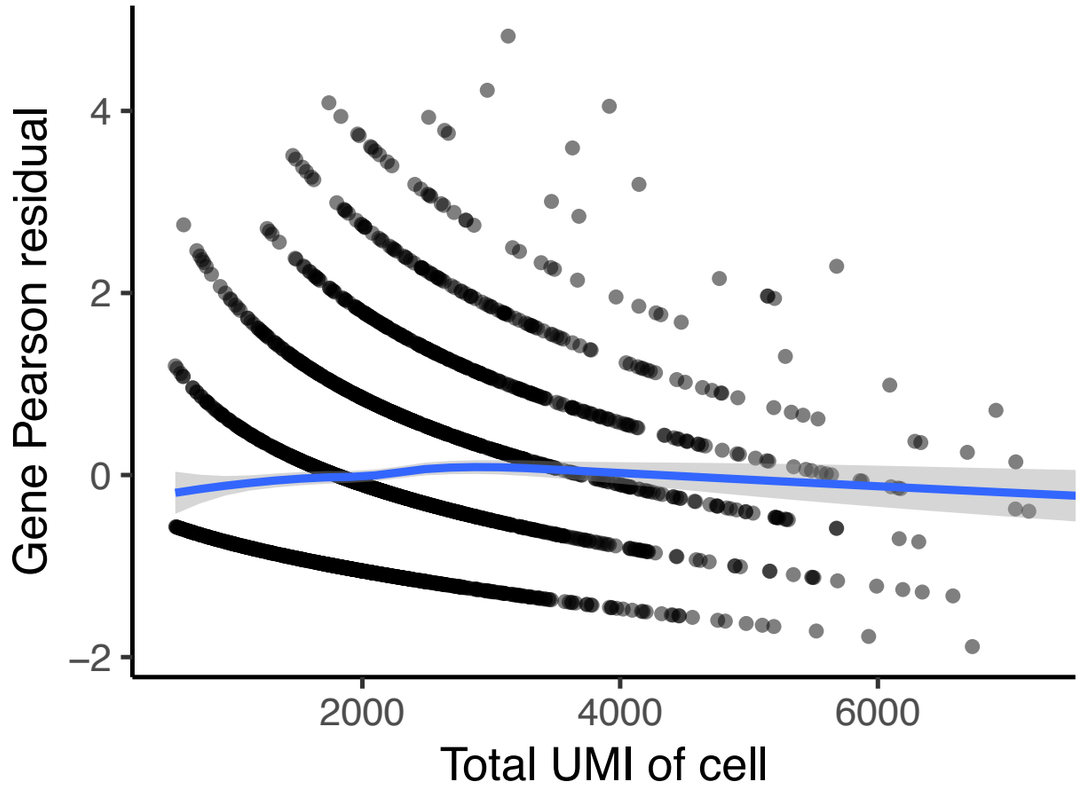
Residual accounts for baseline expression and sequencing depth differences between cells

Pearson residual also accounts for changes in expected variance (how many standard deviations above/below the expected value is the observed value)

Example data: 2700 PBMC
Example gene: NBEAL1

Good fit for this gene

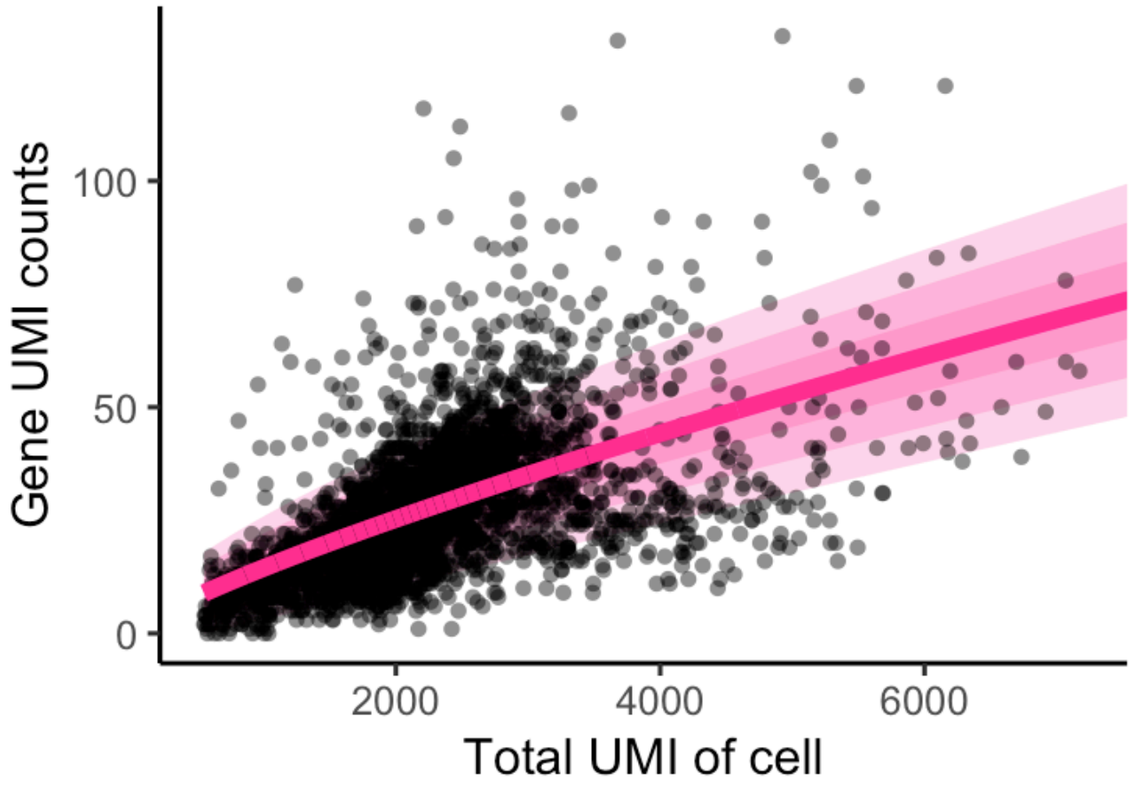


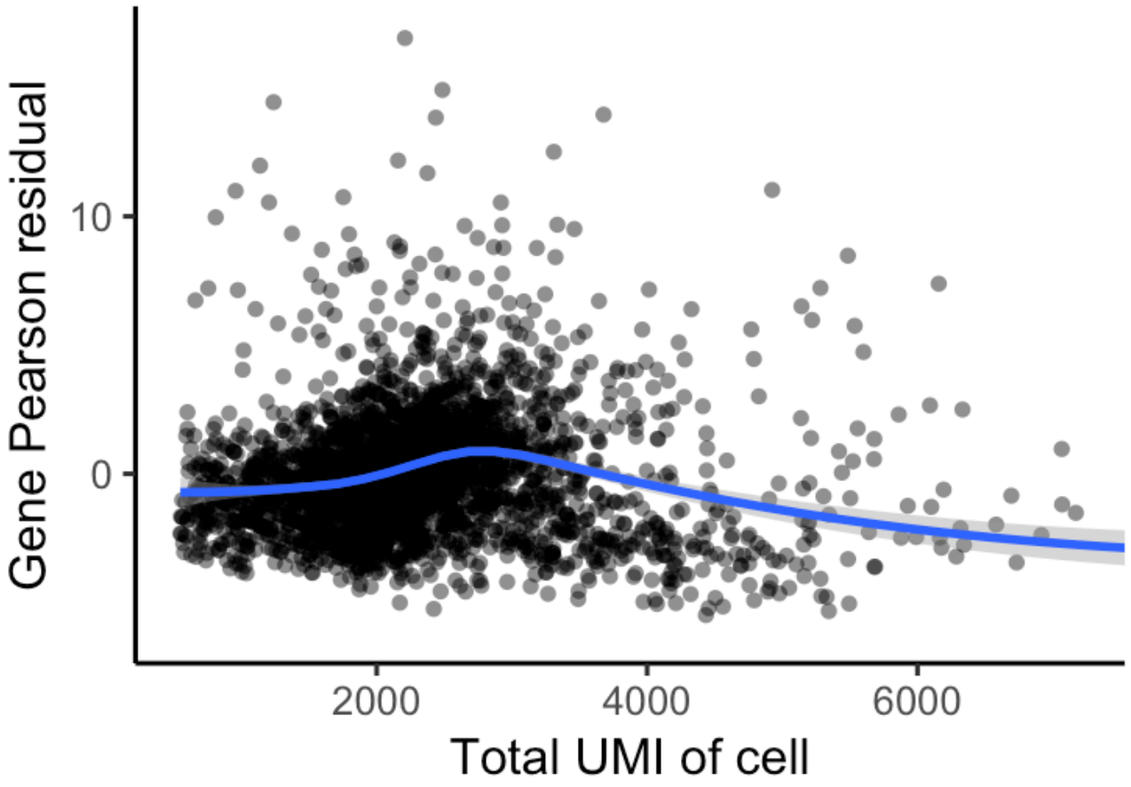


Residuals have mean 0, variance 1 -> good fit

Example data: 2700 PBMC
Example gene: NBEAL1

Bad fit for highly expressed gene



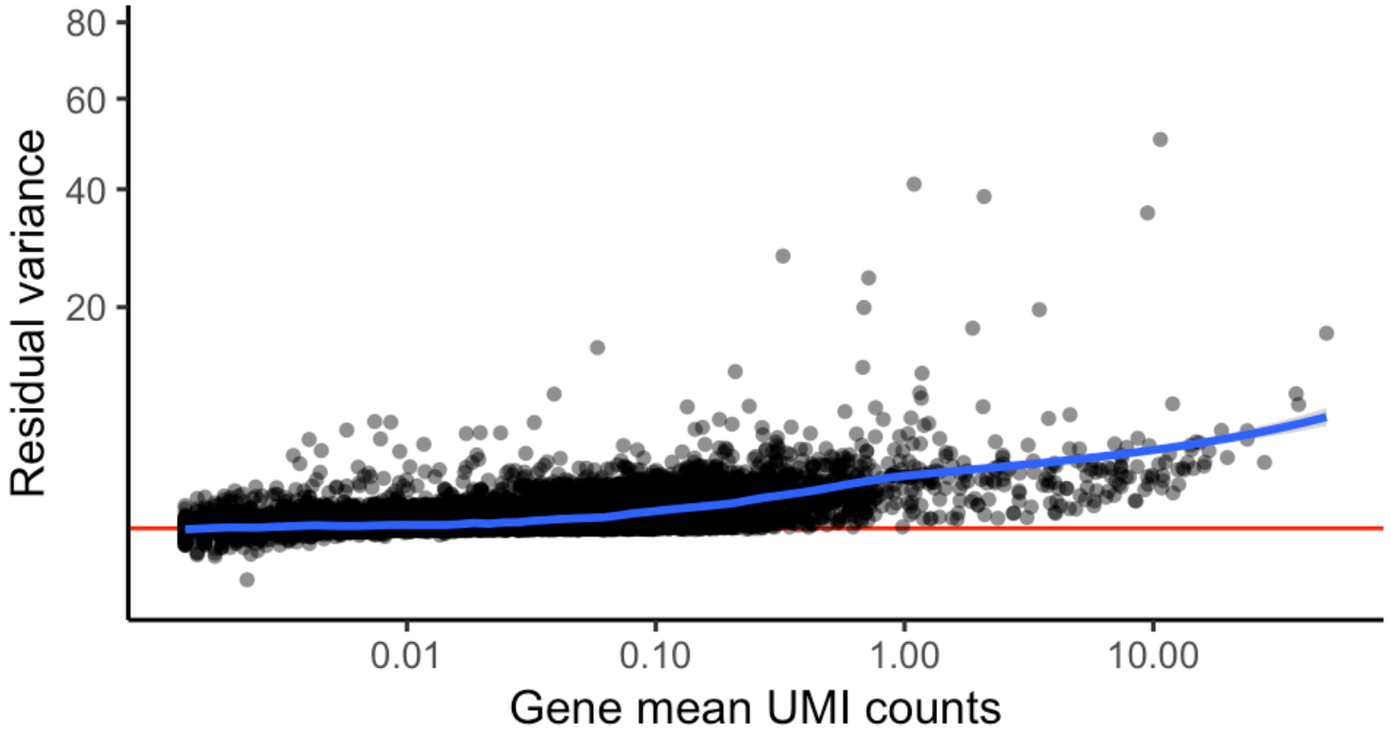


Residual variance is 6.4

Variance of highly expressed genes is higher than mean

Example data: 2700 PBMC
Example gene: RPL13

Poisson model not appropriate for all genes



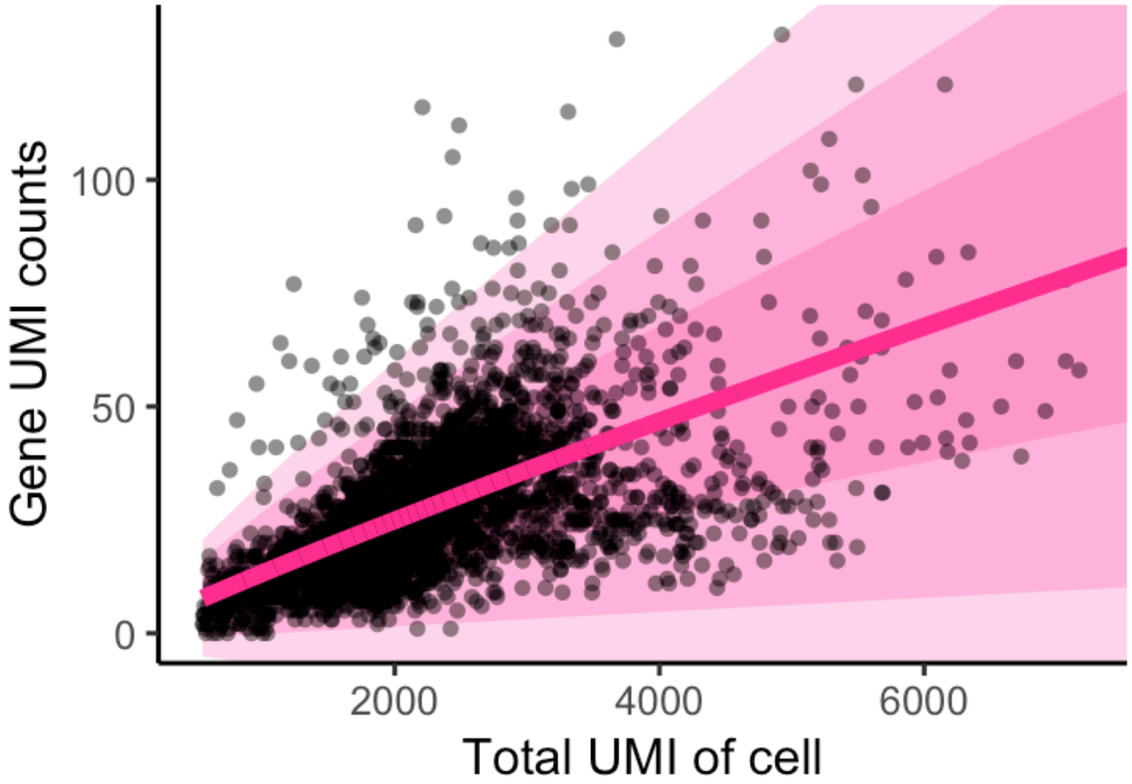
Top 10 variable genes

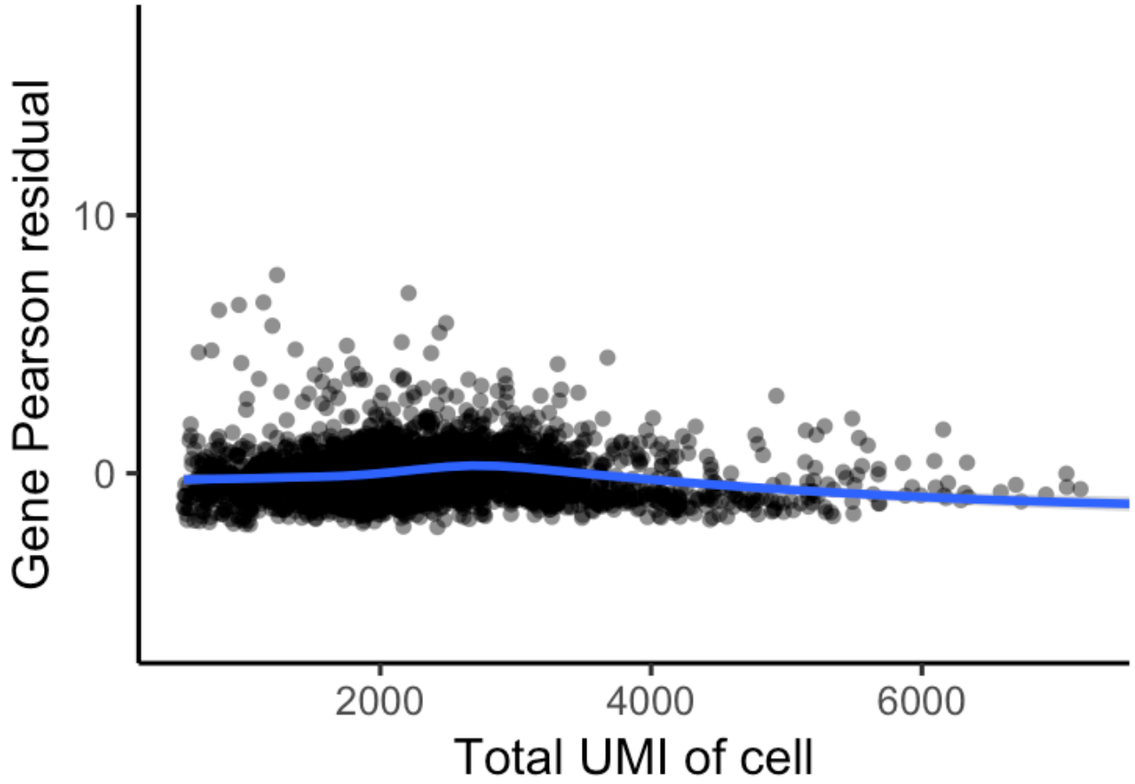
	Detection rate	Mean	Residual variance
FTL	0.987	27.667	50.505
S100A9	0.356	6.047	41.022
LYZ	0.604	10.247	38.562
FTH1	0.988	21.237	35.436
GNLY	0.179	1.574	27.832
S100A8	0.275	3.154	24.326
NKG7	0.302	2.654	19.950
CD74	0.842	8.525	19.653
HLA-DRA	0.556	6.099	17.165
MALAT1	1.000	59.883	16.511

Poisson model does not get rid of mean-variance relationship
Makes boring genes look interesting

Example data: 2700 PBMC

Allow variance > mean: Negative binomial

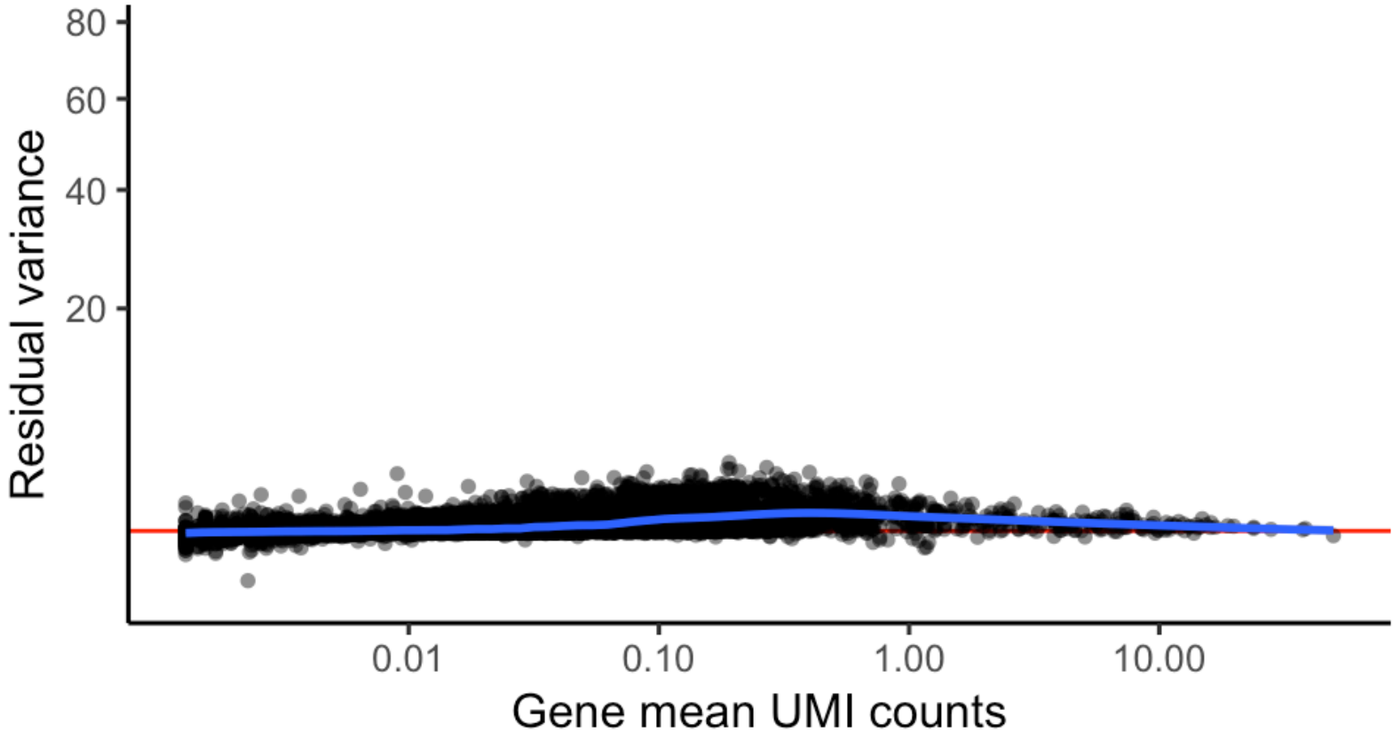




Additional theta parameter: Var = mu + mu^2 / theta
Residuals have mean 0, variance 1 -> good fit

Example data: 2700 PBMC
Example gene: RPL13

Negative binomial fits too well



Top 10 variable genes

	Detection rate	Mean	Residual variance
SIVA1	0.222	0.326	4.278
GIMAP5	0.282	0.470	3.968
NDUFA12	0.219	0.317	3.877
RALY	0.226	0.323	3.871
GZMB	0.121	0.731	3.735
HAGH	0.113	0.169	3.702
SF3B5	0.391	0.628	3.670
PRDX1	0.304	0.522	3.584
C1QB	0.006	0.028	3.583
CWC15	0.168	0.229	3.518

Overfitting

Makes interesting genes look boring

How to avoid overfitting

Problem:

One parameter per gene is not enough

Two parameters per gene are too many

Solution:

Pool information across genes

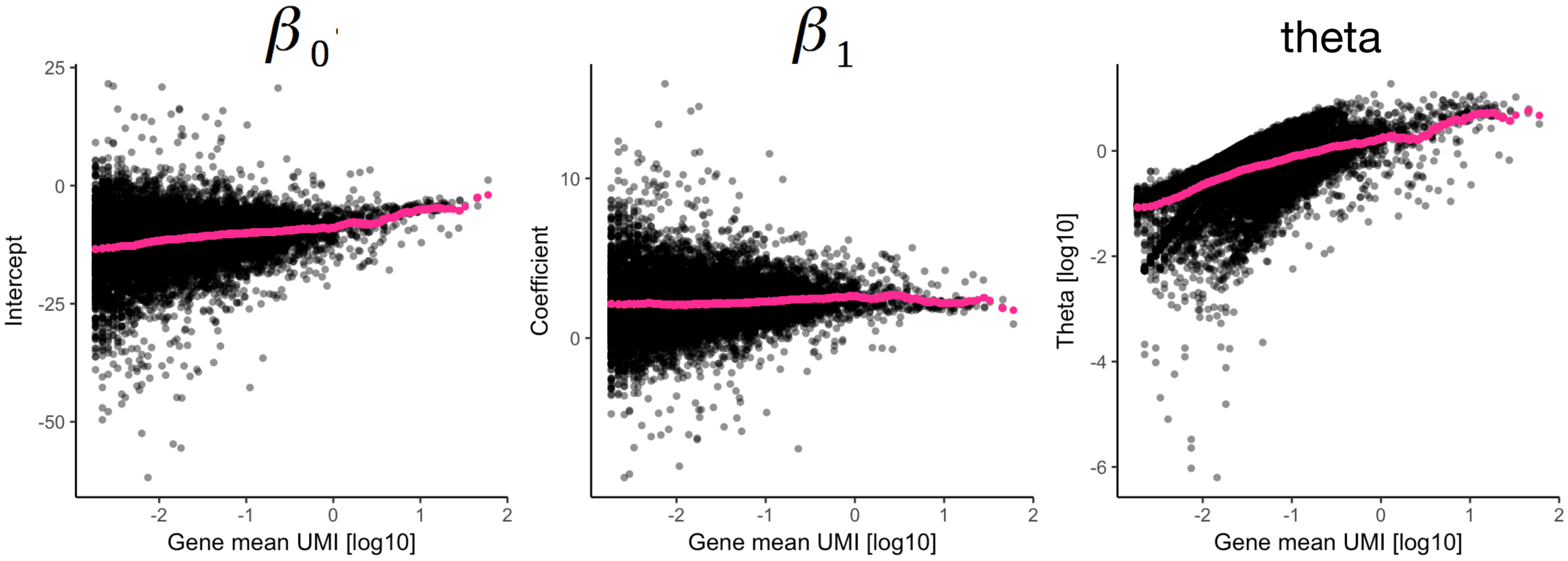
Idea:

Genes expressed at a similar level should have similar models

Assumption:

Relationship between gene mean and model parameter is systematic

Pool information across genes

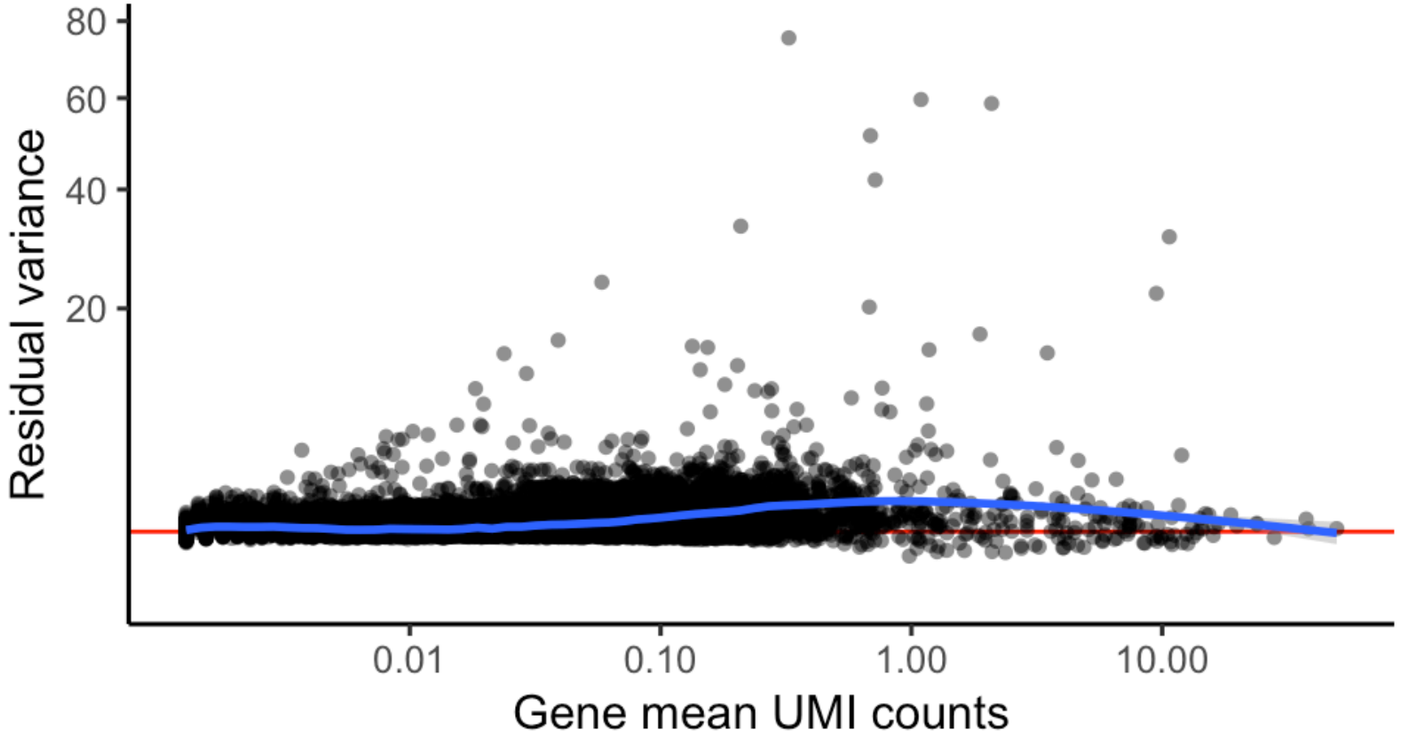


Per-gene parameter estimate

Regularized parameter

$$\log(y) = \beta_0 + \beta_1 x$$

Regularized parameters capture overall trends



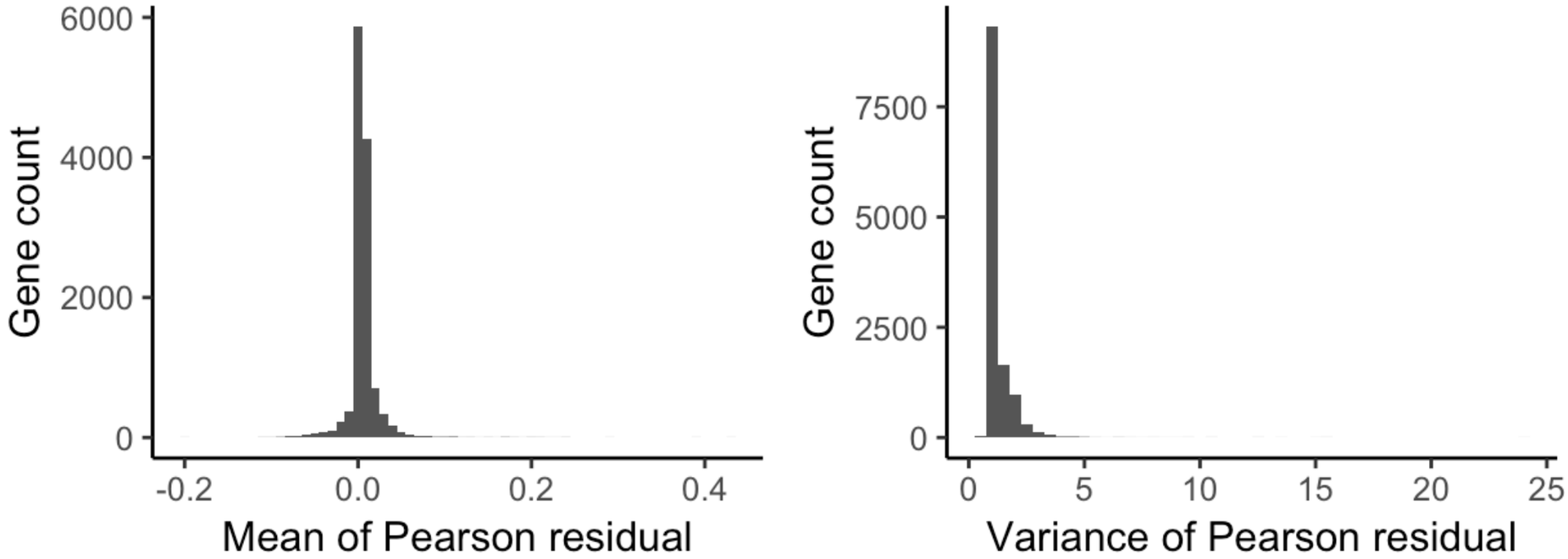
Top 10 variable genes

	Detection rate	Mean	Residual variance
GNLY	0.179	0.325	75.379
S100A9	0.356	1.092	59.645
LYZ	0.604	2.088	58.702
NKG7	0.302	0.687	51.290
S100A8	0.275	0.717	41.883
GZMB	0.121	0.209	33.095
FTL	0.987	10.680	31.216
IGLL5	0.041	0.058	23.817
FTH1	0.988	9.480	22.150
CCL5	0.319	0.680	20.205

No bias

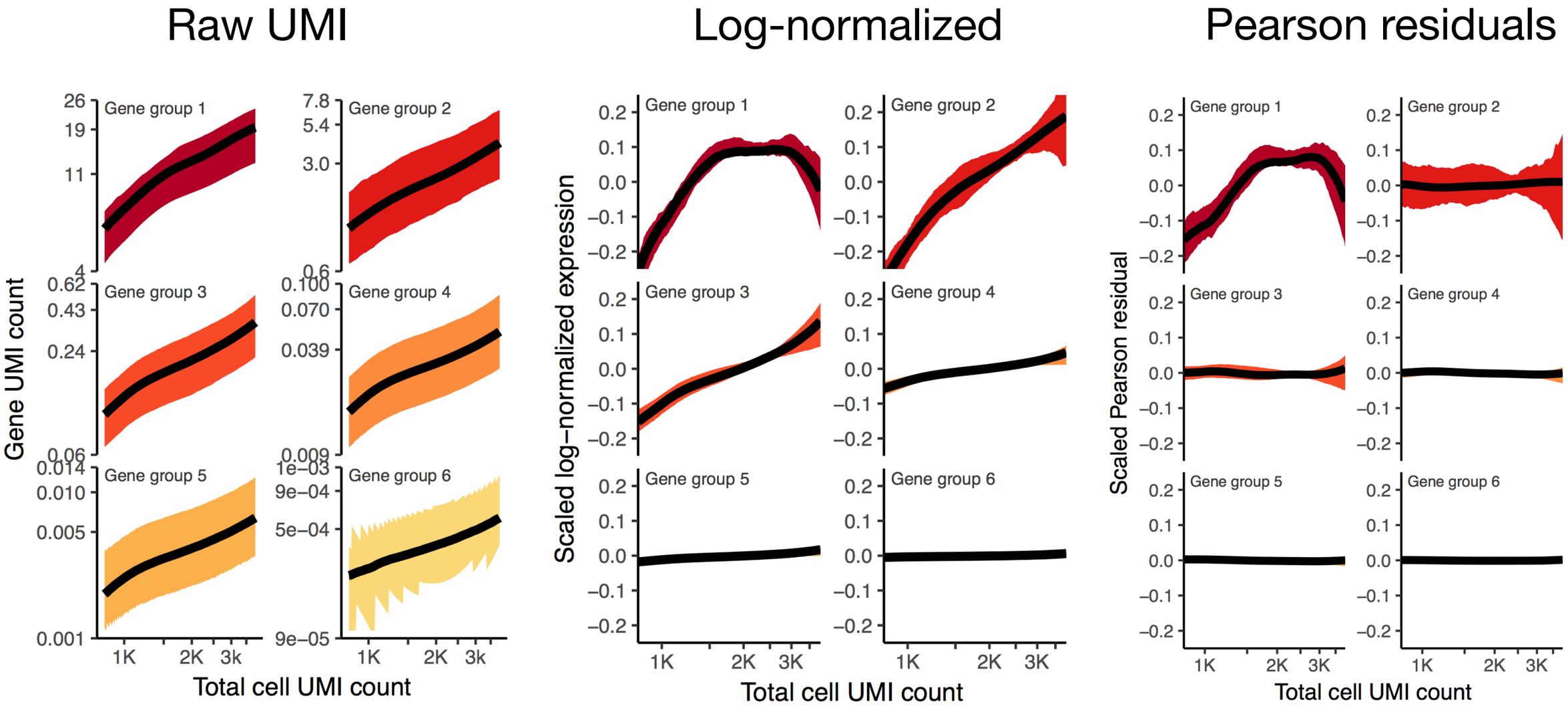
Promising gene list

Regularized parameters capture overall trends



Vast majority of residuals have mean zero, variance one
Overall good fit

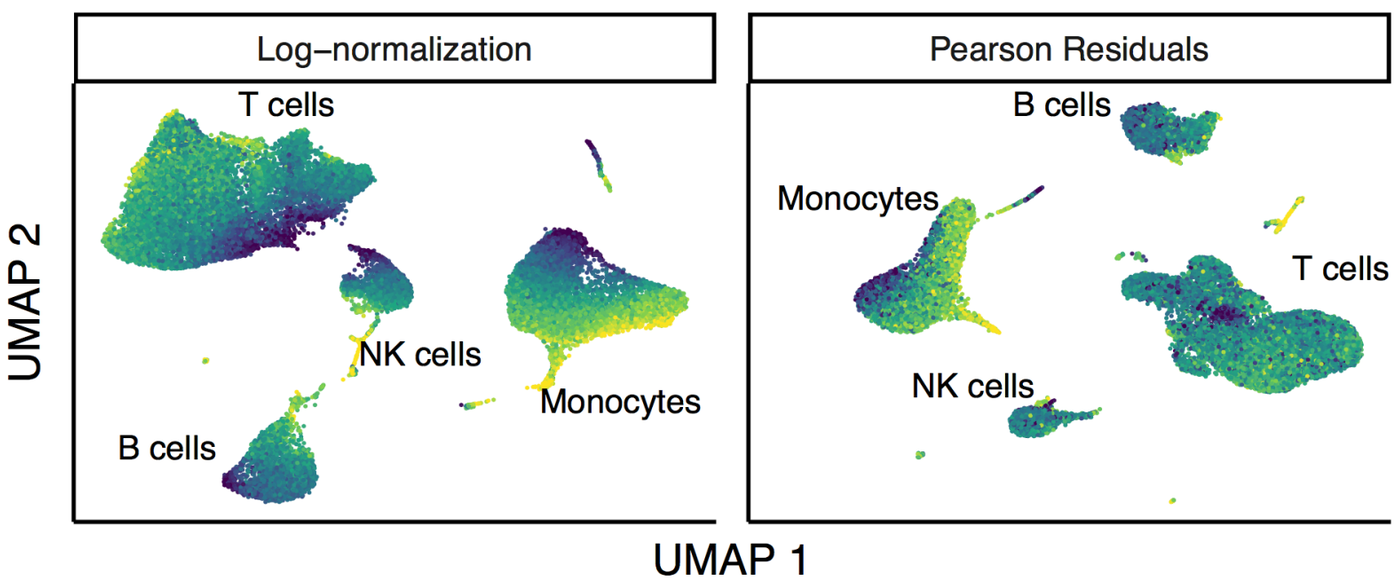
Residuals show little correlation with total cell UMI

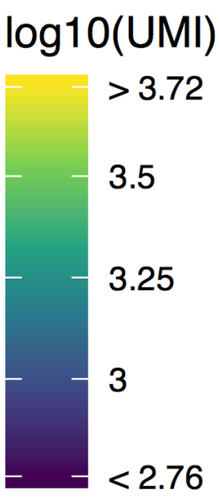


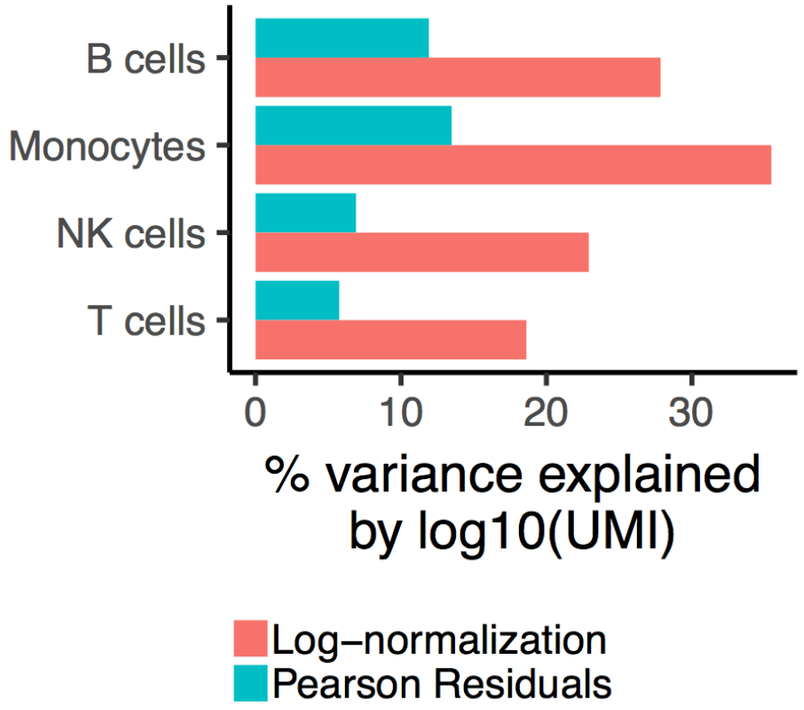
Example data: 33k PBMC

Less variance explained by cell UMI in residuals

Run PCA on Pearson residuals; keep 25 PCs







Example data: 33k PBMC

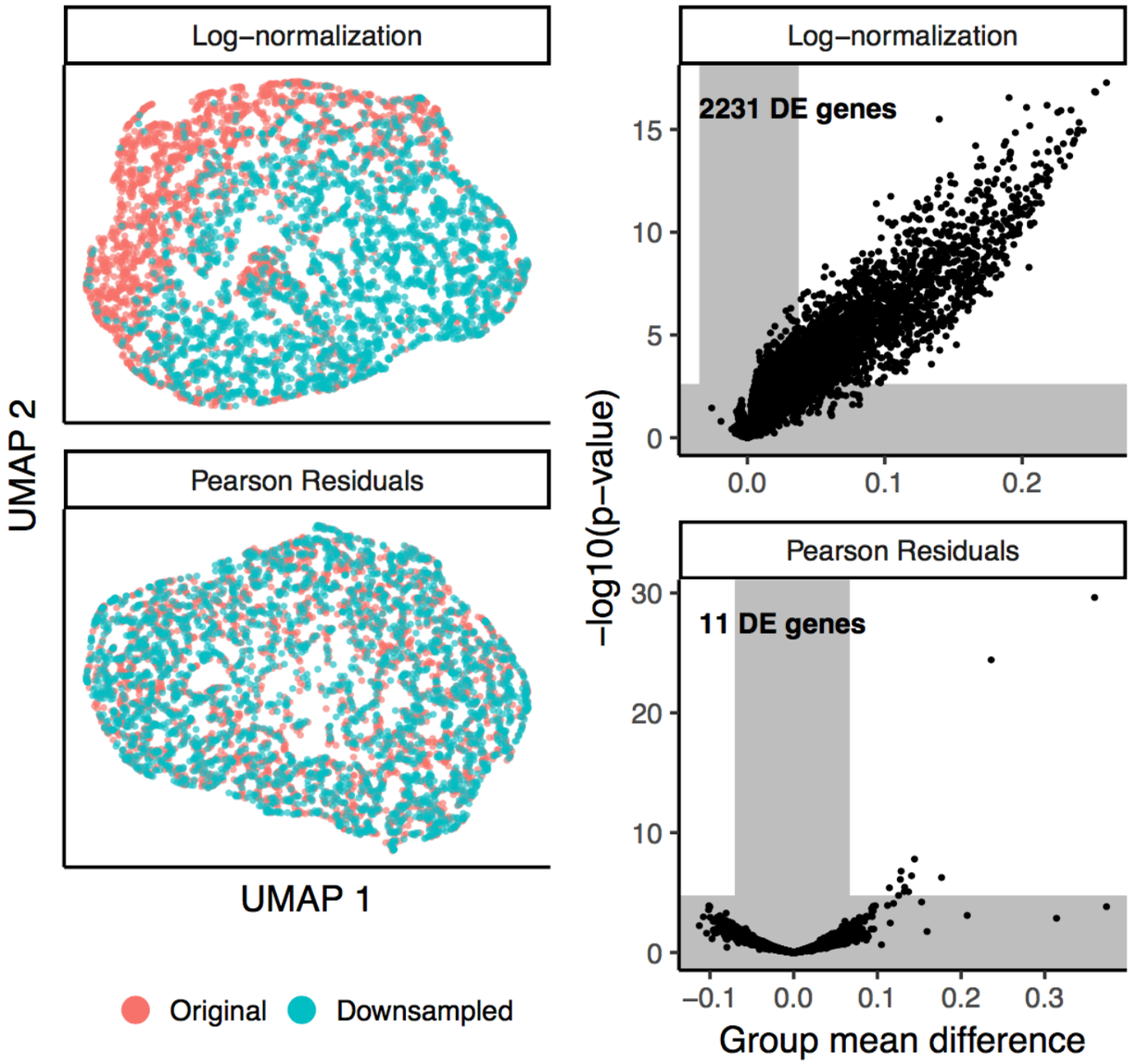
Downsampling does not introduce heterogeneity

5.5k CD14+ Monocytes

50% of the cells are downsampled to 50% total UMI count

The two groups are still biologically equivalent

Test for differential expression after normalization

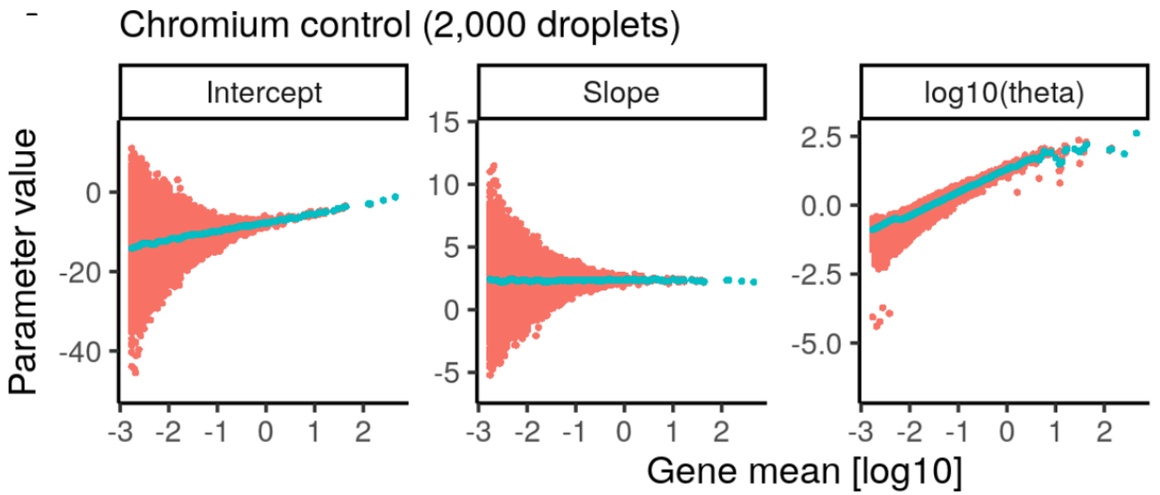


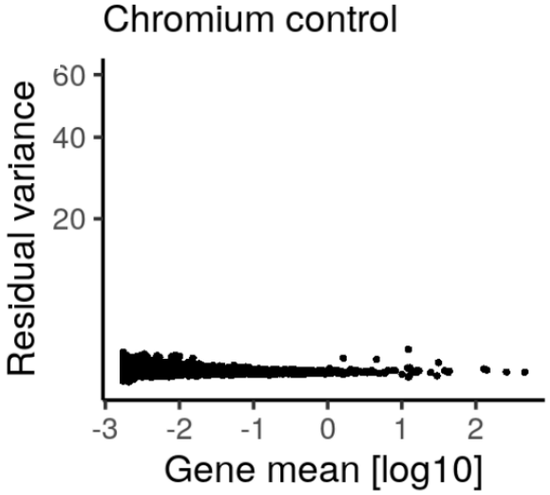
No variable genes in negative control data

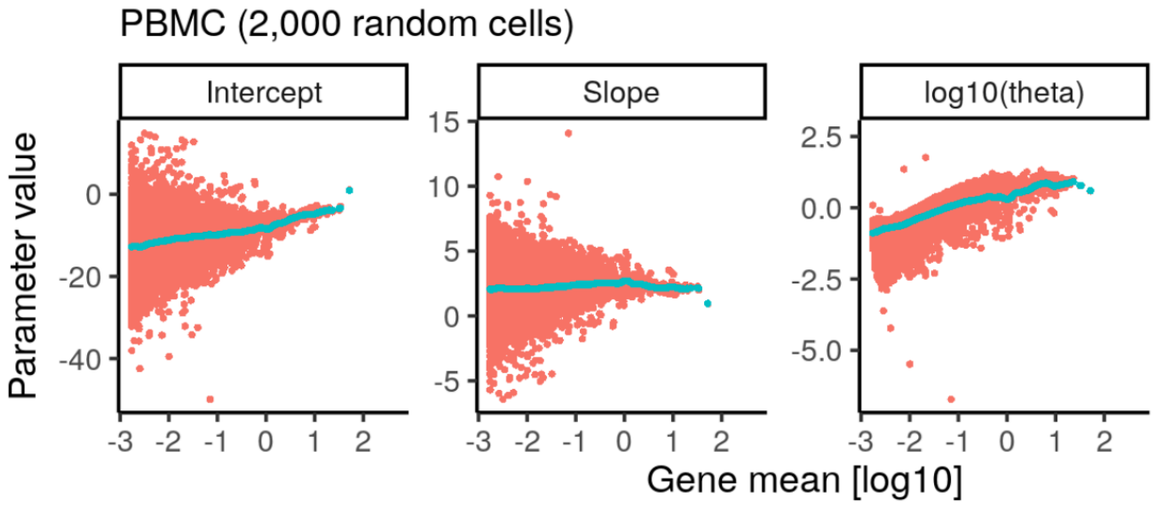
RNA in solution -> microfluidic system

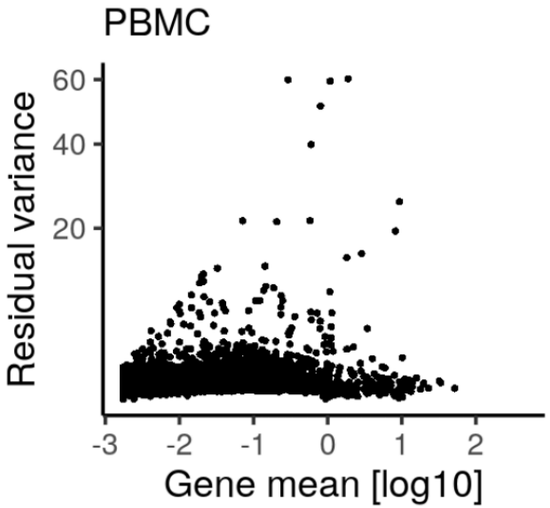
No biological heterogeneity expected

PBMC for comparison









Chromium Control (10X v2) [Svensson et al., 2017], ArrayExpress accession E-MTAB-5480, UMI count matrix available on `https://figshare.com/articles/svensson_chromium_control_h5ad/7860092`, sample 1

Summary

1. How do you define normalization? Why is it important?

2. How do you normalize?

3. How do you demonstrate success? How do you know if you've normalized properly?

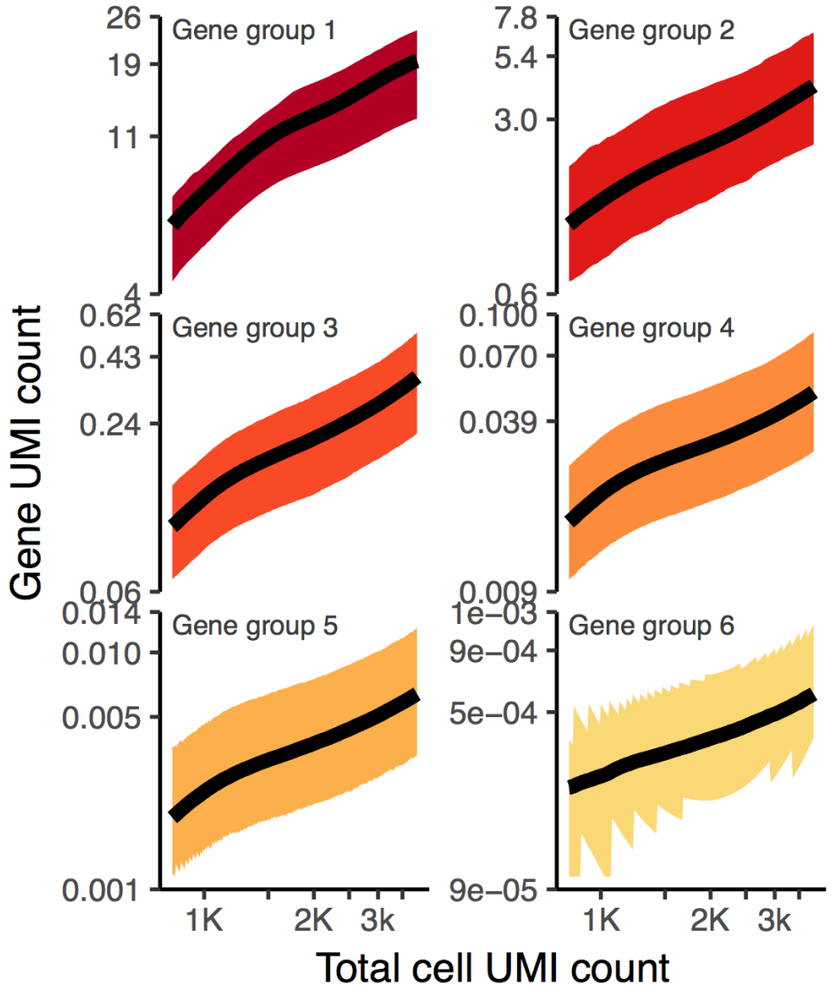
4. Where does your method break? Or where do you see challenges with your method or others?

5. What is your suggestion for how we spend the second day of the workshop to collaborate on computational problems? Are there particular tasks or challenges that you think would be good to address?

How do you define normalization? Why is it important?

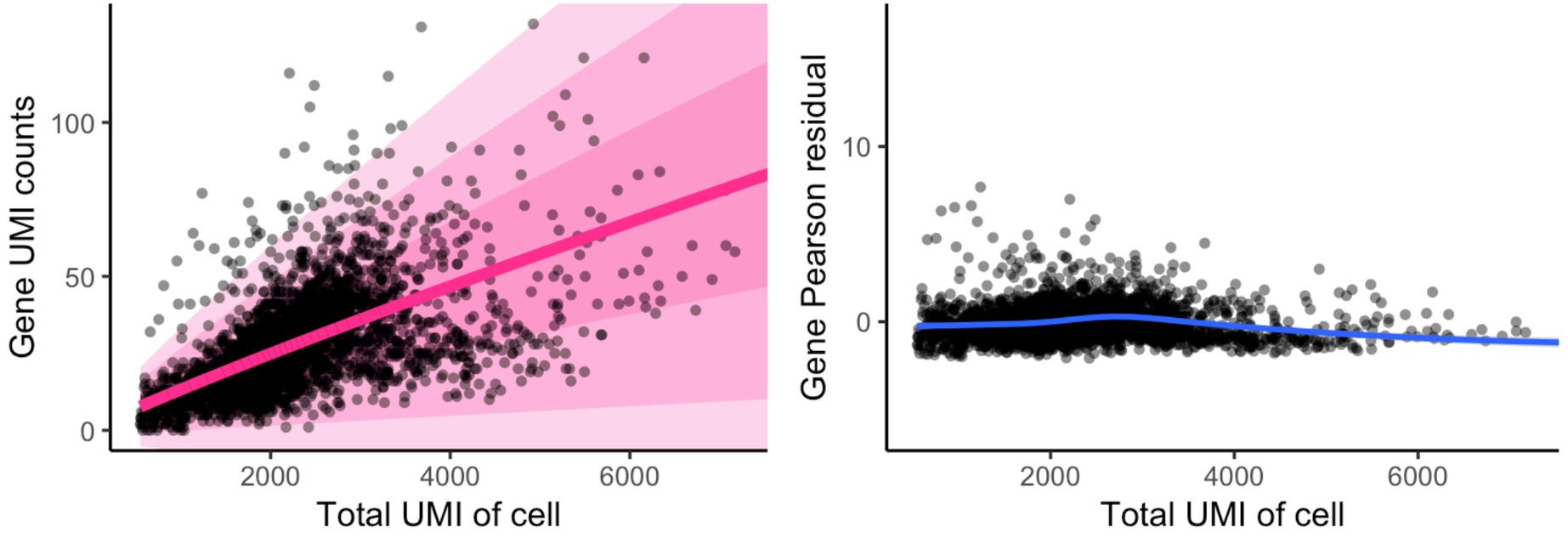
Control for the variability sequencing depth contributes to the observed UMI counts

Goal: focus on biology / cleaner data



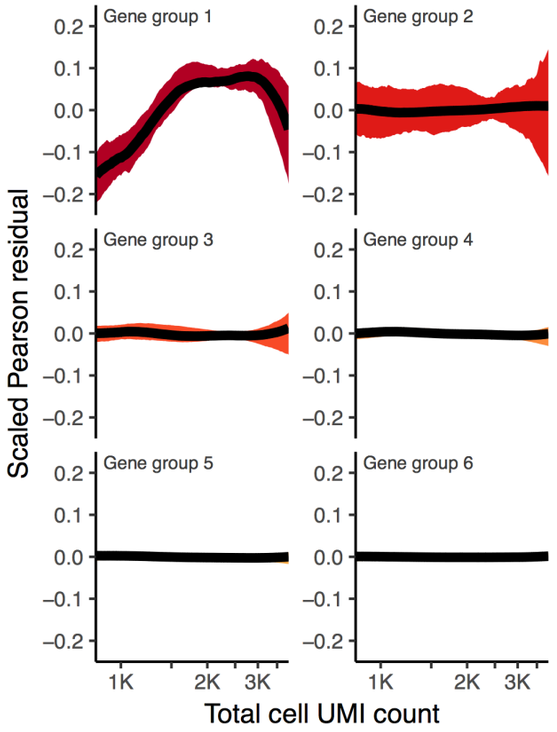
How do you normalize?

Regress out sequencing depth: Generalized linear model with regularized negative binomial error distribution

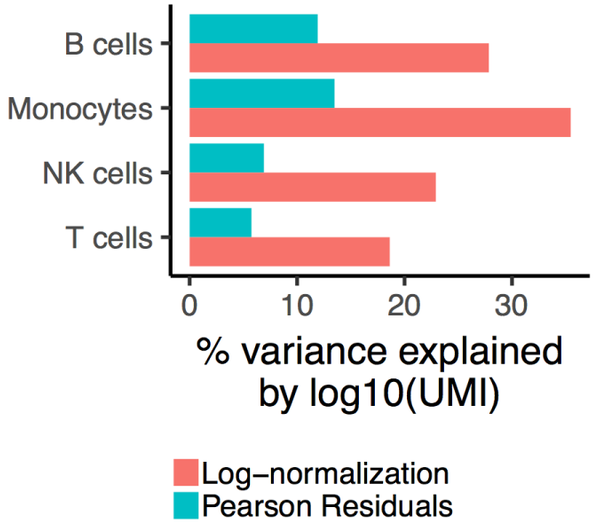


How do you demonstrate success? How do you know if you've normalized properly?

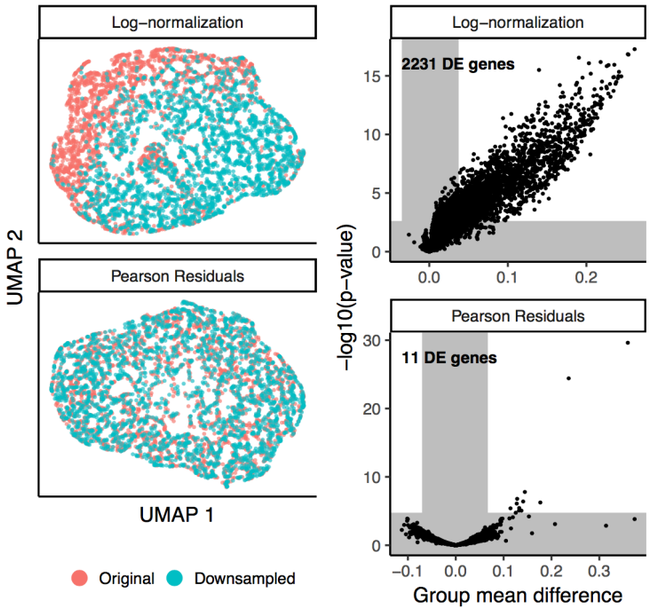
Correlation with total cell UMI



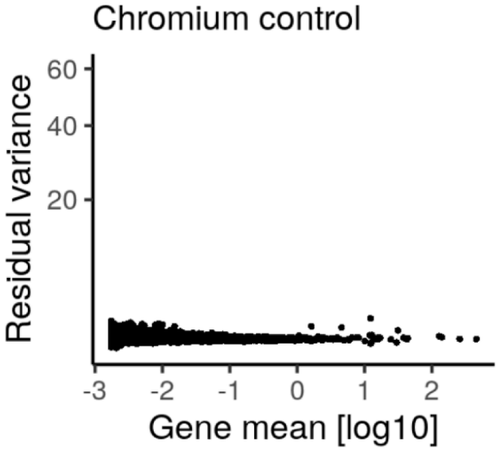
Variance explained by total cell UMI



Heterogeneity after down-sampling



Pool-and-split control data



Where does your method break? Or where do you see challenges with your method or others?

Very highly expressed genes don’t quite fit linear model

How to pool of genes when no other genes with similar mean are present

When biology is correlated with sequencing depth we will reduce signal of interest

What is your suggestion for how we spend the second day?

- Find a consensus: Benchmarking methods
  - How do we define success?
  - What metrics could we use to quantify success?
  - What datasets could we use
    - Real data
    - Simulations: what kind?
- Create / discuss interfaces to the different methods (allow R to call python methods and vice versa)

Acknowledgements

Satija Lab
Rahul Satija
Andrew Butler
Paul Hoffman

NEW YORK GENOME CENTER®

DFG Deutsche Forschungsgemeinschaft

R package:
github.com/ChristophH/sctransform