Don't Normalize

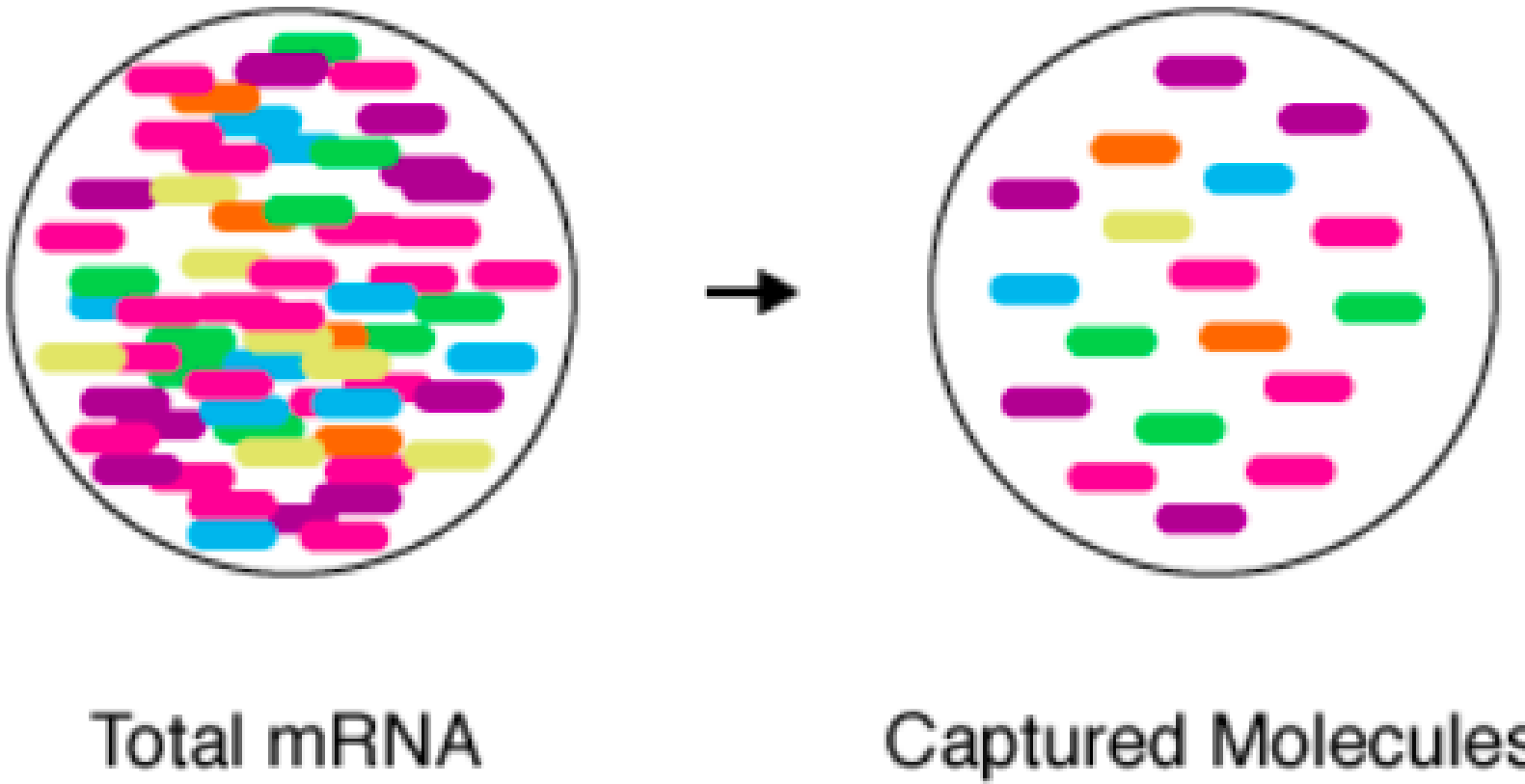
The GLM-PCA approach to normalization

Will Townes

Department of Computer Science, Princeton University

19 November 2019

RNA-seq measures relative abundance



Batson et al 2019 Biorxiv

Normalization as estimation of relative abundance

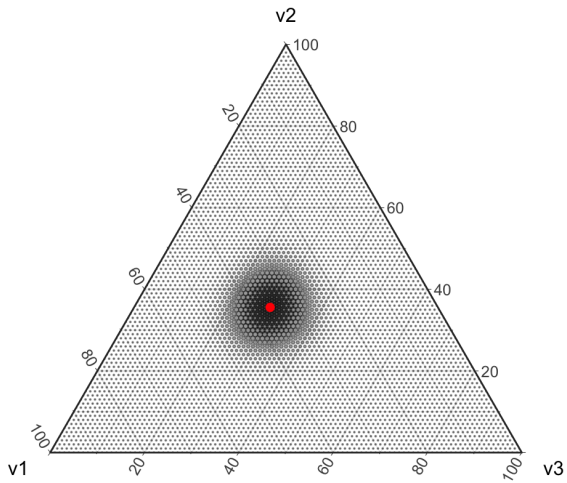
$$y_{ij} \sim Multinomial(n_i, \pi_{ij})$$
$$\hat{\pi}_{ij} = \frac{y_{ij}}{n_i}$$

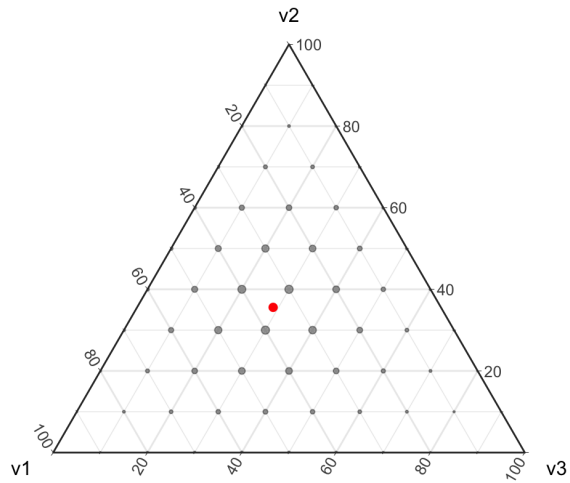
$$\tilde{\pi}_{ij} = \frac{y_{ij} + \alpha_i}{n_i + J\alpha_i}$$
$$\log_2(1 + CPM) = \log_2(\tilde{\pi}_{ij}) + C$$

Poisson approximation when n_i large, π_{ij} small.

Problems with normalization

Small counts limit MLE accuracy.

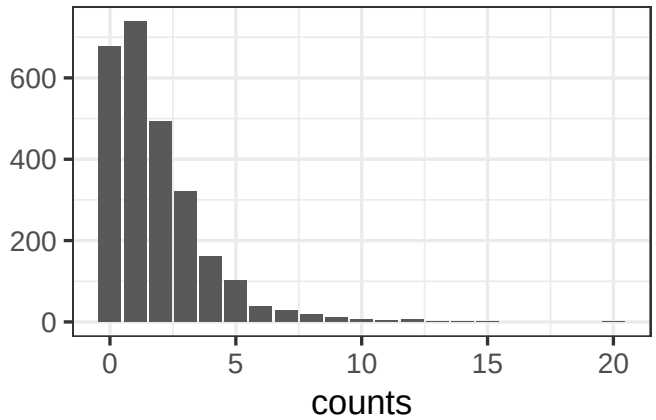


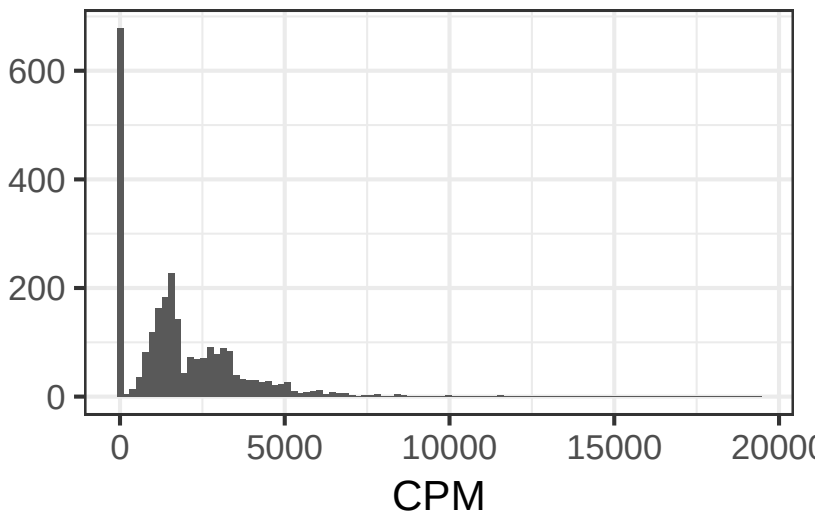


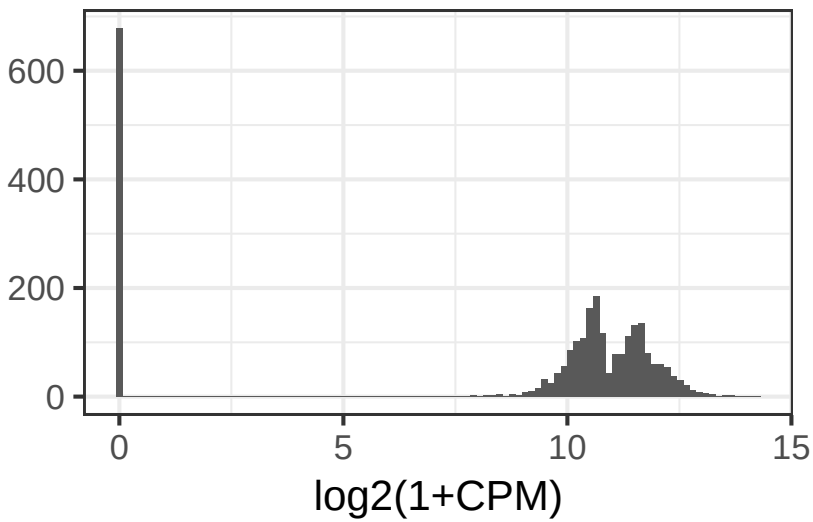
Justin Silverman: `http://www.statsathome.com/2017/09/14/visualizing-the-multinomial-in-the-simplex/`

Problems with normalization

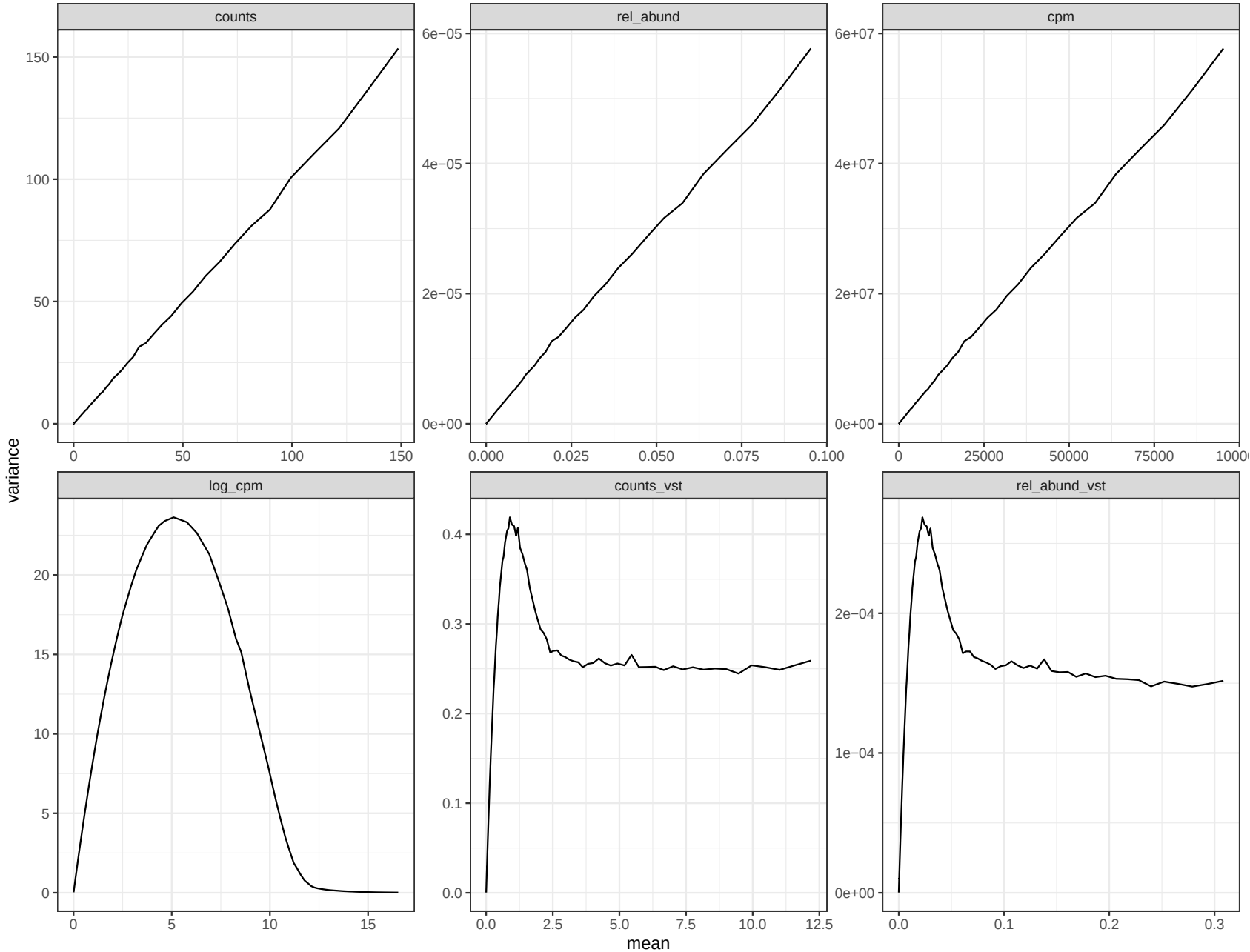
Artificial zero inflation from log-transform.







Variance stabilizing transformations



GLM-PCA: avoid normalization by using models

$$y_{ij} \sim Poi(n_i \pi_{ij}) \approx Mult(n_i,\ \pi_{ij})$$
$$\pi_{ij} = f_j(u_i) = \exp\left\{v_j' u_i\right\}$$

- Improve estimation of π_{ij} by sharing info across cells
- Variance stabilization not necessary with explicit noise model
- ZINB-WAVE, SCVI, linear decoded VAE also doing this

GLM-PCA failure modes

- Nonconvex optimization problem
- Numerical divergences
- Local optima
- Slow computation
- Too many factors?

Maybe normalization is not so bad

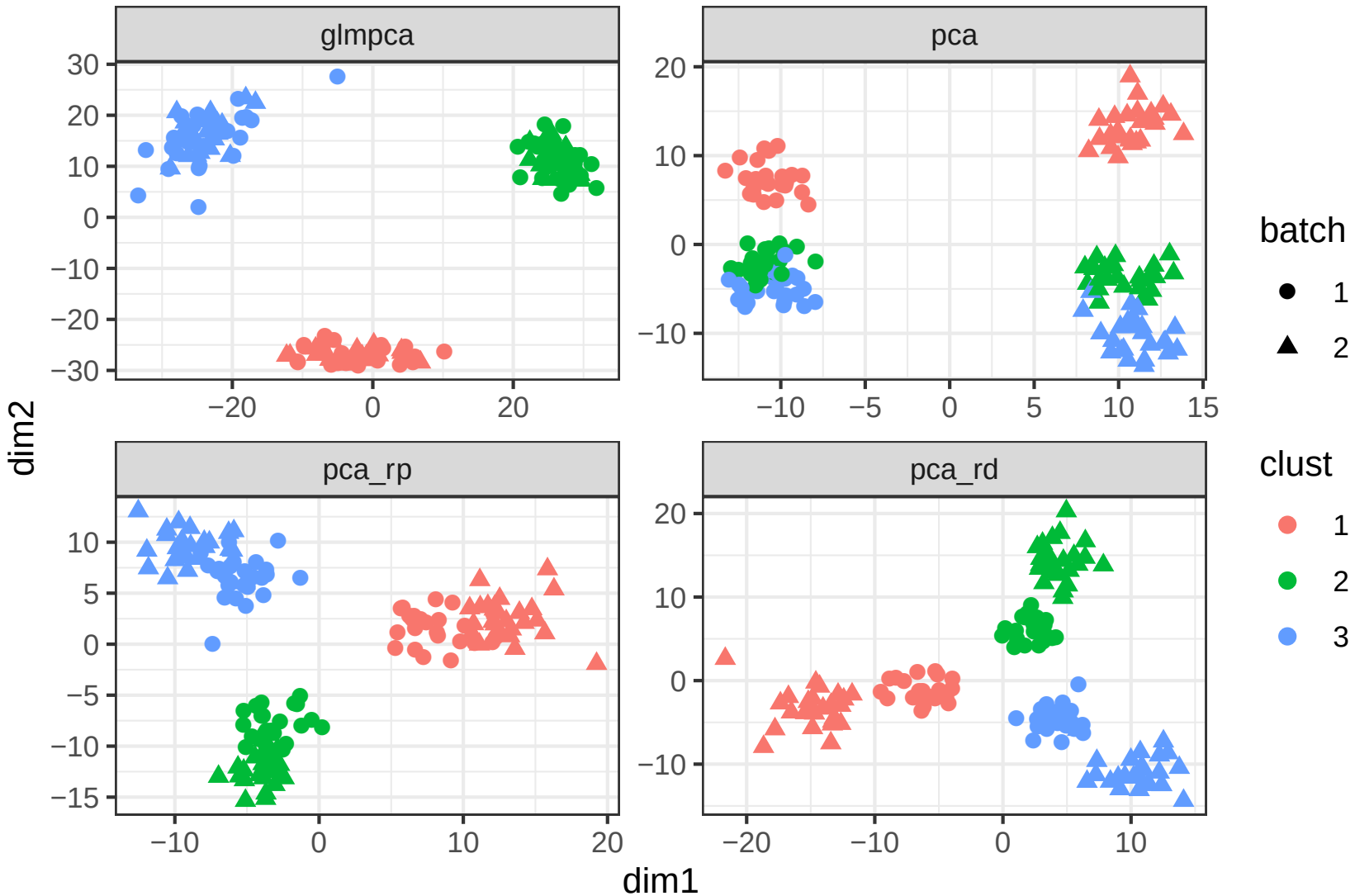
- Linear-Gaussian models (PCA) fast, convenient, interpretable
- PCA requires normally distributed errors
- Transform data to match Gaussian assumption
- Idea: GLM residuals asymptotically normal
- Fit multinomial null model and use deviance residuals:

$$D_j = 2 \sum_i \left[y_{ij} \log \frac{y_{ij}}{n_i \hat{\pi}_j} + (n_i - y_{ij}) \log \frac{(n_i - y_{ij})}{n_i (1 - \hat{\pi}_j)} \right]$$

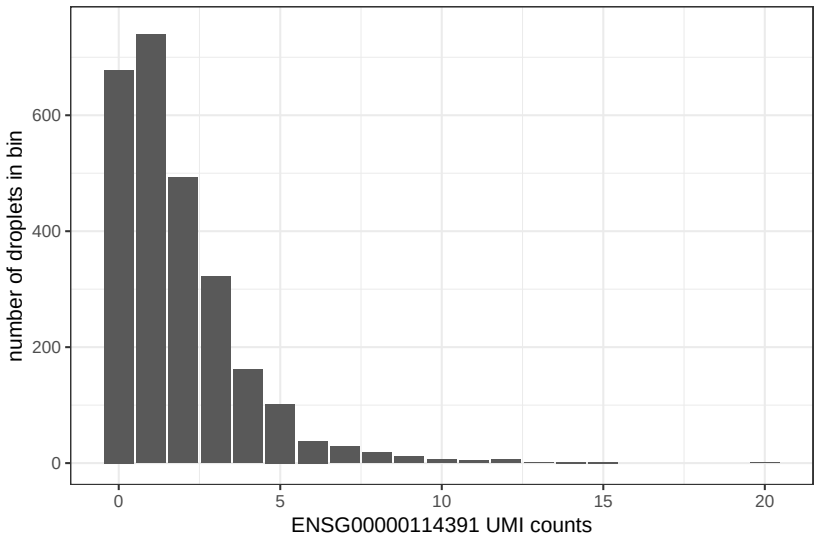
(or Pearson residuals):

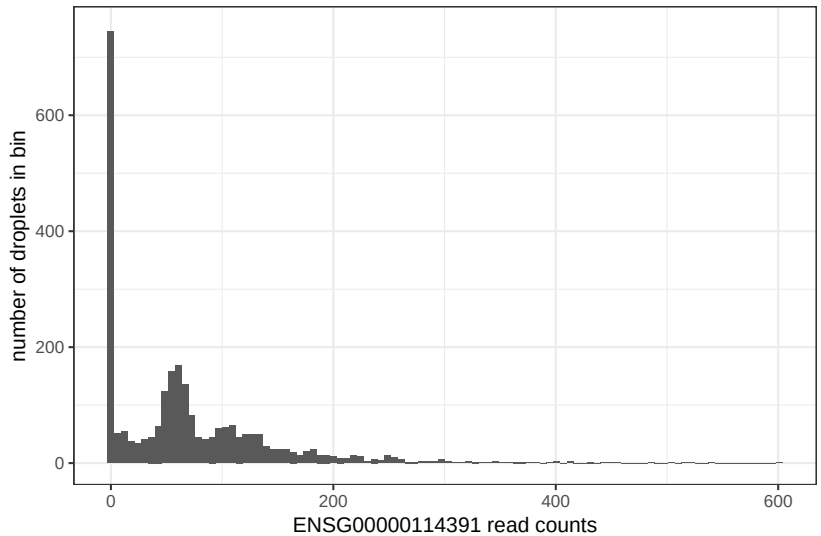
$$\frac{y_{ij} - n_i \hat{\pi}_{ij}}{\sqrt{n_i \hat{\pi}_{ij} (1 - \hat{\pi}_{ij})}}$$

Normalization via null residuals



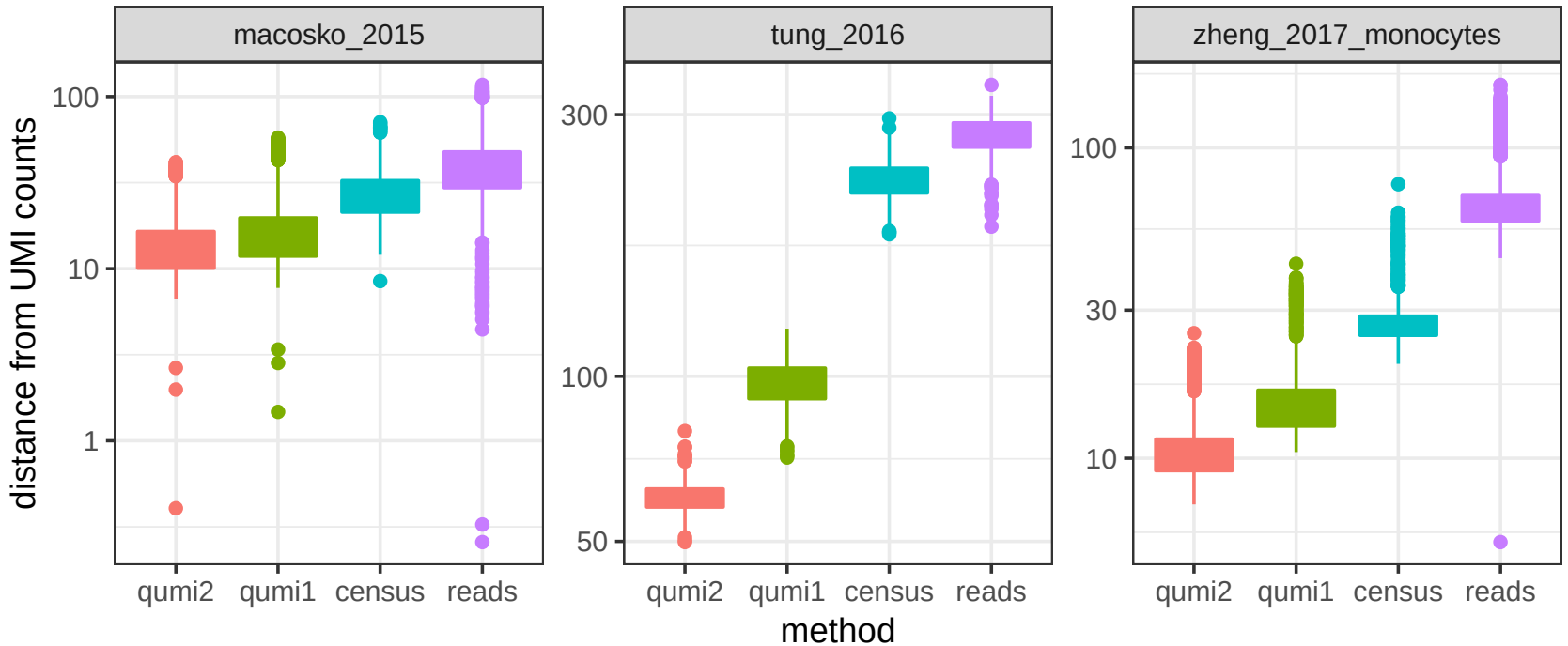
Quantile normalization of read counts





- Sometimes the generative process is too complex for modeling.
- UMI target distribution easier than Gaussian: "quasi-UMIs"
- Quasi-UMI only changes nonzero values

Quasi-UMI normalization accuracy



When does normalization work?

- **Large total UMI counts**
- Better capture efficiency and reverse transcriptase
- Consistently processed samples
- No amplification noise (PCR)

The future of normalization is bright thanks to wet lab innovation!

How to demonstrate success

- Ground-truth negative controls- no biology, verify removal of technical noise and batch effects
- Ground-truth positive controls- known biology, verify preservation of signal
- Denoising/ molecular cross-validation
- Simulations- how to know if correct generative model?
- Posterior predictive checks for Bayesian models

Ideas for tomorrow

- Read counts vs UMI counts- assess separately
- Learn from ecology & metagenomics- e.g. distance metrics
- Denoiser concept (Batson) for comparing implicit normalization of models
- **Negative controls**- Tung 2017, 10x purified cells, Sarkar 2019
- Positive controls- assessments will depend on downstream feature selection, dimension reduction, clustering, etc.
- Speed, memory consumption matter
- Sun et al 2019- comprehensive assessment of dim reduce
- Duo et al 2018- preprocessing, clustering assessments