What is the “right” distribution for modeling technical variability in scRNA-seq data?

Stephanie Hicks, Johns Hopkins
Josh Batson, CZ Biohub

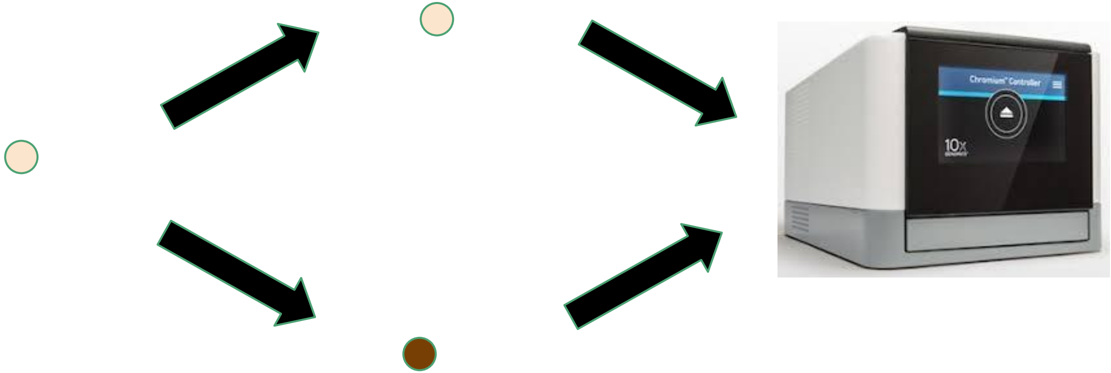
What is the “right” distribution for modeling technical variability in scRNA-seq data?

Stephanie Hicks, Johns Hopkins
Josh Batson, CZ Biohub

Technical variability is the variability between different measurements of the same biological unit.

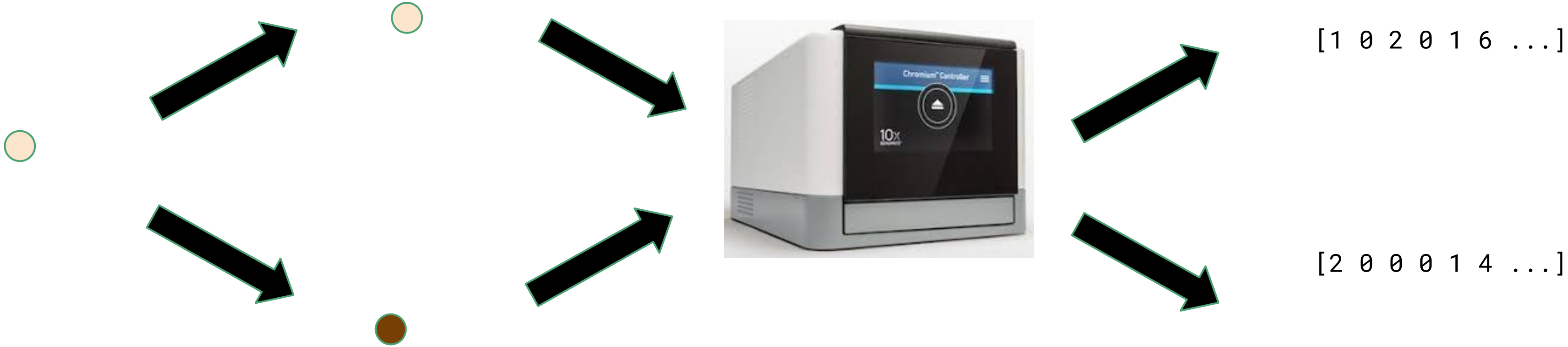
What is the biological unit?

Different measurements of the same **cell**.



What is the biological unit?

Different measurements of the same **cell**.

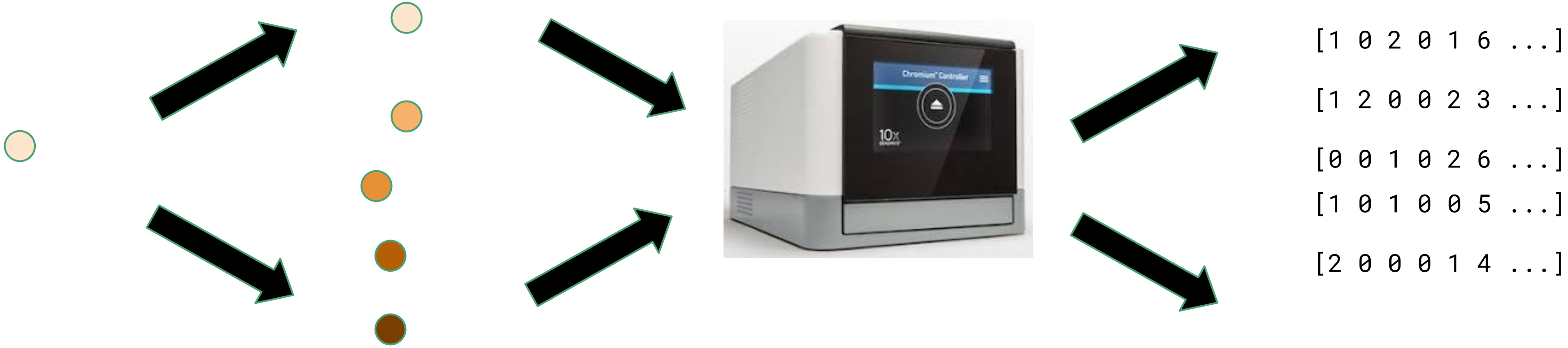


[1 0 2 0 1 6 ...]

[2 0 0 0 1 4 ...]

What is the biological unit?

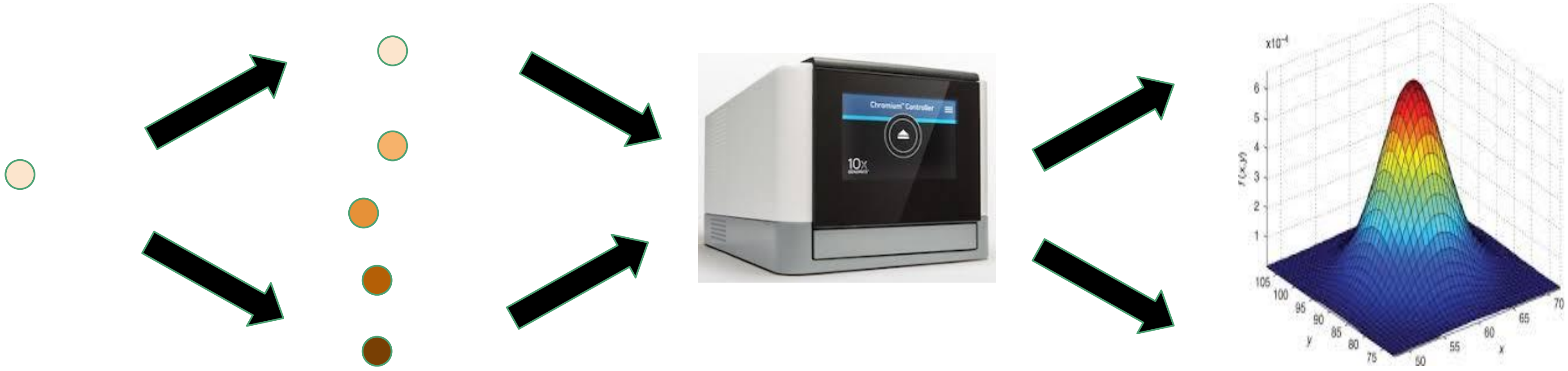
Different measurements of the same **cell**.



```
[1 0 2 0 1 6 ...]
[1 2 0 0 2 3 ...]
[0 0 1 0 2 6 ...]
[1 0 1 0 0 5 ...]
[2 0 0 0 1 4 ...]
```

What is the biological unit?

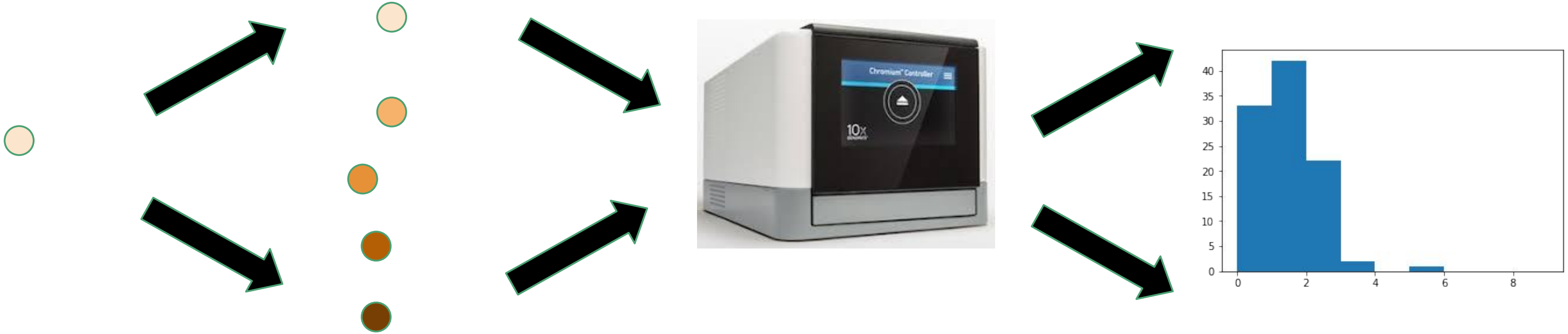
Different measurements of the same **cell**.



Joint Distribution

What is the biological unit?

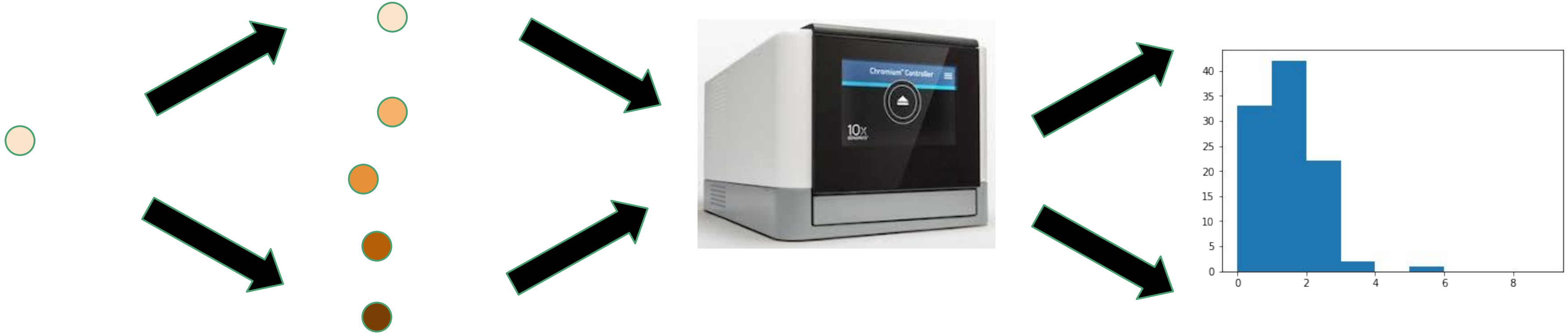
Different measurements of the same **cell**.



Marginal Distribution

What is the biological unit?

Different measurements of the same **cell**.

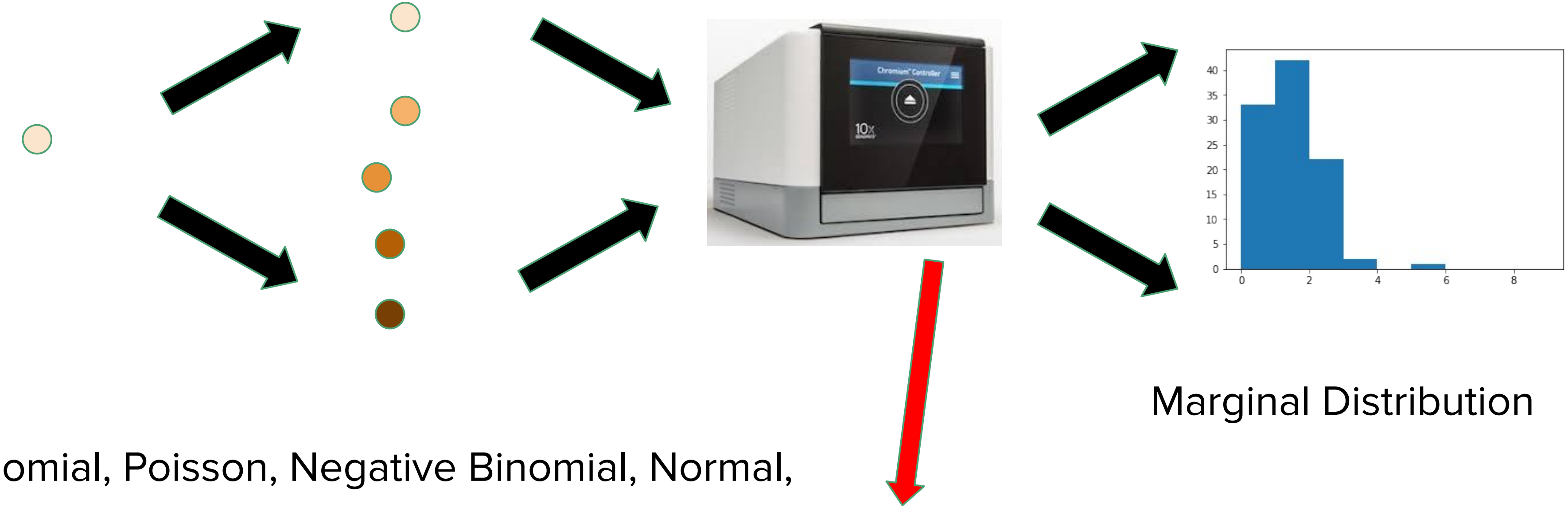


Marginal Distribution

Binomial, Poisson, Negative Binomial, Normal,
Log-Normal, Mixtures of Log-Normal

What is the biological unit?

Different measurements of the same **cell**.



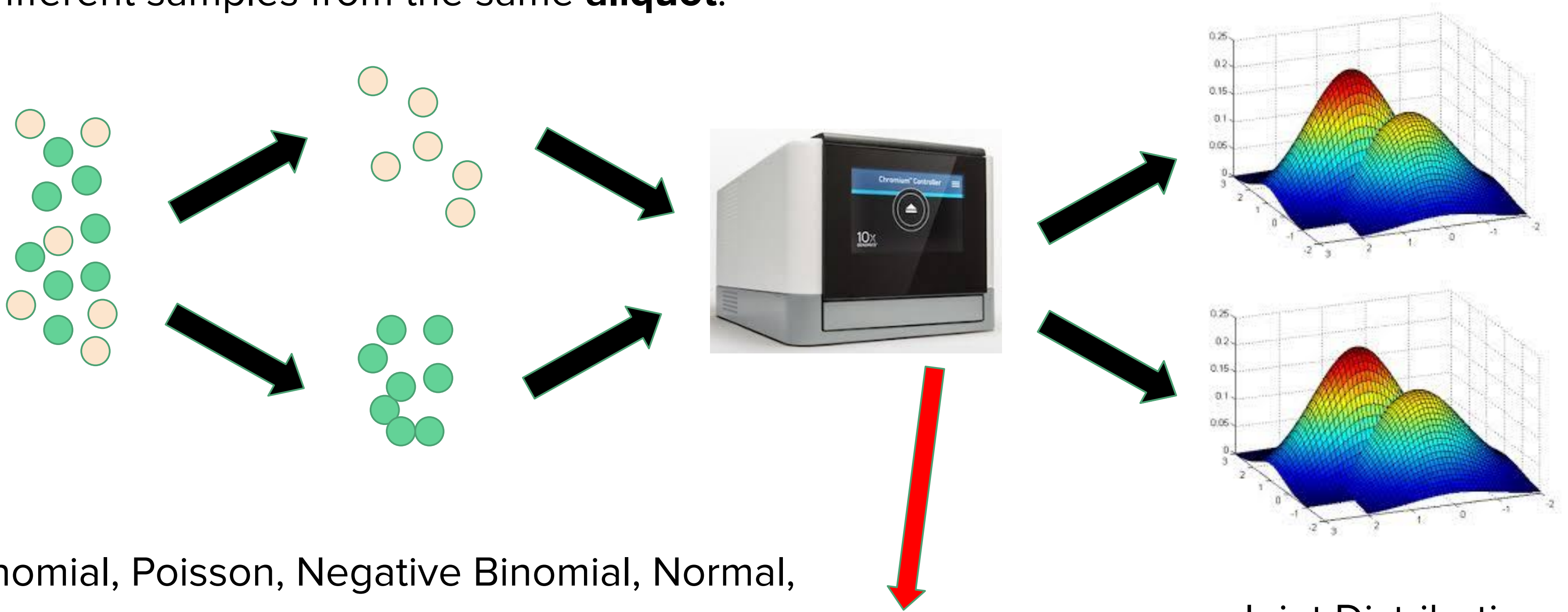
Marginal Distribution

Binomial, Poisson, Negative Binomial, Normal,

Log-Normal, Mixtures of Log-Normal +Zero-inflation

What is the biological unit?

Different samples from the same **aliquot**.

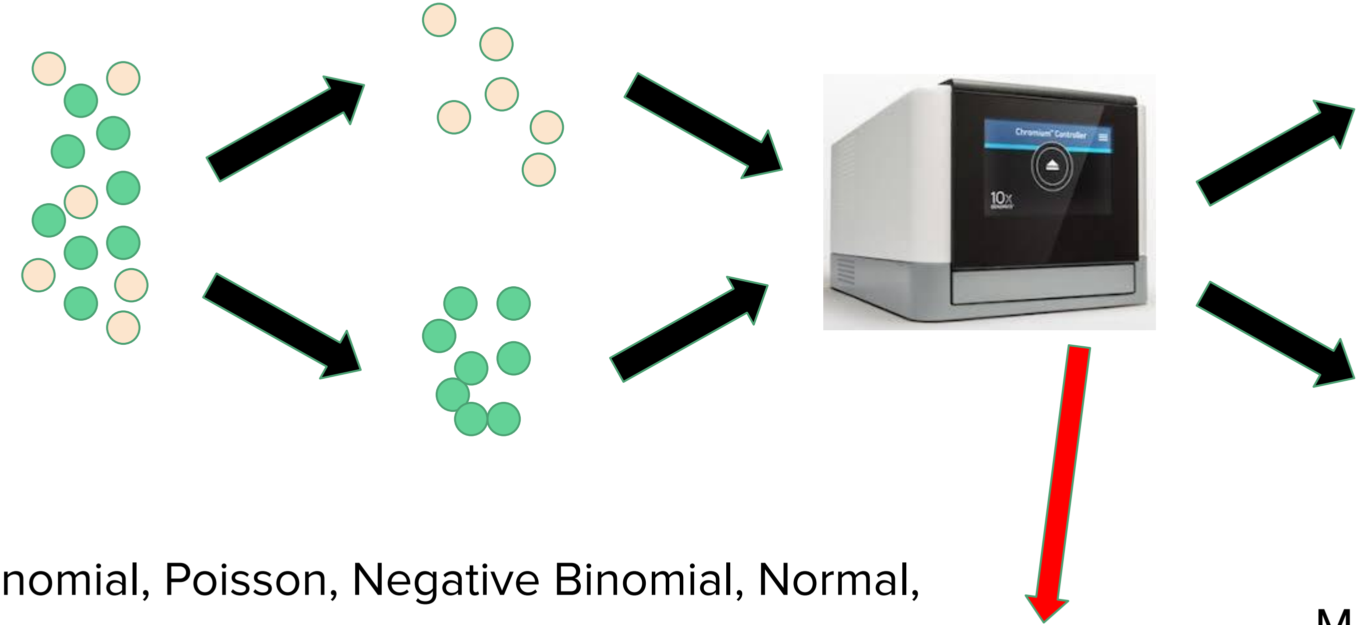


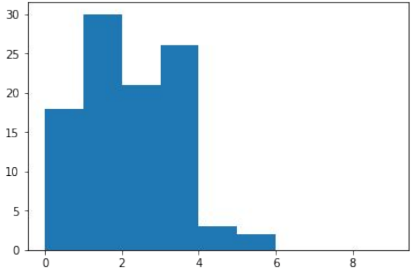
Binomial, Poisson, Negative Binomial, Normal, Log-Normal, Mixtures of Log-Normal +Zero-inflation

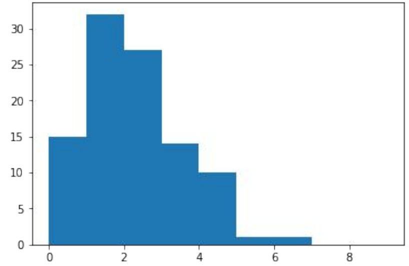
Joint Distributions

What is the biological unit?

Different samples from the same **aliquot**.





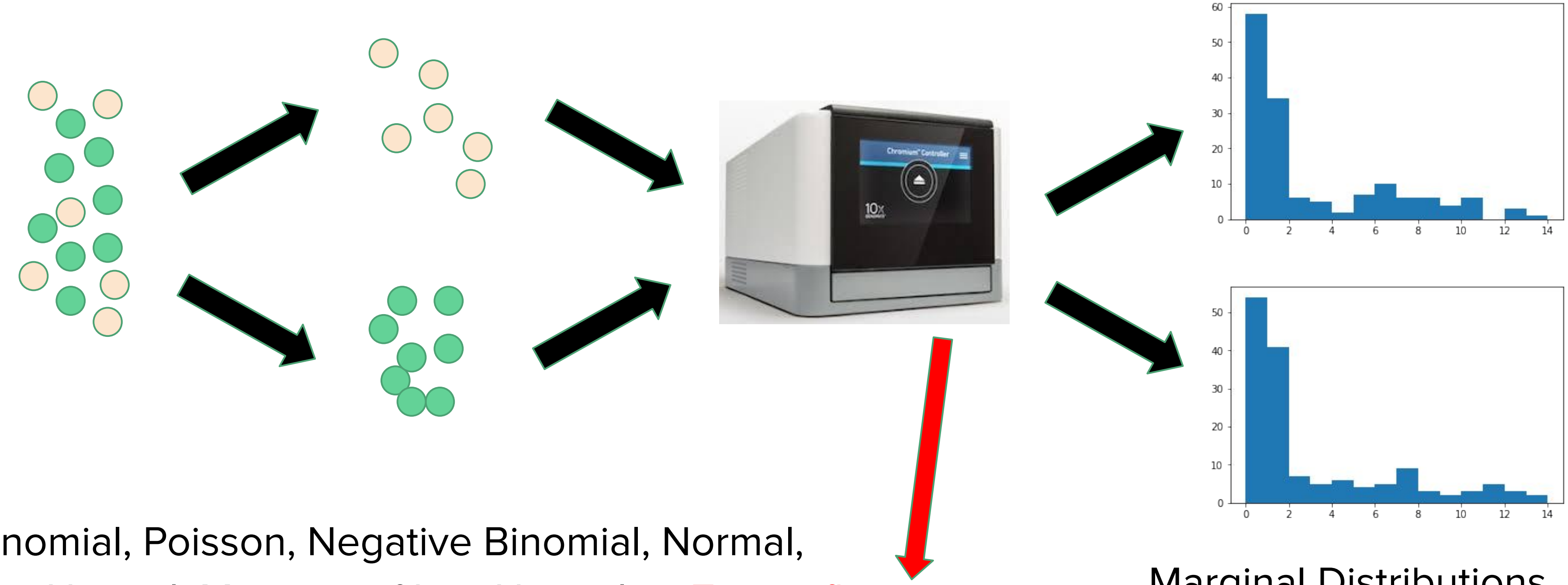


Binomial, Poisson, Negative Binomial, Normal,
Log-Normal, Mixtures of Log-Normal +Zero-inflation

Marginal Distributions

What is the biological unit?

Different samples from the same **aliquot**.



Binomial, Poisson, Negative Binomial, Normal,
Log-Normal, Mixtures of Log-Normal +Zero-inflation

Marginal Distributions

Joint Distributions for Aliquots

Question: Can we disentangle cell-self technical var from unmodelled cell-cell bio var? Should we care?

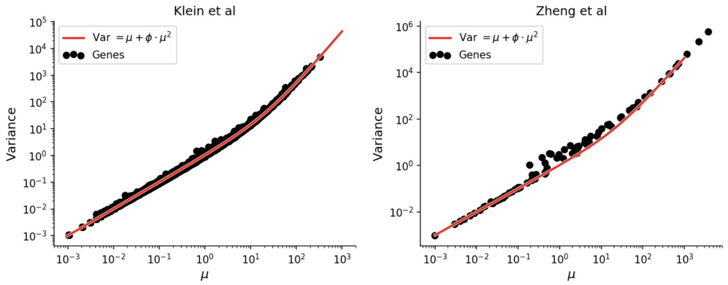
Does the (ZI)NB at the end do both?

Joint Distributions for Aliquots

Latent factors with a per-cell count-based distribution at the end model:

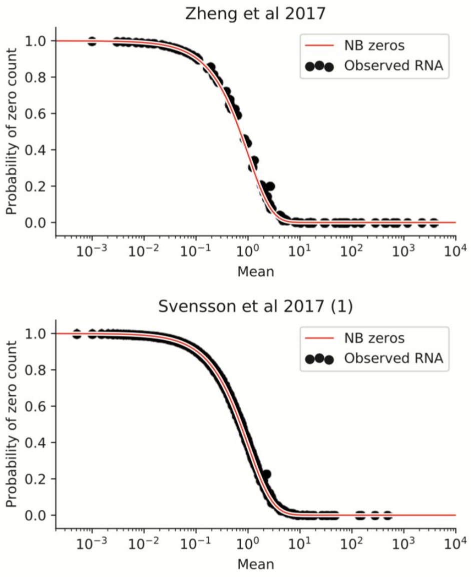
- Normal (or mixture of normals) + with extra component for zero-inflation (ZI)
  - MAST (Finak et al. 2015)
  - CIDR (Lin et al. 2017)
  - ZIFA (Pierson and Yau 2015)
- Count-based. Poisson, NB, ZINB, Multinomial
  - ZINB-WaVE (Risso et al. 2018) - ZINB-based factor analysis
  - DCA (Eraslan et al. 2018) - deep learning based autoencoder using NB model (with or without ZI)
  - scVI (Lopez et al. 2018) - variational autoencoder with ZINB model
  - scRecover (Miao et al. 2018) - ZINB model for imputation
  - GLM-PCA (Townes et al, 2019) - factor analysis with multinomial (P/NB) models

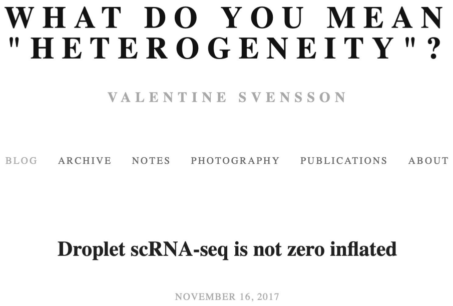
Marginal Distributions for Aliquots



Question: Is (UMI) single-cell RNA-seq data zero-inflated?

- Vieth et al. (2017) - In simulations, found NB to be sufficient for UMI data (aka no ZI)
- Blogpost from Svensson (2017) - observed counts are consistent with NB dist





WHAT DO YOU MEAN "HETEROGENEITY"?

VALENTINE SVENSSON

BLOG ARCHIVE NOTES PHOTOGRAPHY PUBLICATIONS ABOUT

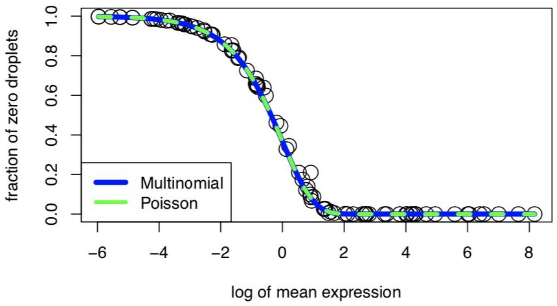
Droplet scRNA-seq is not zero inflated

NOVEMBER 16, 2017

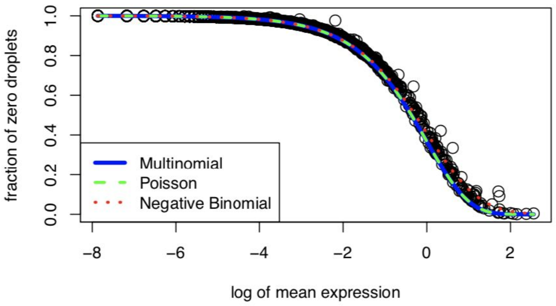
Marginal Distributions for Aliquots

Question: Is (UMI) single-cell RNA-seq data, conditioned on size factor, overdispersed?

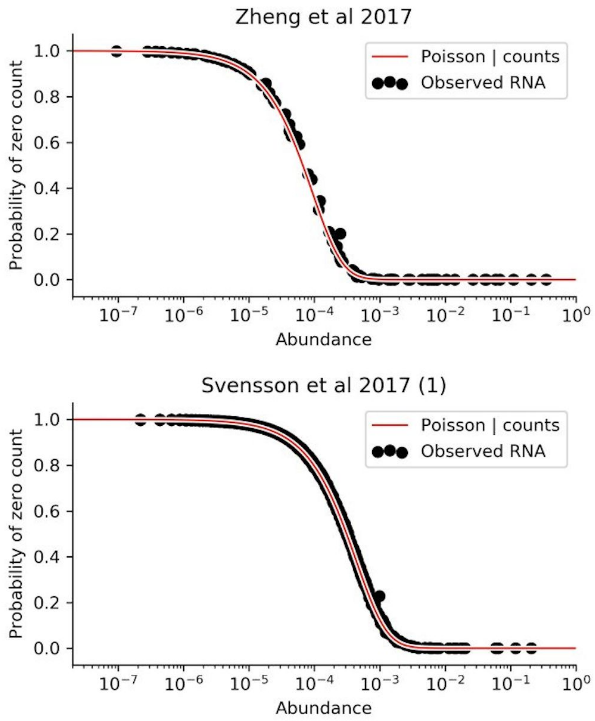
- Townes et al (2019)
- Batson (twitter, 2019)



(c) Technical replicates, per spike-in



(d) Biological replicates, per gene



Discussion Questions

Question: Can we disentangle cell-self technical var from cell-cell bio var? Should we care?

Question: Is single-cell RNA-seq data zero-inflated? (depends on sc protocol used)

Question: Is single-cell RNA-seq data, conditioned on size factor, overdispersed? Is that bio or technical?

Question: Can we estimate what would happen if we ran the same cell twice counterfactually? How could that inform our distributional models?

Proposal for Day 2 of normjam

Define ‘technical variability’. When is it identifiable from bio variability? Propose computational and wetlab (thought) experiments which could measure it.

Assemble the existing datasets that measuring each component and measure what they say about distributions.

eg: same cell with two beads. Same cell 10 min apart. Split sample from same patient. Do library prep twice. Do PCR twice.

Motivation

“All models are wrong, but some are useful.”

--folk wisdom

“….as a field has less theory, it has to leave more to the data. Since you can’t learn anything from data without the armature of statistical analysis, a field without theory tends to grow a thriving statistical community. Thus, the role of statistics grows as soon as the presence of scientific theory wanes.”

--Denny Borsboom, *Theoretical Amnesia*

First, let’s talk about the ***zeros***

Much has been concerned with demonstrating that scRNA-seq data have increased sparsity (or fraction of observed ‘zeros’ where a zero = no UMIs or reads mapping to a given gene in a cell) compared to bulk RNA-seq

Lots of early work: Shalek 2013; McDavid 2013; Kharchenko 2014; Trapnell 2014

Two types of possible zeros:

1. **Biological zeros**. e.g. gene not being expressed
2. **Technical zeros**. e.g. challenges in quantifying small # of mRNA (mRNA degradation during cell lysis, or variation from sampling lowly exp genes)

“**Dropout**” = prev used to describe observed zeros, but does not distinguish btw types of sparsity. I am asking to not use this as the catch-all term for observed zeros

Common statistical models for technical variability

Normal (or mixture of normals) + with extra component for zeros

- MAST (Finak et al. 2015) →

$$logit\left(Pr\left(Z_{ig}=1\right)\right)=\mathrm{X}_{\mathrm{i}}\beta_{\mathrm{g}}^{\mathrm{D}}$$

$$\Pr\left(Y_{ig}=y\middle|Z_{ig}=1\right)=\mathrm{N}\left(\mathrm{X}_{\mathrm{i}}\beta_{\mathrm{g}}^{\mathrm{C}},\ \sigma_{g}^{2}\right)$$

Common statistical models for technical variability

Normal (or mixture of normals) + with extra component for zeros

- MAST (Finak et al. 2015)
- CIDR (Lin et al. 2017)

$$logit\left(Pr\left(Z_{ig}=1\right)\right)=\mathrm{X_i}\beta_{\mathrm{g}}^{\mathrm{D}}$$

$$\mathrm{Pr}\left(Y_{ig}=y\middle|Z_{ig}=1\right)=\mathrm{N}\left(\mathrm{X_i}\beta_{\mathrm{g}}^{\mathrm{C}},\ \sigma_g^2\right)$$

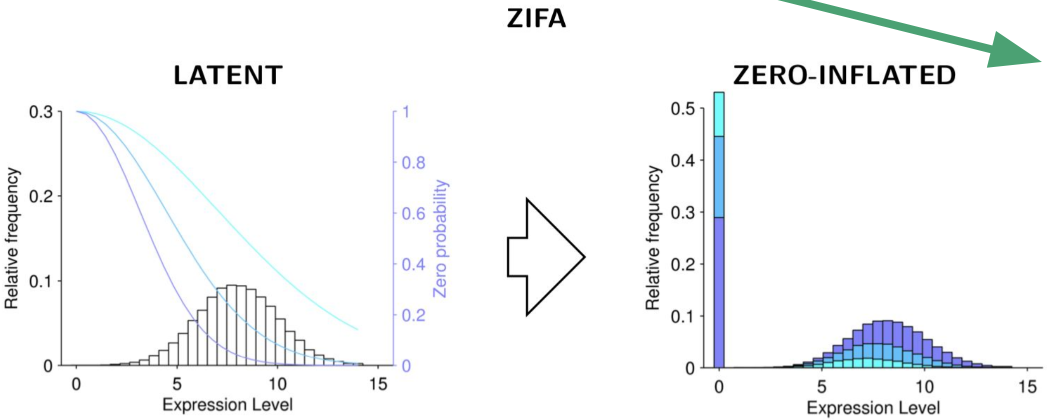
Common statistical models for technical variability

Normal (or mixture of normals) + with extra component for zeros (zero-inflated or ZI)

- MAST (Finak et al. 2015)
- CIDR (Lin et al. 2017)
- ZIFA (Pierson and Yau 2015)

$$logit\left(Pr\left(Z_{ig} = 1\right)\right) = X_i \beta_g^D$$

$$\Pr\left(Y_{ig} = y \middle| Z_{ig} = 1\right) = N\left(X_i \beta_g^C,\ \sigma_g^2\right)$$



$$\begin{aligned}
\mathbf{z}_i &\sim \text{Normal}(0, \mathbf{I}), \\
\mathbf{x}_i | \mathbf{z}_i &\sim \text{Normal}(\mathbf{A}\mathbf{z}_i + \boldsymbol{\mu}, \mathbf{W}), \\
h_{ij} | x_{ij} &\sim \text{Bernoulli}(p_0), \\
y_{ij} &= \begin{cases} x_{ij}, & \text{if } h_{ij} = 0, \\ 0, & \text{if } h_{ij} = 1, \end{cases}
\end{aligned}$$

$$p_0 = \exp(-\lambda x_{ij}^2)$$

Common statistical models for technical variability

What about the count nature of scRNA-seq data?

➔ zero-inflated (ZI) + negative binomial (NB)

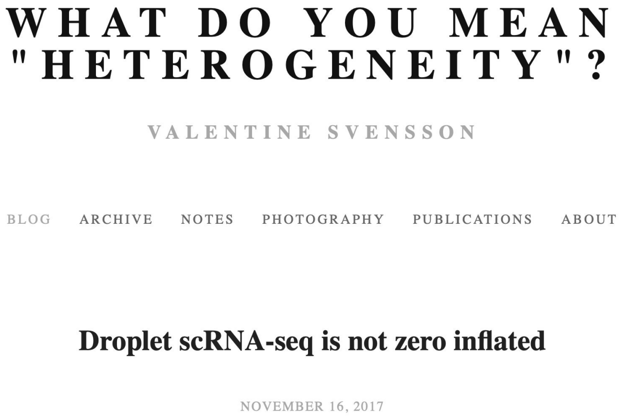
- ZINB-WaVE (Risso et al. 2018) - ZINB-based factor analysis
- DCA (Eraslan et al. 2018) - deep learning based autoencoder using NB model (with or without ZI)
- scVI (Lopez et al. 2018) - variational autoencoder with ZINB model
- scRecover (Miao et al. 2018) - ZINB model for imputation

Common statistical models for technical variability

“Droplet-based scRNA-seq data (with UMI counts) are not zero inflated”

➔ Binomial or negative binomial (NB) -- aka ZI is not needed

- Vieth et al. (2017) - In simulations, found NB to be sufficient for UMI data
- Blogpost from Svensson (2017) - observed counts are consistent with NB dist



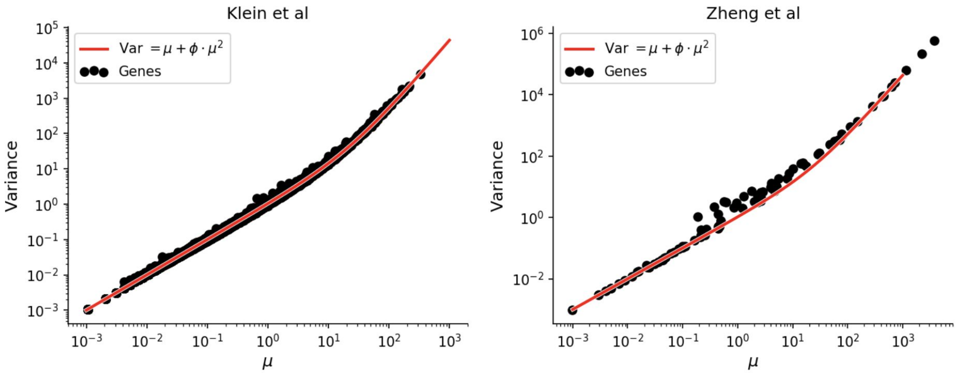
WHAT DO YOU MEAN "HETEROGENEITY"?

VALENTINE SVENSSON

BLOG ARCHIVE NOTES PHOTOGRAPHY PUBLICATIONS ABOUT

Droplet scRNA-seq is not zero inflated

NOVEMBER 16, 2017



Common statistical models for technical variability

“Droplet-based sccRNA-seq data (with UMI counts) are not zero inflated”

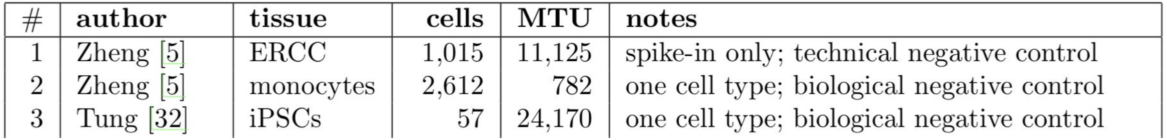
→ Binomial or negative binomial (NB) -- aka ZI is not needed

Many more:

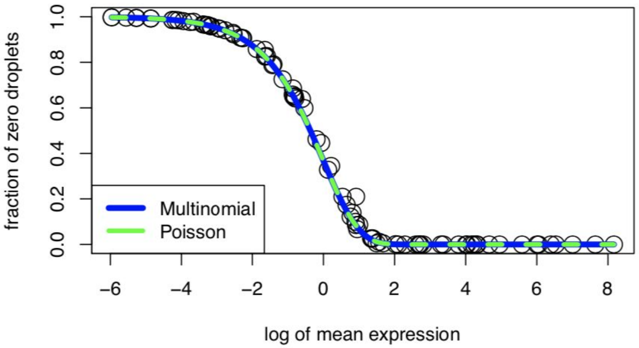
- bayNorm (Tang et al. 2018) - binomial model for imputation, empirical Bayes prior
- SAVER (Huang et al. 2018) - NB model, poisson LASSO regression prior
- sc (Eraslan et al. 2018) - deep learning based autoencoder using NB model (with or without ZI)
- scVI (Lopez et al. 2018) - variational autoencoder with ZINB model
- ....

Common statistical models for technical variability

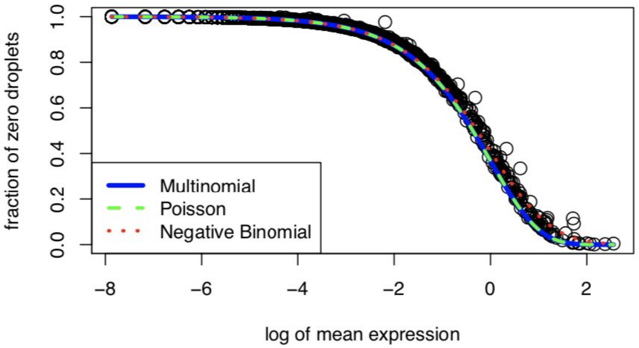
Townes et al. (2019) - UMI count data from negative control scRNA-seq datasets (i.e. identical RNA was added to droplets and sequenced aka we do not expect any biological variation) are well-described by **multinomial distributions**, which can be **approximated by Poisson and negative binomial distributions**



#	author	tissue	cells	MTU	notes
1	Zheng [5]	ERCC	1,015	11,125	spike-in only; technical negative control
2	Zheng [5]	monocytes	2,612	782	one cell type; biological negative control
3	Tung [32]	iPSCs	57	24,170	one cell type; biological negative control



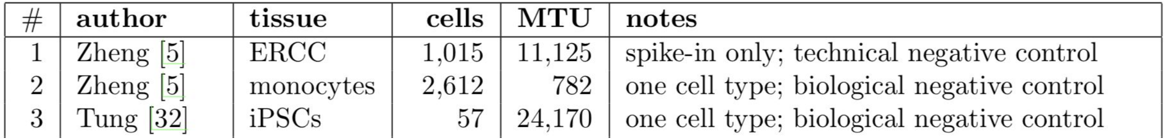
(c) Technical replicates, per spike-in



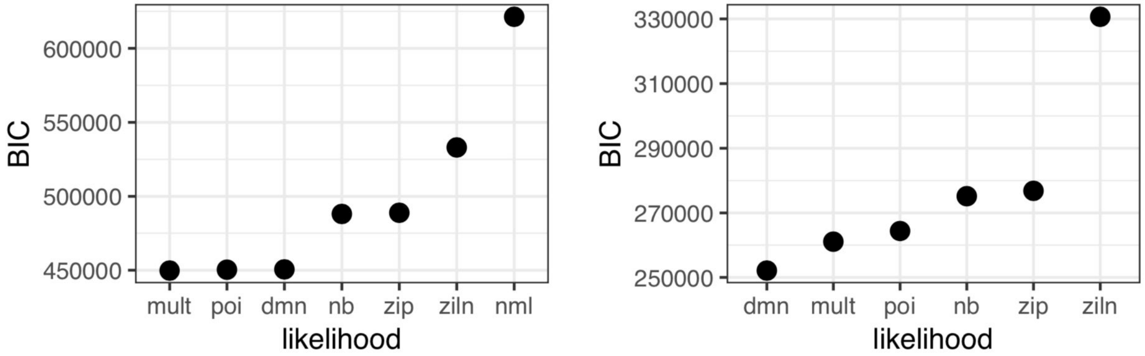
(d) Biological replicates, per gene

Common statistical models for technical variability

Townes et al. (2019) - UMI count data from negative control scRNA-seq datasets (i.e. identical RNA was added to droplets and sequenced aka we do not expect any biological variation) are well-described by **multinomial distributions**, which can be **approximated by Poisson and negative binomial distributions**



#	author	tissue	cells	MTU	notes
1	Zheng [5]	ERCC	1,015	11,125	spike-in only; technical negative control
2	Zheng [5]	monocytes	2,612	782	one cell type; biological negative control
3	Tung [32]	iPSCs	57	24,170	one cell type; biological negative control



(a) Tung UMI counts

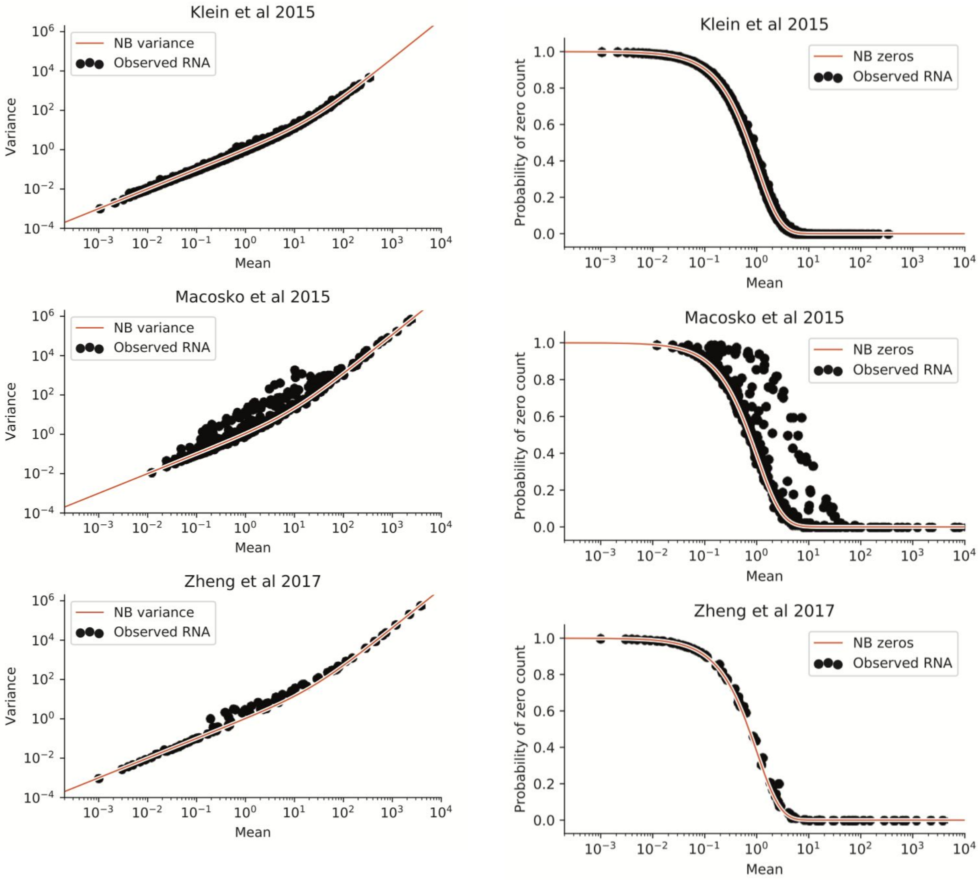
(b) Zheng monocytes UMI counts

Common statistical models for technical variability

March 11 -- **Townes et al. (2019)**

March 14 -- **Hafemeister and Satija (2019)** independently preprinted similar results, with a different error distribution (negative binomial)

March 18 -- **Svensson et al. (2019)** converted the analysis from the 2017 blog post into a preprint



Common statistical models for technical variability

- Normal (or mixture of normals) + with extra component for zeros
- Zero-inflated negative binomial
- Binomial
- Negative binomial
- Multinomial (approx with Poisson)

Zero-inflation

- Depends on sequencing platform (e.g. plate-based vs droplet-based / UMI)

What variability counts as “technical”?

Which counts exactly should follow the distribution?

- Same cell w/ two beads
- Different library preps
- One gene between different cells in a ‘homogeneous population’

What happens if you model it incorrectly?

Big questions

How do we know what is the right model of technical variability for our data?

Choice of method for downstream analysis will likely vary depending on this?

Do we need a formal way (e.g. statistical test) for assessing model assumptions?

Do we need a physical way (e.g. experiment) for assessing model assumptions?