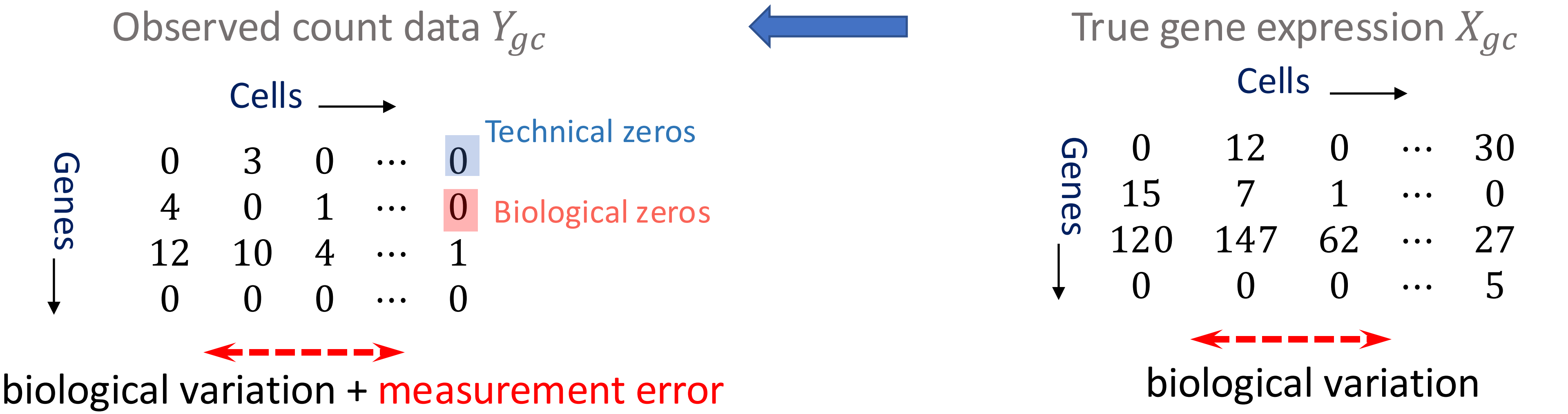
Lecture 3
scRNA-seq noise and signal distributions

Outline

- Modeling technical noise distributions in scRNA-seq count matrix
  - ERCC spike-ins

- Modeling biological variations of gene expressions across cells
  - Distribution deconvolution for scRNA-seq

scRNA-seq count matrix is very noisy



Observed count data Y_{gc}

Cells →

Genes ↓

$$\begin{matrix} 0 & 3 & 0 & \cdots & 0 \\ 4 & 0 & 1 & \cdots & 0 \\ 12 & 10 & 4 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{matrix}$$

Technical zeros

Biological zeros

biological variation + measurement error

True gene expression X_{gc}

Cells →

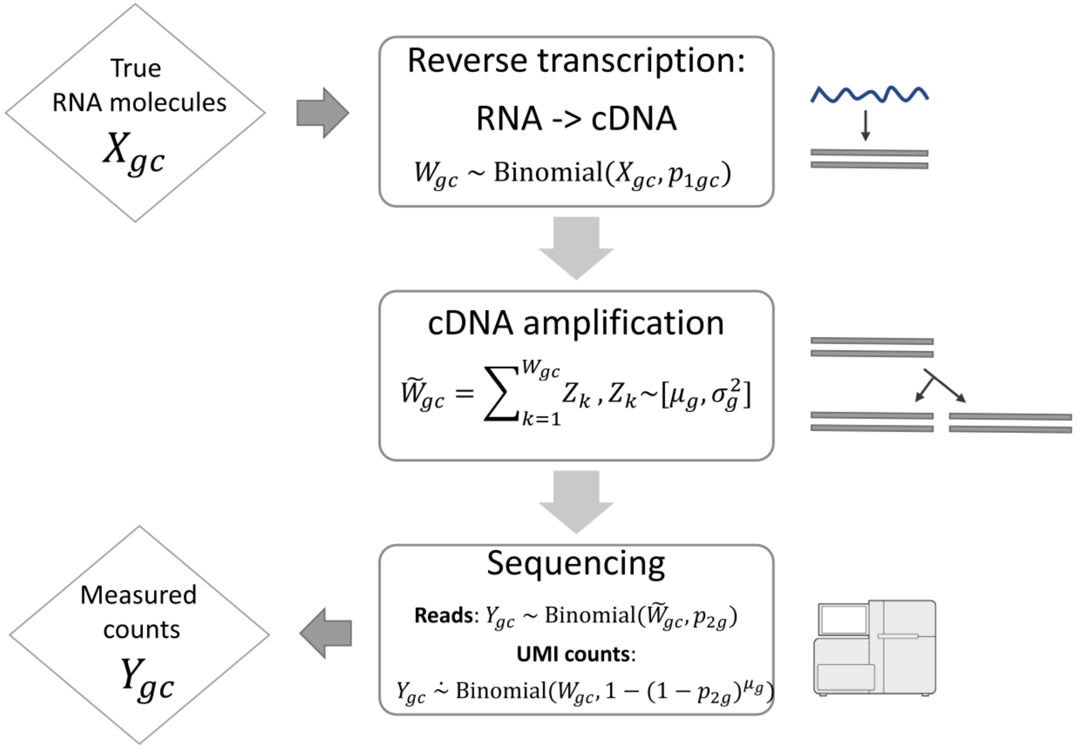
Genes ↓

$$\begin{matrix} 0 & 12 & 0 & \cdots & 30 \\ 15 & 7 & 1 & \cdots & 0 \\ 120 & 147 & 62 & \cdots & 27 \\ 0 & 0 & 0 & \cdots & 5 \end{matrix}$$

biological variation

- Observed count matrix Y is typically extremely sparse
  - Dropout: zeros in the count matrix (a vague concept)
  - Two types of zeros
    - Biological zeros: true mRNA count is zero
    - Technical zeros: true mRNA count is not zero, but observed count is zero
- We will discuss the measurement error distributions and signal distributions (biological variations across cells / gene-gene dependence) separately
  - Why do we care about these?
    - Reasonable statistical / machine learning model to use in analyzing the data
    - How to simulate scRNA-seq data to benchmark different methods?

Measurement error distribution



True RNA molecules X_{gc}

Reverse transcription: RNA -> cDNA

$W_{gc} \sim \text{Binomial}(X_{gc}, p_{1gc})$

cDNA amplification

$\widetilde{W}_{gc} = \sum_{k=1}^{W_{gc}} Z_k, Z_k \sim [\mu_g, \sigma_g^2]$

Sequencing

Reads: $Y_{gc} \sim \text{Binomial}(\widetilde{W}_{gc}, p_{2g})$

UMI counts:

$Y_{gc} \dot{\sim} \text{Binomial}(W_{gc}, 1 - (1 - p_{2g})^{\mu_g})$

Measured counts Y_{gc}

- Both reverse transcription and sequencing can generate technical zeros, which can be theoretically explained by Binomial distributions
$$Y_{gc} \sim \text{Binomial}(X_{gc}, \alpha_c \gamma_g)$$

- Due to low efficiency ($\alpha_c < 10\%$), roughly
$$Y_{gc} \sim \text{Poisson}(\alpha_c \gamma_g X_{gc})$$

- Sequencing depth: total number of reads per cell
  - Refer to p_{2g}: deeper sequencing depth, more reads sampled from the library
  - Roughly controllable by experimenters, depends on the budget

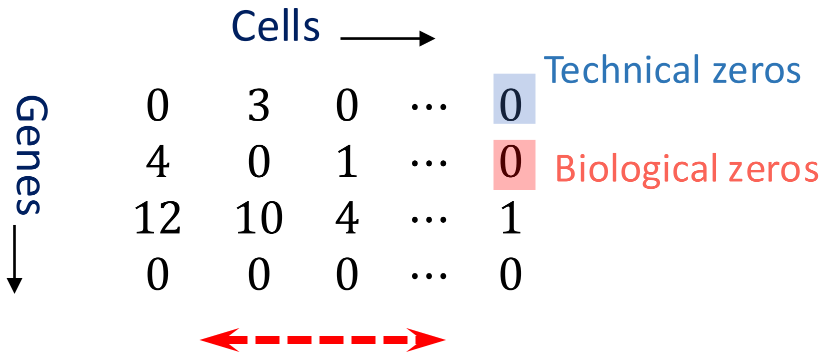
Noise distribution: zero inflation or not?

- Gaussian assumptions on the observed data (even after transformations) usually do not work well
  - scRNA-seq data is extremely sparse

- Because of the extreme sparsity of scRNA-seq data, many earlier papers have used a zero-inflated model: such as zero-inflated Poisson or zero-inflated negative binomial model for scRNA-seq data

  - A zero-inflated model have more parameters to fit, is it worth it?

Observed count data Y_{gc}



$$\begin{array}{ccccc} 0 & 3 & 0 & \cdots & 0 \\ 4 & 0 & 1 & \cdots & 0 \\ 12 & 10 & 4 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{array}$$

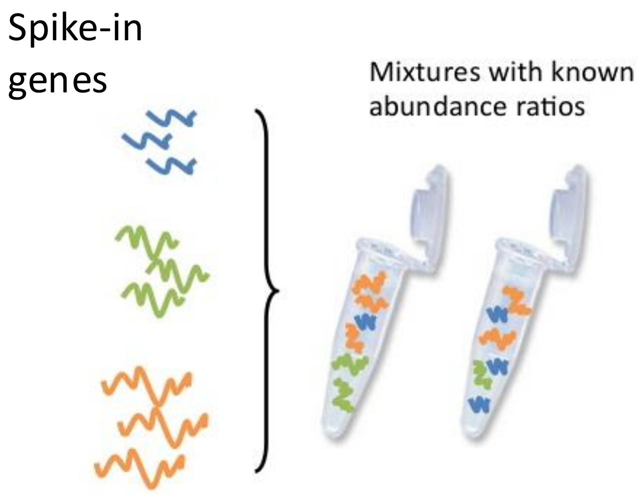
biological variation + measurement error

ERCC spike-ins

- For UMI counts, $Y_{gc} \sim \text{Poisson}(\alpha_c \gamma_g X_{gc})$
  A Poisson distribution + cell-specific efficiency seems sufficient

- The above model is only a simplification, can we find empirical evidence?
  - Typically challenging to separate biological variations from measurement errors
  - Distribution of true gene expression X_{gc} can be complicated (will discuss later)
  - α_c is typically also unidentifiable

- ERCC spike-in 'gene' g (negative controls):



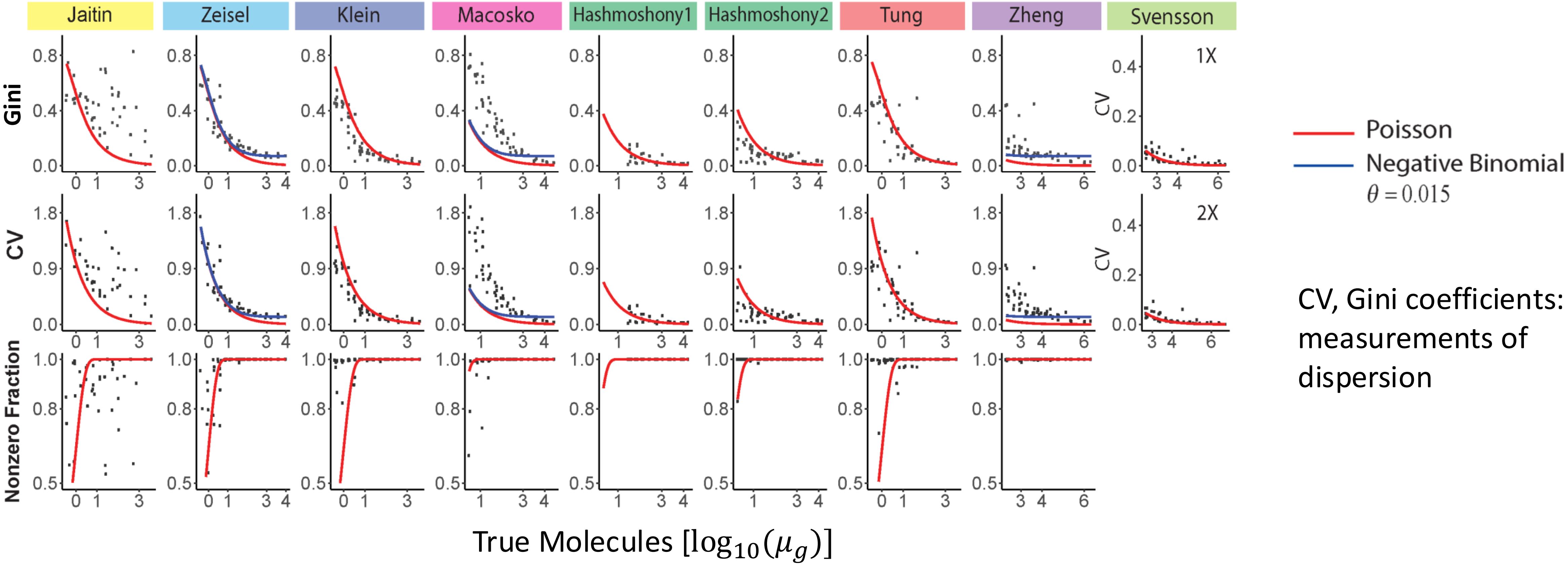
- $X_{gc} \overset{i.i.d}{\sim} \text{Poisson}(\overset{\text{Known}}{\mu_g})$

- Conventionally, researchers treat X_{gc} as constant across cells
  $$\text{Var}(Y_{gc}) = 2\alpha_c \gamma_g \mu_g$$

- Assume $\gamma_g = 1$, then α_c is identifiable

Noise distribution for UMI data is not zero-inflated

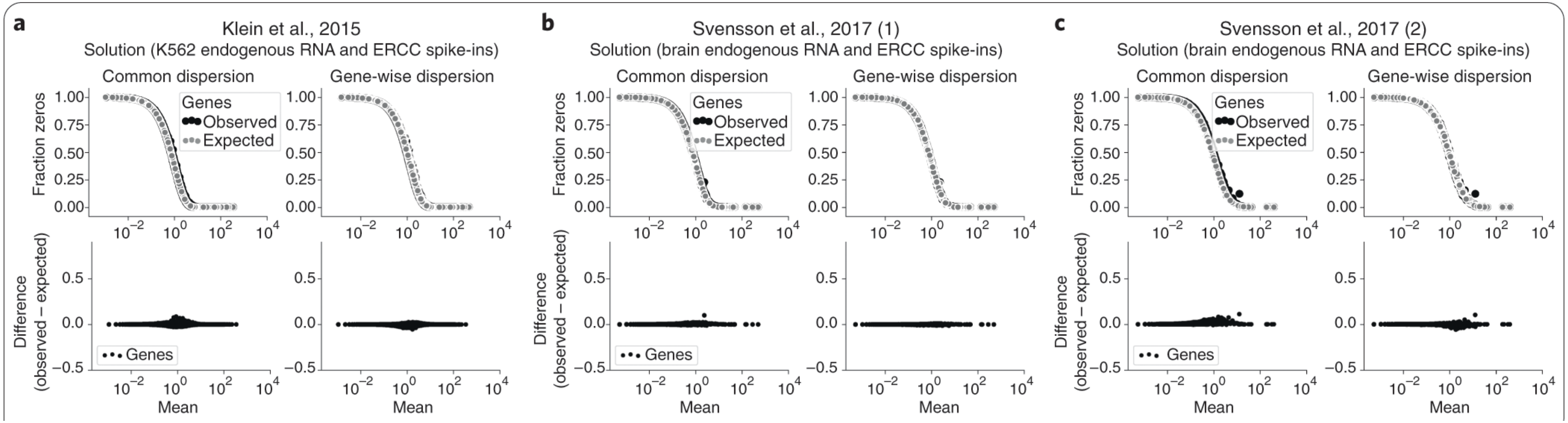
- Some empirical evidence using ERCC spike-ins
  - (Wang et. al. PNAS 2018):
    Assuming the Poisson noise model $Y_{gc} \sim \text{Poisson}(\alpha_c X_{gc})$, used a distribution deconvolution method to estimate the distribution of X_{gc} across cells for each ERCC spike-in gene



CV, Gini coefficients: measurements of dispersion

Noise distribution for UMI data is not zero-inflated

- Some empirical evidence using ERCC spike-ins
  - (Svensson, Nature Biotech, 2020):
    Use Negative-Binomial distribution to model the ERCC spike-ins and $Y_{gc} \sim \mathrm{NB}(\mu_g, \theta_g)$
    check if the observed zero proportion match with the estimated values



Factors affecting the noise distribution

- Batch effect:
  - non-biological factors in an experiment cause changes in the data produced by the experiment
  - Common causes: laboratory conditions, Choice of reagent lot or batch, Personnel differences, Time of day when the experiment was conducted, instruments used to conduct the experiment
  - Long-standing issue for sequencing data
  - New challenge for single-cell sequencing data (more in later lectures)

  - Batch effects introduce both biases and over-dispersion to the noise distribution

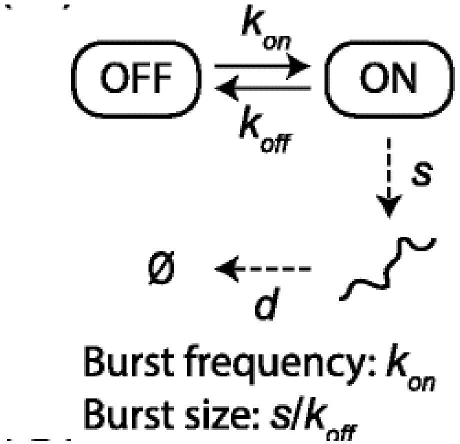
  - With batch effects, the actual noise distribution may be more dispersed than a Poisson model

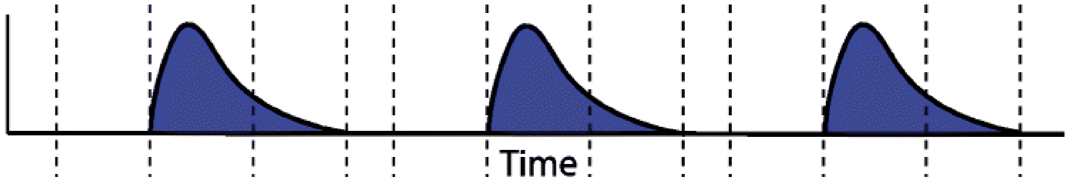
- Researchers have shown that zero-inflation noise model can still benefit non-UMI data

True biological variations

- Distribution of X_{gc} across cells can be really complicated
  - Diversity of cell types
    - many genes are unexpressed in a cell
    - cells of distinct types have different genes expressed

  - Transcriptional bursting





Jiang, Yuchao, Nancy R. Zhang, and Mingyao Li. "SCALE: modeling allele-specific gene expression by single-cell RNA sequencing." *Genome biology* 18 (2017): 1-15.

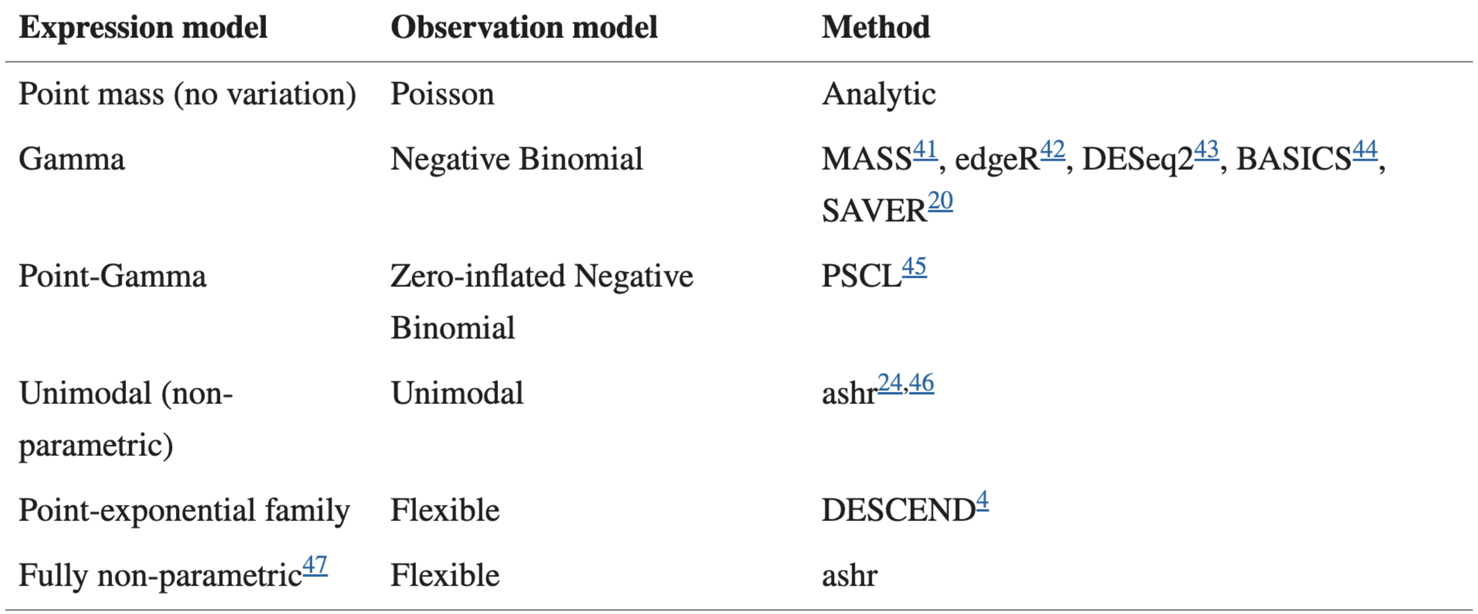
- For a given time interval, number of mRNAs for a gene in a cell follows Poisson-beta distribution (Kepler and Elston, Biophysical J, 2001)

$$Y \sim \text{Poisson}(sp), p \sim \text{Beta}(k_{on}, k_{off})$$

- X_{gc} across cells in a homogenous cell population should also follow a similar distribution

Modeling true gene expression distribution

- True distribution of X_{gc} can be really complicated
  - It is also not identifiable from most scRNA-seq data (as we only know library size l_c instead of efficiency α_c)
  - It is only possible to model the gene expression proportion $p_{gc} = \frac{X_{gc}}{\sum_g X_{gc}}$
    - Without considering batch effects, we may assume $Y_{gc} \sim \mathrm{Poisson}\left(l_c p_{gc}\right)$



Expression model	Observation model	Method
Point mass (no variation)	Poisson	Analytic
Gamma	Negative Binomial	MASS[41], edgeR[42], DESeq2[43], BASICS[44], SAVER[20]
Point-Gamma	Zero-inflated Negative Binomial	PSCL[45]
Unimodal (non-parametric)	Unimodal	ashr[24,46]
Point-exponential family	Flexible	DESCEND[4]
Fully non-parametric[47]	Flexible	ashr

Table 1 of Sarkar and Stephens, Nature Genetics, 2021

- Dependence structure across genes

DSCEND (Wang et. al. PNAS 2018)

- Distribution deconvolution

$$X_{gc} \overset{i.i.d}{\sim} H_g \xrightarrow[\text{Technical noise}]{} Y_{gc} \xrightarrow{\text{Distribution Deconvolution}} \widehat{H}_g \overset{?}{\approx} H_g$$

$$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} \text{Poisson}(l_c X_{gc})$$

- Semi-parametric distributional assumption (G-modeling, Efron Biometrika 2016)

$$h_g(x) = \pi_g \delta_0 + (1 - \pi_g)\exp[Q(x)^T \alpha - g(\alpha)]$$

- $Q(x)$ is non-parametric, and is estimated by cubic splines after discretizing the data
  - For $x \neq 0$, Assume that $x \in \boldsymbol{x} = (x_1, \cdots, x_m)$

    $$\mathbb{P}[X = \boldsymbol{x}] = \exp\{Q^T \alpha - \phi(\alpha)\}$$

    where Q is the 5-degree natural cubic spline matrix at $\boldsymbol{x}$
  - Incorporate covariates in the distribution:
    - Incorporate covariates in both π_g and the non-zero part
    - Non-zero part: assume $X_{gc} = e^{U_c \beta} \tilde{X}_{gc}$ where $\tilde{X}_{gc} \sim H_g$

- Statistical inference: Taylor expansion on the estimating equation

Validation using FISH experiment

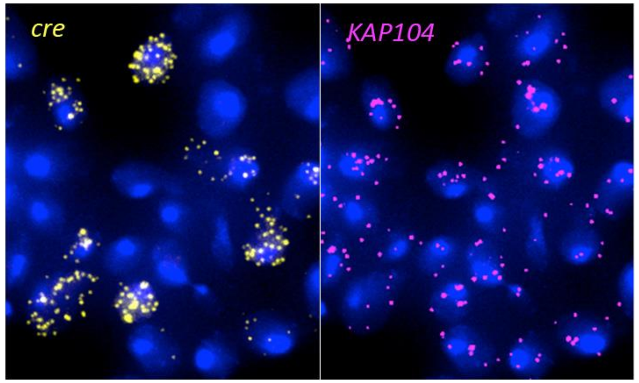
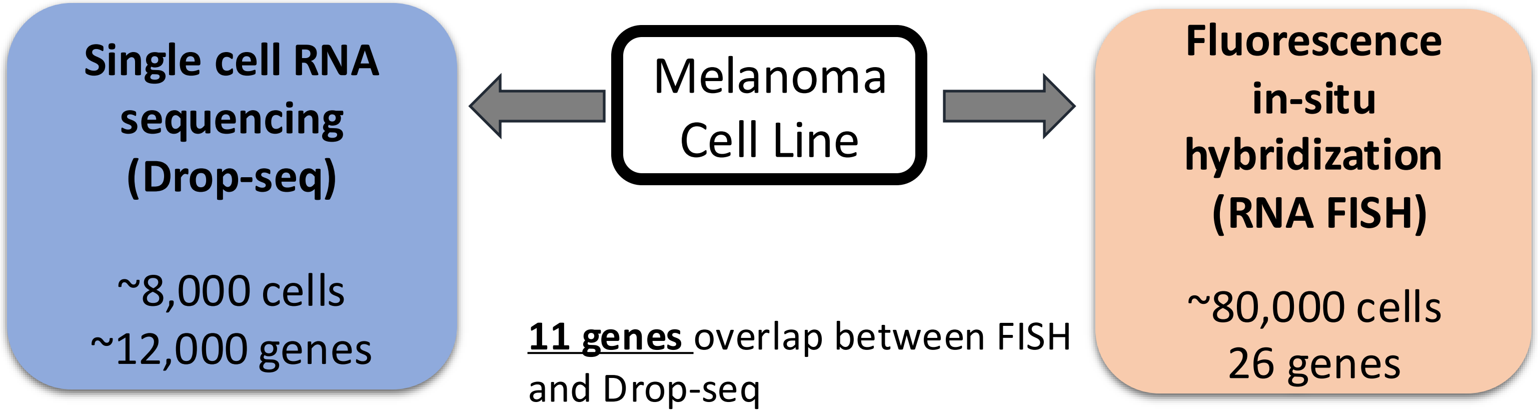


Photo courtesy of Anne Dodson and Professor Jasper Rine



Single cell RNA sequencing (Drop-seq)

~8,000 cells
~12,000 genes

Melanoma Cell Line

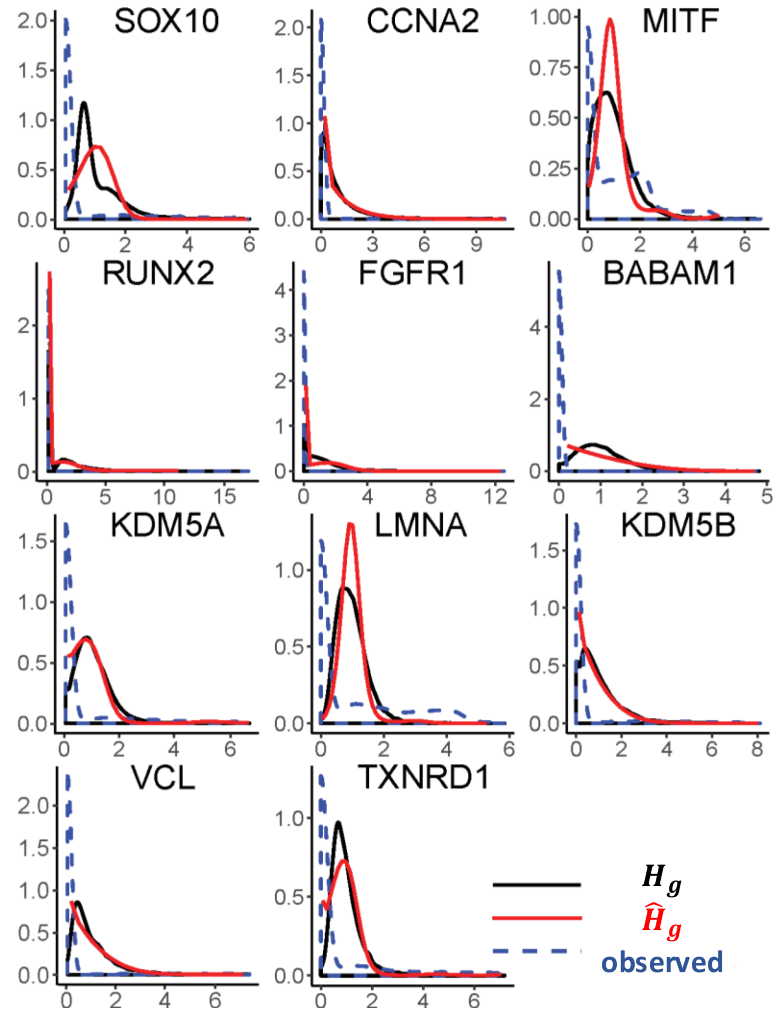
Fluorescence in-situ hybridization (RNA FISH)

~80,000 cells
26 genes

11 genes overlap between FISH and Drop-seq

Much more Accurate

$\hat{H}_g$ V.S. H_g



Modeling distribution of observed counts

- Why do we want to separate the true gene expression variation from the noise distribution?
    - Researchers are interested in the proportion of true zeros
    - Identify changes in gene expression variations instead of in mean

- Sometimes we may just want to model the observed counts
    - Example: test for gene expression mean changes between two cell types

- Complexity in true gene expression can bring in both over-dispersion and zero-inflation in the observed count if we just use a Poisson model with cell-specific library size
    - A common approach is to use a Negative-Binomial distribution or zero-inflated NB distribution
    - (Kim et. al. Genome Biology 2020) showed that Poisson distribution is good enough to model Y_{gc} for a relatively homogenous cell population
    - (Saket and Satija, Genome Biology 2022) showed that Poisson distribution is not enough to model Y_{gc} for a relatively homogenous cell population if sequencing is not shallow and should use a Negative Binomial distribution

- A common approach is to use an autoencoder (latent factor model) to capture gene-gene dependence and cell population heterogeneity use NB likelihood to construct loss function

Related papers

- Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., ... & Zhang, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. Proceedings of the National Academy of Sciences, 115(28), E6437-E6446.
- Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. Nature Biotechnology, 38(2), 147-150.
- Sarkar, A., & Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nature genetics, 53(6), 770-777.
- Kim, T. H., Zhou, X., & Chen, M. (2020). Demystifying “drop-outs” in single-cell UMI data. Genome biology, 21(1), 196.
- Choudhary, S., & Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. Genome biology, 23(1), 27.