

STAT 35510

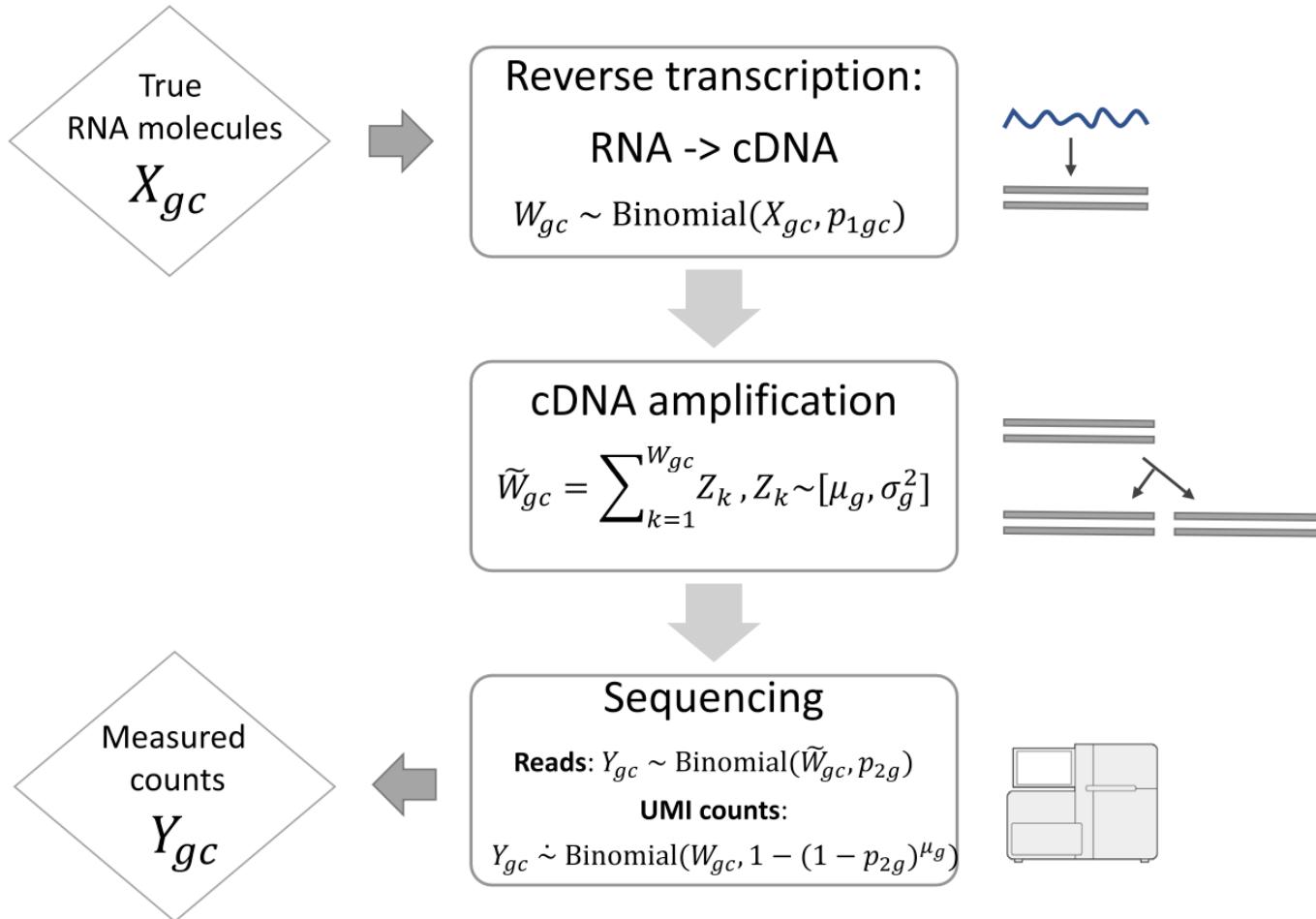
Lecture 2

Spring, 2024
Jingshu Wang

Outline

- Measurement error in scRNA-seq experiments
- Doublet removal and ambient RNA correction
- Biological variations and technical noise distributions in scRNA-seq count matrix

Propagation of measurement error



- A cell c , a gene g
- For UMI counts, roughly $Y_{gc} \sim \text{Binomial}(X_{gc}, \alpha_{gc})$
- For non-UMI reads:
 - $Y_{gc} = 0$ if $W_{gc} = 0$
 - Y_{gc} can be large if W_{gc} due to amplification
- Most of scRNA-seq data nowadays use UMI

library size

- For UMI counts, roughly

$$Y_{gc} \sim \text{Binomial}(X_{gc}, \alpha_{gc})$$

where α_{gc} is the cell-gene-specific efficiency

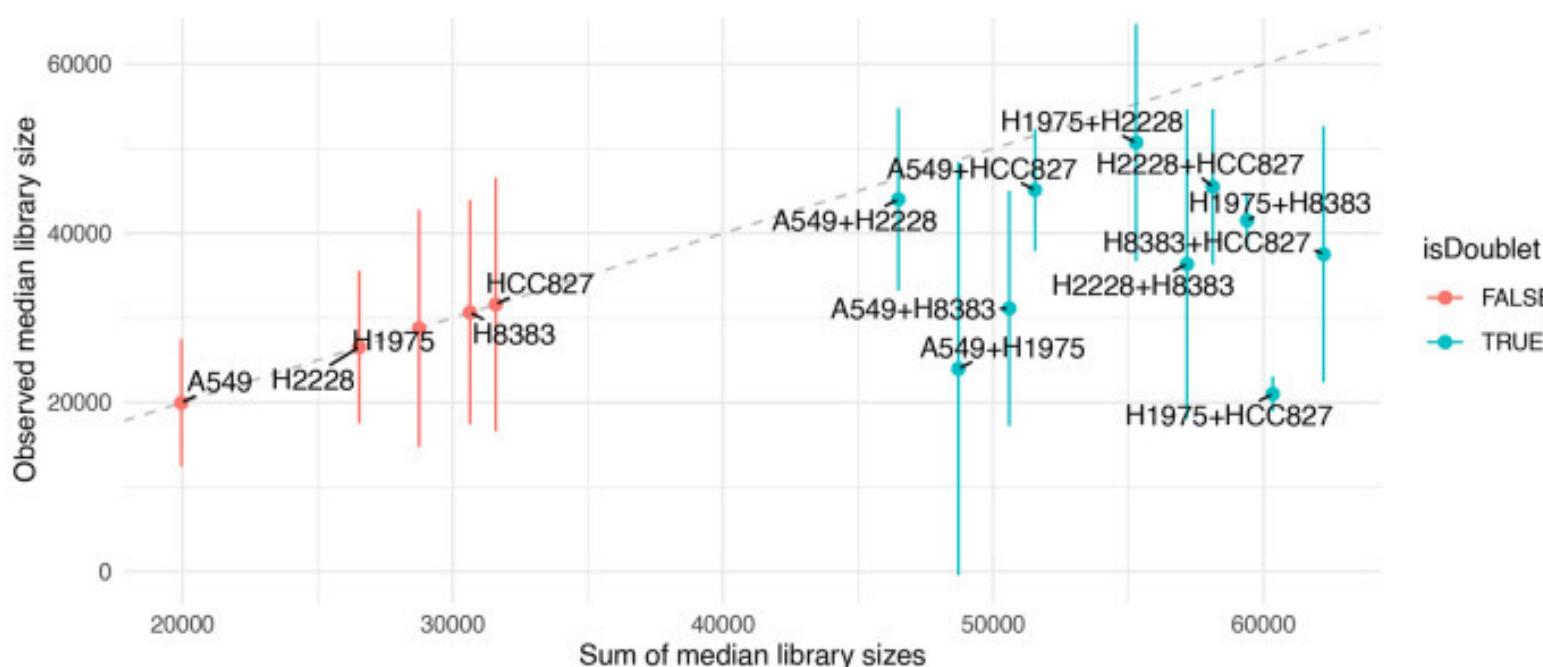
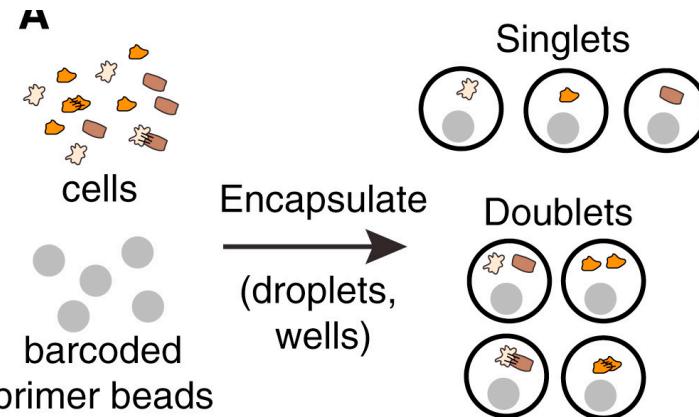
- Assume that $\alpha_{gc} \approx \alpha_c \gamma_g$ where α_c is cell-specific efficiency and γ_g is a gene-specific bias
- Researchers have observed that α_c can vary greatly across cells, but it is typically unidentifiable (will talk more in later slides)
- **Library size of a cell:** total sum of UMI counts across all measured genes in a cell

$$l_c = \sum_g Y_{gc}$$

- Cells with large library size
 - Large cells containing many mRNAs (like neurons), high-quality cells where mRNAs are efficiently captured, doublets
- Cells with small library size
 - Small cells containing few mRNAs, low-quality cells, empty droplets
- Library size normalization: Y_{gc} is not comparable across cells, compare relative proportion Y_{gc}/l_c across cells

Doublets

- It is always possible that two (or more) cells share the same barcode
 - Common to have 10% - 20% doublets in scRNA-seq experiments
 - More cells → higher proportion of doublets
- Doublets or multiplets may have relatively large library size, but removing them simply based on library size is not efficient



Germain, Pierre-Luc, et al.
"Doublet identification in single-cell sequencing data using scDblFinder." *F1000Research* 10 (2021).

Doubllets

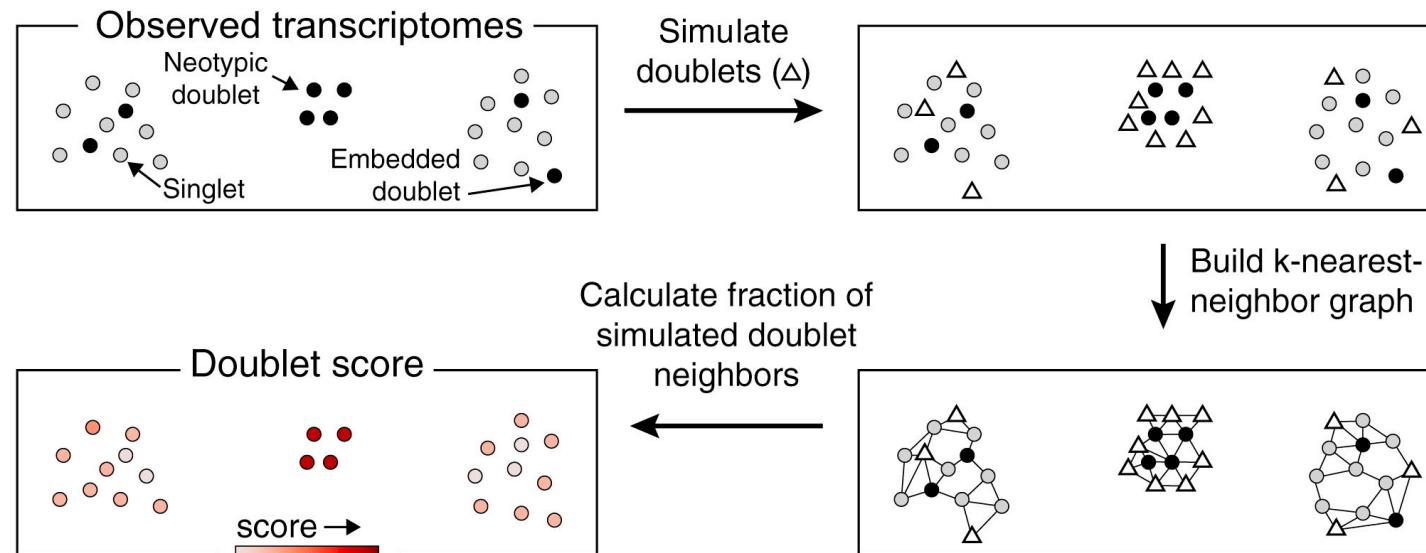
- Two major types of doubllets
 - Homotypic doubllets: formed by cells of the same "type"
 - Transcriptomic profile looks similar to a singlet
 - Hard to identify but also not that harmful for most data analysis purposes
 - Heterotypic doubllets: formed by cells of distinct transcriptional states
 - Possible to identify due to their distinct gene expression profile
- Experimental approaches to identify doubllets
 - Very few false positives, but requires special experimental design (not available for most experiments)
 - Example techniques:
 - species mixture: only works for experiments with multiple species
 - demuxlet (Kang et. al. Nature Biotech 2018): use SNP, works for experiments involving multiple individuals
- Computational approaches: identify doubllets solely based on count matrix

Scublet (Wolock et. al. Cell Systems, 2019)

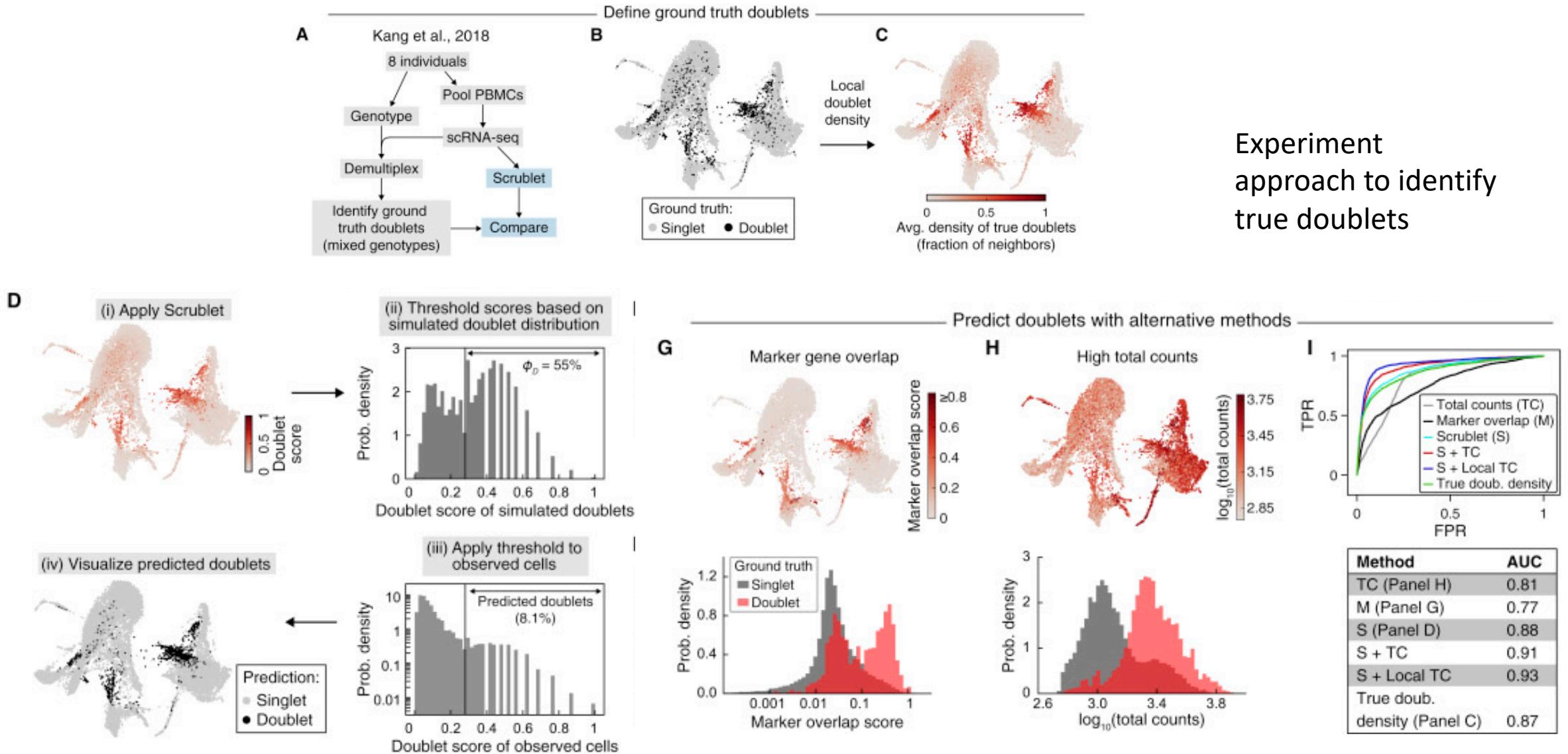
- Core idea:
 - Simulate doublet by combining random pairs of cells
 - Remove cells if they are similar to the simulated doublets
 - Do not rely on library size at all
- Simulate pseudo-doublets:
 - the counts for gene g in doublet i with parent cells a and b is $Y_{gi} = Y_{ga} + Y_{gb}$
- KNN classifier to identify cells similar to the pseudo-doublets
 - Merge observed cells and pseudo-doublets and preprocess the merged data: Normalization, identify highly variable genes, scaling, PCA (more details in Lecture 3)
 - Find k nearest neighbors of each cell using Euclidean distance (by default)
 - q_i : (slightly adjusted) proportion of pseudo-doublets in k nearest neighbors of cell i
$$q_i = \frac{k_d(i)+1}{k_{adj}+2}$$
 - Remove a cell if $q_i > c_0$ where c_0 is some threshold
 - In the paper, they defined some Bayesian likelihood L_i which is monotone increasing in q_i

Scublet (Wolock et. al. Cell Systems, 2019)

- Two key tuning parameters: k and c_0
 - k : they used an adjusted k : $k_{\text{adj}} = \text{round}(k \cdot (1+r))$ where $k = \text{round}(0.5\sqrt{\text{number of cells}})$ and $r \geq 2$ (they found this formula empirically)
 - c_0 The distribution of q_i is empirically bimodal and they define c_0 as valley between two modes

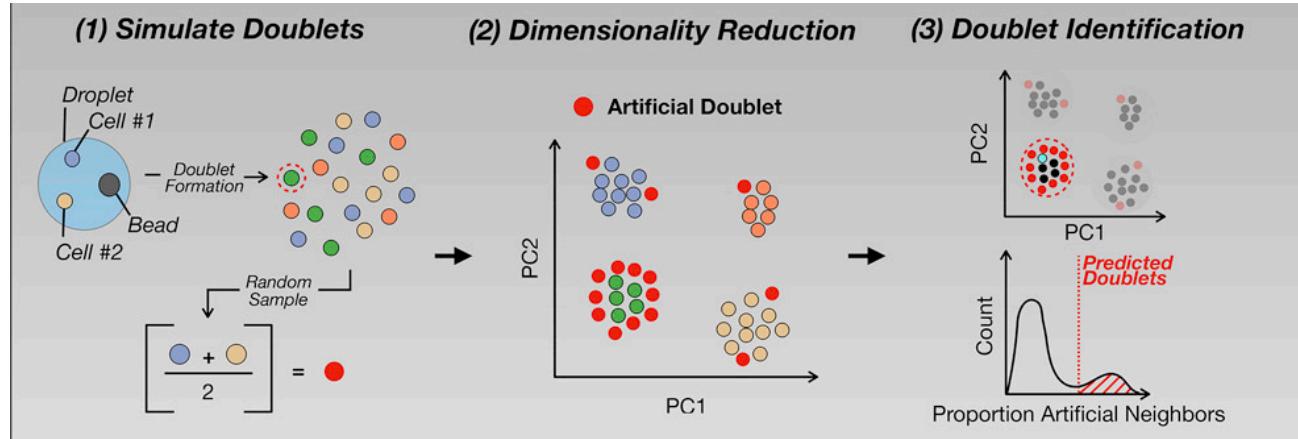


An example



DoubletFinder (McGinnis et. al. Cell Systems, 2019)

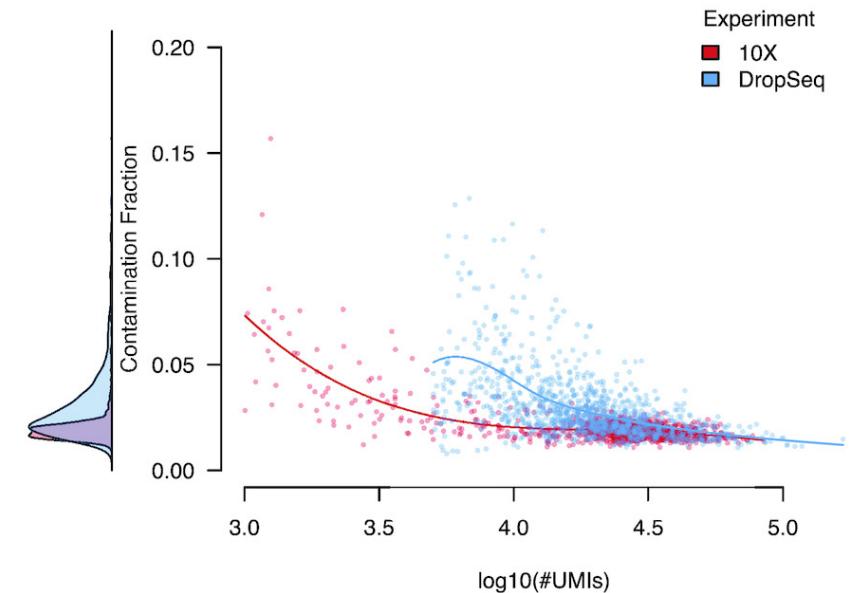
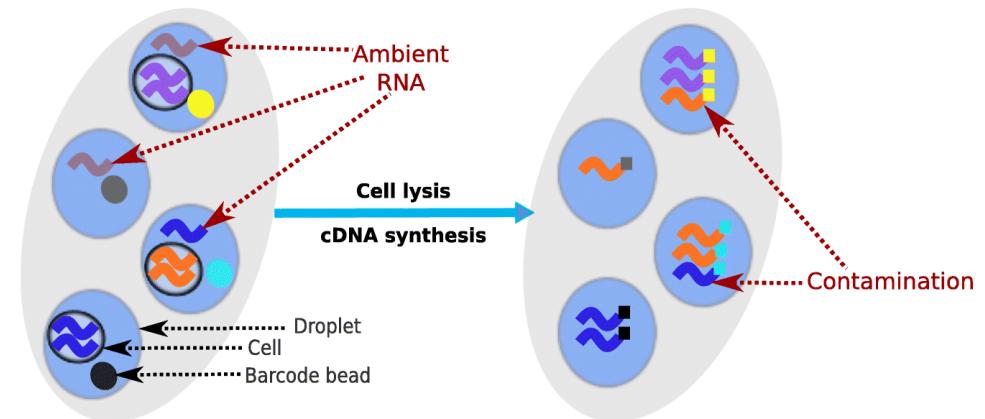
- Same idea as Scublet
 - 25% pseudo-doublets in the merged data



- Different ways to choose tuning parameters: k and c_0
 - k : choose k to maximize the bimodality coefficient of the distribution of q_i
 - Bimodality coefficient (formula from SAS)
$$BC = \frac{\gamma^2 + 1}{\kappa + \frac{3(n-1)^2}{(n-2)(n-3)}} \quad \begin{matrix} \gamma \text{ skewness,} \\ \kappa \text{ kurtosis} \end{matrix}$$
 - Not very ideal, so they used a modified version
 - c_0 : a pre-given proportion of doublets need to be detected
 - DoubletFinder performs slightly better than Scublet in a benchmarking study (Xi and Li, Cell Systems 2021)

Ambient RNA

- In Droplet-based scRNA-seq platforms, a droplet can contain isolated RNAs even if it does not contain a cell
- Ambient RNA: pool of mRNA molecules that have been released in the cell suspension
- Ambient RNA also brings contamination to droplets that contain cells
- Ratio of contaminated RNA on average can be low (~2%, less than 10%), but the contamination rate can vary greatly across cells
- Why may we separate ambient RNA from mRNAs in the cell? → empty droplets serve as negative controls



EmptyDrops (Lun et. al. Genome Biology, 2019)

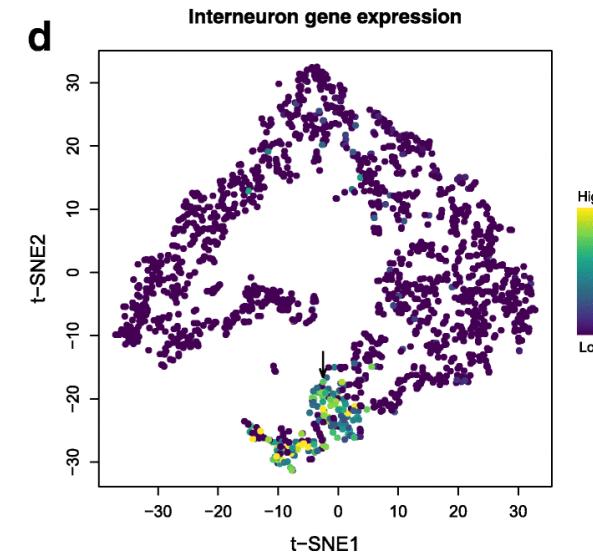
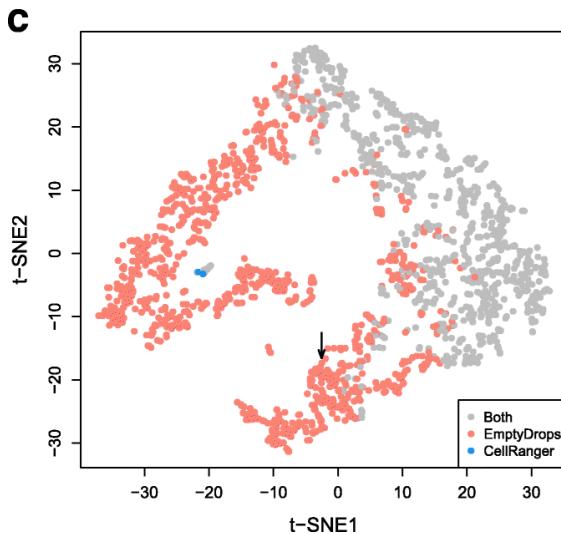
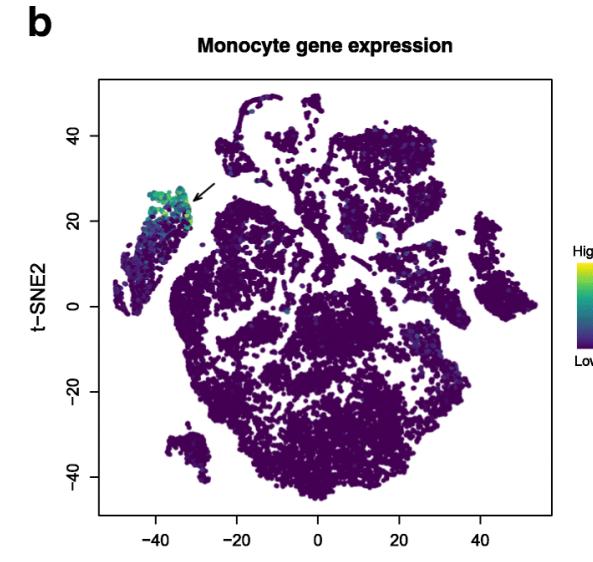
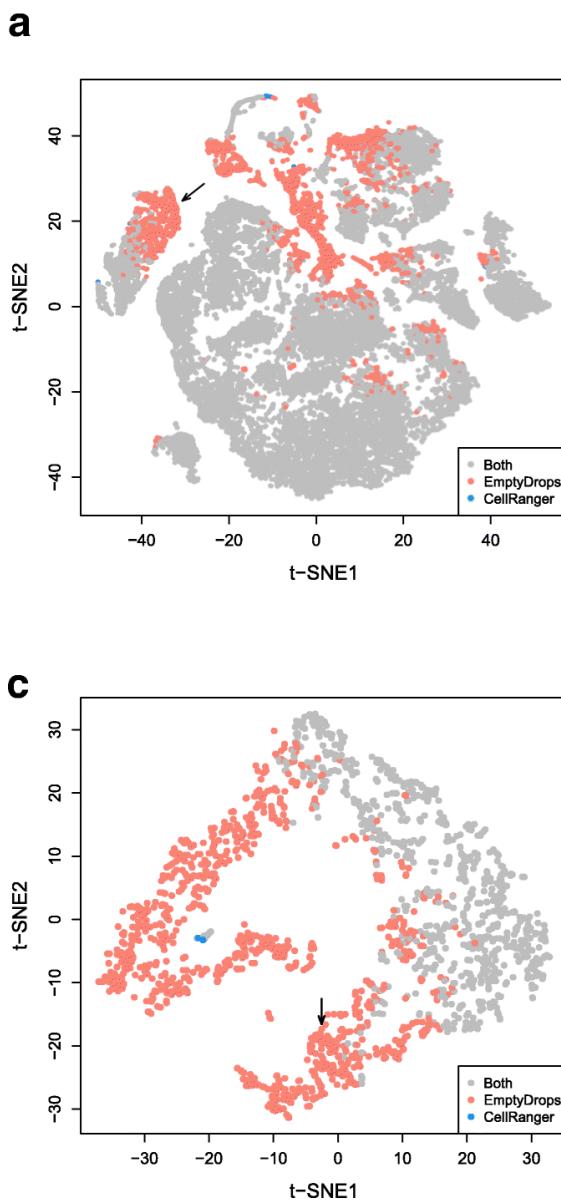
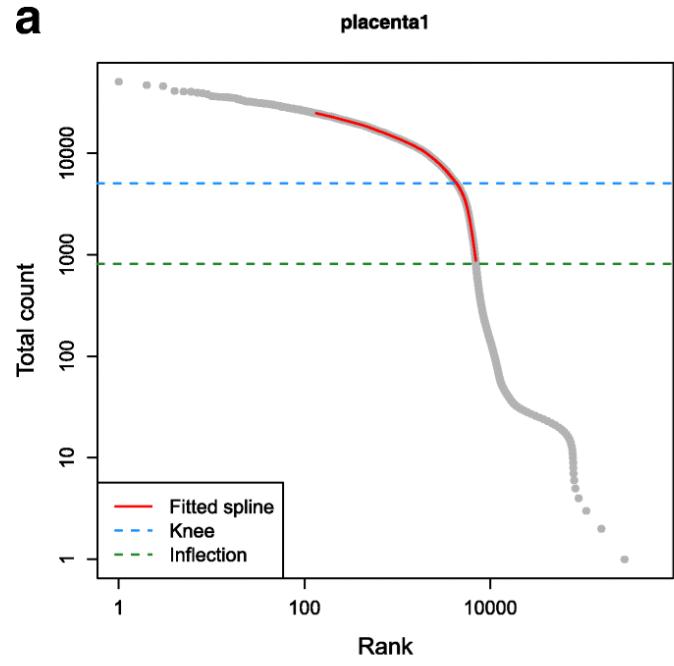
- Typically, we can identify droplets with no cells by the library size (library size too small)
- This paper argued that such method discards small cells with low RNA content
- Goal: rescue true cells with small library size
- This paper only detect empty droplets, it does not correct for ambient RNA in droplets with cells
- Core idea: find empty droplets use both the library size and gene expression profile
 - Learn an initial ambient profile
 - Estimate empty droplet gene expression distribution
 - Compute a p-value for each barcode to test whether the barcode is not an empty droplet
 - Keep barcodes as “cells” if they have small p-values or large enough library size

EmptyDrops (Lun et. al. Genome Biology, 2019)

- Estimate empty droplet gene expression distribution
 - Select barcodes whose library sizes are less than T as an initial pool of empty droplets
 - Assume that gene expressions in an empty droplet i follows
$$(Y_{1i}, \dots, Y_{Gi}) \sim \text{Dirichlet_multinomial}(l_i, (\alpha_0 \tilde{p}_1, \dots, \alpha_0 \tilde{p}_G))$$
[check Wikipedia for the definition]
 - \tilde{p}_g is obtained by some empirical Bayes estimate to avoid reaching 0
 - α_0 estimated by maximum likelihood estimation given an estimated \tilde{p}_g
- Compute p-value to test whether a barcode is not an empty droplet
 - Essentially test whether an observation comes from a known distribution
 - Basically, you check if the observation b is at the tail of the density (likelihood in the paper)
 - Monte Carlo calculation of tail probability
 - Sample N new observations from the above estimated empty droplet distribution, get the density $L_{1b}, \dots L_{Nb}$
 - Calculate p-value as proportion of $L_{1b}, \dots L_{Nb}$ that are smaller than L_b (density of b)
- Barcode selection
 - BH correction of p-values and select a barcode if library size $l_i > U$ where U is a knee point

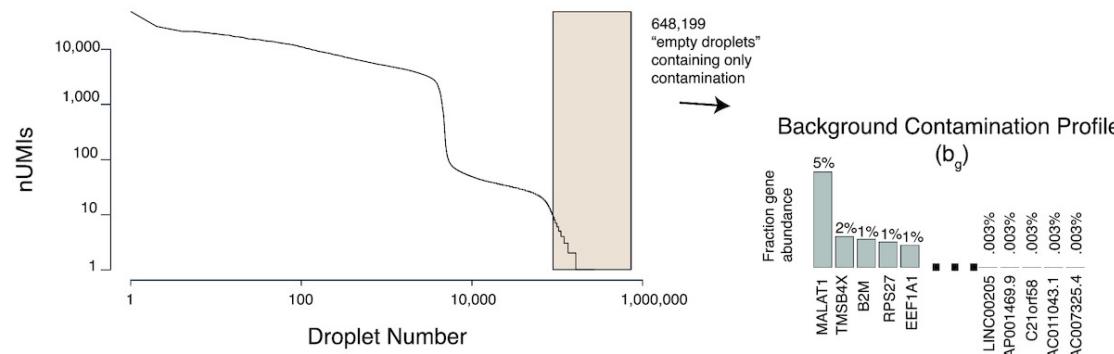
Conventional method

Some results



SoupX (Young et. al. GigaScience, 2020)

- Correct for ambient RNA confounding in cells
- Core idea:
 - Estimate ambient RNA gene expression profile from empty droplet (similar to EmptyDrops)
 1. Determine the expression profile of contamination



- Use marker genes to determine proportion of contamination in each cell
- Remove the estimated ambient RNA count for each gene from the observed counts

SoupX (Young et. al. GigaScience, 2020)

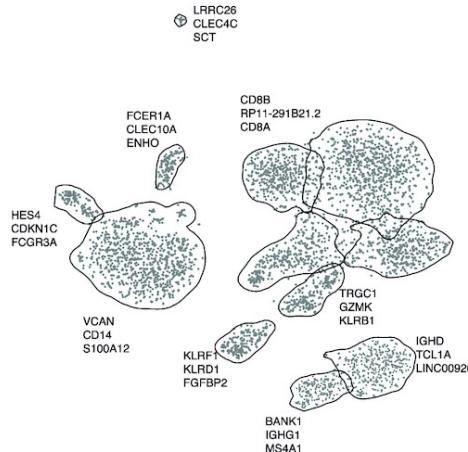
- Use marker genes to determine proportion of contamination in each cell

$$Y_{gc} = m_{gc} + o_{gc}$$

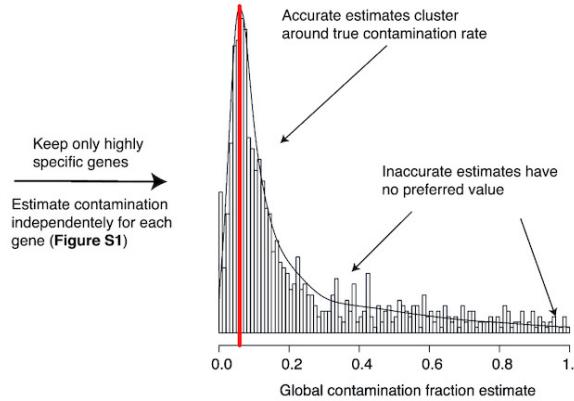
- $o_{gc} = l_c \rho_c b_g$: ρ_c contamination rate in each cell
- “Negative control” genes
 - Assume that the marker genes for one cell cluster has zero expression in other cells
- If gene g is a negative control for the cell, then $m_{gc} = 0$ and $Y_{gc}/(l_c b_g) \approx \rho_c$
- Estimate ρ_c as the mode of the gene-specific estimated rates

2. Estimate or set the global contamination rate

2.1 Marker genes for each cluster identified

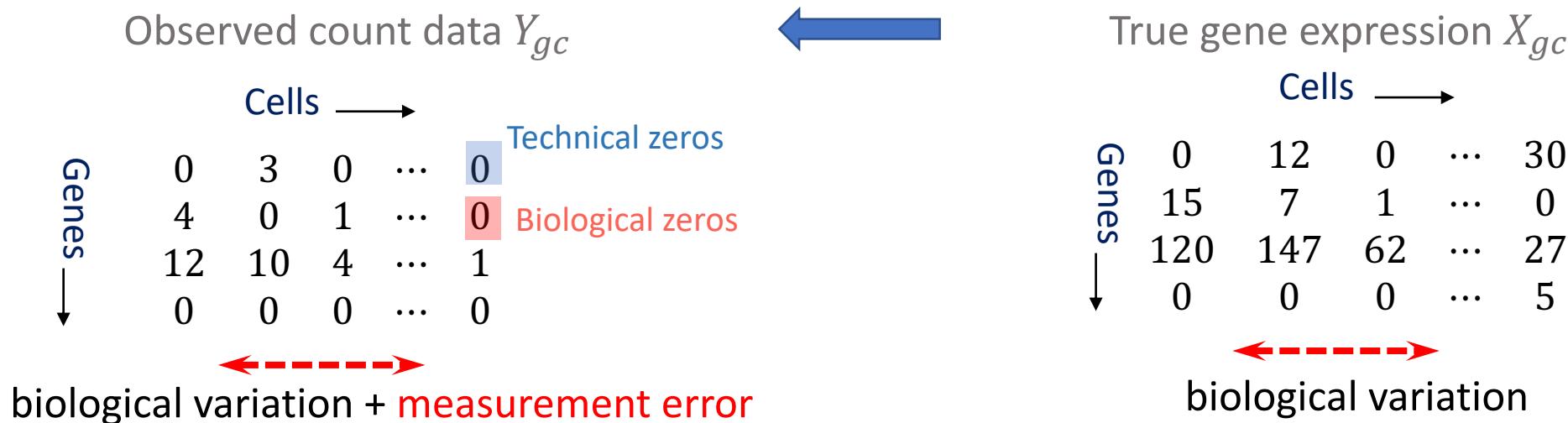


2.2 Set contamination to most common estimate



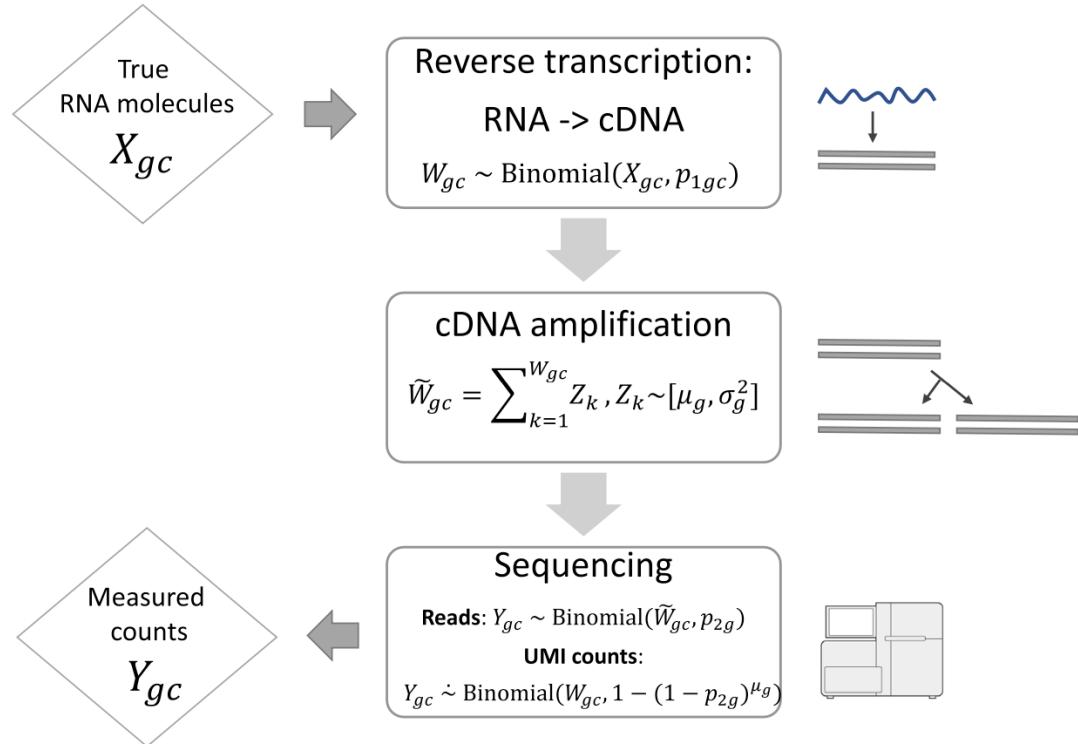
- Some adjustments to provide a good estimate of o_{gc} (need to be an integer, no greater than Y_{gc})

scRNA-seq count matrix is very noisy



- Observed count matrix Y is typically **extremely sparse**
 - About 99% of the entries are zeros
 - Two types of zeros
 - Biological zeros: true mRNA count is zero
 - Technical zeros: true mRNA count is not zero, but observed count is zero
 - Dropouts are not missing at random!

Measurement error distribution



- 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_{2gc} : 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}

		Cells →					
		0	3	0	...	0	Technical zeros
		4	0	1	...	0	Biological zeros
		12	10	4	...	1	
		0	0	0	...	0	

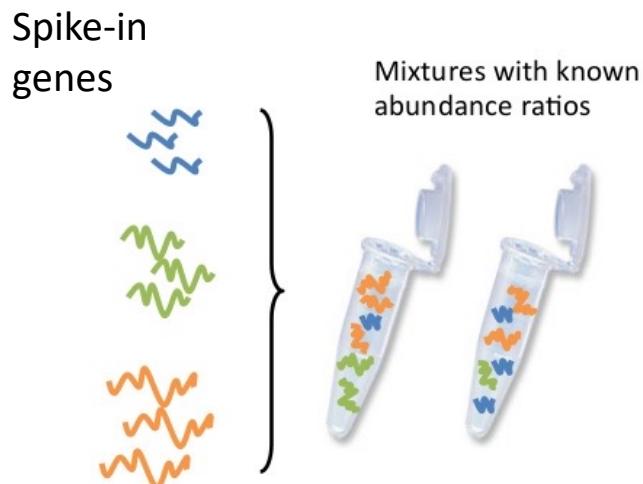
Genes ↓

←-----→

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):



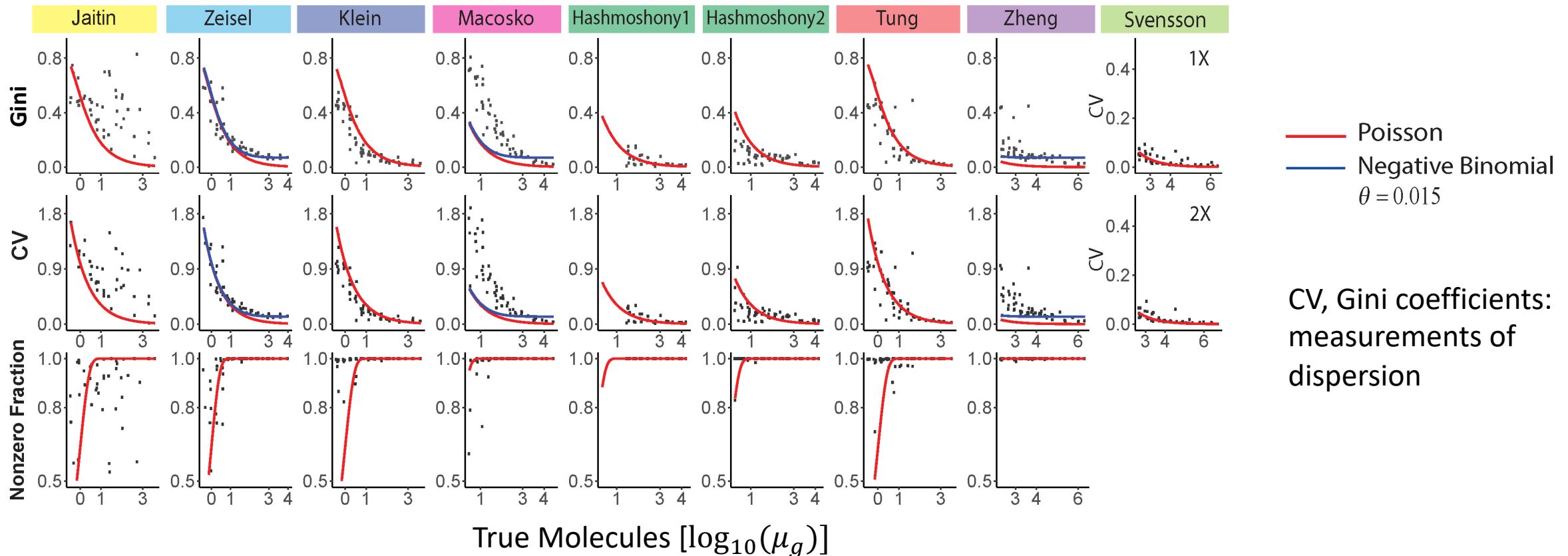
- $X_{gc} \stackrel{i.i.d}{\sim} \text{Poisson}(\mu_g)$ Known
- Conventionally, researchers treat X_{gc} as constant across cells
$$\text{Var}(Y_{gc}) = 2\alpha_c \gamma_g \mu_g$$
- Assume $\gamma_g = 1$, then α_c is identifiable

Noise distribution for UMI data is not zero-inflated

- Some empirical evidence using ERCC spike-ins

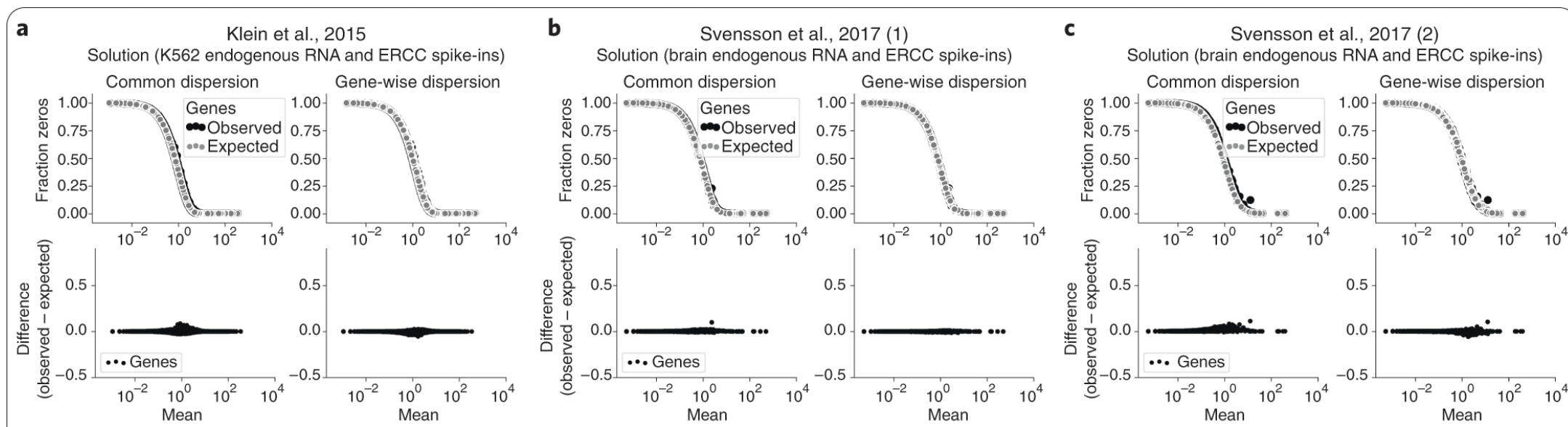
- (Wang et. al. PNAS 2018):

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



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 $Y_{gc} \sim \text{NB}(\mu_g, \theta_g)$
check if the observed zero proportion match with the estimated values

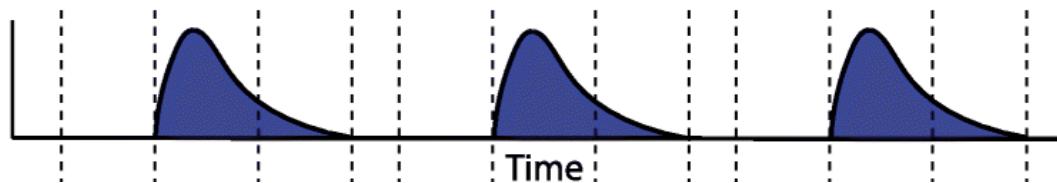
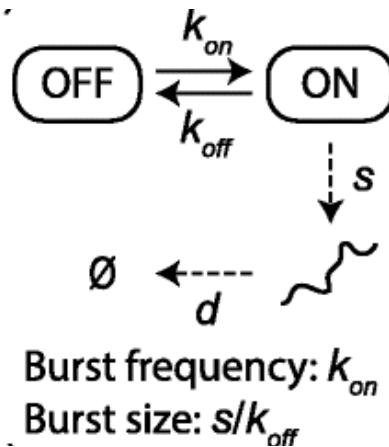


Factors affecting the noise distribution

- 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



- 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 $Y_{gc} \sim \text{Poisson}(l_c p_{gc})$

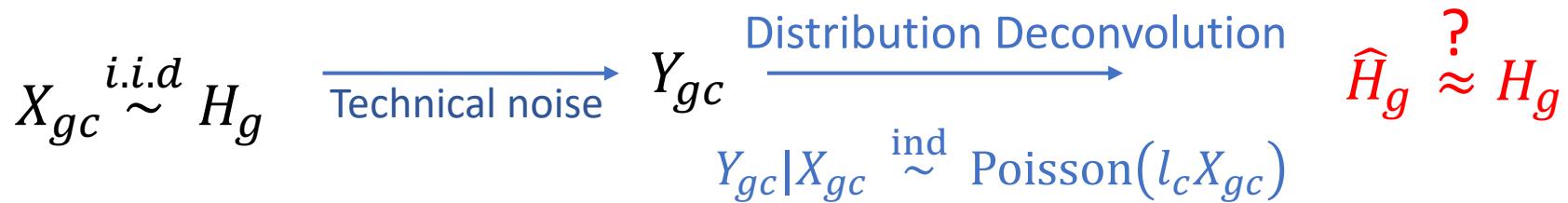
Expression model	Observation model	Method
Point mass (no variation)	Poisson	Analytic
Gamma	Negative Binomial	MASS ⁴¹ , edgeR ⁴² , DESeq2 ⁴³ , BASICS ⁴⁴ , SAVER ²⁰
Point-Gamma	Zero-inflated Negative Binomial	PSCL ⁴⁵
Unimodal (non-parametric)	Unimodal	ashr ^{24,46}
Point-exponential family	Flexible	DESCEND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



- 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 $x \in \boldsymbol{x} = (x_1, \dots, x_m)$

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

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

Single cell RNA sequencing (Drop-seq)

~8,000 cells
~12,000 genes

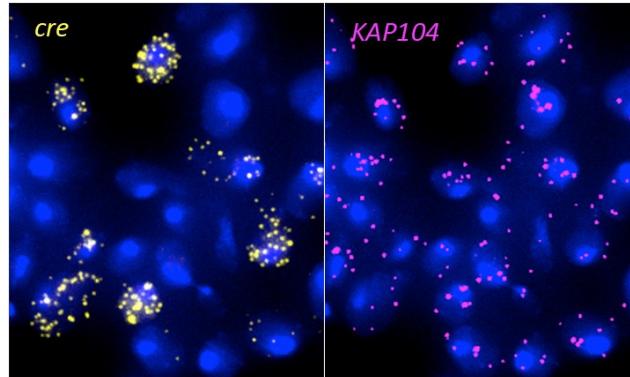
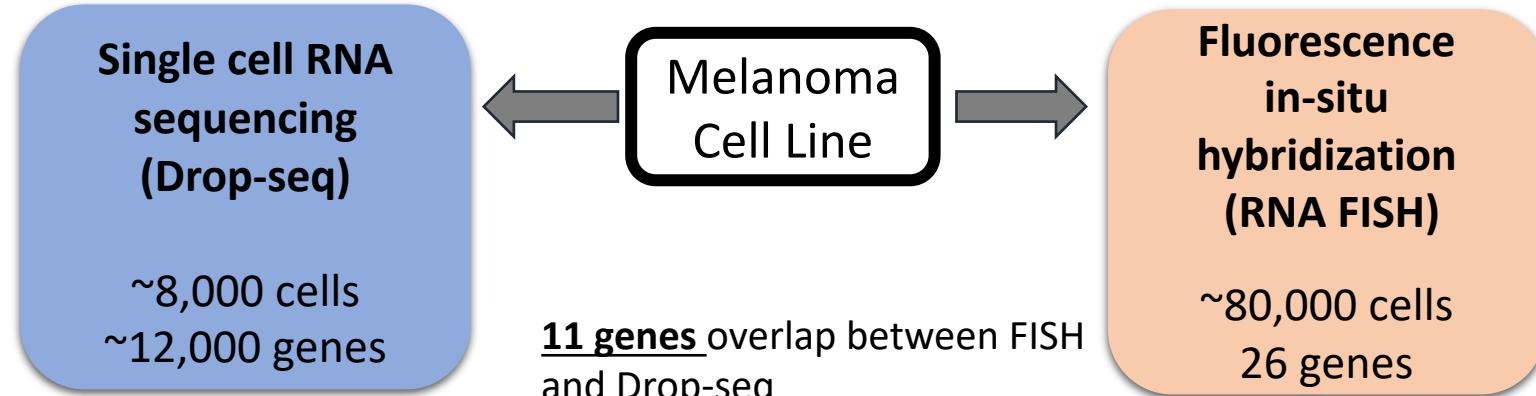
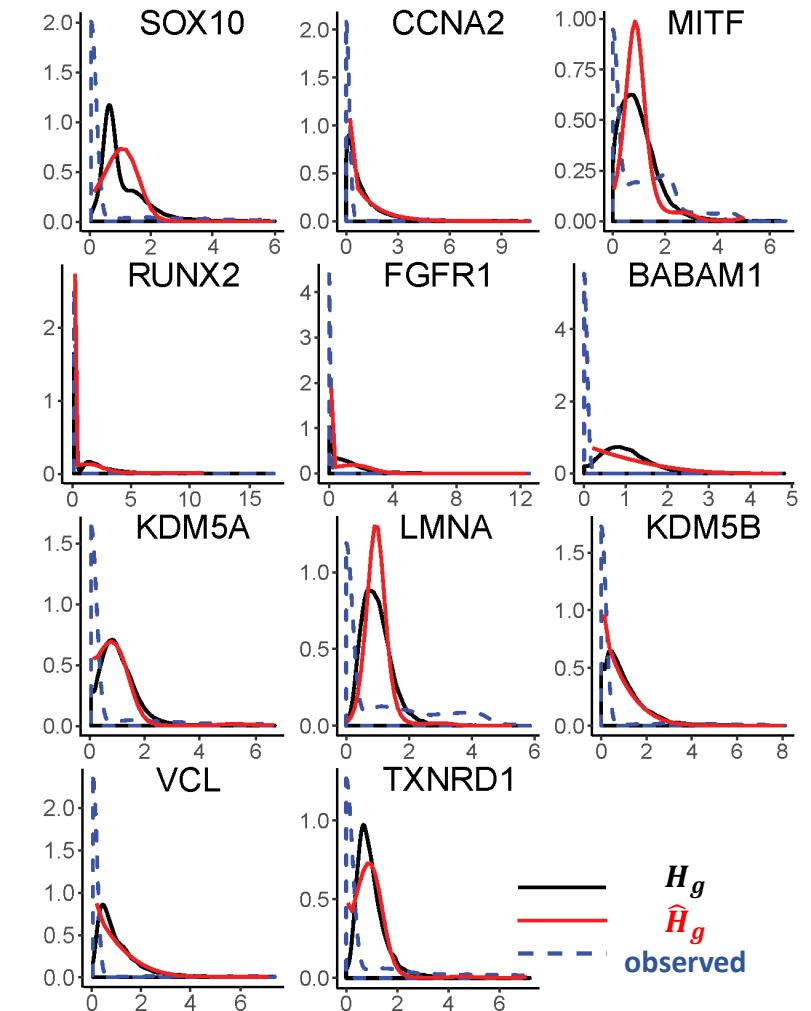


Photo courtesy of Anne Dodson and Professor Jasper Rine

\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 negative-binomial distribution
 - (Kim et. al. Genome Biology 2020) showed that Poisson distribution is good enough to model Y_{gc} for a relatively homogenous cell population
- A common approach is to use an autoencoder to capture gene-gene dependence and use Negative Binomial / Poisson likelihood to construct loss function (will discuss in later lectures)

Related papers

- Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 8(4), 281-291.
- McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell systems*, 8(4), 329-337.
- Lun, A. T., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., Participants in the 1st Human Cell Atlas Jamboree, & Marioni, J. C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome biology*, 20, 1-9.
- Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience*, 9(12), giaa151.
- 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.