Simulation of Single Cell Data

Stephanie DeGraaf, Hector Roux de Bézieux

Introduction

Single-cell RNA sequencing data presents unique challenges for analysis:

- Dropouts, high technical noise, high variability

Many new methods are rapidly being developed:

- 200+ software packages have been created specifically for scRNA-seq data

Simulations are necessary to test performance and prove merits of new methods:

- Often, simulations are poorly documented or not similar to real scRNA data

Introduction

Methods are needed to generate reproducible and accurate simulated datasets.

- BASiCS *(Vallejos C, Marioni J, Richardson S)*

- Splat *(Zappia L, Phipson B, Oshlack A)*

Basics: Bayesian Analysis of Single-Cell Sequencing Data

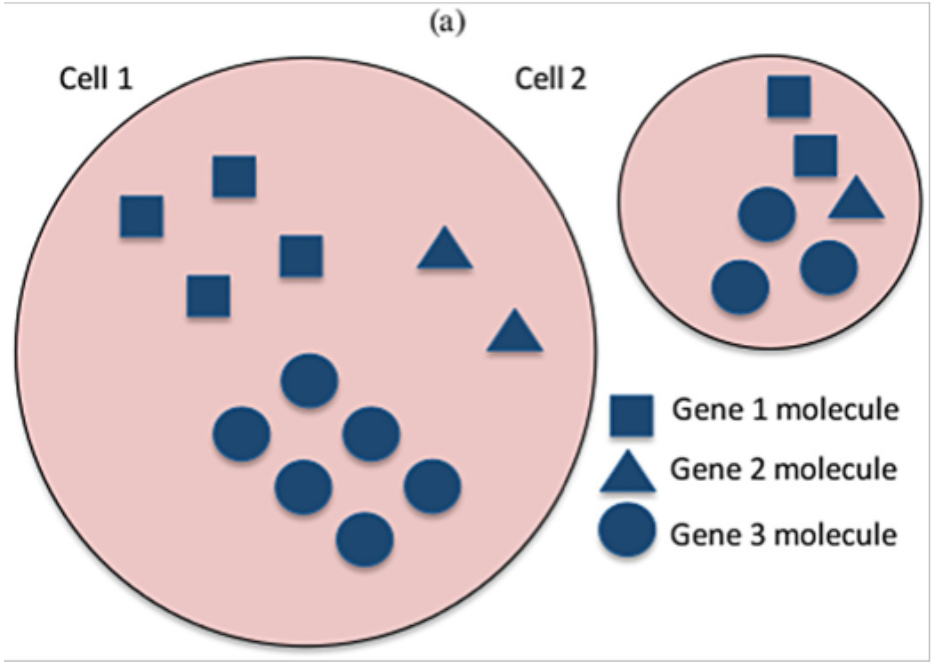
Catalina A. Vallejos John C. Marioni, Sylvia Richardson

- Not initially a simulation method: build to separate technical noise and biological signal

- Build a hierarchical bayesian model

- Relies on spike-in molecules

Spike-in molecules



After normalization, you might think both cells are identical...

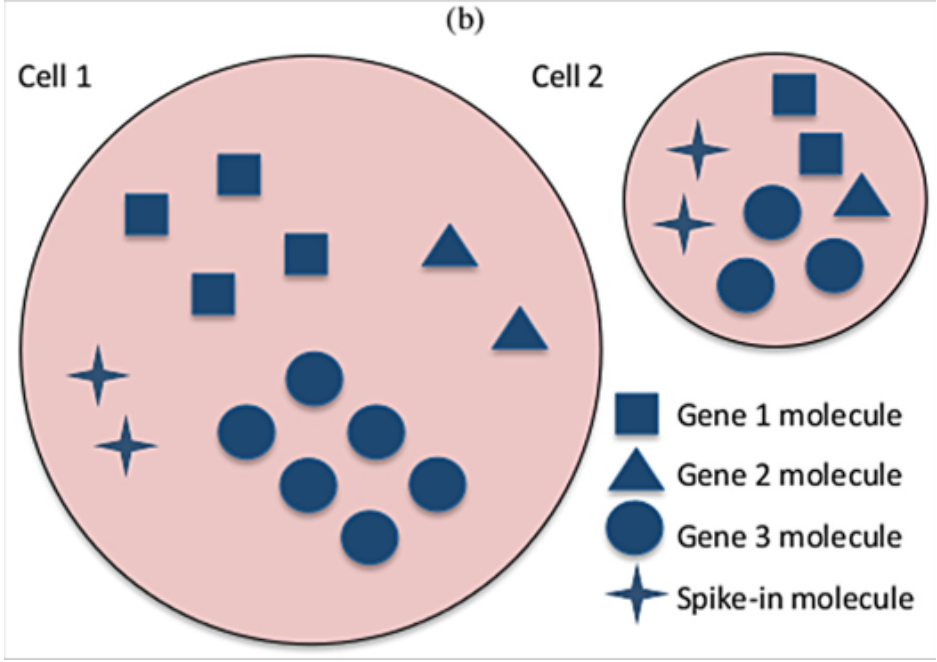
But they are not

Spike-in molecules

- Add spike-in molecules and normalize with them!

Most common is the set of 92 extrinsic molecules derived by the External RNA Controls Consortium (ERCC)

They vary in length and concentration to mimic biological eukaryotic mRNA.



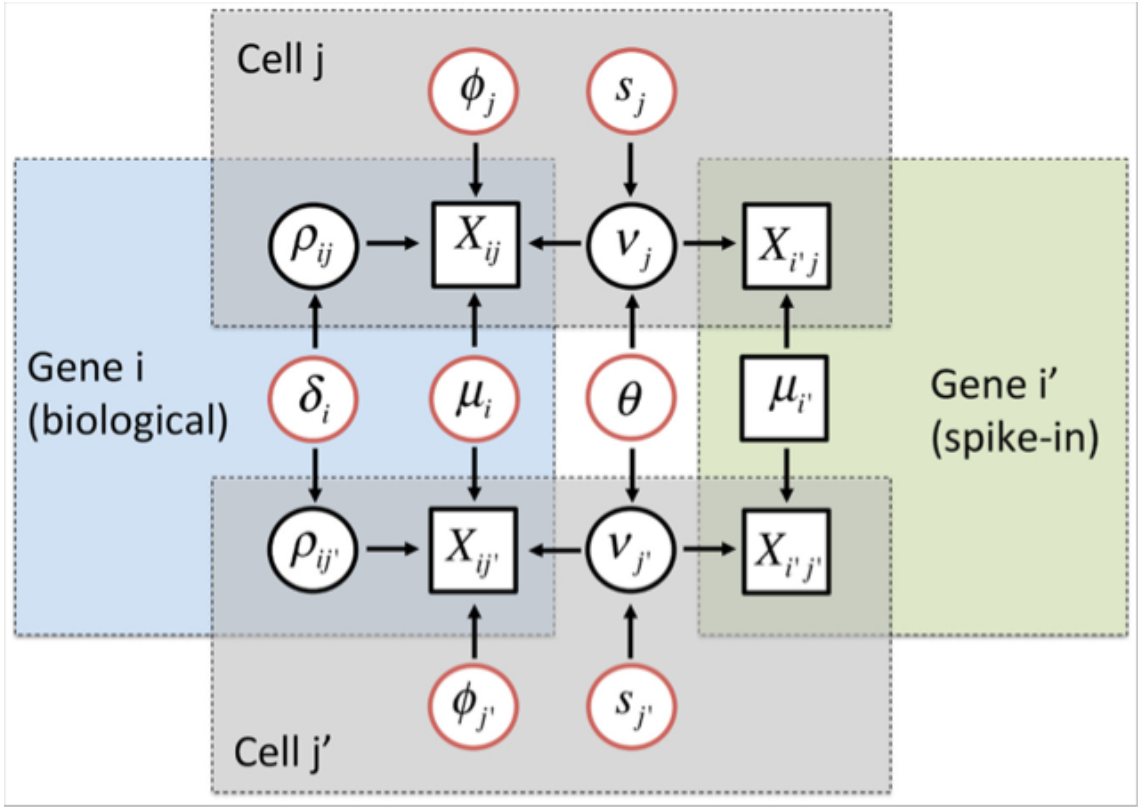
Hierarchical model: overview

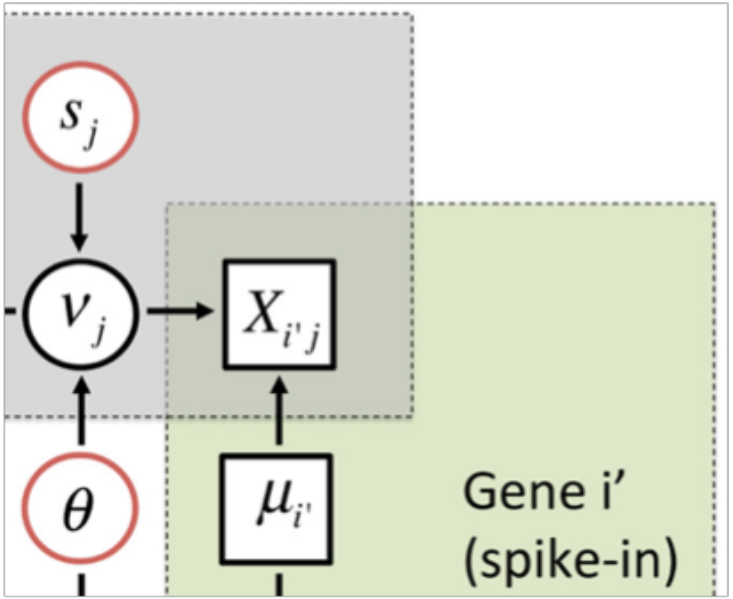
known

unknown

fixed

random





Hierarchical model for spike-in

$$X_{ij} | \mu_i, \nu_j \overset{\text{ind}}{\sim} \text{Poisson}(\nu_j \mu_i)$$

$$\nu_j | s_j, \theta \overset{\text{ind}}{\sim} \text{Gamma}(1/\theta, 1/(s_j \theta))$$

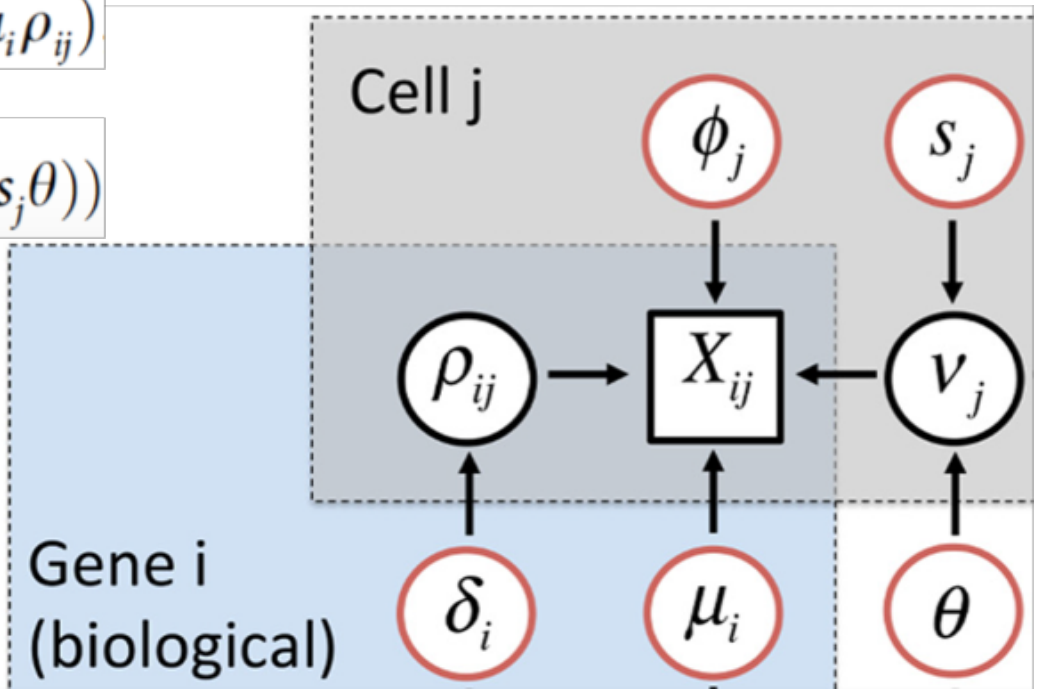
- X_{ij} Count for gene i in cell j
- μ_i Mean concentration of the spike-in i
- ν_j Random effect that fluctuates around the capture efficiency normalising constant sj

Hierarchical model for biological genes

$$X_{ij} | \mu_i, \phi_j, v_j, \rho_{ij} \overset{\text{ind}}{\sim} \text{Poisson}(\phi_j v_j \mu_i \rho_{ij})$$

$$\text{with } v_j | s_j, \theta \overset{\text{ind}}{\sim} \text{Gamma}(1/\theta, 1/(s_j \theta))$$

$$\text{and } \rho_{ij} | \delta_i \overset{\text{ind}}{\sim} \text{Gamma}(1/\delta_i, 1/\delta_i)$$



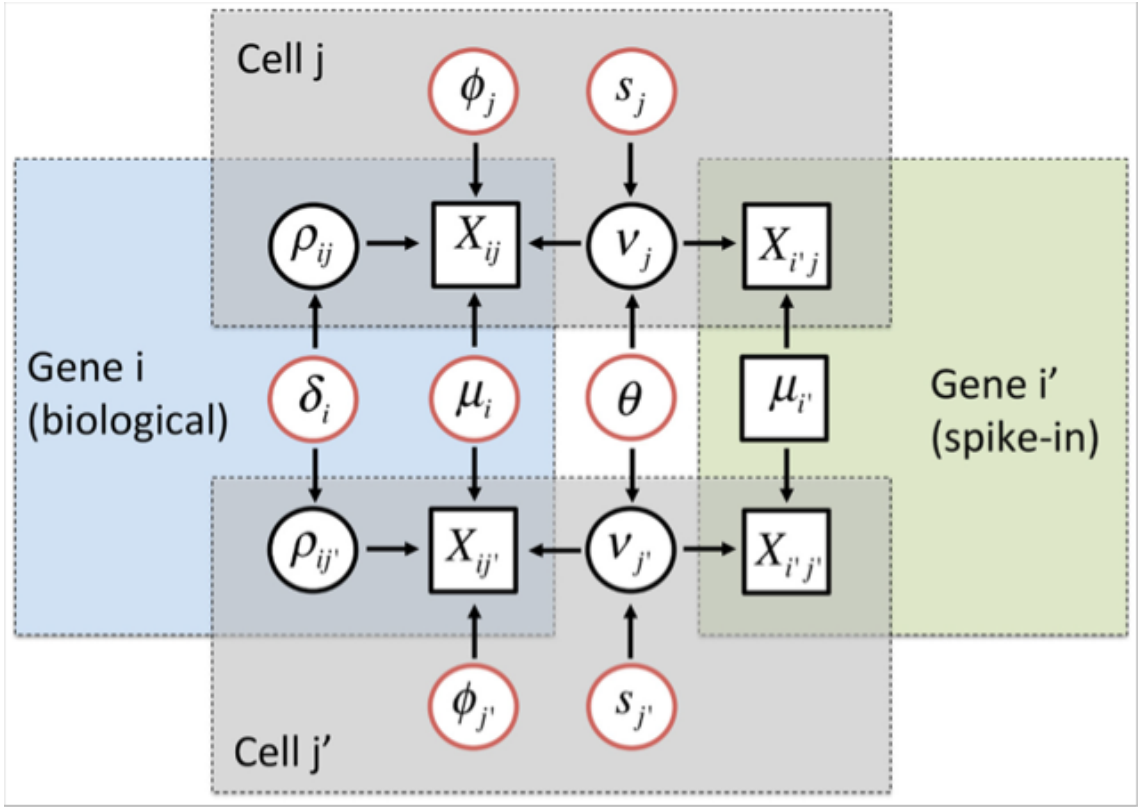
Hierarchical model: recap

known

unknown

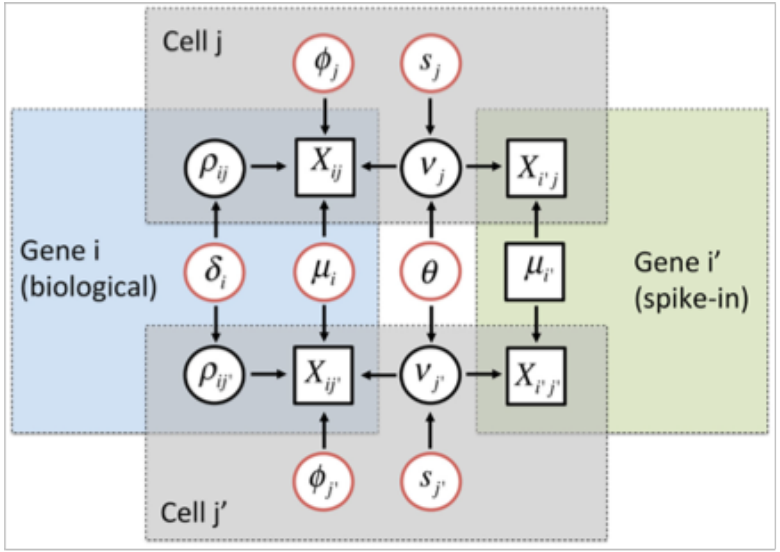
fixed

random



Hierarchical model: framework

- Go from a real dataset
- The μ for the spike-in are known
- Estimate the sj and Θ
- Estimate the δ and φ*μ for the cells and genes
- Simulate more data from the parameters



Basics

Pros:

- Clear interpretation of model
- Clear interpretation of parameters
- Easy to build-upon
- Good results based on real datasets (see after)

Cons:

- A lot of modeling assumptions
- Small positive bias for lowly expressed genes.
- Rely on spiked-end which have some specific bias.
- Same technical variability for all genes, independent of GC content, expression levels,...

Splat simulation

Splat simulation captures features of real scRNA-Seq data

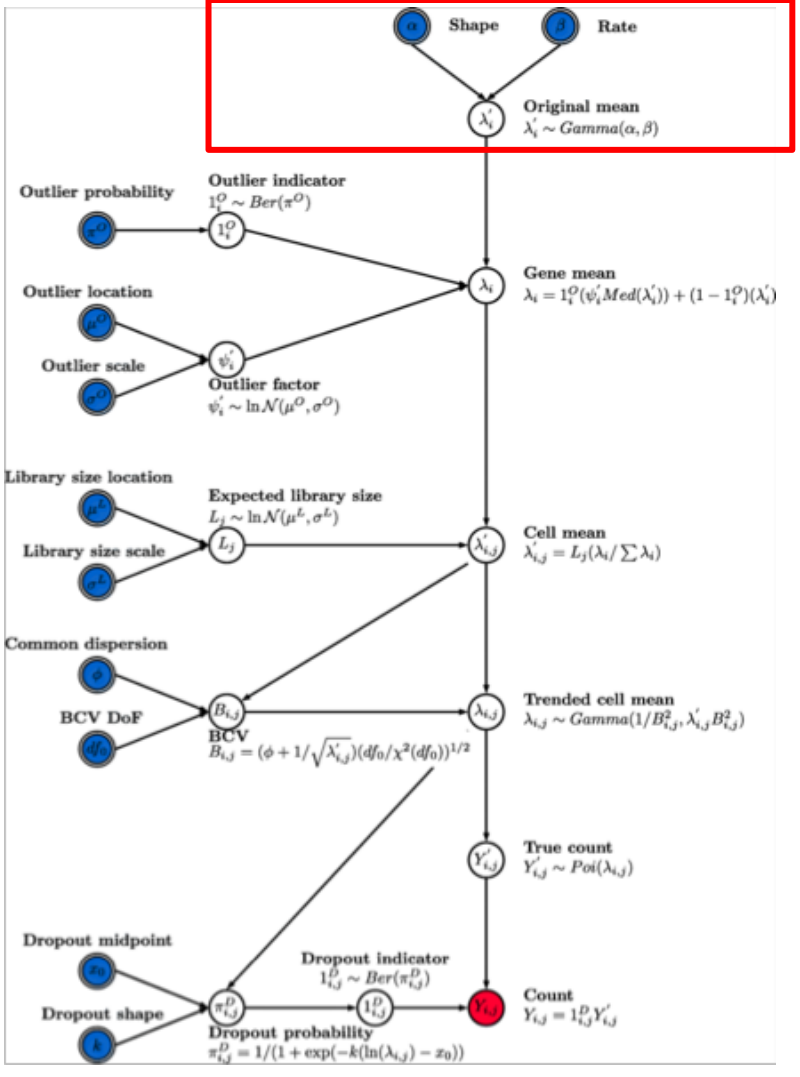
- High expression outlier genes
- Differing sequencing depths (library sizes)
- Trended gene-wise dispersion
- Zero-inflation

Splat uses a gamma-Poisson hierarchical model with hyper-parameters estimated from real data

Splat simulation procedure

Step 1: Generate original means

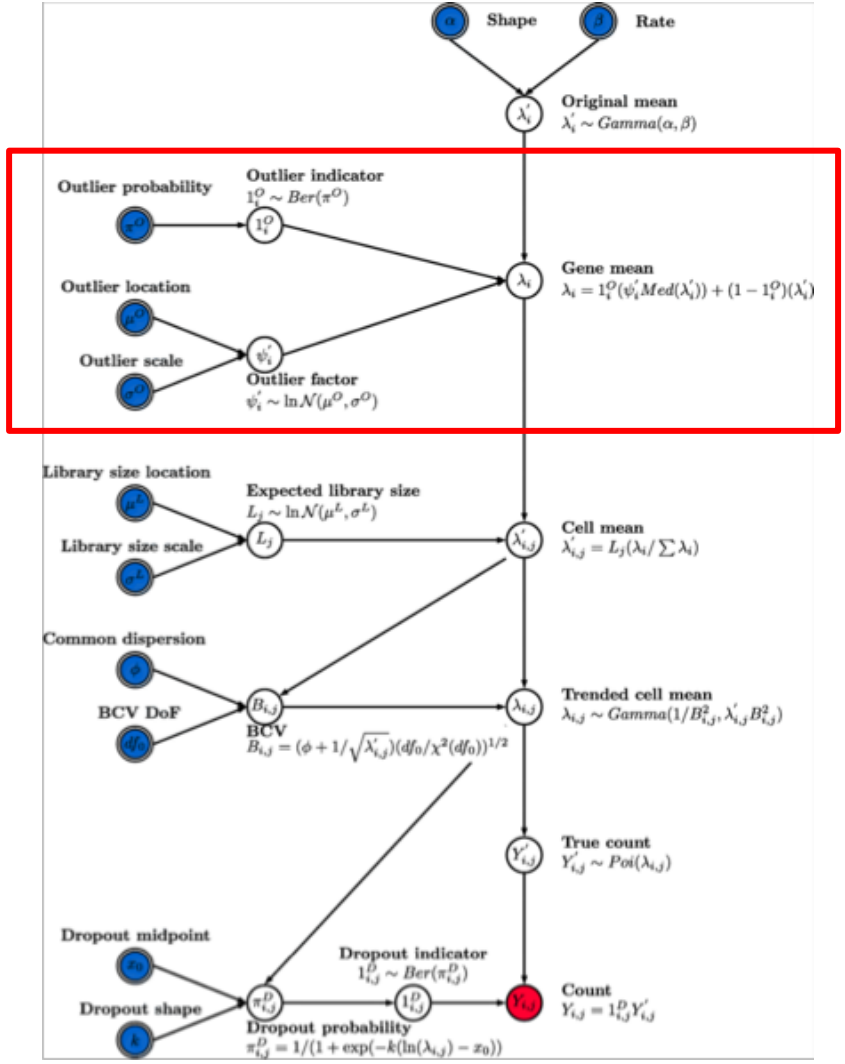
- Gamma distribution with shape parameter alpha, rate parameter beta



Splat simulation procedure

Step 2: Add in outlier counts

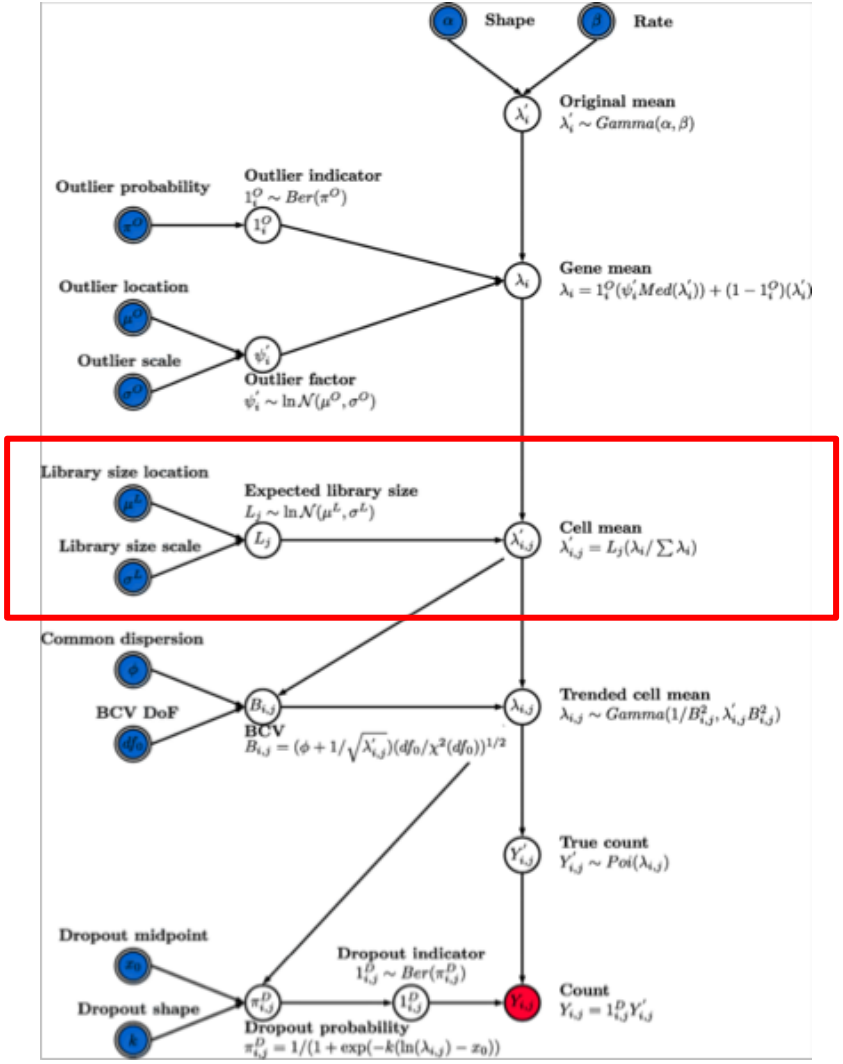
- Gamma distribution does not always capture extreme expression levels
- Specify a probability of outlier
- Add in outliers by replacing the previous mean by an inflated mean
- Determine the inflation factor by sampling from a log-normal distribution



Splat simulation procedure

Step 3: Account for library size

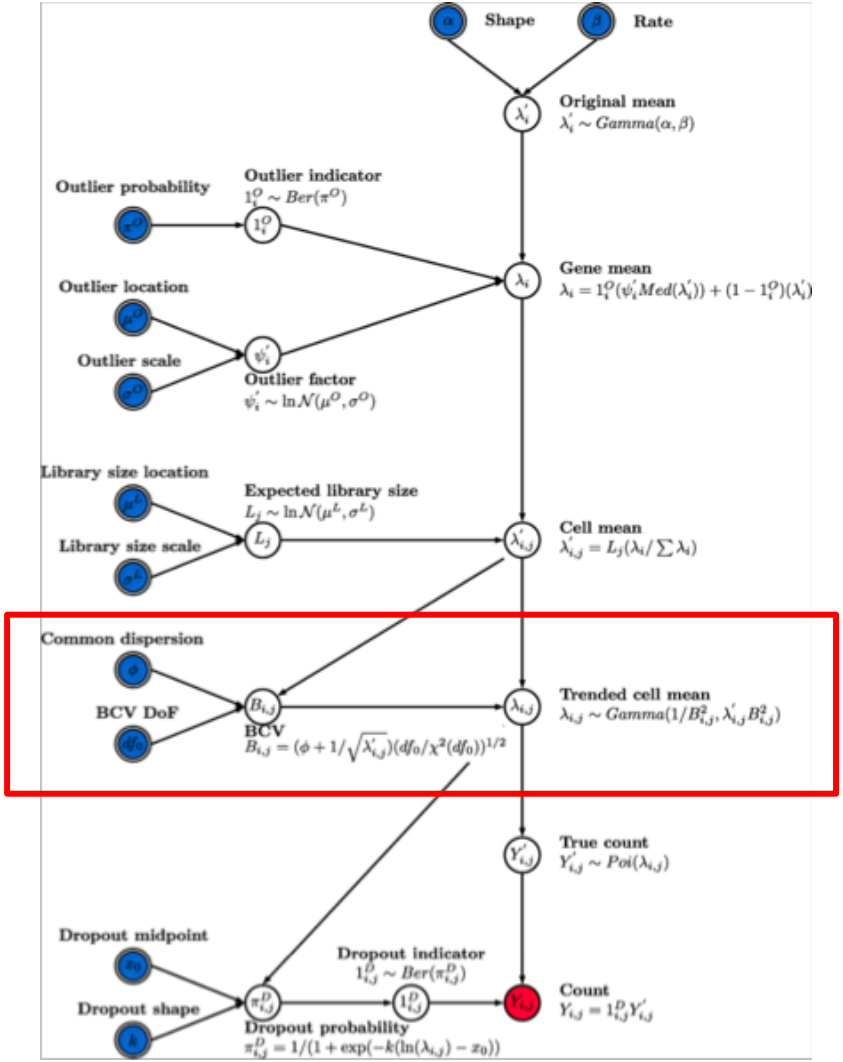
- Library sizes vary within an experiment and between experiments depending on sequencing depth
- Model library sizes with a log-normal distribution
- Proportionally adjust gene means for each cell (independent of underlying gene expression levels)



Splat simulation procedure

Step 4: Enforce mean-variance trend

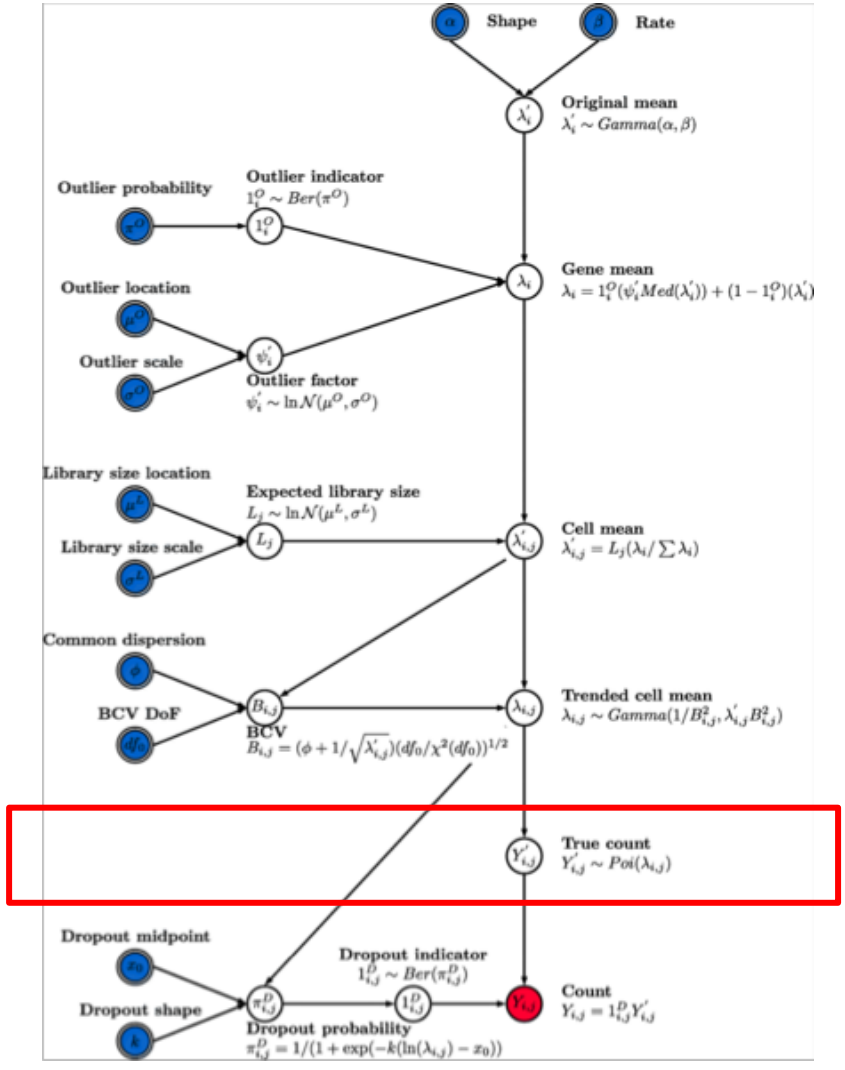
- Mean-variance trend: lowly expressed genes are more variable, highly expressed genes are more consistent
- Simulate the biological coefficient of variation (BCV) for each gene from a scaled inverse chi-squared distribution, where the scaling factor is a function of the gene mean
- Generate new means from a Gamma distribution with parameters dependent on the BCVs and previous means



Splat simulation procedure

Step 5: Generate a count matrix

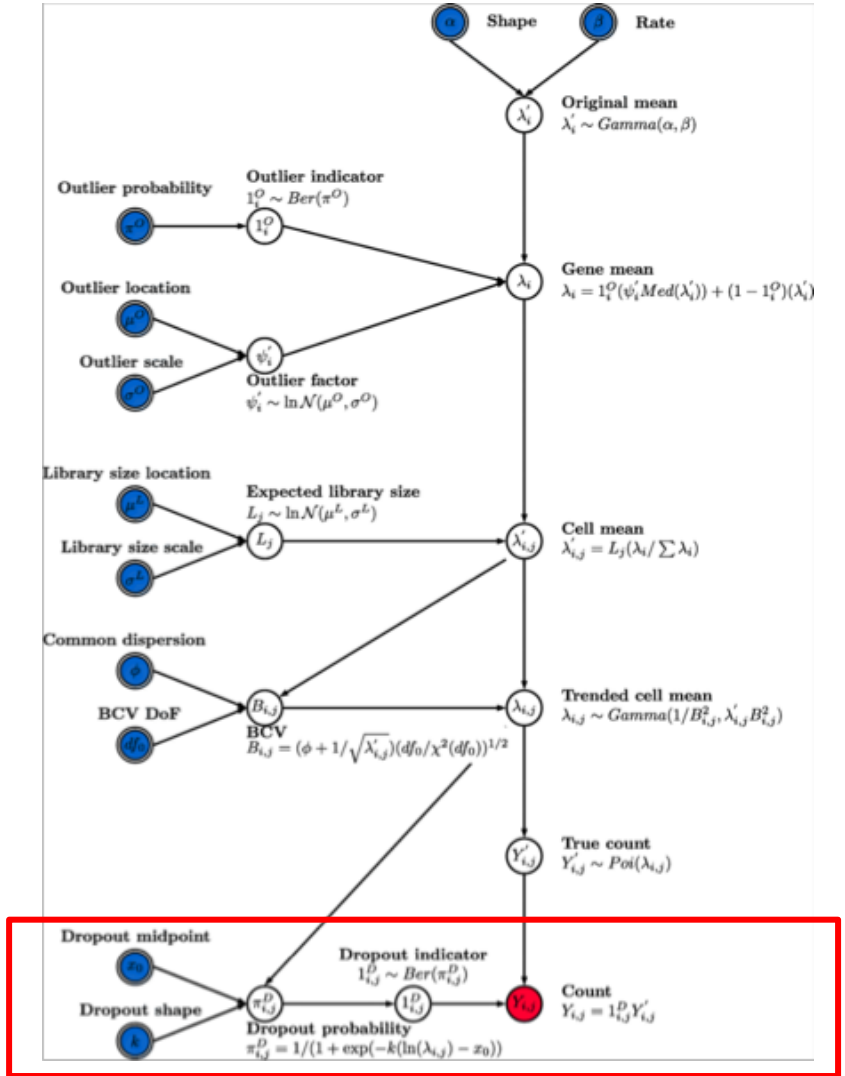
- Sample from a Poisson distribution using the adjusted means



Splat simulation procedure

Step 6: Model technical dropouts

- A key feature of scRNA-seq data is the high proportion of zeros
- One cause is technical dropout
- Generate dropout probabilities with a logistic model based on each gene's mean and proportion of zero counts
- Randomly replace some simulated counts with zeros based on a Bernoulli random variable with that gene's probability of zero



Splat simulation input parameters

Name	Symbol	Description
Mean shape	α	Shape parameter for the mean gene expression gamma distribution
Mean rate	β	Rate parameter for the mean gene expression gamma distribution
Library size location	μ^L	Location parameter for the library size log-normal distribution
Library size scale	σ^L	Scale parameter for the library size log-normal distribution
Outlier probability	π^O	Probability that a gene is an expression outlier
Outlier location	μ^O	Location parameter for the expression outlier factor log-normal distribution
Outlier scale	σ^O	Scale parameter for the expression outlier factor log-normal distribution
Common BCV	φ	Common BCV dispersion across all genes
BCV degrees of freedom	df	Degrees of freedom for the BCV inverse chi-squared distribution
Dropout midpoint	x_0	Midpoint for the dropout logistic function
Dropout shape	k	Shape of the dropout logistic function

Splat

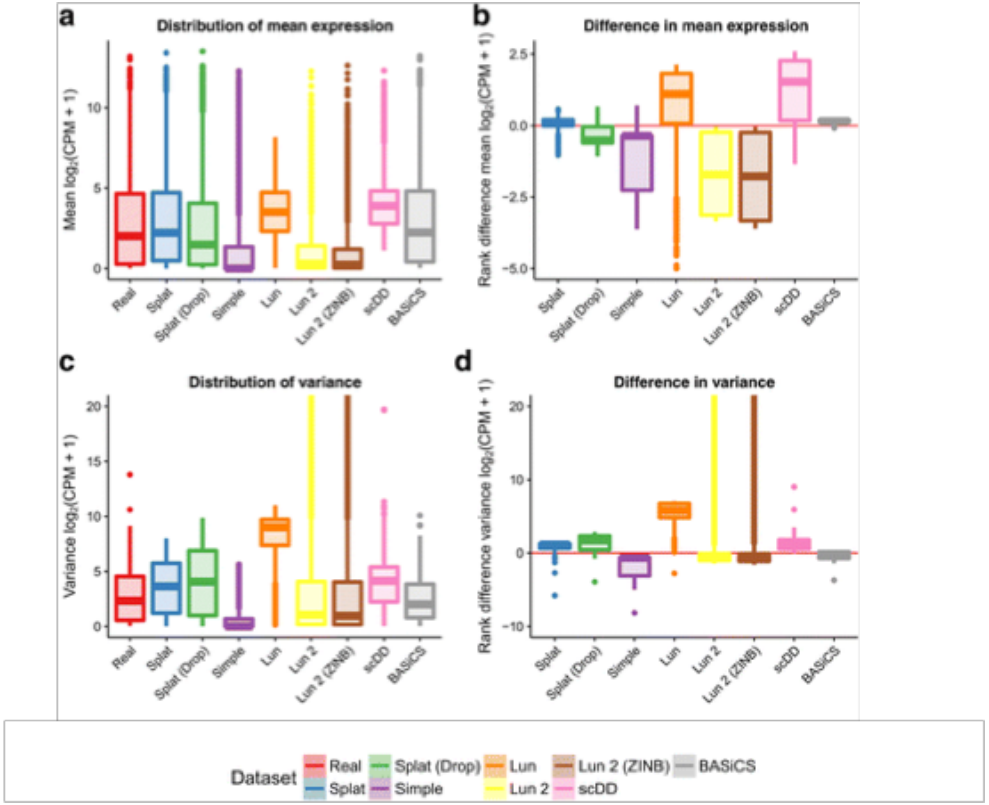
Pros:

- Clear interpretation of model
- Clear interpretation of parameters
- Easy to build-upon
- Good results based on real datasets
- Fast processing time compared to BASiCS (1 min vs 1 day)

Cons:

- A lot of modeling assumptions
- A lot of parameters estimated from data
- Modeling zero inflation is challenging

Comparison of BASiCS and Splat (and others) versus real

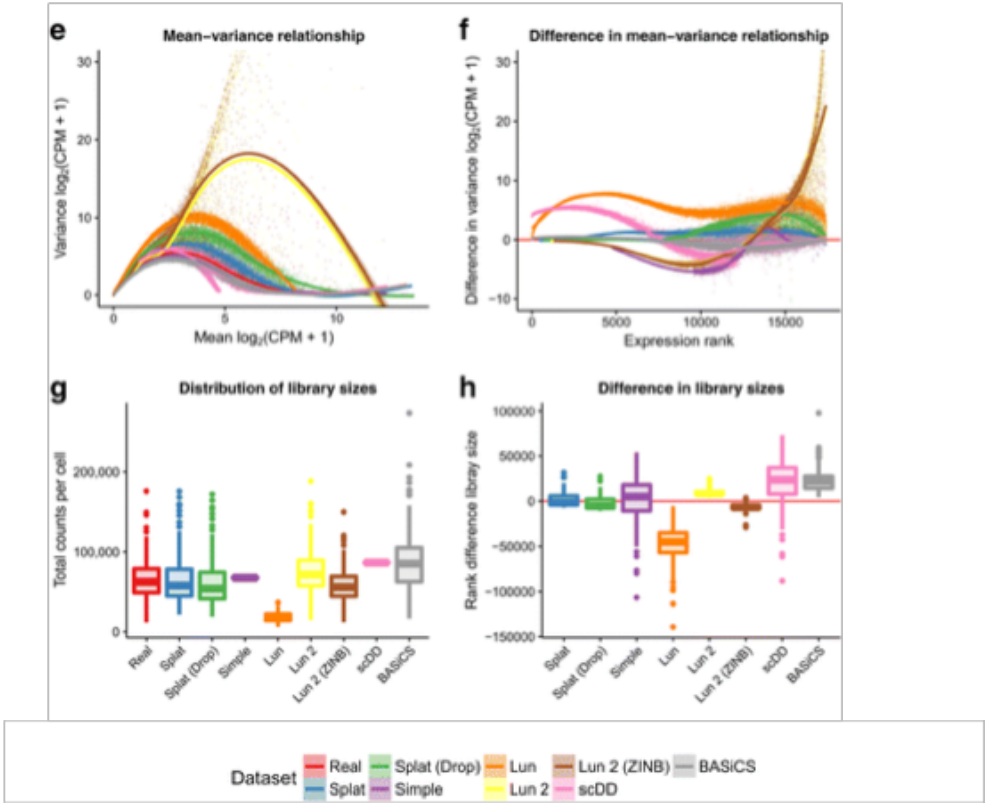


Means and Variances

Splat and BASiCS match the real data well

Other simulations miss lowly expressed genes or are skewed too heavily towards lower expressed genes

Comparison of BASiCS and Splat (and others) versus real



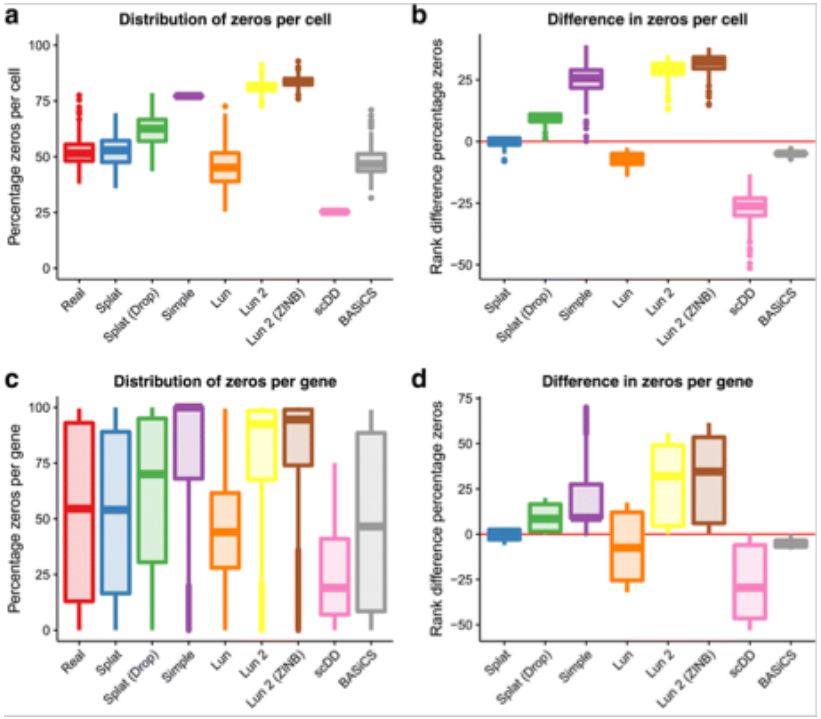
Mean-Variance Relationship

Splat and BASiCS match the mean-variance relationship well

Library Sizes

BASiCS simulation produces too many large library sizes

Comparison of BASiCS and Splat (and others) versus real



Distribution of Zeros

Splat and BASiCS match the real data well

Including the dropout probability in the Splat model overestimates the number of zeros

Summary

- Variability in scRNA-seq data arises from a complex interaction of biological and technical factors
- Simulating accurate scRNA-seq datasets is important
- Splat and BASiCS are both better options than many others
- No simulation method can accurately reproduce all datasets
- Development of analysis methods should rely on simulations that closely match datasets in an appropriate setting for the method

Citations

Vallejos, C. A., Marioni, J. C., & Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS computational biology*, *11*(6), e1004333.

Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, *18*(1), 174.

Thank you for listening

Summary

- Variability in scRNA-seq data arises from a complex interaction of biological and technical factors
- Simulating accurate scRNA-seq datasets is important
- Splat and BASiCS are both better options than many others
- No simulation method can accurately reproduce all datasets
- Development of analysis methods should rely on simulations that closely match datasets in an appropriate setting for the method