

***muscat* detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data**

Helena L. Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra & Mark D. Robinson

Nature Communication (2020)

Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Introduction

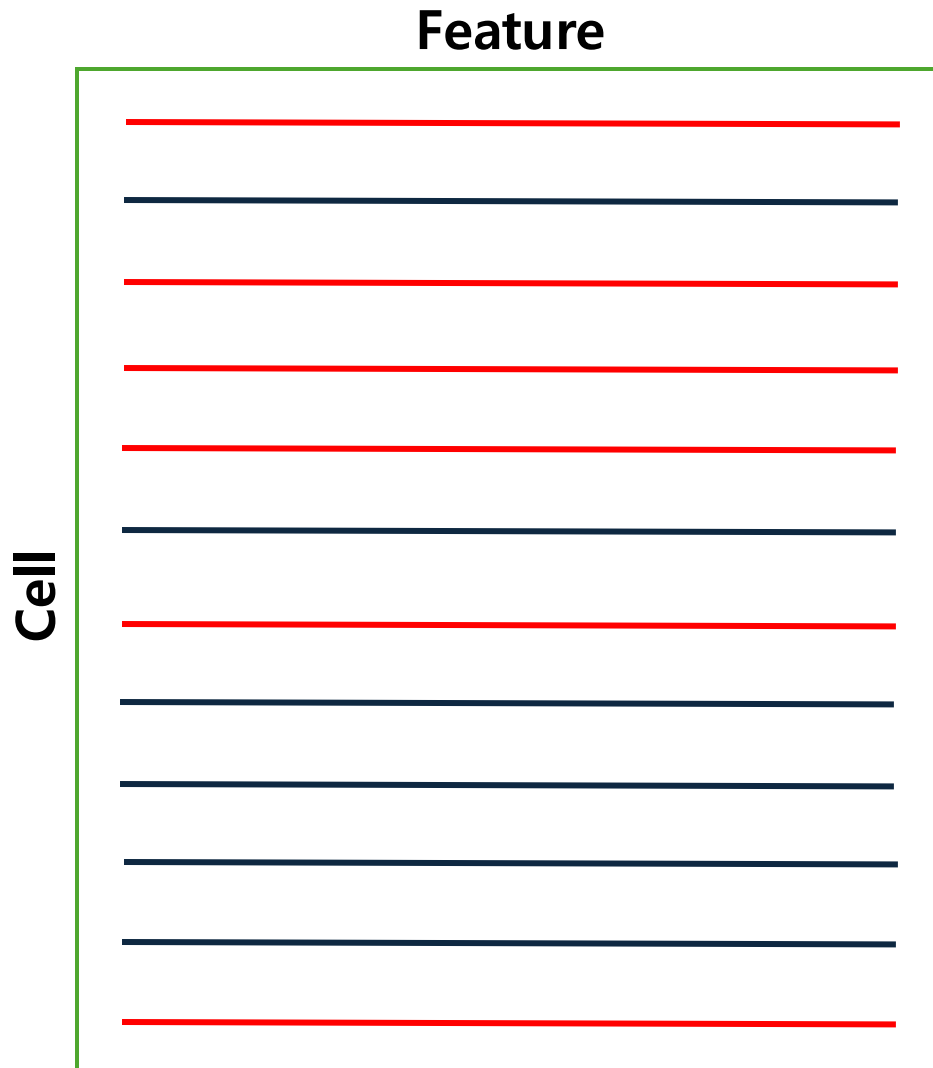
Background

With the rapid adoption of scRNA-seq technologies, many datasets now measure transcriptional profiles across **multiple conditions and at single-cell resolution**. These datasets are highly **complex** due to the presence of multiple samples.

Purpose

The goal of this study is to build **muscat, an integrated framework for scRNA DS analysis** and systematically evaluate **different methods and analytical strategies** for detecting differential states across **a range of realistic scenarios** in such complex scRNA-seq datasets.

Single Cell RNA seq



Common goals in scRNA-seq analysis

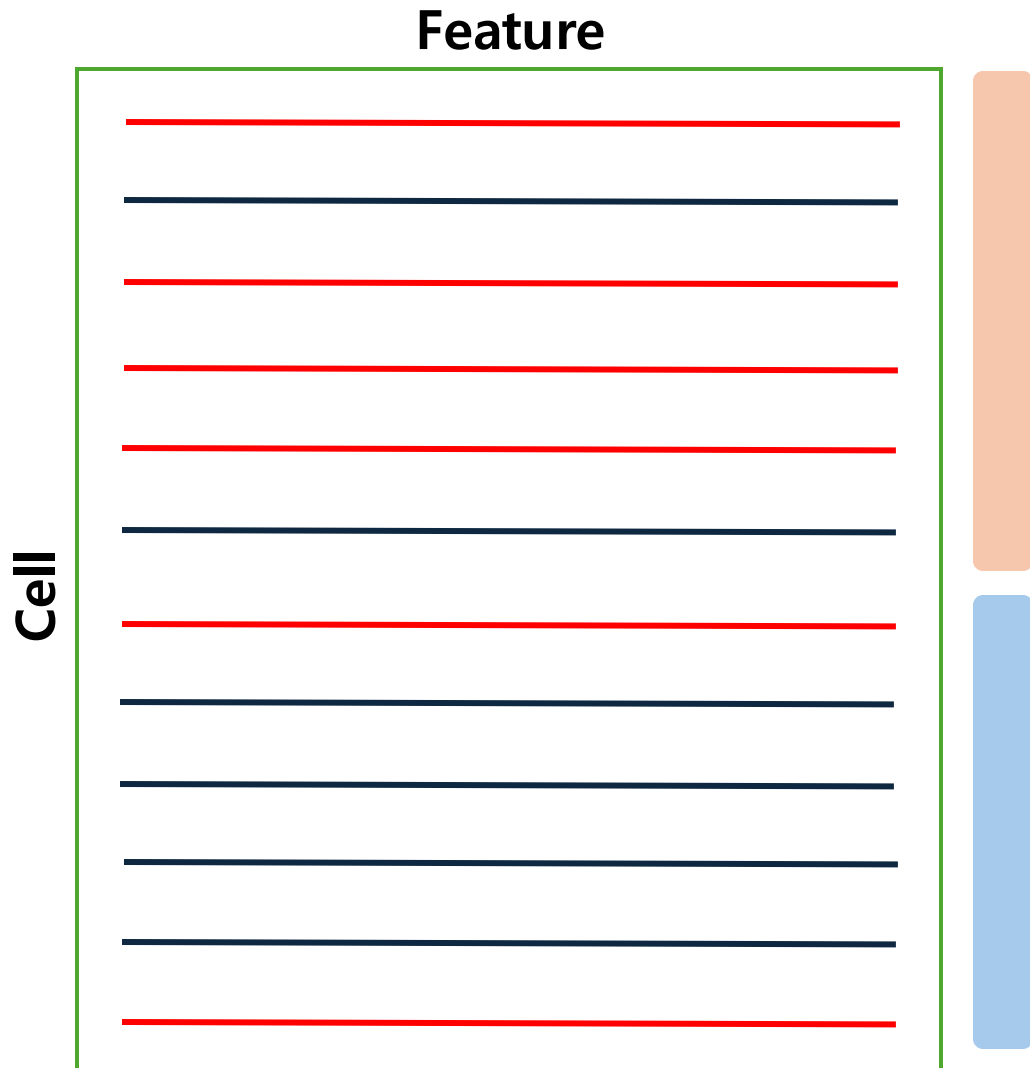
Single cell level analysis

Differential Expression

Cell type Annotation

Differential Abundance of Marker Genes

Single Cell RNA seq



Common goals in scRNA-seq analysis

Single cell level analysis

Differential Expression

Cell type Annotation

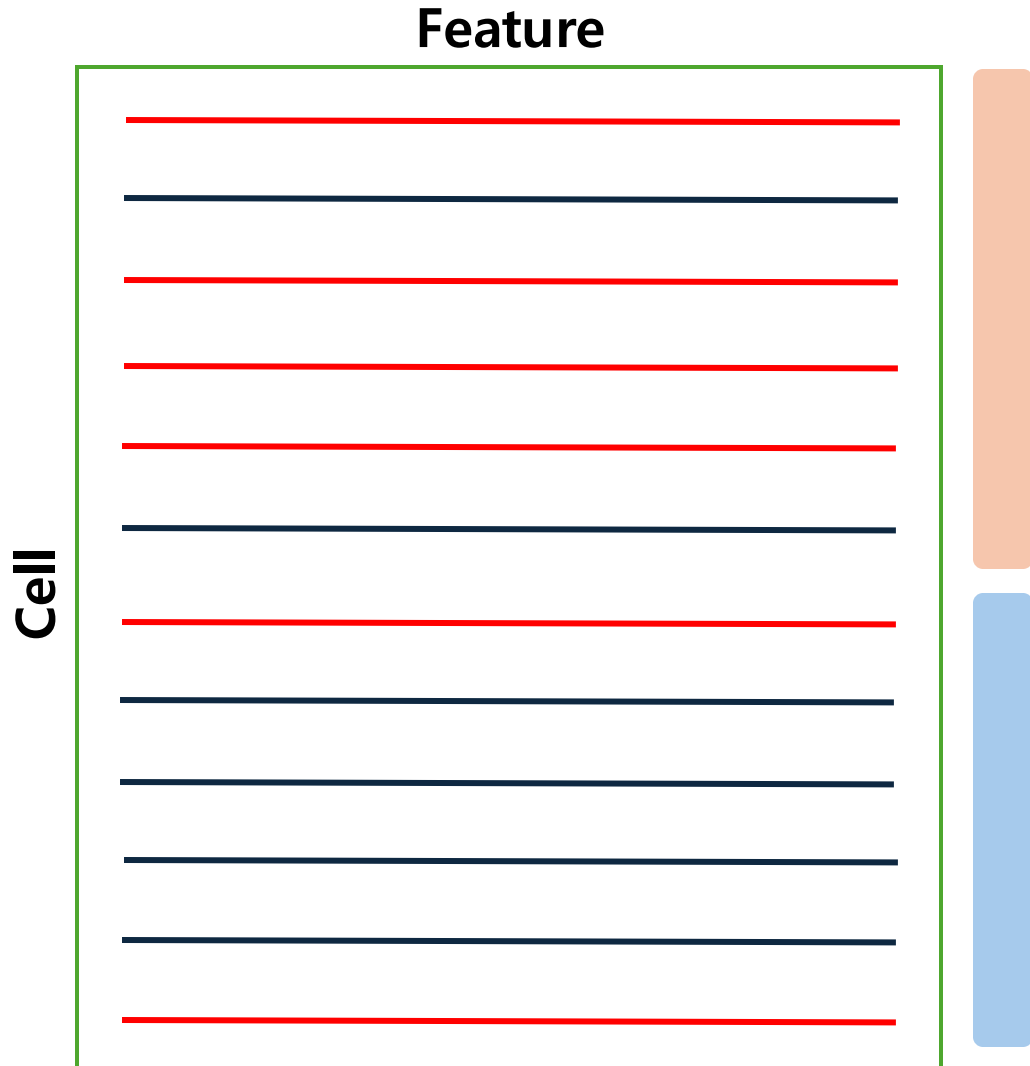
Differential Abundance of Marker Genes

The main target of this paper

Differential State Analysis

- Identification of cell state changes within a given cell subpopulation across conditions

Single Cell RNA seq



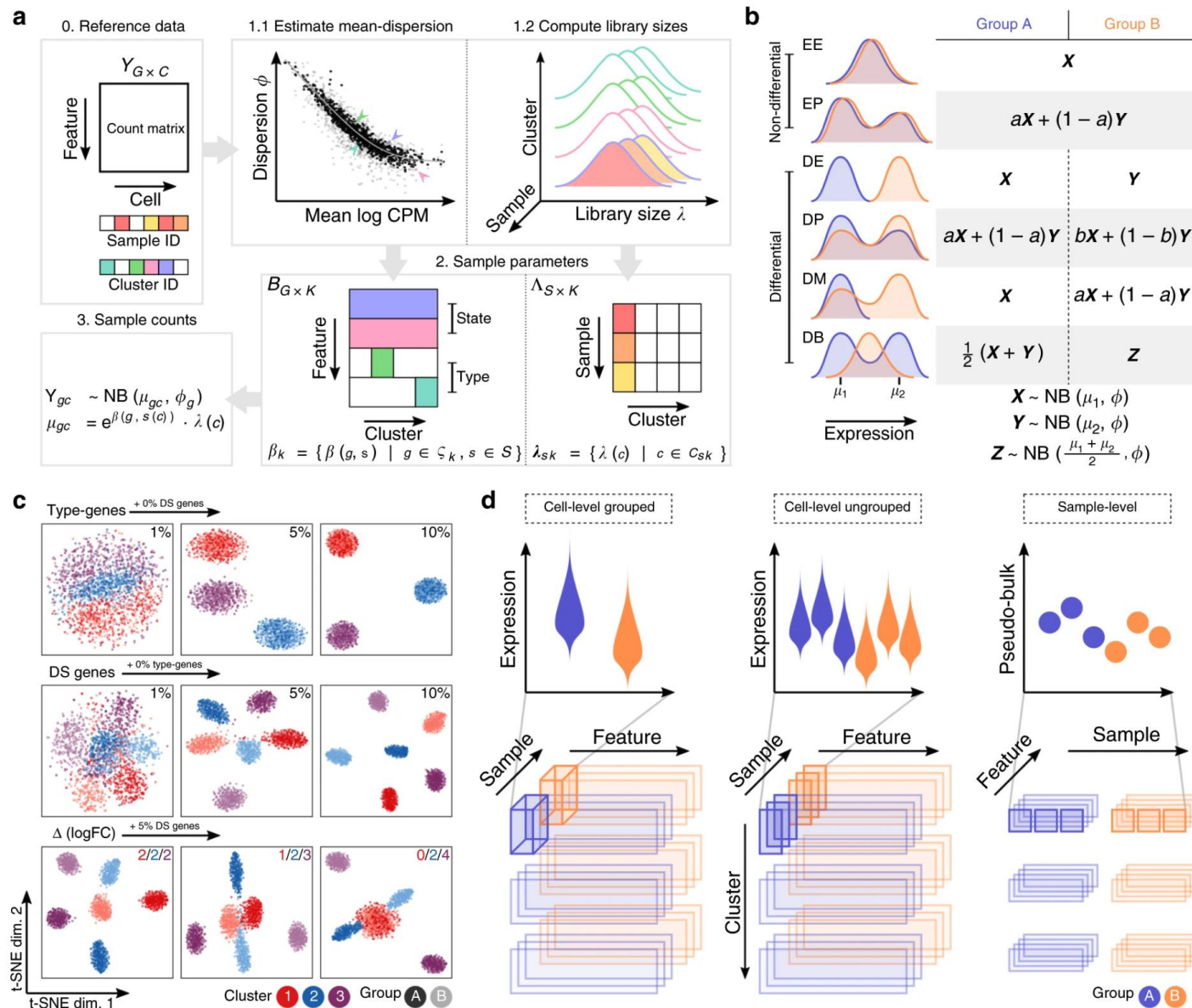
So, in this paper ...

They evaluated **16 different methods** based on **single-cell level vs. pseudobulk level analyses**

within a **simulation framework incorporating 6 distinct scenarios of scRNA seq data**

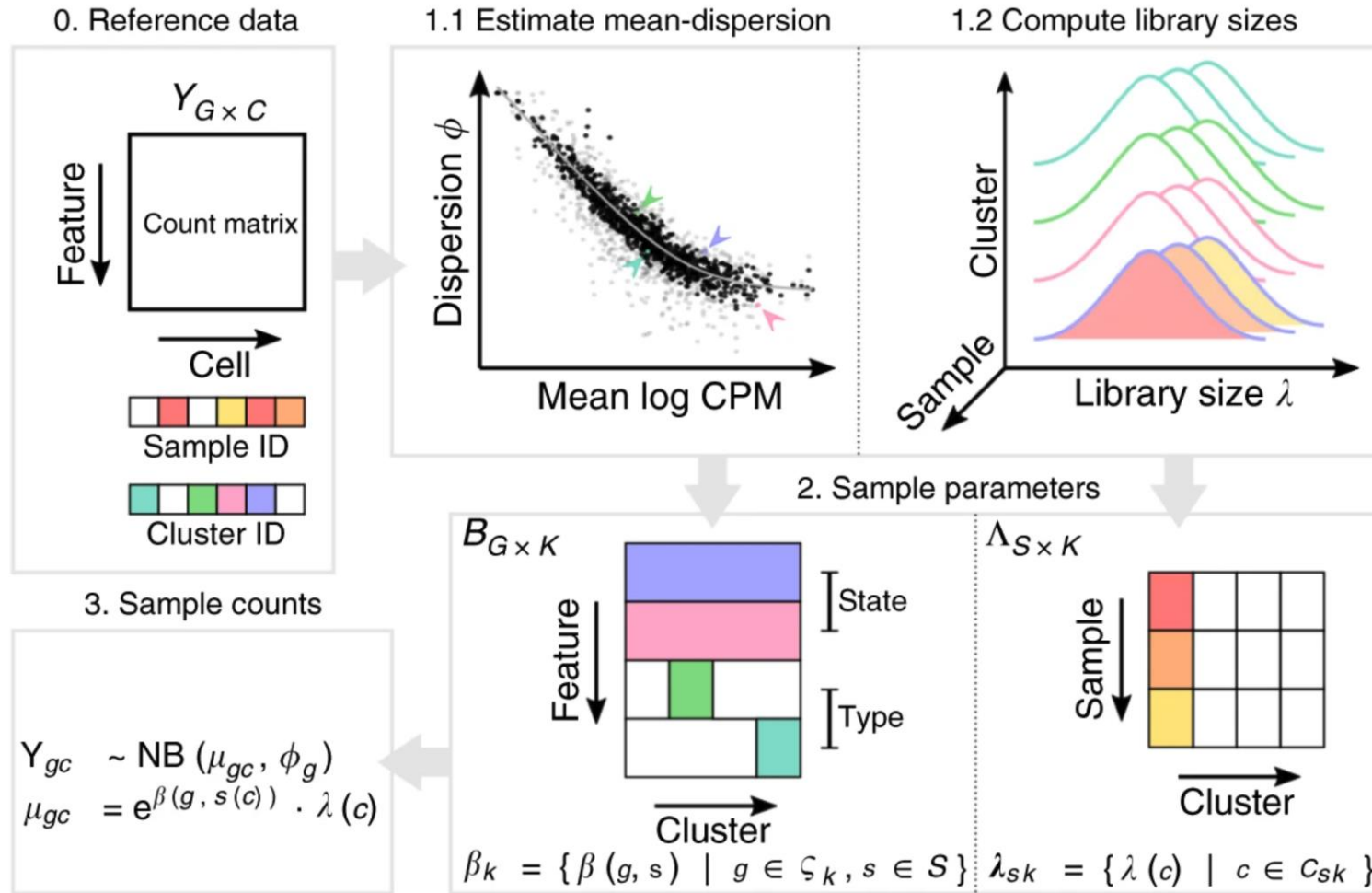
This results can be helpful for CropSeq DEG analysis and scMPRA analysis.

Figure 1. Simulation Pipeline



Simulation Setup

a



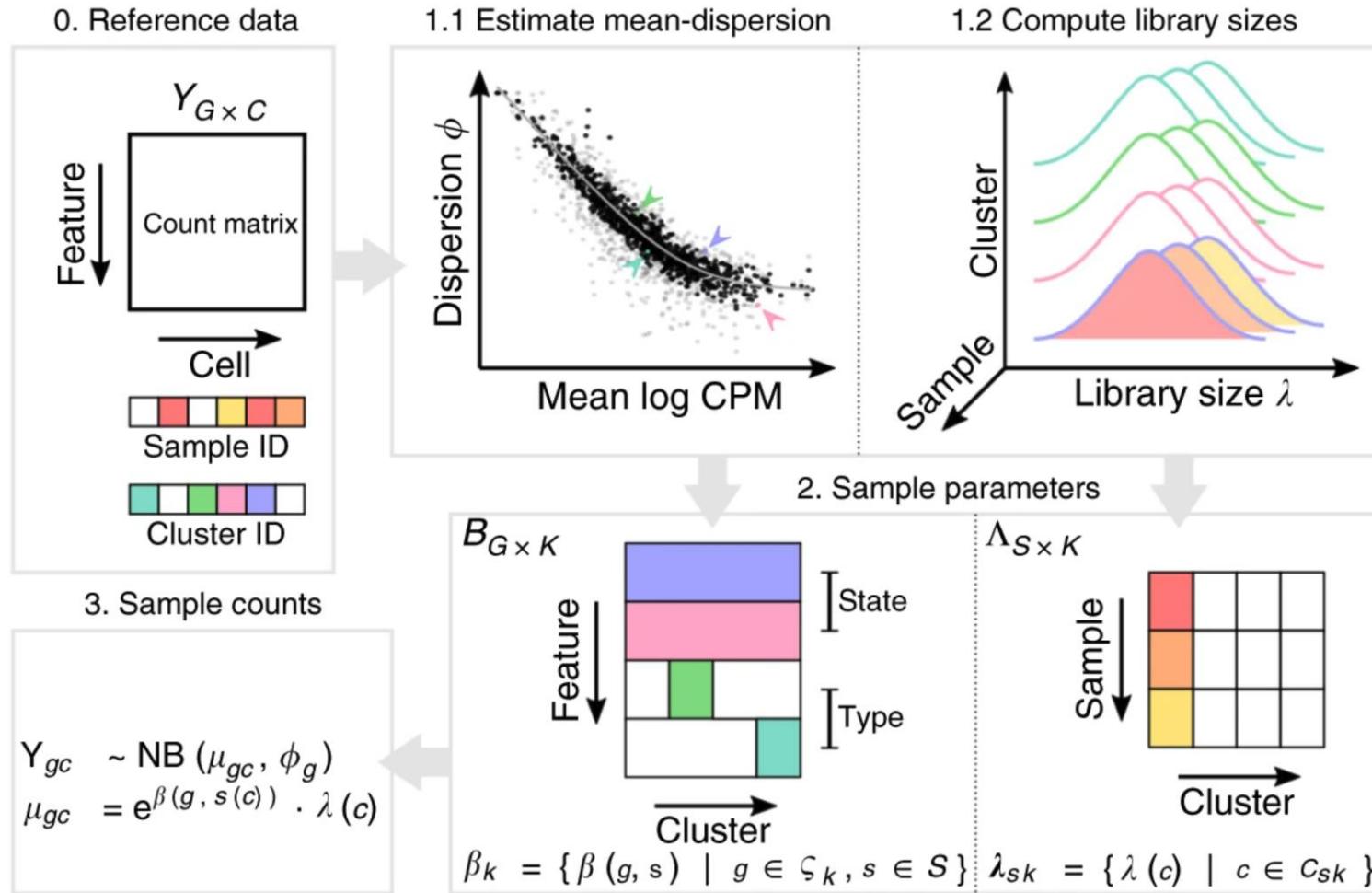
0. Getting Reference Datasets

scRNA-seq data of peripheral blood mononuclear cells (PBMCs) from eight lupus patients measured before and after 6 h-treatment with IFN- β (16 samples in total)

<https://www.nature.com/articles/nbt.4042>

Simulation Setup

a



0. Getting Reference Datasets

scRNA-seq data of peripheral blood mononuclear cells (PBMCs) from eight lupus patients measured before and after 6 h-treatment with IFN- β (16 samples in total)

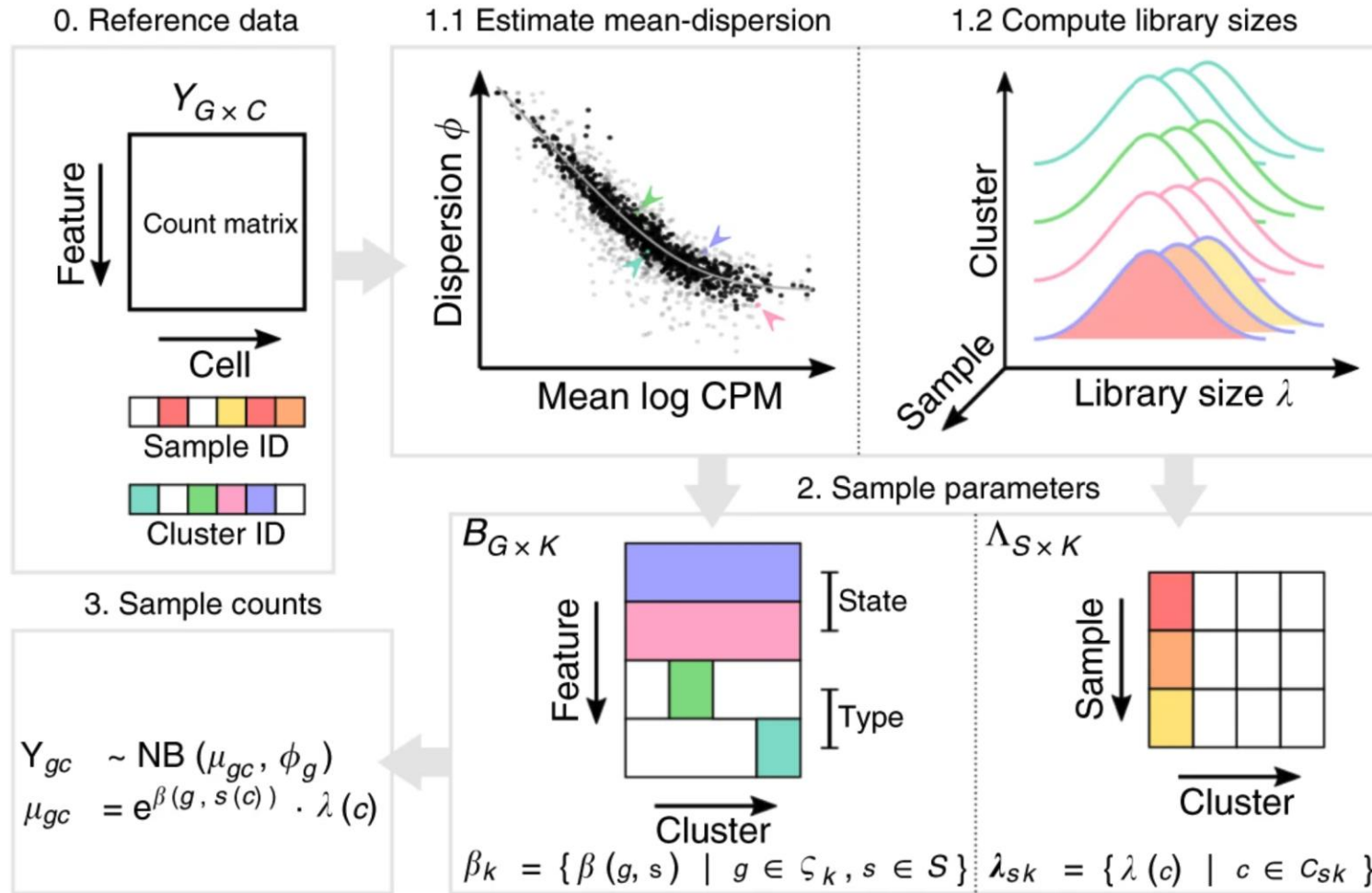
<https://www.nature.com/articles/nbt.4042>

1. Getting Stat of Reference Data

- Dispersion Ratio
- Library Size ($\sim$ Read Depth)

Simulation Setup

a



0. Getting Reference Datasets

scRNA-seq data of peripheral blood mononuclear cells (PBMCs) from eight lupus patients measured before and after 6 h-treatment with IFN- β (16 samples in total)
<https://www.nature.com/articles/nbt.4042>

1. Getting Stat of Reference Data

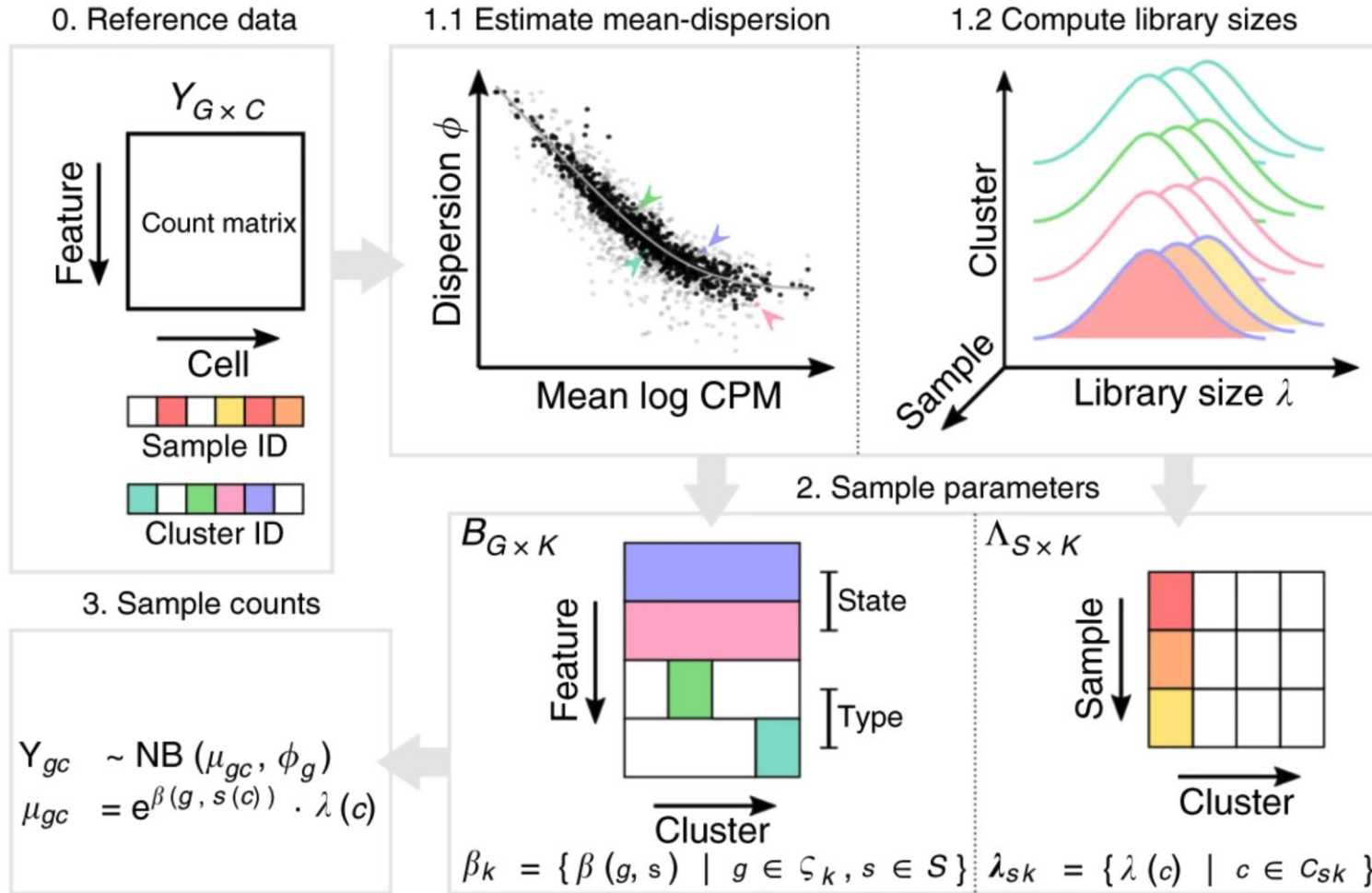
- Dispersion Ratio
- Library Size ($\sim$ Read Depth)

2. Sample Parameters

- Cluster-wise Feature Parameter
(e.g. Gene Expression parameter per Cell types & Condition)
- Library Size ($\sim$ Read Depth)

Simulation Setup

a



0. Getting Reference Datasets

scRNA-seq data of peripheral blood mononuclear cells (PBMCs) from eight lupus patients measured before and after 6 h-treatment with IFN- β (16 samples in total)
<https://www.nature.com/articles/nbt.4042>

1. Getting Stat of Reference Data

- Dispersion Ratio
- Library Size ($\sim$ Read Depth)

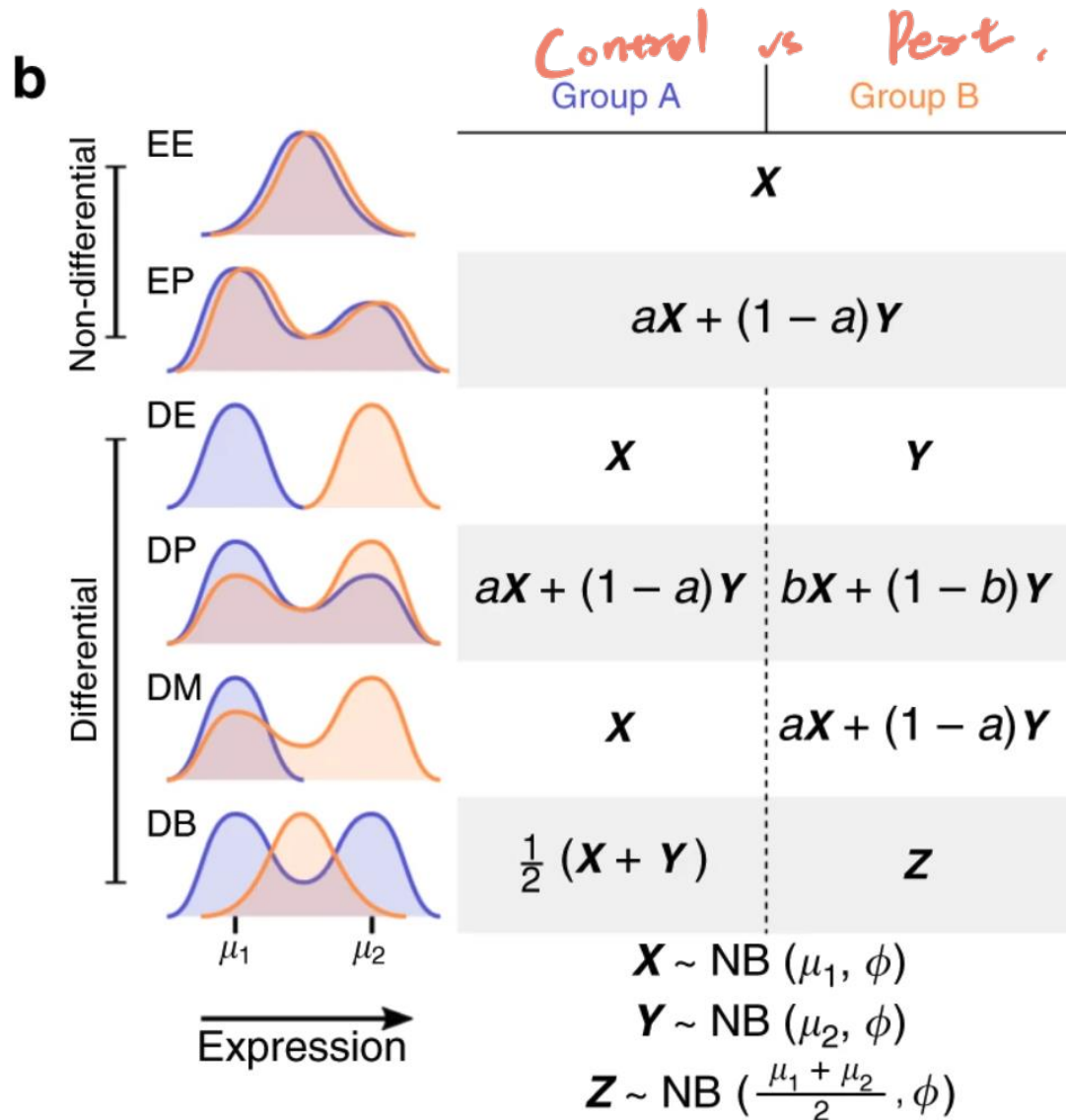
2. Sample Parameters

- Cluster-wise Feature Parameter
(e.g. Gene Expression parameter per Cell types & Condition)
- Library Size ($\sim$ Read Depth)

3. Sample Counts

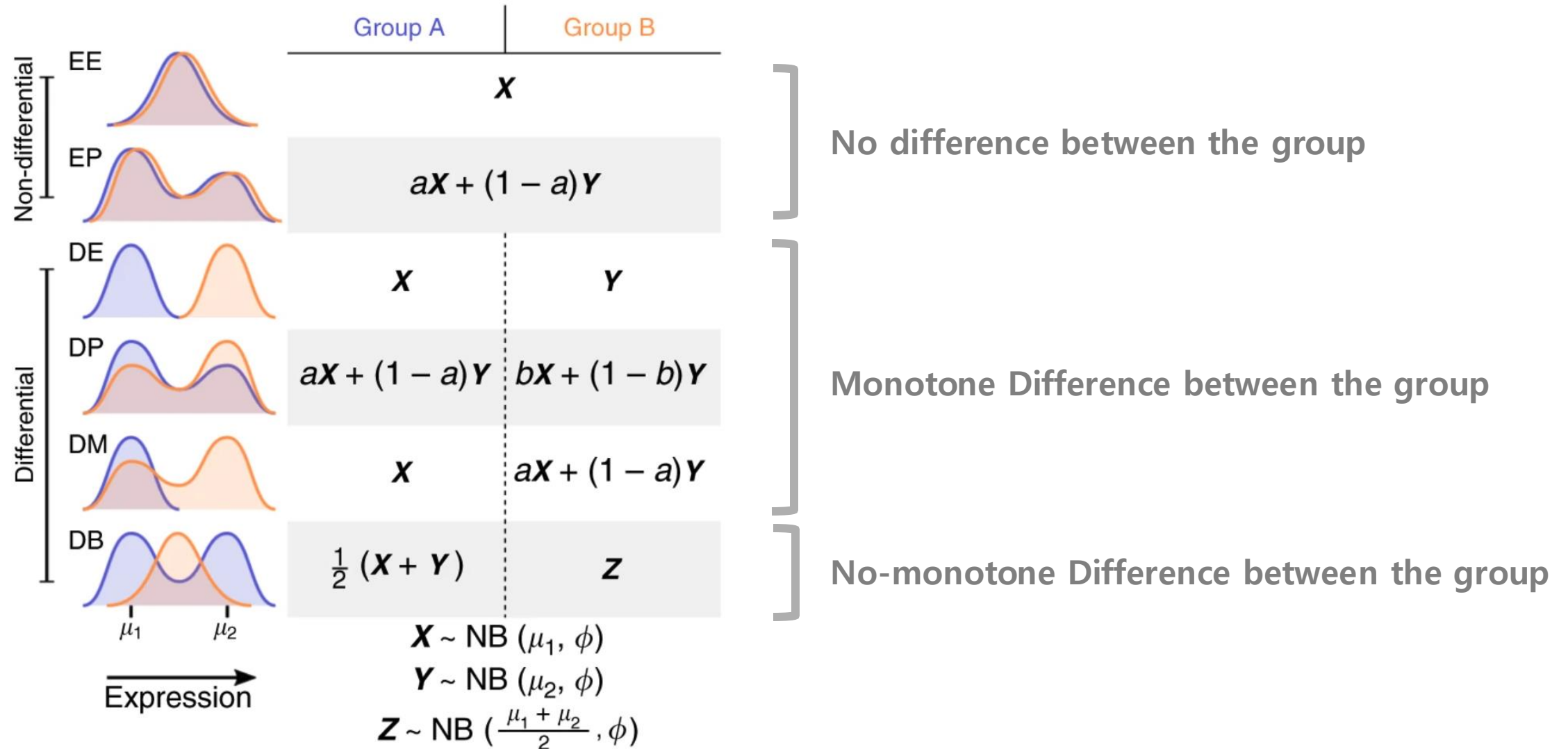
Simulation gene counts are generated from the given parameter and negative binomial distribution

Gene Expression Scenario



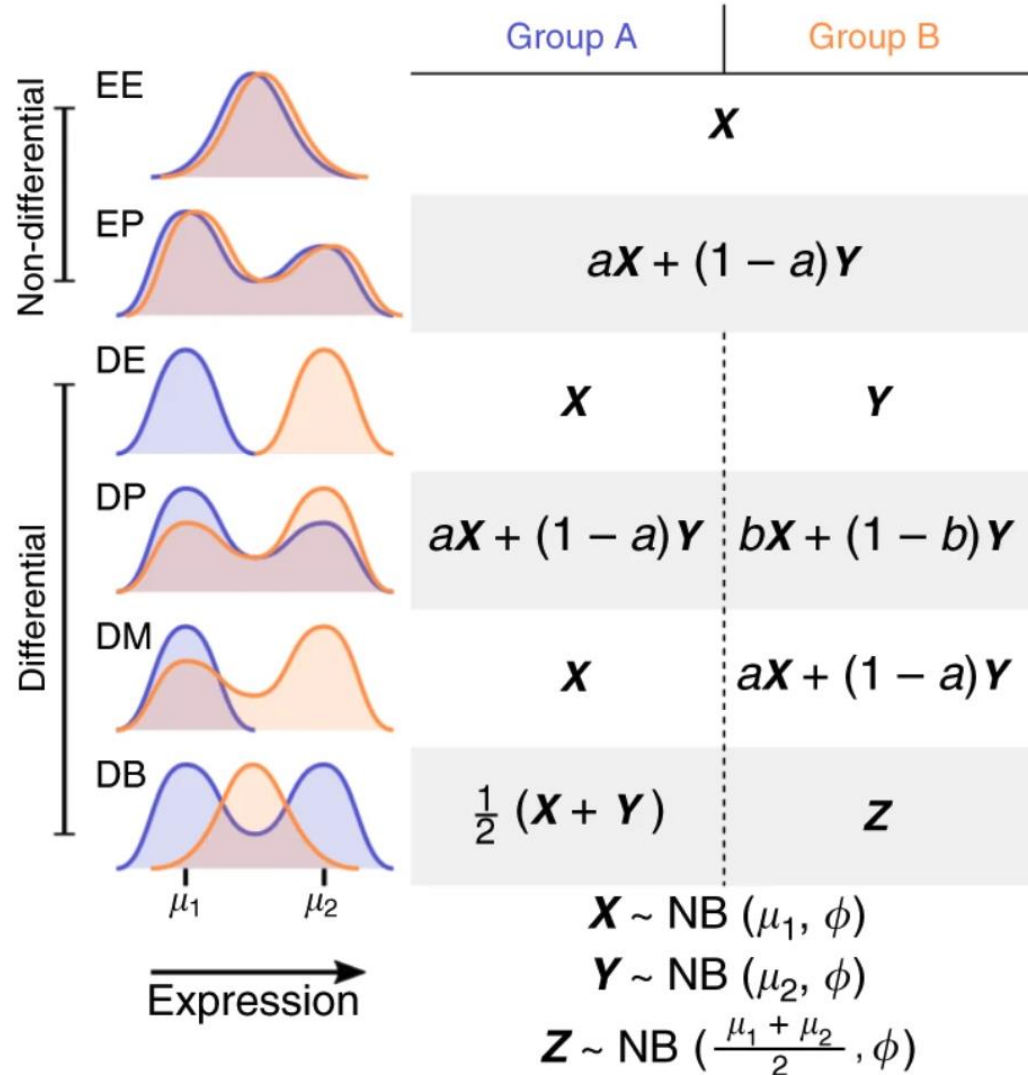
Gene Expression Scenario

b



Gene Expression Scenario

b



Easiest and Simple Form

Bimodal

cf) Source of Bimodality

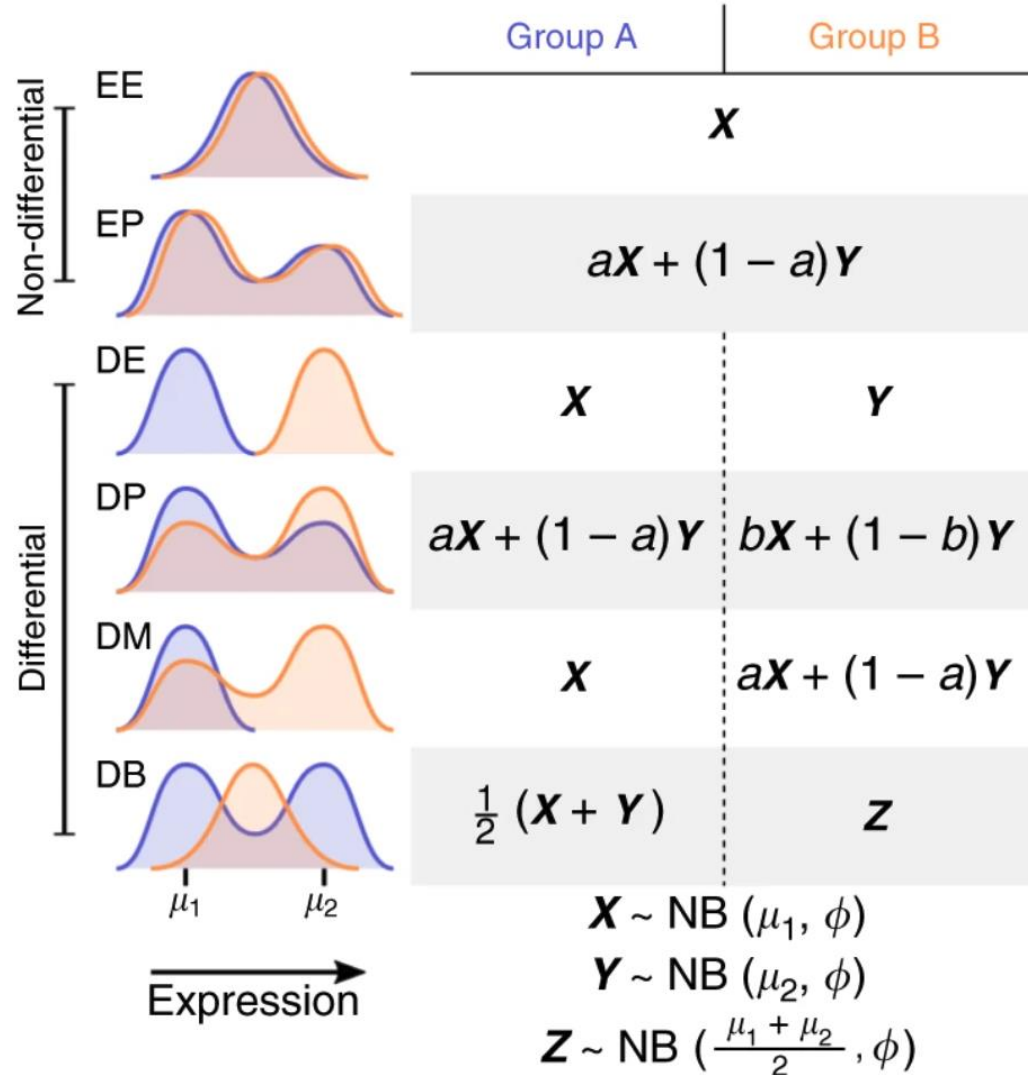
Undetected Confounder

Zero Inflation

(Missing due to lack of measurement power)

Gene Expression Scenario

b



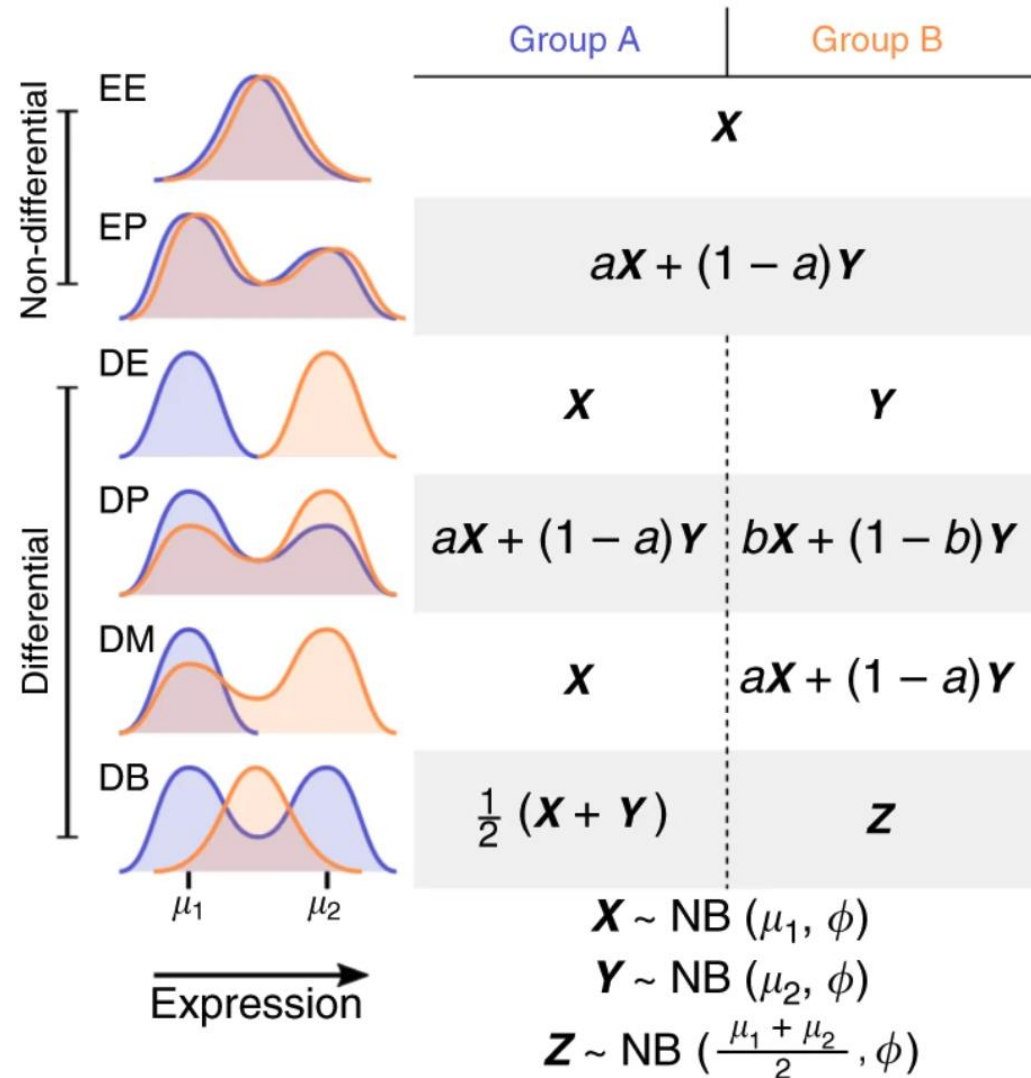
Clear Separation Form

Bimodal but detectable effect
(Most realistic patterns)

Bimodal but clear separation

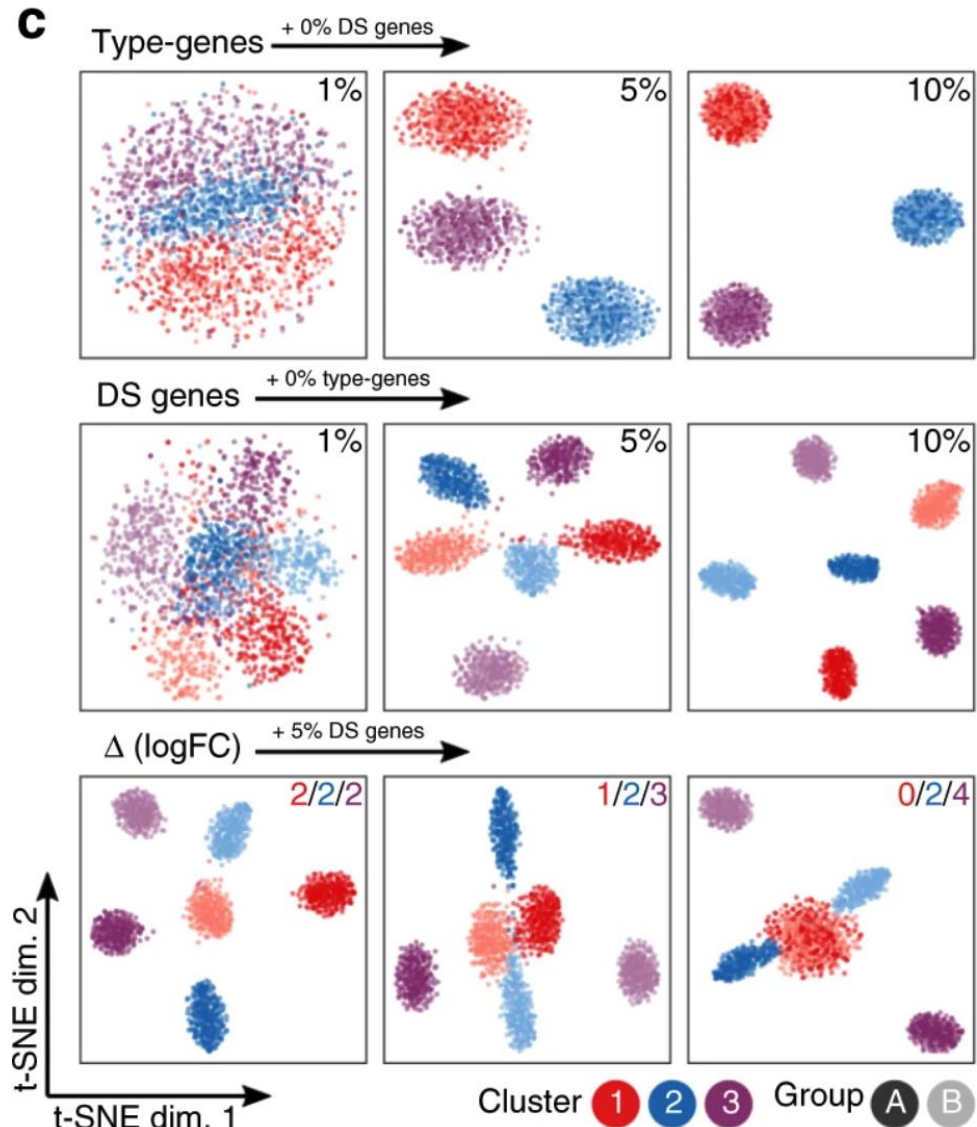
Gene Expression Scenario

b



Clear Separation but non-linear form
(Group means are the same)

Behavior of scRNA Data sets



e.g) Marker Gene of cell type (cluster or subpopulation)

e.g) Genes that behave differently under the condition

e.g) logFC per conditions

Models

6 models

ScDD

MAST

AD

MM

EdgeR

Limma

Models

6 models : 16 variation

ScDD : logcounts, vstresiduals

MAST : logcounts

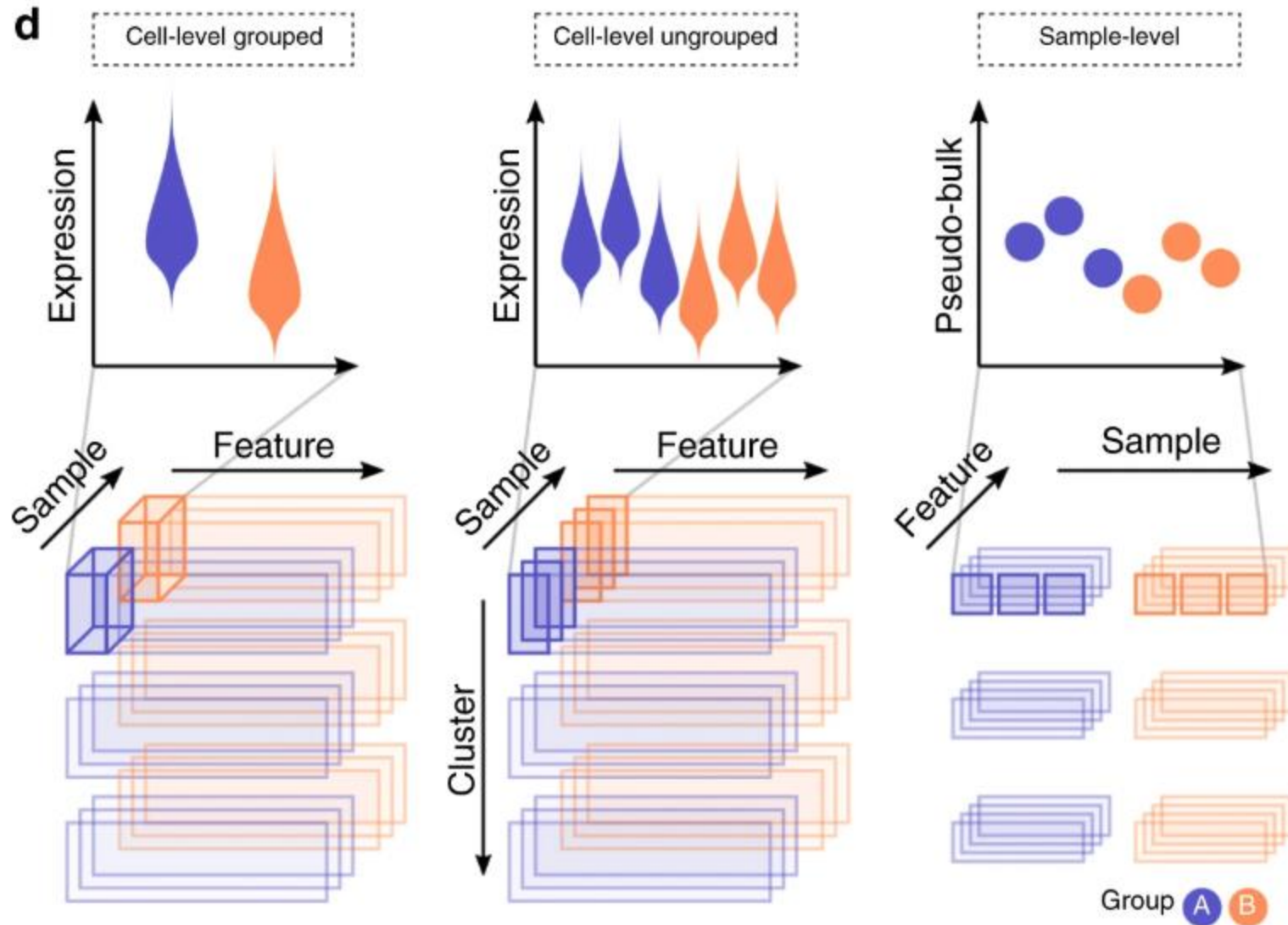
AD : gid.logcounts, gid.vstresiduals, sid.logcounts, sid.vstresiduals

MM : dream, dream2, nbinom, vstresiduals

EdgeR : sum.counts, sum.scalecpm

Limma : voom.sum.counts, trend.mean.logcounts, trend.mean.vstresiduals

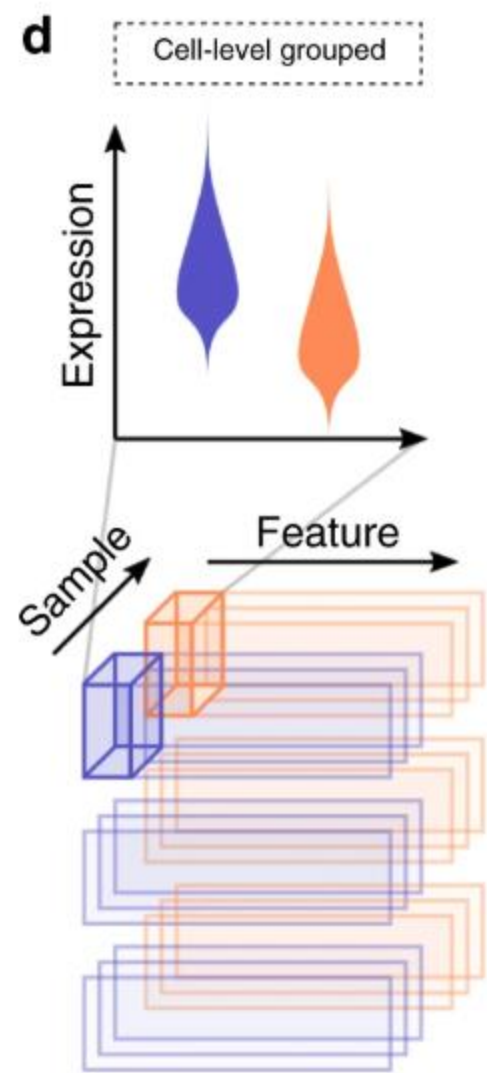
Model Approach



Model Approach

Cell	Sample (Mouse)	Condition	Gene A	Gene B
1A	1	T	-	-
1B	1	T	-	-
1C	1	F	-	-
2A	2	T	-	-
2B	2	F	-	-
2C	2	F	-	-

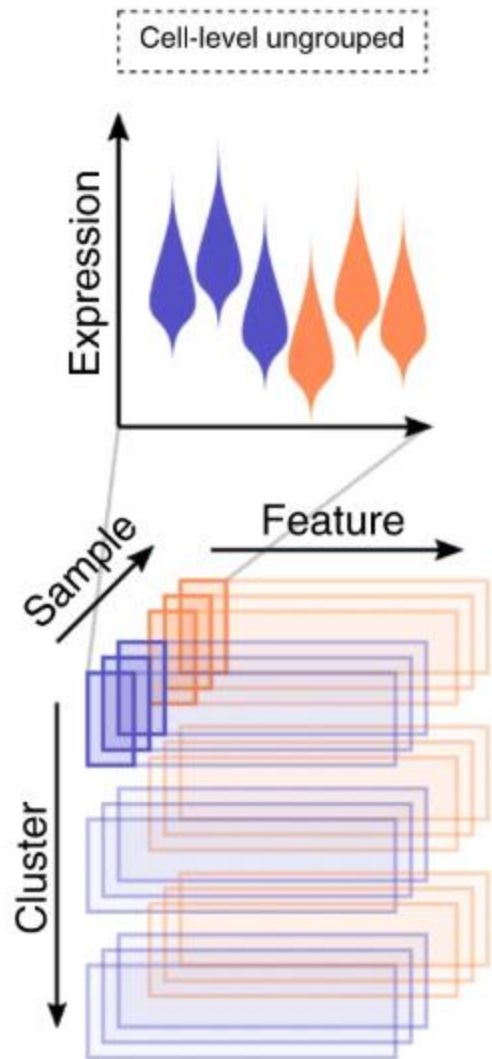
Model Approach



Cell	Sample (Mouse)	Condition	Gene A	Gene B
1A		T	-	-
1B		T	-	-
1C		F	-	-
2A		T	-	-
2B		F	-	-
2C		F	-	-

-> AD (AD-gid), MAST, scDD

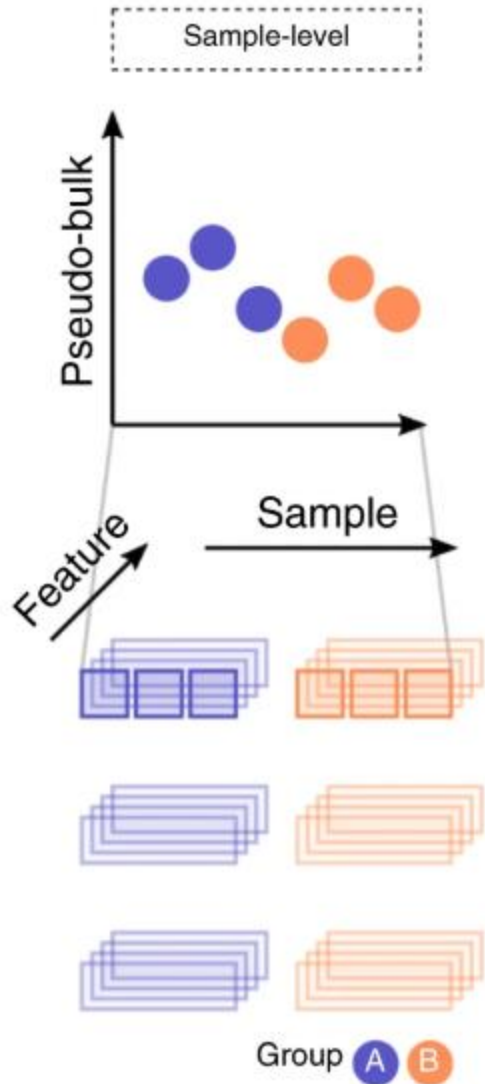
Model Approach



Cell	Sample (Mouse)	Condition	Gene A	Gene B
1A	1	T	-	-
1B	1	T	-	-
1C	1	F	-	-
2A	2	T	-	-
2B	2	F	-	-
2C	2	F	-	-

-> MM, AD (AD-sid)

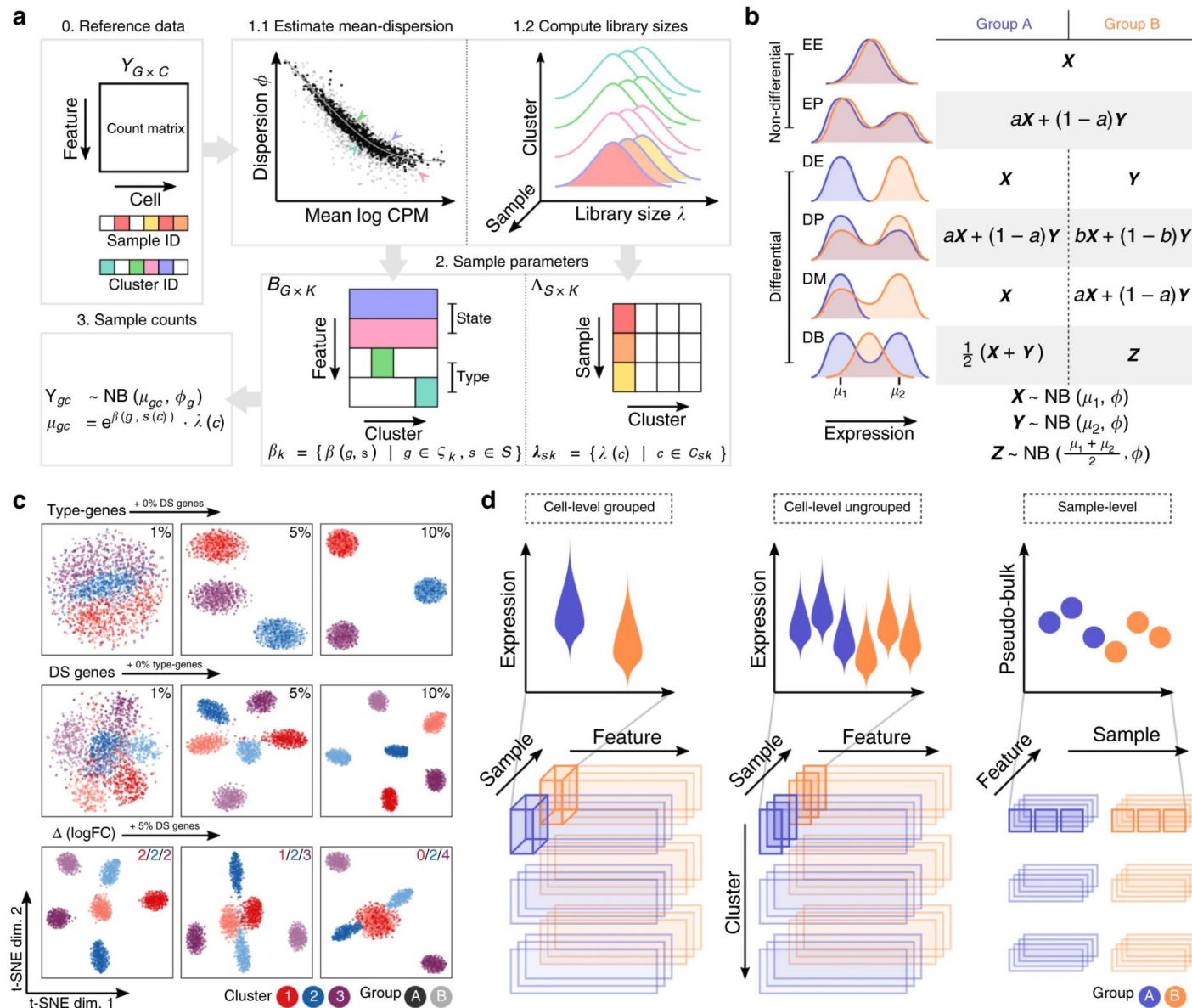
Model Approach



Cell	Sample (Mouse)	Condition	Gene A	Gene B
1A	1	T	-	-
1B	1	T	-	-
1C	1	F	-	-
2A	2	T	-	-
2B	2	F	-	-
2C	2	F	-	-

-> EdgeR, Limma-voom, Limma-trend

Figure 1. Simulation Pipeline



Models

scDD: single cell Differential Distribution (2015)

scDD

MAST

AD

MM

EdgeR

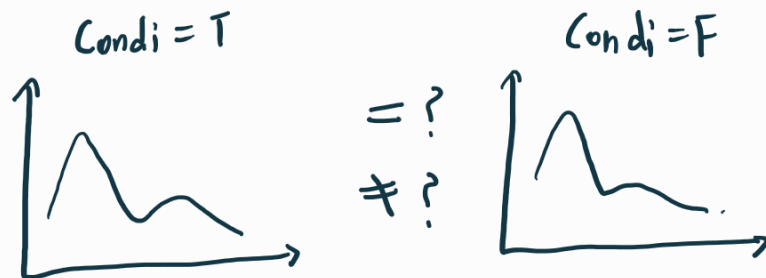
Limma

Cell	Condition	Gene A expression
1A	T	-
1B	T	-
1C	F	-
2A	T	-
2B	F	-
2C	F	-

Dirichlet Process Mixture + Bayes Factor

Distributional Difference

scDD.



How?

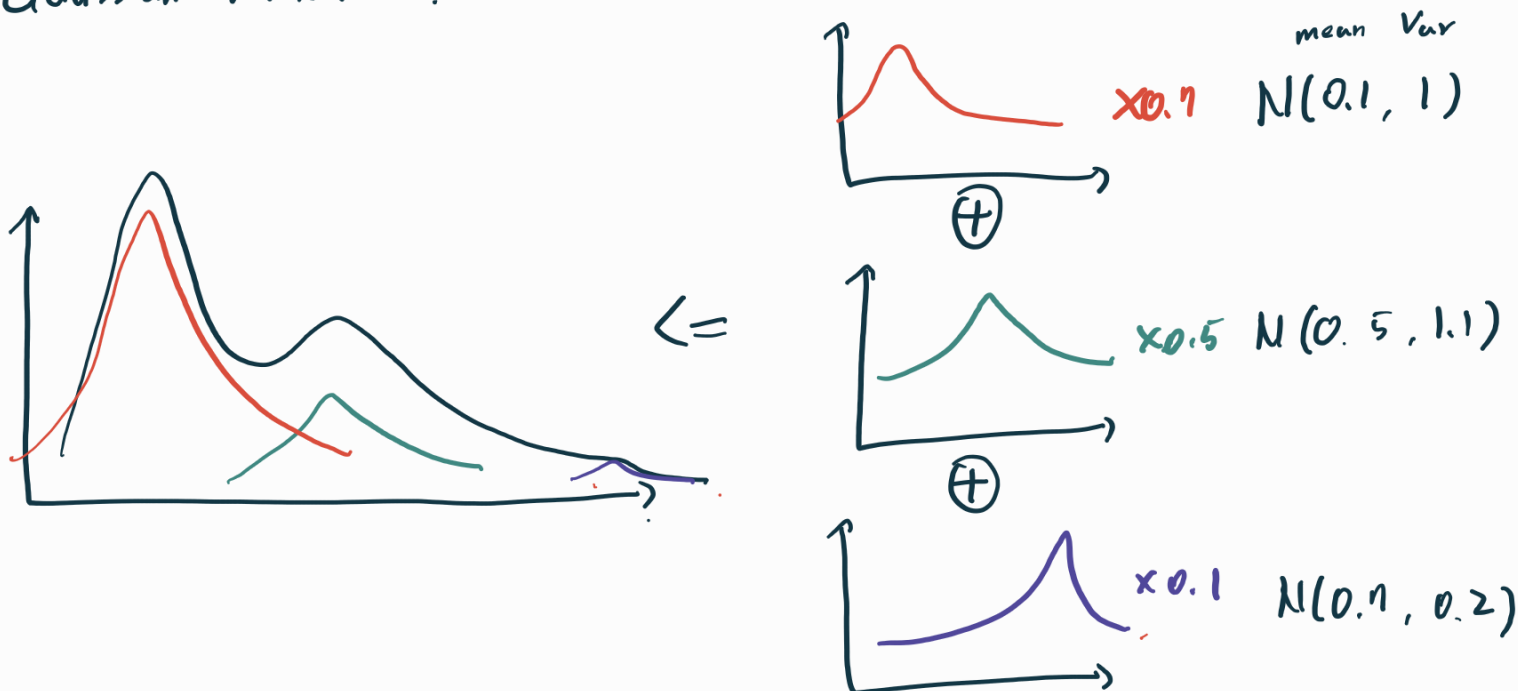
$$H_0: \begin{cases} y_i | G_i = T \sim G \\ y_i | G_i = F \sim G \end{cases} \text{ vs } H_1: \begin{cases} y_i | G_i = T \sim G_T \\ y_i | G_i = F \sim G_F \end{cases}$$

⇒ Bayesian Factor

$$\frac{Pr(D | H_0)}{Pr(D | H_1)} > k$$

G is estimated by Dirichlet Process Mixture
endowed with Gaussian Distribution
(Normal Dist.)

Gaussian Mixture.



For $N(\mu, \sigma^2)$, $\hat{\mu} = \frac{1}{n} \sum x_i$, $\hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

∵ 3 N , we don't know x_i come from which N ,

⇒ 1. Assigning x_i with some rules 2. Estimating N params 3. Evaluating General Fit.
(Dirichlet process)

Repeat 1~3 until we got the best structure.

Models

Interpretation [\[edit \]](#)

A value of $K > 1$ means that M_1 is more strongly supported by the data under consideration than M_2 . Note that classical [hypothesis testing](#) gives one hypothesis (or model) preferred status (the 'null hypothesis'), and only considers evidence *against* it. The fact that a Bayes factor can produce evidence *for* and not just against a null hypothesis is one of the key advantages of this analysis method.^[13]

[Harold Jeffreys](#) gave a scale (**Jeffreys' scale**) for interpretation of K :^[14]

K	dHart	bits	Strength of evidence
$< 10^0$	< 0	< 0	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

The second column gives the corresponding weights of evidence in [decihartleys](#) (also known as [decibans](#)); [bits](#) are added in the third column for clarity. The table continues in the other direction, so that, for example, $K \leq 10^{-2}$ is decisive evidence for M_2 .

An alternative table, widely cited, is provided by Kass and Raftery (1995):^[10]

$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Models

MAST : Model-based Analysis of Single-cell Transcriptomics (2015)

scDD

Regression model with Hurdle Distribution

MAST

**Hurdle Distribution : Two stage Model to handle the zero-inflated counts
(could be come from dropout or biological reason)**

AD

Evaluating the effect of the condition of interest using LRT test

MM

LRT Test

1. Fitting two model with and without condition information
2. Evaluating each model through Likelihood; $\Pr(D|P)$
3. If the condition affects gene expression, then a model that incorporates the condition should explain the data better than the model that does not.
4. Using the theorem that loglikelihood difference follows the chi-square distribution under H_0 : no effect of the condition, we can calculate the p-value of LRT test

EdgeR

Limma

MAST

Raw Data

	Cell	G	X
Condi 1	Y_1	0	
	Y_2	0	
	Y_3	1	
	Y_4	2	
	Y_5	0	
	Y_6	3	
Condi 2	Y_7	1	
	Y_8	2	
	Y_9	0	
	Y_{10}	0	
	Y_{11}	0	
	Y_{12}	1	

$\Rightarrow$

First Stage.

Cell	G	X	R_i
Y_1	0		0
Y_2	0		0
Y_3	1		1
Y_4	2		1
Y_5	0		0
Y_6	3		1
Y_7	1		1
Y_8	2		1
Y_9	0		0
Y_{10}	0		0
Y_{11}	0		0
Y_{12}	1		1

$\Rightarrow$

Second Stage

Cell	G	X	R_i
Y_3	1		1
Y_4	2		1
Y_6	3		1
Y_7	1		1
Y_8	2		1
Y_{12}	1		1

$$I(Y_i = 0) = R_i$$

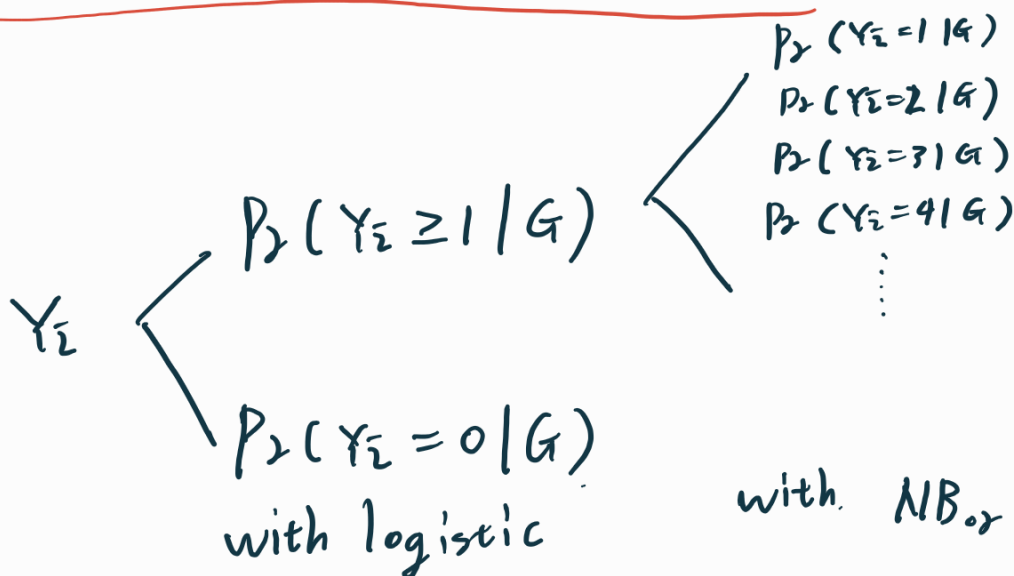
$$D = \{Y_i; Y \geq 1\}$$

$$R_i \sim \text{Bernoulli}(X_i^T \beta)$$

$$Y_i | Y_i \in D \sim \text{NB}(X_i^T \beta)$$

$$\sim \text{Pois}(X_i^T \beta)$$

DGP. (Data Generating Process)



Models

scDD

MAST

AD

MM

EdgeR

Limma

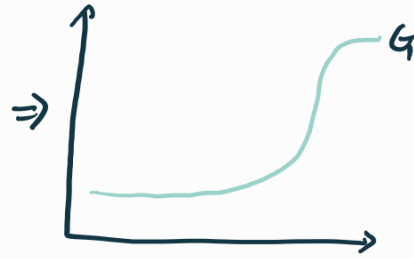
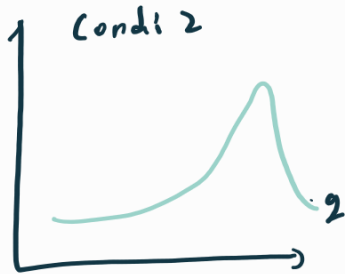
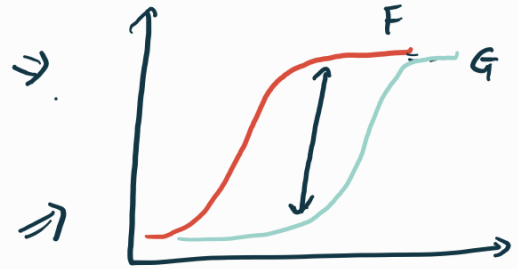
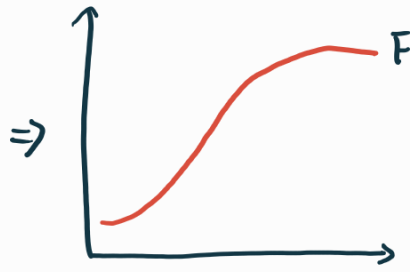
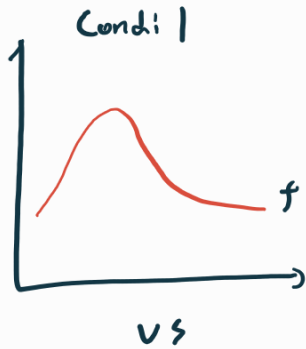
- AD : Anderson Darling Test (1952)
- Somewhat advanced version of Kolmogorov Smilnov Test (1930)

 KS vs CvM vs Anderson–Darling (Quick Comparison)			
Feature	K-S	Cramér–von Mises (CvM)	Anderson–Darling (A-D)
Idea	Max CDF difference	Average CDF difference	Tail-weighted difference
Statistic	$\sup F_n - G_m $	$\int (F_n - G_m)^2 dH$	$\int \frac{(F_n - G_m)^2}{H(1-H)} dH$
Sensitivity	Local (max gap)	Global	Tail-focused
Tail sensitivity	Low	Medium	High
Power (overall)	Lower	Moderate	Higher (esp. tails)
Best for	Sharp local differences	Global shifts	Heavy tails / rare events
Weakness	Misses tail differences	No tail emphasis	Slightly more complex
p-values	Analytic	Approx/table	Approx/table

K-S Index,

Density: $f(t) = P_X(X=t)$

CDF: $F(t) = P_X(X \leq t)$



$$K = \text{K-S Index} = \sup_x |F(x) - G(x)|$$

$$= \|F - G\|_\infty$$

under H_0 , $K \sim \sup_{t \in [0,1]} |B(t)|$

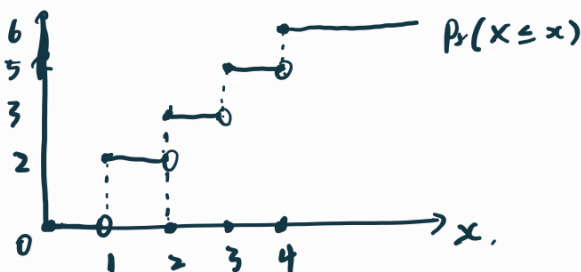
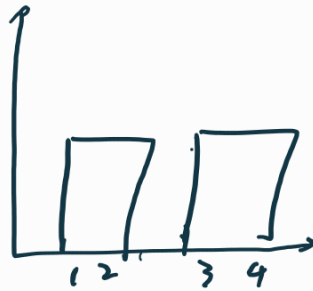
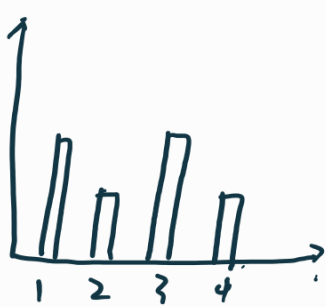
Brownian Bridge.

$(B(t) = (W(t) - W(0)))$

Why CDF? Histogram/KDE depends on bandwidth (Hyper parameter)

Empirical CDF (ECDF) does not.

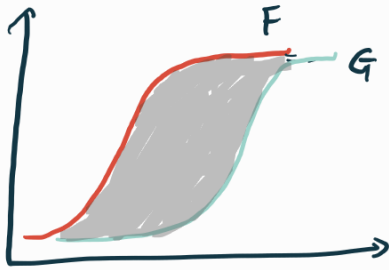
e.g) $X = 1, 1, 2, 3, 3, 4$



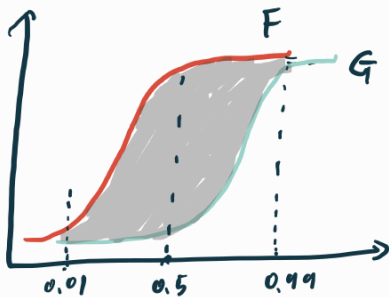
Cramer - von Mises test.

$$w^2 = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dH(x) ; H(x): \text{pooled Distribution.}$$

$$\text{Under } H_0, w^2 \sim \int_0^1 B(t)^2 dt.$$



AD.
$$A^2 = n \int \frac{(F(x) - G(x))^2}{P_x(X \leq x) P_x(X > x)} dH(x).$$



$$\text{Under } H_0, A^2 \sim \int_0^1 \frac{B(t)^2}{t(1-t)} dt.$$

Weight.

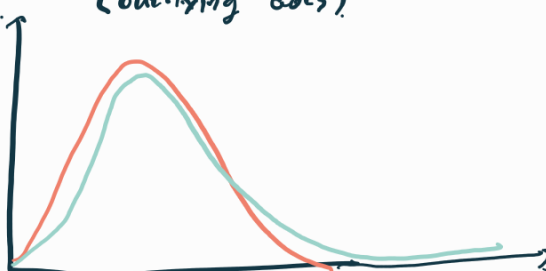
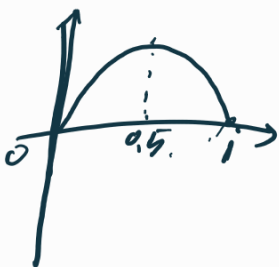
e.g.

- $P_x(X \leq 0.01) P_x(X > 0.01) \Rightarrow 0.1 \times 0.9 \Rightarrow 0.09 \Rightarrow 1/0.09 \approx 11$
- $P_x(X \leq 0.5) P_x(X > 0.5) \Rightarrow 0.5 \times 0.5 \Rightarrow 0.25 \Rightarrow 1/0.25 \approx 4$
- $P_x(X \leq 0.99) P_x(X > 0.99) \Rightarrow 0.9 \times 0.1 \Rightarrow 0.09 \Rightarrow 1/0.09 \approx 11$

$t(1-t)$

So, tail behavior become more important.
(outlying does)

e.g.



Models

scDD

MM : Mixed Model

MAST

Regression Model using Fixed Effect + Random Effect

AD

Fixed effects: coefficients are treated as constant across all observations

MM

Random effects: coefficients are treated as random variables that vary across groups or subjects.

EdgeR

Mixed models are commonly used for hierarchical data or repeated-measures data.

Limma

MM

Target Situation

	log R/D	Rep
A1	2	rep1
A1	1	rep2
A1	100	rep3
A2	-2	rep1
A2	-1	rep2
A2	-100	rep3

 $\Rightarrow$

	β_0	β_1	β_2	β_3	β_4
$\log RD \sim \text{Int} + \text{Allele} + \text{Rep1} + \text{Rep2} + \text{Rep3}$					
2	1	1	1	0	0
1	1	1	0	1	0
100	1	1	0	0	1
-2	1	0	1	0	0
-1	1	0	0	1	0
-100	1	0	0	0	1

$$\beta_i = E(\log RD \mid i\text{-th Covariate} = 1) - E(\log RD \mid i\text{-th Covariate} = 0)$$

What is Rep is 30? we need to estimate 31 of β parameters.

This ruins the power, We have restricted samples (30x2) and trying to get too many info.

$\Rightarrow$ Set $\gamma_i \sim N(\mu, \sigma^2)$, γ_i is r.v. come from same dist.

Then all we need to estimate is $\beta_0, \beta_1, \mu, \sigma^2$.

Other philosophies are in there, but too much ☺

Models

scDD

MAST

AD

MM

EdgeR

Limma

Cell	Sample (Mouse)	Condition	Gene A	Gene B
1A	1	T	-	-
1B	1	T	-	-
1C	1	F	-	-
2A	2	T	-	-
2B	2	F	-	-
2C	2	F	-	-

Models

EdgeR : empirical analysis of DGE in R (2009)

scDD

NB GLM + Bayesian Shrinkage + Quasi-Likelihood Test

MAST

NB GLM : glm on the negative Binomial Distribution

AD

Bayesian Shrinkage

Getting a general dispersion parameter across the genes
using it as prior for gene specific dispersion parameter

MM

Quasi-Likelihood Test

Using Quasi-likelihood method and test based on that one. (QL F-test)

EdgeR

Limma

Pois; $f(x) = P(X=x) = \frac{\exp(-h) h^x}{x!}$; h is parameter.

$$E(X) = \mu = h \quad \Rightarrow h; \text{mean}$$

$$Var(X) = \sigma^2 = h.$$

Negative Binomial; $f(x) = P(X=x) = \frac{(x+r-1)!}{x!(r-1)!} (1-p)^r p^x$; r, p is parameter

$$E(X) = \mu = \frac{r(1-p)}{p}$$

$$Var(X) = \sigma^2 = \frac{r(1-p)}{p^2} =: \mu(1 + \phi\mu)$$

μ ; mean parameter.
 ϕ ; dispersion parameter

eBayes.

• Robust estimation at σ^2 ↗ lack of evidence, unstable due to data

gene A $\Rightarrow NB(\mu_A, \sigma_A^2)$

gene B $\Rightarrow NB(\mu_B, \sigma_B^2)$

gene C $\Rightarrow NB(\mu_C, \sigma_C^2)$

gene D $\Rightarrow NB(\mu_D, \sigma_D^2)$

⋮



• use all genes into.
• enough evidence.

$$\text{eBayes } \tilde{\sigma}_A^2 = w \hat{\sigma}_A^2 + (1-w) \hat{\sigma}_0^2.$$

$\Rightarrow$ Mixing $\hat{\sigma}_A^2$ and $\hat{\sigma}_0^2$.

w is calculated, under the Bayesian Framework.

Models

EdgeR : empirical analysis of DGE in R (2009)

scDD

NB GLM : glm on the negative Binomial Distribution

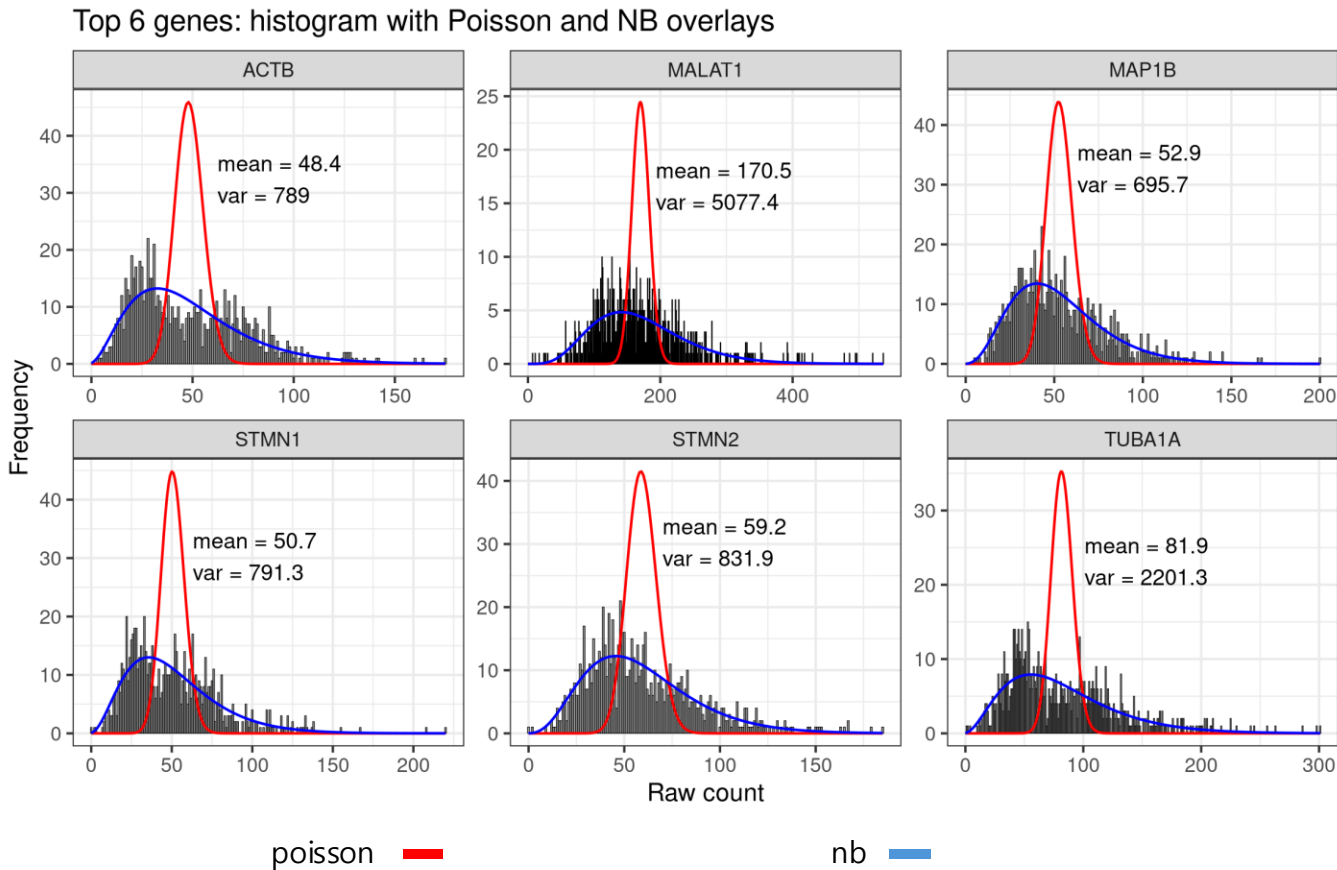
MAST

AD

MM

EdgeR

Limma



Models

Limma : linear models for microarray data (2004)

scDD

Linear Model + empirical Bayes variance Shrinkage

MAST

Limma-voom : Variance modeling at the observational level (2014)

logCPM, Mean-Variance Function, Weighted Least Square

AD

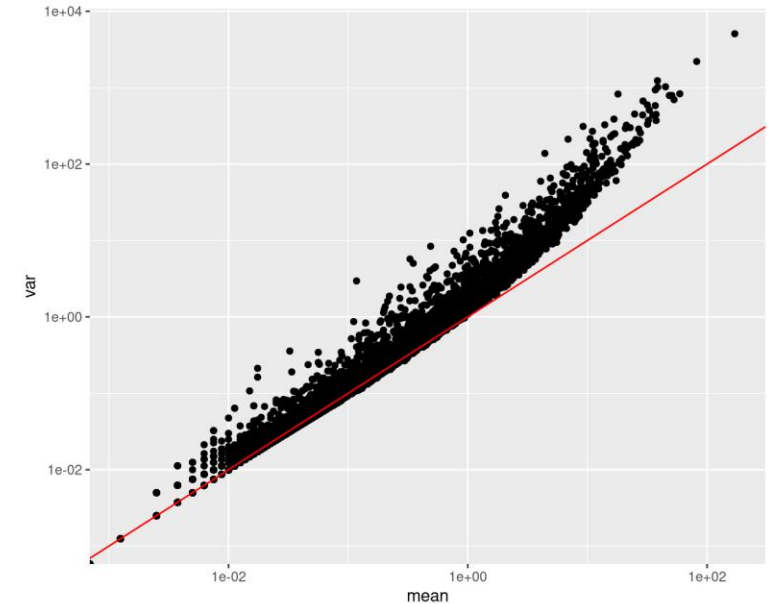
Limma-trend (2017)

logCPM, Mean-Variance Function, eBayes

MM

EdgeR

Limma



Limma.

For each genes.

$$\begin{cases} Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \epsilon \\ \epsilon \sim N(0, \sigma^2), Y_i \sim N(X\beta, \sigma^2) \end{cases}$$

Ordinary β t-test.

eBayes

$$\hat{\beta}_1 = (X^T X)^{-1} X^T Y$$

$$\hat{SE}(\hat{\beta}_1) = [\hat{\sigma}^2 (X^T X)^{-1}]_{1,1}$$

$$t\text{-statistics} = \hat{\beta}_1 / \hat{SE}(\hat{\beta}_1) = \hat{t}$$

$$P\text{-value} = \int_{\hat{t}}^{\infty} dF_t + \int_{-\infty}^{-\hat{t}} dF_t$$

$$\Rightarrow \tilde{SE}(\hat{\beta}_1) = [\tilde{\sigma}^2 (X^T X)^{-1}]_{1,1}$$

$$t\text{-statistic} = \hat{\beta}_1 / \tilde{SE}(\hat{\beta}_1) = \tilde{t}$$

$$P\text{-value} = \int_{\tilde{t}}^{\infty} dF_t + \int_{-\infty}^{-\tilde{t}} dF_t$$

Voom.

• RNA: Count Data. (e.g. 0, 1, 2, 3, 4, ...)

• Limma: Continuous Signal (e.g. -2.3, 6.7, -1.1, -8.8, ...)

• Support (Range) Difference.

• Mean-Variance Relationship of Count Data.

So, Voom is doing.

• $\log \text{CPM}$: $\log\left(\frac{\text{count}}{\sum(\text{count})}\right)$ making it continuous.

$$\sigma_{i2}^2 = \hat{f}(\mu_{i2}) \Rightarrow w_{i2} = \frac{1}{\sigma_{i2}^2} \Rightarrow \text{Weighted Least Square}$$

$$(Y - X\beta)^T W (Y - X\beta)$$

$\Rightarrow$ adjusting the importance of each dot at Model Fitting Procedure.

Trend.

• $\sigma_g = \hat{f}(\mu_g)$ onto eBayes. prior.

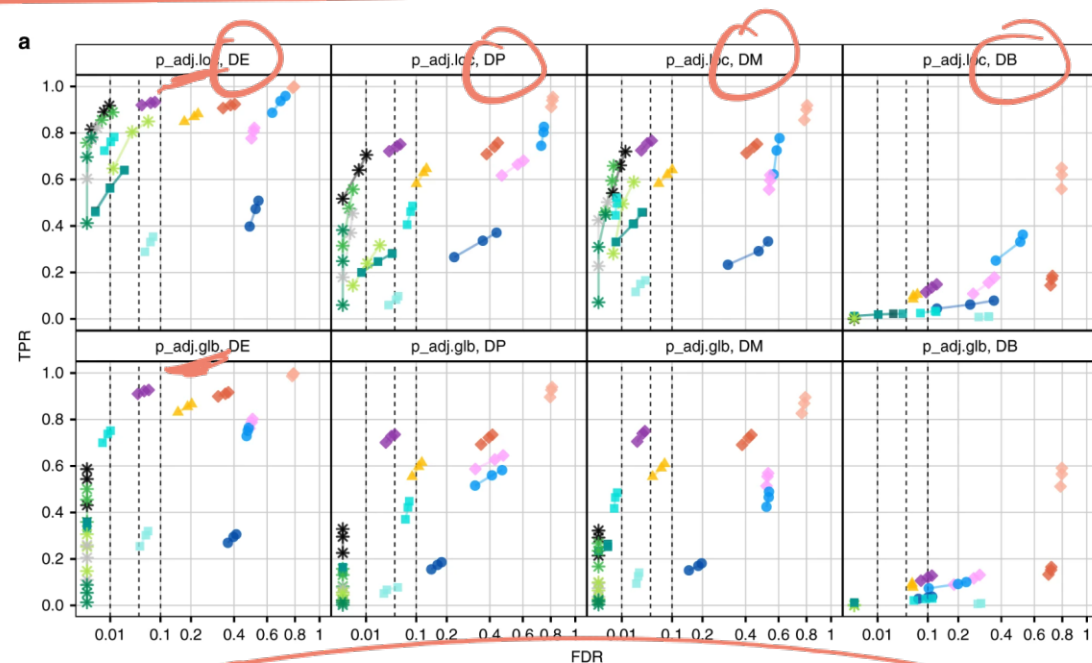
• $\hat{\sigma}_g = w \hat{\sigma}_g + (1-w) \hat{\sigma}_g \Rightarrow SE(\hat{\beta}) \Rightarrow t\text{-test.}$

Models

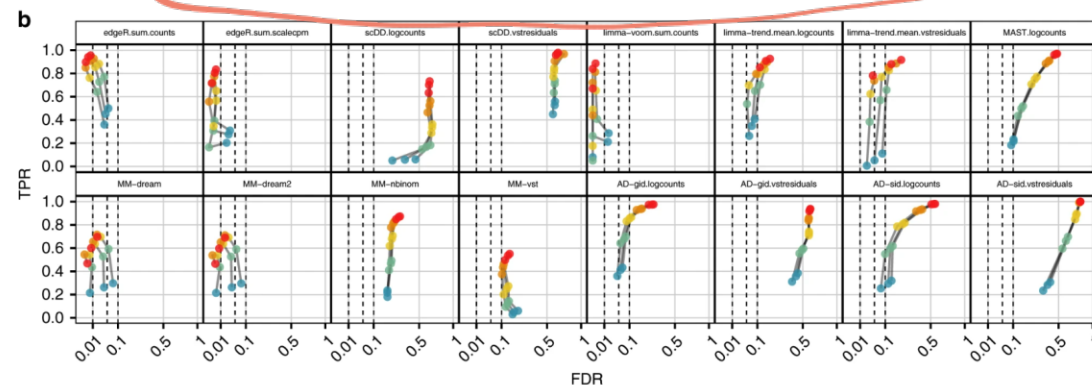
UBT

	Single Cell Approach	Model	Target	Kick
scDD	Cell Grouped	<u>Distribution (GMM)</u>	<u>Distribution Difference</u>	DPM, Bayes Factor
MAST	Cell Grouped	NB regression	<u>Fitness Difference</u>	NB, LRT
AD	Cell Grouped & Cell Ungrouped	Distribution (ECDF)	<u>Distribution Difference</u>	Tail Behavior, Nonparametric
MM	Cell Ungrouped	Linear Mixed Model	Expectation Difference	Random Effect Single Cell analysis
EdgeR	Pseudo Bulk	NB regression	Expectation Difference	eBayes
Limma	Pseudo Bulk	Linear Regression	Expectation Difference	Mean-Variance Function, eBayes, WLS

Figure 2. Result Summary



* edgeR.sum.counts
 edgeR.sum.scalecpm
 scDD.logcounts
 scDD.vstresiduals
 limma-voom.sum.counts
 limma-trend.mean.logcounts
 limma-trend.mean.vstresiduals
 MAST.logcounts
 MM-dream
 MM-dream2
 MM-rbinom
 MM-vst
 AD-gid.logcounts
 AD-gid.vstresiduals
 AD-sid.logcounts
 AD-sid.vstresiduals

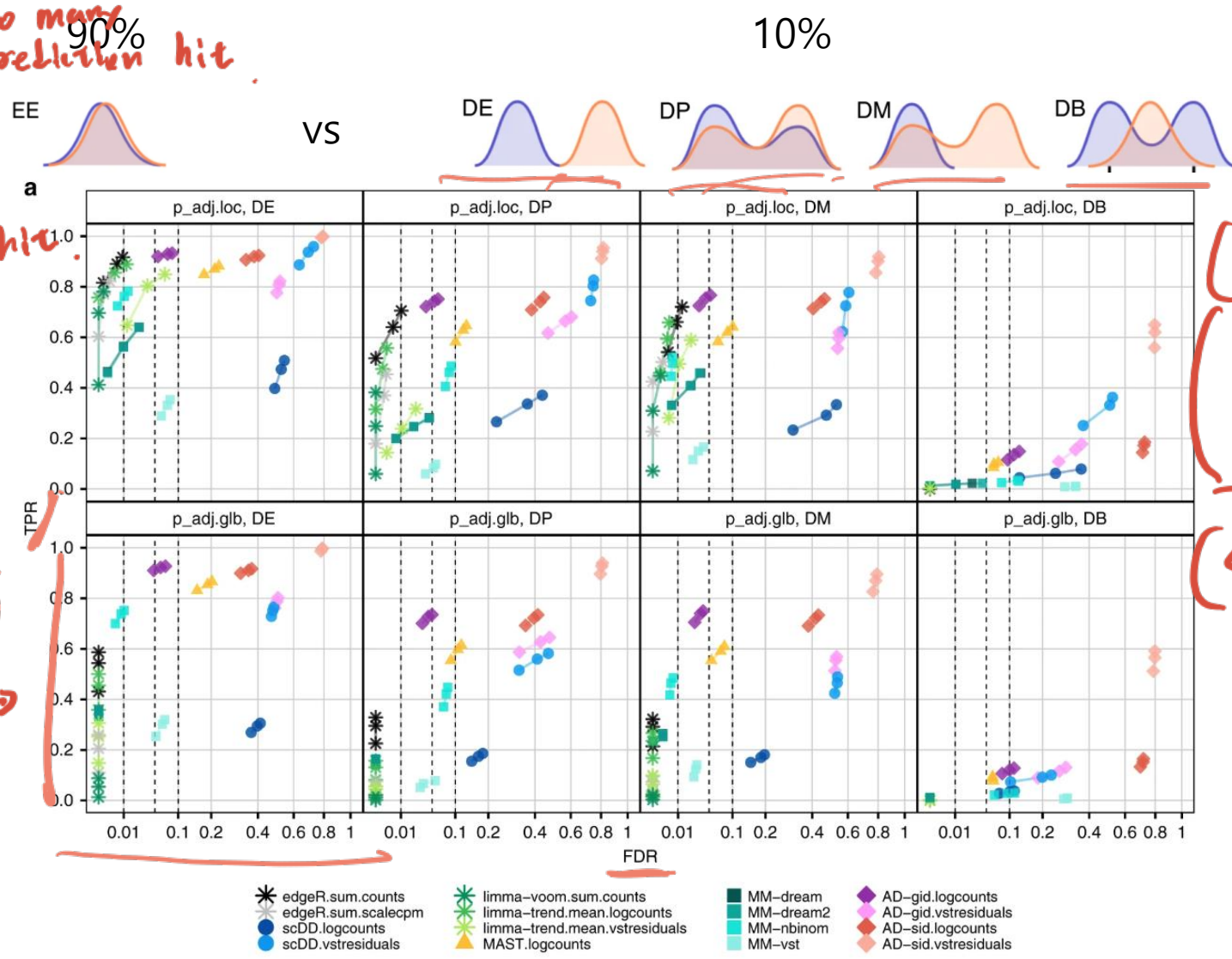


cells/subpopulation-sample

20 50 100 200 400

Too many
rejection hit
90%
to small hit

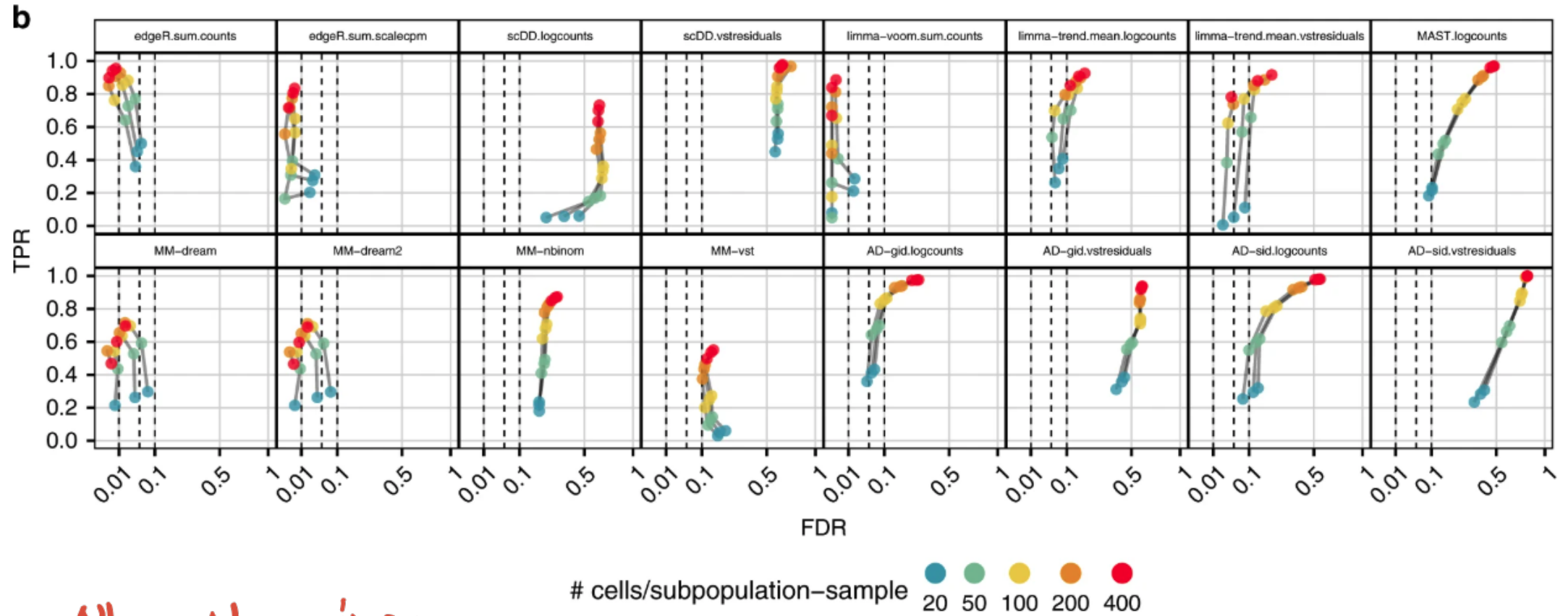
Best
Worst



TPR : Hitted DS genes/Total DS genes
FDR : False DS genes/Predicted DS genes

1% - 5% - 10% FDR threshold
loc: local FDR correction, glb: global FDR correction

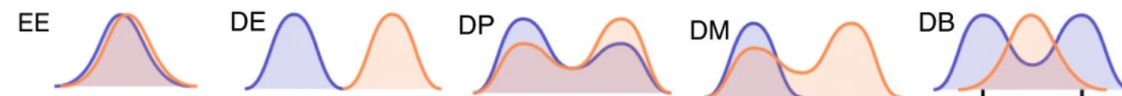
Gene Expression Scenario



• All of them increase in TPR.

• Direction ↖ ↗ ↘

Figure 3.



1. AD
2. edgeR + Limma
3. MM

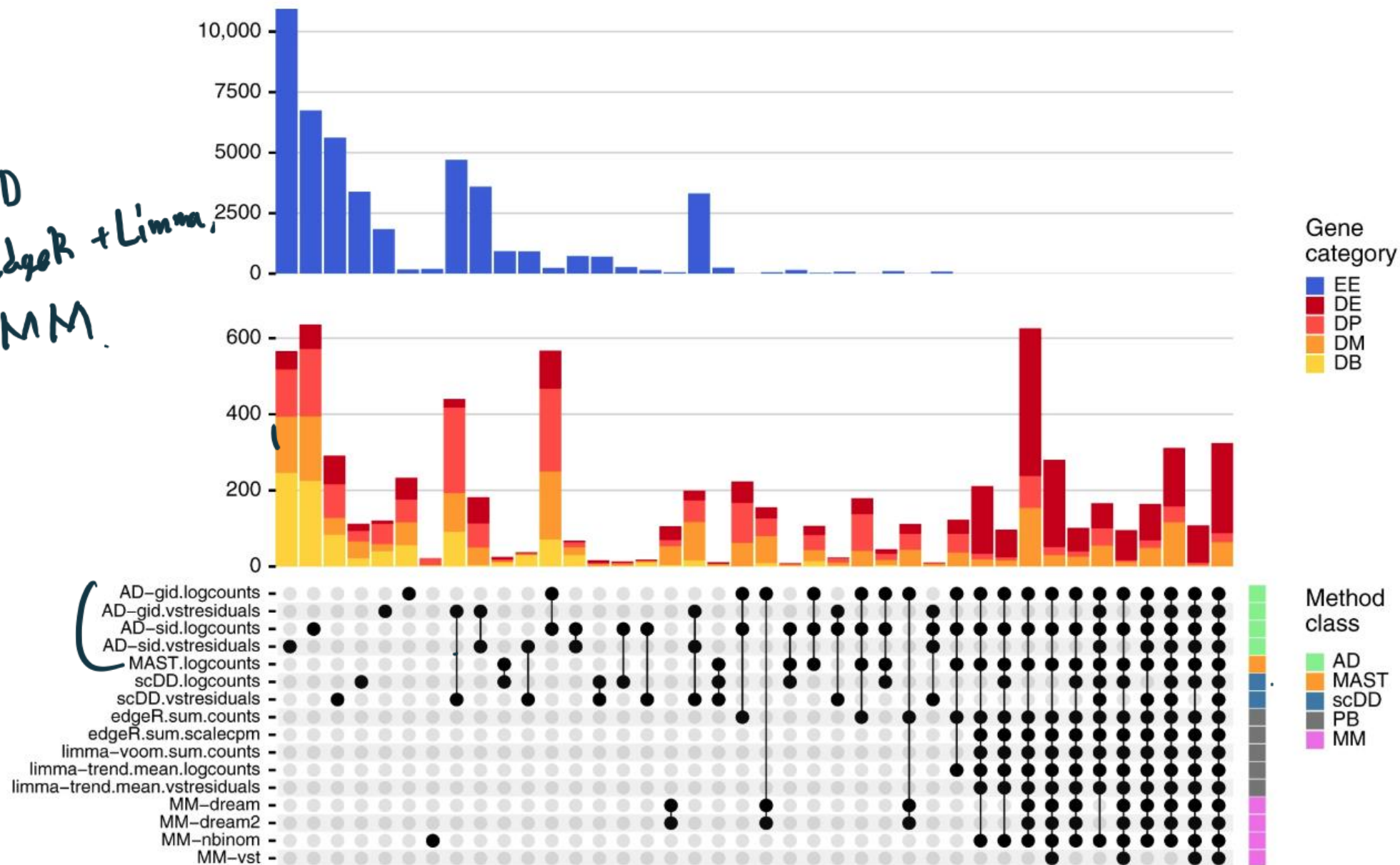
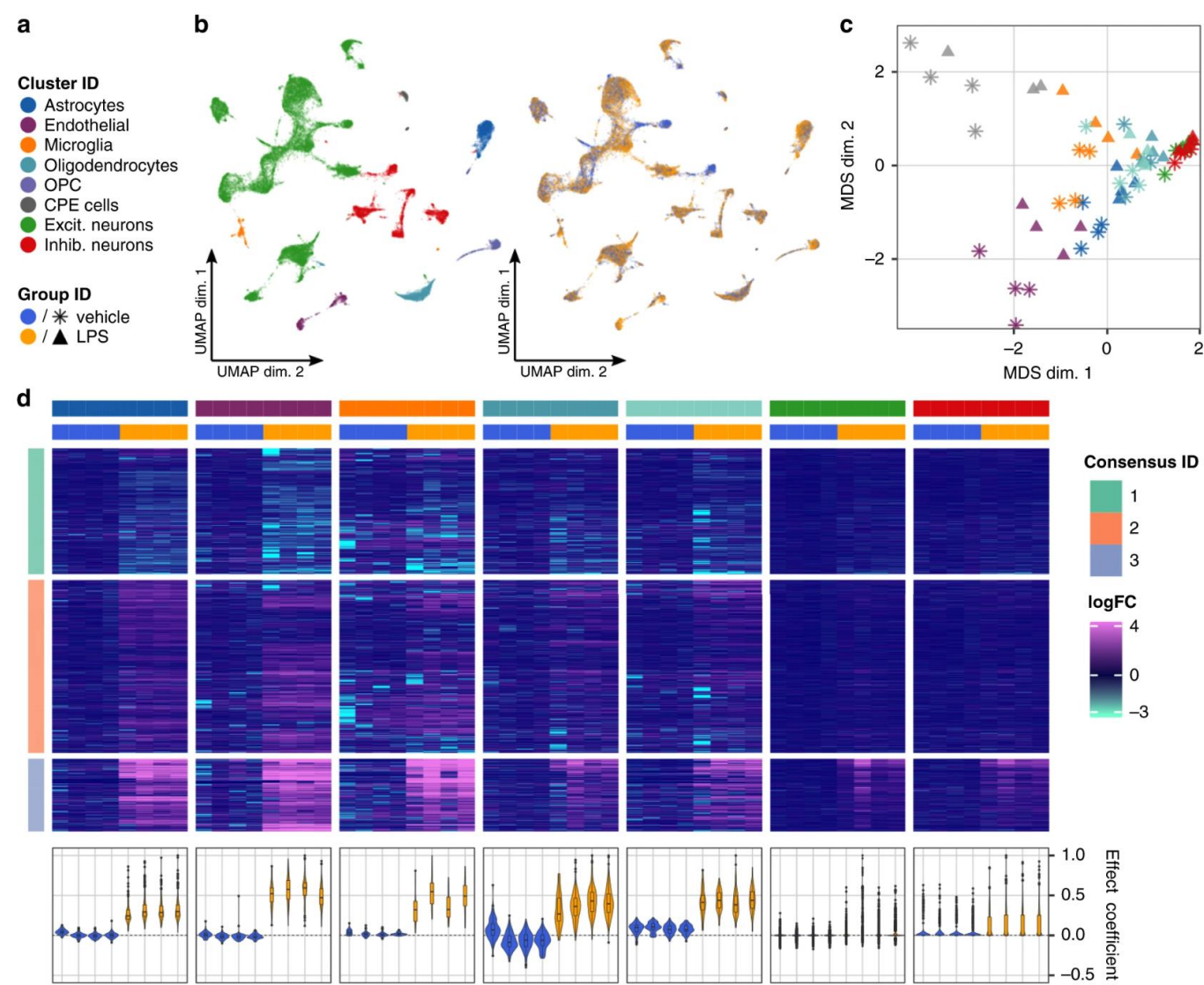


Figure 4. DS analysis of cortex tissue from vehicle- and LPS-treated mice



Expression Scenario

a

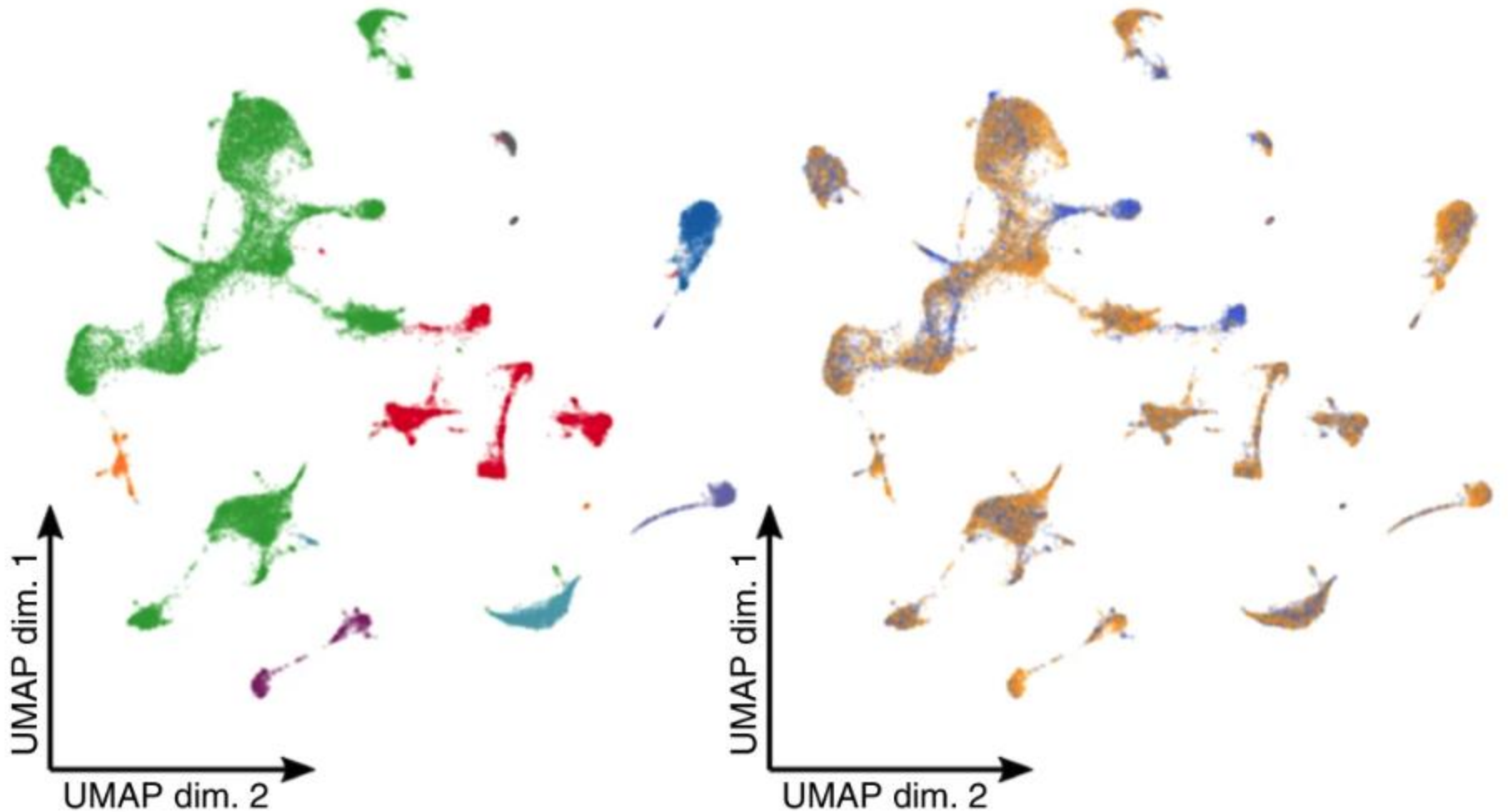
Cluster ID

- Astrocytes
- Endothelial
- Microglia
- Oligodendrocytes
- OPC
- CPE cells
- Excit. neurons
- Inhib. neurons

Group ID

- / * vehicle
- / ▲ LPS

b



Gene Expression Scenario

a

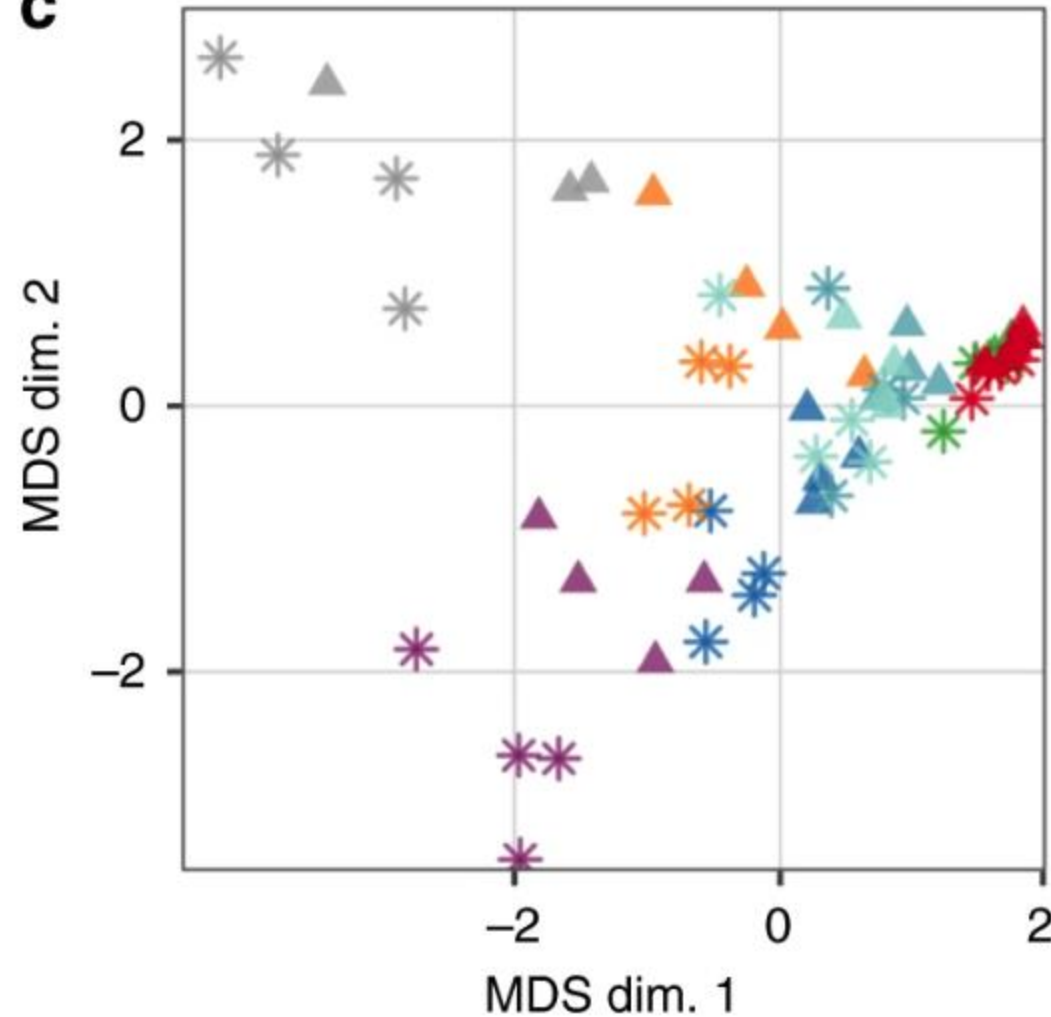
Cluster ID

- Astrocytes
- Endothelial
- Microglia
- Oligodendrocytes
- OPC
- CPE cells
- Excit. neurons
- Inhib. neurons

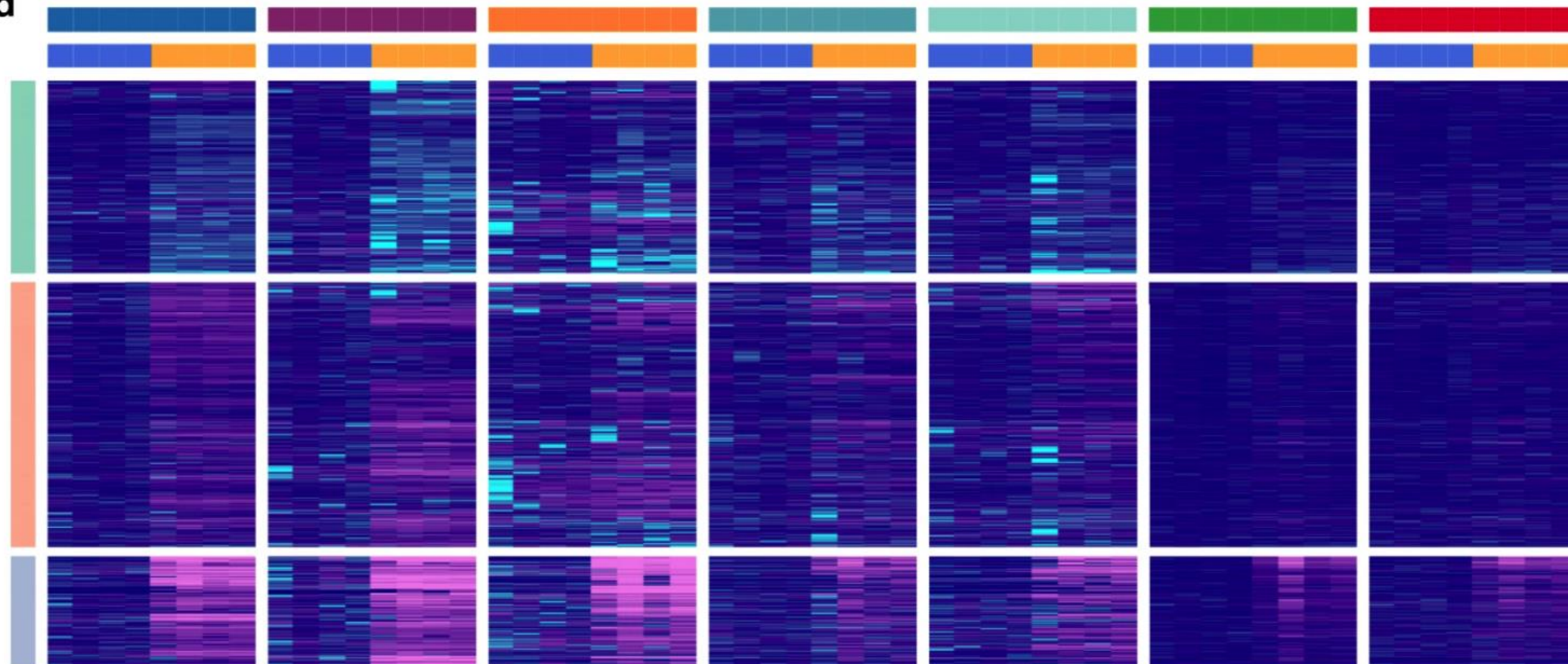
Group ID

- / * vehicle
- / ▲ LPS

c

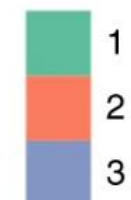


d

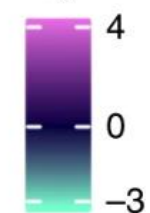


Edge R.

Consensus ID

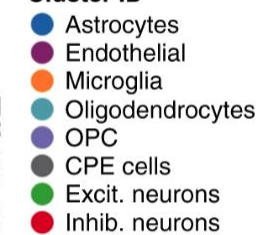


logFC



a

Cluster ID



Group ID

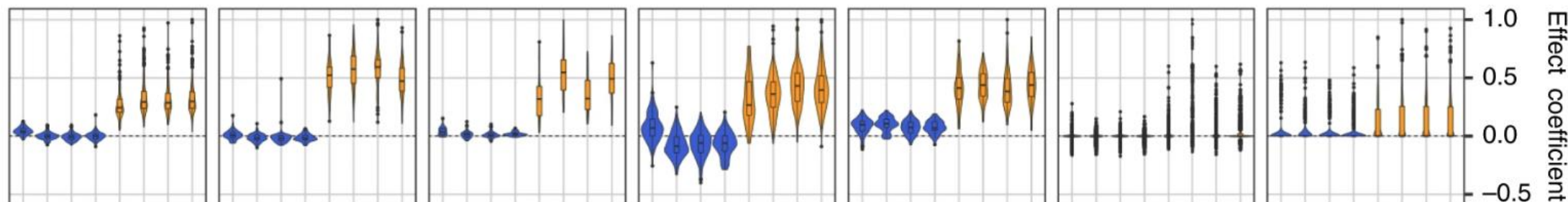
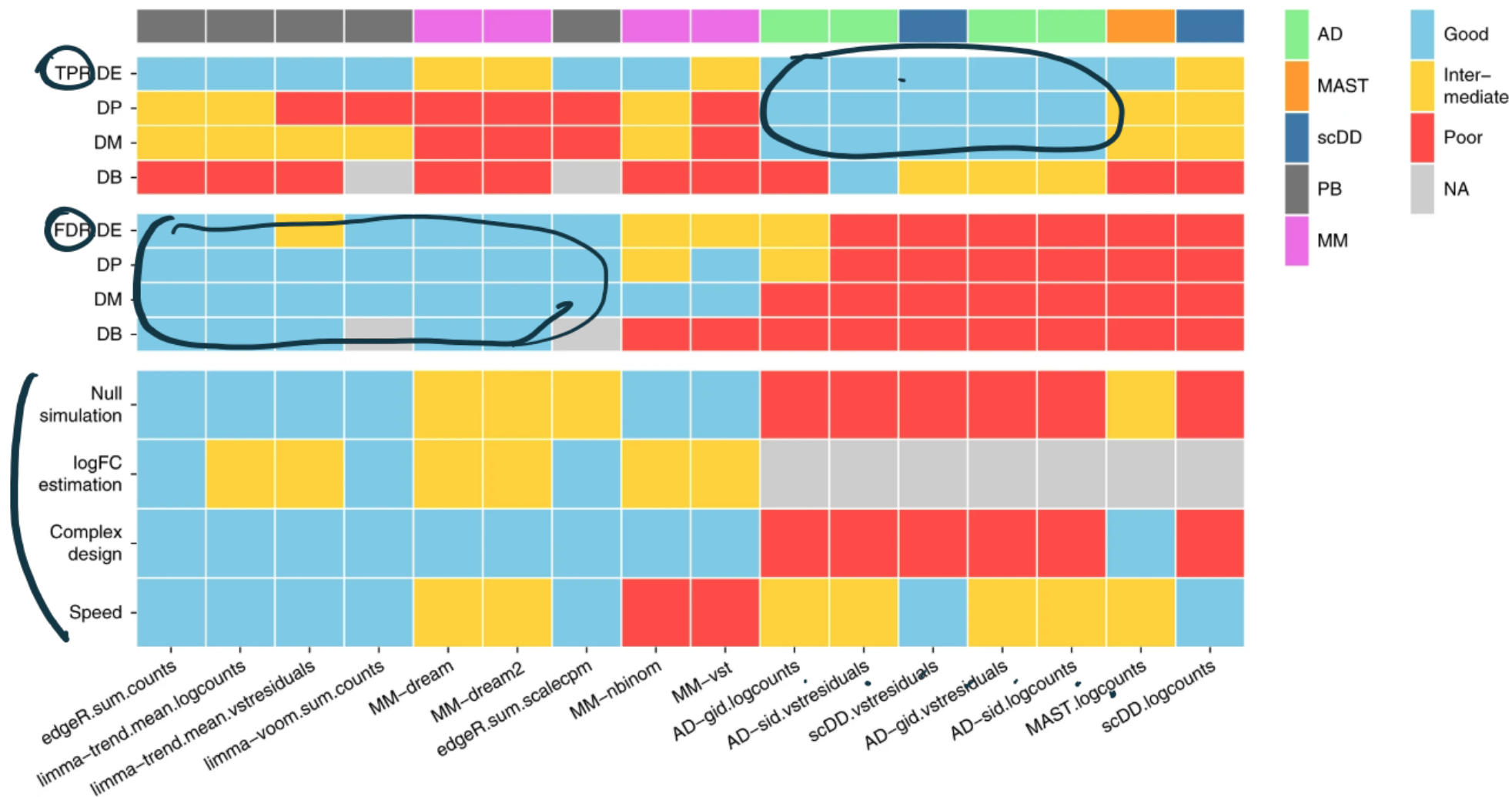


Figure 5.



Q&A