Inference after latent variable estimation for single-cell RNA sequencing data

Presenters:
Joel Mefford & Boyang Fu & Zeyuan Chen

Introduction

A common question to ask in scRNA-seq data:

What genes are differentially expressed among latent space (cell types, cell pseudotime)

One popular strategy to answer this question:

Step 1: Latent variable estimation

Step 2: Differential expression analysis

Notations

$\mathbf{X}$: a random variable describing the data distribution

X: a $n \times p$ realization of $\mathbf{X}$

X_{ij}: the number of reads from the ith cell and jth gene

L: a $n \times k$ latent variable that explains $E[\mathbf{X}]$

$\widehat{L}(X)$: estimated L using X

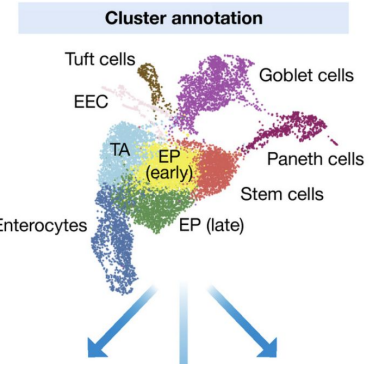
Workflow

X

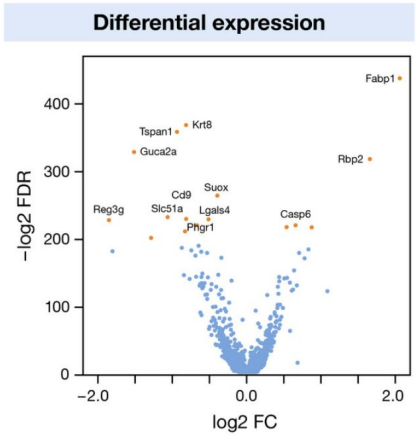
	Gene 1	Gene 2	…	Gene p
Cell 1	3	2	.	13
Cell 2	2		.	1
…			.	
Cell n	25	0	.	18

Latent variable estimation

$\widehat{L}(X)$



Differential expression analysis



Ref: https://hbctraining.github.io/scRNA-seq/lessons/02_SC_generation_of_count_matrix.html

The double dipping problem

Definition: The practice of using the same data *X* to first construct $\widehat{L}(X)$ and then to test *X* for association using $\widehat{L}(X)$.

Problem: Type I error is not guaranteed to be well-controlled.

Intuitive reason: The downstream analysis doesn't take into account the artifact created by the latent variable inference model

Generalized linear models (Poisson regression)

$$\mathbf{X}_{i,j} \sim Poisson(\Lambda_{ij})$$
$$\log(\Lambda_{ij}) = \beta_{0j} + \beta_{1j}^T L_i,\ i = 1, ..., n, j = 1, ..., p$$

A reasonable assumption if:

1. $\mathbf{X}_{i,j}$ is discrete & non negative
2. $\mathbf{X}_{i,j}$ has a mean approximately equal to the variance

Objective:

1. Get coefficients estimate $\hat{\beta}$ of the latent variables (using MLE)
2. Testing $\hat{\beta}_j \neq 0$

Double dipping approach: same data is used for both latent variable estimation and association analysis

X

	Gene 1	Gene 2	…	Gene p
Cell 1	3	2	.	13
Cell 2	2		.	1
…			.	
Cell n	25	0	.	18

Latent variable estimation

$\widehat{L}(X)$

Objective: Is gene *j* differentially expressed across samples. (E.g. Does gene j separate clusters?)

$$\hat{\beta}(L, X_j) \neq 0$$

$$\Pr_{H_0:\beta_1(\hat{L}(X),\mathbf{X}_j)=0}\left(\left|\hat{\beta}_1(\hat{L}(X),\mathbf{X}_j)\right| \geq \left|\hat{\beta}_1\left(\hat{L}(X), X_j\right)\right|\right)$$

Attempt 1: Cell Splitting approach

X

	Gene 1	Gene 2	…	Gene p
Cell 1	3	2	.	13
Cell 2	2		.	1
…			.	
Cell n	25	0	.	18

$X^{train} \in \mathbb{N}_0^{n_1 \times p}$

Latent variable estimation

$\hat{L}(X_{train}, .)$

Project the test dataset $X^{test} \in \mathbb{N}_0^{n_2 \times p}$

$X^{test} \in \mathbb{N}_0^{n_2 \times p}$

$\widehat{L}(X_{train}, X_{test})$

$$\Pr_{H_0:\beta_1(\hat{L}(X^{train},X^{test}),\mathbf{X}_j^{test})=0}\left(\left|\hat{\beta}_1(\hat{L}(X^{train},X^{test}),\mathbf{X}_j^{test})\right| \geq \left|\hat{\beta}_1\left(\hat{L}(X^{train},X^{test}),X_j^{test}\right)\right|\right)$$

Attempt 2: Gene splitting approach

X

	Gene 1	Gene 2	...	Gene p
Cell 1	3	2		13
Cell 2	2			1
...				
Cell n	25	0		18

X^{train} X^{test}

Latent variable estimation

$\widehat{L}(X^{train})$

How much does the genes from the training sets tells the information of the differential expression of the genes in the testing sets?

$$\Pr_{H_0:\ \beta_1\left(\widehat{L}(X^{\text{train}}),\mathbf{X}_j^{\text{test}}\right)=0}\left(\left|\widehat{\beta}_1\left(\widehat{L}(X^{\text{train}}),\mathbf{X}_j^{\text{test}}\right)\right| \geqslant \left|\widehat{\beta}_1\left(\widehat{L}\left(X^{\text{train}}\right),X_j^{\text{test}}\right)\right|\right)$$

Count splitting (1): Generate two sets with transferable latent information

	Gene 1	Gene 2	…	Gene p
Cell 1	3	2	.	13
Cell 2	2		.	1
…			.	
Cell n	25	0	.	18

X

	Gene 1	Gene 2	…	Gene p
Cell 1	2	1	.	8
Cell 2	2		.	0
…			.	
Cell n	15	0	.	9

X^{train}

	Gene 1	Gene 2	…	Gene p
Cell 1	3 - 2	2 - 1	.	13 - 8
Cell 2	2 - 2		.	1 - 0
…			.	
Cell n	25 - 15	0 - 0	.	18 - 9

$X^{test} = X - X^{train}$

Latent variable estimation

$\widehat{L}(X^{train})$

Differential expression analysis

$$Pr_{H_0:\ \beta_1\left(\widehat{L}(X^{\text{train}}),\mathbf{X}_j^{\text{test}}\right)=0}\left(\left|\widehat{\beta}_1\left(\widehat{L}(X^{\text{train}}),\mathbf{X}_j^{\text{test}}\right)\right| \geqslant \left|\widehat{\beta}_1\left(\widehat{L}\left(X^{\text{train}}\right),X_j^{\text{test}}\right)\right|\right)$$

Count splitting (2): Generate two sets with mutually independent property

$$\mathbf{X} = \mathbf{X}^{train} \in \mathbb{N}_0^{n \times p} + \mathbf{X}^{test} \in \mathbb{N}_0^{n \times p}$$

Poisson $\mathbf{X}_{ij}^{train} | \{\mathbf{X}_{ij} = X_{ij}\} \sim \text{Binomial}(X_{ij}, \epsilon)$ $\mathbf{X}_{ij}^{test} | \{\mathbf{X}_{ij} = X_{ij}\} \sim \text{Binomial}(X_{ij}, 1-\epsilon)$

Binomial thinning of Poisson process: $\begin{cases} \mathbf{X}_{ij}^{train} \sim Poisson(\epsilon \Lambda_{ij}) \\ \mathbf{X}_{ij}^{test} \sim Poisson((1-\epsilon)\Lambda_{ij}) \end{cases}$

Real Data

Overdispersion

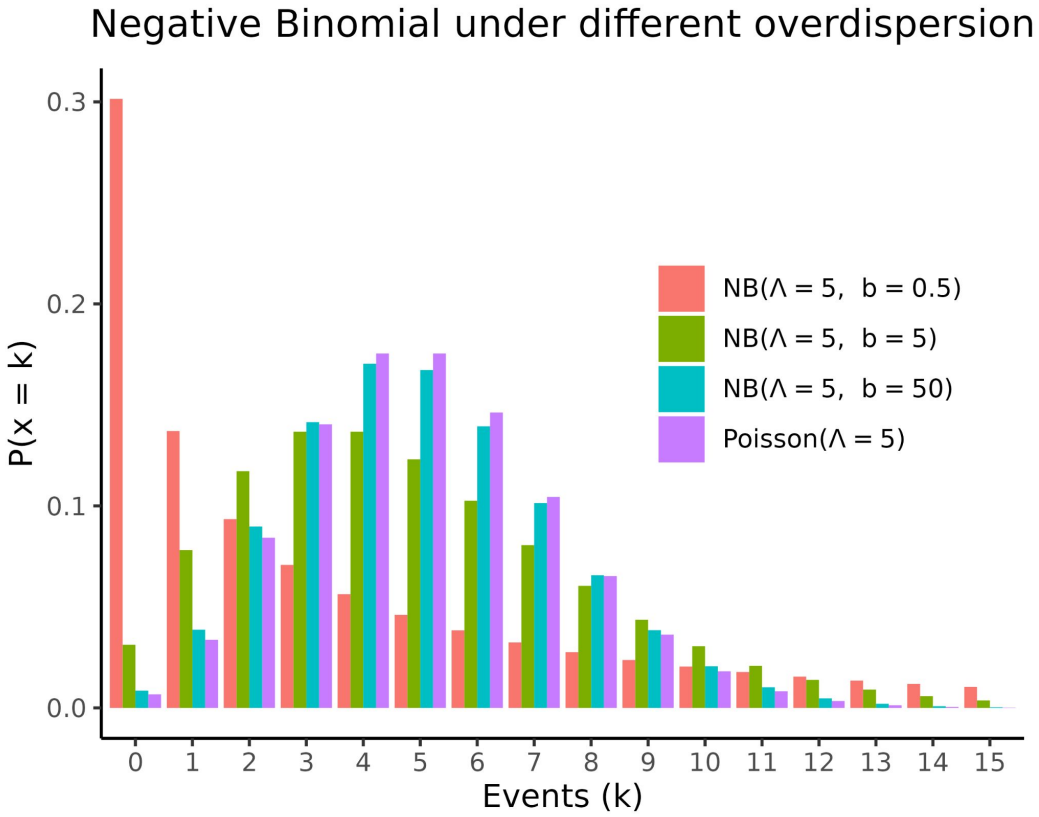
Count Splitting on PBMC 3k and Mouse Brain

Overdispersion (NB parametrization)

$$X \sim Poisson(\Lambda)$$
$$\mathbf{E}(X) = \mathbf{Var}(X) = \Lambda$$

$$X \sim NB(\Lambda, b)$$
$$\mathbf{E}(X) = \Lambda$$
$$\mathbf{Var}(X) = \Lambda + \frac{\Lambda^2}{b}$$

Smaller b indicates larger deviation
Larger b behaves similar to Poisson



Overdispersion (Correlation between Train and Test)

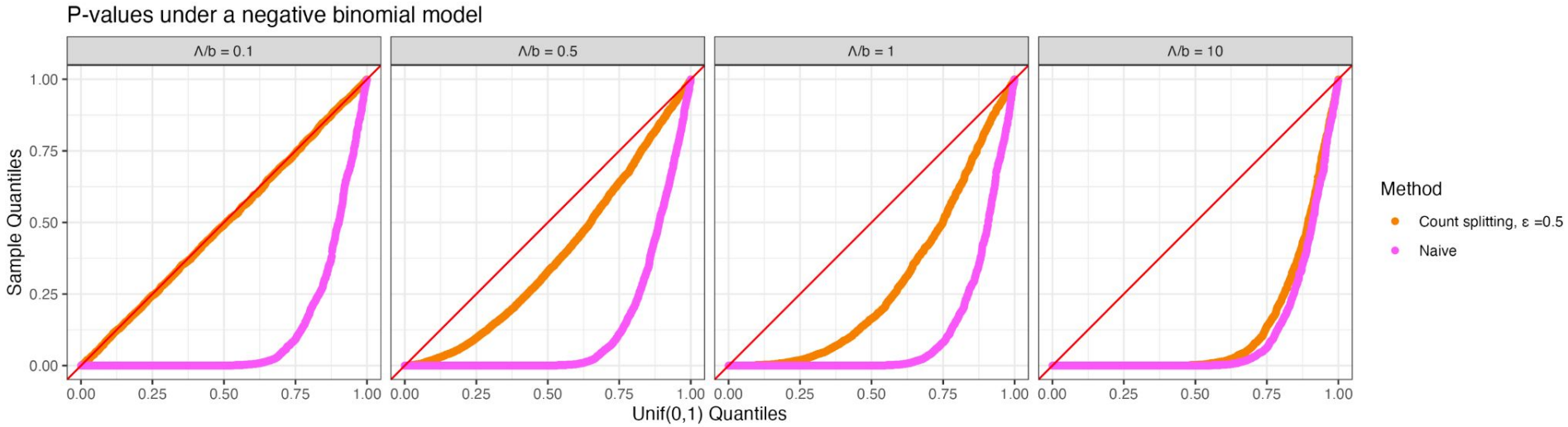
The independence between $\mathbf{X}^{train}$ and $\mathbf{X}^{test}$ no longer holds

Proposition 2 Suppose that $\mathbf{X}_{ij}$ follows a negative binomial distribution with expected value Λ_{ij} and variance $\Lambda_{ij} + \frac{\Lambda_{ij}^2}{b_j}$. If we perform Step 0 of Algorithm 1, then

$$\mathrm{Cor}\left(\mathbf{X}_{ij}^{\mathrm{train}}, \mathbf{X}_{ij}^{\mathrm{test}}\right) = \frac{\sqrt{\epsilon(1-\epsilon)}}{\sqrt{\epsilon(1-\epsilon) + \frac{b_j^2}{\Lambda_{ij}^2} + \frac{b_j}{\Lambda_{ij}}}}. \tag{4.15}$$

$\frac{\Lambda_{ij}}{b_j}$ ↓ $\quad$ $\mathrm{Cor}\left(\mathbf{X}_{ij}^{\mathrm{train}}, \mathbf{X}_{ij}^{\mathrm{test}}\right)$ ↓

Overdispersion - simulation study



Simulated with n = 200, p = 10,

Lambda = 5, b = {50, 10, 5, 0.5}

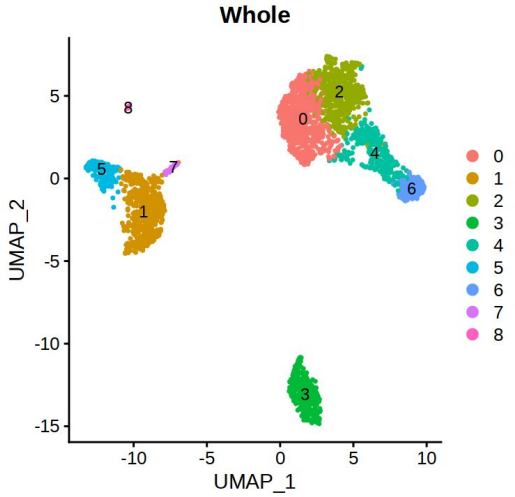
X drawn from same distribution

DE: GLM with NB, Wald test on coefficient related to estimated L

Estimating overdispersion coefficient (double-dipped)

Preprocess X and Learn L:

control coverage
feature selection
Low dim embedding
Clustering/Pseudotime -> L.hat



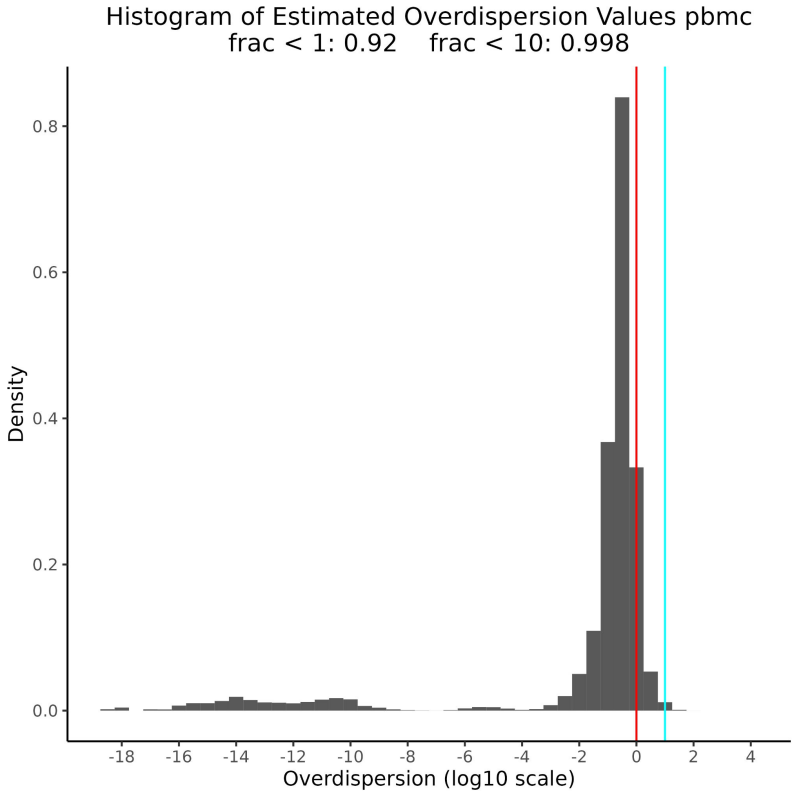
$$X_{ij} \sim NB(\Lambda_{i,j}, b_j)$$

$$\log(\Lambda_{i,j}) = \log \gamma_i + \beta_{0j} + \beta_{1j}^T \hat{L}_i$$

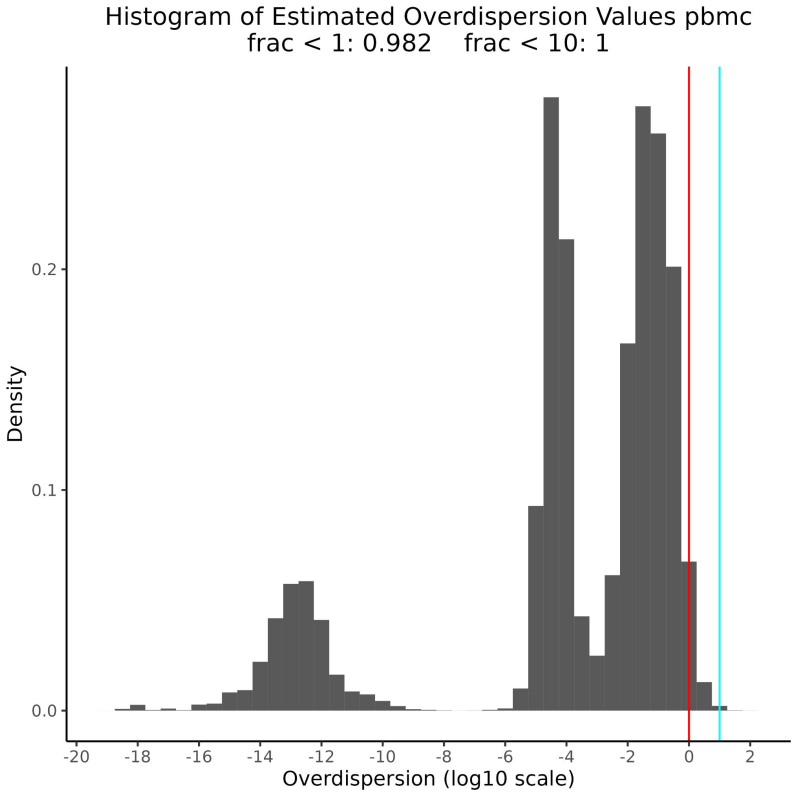
```
MASS::glm.nb(X[,gene_j] ~ L.hat + offset(log(size.factors)))
```

Estimate $\frac{\Lambda_{ij}}{b_j}$

PBMC-3k: overdispersion analysis

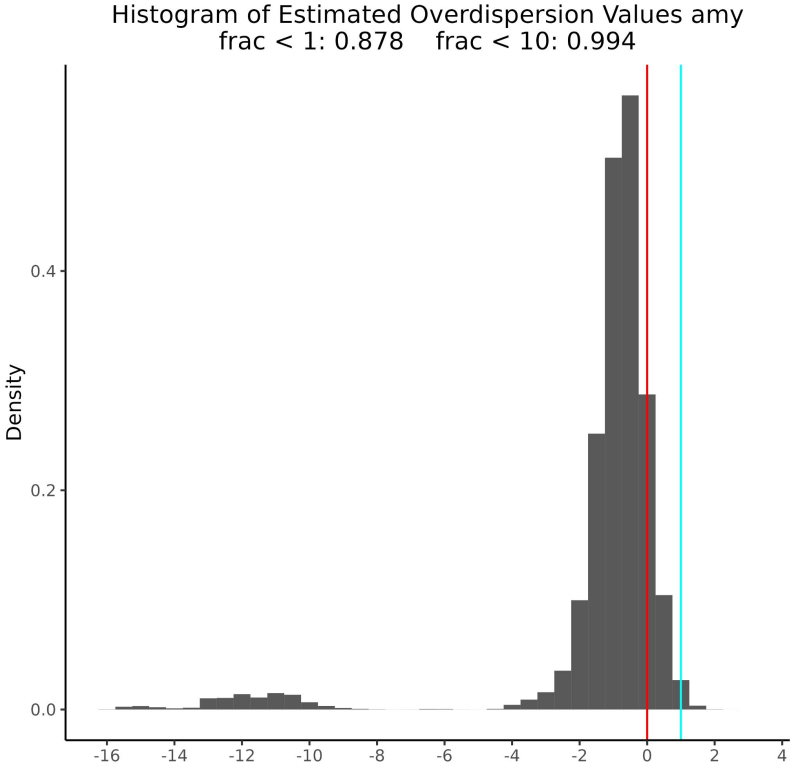


HVG 2000

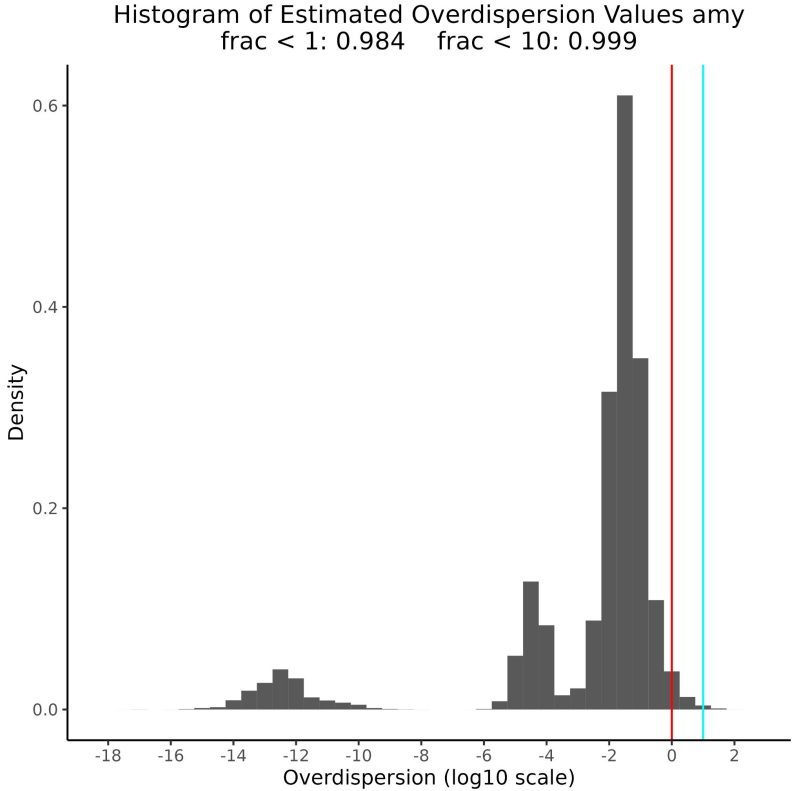


Random 2000

Mouse Brain: over-dispersion analysis

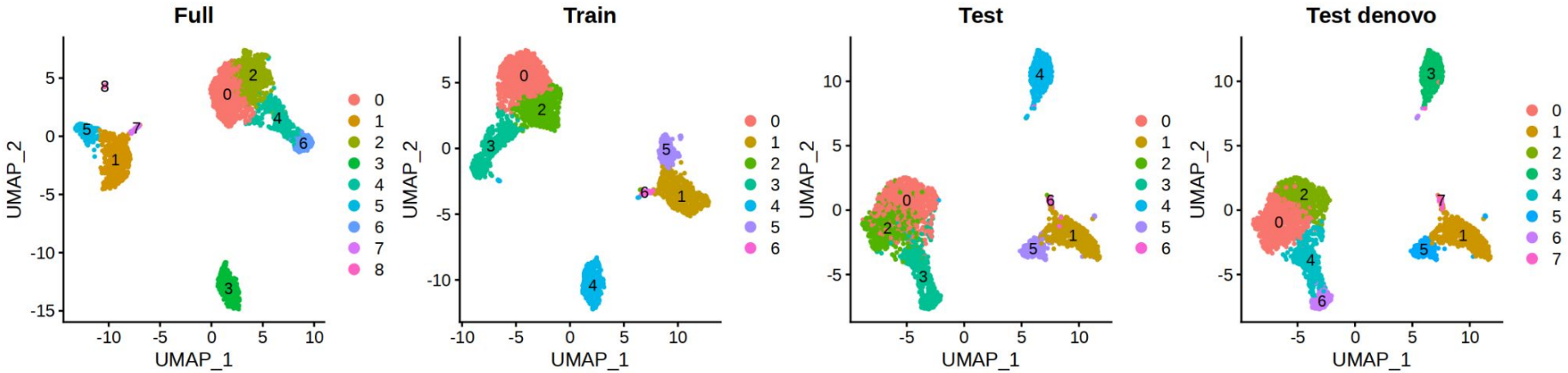


HVG 2000



Random 2000

PBMC-3k: DE analysis

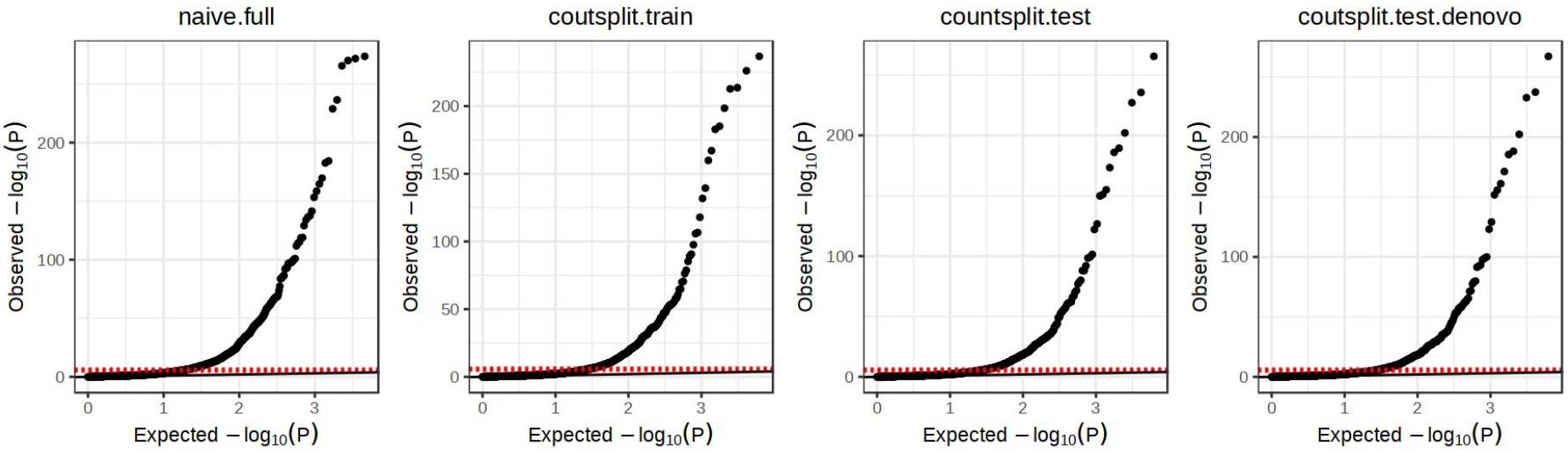


Full: using entire data to cluster and perform DE. so we double dipped entire data
Train: using half of the data (aka the train part) to cluster and perform DE
Test: using the test part of the count but copy the labels/cluster annotation from train
Test denovo: same as test except we also using test to obtain labels. so we double dipped test part of the data

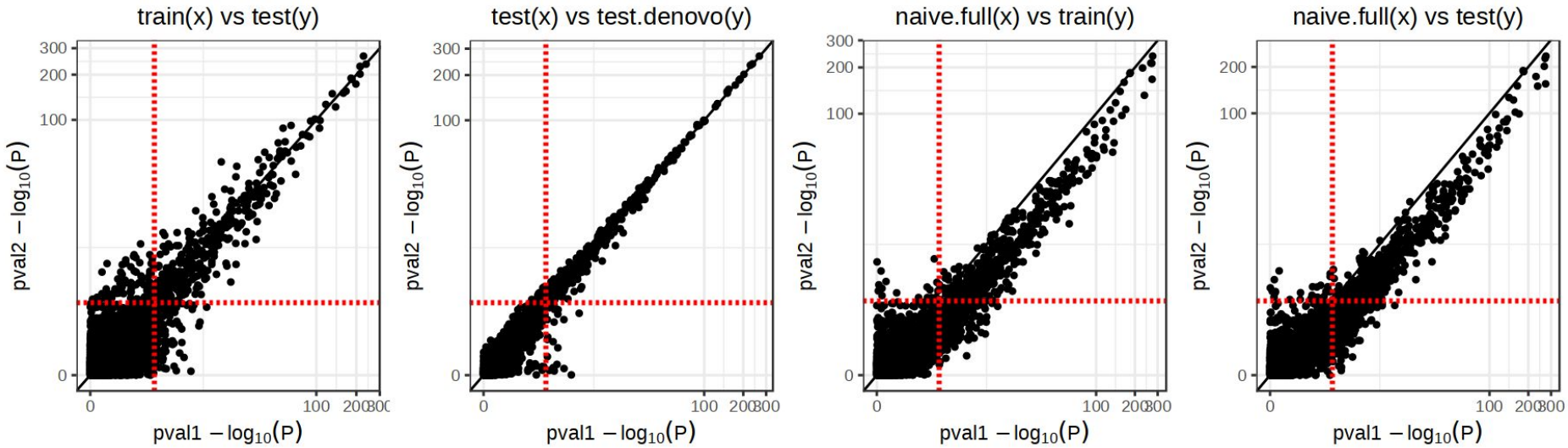
Note:

(1) For distinct cluster 4 (B cells) in train, it is extremely well preserved in test
(2) For CD subtypes: cluster 2 (Memory CD4+) and cluster 0 (Naive CD4+ T) in train, they are less well preserved in test

PBMC-3k: DE analysis (B vs rest)

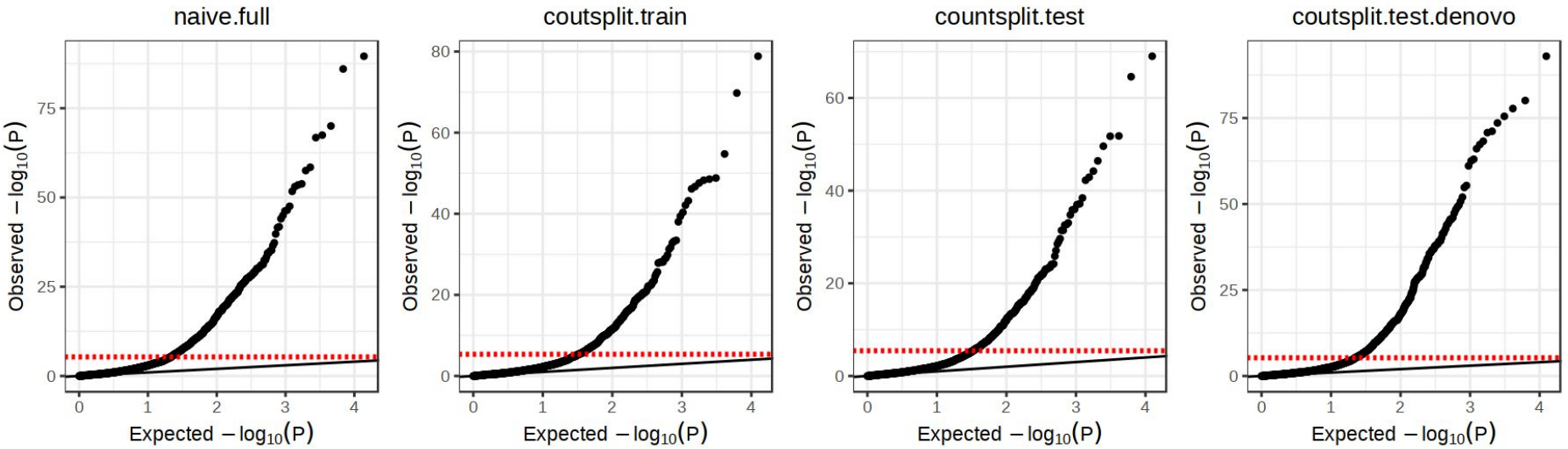


PBMC-3k: DE analysis (B vs rest)

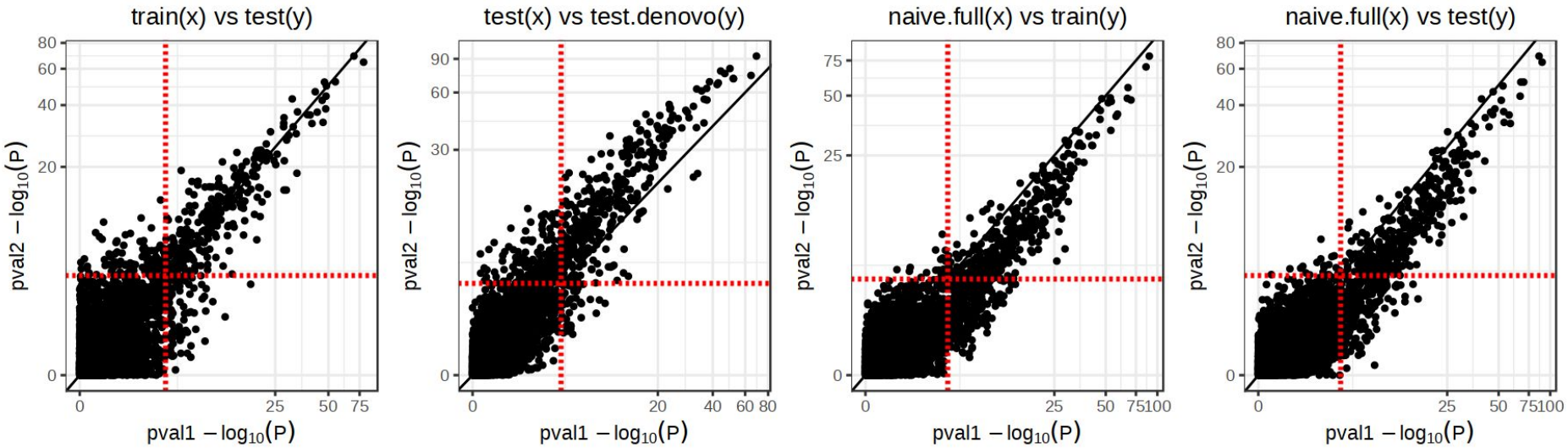


Almost no difference at all

PBMC-3k: DE analysis (memory CD4 + vs rest)

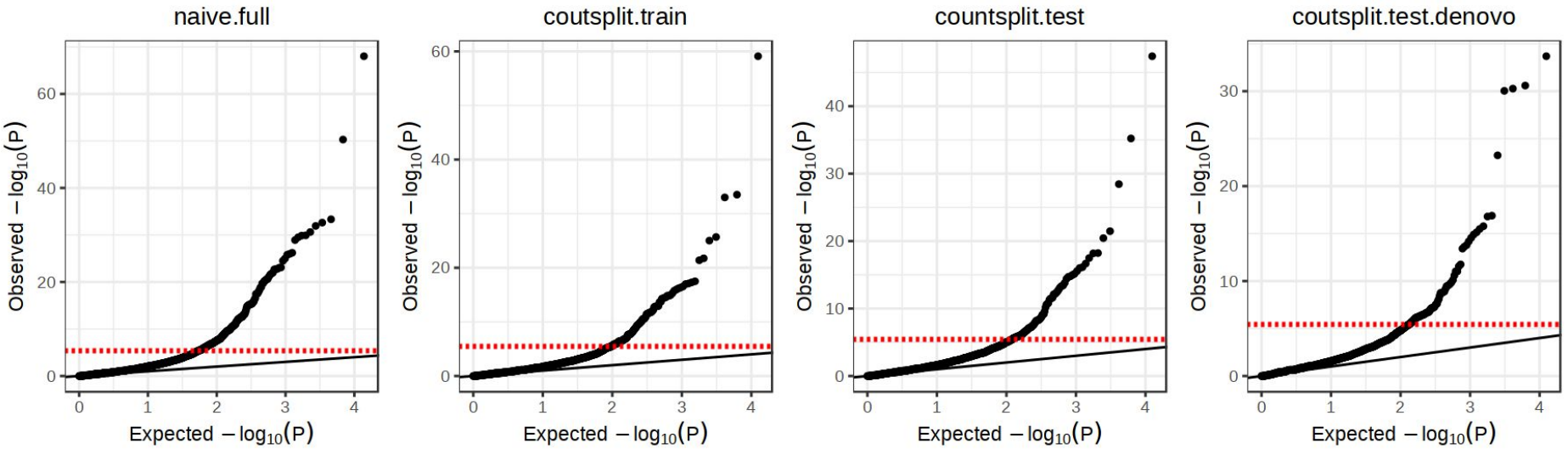


PBMC-3k: DE analysis (memory CD4 + vs rest)

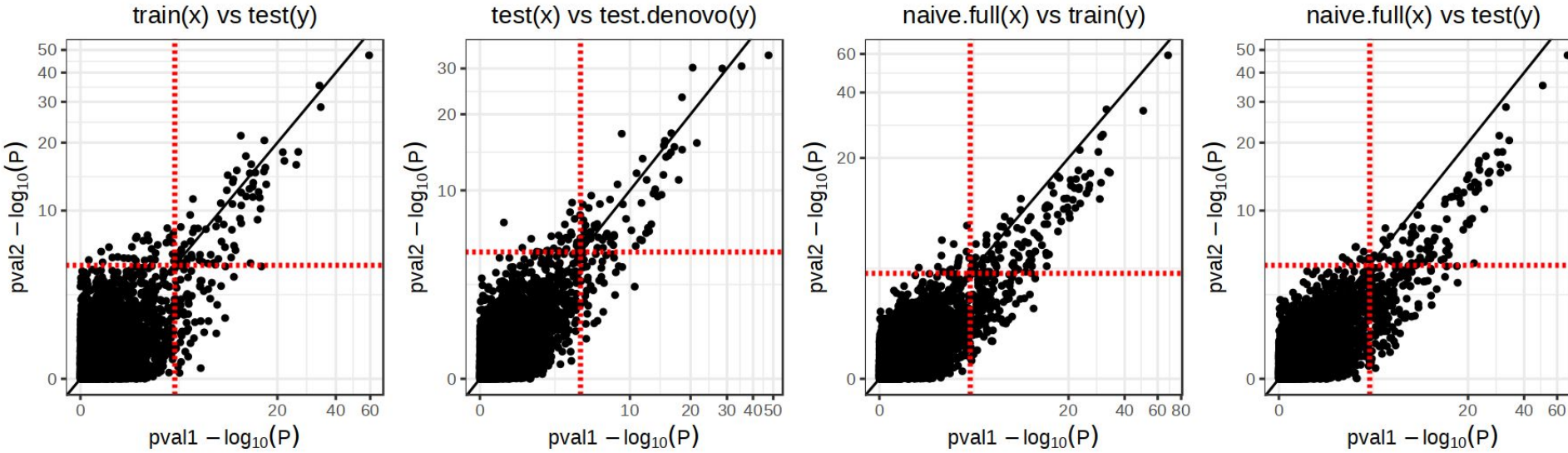


Double dipped test is more inflated

PBMC-3k: DE analysis (memory CD4 + vs naive CD4+)



PBMC-3k: DE analysis (memory CD4 + vs naive CD4+)



Mild difference

Resource

Fitting GLM with Poisson and Negative Binomial Tutorial in R

CountSplit Paper

CountSplit Package Installation and Tutorial

Codes to Replicate Paper Figures

Notebook and codes used in this presentation will also be made available soon !