# Single Cell RNA Sequencing

## Analysis and Applications

Jacob M. Maronge

Department of Statistics
University of Wisconsin–Madison

https://jmmaronge.github.io/
jmmaronge
@jmmaronge

# Why RNA sequencing?

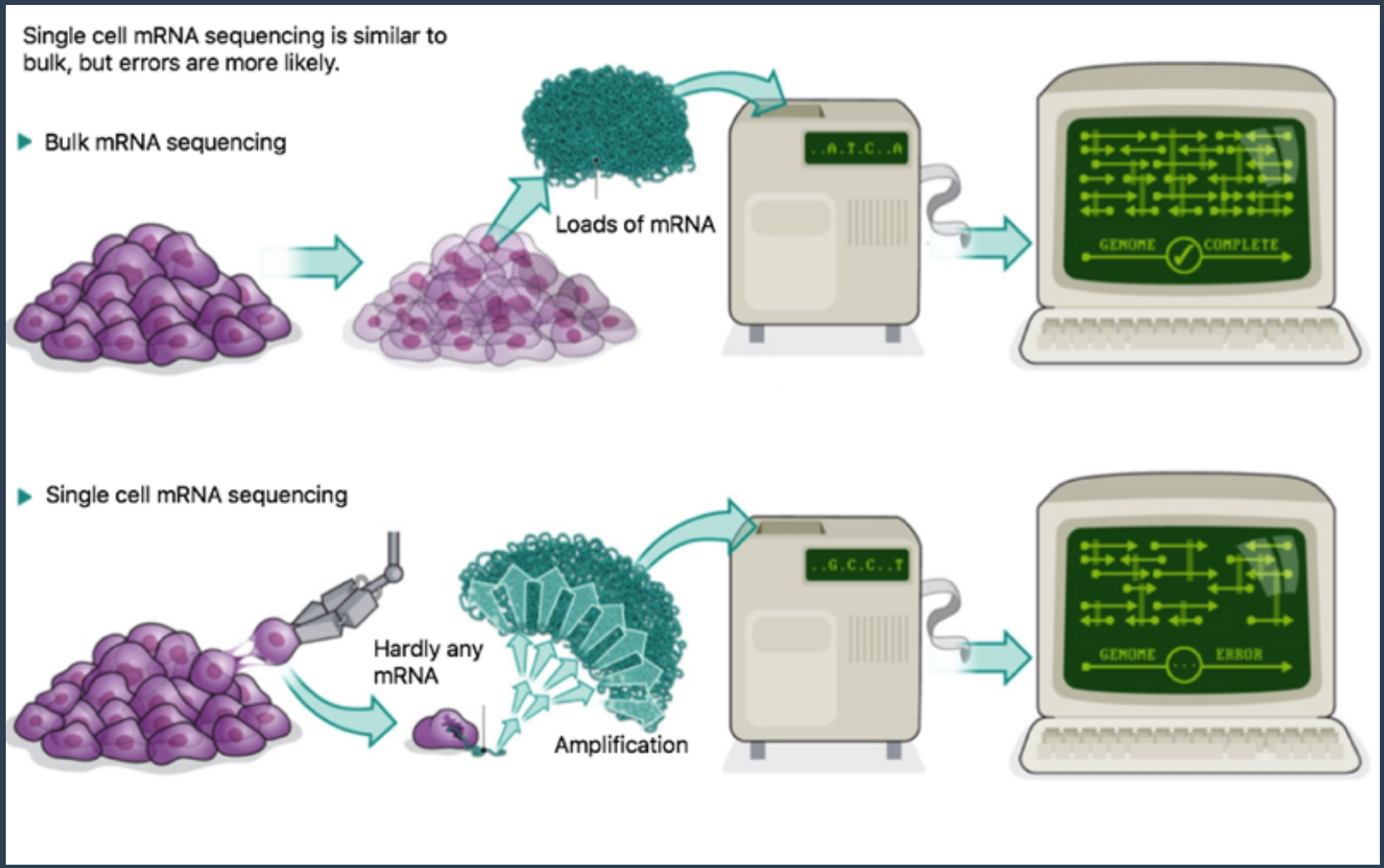
## Central dogma of molecular biology

$$\text{DNA} \rightarrow \text{mRNA} \rightarrow \text{Proteins} (\rightarrow \text{Traits})$$

- Hard to measure proteins
- Measure mRNA as an attempt to get at traits
- 2 ways to measure mRNA gene expression abundance: Bulk and Single Cell

# RNA sequencing

## Difference between bulk and single cell




From: http://web.stanford.edu/class/cs262/presentations/lecture12.pdf

# Single cell RNA-seq

## Data structure

$$
\begin{array}{c}
 & \begin{array}{cccccccc} cell_{1,1} & cell_{2,1} & \dots & cell_{j_1,1} & cell_{1,2} & cell_{2,2} & \dots & cell_{j_2,2} \end{array} \\
\begin{array}{c} gene_1 \\ gene_2 \\ \vdots \\ gene_i \end{array} &
\begin{pmatrix}
n_{1,1,1} & n_{1,2,1} & \dots & n_{1,j_1,1} & n_{1,1,2} & n_{1,2,2} & \dots & n_{1,j_2,2} \\
n_{2,1,1} & n_{2,2,1} & \dots & n_{2,j_1,1} & n_{2,1,2} & n_{2,2,2} & \dots & n_{2,j_2,2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
n_{i,1,1} & n_{i,2,1} & \dots & n_{i,j_1,1} & n_{i,1,2} & n_{i,2,2} & \dots & n_{i,j_2,2}
\end{pmatrix}
\end{array}
$$

- $n_{i,j,k}$ refers to the count of abundance of gene expression in gene $i$, cell $j$, and condition $k$.

# Single cell RNA-seq

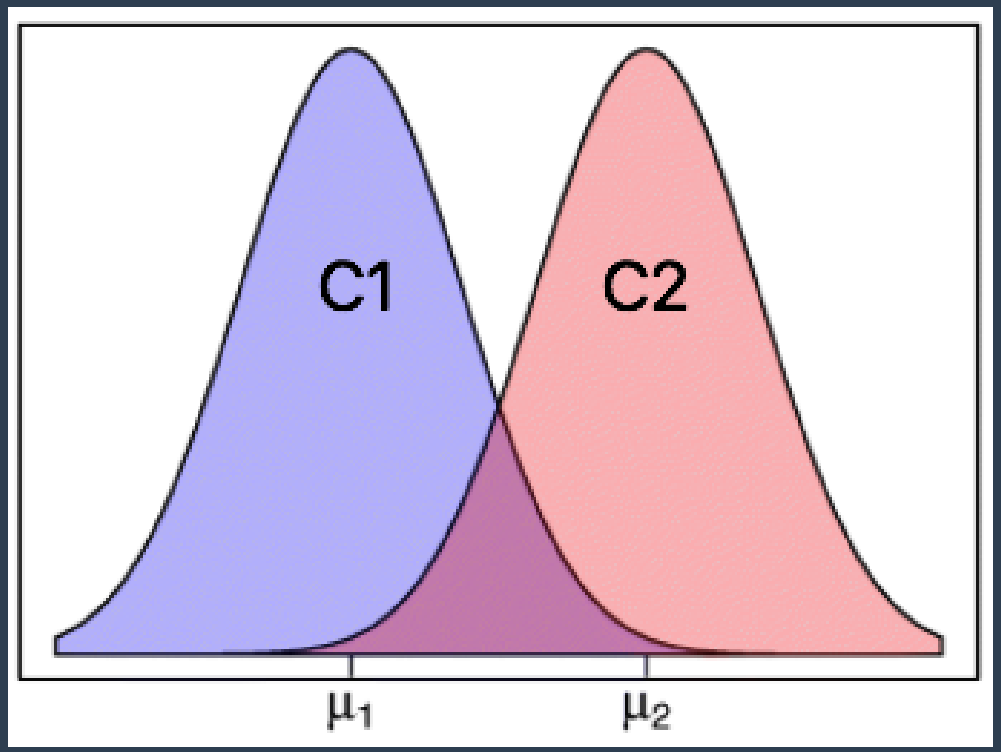
## This semester

- Learned about and found a problem in a statistical analysis technique for single cell RNA sequencing (scDD)
- Learned about the problem of normalization in single cell RNA sequencing.
- Performed a literature review for the problem of identifying cell subpopulations.

# Questions for Statistical Analysis

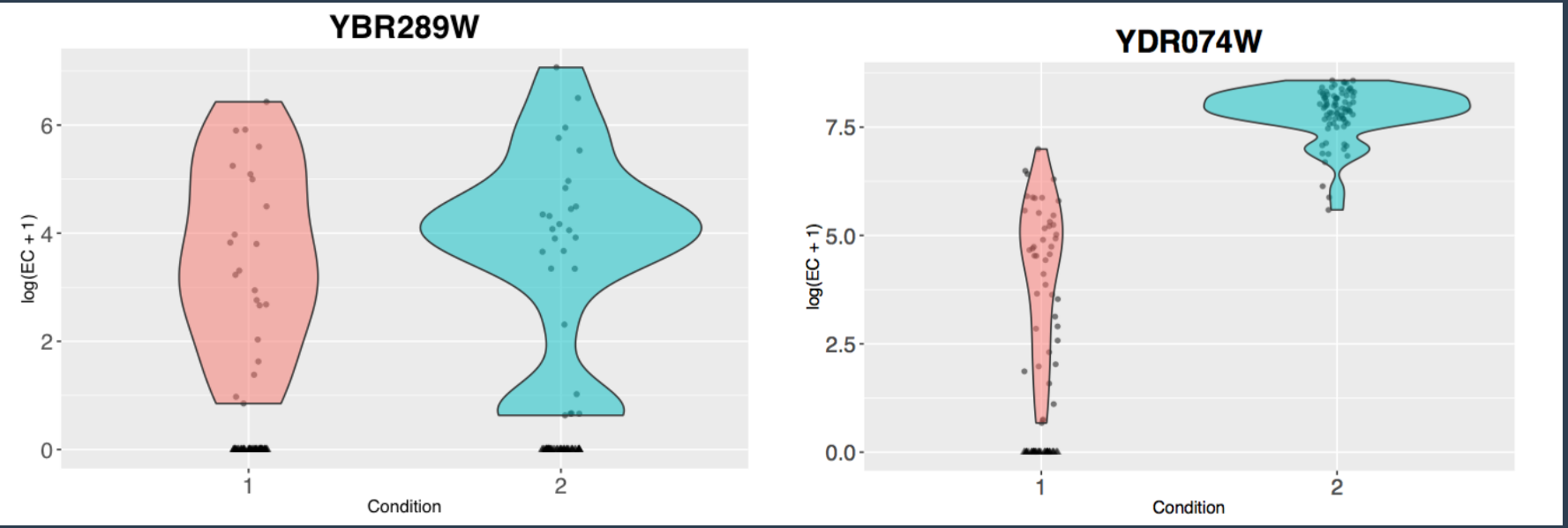
## Differential Expression

- Scientific collaborator comes in - wants to know if there is a difference in gene expression for their favorite gene across conditions.



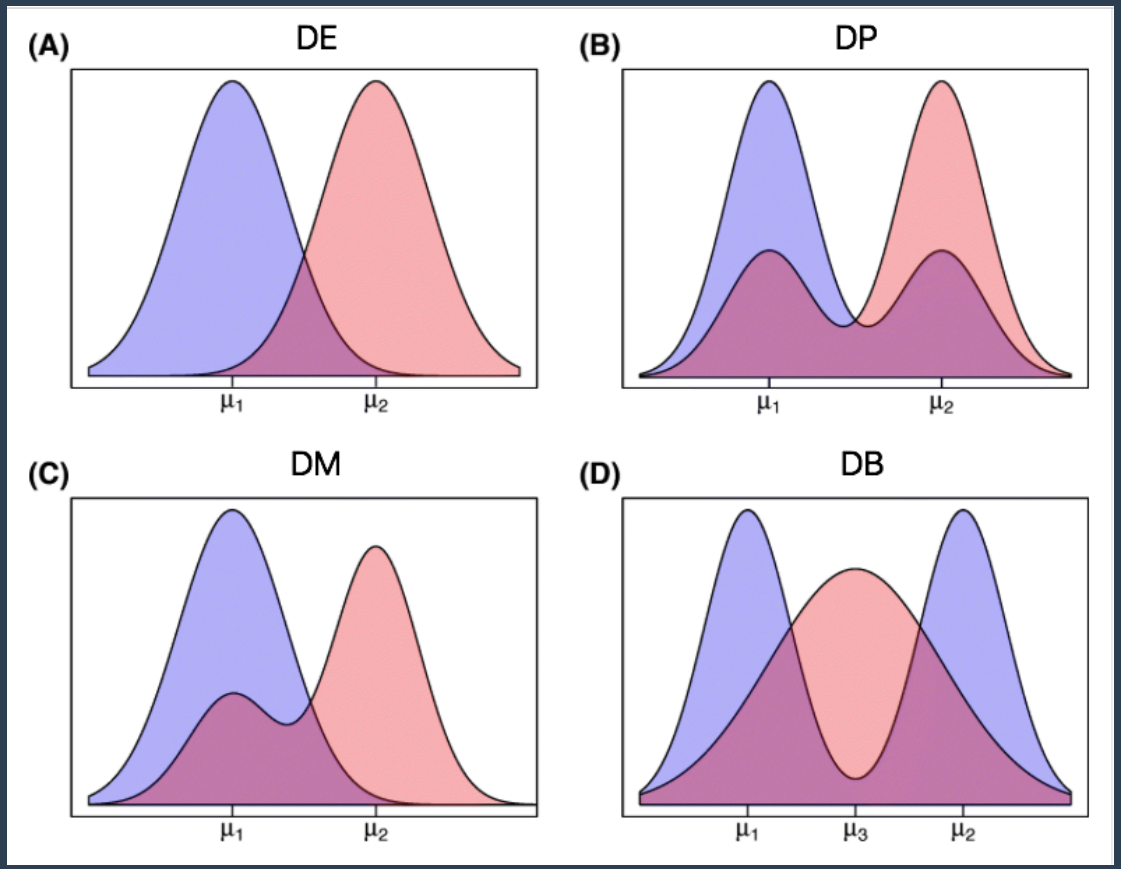
# Analysis

## Challenges for Single Cell Data




- Single cell data present more challenges (lots of zeros, multiple peaks, etc.)

# Single Cell Differential Distribution (scDD)

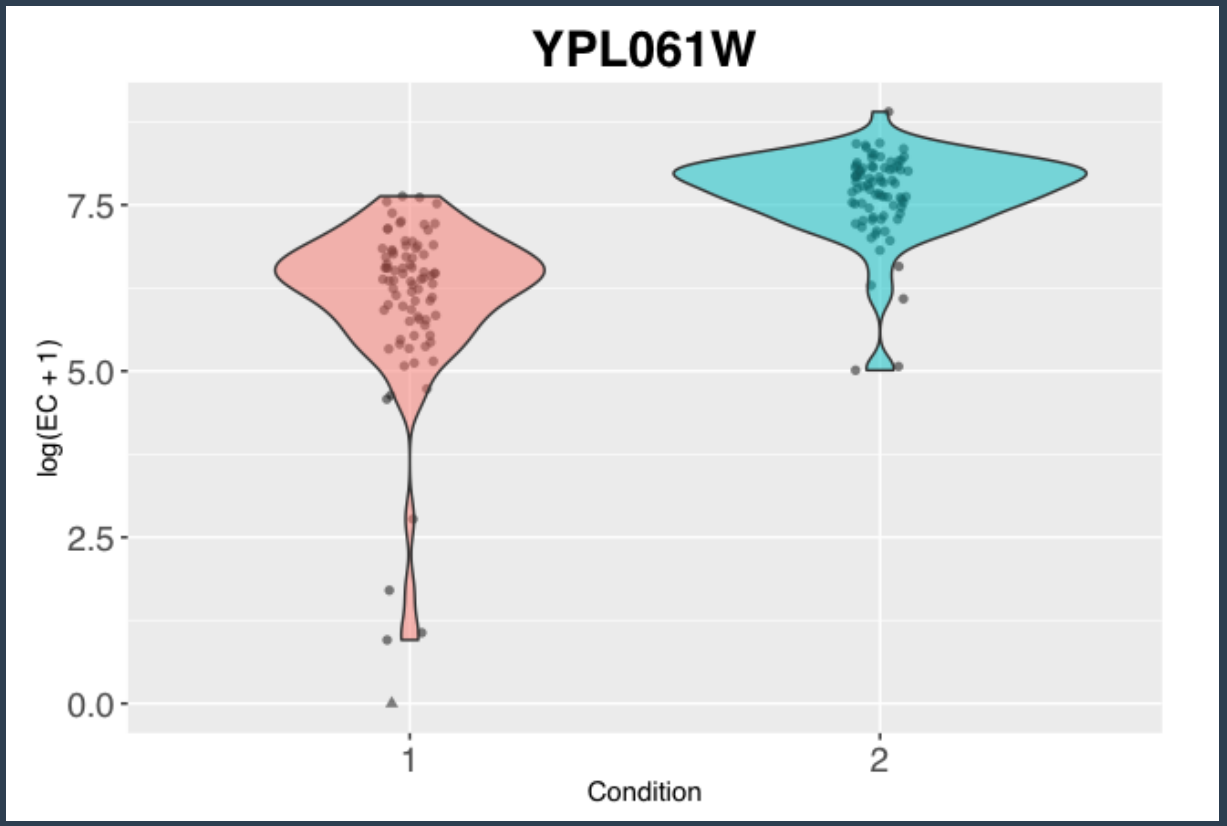



From Korthauer et al. (2016)

- scDD uses 2 stage approach to sort genes into categories: 1.) run permutation test to see if gene is different across conditions; 2.) use Dirichlet Process Mixture Model to sort into one of these 4 categories.

# Potential Problem?

Classified as DM – is it really?




- As a result, a min.size parameter was implemented into the scDD R package to put a threshold on the required number of obs. to be called a cluster.

# Further analysis

## Digging deeper

- What if we look at the same genes in groups of cells and try to classify subpopulations of these cells?
- Ex. pluripotent stem cells: These cells have the ability to turn into blood cells, lung tissue etc.
- Goal: Figure out which pluripotent cells are more likely to become blood, lung, etc.
- Turns out there's many methods implemented for solving similar problems.

# Possible Further Direction?

## ATAC-Seq

- Short for: Assay for Transposase-Accessible Chromatin with high throughput sequencing (Buenrostro et al. (2016)).
- Chromatin is found in cells: consists of DNA, protein, and RNA.
- ATAC-Seq looks at where chromatin is accessible to be transcribed.
- Related to epigenetics - basic idea: actions affect which genes are expressed.

# Thank you!

## Special Thanks

Kendziorski Lab

- Rhonda Bacher
- Jeea Choi
- Prof. Christina Kendziorski
- Ziyue Wang

# References


- Buenrostro J, Wu B, Chang H, Greenleaf W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology / edited by Frederick M Ausubel . [et al]. 2015;109:21.29.1-21.29.9. doi:10.1002/0471142727.mb2129s109.
- Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology. 2016 Oct 25;17(1):222.