First Year RPG Discussion

Xinyi Lin
Supervisor: Joshua Ho
2021/03/31 (updated: 2021-04-07)

Background & Research Goals

- Challenges on single cell data analysis

  - lack of accurate cell clustering method

  - hard to observe dynamic changes on different cell types

  - inaccurate trajectory analysis

  - needs of computational tools for different data types analysis

- PhD focus

  statistical analysis for scalable single cell analysis

  - statistical modeling

  - single cell data analysis

DCATS: Background

- Tissues are cellular heterogeneity

- Composition difference exists

- Interested in changes of proportions in dynamic process
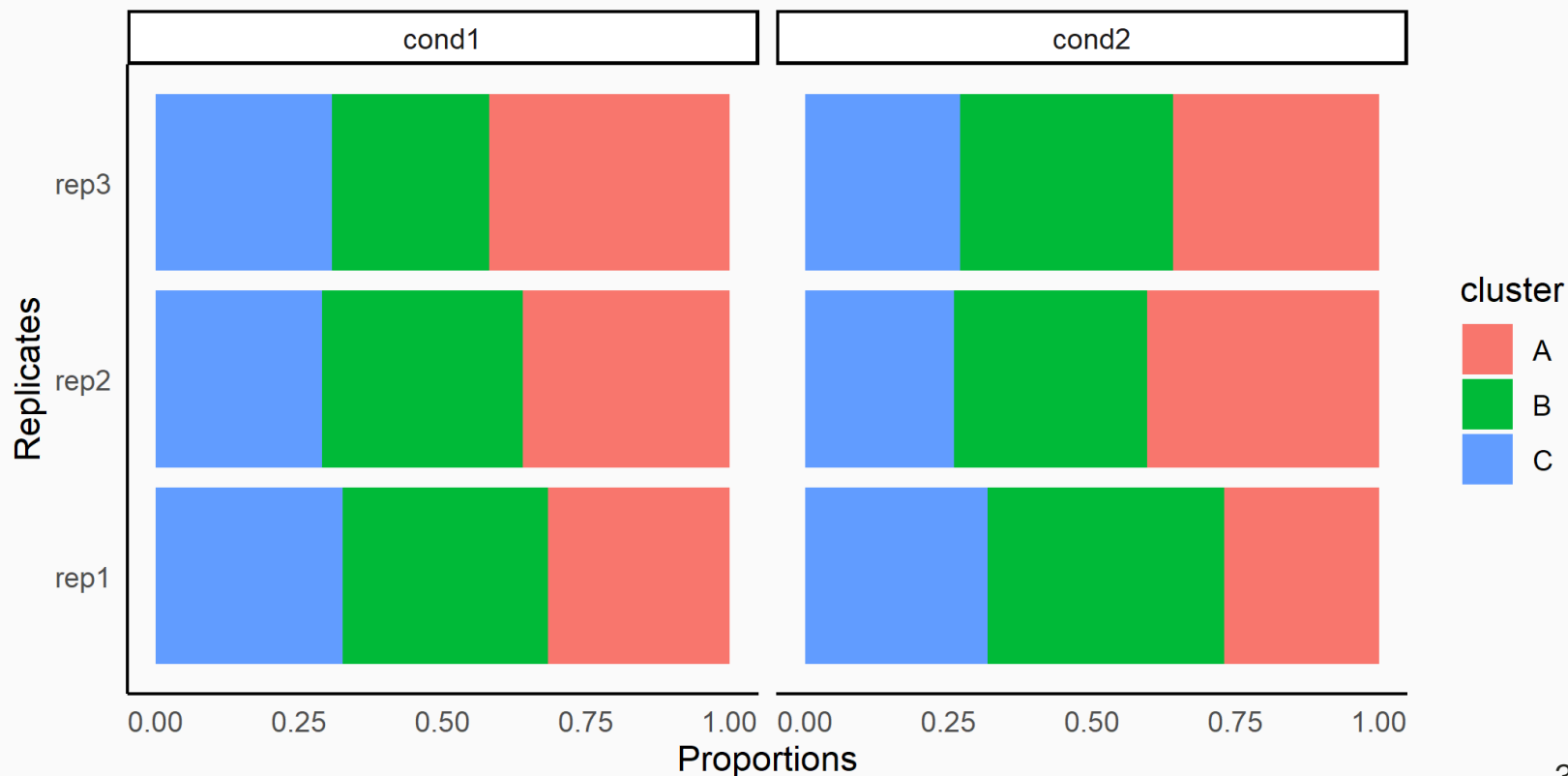
DCATS: Design

- R package designed for differential composition analysis on single cell data

- Basic assumptions

    - misclassification error exists

    - cell counts follow binomial distribution

- Input

    - cell counts matrices of samples coming from different conditions

    - a confusion matrix describing the misclassification error(user-input or calculated by DCATS)
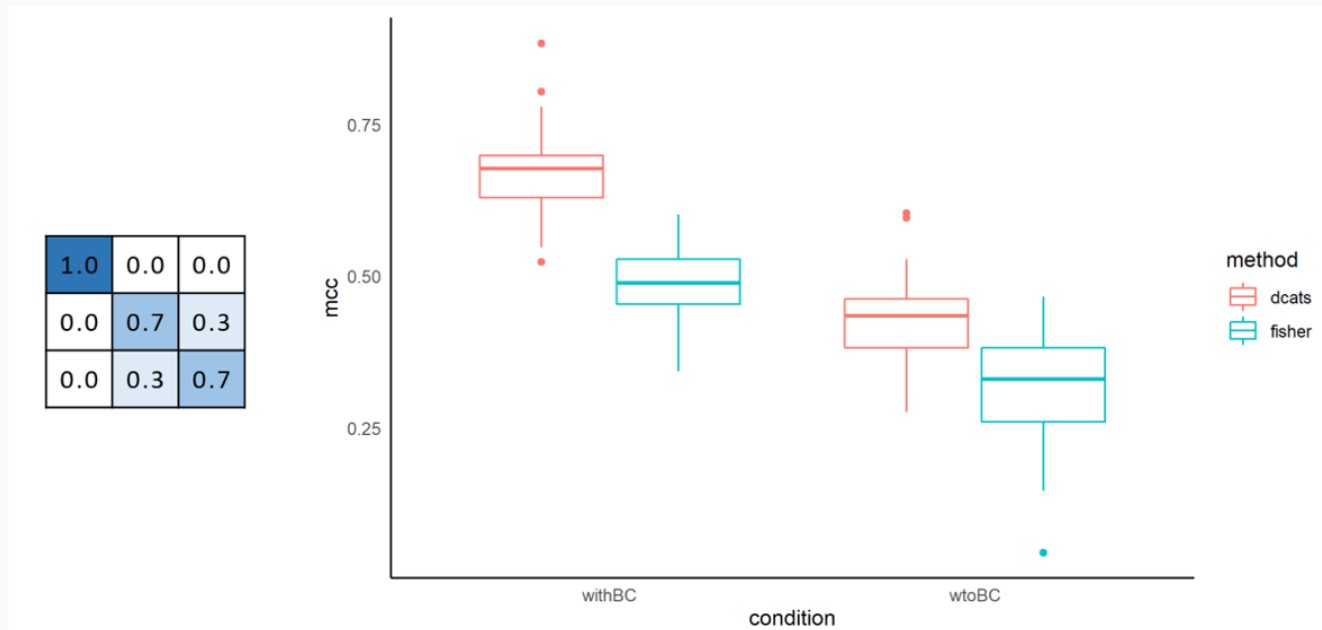
DCATS: Theoretical Simulation

- Given a similarity matrix, only simulate cell count matrices

- For each replicates:
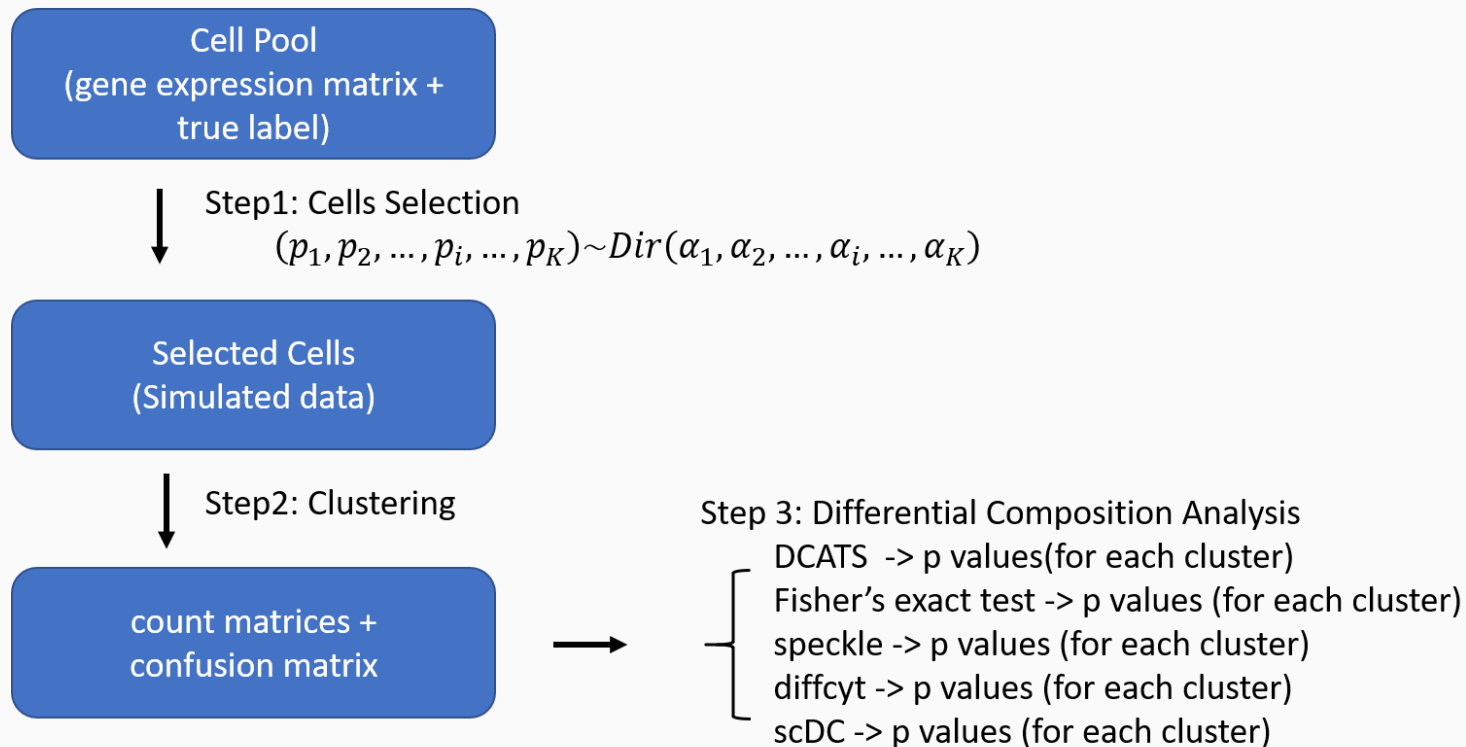
$$(p_1, p_2, p_3) \sim Dir(\alpha_1, \alpha_2, \alpha_3)$$

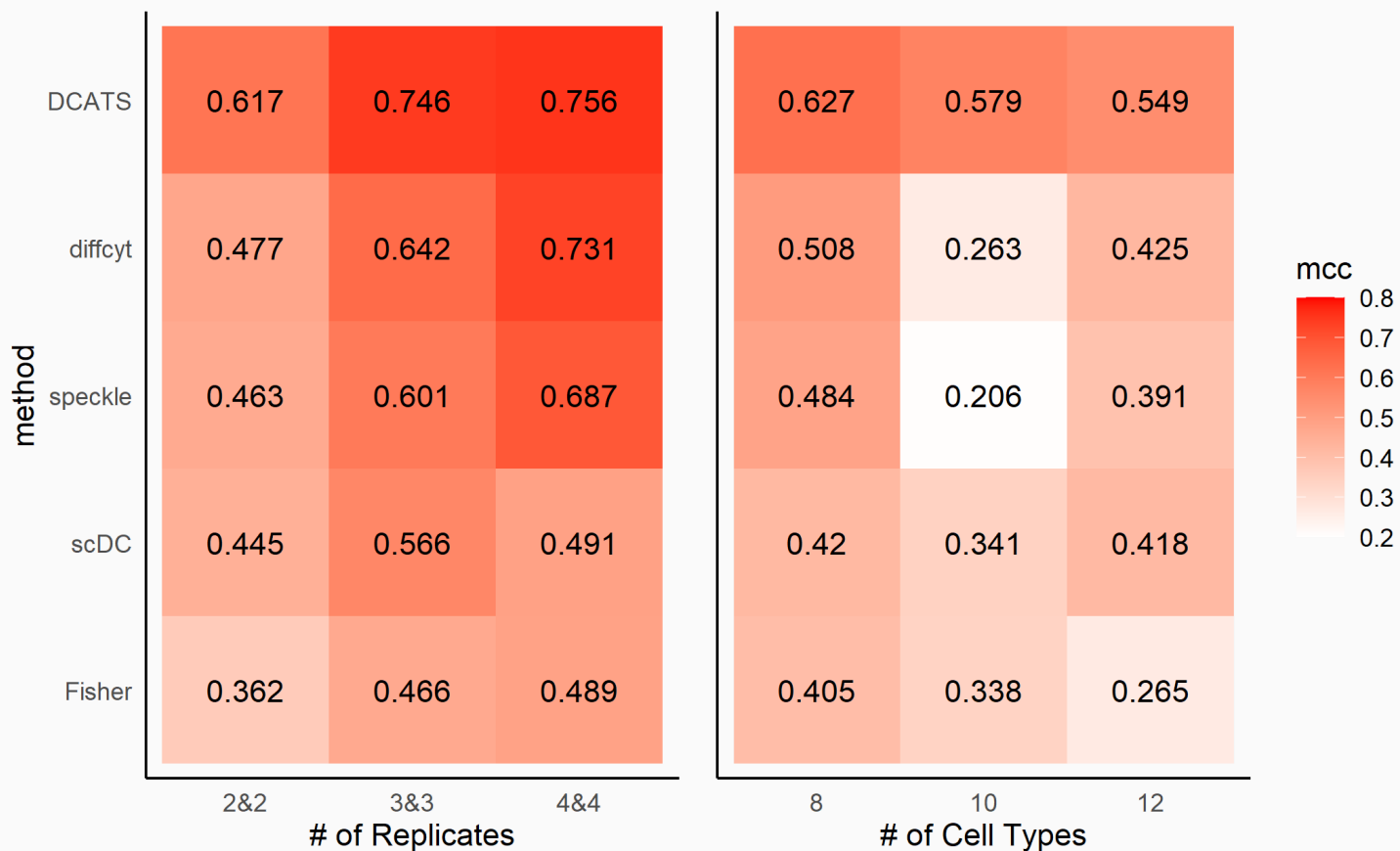$$(c_1, c_2, c_3) \sim multinomial(p_1, p_2, p_3 | M_{sim})$$

- Result

- Simulation process

Cell Pool
(gene expression matrix + true label)

Step1: Cells Selection

$$(p_1, p_2, \dots, p_i, \dots, p_K) \sim Dir(\alpha_1, \alpha_2, \dots, \alpha_i, \dots, \alpha_K)$$

Selected Cells
(Simulated data)

Step2: Clustering

count matrices + confusion matrix

Step 3: Differential Composition Analysis
- DCATS -> p values(for each cluster)
- Fisher's exact test -> p values (for each cluster)
- speckle -> p values (for each cluster)
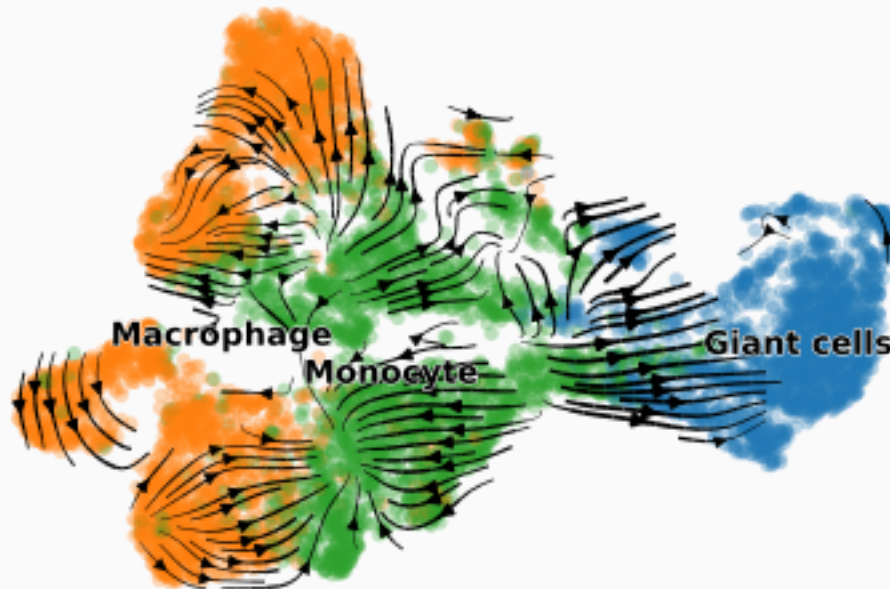- diffcyt -> p values (for each cluster)
- scDC -> p values (for each cluster)

- Simulation results

- Real world data analysis

  - Give consistent conclusions as original papers: *Kang, Hyun Min, et al, 2018* and *Haber, Adam L., et al, 2017*
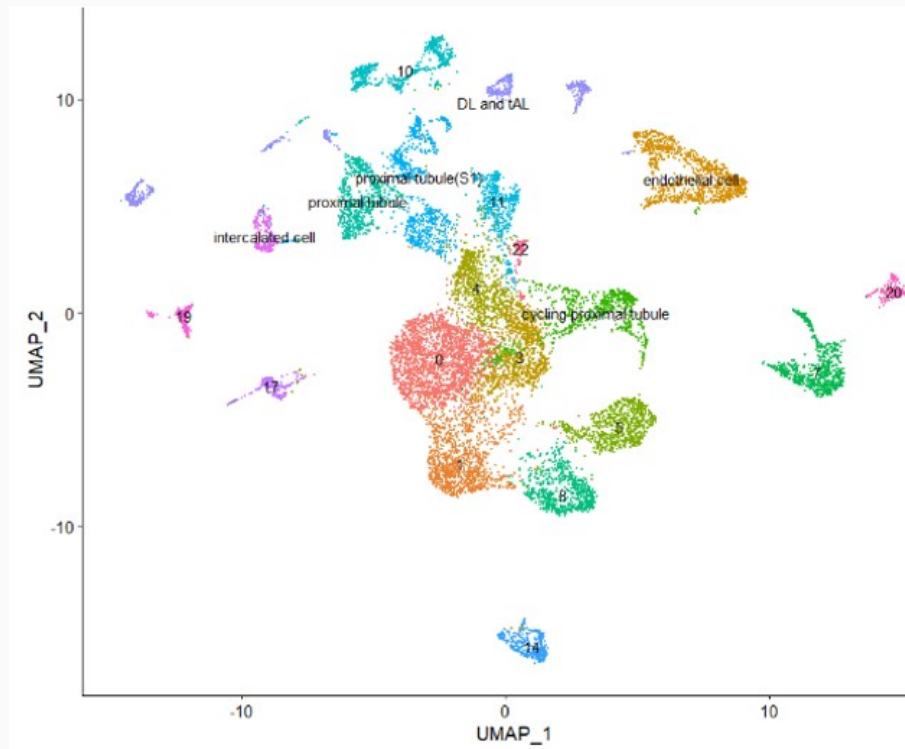
- Giant Cell Tumor of Bones(GCTB)

    - a somatic alterations(*H3F3A*, Gly34Val): IGV, cellSNP

    - trajectory analysis

Trajectory of myeloid cells

- Mouse's Kidneys

    - give a possible clustering result

    - identify some cell types based on the clustering result

Future Plan: current projects

- DCATS: get it published

- Side projects

    - be familiar with computational tools related to scRNA-seq

    - finish related bioinformatics analysis

    - find out some new problems during collaboration

    - learn more biological knowledge

Future Plan: exchange program

- Time: whenever travel restriction eases

- Duration: 3-6 months

- Lab: Xi Lab, Peking University

- Goals: Collaborate with people there on some scRNA-seq related projects

Future Plan: Other possible projects

- cross-species analysis

- time-series analysis

- other possible statistical modeling

Acknowledge

- Dr. Joshua Ho

- Other labmates on Ho Lab, HKU

- Dr. Yuanhua Huang and other people in Huang Lab, HKU

- Shan Han from Dr. Kelvin Cheung's lab

- Professor Sydney Tang and postdoc Stella Yiu

References

[1] L. Zappia, B. Phipson, and A. Oshlack. "Splatter: simulation of single-cell RNA sequencing data". In: *Genome biology* 18.1 (2017), pp. 1-15.

[2] H. M. Kang, M. Subramaniam, S. Targ, et al. "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation". In: *Nature biotechnology* 36.1 (2018), p. 89.

[3] A. L. Haber, M. Biton, N. Rogel, et al. "A single-cell survey of the small intestinal epithelium". In: *Nature* 551.7680 (2017), pp. 333-339.

Thanks!