2019 Fall Lab Rotation Presentation Note

***Page 1:*** Hi guys. My name is Jessie Tong, a first year Biostatistics Ph.D. student. It is my honor to give the final presentation here today. First of all, I would like to thank Prof. Mingyao Li for giving me the opportunity to rotate in her lab this semester. Today, I would like to introduce the project I have been working on during this semester. It is called: Removal of Batch Effects with Matched Mutual Nearest neighbors in Single-cell RNA Sequencing data.

***Page 2 and 3***: In human body, there are diverse cell types across human tissues. Researchers are interested in investigating the mechanisms that give to this diversity in development and how the disease develop with the cells. So, to understand the cells become very important.

To study the cells better, new approaches of observing and measuring are required. Thanks to the technological innovations in the past few decades, now researchers are able to study the cells at increasing resolution with improved tools and techniques.

***Page 4***: To be more specific, with the high resolution, we can study the cells on single cell level, so the single-cell RNA sequencing data tend to be a popular source. The data allow us to have the expression profile of every gene in a one single cell. Different from Bulk data which can only get the average gene expression from a cell population, This procedure is very important, because every cell is different from each other. Simply analyzing the bulk data will mask the heterogeneity of the cells. Compared with the bulk RNA sequencing data, the single-cell RNA sequencing data is more complex but critical to study the cells.

***Page 5***: When analyzing the single-cell RNA sequencing data, one of the most challenging problem is the batch effect. The batch effect will be induced when we produced large-scale single-cell RNA sequencing data set with different technical sources, in different labs or at different times. With batch effect, the expression of the genes could be different from the expression of genes in another batch. Such differences can mask underlying biology in the data. Therefore, to avoid misleading conclusions, the batch effects must be corrected before further analysis.

***Page 6 and 7***: Here, I want to use the data I simulated to show you what is batch effect. Here are two tsne plots. The TSNE plot is a dimension reduction technique to visualize high dimensional data sets. In the TSNE plot, there is no meaning for the x- and y- axis, but the circle here is meaningful. Each circle represents a cell and if the cells are from the same population, there will form a cluster for the cell. The left panel is colored based on the batch and the right panel is colored based on the cell type.

***Page 8***: There are several existing methods to remove the batch effect. However, most of the existing methods for batch correction were specifically designed for bulk RNA sequencing data.

***Page 9***: In 2018, Haghverdi proposed the MNN, short for matching mutual nearest neighbors’ method to conduct batch correction for single-cell RNA sequencing data. This method is one of the most popular methods for batch correction. So far, it has been cited for more than 200 times. The main idea of the method is to remove the discrepancies between biologically related batches according to the presence of MNNs between batches.

***Page 10***: Each cell has expression vector (can be expressed as a vector of 0, 1, . If the one gene is expressed, then the value of that gene is 1.) (a) Batch 1 and batch 2 in high dimensions with an almost orthogonal batch effect difference between them. (b) The algorithm identifies matching cell types by finding MNN pairs of cells (gray box). (c) Batch-correction vectors are calculated between the MNN pairs. (d) Batch 1 is regarded as the reference, and batch 2 is integrated into it by subtraction of correction vectors. (e) The integrated data are considered the reference, and the procedure is repeated for integration of any new batch.

***Page 11***: However, we observed one limitation of the MNN method. The order of the pairwise matching two batches is not clarified in the paper. In other words, we do not know how the author choose the batch as the reference group, we do not know which batch is used as the second batch to match with the reference batch. We dive into their source code of their R function and found that they just simply match the batch sequentially without considering other possible orders of the batches.

***Page 12***: This limitation motivated us to conduct a simulation study to explore if the orders of the batches when we use the MNN method really matter.

Page 13: Here is the simulation design. We simulated a temporal situation with 4 batches. Each batch has different number of cell types. Batch 1 has 3, Batch 2 has 4 and so on. The cell types 1 to 5 have batch effect across the four batches.

Page 14: If we do not do the batch correction, when we do the clustering analysis, there will be total around 19 clusters would be recognized. However, this is not correct, because we know that cell population 1 should be clustered together. Same as population 2, 3, 4, and 5. If we do the batch correction, we expected the best clustering analysis we can get is 7 clusters since we generate 7 different populations.

Page 15: This slide is used to show you how the data look like. The data we analyzed is a large matrix with each column is a cell, and each row is a gene. For example, the zero in the first entry mean that in cell 1 gene 1 is not expressed, and second entry mean in cell 1 gene 2 is expressed.

Page 16: Okay. Let me show you the data I simulated with tsne plot. Similar to the plot I present, the panel on the left is colored based on batches. You can see we have 4 batches. And the panel on the right is colored based on cell types. From the plots we can see very strong batch effect because the same cell type in different batches is clustered in different group.

Page 17: First, I would like to show the performance of MNN method. UMAP plot is similar to T-SNE plot but an improved technique to visualize the high-dimensional data into low dimension

Page 18 and 19 Now, I want to show that the results with different orders. Here I present 4 different order.

Page 20 and 21: To quantified the differences in results when we change the orders, we calculate the correlation between same cell from order 1 and order 2. With this box plot, we can see the correlation between extreme different orders are very low but very high for the almost the same orders.

Page 22: To make a conclusion, the orders of the batches in MNN approach really affect the results for batch effect correction. When apply the MNN approach to correct batch effect in single-cell RNA sequencing data, we should be careful about the orders of the batches if the number of batches is larger than 2. Since there is no natural order of the batches in reality, researchers should consider of correcting the batch effect correction simultaneously instead of pairwise.