Single Cell RNA Sequencing Analysis Review

Lightphenexx

July 6, 2023

1 Introduction

This review will provide some basic introductions on single cell RNA sequencing, as well as some Deep learning models that will be emphasized in subsequent work for predicting gene expression. The first part will introduce some basic concepts and the objective of scRNA. Furthermore, The Best Practices in scRNA Seq, including Data Preprocessing and descriptive statistic or downstream analysis, and the second part will refer to some refrences, which proposed some deep learning model for predicting gene expression. However, due to my limited knowledge, I am currently only able to perform the first step of reproduction.

2 A Brief Introduction Of Single Cell RNA Sequence

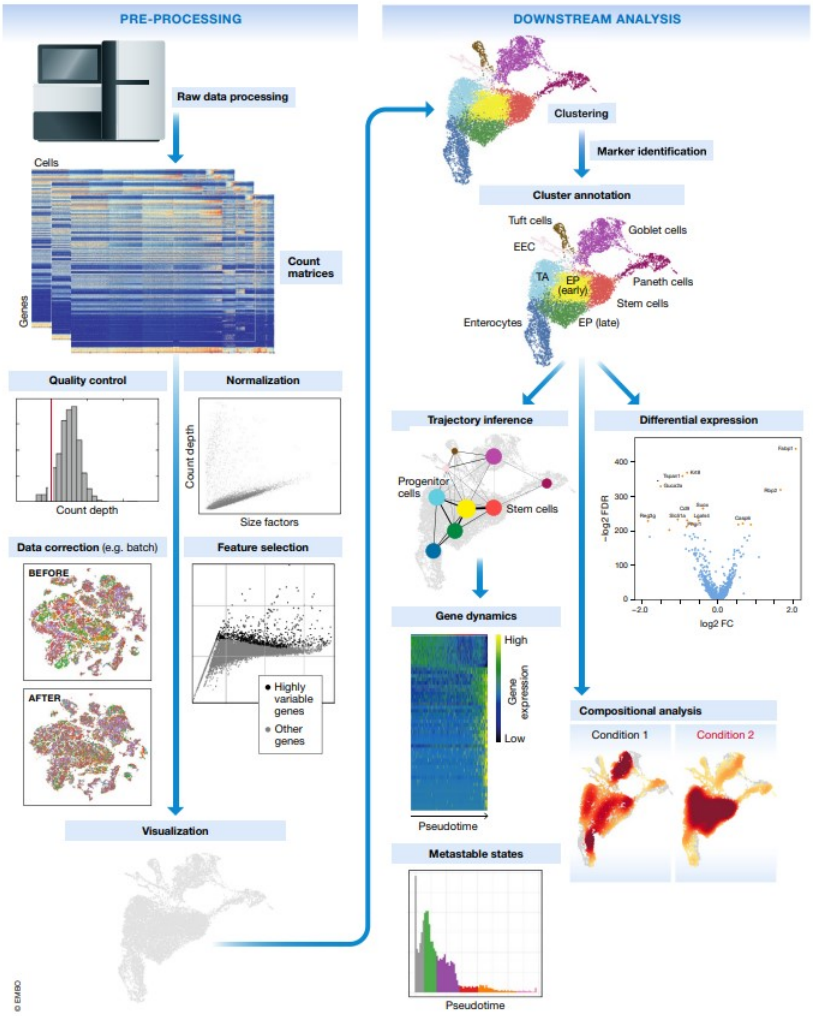
2.1 Objectives Of Single Cell RNA Seq analysis

In recent years, the method of single cell analysis has contributed in various fields, including predicting the risks and the development of certain disease from certain gene expression in transcriptome data, and the connection of each cells. The Main Objective of these studies focus on solving the problems in real applied scenery which could furthermore contribute in

further studies. While the technology of single cell analysis has become gradually mature, researchers has been finding new optimized and efficient method to make the analysis more efficient and outstanding.

2.2 Best Practices in scRNA Sequencing

Up to now, there are many methods to analyze single cell Transcriptome data,Including Data Preprocessing and Descriptive Statistic or Downstream Analysis, and the most mainstream data analysis steps are listed below [1]



2.2.1 Quality control

The very first step of data preprocessing of transciptome data is quality control, we could analyse its quality through the gene counts per barcode, or the distribution of count depth.

Before moving on into the practice of Quality Control, Understanding the basic transcriptome data format is fundamental. For this part infers to Count matrices in which the horizontal axis represents each cells and the vertical axis represents each genes.

Next, we could analyze the three covariates, The count depth of each barcode, the number of genes per barcode, and and the fraction of counts from mitochondrial genes per barcode. Then perform distribution display on each three covariates and identify outlier peaks. Those outlier peaks might represents dying cells, cells with broken membranes and doublets.

2.2.2 Data Normalization

This is also the part of data preprocessing, in order to help further downstream analysis, this method is widely used in training machine learning models. the main objective to data normalization is to reduce the gene bias in a particular datasets.

There are two types of data normalization method: one is within-sample normalization and between-sample normalization[5].

within-sample normalization focus on removing biases, or shrinking the gap between the max and min of gene datasets. Common within-sample normalization methods include RPKM/FPKM and TPM.

In contrast, between-sample normalization tries to create difference between data, which makes it easier to analyze the difference.

2.2.3 Data Imputation

When dealing with transcriptome data, missing gene values could occur. In order to prevent imputing missing values when performing downstream analysis and training deep

neural network, there are many data cleaning method and strategies on dealing with this problem.

Markov affinity-based graph imputation of cells (MAGIC) is a typical data imputation for single cell RNA seq analysis. It would be further addressed in Nvwa Deep learning model.

2.2.4 Batch Effect removal

Batch effect could occur when data are acquired in different time periods or multiple places. while scientists are trying to avoid this from happening, but there are sometimes when we can not prevent batch effect from happening.

One method on dealing with batch effect correction is MNN (mutual nearest neighbor), its main focus is trying to find the most similar cell in different batches of different data.

2.2.5 Dimension Reduction

In Order to Visualize the Transcriptome data and classify cell types using Unsupervised Clustering method, The mainstream method is to perform Dimension Reduction on transcriptome data to help further downstream analysis.

For this review which will Represent two Dimension Reduction Methods, t-SNE[4] and UMAP[?]. The typical method for dimension reduction is Principle Component anaylsis(PCA). Nevertheless, when dealing with high dimension and complex nonlinear data, linear models couldn't perform better than nonlinear models.

The Performance of t-SNE Dimension Reduction model[4] is strong compared with other models, with its strong feature extraction it is most likely to be the best option when dealing with high-dimensional data, also including Single Cell RNA Transcriptome Data. But however, the Publisher itself also proposed that this method is most likely to perform better when $d \leq 3$, for it is only used for visualizing purpose.

2.2.6 Clustering analysis

After performing Dimension Reduction on Transcriptome data, we take the Dimensionality-reduced representations as input to perform Unsupervised Clustering.

The typical unsupervised clustering model KNN is often used in clustering analysis. Indeed, there are many articles introducing this clustering model, and there are many machine learning plus deep learning model which is much optimized than KNN.

The main focus for this section is on clustering annotation, which means identifying cell types when clustering algorithm is been performed.

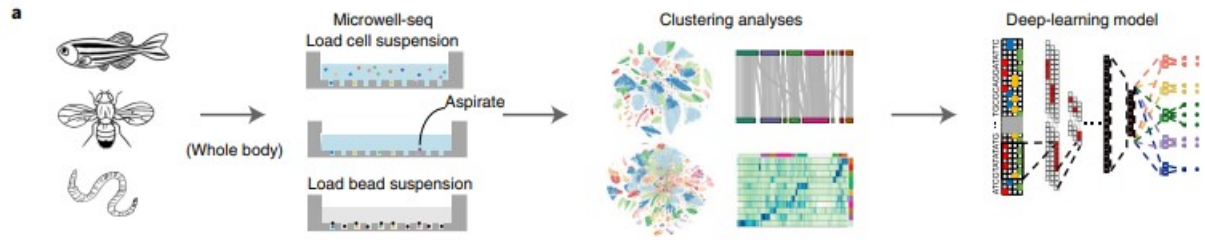
In order to identify cell types, external resources are required when identifying. for example, most researchers will use gene markers as a reference when identifying cell types.

3 Training Deep Learning Models For scRNA Seq

For This Section, which will introduce and analyse a deep learning model called Nvwa[2] and how it is used in predicting gene expression. The Deep Learning model has been trained to predict gene expression and cell landscape of different species. Before Describing how the model has been trained, The following section will first go over how the experiment is been designed.

3.1 Experiment Design

For this experiment, they analyzed three datasets of different species, zebrafish, Drosophila and earthworm. after finishing quality control of the transcriptome data , they Performed Markov affinity-based graph imputation of cells (MAGIC) to impute missing gene expressions first. Next they used the t-SNE dimension reduction method to visualize the transcriptome data and further using scale-normalized expression levels of canonical cell-type-specific markers. Furthermore, they also performed subclustering to identify more specifically.



3.2 Architecture Of Nvwa Deep Learning Model

The Deep learning model is constructed with CNN, RNN, and feed-forward Blocks, The equation below can briefly show how the network is constructed.

$$f_t(x) = Sigmoid(FFN_t(GAP(ReLU(Conv(x)))))$$

The Two Activation Functions are ReLU and Sigmoid. Used in predicted probability calculation.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$
$$f(x) = max(0, x)$$

They first set up a CNN with two convolution layers. The first layer mainly focuses on feature extraction, transforming genomic sequences to feature maps. For the final results, there was a total of 128 filters and the output is a 4×7 weight matrix. When moving on to the second layer, it reflects the degree of interaction between the feature maps of the previous layers.

After Performing Convolution operations, they used the ReLU function to transform the output data, and then use global average pooling (GAP) method to calculate the mean of the overall feature map.

Finally, the output of the GAP was fed into the FNN and then transformed into expression probability using sigmoid function. The FNN Model is constructed with two consistent linear transformations with a ReLU activation in between.

3.3 Training the Deep Learning Model

They first feed the deep learning model with 8 transcriptome data of different species, and then used the average area under the receiver operating characteristic curve (AUROC) to assess the accuracy of the deep learning model, and the overall score of the training result is AUROC of 0.78 and AUPR of 0.59.

When training the deep neural network, they used mean of binary cross entropy loss, Known as BCEloss to measure the differences between the predicted output and the label.

$$BCELoss = -\sum_{i}\sum_{t} y_t^i \log f_t(x_i) + (1 - y_t^i)\log(1 - f_t(x_i))$$

i represents the index of the input samples and t represents the index of the target label. Thus, y_i^t is the target label for sample i, cell t.

In order to prevent overfitting and control the sparsity of the model, They Performed L1 and L2 regularization, which is a common optimization method dealing with overfitting. Thus, the final equation is shown below. λ_1 and λ_2 are hyperparameters, which are Artificial setting.

$$Objective = BCELoss + \lambda_1 L1loss + \lambda_2 L2loss$$

In order to minimize the Objective Function, they used a typical back propagation method known as the adam optimizer.

4 Conclusions

This Review gave the basic introduction on Single Cell RNA Sequencing Analysis and its best practice. For the second part of this review, which haven't proposed Deep learning methods used in data preprocessing. More could be found in this review[6].

The Authors who trained Nvwa model proposed in their article saying that their model needs improvement on the architecture. Thus, future work could focus on transformers and those Large Language Models.

References

[1] Luecken M D, Theis F J. Current best practices in single-cell RNA-seq analysis: a tutorial[J]. Molecular systems biology, 2019, 15(6): e8746.

[2] Li J, Wang J, Zhang P, et al. Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types[J]. Nature Genetics, 2022, 54(11): 1711-1720.

[3] Lakkis J, Wang D, Zhang Y, et al. A joint deep learning model enables simultaneous batch effect correction, denoising, and clustering in single-cell transcriptomics[J]. Genome research, 2021, 31(10): 1753-1766.

[4] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).

[5] Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis[J]. Frontiers in genetics, 2019: 317.

[6] Bao S, Li K, Yan C, et al. Deep learning-based advances and applications for single-cell RNA-sequencing data analysis[J]. Briefings in Bioinformatics, 2022, 23(1): bbab473.