The Comparison between Bulk RNA-seq and Single-cell RNA-seq

Zhang Wanqi
Reading College, Nanjing University of Information Science and Technology, Nanjing, China
hq802899@student.reading.ac.uk

Abstract. Bulk RNA-seq and single cell RNA-seq (Sc-RNA) seq are two well-known methods and are broadly used in biology areas. Even though the two ways are all starting from the mRNA level to do the transcriptional analysis, many differences still show in them, but the differences which are critical for researchers to judge and consider were not addressed clearly in the past studies. Here we summarized the principle, workflow and data presentation of the two methods and compared them comprehensively, which shed a light on the application of high throughput sequencing technique for researchers.

1. Introduction

1.1 The principle and workflow of bulk RNA-seq

The bulk RNA-seq is a transcriptional sequencing technology, which is to measure the sequences of mRNA, small RNA, and NON-coding RNA or some fragments of them by high-throughput sequencing technology, and to detect the genes expression level.

The workflow of the bulk RNA-seq technology is followed.

1.1.1 Prepare a sequencing library. Firstly, the total RNA of the cell or tissue is obtained by RNA extraction, mainly including mRNA and non-coding RNA. Secondly, the RNA samples are processed according to the experimental needs. The processed RNA is fragmented and then reverse transcribed to form cDNA. In this way, the cDNA library is obtained. After attaching the adaptors to the terminals of the cDNA fragments, all the fragments can be amplified by PCR (Figure 1).

![Figure 1. Preparing a cDNA library. The process of obtaining an cDNA library [1].](image)
1.1.2 Sequencing. A sequencing plate can contain more than 400 million fragments, and four different fluorescent probes match well with A, T, C, and G, which is in a complementary way. When probes match with each base, they will show their own unique color (A with red, G with blue, C with green, and T with orange) and then the color is taken and record. After measuring the same row of bases in many sequencing fragments, the original fluorescent row of bases is washed away, and then the lower row of bases is combined with the fluorescent in a rolling way. [2]

1.1.3 Quality control and data analysis. First of all, the quality of sequencing fragments is evaluated by FastQC (Figure.2), so that the sequencing results can be visually seen and evaluated, such as base mass distribution, GC content, undetected base N content and so on. Next, sequence alignment is performed to match the fragments to the reference genome sequence (Figure.3). [3] Then all reads are matched to the genome and each gene has its own number of matched reads, which represents its relative expression level. The last step is the data visualization and this step will be summarized separately on the following.

![Per base sequence quality](image1)

**Figure 2.** The report of FastQC. Green is good, yellow is average and red is poor. The column is the 25%~75% interval mass distribution, the error bar is the 10%~90% interval mass distribution, and the blue line represents the average.

![Sequence alignment principle](image2)

**Figure 3.** Principle of sequence alignment. The reference genes and reads are broken into small fragments for comparison [2].

1.2 The principle and workflow of single-cell RNA-seq

Single-cell RNA sequencing, an extension of bulk RNA-seq invented at 2009 [4], is used to measure each gene expression level in each single cell, which plays an important role in the study of the transcriptional diversity in the cell population.
1.2.1 Prepare the single cell. Limiting dilution (Figure.4a) is a common technique to separate individual cells, in which pipette is used to dilute the cells. Usually, when diluted to a concentration of 0.5 cell per well, only about one third of the holes are distributed for each single cell.

Micromanipulation (Figure.4b) is another classic method of extracting cells from early embryonic or uncultured microorganisms. Microscope-guided capillary pipettes have been used to extract individual cell from suspensions. However, this method is time-consuming and inefficient [5], [6].

Flow-activated cell sorting (FACS) FACS (Figure.4c left) has become the most common strategy for the separation of high-purity single cells. FACS is also the preferred method for target cells with low proportion. In this approach, cells are first labeled with fluorescent monoclonal antibodies which recognize specific surface markers and classify different cell populations [7].

Microfluidic technology (Figure.4c right) refers to the technology of precise control of microliter and milliliter level samples through micron level runner [8]. Microfluidic technology shows huge advantage in cell sorting applications, the available centrifugal gravity, fluid mechanics and electric field force can also be multiple operating units together at the same time, which make the cell capture, cell culture, cell lysis and subsequent analysis into a whole process. This technology is widely used because of its low sample consumption and low analysis cost, the ability to achieve accurate fluid control, and the nanoscale volume required by the technology greatly reduces the risk of external contamination.

Figure 4. Single-cell isolation and library preparation. a. The dilution method isolates individual cells. b. Micromanipulation methods to collect single cells t. c. FACS single cells by tagging cells with fluorescent tagged proteins (Left). Microfluidic technology for single-cell isolation (Right). d. Schematic graph example of droplet-based library preparation. [10]

2. Library preparation and sequencing

There are many ways to prepare the single cell library which mainly due to different platform. For Fluidigm C1, a plate-based method, the full length of transcripts will be got in each single well. However, this method is not good enough for the rare cell population.

Another popular platform is 10× Genomics in which microfluidics is used to capture each single cell and then each single cell is used to prepare the library. The detail of the process is as below:

In a whole view of 10× Genomics, a single cell is mixed with a single gel bead through oil to form tiny droplets coated with water. Then the cell membrane is broken off and the mRNA free from the
cell. The free mRNA mixes with the water in the droplet, contacting reverse transcriptase, nucleic acid primers that bind to gel beads, and dNTP substrates.

After amplification of the cDNA, and then cDNA is fragmented. Specific DNA fragments are seeded on the regel beads. The DNA fragment is composed of three parts: Barcode, UMI and PolyT(Figure.4d) [9]. The Barcode is the length of the 16 bases and each microbead is corresponding to one kind of Barcode. Through these 4 million Barcode, the gel microbeads can be distinguished. UMI is a random sequence, meaning that each DNA molecule has its own UMI sequence. UMI with 10 base lengths has 1 million sequence changes. The role of UMI is to differentiate which reads come from a specific original cDNA molecule.

10× Genomics technology can obtain a large amount of big cell data at the same time, which reduces the time cost and money cost at the same time. However, only part of a transcript is prepared not the full length, which is different to Fluidigm C1 platform. While it has a big advantage on the process large cell population and detect rare cell type. Data analysis will be described in a very detail in the following part.

We just made a detail description of principle and workflow of bulk RNA-seq and ScRNA-seq, and obvious differences can be told between the two technique. Interestingly, the data present is also quite different. Here we will go over the data present ways in the two methods to show this difference. To present the aspects in a straightforward way, we take two studies which both took heart as the research target.

3. Bulk RNA-seq data visualization

3.1 Scatter plot of differential gene expression

This kind of plot is to show the correlation of gene expression level change when compared to the third group. The more points off the diagonal, the lower the correlation of the sample expression and the worse the repeatability. The fewer points off the diagonal, the higher the correlation of the expression between samples and the better the repeatability of the repeated sample (Figure.5).

![Figure 5](image_url)

Figure 5. Representative scatter plot to compare the two RPKM values between two groups. [11]

4. Differential gene clustering heat map

Heat map is very popular data present method in high throughput sequencing search. It is based on the different expressed genes and embed them in a heat map to show the gene clusters which show similar changes in different group as Figure.6 showed that the heat map showed the gene clusters showing similar change in NF, ICM and DCM groups.
Figure 6. The heat map matrix showed the gene clustered by phenotype after adjustment of all expressed genes. Warm colors indicated that the samples with low coefficients have a large difference. The colder colors meant the most similar relationships between samples [11].

In addition to the two methods mentioned above, typical bar plot or scatter plot of RPKM of each gene to show the expression level in different group is also another typical way to show the data.

5. Single-cell RNA-seq data visualization

5.1 Cell cluster analysis
Cluster analysis is used to identify cell subtypes. For example, in R package Seurat, instead of directly conducting cluster analysis on all cells, PCA principal component analysis is first performed, then the principal components with the largest contribution are selected, and the value of selected principal components is used for cluster analysis. There are mainly two methods in Seurat to do the nonlinear dimension reduce to do the cell clustering- t-distributed stochastic neighbor embedding (t-SNE) and Uniform manifold approximation and projection (UMAP)

5.2 t-SNE

t-SNE is a very popular way to do the cell cluster presentation and analysis, and it mainly was used to do the nonlinear dimension reduce to two dimensions for a high-dimension data [12], [13]. Typical graph is shown as Figure 7.

Figure 7. t-SNE clustering of single cells isolated from human hearts. CM, cardiomyocyte; EC, endothelial cell, FB, fibroblast; MP, macrophage; SMC, smooth muscle cell [14].
5.3 UMAP
UMAP is a very effective visualization and scalable dimensionality reduction algorithm. In terms of visual quality, UMAP algorithm has a competitive advantage over T-SNE, but it retains more global structure, has superior performance, and has better scalability. As we can see from the above definition, UMAP is a great tool for big data and high dimensional data such as single cells. Especially in large data sets, they not only aggregate the clusters corresponding to similar cell groups, but also provide useful and intuitive properties and retain more global structures, especially the continuity of cell subsets [15].

5.4 Pseudotime analysis
Pseudotime is a way of how much progress a single cell has made through a process. For example, cell differentiation. cells do not progress synchronously in many biological processes. Regarding the single cell expression study, single cells may be broadly distributed in different progress. This asynchrony generates mainly problems when we try to understand the sequence of regulatory changes that generate as cells transition from one state to another. So based on the transcriptional features, pseudotime analysis can make each single cell on the timeline of a biological process. As shown in Figure.8, all cells were analyzed by pseudotime analysis and embedded in the plot with different color, showing the position of each cell according to its age inferred by the transcriptional features [16].

![Figure.8 Aging trajectories of left atrial and left ventricular cardiomyocytes and violin plot of several genes in different type cells. [14]](image)

5.5 Violin plot
Violin plot is a typical data presentation in ScRNA-seq analysis and it just showed different gene expression level in different cell type which can be seen in Figure.8. Violin plot can give us a very direct impression of the expression pattern of each gene and can help us tell the expression difference between cell types.

6. Conclusion and discussion
With the development of sequencing technique, the sensitivity and cost of the sequencing is becoming more and more acceptable. While the new transcriptional sequencing technique will take the advantages and have a great development.

Bulk RNA-seq and ScRNA-seq are two most popular transcriptional sequencing method which were applied in many areas and used to solve and reveal many key findings. It is meaningful to compare the two methods in terms of principle, workflow and data presentation. Regarding the principle as we discussed above, the difference is not so obvious since both of them are based on the transcriptional features which mRNA is used as a start point. A very different point is
that single cells are needed to be isolated for Sc-RNA seq in which single cell isolation methods and the cell condition are very important and affect the whole process directly.

So for Sc-RNA seq, the cell quality of each single one is critical.

In terms of data presentation where the most difference resides in, traditional data presentation way, such as the heat map and gene expression scatter plot, were also the same in the two techniques, while several new data presentation methods were involved in Sc-RNA seq, for example, cell cluster analysis, pseudotime analysis and violin plot. These new presentation methods conferred the Sc-RNAseq powerful in reveal the new and interesting findings in single cell population in which bulk RNA-seq can not make it.

While new presentation methods need more advanced programming ability which put many challenges to our researchers, programming ability is becoming more and more important in this area. Also with new methods emerging, the validation of these methods are also needed to be addressed.

7. References

[1] Wang Z, Gerstein M and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57-63.

[2] Marioni JC, Mason CE, Mane SM, Stephens M and Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18:1509-17.

[3] Brown J, Pirrung M and McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017;33:3137-3139.

[4] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K and Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377-82.

[5] Brehm-Stecher BF and Johnson EA. Single-cell microbiology: tools, technologies, and applications. Microbiol Mol Biol Rev. 2004;68:538-59, table of contents.

[6] Guo F, Li L, Li J, Wu X, Hu B, Zhu P, Wen L and Tang F. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. Cell Res. 2017;27:967-988.

[7] Julius MH, Masuda T and Herzenberg LA. Pillars article: demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. Proc. Natl. Acad. Sci. 1972. 69: 1934-1938. J Immunol. 2014;193:2048-52.

[8] Whitesides GM. The origins and the future of microfluidics. Nature. 2006;442:368-73.

[9] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P and Linnarsson S. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11:163-6.

[10] Hwang B, Lee JH and Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50:96.

[11] Sweet ME, Cocciolo A, Slavov D, Jones KL, Sweet JR, Graw SL, Reecce TB, Ambardekar AV, Bristow MR, Mestroni L and Taylor MRG. Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure. BMC Genomics. 2018;19:812.

[12] Cieslak MC, Castelfranco AM, Roncalli V, Lenz PH and Hartline DK. t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. Mar Genomics. 2020;51:100723.

[13] Tu D, Zheng J, Jiang Z and Pan H. Multiscale Distribution Entropy and t-Distributed Stochastic Neighbor Embedding-Based Fault Diagnosis of Rolling Bearings. Entropy (Basel). 2018;20.

[14] Wang L, Yu P, Zhou B, Song J, Li Z, Zhang M, Guo G, Wang Y, Chen X, Han L and Hu S. Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. Nat Cell Biol. 2020;22:108-119.

[15] Parra-Hernandez RM, Posada-Quintero JJ, Acevedo-Charry O and Posada-Quintero HF. Uniform Manifold Approximation and Projection for Clustering Taxa through Vocalizations in
[8] a Neotropical Passerine (Rough-Legged Tyrannulet, Phyllomyias burmeisteri). Animals (Basel). 2020;10.

[16] Campbell KR and Yau C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. Nat Commun. 2018;9:2442.