Distribution Agreement

Signature:

________________________________ _______________________________

Yawei Wang Date

Exploration of Normalization Methods on Bulk RNA-seq Data and Single-cell RNA-seq Data

By

Yawei Wang

Master of Science in Public Health

Biostatistics and Bioinformatics

Zhaohui (Steve) Qin, Ph.D.

(Thesis Advisor)

Max Lau, Ph.D.

(Reader)

Exploration of Normalization Methods on Bulk RNA-seq Data and Single-cell RNA-seq Data

By

Yawei Wang

B.S.

Hei LongjiangUniversity

2018

Thesis Advisor: Zhaohui (Steve) Qin, Ph.D.

Reader: Max Lau, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree of
Master of Science in Public Health in
Biostatistics and Bioinformatics
2021

Abstract

Exploration of Normalization Methods on Bulk RNA-seq Data and Single-cell RNA-seq Data

By Yawei Wang

Background: RNA-seq and single-cell RNA-seq are powerful new technologies in biomedical research. To eliminate the inherent technical errors associated with factors like sequencing depth and gene length, RNA-seq data from different samples need to be normalized so that they are comparable. However, the presence of abundant zeros in the data, especially in single-cell RNA-seq data, makes the normalization effect extremely challenging.

Method and Materials: In the bulk RNA-seq normalization section, I used a novel normalization method, named Group method, and compared its performance with other bulk RNA-seq data normalization methods, Upper Quantile, Quantile, Median, TMM, and DESeq, by calculating Spearman correlation between normalized RNA-seq data and TaqMan qRT-PCR data. We also compared their effectiveness on simulated data and differential expression analysis respectively. For the single-cell RNA-seq part, I merge genes based on the KEGG pathway and use the Quantile method to normalize pathway-cell data, which was named the Pathway-Quantile method. I compared this method with log normalization method, scran, and Linnorm on 3k PBMC data (without spike-in genes) and human pancreas data (with spike-in genes) by using the results after UMAP reducing dimension and Seurat package, version 4.0.1 visualizing.

Results: For simulated and real bulk RNA-Seq data, all normalization methods performed similarly in terms of the Spearman correlation between normalized real RNA-Seq data and MAQC TaqMan qRT-PCR data. And Group method does not perform better compared to other methods. For differential expression analysis, all methods showed similar performance. For single-cell RNA-seq data, Pathway-Quantile, is better than pathway-level data, but its performance was inferior to other methods when test on 3k PBMC data.

Conclusion: We found the group method is competitive for normalizing bulk RNA-seq data. However, more studies are needed for normalizing single-cell RNA-seq data using the Group-Quantile method.

Exploration of Normalization Methods on Bulk RNA-seq Data and Single-cell RNA-seq Data

By

Yawei Wang

B.S.

Heilongjiang University

2018

Thesis Advisor: Zhaohui (Steve) Qin, Ph.D.

Reader: Max Lau, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree of
Master of Science in Public Health in
Biostatistics and Bioinformatics
2021

Acknowledgement

First, I would like to thank my thesis advisor, Professor Steve Qin. Steve Qin provided guidance and direction for my thesis. In the process of writing the thesis, he gave careful and timely advice to the difficulties and problems I encountered, put forward many beneficial suggestions for improvement, and put a lot of effort into it. I also thank my thesis reader, Dr. Max Lau, who helped me check and reviewed the thesis.

Besides, I would like to thank the faculty and students of the Biostatistics and Bioinformatics Department of the Rollins School of Public Health, providing me a great and unforgettable time.

At last, I would like to express my gratitude to my parents in China who support me all the time. Without them, I could not have spent two fulfilling years in the United States and realized my dream.

Contents

1. Introduction

Microarrays provide the ability to study many genes in organisms under different biological conditions and significantly reduce cost and time. In recent years, high-throughput sequencing, also named next-generation sequencing (NGS), is applied to a technique for sequencing DNA and RNA in a rapid and low-cost manner and plays an increasingly important role in detecting gene behavior and has been used in related research.[1]

High-throughput transcriptomic sequencing, also known as RNA-seq, first fragmented and then reversely transcribed to cDNA. Then the fragments would be sequenced and compared with known reference genomes or gene transcriptomes or assembled without reference. The number of readings mapped to a gene was used to quantify its expression [2]. During these procedures, many factors may influence the reading counts of gene, including within-sample factors and inter-sample factors. In sample factors, like gene length and GC content, will affect the comparison of different genes within a sample. Inter sample factors, sequencing depth (i.e. library size), influence comparison of same genes' read count between different samples, that is longer transcripts tend to be cut into more fragments than shorter transcripts [3]. Therefore, the number of reads in the transcript is not only directly proportional to its expression level but also directly proportional to its length. The high expression level of the long sequence gene misestimates the true expression level of the gene. Thus, to reduce these noises and make them comparable, normalization, including in sample normalization and inter-sample normalization, are necessary. In this study, we focus on inter-sample normalization [4].

Over the past decade, high-throughput sequencing technology has been widely used in various fields of biology and medicine with many successes. Among them, transcriptomic sequencing (RNA-seq) has been widely used to determine and characterize the expression of genes or transcripts of various species. Bulk RNA-seq, a traditional transcriptomic sequencing technique, that is based on a population of cells with tens of thousands of cells per sample. It mostly reflects the average level of gene expression of a population of cells, and therefore masks the heterogeneity of expression of different cells. In recent years, single-cell RNA-seq (scRNA-seq) technology has been rapidly developed, which enables genome-wide expression of all genes to be revealed at the single-cell level, which is very beneficial for understanding heterogeneity in inter-cell expression [5].

However, it is typical that there are abundant zeros in RNA-seq data, which might influence the results of normalization. In this study, we propose a novel normalization method in bulk RNA-seq data, named the Group method. We compared the performance of the Group method and other bulk RNA-seq normalization methods in the bulk RNA-seq data, simulated data, and their performances on detecting differential gene expression. Single-cell RNA sequencing (scRNA-seq) has the characteristics of low capture rate, high noise, high variability, and more sparse counts, so we also proposed a new method, named the Pathway-Quantile method, which merges genes according to the pathway sets and then use Quantile method to perform normalization. We compared the Pathway-Quantile method with other single-cell normalization methods in 10X Genomics real data set of peripheral blood mononuclear cells (PBMCs) and spike-in Single-cell RNA-seq data of the human pancreas. In this study, we used KEGG pathway gene sets.

2. Data Sources

2.1 Bulk RNA-seq data

2.1.1 Real data

High-throughput RNA-Seq data used in this study were collected from Gene Expression Omnibus (GEO), The dataset series is GSE47774, the data are primary results from the Sequencing Quality Control (SEQC) project. The well-characterized reference RNA Sample A: Universal Human Reference RNA (UHRR) from Stratagene and ERCC Spike-In controls; Sample B: Human Brain Reference RNA (HBRR) from Ambion and ERCC Spike-In controls. Samples C and D were then constructed by combining A and B in mixing ratios, 3:1 and 1:3, respectively. The information of sample data we used is shown in Table 1.

Table 1. Description of RNA-Seq data

Term	Sample	Platform	Site	Library ID	Lane
SRX302874	Sample A (UHR)	Illumina HiSeq 2000	MAY	1	01
SRX302876	Sample A (UHR)	Illumina HiSeq 2000	MAY	1	02
SRX302890	Sample A (UHR)	Illumina HiSeq 2000	MAY	2	01
SRX302130	Sample A (UHR)	Illumina HiSeq 2000	BGI	1	01
SRX302954	Sample B (HBR)	Illumina HiSeq 2000	MAY	1	01
SRX302956	Sample B (HBR)	Illumina HiSeq 2000	MAY	1	02
SRX302970	Sample B (HBR)	Illumina HiSeq 2000	MAY	2	01
SRX302210	Sample B (HBR)	Illumina HiSeq 2000	BGI	1	01
SRX303034	Sample C (A:B=3:1)	Illumina HiSeq 2000	MAY	1	01
SRX303036	Sample C (A:B=3:1)	Illumina HiSeq 2000	MAY	1	02
SRX303050	Sample C (A:B=3:1)	Illumina HiSeq 2000	MAY	2	01
SRX302290	Sample C (A:B=3:1)	Illumina HiSeq 2000	BGI	1	01
SRX303114	Sample D (A:B=1:3)	Illumina HiSeq 2000	MAY	1	01
SRX303116	Sample D (A:B=1:3)	Illumina HiSeq 2000	MAY	1	02
SRX303130	Sample D (A:B=1:3)	Illumina HiSeq 2000	MAY	2	01
SRX302370	Sample D (A:B=1:3)	Illumina HiSeq 2000	BGI	1	01

2.1.2 MAQC TaqMan qRT-PCR data

In this study, we used MAQC TaqMan qRT-PCR data, GSE5350, as the benchmark to qualify the different normalization methods' performance, which can be downloaded from the Gene Expression Omnibus on the platform GPL4097. We chose GSM129638 whose source name is MAQC sample A to be the benchmark to qualify real RNA-Seq data of sample A; GSM129642, source name is MAQC sample B, to be the benchmark to qualify real RNA-Seq data of sample B; GSM129646, source name is MAQC sample C, to be the benchmark to qualify real RNA-Seq data of sample C and GSM129650 with source name sample D to be the benchmark to qualify real RNA-Seq data of sample D. The MAQC TaqMan qRT-PCR data is comprised by 1044 genes from two types of samples (HBR and UHR). We matched the Gene Expression Omnibus genes and genes in MAQC TaqMan qRT-PCR data by gene ID and there were 996 genes left

[5].

2.1.3 Simulated data

We selected SRX302874 (sample A), SRX302954 (sample B), SRX303034 (sample A: sample B=3:1), SRX303114 (sample A: sample B=1:3) to be basic dataset and simulated each genes' read count by Binomial (n, p), n is the read count of each gene and p is the probability we would specify, then simulated datasets will be generated.

2.2 Single-cell RNA-seq data

2.2.1 Peripheral Blood Mononuclear Cell (PBMC) data

This dataset is collected from PBMCs from a Healthy Donor, and it is a Single Cell Gene Expression Dataset from 10X Genomics' latest GemCode platform containing many datasets used in scRNA-seq studies, typically dealing with a larger number of cells at a sparser level. There is no spike-in gene and 2700 PBMCs, 2638 of which were used in this study. This dataset was used to compare the Group-Quantile method and other single-cell RNA-seq normalization methods.

2.2.2 Single-cell RNA-seq data with spike-ins

Spike-in RNA is an RNA transcript with known sequence and quantity, which could be used to calibrate RNA hybridization. Based on assumption that spike-ins and endogenous transcripts have similar expression levels across cells, adding spike-in RNA could eliminate the noise of sing-cell RNA-seq data and improve the efficiency of normalization [6]. In this study, we used a spike-in dataset from ArrayExpress. The dataset, E-MTAB-5061, is a single-cell RNA-seq analysis of the human pancreas from healthy individuals and type 2 diabetes patients. In this study, there are 1936 cells were used. This dataset was also used to compare the Group-Quantile method and other single-cell RNA-seq normalization methods.

2.2.3 KEGG pathway gene sets

KEGG pathway gene sets were used to collect genes from all fully sequenced genomes and some partial genomes of the gene catalog. In this study, we used the KEGG pathway gene sets containing 189 gene sets and 12797 genes. KEGG gene sets with Gene Symbols could download from http://www.gsea-msigdb.org/gsea/downloads.jsp.

3. Methods

3.1 Bulk RNA-seq normalization Methods

To eliminate the sequencing depth (library size) effects between samples, inter-sample normalization is necessary. Usually, in RNA-seq data, many genes have zero expression and would be filtered out by some normalization methods, which possibly generate bias for normalization results. In this study, we used the Group method to reduce the effect of zero read count and compare it with the Upper Quartile method [7], Quantile method [7], Median method [8], Trimmed Mean of M-values (TMM) [9] and DESeq [10]. After normalization, we calculated the Spearman correlation between normalized gene data and MAQC TaqMan qRT-PCR data. We also calculated the Spearman correlation between raw data that do not have been normalized and MAQC TaqMan qRT-PCR data by using 996 matched genes between GEO data and MAQC TaqMan qRT-PCR data. Then we used the Spearman correlation to evaluate different normalization methods on simulated data. At last, we compared the performance of normalization methods on detecting differential expression genes.

3.1.1 Traditional normalization methods with real RNA-Seq data

RC represents the raw read count of genes in RNA-Seq data; the Upper Quartile method is applying the upper quartile (75th) to all read count of genes. It is a scaling normalization method that forces the upper quartile of each lane to be the same. In this study, we used the function “BetweenLaneNormalization” in package “EDASeq” of R studio and specify “upper” in method selection; the Quantile method is one of the widely used preprocessing techniques, which could be used to remove technical noise from genomic data. It is a non-linear full quantile normalization method, which is the same as the Upper Quartile except specifying the method to be “full” in function “BetweenLaneNormalization”; Median approach is also a scaling normalization that forces the median of each lane to be the same; Trimmed Mean of M-values (TMM) is a method used in package “edgeR” of R studio. The assumption of TMM is most of the genes are not differential expression genes. This method removes all genes which do not express and then the sample with a relatively average data trend is used as a reference sample and all others are test samples. For each test sample, the scaling factor is calculated based on the weighted mean of log ratios between the test sample and the reference sample. The weighted average represents the normalized factor for the sample. DESeq method is applied in package “DESeq2” of R studio. Use the “estimateSizeFactors” function to calculate a factor for each sample which is named as size factor, then raw read count of each sample would be divided by the corresponding sample’s size factor and get the normalized result. This method is based on the negative binomial distribution.

3.1.2 Group method

In real RNA-Seq data, it is common that a lot of genes have zero read count. When do normalization, some methods will remove the genes with zero reads count, such as TMM, or genes expression characters will be masked by robust normalized method, like Quantile, which may lead to bias to some extent. To reduce the influence of genes with zero read count for normalization and keep their characteristics at the same time, we try to group genes and use each group's mean of read count to do normalization by using the Quantile method. Then each original RNA-Seq read count will be divided by its corresponding mean of the group's read count and multiply the normalized each mean of the group's read count. Using the mean of the groups' read count to do normalization could reduce the number of zero read count and decline their influences on normalization results. The reason for choosing the Quantile method is it will not remove any gene and it is a robust method when the number of zero read count is small. Using the original RNA-Seq data divided by its corresponding mean of the group's read count is a method to reduce the library size effect. Multiplying the normalized mean of the group' s read count could make results to be more powerful. The model of the Group method is shown as follows.

$$NRC_{ijk} = \frac{RC_{ijk}}{M_{ij}} \times NM_{ij}$$

NRC: normalized gene read count
RC: non-normalized gene read count
M: the mean of group' read count
NM: normalized mean of group' read count
i : index of sample

j : index of group
k : index of each gene read count in certain group and certain sample

In this study, we randomly grouped 43919 transcript ID with its corresponding read count into 1187 groups and the group size is 37. Calculating each group's mean of the read count in each sample and using the Quantile method to normalize the mean of the read count and get the normalized the mean of the read count. Then the original RNA-Seq data was divided by its corresponding mean of the group' read count and multiplied the normalized mean of its mean of the group's read count, we would get the normalized dataset at last.

3.1.3 Normalization with simulated data

SRX302874 (sample A), SRX302954 (sample B), SRX303034 (sample A: sample B=3:1), SRX303114 (sample A : sample B=1:3) are basic dataset and simulate each gene's read count in four samples by Binomial (n, 1), Binomial (n, 0.5), Binomial (n, 0.25), Binomial (n, 0.125), Binomial (n, 0.05), Binomial (n, 0.1) respectively. Use the Group method, upper quantile method, quantile method, median method, trimmed mean of M-values (TMM), and DESeq to do normalization and then calculate Spearman correlation between MAQC TaqMan qRT-PCR data and normalized simulated data.

3.1.4 Differential expression analysis

(1) Simulate SRX302874 (sample A) dataset by Binomial (n, 0.05) five times and get samples named as A1, A2, A3, A4, A5. (2) Randomly select 10% transcript ID (4392) with their corresponding read count from SRX302874 (sample A) as

differential expression genes. (3) For differential expression genes, stimulate read count by Binomial (n, p=0.1) and the remaining 90% genes by using binomial (n, p=0.05) to generate samples named as B1-B5. Here n is the read count for each gene in sample A. (4) Then use Upper Quartile method, Quartile method, Median method, Trimmed Mean of M-values (TMM), and DESeq and Group method to do normalization. (5) Use normalized A1-A5 and B1-B5 to calculate the ratio (B1+B2+…+B5+0.5)/ (A1+A2+…+A5+0.5), adding 0.5 could avoid zero in the denominator. Rank all ratios, select the top 4000 ratios, and count the number of differential expression genes which are correctly detected. (6) Repeat step three to step five 10 times and report mean and standard deviation. (7) Change the percent of differential expression genes from 10% to 20%, 30%. Change the p in Binomial (n, p) from 0.10 to 0.15 and 0.20, so there are 9 combinations and do the same things as before.

3.2 Sing-cell RNA-seq normalization methods

In this section, to compare the performance of different scRNA-seq normalization methods, including log normalization method, linnorm, scran, and our new method—Pathway-Quantile, we employed Uniform Manifold Approximation and Projection (UMAP) [11] analysis, which is a technique used for dimension reduction. It is suitable for large datasets and the analysis is performed by the runUMAP() function within the Seurat package, version 4.0.1. The Dimplot() function in the Seurat package is used to visualize the results.

3.2.1 Log Normalization

We used the NormalizeData() function to perform a global normalization, that is gene counts of each cell are divided by the total counts of the corresponding cell and multiplied the scale factor, then we used log(x+1) to transform the results.

3.2.2 Linear Model and Normality Based Normalizing Transformation Method (Linnorm)

Linnorm.Norm() function was used to do normalization. The assumption of Linnorm is genes are stably expressed across different samples and normalization parameters were calculated by utilizing these stably expressed genes [12].

First, normalizing each sample into a relative scale as the formula shown, where E_{ij} is the expression level of the gene i and sample j, m be the total number of genes and n be the total number of samples.

$$R_{ij} = \frac{E_{ij}}{\sum_{i=1}^{m} E_{ij}} \, (1 \le i \le m, 1 \le j \le n)$$

Then we define gene expression level as G_{ij} as $G_{ij}=ln(\lambda R_{ij})$, where λ represents the median of total counts across all cells. The expression means z_i are expressed as $z_i=a_jX_{ij}+b_j$, where parameters a and b are estimated through the linear model. At last, use the normalization strength coefficient c ($0 \le c \le 1$) and set it to 0.5 by default in the following process, then the normalized data can be generated:

$$a^{updated}{}_j=c(a_j-1)+1$$

$$b^{updated}{}_j=b_j*c$$

$$B_{ij}=\exp\left(a^{updated}{}_j\, G_{ij}+b^{updated}{}_j\right)$$

$$Y_{ij}=\ln (B_{ij}+1)$$

3.2.3 Scran method

Scran uses a variation of the counts per million (CPM) method designed for scRNA-seq. This method eliminates problems from zero counts by combining groups of cells and then calculates normalized factors in a CPM manner. Since a single cell will appear in multiple combined sets (pools), cellular-specific factors can be calculated by linear algebra by deconvolution from nonspecific factors.

This method has better performance on batch correction and difference analysis [13]. This method could be implemented in the “scran” package of R studio.

3.2.4 Group-Quantile method

After mapping with 3k PBMC data, there are 8501 genes and 2700 cells left and after mapping with E-MTAB-5061 data, there are 11711 genes and 1937 cells left. Then group genes and sum up their counts according to corresponding pathways for each cell. Then the gene-cell data was replaced by pathway-cell data. At last, using Quantile method to normalize the pathway-cell data.

4. Result

4.1 Results of bulk RNA-seq data

For real RNA-Seq data, normalization methods, Upper Quartile method, Quantile method, Median method, Trimmed Mean of M-values (TMM), DESeq, and Group method do not result in significant difference on Spearman correlation between normalized real RNA-Seq data and MAQC TaqMan qRT-PCR data compared with the Spearman correlation between raw RNA-Seq data and MAQC TaqMan qRT-PCR data.

The performances of the six normalization methods are similar (Table 1).

Table 1 Spearman correlation between normalized RNA-Seq data and MAQC TaqMan qRT-PCR data

		RC	UQ	Median	TMM	Quantile	DESeq	Group
sampleA	SRX302874	0.873	0.873	0.873	0.873	0.873	0.869	0.873
	SRX302876	0.872	0.872	0.872	0.872	0.872	0.870	0.871
	SRX302890	0.873	0.874	0.873	0.872	0.874	0.871	0.873
	SRX302130	0.875	0.875	0.875	0.874	0.875	0.870	0.875
sampleB	SRX302954	0.871	0.871	0.871	0.870	0.871	0.870	0.871
	SRX302956	0.872	0.872	0.872	0.870	0.871	0.871	0.871
	SRX302970	0.872	0.872	0.872	0.871	0.872	0.869	0.872
	SRX302210	0.869	0.869	0.869	0.868	0.869	0.866	0.868
sampleC	SRX303034	0.843	0.843	0.843	0.843	0.843	0.839	0.843
	SRX303036	0.841	0.841	0.841	0.841	0.872	0.837	0.841
	SRX303050	0.843	0.843	0.843	0.842	0.843	0.839	0.843
	SRX302291	0.845	0.845	0.845	0.844	0.845	0.835	0.845
sampmeD	SRX303114	0.832	0.830	0.830	0.829	0.830	0.827	0.830
	SRX303116	0.832	0.832	0.832	0.8291	0.832	0.830	0.832
	SRX303130	0.830	0.830	0.830	0.828	0.830	0.825	0.829
	SRX302370	0.829	0.829	0.829	0.828	0.829	0.827	0.829

The results of normalization on the simulated dataset are similar among the six methods, we can see that there is no obvious difference in the Spearman correlation between RNA-Seq data and MAQC TaqMan qRT-PCR data, no matter the normalized is the Group method or the traditional methods (Figure 2).

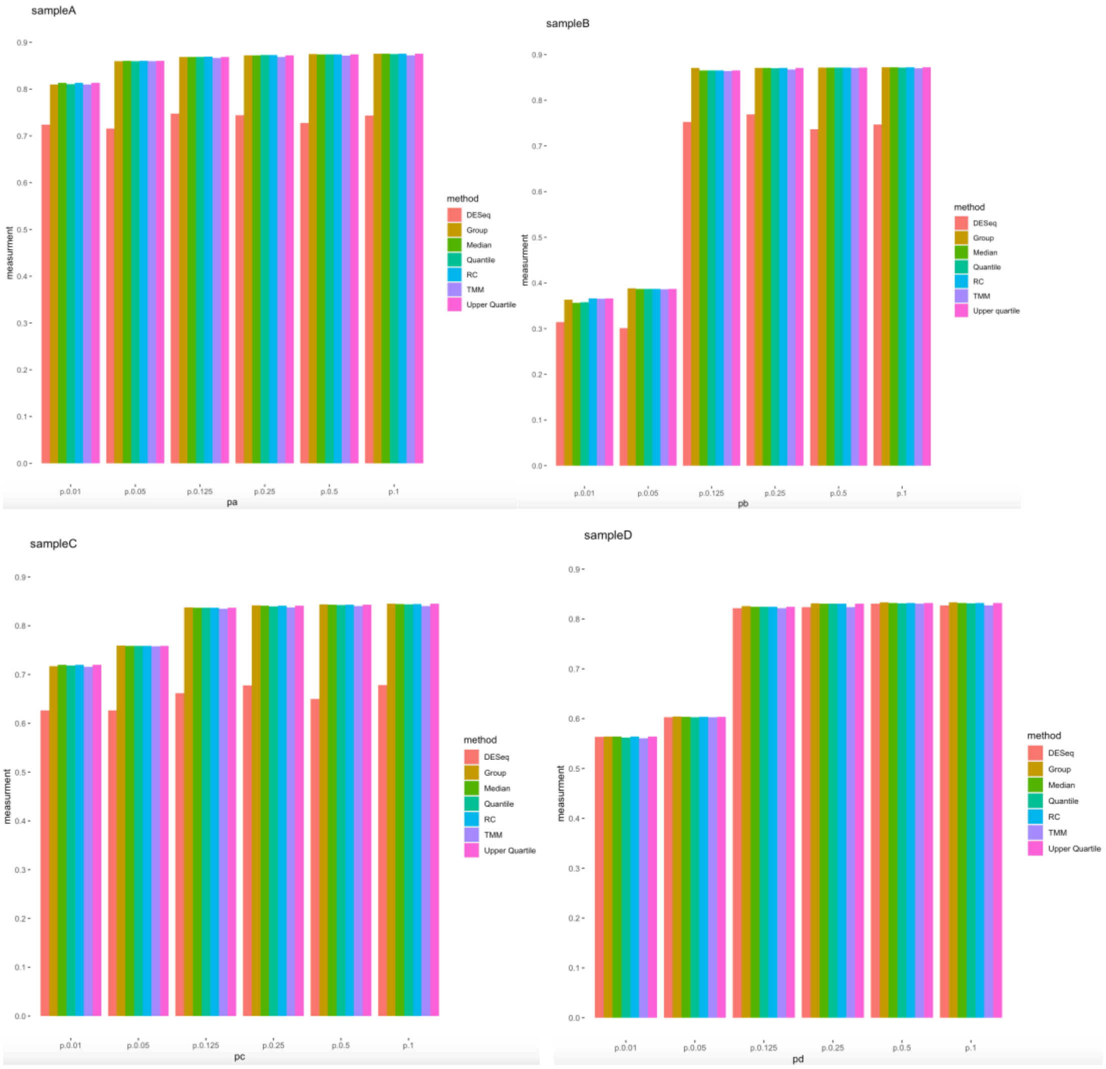


Figure 2

For differential expression genes detection, there is no obvious difference for means of the number of correctly detected differential expression genes among normalization methods, shown in Table 2. Under the same percentage of differential expression genes set (same n for Binomial distribution (n, p)), with the p increased, the number of correct detections of differential expression genes is increased for all methods and with the percentage of differential expression genes increased, the ratio of correct detection has little increase. There is no significant difference observed between the Group method and traditional methods no matter the means of correctly detection, the correct detection ratio, or standard deviation (figure 3).

Table 2. Percentage of differential expression genes correctly detected

Method	10% DE					
	P=0.10		P=0.15		P=0.20	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
UQ	58.49	0.65	62.29	0.57	68.57	0.54
Median	50.84	0.59	63.99	0.57	67.88	1.50
Quantile	49.51	0.86	63.03	0.53	68.49	0.57
TMM	50.39	0.62	63.48	0.40	68.15	0.51
DESeq	50.28	0.57	63.90	0.49	68.73	0.49
Group	50.61	0.67	63.92	0.54	68.67	0.40

Method	20% DE					
	10%:5%		15%:5%		20%:5%	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
UQ	62.44	0.35	71.43	0.28	74.00	0.33
Median	62.01	0.50	70.92	0.28	74.02	0.30
Quantile	64.01	0.42	70.97	0.42	72.98	0.36
TMM	64.26	0.45	71.28	0.40	73.25	0.34
DESeq	63.83	0.42	71.46	0.37	73.78	0.31
Group	64.07	0.46	70.97	0.39	73.11	0.32

Method	30% DE					
	10%:5%		15%:5%		20%:5%	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
UQ	68.67	0.32	73.49	0.25	75.08	0.41
Median	68.64	0.25	73.98	0.18	75.24	0.30
Quantile	67.72	0.33	72.95	0.26	75.20	0.32
TMM	68.22	0.41	72.54	0.23	73.99	0.42
DESeq	67.29	0.32	72.41	0.25	74.33	0.37
Group	67.90	0.36	74.02	0.29	76.34	0.36

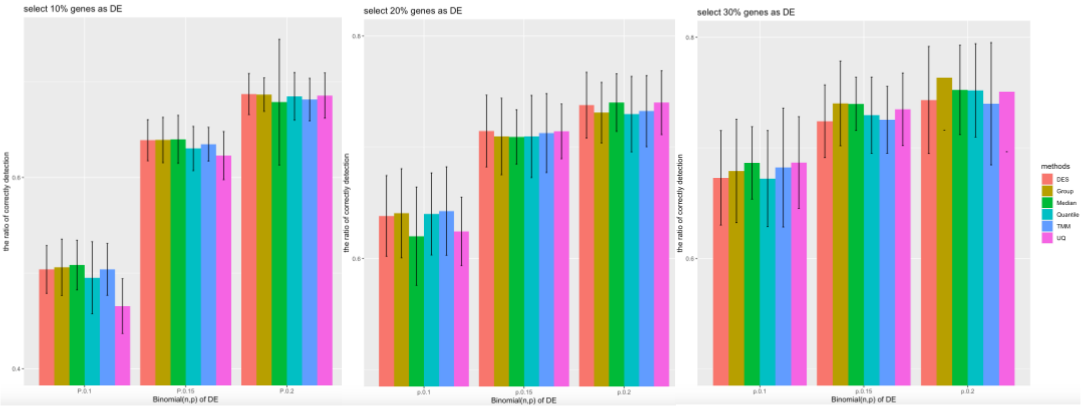
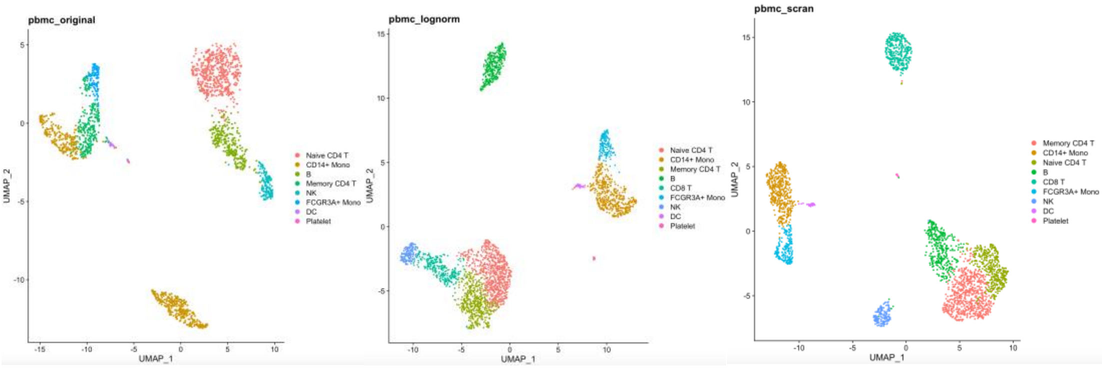


Figure 3

4.2 Results of single-cell RNA-seq data

4.2.1 PBMC data visualization

The results for the UMAP visualization are shown in Figure 4. Single-cell normalization methods, log transform, scran, and Linnorm, have similar performances with relatively clear division among the eight-cell types, but there is no distinct improvement compared with the original data. In comparison, the data which just using the pathway to merge genes and the data normalized by the Pathway-Quantile method were not grouped so clear as the above three methods and original data. However, the Pathway-Quantile method has improved the results compared with the result without the Quantile method normalized. Besides, there are just 6 clusters were grouped after using the pathway to merge genes.



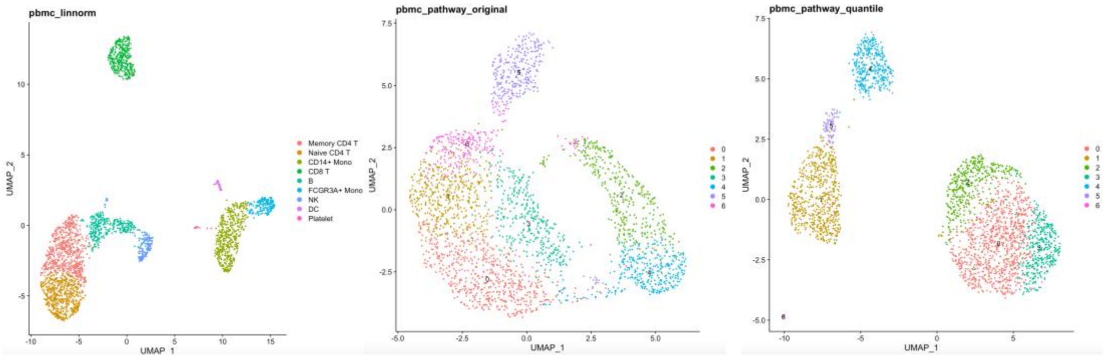
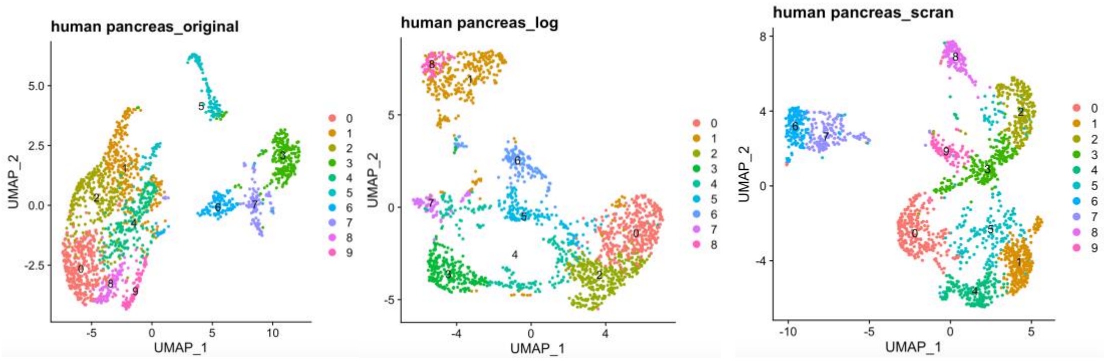


Figure 4

4.2.2 Spike-in single-cell RNA-seq data visualization

As Figure 5 shows, for pancreas data, no method separate conditions very well, though scran comes closest to doing so, which has the best performance with the clearest division among all methods and has improved compared with the original data. The result of data merged according to pathway’s genes sets and the results of the Pathway-Quantile normalization have better performance than the Linnorm method, and the Pathway-Quantile method also has a little bit of contribution to more clearly identify clusters than just using the pathway to merge genes.



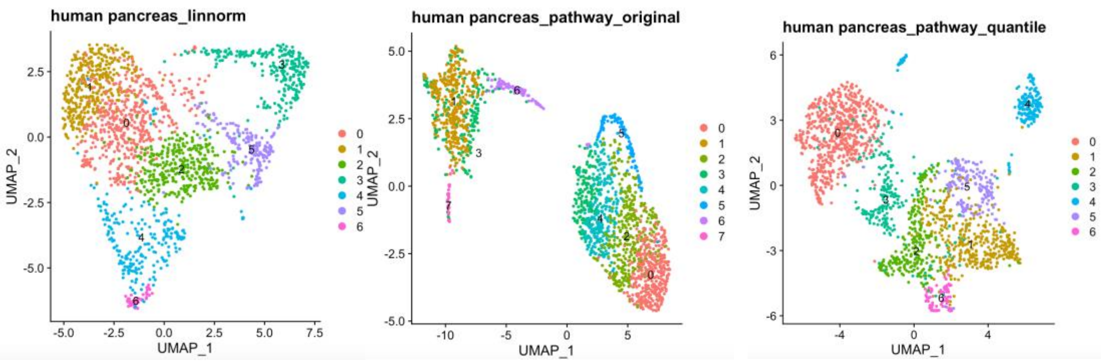


Figure 5

5. Conclusion and Discussion

In bulk RNA-seq part, the Group method does not have a better performance on real RNA-Seq data and simulated data, compared to the Upper Quartile method, Quartile method, Median method, Trimmed Mean of M-values (TMM), and DESeq according to Spearman correlation between normalized real RNA-Seq data and MAQC TaqMan qRT-PCR data. Also, all methods do not improve the Spearman correlation between normalized data and MAQC TaqMan qRT-PCR data compare with the Spearman correlation between raw data and MAQC TaqMan qRT-PCR data. In terms of detecting differential expression genes, all methods have similar performances. The challenging part for the Group method is the number of genes in each group, it should be guaranteed that the means of gene counts in each group for each cell should not be zero, because zero could not be the denominator. So, the group size will be varied in different data.

In the single-cell RNA-seq section, based on dimension reduction and SNN classification used in UMAP function of Seurat package, the Pathway-Quantile method did not contribute to noticeable improvements and did not perform as well as other single-cell RNA-seq normalization methods, log normalization, scran, Linnorm, cross 3k PBMC data. However, the Pathway-Quantile method outperforms only using the pathway to merge genes. Additionally, there are fewer clusters were identified after using pathway, no matter Pathway-Quantile normalized data or just using pathway without Quantile normalization because using pathway genes sets to filter genes will lead to loss of some features. Currently, there is also a limitation for pathway - it is not available to label clusters' cell type based on pathway-cell data.

For human pancreas data, which contains spike-in genes, although no methods could cluster cells clearly, the Pathway-Quantile method and just using pathway without Quantile normalization are more effective than Linnorm in terms of visualization. Similar to the results found in 3k PBMC data, there are fewer clusters were identified after using the pathway to merge genes due to some features lost. There are no advanced known cell makers in this dataset so that we could not effectively label cell types of clusters.

For future work, firstly, besides using visualization to evaluate the efficiency of different methods, we may consider using other statistics such as K-nearest neighbors (KNN) [15], to evaluate the performance of normalization methods. KNN has non-parametric nature and could work beyond two-group classification. Cohen's statistic [16] also could be used to test inter-rater reliability for categorical items. We may also explore the use of machine learning methods, like Random Forest classifiers, trajectory inference, on classifying. Secondly, we plan to evaluate the effectiveness of the Pathway-Quantile method on simulated data with greater differences in expression level between cells and differential expression genes' detecting analysis. Finally, we may use Pathway-Quantile method on other datasets and use different pathway gene sets to merge genes.

Reference

[1] Nkrumah-Elie, Y., Elie, M., & N Reisdorph, N., (2018). *Personalizing Asthma Management for the Clinician*, Elsevier.

[2] Evans, E., Johanna, H., Daniel, S. (2016). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinfor*m. 19: 776–792.

[3] Li, P., Piao, Y., Shon, H.S., Yongjun Piao, Ho Sun Shon & Keun Ho Ryu. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics.* 16: 347

[4] Abbas-Aghababazadeh, F., Li, Q., Fridley, BL. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS ONE*. 13:10.

[5] Haque, A., Engel, J., Teichmann, S.A. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*. 9.75. doi: 10.1186/s13073-017-0467-4

[6] Lun, A., Calero-Nieto, F., Haim-Vilmovsky, L., Göttgens, B., Marioni, J. C. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* 27:11, 1795–1806. doi: 10.1101/gr.222877.117

[7] Bullard J. H., Purdom E., Hansen K. D., Dudoit S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94

[8] Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 19: 185–193. doi: 10.1093/bioinformatics/19.2.185

[9] Robinson M. D., Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11: R25.

[10] Love M. I., Huber W., Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15: 550.

[11] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

[12] Yip, S.H., Wang, P., Kocher, J.-P.A., Sham, P.C., Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Research*, 45: 22, e179. doi: https://doi.org/10.1093/nar/gkx828.

[13] Lun, A., Bach, K., Marioni, J. C. (2016). Pooling across cells to normalize single-cell sequencing data with many zero counts. *Genome Biol.* 17, 75. doi: 10.1186/s13059-016-0947-7

[14] Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric Regression. *Am. Stat*. 46, 3. doi: 10.1080/00031305.1992.10475879

[15] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and*

Psychological Measurement, 20, 37–46. https://doi.org/10.1177/001316446002000104