Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Luxiao Chen Date 04/02/2018

K-mer Based Clustering for UMI Single Cell RNA-seq Data

By

Luxiao Chen

Master of Science in Public Health

Biostatistics and Bioinformatics

_____

Hao Wu, PhD
(Thesis Advisor)

_____

Zhaohui (Steve) Qin, PhD
(Reader)

K-mer Based Clustering for UMI Single Cell RNA-seq Data

By

Luxiao Chen

B.S./M.S.

Nanjing University

2013/2016

Thesis Committee Chair: Hao Wu, PhD

Reader: Zhaohui (Steve) Qin, PhD

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

2018

Abstract

Motivation:

High-throughput RNA sequencing (RNA-seq) is a technology to quantify the gene expression. It has been widely used in various areas of biological and clinical studies. Traditional RNA-seq ("bulk" RNA-seq) operates on the mRNA from a large number of cells, thus the measurement is an averaged expression levels of the input cells. For heterogeneous samples, the bulk RNA-seq fails to provide more detailed information for gene expression variation. Single cell RNA-seq (scRNA-seq) has recently emerged with the technological developments. It profiles the expression for each single cell, thus provide information for understanding the transcriptomic regulation and variation at cellular level. There are a number of data analysis challenges in analyzing the scRNA-seq data, among them the cell clustering is an important one.

Methods:

In this work, we aim to study the possibility of using DNA sequence content (the k-mer counts) instead of gene expression values in cell clustering. We first discussed the relationship between gene counts, UMI RNA-seq transcript counts and k-mer counts by giving out the mathematics expression of gene/k-mer counts related to transcript counts. Then we performed simulation to demonstrate the difficulty of scRNA-seq with low expression counts in particular from unique molecular identifier (UMI), potential advantage of using gene/k-mer count instead of transcript counts and comparison of clustering results between gene counts matrix and k-mer counts matrix.

Results:

We showed that gene/k-mer counts matrix is a transformation of UMI scRNA-seq transcript counts matrix. It can enlarge the value in expression matrix but may lose alternative splicing information stored in transcript counts. By comparing the performance of gene count matrix and k-mer count matrix with different signal noise ratios (SNR). We found long k-mer ($k = 8, 9, 10$) performs better than short ($k = 5, 6, 7$) k-mer. However, under same SNR scenario, gene count matrix still performs better in most scenarios.

Acknowledgement:

During the process of working on thesis, I would like to thank Prof. Hao Wu for his instructions and guidance. I also thank Dr. Hao Feng for his kindly help with my code. Finally, I appreciate the comment of Prof. Steve Qin for my thesis.

Table of Contents

Introduction

1. Genome and transcriptome

A genome is the complete set of genetic information in an organism, which is a store of all the biological information the organism requires to function [1]. In most of living organism, the genome is made of long molecules of DNA called chromosomes. Genes are small sections of DNA distributed on chromosomes. They are codes for the RNA and protein molecules required for all kinds of biological activities in organism. Different species have their own distinctive genomes. Even within same specie, genomes of different individuals are different.

Although a genome contains all the genetic information, on its own it cannot deliver the information to the cell. Utilization of the genetic information requires participation of enzymes and other types of proteins, which consists of complex biochemical reactions called genome expression.

There are two types of product generated by genome expression. The initial one is transcriptome defined as the full range of RNA molecules expressed by a genome. Sometimes the concept of transcriptome is confined to the collection of RNA molecules derived from protein-coding genes. The final product of genome expression is the proteome–full assortment of proteins. The proteome is synthesized under the direct of transcriptome and specifies the nature of the biochemical reactions that the cell is able to carry out [2].

In order to find the relationship between the genome and the cell function, researchers tried to study the proteins as well as expressed RNA. Proteomics is the area of studying the

quantity and diversity of proteins in a certain cell or tissue. However only studying proteome cannot have the whole picture of the biological activities. Also due to the changeability of proteins, it's hard to capture their status at best. What's more, the post-transcriptional modification and the amplification of proteins are still main barriers for scientists to characterize the proteins. Luckily, researchers can still go on studying the biological activity by measuring the transcripts of messenger RNA (mRNA) [3]. It is an intermediate step between genes and proteins that bridges the gap between the genome and the functional molecules in cells.

In multicellular organisms almost all the cells share the same genome, so that they have the same genes. However, the genes are not expressed at the same time in different cells. For each gene they may even have more than one type of mRNA, due to the alternative splicing, RNA editing or alternative transcription termination sites. Actually, for different cells there are different expression patterns of these genes. The variation of the transcriptomes can be related to different physical, biochemical, as well as developmental conditions. It also may lead to the difference between status of health and disease [3].

Thus, by comparing the content of transcriptome between different cells or tissues, researchers can have a deeper understanding what variation of transcription activity may lead to the transformation from normality to disease. It is also possible to know what content of transcriptome contributes to the specificity of certain cells or tissues and the genes driving the development of cells at different stages.

2. High-throughput technology to measure gene expression

Characterizing the gene expression is an important method to know the molecular

composition and status of cells or tissues as well as the feedback to the stimulation of environment. Traditional quantitative analyses of RNA by Northern blots or quantitative PCR are limited to studying individual transcripts at a time. Although these methods can provide expression information of specific interested genes, they cannot meet the demand of having a global view of gene expression changes in a biological system. The development of high-throughput technologies (Microarray and RNA-seq) for transcriptome study brings light to the systematics analysis of gene expression.

The earliest microarray method was raised in 1981 [4] and now microarrays have been applied for various kinds of biological studies because they can provide a cheap and efficient method to evaluate the mRNA levels for thousands of genes at once. The principle of microarray technology is that mRNAs samples labeled with fluorescent dyes are hybridized in parallel to a large amount of DNA sequences probes immobilized on a solid surface. Then the fluorescent dye is stimulated by laser light and gives out fluorescent emission, which represents the hybridization intensity. Finally the relative amounts of the different transcripts can be estimated based on the strength of fluorescent emission [5]. Furthermore, mRNAs from two different samples can be detected with the combination of two different fluorophores in one array [6].

With the application of microarrays, biologists can identify individual gene, whose regulated expression may explain specific biological phenomena. For example, with usage of microarrays in analysis of cancers, scientists can easily identify specific abnormal expressed genes in tumor cells [7]. Scientists can also analyze cell functions based on the global scope of genes expression patterns.

However, microarrays can only detect known sequences, which limits its

application to detect unknown sequences. Another limitation is the inaccurate measurement of microarray that gene expression estimates may be interfered by the background hybridization and probe saturation [8]. Although microarray had been used in various researches, the appearance of RNA-seq almost completely supplanted it.

In contrast to the microarray methods, RNA-seq can determine the cDNA sequences directly. RNA-seq uses the developed deep-sequencing technologies. The workflow of RNA-seq starts with the conversion of pre-treated RNAs to a library of cDNA fragments with adaptors attached to one or both ends. After amplification of these molecules, short sequences from one end or both ends are obtained by sequencing in a high-throughput manner. Generally, the sequences are 30 – 400bp, which depends on the platform used and the types of RNAs [9].

RNA-seq can be used to detect different types of RNA, such as mRNAs, microRNAs, long-noncoding RNAs, both qualitatively and quantitatively. It has several advantages compared to microarray method. First, RNA-seq can be used to detect unknown sequences. Second, because of the DNA sequences can be mapped to unique regions of genome, the background noise is very low for RNA-seq. Third, RNA-seq does not have an upper limit for quantification, which is correlated with the number of sequences obtained [9]. Taking these advantages, RNA-seq has been used to analysis of RNA isoforms to study the alternative splicing. It can also be applied to assemble de novo transcriptomes for organisms without sequenced genomes.

Even though RNA-seq has so many advantages, there are still several challenges for it. For instance, in upstream samples preparation, the fragmentation step during the library construction can lead to the bias of outcome. Also the complex downstream data

analysis requires robust algorithm for normalization, differential expression testing as well as isoform expression estimation [8].

3. Single cell sequencing

Single cell is the fundamental unit of organism [10]. Since the first observation of multicellular structure by Robert Hooke in 1665, Biologists have devoted themselves to study the form and function of cells aiming to give accurate cell classification. It is known that unique transcriptome signatures can characterize cell identity and function [11]. Not only different tissues have distinct gene expression patterns, but also cells in consecutive development process may have significant difference in transcriptome. Most of the RNA-seq studies mentioned above have focused on analyzing bulk tissue samples composed of millions of cells, which are sufficient to do analysis at organism level. However, there may be different subtypes of cells within the same sample, then the resulting expression values for each gene is an average of its expression levels across a large population of different input cells. So, such bulk RNA-seq is not sufficient to depict the details of the cells diversity for many biological questions. For instance, there are many distinct cell types that are difficult to dissect in samples of brain tissue. Then bulk measurements confound changes due to gene regulation with those due to shifts in cell type composition. Another example is for time series studies of gene expression. It is known that transitions from one status into another status of differentiating cells don't start at the same time. Surely sampling a population of them at any time point will contain cells staying at different stages. Tracking the averaged expression of a gene at certain moments may lead to incorrect cell differentiation pattern. So traditional bulk RNA-seq application is constrained by averaging.

The scRNA-seq is required for more detailed study at cell level.

With the more detailed, higher resolution information provided by scRNA-seq, scientists can extend their work in different areas. For instance, single cell RNA-seq can be used in resolving microbial genomes and delineate cell-to-cell diversity within diverse population for microbiology study. The method has also been applied in classifying neurons based on transcriptional profiles instead of morphological features for neurobiology studies. Furthermore, when doing cancer research, scRNA-seq can provide powerful new tools for delineating clonal diversity and understanding the role of rare cells during cancer progression. There are also many other fields where scRNA-seq has proved its unique value, including germline transmission, embryogenesis, tissue mosaicism, organogenesis, immunology and clinical applications [12].

The general scRNA-seq protocol begins with isolating individual cells by laser-capture microdissection (LCM), fluorescence-activated cell sorting (FACS) or microfluidics techniques. Then after lysing the cell, cDNA is obtained by reverse transcription of polyadenylated fraction of mRNA. Finally, sequencing is applied after having enough material by amplifying the cDNA. In order to facilitate quantitative comparisons of the expression level of each gene between cells, usage of extrinsic spike-in molecules is recommended for all scRNA-seq experiments. The most widely used one is the External RNA Control Consortium (ERCC) set of 92 synthetic spikes based on bacterial sequences [13]. In sequencing step, it is possible that reads from 3' or 5' end of the amplified transcripts are much easier sequenced than fragments from other positions. So unique molecular identifiers (UMIs) have been used to mark individual molecules [14] to avoid the biases of amplification–the major source of technical variability.

Compared with the bulk RNA-seq, the differences not only exist in the pipeline of scRNA-seq but also with the data characteristics. First, for traditional bulk RNA-seq data, the sample number is low, which is much smaller than the gene number. However for scRNA-seq, the sample number can reach to 100,000 with the appearance of droplet-based platform [15]. Second one is that due to the low RNA input, the number of transcripts detected is much lower compared to bulk RNA-seq [16]. Some genes even can't be detected although they are expressed in cells – the so-called "dropout" phenomena.

Clustering is a key analysis step in studying cell differentiation. In recent years a large amount of algorithms have been designed to deal with the scRNA-seq data for clustering, such as Monocle [17], Waterfall [18], Wanderlust [19], TSCAN [20]. The ideas of these methods are similar: (1) select informative genes; (2) dimension reduction of GE; (3) cluster the cells based on reduce data; (4) construct a minimum spanning tree (MST) from the clustering results; (5) map cells to the MST. In 2016 Ntranos et al. reported a novel clustering method based on transcript-compatibility-counts [21] and in 2017 Zhe Sun et al. raised a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data [22].

4. K-mer

In genomics k-mer is referred as all the substrings of length k that obtained from a DNA string. When defining the length of a string is L, the amount of k-mer is L-k+1. If for each position, there are n possibilities then the number of possible k-mers is n^k. For example, there are 4 types of nuclear acids, A, T, C and G. Then the number of 3-mer types is 4^3= 64 (AAA, AAT, AAC, … TTT). It is obvious that longer k-mer is more unique that

fewer sequences (transcripts in RNA-seq) can contain it, while shorter k-mer can be shared by multiple different sequences (transcripts in RNA-seq).

K-mer has been applied in different kinds of bioinformatics analysis. It can be used for de novo sequence assembly [23], separating different species in a mixture of genetic material [24], alignment-free sequence analysis [25] and so on.

5. Relationship between gene counts, transcript counts and k-mer counts

5.1 Gene counts and k-mer counts – transformation of transcript counts

In UMI scRNA-seq, one gene in a single cell may have different transcripts sequenced due to the alternative splicing as well as TSS fusion. The transcripts here are sequences got in scRNA-seq. So, each gene may have several different transcripts as the 'MGene' matrix shown below.

$$MGene = \begin{bmatrix} a_{11} & \cdots & a_{1t} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{g1} & \cdots & a_{gt} & \cdots & a_{gn} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mt} & \cdots & a_{mn} \end{bmatrix}, m\ genes\ (row)\ \times\ n\ transcripts\ (col)$$

Suppose there were totally m different genes and n different transcripts in a sample. The rows represent m different genes while the columns represent n different transcripts. The value of element in 'MGene', a_{gt}, could be [0, 1], representing the probability/proportion of Transcripts 't' comes from Gene 'g' if Transcripts 't' were sequenced in scRNA-seq:

If $a_{gt} = 0$, it means Transcript 't' is from Gene 'g'.

If $a_{gt} = 1$, it means all the Transcripts 't' are uniquely from Gene 'g'.

If $a_{gt} \in (0, 1)$, then it means proportion of Transcript 't' derived from Gene 'g' while (1-a_{gt}) represents the proportion of Transcript 't' from other genes.

Suppose there are l cells being sequenced and the count of Transcript 't' in Cell 'c' is d_{tc}. If Transcript 't' is not captured in Cell 'c' during sequencing, then $d_{tc} = 0$. We can use matrix 'MTranscript' to represent the transcripts sequenced in different cells. It is a matrix with rows representing different transcripts and columns representing different cells in a sample as shown below:

$$MTranscript = \begin{bmatrix} d_{11} & \dots & d_{1c} & \dots & d_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{t1} & \dots & d_{tc} & \dots & d_{tl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nc} & \dots & d_{nl} \end{bmatrix}, n\ transcripts\ (row)\ \times\ l\ cells\ (col)$$

Since a gene count of a single cell in UMI scRNA-seq is defined as the sum of counts of transcripts with different UMIs from different locations of the gene [14], we can get the matrix of genes' counts in each cell ('M_Gene_Cell') by multiplying matrix 'MGene' and 'MTranscript' as below:

$$M\_Gene\_Cell = \begin{bmatrix} CG_{11} & \dots & CG_{1c} & \dots & CG_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CG_{g1} & \dots & CG_{gc} & \dots & CG_{gl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CG_{m1} & \dots & CG_{mc} & \dots & CG_{ml} \end{bmatrix}, m\ genes\ (row)\ \times\ l\ cells\ (col)$$

$$M\_Gene\_Cell = MGene \cdot MTranscript$$

$$\begin{bmatrix} CG_{11} & \dots & CG_{1c} & \dots & CG_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CG_{g1} & \dots & CG_{gc} & \dots & CG_{gl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CG_{m1} & \dots & CG_{mc} & \dots & CG_{ml} \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1t} & \dots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{g1} & \dots & a_{gt} & \dots & a_{gn} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mt} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} d_{11} & \dots & d_{1c} & \dots & d_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{t1} & \dots & d_{tc} & \dots & d_{tl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nc} & \dots & d_{nl} \end{bmatrix}$$

In Matrix 'M_Gene_Cell', rows represent different genes and columns represent difference cells. It has m rows (genes) and l columns (cells). Specifically, the Gene 'g' expression in Cell 'c' is $CG_{gc} = \sum_{t=1}^{t=m} a_{gt} d_{tc}$.

With the definition of k-mer, we can cut each transcript into small pieces – k-mers. So, there is a transformation relationship between k-mers and different transcripts sequenced

in scRNA-seq. For example, for transcript 'AT', then the corresponding 1-mer vector is (1, 1, 0, 0) for 'A'=1, 'T'=1, 'C'=0 and 'G'=0. Then for the n different transcripts assumed above, we can have a transformation matrix 'MKmer' to describe the k-mer content in each transcript. The matrix is as below:

$$MKmer = \begin{bmatrix} b_{11} & \dots & b_{1t} & \dots & b_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{k1} & \dots & b_{kt} & \dots & b_{kn} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{p1} & \dots & b_{pt} & \dots & a_{pn} \end{bmatrix}, p\ k-mers\ (row)\ \times\ n\ transcripts\ (col)$$

'MKmer' is a matrix with p rows (k-mer) and n columns (transcript). Due to the characteristics of DNA string, which is consisted of four types of nucleotides, we have $p = 4^k$. The element in 'MKmer', b_{kt}, represents number of k-mer 'k' contained in Transcript 't'. The value of b_{kt} is integer:

If $b_{kt} = 0$, it means k-mer 'k' is not part of Transcript 't'.

If $b_{kt} = 1, 2, 3, \dots$ it means there are b_{kt} k-mer 'k's can be cut out from Transcripts 't'.

Similar to the 'M_Gene_Cell' matrix above, we can easily get the k-mer counts information for each gene in different cells, if we know the transcripts expression information in cells. By multiplying the 'MKmer' and 'MTranscript' matrixes, we can have the matrix 'M_Kmer_Cell':

$$M\_Kmer\_Cell = \begin{bmatrix} CK_{11} & \dots & CK_{1c} & \dots & CK_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CK_{k1} & \dots & CK_{kc} & \dots & CK_{kl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CK_{p1} & \dots & CK_{pc} & \dots & CK_{pl} \end{bmatrix}, p\ k-mers\ (row)\ \times\ l\ cells\ (col)$$

$$M\_Kmer\_Cell = MKmer \cdot MTranscript$$

$$\begin{bmatrix} CK_{11} & \dots & CK_{1c} & \dots & CK_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CK_{k1} & \dots & CK_{kc} & \dots & CK_{kl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ CK_{p1} & \dots & CK_{pc} & \dots & CK_{pl} \end{bmatrix} = \begin{bmatrix} b_{11} & \dots & b_{1t} & \dots & b_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{k1} & \dots & b_{kt} & \dots & b_{kn} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{p1} & \dots & b_{pt} & \dots & b_{pn} \end{bmatrix} \cdot \begin{bmatrix} d_{11} & \dots & d_{1c} & \dots & d_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{t1} & \dots & d_{tc} & \dots & d_{tl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nc} & \dots & d_{nl} \end{bmatrix}$$

We have that $CK_{kc} = \sum_{t=1}^{t=n} b_{kt} d_{tc}$, which means in Cell 'c' the count of K-mer 'k' is CK_{kc}.

By comparing $CG_{gc} = \sum_{t=1}^{t=m} a_{gt} d_{tc}$ and $CK_{kc} = \sum_{t=1}^{t=n} b_{kt} d_{tc}$, we can see that both the gene count value and k-mer count value are transformation of transcripts expression counts in a single cell. The difference is determined by two transformation vectors $[a_{g1} \quad \dots \quad a_{gt} \quad \dots \quad a_{gn}]$ and $[b_{k1} \quad \dots \quad b_{kt} \quad \dots \quad b_{kn}]$.

$$CG_{gc} = [a_{g1} \quad \dots \quad a_{gt} \quad \dots \quad a_{gn}] \cdot \begin{bmatrix} d_{1c} \\ \vdots \\ d_{tc} \\ \vdots \\ d_{nc} \end{bmatrix}$$

$$CK_{kc} = [b_{k1} \quad \dots \quad b_{kt} \quad \dots \quad b_{kn}] \cdot \begin{bmatrix} d_{1c} \\ \vdots \\ d_{tc} \\ \vdots \\ d_{nc} \end{bmatrix}$$

5.2 Disadvantage and potential advantage of gene/k-mer count

CG_{gc}/CK_{kc} is sum of transcripts counts weighted by transformation vector. Due to alternative splicing is a common phenomenon in cell and when length of k-mer is short, different transcript sequences can contain same k-mer, these gene/k-mer counts actually are the combination of different transcripts expression levels, indicating that it mixes expression information from different transcripts. The disadvantage and potential advantage of such transformation can be shown by the following examples.

The disadvantage of using gene/k-mer counts is that while the two transcripts have opposite expression pattern in different cells, then the sum of transcripts counts may mitigate the difference of transcripts between two cells. Suppose two transcripts had different expression patterns in two difference cells. As shown in below 2 by 2 matrix:

$$\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}, \text{2 transcripts A, B (row)} \times \text{2 cells 1, 2 (col)}$$

In cell 1 transcript A has 10 counts (row 1, column 1) and transcript B has 0 count (row 2, column 1) and in cell 2 transcript B has 10 counts (row 2, column 2) while transcript A has 0 count (row 1, column 2). $[1 \quad 1]$ is the transformation vector. By multiplying the transformation vector $[1 \quad 1]$ and transcript expression matrix $\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$, we have gene/k-mer count expression in two cells as following:

$$[1 \quad 1] \cdot \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = [10 \quad 10]$$

We can see the gene/k-mer counts for two different cells are the same that both of them are 10.

It's easy to distinguish two cells with transcripts expression $\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$ while after transformation the expression pattern difference is eliminated as $[10 \quad 10]$.

The potential advantage is that while the two transcripts have same expression pattern in different cells, then the gene/k-mer counts may increase the difference of original transcripts expression counts between two cells. Suppose the two transcripts have the same expression pattern in two different cells. As shown in below 2 by 2 matrix:

$$\begin{bmatrix} 6 & 4 \\ 6 & 4 \end{bmatrix}, \text{2 transcripts A, B (row)} \times \text{2 cells 1, 2 (col)}$$

In cell 1 both transcript A and transcript B have 6 counts and in cell 2 they both have 4 counts, indicating that transcript A and B are higher expressed in cell 1. Similar to above, $[1 \quad 1]$ is the transformation vector. By multiplying the transformation vector $[1 \quad 1]$ and transcript expression matrix $\begin{bmatrix} 6 & 4 \\ 6 & 4 \end{bmatrix}$, we have gene/k-mer count expression in two cells as following:

$$[1 \quad 1] \cdot \begin{bmatrix} 6 & 4 \\ 6 & 4 \end{bmatrix} = [12 \quad 8]$$

We can clearly see that the difference between cell 1 and cell 2 for both transcripts A and B is |6 - 4| = 2. However, after transformation the difference is |12 - 8|=4, which is larger than original transcript difference. The enlarged difference may lead to better clustering, especially for the situation that transcripts difference between different cells is too small. However, this is only the simplest model. In real data analysis, summing counts of two different transcripts sharing same expression pattern in different cells may not improve the clustering accuracy even though it can enlarge the count difference between different cells. Because the clustering result is also affected by the variance of expression counts, sometimes the variance of gene/k-mer count is increased when summing different transcripts counts to increase the expression counts. Furthermore, due to the variability in gene expression, we can't get the exact transcripts expression variability and cell types information in advance. That's why emphasize 'potential' for advantage here.

Not considering the variance change, if we want to enlarge instead of mitigating the original transcripts difference signal, we hope transformation vector can include count information for transcripts sharing same expression pattern. Thus, the criteria to judge whether the transformation is better for clustering is that whether the transformation can make the difference of certain gene/k-mer expression between two different types of cells larger than any difference of transcript expression between two types of cells. That is between two types of cells C1 and C2:

1. For Gene:

$$|CG_{gc1} - CG_{gc2}| > max(\,|d_{1c1} - d_{1c2}| * I(a_{g1} > 0), \ldots\, |d_{nc1} - d_{nc2}| * I(a_{gn} > 0))$$

For K-mer:

$$|CK_{kc1} - CK_{kc2}| > max(\,|d_{1c1} - d_{1c2}| * I(b_{k1} > 0), \ldots\, |d_{nc1} - d_{nc2}| * I(b_{kn} > 0))$$

2. a. The criteria 1 can work for transcripts with larger difference between different types of cells.

   b. The criteria 1 can work in different genes/k-mers as many as possible.

5.3 Characteristics of gene/k-mer count transformation

After giving out the criteria, the next question is whether transformation from transcript counts to gene/k-mer counts can meet the two criteria above based on the characteristics of their own transformation matrix.

For gene, the transformation matrix 'MGene' is fixed. Because the element value is determined by whether a gene has the specific transcript. So, the exact value of a_{gt} (probability/proportion of transcript 't' derived from gene 'g') is determined by the genome content of the cell. Also, the dimension of the 'MGene' (number of genes and number of different transcripts) is fixed because it is also determined by a genome. In transformation vector $[a_{g1} \quad \cdots \quad a_{gt} \quad \cdots \quad a_{gn}]$ elements have values not equal to 0 is confined to the corresponding alternative transcripts for the specific gene. Because we know alternative splicing always happen in cells, so it is sure that using gene count instead of transcript count will lose alternative splicing information. But whether gene count can enlarge the transcript expression difference between different types of cells depends on whether the alternative transcripts sharing the same expression pattern in certain type of cells.

For k-mer, although the transformation matrix 'MKmer' is fixed after the length of k-mer is determined due to the fixed content of sequences, which is similar to 'MGene' matrix. However, the choice of length of k-mer is determined artificially and different length of k-

mer can have different dimension of the 'MKmer' and different element values in 'MKmer'. K-mer with longer length means that the k-mer is more unique for some specific transcripts. For instance, suppose the length of a sequence got from UMI scRNA-seq is 50nt, and k-mer is also 50nt. Then the transformation matrix 'MKmer' is an identity matrix and the k-mer counts in cells matrix 'M_Kmer_Cell' is totally the same as transcript counts in cells matrix 'MTranscript'. This makes no contribution to increase expression counts values for improving clustering accuracy. However long length k-mer can be used to reduce influence of sequencing error often appears at the end of sequences, because it not only can keep the uniqueness of transcript but also can remove the error-sequenced nucleotide at the same time. We can also consider applying k-mer counts for clustering. When the length of k-mer is very short then the probability that many transcripts have it increases. Considering the extreme situation of 1-mer. We are sure that almost all the transcripts contain single nucleotides A, T, C and G. So, for each 1-mer, its expression count in a single cell is influenced by all the expressed counts of transcripts and their 'A, T, C, G' contents. Obviously, the expression information of 1-mer count is too messy to improve the clustering accuracy. Based on the examples of two extreme situations, it's clear that in order to applying k-mer count to increase the expression counts difference requires us to choose appropriate length of k-mer, which is neither too short that messing transcript counts information nor too long that making no contributions to enlarging the expression difference between different cell types.

6. Purpose and content of this work

6.1 Purpose

Clustering the single cell RNA-seq data is the principle step to study the development process of cells and identify subtypes of cells. However, with the small amount of transcriptome in a single cell and low capture, reverse transcription efficiencies, the number of detected transcripts is much lower than bulk RNA-seq. The situation is more sever when UMIs are applied to avoid the amplification bias for getting the original mRNA content in a single cell. So, the resulting UMI counts matrix of single cell RNA-seq is very sparse, which raises more difficult challenge to get accurate clustering results.

Different clustering methods have been reported recent years as described above, but these methods use gene counts directly or the transcript-compatibility counts, so the clustering results can still be affected by the general low gene expression levels. Under the same condition, lower reads number means lower signal. Furthermore, all of these methods require the step of alignment to genome or pesudoalignment, which may be time consuming [15-20].

Audoux et al reported that biological variation of RNA-seq data could be captured by k-mer [21]. By comparing the 31-mer got from raw RNA sequencing data of different samples, they found that the k-mers not only can represent the tissue specificity but also can identify various types of biological events, including differential splicing, differential polyadenylation, intron retention and so on.

Based on the above discussed relationship between k-mer counts, gene counts and transcripts counts in UMI scRNA-seq and findings of Audoux et al that k-mer in RNA-seq can show tissue specificity, we planned to explore the possibility of using k-mer counts

instead of gene counts to do clustering for scRNA-seq data, which aiming at saving the step of alignment while increasing the abundance of expression values in matrix by cutting sequences into small pieces. Also, during the process, we also planned to show the influence of low expression counts to clustering results as well as the influence of summing up expression counts with same/opposite expression patterns in different cells to clustering results.

6.2 Contents

In the following content we are going to describe the data material and methods of data preparation as well as simulation methods in section Methods. In section Results, we are going to show the simulation results for checking relationship between clustering accuracy and low expression counts as well as the influence of summing up expression counts with same/opposite expression patterns in different cells to clustering results. Finally, we are going to show the comparison results of clustering accuracy with k-mer counts and gene counts. In the discussion part we are going to summary all the work has been done and drawbacks in our work.

Methods

1. General pipeline

The whole simulation work mainly has three parts shown as Figure 1.

The first part is material preparation including UMI scRNA-seq download from online database and software packages download from Bioconductor sources.

The second part is simulation data preparation, which includes dealing with raw UMI scRNA-seq data, generating gene count matrix, estimating sequences distribution on each gene, generating k-mer count matrix as well as genes sampling.

The third part is simulation for three purposes. First one is to show the influence of low expression level to clustering accuracy. The second one is to show the influence of summing up expression counts with same/opposite expression patterns in different cells to clustering results. The last one is comparing the performance for clustering between k-mer count matrix and gene count matrix.

2. Materials and software

2.1. Materials

10 Fastq data of UMI scRNA-seq data for mouse embryonic stem cells from dataset GSE46980 on Gene Expression Omnibus:

(SRR1548085, SRR1548086, SRR1548087, SRR1548088, SRR1548089, SRR1548090, SRR1548091, SRR1548092, SRR1548093, SRR1548094)

Mus musculus genome information from UCSC—mm10

2.2. Software and packages

Software: R, Python, bowtie;

Packages: ‘AnnotationHub’, ‘BSgenome’, ‘AnnotationDbi’, ‘combinat’, ‘Mus.musculus’, ‘GenomicRanges’, ‘Biostrings’, ‘GenomicRanges’, ‘BSgenome.Mmusculus.UCSC.mm10’.

3. Simulation data preparation

3.1 Generation of gene expression matrix

3.1.1 Determination of gene number and cell number in simulation

In reality, there may be multiple subtypes of cells in one experiment. However, in order to simplify the simulation, we only do the clustering with two groups of cells, in each there are 100 cells.

Islam et al reported that usually there are approximately 10,000 genes expressed in a single cell and the cDNA molecules capture efficiency is about 48% (s.d. = 5%) [14]. So, in our simulation the gene number is defined as 5000. Thus, the dimension of gene expression matrix is 5000×200.

3.1.2 Gene expression model

Different from the microarray results, the value of UMI scRNA-seq result is gene counts, which is discrete. So, in order to describe the gene expression, we should choose Poisson distribution or Negative Binomial distribution. Islam et al reported that in their UMI scRNA-seq experiments, the distribution of the number of counted molecules for each gene approaches the theoretically optimal Poisson distribution [14]. However, we all know the variance of Poisson distribution is always equal to the mean of Poisson distribution, so that we choose the Negative Binomial distribution to take care of the potential 'over-dispersion' situation.

Parameter Definition:

c: Cell group, in this simulation there are two cell groups c=1 and c=2

i: Gene index, range of 'i' is 1- 5,000

j: Cell index, in each group, range of 'j' is 1-100

X_{ijc}: Gene i expression of cell j in cell group c

μ_{ic}: The mean expression of gene i for cells in group c

ϕ_i: Biological variance of gene 'i' expression for cell group 1 and cell group 2

λ_c: Distribution parameter of exponential distribution. $1/\lambda_c$ is the mean value of mean gene expression in cell group c

fc_i: The fold change of mean expression of gene 'i' between cell group 1 and cell group 2, $fc_i = \frac{\mu_{i1}}{\mu i2}$

sd: The standard deviation of logarithmic fold change for each gene expression in different cells

m: The log-mean of log normal distribution for ϕ_i, it is set as -1 for simulation

τ^2: The log-variance of log normal distribution for ϕ_i, it is set 1 for simulation

The relationship between the parameters above:

$$\begin{cases} X_{ijc} \sim NB(\mu_{ic}, \phi_i) \\ \mu_{ic} \sim exp(\lambda_c) \\ \log(fc_i) \sim N(0, sd) \\ \phi_i \sim \log - normal(m, \tau^2) \end{cases}$$

If the fold change fc_i is fixed then by adjusting the value of λ_c, we can control the gene expression level of both cell groups. Greater λ_c value represents lower gene expression level. By changing the value of sd, we can control the difference of gene 'i' mean expression between two cell groups. Higher value of sd means greater difference (higher signal noise ratio), thus easier to cluster.

For our study, we set λ_c = 0.1, 1, 2 and sd = 0.05, 0.10, 0.15. Then we totally have 3×3=9 scenarios by combing the two sets of parameter values. Under each scenario we

generate 6 gene expression matrixes to mitigate the random sampling effect. So, in our simulation there are totally 54 matrixes with 9 different parameters combinations.

3.2 Estimation of sequences distribution on each gene

As stated in introduction, although the sequences are randomly fragmented, sequences derived from 3' or 5' end of mRNA are easier been sequenced than sequences from other positions. For UMI scRNA-seq, sequences from one gene may come from different TSSs in a single cell due to the alternative splicing. Even though sequences are from the same TSS, because of the TSS fusion they may have sequences with 1-20nt difference [14]. Considering the situation that we can't know the exact sequences distribution on different genes in scRNA-seq, we assumed that the counts of sequences from different positions of a certain gene is a Multinomial distribution. So, we counted the frequencies of sequences from different positions in each gene. Then used the percentage as the estimation of the probability the gene sequenced at corresponding position.

That is, assuming there are M cells ($Cell_1, Cell_2, \dots Cell_m$) and for gene i sequences are sequenced from totally n different positions ($P_1, P_2, \dots P_n$). In Cell k the counts are ($C_{k1}, C_{k2}, \dots C_{kn}$). Then the probability sequence derived from P_i is:

$$\text{Prob}\ (P_i) = \frac{\sum_1^m C_{ki}}{\sum_1^m C_{k1} + \sum_1^m C_{k2} + \dots + \sum_1^m C_{kn}}$$

In this simulation we calculated our estimation based on 10 Fastq data from dataset GSE46980 on Gene Expression Omnibus. This is UMI scRNA-seq data for mouse embryonic stem cells. Finally, we got the sequences distribution for 14115 genes and among them 11540 genes have multiple positions. The mean number of positions for a gene is about 10 and the maximum number of positions for a gene is 605.

3.3 Selection of k-mer length

As described in introduction, if we are aiming at increasing the expression count difference then we should choose the length of k-mer carefully – neither too large nor too small. Here we give a simple definition of coverage to help us to select the appropriate length of k-mer. Assuming in a sample there are totally T different transcripts and the length of transcripts is L, the length of k-mer is K and 'unique ()' is a function calculating number of different values of given variable just as same expression function in R.

$$\boldsymbol{Coverage} = \frac{Number\ of\ \ k-mers\ generated\ from\ all\ different\ \ transcripts}{Number\ of\ \ different\ k-mers\ generated\ from\ all\ different\ transcripts}$$

$$= \frac{T \times (L-K+1)}{unique(T \times (L-K+1))}$$

We can see when K is very small, and then coverage will be super large. For example, K=1, L=50 and T=10,000, then we have $unique\ (T \times (L-K+1))$=4. The coverage for 1-mer is 125,000. However, with the same value of L and T, assuming K=L=50. We can know that due to the uniqueness of long length k-mer that $unique(T \times (L-K+1)) = T \times (L-K+1)$. So, coverage is equal to 1.

In conclusion if coverage value is close to 1 then it means no contribution to increase the expression count values. If the coverage value is too large, it means the k-mer count information is too messy to use. By calculating the coverage of the 14115 genes in our genes pool for both transcript length equal to 50 and 1000, we have the Table 1 shown in appendix. Finally, we choose K=5, 6, 7, 8, 9, 10 for both transcript length equal to 50 and 100 to do the simulation.

3.4 Generation of k-mer expression matrix

For each gene expression matrix, we need to generate a corresponding k-mer counts matrix. The first step is sample 5,000 genes from the 14115 genes pool generated above. Then based on the specific gene count and corresponding sequence distribution of the gene, we random sampled the sequences with length L out and cut them into small pieces – k-mers. At last get the count of each type of k-mer in each cell then combine them into a matrix called k-mer counts matrix.

However, different genes have different sequences and multiple positions to generate the k-mers, which may affect the cluster results. In order to reduce the influence of the process of sampling 5,000 genes from 14115 genes and sampling different positions for generating sequences in a certain gene, we got five samples of genes then applied them for each gene expression matrix.

3.5 Clustering results comparison

After getting the gene count matrix and corresponding k-mer count matrix, we first normalized the expression counts then applied Principle Component Analysis (PCA) to reduce the dimensions. Finally, clustering was done with the selected first K principle components by K-means method. Here the K value is determined based on the method reported in TSCAN [20].

Since we have known the classes of the 200 cells, the accuracy is defined by comparing the clustering result with the original class. Suppose 100 cells are in C1 group and 100 cells are in C2 groups. The clustering result is A C1 group cells and B group C2 cells are clustered into one group. Then the accuracy here is:

$$Accuracy = \frac{max(A + 100 - B,\ B + 100 - A)}{200}$$

Because K-means uses randomly generated seed to determine the staring centroids of the cluster, so for each matrix we do 10,000 times K-means clustering to select the highest accuracy as the result.

4. Simulation methods description

4.1 Influence of low expression count to clustering results

As described in section of gene expression model, we have gene count $X_{ijc} \sim NB(\mu_{ic}, \phi_i), \mu_{ic} \sim exp\ (\lambda_c)$ as well as $fc_i = \frac{\mu_{i1}}{\mu i2}$. By controlling the value of λ_c and fixing the fold-change fc_i, we can change the mean expression level of genes in cells. Larger λ_c means lower expression level of genes in cells. By changing the value of λ_c and keeping other parameters consistent, we checked the clustering accuracy and the average gene count in a single cell with different λ_c values. The values of λ_c are $0.05, 0.10, 0.15, \ldots$ 4,95, 5.00. There are totally 100 different λ_c values and under each value we repeated 50 times to generate gene count matrix and do clustering. Finally, we reported the mean value and median value of clustering results with different λ_c values as well as the corresponding averaged gene count in a single cell with specific λ_c value.

S0: set λ_c=0

S1: Set $\lambda_c = \lambda_c + 0.05$ and n=1, if $\lambda_c > 5.00$ then S6, else then S2

S2: Generate a gene count matrix with specific parameter λ_c and set n=n+1

S3: Record the averaged gene count per cell

S4: Record the clustering result

S5: If n> 50 then S1 else then S2

S6: End the loop and report the averaged clustering results and averaged gene count per cell for each λ_c value

4.2 Influence of summing up expression counts with same/opposite expression patterns in different cells to clustering results

In introduction part, we discussed the potential advantage and disadvantage of summing up expression counts. However, because we did not provide detailed mathematics proof, we tried to use simulation to show the potential regularity. In our simulation we have gene count $X_{ijc} \sim NB\ (\mu_{ic}, \phi_i)$ and μ_{ic} representing the mean expression level of gene 'i' in group c. After generating a gene count matrix with certain parameters, we divided 5000 genes into different two parts – up regulated genes and down regulated genes, by comparing the μ_{i1} and μ_{i2}. If $\mu_{i1} > \mu_{i2}$ then gene 'i' is up regulated gene (in cell 1), otherwise we thought it is down regulated gene (in cell 1). We calculated the difference between μ_{i1} and μ_{i2}. And resort the gene count matrix rows based on the value of $\mu_{i1} - \mu_{i2}$. After resorting the matrix, we only kept the first 2000 rows (having largest values of $\mu_{i1} - \mu_{i2}$) and the last 2000 rows (having largest values of $\mu_{i2} - \mu_{i1}$) of the matrix. We called the matrix 'Original'. Then we got matrix called 'Same', which is generated by summing up the counts of genes that are sharing the same expression pattern and near to each other. For example, the counts of Row 1 in 'Same' matrix are generated by adding counts of Row 1 and Row2 in matrix 'Original'. The matrix having counts by adding up counts of genes having opposite expression pattern (up regulated gene count + down regulated gene count) is called 'Opposite'. For example, the counts in Row 1 in 'Opposite'

are sum of counts in Row 1 and Row 4000 of 'Original'.

The 'Same' matrix counts are sum of 'Original' matrix counts of genes, which are all up regulated or down regulated in same cell. The 'Opposite' matrix counts are sum of 'Original' matrix counts of genes, which having opposite expression pattern in same cells. The dimensions of them are 2000×200, while the dimension of 'Original' is 4000×200.

After getting the three types of matrix, we compared clustering results of them under scenarios with different parameters.

As discussed above the clustering results are also influenced by the variance of expression counts. It is known that gene count has negative binomial distribution that $X_{ijc} \sim NB\ (\mu_{ic}, \phi_i)$. So we changed the ϕ_i, which represents the biological variance in gene expression, to see the clustering accuracy difference between three types of matrixes.

The first simulation is simply setting ϕ_i as constant for all genes. We did the simulation with four ϕ_i values that $\phi_i = 0.1,\ 0.5,\ 1$ and 2. The second simulation is assuming $\phi_i \sim log - normal\ (m, \tau^2)$, setting ϕ_i into four log-normal distributions with different parameters. One is $\phi_i \sim log - normal\ (1, 0.3)$, of which the median of ϕ_i is 2.73. One is $\phi_i \sim log - normal\ (0, 0.3)$, of which the median of ϕ_i values is 1.0. One is $\phi_i \sim log - normal\ (-1, 0.3)$ and its median ϕ_i value is 0.37. The other one is $\phi_i \sim log - normal\ (-2, 0.3)$, which has smallest median ϕ_i value 0.14. Actually, they represent different levels of variance from large to small.

S1: Set ϕ_i value

S2: Generate gene count matrix

S3: Divide genes (rows) into two groups by comparing $\mu_{i1}\ and\ \mu_{i2}$, if $\mu_{i1} > \mu_{i2}$ gene 'i' is up regulated and otherwise it is down regulated

S4: Resort the gene count matrix by rows based on the $\mu_{i1} - \mu_{i2}$ value and only keep the first 2000 rows and last 2000 rows of the resorted matrix. Then call the matrix 'Original'

S5: Get 'Same' matrix by adding up counts of two rows near each other. That is Row N in 'Same' is sum of Row 2N-1and Row 2N of 'Original' matrix

S6: Get 'Opposite' matrix by adding up counts of two rows with different expression pattern in same cell type. That is Row N in 'Opposite' is sum of Row N and Row 4001-N of 'Original' matrix

S7: Get the clustering results of the three matrixes

S8: Repeat S2 to S7 50 times and then go back to S1 set ϕ_i with different parameters

4.3 Clustering results comparison with different parameters

From the gene expression model, we know that λ_c can control the general gene expression level and sd can control the gene expression difference between two cell groups (greater λ_c means lower gene expression level and greater sd means larger gene expression difference between groups). In order to test the performance of k-mer counts matrix in different situations, we designed three different levels for gene expression ($\lambda_c = 0.1, 1, 2$) and three different levels for between groups difference ($sd = 0.05, 0.1, 0.15$) by change the λ_c and sd.

In order to mitigate the random sampling influence on final results, we repeated the above three main steps 6 times (6 matrixes) for each combination of gene expression level and between-groups difference level and also for each matrix we repeated sampling genes 5 times to reduce the influence of different genes content.

S1: Generate the gene expression matrix (5000 genes, 2 groups with 100 cells in each group)

by adjusting the parameters of assumption model (Negative Binomial Distribution)

S2: Generate k-mer counts matrix (k=5, 6, 7, 8, 9, 10) by combing the gene counts information from first step and sequence distribution on genes information

S3: Compare the clustering results on the two matrixes generated above

Results

1. Influence of low expression count to clustering results

In Figure 2 and Table 2 we can see that when λ_c = 0.05, the average gene count is around 19.98 and corresponding clustering accuracy is about 0.90. However, with the increase of λ_c both the clustering accuracy and average gene count per cell decrease. When the λ_c = 5.00, the average gene count decreased close to 0.19 and the clustering accuracy is also decreased to about 0.585. So, when gene expression is low, it's harder to do clustering with such expression matrix, which indicates that we need to raise novel methods to solve this dilemma.

2. Influence of summing up expression counts with same/opposite expression patterns in different cells to clustering results

In Figure 3 and Table 3, we can see that the 'Opposite' matrix always performs worse than 'Same' matrix and 'Original' matrix (exact accuracy information is in table 3), which proves the disadvantage discussed above that by combing expression counts of genes with different expression patterns may mitigate the difference of original transcripts between different cell groups. When constant variance equals to 0.1 and 0.5, we can see that the

‘Same’ matrix performs better than ‘Original’ matrix (median: 0.985 vs. 0.98, 0.728 vs. 0.71), which indicating the potential advantage discussed in introduction. Adding up expression counts of genes sharing same expression pattern may improve the clustering accuracy by enlarging the difference of original transcripts counts in different cell groups. However, when constant variance increases to 1 and 2, the clustering accuracy of ‘Same’ matrix is worse than ‘Original’ matrix (median: 0.63 vs. 0.64, 0.6 vs. 0.605). This indicates that the process of summing up two gene counts may also increase the variance, which may have influence on clustering results.

In Figure 4 and Table 4, we set the biological variance ϕ_i log normal distributed with four sets of parameters. The log standard deviation is consistent in four plots with log-sd equals to 0.3. However, the log-means are different with -2, -1,0 and 1. The corresponding variance medians are 0.14, 0.37, 1.00 and 2.73. The four plots represent different variance levels separately. We can see that when variance is small (log-mean = -2), ‘Same’ matrix has the better performance in clustering accuracy than ‘Original’ matrix (median: 0.97 vs. 0.96) while ‘Opposite’ matrix shows the worst accuracy. And with the increase of variance the performance of ‘Same’ matrix and ‘Original’ matrix becomes worse. Furthermore, the ‘Same’ matrix clustering accuracy is close to or a little lower than ‘Original’ matrix (log-mean = 0, median: 0.622 vs. 0.64; log-mean=1, median: 0.597 vs. 0.595). Again, the result shows that when the variance is small, adding expression counts of genes with the same expression pattern in cells can improve the clustering results. However, when the variance is large, the ‘potential’ advantage disappears.

In conclusion, disadvantage of adding expression counts of genes with different expression patterns is obvious while advantage of adding expression counts of genes

sharing similar expression pattern can only appear when variance is small. So, if we want to use k-mer count to improve clustering accuracy we need to make sure the adding up should appear between transcripts with same expression pattern and also the variance cannot be too large.

3. Clustering results comparison with different parameters

As described in the methods part, parameter λ_c controls the gene expression level. Higher λ_c means that the value in the matrix is larger. For λ_c equals to 0.1, there are about 50,000 gene counts in a cell. For λ_c equals to 1, there are about 5,000 gene counts in a cell. And for λ_c equals to 2, there are only about 2,500 gene counts expressed in a cell. Parameter sd controls the difference of gene expression between two cells. With larger sd, the gene expression difference between two cells are more significant.

From Figure 5 and Table 5, the change tendency of clustering results due to different parameter combinations is consistent with what we described above. With same sd value, the clustering accuracy decreases when λ_c increases. With same λ_c value, clustering accuracy improves with larger sd values.

We can see that in all scenarios, longer length of k-mer (k= 8, 9, 10) has better accuracy than shorter length of k-mer (k=5, 6, 7). But in some scenarios, 10-mer clustering performance is not as good as 8-mer or 9-mer. No matter in which scenario, gene count matrix performs better than k-mer count matrix. When length of k-mer is 5, 6, 7 and 8, shorter sequence length (50nt) is better than longer sequence length (100nt). The reason may be that when length of k-mer is shorter, more transcripts may share the specific k-mer, then the k-mer count mixed counts of more transcripts. So longer length sequence can

generate more k-mers, which may make the k-mer count information messier.

In conclusion, although with the length increase of k-mers, clustering results are improved, the performance of k-mer still cannot be compared with performance of gene counts in all scenarios.

Discussion

Through these work, we showed that it is true low expression values for UMI scRNA-seq is hard to get a good clustering result. We need to raise some method to solve it. After showing the k-mer/gene count matrix is transformation of transcript count matrix, we showed that if we want to improve the clustering results by increasing matrix count values with k-mer, then we should make sure most of the k-mer count is summing of transcript counts sharing same expression pattern and the variance should be small. This is really hard for k-mer count because its transformation matrix is determined by the genome content when k is fixed. Finally, we found that under different scenarios, gene count matrix always performs better than k-mer count matrix.

In our work there are still many aspects can be improved. First, although we estimated the sequences distribution on each gene through real data, the sample size is 10, which is too small to get an accurate estimation. Second, we only studied the clustering performance with k=5, 6, 7, 8, 9, 10. Actually, in order to get an integrative understanding of the k-mer count performance in clustering, we should continue to study other k values. Third, the transformation matrix for k-mer count is determined by genome contents, we should also need to study the k-mer performance for other species due to genomes of different species are different. Fourth, in the second results part, the difference between

‘Same’ and ‘Original’ matrix is very small, in order to make sure the difference truly exists, we should increase the cell numbers for clustering (sample size). Also, we need to repeat the experiments more times to get an accurate estimation and perform suitable statistics test to test whether the difference is significant.

References

1. Nature Education. *Definition of genome in Scitable*. 2014; Available from: https://www.nature.com/scitable/definition/genome-43.
2. Brown TA., *Transcriptomes and Proteomes*, in *Genomes*. 2002, Oxford: Wiley-Liss.
3. Adams, J., *Transcriptome: connecting the genome to gene function*. Nat Educ, 2008. **1**(1): p. 195.
4. TAUB, F., E, J.M. DeLEO, and E.B. THOMPSON, *Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs*. Dna, 1983. **2**(4): p. 309-327.
5. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus*. Nature reviews genetics, 2006. **7**(1): p. 55.
6. Hua, H.B., *High-Throughput Technologies for Gene Expression Analyses: What We Have Learned for Noise-Induced Cochlear Degeneration*. Journal of otology, 2013. **8**(1): p. 25-31.
7. Schulze, A. and J. Downward, *Navigating gene expression using microarrays—a technology review*. Nature cell biology, 2001. **3**(8): p. E190.
8. Tachibana, C., *Transcriptomics today: Microarrays, RNA-seq, and more*. Science, 2015. **349**(6247): p. 544-546.
9. Wang, Z., M. Gerstein, and M. Snyder, *RNA-seq: a revolutionary tool for transcriptomics*. Nature reviews genetics, 2009. **10**(1): p. 57.
10. Turner, W., *The cell theory, past and present*. Journal of anatomy and physiology, 1890. **24**(Pt 2): p. 253.
11. Trapnell, C., *Defining cell types and states with single-cell genomics*. Genome research, 2015. **25**(10): p. 1491-1498.
12. Wang, Y. and N.E. Navin, *Advances and applications of single-cell sequencing technologies*. Molecular cell, 2015. **58**(4): p. 598-609.
13. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics*. Nature Reviews Genetics, 2015. **16**(3): p. 133.
14. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers*. Nature methods, 2014. **11**(2): p. 163.
15. Macosko, E.Z., et al., *Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets*. Cell, 2015. **161**(5): p. 1202-1214.
16. Liu, Z., Y. Tao, and L. Zhu, *Don't Fall for Dropouts: Bayesian Learning on Single-cell RNA-seq Data*.
17. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells*. Nature biotechnology, 2014. **32**(4): p. 381.
18. Shin, J., et al., *Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis*. Cell stem cell, 2015. **17**(3): p. 360-372.
19. Bendall, S.C., et al., *Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development*. Cell, 2014. **157**(3): p. 714-725.
20. Ji, Z. and H. Ji, *TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis*. Nucleic acids research, 2016. **44**(13): p. e117-e117.

21. Ntranos, V., et al., *Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts*. Genome biology, 2016. **17**(1): p. 112.
22. Sun, Z., et al., *DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data*. Bioinformatics, 2017. **34**(1): p. 139-146.
23. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing*. Genome research, 2010. **20**(2): p. 265-272.
24. Ounit, R., et al., *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers*. BMC genomics, 2015. **16**(1): p. 236.
25. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification*. Nature biotechnology, 2016. **34**(5): p. 525.

Appendix

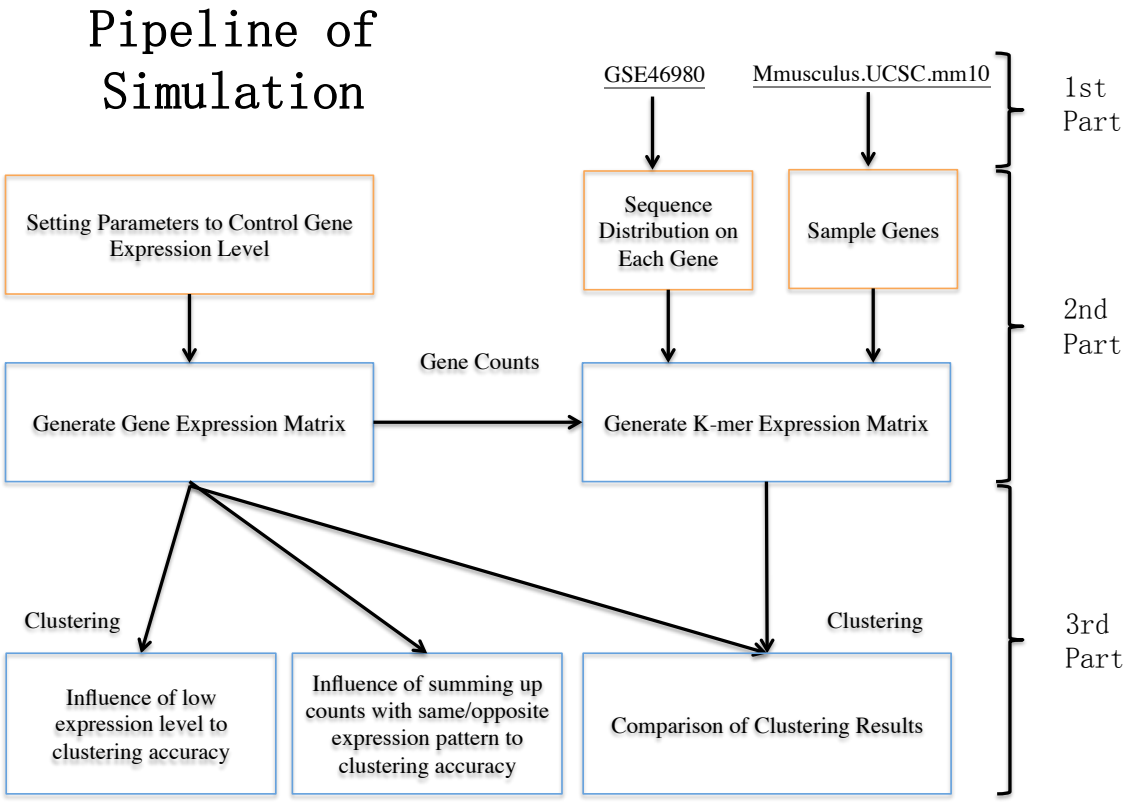

Figure 1. Pipeline of Simulation

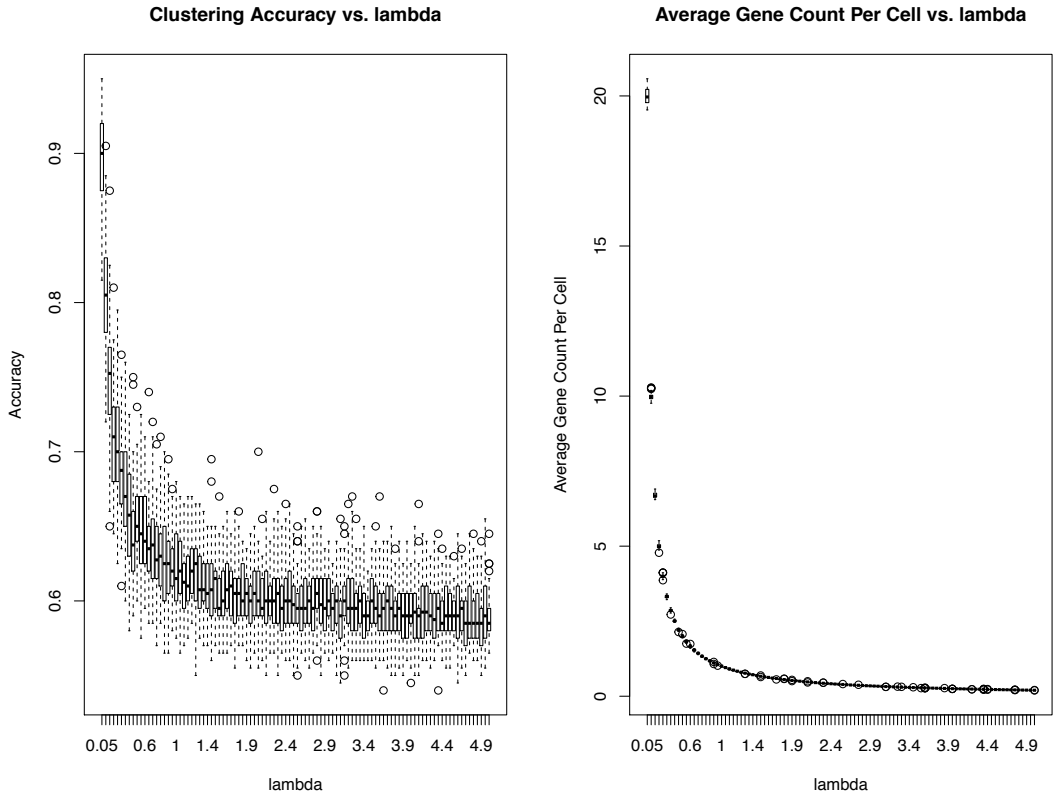

Figure 2. Plot of clustering accuracy and average gene count per cell with different lambda value

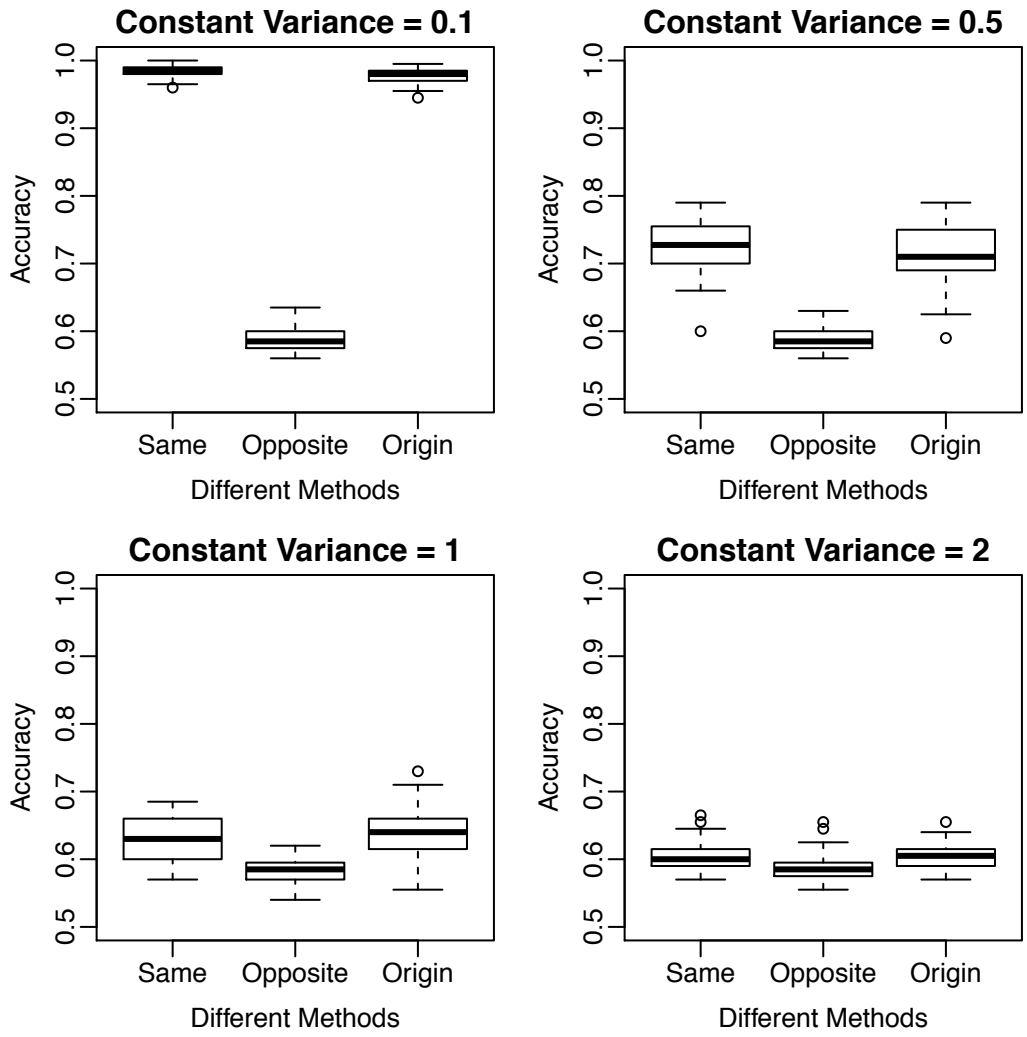

Figure 3. Comparison of clustering results between 'Same', 'Opposite' and 'Original' matrix with different variance (variance is constant)

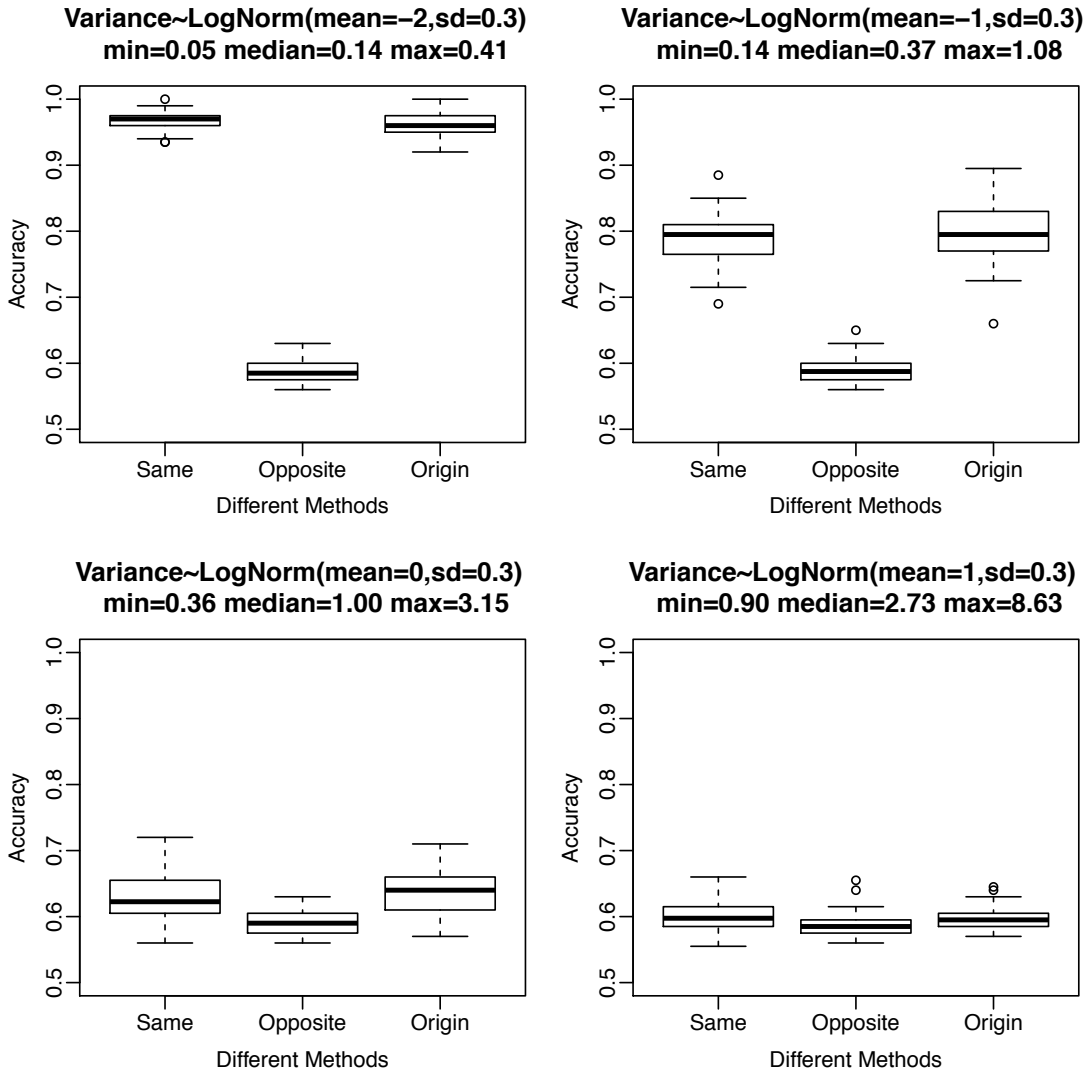

Figure 4. Comparison of clustering results between 'Same', 'Opposite' and 'Original' matrix with different variance (variance is log-normal distributed)

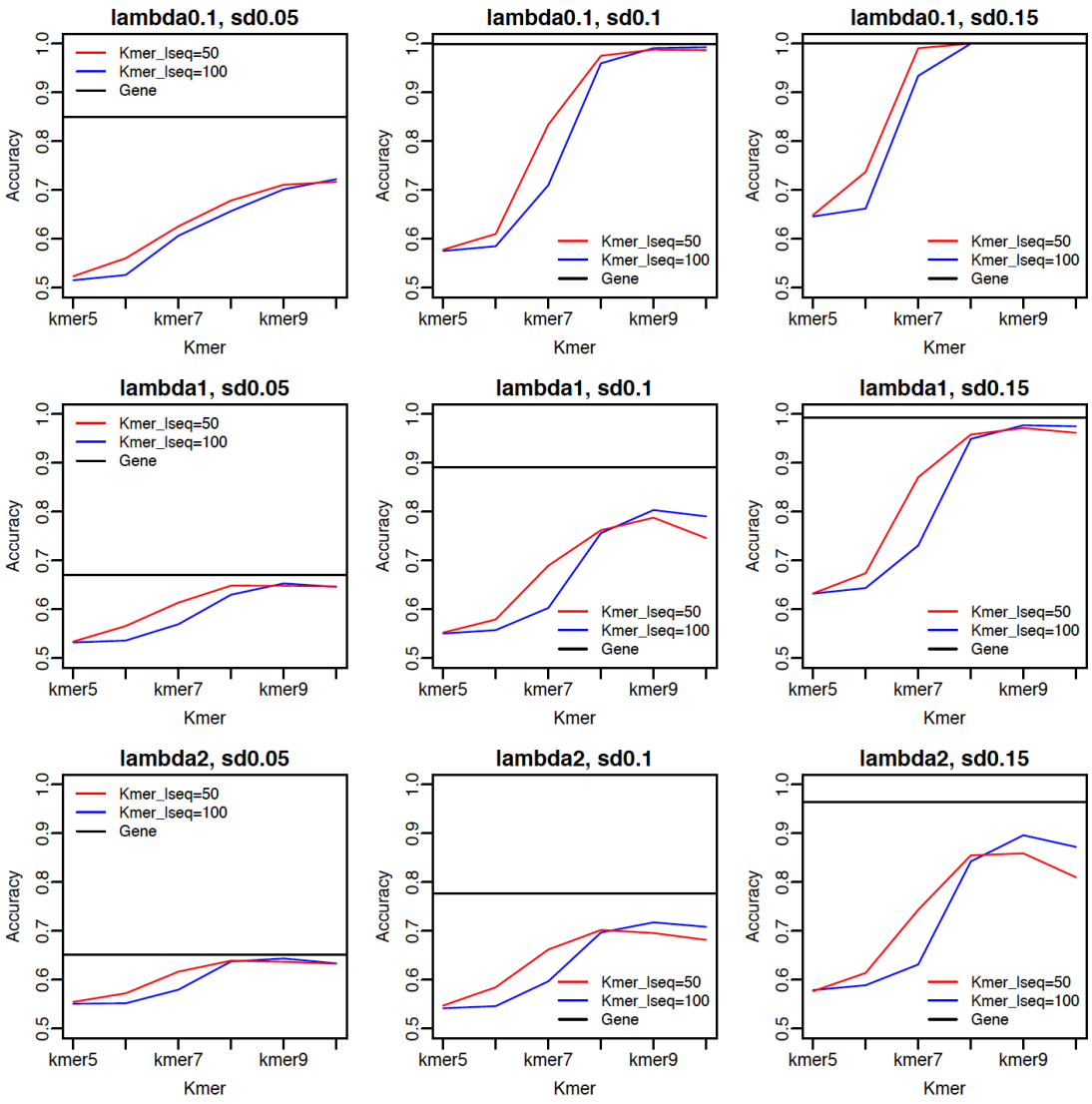

Figure 5. Comparison of clustering results (averaged) of k-mer count matrix and gene count matrix with different expression level and expression difference between cells

Table 1. Coverage for different length of k-mer and different length of transcripts

Lseq=50		Lseq=100	
K	Coverage	K	Coverage
5	6488.2	5	13540.59
6	1586.79	6	3349.89
7	387.88	7	828.66
8	94.85	8	204.98
9	23.92	9	51.15
10	7.19	10	14.12
11	3.03	11	4.99
12	1.97	12	2.61
13	1.67	13	1.96
14	1.57	14	1.76
15	1.53	15	1.69
20	1.46	16	1.66
25	1.4	17	1.64
30	1.35	18	1.62
35	1.3	19	1.61
40	1.24	20	1.6
45	1.17	30	1.51
50	1.05	40	1.43
		50	1.38
		60	1.33
		70	1.28
		80	1.22
		90	1.15
		100	1.02

Table 2. Summary of clustering accuracy and average gene count with different expression level

lambda value	accuracy median	accuracy mean	accuracy sd	gene count median	gene count mean	gene count sd
0.05	0.9	0.8971	0.030490715	19.9653865	19.98339652	0.270321694
0.1	0.805	0.8035	0.043533356	9.981506	9.98549806	0.12171501
0.15	0.7525	0.7511	0.044587566	6.677881	6.6907731	0.087727086
0.2	0.71	0.7101	0.03674915	4.983323	4.99604326	0.073537711
0.25	0.7	0.7021	0.036268415	3.998325	3.9973941	0.059939084
0.3	0.6875	0.6839	0.030359916	3.327595	3.32130218	0.045581769
0.35	0.67	0.6743	0.035123833	2.8580445	2.85609458	0.041364652
0.4	0.6575	0.6559	0.033725936	2.505014	2.50826502	0.027717485
0.45	0.6375	0.6447	0.034588101	2.2116945	2.21172298	0.024359214
0.5	0.65	0.6517	0.025686692	1.998522	1.99996344	0.030332105
0.55	0.645	0.6485	0.035674978	1.8197525	1.81837088	0.029343133
0.6	0.64	0.6445	0.027109942	1.663597	1.66506928	0.026964379
0.65	0.635	0.637	0.026878411	1.5362875	1.53639002	0.026464577
0.7	0.6375	0.6375	0.029141632	1.429675	1.43246998	0.019683082
0.75	0.6275	0.6309	0.026490083	1.329854	1.3308301	0.019239516
0.8	0.63	0.6315	0.027147556	1.2498975	1.24993366	0.014214455
0.85	0.625	0.6303	0.03046259	1.179402	1.17669872	0.015631277
0.9	0.625	0.6252	0.028819636	1.112522	1.11265594	0.014118954
0.95	0.62	0.624	0.023079278	1.052729	1.05366952	0.012780239
1	0.615	0.6206	0.028867042	1.0036075	1.00168664	0.013417995
1.05	0.62	0.6209	0.023919295	0.953709	0.9543666	0.013557758
1.1	0.6125	0.6138	0.022667967	0.9110545	0.91106968	0.01335906
1.15	0.61	0.6146	0.021424762	0.8696775	0.87045146	0.011133928
1.2	0.62	0.6207	0.019430935	0.835152	0.83392318	0.011405145
1.25	0.625	0.6204	0.023858426	0.798608	0.79712528	0.012573076
1.3	0.6075	0.6138	0.024001701	0.769141	0.76968222	0.009087563
1.35	0.6075	0.6117	0.020913085	0.7389905	0.73850434	0.011735581
1.4	0.605	0.609	0.020898198	0.7158615	0.71646248	0.010526849
1.45	0.6075	0.609	0.027404752	0.689807	0.68878058	0.009686217
1.5	0.615	0.608	0.018626293	0.666145	0.66650798	0.00894512
1.55	0.595	0.6033	0.021656172	0.645047	0.6458635	0.009086695
1.6	0.6	0.6047	0.021485876	0.6269095	0.62622406	0.009668941
1.65	0.6075	0.6105	0.021905409	0.6047005	0.60603034	0.008401762
1.7	0.61	0.6075	0.019039433	0.5924175	0.59228974	0.010515641
1.75	0.605	0.603	0.022131333	0.5721125	0.57161606	0.007704895
1.8	0.605	0.6037	0.021590105	0.555842	0.55587478	0.008564423
1.85	0.6	0.6051	0.020012496	0.540695	0.5413057	0.0079613
1.9	0.605	0.6039	0.016972066	0.52546	0.52512248	0.0074126

1.95	0.6	0.5984	0.020786573	0.515187	0.5154657	0.00691855
2	0.605	0.6047	0.018472539	0.5004085	0.50109672	0.007108147
2.05	0.6	0.6048	0.024368766	0.487067	0.48676872	0.007016729
2.1	0.595	0.5984	0.019728773	0.476845	0.47711378	0.00701251
2.15	0.6	0.6028	0.024746882	0.4662135	0.4664127	0.006273036
2.2	0.6	0.5998	0.018843068	0.4537495	0.4531879	0.006268748
2.25	0.6	0.6041	0.025186853	0.4434775	0.44376212	0.006074309
2.3	0.605	0.6004	0.022942697	0.4326245	0.43327716	0.006417873
2.35	0.595	0.5979	0.020431318	0.424858	0.42454802	0.006047635
2.4	0.6	0.6029	0.023301266	0.4162585	0.41730122	0.005920011
2.45	0.6	0.6013	0.021184563	0.406215	0.40718324	0.006583929
2.5	0.5975	0.5991	0.019077982	0.398787	0.39892342	0.006062358
2.55	0.595	0.5959	0.020245685	0.3916415	0.39176888	0.005302709
2.6	0.595	0.5977	0.020707142	0.3845765	0.38464888	0.00503346
2.65	0.595	0.5989	0.020807426	0.3752055	0.37678808	0.005369188
2.7	0.6	0.5975	0.017821393	0.3692825	0.36954762	0.004297734
2.75	0.595	0.5977	0.019954284	0.3647425	0.36473386	0.005891487
2.8	0.605	0.6048	0.020848335	0.3573235	0.35704192	0.005842278
2.85	0.5975	0.5987	0.021184563	0.350276	0.3511381	0.005316771
2.9	0.595	0.5976	0.020183849	0.345305	0.34551778	0.00386511
2.95	0.6	0.6008	0.019122591	0.339179	0.33972932	0.005489921
3	0.595	0.593	0.015452363	0.3336125	0.33373314	0.005228715
3.05	0.6	0.5973	0.020005356	0.3284375	0.32823308	0.003647645
3.1	0.59	0.5903	0.019908208	0.3226035	0.322838	0.00445121
3.15	0.6	0.6009	0.019237771	0.3190595	0.31864994	0.004664771
3.2	0.595	0.5992	0.023611654	0.312256	0.3130834	0.004491101
3.25	0.595	0.6002	0.025494297	0.307312	0.30739284	0.004002414
3.3	0.595	0.5954	0.019028443	0.3027185	0.3029393	0.003947413
3.35	0.6	0.5974	0.019903851	0.299303	0.29965018	0.004767728
3.4	0.59	0.5923	0.019645065	0.2926285	0.29343596	0.003319429
3.45	0.59	0.5912	0.014623157	0.290073	0.29025086	0.003571345
3.5	0.6	0.6001	0.0176268	0.2879275	0.28710464	0.004941792
3.55	0.595	0.5969	0.020351252	0.2822565	0.28183536	0.003836211
3.6	0.59	0.5924	0.020209111	0.278297	0.27777158	0.004042873
3.65	0.595	0.5924	0.019980603	0.2746145	0.27423622	0.003930398
3.7	0.6	0.598	0.021641796	0.270459	0.27070264	0.003142847
3.75	0.595	0.5959	0.020743575	0.2666775	0.26737832	0.004741994
3.8	0.59	0.5896	0.018893904	0.2642285	0.26385438	0.004301197
3.85	0.595	0.5924	0.019065944	0.260214	0.26033538	0.003797247
3.9	0.59	0.5897	0.019702118	0.2561725	0.256064	0.002928717
3.95	0.59	0.5913	0.01778012	0.2531435	0.25284034	0.00316157
4	0.59	0.5918	0.018537331	0.250398	0.25037772	0.003212545
4.05	0.5925	0.5933	0.022397385	0.2471905	0.24783858	0.004030631

4.1	0.59	0.5891	0.02027087	0.244432	0.24450282	0.004283382
4.15	0.5925	0.5947	0.020464255	0.2404315	0.241239	0.003470231
4.2	0.5925	0.5942	0.019596959	0.2394415	0.23908678	0.00286894
4.25	0.59	0.5907	0.017143155	0.2352615	0.23495878	0.003718157
4.3	0.5875	0.5913	0.019081191	0.2320275	0.23241528	0.003859979
4.35	0.595	0.5939	0.019359383	0.230787	0.23032458	0.004019207
4.4	0.585	0.5899	0.016490257	0.227816	0.22828844	0.003911673
4.45	0.59	0.5925	0.017963966	0.2240585	0.22415898	0.003351642
4.5	0.59	0.5919	0.019056575	0.222823	0.22259248	0.004063777
4.55	0.59	0.5932	0.013951812	0.2197795	0.21994322	0.003144291
4.6	0.59	0.5922	0.021716165	0.2173895	0.21750274	0.0034098
4.65	0.595	0.5919	0.017551382	0.215727	0.2155433	0.002789587
4.7	0.585	0.5875	0.019749709	0.2132655	0.212669	0.003424901
4.75	0.585	0.5917	0.018833589	0.210976	0.21076776	0.002626351
4.8	0.585	0.5892	0.018608754	0.2084515	0.20868458	0.00313844
4.85	0.585	0.588	0.018789923	0.205913	0.20598642	0.003137052
4.9	0.585	0.5859	0.019657527	0.2037335	0.20375214	0.002984158
4.95	0.59	0.5925	0.024416664	0.2021095	0.20199728	0.002907537
5	0.585	0.5885	0.017120521	0.199271	0.1997669	0.002577266

Table 3. Clustering results comparison between 'Same', 'Opposite', 'Origin' matrix with different variance values (Constant variance)

Variance	Type	Accuracy median	Accuracy mean	Accuracy sd
0.1	Same	0.985	0.984	0.009
	Opposite	0.585	0.588	0.018
	Origin	0.98	0.978	0.011
0.5	Same	0.728	0.725	0.039
	Opposite	0.585	0.587	0.017
	Origin	0.71	0.715	0.046
1	Same	0.63	0.628	0.032
	Opposite	0.585	0.584	0.02
	Origin	0.64	0.638	0.035
2	Same	0.6	0.604	0.022
	Opposite	0.585	0.587	0.021
	Origin	0.605	0.604	0.021

Table 4. Clustering results comparison between 'Same', 'Opposite', 'Origin' matrix with different variance values (variance is log-normal distributed)

Log-mean	Type	Accuracy median	Accuracy mean	Accuracy sd
-2	Same	0.97	0.968	0.014
	Opposite	0.585	0.588	0.017
	Origin	0.96	0.96	0.016
-1	Same	0.795	0.789	0.038
	Opposite	0.587	0.589	0.018
	Origin	0.795	0.798	0.044
0	Same	0.622	0.632	0.037
	Opposite	0.59	0.591	0.017
	Origin	0.64	0.636	0.034
1	Same	0.597	0.600	0.023
	Opposite	0.85	0.586	0.019
	Origin	0.595	0.600	0.017

Table 5. Summary of clustering results for k-mer and gene count matrix in different expression level and expression difference scenarios

		Mgene			Sequence length = 100		Sequence length = 50	
lambda	sd	mean	sd	k-mer	mean	sd	mean	sd
0.1	0.05	0.849	0.052	5	0.515	0.013	0.523	0.014
				6	0.526	0.015	0.56	0.025
				7	0.606	0.025	0.626	0.018
				8	0.657	0.021	0.678	0.034
				9	0.702	0.038	0.711	0.036
				10	0.722	0.033	0.717	0.047
0.1	0.1	0.999	0.002	5	0.575	0.027	0.578	0.026
				6	0.585	0.024	0.61	0.028
				7	0.71	0.039	0.834	0.048
				8	0.96	0.015	0.975	0.01
				9	0.991	0.007	0.988	0.009
				10	0.993	0.005	0.987	0.008
0.1	0.15	1	0	5	0.646	0.052	0.649	0.053
				6	0.662	0.056	0.737	0.058
				7	0.934	0.051	0.991	0.011
				8	1	0	1	0
				9	1	0	1	0
				10	1	0	1	0
1	0.05	0.671	0.014	5	0.532	0.022	0.534	0.025
				6	0.536	0.025	0.566	0.025
				7	0.569	0.027	0.614	0.024
				8	0.63	0.024	0.648	0.016
				9	0.653	0.019	0.648	0.019
				10	0.646	0.018	0.647	0.023
1	0.1	0.892	0.031	5	0.55	0.031	0.552	0.031
				6	0.557	0.03	0.579	0.034
				7	0.603	0.034	0.689	0.043
				8	0.756	0.044	0.762	0.033
				9	0.803	0.031	0.788	0.033
				10	0.79	0.033	0.746	0.028
1	0.15	0.993	0.003	5	0.632	0.05	0.632	0.051
				6	0.644	0.057	0.674	0.046
				7	0.731	0.055	0.87	0.041
				8	0.949	0.018	0.958	0.018
				9	0.977	0.011	0.971	0.013
				10	0.975	0.015	0.962	0.015
2	0.05	0.651	0.02	5	0.551	0.029	0.554	0.03

				6	0.552	0.029	0.572	0.019
				7	0.58	0.021	0.616	0.017
				8	0.637	0.018	0.639	0.021
				9	0.644	0.017	0.637	0.016
				10	0.634	0.016	0.633	0.019
2	0.1	0.777	0.027	5	0.542	0.021	0.547	0.023
				6	0.546	0.024	0.584	0.024
				7	0.597	0.032	0.662	0.028
				8	0.697	0.034	0.702	0.031
				9	0.718	0.029	0.696	0.029
				10	0.708	0.026	0.682	0.024
2	0.15	0.964	0.007	5	0.579	0.031	0.576	0.031
				6	0.589	0.029	0.614	0.023
				7	0.631	0.027	0.743	0.03
				8	0.842	0.036	0.855	0.028
				9	0.896	0.023	0.859	0.027
				10	0.872	0.033	0.81	0.035