Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Gene expression

ROSeq: A rank based approach to modeling gene expression in single cells

Manan Lalit [1], Aditya Biswas [2], Abhik Ghosh [3,*] and Debarka Sengupta [4,*]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany
[2] Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India
[3] Interdisciplinary Statistical Research Unit , Indian Statistical Institute, Kolkata 700108, India
[4] Centre for Computational Biology and Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Delhi 110020, India

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Single cell transcriptomics provides a window into cell-to-cell variability in complex tissues. Modeling single cell expression is challenging due to high noise levels and technical bias. In the past years, considerable efforts have been made to devise suitable parametric models for single cell expression data. We use Discrete Generalized Beta Distribution (DGBD) to model read counts corresponding to a gene as a function of rank. Use of DGBD yields better overall fit across genes compared to the widely used mixture model comprising Poisson and Negative Binomial density functions. Further, we use Wald's test to probe into differential expression across cell sub-types. We package our implementation as a standalone software called ROSeq. When applied on real data-sets, ROSeq performed competitively compared to the state of the art methods including MAST, SCDE and ROTS.
Software Availability: The Windows, macOS and Linux - compatible softwares are available for download at `https://malaalam.github.io/ROSeq`
Contact: abhianik@gmail.com, debarka@iiitd.ac.in
Supplementary information: Supplementary data is available at *Bioinformatics* online.

1 Introduction

In the past few years, single cell RNA-Sequencing (scRNA-seq) has dramatically accelerated characterization of molecular heterogeneity in healthy and diseased tissue samples (Editorial, 2014). The declining cost of library preparation and sequencing have fostered the adaptation of single-cell transcriptomics as a routine assay in studies arising from diverse domains including stem cell research, oncology, and developmental biology. The field of single-cell transcriptomics suffers severely from various data quality issues, mainly due to the lack of starting RNA material. High levels of noise and technical bias pose significant challenges to single-cell gene expression modeling, which in turn hinders arrival at statistically apt conclusions about cell-type specific gene expression patterns.

A number of parametric and nonparametric methods have already been proposed for modeling expression data and finding differentially expressed genes (DEGs). SCDE (Kharchenko *et al.*, 2014), MAST (Finak *et al.*, 2015) and ROTS (Elo *et al.*, 2008) are notable among these. SCDE and MAST model gene expression using well-known probability density functions and mixed models involving some of those. ROTS, on the other hand, uses a modified, data-adaptive t-type statistic to measure the difference in expression levels. We conjectured that considering ranks instead of absolute expression estimates would make a model less susceptible to the noise and the technical bias, as commonly observed in single-cell data. To realize the same, we employed Discrete Generalized Beta Distribution (DGBD) (Martinez-Mekler *et al.*, 2009) to model the distribution of expression ranks (illustrated in the method section) instead of the raw count. The consideration of rank-ordering distribution was inspired by the seminal work by Martinez-Mekler and colleagues, where they demonstrated the universal applicability of the same in linking frequency estimates and their ranks. We developed ROSeq: a Wald-type test to determine differential expression from scRNA-seq data.

2 Method

ROSeq can be used to compare expression pattern of a gene across two groups of single cells. For gene expression modeling, ROSeq accepts normalized read count data as input. For each gene, ROSeq first defines its range by identifying the minimum and maximum values by pulling the normalized expression estimates across both the groups under study. The algorithm then splits the entire range into $k \times \sigma$ sized bins, where k is a scalar with a default value of 0.05, and σ is the standard deviation of the pulled expression estimates across the cell-groups. Each of these bins corresponds to a rank. Therefore, for each group, cell frequency for each bin maps to a rank. These frequencies are normalized group-wise by dividing by the total cell count within a concerned group. ROSeq uses DGBD to model the distribution of ranks as follows.

$$y_r = A\frac{(N+1-r)^b}{r^a} \tag{1}$$

Here, y_r denotes the fraction of cells in each cell-group corresponding to the rank r and N denotes the total number of bins. a and b are the shape parameters. Finally, A is the normalizing constant ensuring that the sum of the normalized frequencies equals one (see Supplementary Material). For each gene, for a given group, ROSeq estimates the values of a and b by maximizing the Log-Likelihood corresponding to probability mass function depicted in Equation 1.

Let us assume that the input cell subpopulations G_1 and G_2 consist of m and n cells respectively. For any gene g_i, ROSeq first estimates the shape-parameter values $(a_1 = \widehat{a}_1, b_1 = \widehat{b}_1)$ and $(a_2 = \widehat{a}_2, b_2 = \widehat{b}_2)$ respectively. Under the DGBD model, the desired testing for differential gene expressions is equivalent to the test of the null hypothesis $H_0 : a_1 = a_2, b_1 = b_2$ against the omnibus alternative.

Further, ROSeq uses the (asymptotically) optimum two-sample Wald test based on the MLE of the parameters and its asymptotic variances given by the inverse of the Fisher information matrix. We can estimate the asymptotic variance matrices for the MLEs $(\widehat{a}_1, \widehat{b}_1)$ and $(\widehat{a}_2, \widehat{b}_2)$ as $\widehat{V}_1 = I(\widehat{a}_1, \widehat{b}_1)^{-1}$ and $\widehat{V}_2 = I(\widehat{a}_2, \widehat{b}_2)^{-1}$ respectively. The Wald test statistic T for testing H_0 can be written as follows:

$$T = \left(\frac{mn}{m+n}\right)\begin{bmatrix}\widehat{a}_1 - \widehat{a}_2\\ \widehat{b}_1 - \widehat{b}_2\end{bmatrix}^T \left(w\widehat{V}_1 + (1-w)\,\widehat{V}_2\right)^{-1}\begin{bmatrix}\widehat{a}_1 - \widehat{a}_2\\ \widehat{b}_1 - \widehat{b}_2\end{bmatrix}$$

where $w = \frac{n}{m+n}$. The test statistic T follows a central Chi-square distribution χ^2_2 with two degrees of freedom. A gene is considered differentially expressed (rejection of H_0) at 95% level of significance, if the observed value of the test statistic T exceeds the 95% quantile of the χ^2_2 distribution.

3 Applications on Real Datasets

For comparative benchmarking, we used two scRNA-seq datasets from past studies that also performed bulk RNA-seq on the same samples. DEGs called on bulk RNA-seq data were used as the benchmark for single cell differential expression analyses. ROSeq performed competitively as compared to the existing best practice methods including SCDE, MAST and ROTS.

In the first dataset (obtained from Tung *et al.*, 2017), read count data corresponding to 288 single cells each was available from the three human induced pluripotent stem cell lines (NA19098, NA19101 and NA19239). (For computing the results using the SCDE method, 96 single cells were selected randomly without replacement from each subpopulation - this was done in order to stay within the memory limits of a typical personal workstation). The second data-set is from Trapnell *et al.*, 2014. RNA was extracted from primary human myoblasts before and after differentiation (77 and 79 cells respectively) and sequenced in order to investigate for differential expression between these two classes.

Bulk RNA-seq data corresponding to these classes was present in the form of three replicates each, for both the datasets. Differential expression was performed on this read count data in a pair-wise fashion using *DESeq* (see Anders and Huber, 2010), in order to establish the ground truth : genes with an adjusted p-value less than 0.05 and an absolute $\log_2$ fold change greater than two were considered to be differentially expressed. Similarly, genes with an adjusted p-value greater than 0.1 were NOT considered to be differentially expressed.

4 Conclusions

ROSeq produces the largest area under the curve in the comparisons corresponding to the Tung data set (see Figure 1 (b) for the differential expression analysis between individuals NA19098 and NA19101) and is the second-best for the Trapnell data set (see Supplementary Material), which indicates that it performs competitively when compared to other methods. ROSeq also provides a good fit to the actual read count data, as is seen in Figure 1(c) where the coefficient of determination R^2 resulting from fitting a DGBD is high and always above 0.7 in magnitude.

In conclusion, we have developed a software ROSeq that addresses the problem of determining differential expression between two subpopulations. We use a novel approach of employing a DGBD to model read counts corresponding to a gene as a function of ranks. Since read count data arising from any source could be ranked, the usage of ROSeq is in principle extensible to other assays. Parameter tuning for actively adjusting the bin size, procedures for eliminating outliers and dealing with technical dropout noise shall be investigated in the subsequent works.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11:R106**.

Editorial (2014). Method of the year 2013. *Nature Methods*, **11(1)**. https://www.ohio.edu/bioinformatics/upload/Single-Cell-RNA-seq-Method-of-the-Year-2013.pdf.

Elo, L. L., Filen, S., Lahesmaa, R., and Aittokallio, T. (2008). Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **5(3)**, 423–431.

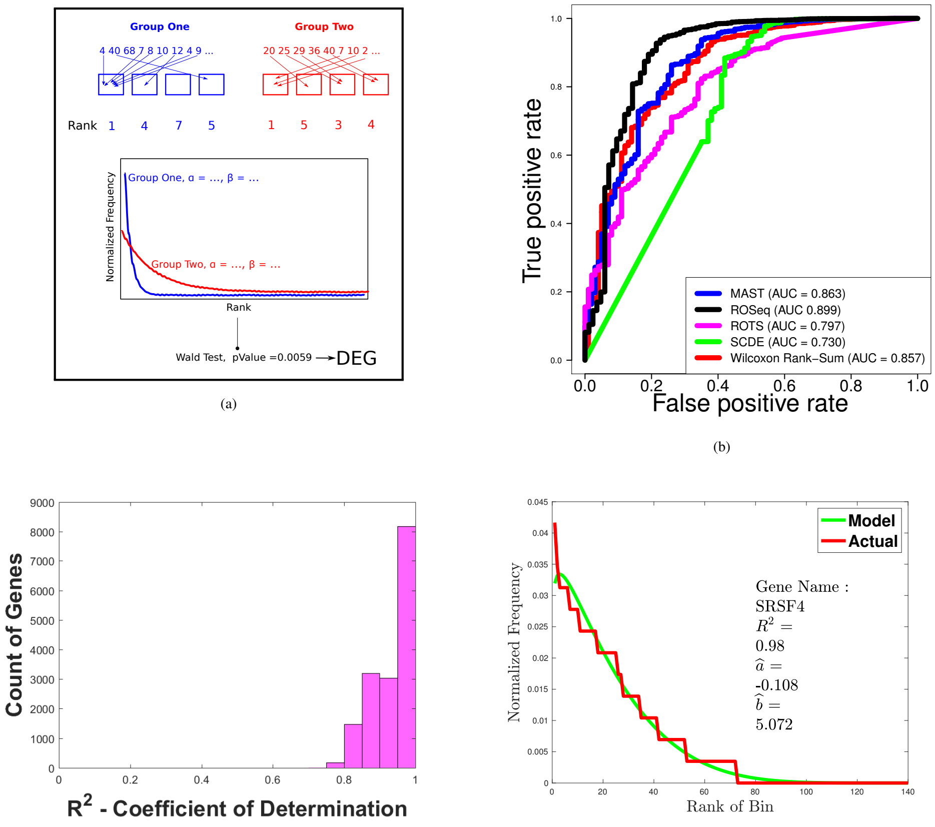
Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, **16:278**.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, **11**, 740–742.

Martinez-Mekler, G., Martinez, R. A., d. Rio, M. B., Mansilla, R., Miramontes, P., and Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE*, **4(3)**.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, **32**, 381–386.

Tung, P., Blischak, J. D., Hsiao, C., Knowles, D. A., Burnett, J. E., Pritchard, J. K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, **7:39921**.



(a)

(b)

(c)

(d)

Fig. 1: (a) An explanatory schematic to ROSeq: for each gene, single cells are assigned to bins (represented as squares in the figure) based on the magnitude of the corresponding, normalized read counts. These bins are subsequently ranked with respect to the number of cells allotted, with rank one being assigned to the bin comprising of the most *number* of cells, and so on. Eventually, a rank distribution is made for both groups and compared - using the Wald test - to determine if the gene is differentially expressed. (b) ROC curve for evaluation of differential expression between individuals NA19098 and NA19101 - ROSeq performs the best in terms of area under the curve (c) Histogram showing the coefficient of determination R^2 from modeling single cell read count data using a Rank-Ordered Distribution on Subpopulation One (NA19098) - Most genes are modeled with a fit higher than 0.7 (d) An example of the model fit on a gene, randomly picked from Subpopulation One (NA 19098)

Document type

Supplementary Material

Title

ROSeq: A rank based approach to modeling gene expression in single cells

Manan Lalit [1], Aditya Biswas [2], Abhik Ghosh [3,*] and Debarka Sengupta [4,*]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany
[2]Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India
[3]Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata 700108, India
[4]Centre for Computational Biology and Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi 110020, India

*To whom correspondence should be addressed.

Contents

1 Extended Introduction

Single Cell RNA-Sequencing (scRNA-seq) has been widely popular since the year 2013 (Nature Methods Editorial, 2014). The dropping cost of sequencing, the arrival of new protocols and the possibility to survey the diversity of cell types has aided this phenomenon and has led to a greater number of cells being sequenced in experiments over the years (For example, Svensson *et al.*, 2018 have shown an exponential increase in the cell numbers reported in publications over time). Due to a higher variability in scRNA-Seq data (see the work of Kharchenko *et al.*, 2014), methods built around bulk cell data are insufficiently able to handle single cell read count data and newer methods are needed for its analysis.

Interesting aspects of single cell read count data include the presence of biological and technical noise. Since the starting quantity of mRNA prior to amplification is less, the *noise* attains a comparable value to the actual gene expression levels, and should not be ignored. As a result, many of the methods (for example, Miao and Zhang, 2017) have focused on modeling the read counts resulting from successful amplification and from noise, individually.

Biological noise arises from the stochastic nature of gene expression. Munsky *et al.*, 2012 suggested that genetically *identical* cells in *identical* environments display variable phenotypes. Others such as Zenklusen *et al.*, 2008 have shown that expression levels vary substantially among cells. Zopf *et al.*, 2013 argue that a substantial portion of the stochastic variability observed in single cell gene expression experiments may be caused by global changes in transcription due to cell cycling.

Technical noise includes non-linear biases which creep in due to the million-fold amplification of the genetic material, and the tendency of the sequencer to miss the read counts for certain genes in a cell-library, because the quantity is lower than a certain observable threshold (this phenomenon is referred to as *drop-outs*).

Different scRNA-seq methods proceed differently, with modeling the read counts corresponding to a gene in a sub-population of cells and handling the above mentioned sources of noise. For example, the SCDE package developed by Kharchenko *et al.*, 2014 models a single cell as a probabilistic mixture of a *negative binomial* to capture successful amplification and a low magnitude *poisson* distribution to represent the technical dropouts and transcriptionally silent genes. In contrast to SCDE, the MAST model developed by Finak *et al.*, 2015 fits a hurdle model across the read counts for a given sub population and solves for the Cellular Detection Rate (CDR), which acts as a proxy for factors (drop outs, cell volume) that influence gene expression.

In this work, we use a Discrete Generalized Beta Distribution (DGBD) in order to provide a non-linear, polynomial fit across the read counts for a given sub population. Previously Zipf, 1949 has shown the existence of a linear relationship between the logarithm of frequency and the logarithm of rank, for the occurrence of words in texts. DGBD is an extension of the original, one parameter Zipf's

Law - Martinez-Mekler *et al.*, 2009 have shown the ability of DGBDs to provide a two parameter (a and b) functional form for rank-ordered distributions, which gives good fits to data from life sciences among other fields of application, and surpasses other two-parameter models.

In order to transform the DGBD into a probability mass function, several *bins* are constructed between the minimum and the maximum value of the read counts corresponding to a gene. Each bin has a width of $k \times \sigma$, where $k = 0.05$ and σ is the standard deviation of the available read count data for each gene. The best fitting parameters $a = \widehat{a}_0$ and $b = \widehat{b}_0$ are found for each sub-population corresponding to a gene by maximizing the Log Likelihood expression as mentioned in Equation 2. Subsequently, in order to evaluate for differential expression between the two sub-populations of single cells, the two-sample Wald Test shall be used to obtain a statistic that varies as a χ_2^2 distribution.

In the next section, mathematical details of the proposed methodology which form the base for determining the statistic shall be highlighted. Subsequently, the performance of the rank-based method, compiled and packaged as a [Windows, macOS, Linux] compatible software called ROSeq (short for '**R**ank-**O**rdered Distributions for scRNA-**Seq** Read Counts') shall be compared with some scRNA-Seq techniques commonly used for evaluating differential expression, namely SCDE (Kharchenko *et al.*, 2014 and Ritchie *et al.*, 2010), MAST (Finak *et al.*, 2015), ROTS (Elo *et al.*, 2008), and Wilcoxon Rank-Sum.

2 Extended Theory

The Discrete Generalized Beta Distribution (DGBD) expresses the normalized frequency y_r or the probability mass function for each *bin* as a function of the rank r of the bin using two parameters a and b. Let N be the total number of bins for a given gene and sub-population. Then the DGBD formulation is expressed as:

$$y_r = A\frac{(N+1-r)^b}{r^a} \tag{1}$$

where A is the normalizing constant ensuring that the sum of the normalized frequencies equals one and is given by:

$$A = \frac{1}{\sum_{r=1}^{r=N} \frac{(N+1-r)^b}{r^a}}$$

For a given data set, we need to properly estimate the parameters a and b so that the fitted DGBD with these estimated parameters are closest to the true data (in a suitable probabilistic sense). The best-fitting parameters $a = \widehat{a}_0$ and $b = \widehat{b}_0$ are determined by maximizing the Log-Likelihood corresponding to the model given by Equation 1. The log-likelihood function, **logL**, for the DGBD expression is specified in the equation 2.

$$\text{logL}(a, b) = -a \times \sum_{r=1}^{r=N} y_r \log(r) + b \times \sum_{r=1}^{r=N} y_r \log(N+1-r) + \left(\sum_{r=1}^{r=N} y_r\right) \log(A) \tag{2}$$

The resulting estimates $(\widehat{a}_0, \widehat{b}_0)$ correspond to the DGBD under which the observed data is most likely to be generated and is commonly known as the maximum likelihood estimator (MLE). They are most efficient (least standard error) and enjoys several optimum properties in a large sample (Casella and Berger, 2002).

Next, let us consider the two sub-populations 1 & 2 with respective number of bins m and n and the problem of testing if the genes in these two sub-populations are differentially expressed in a desired statistical significance level. Let the DGBD parameters corresponding to these sub-populations are denoted by $(a_1,\ b_1)$ and $(a_2,\ b_2)$, respectively, and their MLEs based on the available data are given by $(\widehat{a}_1, \widehat{b}_1)$ and $(\widehat{a}_2, \widehat{b}_2)$. Note that, under the DGBD model, the desired testing for differential gene expressions is equivalent to the test for the null hypothesis $H_0 : a_1 = a_2, b_1 = b_2$ against the omnibus alternative.

For this purpose, here we will use the (asymptotically) optimum two-sample Wald test based on the MLE of the parameters and its asymptotic variances given by the inverse of the Fisher information matrix $I(a,b)$. For the log-likelihood function of the DGBD model given in Equation 2, the estimated Fisher information matrix can be obtained as $I(\hat{a}_0, \hat{b}_0)$, where

$$I(a,b) = \begin{bmatrix} \frac{\partial^2 \text{logL}}{\partial a^2} & \frac{\partial^2 \text{logL}}{\partial a \partial b} \\ \\ \frac{\partial^2 \text{logL}}{\partial b \partial a} & \frac{\partial^2 \text{logL}}{\partial b^2} \end{bmatrix}$$

Note that, for our DGBD model, we have:

$$\begin{aligned} \frac{\partial^2 \text{logL}}{\partial a^2} &= \left(\sum_{r=1}^{r=N} y_r \right) \frac{\partial^2 \log(A)}{\partial a^2} \\ \frac{\partial^2 \text{logL}}{\partial b^2} &= \left(\sum_{r=1}^{r=N} y_r \right) \frac{\partial^2 \log(A)}{\partial b^2} \\ \frac{\partial^2 \text{logL}}{\partial a \partial b} &= \left(\sum_{r=1}^{r=N} y_r \right) \frac{\partial^2 \log(A)}{\partial a \partial b} \end{aligned} \tag{3}$$

So in order to evaluate the above mentioned double derivatives, the first order derivative $\frac{\partial \text{log A}}{\partial a}$ and $\frac{\partial \text{log A}}{\partial b}$ are determined as follows:

$$\begin{aligned} \text{log A} &= -\log \left(\sum_{r=1}^{r=N} \frac{(N+1-r)^b}{r^a} \right) \\ \frac{\partial \text{log A}}{\partial a} &= \frac{1}{\left(\sum_{r=1}^{r=N} \frac{(N+1-r)^b}{r^a} \right)} \times \sum_{r=1}^{r=N} \frac{(N+1-r)^b \log r}{r^a} \\ \frac{\partial \text{log A}}{\partial b} &= \frac{-1}{\left(\sum_{r=1}^{r=N} \frac{(N+1-r)^b}{r^a} \right)} \times \sum_{r=1}^{r=N} \frac{(N+1-r)^b \log(N+1-r)}{r^a} \end{aligned} \tag{4}$$

Re-writing Equation 4 in a more succinct form in the Equation 5 below, we get:

$$
\begin{aligned}
\frac{\partial \text{log A}}{\partial a} = \frac{u_1}{v} \quad &\text{and} \quad \frac{\partial \text{log A}}{\partial b} = \frac{u_2}{v} \\
\text{where, } u_1 &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \log r}{r^a} \\
v &= \left(\sum_{r=1}^{r=N} \frac{(N+1-r)^b}{r^a} \right) \\
u_2 &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \log(N+1-r)}{r^a}
\end{aligned}
\tag{5}
$$

Evaluating the partial derivatives of u_1, v_1, u_2 and v_2 with respect to a and b, in the Equation 6:

$$
\begin{aligned}
\frac{\partial u_1}{\partial a} &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \left(\log r\right)^2}{r^a} \\
\frac{\partial u_1}{\partial b} &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \left[\log r\right] \left[\log (N+1-r)\right]}{r^a} \\
\frac{\partial v}{\partial a} &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \log r}{r^a} \\
\frac{\partial v}{\partial b} &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \log (N+1-r)}{r^a} \\
\frac{\partial u_2}{\partial a} &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \left[\log r\right] \left[\log (N+1-r)\right]}{r^a} \\
\frac{\partial u_2}{\partial b} &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \left[\log (N+1-r)\right]^2}{r^a}
\end{aligned}
\tag{6}
$$

Now, using above formulas, we can easily derive the estimated Fisher Information Matrix for both the sub-populations and hence obtain the asymptotic variance matrices of the MLEs $(\widehat{a}_1, \widehat{b}_1)$ and $(\widehat{a}_2, \widehat{b}_2)$ as given by $\widehat{V}_1 = I(\widehat{a}_1, \widehat{b}_1)^{-1}$ and $\widehat{V}_2 = I(\widehat{a}_2, \widehat{b}_2)^{-1}$, respectively. Then, the two-sample Wald test statistic for testing H_0 is given by:

$$
T = \left(\frac{mn}{m+n} \right) \begin{bmatrix} \widehat{a}_1 - \widehat{a}_2 \\ \widehat{b}_1 - \widehat{b}_2 \end{bmatrix}^T \left(w\widehat{V}_1 + (1-w)\, \widehat{V}_2 \right)^{-1} \begin{bmatrix} \widehat{a}_1 - \widehat{a}_2 \\ \widehat{b}_1 - \widehat{b}_2 \end{bmatrix} \tag{7}
$$

where $w = \frac{n}{m+n}$. If the null hypothesis H_0 is correct, i.e., the genes in the two sub-populations are not differentially expressed, the above test statistics T follows a central chi-square distribution χ_2^2 with two degrees of freedom. Therefore, we conclude that the genes are differentially expressed (i.e., reject H_0) at 95% level of significance, if the observed value of the test statistics T exceeds the 95% quantile of the χ_2^2 distribution (which is approximately 6). The corresponding p-value is given by the probability that a χ_2^2 random variable exceeds the observed value of T.

3 Results

Two data-sets were considered in this study. In the first data-set (obtained from Tung *et al.*, 2017), three replicates from three human induced pluripotent stem cell (iPSC) lines were considered. 96 single cells were considered in each of the three replicates corresponding to one of the three individuals (these individuals shall be referred to by their labels NA19098, NA19101 and NA19239, henceforth), thus leading to a total number of $\frac{\text{96 cells}}{\text{replicate}} \times \frac{\text{3 replicates}}{\text{individual}} \times 3$ individuals $= 864$ cells, which were used for extracting RNA and sequencing to identify read counts for each gene.

In addition to scRNA-Seq read counts, bulk or population RNA-Seq data corresponding to these three individuals was also made available, in the form of 3 replicates each. Differential expression was performed on this bulk cell RNA-Seq read count data in a pair-wise fashion (NA19098 versus NA19101, NA19101 versus NA19239, NA19098 versus NA19239) using a standard, bulk cell Differential Expression technique called *DESeq* (see Anders and Huber, 2010), in order to establish the ground truth : genes with an adjusted p-value less than 0.05 and an absolute $\log_2$ fold change greater than two were considered to be differentially expressed. Similarly, genes with an adjusted p-value greater than 0.1 were NOT considered to be differentially expressed.

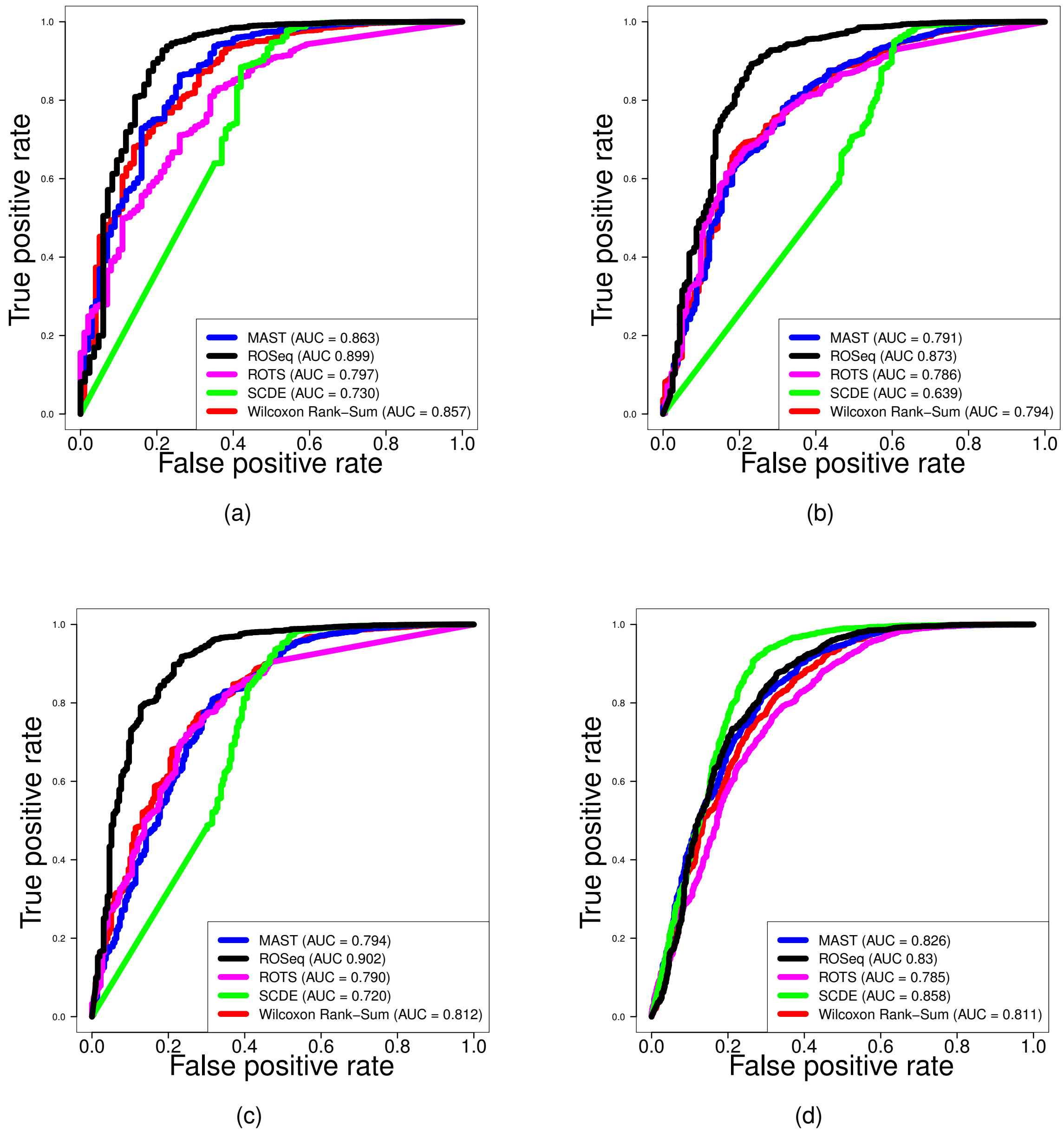
The second data-set is from Trapnell *et al.*, 2014. Single cell transcriptome dynamics was investigated during myogenesis as part of this study. Primary human myoblasts were expanded under high mitogen conditions (GM), and then differentiation was induced by switching to low-mitogen media (DM). RNA was extracted from cells, before and after differentiation and sequenced in order to investigate for differential expression between these two classes. Total number of single cells available for study were 77 and 79 respectively, corresponding to the two classes.

This second data-set also comprised of bulk cell data, which was analyzed in a similar manner as the former study, in order to establish the ground truth. Three replicates each of the two classes were available in this case.

Prior to analyzing the read count data for differential expression, filtering was performed by eliminating any genes with five or less non-zero read counts. This filtering procedure is similar to Hemberg, 2017's filtering procedure, specified on the online scRNA-Seq course. Median Normalization was implemented next for use by methods such as Wilcoxon Rank-Sum and ROSeq (other packages such as MAST, ROTS and SCDE have an inbuilt normalization step in their implementation), in order to equalize coverage across the two sub-populations or classes of cells under investigation.

3.1 Tung Data-set

A receiver operating characteristic curve (ROC) for the three pairwise comparisons is plotted in Figures 1a, 1b and 1c respectively. The figures include the curves based on the prediction from



(a)

(b)

(c)

(d)

Figure 1: (a) ROC curve for evaluation of differential expression between cells from individuals NA19098 and NA19101 (b) ROC curve for evaluation of differential expression between cells from individuals NA19101 and NA19239 (c) ROC curve for evaluation of differential expression between cells from individuals NA19239 and NA19098 (d) ROC curve for evaluation of differential expression between primary human myoblasts before and (24 hours) after differentiation

ROSeq and other methods for evaluating differential expression in scRNA-Seq data sets, namely SCDE, Wilcoxon Rank-Sum, MAST and ROTS. The ROSeq method has the highest Area Under the Curve (AUC) in all the three pairwise comparisons, which indicates that it performs better than other methods at predicting true positives and false positives for the Tung Data-set.

3.2 Trapnell Data-set

A receiver operating characteristic curve was plotted for evaluating differential expression between the two sub-populations of primary human myoblast cells (before and after differentiation) in the Figure 1d. It is seen that the method SCDE performs the best, followed by ROSeq in the second place and the others.

4 Extended Discussion

The four investigations are summarized in terms of rank determined by AUC value in the table below. Analyzing scRNA-Seq read counts using ROSeq by ordering read counts based on rank of the bin and evaluating for differential expression using the Wald's Two Sample Test gives a good magnitude of the area under the ROC curve for the two data-sets. The next best performing method in terms of the AUC magnitude and consistency is MAST.

	MAST	ROSeq	ROTS	SCDE	Wilcoxon Rank-Sum
NA19098 Vs NA19101	2	1	4	5	3
NA19101 Vs NA19239	3	1	4	5	2
NA19239 Vs NA19098	3	1	4	5	2
(Trapnell) Before Vs After Differentiation	3	2	5	1	4

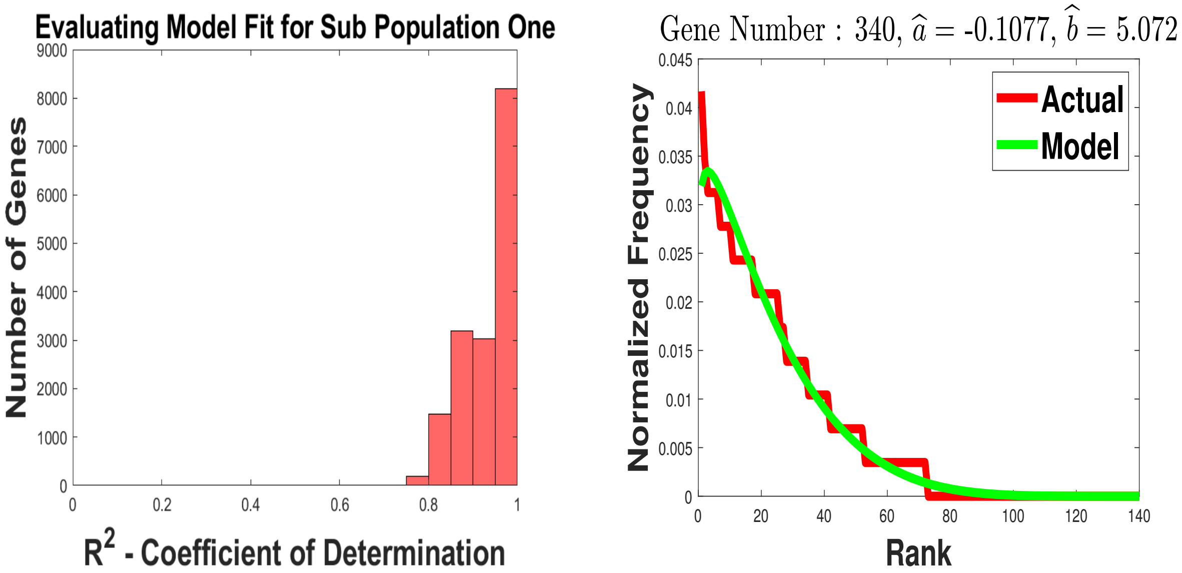
Tabulating performance of different scRNA-Seq methods for the four investigations (3 from Tung Dataset and 1 from Trapnell dataset)

A popular method for the evaluation of differential expression in scRNA-Seq read counts is SCDE (See Kharchenko *et al.*, 2014). For the complete Tung data set, where in any comparison between two individuals involves 576 $(288 \frac{\text{cells}}{\text{sub-population}} \times 2$ sub-populations) cells, SCDE produced an 'out of memory' exception on the personal workstation where the other comparisons were executed. Hence, only 96 single cells for each sub-population were considered for the investigations on Tung dataset above. It is possible that SCDE might perform better if more number of single cells (than 96 currently) are considered in the analysis. For the Trapnell data set, SCDE performs extremely well.

A prediction of the top twenty differentially expressed genes as predicted by ROSeq for the Tung Data-set is plotted as a heat map in the Figures 3a, 3b and 3c; and for the Trapnell Data-set is plotted as a heat map in the Figure 3d. As appears evident upon visual inspection, ROSeq succeeds at identifying differences in the mean levels between two sub-populations.

ROSeq also provides a good fit to the actual read count data, as is seen in Figure 2(a) where the coefficient of determination R^2 resulting from fitting a DGBD to Sub Population One (NA19098), referenced in Tung *et al.*, 2017's dataset is high and always above 0.7 in magnitude. An example of a model fit on a gene, randomly picked from Sub Population One (NA19098) during the differential analysis step between individuals NA19098 and NA19101 is shown in Figure 2.

Currently, as part of the running of ROSeq, the size of the bin has been set equal to $0.5 \times \sigma$, where σ is the standard deviation of the normalized read count data corresponding to a gene. Future work includes performing parametric studies (for example, testing the effect of different bin-sizes on the prediction of differentially expressed genes by ROSeq). Also, a module to enable elimination of outliers and technical drop-out noise using procedures such as trimming or winsorization, is in works



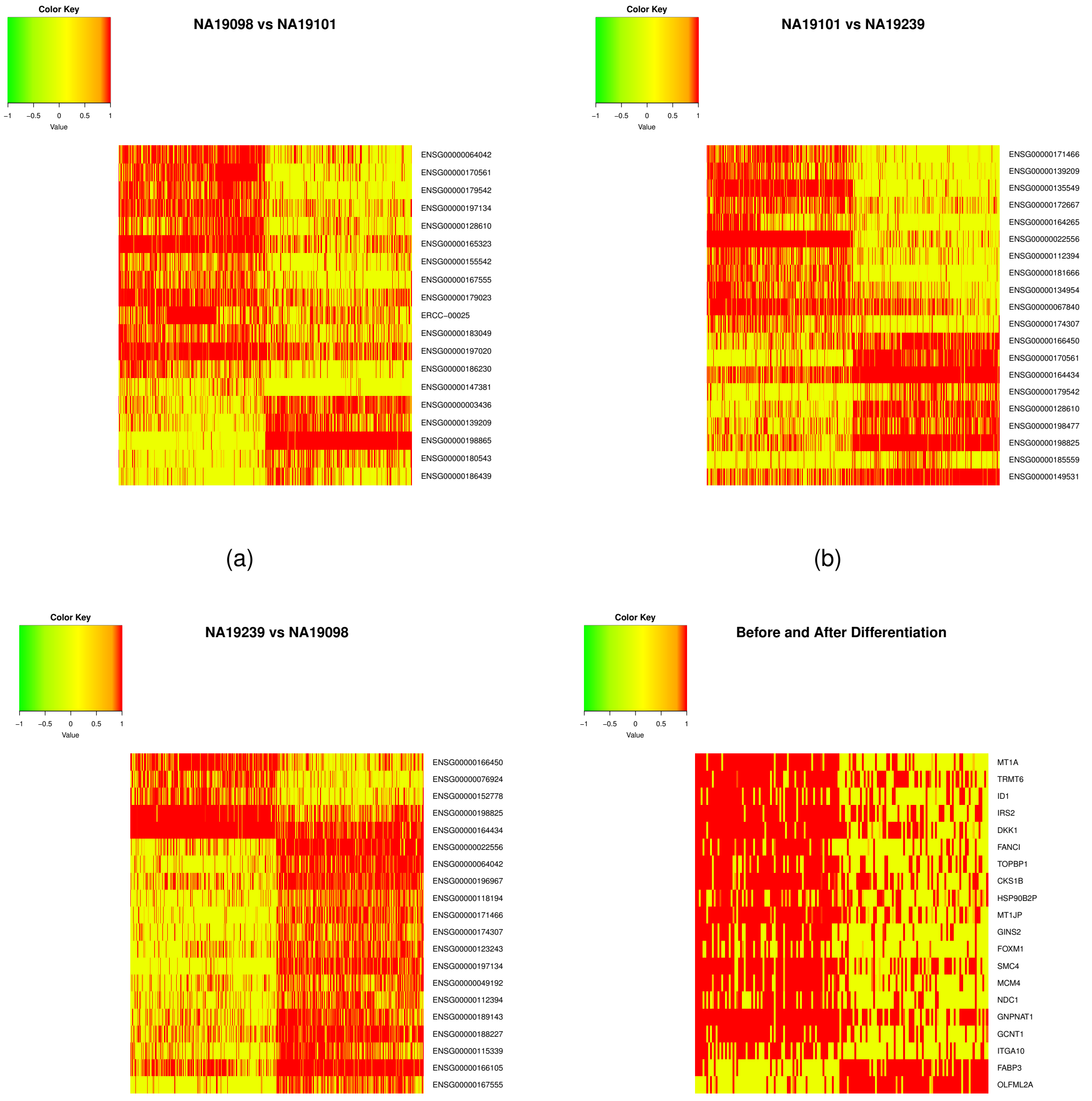
(a) (b)

Figure 2: (a) Histogram showing the coefficient of determination R^2 from modeling single cell read count data using a Rank-Ordered Distribution on Sub Population One (NA19098) - Most genes are modeled with a fit higher than 0.7 (b) An example of the model fit on a gene, randomly picked from Sub Population One (NA 19098)

for the subsequent versions of ROSeq.

Summarizing the results and discussion above, scRNA-Seq read counts were tested for the prediction of differentially expressed genes, as part of two studies - the first was a pairwise comparison between three human induced pluripotent cell lines; and the second comprised of primary human myoblasts sequenced before and after differentiation. Ground Truth was established by applying DESeq on bulk cell data originating from the two classes under investigation. It was noticed that ROSeq provides a very competitive performance in terms of the area under the receiver operating characteristic curve and is recommended as a robust and accurate technique for predicting differential expression between two sub-populations, on account of that.

Since read counts arising from any source (proteomics, ChIP-seq et cetera) can be ordered in a rank wise fashion, ROSeq has implementation value in other fields as well i.e. it has the potential of becoming a general method used for the prediction of differential expression for ANY kind of read count data, irrespective of its origin. Parameter tuning for adjusting the bin size and procedures such as trimming or winsorization for eliminating outliers and technical dropout noise shall be investigated in the subsequent works.



(a)

(b)

(c)

(d)

Figure 3: Heat-maps identifying top twenty most differentially expressed genes, as predicted by ROSeq, between (a) NA19098 (left) and NA19101 (right) (b) NA19101 (left) and NA19239 (right) (c) NA19239 (left) and NA19098 (right) (d) primary human myoblast cells - before (left) and 24 hours after (right) differentiation

5 Installation

ROSeq can be installed as a stand-alone application on a [Windows, macOS or Linux] Operating System. The structure of the available files for installation looks as in Figure 4.

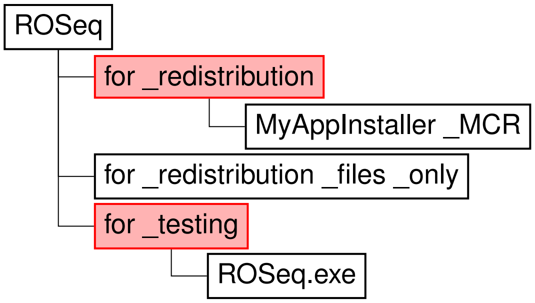


Figure 4: Structure of the ROSeq Directory

The following steps are needed to ensure a successful setup.

- Install by double clicking on “for_redistribution > MyAppInstaller_MCR”.
- Double click on the icon, post the installation process (an icon is placed on the desktop by default).

There are three panels in the software - the *Notes* panel for making notes about the current session, *Status* panel which updates the user if the action went through successfully and the *Pre-Processing* panel which lets the user upload read count data matrix for the desired analysis of differential expression.

If the radio button for Normalized Read Count data file upload is selected, then the path to the respective file needs to be indicated by clicking on the *Open File* button. In this matrix of read counts, each row is indicative of read counts corresponding to one gene while each column is representative of read counts sequenced from one single cell. The *uiimport* function is used to read the number of genes, but that number is changeable in case the user wishes to investigate only a subset of genes.

One of the requirements for the read count matrix is that the two groups of single cells are contiguous. A user input which is desired is the starting and ending column index for each group/subpopulation.

In case a raw read count matrix is uploaded by the user, the minimum number of non-zero read counts for each available gene is specified by the user, as a means to filter out low quality genes. Also, median normalization is used to equalize coverage across all the available single cells.

ROSeq software allows the user to save an existing session through the ‘File > Save’ command and to open an existing session through the ‘File > Open’ command. The ‘Help Menu’ provides a link to open the Supplementary Section Document outside the software.

6 Tutorial

This tutorial below, provides instructions to upload the (un) normalized read count file corresponding to single cells coming from the individuals NA19098 and NA19101, in the Tung *et al.*, 2017 data set; and subsequently analyse the two subpopulations for differential expression.

The number of single cells equals 288 for each individual. After filtering with a threshold equal to six, low quality genes are eliminated and the number of genes considered for differential expression analysis reduces from 19027 to 16087. This is also prompted as a message in the Status Panel.

Following would constitute as a complete set of steps from uploading read count data to saving results as a '*.csv' file:

1. Click on the radio button - 'Upload Raw Read Count file'. This would make visible an 'Open File' button, for the user to upload the respective file.

2. Click on the 'Open File' button, to select the (un) normalized read count file 'single _counts _no _filter _no _norm.csv'. Prior to successful selection of the file, an import wizard window would appear. Click on the buttons 'Next' and 'Finish'.

3. Next, the number of genes in the uploaded file will be suggested (19027). The user could alter the number of genes, in case you wish to look only at the first few (10) genes.

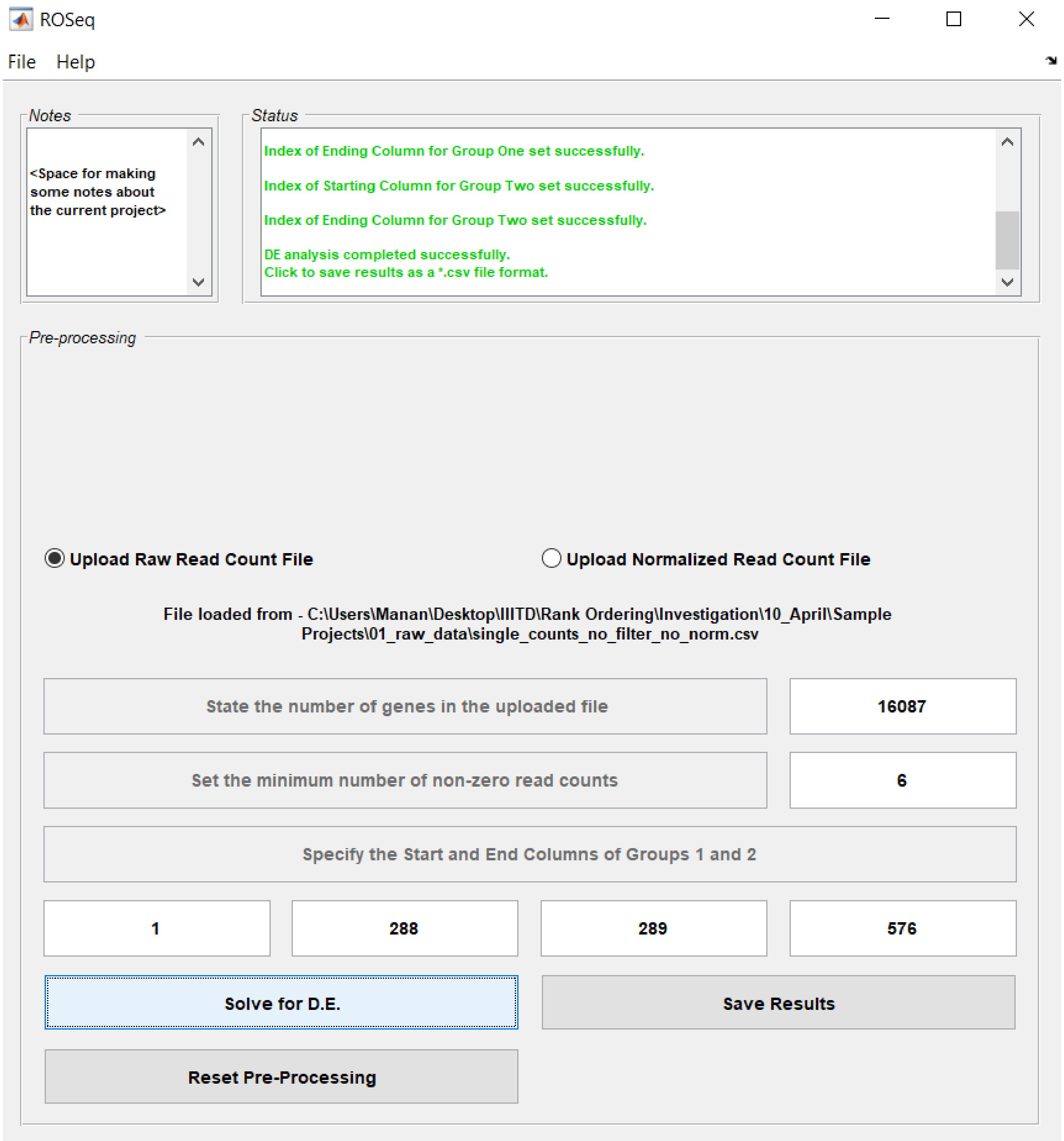
4. Decide on the Filter Threshold. A filter threshold equal to 6, would mean that all genes with the number of non-zero read counts less than 6, would not be considered during the subsequent differential analaysis, as they would be considered low-quality genes.

5. Select the starting (1) and ending (288) column index corresponding to the first group (NA19098). Select the starting (289) and ending (576) column of the second group similarly (NA19101).

6. If all steps were performed so far, then the Solve DE Button becomes enabled. Click on it to run the analysis for differential expression. This would open up a green progress bar for the Subpopulation One, followed by another for Subpopulation Two.

7. Once the analysis is complete, a message ('DE analysis completed successfully.') would appear in the 'Status' Panel. At this stage, the ROSeq software would look as in Figure 5.

8. Lastly click on 'Save Results' button to save the variables [pValue, adjusted pValue and log2(fold count)] as three columns in a '*.csv' file.



<Space for making some notes about the current project>

Index of Ending Column for Group One set successfully.

Index of Starting Column for Group Two set successfully.

Index of Ending Column for Group Two set successfully.

DE analysis completed successfully.
Click to save results as a *.csv file format.

● Upload Raw Read Count File ○ Upload Normalized Read Count File

File loaded from - C:\Users\Manan\Desktop\IIITD\Rank Ordering\Investigation\10_April\Sample Projects\01_raw_data\single_counts_no_filter_no_norm.csv

State the number of genes in the uploaded file: 16087

Set the minimum number of non-zero read counts: 6

Specify the Start and End Columns of Groups 1 and 2

1	288	289	576

Figure 5: Status of the ROSeq software after the successful completion of the differential expression analysis step.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11:R106**.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury, vol 2 edition.

Elo, L. L., Filen, S., Lahesmaa, R., and Aittokallio, T. (2008). Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **5(3)**, 423–431.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, **16:278**.

Hemberg, M. (2017). *Analysis of single cell RNA-seq data*. https://hemberg-lab.github.io/scRNA.seq.course/biological-analysis.html.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, **11**, 740–742.

Martinez-Mekler, G., Martinez, R. A., d. Rio, M. B., Mansilla, R., Miramontes, P., and Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE*, **4(3)**.

Miao, Z. and Zhang, X. (2017). Desingle: A new method for single-cell differentially expressed genes detection and classification. *BioRxiv*.

Munsky, B., Neuert, G., and Oudenaarden, A. V. (2012). Using gene expression noise to understand gene regulation. *Science*, **336(6078)**, 183–187.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2010). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, **43 (7)**.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the last decade. *Nature Protocols*, **13**, 599–604.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature Biotechnology*, **32(4)**, 381–386.

Tung, P., Blischak, J. D., Hsiao, C., Knowles, D. A., Burnett, J. E., Pritchard, J. K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scintific Reports*, **7:39921**.

Zenklusen, D., Larson, D. R., and Singer, R. H. (2008). Single-rna counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, **15(12)**, 1263–1271.

Zipf, G. K. (1949). Human behavior and the principle of least effort. *Cambridge, MA: Addison-Wesley Press*, page 573.

Zopf, C. J., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-cycle dependence of transcription dominates noise in gene expression. *PLOS Computational Biology*.