Survey of differential expression analysis methods available for single-cell (count) data

Nastaran Meftahi

Abstract

Single-cell RNA-sequencing (scRNA-seq) is a breakthrough in biomedical and biological research for studying stochastic gene expressions. Compared to bulk RNA data, it provides higher resolution for identification of novel cell types and quantifying transcript of individual cells. One of the dominant proposed methods for analyzing single-cell data is differential expression (DE) analysis on scRNA-seq data. One difficulty dealing with these data is high dimension and technical noise, which makes analysis difficult or misleading. To deal with these issues, many methods has been developed. Some imputation methods has been used to recover dropout events and reduce the dimensionality. Some DE methods focus on unbiased gene expression and modeling technical and biological variations. The focus of this study is on the DE analysis methods proposed in the literature.

I. Introduction

Bulk RNA sequence (RNA-seq) data is a summarization of read sequences for mapping gene expressions. Even though they are very useful, however, these data are unable to capture gene-specific heterogeneity per cell. In contrast, single-cell RNA-sequencing (scRNA-seq) provides higher resolution compared to bulk RNA-seq data. It is suitable for identification, monitoring and characterization of cell population [1]–[3]. By analyzing gene expression at this level, biomedical researchers and scientists can analyze stochastic gene expression and cellular processes inside each cell [4], [5].

Nevertheless, scRNA-seq data is contaminated by technical and biological noise. The technical variability comes from low RNA counts. Dealing with only one cell, makes it difficult to capture all transcripts and some of them can be missed [6]. To profile the low counts, amplification methods are used for sequencing [7]. However, amplification distorts the relative transcription abundance and causes dropout events in the data. Dropout events happen when we have false gene quantification because of missing transcripts in the reverse-transcript step. So the gene is not detected, high expression in some cells and low expression in others [4], [7].

Biological variability comes from the stochastic nature of gene transcripts. Identifying these variations from the technical variation is essential for unbiased gene expression and differential expression (DE) analysis. In scRNA-seq data we are dealing with technical noise and multiple cell states inside a cell, which causes multi-modality in gene expression. This makes it challenging to perform an unbiased gene DE analysis, which is one of the applications of scRNA-seq data [3], [5], [7].

Variability in gene expressions in different cells as well as stochasticity of within cell gene-expression, results in high dimensionality of scRNA-seq data. Projection of scRNA-seq data to lower dimension and classifying cells can facilitate computational analysis [2], [6].

Many applications has been developed to deal with these issues and perform an unbiased DE analysis. The objective of DE analysis is to differentiate between the technical noise and biological variations [3], [6]. The aim of this study is to present different methods of DE analysis on scRNA-seq data. Many methods has been presented in the literature, some of them transform the scRNA-seq data by imputation and use methods developed for bulk RNA-seq data for scRNA-seq data. Some other methods are specifically developed for scRNA-seq data. In this survey, we focus on the latter methods. The focus the study is on 7 different models. These models are DEseq2, Single-cell differential expression (ScDE), model-based analysis of single-cell transcriptomics (MAST), Single cell differential distribution (ScDD), Single-cell two-phase testing procedure (SC2P), DESingle and Discrete distributional differential expression (D3E).

In the final section of the paper, I present an implementation of DE analysis and compare it with the results of D3E method, which is originally implemented in python, but I have implemented it as a package in R.

II. Methods

In this section, the 7 models for DE analysis are presented. The papers are presented based on the year of publication. All methods are specifically designed to work on scRNA-seq data and identifying differentially expressed genes.

A. DEseq2

In most scRNA-seq studies, the DE genes are selected based on distribution tests and p-values. However, still these genes might not be expressed adequately. Therefore, DEseq2 presents a model which performs a shrinkage estimation for dispersion and fold-change to enhance the performance of DE analysis.

DESeq2 allows for gene-specific shrinkage estimation. Negative binomial (NB) generalized linear model (GLM) by method of moments is used to estimate the gene-wise depression [6]. The read counts are modeled by GLM of NB with dispersion parameter α_i and logarithmic link.

$$K_{ij} = \mathrm{NB}(\mu_{ij}, \alpha_i) \tag{1}$$

where μ_{ij} is the mean and normalized by a constant normalization term s_{ij}, which is estimated by the median of count ratios. The model assumes that the dispersion parameter α_i follows a log-normal prior distribution, centered around the mean normalized read count. The final values of α_i are estimated by performing using empirical Bayes.

To perform a DE analysis, DEseq2 performs a Wald test. The p-values are then passed to a method of Benjamini and Hochberg false detection ratio (FDR) to filter out genes which are below a specific FDR value and accept the rest as differentially expressed [8].

B. Single-cell differential expression (ScDE)

ScDE is a Bayesian approach to scRNA-seq data analysis. It has a bimodal approach towards gene expression analysis, by modeling gene-specific expressions with a mixture of low-magnitude Poisson and zero-inflated NB [9]. The low-level Poisson model accounts for dropouts and background silent gene transcripts. The zero-inflated NB model detects the components which are related to abundance.

The expected probability of differentially expressed genes has been modeled by

$$p_S(x) = \mathbb{E}\left[\prod_{c \in B} p\left(x | r_c, \Omega_c\right)\right] \tag{2}$$

where $p\left(x | r_c, \Omega_c\right)$ is the posterior probability of gene expression at level x in cell subpopulation S. The average expression level is x, c is a cell bootstrapped in B from a population of cells S. The posterior probability of gene expression is calculated as follows

$$p(x | r_c, \Omega_c) = p_d(x) p_{\mathrm{pois}}(x) + (1 - p_d(x)) p_{\mathrm{NB}}(x | r_c) \tag{3}$$

where p_d is the probability of dropout, p_{pois} is the probability of observing expression magnitude of r_c and p_{NB} is the probability of successful gene amplification for an expressed gene at the average level x. After normalization of posterior probabilities of the genes, to identify DE genes, p-values are calculated [7].

C. Model-based analysis of single-cell transcriptomics (MAST)

MAST is a hurdle method, which considers multi-modality associated with scRNA-seq data, due to the low library size and dropout events. This method models dropouts with a bimodal distribution and the inference is improved by using an empirical Bayesian framework. Gene expression level Z, which is a discrete variable is modeled by logistic regression as

$$\text{logit}(p(Z_{ig=1})) = X_i\beta_g^D \tag{4}$$

where $Z_{ig=1}$ is the expression level of gene g in cell i, the non-zero expressions $(Y|Z=1)$ are fitted by a Gaussian generalized linear model (GLM) given as

$$p(Y_{ig}|Z_{ig}=1) = N(X_i\beta_g^C, \sigma_g^2) \tag{5}$$

The cellular detection rate (CDR) is calculated as a proportion of genes expressed in a cell [9]. The DE genes are identified by shrunken variance estimates, derived from an empirical Bayes approach and the genes with p-values less than FDR of 0.01 are selected as differentially expressed [10].

D. Single-cell differential distribution(ScDD)

ScDD considers a multi-modal distribution of a gene expression and uses Bayesian modeling for identifying differential distribution (DD) across conditions. The DD genes are classified based on their expressions as DE genes [11].

ScDD accepts normalized log-transformed nonzero expression measurements of pairs of cells as input. Let $Y_g = (y_{g1}, \ldots, y_{gJ}$, be the log-transformed nonzero measurement of gene g in cell J. The authors model Y_g by conjugate Dirichlet process mixture (DPM). Zero-expressions for non-DD genes are modeled as logistic regression adjusted for the proportion of genes detected in each cell.

The null hypothesis of the model is unconditional equivalent distributions. For evidence of independent condition-specific models for two groups of data, Bayes factor is calculated in the following equation

$$\text{BF}_g = \frac{f(Y_g|M_{DD})}{f(Y_g|M_{ED})} \tag{6}$$

where M_{DD} and M_{ED} are the DD hypothesis and the equivalent distribution hypothesis respectively and $f(Y_g|M)$ denotes the predictive distribution of the observations from gene g under the given hypothesis [1].

The DPM does not have an analytical solution. However, by considering clusters of samples of the mixture components, we can achieve a closed-form solution for $f(Y_g, Z_g|M)$, where Z_g is a partition. It is not possible to integrate out Z_g, so the authors introduced an approximate Bayes factor as follows

$$\text{Score}_g = \log\left(\frac{f_{C1}(Y_g^{C1}, Z_g^{C1}) f_{C1}(Y_g^{C2}, Z_g^{C2})}{f_{C1,C2}(Y_g, Z_g)}\right) \tag{7}$$

where variables $C1$ and $C2$ denote condition 1 and 2, respectively. A high score value presents evidence that a given gene is differentially distributed.

For every permutation of data, the approximate Bayes factor and for each gene, for determining the DD genes, an empirical p-value is calculated [1], [12].

E. Discrete distributional differential expression (D3E)

D3E is a DE analysis method which uses both non-parametric distribution tests such as Kolmogorov-Smirnov test, Anderson-Darling and Cramer-von Mises test to test a null hypothesis of identical distribution for two groups. This model alternatively can perform a Bayesian parameter estimation and parametric test based on likelihood ratio test. Similar to ScDE, the count data is normalized by geometric mean [12].

For Bayesian parameter estimation, D3E models gene expression data based on a mixture of Poisson-Beta distribution, and then performs a likelihood ratio test.

$$\mathrm{PB}(n|\alpha,\beta,\gamma,\lambda) = \mathrm{Pois}\left(n|\frac{\gamma x}{\lambda}\right) \underset{x}{\wedge} \mathrm{Beta}(x|\alpha,\beta) \tag{8}$$

where n is the number of transcripts in a particular gene, x is an auxiliary variable, α and β are rate of promoter activation and inactivation respectively, γ is the rate of transcription for active promoter, λ is the transcript degradation rate. α, β and γ are normalized by the rate of mRNA degradation λ. The assumption of the model is that the samples are drawn from a stationary distribution and λ is constant between samples. Hence, it is possible to estimate the other three parameters by moments matching or Bayesian inference [6], [13].

D3E is based on comparison of two probability distributions for performing DE analysis. A restriction of D3E is that the range of cell groups must be known in advance. Therefore, this method is only applicable to cases where the cell labels are known in advance [13].

Unlike other DE analysis methods which compare the first moment of genes, D3E compares the full distribution of each gene. The DE analysis is done by calculating two-sample p-values. Then the p-values are filtered by Benjamini and Hochberg FDR. The genes which have a lower p-value than FDR are passed as DE genes.

F. DESingle

DESingle uses zero-inflated NB regression model to estimate the proportion of dropouts [6]. This model applies modified median normalization method on raw data. Then gene expression in each cell is estimated by using zero-inflated NB. The parameter of the model are identified by maximum likelihood estimator. For detection of DE genes, likelihood ratio test is performed and similar to D3E, DESingle can only perform a pairwise comparison of distribution of groups of cells [12]. The pmf of the zero-inflated NB model for gene g in cell i follows below formula

$$p(y_{gi} = n|\theta, r, p) = \theta I(n = 0) + (1 - \theta)\binom{n+r-1}{n} p^n (1 - p)^r \tag{9}$$

where θ is the proportion of zero gene expression in cell i,r and p are the size and probability parameters of NB distribution respectively. The likelihood ratio test of the parameters for pairwise cell groups can classify differentially expressed genes into three categories. The first category is DE abundance (DEa), that are DE genes which do not show significant different proportion of real zeros, but are significantly differentially expressed in other cells. The second category are called DE status (DEs). These DE genes are significantly differentially expressed between the two groups under consideration. But they are not DE in other cells. The third category is DE general (DEg). These genes are significantly differentially different in zero-proportion as well as expression abundance [6], [14].

G. Single-cell two-phase testing procedure (SC2P)

SC2P approaches DE analysis from a two-phase perspective [11]. It considers the cell and gene-specific contexts, then that first identifies the phase of expression a gene is in, by taking into account of both cell and gene-specific contexts, in a model-based and data-driven fashion. The authors in [11], consider two phases for genes, one at dropout, when the gene fail to be detected due to technical noise, and another phase where the DE genes are detected and can be identified. This model identifies the genes that go

through different forms of DE, i.e., different frequencies in different populations or different transcription rates.

Phase I is modeled with a zero-inflated Poisson (ZIP) distribution $Y_{gi}|Z_{gi} = 0 \sim \mathrm{ZIP}(p_i, \lambda_i)$. Y_{gi} is the number of counts for gene g and cell i. Variable Z_{gi} is a binary latent variable, which is equal to 1 in phase II, p_i is the extra mass variable which models the zero-inflation in count data and λ_i is the Poisson rate

$$\theta_{gi}|Z_{gi} = 1 \sim \mathrm{LN}(\mu_g, \sigma_g^2), \qquad Y_{gi}|\theta_{gi} \sim \mathrm{Pois}(\theta_{gi} Si), \tag{10}$$

where S_i is the sequencing depth in i. The LNP model captures heterogeneity among cell and gene expression. The marginal distribution of the model follows the next equation

$$p(Y_{gi} = y_{gi}) = (1 - \pi_i)\mathrm{ZIP}(y_{gi}|p_i, \lambda_i) + \pi_i \mathrm{LNP}(y_{gi}|\mu_g, \sigma_g^2), \tag{11}$$

where π_i is the prior probability of gene g in cell i to be in a specific transcription phase. The posterior probability of gene counts can be estimated by considering the set of hyper-parameters of the model and observed count.

The threshold to filter out low count genes is ZIP parameters at 99^{th} percentile [11]. The genes which pass this filter enter phase II for estimation of parameters via empirical Bayesian shrinkage methods. By borrowing information among genes, using the log of count data from phase I and using $\mu_g \sim N(\mu_0, \sigma_0)$ and $\sigma_g^2 \sim \mathrm{Inv}{-}\chi^2(\nu_0, \tau_0)$ parameters as priors, μ_g and σ_g^2 can be estimated. Then it is possible to determine the posterior probability of π_{gi} given the gene counts in each cell.

$$\hat{\pi}_{gi} = p(Z_i = 1|Y_{gi} = y_{gi}) = \frac{\hat{\pi}_i \mathrm{LNP}(y_{gi}|\hat{\mu}_g, \hat{\sigma}_g^2)}{\hat{\pi}_i \mathrm{LNP}(y_{gi}|\hat{\mu}_g, \hat{\sigma}_g^2) + (1 - \hat{\pi}_i)\mathrm{ZIP}(y_{gi}|\hat{\lambda}_i, \hat{p}_i)} \tag{12}$$

where $\hat{\pi}_i$ is the estimated mixture probability for Phase II.

The DE genes can be expressed in the frequency or magnitude of expression. Belonging to phase II which can be determined by checking the posterior probability by considering $\hat{\pi}_i > 0.99$. DE in this phase is determined by logistic regression of $\hat{Z}_i$ [11].

III. Differential expression of scRNA-seq data

In this section, I followed two approaches for DE analysis. In the first approach, I use the full data set and apply dimensionality reduction methods to deal with the high dimension of the data, then used hierarchical clustering to cluster cells. Then use t-test to Identify DE genes and plot them by filtering the genes with changes less than two-fold.

I also apply DE analysis on the first 100 genes, to compare the results with D3E package. The reason for selecting 100 genes is the running time of D3E package and using Bayesian parameter estimation and two-sample Kolmogorov-Smirnov test.

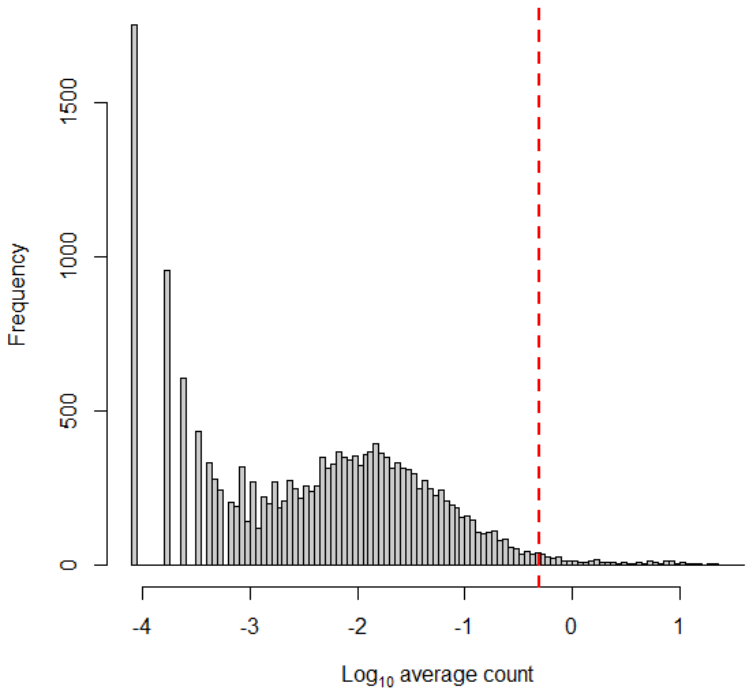


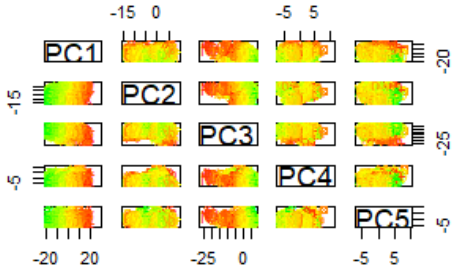
Fig. 1: Histogram of the data after dimensionality reduction.

A. Detailed analysis on full data

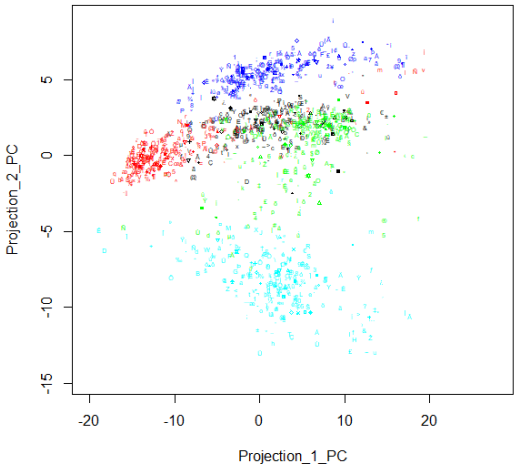
At first, I apply feature selection and dimensionality reduction for clustering and determining DE genes by t-test and FDR. By applying feature selection and filtering gene expressions with low abundance, the number of genes reduced from 17938 to 309. The idea is that genes with low abundance are often caused by technical noise and do not represent biological variability in the data. We can recognize low-abundant genes by calculating the mean and dispersion of gene expression and filtering out the genes with mean lower that 0.5. Fig. 1 shows a histogram of the data after removing genes with low abundance.

In the next step, I checked the sum expression of each gene in all the cells, with the objective of removing the genes which are expressed less than a threshold. However, it did not influence the size of the data as all genes were expressed more than the threshold of 10.

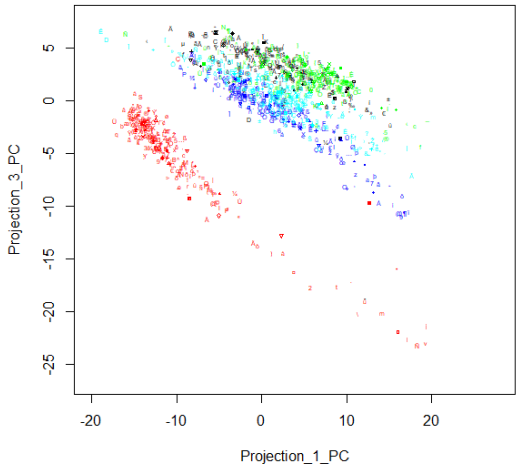
To reduce the dimensionality of the data even further, I used principal component analysis (PCA) on the data. Plotting the first 5 principal component (PC)s is not very informative as it is shown in fig. 2a. Plotting PC 1 and PC 2 is not visually informative either as it is depicted in fig. 2b. The plot of PC 1 and PC 3 in fig. 2c, shows a better picture of the data. The plot of PC 3 and PC 4 in fig. 2d separates the cells with clear boarders.



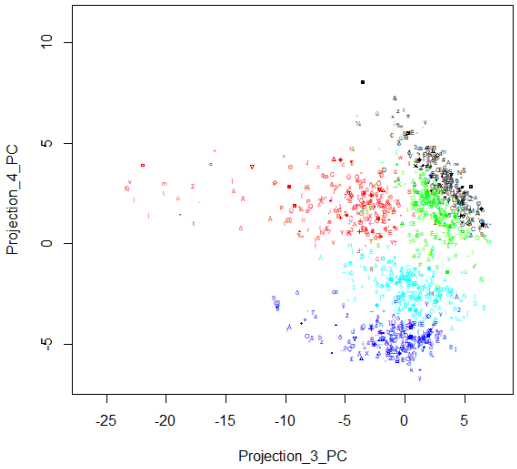
(a) The first 5 PCs.



(b) PC 1 vs PC 2.



(c) PC 1 vs PC 3.



(d) PC 3 vs PC 4.

Fig. 2: PCA analysis of scRNA data

Analyzing the data shows that the first 250 PCs explain about 95% of variations in the data. The plot of cumulative explained variance is presented in fig. 3.

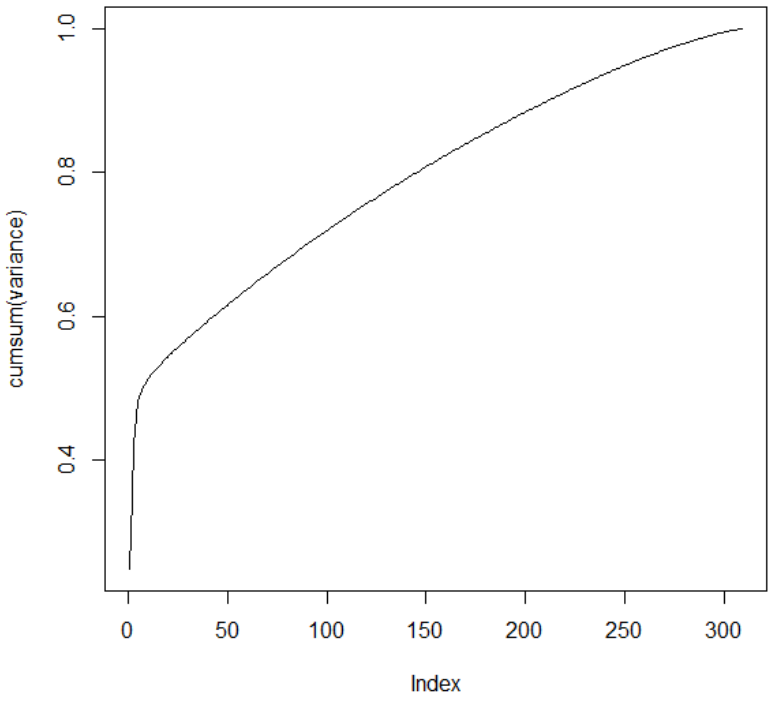


Fig. 3: Cumulative sum of explained variance.

I also used Ward's hierarchical clustering with Euclidean similarity measure to cluster the data. The heatmap of the results can be seen in the following in fig. 4.

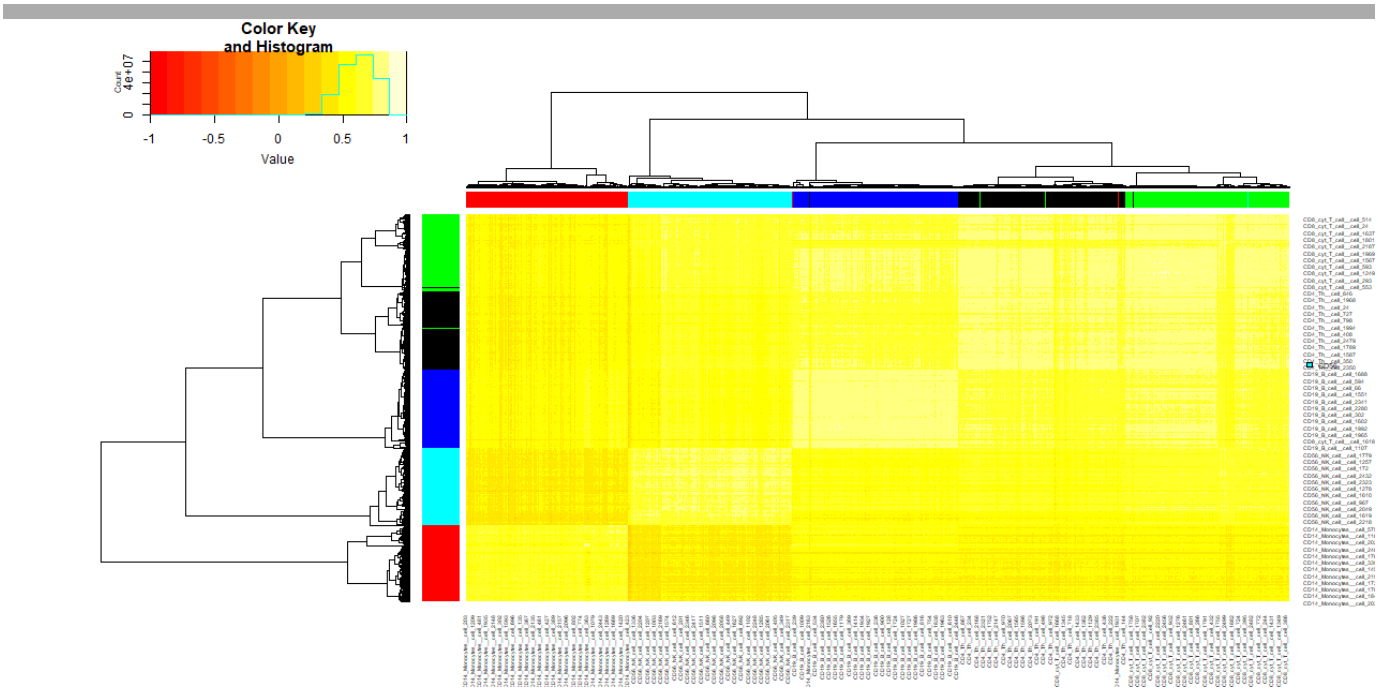
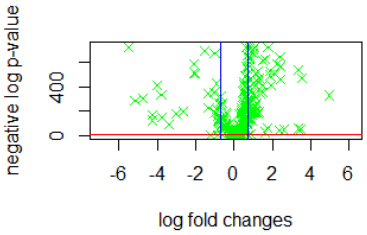


Fig. 4: Hierarchical clustering of gene data.

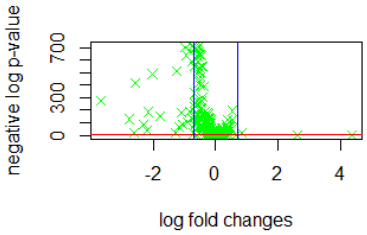
The cells belonging into one group are clustered in similar groups in the data. Five cell types and the

cells belonging to these types are recognized and categorized in a visually clear way.

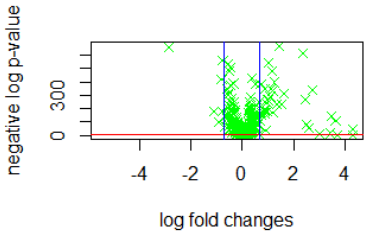
In this step, I perform a t-test to determine DE genes by checking their adjusted p-values and fold changes. We have 5 cells and the cells are compared pairwise. Thus, we have 10 sets of plots as follows in figs. 5a to 5j.



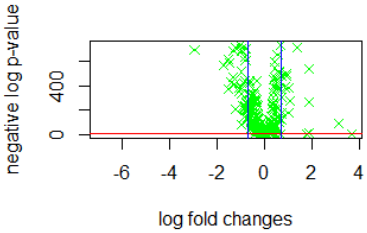
(a) Pairwise gene comparison of cell 1 and cell 2.



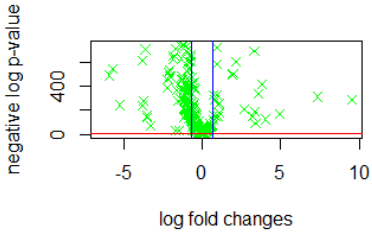
(b) Pairwise gene comparison of cell 1 and cell 3.



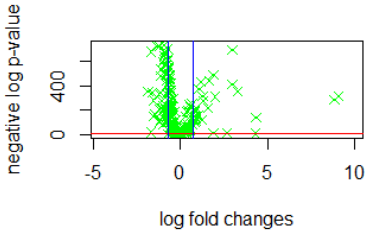
(c) Pairwise gene comparison of cell 1 and cell 4.



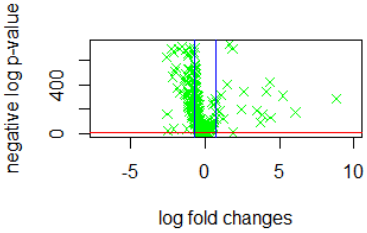
(d) Pairwise gene comparison of cell 1 and cell 5.



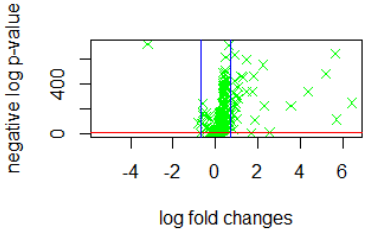
(e) Pairwise gene comparison of cell 2 and cell 3.



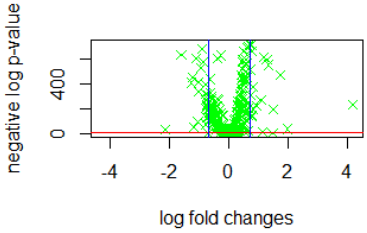
(f) Pairwise gene comparison of cell 2 and cell 4.



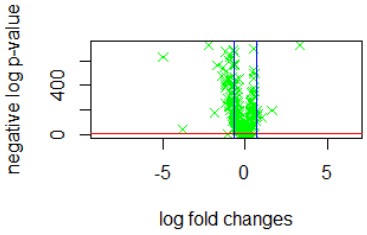
(g) Pairwise gene comparison of cell 2 and cell 5.



(h) Pairwise gene comparison of cell 3 and cell 4.



(i) Pairwise gene comparison of cell 3 and cell 5.



(j) Pairwise gene comparison of cell 4 and cell 5.

Fig. 5: Pairwise comparison of full data set.

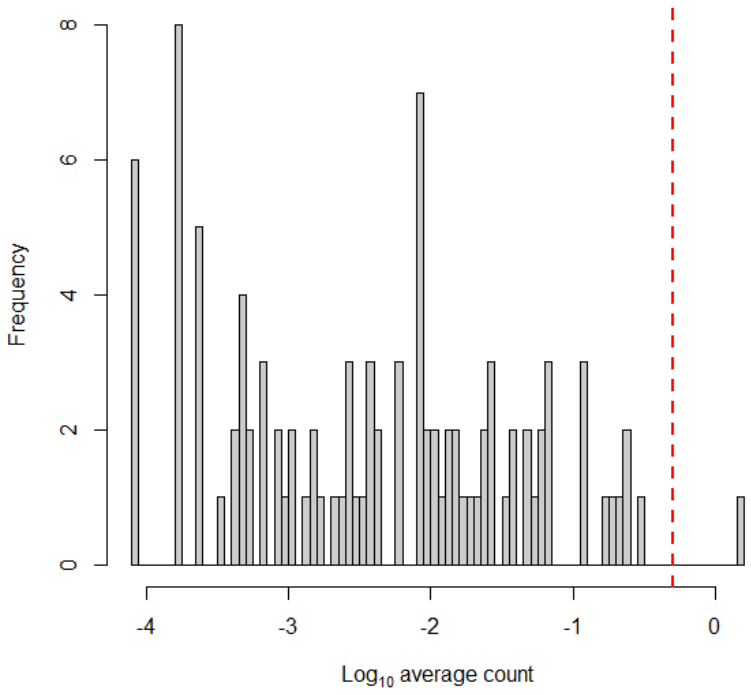


Fig. 6: Histogram of gene abundance of the first 100 genes.

B. Detailed analysis on 100 first genes

In this part, 100 of the genes are analyzed in the same manner as before. The objective is to use these results and compare them with the results from D3E models. In the first step, a histogram of gene abundance of first 100 genes can be seen in fig. 6.

In the second step, the t-test and two-fold expression analysis has been done. The histogram of DE genes can be seen as follows in figs. 7a to 7j.

C. Comparison of the results

In this section, the results of applying D3E for DE analysis on the first 100 genes in the data is presented. The histogram of results can be seen in figs. 8a to 8j.

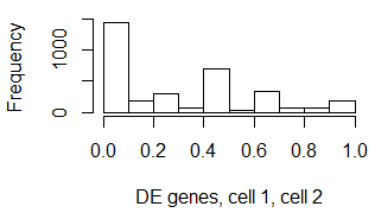
To compare the results with the previous part, I provide the number of DE genes with t-test as well a D3E between cells in table 1.

Test	Cell 1 vs 2	Cell 1 vs 3	Cell 1 vs 4	Cell 1 vs 5	Cell 2 vs 3	Cell 2 vs 4	Cell 2 vs 5	Cell 3 vs 4	Cell 3 vs 5	Cell 4 vs 5
t-test	37	12	21	30	36	32	36	13	23	30
D3E	41	35	43	23	30	33	30	13	26	32

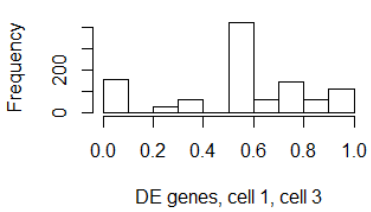
The number of DE genes in D3E is larger than the t-test. But mostly the difference is not huge, except for comparing cell 1 with cell 3 and cell 4. By looking at the DE genes, we can see that most of the genes also overlap.

IV. Conclusion

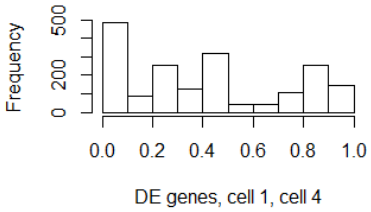
In this paper, I presented 7 different methods, DSeq2, ScDE, MAST, ScDD, D3E, DESingle and SC2p in DE analysis. All the methods have implemented packages in R, but D3E which is implemented in



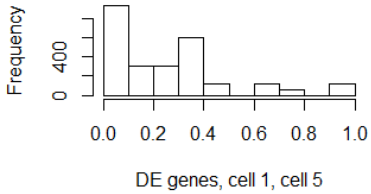
(a) Histogram of genes identified as DE cell 1 and cell 2.



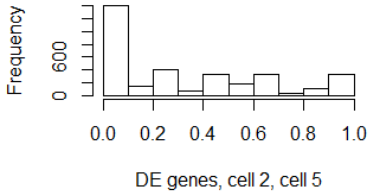
(b) Histogram of genes identified as DE cell 1 and cell 3.



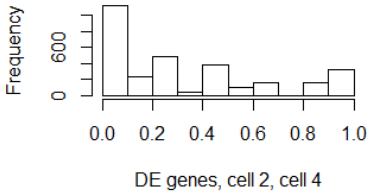
(c) Histogram of genes identified as DE cell 1 and cell 4.



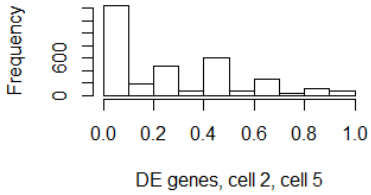
(d) Histogram of genes identified as DE cell 1 and cell 5.



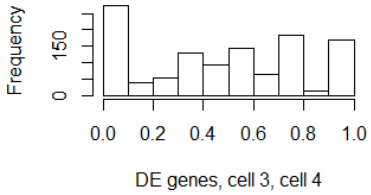
(e) Histogram of genes identified as DE cell 2 and cell 3.



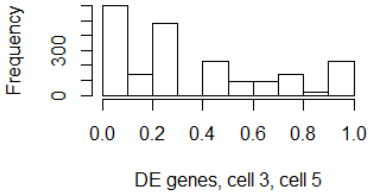
(f) Histogram of genes identified as DE cell 2 and cell 4.



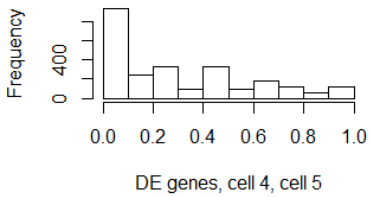
(g) Histogram of genes identified as DE cell 2 and cell 5.



(h) Histogram of genes identified as DE cell 3 and cell 4.



(i) Histogram of genes identified as DE cell 3 and cell 5.

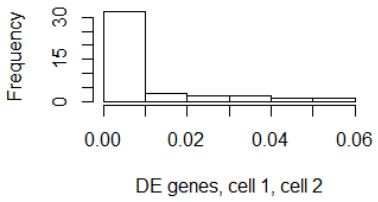


(j) Histogram of genes identified as DE cell 4 and cell 5.

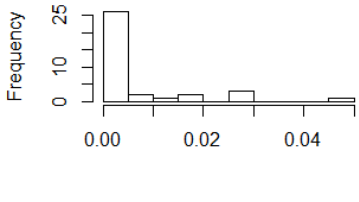
Fig. 7: Pairwise comparison of 100 genes with t-test.

python. In scRNA seq data, the technical noise cannot be modeled, so using a consensus approach is the best way to find DE genes [13]. So, it is a good practice to apply a couple of methods for analysis to identify differentially expressed genes.

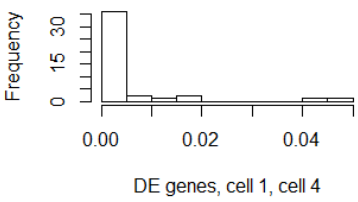
In the final parts of the paper, I implemented two methods for DE analysis, a t-test and D3E method. Based on the results of statistical analysis, we can see that different methods provide different statistical



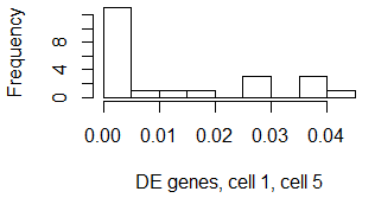
(a) Histogram of genes identified as DE cell 1 and cell 2 by D3E.



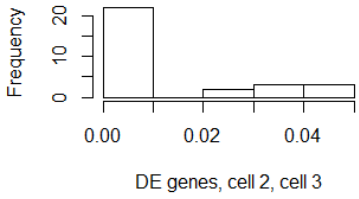
(b) Histogram of genes identified as DE cell 1 and cell 3 by D3E.



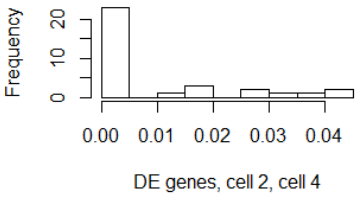
(c) Histogram of genes identified as DE cell 1 and cell 4 by D3E.



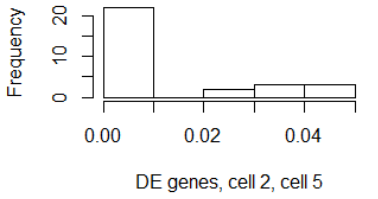
(d) Histogram of genes identified as DE cell 1 and cell 5 by D3E.



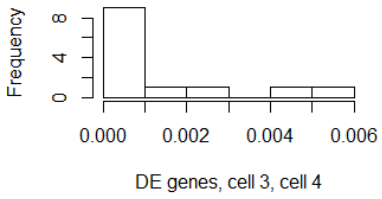
(e) Histogram of genes identified as DE cell 2 and cell 3 by D3E.



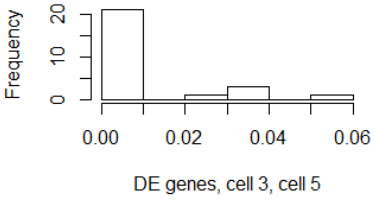
(f) Histogram of genes identified as DE cell 2 and cell 4 by D3E.



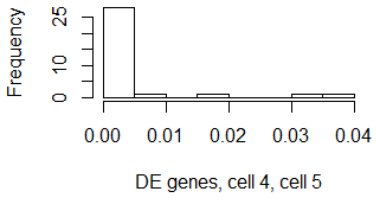
(g) Histogram of genes identified as DE cell 2 and cell 5 by D3E.



(h) Histogram of genes identified as DE cell 3 and cell 4 by D3E.



(i) Histogram of genes identified as DE cell 3 and cell 5 by D3E.



(j) Histogram of genes identified as DE cell 4 and cell 5 by D3E.

Fig. 8: Pairwise comparison of 100 genes with D3E method.

results. However, two methods show some overlap in identification of differentially expressed genes.

REFERENCES

[1] K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski, "A statistical approach for identifying differential distributions in single-cell rna-seq experiments," *Genome biology*, vol. 17, no. 1, p. 222, 2016.

[2] T. S. Andrews and M. Hemberg, "Identifying cell populations with scrnaseq," *Molecular aspects of medicine*, vol. 59, pp. 114–122, 2018.

[3] A. Dal Molin, G. Baruzzo, and B. Di Camillo, "Single-cell rna-sequencing: assessment of differential expression analysis methods," *Frontiers in genetics*, vol. 8, p. 62, 2017.

[4] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, pp. 133–145, 2015.

[5] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell rna-sequencing experiments," *Genome biology*, vol. 17, no. 1, p. 63, 2016.

[6] T. Wang, B. Li, C. E. Nelson, and S. Nabavi, "Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data," *BMC bioinformatics*, vol. 20, no. 1, p. 40, 2019.

[7] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature methods*, vol. 11, no. 7, pp. 740–742, 2014.

[8] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[9] L. Zhang and S. Zhang, "Comparison of computational methods for imputing single-cell rna-sequencing data," *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[10] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic *et al.*, "Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data," *Genome biology*, vol. 16, no. 1, pp. 1–13, 2015.

[11] Z. Wu, Y. Zhang, M. L. Stitzel, and H. Wu, "Two-phase differential expression analysis for single cell rna-seq," *Bioinformatics*, vol. 34, no. 19, pp. 3340–3348, 2018.

[12] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nature methods*, vol. 15, no. 4, p. 255, 2018.

[13] M. Delmans and M. Hemberg, "Discrete distributional differential expression (d 3 e)-a tool for gene expression analysis of single-cell rna-seq data," *BMC bioinformatics*, vol. 17, no. 1, p. 110, 2016.

[14] Z. Miao, K. Deng, X. Wang, and X. Zhang, "Desingle for detecting three types of differential expression in single-cell rna-seq data," *Bioinformatics*, vol. 34, no. 18, pp. 3223–3224, 2018.