An Interactive Dashboard for visualization of RNAseq data of yeast Glaciozyma antarctica PI12

Safinah Sharuddin[1], Nora Muda[2], Nazalan Najimudin[3], Abdul Rahman Othman[4]

[1,2]*Statistics Program, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia*
[3]*School of Biological Sciences, Universiti Sains Malaysia, Penang, Malaysia*
[4]*School of Distance Education, Universiti Sains Malaysia, Penang, Malaysia*

ABSTRACT

Exploration and visualization from the perspective of analytical data are very important nowadays in order to give a comprehensive overview. The purpose of this study is to visualize and explore RNAseq data of yeast Glaciozyma antarctica PI12 in details to facilitate and guide biologists prior to conducting accurate statistical methods using interactive web interface visualization tools such as a dashboard. In this research, we develop a dashboard to visualize a preliminary analysis in an interactive way for data inspection, filtering, and normalization using programming language R with a shiny package. This tool illustrates the shape of data distribution by experimental factor effect for each genomic feature of yeast G. antarctica PI12. Besides, it also enables biologists to see the impact on different cut off values to evaluate and choose suitable types of normalization method to meet the assumptions if required to proceed for the next statistical methods

Keywords: *Glaciozyma antarctica PI12, R, Shiny package, RNAseq, Transcriptome, Web interface, Yeast.*

I. INTRODUCTION

The rapid development of sequencing technology has led us to obtain genomic features such as the whole transcript easier and faster. In line with the development of this technology, the method for analyzing data is increasingly complicated. In the 1990s, microarray technologies have been popular. Subsequently, and until now, the RNA sequencing (RNAseq) has replaced microarray technologies in genomic research [1]-[4]. RNAseq is a next-generation sequencing (NGS) procedure to provide a comprehensive view of the entire transcriptome, which is all set of molecules such as mRNA, rRNA, and other RNAs that are not encoded (noncoding RNA) [5]. Thus, RNAseq data are a type of high dimensional data that can be interpreted as multiple measurements from each sample. In other words, each sample would have gene expression values for thousands of genes. Therefore, data exploration and data visualization are very important in order to understand the nature of the data before conducting statistical analysis. In addition, the diversity of preprocessing methods are introduced such as data transformation and data normalization RNAseq forcing biologists to choose the right option to achieve the objectives of the study. Then, visualization approaches are developed to determine the 'best' way to guide biologists to perform suitable statistical analysis that fulfills the assumptions for each statistical test. Visualization techniques are used to minimize the chance of losing important information (when subsetting the data) [6]. This issue motivates us to develop a web application framework that allows this situation to be more clearly seen by transforming exploratory data analysis in an interactive visualization way using a dashboard. Visualization using a dashboard can be used for purposes of displaying data, diagnostic checking for hypothesis testing or statistical modeling methods (single comparison), and comparing statistical methods (multiple comparisons).

The main purpose of the sample-level analysis is to visualize a sample pattern based on gene expression data summarized from RNA sequencing experiments. It aims to summarize the information provided by a large group of genes into several variables that are easier to manage and then classify samples based on their gene expression profile [15] This can be done by applying multivariate techniques to see the complex structure of RNAseq under different experimental conditions. The preprocessing method should be implemented prior to visualizing high-dimensional data using multivariate techniques. This is because most high-dimensional data exploration such as clustering methods will work better after data transformation and filtering. Clustering samples using data transformation can provide information on sample groups that have similar

expression levels across all genes. It provides a visualization pattern in the gene expression data and investigates the sample similarities after dimension reduction. The issue involved with RNAseq is the mean and variance for RNAseq expression data are said to have a quadratic relationship [16]. Thus, the transformation is necessary to stabilize the variance at higher counts.

In addition, exploration at the gene level can also be visualized using this dashboard. The purpose of exploring data visualization in the gene-level analysis is to provide appropriate biological guidance for the more formal phase of data analysis for each gene. Graphic inspections such as histogram, box-plot, and normal probability plot offer a convenient way to visualize the shape of the underlying distribution. Besides, the interaction plot is to visualize the interaction of factors

II. IMPLEMENTATION

The interactive dashboard for RNAseq visualization for G. antarctica PI12 is developed in Rstudio [7] with shinydashboard package [8]. The dashboard displays results on web interface after running the script in the background server. The dashboard uses RNAseq of Glaciozyma antarctica PI12 to analyze data. Biologists or project teams, who work with this type of data, can apply best practices to develop statistical models. It is also convenient for the user to choose whether a gene-level or sample level analysis that the user wants to investigate.

The visualization to display data exploration analysis with this interactive dashboard replaces static plot with a dynamic plot for the user. It provides a variety of interactive plot according to input entered by the user to evaluate which parameter cutoff is appropriate for the statistical model. It can also extract frequency tables to view measurements of raw data from machine sequences, bar plots to view forms of data distribution as well as support from statistical summary information for each gene. In addition, a dendrogram plot can also be viewed to see the classification at the sample level.

III. MATERIALS DATA

G. antarctica PI12 yeast is classified as psychrophilic yeast, which means "loving cold" organism. This yeast grows well in a cold temperature environment. The growth temperature study found that the optimum temperature for the yeast is 12oC and the maximum temperature is 22oC in Yeast Potato Dextrose (YDP) medium [9]. It can also grow in cold and freeze stress temperatures that are 0oC and -12oC, respectively [10], [11]. Apart from that, the time factor is also studied concerning G. antarctica PI12 yeast. 6 hour exposure time is regarded as an early-stage or self-adaptation level for yeast reaction towards a certain temperature, while 24 hour, on the other hand, is a late-stage towards yeast reaction in certain temperatures [12]. Therefore, the data used in this study take into account both factors involved. In addition, the data used in this study is the experimental result after undergoing sequencing process using NGS technology known as RNAseq data. As in [12], it also uses RNAseq data for G. antarctica PI12 but only for low temperature while the data used in this study include low and high temperatures.

Let say RNAseq experiment is a set of N RNA samples. Each set of RNA samples have G genes. There were G= 7853 annotated genes mapped to whole-genome yeast G. antarctica PI12 [13]. As a result, the measurement is in the form of count read mapped. Thus, the input data is a G x N matrix. RNA samples are usually associated with various treatment conditions. In this study, mRNA is extracted from the yeast, G. Antarctica in the rich growth medium, which is Yeast Potato Dextrose (YPD) under different factors that are Temperature and Time. The study design has two factors namely, the Temperature, which has five levels (-12oC, 0oC, 120C, 16oC, 20oC), and Time, which has two levels (6 hours, 24 hours).

IV. RESULTS

A. Input Data

We illustrate the user interface summarizing the input data using the RNA-seq dataset of yeast Glaciozyma antarctica PI12 as in Fig. 1.

Fig. 1 shows an interactive table using DT package in R programming [14]. Table row represents gene ID and column represents experimental condition (Sample). These interactive tables allow the user to search which gene ID that interests him/her. Moreover, the user can also sort the gene expression values from high to low according to samples

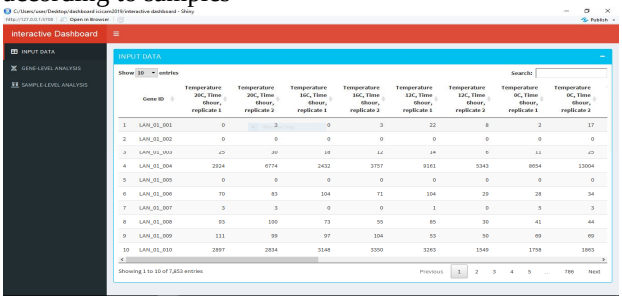

Fig. 1 Web interface for input data panel

B. Sample level analysis

The analyses under sample-level analysis tab can be illustrated as in Fig. 2. It shows the plot before (left) and after data filtering (right). From this plot, we can see that the data distribution has two peaks and is classified as a bimodal distribution because the data tell us the nature

of raw data for RNAseq is bimodal distribution. The right peak distribution indicates the density for low expression value is very high and forms the two-peak distribution. The plot can guide the user to develop statistical modeling for this yeast using RNAseq data at the sample level.

Furthermore, the plot also guides the user to choose whether to use raw data or transformed data for hypothesis testing. For example, in the case of hypothesis testing using Analysis of Variance (ANOVA), one of the assumptions is to rely on the normality of a sample. To test whether the underlying distribution is normal or at least symmetric, this can be done by displaying the distribution graphically. Thus, in the case of RNAseq data, the raw data are not suitable for this analysis. Therefore, data transformation must be conducted if a user wants to use the ANOVA method. When we choose data transformation, the data must be filtered in order to have a symmetrical shape of a distribution. Interestingly, this dashboard offers to visualize the shape of data distribution when changing the different filtering cut off. The user can also use the raw data by applying nonparametric methods to hypothesis testing, which is an alternative to parametric approaches.

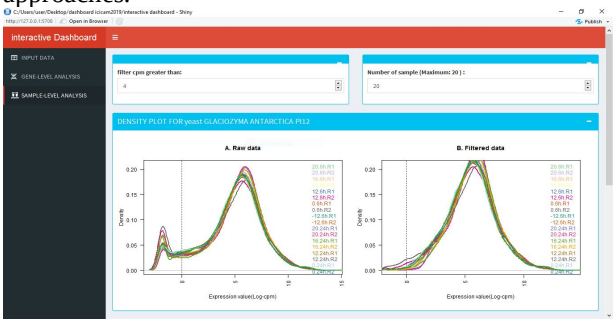

Fig. 2 Web interface for sample level tab panel before and after transformation

The second plot, which is a dendrogram plot, is used to form grouping samples with information from annotated genes of yeast G. antarctica PI12. To visualize in this dashboard, we choose to use the transformed data using the edgeR package [17]. The measurement used in this package is count per million (CPM). Then, the read counts (in CPM) must be filtered out. This is due to the fact that a gene must be expressed at some minimal levels before it can be translated into protein. Therefore, data filtering is necessary to ensure that genes can be considered to be of biological importance. In a statistical point of view, consistent counts do not provide enough statistical evidence to be evaluated as significantly differentially expressed genes. Moreover, filtering out genes that are expressed at low levels prior to differential expression analysis can also reduce the severity of the correction and may improve the power of detection [18].

Fig. 3 illustrates a dendrogram plot using hierarchical clustering with multiscale bootstrap [19] for clustering sample analysis in this dashboard. The output of the dendrogram plot proves that when using the transformed data it will be more accurate to look at the approximate unbiased (AU) p-values from multiscale bootstrap (red values).

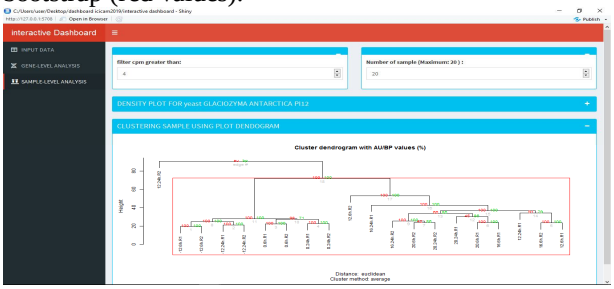

Fig. 3 Web interface for sample level tab panel for sample clustering using dendrogram plot

C. Gene level analysis

Fig. 4 shows the web interface for the visualization of gene-level analysis. On the upper right-hand side of the page, it allows the user to select data of which gene that the user wants to analyze using a drop-down button. On the upper left-hand side of the page, it shows a summary of statistics that explains the nature of the data as well as some information related to the measure of location and measure of spread. To illustrate the summary of statistics graphically, the bar plot and line plot appear below the gene ID box. RNAseq data represents the number of reads (expression values) mapped to a particular gene region under certain conditions, thus, it is appropriate for the plot bar to represent expression values count (counts) by treatment conditions. From this plot, the user can see the selected gene appears more or less at which treatment condition. It also provides a visualization of the shape of data distribution for each gene. The second plot is the line plot to see the relationship between the first variable, temperature and time variable. It can also detect if there is any interaction effect involved in these two variables. The increase and decrease in the reads count of the genes in different conditions can also be graphically explained using this plot

Fig. 4 Web interface for the gene-level tab panel

V. DISCUSSION AND CONCLUSION

Data exploration is the most important part according to a statistical perspective prior to conducting a formal statistical analysis. Therefore, the visualization technique can help us understand the data. As described in the implementation section, the visualization for data exploration purposes using the dashboard is only for displaying and diagnostic checking of RNAseq for yeast G. Antarctica PI12. It will then be updated for users to use their own data from the same technology. It can also be extended to data from various areas such as finance, high-resolution imaging, and website analysis that have multidimensional forms of data (high-dimensional data). Not only for that, the dashboard will then be updated by incorporating the data exploration section for comparison purposes such as hypothesis testing and statistical modeling.

Currently, the application is only hosted in a controlled environment and can be shared with project teams using shiny-server open source. The preliminary analysis of interactive visuals using a web-based framework assists the project team in the selection of downstream analysis for gene expression. This application is not accessible through internet as it is still in a prototype stage since the RNAseq data for this yeast has not been released for the high-temperature condition. In addition, the dashboard will also be published online hosted by the shiny cloud after all the statistical development is completed and it is accessible via the website address.

REFERENCES

[1] M. F. Rai, E. D. Tyksen, L. J. Sandell and R. H. Brophy (2018, January). "*Advantage of RNA-Seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears***"**. J. Orthop. Res. [Online]. 36(1). pp. 484-497. https://www.ncbi.nlm.nih.gov/pubmed/28749036

[2] J. Wang, D. C. Dean, F. J. Hornicek, H. Shi and Z. Duan. (2019, January). "*RNA sequencing (RNA-seq) and its application in ovarian cancer*." Gynecologic Oncology. [Online]. 152(1). pp. 194-201. https://www.sciencedirect.com/science/article/pii/S0090825818312836

[3] K. Bashir, A. Matsui, S. Rasheed and M. Seki. (2019, May). "*Recent advances in the characterization of plant transcriptome in response to drought, salinity, heat and old stress.*" F1000Research. [Online]. 8(658). https://doi.org/10.12688/f1000research.18424.1

[4] P. Gao, D. Xiang, T. D. Quilichini, P. Venglat, P. K. Pandey, E. Wang, C. S. Gillmor and R. Dalta. (2019, March). "*Gene expression atlas of embryo development in Arabidopsis. Plant Reproduction*." [Online]. 32(1). pp. 93-104. https://link.springer.com/article/10.1007%2Fs00497-019-00364-x

[5] R. Hrdlickova, M. Toloue and B. Tian. (2016, May). "*RNA-Seq methods for transcriptome analysis. WIREs RNA"*. [Online]. 8. pp. e1364. https://doi.org/10.1002/wrna.1364

[6] S. Liu, J. McGree, Z. Ge and Y. Xie. "*Computational and statistical methods for analyzing big data with applications***"**. San Diego: Academic Press, 2016.

[7] Rstudio (2012). Rstudio: "*Integrated development for R"* (Version 3.5.2). [Computer software]. Boston, MA. Available: http://www.rstudio.org/

[8] W. Chang and B. B. Ribeiro. (2018). "*shinydashboard: Create dashboard with 'Shiny'*." R package version 0.7.1. [Computer Software]. Available: https://CRAN.R-project.org/package=shinydashboard

[9] Boo et al (2013, March). "*Thermal stress response in Antarctic yeast Glaciozyma antarctica PI12, characterized by real-time quantitative PCR. Polar Biology"*. [Online]. 35(3). pp. 381-389. Available: https://link.springer.com/article/10.1007/s00300-012-1268-2

[10] Bharudin et al. (2019, June). "*Unraveling the adaptation strategies employed by Glaciozyma antarctica PI12 on Antarctic sea ice***"**. Marine Environmental Research [Online]. 137. Pp. 169-176. Available: https://www.sciencedirect.com/science/article/pii/S0141113618300242

[11] C. M. V. L. Wong, S. Y. Boo, C. L. Y. Voo, N. Zainuddin, N. Najimudin. (2019, March). "*A comparative transcriptomic analysis provides insights into the cold-adaptation mechanisms of a psychrophilic yeast, Glaciozyma antarctica PI12***"**. Polar Biology. [Online]. 42(3). pp. 541-553. Available: https://link.springer.com/article/10.1007/s00300-018-02443-7

[12] J. S. P. Koh, C. M. V. L. Wong, N. Najimudin, N. M. Mahadi. (2019, June). "*Gene expression patterns of G. antarctica PI12 in response to cold, and freeze stress. Polar Science***"**. [Online]. 20 (1). pp. 45-54. Available: https://www.sciencedirect.com/science/article/pii/S1873965218301464

[13] M. Firdaus-Raih et al. (2018, January). "*The Glaciozyma Antarctica genome reveals an array of a system that provides sustained responses towards temperature variations in a persistently cold habitat"*. Plos One. [Online]. 13(1). Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0189947

[14] Y. Xie, J. Cheng and X. Tan. (2018). DT: "*A wrapper of the JavaScript library 'DataTables'"*. R package version 0.5. [Computer Software]. Available: https://CRAN.R-project.org/package=DT

[15] D. Amaratunga, J. Cabrera and Z. Shkedy. "*Exploration and analysis of DNA microarray and other high-dimensional data***"**. New Jersey: John Wiley & Sons, 2014, ch. 10.

[16] D. J. McCarthy, Y. Chen and G. K. Smyth. (2012, May). "*Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation*." Nucleic Acids Res. [Online]. 40 (10). pp. 4288-4297. Available: https://www.ncbi.nlm.nih.gov/pubmed/22287627

[17] M. D. Robinson, D. J. McCarthy and G. K. Smyth. (2010, Jan). "*edgeR: a bioconductor package for differential expression analysis of digital gene expression data***"**. Bioinformatics. [Online]. 26(1). pp. 139-140. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/

[18] D. Risso, K. Schwartz, G. Sherlock and S. Dudoit. (2011). "*GC-Content normalization for RNA-seq data. BMC Bioinformatics*." [Online]. 12(1). pp. 480. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-480

[19] R. Suzuki and H. Shimodaira. (June, 2006). "*Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics***"**. [Online]. 22(12). pp. 1540-1542. Available: https://doi.org/10.1093/bioinformatics/btl117