Indian Journal of Science and Technology, *Vol 8(12), 70632, June 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Statistical Analysis of Differential Expression Level of Genes in *Glaciozyma Antarctica PII2*

Nurul Nadia Zulkefri[1], Nora Muda[1]*, Mohd Nazalan Mohd Najimudin[2], Nor Muhammad Mahadi[3,4], Abdul Munir Abdul Murad[3] and Nursyafiqi Zainuddin[2]

[1]School of Mathematical Sciences, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia; nadiazulkefri@gmail.com, noramuda@ukm.edu.my

[2]School of Biological Sciences, Universiti Sains Malaysia, 11800 Georgetown, Penang, Malaysia; nazalan@usm.my, nursyafiqiccb@gmail.com

[3]School of Bioscience and Biotechnology, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia; normm@pc.jaring.my, munir@ukm.my

[4]Malaysia Genome Institute, Ministry of Science, Technology and Innovation, 43000 Kajang, Selangor, Malaysia;

Abstract

The synthesis of small RNA (sRNA) were extracted from *Glaciozyma Antarctica PII2* (*G. Antarctica*) yeast using the Next-Generation Sequencing (NGS) technology. Statistical approach is used in this study to analyze sRNA from *G. Antarctica* in order to increase the production of this yeast given in biological conditions such as temperature, time and medium of growth. Hence, this study uses the analysis of variance (ANOVA) with *F*-test statistic (F_{ANOVA}) and shrinkage *F*-test (F_S) to analyze the NGS data in identifying factors affecting the differential expression level of genes from *G. Antarctica*. F_{ANOVA} statistics are computed on a gene-by-gene basis from the residual sum of squares (SSE). Whereas F_S-test refers to the shrinking of variance estimators from the variance estimators of the gene-by-gene ($\hat{\sigma}_g^2$) *F*-value obtained from ANOVA. Then, the analysis results between F_{ANOVA}-test and F_S-test are compared in order to identify which statistical test is best in analyzing significantly differentially expressed gene based on accuracy value and area under the Receiver Operator Characteristics (ROC) curve. The statistical test with higher accuracy value and has a larger area under the ROC curve is the best statistical test. We found that both F_{ANOVA} and F_S tests show that the majority of genes that are significantly differentially expressed are most affected by the main effects temperature (A) and time (B) and the interaction effect between temperature and time (AB). As for the best test, we found that F_S-test is the best statistical test compared to F_{ANOVA}-test in this study of identifying significantly differentially expressed genes in *G. Antarctica*.

Keywords: ANOVA Model, *F* Statistics, Shrinkage Estimator, Yeast

1. Introduction

The seven years since the introduction of the Next-Generation Sequencing (NGS) technology have seen a major transformation in the way scientists extract genetic information from biological systems, revealing limitless insight in the analysis of transcriptome data. Transcriptome is a set for all the molecules of RiboNucleic Acid (RNA) that consists of Messenger RNA (mRNA), ribosome RNA (rRNA), transport RNA (tRNA) and small RNA (sRNA). In this study, the sRNA is extracted from the *G. Antarctica* yeast using the NGS technology.

Majority of studies are primarily focus on the diversity of the yeasts from Antarctica and its potential in the field of biotechnology. The *Glaciozyma Antarctica* yeast, also known as *G. Antarctica* is widely used in the application of biotechnology, in areas such as food technology and pharmaceutical[7]. The continent of Antarctica which occupies an area of 14 million square kilometers is a major cold habitat, of which about 99% is covered by ice and snow[9]. Apart from being very cold, this continent is considered to be very extreme habitat due to the fact that it is also the driest, windiest and iciest of all known habitats of the world with high solar radiation at least during summer season[10].

*Author for correspondence

The microorganisms that thrive in the extreme environment of Antarctica are cold loving and are referred to as psychrophiles. Psychrophillic yeasts have an optimum temperature for growth about 15°C or lower, a maximum up to 25°C but are still capable of growing at 0°C or below[2]. Yeasts are versatile group of eukaryotic microorganisms which are heterogeneous in the nutritional abilities and are capable of surviving in a range of habitats such as in deep sea, moist and uneven surfaces including polluted waters[16].

2. Data and Materials

This study involves three factors: A: temperature (-12°C, 0°C, 12°C), B: time (6 hours, 12 hours) and C: medium of growth (DOC, YPD) to test the differential level of gene expression in *G. Antarctica* given in different biological conditions. DOC refers to drop out concentration whereas YPD refers to yeast peptone dextrose. Figure 1 shows the classification of data for this study.

Many experiments involve in the study of effects of two or more factors. In general, factorial designs are most efficient for this type of experiment. A factorial design means that each complete trial or replication of the experiment all possible combinations of the levels of the factors are investigated. When factors are arranged in a factorial design, they are often said to be crossed[14]. This study uses $3\times2\times2$ factorial design. The observations in a factorial experiment can be described by the following model

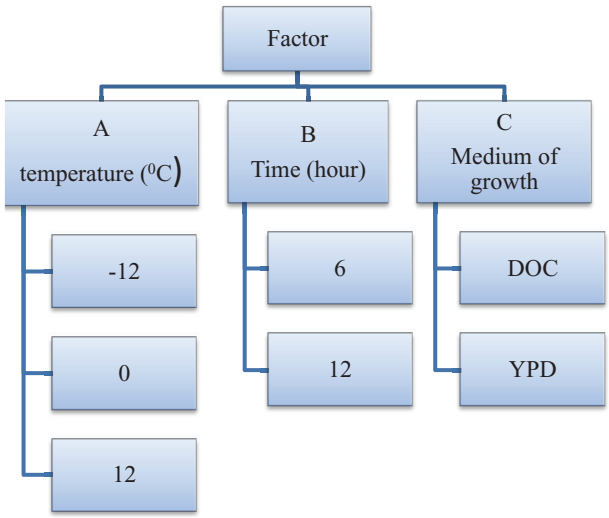

Figure 1. Classification of Data.

$$\begin{aligned} y_{ijk} &= \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} \\ &\quad + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl} \\ &\quad i = 1,2,...,a \\ &\quad j = 1,2,...,b \\ &\quad k = 1,2,...,c \\ &\quad l = 1,2,...,n \end{aligned} \tag{1}$$

where μ is the overall mean effect, τ_i is the effect of ith level of temperature factor, β_j is the effect of the jth level of time factor, γ_k is the effect of the kth level of medium of growth factor, $(\tau\beta)_{ij}$ is the effect of interaction between temperature and time, $(\tau\gamma)_{ik}$ is the effect of interaction between temperature and medium of growth $(\beta\gamma)_{jk}$ is the effect of interaction between time and medium of growth, $(\tau\beta\gamma)_{ijk}$ is the effect of interaction between temperature, time and medium of growth, and ε_{ijkl} is the random error component.

3. Methods

3.1 ANOVA Test (F_{ANOVA})

A procedure for constructing statistical tests by partitioning the total variance into different sources of variance. Statistical tests towards these sources of variance are then tested whether they are statistically significant or otherwise. In general, the purpose of analysis of variance is to test whether the means between treatments are significantly different or not.

With the assumption that the model in equation (1) is adequate and that the error terms ε_{ijkl} are normally and independently distributed with constant variance σ^2, then each of the ratios of mean squares of each treatment to the mean square of error MS_A/MSE, MS_B/MSE, MS_C/MSE, MS_{AB}/MSE, MS_{AC}/MSE, MS_{BC}/MSE and MS_{ABC}/MSE is distributed as F[14].

Hence, the variance estimators for each gene $g = 1,2,...,G$ in *G. Antarctic* yeast are

$$\hat{\sigma}^2_g = \left(MSE_{Error}\right)_g = \left(\frac{SSE}{abc(n-1)}\right)_g \tag{2}$$

Analysis of variance using F_{ANOVA} test statistic was conducted on each gene (gene ID) $g = 1,2,...,G$ of the *G. Antarctic* yeast. This allowed us to identify genes that are

significantly differentially expressed given in different treatments and factors that affect the level of gene expressions.

3.2 *F*-Shrinkage Test (F_S)

Cui et *al.*[4] constructed improved estimators of variance from an ensemble of individual variance estimators by shrinking them toward their common corrected geometric mean. Shrinkage estimators pull individual error estimates toward shrinking targets, with the amount of shrinkage depending on the variability of individual error estimates[3].

The shrinkage estimators effectively pool these estimates when the individual variance estimates are similar where this indicates homogeneity. In addition, the shrinkage estimator gives greater weight to the gene specific contributions when the individual variance estimates are dispersed, indicating heterogeneity. The result is in the expression as in the equation (3)

$$\tilde{\sigma}_g^2 = \left(\prod_{g=1}^{G} \left(X_g / \nu \right)^{1/G} \right) B \times \exp\left[\left(1 - \frac{(G-3)V}{\sum \left(\ln X_g - \overline{\ln X_g} \right)^2} \right)_+ \times \left(\ln X_g - \overline{\ln X_g} \right) \right] \quad (3)$$

where X_g is the residual sum of squares (*SSE*), $\overline{\ln X_g} = \frac{1}{G}\sum \ln\left(X_g \right)$, and $B = \exp(-m)$ is a bias correction. The values of B and V depend on ν degrees of freedom.

Similar to F_{ANOVA}, F_S is also conducted on each gene (gene ID) $g = 1,2,...,G$ of the *G. Antarctic* yeast which allowed us to identify genes that are significantly differentially expressed given in different treatments and factors that affect the level of gene expressions.

3.3 Comparing F_{ANOVA} and F_S

The performance of genes selection can be calculated in terms of specificity and sensitivity. Specificity refers to the true positive rate whereas sensitivity refers to the true negative rate. In the case of binary result which are either genes are significantly differentially expressed and genes are not significantly differentially expressed, the results can be divided into four categories which are True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN)[5].

Table 1. Definition of specificity and sensitivity

Test outcome	Condition	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN
	Sensitivity $= \frac{TP}{TP + FN}$	Specificity $= \frac{TN}{TN + FP}$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{N}} \quad (4)$$

The *F*-test with a higher accuracy value in equation (4) shows a better *F*-test. The perfect method would be a method that produces sensitivity = specificity = 1.

4. Results and Discussion

4.1 Results of F_{ANOVA}

The overall results of F_{ANOVA} analysis found that 214 out of 247 genes are significantly differentially expressed given in different conditions. Table 2 shows the number of genes with factors that are significant in influencing the level of gene expression when the analysis of variance using the F_{ANOVA} is conducted.

From Table 2, we found that the majority of genes that are significantly differentially expressed are of main effects temperature (A), time (B) and interaction effect temperature and time (AB).

Table 2. Comparison between the number of genes with factors that are significant in influencing the level of gene expression with F_{ANOVA}

Factors	Number of significantly expressed genes
A: Temperature	104
B: Time	164
C: Medium of growth	8
AB	118
AC	6
BC	7
ABC	6

4.2 Results of F_S

The overall results of F_S analysis found that 220 out of 247 genes are significantly differentially expressed given in different conditions. Table 3 shows the number of genes with factors that are significant in influencing the level of gene expression when the analysis of variance using the F_S is conducted.

From Table 3, we found that the majority of genes that are significantly differentially expressed are of main effects temperature (A), time (B) and interaction effect temperature and time (AB).

4.3 Comparing F_{ANOVA} and F_S

From Table 4, we can conclude that both methods of the *F*-test statistics shows somewhat similar results where we can see that genes expression are most influenced by the main effects time (B) and temperature (A) and the interaction effect between temperature and time (AB) and are less affected by medium of growth (C) and the interaction effects between temperature and medium of growth (AC), time and medium of growth (BC) and temperature, time and medium of growth (ABC).

To identify which *F*-test is best, we will apply the method in Table 1, Table 5 and Table 6 show the values of TP, FP, FN and TN that can be obtained by using the graph of *p*-value distribution from Figure 2 and Figure 3.

From Table 5 and Table 6, we can calculate the accuracy value from equation (4) for both F_{ANOVA} and F_S test statistics respectively as follows

Table 3. Comparison between the number of genes with factors that are significant in influencing the level of gene expression with F-shrinkage test

Factors	Number of significantly expressed genes
A: Temperature	95
B: Time	162
C: Medium of growth	6
AB	112
AC	5
BC	7
ABC	4

Table 4. Comparison between the number of genes with factors that are significant in influencing the level of gene expression with F_{ANOVA} and F_S

	F_{ANOVA}		F_S	
	No of genes	%	No of genes	%
A : Temperature	104	42.11	95	38.46
B : Time	164	66.40	162	65.59
C : Medium of growth	8	3.24	6	2.43
AB	118	47.77	112	45.34
AC	6	2.43	5	2.02
BC	7	2.83	7	2.83
ABC	6	2.43	4	1.62

Table 5. TP, FP, FN and TN values from F_{ANOVA} test statistic

Test outcome	Condition	
	Differentially expressed	Not differentially expressed
Differentially expressed	136	5
Not differentially expressed	58	48
	Sensitivity = 0.70	Specificity = 0.91

Table 6. TP, FP, FN and TN values from *F*-shrinkage test statistic

Test outcome	Condition	
	Differentially expressed	Not differentially expressed
Differentially expressed	144	5
Not differentially expressed	49	49
	Sensitivity = 0.75	Specificity = 0.91

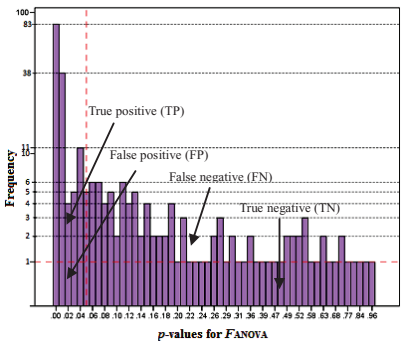

Figure 2. *p*-values distribution from F_{ANOVA} test statistic.

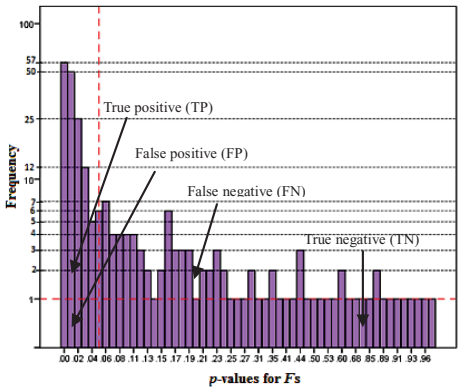

Figure 3. p-values distribution from F_S test statistic.

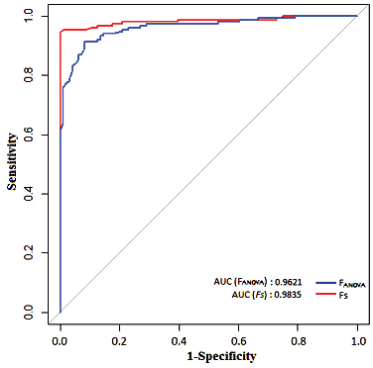

Figure 4. Receiver operating characteristic (ROC) curve.

$$\text{Accuracy } (F_{\text{ANOVA}}) = \frac{136+48}{247} = 0.74$$

$$\text{Accuracy } (F_S) = \frac{144+49}{247} = 0.78$$

Since Accuracy $(F_S) = 78\% >$ Accuracy $(F_{ANOVA}) = 74\%$, we therefore select F_S as the best test statistic for this study.

Figure 4 shows the relationship between genes that are correctly identified as differentially expressed (TPR) on the y-axis and genes that are falsely identified as differentially expressed (FPR) on the x-axis. The Area Under the Curve (AUC) shows the ability of the overall tests to differentiate genes that are significantly differentially expressed and the genes that are not significantly differentially expressed. A perfect test would be a test that has the AUC that is equal to 16. This means that AUC that are closer to 1 results as the best test statistic. From Figure 4, we found that $\text{AUC}_{F_S} = 0.9835 > \text{AUC}_{F_{ANOVA}} = 0.9621$, this concludes that FS test is the best test statistic when compared to F_{ANOVA} test.

Based on the Figure 5 (a), the genes are most significantly differentially expressed at temperature -12°C, while in Figure 5 (b), the genes are most significantly differentially expressed when exposed at time 12 hours. The genes are majority significantly differentially expressed in medium of growth as shown in Figure 5 (c). When measuring in interaction effect, the *F*-shrinkage shows that majority of the expression level of genes are influenced by the interaction effect of temperature and time (AB) at 0°C and 12 hours as shown in Figure 5 (d). Figure 5 (e) shows the interaction effect of temperature

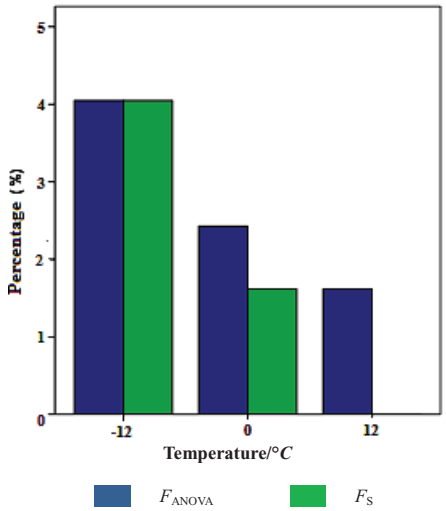

Figure 5.(a) Bar graph showing the percentage of the significantly differential expressed gene based on temperature when F_{ANOVA} and F_S are conducted.

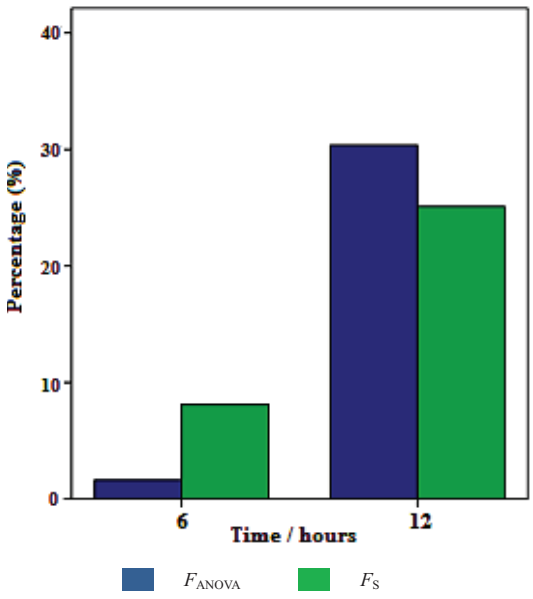

Figure 5.(b) Bar graph showing the percentage of the significantly differential expressed gene based on time when F_{ANOVA} and F_S are conducted.

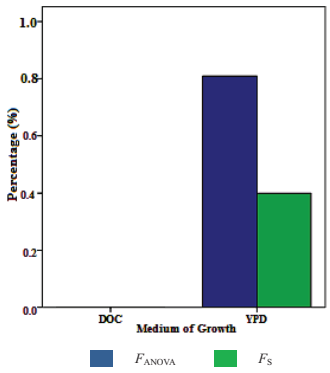

Figure 5. (c) Bar graph showing the percentage of the significantly differential expressed gene based on medium of growth when F_{ANOVA} and F_S are conducted.

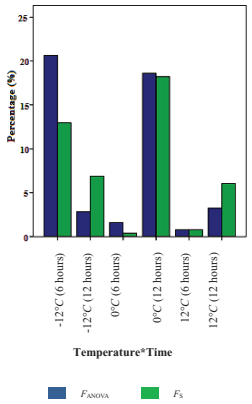

Figure 5. (d) Bar graph showing the percentage of the significantly differential expressed gene on the effect of interaction between temperature and time when F_{ANOVA} and F_S are conducted.

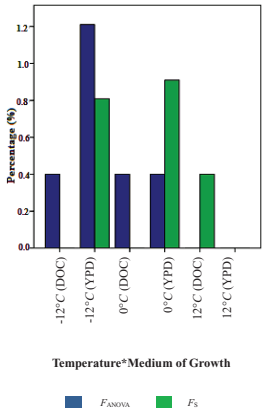

Figure 5. (e) Bar graph showing the percentage of the significantly differential expressed gene on the effect of interaction between temperature and medium of growth when F_{ANOVA} and F_S are conducted.

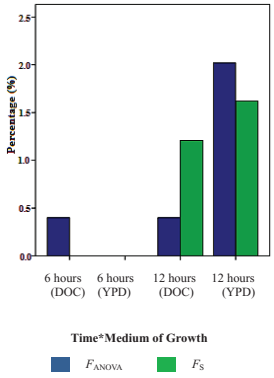

Figure 5. (f) Bar graph showing the percentage of the significantly differential expressed gene on the effect of interaction between time and medium of growth when F_{ANOVA} and F_S are conducted.

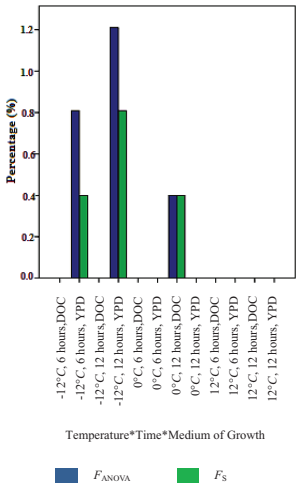

Figure 5. (g) Bar graph showing the percentage of the significantly differential expressed gene on the effect of interaction between temperature, time and medium of growth when F_{ANOVA} and F_S are conducted.

and medium of growth (AC) and the most genes are significantly differentially expressed at temperatures -12°C and 0°C with both using the same medium of growth, which is the YPD medium. The interaction effect of time and medium of growth has shown that the expression level of genes increases which shows a significant expression level when the time given is 12 hours and using the YPD medium as shown in Figure 5 (f). The optimum temperature, time and medium of growth that most of the genes significantly differentially expressed are at temperature -12°C with 12 hours of exposure and using the YPD as the medium of growth as shown in Figure 5 (g).

5. Discussion

Since F_S results as the best test statistic when compared to F_{ANOVA}, we therefore will make a conclusion on the optimum temperature, time and medium of growth that affect the significant genes expression level based on F_S test.

From our analysis, we found that the significantly differentially expressed genes level for the temperature factor are most affected at -12°C, less affected at 0°C and least affected at 12°C (Figure 5(a)). For the time factor, the significantly differentially expressed genes levels are most affected when the time given for the experiment is 12 hours (Figure 5(b)). Moreover, majority of genes were significantly differentially expressed when YPD type of medium of growth is given (Figure 5(c)).

For the interaction effects, F_S test shows that majority of the expression level of genes are influenced by the interaction effect AB at 0°C and 12 hours (Figure 5(d)). Other than that, the interaction effect of AC shows that most genes are significantly expressed at temperatures -12°C and 0°C with both using the same medium of growth, which is the YPD (Figure 5(e)). In addition, the interaction effect of BC has shown that the expression level of genes increases which shows a significant expression level when the time given is 12 hours and using the YPD medium (Figure 5(f)). Lastly, for the three factors interaction effect ABC, the optimum temperature is at -12°C, with the time given 12 hours and using the YPD as the medium of growth (Figure 5(g)).

6. Conclusion

Both *F*-test statistics show that the majority of genes in *G. Antarctica* are most affected by the main effects temperature (A) and time (B), and the effect of interaction between temperature and time (AB). Most genes are significantly expressed when given in those treatments.

However, F_S gives a better performance in identifying the significantly differentially expressed genes when compared to F_{ANOVA} where the accuracy value for F_S is greater than the accuracy value of F_{ANOVA}. In addition, F_S also has a larger AUC and that are closer to 1 compared to F_{ANOVA}.

Since F_S results as the best test statistic when compared to F_{ANOVA}, we therefore will make a conclusion on the optimum temperature, time and medium of growth that affect the significant genes expression level based on F_S test and summarizes all the optimum biological conditions for the significantly differentially expressed genes in *G. Antarctica* yeast in Table 7.

Table 7. The optimum biological conditions for the significantly differentially expressed genes in *G. Antarctica*

Factor	Optimum level(s)
A : Temperature	-12°C
B : Time	12 hours
C : Medium of growth	YPD
AB	(0°C , 12 hours)
AC	(-12°C , YPD) & (0°C, YPD)
BC	(12 hours, YPD)
ABC	(-12°C, 12 hours, YPD)

7. Acknowledgement

This research was supported by the Research Grant DPP-2014-008.

8. References

1. Amaratunga D, Cabrera J. Exploration and analysis of DNA microarray and protein array data. New York: John Wiley & Sons; 2004.
2. Arthur H, Morton H, Watson K. Thermal adaptation in yeast: obligate psychrophiles are obligate aerobes, and obligate thermophiles are facultative anaerobes. Journal of Bacteriology. 1978; 2:815–7.
3. Cui X, Churchill AG. A review of statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003; 4:210–21.
4. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill AG. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics. 2005; 6(1):59–75.
5. Draghici S. Data Analysis Tools for DNA Microarrays. Florida: Chapman & Hall; 2003.
6. Grandi G. Genomics, Proteomics and Vaccines. Chichester: John Wiley & Sons; 2004.
7. Hamid B, Rana RS, Chauhan D, Singh P, Mohiddin FA, Sahay S, Abidi I. Psychrophilic yeasts and their biotechnological applications. Afr J Biotechnol. 2014; 13(22):2188–97.
8. Ho J, Stefani M, Remedios CG, Charleston AC. Differential variability analysis of gene expression and its application to human diseases. Bioinformatics. 2008; 24:390–8.
9. Holdgate MW. Terrestrial ecosystems in the Antarctic. Phil Trans R Soc Lond B. 1977; 279(963):5–25.

10. Hopkins DW, Sparrow AD, Novis PM, Gregorich EG, Elberling B, Greenfield LG. Controls on the distribution of productivity and organic resources in Antarctic dry valley soils. Proc R Soc B 2006; 273:2687–95.
11. Jeffery IB, Higgins DG, Culhan AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics. 2006; 7:359–77.
12. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The Contributions of Sex, Genotype and Age to Transcriptional Variance in Drosophila Melanogaster. Nat Genet. 2001; 29(4):389–95.
13. Kvam V, Liu P. A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. Am J Bot. 2012; 99(2):248–56.
14. Montgomery DC. Design and analysis of experiments. Arizona: John Wiley & Sons; 2009.
15. Pounds S. Computational enhancement of a shrinkage-based analysis of variance f-test proposed for differential gene expression analysis. Biostatistics. 2007; 83:505–6.
16. Shivaji S, Prasad GS. Antarctic yeasts: biodiversity and potential. In: Satyanarayana T, Kunze G, editors. Yeast biotechnology: diversity and applications. New Delhi: Springer; 2004.
17. Tong T, Wang Y. Optimal shrinkage estimation of variances with applications to microarray data analysis. J Am Stat Assoc. 2007; 102(477):113–215.
18. Yang J, Casella G, McIntyre LM. Generalized shrinkage F-like statistics for testing an interaction term in gene expression analysis in the presence of heteroscedasticity. BMC Bioinformatics. 2011; 12(427):1471–2105.