Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 73(3) 2019 251–255

Microarray Data Expression Study for Better Identification of Differentially Expressed Genes

Neeraj Budhlakoti[1], Ravi Shankar[2], Anil Rai[1], Rajeev Ranjan Kumar[1] and A.R. Rao[1]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]CSIR-Institute of Himalayan Bioresource Technology, Palampur

Received 12 March 2018; Revised 03 April 2019; Accepted 28 August 2019

SUMMARY

As whole genome sequencing technologies led to a boom in the availability of genetic information. Microarray has been emerged as one of the most powerful tool in the field of transcriptomics. It analyses the expression of gene in a cell or tissue in given moment of time. It allows the scientist to understand the molecular mechanism in normal and modified conditions. It has provided scientists with a tool to investigate the structure and activity of genes on a wide scale instead of earlier traditional methods. It further permits scientists to understand the molecular mechanisms underlying normal and dysfunctional biological processes. Microarray technology could speed up the screening of thousands of DNA and protein samples simultaneously. This review article focusses on the variation among the results of existing methods of microarray date analysis and also suggest some protocol to add more reliability to the result.

Keywords: Microarray, t-test, Modified t-test, Differentially expressed genes, Phylogenetic replicates.

1. INTRODUCTION

The completion of various genome sequencing projects has resulted the rapid fire in the availability of genetic information. In transcriptomics, the emergence of microarray-based technologies allows high-throughput studies of RNA expression in cell and tissue at a given moment of time [5]. Microarray experiments measure the expression of thousands of genes simultaneously. Still today microarray technologies is one of most popular tool in molecular biology as it is different from traditional methods that it is not limited to analysing one gene at a time [6].Molecular Biology research evolves through the development of the technologies used for carrying them out. It is not possible to research on a large number of genes using traditional methods. Traditional methods lacks sensitivity, are extremely time-consuming and cannot account for unculturable organisms. It has several advantages over traditional technologies, as it fast, can produce on a large scale basis, reproducible, lower costs and precise. Applications ranges from the study of gene expression in yeast under different environmental stress conditions [3, 4] to the variation study of gene expression profiles for tumours from cancer patients [1, 7] and also relevant to drug discovery, toxicological research and many more. In addition to this much of enormous scientific potential, it also helps in understanding gene interactions and gene regulation. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the various genetic causes of anomalies. By these reasons microarrays is being used increasingly in pharmaceutical and clinical diagnostics. Basic principle behind the microarray is that it is a hybridization based expression analysis i.e. the base pair 'A' hybridizes to its complementary base 'T' and 'C' to with 'G'. Here many thousands of gene spots are placed on rectangular array with each spot containing a probe unique to particular gene. Expression of each

Corresponding author: Neeraj Budhlakoti
E-mail address: neeraj35669@gmail.com

gene is determined by hybridization signal, as each copy of mRNA is fluorescently labelled [11]. Whenever we have expression data of gene in two or more than two conditions, one always interested to know which gene are differentially expressed i.e., genes for which read count distributions differ among populations. Several methods and tools have been developed to identify differentially expressed genes among the different experimental condition. Among them, limma; a Bioconductor package based on R programming [11, 12] is very popular.

These methods generally identify the differentially expressed genes based on log FC and p-value (Table 1). However, after more than two decades of research still we do not have satisfactory method. There are lots of variation in results of DEG among different methods [2]. It may be due to several reasons, firstly small no replicates for a experimental condition. Secondly, the distribution of expression data may depart from the usual assumptions, which results in poor performance. For example, if we apply two sample t-test, as its performance is based on sample size and assumption that expression intensity is normally distributed, but this always may not be the case. If data departs from this normality assumption the performance of t-test is very poor. Also the other problems with these methods are that it cannot identify the differential expression of genes which have very low expression value in different conditions. But it can happen that some gene have their different expression range and they cannot express beyond that. So this is one of the limitation of current microarray analysis methods.

In a recent study RNA-Seq samples from range of species were used as replicates i.e. they called it as phylogenetic replicates. Such type of design helps in cost-effective RNA-Seq experiments in the field of biodiversity that may involve hundreds of species under a phylogenetic framework [13]. This phylogenetic replicates may provide additional information in case of lack of sufficient samples in microarray study.

2. METHODOLOGY

In this study microarray data has been collected for four different crops rice, Arabidopsis, Soybean and Tomato from GEO-NCBI [13]. Data has been further grouped according to the tissue, as the expression of genes can vary significantly among the tissue for same species. The data we got from NCBI was simply log transformed. We further applied z-score normalization so that expression value to be in the same scale. We have collected the sample within a tissue specific to a crop having same control to get range of expression value for different experimental condition. Here we have collected the sample for controlled condition from different GSE accession for same tissue say it as modified control. So that we can have some wide range of values for control condition which may add some confidence in the result. Then we have applied

Table 1. List of different methods used for microarray analysis

S. No.	Method	Statistical framework	Advantages	Limitations
1	**Student t-test**	$t_g = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Simple and easy to apply	Require strict statistical assumptions
2	**Satterthwaite-welch t-test**	$t_g = \frac{\bar{x}_2 - \bar{x}_1}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ Where $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	Can be used for unequal population variances	Computationally intensive
3	**Paired sample t-test**	$t = \frac{\bar{x}_d}{s_d/\sqrt{n}}$	Easy to calculate	Cannot be used in data coming from different subjects
4	**ANOVA F-test**	$F1 = \frac{(AKT_0 - AKT_1)/v_0 - v_1}{AKT_1/v_1}$	Easy and precise results in case of balanced data	Not suitable for unbalanced data
5	**Wilcoxon rank sum test**	$z_g = \sum_k Rank(x_{gik})$	Applicable in the case even if data does not follow statistical assumptions	Less power of test

the two sample t-test and moderated t-test based on limma [11, 12]. We compared the results of normal experiment for control vs. treatment to with modified control vs. treatment. We found significant difference among the results. Which suggests that still there is a need of improvement in existing microarray data analysis methods. All this procedure could be simply understood through following flow-diagram.

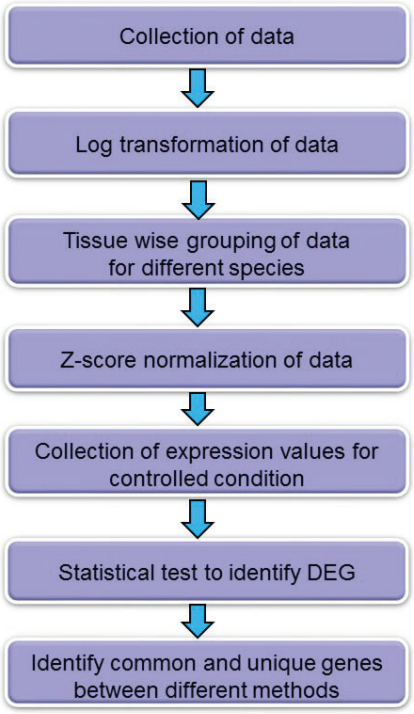

Fig. 1. Basic pipeline of data analysis

Following statistical test has been used for analysis purpose.

2.1 t-test

Let expression value of j^{th} array ($j = 1(1)n$) and g^{th} gene ($g = 1(1)n$) be denoted by x_{gj}. Then t-statistic can be defined as

$$t_{gj} = \frac{\bar{x}_{g2} - \bar{x}_{g1}}{\sqrt{\frac{s_{g1}^2}{n_1} + \frac{s_{g2}^2}{n_2}}}$$

Where $s_{g1}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{m}\left(x_{gj} - \bar{x}_{g1}\right)^2$

and $s_{g2}^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{m}\left(x_{gj} - \bar{x}_{g2}\right)^2$

$\bar{x}_{g1}$ and $\bar{x}_{g2}$ are mean corresponding to each gene for particular condition.

2.2 Modified t-test

In same manner modified t-statistic can be defined as [8]:

$$t_{gj}^{*} = \frac{\hat{\beta}_{gj}}{s_g^{*}\sqrt{v_{gj}}}$$

Follows an approximate t-distribution with d_g degree of freedom and $s_g^{*2} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$, where s_g^{*2} is residual sample variance. The parameter is related through $d_g = f$, $v_g = \frac{1}{n}$, $d_0 = 2v$, $s_0^2 = \frac{a}{(d_0 v_g)}$ & $v_0 = c$ where f, n, v and a are according to Lionnstedt & Speed [9].

2.3 Real data

In this study data has been considered for different crops from GEO-NCBI. The details regarding data is given table below.

Table 2. Details of data used in the study

Crop	Tissue	NCBI -GEO ids
Rice	leaf	GSE53858, GSE7197
	shoot	GSE62124, GSE7475
Arabidopsis	root	GSE41544, GSE48836
	seedling	GSE40574, GSE3874, GSE53621
Soybean	leaf	GSE7108, GSE23128
Tomato	leaf	GSE35618, GSE76332

3. RESULT AND DISCUSSION

In this study we used the two sample t-test and modified t-test. Sample has been analysed on the basis of simple control vs. treatment and modified control vs. treatment for same experiment using above two discussed methodologies. Result of same has been presented and compared in earlier studies (Table 3&4).

Now, it could be easily understood that there are several no. of genes which are contradicting earlier study (Table 3). Differentially expressed genes are simply considered whose p values are less than 0.05. Also it could be understood that there is a significant difference in the result of simple t-test and modified limma based t-test (Table 3). This may be due to the assumption, that it uses before applying any statistical test based on few replicates are very weak. If we apply simple t test, performance becomes more worst when there are one or two replicates in particular conditions.

Table 3. List of DEG for different methods

Crop	Tissue	T-test using general procedure (DEG)	T-test using modified procedure (DEG)	Moderated t-test using general procedure (DEG)	Moderated t- test using modified procedure (DEG)
Rice	leaf	1712	6305	1391	7445
	shoot	6291	9682	6327	10066
Arabidopsis	root	726	343	721	288
	seedling	2021	3150	1927	3171
Soybean	leaf	5283	3908	5534	3273
Tomato	leaf	381	1452	347	1628

Table 4. Comparison of DEGs using two different methods

Crop	Tissue	Common DEG using t-test in both protocol	Common DEG using modified t-test in both protocol	No. of DEG contradicting earlier study
Rice	leaf	1306	1355	6089
	shoot	4887	5194	4871
Arabidopsis	root	161	171	116
	seedling	731	858	2312
Soybean	leaf	2108	2087	1185
Tomato	leaf	168	224	1403

In Table 4 we have summarized the information regarding the overlapping no of genes between different methods. It describes about no of DEG common in t-test and modified t-test using both protocols (control vs. treatment, modified control vs. treatment). It also summarizes the no. of DEG which are contradicting earlier study in different species for particular tissue.

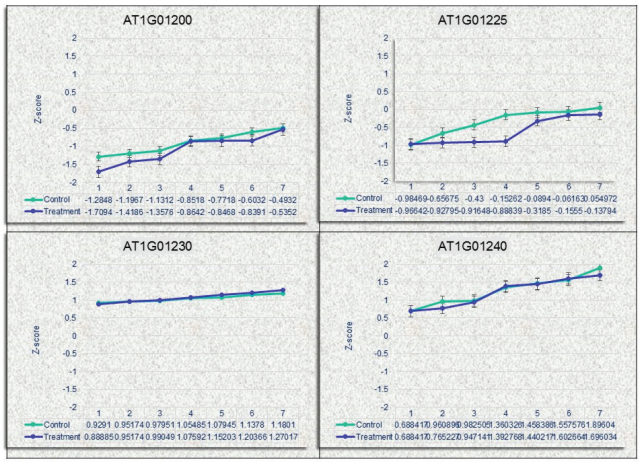

Fig. 2. Expression pattern of few genes depicting restricted expression of genes in *Arabidopsis thaliana*

It can be easily understood that some gene have restricted expression level and they cannot express beyond that (Fig. 2). Then question arises how we could compare the gene expression on global basis. Here expression range for each gene specific to a tissue and crop for controlled conditions has been obtained. It could be further utilized for identifying differentially expressed genes if experiment is conducted in same tissue and species in case of lack of sufficient replicates. Here one can also utilize the concept of phylogenetic replicates. Phylogenetic replicates can be defined as samples of the same tissue from a phylogeny. In this instead of considering only biological replicates, a method based on phylogenetic replicates could be applied to identify the differentially expressed genes under the assumption that same tissue performs same function among the related organism [13].

4. CONCLUSION

Microarrays still offer a number of advantages over other techniques because of the high capacity for multiplexing despite being advent of cheap and accessible next-generation sequencing techniques. In this study we have compared the different methods of expression analysis. We found that there is significant difference among the results. It is discussed previously that every gene has its own expression level range and they cannot express beyond that level. Existing methods does not take care of this, which may result in filtering of some important differentially expressed genes. So there is a need to develop new methodologies which take care of this constraints. Here we suggest to use gene expression data for control condition selected across the different experiment from same tissue and species following similar protocol or one can also

utilize the concept of phylogenetic replicates. In this way we can have wide range of gene expression for controlled condition, which will depict the better picture for identifying DEGs.

5. FUTURE PROSPECTIVE

As a path ahead, new methods could be developed to identify differentially expressed gene which have restricted expression level. Further new methodologies could be developed which does not rely on assumption regarding expression data before applying any statistical test. Also method based on phylogenetic replicates could be used for this purpose.

LIST OF ABBREVIATIONS

GEO : Gene Expression Omnibus

NCBI : National center for Biotechnology Information

FC : Fold change

DEG : Differentially Expressed Genes

RNA :RiboNucleic Acid

Limma : Linear Model for Microarray Analysis

ACKNOWLEDGEMENT

Authors acknowledge the appreciable efforts of anonymous reviewers.

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. Broad (1999). Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, **96**, 6745-6750.

Bokka, S. and Mathur, S.K. (2006). A Nonparametric Likelihood Ratio Test to Identify Differentially Expressed Genes from Microarray Data. *Applied Bioinformatics,* **5**, 267-276.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Her-skowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science,* **282**, 699-705.

DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-685.

Gardiner-Garden, M., Littlejohn, T.G. (2001). A comparison of microarray databases. *Briefings in Bioinformatics*, **2**:143-158.

Georgii, E., Richter, L., Ruckert, U. and Kramer, S. (2005). Analyzing microarray data using quantitative association rules. *Bioinformatics*, **21**, 123-129.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, M., Gaasenbeek, J.P., Mesirov, H., Coller, M. L., Loh, J.R., Downing, M.A., Caligiuri, C.D., Bloomfield, and Lander., E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* **286**, 531-537.

Kuo, W.P., Whipple, M. E., Jenssen, T.K., Todd, R., Epstein, J.B., Ohno-Machado, L., Sonis, S.T. and Park, P.J. (2003) . Microarrays and clinical dentistry. *J. Amer. Dental Assoc.*, **134**, 456-462.

Lonnstedt, I. and Speed, T.P. (2002). Replicated microarray data. *StatisticaSinica*, **12**, 31-46.

Sean, D. and Paul, S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846-1847.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Gen. Mole. Bio.,* **3**, 1-25.

Smyth, G.K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, 397-420.

XunGu. (2016). Statistical detection of differentially expressed genes based on RNA-seq: from biological to phylogenetic replicates. *Briefings in Bioinformatics*, **17**, 243-248.