IOSR Journal of Biotechnology and Biochemistry (IOSR-JBB)
ISSN: 2455-264X, Volume 4, Issue 1 (Jan. - Feb 2018), PP 42-47
www.iosrjournals.org

Analysis of Common Gene Expression Pattern From Human Tuberculosis Microarray Data In The R Pakage

Umar Shittu[1*], Mohammed Yahaya[1], Sunusi Liadi[2], Maryam S. A.[3], Umma Sada[4]

[1*]*Department of Biology, Isa Kaita College of Education Dutsin-ma, Katsina State,.*
[1]*Department of medical microbiology and Parasitology, College of Health Science, Usmanu Danfodiyo University, Sokoto Nigeria.*
[2] *Department of Biology, Isa Kaita College of Education Dutsin-ma, Katsina State, Nigeria.*
[3]*Department of Biology, Isa Kaita College of Education Dutsin-ma, Katsina State, Nigeria.*
[4]*Department of Biology, Isa Kaita College of Education Dutsin-ma, Katsina State, Nigeria.*
**Corresponding Author: Umar Shittu*

Abstract: *Microarray is a collection of microscopic samples commonly from nucleic acids (RNA and DNA) which can be probed with target molecules to produce either gene expression or diagnostic data. Tuberculosis is a bacterial infection usually occurred in human and animal bodies caused by Mycobacterium tuberculosis, this bacteria usually infect the lungs, but it can infect other parts of the body of an organism. The aim of this study is to process the raw data and determine common expression pattern using some tools available for analyzing human TB microarray gene expression data in the Bioconductor R package. The control stimulated samples with phosphate buffer saline (PBS) and experimental unstimulated samples of three different clinical forms of human TB microarray gene expression data, such as latent TB (LTB), pulmonary TB (PTB) and meningealTB (TBM) were collected from GEO-NCBI database and all analysis were performed by using Bioconductor R packages. The results of this study, explore the use of affycoretool for microarray TB image visualization analysis, AffyQCReport tool for TB microarray data quality assessment, GCRMA method for TB microarray data normalization, these pre-processing indicates that the data are of good quality and be use for advance analysis. Hierarchical clustering (hclust) method was used to determine common expression pattern among the three different clinical forms of human TB infection. The finding shows there were more relatedness between expression levels of the arrays from the same clinical group of tuberculosis infection than those arrays with different clinical group and also indicated differences between the arrays. It suggested that hierarchical clustering analysis distinguish different clinical forms of human TB infection. This study recommended that the results generated from these findings can be used in further analysis for detection and control of human TB infection.*

Keywords*: Human tuberculosis, Microarray and Bioconductor R package.*

Date of Submission: 26-01-2018 Date of acceptance: 15-02-2018

I. Introduction

Microarray technology is one of the technologies that contributes toward the development of studies in the field of Molecular Biology. Microarray data are the soft copy microscopic samples commonly collected from nucleic acids (RNA and DNA) which can be probed with target molecules to produce either gene expression or diagnostic data used to determine gene sequence or to identify variations among genes or expression. This technology deals, with the collections of DNA or RNA samples from microarray experiments and uses these samples to measure the expression levels of large numbers of genes or the whole genome of an organism at a time. Microarray experiment involves the growing of an organism, collecting of tissues, extraction of RNA or DNA, hybridization and scanning process to generate microarray data. (Miller *et al.*, 2009). The technology also can use the combination of two complementary single-stranded DNA or RNA to form a single double-stranded through base pairing to a very large number of genes in which each gene is normally represented by more than one probe.

Tuberculosis is a bacterial infection usually occurred in human and animal bodies caused by Mycobacterium tuberculosis, this bacteria usually infect the lungs, but it can infect other parts of the body of an organism. TB disease is a communicable disease that can easily be transmitted between the human being via the inhalation process, when a person with pulmonary TB infection exhaled the bacteria. The individuals nearby or those who live with others having active TB infections, or smoking cigarettes, young children, alcoholics and intravenous drug users, malnutrition, Patients with HIV/AIDS or other immune system deficiencies can become

infected. Tuberculosis is a major cause of illness and death worldwide, but most of the TB infections of about 90% to 95% are latent TB infections which does not show any symptom of the disease. At 2012, 8.6 million meningeal TB cases were estimated all over the world (WHO, 2013). Many cases concerning these infections occurred in those with HIV infections and also most of these cases are occurring in developing countries where the majority of people were having very weak immune system in their bodies due to the lack of proper nutrition. The abilities of immune responses in the body against facultative intracellular pathogens are dependent on the rich food items taken by the individual (Kaufmann, 2002). Despite all the efforts and contributions of molecular technology to human tuberculosis infections. However, still the current issues concern with this disease are detection of the infection and proper way of controlling the infection. The challenge also on the other hand, with microarray is that, it is not easy to analyze microarray data due to the large dimensionalities in microarray data, despite there are many tools available for analyzing microarray image data, but the proper use with such tools to obtain good results depends on the objectives at hand. The overall aim of this research is to process the raw data and determine common expression pattern using some tools available for analyzing human TB microarray gene expression data in the Bioconductor R packages.

II. Materials And Methods

Microarray image data were collected from GEO-NCBI (Gene Expression Omnibus-National Centre for Biotechnical information's) database in a form of CEL file format with Accession number: GSE11199. The data contained 24 samples of tuberculosis infections from the human subjects, 12 samples were stimulated with phosphate buffered saline (PBS) to obtain monocyte-derived macrophages (MDMs) in order to increase the activities of RNA in the human tuberculosis samples and the remaining 12 were unstimulated. In each 12 samples of TB (stimulated and unstimulated) three different clinical forms of tuberculosis infections were found, each clinical group with 4 samples such as meningeal TB (MTB), pulmonary TB (PTB) and latent (LTB).

All the analyses of this research were carried out in the Bioconductor R package. The Bioconductor R package is a free computer software programming language and software environment for statistical computing and graphics. R software was downloaded and installed in the computer system and all the required missing packages in R package for this analysis were also installed in the R environment. Many Bioconductor tools were used for the preprocessing of microaaray human tuberculosis samples such as affycoretool was used for microarray TB image visualization analysis, AffyQCReport tools for TB microarray data quality assessment and GCRMA method for TB microarray data normalization. Hierarchical clustering (hclust) method was also used to determine common expression pattern among the three different clinical forms of human TB infections.

III. Results

Array Index	Array Names (Samples)
1	LTB1stm.CEL
2	LTB2stm.CEL
3	LTB3stm.CEL
4	LTB4stm.CEL
5	PTB1stm.CEL
6	PTB2stm.CEL
7	PTB3stm.CEL
8	PTB4stm.CEL
9	TBM1stm.CEL
10	TBM2stm.CEL
11	TBM3stm.CEL
12	TBM4stm.CEL
13	LTB1unstm.CEL
14	LTB2unstm.CEL
15	LTB3unstm.CEL
16	LTB4unstm.CEL
17	PTB1unstm.CEL
18	PTB2unstm.CEL
19	PTB3unstm.CEL
20	PTB4unstm.CEL
21	TBM1unstm.CEL
22	TBM2unstm.CEL
23	TBM3unstm.CEL
24	TBM4unstm.CEL

Table 1. Samples of stimulated and unstimulated microarray array raw data in a form of CEL file format of three different forms of human tuberculosis infection.

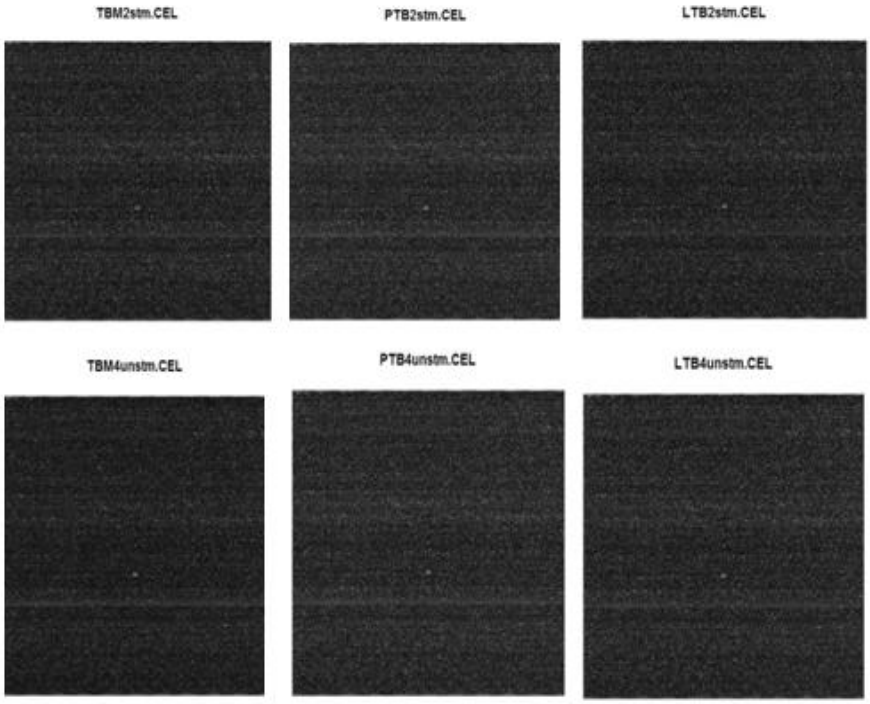

Figure 1: Images of some selected stimulated and unstimulted arrays of human tuberculosis infections.

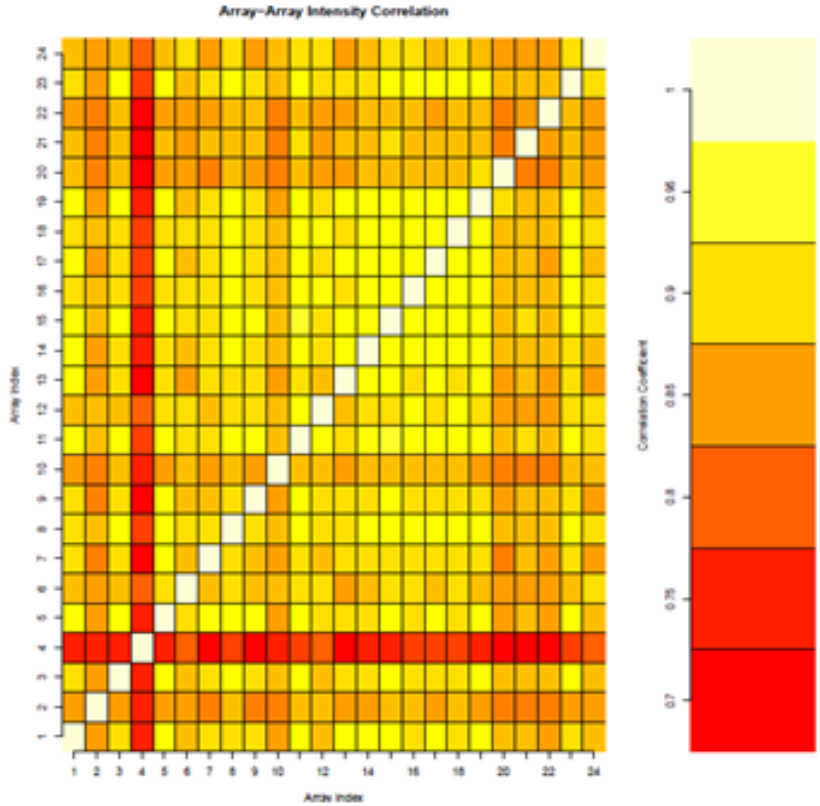

Figure 2: Array to array intensity correlation using stimulated and unstimulated microarray samples of three different clinical forms of human tuberculosis infection.

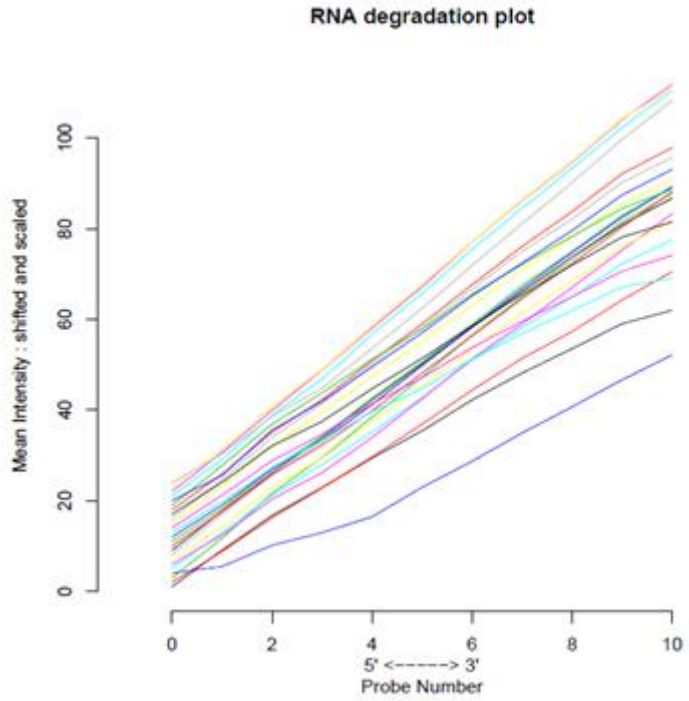

Figure 3: RNA degradation analysis using stimulated and unstimulated microarray samples of three different clinical forms of human tuberculosis infection.

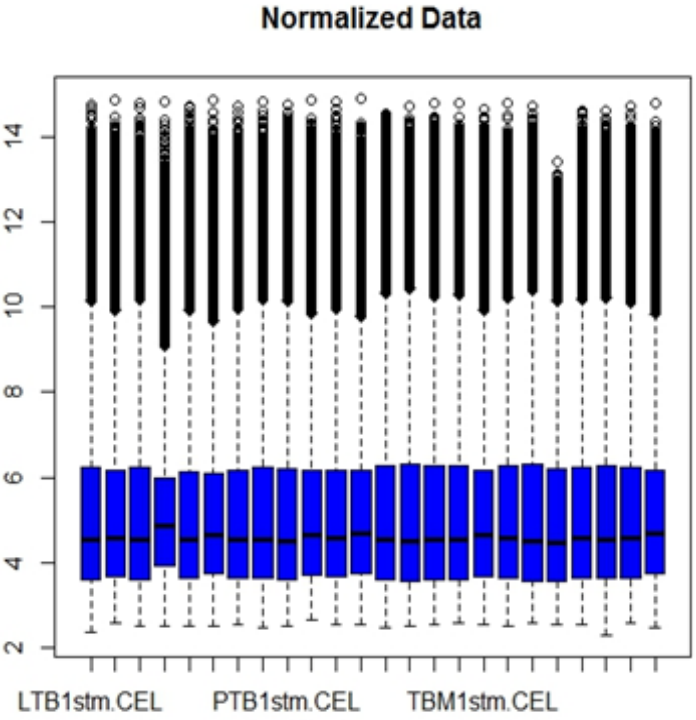

Figure 4: Box plot of normalized samples of stimulated and unstimulated microarray raw data of three different forms of human tuberculosis infection.

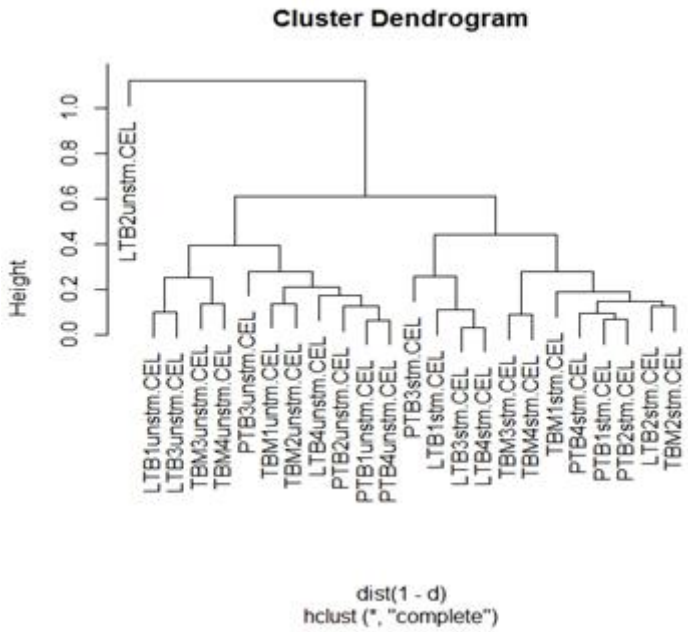

Figure 5: Hierarchical clustering of stimulated and unstimulated microarray samples of human tuberculosis infection.

IV. Discussion

Microarray gene expression data analysis of human tuberculosis infections in the Bioconductor R package allows display of all the array names (samples) as shown in table 1and image visualization of each array as shown in Figure 1, the essence of image visualization of microarray data is for visual inspection and it is possible to find out if there are technical problems eventually occurring only in one region of the array or not. All the selected arrays visualized indicates no much technical problem. The easiest way to identify which array among many arrays associated with the problem is to correlate all the arrays with each other (Wilson *et al*., 2004). In Figure 2 correlation coefficient analysis was used to detect outlier arrays, error in hybridizations, and then to get valuable information about phenotypic characteristics of the raw data (i.e. Replicate and tissue). This heat map of an array to array correlation coefficients indicates a good quality of this data since the smallest correlation coefficient was 0.7, but latent TB sample 4 stimulated (LTB4stm) array has quality problems with very high signals, high variability, and strong background, also appeared different in this plot, correlating poorly with other arrays. It was indicated that, pairs of arrays had a stronger correlation within the tissues than the correlation between the tissues. Samples from similar tissues or treatments tend to have a higher correlation coefficient. RNA degradation analysis was used to assess the quality of RNA and gives a good indication of the quality of the sample that has been hybridized to the array. It occurs when the molecule

begins to break down and is therefore ineffective in determining gene expression. Because RNA degradation usually starts from the 5′ end to 3′ end of the molecule, a strong degradation would result with the values for the probes closer to the 5' end and when the degradation is progressing to 3′ end the probe set should elevate (Jarno, 2008). Figure 3 indicates good quality of RNA in all the samples, because as the degradation was progressing from 5′ end to 3′ end, there was an increase in the probe set number and mean intensity. In figure 4, the plot indicates no intensity-dependent biases associated with plot because it is already corrected with GCRMA normalization method.

Clustering is a method used to determine the common recognition or spatial gene expression patterns. The purpose of clustering is to check whether samples are grouped according to known categories and to identify new classes of biological samples (Nguyen *et al.,* 2008). The results from the hierarchical clustering (hclust) in Figure 5 shown that in unstimulated tuberculosis samples, LTB1 and LTB3 were very similar but LTB4 was distinct from the arrays with the same clinical group. TBM3 and TBM4 were very similar but related to LTB1 and LTB3. TBM1 and TBM2 were very similar but a little bit away from arrays with the same clinical group. PTB1 and PTB4 were very similar and closely related to PTB2 but PTB3 completely away from them. LTB2 was very distinct from all the arrays of the same clinical group and others of different clinical group. For stimulated tuberculosis, LTB1, LTB3 and LTB4 were similar but LTB3 and LTB4 were more closely related. TBM3 and TBM4 were more closely related, but similar with TBM1. PTB1 and PTB2 were also very similar but were related to PTB4. PTB3, LTB2 and TBM2 were distinct from the arrays with the same clinical group. This finding shows there were more relatedness between expression levels of the arrays from the same clinical

group of tuberculosis infection than those arrays with different clinical group. It also indicated differences between the arrays, this may happen because of the analysis used different tools. These results suggested that hierarchical clustering analysis distinguish different clinical forms of human tuberculosis infection.

V. Conclusion

Stimulated and unstimulated microarray data of human tuberculosis infections were successfully analyzed in the Bioconductor R package using different tools. Affycoretools, AffyQCReport tool, GCRMA were able to use for the preprocessing of microarray data. Hierarchical clustering (hclust) method was used to determine common expression pattern among the three different clinical forms of human TB infection, This finding shows there were more relatedness between expression levels of the arrays from the same clinical group of tuberculosis infection than those the arrays with different clinical group and also indicated differences between the arrays. It suggested that, hierarchical clustering analysis distinguish different clinical forms of human TB infection. This study recommended that the results generated from these findings can be used in further analysis for detection and control of human TB infections.

References

[1]. Jarno Tuimala (2008). DNA microarray data analysis using Bioconductor. CSC, the Finnish IT centre for Science, CSC – IT Center for Science Ltd. 2008 ISBN 978-9525520-34-7http://www.csc.fi/oppaat/R.

[2]. Kaufmann S. H. E. (2002). "*Protection against tuberculosis: cytokines, T cells, and Macrophages"*. Department of Immunology, Schumannstrasse 21-22, D-10117 Berlin, Germany. PMCID: PMC1766701.

[3]. Miller M. B., Tang Y. W. (2009). **"***Basic concepts of microarrays and potential Applications in clinical microbiology".*Clinical Microbiology Review, v.22 (4); PMC2772365.

[4]. Nguyen T. Thuong, Sarah J. D., Tran H C., Vesteinn T., Cameron P. S., Nguyen T H. Guy E. T., Nguyen T. N., Martin H, Yik Y. T., Mark S., Alan A., Jeremy J. F., and Thomas R. H (2008). "*Identification of Tuberculosis Susceptibility Genes with Human Macrophage Gene Expression Profiles*". PLoS Pathog. Published online Dec 5,2008. Doi: 10.1371/journal.ppat.1000229PMCID: PMC2585058.

[5]. Wilson CL, SD Pepper, Y Hey, and CJ Miller (2004). Amplification protocols Introduce systematic but reproducible errors into gene expression studies. Biotechniques, 36:498–506, 2004.

[6]. World Health Organization (WHO) (2013). Tuberculosis fact sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/.

IOSR Journal of Biotechnology and Biochemistry (IOSR-JBB) is UGC approved Journal with Sl. No. 4033, Journal no. 44202.

Umar Shittu "Analysis of Common Gene Expression Pattern From Human Tuberculosis Microarray Data In The R Pakage." IOSR Journal of Biotechnology and Biochemistry (IOSR-JBB) 4.1 (2018): PP 42-47.