Journal of
Proteomics & Bioinformatics

ISSN: 0974-276X

Review Article Open Access

Microarray Gene Expression Statistical Data Analysis of Three Different Clinical Forms of Human Tuberculosis Stimulated Samples in the Bioconductor R Package

Umar Shittu[1]*, Mohammed Abu Naser[2], Zainab-L Idris[3] and Maryam SA[1]

[1]*Department of Biology, Isa kaita College of Education Dutsin-Ma, Katsina State, Nigeria*
[2]*Scientific Computing Solutions House No. 19, Road No. 1, Block-B, Section-6, Mirpur, Dhaka-1-207, Bangladesh*
[3]*Department of Biochemistry, Bauchi State University Gadau, Nigeria*

Abstract

The overall aim of this research identified and explores the usage of microarray gene expression statistical tools available in Bioconductor R package for image visualization, data quality control, background correction, summarization, normalization and identification of highly differential gene expression from microarray gene expression data of human tuberculosis infections. The stimulated samples with phosphate buffered saline (PBS) of human tuberculosis microarray gene expression data such include pulmonary TB infection (PTB), meningeal TB infection (TBM) and latent TB infection (LTB) 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 and all the analyses were performed in the R packages. These analyses identified and explore the use of AffyQCReport tool, affycoretools, PCA, MAS 5.0 and GCRMA for microarray gene expression data pre-processing and for the identification of highly significantly expressed genes, LIMMA was used and explore as a statistical tool for such analysis. The statistical analysis from LIMMA indicates that there was a significant difference between the three different forms of human tuberculosis. Therefore, most of the genes significantly expressed in both groups were genes responsible for cellular immune response. The results of three different comparison groups generated from the LIMMA analysis were further analysed using correlation coefficient=1, when p-value<=0.05 and generated Venn diagram, the results from venn diagram shows that majority of the genes were up-regulated indicating less decrease in the rate of gene expression but increase among the regulated genes of stimulated tuberculosis and more genes were observed with higher expression than those with less expression during the three group's comparison. It suggested recommendation that the results obtained from this study can be utilize in further analysis for detection and control of human tuberculosis infections.

Keywords: Human tuberculosis; Microarray; Data analysis

Introduction

Tuberculosis is an infection caused by *Mycobacterium tuberculosis* that usually attacks human body system, it is a communicable disease and can easily be transmitted between the human being through the exhalation and inhalation processes [1]. The infection was found to be of three different clinical forms namely Latent TB, Meningeal TB and Pulmonary TB, but pulmonary TB is a chronic TB infection [2]. TB was first discovered over a very long period of time since 18 centuries by the scientists. Robert Koch discovered the bacteria called *Bacillus* and it was described in 1882 as infectious TB disease (contagious in nature). Koch also disagrees with the study that indicated TB infections of human and cattle are similar, this statement from Koch made many people to agree that, milk from cattle is not a source of human TB infection. During that time, Koch was one of the advisors on health sector to his government and for that reason he became interested in tuberculosis research. At that very time, people widely believed that tuberculosis was an inherited disease. Also, Koch was totally agreed that tuberculosis was caused by a Bacterium and it was an infection, and he tested four postulates using guinea pigs. After conducting these experiments, the result shows that tuberculosis satisfied all the four postulates [3]. At the 19th century, people are looking at TB infection as "romantic disease." because of the high mortality rate in adults, a royal commission was set by the UK government in 1901 [4]. Human tuberculosis infection is often identified from the infected individual only when patients are critically ill or at postmortem investigations [5].

Microarray is a new technique that at a time deals with thousands of RNA or DNA. Microarray technology is usually referred to molecular technology and is one of the technologies that play vital roles toward the development of studies in the field of Molecular Biology. This technology used genetic materials of an organism (RNA or DNA) to identify new genes for a particular trait with their expression levels and functions under different environmental situation. This technology is very useful in providing information about a particular disease and provide the possible ways of controlling that infection and also to measure the expression levels of large numbers of genes or the whole genome of an organism at a time, or to identify differential expressed genes or to determine common expression pattern among the genes. Microarray experiment deals with the collection of tissues or cells from the living organism's extraction of genetic materials from the tissues or cells, hybridization of genetic materials and processing of hybrid in a scanning machine to generate microarray data. [6]. Beginning of the 20th century, medical research settings of UK government was put more concern about the issue of tuberculosis and TB cases was

***Corresponding authors:** Umar Shittu, Department of Biology, Isa kaita College of Education Dutsin-Ma, Katsina State, Nigeria, Tel: +23408065578102; E-mail: shittuumarkk@gmail.com

Received November 11, 2017; **Accepted** February 05, 2018; **Published** February 12, 2018

Citation: Shittu U, Naser MA, Idris ZL, Maryam SA (2018) Microarray Gene Expression Statistical Data Analysis of Three Different Clinical Forms of Human Tuberculosis Stimulated Samples in the Bioconductor R Package. J Proteomics Bioinform 11: 051-056. doi: 10.4172/jpb.1000465

considered as the most urgent issue in the health sector [7]. Study using microarray data to identify early lung immune responses to human tuberculosis using model of mouse was also done [8]. Analysis of microarray image data of meningeal tuberculosis and microarray image data of host immune response to human tuberculosis infections were designed and performed [9]. However, despite all the efforts and contributions of this molecular technology to human tuberculosis, 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 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 selected and explores the usage of a statistical tool and some tools available in R package for pre-processing and identification of highly differential gene expression from microarray gene expression data of human tuberculosis infections.

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 12 Stimulated clinical samples of human tuberculosis, Four (4) samples from each clinical group which include meningeal TB (MTB), pulmonary TB (PTB) and latent (LTB). The samples was 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. All the analyses of this research were carried out in the Bioconductor R package. R package is a free 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. Microarray data were pre-processed with affy, PCA, GCRMA and MAS 5.0 tools. LIMMA was used in identification of differential gene expression, because LIMMA uses the linear model to perform the statistical analysis of microarray data, this analysis requires the formation of compatible target file and designing of matrices. The statistical analysis of microarray data using LIMMA to compare three different forms of stimulated tuberculosis and infections was carried out. B-Statistic was also used in the comparison. The B-statistic (B-values) is based on the Empirical Bayes approach to rank genes and determine if a gene is significantly expressed or not. B-statistic is the log-odds that gene is differentially expressed, for instance if B=1.9, the odds of differential expression (1.9)=5.70, the probability that a gene is differentially expressed is 5.70/(5.70+1) × 100%=85%. In this case there is 85% chance for that gene is differentially expressed, at the B-value=0, there is 50%-50% chances. B-statistic, t-statistic and p-value (probability) are generated with LIMMA method in the R package [10].

Results and Discussion

Human tuberculosis microarray gene expression statistical data analysis in R package allows 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 in the above mentioned figure indicates no much technical problem. A histogram plot of arrays log intensity against density levels in Figure 2, provides information with a graphical representation of the distribution of microarray gene expression data in order to observe whether there is a need for data normalization or not. From the plot the arrays indicates overlapping between the samples, latent TB sample 3 stimulated (LTB3stm) overlapped with having

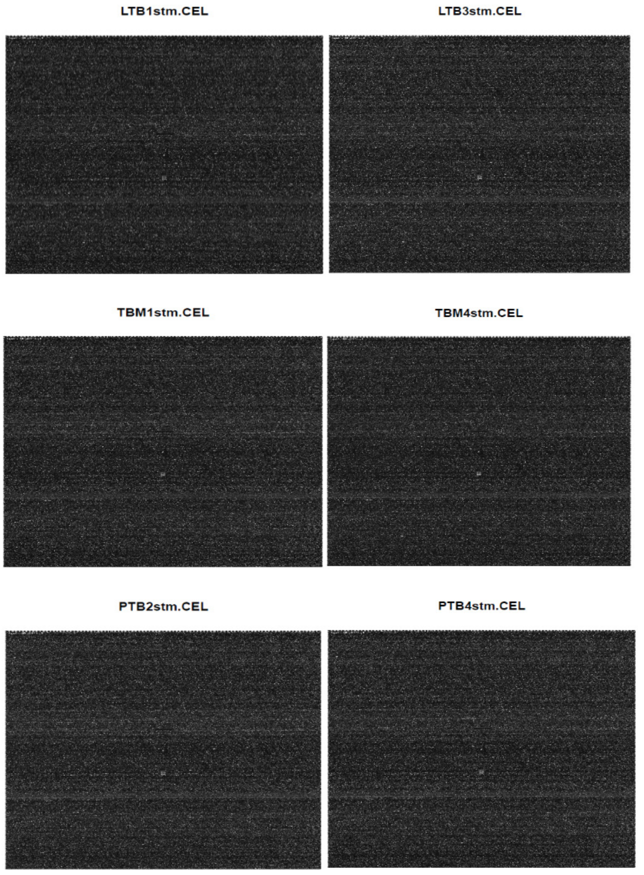

Figure 1: Images of some selected stimulated arrays of human microarray tuberculosis infection.

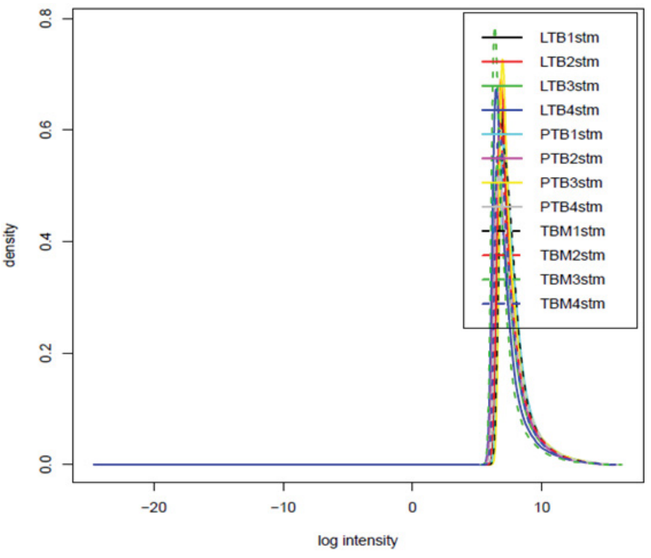

Figure 2: Density plot using stimulated samples of three different clinical forms of human microarray tuberculosis infection.

the highest density value and log intensity value among all the arrays, pulmonary TB sample 3 (PTB3stm) is the second array with highest log intensity value, followed by latent TB sample 2 stimulated (LTB2stm) and meningeal TB sample 4 stimulated (TBM4stm) is among the arrays that shown highest log intensity value as well as highest density value. Because of these problems of overlapping observed associated with the microarray data, it clearly indicates that there is a need for data normalization before applying any advance microarray gene expression data analysis to the data for producing better and reliable results. 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 is usually start 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 [11]. 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 probe set number and means intensity.

The easiest way to identify which array among many arrays associated with the problem is to correlate all the arrays with each other [12]. In Figure 4 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 array to array correlation coefficients indicates a good quality of this data since the smallest correlation coefficient was 0.84, but meningeal TB sample 3 stimulated (TBM3stm) 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. Figure 5 presented principal components analysis (PCA) of the microarray data by performing a covariance analysis between the factors and reduces the dimensionality of the data. PCA identifies basic temporal patterns as the important features that characterize genes [13]. PCA can be used to identify and remove the variables with correlation problem, it represent the samples with a smaller number of variables, can detect dominant patterns of gene expression and also visualize samples and genes. Simpleaffy plot is a general quality control statistic which provides visual representation of the microarray raw data by using beta-actin and GAPDH. Beta-actin

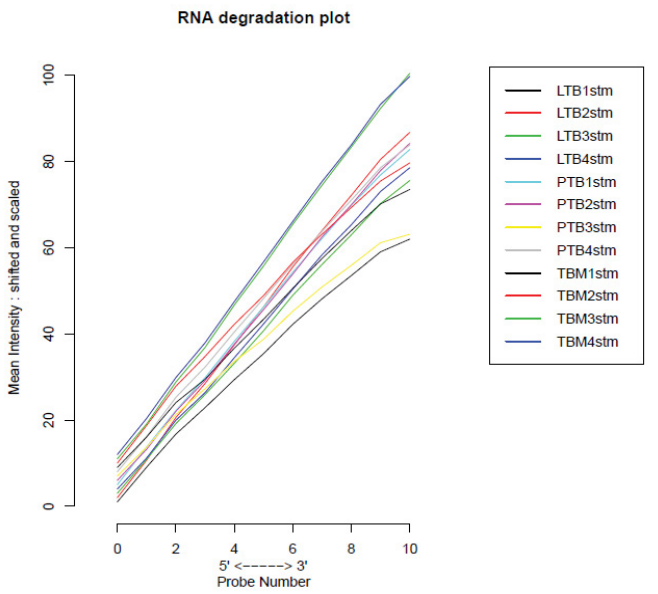

Figure 3: RNA degradation analysis using stimulated microarray samples of three different clinical forms of human tuberculosis infection.

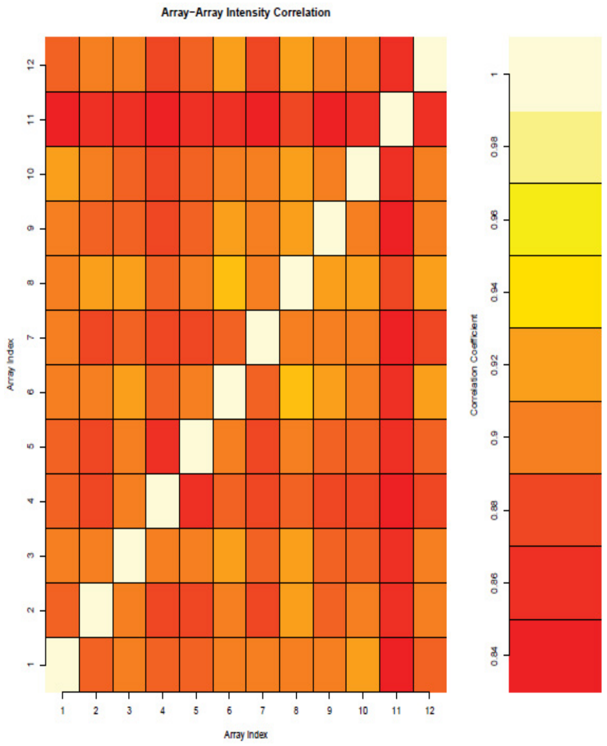

Figure 4: Array to array intensity correlation using stimulated microarray samples of three different clinical forms of human tuberculosis infection.

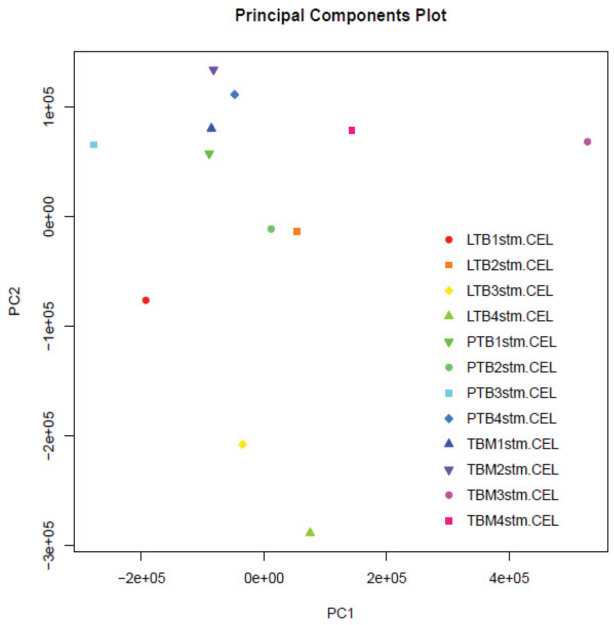

Figure 5: Principle component analysis plot using stimulated microarray samples of three different clinical forms of human tuberculosis infection.

(β-Actin) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were considered as constitutive housekeeping genes for RT-PCR and used to normalize changes in specific gene expression, because they are usually expressed in most of the cell types and are relatively long genes. QC plot in Figure 6 shows the 3′ to 5′ ratios for gene hybridization with specific control to the array type, the percentages of present gene calls and background for β-Actin and GAPDH. Plotted triangles represent Actin values while plotted circles represent GAPDH values. It visualized the first array from the bottom of the plot to the last array at the top which corresponds to the order of the samples with ratios and percentages, the dotted horizontal lines separate this plot into rows one for each array. The dotted vertical lines provide a scale from -3 to 3. A line to the left corresponds to a down-scaling, to the right, to an up-scaling. According to Affymetrix GAPDH and β-Actin values that are considered a potential outlier when the ratio>1.25 and are coloured red, The blue stripe in the image represents the range where scale factors are within 3-fold of the mean for all chips, If any scale factors fall outside this 3-fold region, they are all coloured red and unacceptable but if fall within the scale, are coloured blue and consider acceptable [14]. Microarray data normalization using GCRMA was illustrated in Figure 7 in which the method use to ignore the mismatch intensities and taking into account the probe sequence information and allows an efficient filtering of irrelevant probe sets. According to B-statistics (B-value), the introduction of additional biological replicates with high correlation coefficients tends to produce higher B-values [10]. The higher the B-value, the lower the p-value, the stronger the evidence of the results and the more significant of the result, the t-test was relatively more powerful under symmetric distributions or in a situation where the magnitude of the differences was moderate [14]. LIMMA method was used and identified 54675 genes from the samples; Tables 1-3 illustrate the result of top 20 significantly expressed genes between the three groups of comparison of human stimulated tuberculosis. TACSTD2 (p-value=0. 004 sorted by B-value=8. 27) was the most significantly expressed gene among the three different forms of human stimulated samples of tuberculosis infection. This gene was found in PTB/LTB group only. ETV5, RPS24 and ITGB2-AS1 genes were also significantly expressed in that group. The majority of the genes that were significantly expressed among the top 20 of stimulated samples were found to be in TBM/PTB group such as ETV5, RPS24, PACSIN2, FLT1, SIGLEC10, EPM2AIP1, RTN1, FILIP1L and EPB41L3. Only two genes were found significantly expressed in TBM/LTB group such as FILIP1L and EPB41L3, these genes were also significantly expressed in TBM/PTB. Four genes were found in common in both PTB/LTB and TBM/PTB groups, but only two genes were significantly expressed in both groups such as ETV5 and RPS24, the other two GPANK1 and UBN1 appeared less significance in both two groups. No-one among the genes significantly expressed or with less significance found across in all the three different groups of comparison. Therefore, most of the genes significantly expressed in both groups were genes responsible for cellular immune response. Recent study suggested that lncRNAs might be crucial for regulating the antituberclosis infection mechanisms of macrophages [15]. Venn diagram is a way of comparing genes from different groups showing how many genes are in common and exclusively expressed genes between the groups. Up regulation of genes indicates an increase in the rate of gene expression, Up-regulated genes are genes observed to have higher expression. Down regulation of genes also indicates a decrease in the rate of gene expression, down-regulated genes are genes observed to have lower expression [16]. The analysis in Figure 8 was done by using correlation coefficient=1 and when p-value<=0.05, the results indicated that the number of up regulated genes (41) were higher than down regulated genes (7), but

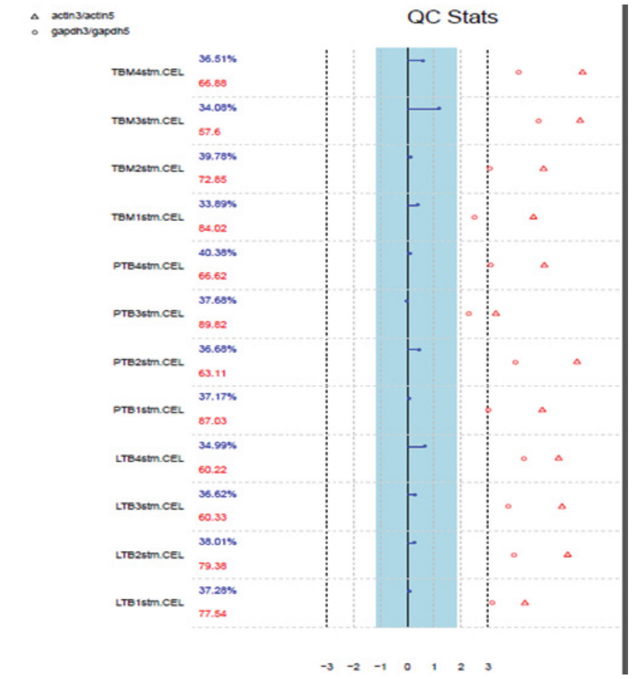

Figure 6: QC plot of microarray stimulated raw data of three different clinical forms of human TB infection from the simple affy package.

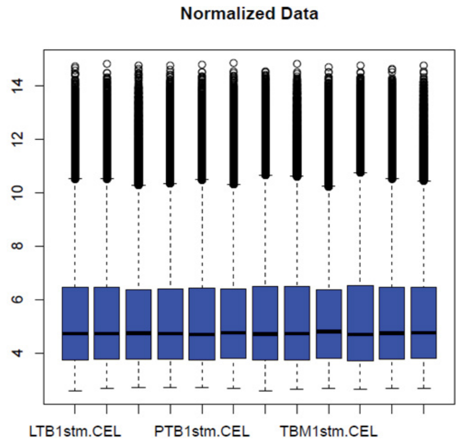

Figure 7: Normalized stimulated microarray samples of three different clinical forms of human tuberculosis infection.

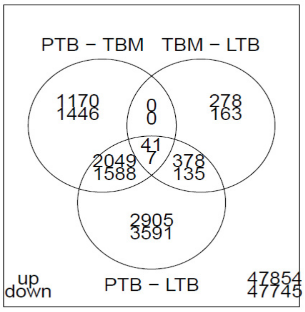

Figure 8: Venn diagram showing 'up' and 'down' regulated genes among the three group comparisons of stimulating samples.

Gene symbol	Function	LogFC	Ave. exp.	T-test	P-value	B-value
TBM/PTB						
ETV5	DNA binding and cellular response	- 0.88	1.78	-10.78	0.03	5. 23
RPS24	Structural constituent of ribosome & poly(A) RNA binding	0.79	2.77	10.33	0.03	4.82
PACSIN2	Cellular component	- 0.18	3.44	- 9.90	0.03	4.42
FLT1	Cellular response to vascular endothelial	- 0.67	1.68	- 9.64	0.03	4.17
SIGLEC10	Immune response as an inhibitory receptor	- 0.45	3.18	- 9.54	0.03	4.07
EPM2AIP1	Interacting gene	0.51	2.41	9.23	0.03	3.77
RTN1	Cellular component and secretion	- 0.47	3.15	- 8.70	0.05	3.22
FILIP1L	Regulator of the antiangiogenic activity	- 0.61	2.59	- 8.28	0.05	2.76
EPB41L3	Suppressor that inhibits cell proliferation	- 0.12	3.53	- 8.27	0.05	2.75
ACO1	Regulation of translation and iron binding	- 0.27	3.03	- 7.68	0.09	2.08
GADD45B	Regulation of growth and apoptosis	0.19	3.10	7.46	0.09	1.82
GCLC	glutathione activity	- 0.29	3.17	- 7.24	0.12	1.56
CLEC5A	Cellular defence response with signalling	- 0.87	3.23	- 6.94	0.12	1.19
GPANK1	Encode protein that play role in immunity	- 0.09	3.18	- 6.71	0.14	0.90
UBN1	DNA and transcription factor binding	- 0.19	2.44	- 6.63	0.14	0.79

Table 1: Identification of differential gene expression between human stimulated microarray samples of meningeal and pulmonary tuberculosis infections.

Gene symbol	Function	LogFC	Ave. exp.	T-test	P-value	B-value
PTB/LTB						
TACSTD2	Growth and cell surface receptor	- 1.63	1.79	-14.59	0.004	8. 27
ETV5	DNA binding and cellular response	- 0.88	1.78	-10.78	0.03	5. 23
RPS24	Structural constituent of ribosome & poly(A) RNA binding	0.79	2.77	10.33	0.03	4.82
ITGB2-AS1	Antisense RNA 1	- 0.39	2.56	- 9.35	0.03	3.88
DOK1	Multimolecular signalling complexes	- 0.17	2.73	- 6.96	0.12	1.21
GPANK1	Encode protein that play role in immunity	- 0.09	3.18	- 6.71	0.14	0.90
UBN1	DNA and transcription factor binding	- 0.19	2.44	- 6.63	0.14	0.79

Table 2: Identification of differential gene expression between human stimulated microarray samples of pulmonary and latent tuberculosis infections.

Gene symbol	Function	LogFC	Ave. exp.	T-test	P-value	B-value
TBM/LTB						
FILIP1L	Regulator of the antiangiogenic activity	- 0.61	2.59	- 8.28	0.05	2.76
EPB41L3	Suppressor that inhibits cell proliferation	- 0.12	3.53	- 8.27	0.05	2.75
ACPP	Catalytic activity and enzyme regulation	- 0.35	2.73	- 6.69	0.14	0.87

Table 3: Identification of differential gene expression between human stimulated microarray samples of meningeal and latent tuberculosis infections.

no gene regulated between PTB\TBMand TBM/LTB. This analysis suggested that there was high increase in the rate of gene expression but less decrease among the regulated genes of stimulated tuberculosis and more genes were observed with higher expression than those with lower expression during the three group's comparison. RT-qPCR was used on pulmonary tuberculosis and many genes were found to be up-regulated, while only one was down-regulated [2].

Conclusion

Stimulated microarray data of human tuberculosis infections were successfully analysed in the R package using different tools. Affycoretools, AffyQCReport tool, GCRMA and MAS 5.0 were able to use for the pre-processing of microarray data. Microarray data statistical analysis was carried out using the LIMMA method. Stimulated samples of tuberculosis were analysed and top 20 significantly expressed genes were identified for each group using a p-value<=0.05 sorted by B-statistics. This statistical analysis from LIMMA indicates that, there was a significant difference between the three different forms of human tuberculosis. Therefore, most of the genes significantly expressed in both groups were genes responsible for cellular immune response. Venn diagrams generated from the results of LIMMA analysis of the samples show that, the majority of the genes were up-regulated indicating less decrease in the rate of gene expression but increase among the regulated genes of stimulated tuberculosis during the three group's comparison. This research, recommended that microarray raw data should be quality assess before carrying out microarray advance analysis. Because quality control provides essential information that serve as a guide for selecting the appropriate normalization method for the data in order to have reliable results. It also recommended that, the results generated from the statistical analysis of LIMMA method for identification of top 20 significantly expressed genes from the samples can be utilize in further analysis for detection and control of human tuberculosis infections.

References

1. Carl Zimmer (2014) “Tuberculosis Is Newer than Thought, Study Says”. New York Times, 21 August 2014. Centers for Disease Control and Prevention.
2. Chen ZL, Wei LL, Shi LY, Li M, Jiang TT, et al. (2017) Screening and identification of lncRNAs as potential biomarkers for pulmonary tuberculosis. Sci Rep 7: 16751.
3. Beresford B, Sadoff JC (2010) Update on research and development pipeline: Tuberculosis vaccines. Clin Infect Dis 50: S178-S183.
4. World Health Organization (WHO) (2013) Tuberculosis fact sheet.
5. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, et al. (2014) Diagnosis of childhood tuberculosis and host RNA expression in Africa. N Engl J Med 370: 1712-1723.
6. Miller MB, Tang YW (2009) Basic concepts of microarrays and potential applications in clinical microbiology. Clinical Microbiology Review 22: 611-633.
7. Paul Kielstra (2014) Ancient enemy, modern imperative: A time for greater action against tuberculosis. In: Zoe Tabary (ed) Economist Insights (The Economist Group). Retrieved 1 August 2014.
8. Kang DD, Lin Y, Moreno JR, Randall TD, Khader SA, et al. (2011) Profiling early lung immune responses in the mouse model of tuberculosis. PLoS ONE 6: e16161.
9. Gonzalez-Juarrero M, Kingry LC, Ordway DJ, Henao-Tamayo M, Harton M, et al. (2009) Immune response to Mycobacterium tuberculosis and identification of molecular markers of disease. Am J Respir Cell Mol Biol 40: 398-409.
10. Thomas Girke, UC Riverside (2010) R & Bioconductor Manual. Institute for Integrative, Genome Biology.
11. Jarno Tuimala (2008) DNA microarray data analysis using bioconductor. CSC, the Finnish IT Centre for Science, CSC – IT Center for Science Ltd.
12. Wilson CL, SD Pepper, Y Hey, CJ Miller (2004) Amplification protocols Introduce systematic but reproducible errors into gene expression studies. Biotechniques 36: 498-506.
13. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cylce-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell 9: 3273-3297.
14. Thuong NTT, Dunstan SJ, Chau TTH, Thorsson V, Simmons CP, et al. (2008) Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. PLoS Pathogens 4: e1000229.
15. Yang X, Yang J, Wang J, Wen Q, Wang H, et al. (2016). Microarray analysis of long noncoding RNA and mRNA expression profiles in human macrophages infected with *Mycobacterium tuberculosis*. Sci Rep 2016 6: 38963.
16. Kauffmann A, Gentleman R, Huber W (2009) Array quality metrics a bioconductor package for quality assessment of microarray data. Bioinformatics 25: 415-416.