Analyzing time course microarray data of *Toxoplasma gondii* and study the impact on host transcript levels using Bioconductor

Budhayash Gautam¹, Gurmit Singh², Atul Kumar Singh³ and Gulshan Wadhwa³*  
¹Department of Computational Biology and Bioinformatics and  
²Department of Computer Science and Information Technology  
Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad 211 007, India  
³Apex Bioinformatics Centre, Department of Biotechnology, Ministry of Science and Technology, New Delhi 110 003, India

*Author for correspondence:  
Tel: +91-9811301820; Fax: +91-11-24362884  
E-mail: gulshan@dbt.nic.in

*For ‘Supplementary Matter’ see the papers at www.nopr.niscair.res.in. or www.niscair.res.in.

*Toxoplasma gondii* is an obligate, intracellular, apicomplexan parasite that can infect a wide range of warm-blooded animals including humans. In humans and other intermediate hosts, *Toxoplasma* develops into chronic infection that cannot be eliminated by host’s immune response or by currently used drugs. The ability of the parasite to convert to the bradyzoite stage and to live inside slow-growing cysts that can go unnoticed by the host immune system allows for the persistence of parasite throughout the life of the infected host. Little is known, however, about how bradyzoites manipulate their host cell. Large scale microarray experiments are becoming increasingly routine, particularly those which track a number of different cell lines through time. This time course information provides valuable insight into dynamics of various biological processes. The proper statistical analysis, however, requires the use of more sophisticated tools and complex statistical models. In the current study, the open-source R programming environment in conjunction with the open-source Bioconductor software was used to analyze microarray data of *T. gondii*. Several statistical analysis procedures like (log) fold changes in conjunction with ordinary and moderated t-statistics were used to determine differentially expressed genes. The differentially expressed genes were subjected to cluster analysis, followed by the annotation of the up and down regulated genes based on the gene ontology. The findings in the present study suggest the overall effect of the gene expression changes is to modulate the key metabolic pathways leading to compromised host immune response, enhancement in programmed cell death, depression in cell proliferation process and induction of various diseases.

**Keywords:** Bioconductor, cluster analysis, differential gene expression, time course microarray, *Toxoplasma gondii*, t-tests

**Introduction**

*Toxoplasma gondii*, an intracellular pathogen, has the potential to infect nearly every warm-blooded animal but rarely causes morbidity. It is an extremely common parasite in humans and animals. Although sexual reproduction of this intracellular protozoan takes place only within felines, the intermediate hosts (many species of mammals and birds) support asexual reproduction consisting of two stages: tachyzoites and bradyzoites. Tachyzoites replicate rapidly, disseminate through the host and cause tissue damage. Most are then cleared by the host immune response but not before some have converted into the bradyzoite stage. Bradyzoites replicate slowly, form a cyst within the host cell and sustain a chronic infection for the life of the mammalian host. These bradyzoites latently persist and cause little pathology in a healthy host but, in an immune-compromised animal, they can reconvert into the tachyzoite stage and cause potentially fatal encephalitis. This intracellular survival likely necessitates host cell modulation and tachyzoites are known to modify a number of signaling cascades within the host to promote parasite survival. Little is known, however, about how bradyzoites manipulate their host cell⁵,².

In the present study, the human cDNA microarrays were used to investigate whether and how the changes in host gene expression during infection with bradyzoites differed from those during infection with tachyzoites of the parasite *T. gondii*³. For this study, microarrays data were analyzed to identify profile changes in host gene expression. Open-source R programming environment in conjunction with the open-source Bioconductor software was used to analyze microarray data of *T. gondii*. Several statistical analysis procedures like (log) fold changes in conjunction with ordinary and moderated
t-statistics have been used for determining differentially expressed genes. The differentially expressed genes were subjected to cluster analysis followed by the annotation of the up and down regulated genes based on the gene ontology.

Materials and Methods

The microarray data used in the present study was obtained from the Array-Express, which is Database of gene expression, and other microarray data at the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/arrayexpress/). The microarray data was of human foreskin fibroblasts (HFFs) with four experimental conditions that include uninfected "standard", uninfected "stress", tachyzoite-infected "standard+TZ", and bradyzoite-infected "stress+BZ". Analysis was performed using the limma package of Bioconductor. Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data (http://www.bioconductor.org).

Analysis of Microarray Data

Before any formal statistical analysis, it is important to check the data quality. Initially, data was examined at perfect and mismatch probe-level to detect anomalies. Fundamental to the task of analyzing gene expression data is the need to identify genes whose patterns of expression differ according to phenotype or experimental condition. A simple approach is to select genes using a fold-change criterion. This may be the only possibility in cases where no, or very few replicates, are available. An analysis solely based on fold change, however, does not allow the assessment of significance of expression differences in the presence of biological and experimental variation, which may differ from gene to gene. This is the main reason for using statistical tests to assess differential expression. In the present study, the statistical analysis was performed through moderate t-test using Limma package available in Bioconductor software as mentioned in the following steps.

The basic statistic used for significance analysis was the moderated t-statistic, which was computed for each probe and for each contrast. This has the same interpretation as an ordinary t-statistic except that the standard errors have been moderated across genes, i.e., shrunk towards a common value, using a simple Bayesian model. For extracting the true positive results from fit.ebayes object, multiple testing procedures were used. Multiple testing procedures allow one to assess the overall significance of the result of a family of hypothesis tests. They focus on specificity by controlling type I (false positive) error rates, such as, the family-wise error rate or the false discovery rate. In the present study, false discovery rate corrections were employed using the Benjamini and Hochberg procedure. Till now, there are no specific guidelines describing that up to what level of p-value of the differentially expressed genes are biologically significant. However a naive approach is to select all the genes having p-value less than 0.001 to keep the subset small, significant and manageable for further analysis. R code was used to extract all the probeset ids or the ids from the original normalized data matrix using the p-values of less than 0.00001 from the fit.ebayes object.

Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may help in further understanding of the functions of many genes for which information has not been previously available. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation.

In the present study, gene-based clustering was performed with all the clustering algorithms used (hierarchical clustering and partitioning around medoid); however, heatmaps produced by the hierarchical clustering also depicts sample-based clustering as well.

Annotation Based on Gene Ontology

Annotating the genes or, in other words, combining the gene expression data with other knowledge, is typically carried out after statistical testing. Bioconductor project produces annotation packages for many chip-types and these can be directly used for annotating the results. As an input, the annotation process takes a vector of gene names. Those can typically be extracted from a matrix of limma results. Output of the process is a text or an HTML file containing the annotations. In the present study, annotation was done manually from the supplemented data, which was provided by Fouts and Boothroyd
Results and Discussion

Analysis of Microarray Data

Quality Assessment Before and After Normalization

This step was used to determine any anomalies or defects in the probe level data before proceeding further for any analysis. In order to either correct the abnormalities or filter out the defected probe data, in the present study, the following steps were involved in quality assessment of the probe level data. Density plots and MA plots are the effective way to estimate errors in the probe level data. The MA plot gives a quick overview of the distribution of the data. Quality problems are most apparent from an MA plot in cases where the lowess smoother oscillates wildly or if the variability of the M values appears to be greater in one or more arrays relative to the others. However, these anomalies did not occur in all the six arrays used in the present study, suggesting the good quality of the chips used (Ψ Suppl Fig. S1)8. It depicts the distribution of intensity ratio (M) of the genes plotted by the average intensity (A). Ψ Suppl Fig. S2 shows the density plot for the distribution of probe level intensities of 20 arrays. Again anomalies did not occur in all the six arrays used in the present study, suggesting the good quality of the chips used8-10.

Statistical Analysis

Empirical Bayes is a better analysis method than, say, traditional t-test for DNA microarray data, since it gives us more precise estimates of the statistical significance of the genes. The results of the statistical analysis was in the form of list of top genes which were expressed most (Ψ Suppl Table S1), based on the logFC, adj.P.Val, p-value and false discovery rate9-21. All the genes that have the unadjusted p-value at most 0.00001 were extracted7,9,20,21. The total number of differentially expressed genes was 96. The volcano plots were plotted for these 96 genes (Fig. 1). The plot was showing the number of up regulated and down regulated genes, which were depicting a twofold change in there expressions. Of 96 genes, there were 32 genes which were up regulated (Ψ Suppl Table S2). The other 64 genes were down-regulated during the T. gondii infection (Ψ Suppl Table S3).

Clustering

Hierarchical clustering and soft clustering was performed using R of these top 96 genes, which were showing two fold changes in their expression levels. For clustering, Amap and Cluster libraries were used for hierarchical and PAM, respectively. The results of these clustering were as follows.

The clustering by PAM also shows 2 clusters of genes, in which 34 genes belong to one cluster and 62 belong to other cluster (Fig. 2), which is almost same as the hierarchical clustering of the 96 genes (Fig. 3). In the hierarchical clustering there are 2 gene clusters, among which 32 genes grouped to form the one cluster and the rest (64) grouped to form the other gene cluster (Fig. 2). Thus, classification and number of genes have been done properly by these two methods7,8.

Annotation of Genes

Annotating the genes is typically carried out after statistical testing. Bioconductor project produces annotation packages for many chip types, and these can be directly used for annotating the results. Suppl Tables S2 and S3 were manually prepared from the list of top differentially expressed genes. Annotation
of all the genes was prepared by the supplementary data provided by the SMD. There are some gene products which are necessary for host immune system but, due to the *T. gondii* infection, their normal level of production decreased, leading to the escape of the pathogen from the host immune system. This condition also helps the pathogen to reside in the host cell for a longer period of time and thus able to again infect the host or can transmitted from the host to another host.

Of 32 up regulated genes, the genes of our interest include the genes which codes for the following gene products: (1) Cyclin-dependent kinase 7 (MO15 homolog, Xenopus laevis, cdk-activating kinase): The protein encoded by this gene is a member of the cyclin-dependent protein kinase (CDK) family. CDK-family members are known to be important regulators of cell cycle progression. This protein is thought to serve as a direct link between the regulation of transcription and the cell cycle. Thus overexpression of this gene causes defective regulation of transcription and the cell cycle. (2) Hydroxymethylbilane synthase: It is an enzyme involved in the third step of the metabolism of porphyrin, which converts porphobilinogen into hydroxymethylbilane. The enzyme has the unique
cofactor dipyrromethane. Defective activity of this enzyme can lead to the disorder acute intermittent porphyria. These gene products are important enzymes and proteins as they are actively carrying out the cell metabolism. Thus, defective activity can severely affect host metabolism.

The other genes of our interest are from down regulated genes, which code for the following proteins: Interleukin 15, TNF receptor-associated protein 1, CD69 antigen, Chemokine receptor 1. All the above gene products are actively involved in host immune system, e.g., (1) Interleukin 15 (IL-15) is a cytokine with structural similarity to IL-2. Like IL-2, IL-15 binds to and signals through a complex composed of IL-2/IL-15 receptor beta chain (CD122) and the common gamma chain (gamma-C, CD132). This cytokine induces cell proliferation of natural killer cells; cells of the innate immune system whose principal role is to kill virally infected cells. (2) CD69 (Cluster of differentiation 69) is a human transmembrane C-type lectin protein encoded by the CD69 gene. The activation of T lymphocytes, both in vivo and in vitro, induces expression of CD69. This molecule is involved in lymphocyte proliferation and functions as a signal-transmitting receptor in lymphocytes, including natural killer (NK) cells and platelets. (3) Chemokines and their receptors, which mediate signal transduction, are critical for the recruitment of effector immune cells to the site of inflammation. Knockout studies of the mouse homolog suggested the roles of this gene in host protection from inflammatory response, and susceptibility to virus and parasite. This indicates that these gene products are down regulated so that the pathogen may evade the immune system of the host. Thus, these gene products play major roles in various immune processes.

Conclusion
The analysis of above results reveals that most of the genes in the samples were expressed at similar frequency and variance, with an exception of a group of genes those were differentially expressed in host, i.e., humans. Few genes were found highly over-expressed, while few others were found under-expressed. The up and down regulation of the genes were affecting key immune system leading to host compromised immunity and escape of the pathogen from the host immune system. As a result, the Toxoplasma remains undetected in the host for a longer period of time. Genes from these groups can be further used for prediction and analysis of the gene products, which will help in designing new diagnostic and treatment strategies for the toxoplasmosis. These results may be validated through wet lab.

Acknowledgment
The authors are grateful to the Sam Higginbottom Institute of Agriculture, Technology and Sciences (Deemed to be University), Allahabad for providing the facilities and support to complete the present research work.

References
2 Goebel S, Gross U & Luder C G, Inhibition of host cell apoptosis by Toxoplasma gondii is accompanied by reduced activation of the caspase cascade and alterations of poly (ADP-ribose) polymerase expression, J Cell Sci, 114 (2001) 3495-3505.
3 Fouts A E & Boothroyd J C, Infection with Toxoplasma bradyzoites has a diminished impact on host cell-transcript levels relative to Tachyzoite infection, Infect Immunol, 75 (2007) 634-642.


