Computational Analysis Of Differentially Expressed Genes In Mycobacterium Tuberculosis Infection

[1]Abayomi Mosaku, [3]Solomon Rotimi, [4]Samuel Ndueso John, [1, 2]Ezekiel Adebiyi
[1]Covenant University Bioinformatics Research (CUBRe), [2]Department of Computer and Information Sciences, [3]Department of Biological Sciences, Department of [4]Electrical and Information Engineering
Covenant University
Ota, Nigeria

Joao Pedro Leonor Fernandes Saraiva, Rainer Koenig
Jena University Hospital
Center for Sepsis Control and Care
Jena, Germany

***Abstract*—Tuberculosis remains a serious social and public health problem, affecting millions of people annually, and is reported at the end of 2014 by the World Health Organization as one of the world's deadliest communicable diseases. The most challenging being the multi-drug resistant strains of the mycobacterium. Another major challenge frustrating the effective control of this disease, especially in poor countries, is the long time taken to diagnose it, the standard diagnosis of TB is by microscopy, but this does not give any information on drug-resistance – the cell culture tests take two weeks, by which time it might have spread to many other people. In this project, the authors utilized various Statistical and Computational techniques to analyze and discover genes that are differentially expressed in human blood cell (Peripheral blood mononuclear cells, PBMCs) subsequent to its stimulation with heat-killed Mycobacterium on comparison with an Roswell Park Memorial Institute (RPMI) culture medium as a control. Using this in-silico technique, some unique biomarkers were discovered which are further discussed in details. These biomarkers identified as differentially expressed in the human blood cell will not only enhance our understanding of the pathogen, but is also a spring board for the completion of an Electronic hand-held, DNA-Based Tuberculosis diagnosis device. Our anticipated new technology is at the intersection of genetics and computer science that will be used for rapid and early detection of *Mycobacterium Tuberculosis* infection, a perfect alternative to all existing symptom based diagnostic tool.**

Index Terms*— *Tubercuosis, Mycobacterium tubercuosis, Illumina Micro Array, Gene Expression Data, Big Data, Bioinformatics, Computational Biology, Diagnosis, Hand-held device, Personalized Medicine

I. Introduction

Tuberculosis (TB) remains a major global health problem, responsible for ill health among millions of people each year. As at the end of 2014, the World Health Organization has called tuberculosis one of the world's deadliest communicable diseases and reported that of all infectious diseases, only the human immunodeficiency virus (HIV) which causes AIDS, kills more people than TB [2]. The extremely high death toll from this disease still poses the question: how is it that a staggering number of lives are being lost to a curable disease? TB is present in all regions of the world and the WHO Global Tuberculosis Report 2014 includes data from 202 countries and territories. In 2013, an estimated 9.0 million people developed TB and 1.5 million died from the disease. African region accounts for about four out of every five HIV-positive TB cases and TB deaths among people who have HIV [15]. An estimate of 9 million people developed TB in 2013, more than half (56%) of these were in the South-East Asia and Western Pacific regions. A further one quarter was in the African region, which also had the highest rate of cases and deaths relative to population. India and China alone accounted for 24% and 11% respectively of total cases.

The Pathogen, *Mycobacterium tuberculosis* TB is an infectious disease caused by the *Bacillus Mycobacterium tuberculosis*. It typically affects the lungs (pulmonary TB) but can affect other sites as well (extra pulmonary TB). The disease is spread in the air when people who are sick with pulmonary TB expel bacteria mostly by coughing [6]. A relatively small proportion of people infected with *M. tuberculosis* will develop TB disease. However, the probability of developing TB is much higher among people infected with HIV. TB is also more common among men than women, and affects mainly adults in the most economically productive age groups. The most common method for diagnosing TB worldwide is sputum smear microscopy (developed more than 100 years ago) in which bacteria are observed in sputum samples examined under a microscope [16].

The Need For Early and Rapid Diagnosis: Given that most deaths from TB are preventable, the death toll from the disease is still unacceptably high, rapid diagnosis of TB is key to ensuring that early and prompt attention is given to infected individuals. The urgency of this need is also reflected in the global TB strategy developed by WHO titled: *Stop TB Strategy* which emphasizes the need for participation in research to develop new diagnostics. Early diagnosis of TB is emphasized as a universal need for enhanced patient care. Sputum smear microscopy has been the primary method for detecting TB. Microscopy is

not a sensitive test, particularly in people living with HIV and in children; it cannot distinguish between *Mycobacterium tuberculosis* complex and non-tuberculosis *Mycobacterium* [22]. Though diagnosis based on culture is considered the reference standard, results take weeks to obtain and testing requires well-equipped laboratory, highly trained staff, and efficient transport system to ensure viable specimens. Following research and development in the past decade, rapid and more sensitive tests such as the Molecular based Xpert MTB/RIF device are now coming up to replace or complement existing conventional tests. Nevertheless, of the 4.9 million incident pulmonary TB patients globally in 2013, only 2.8 million (58%) were bacteriologically confirmed, i.e were smear or culture-positive. The remaining 42% were diagnosed clinically i.e. based on symptoms, chest X-ray abnormalities or suggestive histology [2]. The greatest challenge with symptom based diagnosis of TB is that common symptoms of TB combined with the poor specificity of X-ray screening may result in false diagnoses and people without TB being enrolled on TB treatment when it is not needed.

Our Research Objective: In our research, we used R Statistical Programming Language and "Bioconductor open source tools for the analysis and comprehension of high throughput genomic data" in analyzing gene expression data of *Mycobacterium tuberculosis* infected human host cell. Gene Set Enrichment Analysis (GSEA) was carried out, functional annotation and further computation was done using R/Bioconductor PIANO (Platform for integrative analysis of omics data) package, a list of biomarkers unique TB infected human blood cells were discovered.

In this study, we have taken an alternative, unbiased approach to this biological problem, using a combination of diverse computational technique, expression profile analysis and system biology to infer interesting genes that are differentially expressed in Tuberculosis infected human blood cell. Given human blood sample at any other time, the expression levels of this genes in the blood will be electronically computed by our computational algorithm that will be embedded in a hand-held device which will automatically diagnose the presence of tuberculosis infection in any patient. This will eradicate the diagnostic delays experienced in all presently existing methods.

II. MATERIALS AND METHODS

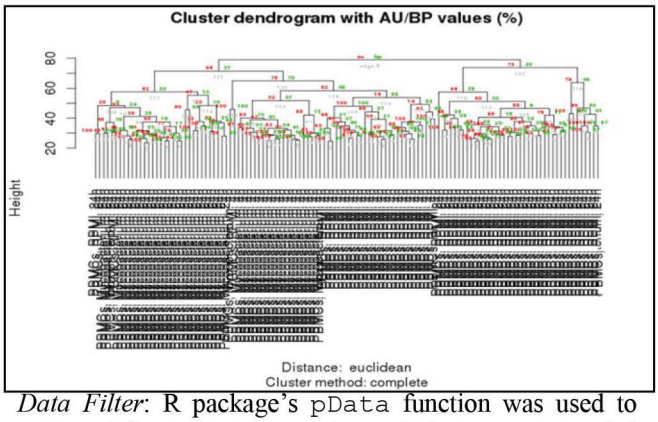
Pretreatment Of Experimental Data: Tuberculosis data was gotten from the Gene Expression Omnibus (GEO) from Smeeken P, et al's Laboratory Experiment. Total RNA of Peripheral blood mononuclear cells (PBMCs) were extracted from healthy human volunteers. PBMCs were stimulated with heat-killed *Mycobacterium tuberculosis* (MTB), and some other non-fungal inflammatory stimuli and RPMI culture medium as a control. A large number of biological replicates (>20) were included per stimulation condition and duration, resulting in overall 299 samples made up of *Candida albican, Escherichia coli-derived lipopolysaccharide* (LPS), *Borrelia burgdorferi, Mycobacterium tuberculosis* (MTB) and RPMI culture medium as a control. There were 59 Samples of PBMC stimulated with M.TB and 65 Samples of RPMI Control. The full Gene Expression data was downloaded from the GEO. A series record entity type of the expression profile array was downloaded as a GSE series Matrix files, it contains the full description of the experiment as a whole, together with all tables describing extracted data, summary conclusions, and analyses. Expression values of the data was retrieved using as a matrix of the gene expression measurement with rows as genes but represented in the data with their probe ids and columns containing the experimental samples, but represented in the data with their GEO sample IDs.

Illumina Data Variance Stabilization: The `lumiT` function which is an interface of difference variance stabilizing transformation. of Illumina data was used to stabilize the expression variance based on the bead level expression variance and mean relations, using the vst (Variance Stabilizing Transformation) method. The variance-stabilizing transformation takes the advantage of larger number of technical replicates available on the Illumina microarray. It models the mean-variance relationship of the within-array technical replicates at the bead level of Illumina microarray. An arcsinh transform is then applied to stabilize the variance. This results in the transformed (variance stabilized) gene expression value, which we later normalized

Normalization Of Data: We carried out a between chip normalization of the gene expression data using Robust Spline Normalization (RSN) algorithm, RSN combines the features of quantile and loess normalization. It is designed to normalize the variance-stabilized data. This produced a normalized gene expression values. The loess normalization is very reliable especially when local data are not appropriately modeled by linear regression, It is a robust data normalization technique capable of fitting complex non-linear function.

Gene Annotation: This was done in order to replace the probe-ids with the actual gene-ids for clearer gene identification. The `fData` function was used to access the feature data of the experiment, from where the data frame containing the entrez-ids of the probes were retrieved. This was mapped with the rows of the expression data to replace the probe-ids with the corresponding gene-id.

Computation Of Mean Expression Values: R package's `avereps` function was used to condense the microarray data object so that values for within-array replicate probes are replaced with their average. The resulting output was such that a new data object is computed in which each gene id is represented by the average of its replicate spots or features on the microarray.



Data Filter: R package's `pData` function was used to retrieve information on experimental phenotypes recorded in the expression set. These generic functions accesses the phenotypic data (e.g., covariates) and meta-data (e.g., descriptions of covariates) associated with the experiment. The column names which were previously labeled by their sample ids were replaced with the corresponding experimental titles noted by their time point (e.g pbmc_mycobacterium_24h, pbmc_mycobacterium_4h, pbmc_rpmi_4h, pbmc_rpmi_24h e.t.c). We carried out hierarchical clustering analysis and computed p-values for each cluster via multi-scale bootstrap re-sampling. Two types of p-values were computed: approximately unbiased (AU) p-value and bootstrap probability (BP) value. Multi-scale bootstrap re-sampling is used for the calculation of AU p-value, which has superiority in bias over BP value calculated by the ordinary bootstrap re-sampling [23]. In addition, the computation time was enormously decreased due to the available parallel computing option.

The hierarchical clustering was done for the experimental data using *complete linkage* method algorithm to find similar clusters, and the cluster distance between the experimental samples in the microarray data was computed using Euclidean distance algorithm as shown in the equation:

$$\mathrm{Ed}_{\mathrm{MTB,RPMI}} = \sqrt{\sum_{i=1}^{n}(MTB_i + RPMI_i)^2} \qquad (1)$$

Complete pairwise correlation was used in computing correlation in clustering the samples via Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}\mathrm{MTB}i\mathrm{RPMI}i - [(\sum_{i=1}^{n}\mathrm{MTB}i).(\sum_{i=1}^{n}\mathrm{RPMI}i)/\mathrm{n}]}{[\sum_{i=1}^{n}\mathrm{MTB}i2 - (\sum_{i=1}^{n}\mathrm{MTB}i)2/\mathrm{n}][\sum_{i=1}^{n}\mathrm{RPMI}i2 - (\sum_{i=1}^{n}\mathrm{RPMI}i)2/\mathrm{n}]} \quad (2)$$

The Pearson correlation coefficient is more sensitive to outliers than the non-parametric Spearman correlation coefficient [24]. Bootstrap re-sampling was done using over 1000 bootstrap replications, the cluster dendogram was generated and p-values were computed for each of the clusters. Intensity filter and variance filtering was carried out, the experimental sample distribution was found to have a normal distribution.

Differential Expression Analysis: The differential expression analysis was carried out, and a table of genes with their respective p-values, fold change, t-value, significance score, and level of their differential expression was generated. The algorithm used for the differential expression analysis is as shown in the flow chart in Fig. 3 The gene expression values for all the experimental was stored in an R variable as a numeric matrix named `fullData`, As at this stage the gene expression data still contained 299 columns of expression values for borrelia, candida, LPS, MTB, and RPMI but there was a need to filter out the array of expression values for Mycobacterium TB and RPMI alone. This was done and stored in another variable `mtnNrpmi` as shown below :

```
mtbNrpmi=fullData[,sampleAnn=='MTb'|sa
mpleAnn=='RPMI'];
```

We separated the 59 columns samples of *Mycobacterium tuberculosis* alone and stored as another variable `MTB` of numeric matrix type.

```
MTB = fullData[,sampleAnn=='MTb'];
```

We separated the 65 columns samples of RPMI alone and stored as another variable `NORMAL` of numeric matrix type.

```
NORMAL = fullData[,sampleAnn=='RPMI']
```

A character vector of 124 elements (`mtbRpmiVect`) made up of 59 elements as MTB and 65 elements as RPMI was created. The R package's `rowttests` function was used to carry out two-sided, two-class t-test with equal variances for each of the genes in the data. `rowttests` are implemented in C-programming Language and were found to be reasonably fast and memory-efficient compared to `fastT` which is an alternative implementation, in Fortran. `This` performs for each row of `the gene` a two-sided, two-class t-test with equal variances. It requires a factor whose length `must equal x (the number of genes in the sample)` with two levels, corresponding to the two groups (MTB and RPMI) in order to compute the statistics, corresponding to the two groups. The `mtbNrpmiVect` was encoded as a factor to be used in the statistics and stored as another variable `mtbRpmiFact`

```
mtbRpmiVect =
sampleAnn[sampleAnn=='MTb'|sampleAnn==
'RPMI'];
levels=c('MTb','RPMI');
mtbRpmiFact                          =
factor(mtbRpmiVect,levels)
test = rowttests(mtbNrpmi,mtbRpmiFact)
```

The p-values were computed and the adjusted P-value was further computed using the p-values resulting from the t-statistics by applying Benjamini & Hochberg false discovery rate method [25]. The path fold between the Mycobacterium infected expression values and the RPMI was also computed

```
fold    <-    as.numeric(apply(MTB,    1,
median) - apply(NORMAL, 1, median))
```

Differentially Expressed Genes Selection: A Matrix of differentially expressed genes were gotten. Genes with adjusted p-values less than 0.05 were differentially expressed and genes with adjusted p-values greater than or equal to 0.05 are not differentially expressed.

```
diff=adjPval
diff[diff<0.05]<-'diffexp'
diff[diff!='diffexp'] <- 'no'
```

Computation Of Fold Change: The quantile statistics was first computed for the adjusted p-values. Genes with adjusted p-values greater than 0.05 had a significance score of zero (0). Genes with adjusted P-values less than or equal to 1st quantile P-values had a significance score of 0.9 and those falling in this category but with negative fold change had a significance score of -0.9. Genes whose adjusted p-values were less than 2nd quantile p-values had a significance score of 0.7 and those falling in this category but with negative fold change had a significance score of -0.7. Genes whose adjusted p-values were less than 3rd quantile p-values had a significance score of 0.5 and those falling in this category but with negative fold change had a significance score of -0.5. Genes whose adjusted p-values were less than or equal to 4th quartile p-values had a significance score of 0.3 and those falling in this category but with negative fold change had a significance score of -0.3.

Detection Of Up-Regulated And Down-Regulated Genes: Genes whose significance score is less than zero (0) are down regulated, genes whose significance score is greater than zero (0) are up-regulated while genes whose significance score are equal to zero (0) are neither up-regulated nor down-regulated.

Console

```
> head(geneTable, n=30)
   geneID      mean       foldCH       adjP diffExp regulation Significance Score
1    1915 14.488130  0.066621840 0.034852900 diffexp         up               0.3
2    2597 10.706179  0.185974690 0.043823447 diffexp         up               0.3
3    9906  6.688651 -0.026590029 0.035176212 diffexp       down              -0.3
4   56940  6.649619 -0.008682956 0.300334555       no         no               0.0
5    6234 13.135512 -0.060891080 0.101214415       no         no               0.0
6    9670  7.167485  0.007032312 0.401855619       no         no               0.0
7  653113  6.823008 -0.008142880 0.518451427       no         no               0.0
8   60312  6.646747  0.003703770 0.161761829       no         no               0.0
9   81620  6.582207  0.010658911 0.645731091       no         no               0.0
10   7442  6.542083 -0.032603192 0.025314788 diffexp       down              -0.3
11   4026  8.867540  0.124397022 0.011104112 diffexp         up               0.5
12   9134  6.487206  0.050158884 0.004133614 diffexp         up               0.5
13   3182 10.029795 -0.013281365 0.940881614       no         no               0.0
14 118426  6.855693  0.040430515 0.745319195       no         no               0.0
15  79753  7.961658  0.052214555 0.052806350       no         no               0.0
16   1308  6.796847  0.025500773 0.755795628       no         no               0.0
17 255877  6.872968 -0.027899158 0.422716342       no         no               0.0
18  29841  6.607280 -0.000235164 0.394418687       no         no               0.0
19 339760  6.828657  0.000443113 0.264093725       no         no               0.0
20  84239  6.651422  0.025374927 0.149902275       no         no               0.0
21   1730  6.546633 -0.007016975 0.092892029       no         no               0.0
22  10362  8.300044 -0.098919200 0.575478669       no         no               0.0
23    969  7.988244  0.508889085 0.019063665 diffexp         up               0.5
24  81893  7.156525 -0.036000589 0.906312018       no         no               0.0
25  23417  6.776647 -0.047852888 0.515539449       no         no               0.0
26  23074  6.732103  0.030653133 0.178704864       no         no               0.0
27 401010  6.811761  0.005162677 0.331730675       no         no               0.0
28  22801  6.853799 -0.008618296 0.961447611       no         no               0.0
29 353376  7.490028  0.063109200 0.829081470       no         no               0.0
30   9792  9.102956 -0.201187829 0.004938974 diffexp       down              -0.5
>
```

Fig. 2. List of up-regulated,down-regulated Genes

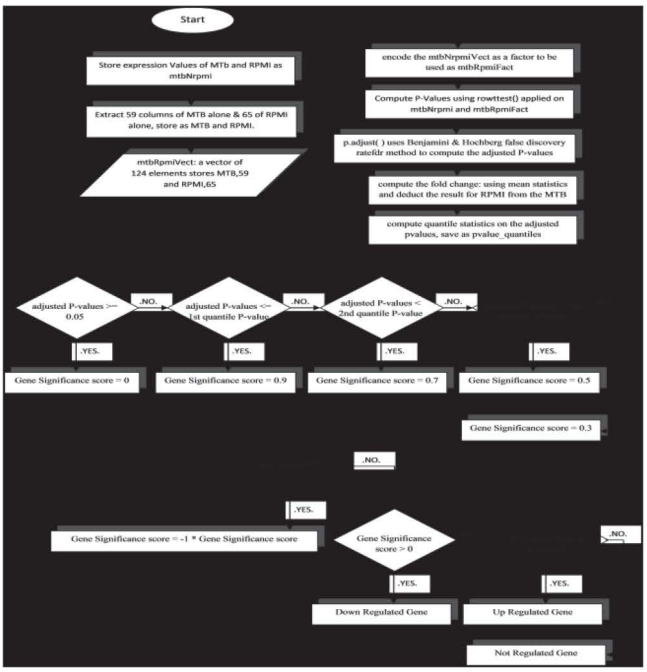


Fig. 3. Flow chart for the analysis of differentially expressed genes

Gene Set Enrichment Analysis and Pathview: After having a ranked list of up-regulated and down-regulated individual genes and their respective differential expression values, single-gene analysis may miss important effects on pathway. since cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.[26]. GSEA (Gene Set Enrichment Analysis) features a number of advantages when compared with single-gene methods. First, it eases the interpretation of a large-scale experiment by identifying pathways and processes. Rather than focus on high scoring genes (which can be poorly annotated and may not be reproducible), researchers can focus on gene sets, which tend to be more reproducible and more interpretable. Second, when the members of a gene set exhibit strong cross-correlation, GSEA can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes. Third, the leading-edge analysis can help define gene subsets to elucidate the results. [26][27].

Gene set collection file was loaded into the system for further analysis in order to have the direct mapping of each gene into the respective biological pathways. Gene set enrichment analysis was done using mean, and wilcoxon statistics on R/Bioconductor PIANO package by combining the results of multiple runs of gene set analyses. We computed the consensus scores based on rank aggregation for each directionality class and the result was visualized using a consensus heat map plot of the as shown in Fig. 4. This consensus heat map shows the biological pathways that were up-regulated or down regulated in the mixed directionality class, distinct directional class, and the non directional class. Since the use a consensus scoring approach, based on multiple

GSEA in combination with the directionality classes constitutes a more thorough basis for an enriched biological interpretation, we went further to implement the bioconductor Pathview function for pathway based gene data integration and visualization. It maps and renders our gene list on pathway graphs of the up-regulated pathways as discovered by the consensus heat map representation. Pathview automatically downloaded the pathway graph data, parsed the gene list data file, mapped it to the pathway, and render pathway graph, generating both native KEGG view and Graphviz views for pathways as shown in Fig. 5. below. Genes that are up-regulated are shown in red on the Tuberculosis pathway, while those that are down-regulated are shown on in green, and the non-regulated genes are shown in grey as shown in the KEGG view of the Tuberculosis pathway shown in Fig. 5. List of Potential Biomarkers found are displayed in TABLE I. below

TABLE I. BIOMARKER LIST

Gene Name	Gene Definition
IL-10	Interleukin 10
TNFα (TNFA)	Tumor necrosis factor superfamily, member 2
INFγR2	Interferon gamma receptor 2
IL10RB	Interleukin 10 receptor beta
CLEC4E	C-type lectin domain family 4 member E
FcRγ (FCER1G)	Fc receptor, IgE, high affinity I, gamma polypeptide
Src	SRC proto-oncogene, non-receptor tyrosine kinase
CD14	CD14 molecule
TLR2	Toll-like receptor 2
BID	BH3 interacting domain death agonist
STAT	Signal transducer & activator of transcription 1
Cyp27b1	25-hydroxyvitamin D3 1alpha-hydroxylase
v-aTPase	V-type H+-transporting ATPase subunit a
LAMP1(LAMP1_2)	lysosomal-associated membrane protein ½
C3b	complement component 3
Myd88	myeloid differentiation primary response protein MyD88
p38	p38 MAP kinase
CEBPB	CCAAT/enhancer binding protein

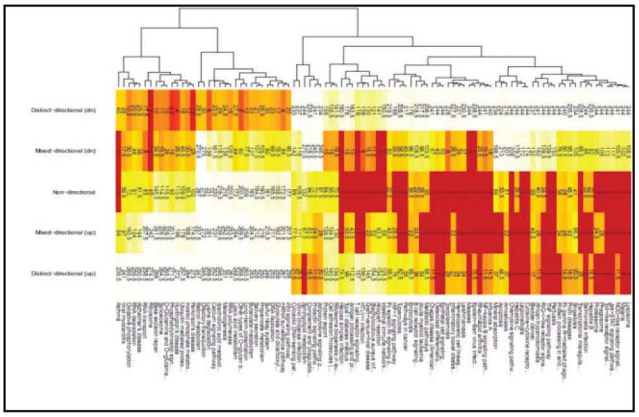


Fig. 5. The Consensus heatmap

III. RESULTS AND DISCUSSIONS

Biological markers are considered to be cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids. They can also include biological characteristics that can be objectively measured and evaluated as an indicator of pathogenic processes (Tudela et al., 2012). In the research, we used computational technique to identify potential biomarkers for possible of diagnosis of *M. tuberculosis* infection from the blood.

During the first contact, the microbial biochemical components, such as outer coat mannosylatedlipoarabinomannan, trehalosedimycolate and N-glycolymuramyl dipeptide, recognized by the Toll-like receptors (TLRs) of immune system trigger an intracellular signaling cascade, which leads to a phagocytic activity by the host immune system. This then result into engulfing of the microbe into cytosolic vesicles- the phagolysozomes and secretion of pro-inflammatory cytokines, such as tumor-necrosis factor alpha (TNFα) [31] TLR-2 has an intracellular domain that is activated to initiate a signaling cascade via adapter proteins such as MyD88, which results in the recruitment of interleukin-1 (IL-1) receptor-associated kinase (IRAK) 4. This leads to the activation of nuclear NF-kB, which is the main nuclear activator of proinflammatory cytokines. NK cells, which are large granular circulating lymphocytes, are attracted to the sites of bacterial infections, where they specialize in recognizing and destroying infected host cells. During this process they secrete interferon gamma (IFNγ), which activates macrophages, inducing them to secrete the cytokines IL-12, IL-15 and IL-18, which activate CD8+T-cells, thus forming the link to the adaptive immune system [18]

This study identified several of these acute phase and inflammatory proteins that are upregulated as part of host response to infection, especially tuberculosis. However, a number of this upregulated protein has also been found to be upregulated in other pathological conditions. This therefore limits their use in diagnosis of *M. tuberculosis* infection.

The alteration of host intermediary metabolism is a common occurrence during infection and where this has been characterized, it can serve as a good biomarker for the pathological condition. This research identified the up-

regulation of 25-hydroxyvitamin D3 1alpha-hydroxylase. This enzyme is part of the Cytp450 family and also participates in the synthesis of Vitamin D by catalyzing the hydroxylation of Calcifediol to calcitriol (the bioactive form of Vitamin D). There are evidences that vitamin D modulates macrophag responses to *M. tuberculosis* infection and there is a correlation between vitamin D deficiency and tuberculosis susceptibility [17] [19] [21] [31]. Although the mechanisms underlying vitamin D signaling and control of *M. tuberculosis*infection are not well understood the up-regulation of vitamin D synthesis in macrophages and dendritic cells upon exposure to *M. tuberculosis* has been reported earlier [28]

The combination of 25-hydroxyvitamin D3 1alpha-hydroxylase with other up-regulated proteins could sever as biomarkers for diagnosis of *M. tuberculosis* infection.

ACKNOWLEDGMENT

AM is supported by H3AbioNet via a NHGRI grant number U41HG006941

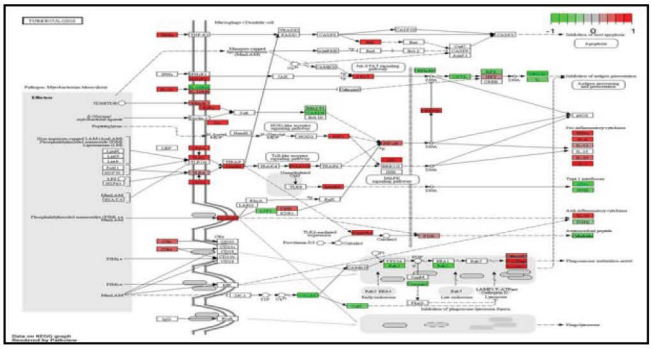


Fig. 6. Tuberculosis Pathway reflecting important

REFERENCES

[1] Sanne P. Smeekens et al. (2013). Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. Nature Communications, doi: 10.1038/ncomms2343

[2] World Health Organisation Global tuberculosis report 2014, ISBN 978 92 4 156480 9, NLM classification: WF 300

[3] Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C. Tuberculosis. Lancet. 2003;362(9387):887-99.

[4] Cegielski JP, Chin DP, Espinal MA, Frieden TR, Rodriquez Cruz R, Talbot EA,Weil DE, Zaleskis R, Raviglione MC. The global tuberculosis situation. Progress and problems in the 20th century, prospects for the 21st century. Infect Dis Clin North Am. 2002;16(1):1-58.

[5] Nunn P. The global control of tuberculosis: what are the prospects? Scand J Infect Dis. 2001;33(5):329-32.)

[6] Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review. Emerg Infect Dis. 2002;8(8):843-9.

[7] Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains. Trends Microbiol. 2002 Jan;10(1):45-52

[8] Hatfull GF, Jacobs WR Jr, editors. Molecular genetics of mycobacteria. Washington, D.C.: ASM Press; 2000.

[9] Mukherjee JS, Rich ML, Socci AR, Joseph JK, Viru FA, Shin SS, Furin JJ, Becerra MC, Barry DJ, Kim JY, Bayona J, Farmer P, Smith Fawzi MC, Seung KJ. Programmes and principles in treatment of multidrug-resistant tuberculosis. Lancet. 2004; 363(9407):474-81.)

[10] Mukherjee JS, Rich ML, Socci AR, Joseph JK, Viru FA, Shin SS, Furin JJ, Becerra MC, Barry DJ, Kim JY, Bayona J, Farmer P, Smith Fawzi MC, Seung KJ. Programmes and principles in treatment of multidrug-resistant tuberculosis. Lancet. 2004; 363(9407):474-81.)

[11] Dye C, Williams BG, Espinal MA, Raviglione MC. Erasing the world's slow stain: strategies to beat multidrug-resistant tuberculosis. Science. 2002 ; 295(5562):2042-6.)

[12] Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C.,

[13] Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J.Y. and Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10): R80. doi: 10.1093/nar/gkt111

[14] Väremo L, Nielsen J and Nookaew I (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Research, 41(8), pp. 4378-4391.

[15] Bassam H. Mahboub & Mayank G. Vats. (2013) Tuberculosis - Current Issues in Diagnosis and Management, ISBN 978-953-51-1049-1, 478 pages, DOI: 10.5772/56396

[16] Molicotti P1, Bua A, Zanetti S. (2014), Cost-effectiveness in the diagnosis of tuberculosis: choices in developing countries; Pubmed PMID: 24423709, 8(1):24-38. doi: 10.3855/jidc.3295.

[17] Kovalenko, V. M., Bagnyukova, T. V., Sergienko, O. V., Bondarenko, L. B., Shayakhmetova, G. M., Matvienko, A. V., & Pogribny, I. P. (2007). Epigenetic changes in the rat livers induced by pyrazinamide treatment. *Toxicol Appl Pharmacol, 225*(3), 293-299. doi: 10.1016/j.taap.2007.08.011

[18] Natarajan, K., Kundu, M., Sharma, P., & Basu, J. (2011). Innate immune responses to M-tuberculosis infection. *Tuberculosis, 91*(5), 427-431. doi: 10.1016/j.tube.2011.04.003

[19] Selvaraj, P., Harishankar, M., Singh, B., Banurekha, V., & Jawahar, M. (2012). Effect of vitamin D-3 on chemokine expression in pulmonary tuberculosis. *Cytokine, 60*(1), 212-219. doi: 10.1016/j.cyto.2012.06.238

[20] Shapira, Y., Agmon-Levin, N., & Shoenfeld, Y. (2010). Mycobacterium Tuberculosis, Autoimmunity, and Vitamin D. *Clinical Reviews in Allergy & Immunology, 38*(2-3), 169-177. doi: 10.1007/s12016-009-8150-1

[21] Skodric-Trifunovic, V., Blanka, A., Stjepanovic, M., Igniatovic, S., Mihailovic-Vucinic, V., Sumarac, Z., . . . Ilic, K. (2014). THE HEALTH BENEFITS OF VITAMIN D RELEVANT FOR TUBERCULOSIS. *Journal of Medical Biochemistry, 33*(4), 301-306. doi: 10.2478/jomb-2014-0032

[22] Ira Shah, Yashashree Gupta. (2015). Role of Molecular Tests for Diagnosis of Tuberculosis in Children, doi: 10.7199/ped.oncall.2015.16

[23] Ryota Suzuki,Hidetoshi Shimodaira1. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering, Vol. 22 no. 12 2006, pages 1540–1542, doi:10.1093/bioinformatics/btl117

[24] David Mount, 2004 Bioinformatics: Sequence and Genome Analysis, Second Edition ISBN 978-087969712-9

[25] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289–300

[26] Pablo Tamayo, et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles doi: 10.1073/pnas.0506580102

[27] Leif Väremo, Jens Nielsen and Intawat Nookaew. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods

[28] Sundaramurthy, V., & Pieters, J. (2007). Interactions of pathogenic mycobacteria with host macrophages. *Microbes and Infection, 9*(14-15), 1671-1679. doi: 10.1016/j.micinf.2007.09.007

[29] Sweet, L., & Schorey, J. S. (2006). Glycopeptidolipids from Mycobacterium avium promote macrophage activation in a TLR2- and MyD88-dependent manner. *J Leukoc Biol, 80*(2), 415-423. doi: 10.1189/jlb.1205702

[30] Tudela, P., Prat, C., Lacoma, A., Modol, J., Dominguez, J., Gimenez, M., & Tor, J. (2012). Biological markers for predicting bacterial infection, bacteremia, and severity of infection in the emergency department. *Emergencias, 24*(5), 348-356.

[31] Vankayalapati, R., & Barnes, P. (2009). Innate and adaptive immune responses to human Mycobacterium tuberculosis infection. *Tuberculosis, 89*, S77-S80.

[32] Venturini, E., Facchini, L., Martinez-Alier, N., Novelli, V., Galli, L., de Martino, M., & Chiappini, E. (2014). Vitamin D and tuberculosis: a multicenter study in children. *Bmc Infectious Diseases, 14*. doi: 10.1186/s12879-014-0652-7