BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 17(6)

Research Article

Up regulated virulence genes in *M. tuberculosis* H37Rv gleaned from genome wide expression profiles

Rohini Kumari & Pramod Katara*

Computational Omics Lab, Centre of Bioinformatics, University of Allahabad, Prayagraj, Uttar Pradesh - 211002 India; *Corresponding author; Pramod Katara E-mail: pmkatara@gmail.com & pkatara@allduniv.ac.in

Received May 13, 2021; Revised May 30, 2021; Accepted May 30, 2021, Published June 30, 2021

DOI: 10.6026/97320630017608

Declaration on Publication Ethics:
The authors state that they adhere to COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any misleading information to the publisher regarding this article.

Author responsibility:
The authors are responsible for the content of this article.

Declaration on official E-mail:
The corresponding author declares that an official e-mail from the institution is not available for all authors.

Abstract:
Identification of up regulated virulence genes in *M. tuberculosis* H37Rv using genome wide expression profiles is of interest in drug discovery for the disease. Hence, we report 17 up-regulated PPIN (Protein-Protein Interaction Network) enriched potential virulence linked genes using expression data available at the *Gene* Expression Omnibus (GEO) database for further consideration.

Keywords: Tuberculosis, expression profiling and virulence genes

Background:
M. tuberculosis H37Rv causes tuberculosis is a pathogen of global interest **[1].** Drugs (isoniazid, rifampicin, amikacin, capreomycin, kanamycin, etc.,) are effective in controlling the disease. However, the emergence of drug resistance tuberculosis (DR-TB) is a known paradox in this context **[2].** India, China, and the Russia share largest numbers of MDR-TB cases (47% of the global total) with increased mortality **[1, 3].** The M. tuberculosis genomes are available for multiple strains. The genome for the most explored strain H37Rv is known since 1998 **[4].** It is of interest to explore the genome data to glean valuable information on the disease causing virulence factors. Expression profiling using data at the Gene Expression Omnibus (GEO) helps to get clues for gene function **[5, 6].** Therefore, it is of interest to document data on the identification of virulence genes in *M. tuberculosis* H37Rv using genome wide expression profiles.

Materials & Methods:

Selection of gene expression data:
cDNA gene expression data for *M. tuberculosis H37Rv* was downloaded from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database. Considering the

objective of the work three-experiment series GSE93316 (20 GSM; source - pulmonary tuberculosis patients [7]), GSE49760 (24 GSM; source - whole blood from HIV- and HIV+ [8]), and GSE58466 (6 GSM; source - Replicating in Human Type II Alveolar Epithelial A549 cell line [9]) were used, all these three data series are based on the same platform, i.e., GPL4057.

Data normalization and filtration:
High throughput data need to be pre processed and normalized. Data was directly normalized using the data transformation (log2) GEO2R application at NCBI. Data filtration was done for the presence of empty/duplicated and incomplete sets manually **[10].**

Expression profiling:
The fold change (FC) parameter was used to select differentially expressed genes, further genes with less than two FC was filtered out from the further analysis. Grouping of similarly expressed genes, hierarchical clustering with average linkage was completed using the cluster software. In the end, for visualization of hierarchical clustering, Z-score (+/- 4) based Heatmap was developed through heatmapper *(*http://www.heatmapper.ca/*)*.

Virulence gene enrichment using the PPIN analysis:
The systemic view of functional linkage among the predicted genes and their protein-protein interaction network was drawn. For PPIN purpose protein IDs of all genes were used as an input for STRING database (*https://string-db.org/*), which offer PPIN based functional enrichment analysis. Gene ontology terms for molecular function, cellular component, and biological process were also observed for complete set genes, GO terms with FDR < 0.001 were considered for enrichment. To observe the sequence-based similarity of predicted and reported virulence genes, sequence alignment (blast) with VFDB database done (**Figure 1**).

Results and Discussion:
Transcriptome analysis helps to deduce genes affected by a process involved in particular conditions using expression data **[11].** Hence, differential gene expression analysis was completed to glean genes with greater than two-fold **[12].** Over expression of the genes indicates that they are actively involved in the pathophysiology of tuberculosis during the infection process. Genes with over expression of at least 70% GSM were considered as disease-related genes (442 in numbers). Gene-cluster, hierarchical clustering was done with an average linkage distance. Z-score-based heat-map of hierarchical clustering and across the experiment (GSE) was also completed. This indicates that all the up regulated DEG genes show considerable variability in their gene expression patterns in all datasets grouped in six key clusters (A-E, **Figure 2**).

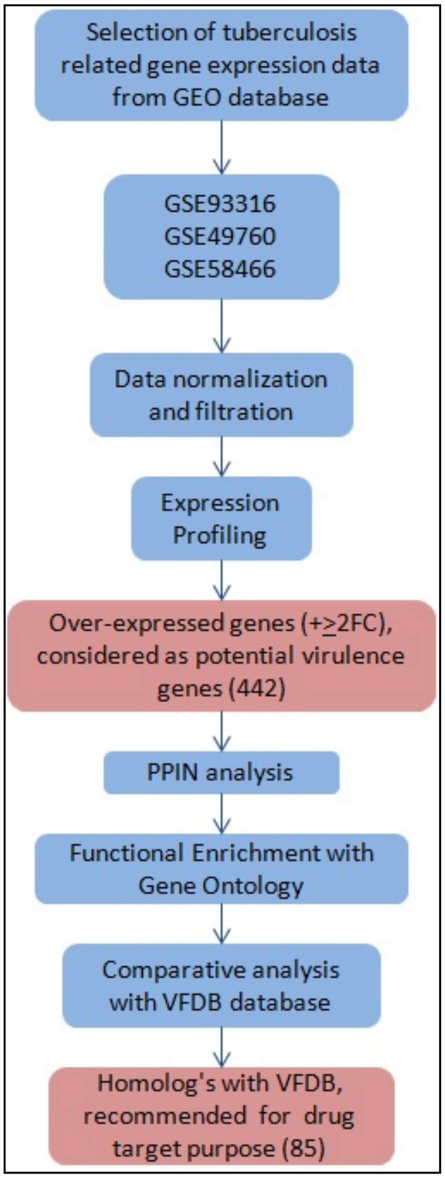
Selection of tuberculosis related gene expression data from GEO database
GSE93316
GSE49760
GSE58466
Data normalization and filtration
Expression Profiling
Over-expressed genes (+≥2FC), considered as potential virulence genes (442)
PPIN analysis
Functional Enrichment with Gene Ontology
Comparative analysis with VFDB database
Homolog's with VFDB, recommended for drug target purpose (85)

Figure 1: Flow chart, describing complete flow of the work

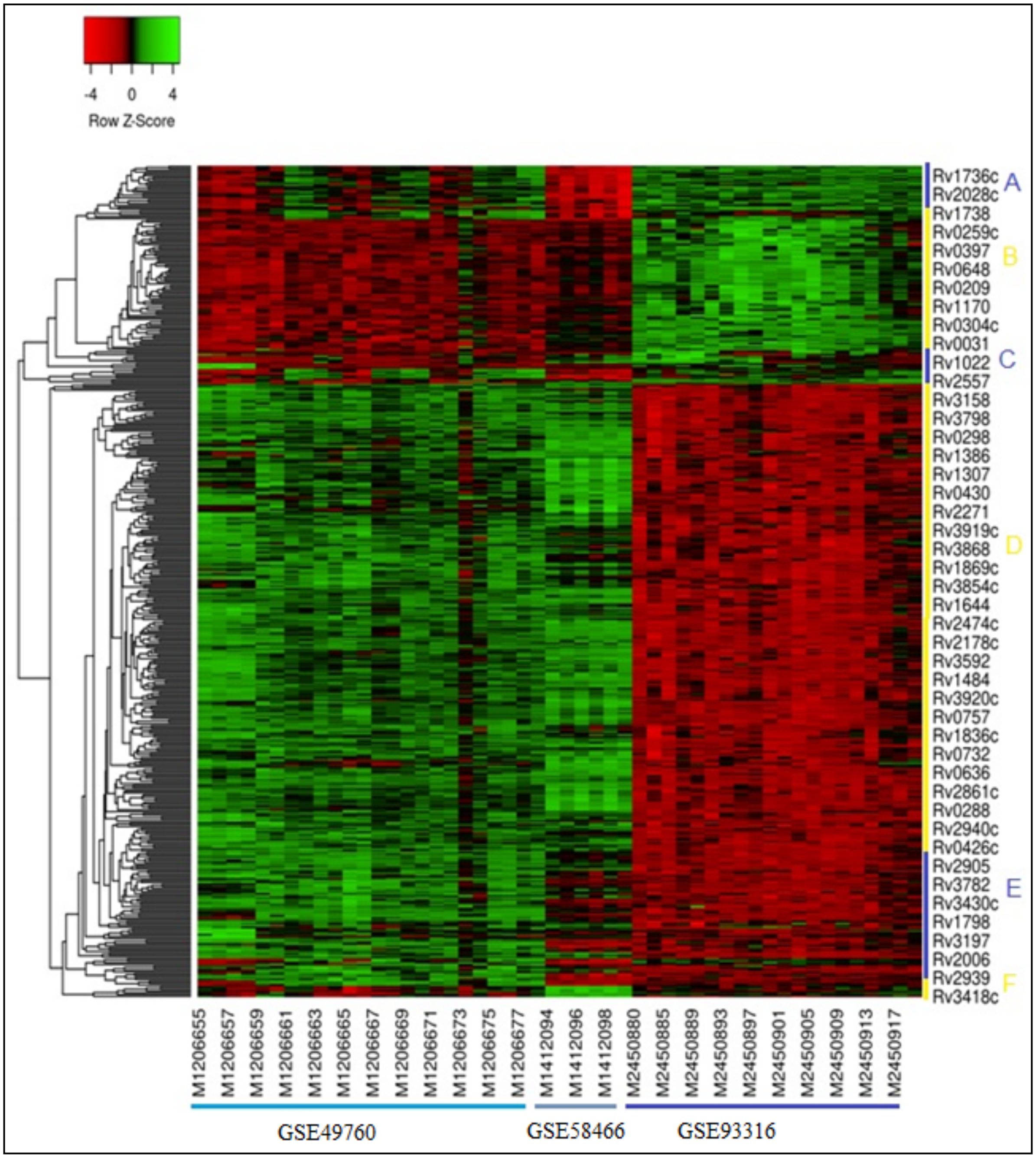

Figure 2: A heat-map for hierarchical clustering of over expressed genes in three different datasets showing gene expression patterns of differential expressed genes of *M. tuberculosis*. We show the co-expressed genes are clustered in six (A-E) major clusters.

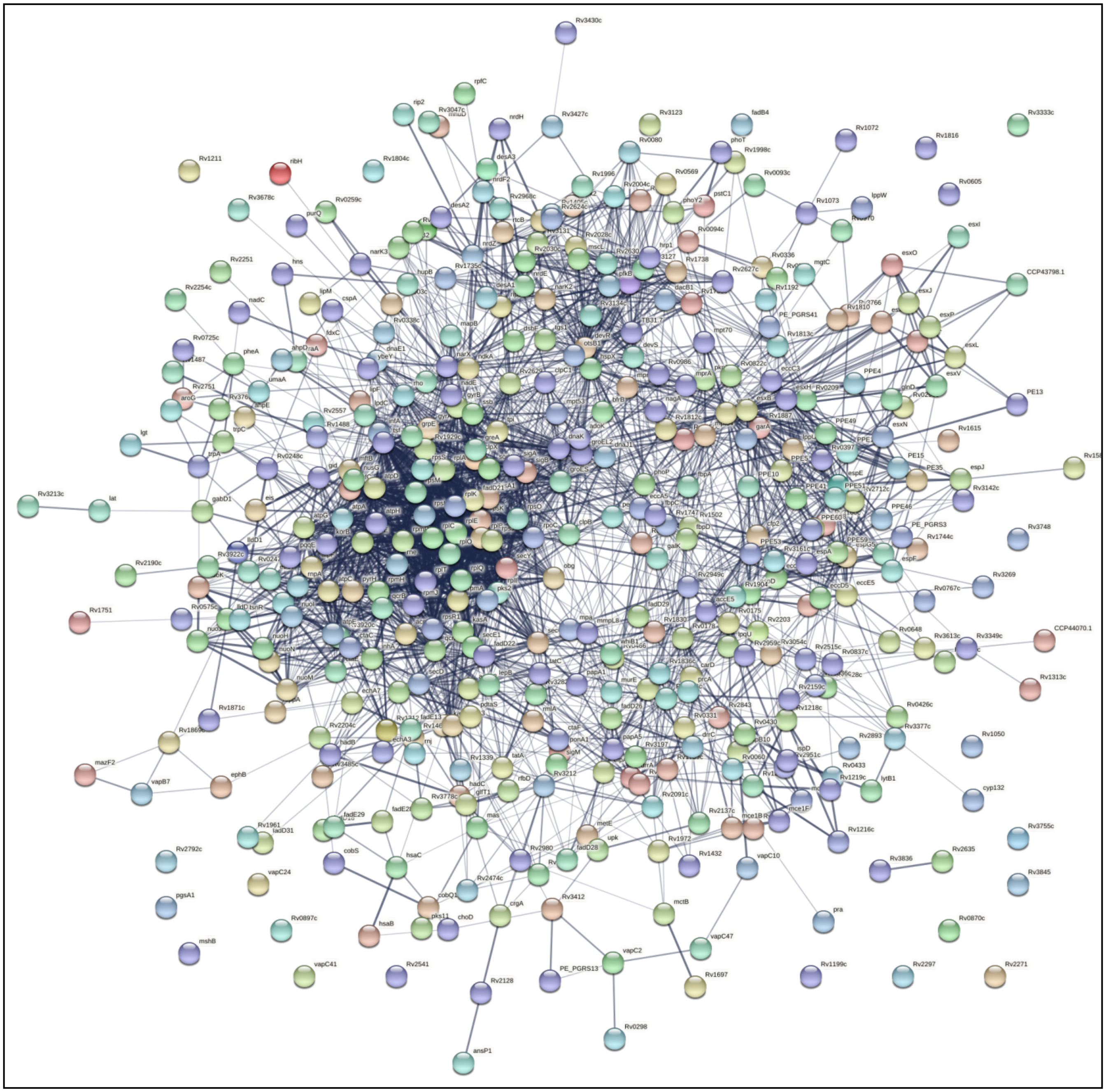

Figure 3: Protein-protein interaction network between potential virulence and associated gene product of *M. tuberculosis*. Nodes represent proteins and edges represent association.

Table 1: List of up regulated genes using gene annotation data

S.No.	Term	Gene count	Genes
1	Secreted	38	Rv0129c, Rv0288, Rv0350, Rv0440, Rv1038c, Rv1195, Rv1196, Rv1197, Rv1198, Rv1386, Rv1793, Rv1980c, Rv2004c, Rv2031c, Rv2346c, Rv2347c, Rv2376c, Rv2416c, Rv2430c, Rv2602, Rv2626c, Rv2780, Rv2875, Rv2878c, Rv2986c, Rv3614c, Rv3615c, Rv3616c, Rv3619c, Rv3620c, Rv3782, Rv3803c, Rv3804c, Rv3846, Rv3872, Rv3874, Rv3875, Rv3878
2	Ribosomal protein	21	Rv0055, Rv0056, Rv0640, Rv0641, Rv0682, Rv0700, Rv0701, Rv0704, Rv0705, Rv0716, Rv0722, Rv0723, Rv1298, Rv1643, Rv2441c, Rv2785c, Rv3456c, Rv3459c, Rv3460c, Rv3461c, Rv3924c
3	ATP synthesis	8	Rv1304, Rv1305, Rv1306, Rv1307, Rv1308, Rv1309, Rv1310, Rv1311
4	Hydrogen ion transport	8	Rv1304, Rv1305, Rv1306, Rv1307, Rv1308, Rv1309, Rv1310, Rv1311
5	rRNA-binding	16	Rv0055, Rv0056, Rv0640, Rv0641, Rv0682, Rv0701, Rv0704, Rv0705, Rv0716, Rv0723, Rv1298, Rv1643, Rv2785c, Rv3459c, Rv3460c, Rv3462c
6	Chaperone	12	Rv0350, Rv0351, Rv0352, Rv0384c, Rv0440, Rv1794, Rv2031c, Rv2115c, Rv2457c, Rv3418c, Rv3596c, Rv3874
7	Fatty acid biosynthesis	8	Rv0824c, Rv1094, Rv1484, Rv2244, Rv2245, Rv2246, Rv3229c, Rv3825c
8	Transport	39	Rv0284, Rv0638, Rv0732, Rv0820, Rv0821c, Rv0935, Rv0985c, Rv0986, Rv1177, Rv1304, Rv1305, Rv1306, Rv1307, Rv1308, Rv1309, Rv1310, Rv1311, Rv1440, Rv1698, Rv1736c, Rv1737c, Rv1739c, Rv1747, Rv1795, Rv1797, Rv2093c, Rv2094c, Rv2127, Rv2195, Rv2196, Rv2200c, Rv2587c, Rv2846c, Rv2936, Rv2937, Rv2938, Rv3155, Rv3158, Rv3823c
9	Virulence	27	Rv0288, Rv0462, Rv0931c, Rv0981, Rv0982, Rv1195, Rv1196, Rv1198, Rv1386, Rv1884c, Rv2109c, Rv2115c, Rv2346c, Rv2430c, Rv2623, Rv3131, Rv3409c, Rv3543c, Rv3544c, Rv3583c, Rv3614c, Rv3615c, Rv3616c, Rv3872, Rv3874, Rv3875, Rv3878
10	CF(1)	5	Rv1307, Rv1308, Rv1309, Rv1310, Rv1311
11	Ion transport	10	Rv0985c, Rv1304, Rv1305, Rv1306, Rv1307, Rv1308, Rv1309, Rv1310, Rv1311, Rv1698
12	RNA-binding	19	Rv0055, Rv0056, Rv0640, Rv0641, Rv0682, Rv0701, Rv0704, Rv0705, Rv0716, Rv0723, Rv1297, Rv1298, Rv1643, Rv2752c, Rv2785c, Rv3459c, Rv3460c, Rv3462c, Rv3923c
13	Fatty acid metabolism	14	Rv0824c, Rv1094, Rv1185c, Rv1484, Rv1665, Rv2244, Rv2245, Rv2246, Rv2930, Rv2941, Rv2948c, Rv2950c, Rv3229c, Rv3825c
14	Ubl conjugation	15	Rv0440, Rv0467, Rv0640, Rv0684, Rv1094, Rv1185c, Rv1308, Rv1996, Rv2029c, Rv2031c, Rv2115c, Rv2624c, Rv2752c, Rv3418c, Rv3846
15	Isopeptide bond	15	Rv0440, Rv0467, Rv0640, Rv0684, Rv1094, Rv1185c, Rv1308, Rv1996, Rv2029c, Rv2031c, Rv2115c, Rv2624c, Rv2752c, Rv3418c, Rv3846
16	Translocation	6	Rv0638, Rv0732, Rv1440, Rv2093c, Rv2094c, Rv2587c
17	CF(0)	4	Rv1304, Rv1305, Rv1306, Rv1307

Heatmap shows that all genes almost show similar expression variation in GSE49760 and GSE58466, which shows the gene expression in 'whole blood from HIV +/- patients' and H'uman Type II Alveolar Epithelial A549 cell line.' Interestingly, GSE93316, which characterize expression from pulmonary tuberculosis patients, shows an almost opposite expression pattern compare to above mention GSE's. The difference in the sample source, as GSE9336, utilized the sample from patients and GSE49760 and GSE58466 took the transcriptome sample from lab sources, i.e., cultured blood and cell lines, thus differ in the biological milieu **[13].** Functional association among the predicted genes was observed through protein-protein interaction network (PPIN) provided by STRING. Graphical representation of the predicted PPIN provides a weight-based functional linkage among the given proteins and helps to understand the systemic view of biological processes **[14].**

Network statics of the predicted PPIN shows 2938 edges between 440 protein nodes with an average of 13.3 degrees per node (figure 3). As the PPIN is a scale-free network whose degree distribution follows power low, where few nodes show very less degree and few are with very high **[15]**. Predicted network visibly present that network-nodes shows a range of degree distribution which ranging from 01 to 67. The predicted network is also showing the presence of multiple sub-networks, out of them, four are very dense and show high order functional linkage among the member nodes. In total 21 nodes are there in a network with > 50 degrees, top of them are Rv3418c (67), Rv1307 (66), Rv0732 (61), and Rv0704 (60). Gene ontology results from STRING were also observed for the mining of knowledge for considered proteins; all three types of gene ontology have been observed for 442 genes. Molecular function ontology, in total, provides 20 function terms with FDR < 0.001; the most observed gene counts have belonged to binding [protein binding, rRNA binding, enzyme binding, RNA binding, ribonucleotide binding, zymogen binding and cyclic compound binding]. Other important terms belong to transporter activities (proton transmembrane transporter activity; proton-transporting ATP synthase activity, rotational mechanism; active transmembrane transporter activity; transmembrane transporter activity; ATPase-coupled transmembrane transporter activity; inorganic molecular entity transmembrane transporter activity). As Molecular functions generally correspond to activities that can be performed by individual gene products, i.e., a protein/ RNA, here it is clear that these predicted virulence genes mainly belong to binding and transportation **[16]**. Cellular component terms indicate that predicted genes mainly performed their function in 17 components.

The major cellular component term, which shared by more than 100, genes are cell periphery, plasma membrane, external encapsulating structure, cell wall, extracellular region, and cytoplasm. Biological process terms indicate that predicted virulence genes are mainly involved in 21 processes. The most common observed terms are cellular process, growth, potential metabolic process, and response to abiotic stimulus, metabolic process, and transport. All these processes process are very crucial and well reported for their connection with infection and pathogenecity **[17-20]**. To get more functional insights about the genes, their term description has been mined which more clearly present the functional aspects of the genes (**Table 1**). Overall 17 terms were observed for the considered genes, as observed through gene ontology, mainly these terms belong to secretion, binding, and transport. All these three terms are well known to relate to microbial pathogenicity, interestingly out of all, 27 genes are already linked to virulence **[21, 22].**

On the basis of differential gene expression followed by Gene ontology and annotation, it is hypothesized that considered genes participate in disease physiology. Gene ontology and annotation terms are not available for all considered genes. In such conditions, sequence homology-based comparative analysis with VFDB database, collections of virulence genes across different pathogens add credence in predicted virulence genes **[23]**. For this purpose, all genes were compared with this database for their sequence-based similarity search. VFDB core datasets were chosen as target databases for comparison. Result analysis of sequence comparison shows that out of 442 considered genes 27 are already documented in VFDB database and 58 shows considerable similarities with >80% sequence coverage.

Conclusion:

We report 17 up regulated PPIN (Protein-Protein Interaction Network) enriched virulence linked genes using expression data available at the Gene Expression Omnibus (GEO) database for further consideration in the context of *M.* tuberculosis infected DR-TB, MDR-TB, and XDR-TB.

Acknowledgment:

Rohini Kumari, acknowledges UGC for financial support during this study (RGNF-2014-15-SC-BIH-62160).

Conflict of interest: There is no conflict of interest exists.

References:

[1] https://www.who.int/news-room/fact-sheets/detail/tuberculosis
[2] Dheda K *et al. Lancet Respir Med*. 2014 **2:321** [PMID: 24717628]
[3] Prasad R *et al. Indian J Med Res*. 2017 **145:271** [PMID: 28749390]
[4] Cole ST *et al. Nature* 1998 **393:537** [PMID: 9634230]
[5] Katara P *et al. Bioinformation* 2010 **5:31** [PMID: 21346876]
[6] Lowe R *et al. PLoS Comput Biol*. 2017 **13:e1005457** [PMID: 28545146]
[7] Sharma S *et al. PLoS One* 2017 **12:e0173508**. [PMID: 28282458]
[8] Ryndak MB *et al. PLoS One* 2014 **9:e94939** [PMID: 24755630]
[9] Ryndak MB *et al. PLoS One* 2015 **10:e0123745** [PMID: 25844539]
[10] Beyene J *et al. BMC Proc*. 2007 **S150** [PMID: 18466495]
[11] Tarca AL *et al. Am J Obstet Gynecol*. 2006 **195:373** [PMID: 16890548]
[12] Makhijani RK *et al. IET Syst Biol*. 2018 **12:213** [PMID: 30259866]
[13] Choi JK and Kim SC *Genetics* 2007 **175:1607** [PMID: 17237506]
[14] Szklarczyk D *et al. Nucleic Acids Res*. 2019 **47(D1)** [PMID: 30476243]
[15] Barabási AL *Science* 2009 **325:412** [PMID: 19628854]
[16] Jain M & Cox JS. *PLoS Pathog* 2005 **1:e2** [PMID: 16201014]
[17] Vincze E *et al. Acta TubercScand*. 1960 **38:26** [PMID: 13842369]
[18] Fitzgerald RJ and Bernheim F *Am Rev Tuberc*. 1948 **58:210** [PMID: 18884138]
[19] Rosa A. *Riv Patol Clin* 1962 **17:123** [PMID: 13974733]
[20] Smith LJ *et al. Nucleic Acids Res*. 2017 **45:6600** [PMID: 28482027]
[21] Ganguly *et al. Tuberculosis* (Edinb) 2008 **88:510** [PMID: 18640874]
[22] Geffken *et al. Methods Mol Biol*. 2015 **015:1285** [PMID: 25779328]
[23] Liu B *et al. Nucleic Acids Res*. 2019 **47:D687** [PMID: 30395255]

Edited by P Kangueane
Citation: Kumari & Katara, Bioinformation 17(6): 608-615 (2021)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article for FREE of cost without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Agro Informatics Society

BIOMEDICAL INFORMATICS

BMIS

AGIS

OPEN ACCESS GOLD

Biomedical Informatics Society

since 2005

BIOINFORMATION
Discovery at the interface of physical and biological sciences

indexed in

PMC

PubMed

INDEXED IN EMERGING SOURCES CITATION (Web of Science) CLARIVATE ANALYTICS

WEB OF SCIENCE

EBSCO

Web of Science Group

Crossref

doi®

ResearchGate

RG

publons