Current Trends in Biotechnology and Pharmacy
Vol. 15 (6) 67 - 79, 2021, ISSN 0973-8916 (Print), 2230-7303 (Online)
10.5530/ctbp.2021.6.13

An Integrative Bioinformatics Analysis for Identifying Hub Genes in Human Immunodeficiency Virus

Kartik Jadeja[1], Anukriti Verma[1], Bhawna Rathi[1]*

[1]Amity Institute of Biotechnology, J-3 Block, Amity University Campus
Sector-125, Noida - 201313 (U.P.), India.
Corresponding author: brathi@amity.edu

Abstract

The inability of modern medicine to find a cure for human immunodeficiency virus (HIV) since its discovery as the causative agent of acquired immunodeficiency syndrome (AIDS), has made HIV as one of the most dreadful pathogens of this century. The present study presents the putative drug targets and biomarkers of HIV using bioinformatic tools. Microarray data analysis of datasets that involved HIV cases vs controls was done in order to scrutinize the differentially expressed genes (DEGs). From this data a total of 57 common up-regulated DEGs and 108 common down-regulated genes were obtained for HIV. The functional annotation and pathway analysis of the DEGs was done in order to get the biological processes and pathwaysin which the DEGs were enriched. From the analysis of the network the significant genes that were obtained for HIV datasets were 'Protein Kinase C Alpha (**PRKCA**)', 'Signal Transducer and Activator of Transcription 5B (**STAT5B**)', 'Krupple Like Factor-6 (**KLF6**)', 'Granzyme K (**GZMK**)', 'T-cell Immunoreceptor With Ig and ITIM Domains (**TIGIT**)', 'Ectonucleoside Triphosphate Diphosphohydrolase 1 (**ENTPD1**)', 'Regulator of G protein signaling 1 (**RGS1**)', 'CD48 Molecules (**CD48**)', 'Cytotoxicity and Regulatory T-cell Molecule (**CRTAM**)', and 'Neutrophil Cytosolic factor 4 (**NCF4**)'. The topological analysis was done for these genes for better understanding of their interaction.

The present study also suggests that the significant DEGs and Biological processes and pathways that accompany them might have the potential to be exploited as possible drug targets and biomarkers in the diagnosis, prognosis as well as treatment of HIV and its comorbidities and warrants for further experimental validation.

Keywords HIV, Microarray analysis, PPI interactions, Functional annotation, Pathway analysis, Topological analysis.

Introduction

The human immunodeficiency virus (HIV) is a virus that is responsible for causing Acquired immunodeficiency syndrome (AIDS).There are roughly around 35 million people that have been reported to be infected by HIV with an estimated 2 million cases of HIV infection occurring annually around the globe (1). Generally, a human body has immune system that attacks the invading microbes or any foreign particles. White blood cells are present in our immune system that are responsible for protecting us from infections. CD4+ cells present in the white blood cells that are also known as T cells or helper cells. Infection due to HIV takes advantage of body's immune system and causes several health problems to such extend that it even leads to fatality of the infected person. HIV causes inability to protect against diseases and also causes reduction in CD4 cells. AIDS has no cure but the disease can be stalled so that a patient can stay healthy for longer period of time (2).

Due to the lack of medication for HIV it has become necessary to identify and evaluate the presumptive drug targets as well as biomarkers and enhance the current therapy and possible drug for HIV treatment (3). High-throughput techniques have hustled the discovery of therapeutic targets of drugs as well as biomarkers in diseases (4). Computational analysis is a high-throughput method that designates presumed targets for the association as well as defines a directed pathway for analyzing the experimental framework of the study.

For this sort of analysis, high-throughput data of patient samples obtained from databases of National Centre for Biotechnology Information (NCBI) as well as ArrayExpress of European Bioinformatics Institute (EBI) that contains extensive quantitative assessment of transcriptomic information and/or gene expression can be used (5). The ArrayExpress Archive of Functional Genomics Data is regarded as one of the preeminent international repositories for functional genomics high throughput data that stores functional genomics data derived from microarray-based experiments and high throughput sequencing (HTS) (6).

The primary objective of this study was to analyze the significant genes, functions and pathways of HIV and its associative disease that might play a role as a potential trigger in causing HIV infection. Therefore, high-throughput transcriptomic datasets of HIV were employed to execute a comparative transcriptomic analysis between the diseased subjects and the healthy ones. Functional annotation, pathway and network analysis of the differentially expressed genes (DEGs) has been performed providing a deep insight for presumed drug targets and biomarkers of HIV.

Materials and Methods

Microarray data selection

ArrayExpress of European Bioinformatics Institute (EBI) that contains extensive quantitative assessment of transcriptomic information and/or gene expression which is also linked with Gene Expression Omnibus (GEO) database of National Centre for Biotechnology Information (NCBI) (5), was iterated for retrieval of gene expression profiles of human immunodeficiency virus (HIV). The criteria laid down for selection of the datasets were as follows

(1) Datasets with HIV case vs control.

(2) Patient cohorts not undergoing any sort of treatment.

(3) Datasets with RAW files.

(4) Datasets consisting of replicates.

(5) Datasets whose transcriptional analysis has been performed using Affymetrix microarray.

(6) Datasets that are published in journals.

2.2 Pre-processing

The pre-processing steps were operated using Bioconductor which is an open source as well as open development software, based on the R programming language (7). Screening of the differentially expressed genes (DEGs) was done during the pre-processing step, within every dataset. Determination of expression levels, background correction, and normalization was performed using a robust multi-array average (RMA) analysis method (8).

2.3 Statistical analysis and identification of differentially expressed genes

Linear Models for Microarray Data (LIMMA) is an R package in which linear models are used to analyze microarray experiments (9). The information across genes is borrowed using Empirical Bayes and other shrinkage methods, thus, making the analyses stable even for experiments that has small number of arrays (10). The adj.p.Val, p.Value, , t- value, B-value as well as logFC of every single gene for all the obtained datasets were calculated. t-statistics was employed with log-odds of the differential expression simultaneously. Genes with 1.5-fold change and a p value less than 0.05 [$p < 0.05$ and $FC \geq 1.5$ ($|\log2\ FC| \geq 0.58$)] as the cutoff criterion for the up-regulated and down-regulated DEG's were selected for subsequent analysis. Since p-value related to the probability that chance could have generated the variation observed, a lesser p-value indicates higher significance of results.

2.4 Functional Annotation, pathway analysis and gene ontologies

The functional annotation of the DEG's for HIV dataset was carried out using the DAVID (The Database for Annotation, Visualization and Integrated Discovery [https://david.ncifcrf.]) gene functional classification tool as mentioned in figure 1.

It annotates lists of gene or protein identifiers rapidly and summarizes according to shared categorical data for biochemical pathway membership, gene ontology, and protein domain (11). The Kyoto Encyclopedia of Genes and Genomes (KEGG [https://www.genome.jp/kegg/]) is a publicly available knowledgebase of genomic and molecular information was iterated for more advanced characterization of the DEG's (12).

2.5 Network analysis

The DEGs were integrated into network using the version 11.0 of the STRING. The aim of the STRING database (https://string-db.org/) is to provide a crucial evaluation as well as integration of protein-protein interactions (PPIs). This also includes their functional as well as physical associations, which are also regarded as indirect and direct association respectively. The latest version of STRING i.e v 11.0 calls for novel and scalable algorithms for the purpose of transferring information on interaction among organisms, it covers more than double the number of organisms covered as compared to the previous version bringing the number up to 5090 (13). The network created using STRING was evaluated on version 2.8 of the CYTOSCAPE app. It works on a principle of a network graph, with biological entities such as genes, proteins or cells which are represented as nodes and biological interactions are represented as edges between nodes. The data is unified with the network on the bases of various attributes that maps node or edge to explicit data values such as protein functions or gene expression levels (14). It was used for the evaluating the topological properties of the network such as nodes, edges, number of directed edges, closeness centrality, neighborhood connectivity, clustering coefficient, betweenness centrality, average shortest path length, stress, degree, radiality, eccentricityand topological coefficient. The MCODE app of CYTOSCAPE was used to determine the 10 significant genes amongst the network of DEGs obtained for HIV dataset. The topological properties of the network of these 10 genes as well were also determined which helped us in understanding the mechanism of the network.

3 RESULTS

3.1 Microarray data selection

Fifteen datasets (GSE76403, GSE10038, GSE56619, GSE44460, GSE17372, GSE30536, GSE28160, GSE16593, GSE28686, GSE18468, GSE17189, GSE14278, GSE14245, GSE7224 and GSE9927) which contained gene expression for HIV or diseases occurring in association with HIV were obtained based on the criteria mentioned in section 2.1.

The criteria were set in order to avoid any inconsistency. The detailed information is mentioned in table 1.

3.2 *Pre-processing*

Robust Multi-array Average (RMA) is an algorithm that was adopted inorder to shape an expression matrix from the Affymetrix data. The raw affymetrix intensity values are background corrected, log2 transformed and then quantile normalized. To provide similar empirical distribution of intensity to each array is the major goal of the quantile normalization. Next, to the normalized data a linear model is fit in order obtain an expression measure for each probe set on each array (15). This is represented as follows:

$Yijn = \mu in + \alpha jn + \varepsilon inj$,

where,

$Yijn$ = The quantile normalized probe value

μin =Log scale expression level (RMA measure).

αjn = Probe affinity affect.

εinj = Independent identically distributed error term (with mean 0) +.

3.3 Statistical analysis of differentially expressed genes

Moderated t-statistic is the elemental statistic used for significance analysis, which was computed for each probe.

The empirical Bayes shrinkage method was engaged in order to moderate the standard error across the genes. This has the effect of borrowing information from the ensemble of genes to aid with inference about each individual gene (10).

A number of summary statistics are given for the probes.

Table 1: Datasets of HIV retrieved from ArrayExpress

Accession	Title	Assays
E-GEOD-76403	Circulating leukocyte transcriptional signatures in HIV patients with impaired gas exchange	39
E-GEOD-10038	Upregulation of Expression of Matrix Metalloproteinases in Alveolar Macrophages of HIV1+ Smokers with Early Emphysema	11
E-GEOD-56619	Gene expression profiles of peripheral blood mononuclear cells from HIV and HCV coinfected patients before and after 24 weeks of combined antiretroviral therapy	16
E-GEOD-44460	Induction of IL-17+ T-cells by HIV-Tat protein is mediated via Vascular Endothelial Growth Factor Receptor-2	12
E-GEOD-17372	Molecular Classification of AIDS-Related Lymphomas [including third-party data]	17
E-GEOD-30536	Expression data from IFN alpha 2-treated macrophages infected with HIV	17
E-GEOD-28160	Significant Effects of Antiretroviral Therapy on Global Gene Expression in Brain Tissues of Patients with HIV-Associated Neurocognitive Disorders	35
E-GEOD-16593	Differential expression associated with GB virus C in HCV/HIV co-infection	10
E-GEOD-28686	Peripheral blood RNA expression profiling in illicit methcathinone users reveals effect on immune system	40
E-GEOD-18468	Fatigue-related HIV disease gene-networks identified in CD14+ cells isolated from HIV-infected patients	15
E-GEOD-17189	Transcription profiling by array for molecular classification of human AIDS-related lymphomas	17
E-GEOD-14278	Transcription profiling by array of human CD4+ T cells from HIV-resistant and HIV-susceptible individuals	18
E-GEOD-14245	Multiple Salivary Biomarkers for Early Detection of Pancreatic Cancer	24
E-GEOD-7224	Transcription profiling of human tonsil and oral epithelia	13
E-GEOD-9927	Transcription profiling of human CD4+ T cells from HIV infected individuals vs. controls reveals type I interferon-mediated disruption of T cell dynamics	20

Moderated t-statistic is represented by the column t. p-value is the associated p value is the associated p value after adjustment for multiple testing. The *p* value was computed using Benjamini and Hochberg's method to contain the false discovery rate, they are called *Adj.p.value*or *fdradj.p.value*. The B-statistic or B is the log-odds that the gene is differentially expressed hence it was employed to evaluate the differential expression of the genes. *logFC*is the log-folds change which is calculated using hurdle model component. It was engaged to estimate log2 folds change between the healthy and the diseased subjects for the purpose of distinguishing the up- as well as the down-regulated values (16).

To a given gene h, LIMMA fits a linear model to test the null hypothesis. After fitting a linear model if desired, the empirical Baves shrinkage method assumes an inverse Chi-square prior for the σ_h^2 σ_h^2 with mean $s_0^2 s_0^2$ and degrees of freedom $f_0 f_0$. The posterior values for the residual variances are given by:

$$\tilde{s}_h^2 = \frac{f_0\, s_0^2 + f_h\, s_h^2}{f_0 + f_h} \tilde{s}_h^2 = \frac{f_0\, s_0^2 + f_h\, s_h^2}{f_0 + f_h}$$

Where,

$\sigma_h^2 \sigma_h^2$ is the residual value of $h^{th} h^{th}$ gene

$s_0^2 s_0^2$ is the sample value of the $h^{th} h^{th}$ gene

$f_h f_h$ is the residual degrees of freedom for the $h^{th} h^{th}$ gene.

3.4 Identification of differentially expressed genes

Differentially expressed genes (DEGs) that responds to a signal were obtained as a result of the statistical evaluation of the datasets.

They also play an important role in the regulation of genes (17). A total of 3,750 DEGs were identified, 250 DEGs were selected per datasets (Table 1). log2 fold change was taken into account along with *p* value in order to obtain highly robust DEGs. 1.5-fold change was illustrated by a log2 ratio of 0.5 for up-regulation or − 0.5 for down-regulation. The *p* values below 0.05 were considered substantial (17). Among the identified DEGs, 1,775 were up-regulated genes and 1,975 were down-regulated (Table 2 and 3). From the given DEGs common genes were identified and selected, which gave 56 co-upregulated genes and 105 co-downregulated genes as shown in table. The obtained probes were annotated using the GeneAnnot database (https://genecards.weizmann.ac.il/geneannot/index.shtml). The obtained common up- and down-regulated genes were utilized for further analysis.

3.5 Functional annotation and pathway analysis

To gain a better understanding, all common DEGs were uploaded to the DAVID database. DAVID database was used for functional annotation of the DEGs. The up-regulated DEGs were significantly enriched in BPs, which includes 'Immune response', 'Signal transduction', 'Negative regulation of transcription from RNA polymerase II promoter', 'Regulation of cell growth', 'Transcription, DNA templated', 'Spermatogenesis', 'Wntsignaling pathway', 'Angiogenesis'. The Down-regulated DEGs were significantly enriched in 'Immune response'. 'Response to virus', 'Defense response to virus', 'Negative regulation of viral genome replication', 'Interferon-gamma mediated pathway', 'Response to drug', 'Type-1 interferon signaling pathway', 'Signal transduction', 'Apoptic process', 'Mitotic nuclear division', 'Cell proliferation', 'Viral process', 'DNA repair', 'Regulation of transcription . DNA templated', 'Protein complex assembly', 'Negative reulation of transcription from RNA polymerase promoter', 'Cellular response to

Table 2: Common up-regulated DEGs

Gene symbol	GEO accession	Gene title
TSHZ2	GSE76403, GSE56619, GSE28686	teashirt zinc finger homeobox 2
MID2	GSE76403. GSE28686	midline 2
ZFP28	GSE76403. GSE9927	ZFP28 zinc finger protein
IL7R	GSE76403, GSE18468	interleukin 7 receptor
LTBP3	GSE76403, GSE56619	latent transforming growth factor beta binding protein 3
TRIM2	GSE76403, GSE44460, GSE7224, GSE9227	tripartite motif containing 2
TXNIP	GSE56619, GSE16593	thioredoxin interacting protein
SH3YL1	GSE76403, GSE56619	SH3 and SYLF domain containing 1
PDE8A	GSE76403. GSE9927	phosphodiesterase 8A
NOG	GSE76403, GSE28686	noggin
CD40LG	GSE76403, GSE28686	CD40 ligand
AK5	GSE76403, GSE28686	adenylate kinase 5
SFMBT1	GSE76403, GSE44460	Scm-like with four mbt domains 1
TRABD2A	GSE76403, GSE28686	TraB domain containing 2A
MAL	GSE76403, GSE56619	mal, T-cell differentiation protein
EDAR	GSE76403, GSE28686, GSE18468	ectodysplasin A receptor
PDK4	GSE10038, GSE44460	pyruvate dehydrogenase kinase 4
TGFB3	GSE10038, GSE28160	transforming growth factor beta 3
CYYR1	GSE10038, GSE28160	cysteine and tyrosine rich 1
EXOG	GSE56619, GSE28160	endo/exonuclease (5'-3'), endonuclease G-like
MBP	GSE56619, GSE28160	myelin basic protein
MVK	GSE44460, GSE28160	mevalonate kinase
PRKCA	GSE44460, GSE9927	protein kinase C alpha
ATP6V0D2	GSE44460, GSE17189	ATPase H+ transporting V0 subunit d2
FLVCR2	GSE44460, GSE17189	feline leukemia virus subgroup C cellular receptor family member 2
FGF9	GSE44460, GSE9927	fibroblast growth factor 9
CCDC65	GSE4460, GSE17189	coiled-coil domain containing 65
EPHB1	GSE28160, GSE18468	EPH receptor B1
SIAE	GSE28160, GSE7224	sialic acid acetylesterase
MED25	GSE28160, GSE16593	mediator complex subunit 25
HBB	GSE28169, GSE17189	hemoglobin subunit beta
PRKAB1	GSE28160, GSE16593, GSE14278	protein kinase AMP-activated non-catalytic subunit beta 1
CCND3	GSE16593, GSE14278	cyclin D3
PAG1	GSE16593, GSE7224	phosphoprotein membrane anchor with glycosphingolipid microdomains 1
CD48	GSE16593, GSE7224	CD48 molecule
MYLIP	GSE16593, GSE9927	myosin regulatory light chain interacting protein
TAGAP	GSE16593, GSE7224, GSE9227	T-cell activation RhoGTPase activating protein
GZMK	GSE16593, GSE7224	granzyme K
KLF6	GSE16593, GSE14278	Kruppel like factor 6
MARCKSL1	GSE16593, GSE28686	MARCKS like 1
NCF4	GSE16593, GSE7224	neutrophil cytosolic factor 4
VIM	GSE16593, GSE14245, GSE9927	vimentin
AOC3	GSE16593, GSE28686	amine oxidase, copper containing 3
ZSCAN18	GSE28686, GSE9927	zinc finger and SCAN domain containing 18
SLC16A10	GSE28686, GSE9927	solute carrier family 16 member 10
NOV	GSE28686, GSE14278	nephroblastoma overexpressed
LMO7	GSE18468, GSE7224	LIM domain 7
CHPT1	GSE18468, GSE9927	choline phosphotransferase 1
PDLIM3	GSE18468, GSE7224	PDZ and LIM domain 3
SPEF2	GSE18468, GSE17189	sperm flagellar 2
GLCCI1	GSE17189, GSE7224	glucocorticoid induced 1
CDH26	GSE17189, GSE7224	cadherin 26
H2AFY	GSE14278, GSE14245	H2A histone family member Y
WDR48	GSE14278, GSE14245	WD repeat domain 48
LINC01215	GSE14245, GSE7224	long intergenic non-protein coding RNA 1215

DNA damage stimulus'. KEGG pathway was utilized for the pathway analysis of the DEGs.

Ensuing KEGG pathway enrichment analysis, the common up-regulated DEGs were identified to be significantly enriched in 'Cytokine-cytokine receptor interaction', 'T-cell receptor pathway', 'Natural killer mediated cytotoxicity', 'Fc gamma R-mediated phagocytosis', 'Purine metabolism', 'Biosynthesis of antibiotics', 'Fox0 signalling pathway'. The common down- regulated DEGs were enriched in 'Cell cycle', 'Herpes simplex infection', 'p53 signaling pathway', 'Influenza A', 'Measles', 'Pyrimidine metabolism', 'NOD-like receptor signaling pathway', 'Progesterone-mediated oocyte maturation', 'Viral carcinogenesis', 'Hepatitis C', 'Hepatitis B', 'Cellular senescence', 'Glutathione metabolism', 'Oocyte meiosis', 'Epstein-Barr virus infection'. The identified functions and pathways might potentially serve as a trigger for pathophysiological analysis of HIV.

3.6 Protein interaction network analysis

Information flow can be argued to be the common denominator of the various forms of protein–protein interactions, the flow of information through the cell are allowed by biologically meaningful interfaces that have evolved for the same, and ultimately, they are essential for carrying out a functional system. Hence, all types of protein–protein interactions are desirable to be collected and integrated under a single framework (18). The common up-regulated and down-regulated DEGs were used for string database. The network for the common DEGs (figure 2 and 3) were created using STRING database. For the common up-regulated DEGs, 'Number of nodes were 54', 'Number of edges were 77', 'Average node degree was 2.85', 'Average local clustering coefficient was 0.372', 'Expected number of edges were 66', and 'PPI enrichment p-value was 0.0318'. For the common down-regulated DEGs 'Number of nodes was 100', 'Number of edges were 1105', 'Average node degree was 22.1', 'Average local clustering coefficient was 0.627', 'Expected number of edges were 448', and 'PPI enrichment p-value was <1.0e-16'. Both the network had significantly more interactions then expected which meant that the proteins involved in the interaction are at least partially connected biologically.

The STRING database includes physical interactions which are obtained from experimental data as well as functional associations which are obtained from curated pathways, prediction methods, and automatic text mining. However, inspection of small networks and their underlying evidence is the only intent of the STRING database. On the other hand, the Cytoscape software in terms of its working with a large network is much better suited as well as it offers greater flexibility for the analysis of network, and import as well as visualization of additional data (19).

The STRING app of CYTOSCAPE was used to import the network previously created, this helped us include the resource of CYTOSCASPE as well as STRING in same workflow. The network was analyzed in the MCODE app of CYTOSCAPE which gave us 10 significant genes from the network as shown in figure 4. The initial stage of the MCODE us vertex weighting, the highest *k*-core of the vertex neighbourhood is used to weigh all the vertices based on their local network density. The following stage is molecular complex prediction which takes the vertex weighted graph as an input and then the complex with the highest weighted vertex is seeded and loops

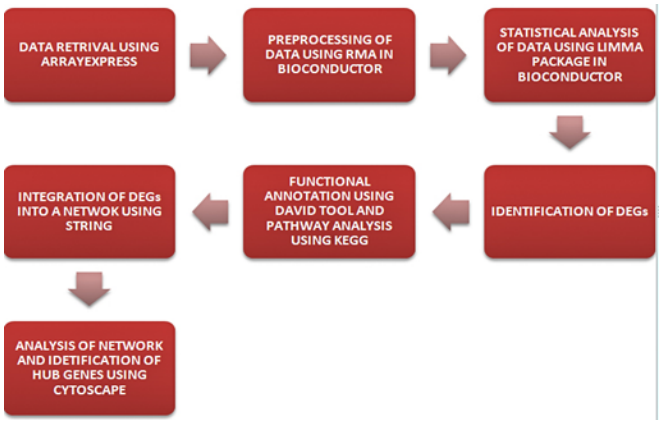

Figure 1: The workflow of the current study

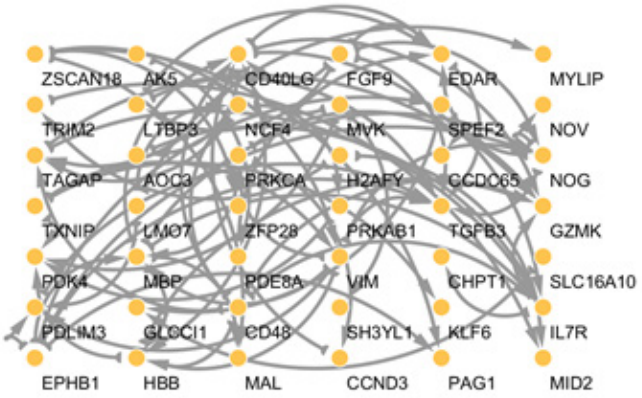

Figure 2: Network of common up-regulated DEGs

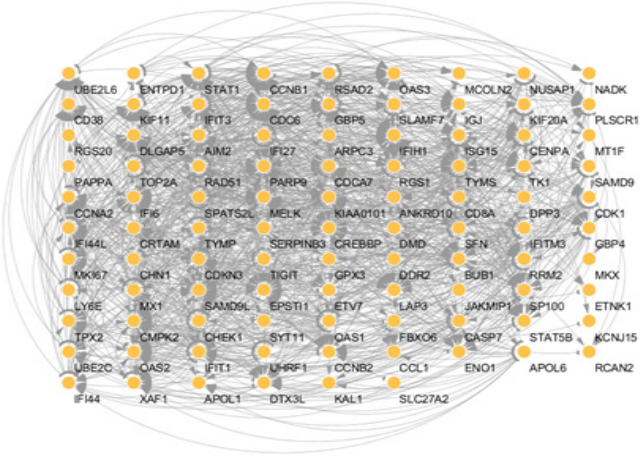
Figure 3: Network of common down-regulated DEGs

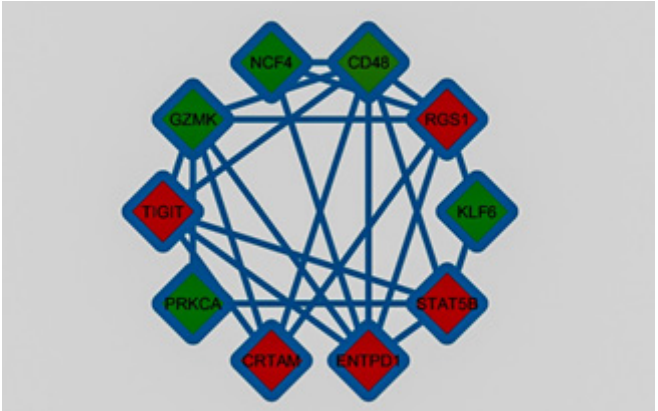

Figure 4: Network of the significant genes

Table 3: Common down-regulated DEGs

Gene symbol	GEO accession	Gene title
CDCA7	E76403, GSE56619, GSE28686, GSE9227	cell division cycle associated 7
FBXO6	GSE76403, GSE44460, GSE28686	F-box protein 6
SLAMF7	GSE76403, GSE28686	SLAM family member 7
CDKN3	GSE76403,GSE56619	cyclin-dependent kinase inhibitor 3
CENPA	GSE76403, GSE9227	centromere protein A
RCAN2	GSE76403, GSE28686	regulator of calcineurin 2
CCNB1	GSE76403, GSE9227	cyclin B1
TIGIT	GSE76403, GSE28686	T-cell immunoreceptor with Ig and ITIM domains
CDC6	GSE76403, GSE56619	cell division cycle 6
MKI67	GSE76403, GSE56619	marker of proliferation Ki-67
MT1F	GSE76403, GSE28686	metallothionein 1F
UBE2C	GSE76403, GSE56619	ubiquitin-conjugating enzyme E2C
MCOLN2	GSE76403, GSE28686	mucolipin 2
CD8A	GSE76403, GSE28686	CD8a molecule
TOP2A	GSE76403, GSE56619, GSE9227	topoisomerase (DNA) II alpha
TYMS	GSE76403, GSE56619, GSE28686, GSE9227	thymidylate synthetase
RRM2	GSE76403, GSE56619	ribonucleotide reductase M2
APOL1	GSE76403, GSE17189	apolipoprotein L1
MELK	GSE76403, GSE9227	maternal embryonic leucine zipper kinase
CASP7	GSE76403, GSE28686	caspase 7
CRTAM	GSE76403, GSE28686	cytotoxic and regulatory T-cell molecule
GBP4	GSE76403, GSE28686	guanylate binding protein 4
STAT1	GSE76403, GSE28686, GSE28160, GSE18468, GSE9227	signal transducer and activator of transcription 1
PTTG1 /// PTTG3P	GSE76403, GS56619	pituitary tumor-transforming 1 /// pituitary tumor-transforming 3, pseudogene
BUB1	GSE76403, GSE9227	BUB1 mitotic checkpoint serine/threonine kinase
CHEK1	GSE76403, GSE56619, GSE9227	checkpoint kinase 1
PARP9	GSE76403, GSE56619, GSE44460, GSE28160, GSE9227	poly(ADP-ribose) polymerase family member 9
IGKV1-12 /// IGKV1D-12	GSE76403, GSE56619	immunoglobulin kappa variable 1-12 /// immunoglobulin kappa variable 1D-12
AIM2	GSE76403, GSE56619	absent in melanoma 2
SLC1A4	GSE76403, GSE56619	solute carrier family 1 (glutamate/neutral amino acid transporter), member 4
CD38	GSE76403, GSE9227	CD38 molecule
DLGAP5	GSE76403, GSE56619	discs, large (Drosophila) homolog-associated protein 5
KIF11	GSE76403, GSE56619, GSE9227	kinesin family member 11
CCNB2	GSE76403, GSE56619, GSE28686, GSE9227	cyclin B2
JAKMIP1	GSE76403, GSE28686, GSE17189	janus kinase and microtubule interacting protein 1

CCNA2	GSE76403, GSE28686, GSE9227	cyclin A2
CDK1	GSE76403, GSE56619	cyclin-dependent kinase 1
UHRF1	GSE76403, GSE56619	ubiquitin-like with PHD and ring finger domains 1
IFI27	GSE76403, GSE56619	interferon, alpha-inducible protein 27
ETV7	GSE76403, GSE56619, GSE28686, GSE14278	ets variant 7
PAPPA	GSE10038, GSE14245	pappalysin 1
LCN15	GSE10038, GSE17372	lipocalin 15
DDR2	GSE10038, GSE28160	discoidin domain receptor tyrosine kinase 2
ENO1	GSE10038, GSE17372	enolase 1
ANKRD10	GSE10038, GSE17372	ankyrin repeat domain 10
JCHAIN	GSE56619, GSE28686	joining chain of multimeric IgA and IgM
IFITM3	GSE56619, GSE28686	interferon induced transmembrane protein 3
TYMP	GSE56619, GSE9227	thymidine phosphorylase
ISG15	GSE56619, GSE9227	ISG15 ubiquitin-like modifier
ENTPD1	GSE56619, GSE17189	ectonucleoside triphosphate diphosphohydrolase 1
PLSCR1	GSE56619, GSE28160, GSE28686, GSE9227	phospholipid scramblase 1
KIAA0101	GSE56619, GSE28686, GSE9227	KIAA0101
NUSAP1	GSE56619, GSE9227	nucleolar and spindle associated protein 1
NADK	GSE56619, GSE17189	NAD kinase
RAD51	GSE26619, GSE28686	RAD51 recombinase
LY6E	GSE56619, GSE28686, GSE9227	lymphocyte antigen 6 complex, locus E
CREBBP	GSE56619, GSE17372	CREB binding protein
GBP5	GSE44460, GSE28686	guanylate binding protein 5
CCL1	GSE44460, GSE17189	C-C motif chemokine ligand 1
RGS1	GSE44460, GSE 9227	regulator of G-protein signaling 1
DTX3L	GSE44460, GSE28160	deltex E3 ubiquitin ligase 3L
ARPC3	GSE44460, GSE14245	actin related protein 2/3 complex subunit 3
DMD	GSE17372, GSE14245	dystrophin
MKX	GSE17372, GSE14245	mohawk homeobox
SERPINB3	GSE17372, GSE7224	serpin family B member 3
TSGA13	GSE17372, GSE14245	testis specific 13
KCNJ15	GSE17372, GSE7224	potassium voltage-gated channel subfamily J member 15
ANOS1	GSE17372, GSE28160	anosmin 1
CHN1	GSE30536, GSE28686	chimerin 1
SFN	GSE30536, GSE14278, GSE7224	stratifin
MX1	GSE28160, GSE9227	MX dynamin like GTPase 1
CMPK2	GSE28160, GSE9227	cytidine/uridine monophosphate kinase 2
UBE2L6	GSE28160, GSE28686, GSE9227	ubiquitin conjugating enzyme E2 L6
EPSTI1	GSE28160, GSE28686, GSE9227	epithelial stromal interaction 1 (breast)
SP100	GSE28160, GSE9227	SP100 nuclear antigen
IFIT1	GSE28160, GSE9227	interferon induced protein with tetratricopeptide repeats 1

IFI44L	GSE28160, GSE28686, GSE9227	interferon induced protein 44 like
RSAD2	GSE28160, GSE28686, GSE9227	radical S-adenosyl methionine domain containing 2
APOL6	GSE28160, GSE28686, GSE18468, GSE17189	apolipoprotein L6
IFIT3	GSE28160, GSE28686, GSE9227	interferon induced protein with tetratricopeptide repeats 3
IFI44	GSE28160, GSE28686, GSE9227	interferon induced protein 44
IFIH1	GSE28160, GSE9227	interferon induced with helicase C domain 1
ETNK1	GSE28160, GSE9227	ethanolamine kinase 1
OAS3	GSE28160, GSE28686, GSE9227	2'-5'-oligoadenylate synthetase 3
OAS1	GSE28160, GSE28686, GSE9227	2'-5'-oligoadenylate synthetase 1
SAMD9L	GSE28160, GSE9227	sterile alpha motif domain containing 9 like
LAP3	GSE28160, GSE28686, GSE9227	leucine aminopeptidase 3
SAMD9	GSE28160, GSE14278	sterile alpha motif domain containing 9
SPATS2L	GSE28160, GSE9227	spermatogenesis associated serine rich 2 like
SLC27A2	GSE28686, GSE18468, GSE9227	solute carrier family 27 member 2
CDK1	GSE28686, GSE9227	cyclin dependent kinase 1
DPP3	GSE28686, GSE14245	dipeptidyl peptidase 3
SYT11	GSE28686, GSE9227	synaptotagmin 11
XAF1	GSE28686, GSE9227	XIAP associated factor 1
SLC1A4	GSE28686, GSE17189	solute carrier family 1 member 4
IFI6	GSE28686, GSE9227	interferon alpha inducible protein 6
TPX2	GSE28686, GSE9227	TPX2, microtubule nucleation factor
KIF20A	GSE28686, GSE9227	kinesin family member 20A
OAS2	GSE28686, GSE9227	2'-5'-oligoadenylate synthetase 2
TK1	GSE18468, GSE9227	thymidine kinase 1
STAT5B	GSE18468, GSE9227	signal transducer and activator of transcription 5B
RGS20	GSE14278, GSE7224	regulator of G-protein signaling 20
GPX3	GSE14245, GSE7224	glutathione peroxidase 3
SRRD	GSE14245, GSE7224	SRR1 domain containing

outward from the seed vertex, including vertices whose weight is above a given limit in the complex. The final stage is post-processing in which the complexes that do not contain at least 2-core are filtered (20).The networks are then scored. The analysis criteria was kept default i.e 'Degree cutoff was kept 2', 'Node score cutoff was kept 0.2', 'K-core was kept 2', and 'Max. depth being 100'. As a result, we obtained 3 networks with the score of 28.30, 6.54, and 5.11 respectively.

The network with the score of 5.11 was selected which gave us the 10 Significant genes which are 'Protein Kinase C Alpha (**PRKCA**)', 'Signal Transducer and Activator of Transcription 5B (**STAT5B**)', 'Krupple Like Factor-6 (**KLF6**)', 'Granzyme K (**GZMK**)', 'T-cell Immunoreceptor With Ig and ITIM Domains (**TIGIT**)', 'Ectonucleoside Triphosphate Diphosphohydrolase 1 (**ENTPD1**)', 'Regulator of G protein signaling 1 (**RGS1**)', 'CD48 Molecules (**CD48**)', 'Cytotoxicity and Regulatory T-cell Molecule (**CRTAM**)', and 'Neutrophil Cytosolic factor 4 (**NCF4**)'. The network of 10 significant genes obtained for HIV dataset was then analyzed and the following topological characters

Table 4: Topological analysis of the significant DEGs

Genes	STAT5B	RGS1	KLF6	GZMK	CD48	CRTAM	ENTPD1	NCF4	TIGIT	PRKCA
BetweennessCentrality	0.136111	0.146759	0.009259	0.127778	0.106019	0.006944	0.071296	0	0.027778	0.006944
ClosenessCentrality	0.692308	0.75	0.5625	0.75	0.818182	0.642857	0.75	0.5625	0.692308	0.529412
ClusteringCoefficient	0.3	0.466667	0	0.53	0.571429	0.83	0.6	1	0.7	0
Degree	5	6	2	6	7	4	6	3	5	2
Neighborhood Connectivity	4.4	4.666667	5.5	5	5	6	5.3	6.3	5.6	5.5
Stress	18	18	2	20	22	2	16	0	6	2
TopologicalCoefficient	0.571429	0.5625	0.642857	0.604167	0.555556	0.666667	0.592593	0.791667	0.62	0.75

were found:

1. Degree distributions.

The number of edges linked to a node n defines the node degree of a node n in an undirected network. For the node degree, a self-loop of a node is considered as two edges on the same node. The numbers of nodes with degree h are givenby the node degree distribution forwhich h = 0,1, … In directed networks, number of incoming edges defines the in-degree of a node n and the number of outgoing edges defines the out-degree of a node n.A node with a high degree is considered as a hub.

2. The neighbourhood connectivity.

It is defined by the average connectivity of all neighbours of n to a node n. the average of the neighbourhood connectivity of all nodes n with h neighbours are given by the neighbourhood connectivity distribution for which h = 0,1, …

3. Clustering coefficients.

For undirected networks, the clustering coefficient C_nC_n of a node n can be defined as,

$$C_n = 2e_n/(h_n\,(h_n - 1))$$

where,

h_nh_n is the number of neighbours of n

e_ne_n the number of edges between all neighbours of n. For directed networks, the definition becomes slightly different which is defined as,

$$C_n = e_n/(h_n(h_n - 1))$$

In any case clustering coefficient can be defined as the ration of the number of edges between the neighbours to the maximum number of maximum edges that could possibly exist between neighbours. This can be defined by,

$$n = N/M$$

Where,

N= number of edges between neighbour of n,

M= maximum number of possible edges between neighbours.

4. Topological coefficient.

It can be defined as a relative measure for the extent to which a node shares neighbours with other nodes. For a node n with neighbourk_n k_nIt can be computed as follows,

$$T_n = avg(J(n,m))/k_n$$

Where,

J(n,m) is the number of nodes share by the nodes n and m,

T_nT_n is the topological coefficient.

5. Stress.

It is the number of shortest paths passing through a node n.

6. Closeness centrality.

It is a measure of how promptly information spreads from a given node to other accessible nodes in the network. It is the reciprocal of the average shortest path length.

7. Betweenness centrality.

It can be computed as follows,

$$C_b(n) = \sum_{p\neq n\neq q} (\sigma_{pq}(n)/\sigma_{pq})$$

Where,

$C_b(n)C_b(n)$ is the betweenness centrality for a node n,

p and q are nodes from the different network then n,

$\sigma_{pq}\,\sigma_{pq}$ denotes the number of shortest path from p to q,

$\sigma_{pq}(n)\sigma_{pq}(n)$ denotes the number of shortest path from p to q in which n lies.

In a term the amount of control that a node exerts over the interactions of other nodes in the network is reflected by the betweenness centrality of that node as mentioned in table 4.

4. Discussion

Computational analysis of genes and their phenotypes has developed into a valuable tool for disease research, drug development and biomarker research in recent years. The broad understanding of gene function, regulation and interactions can be obtained by using vast number of data that are guaranteed by these transcriptional profiling techniques. Its most dynamic application involves the study of gene expression and their patterns across many experiments that survey a broad array of cellular responses, phenotypes and conditions.

Microarrays are considered to be an essential breakthrough in experimental molecular biology for two decades now, hence allowing the monitoring of gene expression of numerous genes in parallel. They have been producing huge amounts of valuable data for a long time.

Understanding of analysis and handling of such data has been one of the major leverages in the utilization of the technology (21).

The necessity to coordinate molecular interactions is entailed by the structural and functional relationships underlying the organization of living systems, chiefly the gene involved in expression and protein activity. The dynamic changes in gene expression and therefore the protein content is dependent on the functional state of the cell, despite the fact that the genome in each cell of any given organism is practically the same.Genetic and protein-proteininteraction (PPI) networks are facilitated to be modelled by both, the development of bioinformatic approaches along with Genome-wide expression profiles using DNA arrays, andthereby help in understanding how the biological networks operate (22).

Microarray technology is being used widely in various biomedical research areas, the corresponding microarray data analysis is an essential step toward the best utilizing of array technologies. Gene regulatory networks have become necessary for various biological research areas such as drug discovery and design, that provide clear insights and understanding of the cellular process in living cells as interactions among the genes and their products have known to play a vital role in many molecular processes. The protein-protein interaction network acts as a blueprint which can beobserved in order to derive the relationships among genes. Keeping in sight the importance of microarray data and PPI network, few computational approaches were designed in order to deduce gene regulatory networks from gene expression data. With each passing year and every research carried out, the biological data are erupting, both in size and complexity. The high-throughput techniques are now regularly used in several laboratories and institutes around the world in basic science applications and majorly in efforts to get a better understanding of human disease and a find a possible cure for some major diseases like HIV, Ebola etc. (23). The pathogenesis of disease like arthritis, which do not have transparent causes or etiological agents, may be elucidated by studying genetic expression patterns and may be elucidated by studying genetic expression patterns and whole-genome association studies in patients and controls, revealing 'genetic signatures' of the disease in question. In such diseases, the dearth of accurate diagnostic technology precludes the need for study of putative biomarkers of disease. Further, the spondyloarthropathies are treated with non-steroidal anti-inflammatory drugs (NSAIDs) which only provide symptomatic relief for pain associated with joint inflammation and do not treat the underlying cause of the disease. Therefore, the identification of potential drug targets by finding genes linked to the disorder holds great value for the development of targeted treatments. Numerous studies have constructed and utilized computational pipelines for the selection of promising genetic candidates in a variety of diseases (24-27).

Most of these operate by integrating publicly available disease datasets to identify differentially expressed genes (DEGs) using different statistical techniques. DEGs are the genes which have are significantly over or under-expressed in patients as compared to healthy people. These genes are found through transcriptomic analysis of whole blood, organ biopsy, or particular cells obtained from both patients and matched controls. These may be key genes to the pathogenesis of disease, (1) providing a starting point for studying the disease mechanisms (2) biomarkers for accurate diagnosis, and (3) targets for novel drugs. The detection of DEGs may be done by employing statistical methods such as the t-test, B-test, SAM (significance analysis of microarrays) and fold-change rule. In general, these models set a threshold based on gene data distribution and genes above or below this threshold are considered to be differentially expressed. A key limitation which limits comparison of different datasets is the non-interoperability of the statistical analyses- studies have found that when tested with the same dataset, different statistical models output different lists of DEG's (28-29).

To overcome this limitation, in this study Linear Models of Microarray Analysis (limma) was implemented through Bioconductor R on the raw unprocessed data from 26 different datasets to cross-compare the differentially expressed genes in these datasets. Limma incorporates log2fold changes, moderated t-statistic and B-statistic. The moderated t-statistic in limma is an improvement on the simple t-test as it moderates the standard error across genes using a Bayesian model, thus increasing the degrees of freedom and increasing reliability (21). Using Bioconductor R to implement limma in the selected datasets, we selected DEGs with log2FC greater than 0.58 or lesser than -0.58, which were also statistically significant with p value less than 0.05. In present study, we have identified a set of putative biomarkers that may play a crucial role in the progression of HIV.

These genes could be utilised as research subjects for exploring their roles in the disease process, and also further extend the knowledge of the molecular mechanism of HIV, and also as putative prognostic biomarkers for clinical validation studies to understand their prognostic effects. However, this primary study with *in-silico* analysis needs to be bolstered by larger experimental and epidemiological studies to produce truly actionable findings.

From the selected database for HIV the significant genes which were obtained are 'Protein Kinase C Alpha (**PRKCA**)', 'Signal Transducer and Activator of Transcription 5B (**STAT5B**)', 'Krupple Like Factor-6 (**KLF6**)', 'Granzyme K (**GZMK**)', 'T-cell Immunoreceptor With Ig and ITIM Domains (**TIGIT**)', 'Ectonucleoside Triphosphate Diphosphohydrolase 1 (**ENTPD1**)', 'Regulator of G protein signaling 1 (**RGS1**)', 'CD48 Molecules (**CD48**)', 'Cytotoxicity and Regulatory T-cell Molecule (**CRTAM**)', and 'Neutrophil Cytosolic factor 4 (**NCF4**)'. The DEGs identified from the current study indicates their potential role in molecular pathogenesis as well as coordinating them as presumptive biomarkers for the diagnosis and prediction of HIV infection.

Transducer and Activator of Transcription 5B (**STAT5B**) were reported to be reduced in HIV infection, the reduction was not observed in first 3 days as the STAT5B modulation are generally occurs late into the viral infection cycle, however after 8 days when the infection cycle was well advanced there was radical decrease in the expression of STAT5B (30). However the reduction in expression of STAT5B is only seen in T-tropic form of HIV as reported by Pericle et al. Regulator of G protein signaling 1 (**RGS1**) is one of the genes which encodes molecules which are responsible for differentiation as well as activation of B-cells, this gene along with the other are targeted and affected by the HIV envelop protein gp120 (31) as reported by Jelicic et al. Ectonucleoside Triphosphate Diphosphohydrolase 1 (ENTPD1) might be important in keeping an sufficient balance between activation and regulation of effector immune response in the setting of HIV-1 infection (32) as reported by *Lévy, Y*. Cytotoxicity and Regulatory T-cell Molecule (CRTAM) and it's hyterotypic relation with Necl2 plays an important role in interaction, adhesion as well as migration of Natural killer cells and CD8+ cell when they are stimulated (33) as reported by *Arase et al*. T-cell Immunoreceptor With Ig and ITIM Domains (**TIGIT**) was found to a novel marker of dysfunctional HIV-specific T cells, it also suggested that TIGIT along with some other checkpoint receptors might be a novel curative HIV targets to reverse T cell exhaustion (34) as reported by *Chew et al*.

Krupple Like Factor-6 (**KLF6**) which as reported by Mallipatu et al. have a notable reduction in its expression in HIV-infected human podocytes, a decrease in mitochondrial membrane potential was also observed as a result of KLF6 loss (35). CD48 Molecules (**CD48**) along with NTB-A, which are the coreceptor of natural killer cells are downmodulated in relation with induction of NKG2D ligands as well as a decrease in HLA-A and –B aids HIV in tempering the complete activating potential of Natural Killer cells upon target recognition (36) as reported by *Ward et al*.

Protein Kinase C Alpha (**PRKCA**) along with MAPK3 and IFNG as reported by *Ptak et al*. have been described to be interacting with nine different genes of HIV each, hence suggesting that these proteins play an vital role in the pathogenesis as well as replication of the HIV (37). Granzyme K (**GZMK**) has been previously reported as an upregulated gene in HIV infection by *Genin et al* (38). Though there is no evidence of direct involvement of Neutrophil Cytosolic factor 4 (**NCF4**) in HIV infection, there has been numerous publications which suggest NCF4 being a major protein in opportunistic diseases occurring in immunocompromised individuals. The DEGs were further analysed for possible connections to particular biological processes and pathways using GO enrichment analysis. *STAT5B* was found to be more biologically significant as compared to other DEG's according to the network topological properties.

5. Conclusion

The afore mentioned differentially expressed genes and their network can provide a progressive understanding of the disease-associated candidate genes biological processes, as well as the mechanism of HIV infection. These DEGs along with the identified biological processes and pathways can be exploited as putative targest for the diagnosis and treatment of HIV infection. These candidate genes provide a valuable way of finding HIV much before than was originally possible. These novel putative targets can be used to improve the diagnosis and treatment regime in the future for HIV and warrants for further experimental analysis.

Acknowledgment

The authors are grateful to Amity Institute of Biotechnology, Amity University Uttar Pradesh, Noida for providing the facility and technical support during the preparation of the manuscript.

Conflicts of Interest

Authors declare no conflicts of interest.

References

1. Downs, A.M. and De I. Vincenzi (1996). Probability of heterosexual transmission of HIV: relationship to the number of unprotected sexual contacts. Europeon study Group in heterosexual transmission of HIV. J. Acquir Immune Defic Syndr Hum Retroviral, 11(4): 388-395.
2. Kapila, A., Chaudhary, S., Sharma, RB., Vashisht, H., Sisodia, SS. (2016). A review on: HIV AIDS. Indian J. Pharm. Biol. Res, 4(3):69-73
3. Daar ES. Novel approaches to HIV therapy (2017). F1000Res, 6:759.
4. Yang X, Kui L, Tang M, Li D, Wei K, Chen W, Miao J, Dong Y (2020). High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. Front Genet., 11:19.
5. Verma A, Somvanshi P, Haque S, Rathi B, Sharda S (2019). Association of Inflammatory Bowel Disease with Arthritis: Evidence from In Silico Gene Expression Patterns and Network Topological Analysis. Interdiscip Sci, 11(3):387-396.
6. Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A. (2013). ArrayExpress update—trends in database growth and links to data analysis tools. Nucleic Acids Research, (41)D1:D987–D990.
7. Zhang, Y., Szustakowski, J., & Schinke, M. (2009). Bioinformatics Analysis of Microarray Data. Cardiovascular Genomics, 259–284.
8. Reimers, M., & Carey, V. J. (2006). Bioconductor: An Open Source Framework for Bioinformatics and Computational Biology. DNA Microarrays, Part B: Databases and Statistics, 119–134.
9. Qu Y, He F, Chen Y (2010). Different effects of the probe summarization algorithms PLIER and RMA on high-level analysis of Affymetrix exon arrays. BMC Bioinformatics. 11:211.
10. Smyth, G. K. (n.d.). limma: Linear Models for Microarray Data. Statistics for Biology and Health, 397–420.
11. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4(5):P3.
12. Aoki-Kinoshita, K. F., &Kanehisa, M. (2007). Gene Annotation and Pathway Mapping in KEGG. Methods in Molecular Biology,

71–91.

13. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res, 47(D1):D607-D613.

14. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., &Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics, 27(3):431-432.

15. Irizarry, R. A. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2):249–264.

16. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7):e47–e47.

17. Ning LF, Yu YQ, Guoji ET, Kou CG, Wu YH, Shi JP, Ai LZ, Yu Q (2015). Meta-analysis of differentially expressed genes in autism based on gene expression data. Genet Mol Res, 14(1):2146–2155.

18. Huang,J.K., Carlin,D.E., Yu,M.K., Zhang,W., Kreisberg,J.F., Tamayo,P. and Ideker,T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst., 6: 484–495.

19. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ (2019). Cytoscape String App: Network Analysis and Visualization of Proteomics Data. J Proteome Res, 18(2):623-632.

20. Bader GD, Hogue CW (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4:2

21. Quackenbush, J. (2001). Computational analysis of microarray data. Nature Reviews Genetics, 2(6),

22. Goñi J, Esteban FJ, de Mendizábal NV, Sepulcre J, Ardanza-Trevijano S, Agirrezabal I, Villoslada P (2008). A computational analysis of protein-protein interaction networks in neurodegenerative diseases. BMC Syst Biol, 2:52.

23. Berger, B., Peng, J., & Singh, M. (2013). Computational solutions for omics data. Nature Reviews Genetics, 14(5): 333–346.

24. Tuller, T., Atar, S., Ruppin, E., Gurevich, M., and Achiron, A. (2013). Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases. Genes and Immunity, 14(2): 67–82.

25. Xiong, Q., Ancona, N., and Hauser, E. (2012). Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Research, 22(2): 386–397.

26. Hebels, D. G. A. J., Sveje, K. M., de Kok, M. C., van Herwijnen, M. H. M., Kuhnle, G. G. C., Engels, L. G. J. B., et al. (2011). N-nitroso compound exposure-associated transcriptomic profiles are indicative of an increased risk for colorectal cancer. Cancer Letters, 309(1): 1–10.

27. Rybaczyk, L., Rozmiarek, A., Circle, K., Grants, I., Needleman, B., and Wunderlick, J. E. (2009). A New Bioinformatics Approach to Analayze Gene Expressions and Signaling Pathways Reveals Unique Purine Gene Dysregulation Profiles that Distinguish between CD and UC. Inflammatory Bowel Diseases, 15(7): 971–984.

28. Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics (Oxford, England), 18(4): 546–554.

29. Lee, J. W., and Sohn, I. S. (2006). Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Statistical Methods in Medical Research, 15: 3–20.

30. Pericle, F., Pinto, L., Hicks, S., Kirken, R., Sconocchia, G., Rusnak, J., Dolan, M., Shearer, G., Segal, D. (1998). Cutting Edge: HIV-1 Infection Induces a Selective Reduction in STAT5 Protein Expression. J Immunol, 160 (1): 28-31.

31. Jelicic, K., Cimbro, R., Nawaz, F., Huang, D. W., Zheng, X., Yang, J., Lempicki, R. A., Pascuccio, M., Van Ryke, D., Schwing, C., Hiatt, J., Okwara, N., Wei, D., Roby, G., David, A., Hwang, II.Y., Kehrl, J. H., Arthos, J., Cicala, C., Fauci, A. S. (2013). The HIV-1 envelope protein gp120 impairs B cell proliferation by inducing TGF-β1 production and FcRL4 expression. Nature Immunology, 14(12): 1256–1265.

32. Lévy Y. Role of regulatory T cells in the pathogenesis of HIV-1 infection (2009). Retrovirology. 6(Suppl 2): I22-I22.

33. Noriko Arase, Arata Takeuchi, Midori Unno, Satoshi Hirano, Tadashi Yokosuka, Hisashi Arase, Takashi Saito (2005). Heterotypic interaction of CRTAM with Necl2 induces cell adhesion on activated NK cells and $CD8^+$ T cells. International Immunology, 17:9, 1227–1237.

34. Chew, G. M., Fujita, T., Webb, G. M., Burwitz, B. J., Wu, H. L., Reed, J. S., Hammond, K. B., Clayton, K. L., Ishii, N., Abdel-Mohsen, M., Liegler, T., Mitchell, B. I., Hecht, F. M., Ostrowski, M., Shikuma, C. M., Hansen, S. G., Maurer, M., Korman, A. J., Deeks, S. G., Sacha, J. B. Ndhlovu, L. C. (2016). TIGIT Marks Exhausted T Cells, Correlates with Disease Progression, and Serves as a Target for Immune Restoration in HIV and SIV Infection. PLoS pathogens, 12(1):e1005349.

35. Mallipattu SK, Horne SJ, D'Agati V, Narla G, Liu R, Frohman MA, Dickman K, Chen EY, Ma'ayan A, Bialkowska AB, Ghaleb AM, Nandan MO, Jain MK, Daehn I, Chuang PY, Yang VW, He JC (2015). Krüppel-like factor 6 regulates mitochondrial function in the kidney. J Clin Invest, 125(3):1347-61.

36. Ward, J., & Barker, E. (2008). Role of natural killer cells in HIV pathogenesis. Current HIV/AIDS Reports, 5(1):44–50.

37. Ptak, R. G., Fu, W., Sanders-Beer, B. E., Dickerson, J. E., Pinney, J. W., Robertson, D. L., Rozanov, M. N., Katz, K. S., Maglott, D. R., Pruitt, K. D., Dieffenbach, C. W. (2008). Short Communication: Cataloguing the HIV Type 1 Human Protein Interaction Network. AIDS Research and Human Retroviruses, 24(12):1497–1502.

38. P. Genin, Y. Mamane, H. Kwon, C. LePage, M.A. Wainberg, J (1999). Hiscott; Differential regulation of CC chemokine gene expression in human immunodeficiency virus-infected myeloid cells. Virology, 261(2):205-215.