iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36331

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Transcriptomics Analysis using Galaxy and other Online Servers for Rheumatoid Arthritis

Sourabh Parmar[1], Kandimalla Gunasankar[2], Pritish Kr Singh[3], Vinay Dev Manhas[4], Kaushal Kumar[5], Sugunakar Vuree[6]
[1, 2, 3, 4, 5, 6]*Department of Biotechnology, Lovely Professional University, Punjab, India*

Abstract: ***Researchers use transcriptomics analyses for biological data mining, interpretation, and presentation. Galaxy-based tools are utilized to analyze various complex disease transcriptomic data to understand the pathogenesis of the disease, which are user-friendly. This work provides simple methods for differential expression analysis and analysis of these results in gene ontology and pathway enrichment tools like David, WebGestalt. This method is very effective in better analysis and understanding the transcriptomic data. Transcriptomics analysis has been made on rheumatoid arthritis sra data. Rheumatoid arthritis (RA) is a systemic autoimmune disease. T cells and autoantibodies mediate the pathogenesis. This article discusses the genes which are differentially expressed between the healthy (n=50) and diseased (n=51) and the functions of those genes in the pathogenesis of RA.***

Keywords: Galaxy server, Kallisto, DeSEQ2, Differentially expressed genes, David, Kegg, WebGestalt, Rheumatoid arthritis.

I. INTRODUCTION

Rheumatoid arthritis is a systemic autoimmune disease and is most commonly observed in 40-60 age groups. Around the globe, 0.5-1% are diseased with RA. When compared to males, females are more prone to diseased due to the stimulatory effect of estrogen, it is yet controversial. The pathogenesis and etiology of RA are very complex and it involves many types of cells like macrophages, T and B cells, fibroblasts, dendritic cells. During preclinical development of RA, citrullination takes place in MHC II and recognizes by the antigen-presenting cells. This triggers the immune response and activates the CD4+ T cells which further activates B cells and recognizes the self-proteins. In the pathogenesis of rheumatoid arthritis, Due to dysfunction of metabolic activities in naïve T cells, PPP (pentose phosphate pathway) shunting occurs and ATM is altered. This leads to bypass naïve T cells from the G2/M checkpoint, causing hyperproliferation and naïve T cells changes to memory state early [1]. These malfunctioned CD4+ T cells infiltrate into the synovitis and activates the macrophages, B cells, neutrophils and these help in the expression of RANKL and osteoclasts are activated which cause bone destruction. For transcriptomic analysis of RA, RNA-seq from SRA study (**SRP155483)** of white blood cells from RA patients and healthy donors has taken. By using galaxy tools differentially expressed genes (DEGs) have to be found and those results are pasted in further sections.

The emergence of Next Generation Sequencing (NGS) significantly played a significant role in shifting in-vivo research to in-silico research. Transcriptome analysis is one of the technologies used to understand the pathogenesis of various disorders more deeply. Transcriptomic studies are being used to determine a differential expression of genes and offer detection and quantification of the expression of non-coding transcripts. However, RNA Sequencing (RNA-seq) is the most preferred technology of transcriptomic studies. The analysis of RNA-seq datasets can be performed by various tools present in the R package or other software. In this respect, we provide various tools from the Galaxy platform for RNA-seq analysis.

Galaxy is easy to use, web-based data analysis platform for in silico analysis and it is reproducible. It provides a user-friendly platform as there is no complex installation process for using Galaxy as one can do all kinds of analysis with just simple clicks using our mouse and keyboards. Galaxy also provides workflows and histories of all the works that have been performed on the platform. The output of one tool can be used as input of other tools. These tools are present as independent modules but can be interlinked. Galaxy provides storage for data so that one can efficiently perform the same task again. There are three servers named Galaxy Main (https://usegalaxy.org/), Galaxy Europe (https://usegalaxy.eu/), and Galaxy Australia (https://usegalaxy.org.au/) that can be used. The following sections, inputs, outputs, and main features of each galaxy tool, such as FASTQ, Kallisto quant, DESeq2., available for differential gene expression, are detailed. For better understanding, a brief explanation of tools for transcriptomic analyses is given [2]. Further, more interpretation of results is made by gene ontology and pathway enrichment tools like DAVID, PantherDB, WebGestalt, KEGG, Reactome, Enrichr. Finally, in the “Result” section, the outcomes were given.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

II. BACKGROUND

As for the analysis of SRA data and identifying the DEGs, different tools in the galaxy server are used like FastQC, Cutadapt, Kallisto-quant, DESeq2. Even modern technologies are not perfect for doing sequencing. Due to limitations of the instrument, they give different types of errors in the result, such as wrong sequencing of nucleotide. So, in the downstream analysis of these results, we perform quality checks using the FASTQC tool to identify the errors and to exclude them we use the cutadapt tool. KALLISTO-quant is a program for quantifying RNA-seq, which is faster than previous approaches like TOPHAT2, CUFFLINKS, HISAT, BOWTIE, SAILFISH and also with similar accuracy. Kallisto uses T-DBG (Transcriptome de Bruijn graph) constructed from k-mers. Kallisto quant uses the following equation Eq.(1) of the likelihood function for rapidly quantifying the abundances of transcripts from the pseudo alignment of RNA-seq [3].

$$L(\alpha) \propto \prod_{f\in F}\sum_{t\in T} Y_{f,t}\frac{\alpha_t}{l_t} = \prod_{e\in E}\left(\sum_{t\in e}\frac{\alpha_t}{l_t}\right)^{c_e} \quad (1)$$

Where F and T are set of fragments & transcripts. l_t Is the length of transcript t, $Y_{f,t}$ is the compatibility matrix. α_t is the selecting probability of fragments from transcripts. c_e is the number of counts from the equivalence class (e) observed.

DESeq2 package is used to analyze differentially expressed genes by using the negative binomial distribution. It analysis the count matrix of one row of gene and it unambiguously mapped with a column of sample. Normalized gene counts, visualization of genes as graphs (heatmap, variance plot, volcano plot), and differential analysis of count data result. , Estimating the log fold change and dispersion uses the empirical Bayes techniques [4], [5].

For protein-protein interactions, string web server is used. This database gives protein-protein interactions collected and stored from various literature available in PubMed and OMIM. For enrichment analysis, gene ontology (GO) and KEGG classification systems are implemented [6].

For pathways, KEGG has been used in the DAVID web server. KEGG is an encyclopedia of genes and genomes. It is an integrated database resource categorized into systems information (such as pathways, hierarchies, tables), Genomic information (orthologs, genomes, genes, proteins, sequence similarity), chemical information (compound, glycans, reactions and its class, enzyme nomenclature), health information (gene variants, diseases, drugs, and its groups, health-related substances). It gives the pathway maps of the gene list [7].

III. MATERIALS AND METHODS

A. Data Availability

To find the pathways and gene ontologies of differentially expressed genes in rheumatoid arthritis, SRA study **SRP155483** is used. It consists of white blood cells taken from peripheral blood from healthy and patients with arthritis. From this SRA study, 50 healthy as controls and 51 patients with rheumatoid arthritis were taken for analysis.

B. Retrieving data:

In Galaxy, we can retrieve the data using the tool “Faster download and extract reads in fastq format from NCBI SRA” using SRA ids. It creates separate datasets like single-end data and paired-end data. We can use this for further quality control checks to find differentially expressed genes using DESeq2. To find the differentially expressed genes, we also need reference genome and reference transcriptome. We can upload the reference directly or can fetch it from the FTP link.

C. Steps used in GALAXY server to find DEGs:

1) First of all, create a user account in Galaxy.
2) Save all the required SRA IDs in a notepad and upload the data.
3) In tools open ‘Faster Download and Extract Reads in FASTQ format from NCBI SRA’ and select the uploaded data.
4) Quality check for the retrieved SRA data using the tool “FASTQC read quality reports.” It generates two files, of which one is RAW data, and another is WEB page data. In WEB page data, we can visualize the results.
5) If terrible reads are observed, we can use the tool “Cutadapt” to trim and filter the lousy quality reads.
6) Following this, the Kallisto Quant tool is used to quantify abundance (RNA-transcripts). In this, submit SRA files and reference transcriptome. This generates two datasets. One is tabular, and another is HDF5.
7) The tabular results are used in the DESeq2 tool.

8) In the DESeq2 tool, select the Kallisto results for SRA ids separately as per our requirement (like treated vs. untreated) in the factor level. Select input as TPM values (e.g., from Kallisto, sailfish, or salmon).
9) DESeq2 generates different datasets of results as per our selection.
10) It generates a normalized count file of RNA transcripts of each gene with the sample, a graphical summary of the results, a summary of the file in tabular format, which contains Gene id, base means, Log2(FC), stderr, wald-status, P-value, p-adj.
11) Open tool 'Filter data on any column using simple expressions' and select the DESeq2 result file and give condition c7 < 0.05 [8]. It gives the result of differentially expressed genes.
12) Open tool 'Annotate DESeq2/DEXSeq output tables Append annotation from GTF to differential expression tool outputs'. Select the filtered result, input file type DESeq2/edgeR/limma and give the human reference. It appends the DESeq2 results with a gene symbol, chromosome number, strand and feature.
13) Again, open the tool 'Filter data on any column using simple expressions' and select the annotate results as input. Give condition c3 > 0 for upregulated genes and c3 < 0 for downregulated genes, different times.
14) Save results. These results are further used in David and webGestalt.

D. *WebGestalt*: To find Gene Ontology of DEGs.
1) Go to the WebGestalt web server (http://www.webgestalt.org/).
2) Select organism of interest and method of interest (Homo sapiens and Gene set enrichment analysis (GSEA).
3) Select functional database (gene ontology, pathways, interactions and many more….)
4) Upload gene list. (Upload along with rank scores i.e., fold change values for GSEA analysis).
5) Submit the data.
6) Analyze the result.

E. *DAVID*: To find enriched pathways of DEGs.
1) Go to the DAVID web server (https://david.ncifcrf.gov/) [9] and select start analysis.
2) In the upload section, upload all the DEGs of official gene ids.
3) Select identifier and organism.
4) Select list type and submit the list.
5) Go to the List section, select the list to use.
6) It shows the annotation summary results. In those select the required annotation options for the results or use defaults.
7) On the bottom, three types of results have been shown, click on the required result and observe it.

F. *STRING*: To find PPIs of DEGs.
1) Go to String-DB (https://string-db.org/)
2) Select multiple proteins (in case you have a single protein, then select protein by name)
3) In multiple proteins, paste the protein list or upload.
4) Select the organism and click on search.
5) It shows the mapped list of proteins, goes through it, and if there is any change of gene, then change it.
6) Click continue
7) It shows the results of interactions between the proteins.

IV. RESULTS and DISCUSSION

The analysis of differentially expressed genes by utilizing Rheumatoid arthritis (RA) samples gives information about the functional characterization of genes among the Healthy and Patient samples. Without performing adjustments for multiple testing, any microarray data analysis is incomplete. Benjamini and Hochberg method is used and multiple testing adjustment is utilized to avoid the random test scores when a stricter p-value threshold e.g., 0.001 is allowed. To control the False Discovery Rate (FDR), adjusted p-value cut-off of 0.05 is used.

By using Galaxy tools on SRA study **SRP155483**, a total of 1081 genes that are differentially expressed are found (abj-P <0.05), in this total of 335 are upregulated and 746 are downregulated genes (abs (Log2(FC)) > 0), These genes are used to find Gene ontology and the enrichment pathways. Top upregulated and downregulated genes are mentioned in table 1.

Gene name	Base mean	log2(FC)	P-adj
Top Upregulated significant gene IDs			
ZDHHC1	34.54879	0.67019	2.84E-05
SLC4A10	58.39845	0.602315	0.002841
H1-2	716.8165	0.530965	0.002272
B4GALNT4	9.1778	0.509485	0.02074
CDKN1C	39.46934	0.491705	0.027975
AL928654.3	7.967808	0.490894	0.026017
MT1E	14.98762	0.47993	0.018677
PLEKHD1	6.8812	0.475086	0.027106
CUX2	8.858752	0.471036	0.023487
HLA-F	414.9672	0.46837	0.031823
ADORA1	4.131262	0.466941	0.037885
GATA6	2.829328	0.465239	0.038748
SNAI1	16.08507	0.454088	0.033486
ROR2	2.902459	0.441918	0.037031
Top Downregulated significant gene IDs			
ZNF608	58.66518	-1.03151	1.54E-09
OLAH	19.23331	-0.95253	1.71E-06
RIMBP3	113.5881	-0.93642	1.71E-06
EDN1	21.7714	-0.89619	7.04E-06
FKBP5	2324.192	-0.85075	5.54E-06
DAAM2	143.6491	-0.84668	1.89E-05
ADAMTS2	10.53182	-0.83491	9.83E-06
ARG1	225.7799	-0.82264	3.45E-05
IL1R2	1639.384	-0.80191	1.14E-05
INSYN2B	12.64698	-0.79935	3.74E-05
TMIGD3	17.72803	-0.7507	0.000234
BAIAP2L1	22.20195	-0.72553	0.000171
FAM20A	20.35626	-0.71502	0.000155
FLRT2	8.717469	-0.70987	0.000399
AZU1	58.81551	-0.70662	0.00071

Table 1 Top Upregulated and Downregulated differentially expressed genes

A. Gene Ontology (GO)

To find the Gene Ontology of the DEGs, WebGestalt online server is used for gene set enrichment analysis and submitted the DEGs along with scores (Log2(FC)). Out of 1081 DEGs, 1032 gene IDs are unambiguously mapped with the unique Entrez gene IDs and the remaining 49 IDs can not be mapped in the WebGestalt database. The GO slim summary of uploaded genes is represented in figure 1 with a red, green and blue bar.

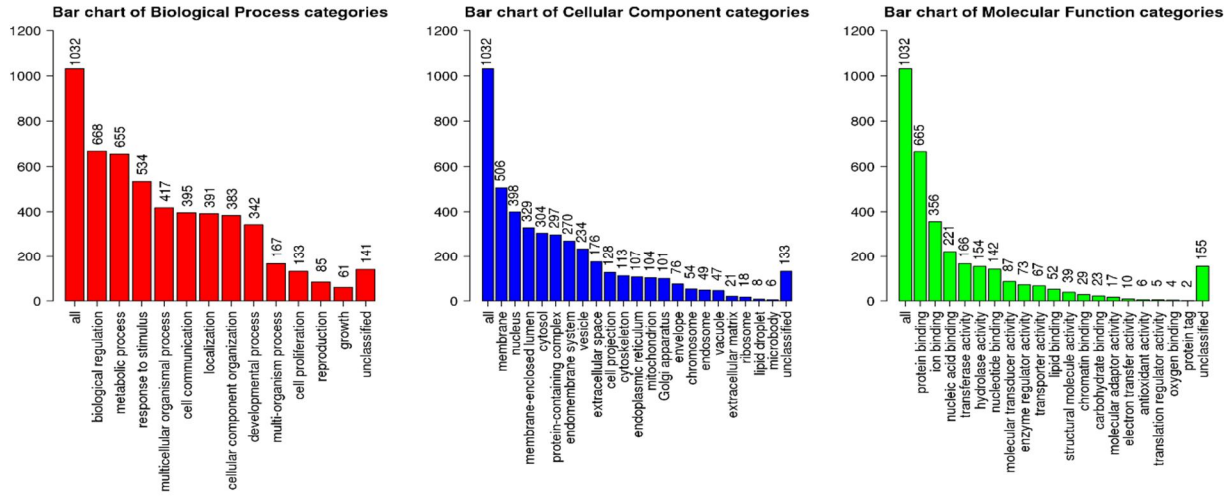


Figure 1 GO slim summary result for uploaded gene IDs: Red bar represents Biological categories, the Blue bar represents cellular component categories and Green bar represents Molecular function categories

GO analysis of DEGs demonstrated that among 1032 unique genes and under threshold FDR 0.05, 837 gene IDs are annotated for the biological process category and 562 gene IDs are annotated for the cellular component category. Ribonucleoprotein complex biogenesis (GO:0022613), mitochondrial gene expression (GO:0140053), ncRNA processing (GO:0034470), protein acylation (GO:0043543), rRNA metabolic process (GO:0016072), defense response to other organisms (GO:0098542), ribonucleoprotein complex subunit organization (GO:0071826), nucleic acid phosphodiester bond hydrolysis (GO:0090305), peptidyl-lysine modification (GO:0018205), covalent chromatin modification (GO:0016569), regulation of peptide secretion (GO:0002791), viral life cycle (GO:0019058), coagulation (GO:0050817)are most enriched Biological processes and in these 9 are positively related (for upregulated genes) and 4 are negatively related (for downregulated genes) categories which are shown in figure 2a. Mitochondrial protein complex (GO:0098798), mitochondrial inner membrane (GO:0005743), mitochondrial matrix (GO:0005759), vesicle lumen (GO:0031983), extracellular matrix (GO:0031012), nuclear chromatin (GO:0000790), receptor complex (GO:0043235) are the most enriched cellular components and 4 positive related and 3 negative related categories which are shown in figure 2b.

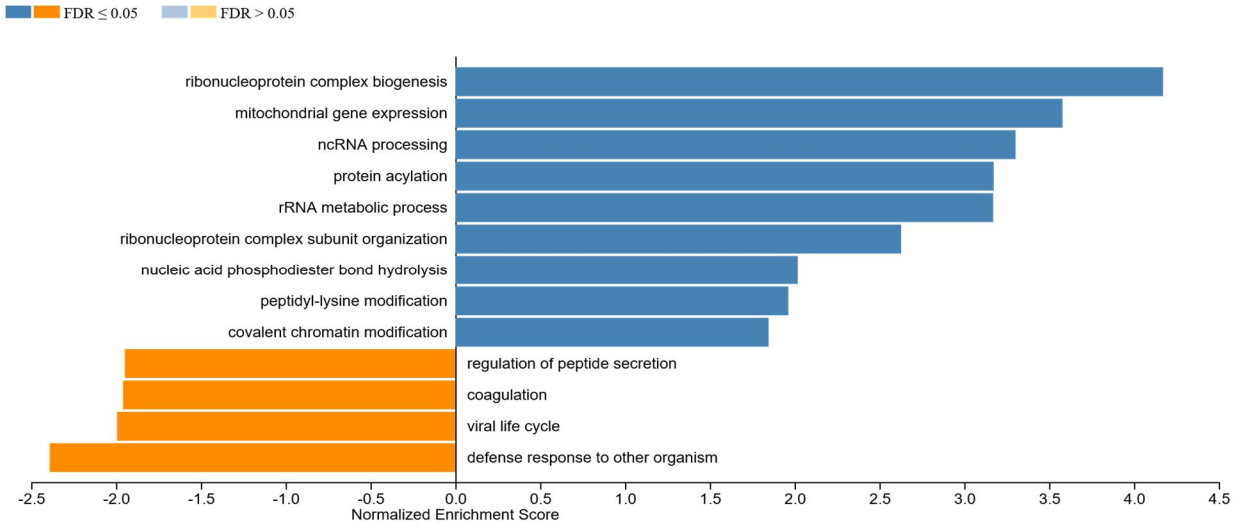


Figure 2a Top enriched Biological process category in Gene Ontology

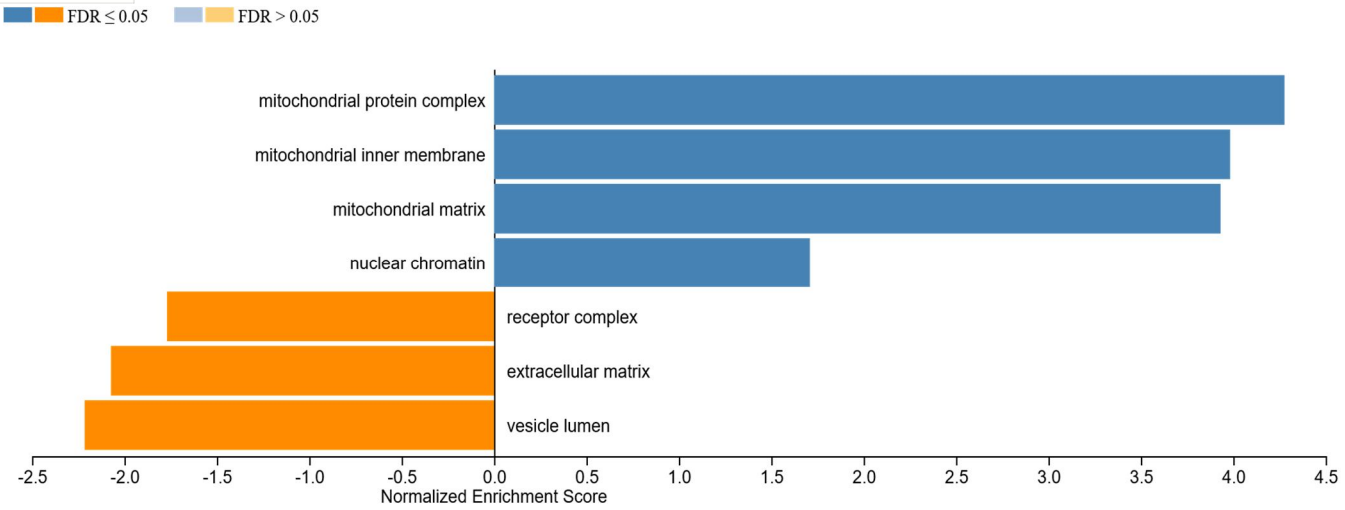


Figure 2b Top enriched Cellular component category in Gene Ontology

B. Kegg Pathways

DAVID tool is used to find the kegg pathways of the genes. Top enriched kegg pathways are mentioned in table 2. In the Hematopoietic cell lineage pathway, many of the CD receptors are found. In RA patients some of these are upregulated and downregulated. GYPA, CR1, ITGAM, IL4R, MME, IL1R1, ITGB3, IL1R2, ITGA1, ITGA2B, GP5, CSF2RA, GP9, CD59, CD24, CD55, CD33 are the differentially expressed genes in this pathway. HIF-1 signaling pathway functioned as cellular responses to low oxygen concentration and the HIF-1 encoded proteins of targeted genes increase the O2 delivery and helps to mediate the adaptive responses to low oxygen concentration. This O2 deprivation may also affect glucose metabolism in T cells, TCA cycle in the production of ROS and angiogenesis. CREBBP, EDN1, CDKN1B, PDHA1, EGLN2, PFKFB3, EGF, HK2, IGF1R, MKNK1, BCL2, PLCG2, EIF4EBP1, EP300, EIF4E2, TLR4, PDK1 are the DEGs in this pathway. Similarly, the remaining top enriched Kegg pathways like Platelet activation, Legionellosis, cGMP-PKG signaling pathway, Regulation of actin cytoskeleton, Ubiquitin mediated proteolysis, Complement and coagulation cascades, Thyroid hormone signaling pathways are may be involved in Rheumatoid arthritis.

Term	Count	%	Fold Enrichment	Benjamini	FDR
hsa04640:Hematopoietic cell lineage	17	1.686508	3.04801	0.028262265	0.028262
hsa04066:HIF-1 signaling pathway	17	1.686508	2.762259	0.046357334	0.046357
hsa04611:Platelet activation	17	1.686508	2.039822	0.664927709	0.664928
hsa05215:Prostate cancer	13	1.289683	2.304344	0.664927709	0.664928
hsa05134:Legionellosis	9	0.892857	2.599773	0.854231268	0.854231
hsa04022:cGMP-PKG signaling pathway	18	1.785714	1.77706	0.854231268	0.854231
hsa04810:Regulation of actin cytoskeleton	22	2.18254	1.634143	0.854231268	0.854231
hsa04120:Ubiquitin mediated proteolysis	16	1.587302	1.821739	0.854231268	0.854231
hsa04610:Complement and coagulation cascades	10	0.992063	2.260672	0.854231268	0.854231
hsa04919:Thyroid hormone signaling pathway	14	1.388889	1.898965	0.854231268	0.854231

Table 2 Top enriched KEGG pathways for differentially expressed genes

C. Protein-Protein interactions on String

Interactions between the proteins of identified DEGs of RA patients have constructed a total of 982 nodes and 1896 edges were screened in figure 3. The top 10 hub nodes with the highest degree of interaction are EGF, PTEN, TLR4, ITGAM, ESR1, EP300, TLR2, CREBBP, MAPK14 and NHP2L1. These top hub nodes may be used as a significant biomarker in RA. The degree of a node, centralities of the top 10 DEGs are mentioned in table 4.

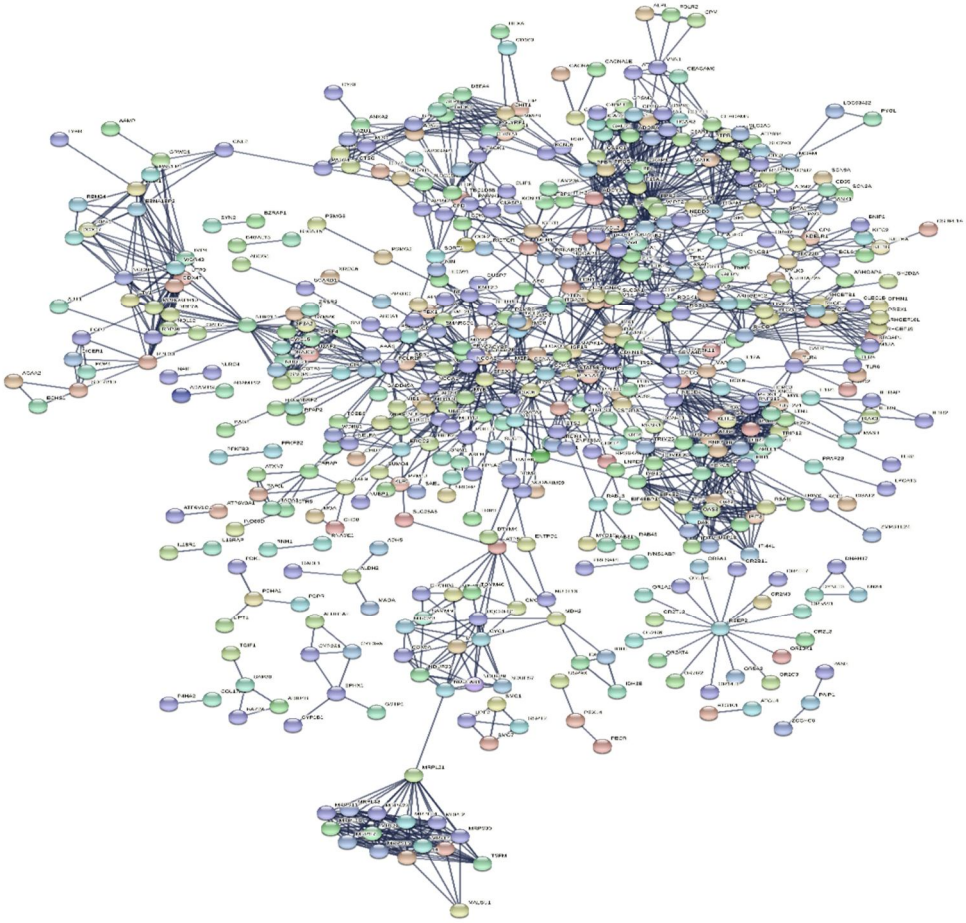


Figure 3 Protein-Protein interactions of DEGs in rheumatoid arthritis

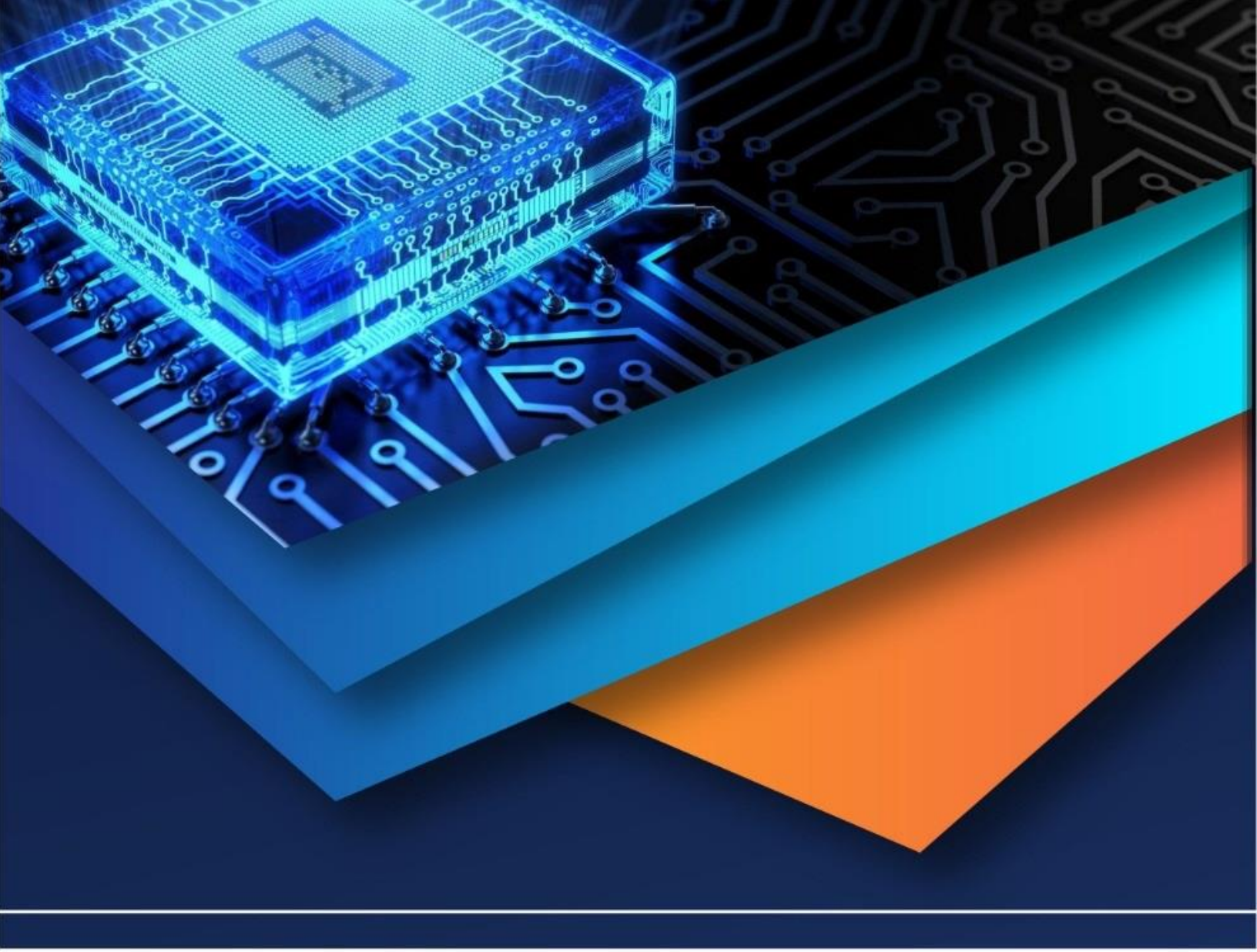
Name	Node Degree	Betweenness Centrality	Closeness Centrality	Stress centrality	Clustering Coefficient
EGF	81	0.055511645	0.397374179	466722	0.122222222
PTEN	77	0.044719691	0.401947764	393886	0.152768284
TLR4	76	0.034794692	0.386547467	393182	0.16245614
ITGAM	75	0.0307008	0.37304848	300576	0.173693694
ESR1	74	0.033011801	0.395126197	315026	0.144020733
EP300	71	0.030752343	0.386382979	281100	0.158551308
TLR2	61	0.017506406	0.37304848	243506	0.204371585
CREBBP	58	0.017620851	0.376607217	176220	0.186327889
MAPK14	55	0.034373545	0.382315789	297274	0.136700337
NHP2L1	53	0.022617343	0.34148176	173382	0.237300435

Table 3 Top 10 hub nodes with the highest degree of interaction in rheumatoid arthritis

In conclusion, the present study provided a comprehensive transcriptomic analysis of rheumatoid arthritis by finding the DEGs, gene ontology, Kegg pathways and interaction between the proteins. By analyzing these altered genes gives information on molecular mechanisms and the significant biomarkers for the diagnosis and treatment of RA. Further molecular biological experiments are needed to confirm the pathways and functions of differentially expressed genes.

REFERENCES

[1] V. Pucino et al., "Metabolic Checkpoints in Rheumatoid Arthritis," Front. Physiol., vol. 11, no. April, pp. 1–12, 2020, doi: 10.3389/fphys.2020.00347.
[2] E. Afgan et al., "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," Nucleic Acids Res., vol. 46, no. W1, pp. W537–W544, Jul. 2018, doi: 10.1093/nar/gky379.
[3] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," Nat. Biotechnol., vol. 34, no. 5, pp. 525–527, 2016, doi: 10.1038/nbt.3519.
[4] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," Genome Biol., vol. 15, no. 12, pp. 1–21, 2014, doi: 10.1186/s13059-014-0550-8.
[5] S. Anders and W. Huber, "Differential expression analysis for sequence count data," Genome Biol., vol. 11, no. 10, p. R106, 2010, doi: 10.1186/gb-2010-11-10-r106.
[6] D. Szklarczyk et al., "STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," Nucleic Acids Res., vol. 47, no. D1, pp. D607–D613, 2019, doi: 10.1093/nar/gky1131.
[7] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," Nucleic Acids Res., vol. 45, no. D1, pp. D353–D361, 2017, doi: 10.1093/nar/gkw1092.
[8] R. Liu et al., "Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses," Nucleic Acids Res., vol. 43, no. 15, Apr. 2015, doi: 10.1093/nar/gkv412.
[9] X. Jiao et al., "DAVID-WS: a stateful web service to facilitate gene/protein list analysis," Bioinformatics, vol. 28, no. 13, pp. 1805–1806, Jul. 2012, doi: 10.1093/bioinformatics/bts251.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY