iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IV **Month of publication:** April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4087

www.ijraset.com

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

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Enrichment Analysis for Gene Dataset

Amitha Supragna Sandur
Bachelor of Engineering, Department of Biotechnology, R. V. College of Engineering, Bangalore, India.

Abstract: Enrichment Analysis is carried out for gene expression analysis. Many genes do not obey the central dogma, there are variations in level of expression at each step of the central dogma. Upregulated and downregulated genes can be found out specific for a particular disease which will further be helpful in therapeutic effects of such regulation of genes. In this paper, a simple procedure for carrying out gene enrichment analysis is showed and how it can be interpreted is discussed. There are various tools developed which work on the same principle. EnrichR tool is used for the work being presented here.

I. INTRODUCTION

Enrichment analysis using EnrichR tool works on the basic principle that there are two gene set list which are compared, one gene set is called the query set and the other is called sample set. Query gene set is the one which has been previously annotated and used as reference for future analysis. It is curated by researchers. Sample/ diseased/ perturbed gene set is the one which has not been annotated and it will be compared against the annotated or control gene set. It is entered by the user. This analysis computes the overlap between the two sets and gives results of the intersection in numeric format by assigning a score.

There are various scores considered which forms the basis of enrichment analysis. Four popular scoring schemes used are P, Z, Q and Combined scores. P value is based on Fisher's exact test or Hypergeometric test. Q value is a slightly modified P value for correcting the multiple hypotheses testing. Z score uses a modification of Fisher's exact test and calculates the deviation from an expected rank value. Combined score uses both P and Z score in a mathematical formula and this is considered as the best scoring scheme among all.

The results showed on EnrichR after enrichment analysis can be viewed in many convenient forms: Bar Graph, Table Grid, Network, Clustergram (new). The results can be downloaded in various formats for further analysis.

II. METHODOLOGY

First step is to obtain the gene list. There are various resources from which gene dataset can be downloaded. Some resources are Huge Navigator, BioMart Portal- HGNC, and so on.

Further, two or more gene set can be downloaded and overlap between the two can be used as input in EnrichR software for the analysis. The overlap can be found out using an online software called Venny which works on the principle of Venn Diagrams to find out the intersection. The two gene list are uploaded on Venny and clicking on intersection will give a list of overlapping genes.

Second step is to feed this gene list in enrichment analysis tool. Filters can be applied to get a refined set. In this work, the later resource as mentioned above (HGNC- HUGO Gene Nomenclature Committe) was used to obtain a gene set consisting of 6759 genes. This geneset is downloaded in the link provided and opened using excel sheet (in .csv format). The gene list is copied onto the box provided in the online EnrichR tool on website (Venny may be used in case of two or more gene dataset). Clicking on submit button and waiting for a few seconds will direct us to the results page. The results can be viewed specific to either of the following: Transcription, Pathways, Ontologies, Disease/Drugs, Cell Types. Pathways is suitable in our study, which provides results in many resources (KEGG, KEA, BioCarta, WikiPathways, Reactome and so on). Results according to KEGG in the form of excel sheet is attached in this paper for reference.

This gives information on pathways that are enriched and other parameters involved include: overlapping of genes, P value, Adjusted P value, Old P value, Adjusted Old P value, Z score, Combined score, and the genes associated to each pathway.

III. DISCUSSION

Interpretation of results is based on one of the scores. Most commonly used is combined score. Negative combined score pathways are removed from the set unless we are missing out on some important pathway, it is always better to check the entire list before proceeding so that nothing is missed out for further analysis. Future work can be based on extracting the upregulated and downregulated genes from the above analysis and used in targeting and finding therapeutic effects of such regulation.

REFERENCES

[1] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan; Enrichr: interactive and

collaborative HTML5 gene list enrichment analysis tool; BMC Bioinformatics. 2013; 14: 128.

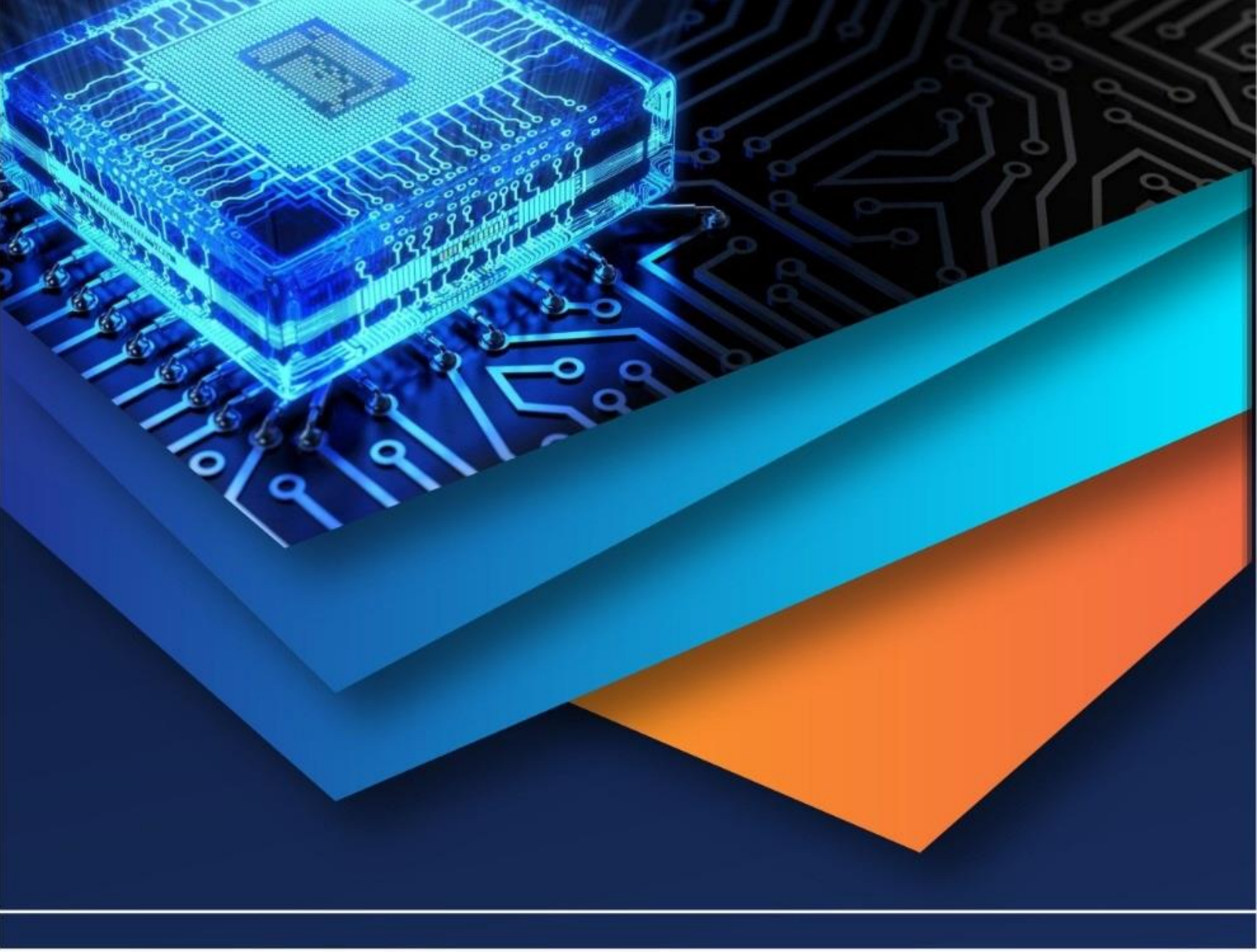
[2] Aravind Subramaniana,b, Pablo Tamayoa,b, Vamsi K. Moothaa,c, Sayan Mukherjeed, Benjamin L. Eberta,e, Michael A. Gillettea,f, Amanda Paulovichg, Scott L. Pomeroyh, Todd R. Goluba,e, Eric S. Landera,c,i,j,k, and Jill P. Mesirova,k; Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles; PNAS; 2005; vol. 102 no. 43.

SUPPLEMENT INFORMATION

A. *..\..\..\Downloads\KEGG_2016_Table (5).Txt*

B. *..\..\..\Downloads\Results (1).Txt*

To be opened with .csv extension to view the file correctly.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY