

RESEARCH

Pathway Enrichment Analysis User Guide

Giuseppe Agapito^{1,3*} and Mario Cannataro^{2,3}

*Correspondence: agapito@unicz.it

¹Department of Legal, Economic and Social Sciences, University "Magna Graecia", Catanzaro, Italy
Full list of author information is available at the end of the article

To perform pathway enrichment analysis, the users need internet connection, a pathway enrichment analysis framework, Java, R, and a web-browser installed on his/her computer.

- To obtain genes differentially expressed from microarray data, users can use GEO (<https://www.ncbi.nlm.nih.gov/geo>), a web-repository that collects several microarray data sets for many different diseases.
- To obtain human cancer samples to identify relevant genomic changes that may play a role in cancer development, users can use TCGA (<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>), a web repository that collects several types of human cancer, including nine rare tumors.
- To perform pathway enrichment analysis, users could use BiP, CePa, pathDIP, or SPIA.

Data Sets Download

GEO data sets download

- G.1 Upon connecting to GEO, the user will input the disease of choice in the Search box, paying attention to select GEODataset from the drop-down menu located at the left of the Search box. As second step, clicking Search the number of founded items will be visualized. Clicking on them will open the search results page.
- G.2 It is then possible to filter the results according to the researcher interest, making it easier to find the data set user is looking for.
- G.3 At the information data set page, clicking "Analyze with GEO2R" will open the page needed to obtain the differential genes.
- G.4 In GEO2R page, the user will need to first set the groups to use in the analysis, clicking on *define groups*. The user will then select the appropriate samples and link them to their group.
- G.5 Click on "Analyze" button located at the bottom to run the differential gene expression analysis.
- G.6 Top results are shown in the table at the bottom of the page. Selecting "Download full table" to obtain the results.

The main steps listed above are shown in Figure 1.

TCGA data sets download

- T.1 Upon connecting to Genomic Data Commons Portal (<https://portal.gdc.cancer.gov>), the user will input the disease he/she is looking for in the Search box, or by clicking on the human vignette situated on the right corner. As a result, the Explorer page will be open.

- T.2 From the Explorer page, it is then possible to download the genes list selecting the Gene tab.
- T.3 At the cBioPortal (<https://www.cbioportal.org>) web page, user can annotate gene lists (if available) by selecting the data set of interest from those available listed in the main page. After selecting the annotation data set, it is then possible to click the "Query By Gene" button located at the bottom of the page. Clicking "Query By Gene" will open the query building page.
- T.4 In the Query building page, user can paste into the text area the previous downloaded gene lists to annotate, then click "Submit Query".
- T.5 The results are shown in the table at the bottom of the page. Select "Download full table" to obtain the results.

The main steps listed above are shown in Figure 2.

Enrichment

The gene lists obtained from the previous steps are going to be used to perform pathway enrichment analysis (PEA) by using an enrichment tool.

BiP

To perform PEA by using BiP, user must launch BiP and then load the genes or proteins list, and selecting the pathway database to compute the enrichment. Gene list can contain Gene Symbols, or UniProt IDs. User can choose if using any downloaded pathway data in BioPAX format for the analysis. Results will be visualized in a tabular format, that will be saved in a Comma Separated Value (CSV) or txt file. A more detailed vignette of the full BiP analysis capabilities is available at <https://gitlab.com/giuseppeagapito/bip>.

CePa

To perform PEA by using CePa it is necessary to write a simple R script. User must run R, load CePa package and then load the gene list, using the "read.csv" command. Gene list can contain Gene Symbols, or UniProt IDs. The pathway enrichment analysis can be performed by using the "cepa.all()" function. CePa will use the embedded KEGG pathway database to compute the enrichment. In the following we show a simple R script to compute pathway enrichment by using CePa. CePa is available at <http://cran.r-project.org/web/packages/CePa/>.

```
library("CePa")
#read the disease-genes input file
genes <- read.csv("/genes.txt", sep = "\t")
colnames(genes) <- "list" #add the name to the column
res = cepa.all(dif = gene.list$dif) #run the PEA analysis
plot(res) #display the PEA results
```

pathDip

To perform PEA by using pathDIP, user must connect to the pathDip web-site and then paste the gene list into the Search box. Gene list can contain Gene Symbols,

Entrez Gene IDs or UniProt IDs. It is important to choose the correct pathway sources to use for the analysis. Before run the analysis, user can choose if download or visualize the results. If the user chooses to download the results, they will be included in a txt file. A more detailed vignette of the full pathDip analysis capabilities is available at <http://ophid.utoronto.ca/pathDIP/>.

SPIA

To perform PEA by using SPIA user must run R, load SPIA package and then write the R scripts. User must load the gene list containing Entrez IDs, using the "read.csv()" command. The pathway enrichment analysis is executed through the "spia()" function. Following we show a simple R script to compute pathway enrichment by using SPIA. SPIA is available at <http://bioconductor.org/packages/SPIA/>.

library(SPIA)

```
#read the disease-genes input
df <- read.csv("/dataset.txt", sep = "\t", header = TRUE)
decolon <- df$log2FC
#add the name to the column
names(decolon) <- as.vector(df$enterez)
allcolen <- df$enterez
#run the PEA analysis
res <- spia(de=decolon, all=allcolen, organism="hsa", nB=2000)
plot(res) #display the PEA results
```

The enriched pathways can be used by researchers to give a biological meaning to huge lists of genes proteins of interest detached from their biological context, making easier to use into clinical and therapeutic scenarios.

Acknowledgements

This work is the extended version of "Using BioPAX-Parser (BiP) to annotate lists of biological entities with pathway data" accepted at the "1st International Workshop on Conceptual Modeling for Life Sciences (CMLS 2020)", in conjunction with the "39th International Conference on Conceptual Modeling (ER 2020)", and published on the Lecture Notes on Computer Science (LNCS 12584) proceedings, Springer Nature.

Funding

Not applicable.

Availability of data and materials

TCGA database link:

<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>

GEO database link: <https://www.ncbi.nlm.nih.gov/geo>

Reactome database link : <https://reactome.org/download-data>

KEGG database link: <https://www.kegg.jp>

BiP software tool link: <https://gitlab.com/giuseppeagapito/bip>

CePa software tool link: <http://cran.r-project.org/web/packages/CePa/>

pathDIP software tool link: <http://ophid.utoronto.ca/pathDIP>

SPIA software tool link: <http://bioconductor.org/packages/SPIA/>

Also, all the links to the datasets and materials have been provided through the manuscript.

Ethics approval and consent to participate

No ethics approval was required for the study.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Legal, Economic and Social Sciences, University "Magna Graecia", Catanzaro, Italy. ²Department of Medical and Surgical Sciences, University "Magna Graecia", Catanzaro, Italy. ³Data Analytics Research Center, University "Magna Graecia", Catanzaro, Italy.

References

NCBI Resources | Documentation | Query & Browse | Email GEO | Sign In to NCBI

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started | **Tools** | **Search for Studies at GEO DataSets**

G.1 colorectal cancer

There are 47844 results for "colorectal cancer" in the GEO DataSets Database. There are 1236068 results for "colorectal cancer" in the GEO Profiles Database.

G.2 Filters: Manage Filters

Top Organisms [Tree]

- Homo sapiens (1107)
- Mus musculus (213)
- Rattus norvegicus (18)
- synthetic construct (5)
- Sus scrofa (1)
- More...

Find related data

Database: [Select]

Search details

("colorectal neoplasms"[MeSH Terms] OR colorectal cancer[All Fields]) AND "Expression profiling by array"[Filter]

G.3 Analyze with GEO2R

Expression data from APC min/+ mice treated with Lon protease

(Submitter supplied) MYC has been named the quintessential oncogene and is deregulated in the majority of human cancers. Still, finding c-MYC inhibitors for therapeutic use has been problematic and MYC itself has long been viewed as "undruggable". Here we present a novel strategy for achieving c-MYC inhibition, involving specific bacterial effector molecules. We made the surprising observation that uropathogenic E. coli activate c-MYC degradation and attenuate MYC expression in host cells and tissues. more...

Organization: Mus musculus
Type: Expression profiling by array
Platform: GPL11180 26 Samples
Download data: CEL

Expression data from mice bladder with MB49-induced cancer treated with Lon protease

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed

GEO accession GSE41011 **Set** Dynamic Transcriptome Analysis Reveal New Prognosis

Define groups **G.4**

Enter a group name: [List]

Group	Accession	Source name
-	GSM1006732	distant normal_stage 1_#1
-	GSM1006733	distant normal_stage 1_#3
-	GSM1006734	distant normal_stage 1_#4
-	GSM1006735	distant normal_stage 2_#1
		stage 2#2
		stage 2#3
		stage 2#4
		stage 3#1
		stage 3#2
		stage 3#3
		stage 3#4
		stage 4#1

Quick start

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare.
- Assign Samples to each group. Highlight Sample rows then click the group name.
- Click 'Analyze' to perform the calculation with default settings.
- You may change settings in the Options tab.

How to use

Analyze **G.5**

Visualization

G.6

Top differentially expressed genes

ID	adj.P.Val	P.Value	t	B	logFC	GB_ACC	SPOT_ID
SPINK5	0.00864	0.000011	-130.36	3.346	-1.508	NM_006846.3	
CRP	0.05008	0.00017	-41.77	2.201	-7.303	NM_000567.2	
HLA-DQB1	0.05008	0.000342	-31.18	1.589	-4.93	NM_002123.3	
MAPK11	0.05008	0.000361	-30.48	1.536	-2.017	NM_002751.5	
ANKA1	0.05008	0.000381	-29.82	1.484	-2.864	NM_000700	NA
MSR1	0.05008	0.000412	28.85	1.404	0.79	NM_002445.3	
AMBIP	0.05008	0.000447	-27.89	1.321	-4.289	NM_001633.3	NA
CCL4	0.0619	0.000632	24.15	0.952	0.757	NM_002984.2	NA
MERTK	0.06953	0.000845	21.38	0.622	0.831	NM_006343.2	
DDX58	0.06953	0.000887	-20.95	0.565	-2.422	NM_014314.3	

Figure 1 GEO data sets download user guide.

The image is a screenshot of the cBioPortal website, which is a platform for cancer genomics data. The interface is divided into several sections:

- Top Navigation:** Includes the NIH logo, "NATIONAL CANCER INSTITUTE CDC Data Portal", and navigation tabs for Home, Projects, Exploration, Analysis, and Repository. There are also search and user account options.
- Harmonized Cancer Data Commons Data Portal:** A central banner with a search bar and a "Data Portal Summary" box. The summary shows 68 projects, 67 primary sites, 615,761 files, 23,535 genes, 84,591 cases, and 3,461,256 mutations.
- Search and Filter Section:** A "Search Cases" box with filters for Primary Site, Program, and other criteria. Below it is a "Cases (2,063)" table with columns for Case ID, Project, Primary Site, Gender, Files, Seq, and Available Data Categories.
- Visualization Tools:** A "Select Studies for Visualization & Analysis" section with a "Quick select" dropdown and a list of studies categorized by type (e.g., Pan-Cancer Studies, Pediatric Cancer Studies, Immunogenomic Studies).
- Query Builder:** A "Submit Query" section with fields for "Selected Studies", "Select Genomic Profiles", "Select Patient/Case Set", and "Enter Genes".
- Results and Analysis:** A "Breast Cancer (MSK, Cancer Cell 2018)" results page showing a "Summary" tab, "Clinical Data", and "CN Segments". It features various charts like "Mutation Counts per Position", "Copy Number Alterations", and "Gene Fusion Profiles".

Figure 2 GEO data sets download user guide.