richPathR

Tutorial

1. Download and install the package

This package can be downloaded from GitHub account: https://github.com/regmibinod30-bio/richPathR and installed locally or it can be installed directly from the GitHub.

>install(devtools)
>Install_github("Binod Regmi/richPathR"

(NOTE: currently the package in GitHub is set private. Install.github functionality does not work for now.)

==========

NOTE: in developing and testing stage, make sure all the packages are available in your machine including devtools and roxygen. Navigate to the directory, click build icon in Rstudio and click load_all. All the functions are loaded. The test data is available in /data dir for testing the package. Tutorial makes navigation easy.

==========

2. Required packages and dependencies

This package was written in MacBookPro 2019 and tested in Windows. Make sure the following packages are up to date in your computing environment. Install the following or later versions of the R packages.

1. dplyr(1.0.8)
2. enrichR(3.0)
3. filestrings(3.2.2)
4. ggplot2(3.3.5)
5. pheatmap(1.0.12)
6. plotly(4.10.0)
7. purrr(0.3.4)
5. readxl(1.4.0)
6. tidyr(1.2.0)
7. VennDiagram(1.7.3)
8. xlsx(0.6.5)
9. R(4.1.3)

The following versions of R development tools were used to write this package

10. devtools(2.4.3)
11. roxygen2(7.2.0)

The easiest way of calling the required package is using the following function. Run the following command:

>call_required_packages()

3. Obtaining the most recent database from *enrichr*

It is highly recommended to download the current database from *enrichr* web application. Use the following command to obtain the library database. It implements listEnrichrDbs functionality hosted in *enrichr* package.

>db_EnrichR_lib()

This command will yield a table something like this:

geneCoverage	genesPerTerm	libraryName	link	numTerms	appyter	categoryId
13362	275	Genome_Browser_PWMs	http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/	615	ea115789fcbf12797fd692cec6df0ab4dbc79c6a	1
27884	1284	TRANSFAC_and_JASPAR_PWMs	http://jaspar.genereg.net/html/DOWNLOAD/	326	7d42eb43a64a4e3b20d721fc7148f685b53b6b30	1
6002	77	Transcription_Factor_PPIs		290	849f222220618e2599d925b6b51868cf1dab3763	1
47172	1370	ChEA_2013	http://amp.pharm.mssm.edu/lib/cheadownload.jsp	353	7ebe772afb55b63b41b79dd8d06ea0fdd9fa2630	7
47107	509	Drug_Perturbations_from_GEO_2014	http://www.ncbi.nlm.nih.gov/geo/	701	ad270a6876534b7cb063e004289dcd4d3164f342	7
21493	3713	ENCODE_TF_ChIP-seq_2014	http://genome.ucsc.edu/ENCODE/downloads.html	498	497787ebc418d308045efb63b8586f10c526af51	7
1295	18	BioCarta_2013	https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways	249	4a293326037a5229aedb1ad7b2867283573d8bcd	7
3185	73	Reactome_2013	http://www.reactome.org/download/index.html	78	b242984a1b68483b0123b08650201c9b313d5c66	7

4. Implementing *enrichr* and generating two data frames

Put the csv file of gene set to be explored in /data dir. Any number of gene sets and libraries can be used as input for implementing this package and quickly obtaining large data frames and generate tables of the most common and unique *terms*. However, for visualizations purpose, we recommend a maximum of five gene sets and an equal number of libraries. First specify the list of database libraries of interest as follows.

>dbs <- c("Cancer_Cell_Line_Encyclopedia", "NCI-60_Cancer_Cell_Lines", "NCI-Nature_2016", "UK_Biobank_GWAS_v1", "KEGG_2021_Human")

Use the following command to implement *enrichr* plug in to obtain the excel files. The output of this command will generate /results directory in /data directory. For each gene list, the result directory contains a separate excel file. For each library, there is a separate sheet in the file.

>implement_enrichr(dbs=dbs, gene_list = "data/prc_gene_set.csv")

The following two commands will generate the data frames, the second command will generate the larger data frame.

enrichr_df <- enrichr_df()
expanded_enrichr_df <- expanded_enrichr_df()

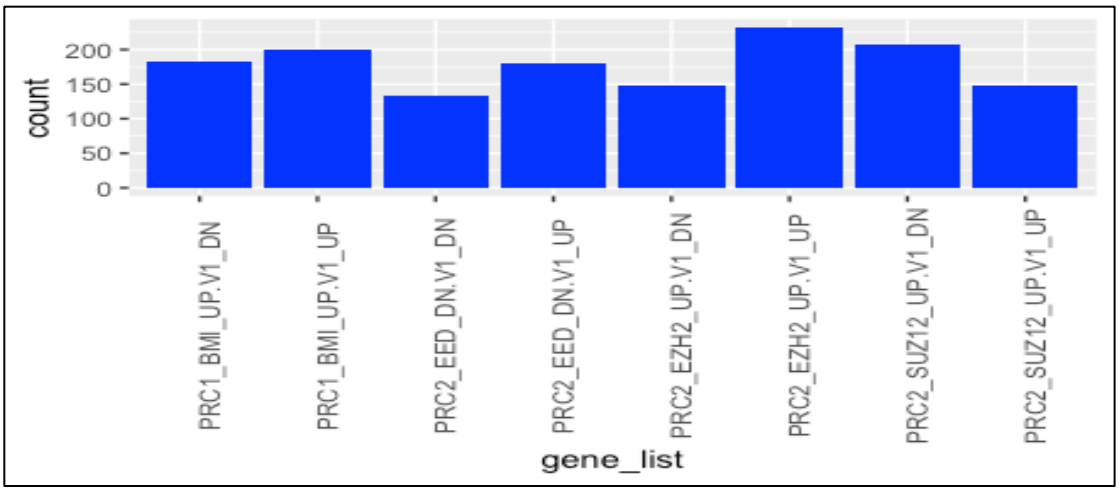
5. Visualization

Obtaining the large data frame and generating the most unique and ubiquitous terms are very useful functions of this package. Once a data frame with multiple gene lists and libraries is obtained, the users can use their script for visualizations. The followings are the exploratory visualization tools provided in the package.

a. Bar plots

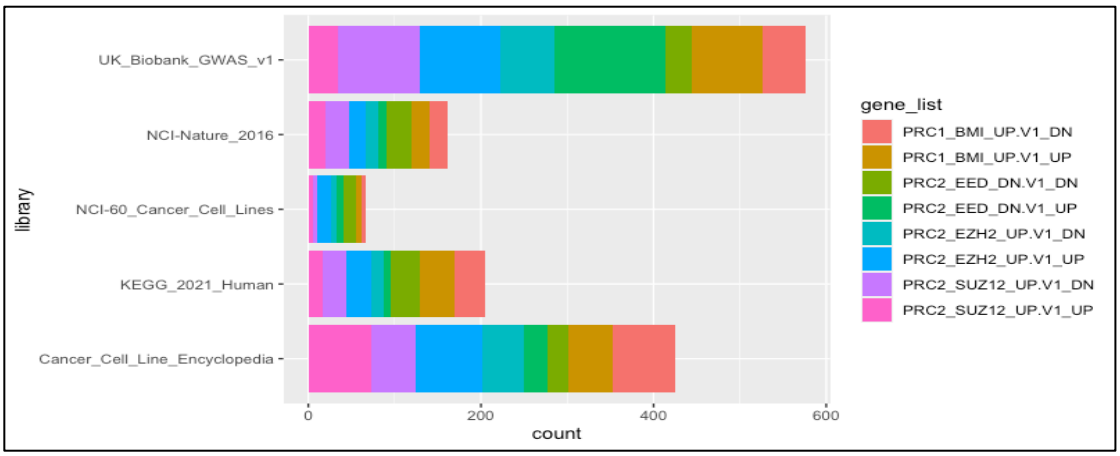
Use the following function to generate a bar plot of term count distributed across the gene list and libraries. The minimum_combined score can be any positive integer, the default is 5. Make sure enrichr_df generated by enrichr_df() is available in the computing environment.

bar_plot_count(enrichr_df = enrichr_df , minimum_combined_score)
Implementing this function generates the plot something like this.



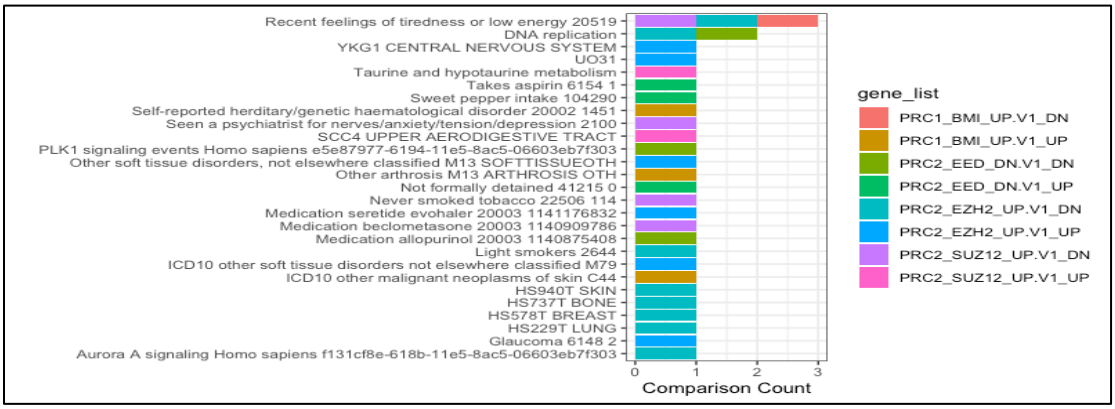
Use the following command to generate a mixed bar plot of the distribution of counts across gene sets and libraries. The minimum_combined score can be any positive integer; the default is 5.

>bar_plot_genelist_library(enrichr_df = enrichr_df, minimum_combined_score)



Use the following command to visualize the top 30 (combined_score sorted) terms distributed across the gene list. This visualization is particularly useful to see the unique and common most significant terms distributed. The minimum_combined score can be any positive integer; the default is 5.

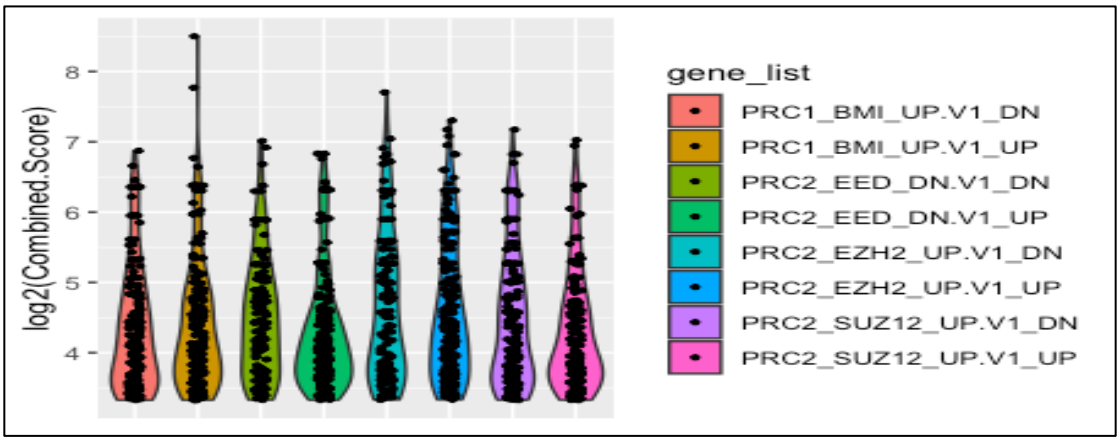
>bar_genelist_terms(enrichr_df = enrichr_df, minimum_combined_score)



b. Violin plots

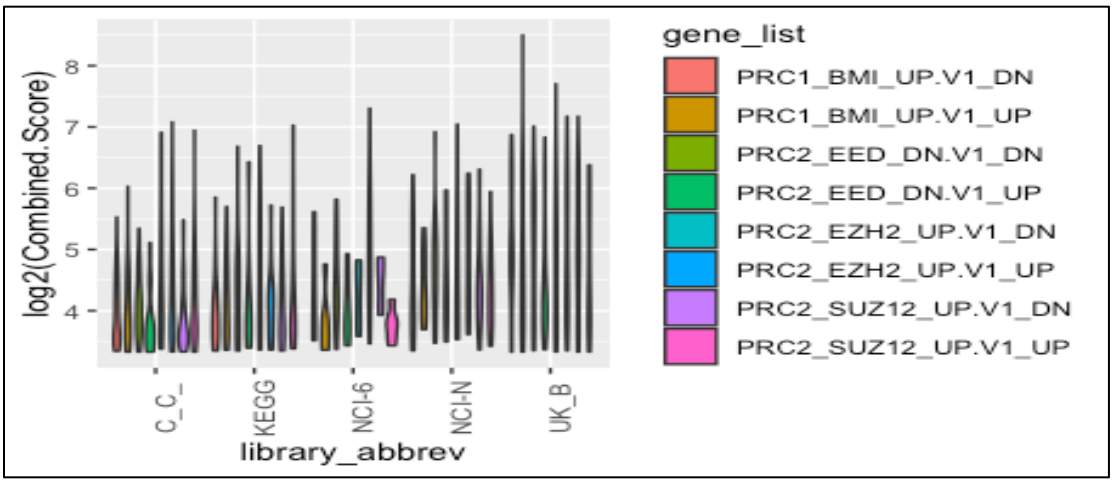
Use the following function to visualize the log_combined score as the violin plots. The combined score can be any positive integer, default is 5. This plot is extremely useful to scan the most significant terms across multiple gene lists and libraries.

>violin_plot_genelist(enrichr_df = enrichr_df, minimum_combined_score)



Use the following command to generate violin plots split into libraries.

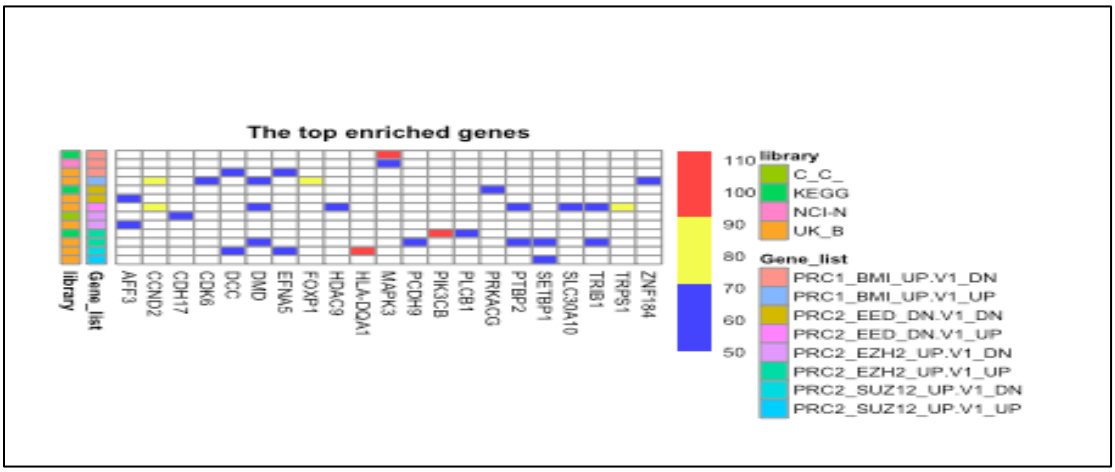
>violin_plot_genelist_library(enrichr_df = enrichr_df,
minimum_combined_score)



c. Heat map

Use the following command to generate the heat map. Before using this functionality, generate expanded_enrichr_df and make it available in the computing environment. The minimum_combined score can be any positive integer; the default is 5. This functionality breaks down genes from the top 30 most significant gene lists and visualizes the top hit genes distributed across gene lists and libraries.

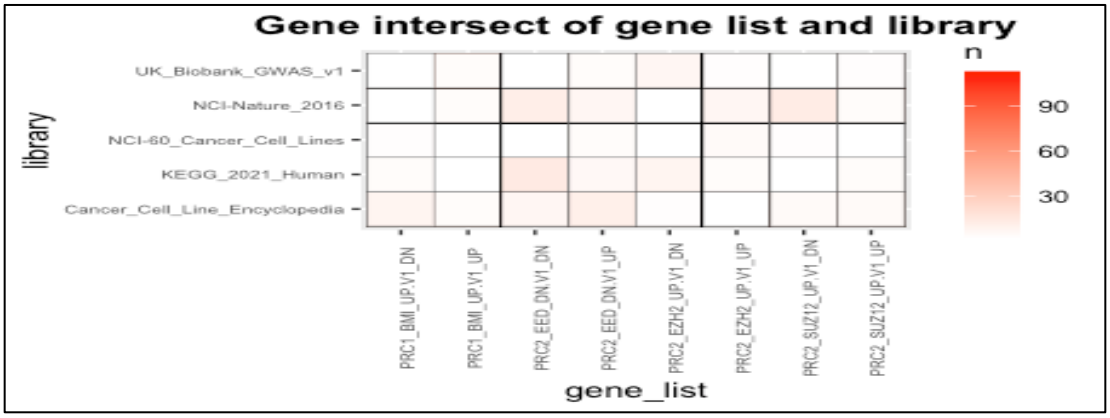
>enrichr_heat_map(expanded_enrichr_df = enrichr_df, minimum_combined_score)



d. Tile plot

To explore which gene lists hit most in which library, use the tile_plot function. Before implementing this function, expanded *enrichr* data frame should be available in computing environment. The minimum_combined score can be any positive integer; the default is 5.

>tille_plot(expanded_enrichr_df = expanded_enrichr_df, minimum_combined_score)



6. Generating tables

This is very useful function to explore the unique terms distributed across multiple gene list and libraries. Before implementing this function, *enrichr* data frame should be available in the computing environment. The combined score can be any positive integer, the default is 5. This table contains more rows than generating ubiquitous terms because all the terms with count one are listed in this table which is usually more than the max count.

>enrichr_unique_terms(enrichr_df = enrichr_df, minimum_combined_score)

	Term	Overlap	P.value	Adjusted.P.value	Odds.Ratio	Combined.Score	Genes	library	gene_list	term_abbrev	library_abbrev
1	C3A LIVER	14/387	1.7603E-05	0.01376583	4.21763858	46.1722649	FN3K;FGA;LIN	Cancer_Cel	PRC1_BMI_	C3ALI	C_C_
2	HEPG2 LIVER	13/407	0.00012444	0.04865708	3.68234728	33.1104351	FGA;APOA1;SE	Cancer_Cel	PRC1_BMI_	HEPGL	C_C_
3	SW403 LARGE	9/262	0.00082978	0.16141404	3.9090465	27.7321566	CYP27A1;LRR(	Cancer_Cel	PRC1_BMI_	SW40LI	C_C_
4	COV318 OVAR	8/230	0.00148378	0.16141404	3.94403342	25.688119	GABARAPL3;F	Cancer_Cel	PRC1_BMI_	COV31O	C_C_
5	NCIH1944 LUI	7/184	0.00175794	0.16141404	4.31424984	27.3679393	FGA;HYAL1;HF	Cancer_Cel	PRC1_BMI_	NCIH194L	C_C_
6	NCIH1395 LUI	9/292	0.00176053	0.16141404	3.48930004	22.1296353	GABARAPL3;F	Cancer_Cel	PRC1_BMI_	NCIH139L	C_C_
7	FU97 STOMAC	10/362	0.00222564	0.16141404	3.12355547	19.0777748	FGA;SOAT2;HY	Cancer_Cel	PRC1_BMI_	FU97S	C_C_
8	VMRCLCD LUN	10/363	0.00227053	0.16141404	3.11454682	18.9605575	GABARAPL3;A	Cancer_Cel	PRC1_BMI_	VMRCLCDL	C_C_
9	NCIH1930 LUI	9/310	0.00263481	0.16900067	3.27761395	19.4655705	ATP8A2;KCNC	Cancer_Cel	PRC1_BMI_	NCIH193L	C_C_
10	SNU668 STOM	7/200	0.00280947	0.16900067	3.95336788	23.2250795	GABARAPL3;S	Cancer_Cel	PRC1_BMI_	SNU66S	C_C_
11	TE14 OESOPH/	5/110	0.00375682	0.1958556	5.15646259	28.7946264	SERPINB3;CAN	Cancer_Cel	PRC1_BMI_	TE14O	C_C_
12	RDES BONE	8/290	0.00607736	0.27955845	3.09536828	15.7962379	SV2B;DCC;DD	Cancer_Cel	PRC1_BMI_	RDESB	C_C_
13	BCP1 HAEMAT	6/181	0.00720075	0.28154922	3.71990529	18.3524149	PAGE4;SOAT2	Cancer_Cel	PRC1_BMI_	BCHALT	C_C_
14	JHH5 LIVER	7/245	0.0083768	0.29775721	3.19852941	15.2962925	FGA;SOAT2;HY	Cancer_Cel	PRC1_BMI_	JHH5L	C_C_
15	PANC0504 PA	4/87	0.00901635	0.29816395	5.19586543	24.4658554	GJC2;HYAL1;P	Cancer_Cel	PRC1_BMI_	PANC05P	C_C_
16	MFE280 ENDC	6/196	0.01042945	0.29816395	3.42361151	15.6223558	FLG;NPFFR1;C	Cancer_Cel	PRC1_BMI_	MFE28E	C_C_
17	MFE296 ENDC	4/91	0.01052008	0.29816395	4.9559701	22.5718128	FN3K;APOBEC	Cancer_Cel	PRC1_BMI_	MFE29E	C_C_
18	TE617T SOFT T	6/197	0.01067595	0.29816395	3.40551329	15.4602199	KCNJ4;CHRNA	Cancer_Cel	PRC1_BMI_	TE6ST	C_C_
19	SKMEL24 SKIN	6/202	0.01197073	0.3158742	3.31779231	14.6821948	CIITA;TBXAS1;	Cancer_Cel	PRC1_BMI_	SKMEL24S	C_C_

This is a very useful function to explore the most ubiquitous terms distributed across multiple gene lists and libraries. Before implementing this function, *enrichr* data frame should be available in the computing environment. The combined score can be any positive integer, the default is 5.

The terms listed in the table are screened based on max(). Therefore, this table yields one or few rows.

>enrichr_ubiquitous_terms(enrichr_df = enrichr_df, minimum_combined_score)

	Term	Overlap	P.value	Adjusted.P.value	Odds.Ratio	Combined.Score	Genes	library	gene_list	term_abbrev	library_abbrev
1	Self-reported	2/14	0.00734926	0.54137121	17.8387387	87.644497	FGA;PACSIN3	UK_Bioban	PRC1_BMI_	Spe(21	UK_B
2	Hypertrophic	4/90	0.01012998	0.3417055	5.01385182	23.0248912	TNNT2;ATP2A	KEGG_202	PRC1_BMI_	Hyprc	KEGG
3	Self-reported	1/14	0.12203167	0.41283056	8.27842809	17.4134636	SLC19A2	UK_Bioban	PRC1_BMI_	Spe(21	UK_B
4	Hypertrophic	5/90	0.00149409	0.09840231	6.44771242	41.9503562	EDN1;ACE;MY	KEGG_202	PRC1_BMI_	Hyprc	KEGG
5	Self-reported	1/14	0.12698281	0.6381722	7.93028846	16.3657645	PACSIN3	UK_Bioban	PRC2_EED_	Spe(21	UK_B
6	Self-reported	1/14	0.12574746	0.24215369	8.0145749	16.6180579	SERINC5	UK_Bioban	PRC2_EED_	Spe(21	UK_B
7	Hypertrophic	3/90	0.05579244	0.64786113	3.59806605	10.3844392	ACTC1;CACNA	KEGG_202	PRC2_EZH2	Hyprc	KEGG
8	Self-reported	1/14	0.12636534	0.28038397	7.97221104	16.4911407	SLC19A2	UK_Bioban	PRC2_EZH2	Spe(21	UK_B
9	Hypertrophic	3/90	0.05579244	0.41531564	3.59806605	10.3844392	EDN1;ITGA2;D	KEGG_202	PRC2_EZH2	Hyprc	KEGG
10	Self-reported	1/14	0.12636534	0.46116835	7.97221104	16.4911407	FGA	UK_Bioban	PRC2_SUZ1	Spe(21	UK_B
11	Hypertrophic	4/90	0.01107873	0.21825094	4.87926769	21.9700173	TNNT2;ATP2A	KEGG_202	PRC2_SUZ1	Hyprc	KEGG
12	Hypertrophic	5/90	0.00149409	0.13222686	6.44771242	41.9503562	EDN1;ACE;ITG	KEGG_202	PRC2_SUZ1	Hyprc	KEGG