Enrichment Analysis for Bladder Cancer: Stage1 vs Precancerous

Tazeen Shaukat

R Markdown - Enrichment anlalysis using EnrichR R package

Summary

- I have already shortlisted the genes in my . . . _Step1.rmd file based on p-value cut off of 0.05
- I also have created a clean list of gene names in the same file above
- However, I will be redoing the filtering and gene name clean up to ensure I get the same results, and also to make sure the gene names are given to the EnrichR function in the correct/expected format
- After discussing with my group, I will also be adding another layer of filtering the results on fdr cut off of 0.01 and limit the result set to a list of no more than a 1000 genes
- Then I will connect to the EnrichR database and run Enrichment

Import T-test results so that we can short list further

```
#folder that contains group comparison results
fileName <- "input/Tazeen_Stage1_Precancerous_Ttest_Shortlisted.csv"

# read in group comparison results
result1 <- read.csv(file = fileName,
                        header = T, stringsAsFactors = F, row.names = 1)
```

Filtering/shortlisting Short list results based in p-value cut off (pvalue <= 0.01) This is already done but redoing just to be sure

```
pValueCutOff <- 0.05
fdrCutOff <- 0.01
which1 <- (as.numeric(result1$Pvalue) <= pValueCutOff) & (as.numeric(result1$FDR) <= fdrCutOff)
table(which1) #11292 genes with p <= 0.05 and fdr <= 0.01
```

```
## which1
## FALSE  TRUE
##  7103 11292
```

```
resultShort <- result1[which1, ] #short listed genes
resultShort <- head(resultShort[order(as.numeric(resultShort$FDR)), ], 1000) #keeping only the top 1000
```

Clean gene names

Note - the gene names are in the form of "reporterid_genename". So need to split this.

```
funcSplit <- function(rep_gene) {
    rep_gene_split <- unlist(strsplit(x = rep_gene,
                                      split = "|",
                                      fixed = TRUE))
    gene <- rep_gene_split[2]
    return(gene)
}
geneListSplit <- apply(X = as.matrix((resultShort$Feature)),
                       MARGIN = 1, FUN = funcSplit )
head(geneListSplit) #cleaned gene names
```

```
## [1] "MFAP4"   "CFD"     "COL16A1" "DCN"     "ACTG2"   "AEBP1"
```

```
#remove duplicates
geneListSplit1 <- unique(geneListSplit)

# remove NA value
geneListFinal <- na.omit(geneListSplit1)

head(geneListFinal)
```

```
## [1] "MFAP4"   "CFD"     "COL16A1" "DCN"     "ACTG2"   "AEBP1"
```

```
#print number of unique genes
# left with 950 after cleanup
length(geneListFinal)
```

```
## [1] 950
```

```
geneList_final_df <- data.frame(Gene = geneListFinal)
head(geneList_final_df)
```

```
##      Gene
## 1   MFAP4
## 2     CFD
## 3 COL16A1
## 4     DCN
## 5   ACTG2
## 6   AEBP1
```

Load Databases for Enrichr R package , and check connection

```
#checking if EnrichR website and packing are working
#dbs <- enrichR::listEnrichrDbs()  #total number of databases available = 200+

#testing if EnrichR package is working
testOutput <- enrichR::enrichr(genes = c("Runx1", "Gfi1", "Gfi1b", "Spi1", "Gata1", "Kdr"), databases =
## Uploading data to Enrichr... Done.
##   Querying KEGG_2021_Human... Done.
## Parsing results... Done.
```

```r
head(testOutput[[1]])
```

```
##                                      Term Overlap      P.value Adjusted.P.value
## 1                  Acute myeloid leukemia    2/67 0.0001643951      0.002794717
## 2 Transcriptional misregulation in cancer   2/192 0.0013407651      0.011396503
## 3                      Pathways in cancer   2/531 0.0098313553      0.055711013
## 4                  VEGF signaling pathway    1/59 0.0175720140      0.074681059
## 5                Chronic myeloid leukemia    1/76 0.0225871297      0.076796241
## 6               Th17 cell differentiation   1/107 0.0316774286      0.079200112
##   Old.P.value Old.Adjusted.P.value Odds.Ratio Combined.Score      Genes
## 1           0                    0  153.30000     1335.73937 SPI1;RUNX1
## 2           0                    0   52.11579      344.72067 SPI1;RUNX1
## 3           0                    0   18.39792       85.03847 SPI1;RUNX1
## 4           0                    0   68.74483      277.82863        KDR
## 5           0                    0   53.11733      201.33461      RUNX1
## 6           0                    0   37.52453      129.54033      RUNX1
```

```r
#List of databases for which enrichment analysis will be run
dblist1 <- read.csv(file = "input/2023-EnrichR-Databases.txt",
                    header = F, stringsAsFactors = F)

head(dblist1)
```

```
##                           V1
## 1            KEGG_2021_Human
## 2     WikiPathway_2021_Human
## 3 GO_Biological_Process_2023
## 4              Reactome_2022
## 5             BioPlanet_2019
## 6               ClinVar_2019
```

Call function to run Enrichment

```r
# set output file name
outputFileName1 <- paste("output/Tazeen-Team4-Step2-Stage1_Precancerous.", "_EnrichR.xlsx", sep="")

#Load R script into the environment
source(file = "functionEnrichment.R")

#call function to run Enrichment
functionEnrichment(dblist1, geneList_final_df, outputFileName1)
```

```
## Uploading data to Enrichr... Done.
##   Querying KEGG_2021_Human... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying WikiPathway_2021_Human... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying GO_Biological_Process_2023... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Reactome_2022... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying BioPlanet_2019... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying ClinVar_2019... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Transcription_Factor_PPIs... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying TRANSFAC_and_JASPAR_PWMs... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying TargetScan_microRNA... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying miRTarBase_2017... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying DisGeNET... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying OMIM_Disease... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Jensen_DISEASES... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Chromosome_Location... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying VirusMINT... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Virus-Host_PPI_P-HIPSTer_2020... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying HMDB_Metabolites... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying dbGap... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying MSigDB_Hallmark_2020... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying ProteomicsDB_2020... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying GWAS_Catalog_2023... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying InterPro_Domains_2019... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying CCLE_Proteomics_2020... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Proteomics_Drug_Atlas_2023... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying PheWeb_2019... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Pfam_Domains_2019... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying ChEA_2022... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying SILAC_Phosphoproteomics... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying Azimuth_2023... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying MAGNET_2023... Done.
## Parsing results... Done.
## Uploading data to Enrichr... Done.
##   Querying GeDiPNet_2023... Done.
## Parsing results... Done.
```

```
#NEED INTERNET CONNECTION
```

Note - you will need internet connection to complete the above step.