Chapter 1: R preliminaries

Hyemin Gu

2020-12-11

Table of Contents

Getting started

유전체 연구에 있어 바이오 인포매틱스는 데이터를 기반으로 하여 생물학적 기능과 더불어 생체 시스템의 조직에 대한 인사이트를 주는 데에 의의가 있다. 방대한 유전 정보를 다루기 위해서는 강력한 통계 계산 툴과 이를 사용하기 위한 환경이 필요하다. 오픈 소스로 다양한 통계적 분석기법을 제공하며 바이오 인포매틱스를 위한 커뮤니티가 활성화 되어있는 R 을 이용하여 유전체 연구, 그 중에서도 differential expression analysis (DEA)를 수행하도록 하자. R 설치 및 기본적인 사용 방법은 1 권에서 충분히 다루었으므로 이 장에서는 R 을 바이오 인포매틱스에 활용하기 위해 필수적인 몇 가지 사항을 리뷰하도록 한다.

working directory 설정

R 세션을 열고 나면 현재 작업이 이루어지는 디렉토리가 어디인지를 확인하는 것이 우선이다. 현재 작업 디렉토리 내에 있는 파일은 파일 이름만으로 열람할 수 있으나, 그 외의 파일은 그 파일이 위치한 디렉토리와 파일 이름을 함께 명시한 파일 경로를 알려 줘야 열람할 수 있다.

보통, 작업 디렉토리와 데이터 저장소, 결과물 저장소, 함수 파일 저장소는 분리되어있는 경우가 흔하다. 한편, 현재 작업 디렉토리를 알고 있으면 각각의 디렉토리를 매번 절대 주소를 통해 접근하지 않고, 현재 작업 디렉토리로부터의 상대적인 위치를 통해 접근할 수 있게 되어 작업하는 컴퓨터가 달라질 때마다 이들 디렉토리 각각의 절대 주소를 바꿔주는 수고를 덜 수 있다.

우선 현재 작업 디렉토리를 확인하고 새로 설정하는 방법을 알아보자. 그리고 현재 디렉토리의 내용물을 확인할 수도 있다.

일반적인 working directory 설정

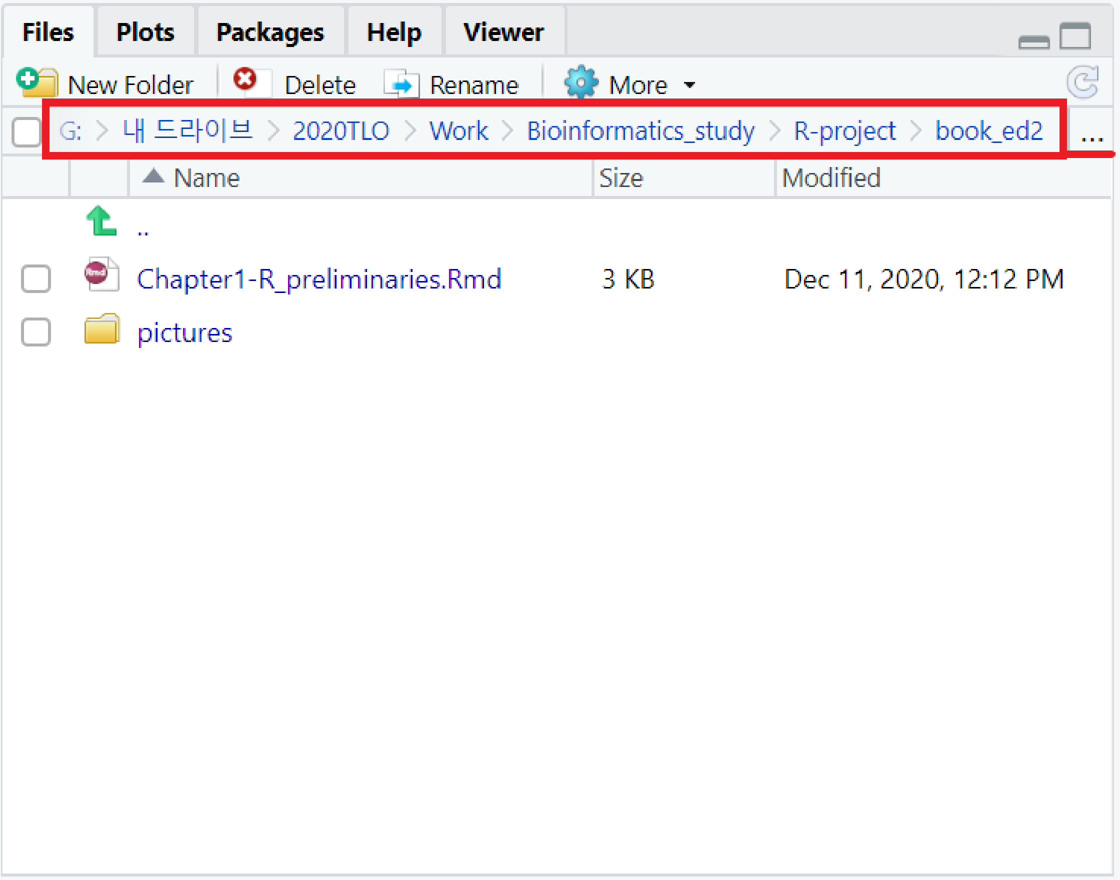
```
getwd()  # 현재 작업 디렉토리 파악

## [1] "G:/내 드라이브/2020TLO/Work/Bioinformatics_study/R-project/book_ed2
"

setwd("G:/내 드라이브/2020TLO/Work/Bioinformatics_study/R-project/book_ed2
")  # 지정된 디렉토리를 작업 디렉토리로 설정
```

```
dir()  # 현재 작업 디렉토리의 내용물

##  [1] "Appendix.Rmd"
##  [2] "Chapter1-R_preliminaries.docx"
##  [3] "Chapter1-R_preliminaries.Rmd"
##  [4] "Chapter2-Getting_started_with_DEA.docx"
##  [5] "Chapter2-Getting_started_with_DEA.Rmd"
##  [6] "Chapter3-DEA_practices.Rmd"
##  [7] "data"
##  [8] "metadata_tcga.Rdata"
##  [9] "pictures"
## [10] "SNCA_survival_ct3VK.pdf"
## [11] "SNCA_survival_tZi8G.pdf"
```

최초 작업 디렉토리는 기본 *내 문서* 또는 연 스크립트 파일의 위치로 세팅되어 있다.

RStudio 에서 working directory 설정

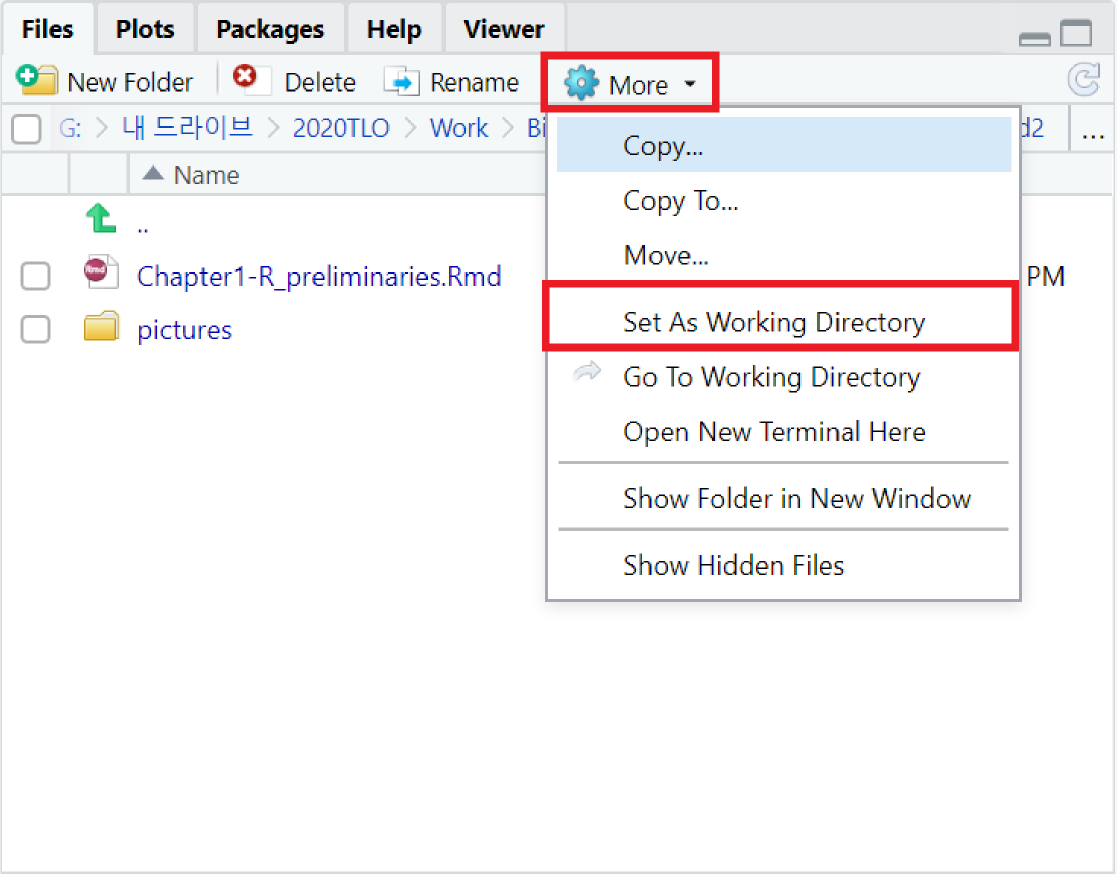
GUI 환경의 RStudio 에서는 우측 하단에 있는 Files 탭에 디렉토리 경로가 표시되며 ...을 클릭하여 다른 디렉토리를 열람할 수 있다. 열람한 디렉토리를 작업 디렉토리로 설정하기 위해서는 More 탭의 Set As Working Directory 를 클릭할 수도 있다.



G: > 내 드라이브 > 2020TLO > Work > Bioinformatics_study > R-project > book_ed2

	Name	Size	Modified
	..		
□	Chapter1-R_preliminaries.Rmd	3 KB	Dec 11, 2020, 12:12 PM
□	pictures		

Files 탭에서 디렉토리 열람



열람 중인 디렉토리를 작업 디렉토리로 설정

작업 디렉토리로부터의 상대 경로를 표기하는 방법은 다음과 같다.

- ./ : 현재 디렉토리
- ../ : 현재 디렉토리로부터 1 레벨 상위 디렉토리
- ~/ : 사용자의 home directory (Windows 의 경우 *내 문서*)
- / : 컴퓨터의 root directory (ex: C drive)

```
dir("./")  # 현재 디렉토리의 내용물 확인

##  [1] "Appendix.Rmd"
##  [2] "Chapter1-R_preliminaries.docx"
##  [3] "Chapter1-R_preliminaries.Rmd"
##  [4] "Chapter2-Getting_started_with_DEA.docx"
##  [5] "Chapter2-Getting_started_with_DEA.Rmd"
##  [6] "Chapter3-DEA_practices.Rmd"
##  [7] "data"
##  [8] "metadata_tcga.Rdata"
##  [9] "pictures"
## [10] "SNCA_survival_ct3VK.pdf"
## [11] "SNCA_survival_tZi8G.pdf"
```

```
dir("../")  # 상위 1 레벨 디렉토리의 내용물 확인

##  [1] "book_ed2"            "CAF_validation"       "functions"

##  [4] "GDCdata"             "geo_data"             "geo_Rdata"

##  [7] "gepiaResults"         "PD-L1_resistance"      "PD-L1_search_from_g
eo"
## [10] "TAZ_Expression"        "TCGA_validation"

dir("../../")  # 상위 2 레벨 디렉토리의 내용물 확인

##  [1] "~$1.Bioconductor 로_TCGA 데이터_접근하기.pptx"

##  [2] "~$2.Colorectal_Cancer_선행논문_workflow 분석_및_연구주제_workflow_수
립.pptx"
##  [3] "~$자료 분석.pptx"

##  [4] "1008-papers-cho"

##  [5] "1104-papers-moon"

##  [6] "1110-papers-cho"

##  [7] "1117-papers-cho"

##  [8] "1120-papers-Jeon"

##  [9] "ppt-김이준교수님"

## [10] "presentations"

## [11] "R-project"

## [12] "references"

## [13] "책작업"

dir("../../functions/")  # 현재 디렉토리와 같은 레벨에 있는 functions 의 내용물
확인

## character(0)
```

즉, 디렉토리 구조가 동일하다는 전제 하에, R-project/PD-L1_resistance_geo/ 안에 있는 gse117358.R 이라는 R script 를 열면 R-project/PD-L1_resistance_geo/

디렉토리가 현재 폴더로 자동 설정되고, 함수 파일이 저장되어 있는 R-project/functions/dea.R 은 ../functions/dea.R 와 같이 상대 경로로 접근할 수 있다. 이 경우, 사용하는 컴퓨터가 바뀜에 따라 절대 경로를 바꿔줄 필요가 없다.

R 코드 실행

R 은 interactive programming language 로, Console 에 명령어를 키보드로 입력하고 Enter 을 눌러 실행시키면 계산을 하고 결과값을 기록하는 등의 과정을 사용자가 직접 확인하면서 수행할 수 있다. Differential expression analysis 를 수행하기 위해서는 데이터를 다운받아 불러오고 그룹을 나눠 그룹간 expression level 차이가 큰 유전자를 뽑아내는 등의 작업을 순차적으로 진행해야 한다. 이 과정은 주로 긴 연산 과정을 수반하고 DEA 를 수행할 때마다 반복된다는 특징이 있다. 따라서, DEA 를 수행하는 데에 필요한 순차적 명령어 집합을

- 스크립트로 기록하여 일괄적으로 실행시키고
- 유사한 기능을 하는 서브루틴들을 함수 파일로 두고 필요할 때 불러오도록 하여

작업의 효율성과 비전공자의 손쉬운 사용성을 높였다.

앞서 언급한 스크립트는 DEA 를 수행할 때마다 R 에서 열고 불러올 데이터의 레이블 또는 전처리 과정 정도만 바꿔 실행시킬 것이다. 그리고 함수 파일은 R-project/functions 디렉토리에 위치시키며 필요한 경우에만 위의 스크립트에서 불러다 쓰게 될 것이며 내용을 확인하거나 수정할 일이 거의 없을 것이다.

스크립트 실행과 함수 파일 로드를 이해하기 위해 아래의 설명을 덧붙인다.

R 파일 확장자

R 에서 작성된 파일의 종류에 따른 확장자 몇 가지를 소개하면 다음과 같다.

- R 스크립트 또는 함수: *.R
- R Markdown: *.Rmd (R 을 통한 문서화)
- R 데이터: *.Rdata (R 작업환경과 그 안의 변수 저장)

- R history: .Rhistory (해당 작업폴더의 내용을 저장하고 종료하면 실행한 커맨드 기록 저장 -> 텍스트 에디터(메모장)으로 열거나 R 에 로드시킬 수 있음)

R 스크립트 실행 또는 함수 파일 로드

R 스크립트와 함수의 차이는 코드의 실행 여부에 있다. 스크립트를 기준으로 설명하면, R Console 에 코드를 키보드로 입력하여 결과를 얻는 것과 마찬가지로, 실행할 코드를 .R 파일에 기록하면 코드를 순차적으로 실행시킬 수 있다.

- 특정 라인에 커서를 놓고 Ctrl+Enter 을 쳐서 해당 라인 실행
- 블록을 잡고 Ctrl+Enter 을 쳐서 해당 블록 실행
- source("*.R")로 전체 스크립트 실행

스크립트와 함수를 비교하기 위해 스크립트의 예제인 install_libraries.R 과 함수 파일의 예제인 DEA.R 을 비교해보자.

script 예제

```
## Install general packages for bioinformatics
## Need to load the packages before using them
# EXAMPLE: library(dplyr)

## cf) you can instead try
## if (!require(package_name))
##   install.packages("package_name")

## R packages
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")  # Bioconductor access
if (length(rownames(installed.packages()) == "dplyr")<1)
  install.packages("dplyr")  # handling data frames in R
```

스크립트 안에는 R Console 에서 바로 실행가능한 코드가 들어있다. 자주 쓰는 코드를 파일에 기록해두고 필요할 때 불러와서 명령을 실행할 수 있다.

함수 예제

```
## DEG analysis

analyze_DEG <- function(grp1, grp2, filtering, download_path=NULL) {
  ## DEG analysis for expression level matrices
  ## EXAMPLES :
  ## res_filt <- analyze_DEG(up, down, "adj.P.Val < 0.05 & logFC>=1")
  ## res_filt <- analyze_DEG(up, down, "adj.P.Val < 0.05 & abs(logFC)>=1")
  ## res_filt <- analyze_DEG(up, down, "adj.P.Val < 0.05 & logFC>=1", "../r
esults/gse11111_CD274_up_down-adjpval0_05-logFC1.csv")
  library(limma)
  grp_names <- c(deparse(substitute(grp1)), deparse(substitute(grp2)))
  grp <- c(rep(grp_names[1], ncol(grp1)), rep(grp_names[2], ncol(grp2)))
  design <- model.matrix(~grp+0)
  colnames(design) <- grp_names

  data <- cbind(grp1, grp2)
  fit <- lmFit(data,design)
  x <- paste(grp_names[2], grp_names[1], sep='-')
  cont <- makeContrasts(contrasts=x,levels=design)

  fit.cont <- contrasts.fit(fit,cont)
  fit.cont2 <- eBayes(fit.cont)
  res <- topTable(fit.cont2,number=Inf)
  res_filt <- topta(res, eval(parse(text=filtering)))

  if (!is.null(download_path))
    write.csv(res_filt, download_path)
  return(res_filt)
}
```

함수는 함수_이름 <- function(함수_파라미터) { expression }과 같이 정의된다. 스크립트 안에 함수를 정의하는 것 만으로는 실행되지 않으며, 정의된 함수가 우측 상단의 Environment 탭에 로드될 뿐이다. 함수를 실행 시키려면 함수 파라미터에 아규먼트를 넣어 함수_이름(아규먼트) 또는 결과변수 <- 함수_이름(아규먼트)와 같이 입력해야 한다.

```
source("../functions/dea.R")
set.seed(1)
grp1 <- matrix(rep(1:10, 5)+0.01*rnorm(50), ncol=5)
grp2 <- matrix(rep(sample(1:10, 10), 7)+0.01*rnorm(70), ncol=7)
res_filt <- analyze_DEG(grp1, grp2, "adj.P.Val < 0.05 & logFC>=1")
res_filt

##      logFC  AveExpr        t      P.Value    adj.P.Val        B
## 5 5.001765 7.917713 998.0877 5.711428e-95 2.855714e-94 207.8655
## 3 4.004519 5.335374 787.0524 1.463631e-90 3.659079e-90 197.7635
## 2 2.999174 3.751517 652.2659 4.482850e-87 7.471416e-87 189.6978
## 6 2.009725 7.168251 422.5479 5.107742e-79 7.296774e-79 170.9041
## 4 2.004217 5.164915 345.7463 2.693408e-75 3.366760e-75 162.1792
```

Installing libraries

CRAN R 은 계산에 필요한 다양한 라이브러리를 제공한다. 특히, 바이오 인포매틱스 관련 라이브러리는 Bioconductor 에서 제공하는 BiocManager 라이브러리를 통해 설치할 수 있는 경우가 많다.

R packages repositories

- CRAN
- Bioconductor
- GitHub, and etc

필수 라이브러리가 아닌 경우, 라이브러리를 직접 설치한 뒤 로드하여 그 안의 함수를 사용할 수 있다. 또는 라이브러리::함수()와 같이 라이브러리의 특정 함수를 일회성으로 호출할 수도 있다.

라이브러리 설치 & 로드 방법

```
install.packages("BiocManager")  # 라이브러리 설치

# way 1
library(BiocManager)  # 라이브러리 로드
install("TCGAbiolinks")  # BiocManager 안의 install 함수 실행

# way 2
BiocManager::install("TCGAbiolinks")  # 패키지 안의 함수 일회성 실행
```

다음은 functions 디렉토리의 install_libraries.R 파일의 내용이다. 스크립트 수행 전에 필요한 전체 패키지 리스트가 설치되어있는지 확인하고, 빠진 것이 있다면 설치하는 코드이다. **새 컴퓨터에서 새 프로젝트로 작업할 때**, 적어도 한번 **source("../functions/install_libraries.R")**을 불러주는 것을 권장한다.

install_libraries.R

```
## Install general packages for bioinformatics
## Need to load the packages before using them
# EXAMPLE: library(dplyr)

## cf) you can instead try
## if (!require(package_name))
##   install.packages("package_name")

## R packages
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")  # Bioconductor access
```

```
if (length(rownames(installed.packages()) == "dplyr")<1)
  install.packages("dplyr")  # handling data frames in R
if (length(rownames(installed.packages()) == "stringr")<1)
  install.packages("stringr")  # handling strings in R
if (length(rownames(installed.packages()) == "survival")<1)
  install.packages("survival")  # survival analysis
if (length(rownames(installed.packages()) == "survminer")<1)
  install.packages("survminer")  # survival plot
if (length(rownames(installed.packages()) == "reticulate")<1)
  install.packages("reticulate")  # running Python script in RStudio
if (length(rownames(installed.packages()) == "png") <1)
  install.packages("png")  # save in png file

## packages in Bioconductor
if (length(rownames(installed.packages()) == "TCGAbiolinks")<1)
  BiocManager::install("TCGAbiolinks")  # TCGA data access
if (length(rownames(installed.packages()) == "affy")<1)
  BiocManager::install("affy")  # handling matrices for TCGAbiolinks
if (length(rownames(installed.packages()) == "SummarizedExperiment")<1)
  BiocManager::install("SummarizedExperiment")  # handling matrices for TCG
Abiolinks
if (length(rownames(installed.packages()) == "EDASeq")<1)
  BiocManager::install("EDASeq")  # matrix normalization for TCGAbiolinks
if (length(rownames(installed.packages()) == "GEOquery")<1)
  BiocManager::install("GEOquery")  # GEO data access
if (length(rownames(installed.packages()) == "Biobase")<1)
  BiocManager::install("Biobase")  # handling matrices for GEO data
if (length(rownames(installed.packages()) == "limma")<1)
  BiocManager::install("limma")  # DEG analysis
if (length(rownames(installed.packages()) == "edgeR")<1)
  BiocManager::install("edgeR")  # DEG analysis of RNA-seq data
if (length(rownames(installed.packages()) == "recount")<1)
  BiocManager::install("recount")  # access all meta data in GDC portal
if (length(rownames(installed.packages()) == "mygene")<1)
  BiocManager::install("mygene")  # gene id conversion
```

install_libraries.R 실행

```
source("../functions/install_libraries.R")
```

유용한 라이브러리

Useful libraries list

Library.name	Application	Repository
affy	handling matrices for TCGAbiolinks	Bioconductor
annotate	support user actions that rely on the different metadata packages	Bioconductor
AnnotationDbi	interface and database connection functions for annotation data packages	Bioconductor

ArrayQualityMetrics	generates a quality report for the microarray data	Bioconductor
Biobase	handling matrices for GEO data	Bioconductor
BiocManager	Bioconductor access	CRAN
biomaRt	converting to gene symbols	Bioconductor
dplyr	handling data frames in R	CRAN
EDASeq	matrix normalization for TCGAbiolinks	Bioconductor
edgeR	DEG analysis of RNA-seq data	Bioconductor
GEOquery	GEO data access	Bioconductor
ggplot2	plotting data	CRAN
GO.db	Annotation maps for Gene Ontology (GO)	Bioconductor
gosim	the computation of functional similarities between GO terms and a gene product	Bioconductor
GOstats	interact with GO and microarray data	Bioconductor
GSEABase	Gene Set Enrichment Analysis (GSEA)	Bioconductor
igraph	simple graphs and networks as well as for graph analysis and plotting	CRAN
KEGG.db	Annotation maps for KEGG	Bioconductor
KEGGgraph	interface between the KEGG pathway and R and the required analysis functions	Bioconductor
limma	DEG analysis	Bioconductor
recount	access all meta data in GDC portal	Bioconductor
stringr	handling strings in R	CRAN
SummarizedExperiment	handling matrices for TCGAbiolinks	Bioconductor
survival	survival analysis	CRAN
survminer	survival plot	CRAN
TCGAbiolinks	TCGA data access	Bioconductor
topGO	test the GO terms	Bioconductor
xlsx	read/write/format Excel file formats	CRAN

Reading and writing data

csv, txt 양식

테이블 읽기

csv 파일은 컴마(,)로 필드를 구분한 테이블 저장 양식이다. 따라서 csv 파일을 읽어오는 `read.csv` 의 기본 설정은 header=TRUE, sep=“,”이다.

한편, 일반적인 파일 확장자로 된 테이블을 읽어오는 `read.table` 의 기본 설정은 header=FALSE, sep=""이다.

[참고] 줄글로 된 일반적인 text 는 `readline` (한 줄씩) 또는 `readLines` (처음부터 끝까지)를 통해 읽어온다.

```
# read csv
ch1_table1_csv <- read.csv("./data/ch1-useful_libraries_csv.csv")
head(ch1_table1_csv, 3)
```

```
##    Library.name
## 1          affy
## 2      annotate
## 3 AnnotationDbi
##                                                            Application
## 1                                    handling matrices for TCGAbiolinks
## 2       support user actions that rely on the different metadata package
s
## 3 interface and database connection functions for annotation data packag
es
##     Repository
## 1 Bioconductor
## 2 Bioconductor
## 3 Bioconductor
```

```
# read general table
ch1_table1 <- read.table("./data/ch1-useful_libraries.txt", header=T, sep=
"\t")
head(ch1_table1)
```

```
##          Library.name
## 1               affy
## 2           annotate
## 3      AnnotationDbi
## 4 ArrayQualityMetrics
## 5            Biobase
## 6        BiocManager
##                                                            Application
## 1                                    handling matrices for TCGAbiolinks
## 2       support user actions that rely on the different metadata package
s
## 3 interface and database connection functions for annotation data packag
es
## 4                     generates a quality report for the microarray data
## 5                                        handling matrices for GEO data
## 6                                                  Bioconductor access
##     Repository
## 1 Bioconductor
## 2 Bioconductor
## 3 Bioconductor
## 4 Bioconductor
## 5 Bioconductor
## 6         CRAN
```

테이블 저장하기

테이블(matrix 또는 data.frame)을 csv 또는 txt 양식으로 저장할 때는 `write.csv` 또는 `write.table` 을 사용한다. 기본 세팅은 append=FALSE (기존 파일의 뒤에 추가 안함), sep=" ", row.names=TRUE (행 이름으로 번호가 붙음)이다. 행 이름으로 번호를 붙이고 싶지 않다면 row.names = F 또는 row.names = FALSE 를 추가해야 한다.

```
ch1_table1 <- read.table("./data/ch1-useful_libraries.txt", header=T, sep="\t")

# write in csv
write.csv(ch1_table1, file="./data/ch1-useful_libraries_csv.csv", row.names = F)

# write in general txt
write.csv(ch1_table1, file="./data/ch1-useful_libraries.txt", row.names = F)
```

excel 양식

엑셀 파일을 읽고 쓸 때는 **openxlsx** 패키지를 이용해야 한다.

테이블 저장하기

`write.xlsx` 의 주요 기본 설정은 overwrite = TRUE (현재 파일에 덮어쓰기), colNames = FALSE, rowNames = FALSE, xy = c(1,1) (쓰기 시작 셀 위치 c(startCol, startRow))이다. 행/열 이름을 같이 저장하고 싶다면 rowNames 또는 colNames 를 TRUE 로 바꾸면 된다.

테이블 읽기

`read.xlsx` 의 주요 기본 설정은 rows = NULL, cols = NULL (A numeric vector specifying which rows in the Excel file to read. If NULL, all rows are read.), sheet = 1, startRow = 1 (읽기 시작 행), colNames = FALSE, rowNames = FALSE, na.strings = "" (NA 로 처리할 셀의 표기 eg:“NA”)이다. 행/열 이름을 지정하고 싶다면 rowNames 또는 colNames 를 TRUE 로 바꾸면 된다.

```
ch1_table1 <- read.csv("./data/ch1-useful_libraries_csv.csv")

if (!require(openxlsx))
  install.packages("openxlsx")

## Loading required package: openxlsx

## Warning: package 'openxlsx' was built under R version 4.0.3

# write in excel
openxlsx::write.xlsx(ch1_table1, "./data/ch1-useful_libraries_excel.xlsx",
                     sheetName="1", colNames = T)
rm(ch1_table1)

# read excel
ch1_table1 <- openxlsx::read.xlsx("./data/ch1-useful_libraries_excel.xlsx
",
                                  sheet = 1, startRow = 1,
                                  colNames = TRUE, rowNames = FALSE,
                                  na.strings = "NA")
```

Rdata 양식

R 전용 데이터 저장양식이고 여러 객체를 한꺼번에 저장, 읽어올 수 있는 .Rdata 로 데이터를 읽고 쓸 수도 있다.

```
ch1_table1_csv <- read.csv("./data/ch1-useful_libraries_csv.csv")
ch1_table1 <- read.table("./data/ch1-useful_libraries.txt", header=T, sep=
"\t")

# write in Rdata
save(list = c("ch1_table1"), "./data/ch1-useful_libraries.Rdata")  # 지정된
 객체 저장
save(list = c("ch1_table1", "ch1_table1_csv"), "./data/ch1-useful_librarie
s.Rdata")  # 지정된 여러 개 객체 저장
save(list = ls(), "./data/ch1-useful_libraries.Rdata")  # 전체 객체 저장

# read Rdata
open("./data/ch1-useful_libraries.Rdata")
```

Subsetting data

데이터 프레임은 2 차원 array 중 각 열마다 서로 다른 타입의 데이터를 포함할 수 있는 자료구조이다. 테이블은 기본적으로 데이터프레임 양식으로 읽어올 수 있다.

data frame 다루기

데이터를 열고 확인하기

- head(데이터프레임) : 데이터 앞부분 출력
- tail(데이터프레임) : 데이터 뒷부분 출력
- View(데이터프레임) : 뷰어 창에서 데이터 확인
- dim(데이터프레임) : 데이터 차원 출력
- str(데이터프레임) : 데이터 속성 출력 (각 열의 변수 타입)
- summary(데이터프레임) : 요약통계량 출력
- names(데이터프레임) : 데이터의 행, 열 이름
- rownames(데이터프레임) : 데이터의 행 이름
- colnames(데이터프레임) : 데이터의 열 이름

```
ch1_table1_csv <- read.csv("./data/ch1-useful_libraries_csv.csv")
class(ch1_table1_csv)

## [1] "data.frame"

head(ch1_table1_csv, 3)  # 또는 head(ch1_table1_csv)

##    Library.name
## 1          affy
## 2      annotate
## 3 AnnotationDbi
##                                                            Application
## 1                                    handling matrices for TCGAbiolinks
## 2       support user actions that rely on the different metadata package
s
## 3 interface and database connection functions for annotation data packag
es
##     Repository
## 1 Bioconductor
## 2 Bioconductor
## 3 Bioconductor

tail(ch1_table1_csv, 3)  # 또는 tail(ch1_table1_csv)

##    Library.name                          Application   Repository
## 26 TCGAbiolinks                     TCGA data access Bioconductor
## 27        topGO                    test the GO terms Bioconductor
## 28         xlsx read/write/format Excel file formats         CRAN
```

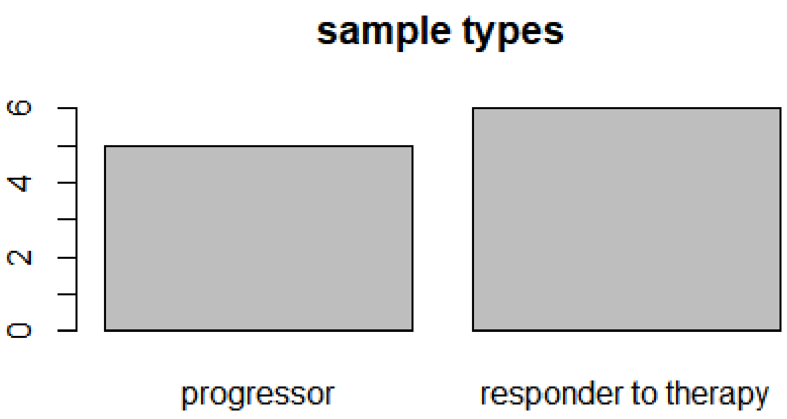
```
View(ch1_table1_csv)

str(ch1_table1_csv)

## 'data.frame':    28 obs. of  3 variables:
##  $ Library.name: chr  "affy" "annotate" "AnnotationDbi" "ArrayQualityMet
rics" ...
##  $ Application : chr  "handling matrices for TCGAbiolinks" "support user
 actions that rely on the different metadata packages" "interface and datab
ase connection functions for annotation data packages" "generates a quality
 report for the microarray data" ...
##  $ Repository  : chr  "Bioconductor" "Bioconductor" "Bioconductor" "Bioc
onductor" ...

summary(ch1_table1_csv)

##  Library.name       Application         Repository
##  Length:28          Length:28          Length:28
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character

names(ch1_table1_csv)

## [1] "Library.name" "Application"  "Repository"

rownames(ch1_table1_csv)

##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14
" "15"
## [16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28"

colnames(ch1_table1_csv)

## [1] "Library.name" "Application"  "Repository"
```

변수 값의 분포 확인

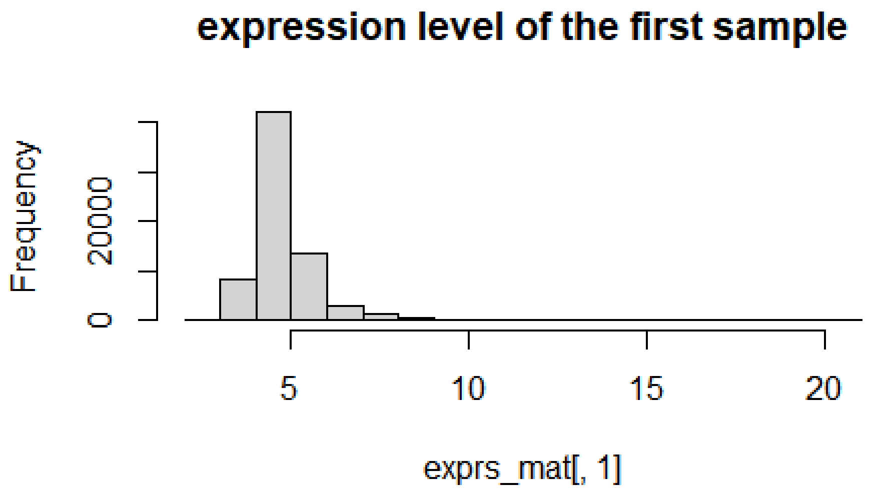
- hist(연속형변수) : 히스토그램
- boxplot(연속형 변수의 벡터 또는 행렬) : 박스플롯 1 개 또는 여러 개
- table(범주형 변수) : 빈도
- barplot(table(범주형변수)) : 막대그래프

```
source("../functions/geo_data.R")
geo_series_idx <- "gse111636"
gse <- download_gse(geo_series_idx) # geo data 로드

## Loading required package: Biobase

## Loading required package: BiocGenerics

## Loading required package: parallel
```

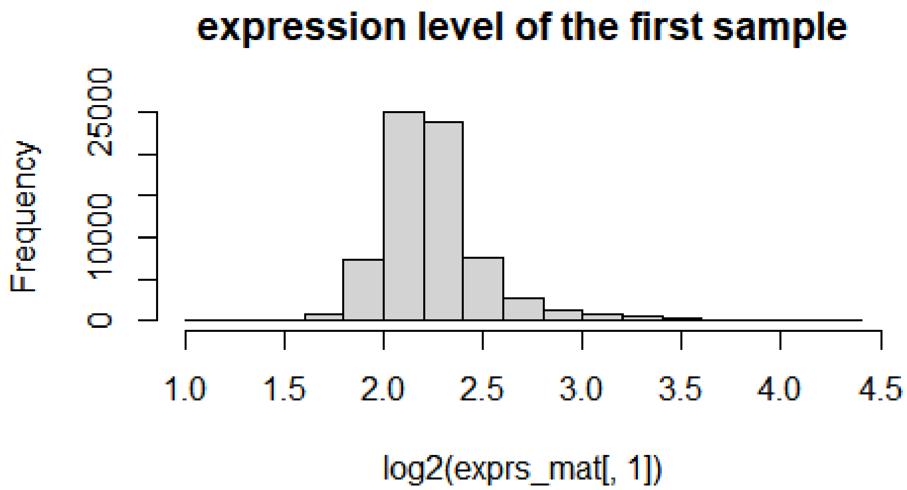
```
##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following object is masked from 'package:limma':
##
##     plotMA

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Setting options('download.file.method.GEOquery'='auto')

## Setting options('GEOquery.inmemory.gpl'=FALSE)

##
## -- Column specification -----------------------------------------------
---------
## cols(
##   ID_REF = col_character(),
##   GSM3036125 = col_double(),
##   GSM3036126 = col_double(),
##   GSM3036127 = col_double(),
##   GSM3036128 = col_double(),
##   GSM3036129 = col_double(),
##   GSM3036130 = col_double(),
##   GSM3036131 = col_double(),
##   GSM3036132 = col_double(),
##   GSM3036133 = col_double(),
##   GSM3036134 = col_double(),
##   GSM3036135 = col_double()
## )
## File stored at:

## C:\Users\Public\Documents\ESTsoft\CreatorTemp\RtmpqUo95R/GPL17586.soft

## Warning: 5990 parsing failures.
##   row   col expected actual         file
## 67363 start a double    --- literal data
## 67363 stop  a double    --- literal data
## 67364 start a double    --- literal data
## 67364 stop  a double    --- literal data
## 67365 start a double    --- literal data
## ..... ..... ........ ...... ............
## See problems(...) for more details.

data <- extract_gse(gse, "../geo_Rdata", geo_series_idx) # data 는 exprs_mat, gene_info, annot_data 를 담고있는 리스트
attach(data)

# factor variables
drug_resp <- table(annot_data$"treatment response:ch1")
barplot(drug_resp, main = "sample types")
```



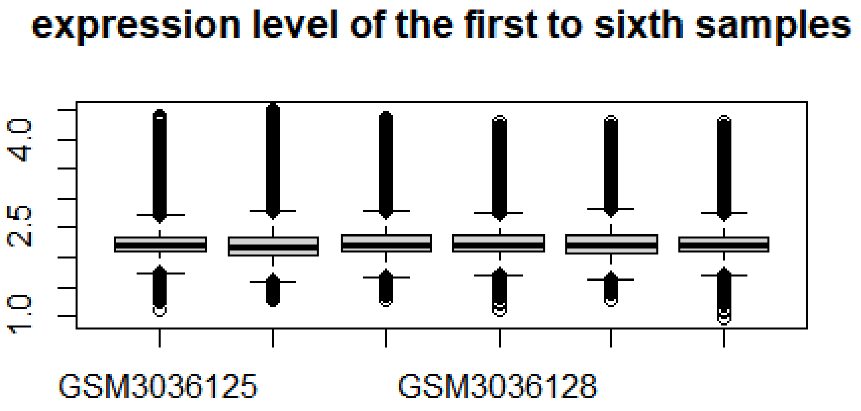
```
# numeric variables
hist(exprs_mat[,1], main="expression level of the first sample")  # normal distribution 에 가까운지 확인 가능
```



```
hist(log2(exprs_mat[,1]), main="expression level of the first sample")  # 
log2 transform 시, normal distribution 에 가까움
```



```
exprs_mat <- log2(exprs_mat)
boxplot(exprs_mat[,1:6], main="expression level of the first to sixth sampl
es")  # 1~6 번 sample 의 boxplot
```



자주 사용하는 요약통계량 함수

- mean(): 평균
- sd(): 표준편차
- sum(): 합계
- median(): 중앙값
- min(): 최솟값
- max(): 최댓값
- n(): 빈도

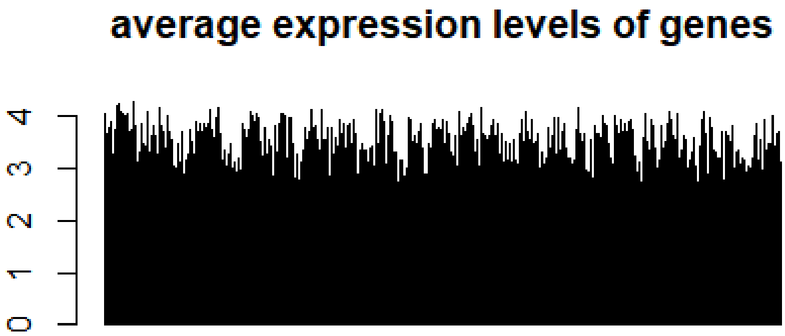
데이터프레임 조작하기

데이터프레임에 파생변수를 생성할 때는 데이터프레임$새변수이름 <- 결과값과 같이 입력한다.

주요 함수:

- rowMeans(행렬): 같은 행끼리 평균 산출
- colMeans(행렬 또는 데이터프레임): 같은 열끼리 평균 산출
- ifelse(조건문, 참일 때 결과, 거짓일 때 결과) : 범주형 변수 만드는 방법

```
gene_info$average_expression <- rowMeans(exprs_mat)
barplot(gene_info$average_expression, main="average expression levels of genes")
```



```
annot_data$group <- ifelse(annot_data$`treatment response:ch1` == "progress
or", "PR", "R")
table(annot_data$group)

##
## PR  R
##  5  6
```

데이터 추출하기

Way 1) 데이터프레임에 인덱싱 기호([]) 붙여 인덱싱

```
dim(exprs_mat)

## [1] 70523    11

r_subset <- exprs_mat[c(1:4),]  # 행 추출
dim(r_subset)

## [1]  4 11

r_subset <- exprs_mat[sample(1:nrow(exprs_mat), 2),]  # 2 개 행 랜덤 샘플링
dim(r_subset)

## [1]  2 11

c_subset <- exprs_mat[, "GSM3036125"]  # 열 추출
dim(c_subset)

## NULL
```

```
c_subset <- exprs_mat[,annot_data$`treatment response:ch1`=="progressor"]
 # 조건 만족 열 추출
dim(c_subset)

## [1] 70523     5

median_exprs_level <- median(exprs_mat[1,])
c_subset <- exprs_mat[,exprs_mat[1,] >= median_exprs_level]  # 조건 만족 열
추출
dim(c_subset)

## [1] 70523     6

median_exprs_level2 <- median(exprs_mat[2,])
c_subset <- exprs_mat[,exprs_mat[1,] >= median_exprs_level &
                      exprs_mat[2,] >= median_exprs_level2]  # 여러 조건 동
시에 만족 열 추출
dim(c_subset)

## [1] 70523     4

c_subset <- exprs_mat[,exprs_mat[1,] >= median_exprs_level |
                      exprs_mat[2,] >= median_exprs_level2]  # 여러 조건 중
하나라도 만족 열 추출
dim(c_subset)

## [1] 70523     8
```

Way 2) `subset()` 이용한 추출 데이터프레임 형식으로 된 자료의 경우 base::subset()함수를 이용하면 좀더 직관적으로 인덱싱을 할 수 있다.

```
drug <- read.csv("../PD-L1_resistance/drug.csv", stringsAsFactors = T)  # T
CGA-COAD clinical data 중 drug info, 문자열은 factor 변수 타입으로 처리
dim(drug)

## [1] 595  29

# drugname 이 Avastin 인 행 추출
Avastin <- subset(drug, clinical_drug_coad.pharmaceutical_therapy_drug_nam
e == "Avastin")
dim(Avastin)

## [1] 14 29
```

[참고] dplyr 을 이용한 데이터 추출 -> 1 권 4 장 확인 filter(): 행 추출 select(): 열(변수) 추출 arrange(): 정렬 mutate(): 변수 추가 summarize(): 통계치 산출 group_by(): 집단별로 나누기 left_join(): 데이터 합치기(열) bind_rows(): 데이터 합치기(행) n(): 레코드 수 산출

문자열 매칭 : grep(), grepl()

해당 문자열(패턴)을 포함하는 원소를 찾을 때 `grep()` 또는 `grepl()` 함수가 유용하다.

- `grep(패턴, 변수)`: 패턴이 들어있는 문자열 변수의 인덱스 반환 (원래 변수의 길이보다 작거나 같음)
- `grepl(패턴, 변수)`: 패턴이 들어있으면 TRUE, 없으면 FALSE 인 원래 변수와 같은 길이의 벡터 반환

주요 옵션인 ignore.case 의 기본값은 ignore.case = FALSE 이며 대소문자를 구분한다는 뜻이다.

subset 함수에는 grepl 함수를 적용한다.

```
# grep 사용
grep("base", rownames(installed.packages()))  # base 패턴을 포함하는 설치된
패키지 인덱스

## [1]  12  14 250 315

# grepl 과 subset 사용
# drug name 으로 beva 또는 avastin 을 대소문자 구분없이 포함하는 행 추출
grepl("(beva|avastin)", drug$clinical_drug_coad.pharmaceutical_therapy_dru
g_name, ignore.case = T)[1:30]

##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE F
ALSE
## [13]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FA
LSE
## [25] FALSE  TRUE FALSE FALSE FALSE FALSE

beva <- subset(drug, grepl("(beva|avastin)", clinical_drug_coad.pharmaceut
ical_therapy_drug_name, ignore.case = T))
```

Trouble shootings with R

R 을 이용하면서 자주 생기는 문제상황에 대한 솔루션이다.

패키지 설치가 안될 때

패키지를 설치하려고 할 때 **Permission denied** 등의 문구가 뜨며 설치가 완료되지 않는 경우가 있다. 이런 때 해볼 수 있는 조치는 순서대로 다음과 같다.

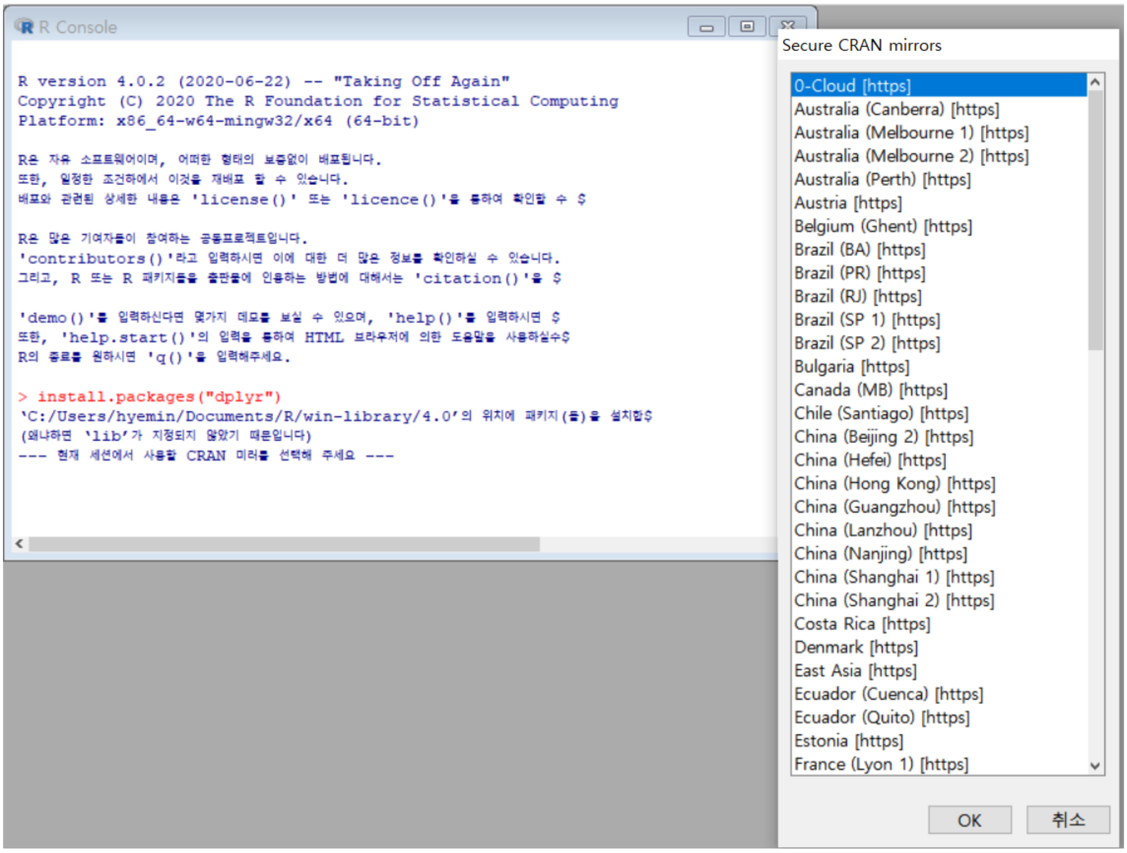
1. 00LOCK 디렉토리 삭제

내문서/R/win-library/4.0 디렉토리에 00LOCK 이라는 이름의 폴더가 새로 생겼을 수 있다. 00LOCK 폴더를 삭제하고 다시 설치를 진행해보자.

2. R 또는 RStudio 를 관리자 권한으로 실행

R 또는 RStudio 아이콘을 오른쪽 클릭하여 *관리자 권한으로 실행*한 뒤 설치를 진행해보자.

3. R 로 열어 설치할 때, CRAN mirror 을 0-Cloud 로 지정

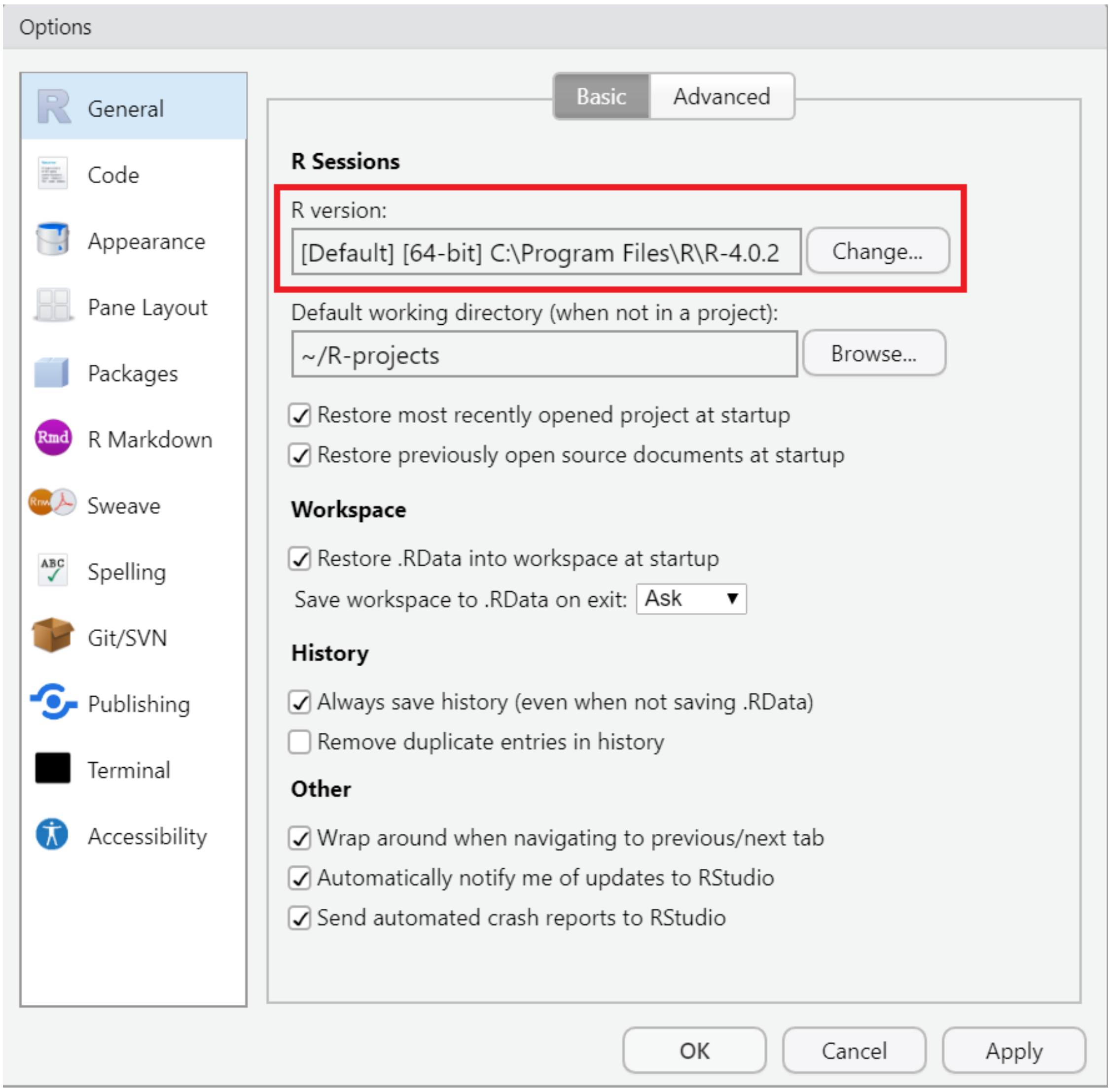


```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R은 자유 소프트웨어이며, 어떠한 형태의 보증없이 배포됩니다.
또한, 일정한 조건하에서 이것을 재배포 할 수 있습니다.
배포와 관련된 상세한 내용은 'license()' 또는 'licence()'을 통하여 확인할 수 $

R은 많은 기여자들이 참여하는 공동프로젝트입니다.
'contributors()'라고 입력하시면 이에 대한 더 많은 정보를 확인하실 수 있습니다.
그리고, R 또는 R 패키지들을 출판물에 인용하는 방법에 대해서는 'citation()'을 $

'demo()'를 입력하신다면 몇가지 데모를 보실 수 있으며, 'help()'를 입력하시면 $
또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을 사용하실수$
R의 종료를 원하시면 'q()'을 입력해주세요.

> install.packages("dplyr")
'C:/Users/hyemin/Documents/R/win-library/4.0'의 위치에 패키지(들)을 설치합$
(왜냐하면 'lib'가 지정되지 않았기 때문입니다)
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---
```

4. R, RStudio 삭제 후 재설치한 뒤 패키지 설치

R version not available 과 같은 문구가 뜨는 경우에는 현재 설치된 R version 이 패키지의 R version requirement 와 맞지 않는 경우이다. 꼭 필요한 경우에는 버전에 맞는 R 을 추가 설치하고 (기존 R 버전 지울 필요 없음) interpreter 을 버전에 맞게 지정할 수 있다.

RStudio 에서 Tools > Global Options… 또는 Tools > Project Options…를 열어 기본 또는 해당 프로젝트의 General 탭에서 R version 을 지정할 수 있다.



R Sessions

R version:
[Default] [64-bit] C:\Program Files\R\R-4.0.2

Default working directory (when not in a project):
~/R-projects

☑ Restore most recently opened project at startup
☑ Restore previously open source documents at startup

Workspace

☑ Restore .RData into workspace at startup
Save workspace to .RData on exit: Ask

History

☑ Always save history (even when not saving .RData)
□ Remove duplicate entries in history

Other

☑ Wrap around when navigating to previous/next tab
☑ Automatically notify me of updates to RStudio
☑ Send automated crash reports to RStudio

Tools > Global Options… > General

위와 같은 순서로도 해결되지 않을 때에는 해당 패키지를 포기할 수 밖에 없다.

함수 실행 중 문제가 생겼을 때

help(함수명) 또는 ? 함수명을 입력하여 함수의 Usage, 실행에 필요한 Arguments, 옵션, Examples 를 확인할 수 있다.

```
# 함수 documentation 확인
help(grep)
? grep
```

```
# package documentation 확인
?? dplyr
```

해당 함수를 구글링하여 필요한 패키지 설치가 안되어있다면 패키지 설치, 로드 후 함수 실행을 해야할 수도 있다.

변수 이름 설정 시 유의사항, 예약어 목록

변수 이름에는 영문, 언더바(_), 온점(.), 숫자만 사용되며, 영문으로 시작해야 한다.

변수 할당은 변수명 `<-` 내용물 또는 변수명 `=` 내용물과 같이 한다.

변수명은 대소문자를 구분한다.

아래는 R 에서 이미 사용 중인 예약어로, 변수 이름으로 사용할 수 없다.

```
## Warning in kable_pipe(x = structure(c("if", "function", "break", "Inf",
 : The
## table should have a header (column names)
```

R 예약어

if	else	repeat	while
function	for	in	next
break	TRUE	FALSE	NULL
Inf	NaN	NA	NA_integer_
NA_real_	NA_complex_	NA_character_	...
..1	..2		

코드의 재현성을 위한 팁

같은 코드가 항상 똑같은 결과값을 제공하는지 확인하기 위해서 코드 맨 첫 줄에 작업 공간을 비우는 라인을 추가하거나, sample()등의 함수를 이용하여 랜덤 추출을 할 때, random number generator 을 fix 할 수 있다.

```
rm(list = ls())  # R 작업 공간(메모리) 클리어

x <- 1:20
sample(x, 5)  # 사용법: sample(전체집단벡터, 샘플수)
```

```
## [1]  8  7  4 17 10
```

```
sample(x, 5)  # 이전에 나온 결과값과 다르다
```

```
## [1]  4  9 19 15 18
```

```
# set seed(숫자)를 입력한 이후 랜덤 숫자 생성 순서가 고정됨
set.seed(1)
sample(x, 5)  # A
```

```
## [1]  4  7  1  2 13
```

```
sample(x, 5)  # B
```

```
## [1] 11 14 18  1  5
```

```
set.seed(1)
sample(x, 5)  # A'
```

```
## [1]  4  7  1  2 13
```

```
sample(x, 5)  # B'
```

```
## [1] 11 14 18  1  5
```

함수 결과 출력이 안될 때, 객체를 찾을 수 없을 때

함수 결과물을 <-을 이용하여 따로 객체에 저장하지 않으면 기본적으로 결과값을 출력만 하고 재사용이 안된다.

한편, 함수 결과물을 객체에 저장하면 함수 실행 시 결과값이 콘솔창에 출력되지 않는다.

다음 예제를 보자.

```
rm(list = ls())  # R 작업 공간(메모리) 클리어
```

```
read.csv("./data/ch1-useful_libraries_csv.csv")
```

```
##           Library.name
## 1                 affy
## 2             annotate
## 3        AnnotationDbi
## 4  ArrayQualityMetrics
## 5              Biobase
## 6          BiocManager
## 7              biomaRt
## 8                dplyr
## 9               EDASeq
## 10               edgeR
## 11            GEOquery
## 12             ggplot2
## 13               GO.db
## 14               gosim
## 15             GOstats
## 16            GSEABase
## 17              igraph
## 18             KEGG.db
## 19           KEGGgraph
## 20               limma
## 21             recount
## 22             stringr
## 23 SummarizedExperiment
## 24            survival
## 25           survminer
## 26        TCGAbiolinks
## 27               topGO
## 28                xlsx
##                                                                   Applicat
ion
## 1                                           handling matrices for TCGAbio
links
## 2              support user actions that rely on the different metadata 
packages
## 3       interface and database connection functions for annotation data 
packages
## 4                            generates a quality report for the microarr
ay data
## 5                                              handling matrices for GEO
 data
## 6                                                        Bioconductor ac
cess
## 7                                                  converting to gene sy
mbols
## 8                                                   handling data frames
 in R
## 9                                         matrix normalization for TCGAbi
olinks
## 10                                               DEG analysis of RNA-seq
 data
## 11                                                           GEO data ac
cess
## 12                                                             plotting 
data
## 13                                      Annotation maps for Gene Ontolog
y (GO)
## 14 the computation of functional similarities between GO terms and a gen
e product
## 15                                        interact with GO and microarra
y data
## 16                                         Gene Set Enrichment Analysis 
```

```
(GSEA)
## 17         simple graphs and networks as well as for graph analysis and 
plotting
## 18                                                   Annotation maps for
 KEGG
## 19   interface between the KEGG pathway and R and the required analysis 
functions
## 20                                                              DEG anal
ysis
## 21                                          access all meta data in GDC p
ortal
## 22                                                      handling strings 
in R
## 23                                          handling matrices for TCGAbio
links
## 24                                                         survival anal
ysis
## 25                                                             survival 
plot
## 26                                                          TCGA data ac
cess
## 27                                                         test the GO t
erms
## 28                                        read/write/format Excel file f
ormats
##      Repository
## 1  Bioconductor
## 2  Bioconductor
## 3  Bioconductor
## 4  Bioconductor
## 5  Bioconductor
## 6          CRAN
## 7  Bioconductor
## 8          CRAN
## 9  Bioconductor
## 10 Bioconductor
## 11 Bioconductor
## 12         CRAN
## 13 Bioconductor
## 14 Bioconductor
## 15 Bioconductor
## 16 Bioconductor
## 17         CRAN
## 18 Bioconductor
## 19 Bioconductor
## 20 Bioconductor
## 21 Bioconductor
## 22         CRAN
## 23 Bioconductor
## 24         CRAN
## 25         CRAN
## 26 Bioconductor
## 27 Bioconductor
## 28         CRAN
```

```r
ls()  # 현재 작업 공간 내 변수 출력
```

```
## character(0)
```

```r
ch1_table1 <- read.csv("./data/ch1-useful_libraries_csv.csv")
ls()  # 현재 작업 공간 내 변수 출력
```

```
## [1] "ch1_table1"
```

```r
head(ch1_table1, 3)
```

```
##    Library.name
## 1          affy
## 2      annotate
## 3 AnnotationDbi
##                                                             Application
## 1                                    handling matrices for TCGAbiolinks
## 2        support user actions that rely on the different metadata package
s
## 3 interface and database connection functions for annotation data packag
es
##     Repository
## 1 Bioconductor
## 2 Bioconductor
## 3 Bioconductor
```

Chapter 2: Getting started with Differential Expression Analysis

Hyemin Gu

2020-12-15

Table of Contents

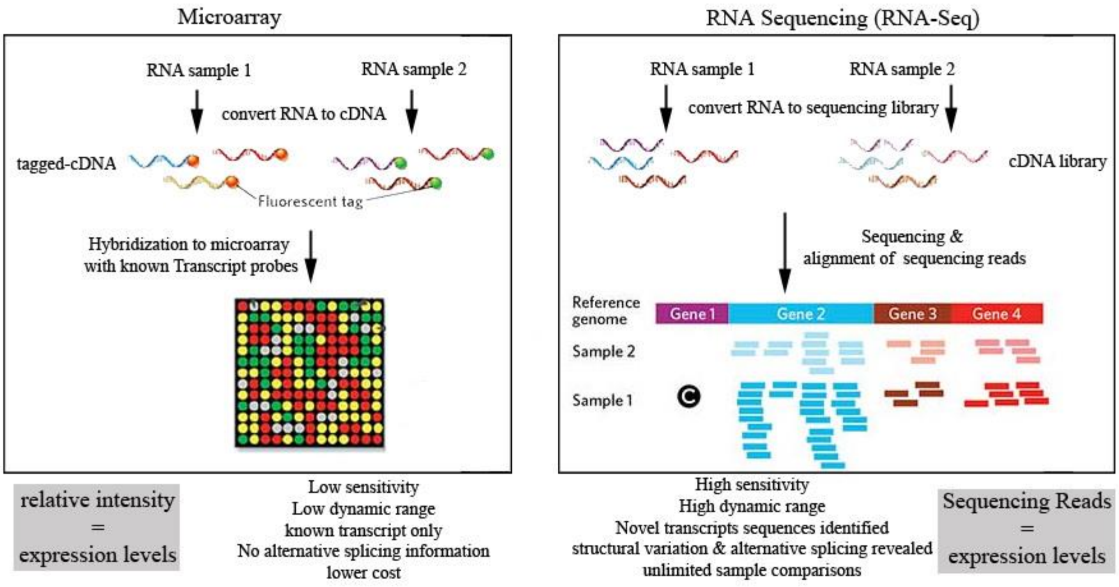
Getting started

Differential Expression Analysis(DEA)란 샘플간 유전자의 발현량 차이를 통계적으로 비교하여 유의한 차이를 보이는 유전자, Differentially expressed genes(DEGs)를 선별하여 기능 분석을 하는 등의 연구 방법이다.

실험의 목적에 따라 샘플을 실험군과 대조군, 또는 여러 개의 그룹으로 나누어 집단들 간의 유전자 발현량 차이를 pairwise 하게 비교하는 single-factor comparison 방법이 일반적이다. 다른 방법론으로 time series analysis 도 있다.

유전자의 발현량(gene expression level)은 유전자 발현량의 프로파일링 방법에 따라 다르게 정의 된다. Microarray data 의 경우 실수값을 갖는 relative intensity 로, Microarray 의 단점을 보완하기 위한 최신의 NGS 방식으로 얻어진

RNA-Seq data 의 경우 정수값을 갖는 sequencing read count 로 정의된다.



실험 세팅이 동일하더라도 프로파일링 방법이 다르면 서로 다른 파이프라인으로 DEG 를 계산해야 한다. 이 장에서는 공개 데이터 포털에서 이들 데이터를 얻는 방법과 전반적인 Workflow 를 소개하도록 한다.

[참고] Seeing the Unseen: Microarray-Based Gene Expression Profiling in Vision

[참고] https://www.otogenetics.com/rna-sequencing-vs-microarray/

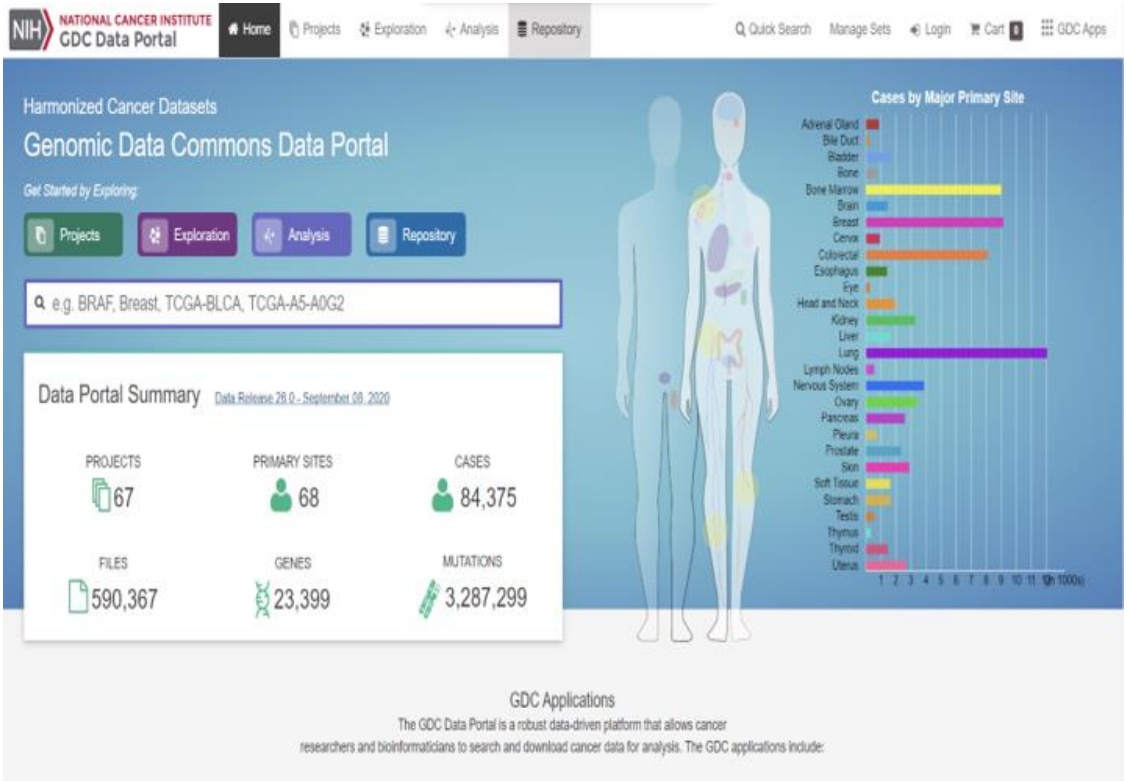
Data access

바이오 인포매틱스 분야의 논문을 보면 공개 데이터를 연구에 메인으로 이용하거나 validation 을 위해 사용하기도 한다. 그리고 연구에 사용한 데이터는 아카이브에 업로드하여 공개한다. 따라서 연구 데이터를 찾을 때 일차적으로 **관심 주제의 논문을 구글 또는 구글 학술검색으로 찾아 dataset 이 공개된 아카이브와 그 accession number 을 얻고 해당 아카이브에서 데이터를 검색**하면 양질의 데이터를 얻을 수 있다.

Searching for open data from archives

위에서 알아본 바와 같이 사용할 dataset 의 accession number 을 알고 있거나 연관 검색어를 통해 아카이브에서 dataset 을 찾는 방법을 알아보자. Gene expression data 를 보유한 아카이브는 GDC portal, GEO, SRA, Array Express, ENA, DRA 등이 있다. 각 아카이브별 특징과 dataset 검색 방법을 알아보자.

GDC portal



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Data Portal Summary Data Release 26.0 - September 08, 2020

PROJECTS 67
PRIMARY SITES 68
CASES 84,375
FILES 590,367
GENES 23,399
MUTATIONS 3,287,299

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

GDC portal main

GDC portal(https://portal.gdc.cancer.gov/)은 미국 국립보건원 (National Institutes of Health, NIH) 산하에 있는 National Cancer Institute 에서 운영하는 데이터 포털이며 cancer data 를 검색 및 다운로드할 수 있다. 대표적으로 **The Cancel Genome Atlas(TCGA)** 에서 33 개의 암종에 대해 다양한 프로젝트를 진행하여 유전체/전사체/단백체 데이터를 수집 및 분석했다.

다양한 범주의 데이터 분류 코드 테이블을 조회할 수 있는 링크이다. https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables

BCR Batch Codes, Center Codes, Data Levels, Data Types, Platform Codes, Portion / Analyte Codes, Sample Type Codes, TCGA Study Abbreviations, Tissue Source Site Codes 가 수록되어 있다. 일례로 Sample Type Codes 를 보면 다음과 같다.

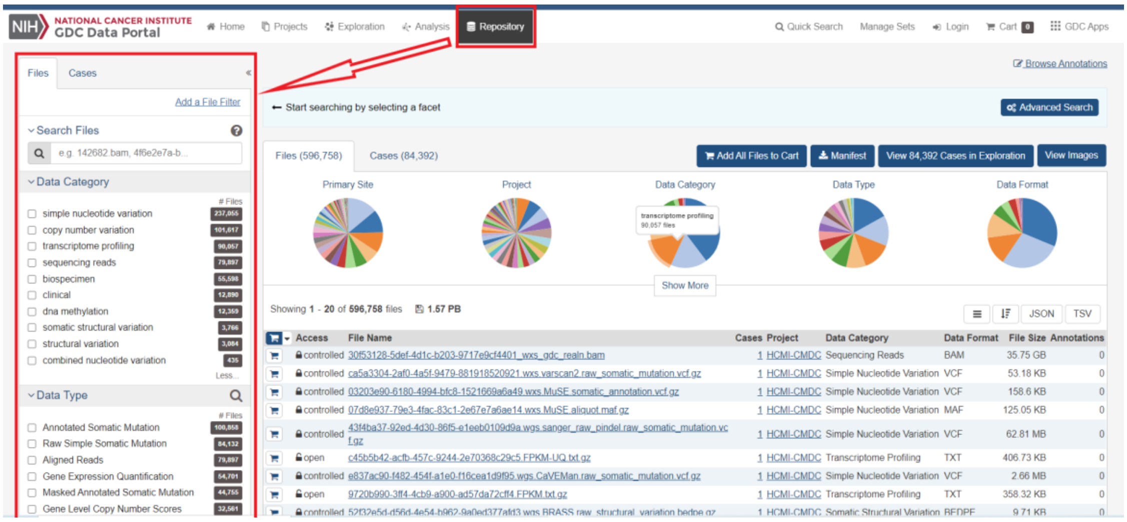
TCGA sample type codes

Code	Definition	Short.Letter.Code
1	Primary Solid Tumor	TP
2	Recurrent Solid Tumor	TR
3	Primary Blood Derived Cancer - Peripheral Blood	TB
4	Recurrent Blood Derived Cancer - Bone Marrow	TRBM

5	Additional - New Primary	TAP
6	Metastatic	TM
7	Additional Metastatic	TAM
8	Human Tumor Original Cells	THOC
9	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
15	sample type 15	15SH
16	sample type 16	16SH
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP
61	Cell Line Derived Xenograft Tissue	XCL
99	sample type 99	99SH

한편, 데이터의 분류 방법이 개편되면서 harmonized database 와 기존 방식으로 분류된 legacy archive 가 별개로 존재하게 되었다. harmonized database 는 https://portal.gdc.cancer.gov/repository 에서, legacy archive 는 https://portal.gdc.cancer.gov/legacy-archive/search/f 에서 액세스할 수 있다. **repository 에서 조회한 데이터는 샘플별로 카트에 담아 다운받을 수 있다. 추후에, R 을 이용하여 대량의 데이터를 덤프하는 방법을 알아볼 것이다.**

harmonized database



Files Cases

Add a File Filter

Search Files

e.g. 142682.bam, 4f6e2e7a-b...

Data Category

Data Category	# Files
□ simple nucleotide variation	237,055
□ copy number variation	101,617
□ transcriptome profiling	90,057
□ sequencing reads	79,897
□ biospecimen	55,598
□ clinical	12,890
□ dna methylation	12,359
□ somatic structural variation	3,766
□ structural variation	3,084
□ combined nucleotide variation	435

Less...

Data Type

Data Type	# Files
□ Annotated Somatic Mutation	100,858
□ Raw Simple Somatic Mutation	84,132
□ Aligned Reads	79,897
□ Gene Expression Quantification	54,701
□ Masked Annotated Somatic Mutation	44,755
□ Gene Level Copy Number Scores	32,561

Showing 1 - 20 of 596,758 files 1.57 PB

JSON TSV

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
controlled	30f53128-5def-4d1c-b203-9717e9cf4401_wxs_gdc_realn.bam	1	HCMI-CMDC	Sequencing Reads	BAM	35.75 GB	0
controlled	ca5a3304-2af0-4a5f-9479-881918520921.wxs.varscan2.raw_somatic_mutation.vcf.gz	1	HCMI-CMDC	Simple Nucleotide Variation	VCF	53.18 KB	0
controlled	03203e90-6180-4994-bfc8-1521669a6a49.wxs.MuSE.somatic_annotation.vcf.gz	1	HCMI-CMDC	Simple Nucleotide Variation	VCF	158.6 KB	0
controlled	07d8e937-79e3-4fac-83c1-2e67e7a6ae14.wxs.MuSE.aliquot.maf.gz	1	HCMI-CMDC	Simple Nucleotide Variation	MAF	125.05 KB	0
controlled	43f4ba37-92ed-4d30-86f5-e1eeb0109d9a.wgs.sanger_raw_pindel.raw_somatic_mutation.vcf.gz	1	HCMI-CMDC	Simple Nucleotide Variation	VCF	62.81 MB	0
open	c45b5b42-acfb-457c-9244-2e70368c29c5.FPKM-UQ.txt.gz	1	HCMI-CMDC	Transcriptome Profiling	TXT	406.73 KB	0
controlled	e837ac90-f482-454f-a1e0-f16cea1d9f95.wgs.CaVEMan.raw_somatic_mutation.vcf.gz	1	HCMI-CMDC	Simple Nucleotide Variation	VCF	2.66 MB	0
open	9720b990-3ff4-4cb9-a900-ad57da72cff4.FPKM.txt.gz	1	HCMI-CMDC	Transcriptome Profiling	TXT	358.32 KB	0
controlled	52f32e5d-d56d-4e54-b962-9a0ed377afd3.wgs.BRASS.raw_structural_variation.bedpe.gz	1	HCMI-CMDC	Somatic Structural Variation	BEDPE	9.71 KB	0

Repository of GDC harmonized database

Link to GDC harmonized database : https://portal.gdc.cancer.gov/repository

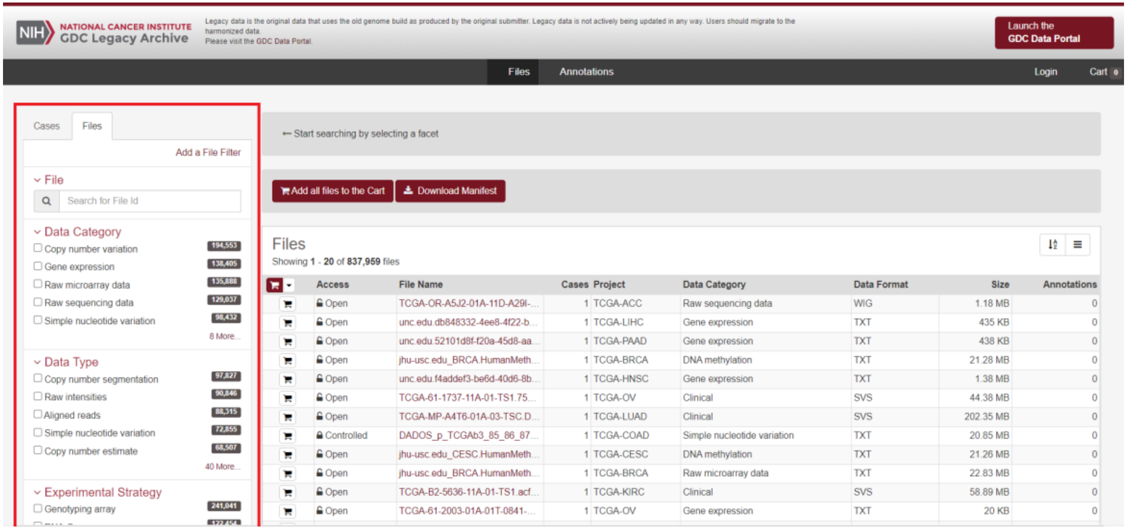
GDC portal 메인에서 Repository 탭을 클릭해서 들어가면 Files(Data category, Data type, Experimental Strategy 등) 또는 Cases(Project, Disease type, clinical 등)에 따라 구분된 데이터를 찾을 수 있다.

Gene expression data 는 RNA-Seq 타입만 제공되며 다음과 같이 분류된다.

- Data Category: Transcriptome Profiling
  - Data Type: Gene Expression Quantification
    - Experiment Strategy: RNA-Seq

이에 속하는 Workflow type 으로는 HTSeq - Counts, HTSeq - FPKM, HTSeq - FPKM-UQ, STAR - Counts 이 있다.

legacy archive


NIH
NATIONAL CANCER INSTITUTE
GDC Legacy Archive
Launch the GDC Data Portal
Files
Annotations
Login
Cart 0
Cases
Files
Add a File Filter
File
Search for File Id
← Start searching by selecting a facet
Add all files to the Cart
Download Manifest

Legacy data is the original data that uses the old genome build as produced by the original submitter. Legacy data is not actively being updated in any way. Users should migrate to the harmonized data.
Please visit the GDC Data Portal.

Data Category	
□ Copy number variation	194,553
□ Gene expression	138,405
□ Raw microarray data	135,888
□ Raw sequencing data	129,037
□ Simple nucleotide variation	98,432
8 More...	

Data Type	
□ Copy number segmentation	97,827
□ Raw intensities	90,846
□ Aligned reads	88,315
□ Simple nucleotide variation	72,855
□ Copy number estimate	68,507
40 More...	

Experimental Strategy	
□ Genotyping array	241,041

Files
Showing **1** - **20** of **837,959** files

	Access	File Name	Cases	Project	Data Category	Data Format	Size	Annotations
	Open	TCGA-OR-A5J2-01A-11D-A29I-...	1	TCGA-ACC	Raw sequencing data	WIG	1.18 MB	0
	Open	unc.edu.db848332-4ee8-4f22-b...	1	TCGA-LIHC	Gene expression	TXT	435 KB	0
	Open	unc.edu.52101d8f-f20a-45d8-aa...	1	TCGA-PAAD	Gene expression	TXT	438 KB	0
	Open	jhu-usc.edu_BRCA.HumanMeth...	1	TCGA-BRCA	DNA methylation	TXT	21.28 MB	0
	Open	unc.edu.f4addef3-be6d-40d6-8b...	1	TCGA-HNSC	Gene expression	TXT	1.38 MB	0
	Open	TCGA-61-1737-11A-01-TS1.75...	1	TCGA-OV	Clinical	SVS	44.38 MB	0
	Open	TCGA-MP-A4T6-01A-03-TSC.D...	1	TCGA-LUAD	Clinical	SVS	202.35 MB	0
	Controlled	DADOS_p_TCGAb3_85_86_87...	1	TCGA-COAD	Simple nucleotide variation	TXT	20.85 MB	0
	Open	jhu-usc.edu_CESC.HumanMeth...	1	TCGA-CESC	DNA methylation	TXT	21.26 MB	0
	Open	jhu-usc.edu_BRCA.HumanMeth...	1	TCGA-BRCA	Raw microarray data	TXT	22.83 MB	0
	Open	TCGA-B2-5636-11A-01-TS1.acf...	1	TCGA-KIRC	Clinical	SVS	58.89 MB	0
	Open	TCGA-61-2003-01A-01T-0841-...	1	TCGA-OV	Gene expression	TXT	20 KB	0

Repository of GDC legacy data

Link to GDC legacy archive : https://portal.gdc.cancer.gov/legacy-archive/search/f

Legacy archive 역시 Files(Data category, Data type, Experimental Strategy 등) 또는 Cases(Project, Disease type, clinical 등)에 따라 구분된 데이터를 찾을 수 있다.

Gene expression data 는 RNA-Seq 와 microarray 모두 제공되며 다음과 같이 분류된다.

- Data Category: Gene expression
  - Data Type: Gene expression quantification
    - Experiment Strategy: RNA-Seq
    - Experiment Strategy: Gene expression array
  - Data Type: Isoform expression quantification
    - Experiment Strategy: RNA-Seq
  - Data Type: Exon quantification
    - Experiment Strategy: RNA-Seq
  - Data Type: Exon junction quantification
    - Experiment Strategy: RNA-Seq

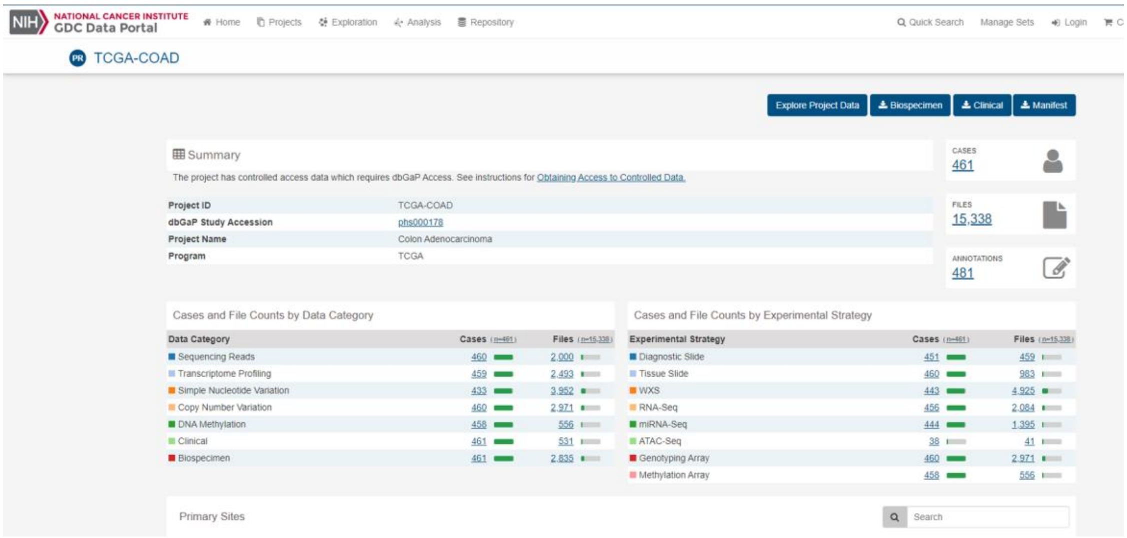
- Data Category: Raw microarray data
  - Data Type: Raw intensities
    - Experiment Strategy: Protein expression array
    - Experiment Strategy: Gene expression array
  - Data Type: Normalized intensities
    - Experiment Strategy: Gene expression array
  - Data Type: Intensities
    - Experiment Strategy: Protein expression array

        - Experiment Strategy: Gene expression array
- Data Category: Raw sequencing data
    - Data Type: Aligned reads
        - Experiment Strategy: RNA-Seq
    - Data Type: Unaligned reads
        - Experiment Strategy: RNA-Seq
- Data Category: Protein expression
    - Data Type: Protein expression quantification
        - Experiment Strategy: Protein expression array
- Data Category: Processed microarray data
    - Data Type: Processed intensities
        - Experiment Strategy: Gene expression array

다른 Data Category 와는 다르게 Processed microarray data 는 TCGAbiolinks 라는 R 라이브러리에서 제공하지 않는다.

Data exploration

위의 내용을 통해 GDC portal 의 데이터 저장소로부터 데이터를 찾고 샘플별로 다운받을 수 있었다. 데이터를 다운받기 전에 연구 프로젝트별로 data exploration 을 진행하려고 한다. 다음은 TCGA 데이터에서 Colon adenocarcinoma 를 연구한 TCGA-COAD 프로젝트를 조회하는 예제이다.



TCGA-COAD

Summary

The project has controlled access data which requires dbGaP Access. See instructions for Obtaining Access to Controlled Data.

Project ID	TCGA-COAD
dbGaP Study Accession	phs000178
Project Name	Colon Adenocarcinoma
Program	TCGA

CASES 461

FILES 15,338

ANNOTATIONS 481

Cases and File Counts by Data Category

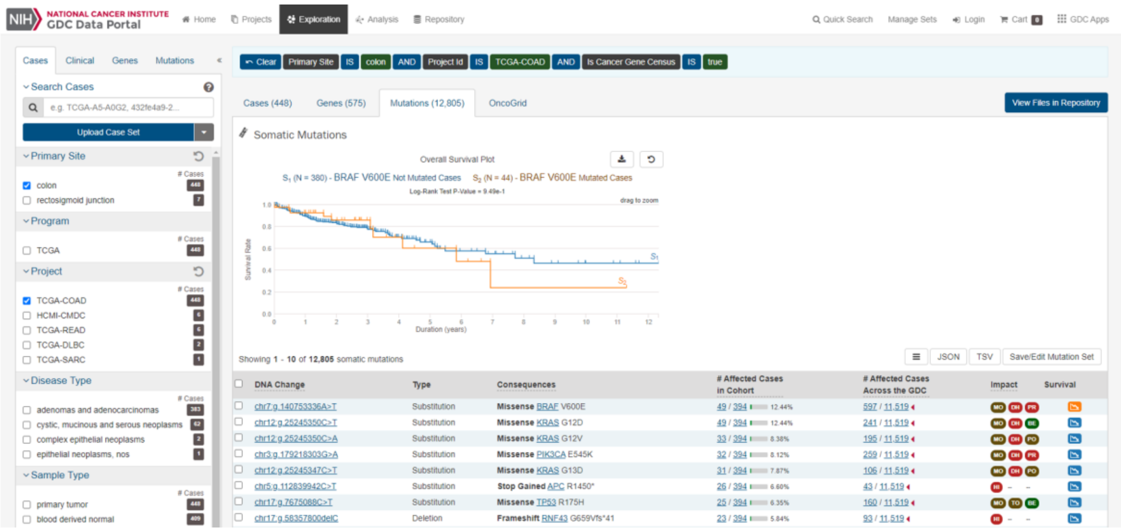
Data Category	Cases (n=461)	Files (n=15,338)
Sequencing Reads	460	2,000
Transcriptome Profiling	459	2,493
Simple Nucleotide Variation	433	3,952
Copy Number Variation	460	2,971
DNA Methylation	458	556
Clinical	461	531
Biospecimen	461	2,835

Cases and File Counts by Experimental Strategy

Experimental Strategy	Cases (n=461)	Files (n=15,338)
Diagnostic Slide	451	459
Tissue Slide	460	983
WXS	443	4,925
RNA-Seq	456	2,084
miRNA-Seq	444	1,395
ATAC-Seq	38	41
Genotyping Array	460	2,971
Methylation Array	458	556

Primary Sites

TCGA-COAD 프로젝트 설명



Primary Site

	# Cases
☑ colon	448
□ rectosigmoid junction	7

Program

	# Cases
□ TCGA	448

Project

	# Cases
☑ TCGA-COAD	448
□ HCMI-CMDC	6
□ TCGA-READ	6
□ TCGA-DLBC	2
□ TCGA-SARC	1

Disease Type

	# Cases
□ adenomas and adenocarcinomas	383
□ cystic, mucinous and serous neoplasms	62
□ complex epithelial neoplasms	2
□ epithelial neoplasms, nos	1

Sample Type

	# Cases
□ primary tumor	448
□ blood derived normal	409

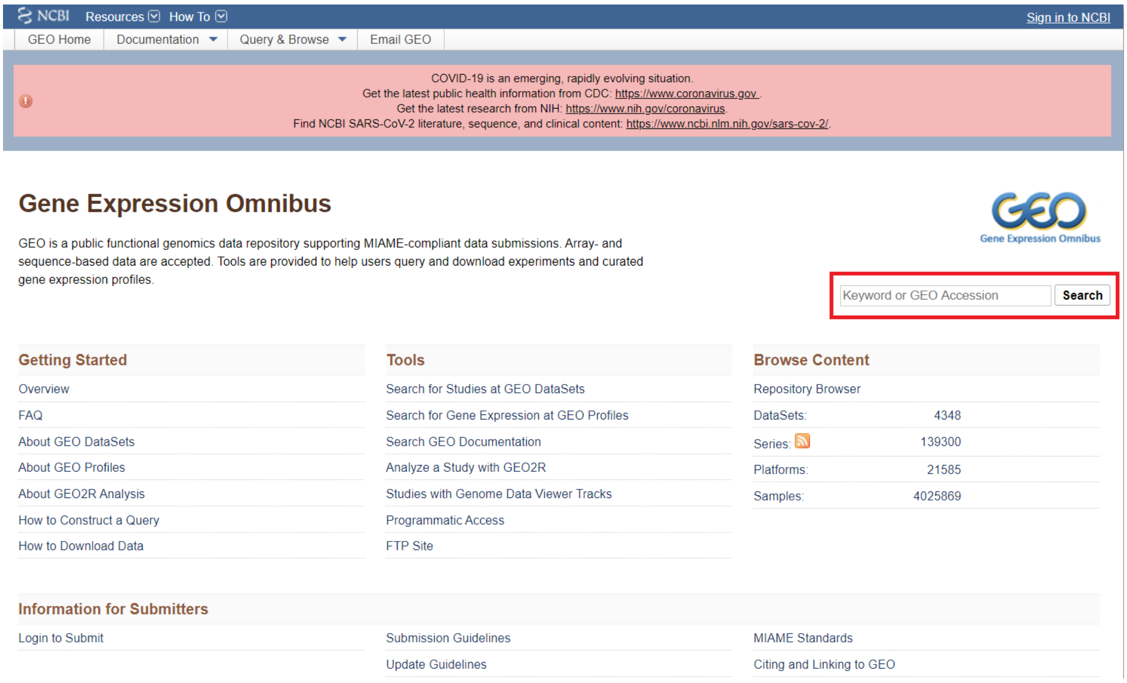
Showing 1 - 10 of 12,805 somatic mutations

□	DNA Change	Type	Consequences	# Affected Cases in Cohort	# Affected Cases Across the GDC	Impact	Survival
□	chr7:g.140753336A>T	Substitution	Missense BRAF V600E	49 / 394 12.44%	597 / 11,519	MO DH PR	
□	chr12:g.25245350C>T	Substitution	Missense KRAS G12D	49 / 394 12.44%	241 / 11,519	MO DH BE	
□	chr12:g.25245350C>A	Substitution	Missense KRAS G12V	33 / 394 8.38%	195 / 11,519	MO DH PO	
□	chr3:g.179218303G>A	Substitution	Missense PIK3CA E545K	32 / 394 8.12%	259 / 11,519	MO DH PR	
□	chr12:g.25245347C>T	Substitution	Missense KRAS G13D	31 / 394 7.87%	106 / 11,519	MO DH PO	
□	chr5:g.112839942C>T	Substitution	Stop Gained APC R1450*	26 / 394 6.60%	43 / 11,519	HI - -	
□	chr17:g.7675088C>T	Substitution	Missense TP53 R175H	25 / 394 6.35%	160 / 11,519	MO TO BE	
□	chr17:g.58357800delC	Deletion	Frameshift RNF43 G659Vfs*41	23 / 394 5.84%	93 / 11,519	HI - -	

TCGA-COAD data exploration

Exploration 탭에서 Primary Site=colon, Project=TCGA-COAD 를 설정하면 Cases, Genes, Mutations, OncoGrid 로 탭을 넘겨가며 데이터의 개략적 정보를 파악할 수 있다.

NCBI: GEO & SRA



COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

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
- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools
- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

Browse Content

Repository Browser	
DataSets:	4348
Series:	139300
Platforms:	21585
Samples:	4025869

Information for Submitters
- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO

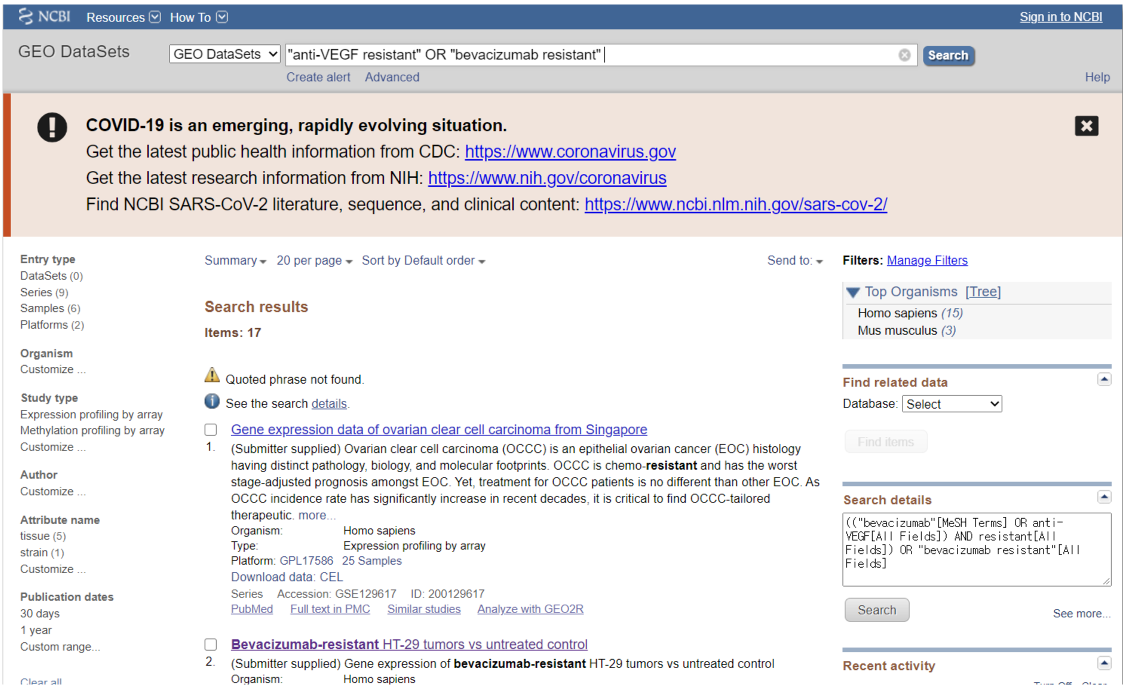
The Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/)는 The National Institutes of Health (NIH) 산하에 있는 U.S. National Library of Medicine 에서 운영하는 The National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) data portal 내에 구축된 Microarray data archive 이다. GEO 는 Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra)와 함께 NCBI 의 하위 archive 이다. RNA-Seq data 는 주로 SRA 에 저장되어 있다. 한편, RNA-Seq data 중에는 raw data 를 전처리한 데이터가 GEO 에 올라와 있는 경우도 있다.

GEO 에서 통용되는 accession code 는 샘플 단위의 GSM 또는 series data 의 경우 GSE 로 시작한다.

GEO 에서 데이터 찾기

GEO 메인 화면의 검색창에 키워드를 넣어 조건에 맞는 데이터를 검색할 수 있다. 키워드가 여러 개라면 각각을 AND 또는 OR 로 연결하여 쿼리를 입력한다.

예시: “anti-VEGF resistant” OR “bevacizumab resistant”



"anti-VEGF resistant" OR "bevacizumab resistant"

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov
Get the latest research information from NIH: https://www.nih.gov/coronavirus
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/

Entry type
DataSets (0)
Series (9)
Samples (6)
Platforms (2)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (5)
strain (1)
Customize ...

Publication dates
30 days
1 year
Custom range...

Clear all

Search results

Items: 17

⚠ Quoted phrase not found.

ⓘ See the search details.

□ 1. Gene expression data of ovarian clear cell carcinoma from Singapore
(Submitter supplied) Ovarian clear cell carcinoma (OCCC) is an epithelial ovarian cancer (EOC) histology having distinct pathology, biology, and molecular footprints. OCCC is chemo-**resistant** and has the worst stage-adjusted prognosis amongst EOC. Yet, treatment for OCCC patients is no different than other EOC. As OCCC incidence rate has significantly increase in recent decades, it is critical to find OCCC-tailored therapeutic. more...
Organism: Homo sapiens
Type: Expression profiling by array
Platform: GPL17586 25 Samples
Download data: CEL
Series Accession: GSE129617 ID: 200129617
PubMed Full text in PMC Similar studies Analyze with GEO2R

□ 2. **Bevacizumab-resistant** HT-29 tumors vs untreated control
(Submitter supplied) Gene expression of **bevacizumab-resistant** HT-29 tumors vs untreated control
Organism: Homo sapiens

Filters: Manage Filters

▼ Top Organisms [Tree]
Homo sapiens *(15)*
Mus musculus *(3)*

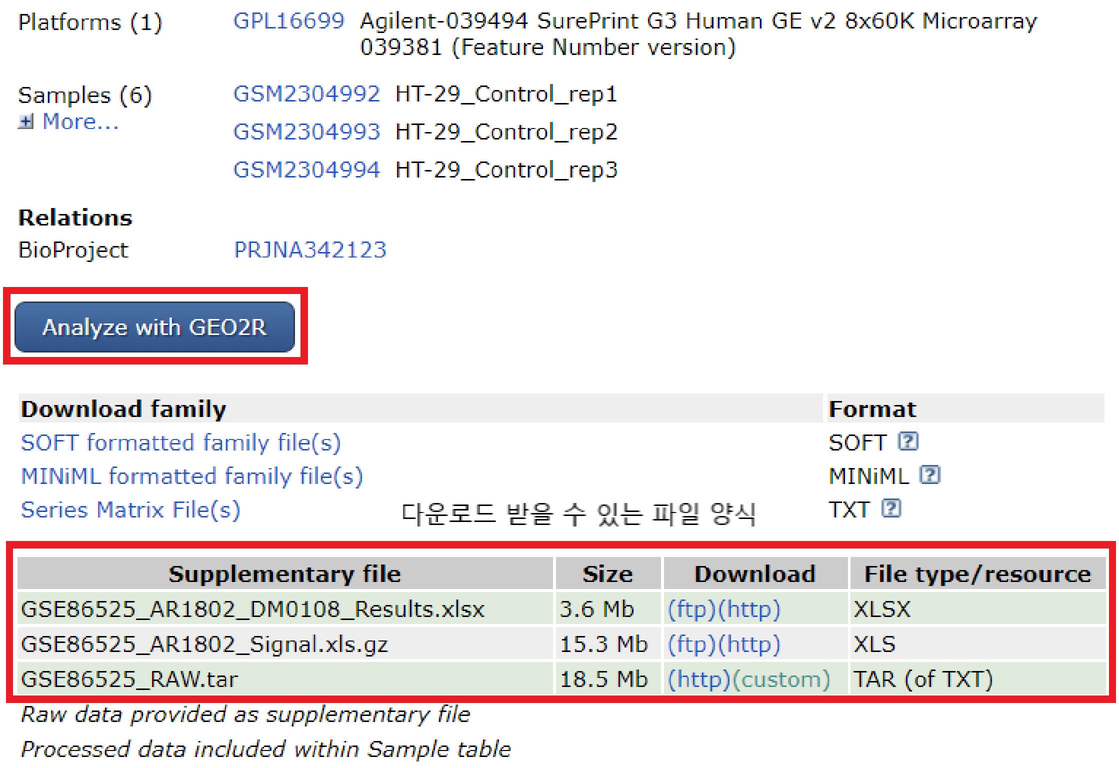
Find related data
Database: Select

Search details

```
(("bevacizumab"[MeSH Terms] OR anti-VEGF[All Fields]) AND resistant[All Fields]) OR "bevacizumab resistant"[All Fields]
```

GEO data 검색하기

각 데이터베이스 링크를 클릭하면 데이터 상세 페이지에 들어가진다. 여기에서 실험 상세 정보를 확인하고 데이터마다 지원하는 양식으로 자료를 다운받을 수도 있다. 한편, GEO2R 이라는 data exploration 툴로 gene expression 데이터를 살펴보고 웹상에서 DEA 를 할 수도 있다.



Platforms (1)	GPL16699	Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381 (Feature Number version)
Samples (6) ⊞ More...	GSM2304992	HT-29_Control_rep1
	GSM2304993	HT-29_Control_rep2
	GSM2304994	HT-29_Control_rep3

Relations

BioProject PRJNA342123

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

다운로드 받을 수 있는 파일 양식

Supplementary file	Size	Download	File type/resource
GSE86525_AR1802_DM0108_Results.xlsx	3.6 Mb	(ftp)(http)	XLSX
GSE86525_AR1802_Signal.xls.gz	15.3 Mb	(ftp)(http)	XLS
GSE86525_RAW.tar	18.5 Mb	(http)(custom)	TAR (of TXT)

Raw data provided as supplementary file

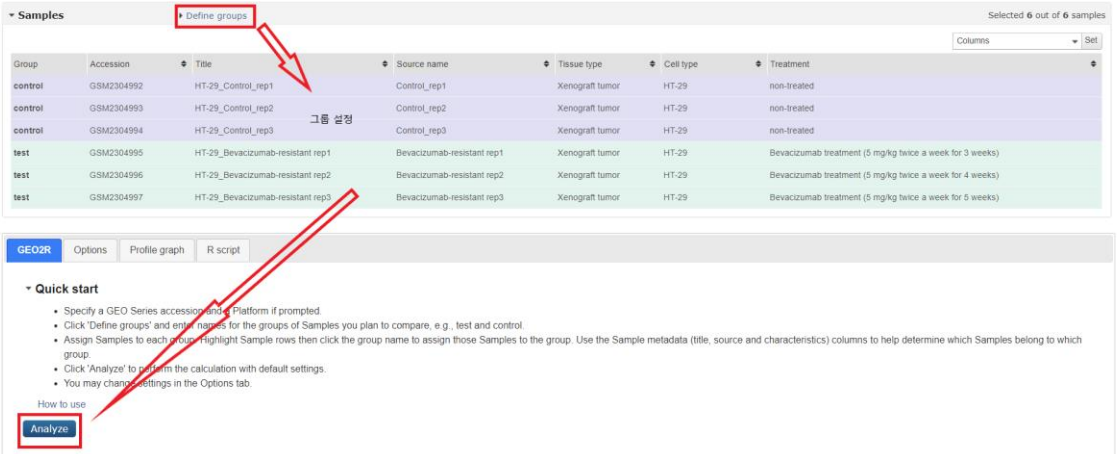
Processed data included within Sample table

데이터 상세 페이지

GEO2R 를 이용한 data exploration

GEO2R 을 클릭하면 데이터의 샘플을 레코드로 갖는 테이블이 나온다. data exploration 을 위해서는 샘플들을 비교할 그룹을 설정해야 하고 Analyze 를 클릭해야 한다. 순서대로 보면 다음과 같다.

2. 테이블 상단의 Define groups 에서 2 개 이상의 그룹을 정의한다.
3. 같은 그룹에 넣고싶은 샘플끼리 블록을 잡고 Define groups 의 해당 그룹을 클릭하면 그룹이 설정되고 색깔이 입혀진다.
4. 그룹 설정이 끝나면 테이블 아래의 GEO2R 탭에서 Analyze 를 클릭한다. OPtions 탭에서 DEA 를 위한 옵션을 설정할 수도 있다.



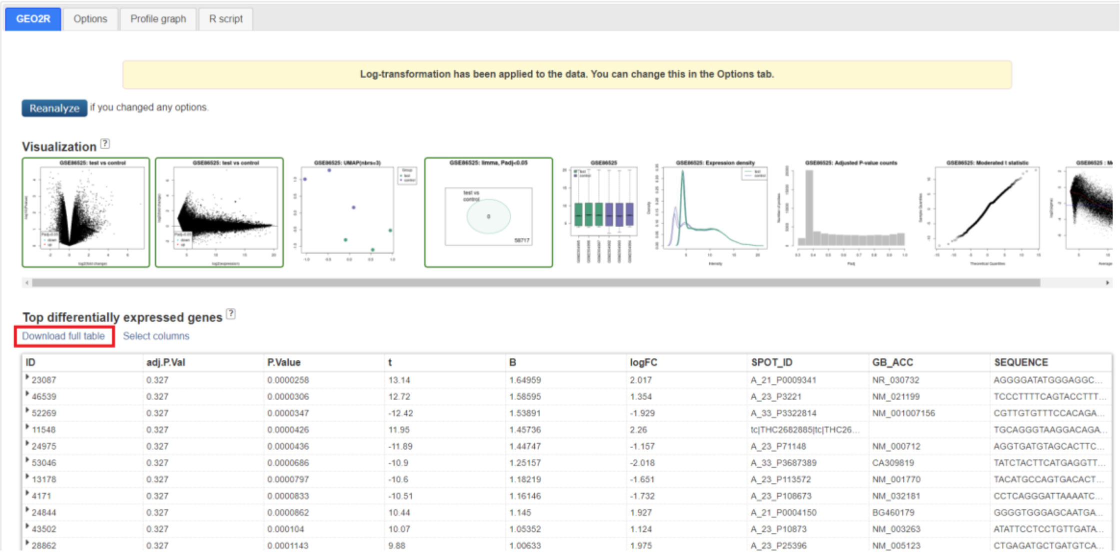
Group	Accession	Title	Source name	Tissue type	Cell type	Treatment
control	GSM2304992	HT-29_Control_rep1	Control_rep1	Xenograft tumor	HT-29	non-treated
control	GSM2304993	HT-29_Control_rep2	Control_rep2	Xenograft tumor	HT-29	non-treated
control	GSM2304994	HT-29_Control_rep3	Control_rep3	Xenograft tumor	HT-29	non-treated
test	GSM2304995	HT-29_Bevacizumab-resistant rep1	Bevacizumab-resistant rep1	Xenograft tumor	HT-29	Bevacizumab treatment (5 mg/kg twice a week for 3 weeks)
test	GSM2304996	HT-29_Bevacizumab-resistant rep2	Bevacizumab-resistant rep2	Xenograft tumor	HT-29	Bevacizumab treatment (5 mg/kg twice a week for 4 weeks)
test	GSM2304997	HT-29_Bevacizumab-resistant rep3	Bevacizumab-resistant rep3	Xenograft tumor	HT-29	Bevacizumab treatment (5 mg/kg twice a week for 5 weeks)

▾ 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, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong to which group.
- Click 'Analyze' to perform the calculation with default settings.
- You may change settings in the Options tab.

GEO2R 설정

Volcano plot, MA plot 등의 DEG 와 관련된 plot 들과 전체적인 데이터의 분포를 Visualization 탭에서 확인할 수 있다. DEG list 는 Download full columns 를 통해 .tsv 파일로 저장할 수 있다. tsv 파일은 일반적인 텍스트 에디터(메모장 등)에서 읽을 수 있으며 엑셀로 가져오면 각 셀이 구분된 상태로 확인할 수 있다. DEG list 파일을 저장해뒀다가 Enrichment analysis 등의 추후 분석을 하면 된다.



GEO2R
Options
Profile graph
R script
Reanalyze
Visualization
GSE86525: test vs control
GSE86525: test vs control
GSE86525: UMAP(nbrs=3)
GSE86525: limma, Padj<0.05
test vs control
0
58717
GSE86525
GSE86525: Expression density
GSE86525: Adjusted P-value counts
GSE86525: Moderated t statistic
Download full table
Select columns

Log-transformation has been applied to the data. You can change this in the Options tab.

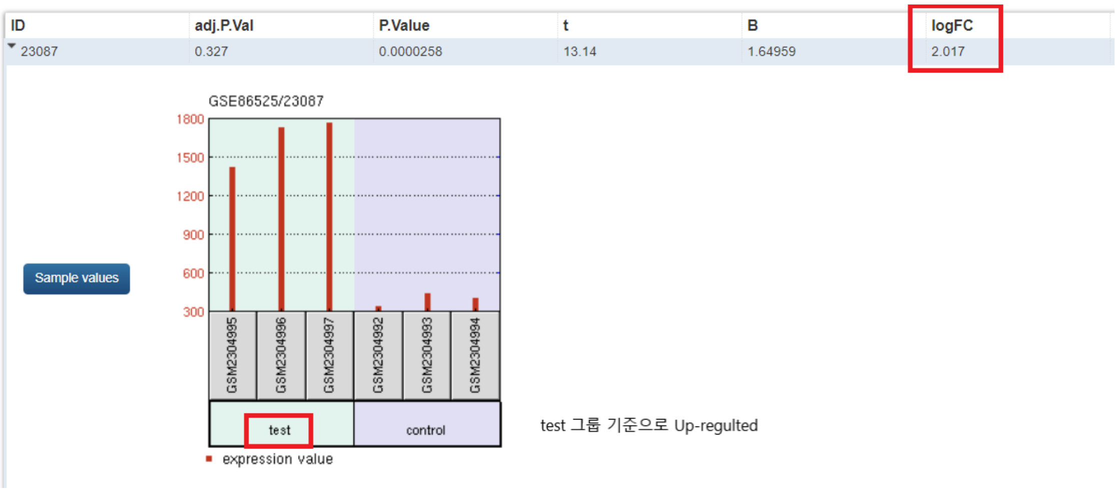
Reanalyze if you changed any options.

Top differentially expressed genes

ID	adj.P.Val	P.Value	t	B	logFC	SPOT_ID	GB_ACC	SEQUENCE
23087	0.327	0.0000258	13.14	1.64959	2.017	A_21_P0009341	NR_030732	AGGGGATATGGGAGGC...
46539	0.327	0.0000306	12.72	1.58595	1.354	A_23_P3221	NM_021199	TCCCTTTTCAGTACCTTT...
52269	0.327	0.0000347	-12.42	1.53891	-1.929	A_33_P3322814	NM_001007156	CGTTGTGTTTCCACAGA...
11548	0.327	0.0000426	11.95	1.45736	2.26	tc|THC2682885|tc|THC26...		TGCAGGGTAAGGACAGA...
24975	0.327	0.0000436	-11.89	1.44747	-1.157	A_23_P71148	NM_000712	AGGTGATGTAGCACTTC...
53046	0.327	0.0000686	-10.9	1.25157	-2.018	A_33_P3687389	CA309819	TATCTACTTCATGAGGTT...
13178	0.327	0.0000797	-10.6	1.18219	-1.651	A_23_P113572	NM_001770	TACATGCCAGTGACACT...
4171	0.327	0.0000833	-10.51	1.16146	-1.732	A_23_P108673	NM_032181	CCTCAGGGATTAAAATC...
24844	0.327	0.0000862	10.44	1.145	1.927	A_21_P0004150	BG460179	GGGGTGGGAGCAATGA...
43502	0.327	0.000104	10.07	1.05352	1.124	A_23_P10873	NM_003263	ATATTCCTCCTGTTGATA...
28862	0.327	0.0001143	9.88	1.00633	1.975	A_23_P25396	NM_005123	CTGAGATGCTGATGTCA...

GEO2R result

중요한 점은 logFC 값의 양수, 음수 구분이 어떤 그룹을 기준으로 Up-regulated 또는 Down-regulated 된 것인지 상위 DEG 의 레코드를 열어 필히 확인해 봐야한다는 점이다.

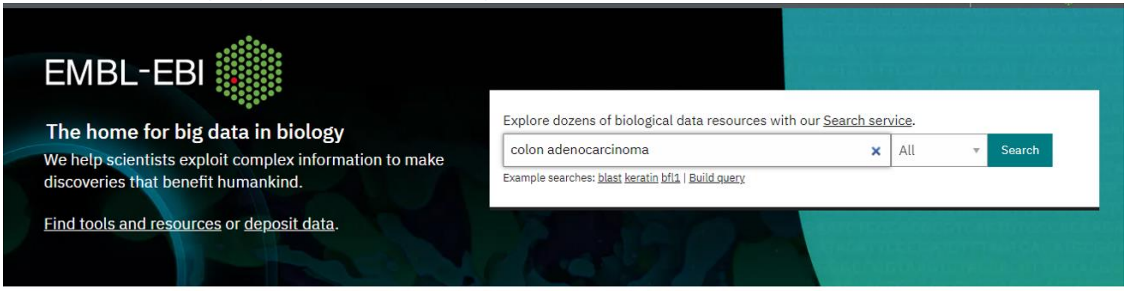


ID	adj.P.Val	P.Value	t	B	logFC
23087	0.327	0.0000258	13.14	1.64959	2.017

test 그룹 기준으로 Up-regulted

logFC 와 Up/Down-regulated group 매치

EMBL-EBI: Array Express & The European Nucleotide Archive (ENA)



EMBL-EBI

The home for big data in biology

We help scientists exploit complex information to make discoveries that benefit humankind.

Find tools and resources or deposit data.

Explore dozens of biological data resources with our Search service.

colon adenocarcinoma

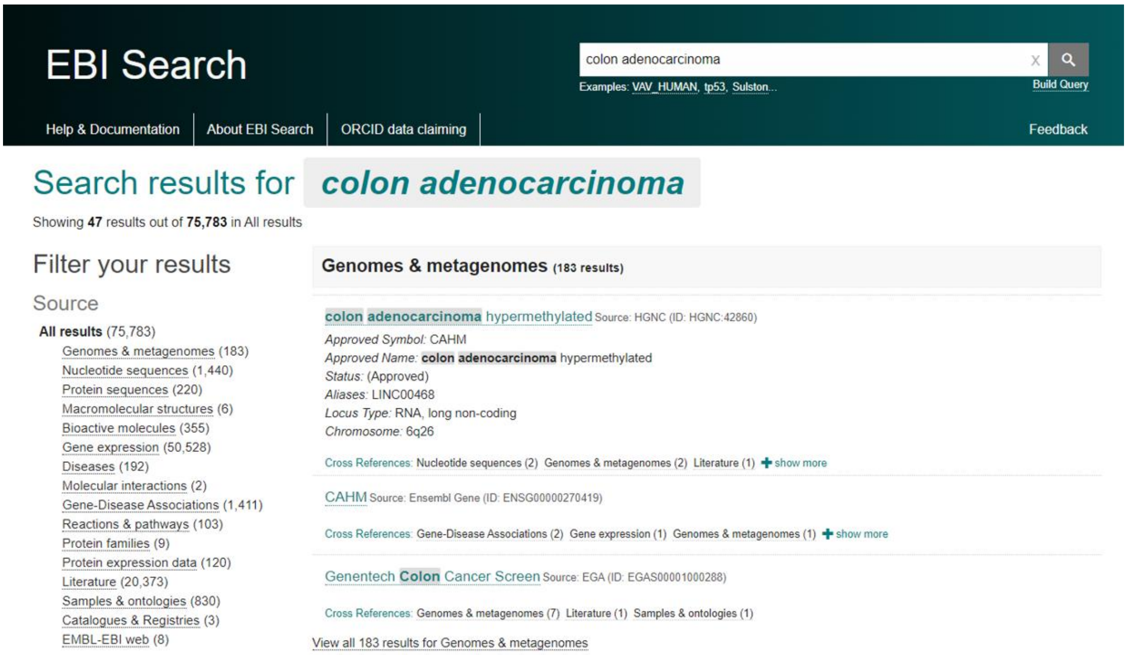
Example searches: blast keratin bfl1 | Build query

EBI main

The European Molecular Biology Laboratory's European Bioinformatic's Institute (EMBL-EBI, https://www.ebi.ac.uk/)는 유럽에서 primary data deposit location 으로 사용되며 NCBI 와 마찬가지로 mycroarray data 를 보유한 Array Express (https://www.ebi.ac.uk/arrayexpress/)와 raw RNA-Seq data 를 보유한 The European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/)를 운영하고 있다.

Array Express 는 매주 GEO 의 데이터를 replicate 해온다. Array Express 에서 통용되는 accession code 중 E-GEOD-로 시작하는 것들이 GEO 에서 GSE 데이터를 replicate 해온 것이다. 즉, E-GEOD-11111 은 GSE11111 과 동일한 자료이다.

EBI 는 data exploration tool 은 제공하지 않지만 원하는 키워드로 검색 시, 데이터 뿐만 아니라 논문 검색 결과도 제공한다.



colon adenocarcinoma

Search results for **colon adenocarcinoma**

Showing **47** results out of **75,783** in All results

Filter your results

Source

All results (75,783)

- Genomes & metagenomes (183)
- Nucleotide sequences (1,440)
- Protein sequences (220)
- Macromolecular structures (6)
- Bioactive molecules (355)
- Gene expression (50,528)
- Diseases (192)
- Molecular interactions (2)
- Gene-Disease Associations (1,411)
- Reactions & pathways (103)
- Protein families (9)
- Protein expression data (120)
- Literature (20,373)
- Samples & ontologies (830)
- Catalogues & Registries (3)
- EMBL-EBI web (8)

Genomes & metagenomes (183 results)

colon adenocarcinoma hypermethylated Source: HGNC (ID: HGNC:42860)
Approved Symbol: CAHM
Approved Name: **colon adenocarcinoma** hypermethylated
Status: (Approved)
Aliases: LINC00468
Locus Type: RNA, long non-coding
Chromosome: 6q26

Cross References: Nucleotide sequences (2) Genomes & metagenomes (2) Literature (1) ✚ show more

CAHM Source: Ensembl Gene (ID: ENSG00000270419)

Cross References: Gene-Disease Associations (2) Gene expression (1) Genomes & metagenomes (1) ✚ show more

Genentech **Colon** Cancer Screen Source: EGA (ID: EGAS00001000288)

Cross References: Genomes & metagenomes (7) Literature (1) Samples & ontologies (1)

View all 183 results for Genomes & metagenomes

EBI 검색결과

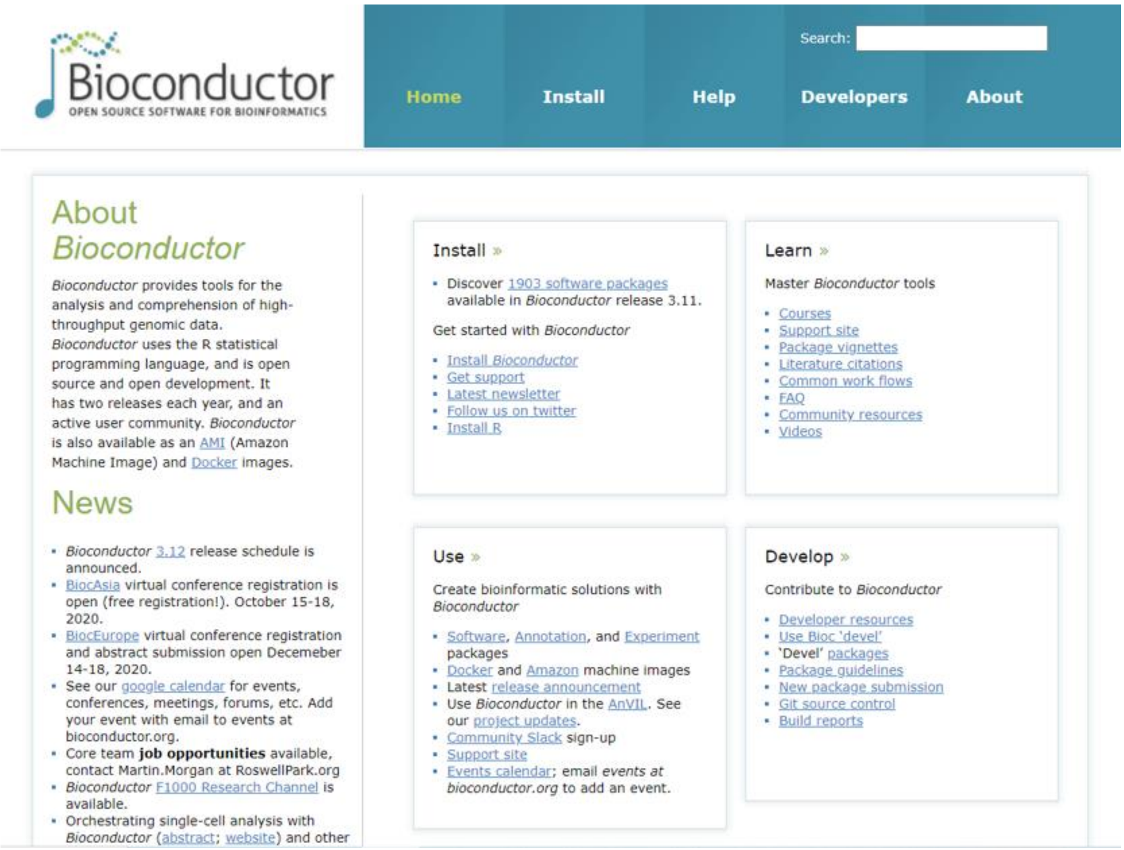
DDBJ: DDBJ Sequence Read Archive (DRA)

The DNA Data Bank of Japan 에서 운영하는 DDBJ Sequence Read Archive (DRA, https://www.ddbj.nig.ac.jp/dra/index-e.html)에서는 RNA-Seq data 를 열람할 수 있으나 포함된 고유 데이터는 다른 아카이브보다 적은 편이다.

[참고] https://www.ccdatalab.org/blog/2019/3/29/gene-expression-repos-explained

Bioconductor in R

Bioconductor (http://www.Bioconductor.org)는 의생명 분야의 유전체 데이터를 분석할 수 있는 오픈 소스 및 개방형 개발을 지향하는 R 환경에서의 소프트웨어 개발 프로젝트이다.



About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. *Bioconductor* uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as an AMI (Amazon Machine Image) and Docker images.

News

- *Bioconductor* 3.12 release schedule is announced.
- BiocAsia virtual conference registration is open (free registration!). October 15-18, 2020.
- BiocEurope virtual conference registration and abstract submission open Decemeber 14-18, 2020.
- See our google calendar for events, conferences, meetings, forums, etc. Add your event with email to events at bioconductor.org.
- Core team **job opportunities** available, contact Martin.Morgan at RoswellPark.org
- *Bioconductor* F1000 Research Channel is available.
- Orchestrating single-cell analysis with *Bioconductor* (abstract; website) and other

Install »

- Discover 1903 software packages available in *Bioconductor* release 3.11.

Get started with *Bioconductor*

- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment packages
- Docker and Amazon machine images
- Latest release announcement
- Use *Bioconductor* in the AnVIL. See our project updates.
- Community Slack sign-up
- Support site
- Events calendar; email *events at bioconductor.org* to add an event.

Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'
- 'Devel' packages
- Package guidelines
- New package submission
- Git source control
- Build reports

Bioconductor main

microarray 데이터를 비롯한 RNA-seq 등의 다양한 유전체 데이터를 분석할 수 있는 패키지들을 포함하고 있다.

- Learn > Courses 를 통해 필요한 내용을 습득할 수 있다.
- Learn > Common Work Flows 를 통해 특정 목표를 위한 패키지 이용 순서를 알아볼 수 있다.
- Use > Software – find packages 에서 필요한 패키지를 찾아 Documentation, details 등의 문서를 얻을 수 있다.

TCGA 데이터 또는 GEO 데이터를 다운받거나 DEA 를 하는 등의 작업을 R 에서 직접 커스터마이즈하여 수행하기 위해서는 Bioconductor 의 BiocManager 라이브러리 설치가 필수적이다.

R 에서 BiocManager 를 설치하는 코드이다. 1 장의 install_libraries.R 을 통해 설치했다면 건너뛰어도 된다.

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
```

Gene ID conversion in R with mygene

종종 Gene ID 를 다른 종류로 바꿔야할 일이 생긴다. Entrez_ID 등을 Gene symbol 로 바꾸는 등의 경우이다. R 에서 mygene 패키지를 이용하면 이를 해결할 수 있다. 우선 mygene 패키지를 설치한다.

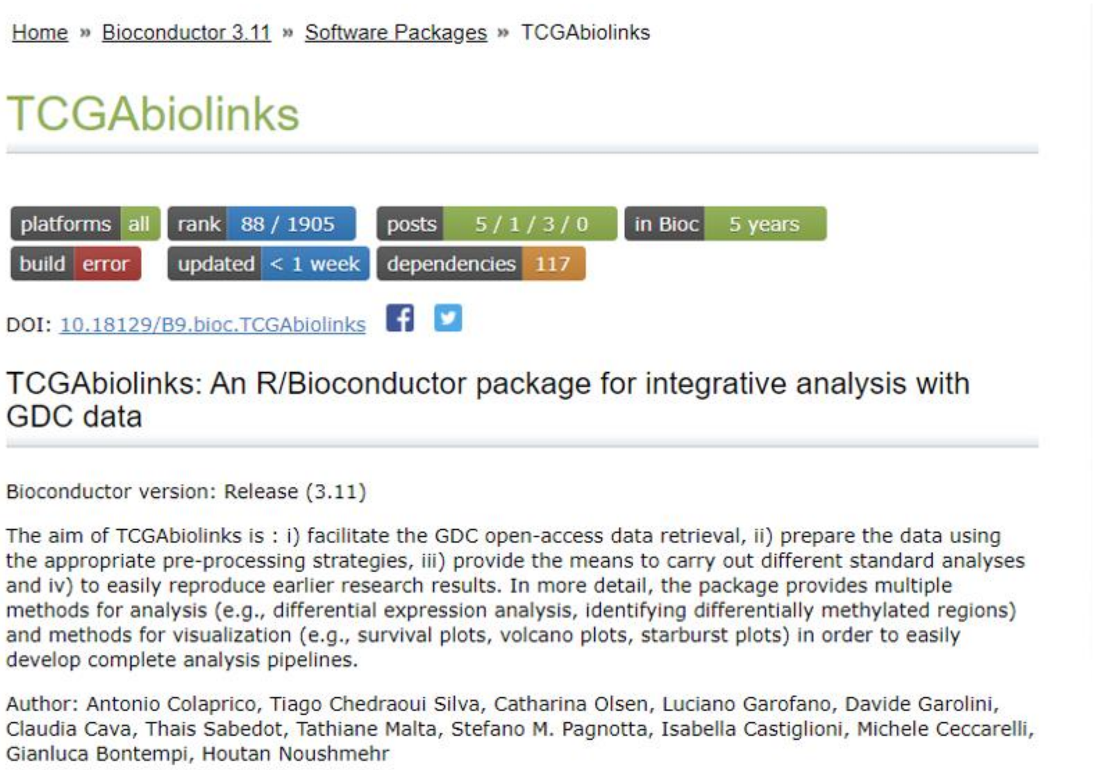
mygene 패키지의 queryMany 함수에 변환을 원하는 gene list 를 벡터 형태로 넣고, source ID 를 scope 파라미터에, sink ID (결과물)를 fields 파라미터에 원하는 종류만큼 넣고, species 를 명시하면 source ID 를 인식하여 sink ID 로 변환시켜 준다.

```
gene_list <- c("ENSG00000000003", "ENSG00000000005", "ENSG00000000419",
               "ENSG00000000457", "ENSG00000000460")

gene_list_conv <- queryMany(gene_list, scopes="ensembl.gene",
                     fields=c("symbol", "entrezgene"), species="human")

## Finished

gene_list_conv

## DataFrame with 5 rows and 5 columns
##             query         _id   X_score  entrezgene      symbol
##       <character> <character> <numeric> <character> <character>
## 1 ENSG00000000003        7105   23.0829        7105      TSPAN6
## 2 ENSG00000000005       64102   23.0788       64102        TNMD
## 3 ENSG00000000419        8813   22.2936        8813        DPM1
## 4 ENSG00000000457       57147   23.0829       57147       SCYL3
## 5 ENSG00000000460       55732   23.0855       55732    C1orf112

gene_list_conv$symbol

## [1] "TSPAN6"   "TNMD"     "DPM1"     "SCYL3"    "C1orf112"
```

TCGA data access in R with TCGAbiolinks

GDC portal 에서 TCGA data 를 다운받으려면 샘플별로 일일이 선택해야하는 어려움이 있었다. 원하는 조건을 만족시키는 모든 데이터를 일괄적으로 다운받고 싶다면 R 에서 Bioconductor 패키지 중 대표적으로 TCGAbiolinks 을 이용하는 것이 훨씬 수월하다.

Use > Software – Find packages 에서 TCGAbiolinks 를 검색하거나 https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html 에 들어가서 관련 정보와 documentation 을 얻을 수 있다.



Home » Bioconductor 3.11 » Software Packages » TCGAbiolinks

TCGAbiolinks

platforms all | rank 88 / 1905 | posts 5 / 1 / 3 / 0 | in Bioc 5 years
build error | updated < 1 week | dependencies 117

DOI: 10.18129/B9.bioc.TCGAbiolinks

TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

Bioconductor version: Release (3.11)

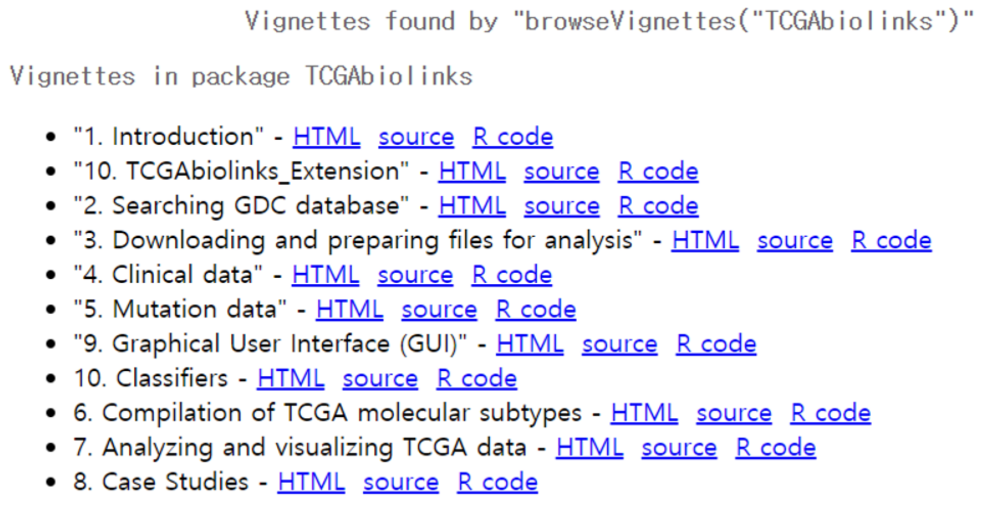
The aim of TCGAbiolinks is : i) facilitate the GDC open-access data retrieval, ii) prepare the data using the appropriate pre-processing strategies, iii) provide the means to carry out different standard analyses and iv) to easily reproduce earlier research results. In more detail, the package provides multiple methods for analysis (e.g., differential expression analysis, identifying differentially methylated regions) and methods for visualization (e.g., survival plots, volcano plots, starburst plots) in order to easily develop complete analysis pipelines.

Author: Antonio Colaprico, Tiago Chedraoui Silva, Catharina Olsen, Luciano Garofano, Davide Garolini, Claudia Cava, Thais Sabedot, Tathiane Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr

Bioconductor 홈페이지에서 제공하는 TCGAbiolinks 안내 페이지

R Console 에서 다음 코드 입력하여도 documentations 를 열람할 수 있다.

```
browseVignettes("TCGAbiolinks")
```



Vignettes found by "browseVignettes("TCGAbiolinks")"

Vignettes in package TCGAbiolinks

- "1. Introduction" - HTML source R code
- "10. TCGAbiolinks_Extension" - HTML source R code
- "2. Searching GDC database" - HTML source R code
- "3. Downloading and preparing files for analysis" - HTML source R code
- "4. Clinical data" - HTML source R code
- "5. Mutation data" - HTML source R code
- "9. Graphical User Interface (GUI)" - HTML source R code
- 10. Classifiers - HTML source R code
- 6. Compilation of TCGA molecular subtypes - HTML source R code
- 7. Analyzing and visualizing TCGA data - HTML source R code
- 8. Case Studies - HTML source R code

R 에서 TCGAbiolinks 의 documetation 열람

이제 본격적으로 TCGAbiolinks 와 관련 패키지를 다운받아 R 에서 TCGA 데이터를 다운받고 DEG 분석을 해보자.

먼저, 필요한 패키지들을 설치한다.

```
##### libraries installation #####
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("TCGAbiolinks")
# update all/some/none? [a/s/n]: type a and then enter
BiocManager::install("EDASeq")
BiocManager::install("edgeR")
BiocManager::install("SummarizedExperiment")

library(TCGAbiolinks)
# If TCGAbiolinks is not loaded, type below:
# install.packages("openssl")
```

gene expression data 다운로드

legacy archive 에서 TCGA-COAD 프로젝트의 RNA-Seq gene expression 데이터를 전부 다운받는 코드이다. 앞서 살펴본 GDC portal 에서의 분류기준과 barcode 가 GDEquery 의 아규먼트로 들어간 것을 확인할 수 있다.

TCGAbiolinks 의 GDCquery, GDCdownload 함수가 사용되었다. GDCquery 에 대한 보다 자세한 옵션은 https://rdrr.io/bioc/TCGAbiolinks/man/GDCquery.html 를 참고하자.

```
##### data download #####
library(TCGAbiolinks)
query <- GDCquery(project = "TCGA-COAD",  # colon-adenocarcinoma
                  data.category = "Gene expression",
                  data.type = "Gene expression quantification",
                  experimental.strategy = "RNA-Seq",
                  platform = "Illumina HiSeq",
                  file.type = "results",
                  legacy = TRUE)  # data from legacy archive

# Download a list of barcodes with platform IlluminaHiSeq_RNASeqV2
GDCdownload(query, directory = "../GDCdata")
```

Preparing expression matrix

다운받은 데이터를 로드하고 raw_count matrix 로 만드는 과정이다. SummarizedExperiment 패키지가 필요하므로 없다면 설치하자.

TCGAbiolinks 의 GDCprepare 함수와 SummarizedExperiment 의 assay 가 사용되었다.

```
COADMatrix <- SummarizedExperiment::assay(COADRnaseqSE,"raw_count")
```

Normalizing and quantile filtering expression matrix

Raw count matrix 를 normalize 하고 gene expression 이 상대적으로 적은 (하위 25%) gene 은 걸러내는 과정이다.

TCGAbiolinks 의 `TCGAanalyze_Normalization`, `TCGAanalyze_Filtering` 함수가 사용되었다.

```
library(EDASeq)

dataNorm <- TCGAanalyze_Normalization(tabDF = COADRnaseqSE, geneInfo =  gen
eInfo)

## I Need about  81 seconds for this Complete Normalization Upper Quantile
 [Processing 80k elements /s]

## Step 1 of 4: newSeqExpressionSet ...

## Step 2 of 4: withinLaneNormalization ...

## Step 3 of 4: betweenLaneNormalization ...

## Step 4 of 4: exprs ...

# quantile filter of genes
dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
                                  method = "quantile",
                                  qnt.cut =  0.25)
```

Finding differentially expressed genes

샘플 바코드가 NT (normal)인 것과 TP (tumor)인 것으로 그룹을 나누어 정상 그룹에 비해 COAD 환자 그룹에서 up-regulated 된 DEG 를 추출하는 과정이다 (logFC.cut = 1). 유전자들 간의 다중비교에서의 통계적 유의성을 보장하기 위해 adjusted P value < 0.01 인 결과만 가져왔다 (fdr.cut = 0.01). 다중비교의 방법으로 glmLRT 를 사용하였다 (method = “glmLRT”).

TCGAbiolinks 의 `TCGAanalyze_DEA`, `TCGAanalyze_LevelTab` 함수가 사용되었다.

```
# selection of normal samples "NT"
samplesNT <- TCGAquery_SampleTypes(barcode = colnames(dataFilt),
                                   typesample = c("NT"))

# selection of tumor samples "TP"
samplesTP <- TCGAquery_SampleTypes(barcode = colnames(dataFilt),
                                   typesample = c("TP"))

# Diff.expr.analysis (DEA)
```

```
dataDEGs <- TCGAanalyze_DEA(mat1 = dataFilt[,samplesNT],
                          mat2 = dataFilt[,samplesTP],
                          Cond1type = "Normal",
                          Cond2type = "Tumor",
                          fdr.cut = 0.01 ,
                          logFC.cut = 1,
                          method = "glmLRT")
```

```
## Batch correction skipped since no factors provided

## ----------------------- DEA -------------------------------

## there are Cond1 type Normal in  41 samples

## there are Cond2 type Tumor in  285 samples

## there are  14893 features as miRNA or genes

## I Need about  162 seconds for this DEA. [Processing 30k elements /s]

## ----------------------- END DEA -------------------------------
```

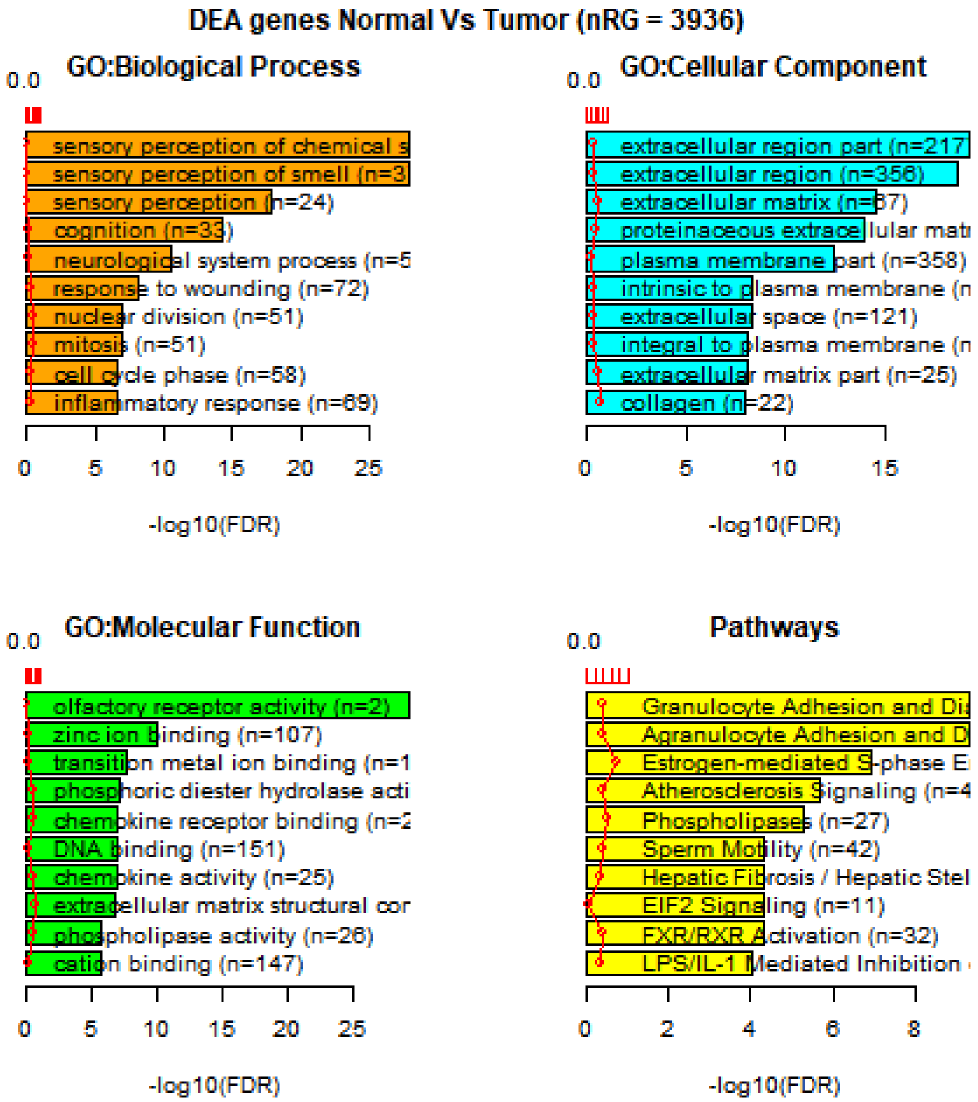
```
# DEGs table with expression values in normal and tumor samples
dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGs,"Tumor","Normal",
                                      dataFilt[,samplesTP],dataFilt[,sampl
esNT])

head(dataDEGsFiltLevel[order(dataDEGsFiltLevel$FDR),], 10) # DEGs ordered b
y FDR
```

```
##              mRNA     logFC           FDR       Tumor      Normal       Delt
a
## BEST4       BEST4 -6.399300 2.689266e-265   39.242105  3605.48780   251.1
2202
## UGP2         UGP2 -1.926551 8.591576e-202 4495.694737 18450.41463  8661.1
8335
## SULT1A2   SULT1A2 -4.098707 6.659248e-176   61.971930  1166.58537   254.
00480
## SLC25A34 SLC25A34 -4.145745 2.876465e-173   45.919298   866.65854   190.
36968
## CLEC3B     CLEC3B -3.892536 4.990199e-171  214.231579  3291.73171   833.
90403
## FAM151A   FAM151A -5.216771 2.083841e-166    8.098246   365.14634    42.2
4669
## CUBN         CUBN -5.267365 6.107065e-165   24.670175  1306.39024   129.9
4683
## CDH3         CDH3  6.228599 1.027715e-149 4692.719298    63.97561 29229.0
6556
## ABCG2       ABCG2 -5.012779 3.582667e-143  107.726316  3855.75610   540.0
0825
## PHLPP2     PHLPP2 -2.478031 2.555206e-139  869.554386  5232.85366  2154.
78295
```

Enrichment analysis

TCGAbiolinks 의 `TCGAvisualize_EAbarplot` 함수는 DEG 테이블을 가지고 GO analysis 및 Pathway analysis 를 수행하여 유의한 functional profile 을 –log(FDR)값 순위에 따른 barplot 으로 그려 확인시켜준다.

```
Genelist <- rownames(dataDEGsFiltLevel)
ansEA <- TCGAanalyze_EAcomplete(TFname="DEA genes Normal Vs Tumor",Genelis
t)

## [1] "I need about  1 minute to finish complete  Enrichment analysis GO[B
P,MF,CC] and Pathways... "
## [1] "GO Enrichment Analysis BP completed....done"
## [1] "GO Enrichment Analysis MF completed....done"
## [1] "GO Enrichment Analysis CC completed....done"
## [1] "Pathway Enrichment Analysis completed....done"

library(png)
png("./pictures/ch2_18.png")
TCGAvisualize_EAbarplot(tf = rownames(ansEA$ResBP),
                        GOBPTab = ansEA$ResBP,
                        GOCCTab = ansEA$ResCC,
                        GOMFTab = ansEA$ResMF,
                        PathTab = ansEA$ResPat,
                        nRGTab = Genelist,
                        nBar = 10,
                        filename = NULL)  # default: save as pdf
dev.off()

## png
##   2
```



Enrichment analysis result

Survival plot

TCGAbiolinks 의 `TCGAAanalyze_survival` 함수는 cancer project 의 clinical data 만을 다운받아 group 에 따른 survival plot 을 그려준다. 전체 옵션은 다음과 같다.

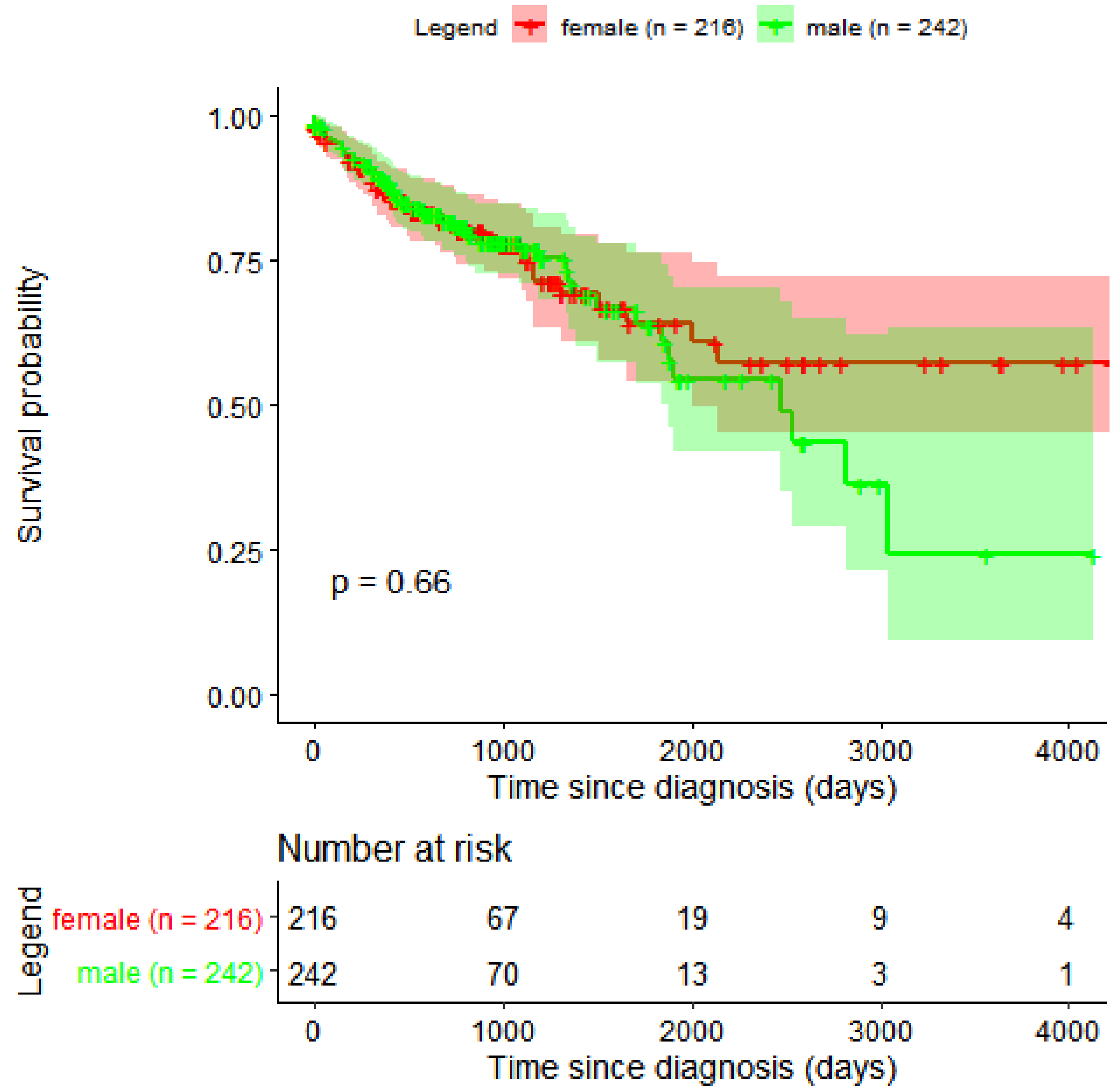
- clinical_patient: TCGA Clinical patient with the information days_to_death
- clusterCol: Column with groups to plot. This is a mandatory field, the caption will be based in this column
- legend: Legend title of the figure
- xlim: xlim x axis limits e.g. xlim = c(0, 1000). Present narrower X axis, but not affect survival estimates.
- main: main title of the plot
- ylab: y-axis text of the plot
- xlab: x-axis text of the plot

- filename: The name of the pdf file
- color: Define the colors of the lines.
- pvalue: Show pvalue in the plot.
- risk.table: Show or not the risk table
- conf.int: Show confidence intervals for point estimates of survival curves.

실행 예제를 보자.

```r
png("./pictures/ch2_19.png")

clin.coad <- GDCquery_clinic("TCGA-COAD","clinical")
TCGAanalyze_survival(clin.coad, "gender",  # mandatory arguments
                   main = "TCGA Set\n COAD", height = 10, width=10,
                   p.value = T, risk.table = T,
                   filename = NULL)  # Default: save as pdf
dev.off()
```

```
## png
##   2
```



Survival plot result

Extracting clinical data

다음은 TCGA 데이터의 clinical information 을 추출하는 과정이다. Gene expression 데이터를 다운받을 때와 유사하게 TCGAbiolinks 의 GDCquery 와 GDCdownload 를 사용하되 쿼리 안에 data.category 와 data.type, data.format 이 다른 것을 확인할 수 있다.

```
library(TCGAbiolinks)
query <- GDCquery(project = "TCGA-COAD",
                  data.category = "Clinical",
                  data.type = "Clinical Supplement",
                  data.format = "BCR Biotab")

GDCdownload(query, directory = "../GDCdata")
```

```
names(clinical.BCRtab.all)

## [1] "clinical_nte_coad"              "clinical_follow_up_v1.0_coad"
## [3] "clinical_omf_v4.0_coad"          "clinical_patient_coad"
## [5] "clinical_follow_up_v1.0_nte_coad" "clinical_drug_coad"
## [7] "clinical_radiation_coad"

nte <- as.data.frame(clinical.BCRtab.all[1])
colnames(nte) <- nte[1,]
nte <- nte[-c(1:2),]
follow_up <- as.data.frame(clinical.BCRtab.all[2])
colnames(follow_up) <- follow_up[1,]
follow_up <- follow_up[-c(1:2),]
omf <- as.data.frame(clinical.BCRtab.all[3])
colnames(omf) <- omf[1,]
omf <- omf[-c(1:2),]
patient <- as.data.frame(clinical.BCRtab.all[4])
colnames(patient) <- patient[1,]
patient <- patient[-c(1:2),]
follow_up_nte <- as.data.frame(clinical.BCRtab.all[5])
colnames(follow_up_nte) <- follow_up_nte[1,]
follow_up_nte <- follow_up_nte[-c(1:2),]
drug <- as.data.frame(clinical.BCRtab.all[6])
colnames(drug) <- drug[1,]
drug <- drug[-c(1:2),]
radiation <- as.data.frame(clinical.BCRtab.all[7])
colnames(radiation) <- radiation[1,]
radiation <- radiation[-c(1:2),]


head(nte, 3)

##                     bcr_patient_uuid bcr_patient_barcode
## 3 CE00896A-F7D2-4123-BB95-24CB6E53FC32       TCGA-5M-AAT6
## 4 A54B322B-80D3-435C-8D14-25841B741F6C       TCGA-5M-AATE
## 5 DD7A53EE-CB08-40C4-9935-51D1CA17E9E8       TCGA-A6-A565
##   days_to_new_tumor_event_after_initial_treatment
## 3                                         219
## 4                                         810
## 5                                         301
##   site_of_additional_surgery_new_tumor_event_mets additional_radiation_
therapy
## 3                              [Not Available]                         N
O
## 4                              [Not Available]                         N
O
## 5                              [Not Available]                         N
O
##   additional_pharmaceutical_therapy
## 3                            YES
## 4                            YES
## 5                            YES
##   days_to_new_tumor_event_additional_surgery_procedure
## 3                                   [Not Available]
## 4                                   [Not Available]
## 5                                   [Not Available]
##   new_neoplasm_event_occurrence_anatomic_site new_neoplasm_event_type
## 3                           [Not Available]             [Unknown]
## 4                           [Not Available]             [Unknown]
## 5                           [Not Available]             [Unknown]
##   new_neoplasm_occurrence_anatomic_site_text
## 3                          [Not Available]
## 4                          [Not Available]
## 5                          [Not Available]
##   new_tumor_event_additional_surgery_procedure progression_determined_b
y
## 3                                       NO          [Not Available]
## 4                                       NO          [Not Available]
## 5                                       NO          [Not Available]
##   residual_disease_post_new_tumor_event_margin_status
## 3                                  [Not Available]
## 4                                  [Not Available]
## 5                                  [Not Available]
```

```
head(follow_up, 3)
```

```
##                       bcr_patient_uuid bcr_patient_barcode bcr_followup_b
arcode
## 3 A94E1279-A975-480A-93E9-7B1FF05CBCBF       TCGA-3L-AA1B  TCGA-3L-AA1B
-F67516
## 4 A94E1279-A975-480A-93E9-7B1FF05CBCBF       TCGA-3L-AA1B  TCGA-3L-AA1B
-F70121
## 5 92554413-9EBC-4354-8E1B-9682F3A031D9       TCGA-4N-A93T  TCGA-4N-A93T
-F67783
##                     bcr_followup_uuid form_completion_date
## 3 8AEC4573-0E07-4BFC-A1A7-FF843D3D447A           2014-11-6
## 4 9F19C5A6-1A27-409D-B136-B731A79B9F2C           2015-2-26
## 5 B614BDCE-3651-4441-BA59-DDDD53D595FF          2014-11-11
##   followup_case_report_form_submission_reason lost_follow_up radiation_t
herapy
## 3             Scheduled Follow-up Submission           NO
 NO
## 4             Scheduled Follow-up Submission           NO
 NO
## 5             Scheduled Follow-up Submission           NO
 NO
##   postoperative_rx_tx primary_therapy_outcome_success vital_status
## 3                NO     Complete Remission/Response       Alive
## 4                NO     Complete Remission/Response       Alive
## 5               YES                Stable Disease       Alive
##   days_to_last_followup    days_to_death person_neoplasm_cancer_status
## 3                   349 [Not Applicable]                   [Unknown]
## 4                   475 [Not Applicable]                   [Unknown]
## 5                   146 [Not Applicable]                  WITH TUMOR
##   new_tumor_event_after_initial_treatment  followup_treatment_success
## 3                                   NO                   [Unknown]
## 4                                   NO Complete Remission/Response
## 5                                   NO             Stable Disease

head(omf, 3)

##                      bcr_patient_uuid bcr_patient_barcode     bcr_omf_bar
code
## 3 d976782c-90c1-421d-b83c-7fc2617e2709        TCGA-A6-2677 TCGA-A6-2677-
012677
## 4 7d8eab0a-e6c8-4449-9ebf-50c41db94a06        TCGA-A6-2681 TCGA-A6-2681-
044803
## 5 7d8eab0a-e6c8-4449-9ebf-50c41db94a06        TCGA-A6-2681 TCGA-A6-2681-
044804
##                         bcr_omf_uuid form_completion_date
## 3 84EE676C-60D4-4B41-A440-657537A18D6A            2011-6-6
## 4 30378729-C13A-4D17-BA8B-3E73BDAD7B9E           2013-6-27
## 5 2099E22F-A107-47BC-A4A5-7AA4E4283032           2013-6-27
##         malignancy_type days_to_other_malignancy_dx surgery_indicator
## 3 Synchronous Malignancy            [Not Available]   [Not Available]
## 4      Prior Malignancy                       -365             YES
## 5      Prior Malignancy           [Not Available]              YES
##                           surgery_type days_to_surgical_resection
## 3                     [Not Available]                       59
## 4 Excision skin lesion Right upper back                     -365
## 5                          Mastectomy           [Not Available]
##   drug_tx_indicator  drug_tx_extent      drug_name days_to_drug_therapy
_start
## 3               NO [Not Available] [Not Available]           [Not Availa
ble]
## 4               NO [Not Available] [Not Available]           [Not Availa
ble]
## 5               NO [Not Available] [Not Available]           [Not Availa
ble]
##   radiation_tx_indicator radiation_tx_extent rad_tx_to_site_of_primary_t
umor
## 3                  YES        Locoregional                           NO
## 4                   NO     [Not Available]                [Not Availabl
e]
## 5                   NO     [Not Available]                [Not Availabl
e]
##   days_to_radiation_therapy_start  system_version    pathologic_T
## 3                           514           6th            T3a
## 4                [Not Available] [Not Available] [Not Available]
## 5                [Not Available] [Not Available] [Not Available]
##      pathologic_N    pathologic_M pathologic_stage   clinical_stage
## 3             N1             MX       Stage III [Not Applicable]
## 4 [Not Available] [Not Available]  [Not Available]  [Not Available]
## 5 [Not Available] [Not Available]  [Not Available]  [Not Available]
##   other_malignancy_anatomic_site other_malignancy_anatomic_site_text
## 3                       Kidney                  [Not Applicable]
## 4                         Back                  [Not Applicable]
## 5                        Breast                    [Not Applicable]
##   other_malignancy_histological_type other_malignancy_histological_type
_text
## 3  Kidney Clear Cell Renal Carcinoma                        [Not Applicabl
e]
## 4                     Other, specify         Squamous cell Carcinoma in si
tu
## 5                    [Not Available]                        [Not Applicabl
e]
##   other_malignancy_laterality     stage_other
## 3                        Left [Not Available]
## 4            [Not Applicable] [Not Available]
## 5                        Left [Not Available]
```

```r
head(patient, 3)
```

```
##                     bcr_patient_uuid bcr_patient_barcode form_completio
n_date
## 3 A94E1279-A975-480A-93E9-7B1FF05CBCBF        TCGA-3L-AA1B            201
4-4-22
## 4 92554413-9EBC-4354-8E1B-9682F3A031D9        TCGA-4N-A93T            201
4-10-1
## 5 A5E14ADD-1552-4606-9FFE-3A03BCF76640        TCGA-4T-AA8H             20
14-6-5
##               histological_type tissue_prospective_collection_indicator
## 3          Colon Adenocarcinoma                                     YES
## 4          Colon Adenocarcinoma                                     YES
## 5 Colon Mucinous Adenocarcinoma                                      NO
##   tissue_retrospective_collection_indicator gender days_to_birth
## 3                                        NO FEMALE        -22379
## 4                                        NO   MALE        -24523
## 5                                       YES FEMALE        -15494
##                        race              ethnicity other_dx
## 3 BLACK OR AFRICAN AMERICAN NOT HISPANIC OR LATINO       No
## 4 BLACK OR AFRICAN AMERICAN NOT HISPANIC OR LATINO       No
## 5 BLACK OR AFRICAN AMERICAN NOT HISPANIC OR LATINO       No
##   history_of_neoadjuvant_treatment year_of_initial_pathologic_diagnosis
## 3                               No                                 2013
## 4                               No                                 2013
## 5                               No                                 2013
##   system_version pathologic_T pathologic_N pathologic_M pathologic_stage
## 3            7th           T2           N0           M0          Stage I
## 4            7th          T4a          N1b           M0       Stage IIIB
## 5            7th           T3           N0           MX        Stage IIA
##   residual_tumor primary_lymph_node_presentation_assessment
## 3             R0                                        YES
## 4             R0                                        YES
## 5             R0                                        YES
##   lymph_node_examined_count number_of_lymphnodes_positive_by_he
## 3                        28                                   0
## 4                        25                     [Not Available]
## 5                        24                                   0
##   number_of_lymphnodes_positive_by_ihc vital_status days_to_last_followu
p
## 3                                 0        Alive                  154
## 4                                 2        Alive                   8
## 5                    [Not Available]        Alive                  160
##      days_to_death person_neoplasm_cancer_status
## 3 [Not Applicable]                   TUMOR FREE
## 4 [Not Applicable]                   WITH TUMOR
## 5 [Not Applicable]                   TUMOR FREE
##   preoperative_pretreatment_cea_level non_nodal_tumor_deposits
## 3                   [Not Available]                      NO
## 4                               2.0                    YES
## 5                   [Not Available]                      NO
##   circumferential_resection_margin venous_invasion lymphatic_invasion
## 3                 [Not Available]             NO                NO
## 4                             30             NO                NO
## 5                             20             NO                NO
##   perineural_invasion_present microsatellite_instability number_of_loci_
tested
## 3                        NO                       NO       [Not Availabl
e]
## 4                        NO                [Unknown]       [Not Availabl
e]
## 5                        NO                       NO       [Not Availabl
e]
##   number_of_abnormal_loci loss_expression_of_mismatch_repair_proteins_b
y_ihc
## 3        [Not Available]                                             YES
## 4        [Not Available]                                             YES
## 5        [Not Available]                                             YES
##     loss_expression_of_mismatch_repair_proteins_by_ihc_result
## 3 MLH1-Expressed|MSH2-Expressed|PMS2-Expressed|MSH6-Expressed
## 4 MLH1-Expressed|MSH2-Expressed|PMS2-Expressed|MSH6-Expressed
## 5 MLH1-Expressed|MSH2-Expressed|PMS2-Expressed|MSH6-Expressed
##   kras_gene_analysis_performed kras_mutation_found kras_mutation_codon
## 3                         NO     [Not Available]     [Not Available]
## 4                         NO     [Not Available]     [Not Available]
## 5                         NO     [Not Available]     [Not Available]
##   braf_gene_analysis_performed braf_gene_analysis_result
## 3                         NO          [Not Available]
## 4                         NO          [Not Available]
## 5                         NO          [Not Available]
##   synchronous_colon_cancer_present history_of_colon_polyps colon_polyps_
present
## 3                             NO                   YES                 YE
S
## 4                            YES                    NO                 YE
S
## 5                             NO                    NO                  N
O
##    weight height number_of_first_degree_relatives_with_cancer_diagnosis
## 3    63.3    173                                                  0
## 4     134 167.64                                                  0
## 5 107.956  167.6                                                  0
##   radiation_therapy postoperative_rx_tx primary_therapy_outcome_success
## 3               NO                 NO    Complete Remission/Response
## 4               NO                YES                Stable Disease
## 5               NO                 NO    Complete Remission/Response
##   new_tumor_event_after_initial_treatment age_at_initial_pathologic_dia
gnosis
## 3                                   NO                                61
## 4                                   NO                                67
## 5                                   NO                                42
##   anatomic_neoplasm_subdivision cancer_diagnosis_cancer_type_icd9_text_
name
## 3                       Cecum                           [Not Available]
## 4              Ascending Colon                           [Not Available]
## 5             Descending Colon                           [Not Available]
##         clinical_M       clinical_N       clinical_T   clinical_stage
## 3 [Not Applicable] [Not Applicable] [Not Applicable] [Not Applicable]
## 4 [Not Applicable] [Not Applicable] [Not Applicable] [Not Applicable]
## 5 [Not Applicable] [Not Applicable] [Not Applicable] [Not Applicable]
##   days_to_form_completion days_to_initial_pathologic_diagnosis
## 3         [Not Available]                                 0
## 4         [Not Available]                                 0
## 5         [Not Available]                                 0
##   days_to_patient_progression_free days_to_tumor_progression death_cause
_text
## 3                  [Not Available]          [Not Available]  [Not Availab
le]
## 4                  [Not Available]          [Not Available]  [Not Availab
le]
## 5                  [Not Available]          [Not Available]  [Not Availab
le]
##      disease_code eastern_cancer_oncology_group extranodal_involvement
## 3 [Not Available]              [Not Available]       [Not Applicable]
## 4 [Not Available]              [Not Available]       [Not Applicable]
## 5 [Not Available]              [Not Available]       [Not Applicable]
##   family_member_relationship_type icd_10 icd_o_3_histology icd_o_3_site
## 3                [Not Available]  C18.0           8140/3       C18.0
## 4                [Not Available]  C18.2           8140/3       C18.2
## 5                [Not Available]  C18.6           8480/3       C18.6
##   informed_consent_verified init_pathology_dx_method_other
## 3                       YES              [Not Available]
## 4                       YES              [Not Available]
## 5                       YES              [Not Available]
##   initial_pathologic_diagnosis_method karnofsky_performance_score
## 3                     [Not Available]            [Not Available]
## 4                     [Not Available]            [Not Available]
## 5                     [Not Available]            [Not Available]
##    lost_follow_up measure_of_response   number_cycles number_pack_years_
smoked
```

```
## 3 [Not Available]     [Not Available] [Not Available]          [Not Avail
able]
## 4 [Not Available]     [Not Available] [Not Available]          [Not Avail
able]
## 5 [Not Available]     [Not Available] [Not Available]          [Not Avail
able]
##   patient_death_reason patient_id   pharm_regimen pharm_regimen_other
## 3      [Not Available]       AA1B [Not Available]     [Not Available]
## 4      [Not Available]       A93T [Not Available]     [Not Available]
## 5      [Not Available]       AA8H [Not Available]     [Not Available]
##      project_code regimen_indication relative_family_cancer_history
## 3 [Not Available]    [Not Available]               [Not Available]
## 4 [Not Available]    [Not Available]               [Not Available]
## 5 [Not Available]    [Not Available]               [Not Available]
##       stage_other stem_cell_transplantation stopped_smoking_year
## 3 [Not Available]           [Not Available]      [Not Available]
## 4 [Not Available]           [Not Available]      [Not Available]
## 5 [Not Available]           [Not Available]      [Not Available]
##   tissue_source_site tobacco_smoking_history tumor_tissue_site
## 3                3L         [Not Available]             Colon
## 4                4N         [Not Available]             Colon
## 5                4T         [Not Available]             Colon
##   year_of_tobacco_smoking_onset
## 3              [Not Available]
## 4              [Not Available]
## 5              [Not Available]
```

```
head(follow_up_nte, 3)
```

```
##                      bcr_patient_uuid bcr_patient_barcode bcr_followup_b
arcode
## 3 CE00896A-F7D2-4123-BB95-24CB6E53FC32        TCGA-5M-AAT6  TCGA-5M-AAT6
-F70107
## 4 A54B322B-80D3-435C-8D14-25841B741F6C        TCGA-5M-AATE  TCGA-5M-AATE
-F70110
## 5 565e2726-4942-4726-89d3-c5e3797f7204        TCGA-A6-2671   TCGA-A6-267
1-F6018
##   days_to_new_tumor_event_after_initial_treatment
## 3                                          219
## 4                                          810
## 5                                          535
##   site_of_additional_surgery_new_tumor_event_mets additional_radiation_
therapy
## 3                               [Not Available]                         N
O
## 4                               [Not Available]                         N
O
## 5                               [Not Available]                         N
O
##   additional_pharmaceutical_therapy
## 3                             YES
## 4                             YES
## 5                            YES
##   days_to_new_tumor_event_additional_surgery_procedure
## 3                                   [Not Available]
## 4                                   [Not Available]
## 5                                   [Not Available]
##   new_neoplasm_event_occurrence_anatomic_site new_neoplasm_event_type
## 3                          [Not Available]              [Unknown]
## 4                          [Not Available]              [Unknown]
## 5                          [Not Available]              [Unknown]
##   new_neoplasm_occurrence_anatomic_site_text
## 3                         [Not Available]
## 4                         [Not Available]
## 5                         [Not Available]
##   new_tumor_event_additional_surgery_procedure progression_determined_b
y
## 3                                       NO          [Not Available]
## 4                                       NO          [Not Available]
## 5                                       NO          [Not Available]
##   residual_disease_post_new_tumor_event_margin_status
## 3                                  [Not Available]
## 4                                  [Not Available]
## 5                                  [Not Available]
```

```
head(drug, 3)
```

```
##                      bcr_patient_uuid bcr_patient_barcode    bcr_drug_bar
code
## 3 92554413-9EBC-4354-8E1B-9682F3A031D9       TCGA-4N-A93T TCGA-4N-A93T-
D65957
## 4 565e2726-4942-4726-89d3-c5e3797f7204       TCGA-A6-2671  TCGA-A6-2671
-D6020
## 5 565e2726-4942-4726-89d3-c5e3797f7204       TCGA-A6-2671  TCGA-A6-2671
-D6055
##                         bcr_drug_uuid form_completion_date         drug_
name
## 3 25E5CD87-BB8D-4D2A-8219-1660D2F13AEE           2014-10-1            Xe
loda
## 4 3aefce87-3734-4b5f-a4e2-c62dbba5badf           2011-1-11 Study drug AM
G 655
## 5 409b319f-5642-4f38-8e5d-d948107d0425           2011-1-12
5 FU
##   clinical_trail_drug_classification           therapy_type
## 3                   [Not Available]           Chemotherapy
## 4                   [Not Available] Other, specify in notes
## 5                   [Not Available]           Chemotherapy
##   days_to_drug_therapy_start therapy_ongoing days_to_drug_therapy_end
## 3                         68            YES         [Not Available]
## 4                         96             NO                     270
## 5                         96             NO                     270
##   measure_of_response days_to_stem_cell_transplantation   pharm_regimen
## 3    [Not Applicable]                  [Not Available] [Not Available]
## 4     [Not Available]                  [Not Available] [Not Available]
## 5     [Not Available]                 [Not Available] [Not Available]
##   pharm_regimen_other   number_cycles   therapy_type_notes
## 3     [Not Available] [Not Available]      [Not Available]
## 4     [Not Available]            12 Protocol AMG 20060464
## 5     [Not Available]            12       [Not Available]
##   prescribed_dose_units total_dose_units prescribed_dose  regimen_number
## 3       [Not Available]  [Not Available] [Not Available] [Not Available]
## 4                mg/kg           mg/kg [Not Available]              1
## 5                  mg              mg        450-735              1
##   route_of_administration stem_cell_transplantation
## 3         [Not Available]          [Not Available]
## 4                    IV           [Not Available]
## 5                    IV           [Not Available]
##   stem_cell_transplantation_type regimen_indication regimen_indication_n
otes
## 3               [Not Available]    [Not Available]        [Not Applicabl
e]
## 4               [Not Available]        PALLIATIVE         [Not Applicabl
e]
## 5               [Not Available]        PALLIATIVE         [Not Applicabl
e]
##       total_dose tx_on_clinical_trial
## 3 [Not Available]                 NO
## 4 [Not Available]     [Not Available]
## 5           7185     [Not Available]
```

```r
head(radiation, 3)
```

```
##                      bcr_patient_uuid bcr_patient_barcode
## 3 e6ec5a68-7555-4f26-bd7e-9cdb4c5f7004       TCGA-AA-3549
## 4 bce3ce45-4fb3-4d8e-9ec7-d24427c2ba4d       TCGA-AA-3692
## 5 bce3ce45-4fb3-4d8e-9ec7-d24427c2ba4d       TCGA-AA-3692
##   bcr_radiation_barcode                 bcr_radiation_uuid
## 3   TCGA-AA-3549-R38338 B72A855F-225F-4537-A74F-8485ABDBA0D0
## 4   TCGA-AA-3692-R38345 2054309D-1EBC-4311-BF8A-621F6447F385
## 5   TCGA-AA-3692-R38346 1081C34F-DA75-4856-966C-8F9B10E784AA
##   form_completion_date radiation_type anatomic_treatment_site radiation_
dosage
## 3          2012-12-13       External      Distant Recurrence
  9
## 4          2012-12-13       External      Distant Recurrence
 39
## 5          2012-12-13       External      Distant Recurrence
 38
##   units    numfractions days_to_radiation_therapy_start
## 3    Gy [Not Available]                         1126
## 4    Gy [Not Available]                           31
## 5    Gy [Not Available]                          365
##   radiation_treatment_ongoing days_to_radiation_therapy_end
## 3                        NO                         1126
## 4                        NO                          426
## 5                        NO                          761
##                 measure_of_response   course_number radiation_type_notes
## 3 Radiographic Progressive Disease [Not Available]     [Not Applicable]
## 4 Radiographic Progressive Disease [Not Available]     [Not Applicable]
## 5 Radiographic Progressive Disease [Not Available]     [Not Applicable]
##   regimen_indication regimen_indication_notes
## 3    [Not Available]          [Not Available]
## 4    [Not Available]          [Not Available]
## 5    [Not Available]          [Not Available]
```

clinical information 을 기준으로 원하는 데이터만 추출하는 데에 활용할 수 있다.

TCGA 전체 프로젝트의 메타 데이터 조회

특정 프로젝트가 아닌 전체 TCGA 데이터에서 관심있는 약물이 있는지 보고싶다면 어떻게 할까? recount 패키지를 설치하여 tcga 의 모든 메타데이터를 가져와서 해당 약물이 들어있는 프로젝트를 찾을 수도 있다.

```
library(recount)

# Get all metadata
if (!file.exists("./data/metadata_tcga.Rdata")) {
  metadata_tcga <- recount::all_metadata("tcga")
  save(list=c("metadata_tcga"), file="./data/metadata_tcga.Rdata")
} else {
  load(file="./data/metadata_tcga.Rdata")
}

# Get a drugs list
drug_columns <- grep("drug", names(metadata_tcga@listData))
names(metadata_tcga@listData)[drug_columns] # cgc_drug_therapy_drug_name

## [1] "cgc_case_drug_therapy"
## [2] "cgc_drug_therapy_drug_name"
## [3] "cgc_drug_therapy_pharmaceutical_therapy_type"
## [4] "cgc_drug_therapy_id"
## [5] "xml_has_drugs_information"

all_drugs <- levels(as.factor(metadata_tcga@listData$cgc_drug_therapy_drug
_name))
length(all_drugs)  # 서치된 drugs 수

## [1] 494

# head(all_drugs, 20)  # 20 개 결과 확인


write.table(all_drugs, file="./data/drugs_all_tcga.csv")


# 관심 약물이 있는지 확인
query_drug_idx <- grep("ipilimumab", metadata_tcga$cgc_drug_therapy_drug_n
ame,
                     ignore.case = T)
unique(metadata_tcga$cgc_drug_therapy_drug_name[query_drug_idx])

## [1] "ipilimumab" "Ipilimumab"

query_drug_idx <- grep("Nivolumab", metadata_tcga$cgc_drug_therapy_drug_na
me,
                     ignore.case = T)
unique(metadata_tcga$cgc_drug_therapy_drug_name[query_drug_idx])

## [1] "Nivolumab" "nivolumab"

# 관심 약물에 관련된 cancer project 확인
# ipilimumab
query_proj_idx <- grep("Ipilimumab", metadata_tcga$cgc_drug_therapy_drug_n
ame,
                     ignore.case = T)
Ipili <- metadata_tcga[query_proj_idx,]
unique(Ipili$gdc_cases.project.project_id)  # "TCGA-SKCM"

## [1] "TCGA-SKCM"

# nivolumab
query_proj_idx <- grep("Nivolumab", metadata_tcga$cgc_drug_therapy_drug_na
me,
                     ignore.case = T)
Nivo <- metadata_tcga[query_proj_idx,]
unique(Nivo$gdc_cases.project.project_id)  # "TCGA-BLCA" "TCGA-SKCM"

## [1] "TCGA-BLCA" "TCGA-SKCM"
```

GEO data access in R with GEOquery

geo_data.R 로 gene expression data 다운받고 전처리하기기

GEO excession code 를 이용하여 series data 를 다운받기 위해서는 GDCquery 패키지가 필요하다. download_gse 함수를 통해 메인 코드에서 한 줄로 다운을 받을 수 있다.

Biobase 패키지를 설치하고 나면 extract_gse 함수를 이용하여 다운받은 파일로부터

- exprs(ExpressionSet) 으로 얻어지는 expression matrix
- pdata(ExpressionSet) 으로 얻어지는 clinical data
- fdata(ExpressionSet) 으로 얻어지는 gene information

을 각각 exprs_mat, annot_data, gene_info 라는 변수에 담고, 한꺼번에 리스트로 묶어 반환할 수 있다.

```r
## download geo on your computer, or if exist, load them on R
## load GEO series data as an ExpressionSet
download_gse <- function(gse_serial_no, download_dir="../geo_data") {
  ## Example: gse <- download_gse("gse11111")
  ## gse <- download_gse("gse11111", "../geo_data")
  if (!is.character(gse_serial_no))
    return("Argument Error: Ex) gse <- download_gse('gse11111','../geo_data
')")

  if (!file.exists(download_dir)) {dir.create(download_dir)}

  library(GEOquery)
  file_idx <- grep(gse_serial_no, dir(download_dir), ignore.case = T)
  if (length(file_idx)>0) {
    filename <- file.path(download_dir, dir(download_dir)[file_idx])
    gse <- getGEO(filename=filename)
  } else {
    gse <- getGEO(toupper(gse_serial_no), GSEMatrix = TRUE, destdir=downloa
d_dir)
  }
  return(gse) # result: ExpressionSet
}


## extract exprs_mat, annot_data, gene_info from ExpressionSet
extract_gse <- function(gse, save_dir="../geo_Rdata", save_name) {
  ## Example:
  ## data <- extract_gse(gse, "../geo_Rdata", "gse11111")
  library(Biobase)
  exprs_mat <- exprs(gse)  # expression matrix
  gene_info <- fData(gse)  # gene information
  annot_data <- pData(gse)  # clinical data

  if (!file.exists(save_dir)) {dir.create(save_dir)}

  save(list=c("annot_data", "exprs_mat", "gene_info"),
       file=paste0(save_dir, "/", save_name, ".Rdata"))

  data <- list('exprs_mat'=exprs_mat, 'gene_info'=gene_info, 'annot_data'=a
nnot_data)
  return(data)
}
```

위의 함수들은 functions/geo_data.R 에 저장되어 있다.

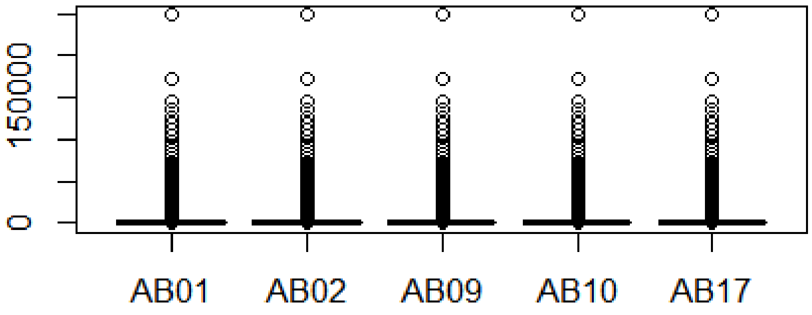
`source("../functions/geo_data.R")` 또는 `source(geo_data.R` 까지의 경로)으로 불러온 뒤 함수를 사용할 수 있다.

```
source("../functions/geo_data.R")
geo_series_idx <- "gse111636"
gse <- download_gse(geo_series_idx) # geo data 로드

##
## -- Column specification ------------------------------------------------
---------
## cols(
##   ID_REF = col_character(),
##   GSM3036125 = col_double(),
##   GSM3036126 = col_double(),
##   GSM3036127 = col_double(),
##   GSM3036128 = col_double(),
##   GSM3036129 = col_double(),
##   GSM3036130 = col_double(),
##   GSM3036131 = col_double(),
##   GSM3036132 = col_double(),
##   GSM3036133 = col_double(),
##   GSM3036134 = col_double(),
##   GSM3036135 = col_double()
## )

## File stored at:

## C:\Users\Public\Documents\ESTsoft\CreatorTemp\RtmpmCb1AG/GPL17586.soft

## Warning: 5990 parsing failures.
##    row   col expected actual         file
## 67363 start a double    --- literal data
## 67363 stop  a double    --- literal data
## 67364 start a double    --- literal data
## 67364 stop  a double    --- literal data
## 67365 start a double    --- literal data
## ..... ..... ........ ...... ............
## See problems(...) for more details.

data <- extract_gse(gse, "../geo_Rdata", geo_series_idx) # data 는 exprs_ma
t, gene_info, annot_data 를 담고있는 리스트
attach(data)
```

DEG 찾기

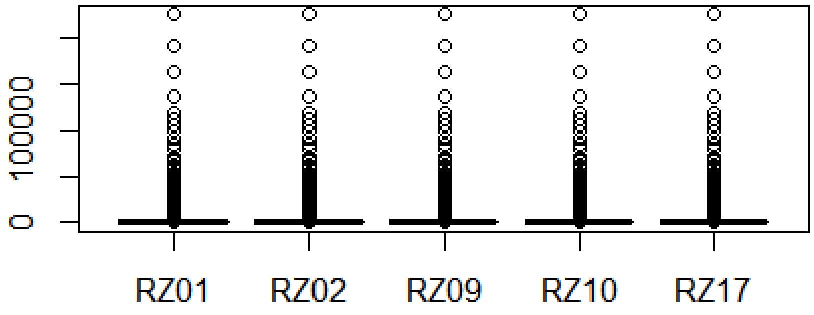
다음은 RNA-Seq 으로부터 얻은 count data 로 된 expression matrix 에서 DEG 를 얻는 예제이다.

```
rm(list=ls())
gse_serial_no <- "gse117358"
gene_counts <- read.csv("../geo_data/GSE117358_genecounts.csv")
head(gene_counts[,1:10])

##   AB01  AB02 AB09  AB10 AB17 AB18 AB25  AB26 AB33  AB34
## 1 5344  6003 5392  5452 4652 4620 4510  5077 4313  6617
## 2    0     0    0     0    0    0    0     0    0     0
## 3  812   967  841   882  658  434  686   867  568  1247
## 4 8346 14646 6320 16274 5830 5422 4158 13197 4356 13387
## 5   88   142   83   128   70   75   51   107   45   143
## 6    2     3    0     0    0    5    4     1    1     5

library(dplyr)

# sample can be grouped into (AB responder, AB nonresponder, RZ responder,
RZ nonresponder)
AB <- gene_counts %>% select(colnames(gene_counts[grep("AB", colnames(gene_
counts))]))
rownames(AB) <- gene_counts$Symbol
AB <- as.matrix(AB)

source("../functions/preprocess_expression_mat.R")
######### DEG analysis on AB
# normalization of genes
AB <- qnormalize(AB)
```



```
# quantile filter of genes
AB1_Filt <- qfilter(AB, qnt.cut =  0.25)

source("../functions/dea.R")
AB1_R <- AB1_Filt[, seq(1, ncol(AB1_Filt), by=2)]
AB1_NR <- AB1_Filt[, seq(2, ncol(AB1_Filt), by=2)]
res_filt_up1 <- analyze_DEG_cnt(AB1_R, AB1_NR, "logFC > 1 & FDR < 0.01")
```

```
## Loading required package: limma

##
## Attaching package: 'limma'

## The following object is masked from 'package:BiocGenerics':
##
##     plotMA

## Disp = 0.10439 , BCV = 0.3231

res_filt_down1 <- analyze_DEG_cnt(AB1_R, AB1_NR, "logFC < -1 & FDR < 0.01")

## Disp = 0.10439 , BCV = 0.3231

head(res_filt_up1, 10)

##               logFC    logCPM        PValue           FDR
## Rpl36a-ps3 6.776259 3.1668831 1.269852e-132 3.535267e-128
## Tpt1-ps5   4.647967 3.5673318  4.371627e-89  1.106419e-85
## Rps11-ps3  4.960140 1.3504531  9.041029e-75  8.390075e-72
## Lcn2       3.894966 4.9180481  4.200203e-73  3.340961e-70
## Gm15056    3.755328 5.4481528  3.232379e-69  2.249736e-66
## Rpl27-ps1  4.713970 0.7153889  2.398134e-62  1.236371e-59
## Rpl36-ps2  4.122101 1.3215682  2.800764e-60  1.392380e-57
## Pdcd1lg2   3.277930 5.1652066  2.369808e-55  9.425065e-53
## Gm2225     3.399671 2.6806590  1.599887e-53  5.498872e-51
## Gm3076     4.440987 0.2619333  7.353458e-52  2.201293e-49

head(res_filt_down1, 10)

##                logFC   logCPM        PValue           FDR
## Krt77      -8.193637 2.119829 1.272916e-126 1.771900e-122
## Gm8623     -7.211263 2.349354 2.608129e-123 2.420344e-119
## Gm5879     -5.457276 3.661994 1.408424e-110 9.802634e-107
## Rpl31-ps13 -5.751612 2.601681 4.600136e-106 2.561356e-102
## Krtap3-2   -5.113961 3.381023 6.311631e-100  2.928597e-96
## Krt33b     -4.777194 4.420790  2.616557e-96  1.040642e-92
## Gm11937    -5.030472 2.977936  7.419042e-95  2.581827e-91
## Krt33a     -4.599164 5.033335  9.030293e-93  2.793371e-89
## Krt34      -4.500810 4.953944  1.106058e-89  3.079266e-86
## Krt31      -4.487836 4.294202  1.216148e-87  2.821463e-84
```



```
## Disp = 0.02767 , BCV = 0.1664

## Disp = 0.02767 , BCV = 0.1664
```

Workflow

DEA 의 작업 순서는 다음과 같다.

5. preprocessing the data
6. quantile normalization
7. Microarray data: log2 transform
8. quantile filtering
9. Finding DEGs
10. Functional enrichment
11. Network analysis
12. Visualization

Microarray data 의 경우 실수값을 갖는 relative expression level matrix 에, RNA-Seq 를 전처리하여 얻은 count data 의 경우 정수값을 갖는 count matrix 에 대해 다음의 과정을 진행하게 된다. **서로 차이가 나는 부분은 log2 tranform 의 적용 유무와 DEG 를 찾는 부분이다.**

- Microarray data 의 경우 raw data 의 경우 log2 transform 을 적용하고, DEG 를 찾기 위해 R 의 limma 패키지를 사용한다.
- RNA-Seq count data 의 경우 log2 transform 을 적용하지 않고, DEG 를 찾기 위해 edgeR 패키지를 이용한다.

TCGA data 의 경우, 앞서 TCGAbiolinks 를 통한 tcga data access 부분에서 살펴봤듯이 TCGAbiolinks 의 함수를 사용하여 전반적인 분석을 할 수 있다. 지금부터 설명하는 내용은 TCGAbiolinks 와 같은 분석 라이브러리를 지원하지 않는 geo data 또는 일반적인 데이터에 적용할 수 있도록 R 로 직접 구현한 것이다.

preprocessing the data

전처리 작업은 샘플마다 expression level 검측에 생길 수 있는 오차, outlier 를 일부 제거하고 통계적 유의성을 확보하게 해준다.

TCGAbiolinks 에서 진행한 것과 같은 gene expression matrix 의 전처리 과정이 function/preprocess_expression_mat.R 에 함수로 들어가있다.

- `qnormalize(mat)`: quantile normalization on expression matrix and visualize in boxplot
- `log2_transform(mat)`: log2 transform if expresssion values are skewed from normal
- `qfilter(mat, qnt.cut)`: quantile filtering genes with relatively low expression

아래의 코드와 같이 expression set 에서 뽑아낸 expression matrix (exprs_mat)에 세 함수를 순차적으로 적용하면 전처리 작업이 끝난다. 전체적인 진행은 앞에서 살펴본 geo 데이터를 이용한 DEG 분석 부분을 참고하자.

```
exprs_mat <- qnormalize(exprs_mat)
exprs_mat <- log2_transform(exprs_mat)
exprs_mat <- qfilter(exprs_mat)
```

각각의 함수를 살펴보며 실제로 무슨 작업이 이루어지는지 알아보자.

quantile normalization

expression matrix 는 각 열이 샘플 하나의 전체 gene expression level 을 담고 있다. 이를 expression level 순으로 줄세워 같은 순위에 있는 expression level 의 평균을 낸다. 그리고 원래 expression level 자리에 들어갈 값을 이 평균값으로 대체한다. 이를 통해 얻을 수 있는 효과는 샘플마다 expression level 의 분포가 동일지며, 개별 샘플의 expression level 의 절대적인 값에 의존하지 않고 gene 들의 expression level 순위에만 결과가 영향을 받는다는 것이다.

```
## quantile normalization on expression matrix
qnormalize <- function(mat) {
  qs = matrix(ncol=ncol(mat), nrow=nrow(mat))  # sorted expression level
  qr = matrix(ncol=ncol(mat), nrow=nrow(mat))  # rank of expression level

  for (i in 1:ncol(mat)){
    qs[,i] = sort(mat[,i])  # sort expression level for each sample
    qr[,i] = rank(mat[,i])  # rank expression level for each sample
  }

  qm = apply(qs, 1, mean)       # mean expression level of same rank

  qn = matrix(ncol=ncol(mat), nrow=nrow(mat))  # quantile normalization res
ult
  for (i in 1:length(qr)){
    r = qr[i]       # (i, j)th rank
    qn[i] = qm[r]  # mean expression level of the rank
  }
  dimnames(qn) <- list(rownames(mat), colnames(mat))

  boxplot(qn[,1:5])
  return(qn)
}
```

log2 transform

Microarray data 의 gene expression level 의 분포는 주로 right-skewed 인 경우가 많다. 일반적으로 통계적 test 의 모수적 방법은 normal 분포를 가정하므로 log2 transform 을 하여 normal 분포에 가깝도록 한다. 아래 함수에서는 shapiro.test 를 통해 데이터가 normal 인지 아닌지를 판단한다.

- p value 가 0.05 보다 크면 변환을 하지 않은 distribution 을 그려줌과 동시에 행렬을 그대로 반환
- 그렇지 않다면, log2 변환을 해서 원래와 변환 후 distribution 을 그려주고 변환한 행렬을 반환한다.

```
## log2 transform if expresssion values are skewed from normal
log2_transform <- function(mat) {
  if (length(mat[,1])>5000) {
    test <- mat[sample(1:nrow(mat), 5000), 1]
  } else {
    test <- mat[,1]
  }
  normality <- shapiro.test(test)
  par(mfrow=c(1,2))
  if (normality$p.value > 0.05) { # if the data is normal
    par(mfrow=c(1,1))
    plot(sort(mat[,1]), type="b", main='distribution of original values')
    return(mat)
  }

  plot(sort(mat[,1]), type="b", main='distribution of original values')
  mat <- log2(mat)

  plot(sort(mat[,1]), type="b", main='distribution after log transform')
  par(mfrow=c(1,1))
  return(mat)
}
```

quantile filtering

간혹 어떤 샘플에서도 expression level 이 매우 적은 gene 이 있을 수 있다. 이런 gene 은 분석에 유의미하지 않으며 불필요한 차원을 늘리는 역할을 하므로 average gene expression level 이 약 하위 25%에 속하는 gene 들은 gene expression 분석에 앞서 의도적으로 제외하는 것이다.

```
# quantile filtering genes with relatively low expression
qfilter <- function (mat, qnt.cut = 0.25) {
  GeneThresh <- as.numeric(quantile(rowMeans(mat), qnt.cut))
  geneFiltered <- which(rowMeans(mat) > GeneThresh)
  mat_Filt <- mat[geneFiltered, ]
  return(mat_Filt)
}
```

Finding DEGs

TCGAbiolinks 나 GEO2R 로 DEG 를 찾기 힘든 경우나 일반적으로 exprs_mat, gene_info (=fData(expression_set) or rowData(expression_set)), annot_data (=pData(expression_set) or colData(expression_set)) 을 갖고 있는 경우에 DEG 분석을 하는 방법은 데이터 종류에 따라 크게 2 가지로 나뉜다.

- RNA-Seq count data: edgeR 패키지 사용
- Microarray data 의 relative expression level: limma 패키지 사용

각각의 경우에 따른 작업 순서를 함수로 구현하여 /functions/dea.R 에 담아두었다.

- `analyze_DEG(grp1, grp2, filtering, download_path)` : Microarry data 에 대해 2 개 그룹간 DEG 추출
- `analyze_DEG2(grps, filtering, download_path)` : Microarry data 에 대해 여러 개 그룹간 DEG 추출 (`grps <- list('grp1' = grp1, 'grp2' = grp2, 'grp3' = grp3 )` 과 같이 선언)

- analyze_DEG_cnt(grp1, grp2, filtering, download_path) : RNA-Seq 의 count data 에 대해 2 개 그룹간 DEG 추출

먼저 함수의 사용 예시를 보자. 여러 개 그룹간 DEG 를 추출하는 analyze_DEG2 함수의 사용 예시의 경우 3 장에서 살펴보기로 한다.

```
source("../functions/dea.R")
AB1_R <- AB1_Filt[, seq(1, ncol(AB1_Filt), by=2)]
AB1_NR <- AB1_Filt[, seq(2, ncol(AB1_Filt), by=2)]
res_filt_up1 <- analyze_DEG_cnt(AB1_R, AB1_NR, "logFC > 1 & FDR < 0.01")

## Disp = 0.10439 , BCV = 0.3231

res_filt_down1 <- analyze_DEG_cnt(AB1_R, AB1_NR, "logFC < -1 & FDR < 0.01")

## Disp = 0.10439 , BCV = 0.3231

head(res_filt_up1, 10)

##                logFC    logCPM        PValue           FDR
## Rpl36a-ps3 6.776259 3.1668831 1.269852e-132 3.535267e-128
## Tpt1-ps5   4.647967 3.5673318  4.371627e-89  1.106419e-85
## Rps11-ps3  4.960140 1.3504531  9.041029e-75  8.390075e-72
## Lcn2       3.894966 4.9180481  4.200203e-73  3.340961e-70
## Gm15056    3.755328 5.4481528  3.232379e-69  2.249736e-66
## Rpl27-ps1  4.713970 0.7153889  2.398134e-62  1.236371e-59
## Rpl36-ps2  4.122101 1.3215682  2.800764e-60  1.392380e-57
## Pdcd1lg2   3.277930 5.1652066  2.369808e-55  9.425065e-53
## Gm2225     3.399671 2.6806590  1.599887e-53  5.498872e-51
## Gm3076     4.440987 0.2619333  7.353458e-52  2.201293e-49

head(res_filt_down1, 10)

##                 logFC   logCPM        PValue           FDR
## Krt77      -8.193637 2.119829 1.272916e-126 1.771900e-122
## Gm8623     -7.211263 2.349354 2.608129e-123 2.420344e-119
## Gm5879     -5.457276 3.661994 1.408424e-110 9.802634e-107
## Rpl31-ps13 -5.751612 2.601681 4.600136e-106 2.561356e-102
## Krtap3-2   -5.113961 3.381023 6.311631e-100  2.928597e-96
## Krt33b     -4.777194 4.420790  2.616557e-96  1.040642e-92
## Gm11937    -5.030472 2.977936  7.419042e-95  2.581827e-91
## Krt33a     -4.599164 5.033335  9.030293e-93  2.793371e-89
## Krt34      -4.500810 4.953944  1.106058e-89  3.079266e-86
## Krt31      -4.487836 4.294202  1.216148e-87  2.821463e-84
```

filtering 의 파라미터의 경우 DEG 로 분류할 gene 을 골라내는 과정으로, adj.P.Val 또는 FDR 이 0.05 또는 0.01 미만으로 통계적 유의수준을 잡고, 이를 만족하는 gene 중 logFC > 1 인 것을 Up-regulated 로, logFC < -1 인 것을 Down-regulated 로 잡는다. 이와 관련하여 몇 가지 유의할 할 점이 있다. 구체적인 사항은 각각의 함수에 주석으로 달린 사용 예시 (EXAMPLE)을 확인하자.

1. “logFC > 1 & FDR < 0.01”와 같이 조건문을 쌍따옴표로 감싸 문자열로 입력하되 대소문자를 바꿔쓰지 않도록 유의해야 한다.

2. Microarray data 에 적용하는 위의 두 함수의 경우 FDR 대신 adj.P.Val 을 써야 한다. ex) “logFC > 1 & adj.P.Val < 0.01”

또한 Up-regulation 과 Down-regulation 의 기준이 되는 dataset 은 grp1 이며, grp2 는 대조군이 되는 dataset 이다.

```
## DEG analysis

analyze_DEG <- function(grp1, grp2, filtering, download_path=NULL) {
  ## DEG analysis for expression level matrices
  ## EXAMPLES :
  ## res_filt <- analyze_DEG(up, down, "adj.P.Val < 0.05 & logFC>=1")
  ## res_filt <- analyze_DEG(up, down, "adj.P.Val < 0.05 & abs(logFC)>=1")
  ## res_filt <- analyze_DEG(up, down, "adj.P.Val < 0.05 & logFC>=1", "../r
esults/gse11111_CD274_up_down-adjpval0_05-logFC1.csv")
  library(limma)
  grp_names <- c(deparse(substitute(grp1)), deparse(substitute(grp2)))
  grp <- c(rep(grp_names[1], ncol(grp1)), rep(grp_names[2], ncol(grp2)))
  design <- model.matrix(~grp+0)
  colnames(design) <- grp_names

  data <- cbind(grp1, grp2)
  fit <- lmFit(data,design)
  x <- paste(grp_names[2], grp_names[1], sep='-')
  cont <- makeContrasts(contrasts=x,levels=design)

  fit.cont <- contrasts.fit(fit,cont)
  fit.cont2 <- eBayes(fit.cont)
  res <- topTable(fit.cont2,number=Inf)
  res_filt <- subset(res, eval(parse(text=filtering)))

  if (!is.null(download_path))
    write.csv(res_filt, download_path)
  return(res_filt)
}

analyze_DEG2 <- function(grps, filtering, download_path=NULL) {
  ## DEG analysis on several groups with order
  ## Ex) grps <- list('low'=low, 'medium'=medium, 'high'=high)
  ## EXAMPLES :
  ## res_filt <- analyze_DEG2(grps, "adj.P.Val < 0.05 & logFC>=1")
  ## res_filt <- analyze_DEG2(grps, "adj.P.Val < 0.05 & abs(logFC)>=1")
```

```r
  ## res_filt <- analyze_DEG2(grps, "adj.P.Val < 0.05 & logFC>=1", "../results/gse11111_CD274_up_down-adjpval0_05-logFC1.csv")
  library(limma)
  n <- length(grps)
  grp_names <- names(grps)
  grp <- c()
  data <- matrix(nrow=nrow(grps[[1]]), ncol=0)
  for (i in 1:n) {
    grp <- c(grp, rep(grp_names[i], ncol(grps[[i]])))
    data <- cbind(data, grps[[i]])
  }
  design <- model.matrix(~grp+0)
  colnames(design) <- grp_names

  fit <- lmFit(data,design)
  x <- paste(grp_names[1:n-1], grp_names[2:n], sep='-')
  cont <- makeContrasts(contrasts=x,levels=design)

  fit.cont <- contrasts.fit(fit,cont)
  fit.cont2 <- eBayes(fit.cont)
  res <- topTable(fit.cont2,number=Inf)

  res_filt <- subset(res, eval(parse(text=filtering)))

  if (!is.null(download_path))
    write.csv(res_filt, download_path)
  return(res_filt)
}

analyze_DEG_cnt <- function(grp1, grp2, filtering, download_path=NULL) {
  ## DEG analysis for count matrices
  ## EXAMPLES :
  ## res_filt <- analyze_DEG_cnt(up, down, "logFC > 1 & FDR < 0.01")
  ## res_filt <- analyze_DEG_cnt(up, down, "logFC < -1 & FDR < 0.01")
  ## res_filt <- analyze_DEG_cnt(up, down, "logFC < -1 & FDR < 0.01", "../results/gse11111_CD274_up_down-adjpval0_05-logFC1.csv")
  library(edgeR)
  grp_names <- c(deparse(substitute(grp1)), deparse(substitute(grp2)))
  grp <- c(rep(grp_names[1], ncol(grp1)), rep(grp_names[2], ncol(grp2)))

  data <- DGEList(counts=cbind(grp1, grp2),group=factor(grp))
  cDisp <- estimateCommonDisp(data, verbose=T)
  res <- exactTest(cDisp, pair=c(1,2))
  res_sort <- topTags(res, n = Inf, adjust.method = "BH", sort.by = "PValue")
  res_filt <- subset(res_sort$table, eval(parse(text=paste0(filtering))))

  if (!is.null(download_path))
    write.csv(res_filt, download_path)
  return(res_filt)
}
```

Functional enrichment

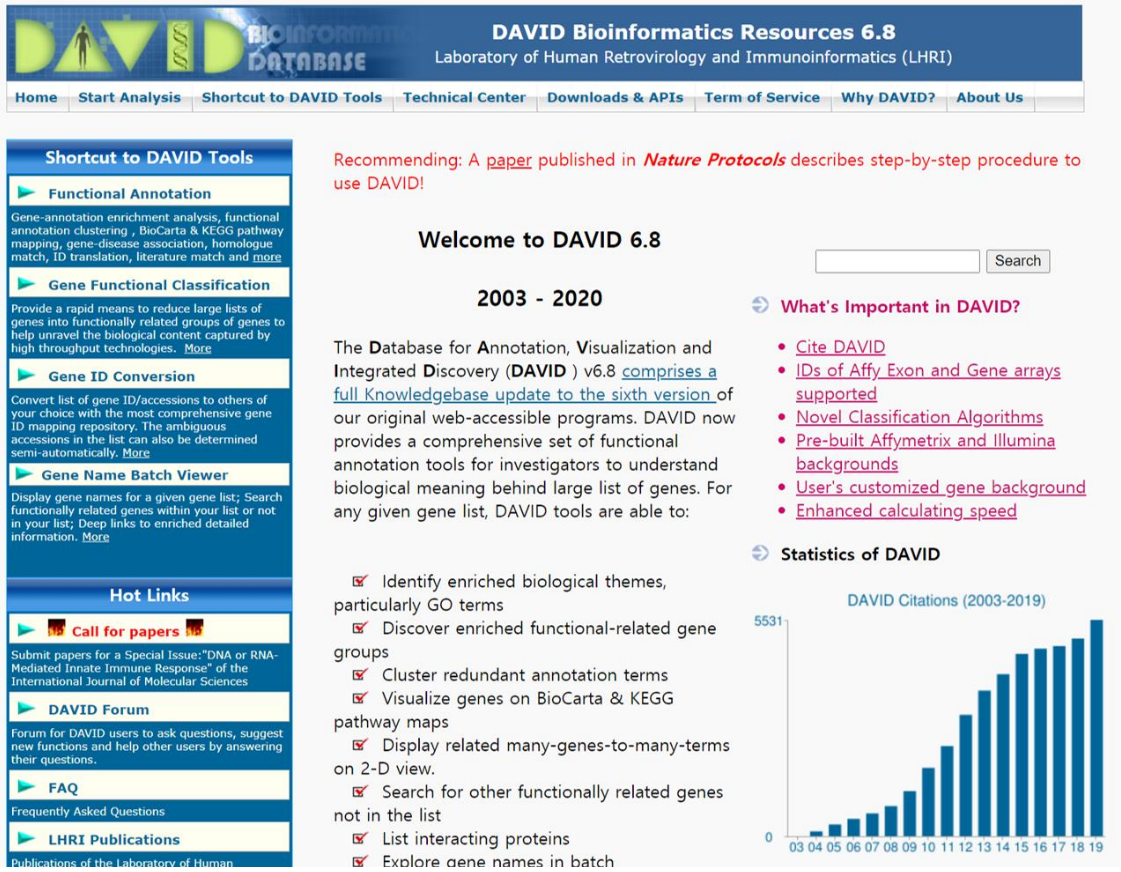
DEG 를 찾고 나면 해당 gene 들이 어떤 역할을 하는지 궁금할 것이다. DEGs set 을 biological role 을 알아보기 위해 GO 또는 KEGG functional enrichment analysis 를 하려고 한다.

- Gene Ontology (GO)는 Biological Processes (BP), Cellular components (CC), and molecular functions (MF)와 같이 3 가지 구분된 요소를 가진다.
- Kyoto Encyclopedia of Genes and Genomes (KEGG)는 다양한 pathway (예를 들면, signaling pathways, metabolic pathways 등)를 위한 데이터베이스를 구축하고 있다. 자세한 정보는 http://www.genome.jp/kegg/를 참고하자.

Enrichment analysis 는 주로 DAVID (http://David.ncifcrf.gov/home.jsp/)를 통해 수행하고 결과파일을 다운받아 R 등으로 Barplot 을 그린다. 한편 클릭만으로 보고서를 만들어주는 Metascape 도 있다.

DAVID

DAVID 에서 functional enrichment analysis



DAVID Bioinformatics Resources 6.8

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Hot Links

Call for papers

Submit papers for a Special Issue:"DNA or RNA-Mediated Innate Immune Response" of the International Journal of Molecular Sciences

DAVID Forum

Forum for DAVID users to ask questions, suggest new functions and help other users by answering their questions.

FAQ

Frequently Asked Questions

LHRI Publications

Publications of the Laboratory of Human

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2020

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

☑ Identify enriched biological themes, particularly GO terms
☑ Discover enriched functional-related gene groups
☑ Cluster redundant annotation terms
☑ Visualize genes on BioCarta & KEGG pathway maps
☑ Display related many-genes-to-many-terms on 2-D view.
☑ Search for other functionally related genes not in the list
☑ List interacting proteins
☑ Explore gene names in batch

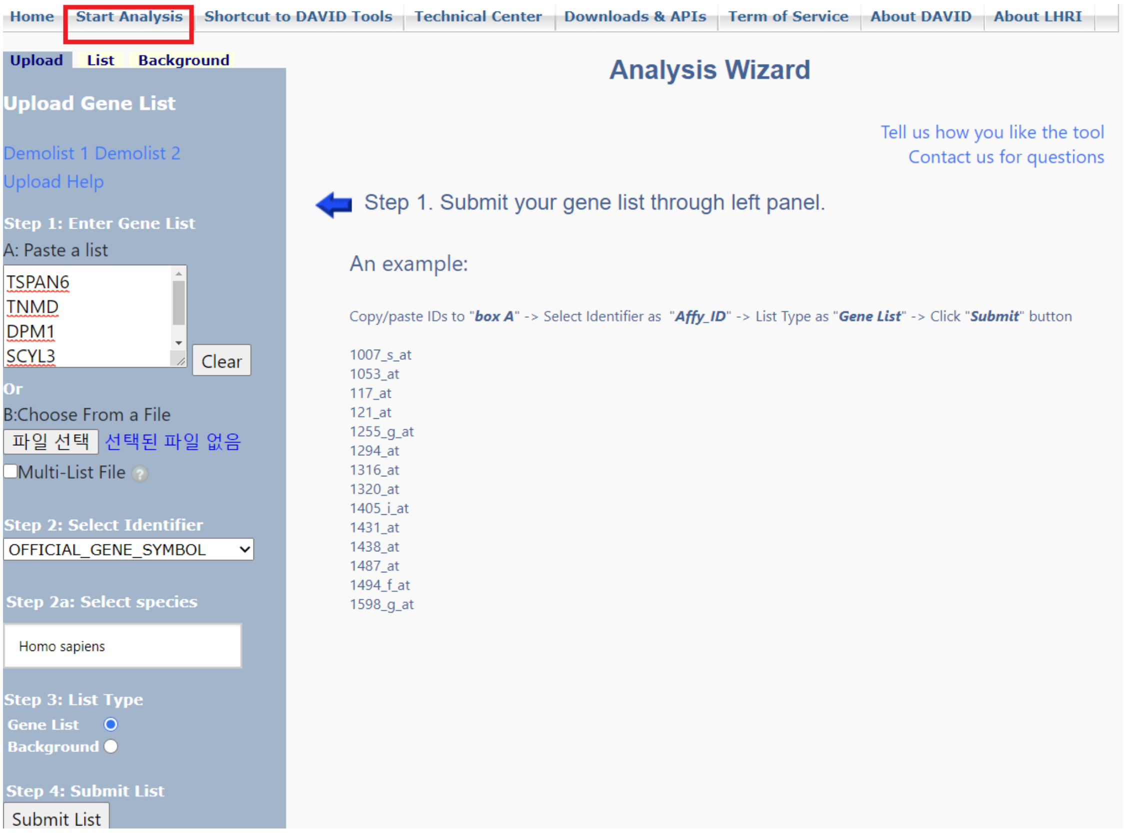
What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID main

DAVID (http://David.ncifcrf.gov/home.jsp/)는 GO 와 KEGG analysis 모두 가능한 웹사이트이다.



Upload Gene List

Step 1: Enter Gene List

A: Paste a list

TSPAN6
TNMD
DPM1
SCYL3

Or

B:Choose From a File

□ Multi-List File

Step 2: Select Identifier

OFFICIAL_GENE_SYMBOL

Step 2a: Select species

Homo sapiens

Step 3: List Type

Gene List ●
Background ○

Step 4: Submit List

Analysis Wizard

Step 1. Submit your gene list through left panel.

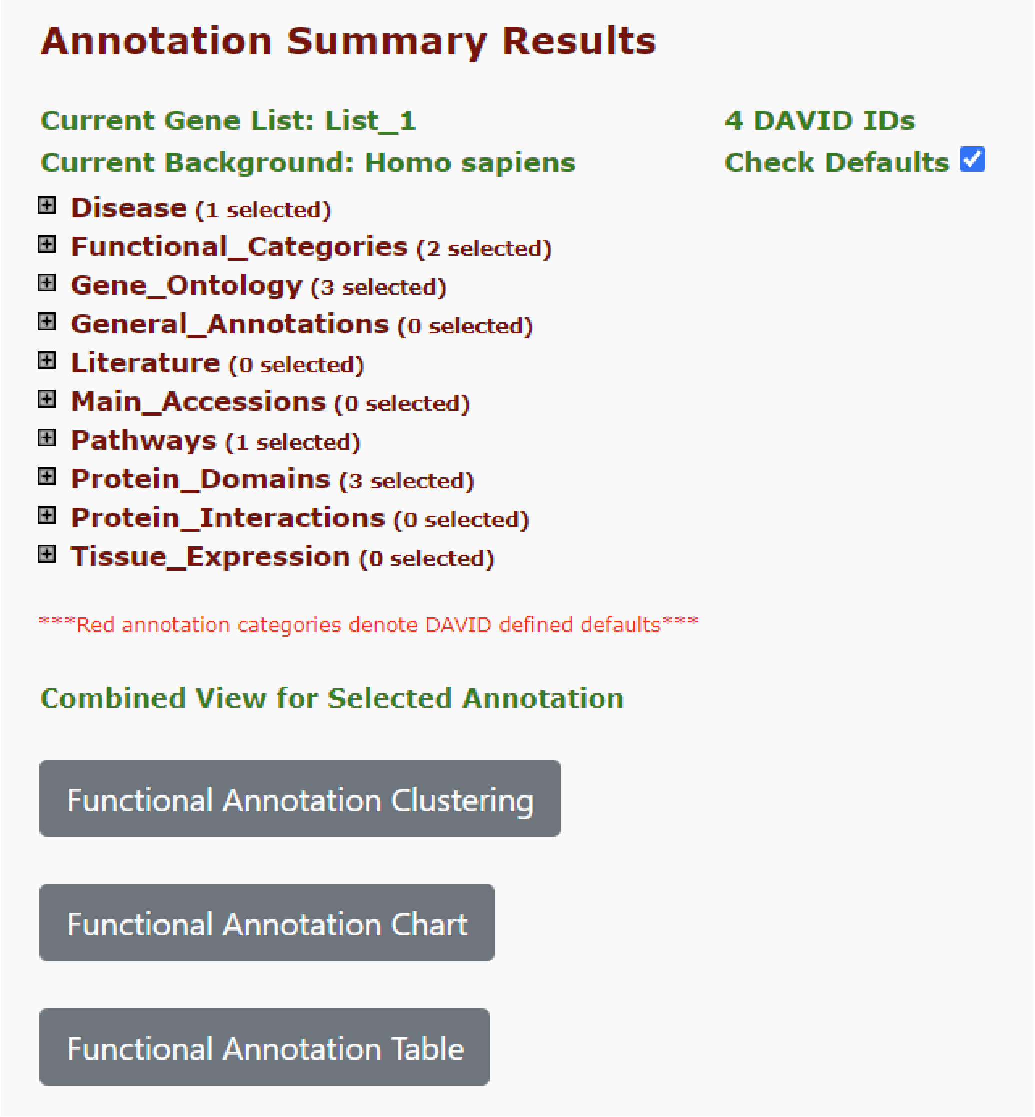
An example:

Copy/paste IDs to "***box A***" -> Select Identifier as "***Affy_ID***" -> List Type as "***Gene List***" -> Click "***Submit***" button

1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at

DAVID initiating enrichment analysis

상단의 Start Analysis 를 클릭하고 엔터로 구분된 gene ID 를 넣은 뒤, 알맞은 identifier 을 지정하고, species 를 'Homo sapiens'와 같이 설정, List Type 은 Gene List 로 주고 Submit List 를 클릭한 뒤 나오는 페이지에서 Functional Annotation Tool 을 클릭하면 다음과 같은 정보를 얻을 수 있다.



Annotation Summary Results

Current Gene List: List_1 — 4 DAVID IDs

Current Background: Homo sapiens — Check Defaults ☑

⊞ Disease (1 selected)
⊞ Functional_Categories (2 selected)
⊞ Gene_Ontology (3 selected)
⊞ General_Annotations (0 selected)
⊞ Literature (0 selected)
⊞ Main_Accessions (0 selected)
⊞ Pathways (1 selected)
⊞ Protein_Domains (3 selected)
⊞ Protein_Interactions (0 selected)
⊞ Tissue_Expression (0 selected)

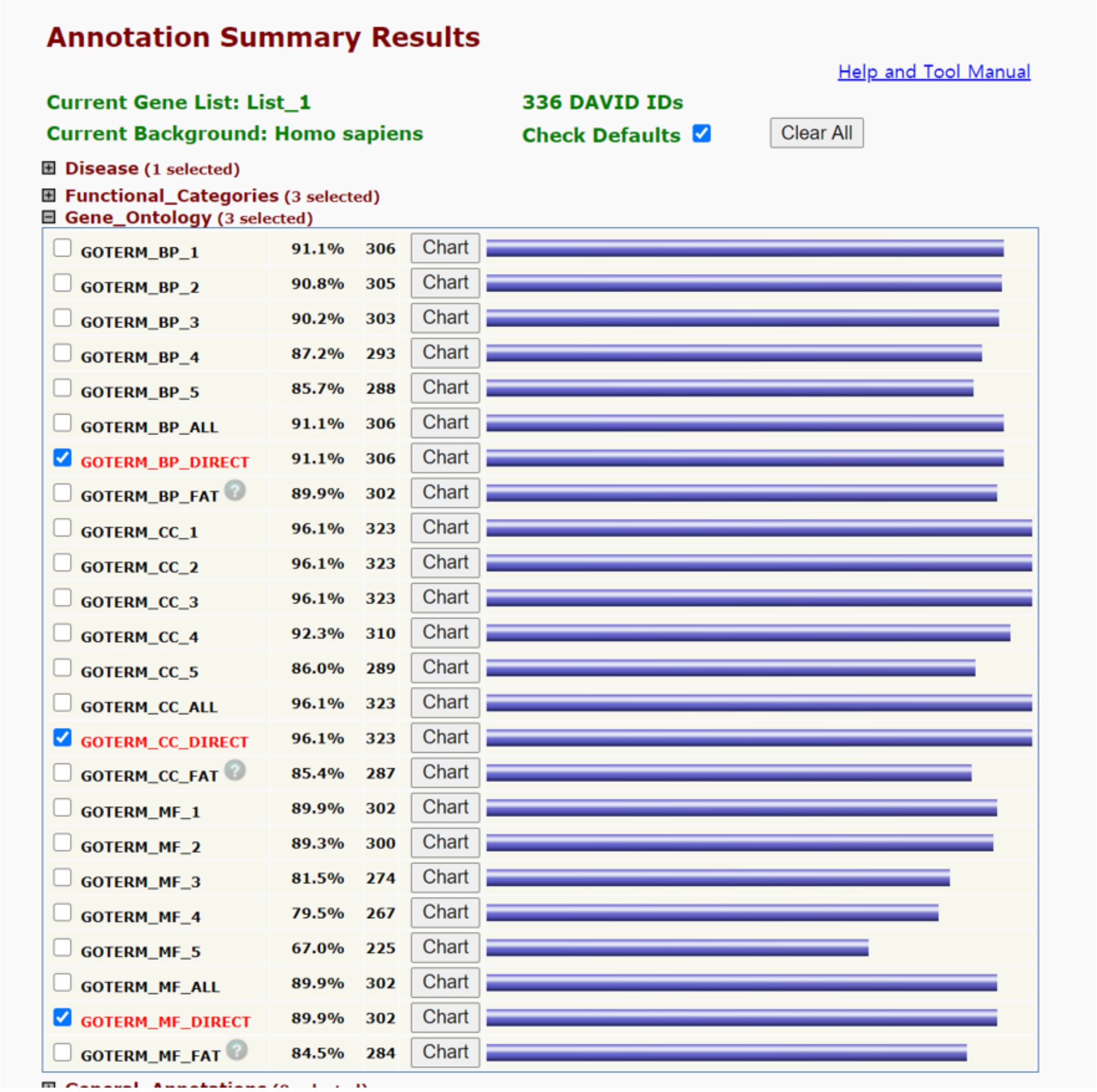
Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

Functional Annotation Tool

각각을 extend 해 Gene_Ontology 또는 Pathways > KEGG pathway 에 따른 annotation table 을 확인할 수 있다. 중요하다고 표시된 빨간색 레코드의 오른쪽에 있는 Chart 를 클릭하면 새 창으로 annotation table 이 열린다. 이를 다운받을 수 있다.



Annotation Summary Results

Help and Tool Manual

Current Gene List: List_1 336 DAVID IDs

Current Background: Homo sapiens Check Defaults ☑ Clear All

⊞ Disease (1 selected)

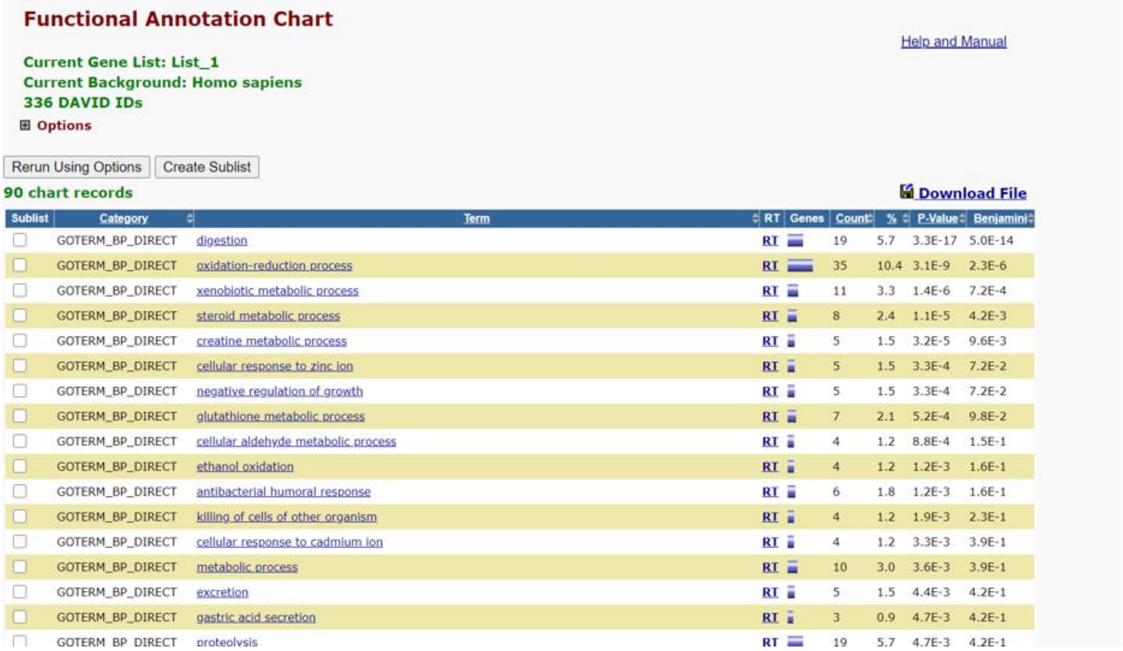
⊞ Functional_Categories (3 selected)

⊟ Gene_Ontology (3 selected)

☐ GOTERM_BP_1	91.1%	306	Chart
☐ GOTERM_BP_2	90.8%	305	Chart
☐ GOTERM_BP_3	90.2%	303	Chart
☐ GOTERM_BP_4	87.2%	293	Chart
☐ GOTERM_BP_5	85.7%	288	Chart
☐ GOTERM_BP_ALL	91.1%	306	Chart
☑ GOTERM_BP_DIRECT	91.1%	306	Chart
☐ GOTERM_BP_FAT ?	89.9%	302	Chart
☐ GOTERM_CC_1	96.1%	323	Chart
☐ GOTERM_CC_2	96.1%	323	Chart
☐ GOTERM_CC_3	96.1%	323	Chart
☐ GOTERM_CC_4	92.3%	310	Chart
☐ GOTERM_CC_5	86.0%	289	Chart
☐ GOTERM_CC_ALL	96.1%	323	Chart
☑ GOTERM_CC_DIRECT	96.1%	323	Chart
☐ GOTERM_CC_FAT ?	85.4%	287	Chart
☐ GOTERM_MF_1	89.9%	302	Chart
☐ GOTERM_MF_2	89.3%	300	Chart
☐ GOTERM_MF_3	81.5%	274	Chart
☐ GOTERM_MF_4	79.5%	267	Chart
☐ GOTERM_MF_5	67.0%	225	Chart
☐ GOTERM_MF_ALL	89.9%	302	Chart
☑ GOTERM_MF_DIRECT	89.9%	302	Chart
☐ GOTERM_MF_FAT ?	84.5%	284	Chart

⊞ General_Annotations (0 selected)

GO terms selection



Functional Annotation Chart

Help and Manual

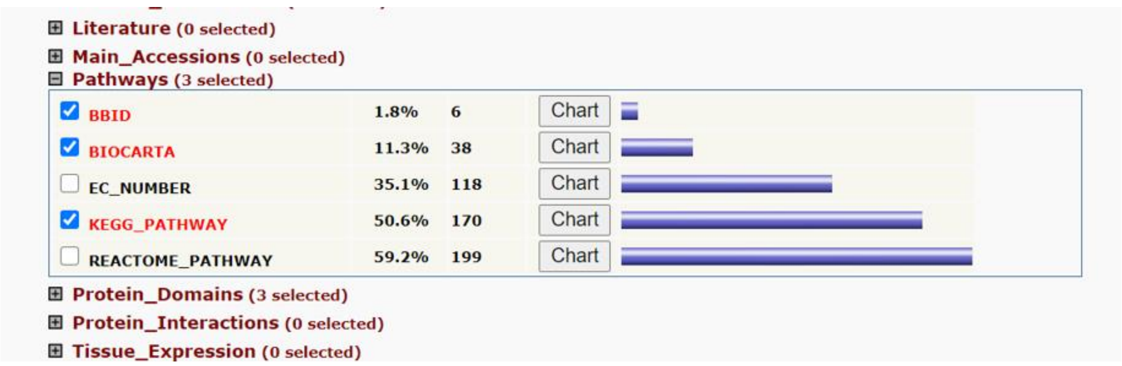
Current Gene List: List_1
Current Background: Homo sapiens
336 DAVID IDs

⊞ **Options**

90 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	digestion	RT	▬	19	5.7	3.3E-17	5.0E-14
□	GOTERM_BP_DIRECT	oxidation-reduction process	RT	▬	35	10.4	3.1E-9	2.3E-6
□	GOTERM_BP_DIRECT	xenobiotic metabolic process	RT	▬	11	3.3	1.4E-6	7.2E-4
□	GOTERM_BP_DIRECT	steroid metabolic process	RT	▬	8	2.4	1.1E-5	4.2E-3
□	GOTERM_BP_DIRECT	creatine metabolic process	RT	▬	5	1.5	3.2E-5	9.6E-3
□	GOTERM_BP_DIRECT	cellular response to zinc ion	RT	▬	5	1.5	3.3E-4	7.2E-2
□	GOTERM_BP_DIRECT	negative regulation of growth	RT	▬	5	1.5	3.3E-4	7.2E-2
□	GOTERM_BP_DIRECT	glutathione metabolic process	RT	▬	7	2.1	5.2E-4	9.8E-2
□	GOTERM_BP_DIRECT	cellular aldehyde metabolic process	RT	▬	4	1.2	8.8E-4	1.5E-1
□	GOTERM_BP_DIRECT	ethanol oxidation	RT	▬	4	1.2	1.2E-3	1.6E-1
□	GOTERM_BP_DIRECT	antibacterial humoral response	RT	▬	6	1.8	1.2E-3	1.6E-1
□	GOTERM_BP_DIRECT	killing of cells of other organism	RT	▬	4	1.2	1.9E-3	2.3E-1
□	GOTERM_BP_DIRECT	cellular response to cadmium ion	RT	▬	4	1.2	3.3E-3	3.9E-1
□	GOTERM_BP_DIRECT	metabolic process	RT	▬	10	3.0	3.6E-3	3.9E-1
□	GOTERM_BP_DIRECT	excretion	RT	▬	5	1.5	4.4E-3	4.2E-1
□	GOTERM_BP_DIRECT	gastric acid secretion	RT	▬	3	0.9	4.7E-3	4.2E-1
□	GOTERM_BP_DIRECT	proteolysis	RT	▬	19	5.7	4.7E-3	4.2E-1

GO result

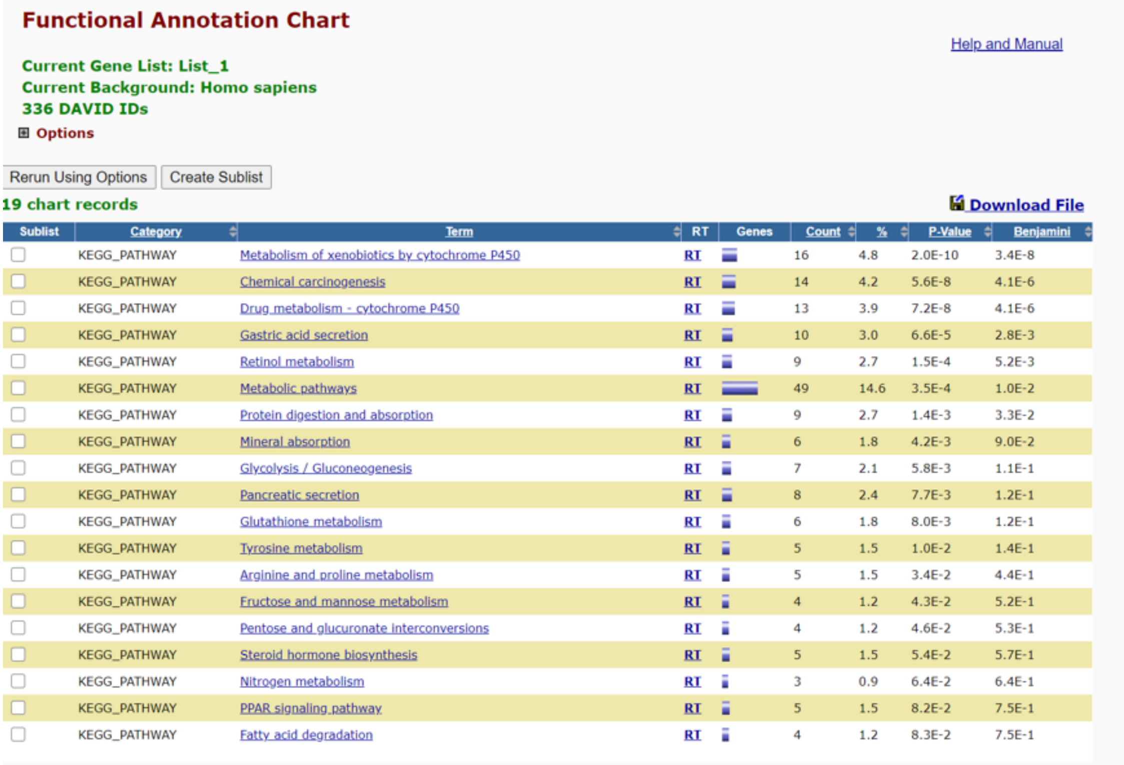


⊞ **Literature** (0 selected)
⊞ **Main_Accessions** (0 selected)
⊟ **Pathways** (3 selected)

☑ BBID	1.8%	6	Chart	▬
☑ BIOCARTA	11.3%	38	Chart	▬
□ EC_NUMBER	35.1%	118	Chart	▬
☑ KEGG_PATHWAY	50.6%	170	Chart	▬
□ REACTOME_PATHWAY	59.2%	199	Chart	▬

⊞ **Protein_Domains** (3 selected)
⊞ **Protein_Interactions** (0 selected)
⊞ **Tissue_Expression** (0 selected)

Pathway terms selection



Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
336 DAVID IDs
⊞ **Options**

19 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT	▬	16	4.8	2.0E-10	3.4E-8
□	KEGG_PATHWAY	Chemical carcinogenesis	RT	▬	14	4.2	5.6E-8	4.1E-6
□	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT	▬	13	3.9	7.2E-8	4.1E-6
□	KEGG_PATHWAY	Gastric acid secretion	RT	▬	10	3.0	6.6E-5	2.8E-3
□	KEGG_PATHWAY	Retinol metabolism	RT	▬	9	2.7	1.5E-4	5.2E-3
□	KEGG_PATHWAY	Metabolic pathways	RT	▬	49	14.6	3.5E-4	1.0E-2
□	KEGG_PATHWAY	Protein digestion and absorption	RT	▬	9	2.7	1.4E-3	3.3E-2
□	KEGG_PATHWAY	Mineral absorption	RT	▬	6	1.8	4.2E-3	9.0E-2
□	KEGG_PATHWAY	Glycolysis / Gluconeogenesis	RT	▬	7	2.1	5.8E-3	1.1E-1
□	KEGG_PATHWAY	Pancreatic secretion	RT	▬	8	2.4	7.7E-3	1.2E-1
□	KEGG_PATHWAY	Glutathione metabolism	RT	▬	6	1.8	8.0E-3	1.2E-1
□	KEGG_PATHWAY	Tyrosine metabolism	RT	▬	5	1.5	1.0E-2	1.4E-1
□	KEGG_PATHWAY	Arginine and proline metabolism	RT	▬	5	1.5	3.4E-2	4.4E-1
□	KEGG_PATHWAY	Fructose and mannose metabolism	RT	▬	4	1.2	4.3E-2	5.2E-1
□	KEGG_PATHWAY	Pentose and glucuronate interconversions	RT	▬	4	1.2	4.6E-2	5.3E-1
□	KEGG_PATHWAY	Steroid hormone biosynthesis	RT	▬	5	1.5	5.4E-2	5.7E-1
□	KEGG_PATHWAY	Nitrogen metabolism	RT	▬	3	0.9	6.4E-2	6.4E-1
□	KEGG_PATHWAY	PPAR signaling pathway	RT	▬	5	1.5	8.2E-2	7.5E-1
□	KEGG_PATHWAY	Fatty acid degradation	RT	▬	4	1.2	8.3E-2	7.5E-1

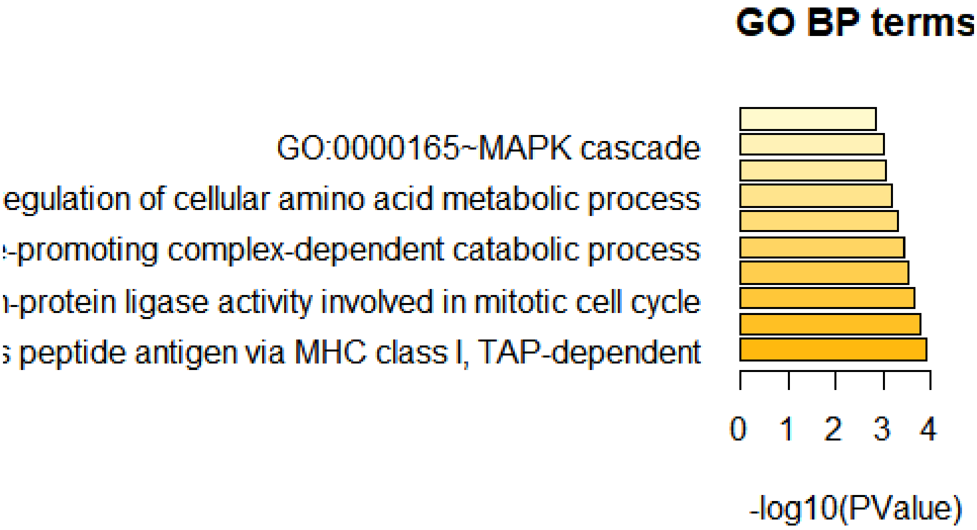
KEGG pathway result

Visualization in R

저장한 텍스트 파일을 R 에서 읽어와 barplot 으로 그리는 예제이다. enrich_term 으로 읽어올 파일의 이름과 bar 의 개수 cnt 를 원하는 대로 지정해서 사용하면 된다.

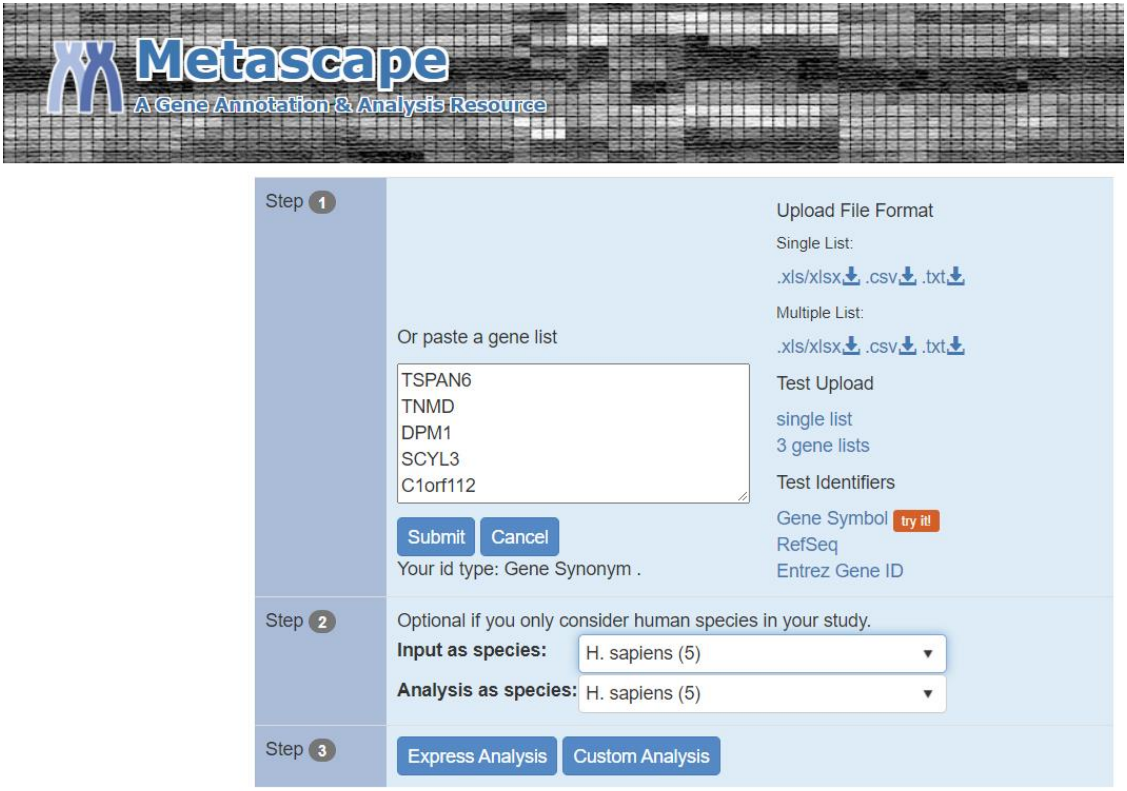
`colorRampPalette(c("color A", "color B"))`는 color A 부터 color B 까지 그라데이션 컬러를 정의하는 함수이다. 원하는 두 color 을 입력하여 새로 만든 `colfunc` 함수에 cnt 를 아규먼트로 넣으면 cnt 개수의 컬러가 구현된다. ex) `colfunc(cnt)`

```
enrich_term <- read.table("./data/ch2-DAVID_output.txt", header=T, sep = "
\t")
cnt <- 10

colfunc <- colorRampPalette(c("darkgoldenrod1", "lemonchiffon"))
par(mar=c(4,19,4,0))
barplot(-log10(enrich_term$PValue[1:cnt]), names = enrich_term$Term[1:cn
t],
        horiz = T, las=1, # horizontal barplot
        xlab = "-log10(PValue)", xlim = c(0, max(-log10(enrich_term$PValue
[1:10]))+1),
        col = colfunc(cnt), main = "GO BP terms")
```



```
par(mar=c(5.1, 4.1, 4.1, 2.1))
```

Metascape



Or paste a gene list

```
TSPAN6
TNMD
DPM1
SCYL3
C1orf112
```

Your id type: Gene Synonym .

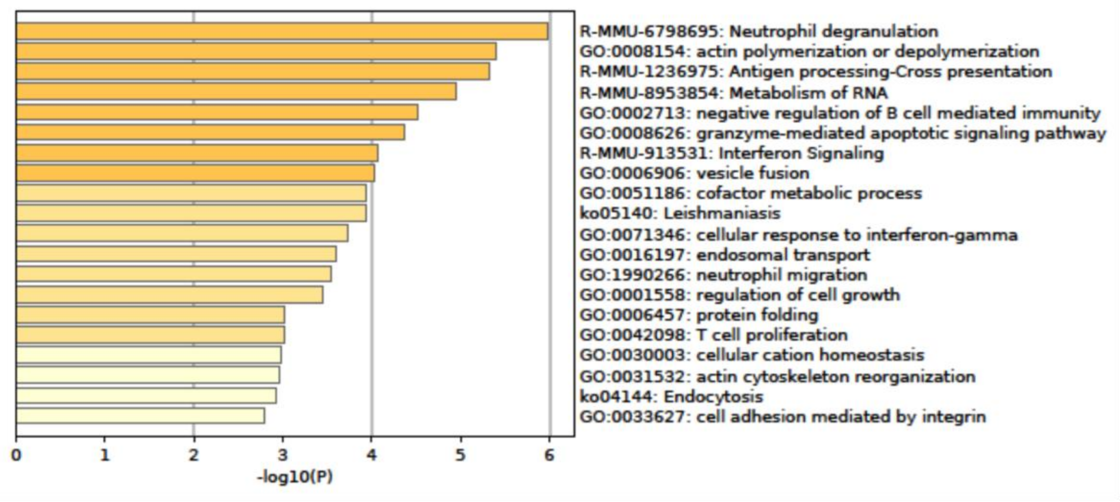
Optional if you only consider human species in your study.

Input as species: H. sapiens (5)

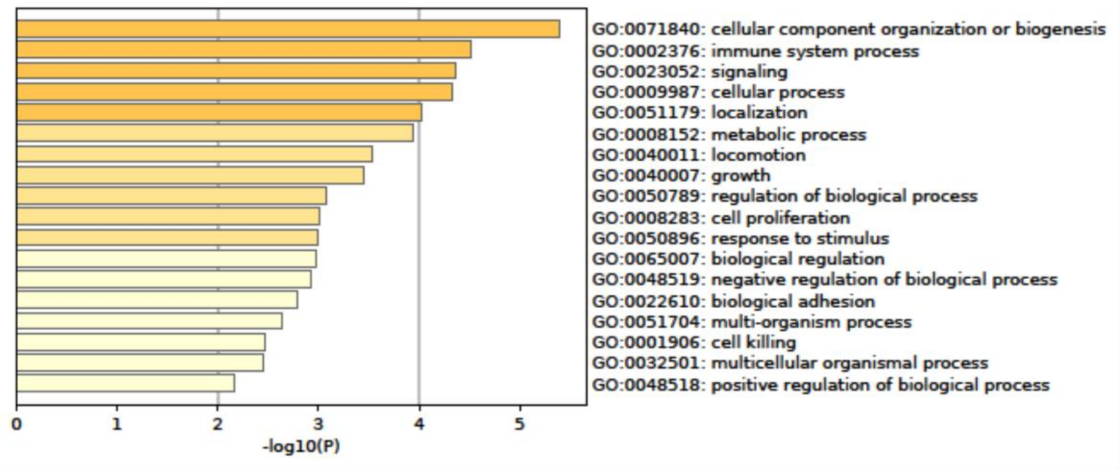
Analysis as species: H. sapiens (5)

enrichiment analysis with Metascape

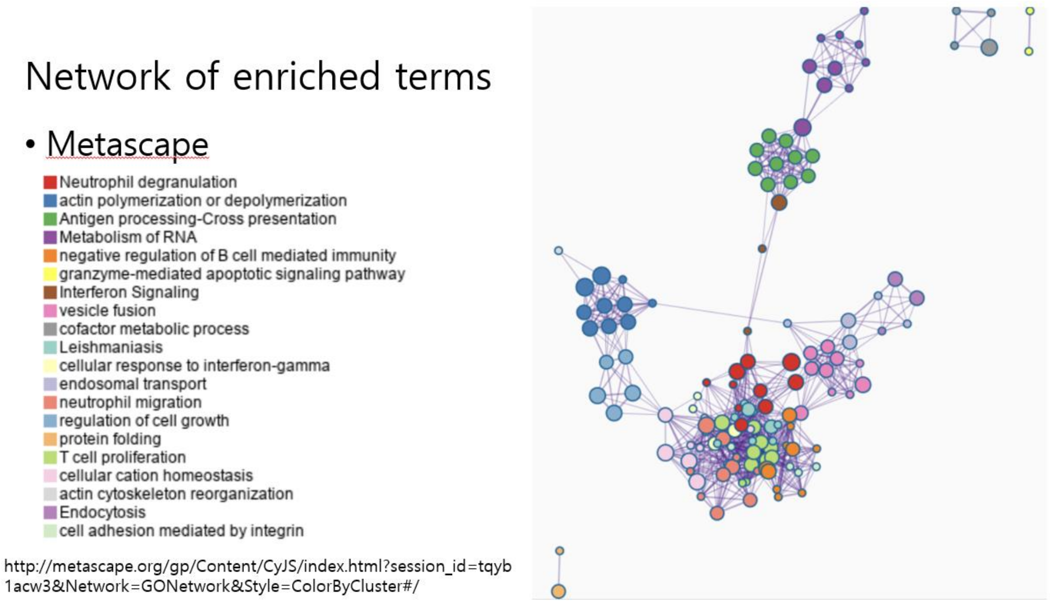
Metascape (https://metascape.org/gp/index.html#/main/step1) 역시 functional enrichment analysis 를 위한 웹사이트이다. 사용방법은 DAVID 와 유사하며, gene annotation 뿐만 아니라 PPI network 를 도식화하여 함께 제공한다.



GSE138224 upregulated genes annotation1



GSE138224 upregulated genes annotation2



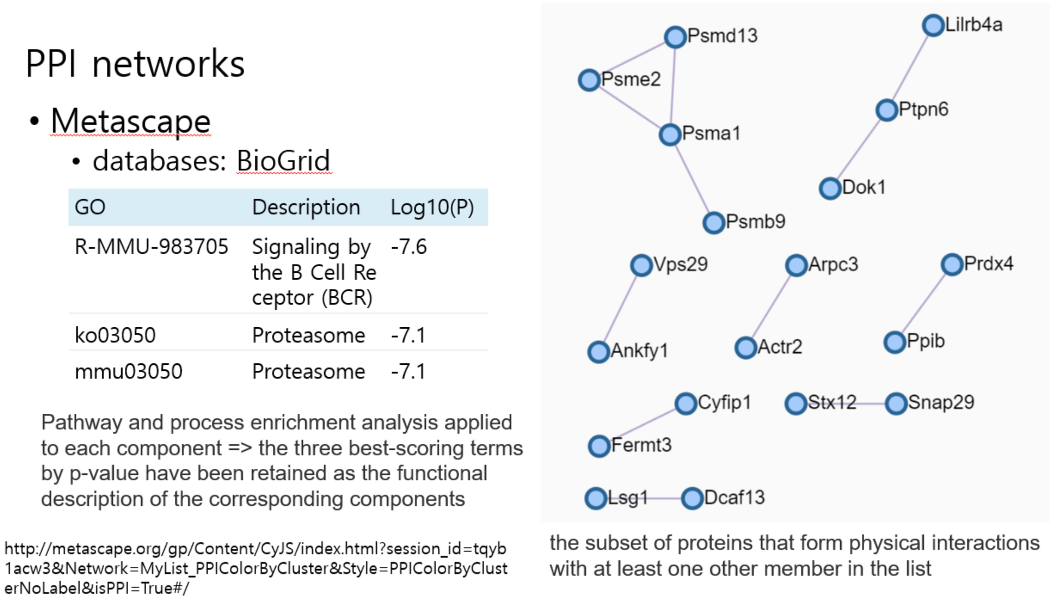
Network of enriched terms

- Metascape

🟥 Neutrophil degranulation
🟦 actin polymerization or depolymerization
🟩 Antigen processing-Cross presentation
🟪 Metabolism of RNA
🟧 negative regulation of B cell mediated immunity
🟨 granzyme-mediated apoptotic signaling pathway
🟫 Interferon Signaling
■ vesicle fusion
■ cofactor metabolic process
■ Leishmaniasis
■ cellular response to interferon-gamma
■ endosomal transport
■ neutrophil migration
■ regulation of cell growth
■ protein folding
■ T cell proliferation
■ cellular cation homeostasis
■ actin cytoskeleton reorganization
■ Endocytosis
■ cell adhesion mediated by integrin

http://metascape.org/gp/Content/CyJS/index.html?session_id=tqyb1acw3&Network=GONetwork&Style=ColorByCluster#/

GSE138224 upregulated genes annotation3



PPI networks

- Metascape
  - databases: BioGrid

GO	Description	Log10(P)
R-MMU-983705	Signaling by the B Cell Receptor (BCR)	-7.6
ko03050	Proteasome	-7.1
mmu03050	Proteasome	-7.1

Pathway and process enrichment analysis applied to each component => the three best-scoring terms by p-value have been retained as the functional description of the corresponding components

http://metascape.org/gp/Content/CyJS/index.html?session_id=tqyb1acw3&Network=MyList_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True#/

the subset of proteins that form physical interactions with at least one other member in the list

GSE138224 upregulated genes annotation4

Network analysis

Protein-Protein Interaction network

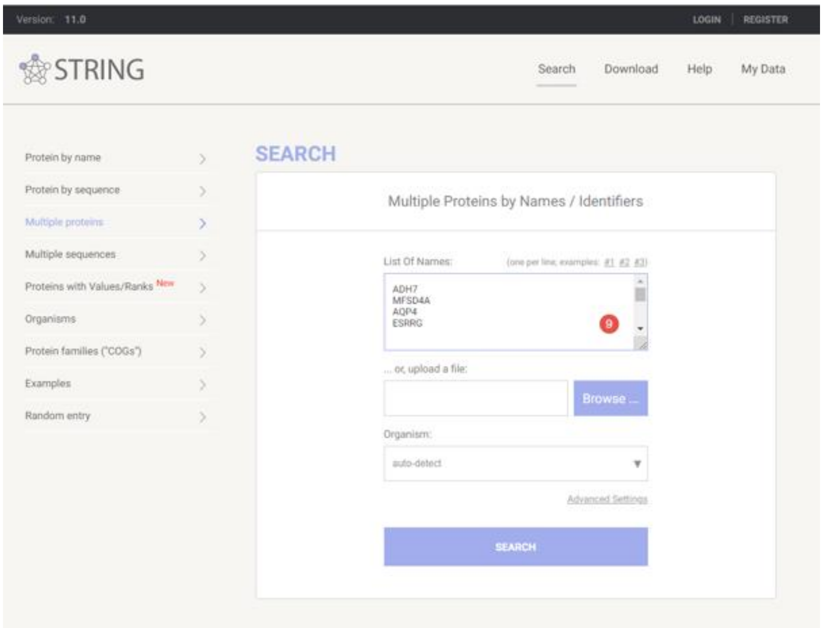
다음은 Protein-Protein Interaction network 에 대한 Wikipedia 의 정의이다.

physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect

PPI network 를 그리기 위한 툴은 Cytoscape 프로그램 - String App, 그리고 웹 버전인 STRING (https://string-db.org/)이 있다.

STRING (https://string-db.org/)에서 PPI network 를 그리는 방법은 다음과 같다.

3. DEGs list 를 입력한다.



Multiple Proteins by Names / Identifiers

List Of Names: (one per line, examples: #1 #2 #3)

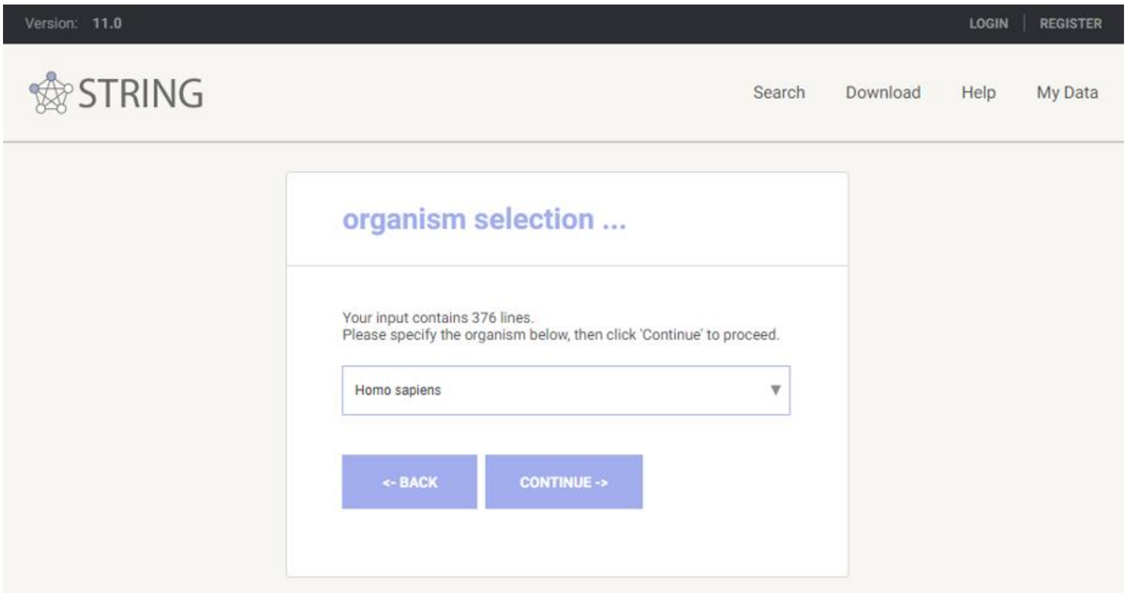
ADH7
MFSD4A
AQP4
ESRRG

... or, upload a file:

Organism: auto-detect

How to use STRING 1

4. organism 을 선택한다

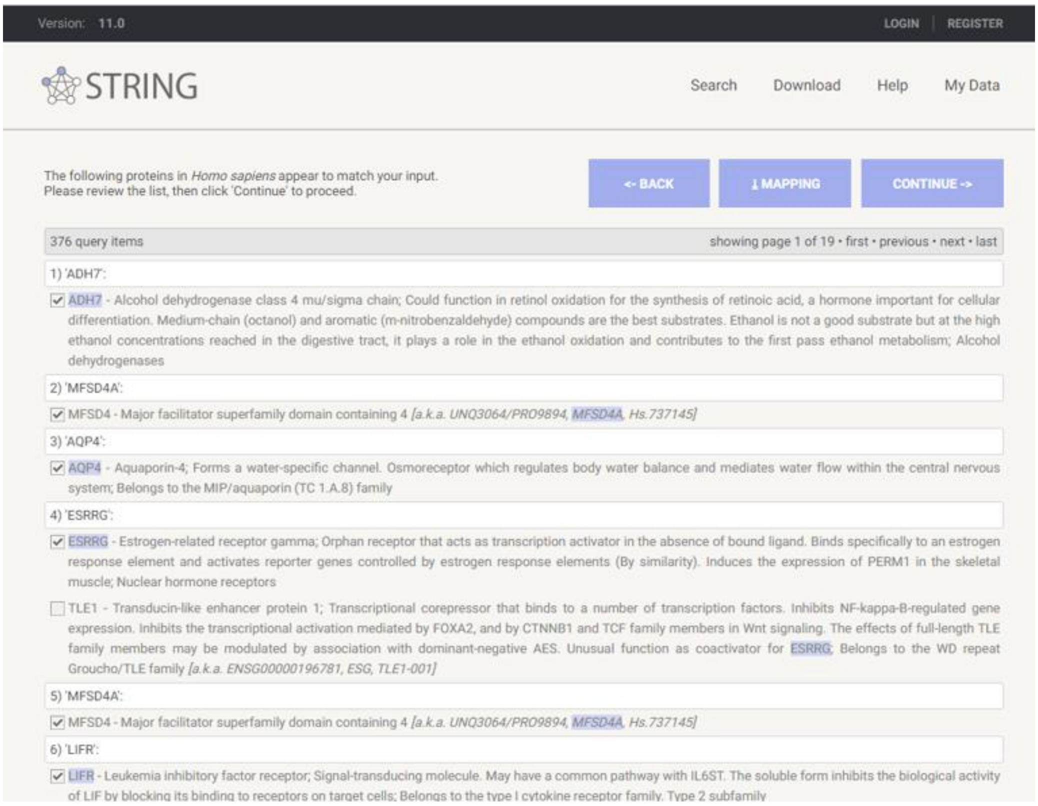


Your input contains 376 lines.
Please specify the organism below, then click 'Continue' to proceed.

Homo sapiens

How to use STRING 2

4. input 으로 넣은 gene 이 원하는 protein 에 매핑되었는지 확인하고 Continue 를 클릭한다.



The following proteins in *Homo sapiens* appear to match your input.
Please review the list, then click 'Continue' to proceed.

376 query items — showing page 1 of 19 • first • previous • next • last

1) 'ADH7':

☑ ADH7 - Alcohol dehydrogenase class 4 mu/sigma chain; Could function in retinol oxidation for the synthesis of retinoic acid, a hormone important for cellular differentiation. Medium-chain (octanol) and aromatic (m-nitrobenzaldehyde) compounds are the best substrates. Ethanol is not a good substrate but at the high ethanol concentrations reached in the digestive tract, it plays a role in the ethanol oxidation and contributes to the first pass ethanol metabolism; Alcohol dehydrogenases

2) 'MFSD4A':

☑ MFSD4 - Major facilitator superfamily domain containing 4 *[a.k.a. UNQ3064/PRO9894, MFSD4A, Hs.737145]*

3) 'AQP4':

☑ AQP4 - Aquaporin-4; Forms a water-specific channel. Osmoreceptor which regulates body water balance and mediates water flow within the central nervous system; Belongs to the MIP/aquaporin (TC 1.A.8) family

4) 'ESRRG':

☑ ESRRG - Estrogen-related receptor gamma; Orphan receptor that acts as transcription activator in the absence of bound ligand. Binds specifically to an estrogen response element and activates reporter genes controlled by estrogen response elements (By similarity). Induces the expression of PERM1 in the skeletal muscle; Nuclear hormone receptors

☐ TLE1 - Transducin-like enhancer protein 1; Transcriptional corepressor that binds to a number of transcription factors. Inhibits NF-kappa-B-regulated gene expression. Inhibits the transcriptional activation mediated by FOXA2, and by CTNNB1 and TCF family members in Wnt signaling. The effects of full-length TLE family members may be modulated by association with dominant-negative AES. Unusual function as coactivator for ESRRG; Belongs to the WD repeat Groucho/TLE family *[a.k.a. ENSG00000196781, ESG, TLE1-001]*

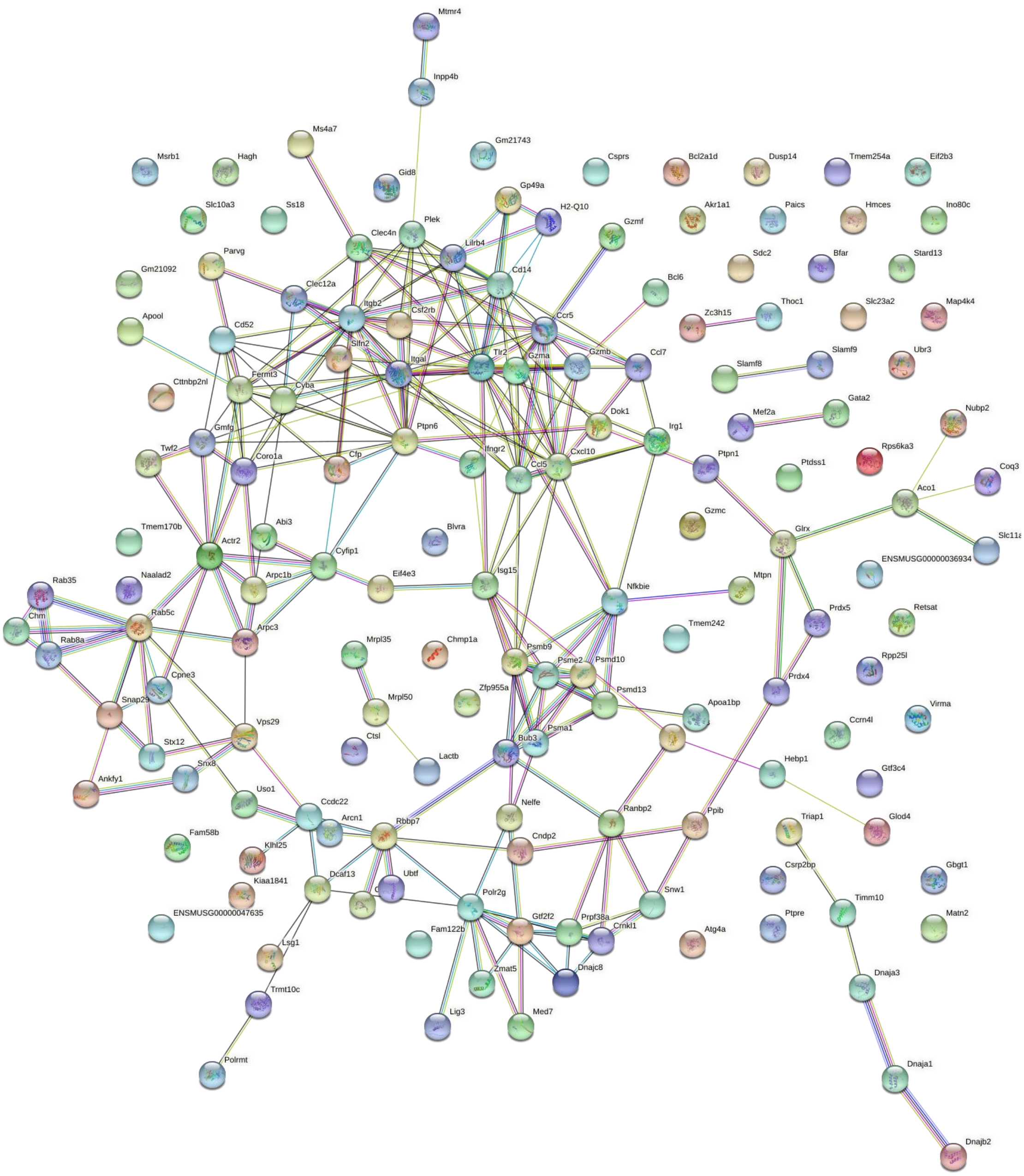
5) 'MFSD4A':

☑ MFSD4 - Major facilitator superfamily domain containing 4 *[a.k.a. UNQ3064/PRO9894, MFSD4A, Hs.737145]*

6) 'LIFR':

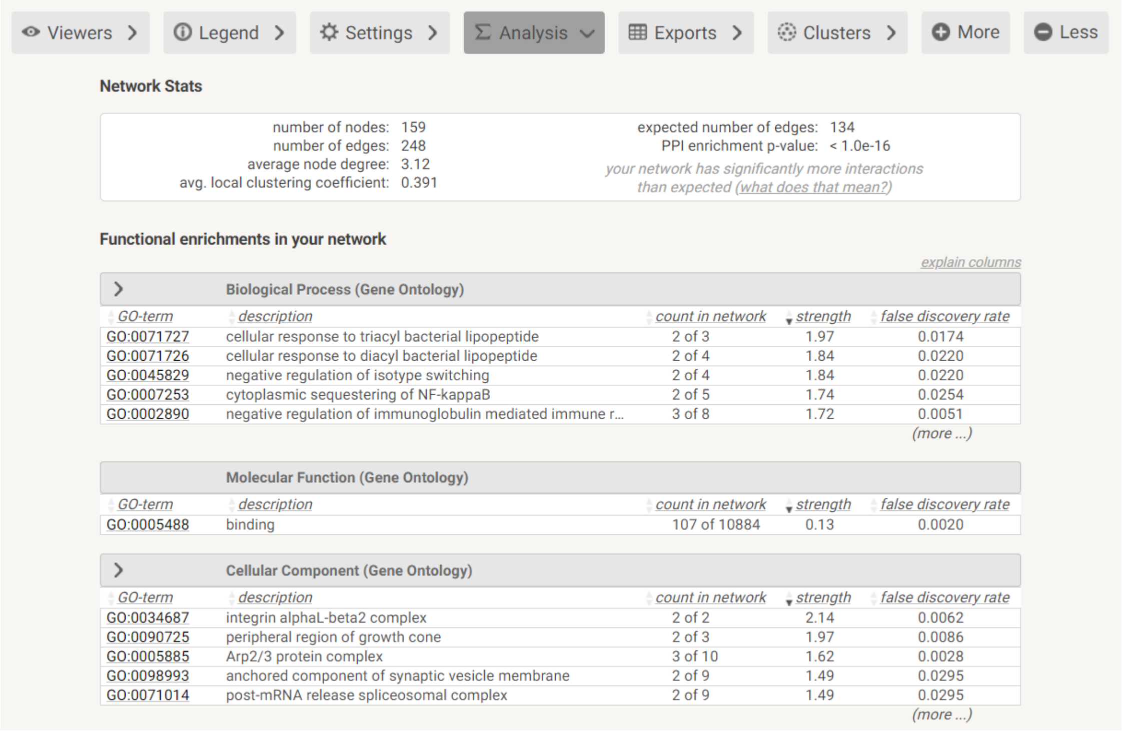
☑ LIFR - Leukemia inhibitory factor receptor; Signal-transducing molecule. May have a common pathway with IL6ST. The soluble form inhibits the biological activity of LIF by blocking its binding to receptors on target cells; Belongs to the type I cytokine receptor family. Type 2 subfamily

How to use STRING 1



STRING PPI network

결과물로 위와 같은 PPI network 가 그려진다.



Network Stats

number of nodes:	159	expected number of edges:	134
number of edges:	248	PPI enrichment p-value:	< 1.0e-16
average node degree:	3.12		
avg. local clustering coefficient:	0.391		

your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0071727	cellular response to triacyl bacterial lipopeptide	2 of 3	1.97	0.0174
GO:0071726	cellular response to diacyl bacterial lipopeptide	2 of 4	1.84	0.0220
GO:0045829	negative regulation of isotype switching	2 of 4	1.84	0.0220
GO:0007253	cytoplasmic sequestering of NF-kappaB	2 of 5	1.74	0.0254
GO:0002890	negative regulation of immunoglobulin mediated immune r...	3 of 8	1.72	0.0051

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0005488	binding	107 of 10884	0.13	0.0020

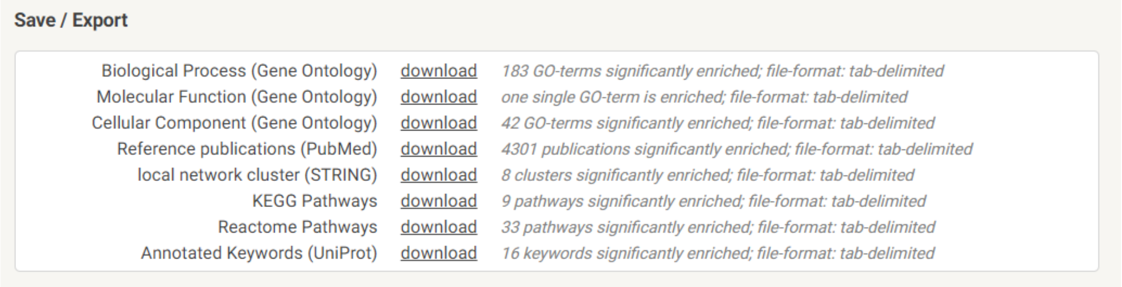
Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0034687	integrin alphaL-beta2 complex	2 of 2	2.14	0.0062
GO:0090725	peripheral region of growth cone	2 of 3	1.97	0.0086
GO:0005885	Arp2/3 protein complex	3 of 10	1.62	0.0028
GO:0098993	anchored component of synaptic vesicle membrane	2 of 9	1.49	0.0295
GO:0071014	post-mRNA release spliceosomal complex	2 of 9	1.49	0.0295

(more ...)

STRING enrichment terms

`Analysis` 탭에서 enrichment analysis 결과를 확인할 수 있다.



Save / Export

Biological Process (Gene Ontology)	download	*183 GO-terms significantly enriched; file-format: tab-delimited*
Molecular Function (Gene Ontology)	download	*one single GO-term is enriched; file-format: tab-delimited*
Cellular Component (Gene Ontology)	download	*42 GO-terms significantly enriched; file-format: tab-delimited*
Reference publications (PubMed)	download	*4301 publications significantly enriched; file-format: tab-delimited*
local network cluster (STRING)	download	*8 clusters significantly enriched; file-format: tab-delimited*
KEGG Pathways	download	*9 pathways significantly enriched; file-format: tab-delimited*
Reactome Pathways	download	*33 pathways significantly enriched; file-format: tab-delimited*
Annotated Keywords (UniProt)	download	*16 keywords significantly enriched; file-format: tab-delimited*

STRING downloadable outputs

`Exports` 탭에서 다운받을 수 있는 파일의 리스트이다.

Visualization

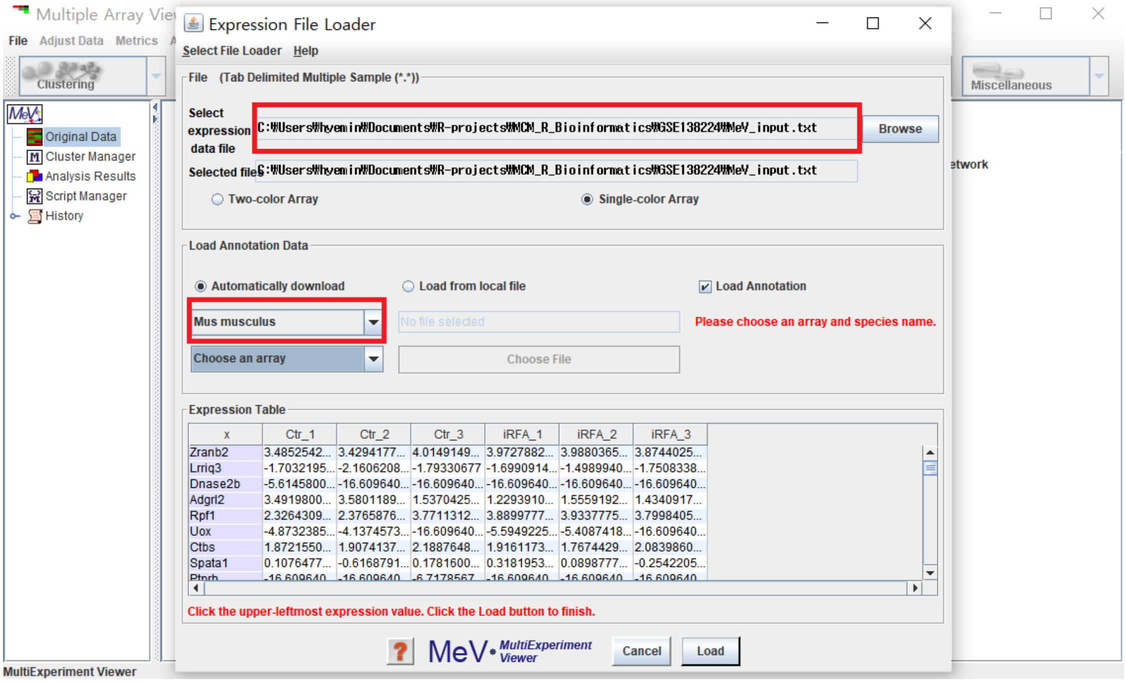
heatmap of DEGs

expression level matrix 를 heatmap 으로 그려 서로 다른 그룹의 expression level 차이를 색깔로 표현할 수 있다. R 에서 heatmap 함수를 이용해도 되지만 대규모의 데이터도 쉽게 그려주는 MeV 라는 프로그램을 이용하려고 한다.

MeV

MeV 는 대규모의 expression level matrix 를 heatmap 으로 그려주는 프로그램이다. https://sourceforge.net/projects/mev-tm4/에서 MeV 프로그램을 다운받을 수 있다.

MeV 프로그램을 열고 상단의 `File > Load Data` 를 클릭하면 expression matrix 데이터를 업로드할 수 있는 창이 뜬다.



File (Tab Delimited Multiple Sample (*.*))

Select expression data file: C:₩Users₩hyemin₩Documents₩R-projects₩MCM_R_Bioinformatics₩GSE138224₩MeV_input.txt

Selected files: C:₩Users₩hyemin₩Documents₩R-projects₩MCM_R_Bioinformatics₩GSE138224₩MeV_input.txt

○ Two-color Array ● Single-color Array

Load Annotation Data

● Automatically download ○ Load from local file ☑ Load Annotation

Mus musculus

Choose an array

No file selected

Please choose an array and species name.

Expression Table

x	Ctr_1	Ctr_2	Ctr_3	iRFA_1	iRFA_2	iRFA_3
Zranb2	3.4852542...	3.4294177...	4.0149149...	3.9727882...	3.9880365...	3.8744025...
Lrriq3	-1.7032195...	-2.1606208...	-1.79330677	-1.6990914...	-1.4989940...	-1.7508338...
Dnase2b	-5.6145800...	-16.609640...	-16.609640...	-16.609640...	-16.609640...	-16.609640...
Adgrl2	3.4919800...	3.5801189...	1.5370425...	1.2293910...	1.5559192...	1.4340917...
Rpf1	2.3264309...	2.3765876...	3.7711312...	3.8899777...	3.9337775...	3.7998405...
Uox	-4.8732385...	-4.1374573...	-16.609640...	-5.5949225...	-5.4087418...	-16.609640...
Ctbs	1.8721550...	1.9074137...	2.1887648...	1.9161173...	1.7674429...	2.0839860...
Spata1	0.1076477...	-0.6168791...	0.1781600...	0.3181953...	0.0898777...	-0.2542205...
Ptprh	-16.609640	-16.609640	-6.7178567	-16.609640	-16.609640	-16.609640

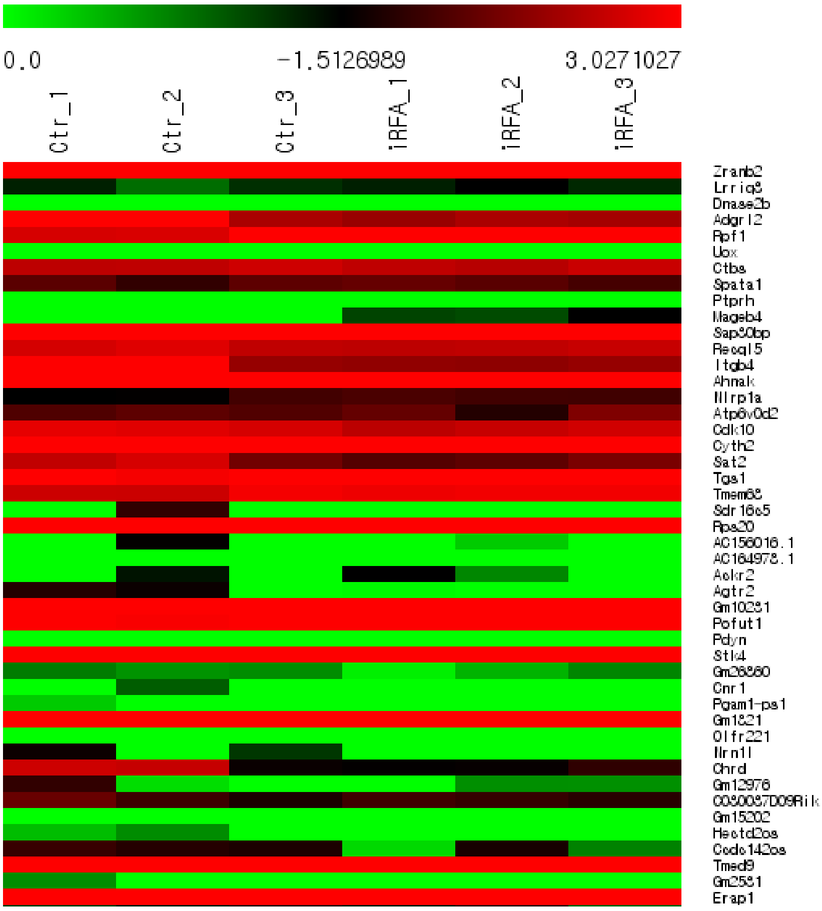
Click the upper-leftmost expression value. Click the Load button to finish.

MeV loading data

다음과 같이 탭으로 필드가 구분되고 header 가 있는 양식의 expression data file 을 업로드한다.

```
##           x     Ctr_1      Ctr_2      Ctr_3     iRFA_1     iRFA_2     iRFA_3
## 1  Zranb2  3.485254   3.429418   4.014915   3.972788   3.988037   3.87440
3
## 2  Lrriq3 -1.703220  -2.160621  -1.793307  -1.699091  -1.498994  -1.7508
34
## 3 Dnase2b -5.614580 -16.609640 -16.609640 -16.609640 -16.609640 -16.6096
40
## 4  Adgrl2  3.491980   3.580119   1.537043   1.229391   1.555919   1.43409
2
## 5    Rpf1  2.326431   2.376588   3.771131   3.889978   3.933778   3.79984
1
## 6     Uox -4.873239  -4.137457 -16.609640  -5.594923  -5.408742 -16.6096
40
```

그리고 species 를 설정한 뒤 아래에 테이블이 잘 로드되었는지 확인할 수 있다. 그러고 나서 `Load` 를 클릭하면 대규모의 expression level matrix 로 된 heatmap 을 얻을 수 있다. `File > Save Image` 를 통해 그림을 저장할 수도 있다.

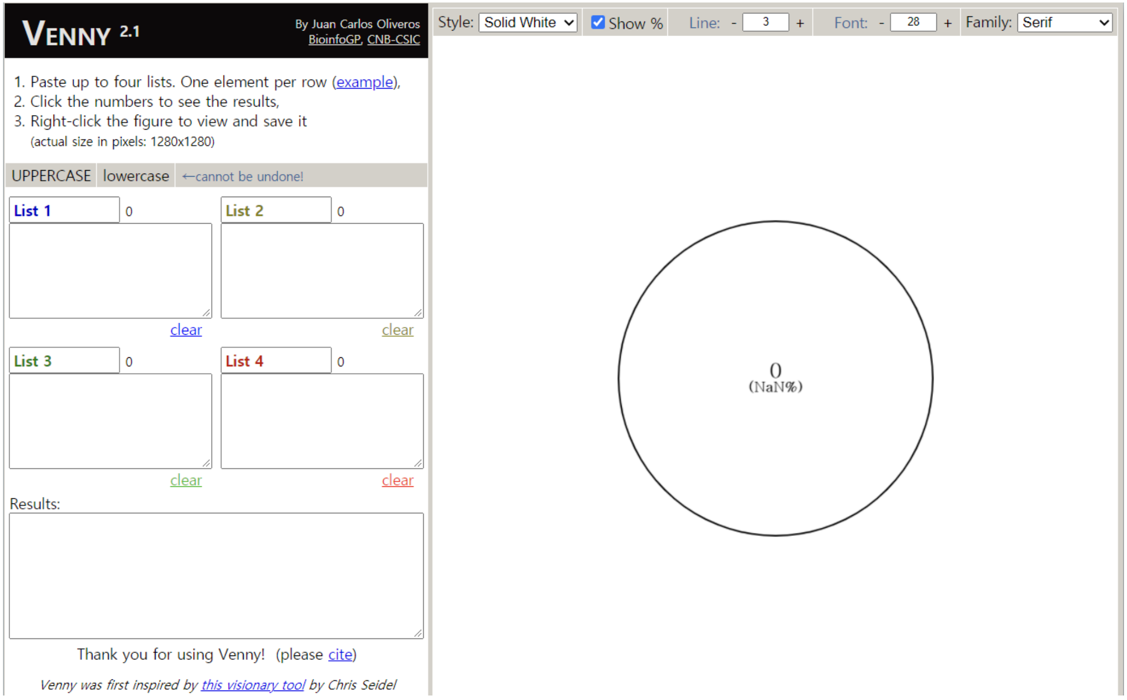


MeV heatmap of GSE138224

Venn-diagram

여러 건의 실험/분석에 대해 독립적으로 DEG 를 찾은 뒤, 공통적으로 등장하는 gene 을 확인하고 싶은 경우에 벤다이어그램을 이용하면 효과적이다. 웹사이트인 Venny 또는 R 의 VennDiagram 패키지를 통해 벤다이어그램을 그릴 수 있다.

Venny



VENNY 2.1

By Juan Carlos Oliveros
BioinfoGP, CNB-CSIC

1. Paste up to four lists. One element per row (example),
2. Click the numbers to see the results,
3. Right-click the figure to view and save it
(actual size in pixels: 1280x1280)

Style: Solid White ☑ Show % Line: - 3 + Font: - 28 + Family: Serif

Results:

Thank you for using Venny! (please cite)

Venny was first inspired by this visionary tool by Chris Seidel

Venny main

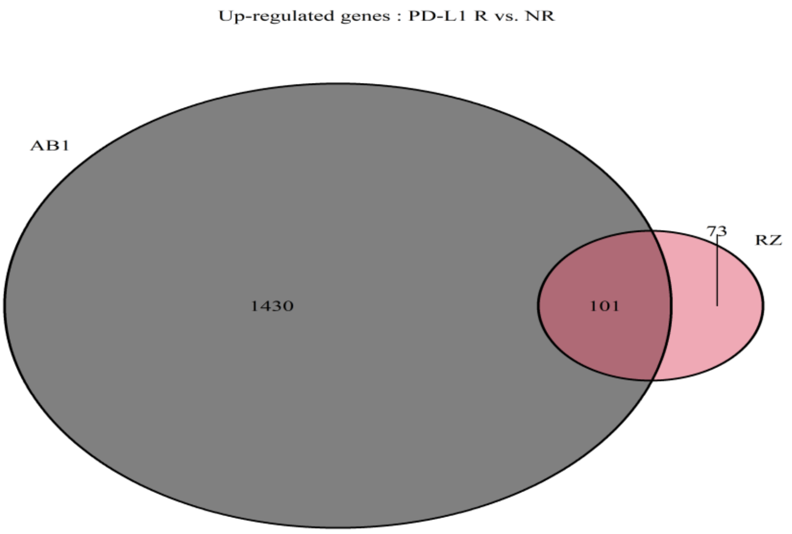
Venny (https://bioinfogp.cnb.csic.es/tools/venny/)는 실험/분석 당 얻은 각 Gene 을 엔터로 구분한 리스트를 각 그룹에 넣으면 총 4 개 그룹까지 벤다이어그램을 그려주는 웹사이트이다. 겹치는 영역을 클릭하면 해당 영역에 존재하는 gene list 를 얻을 수 있다.

R

- venn.diagram 함수를 이용하여 *.tiff 로 저장

```
### Venn diagram for DEG list
library(VennDiagram)

## Warning: package 'VennDiagram' was built under R version 4.0.3

## Loading required package: grid

## Loading required package: futile.logger

ups <- list(AB1 = rownames(res_filt_up1), RZ = rownames(res_filt_up2))
ups_list <- get.venn.partitions(ups)
venn.diagram(ups, filename="./pictures/up-regulated_genes.tiff",
             fill=c(1:2), alpha = rep(0.5,2), # transparency
             main = "Up-regulated genes : PD-L1 R vs. NR")

## [1] 1

ups_list[1, "..values.."]  # common genes
```

```
## $`1`
##   [1] "Gm15056"       "Gm18853"       "Dnase1l3"      "Acod1"        
##   [5] "Gm3756"        "Tlr12"         "RP24-499N24.4" "Cd8a"         
##   [9] "Klrc1"         "Gm5526"        "Klrc2"         "Cd8b1"        
##  [13] "AW112010"      "Ly6i"          "Ido1"          "Fam26f"       
##  [17] "Gm8451"        "Gm19585"       "Nkg7"          "Klrc3"        
##  [21] "Pdcd1"         "Tbx21"         "Gbp4"          "Sh2d2a"       
##  [25] "Gm16213"       "Gzmk"          "9130208D14Rik" "Ccl4"         
##  [29] "Trbv31"        "Muc20"         "Crtam"         "Il27"         
##  [33] "Tmem163"       "Gm17767"       "Fasl"          "Gbp11"        
##  [37] "D630039A03Rik" "Rgs8"          "Cxcl9"         "Trbv29"       
##  [41] "Gm38247"       "Serpina3g"     "Itk"           "Fam71b"       
##  [45] "Gm12791"       "Ubd"           "Ltf"           "Slc17a6"      
##  [49] "C6"            "Gbp10"         "Cxcl11"        "Gm15433"      
##  [53] "Olfr753-ps1"   "Gm28068"       "Wdr95"         "Olfr92"       
##  [57] "Gm20497"       "Chrm3"         "Gbp6"          "Gbp2b"        
##  [61] "Chil3"         "Gm12250"       "Gm43302"       "Trav7-3"      
##  [65] "Art2a-ps"      "Trbv17"        "Pla2g2d"       "Lag3"         
##  [69] "Gm44174"       "Art2b"         "Trbv15"        "Tgtp1"        
##  [73] "Cxcl10"        "Trav7d-4"      "Tgtp2"         "Cxcr6"        
##  [77] "Vhl-ps1"       "Trav16n"       "Trav7-4"       "RP23-114B10.3"
##  [81] "Il10"          "RP23-313P23.4" "Jchain"        "Nrxn3"        
##  [85] "Gm20513"       "Trav12-3"      "Gm44175"       "Igkv8-30"     
##  [89] "Igkv14-111"    "Gm38346"       "Trav7n-4"      "Igkv5-39"     
##  [93] "Gm16242"       "Trav9-4"       "Trav12-1"      "Gm20429"      
##  [97] "Trbj2-4"       "Igkv10-96"     "Gm156"         "Igkv2-109"    
## [101] "Trav8n-2"
```

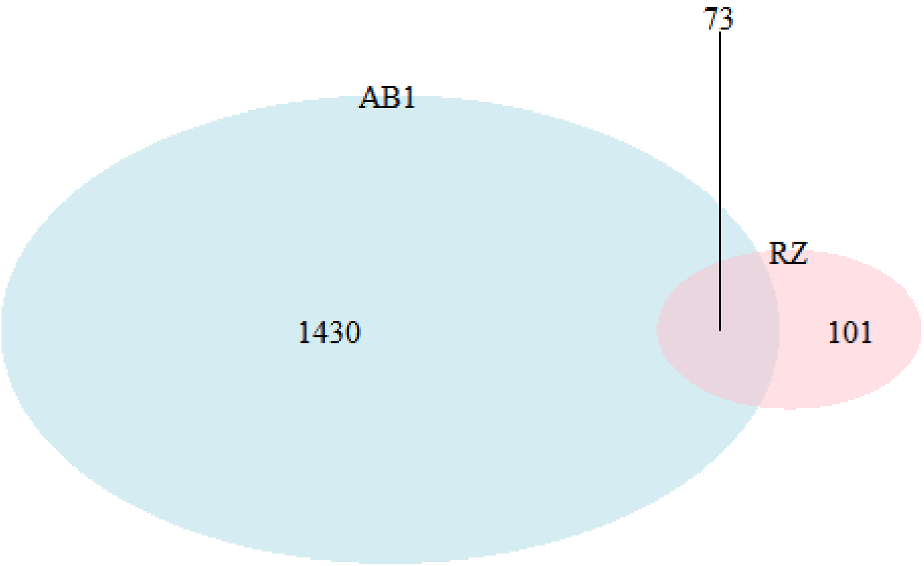


Up-regulated genes in two experiments

- draw.pairwise.venn 함수를 이용하여 *.tiff 로 저장하지 않고 바로 플롯

```
### Venn diagram for DEG list
library(VennDiagram)
ups <- list(AB1 = rownames(res_filt_up1), RZ = rownames(res_filt_up2))
ups_list <- get.venn.partitions(ups)
```

```
grid.newpage()
draw.pairwise.venn(area1 = ups_list$..count..[3]+ups_list$..count..[2],
                   area2 = ups_list$..count..[1]+ups_list$..count..[2],
                   cross.area = ups_list$..count..[2],
                   category = c("AB1", "RZ"),
                   fill = c("light blue", "pink"),
                   lty = "blank",
                   alpha = rep(0.5, 2),
                   cat.pos = c(0,0), # category label position
                   cat.dist = c(0,0)) # category label distance from circle
```

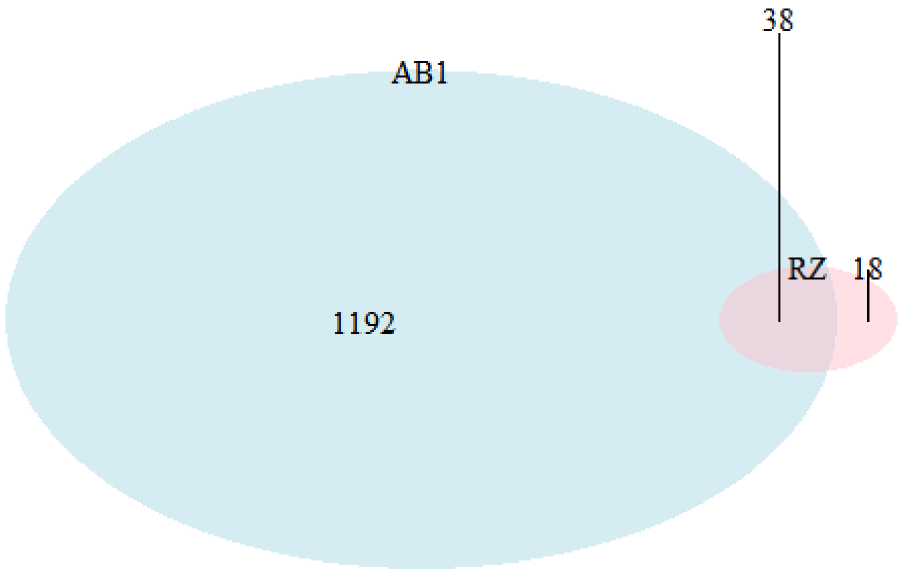


```
## (polygon[GRID.polygon.201], polygon[GRID.polygon.202], polygon[GRID.pol
ygon.203], polygon[GRID.polygon.204], text[GRID.text.205], text[GRID.text.
206], text[GRID.text.207], lines[GRID.lines.208], text[GRID.text.209], tex
t[GRID.text.210])
```

```
ups_list[1, "..values.."]  # common up-regulated genes
```

```
## $`1`
##   [1] "Gm15056"       "Gm18853"       "Dnase1l3"      "Acod1"
##   [5] "Gm3756"        "Tlr12"         "RP24-499N24.4" "Cd8a"
##   [9] "Klrc1"         "Gm5526"        "Klrc2"         "Cd8b1"
##  [13] "AW112010"      "Ly6i"          "Ido1"          "Fam26f"
##  [17] "Gm8451"        "Gm19585"       "Nkg7"          "Klrc3"
##  [21] "Pdcd1"         "Tbx21"         "Gbp4"          "Sh2d2a"
##  [25] "Gm16213"       "Gzmk"          "9130208D14Rik" "Ccl4"
##  [29] "Trbv31"        "Muc20"         "Crtam"         "Il27"
##  [33] "Tmem163"       "Gm17767"       "Fasl"          "Gbp11"
##  [37] "D630039A03Rik" "Rgs8"          "Cxcl9"         "Trbv29"
##  [41] "Gm38247"       "Serpina3g"     "Itk"           "Fam71b"
##  [45] "Gm12791"       "Ubd"           "Ltf"           "Slc17a6"
##  [49] "C6"            "Gbp10"         "Cxcl11"        "Gm15433"
##  [53] "Olfr753-ps1"   "Gm28068"       "Wdr95"         "Olfr92"
##  [57] "Gm20497"    "Chrm3"         "Gbp6"       "Gbp2b"        
##  [61] "Chil3"      "Gm12250"       "Gm43302"    "Trav7-3"      
##  [65] "Art2a-ps"   "Trbv17"        "Pla2g2d"    "Lag3"         
##  [69] "Gm44174"    "Art2b"         "Trbv15"     "Tgtp1"        
##  [73] "Cxcl10"     "Trav7d-4"      "Tgtp2"      "Cxcr6"        
##  [77] "Vhl-ps1"    "Trav16n"       "Trav7-4"    "RP23-114B10.3"
##  [81] "Il10"       "RP23-313P23.4" "Jchain"     "Nrxn3"        
##  [85] "Gm20513"    "Trav12-3"      "Gm44175"    "Igkv8-30"     
##  [89] "Igkv14-111" "Gm38346"       "Trav7n-4"   "Igkv5-39"     
##  [93] "Gm16242"    "Trav9-4"       "Trav12-1"   "Gm20429"      
##  [97] "Trbj2-4"    "Igkv10-96"     "Gm156"      "Igkv2-109"    
## [101] "Trav8n-2"
```

```
downs <- list(AB1 = rownames(res_filt_down1), RZ = rownames(res_filt_down
2))
downs_list <- get.venn.partitions(downs)

grid.newpage()
draw.pairwise.venn(area1 = downs_list$..count..[3]+downs_list$..count..
[2],
                  area2 = downs_list$..count..[1]+downs_list$..count..[2],
                  cross.area = downs_list$..count..[2],
                  category = c("AB1", "RZ"),
                  fill = c("light blue", "pink"),
                  lty = "blank",
                  alpha = rep(0.5, 2),
                  cat.pos = c(0,0), # category label position
                  cat.dist = c(0,0)) # category label distance from circle
```

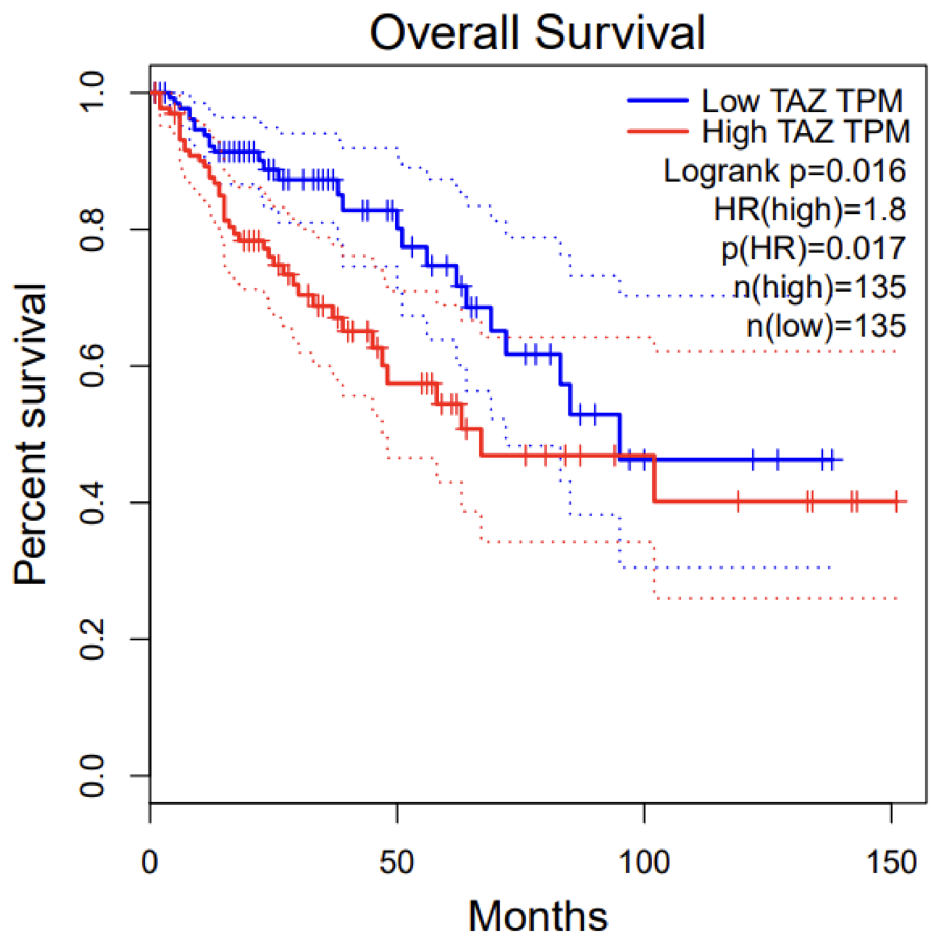


```
## (polygon[GRID.polygon.211], polygon[GRID.polygon.212], polygon[GRID.pol
ygon.213], polygon[GRID.polygon.214], text[GRID.text.215], text[GRID.text.
216], lines[GRID.lines.217], text[GRID.text.218], lines[GRID.lines.219], t
ext[GRID.text.220], text[GRID.text.221])
```

```
downs_list[1,"..values.."]  # common down-regulated genes

## $`1`
##  [1] "Krt77"         "Gm11808"      "Lor"          "Krt5"
##  [5] "Krtdap"        "Flg2"         "Krt15"        "Spink5"
##  [9] "Psapl1"        "Elovl4"       "Gm7429"       "Gm7993"
## [13] "Gm7816"        "Mucl1"        "Gm11949"      "Rpl35a-ps2"
## [17] "C130079G13Rik" "Olfr111"
```

Survival plot

위 과정을 통해 얻은 DEGs 에 대한 clinical validation 을 위해 reference dataset 에서 유전자의 발현 정도에 따른 survival plot 을 그리려고 한다. survival plot 의 기본 개념을 먼저 소개한다.



An example survival plot

Survival plot 에 대한 기본적인 개념은 다음과 같다. Survival plot 은 특정 유전자가 up-regulated 된 그룹과 down-regulated 된 그룹의 시간(Months)에 따른 생존률을

비교하는 정보를 담고 있다. 생존기간을 정의하는 방법에 따라 전체생존율 (Overall survival, OS)과 무질병생존율 (Disease free survival, DFS)로 구분할 수 있다.

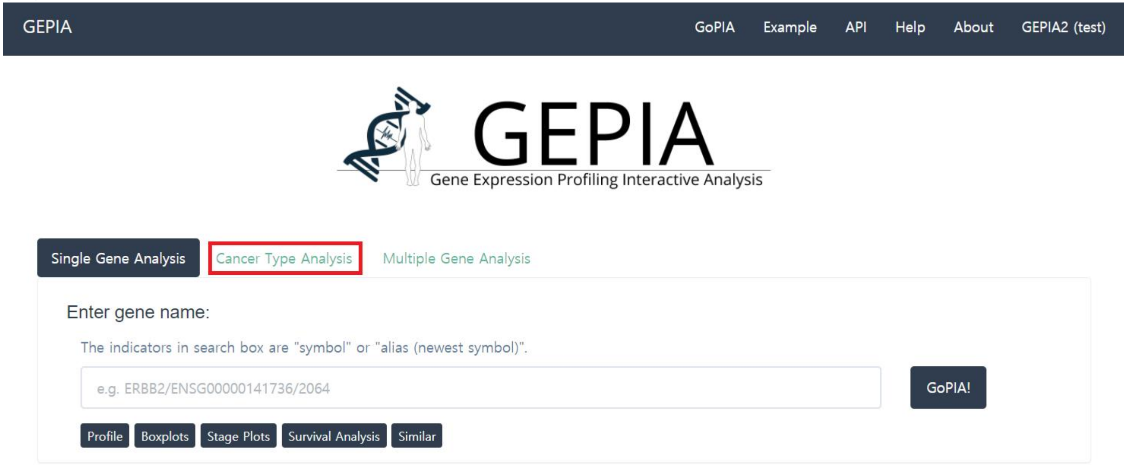
- OS: 조직학적 진단을 시행한 시점에서 마지막 추적관찰 시기까지의 기간
- DFS: 조직학적 진단을 시행한 시점에서 이후 재발, 진행, 사망까지의 기간 혹은 재발이나 진행 없이 마지막 추적관찰 시기까지의 기간

값을 읽는 방법은 x 축의 시간에 따른 y 축의 probability of survivals 값의 추이를 보고 집단 간의 비교를 하는 것이다. 일정 시간이 지난 뒤 높은 값을 유지하는 집단의 특성이 반대 집단의 특성에 비해 생존에 유리하다고 할 수 있다.

survival plot 상에서의 높이 차이도 중요하지만 해당 결과의 통계적 유의성을 확보하기 위해서는 Logrank p 값이 0.05 또는 0.01 미만의 기준을 만족하는지 확인해야 한다.

이제, reference dataset 을 보유하고 있는 웹사이트에서 또는, R 에서 직접 원하는 데이터를 불러와서 survival plot 을 그리는 방법을 각각 소개한다.

gepia



Enter gene name:

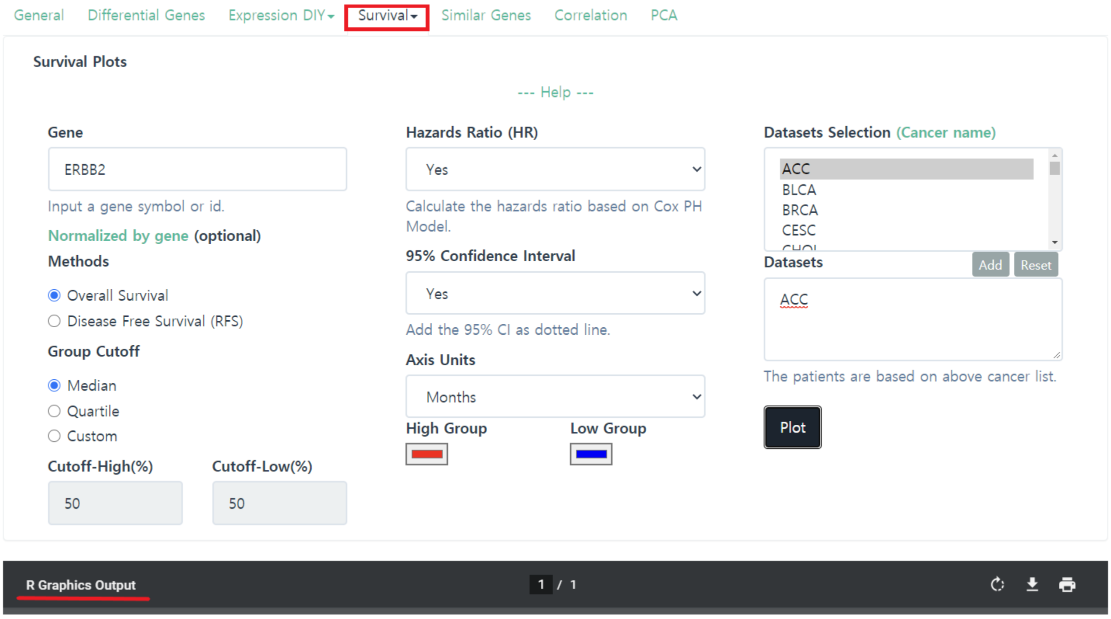
The indicators in search box are "symbol" or "alias (newest symbol)".

e.g. ERBB2/ENSG00000141736/2064

gepia main

gepia (http://gepia.cancer-pku.cn/)는 survival plot 뿐만 아니라 다양한 visualization 이 가능한 웹사이트다. Reference dataset 으로 암환자 샘플의 경우 TCGA, 정상 환자 샘플의 경우 Genotype-Tissue Expression (GTEx, https://gtexportal.org/home/)를 사용한다.

survival plot 을 그리려면 **Cancer Type Analysis > Differential genes analysis > Survival** 순으로 탭을 클릭한 뒤, 원하는 Gene 과 암종의 dataset 을 `Add` 하여 `Plot` 하면 된다. 웹사이트 하단에 R graphics output 으로 survival plot 이 출력된다.



Survival Plots

Gene
ERBB2
Input a gene symbol or id.
Normalized by gene (optional)

Methods
● Overall Survival
○ Disease Free Survival (RFS)

Group Cutoff
● Median
○ Quartile
○ Custom

Cutoff-High(%) 50
Cutoff-Low(%) 50

Hazards Ratio (HR)
Yes
Calculate the hazards ratio based on Cox PH Model.

95% Confidence Interval
Yes
Add the 95% CI as dotted line.

Axis Units
Months

High Group 🟥
Low Group 🟦

Datasets Selection (Cancer name)
ACC
BLCA
BRCA
CESC
CHOL

Datasets
ACC
The patients are based on above cancer list.

Survival plot

주요 기능

- Survival plot 이외에도 boxplot 등의 다양한 visualization 가능
- multiple genes 에 대한 survival plot 도 지원
- gepia2 에서 Python API 를 제공하여 batch processing (여러 개의 결과물을 일괄적으로 처리) 가능

Python API for gepia2

gepia 에서는 인풋을 지정할 때의 반복적인 수작업을 줄이기 위해 Python 3 에서 동작하는 API 를 구현해놓았다. Python 3 유저는 Python terminal 을 통해 실행하면 되지만 R 과 RStudio 유저를 위해 python script (*.py)를 RStudio 에서 실행하는 방향으로 설명한다.

5. 우선 다음 사이트에서 파이썬 3 버전을 설치한다.
   https://www.python.org/downloads/

설치가 제대로 되었는지 확인하는 방법은 윈도우 cmd 또는 맥 터미널을 열어 다음을 입력하고 설치한 파이썬 버전이 출력되는지를 보는 것이다. 설치가 안되었다면 해당 코드가 실행되지 않을 것이다.

```
python --version
```

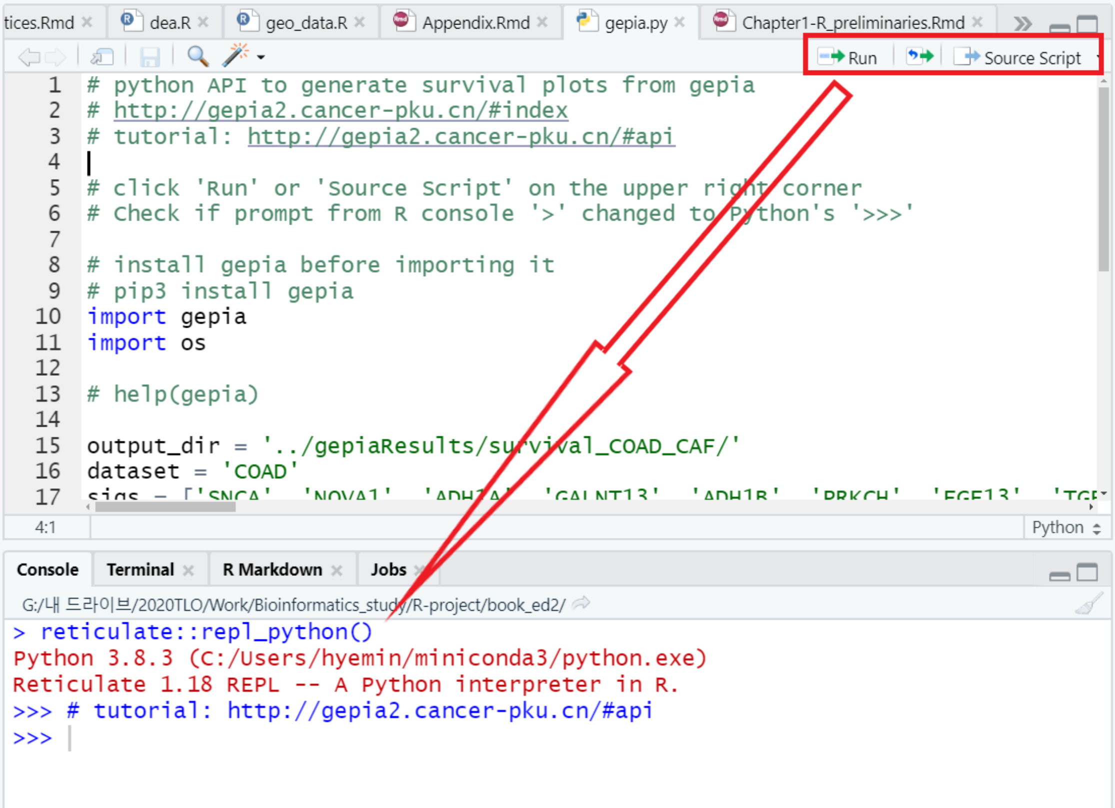
6. 곧바로 터미널 창에서 gepia api 를 설치한다.

```
pip install gepia
```

5. RStudio 에서 function/gepia.py 을 열고 Lines 15-18 의 **output_dir (출력물을 담을 폴더/없으면 현재 폴더), dataset(암종 dataset), sigs (survival plot 을 그리기 위한 유전자의 목록) 설정을 수정한다. 유전자 개별적으로 survival plot 을 그리려면 [] 안에 각 유전자를 ,로 구분하여 넣는다. 여러 유전자가 동시에 up-regulated 또는 down-regulated 되었을 때를 비교하여 그리려면 [] 안에 [] 하나 더 넣어 그 안에 같은 그룹의 유전자들을 적는다. ex) sigs = [[group1 의 유전자들], [group2 의 유전자들]]

```
output_dir = './' #../gepiaResults/survival_COAD_CAF/'
dataset = 'COAD'
sigs = ['SNCA', 'NOVA1', 'ADH1A', 'GALNT13', 'ADH1B', 'PRKCH', 'FGF13', 'TG
FBR3', 'CFD', 'ADORA1', 'ATP8B4', 'PBX1', 'AKR1C3', 'AKR1C2', 'SMPDL3A', 'T
CF21', 'ATOH8', 'GREM2', 'MASP1', 'PPARG', 'FGFR4', 'METTL7A', 'CACNB2', 'S
HC3', 'LIPG', 'CLIC6', 'TRPC6', 'PTGER2', 'SLIT3', 'ITPR3', 'ZNF536', 'IL33
', 'PF4V1', 'ISLR', 'CYTL1', 'COL14A1', 'GALNT15', 'STK32B', 'TMEM119', 'ED
NRB', 'MOXD1', 'IL13RA2', 'S100A4', 'IMPA2', 'PLPP3', 'PTHLH', 'MMP3', 'FBN
2', 'CXCL12', 'RASSF2', 'SHISA3', 'CCL13', 'STEAP1', 'SST', 'TFPI2', 'CCL8
', 'RSPO3']
# sigs = [], ['ADH1A', 'GALNT13', 'ADH1B'], ]    # for multiple genes resul
t
```

4. python console 로 넘어가기 위해 오른쪽 상단의 Run' 이나 'Source Script' 클릭한다. 제대로 실행되었다면 R console 의 프롬프트 '>'가 Python 의'>>>'로 바뀐 것 확인할 수 있다.



```
# python API to generate survival plots from gepia
# http://gepia2.cancer-pku.cn/#index
# tutorial: http://gepia2.cancer-pku.cn/#api

# click 'Run' or 'Source Script' on the upper right corner
# Check if prompt from R console '>' changed to Python's '>>>'

# install gepia before importing it
# pip3 install gepia
import gepia
import os

# help(gepia)

output_dir = '../gepiaResults/survival_COAD_CAF/'
dataset = 'COAD'
sigs = ['SNCA', 'NOVA1', 'ADH1A', 'GALNT13', 'ADH1B', 'PRKCH', 'FGF13', 'TGF
```

```
G:/내 드라이브/2020TLO/Work/Bioinformatics_study/R-project/book_ed2/
> reticulate::repl_python()
Python 3.8.3 (C:/Users/hyemin/miniconda3/python.exe)
Reticulate 1.18 REPL -- A Python interpreter in R.
>>> # tutorial: http://gepia2.cancer-pku.cn/#api
>>> 
```

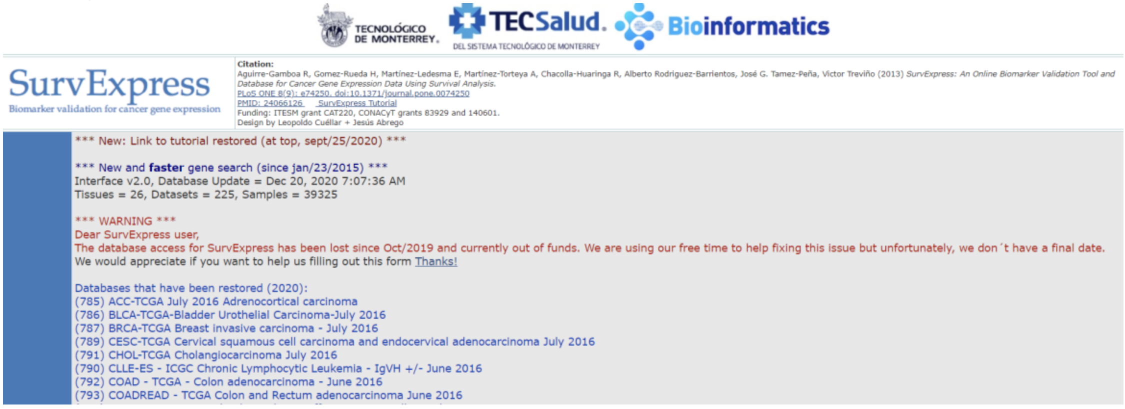
RStudio 에서 Python 스크립트 gepia.py 실행

function/gepia.py 안의 내용은 아래 코드와 같다. 지정한 디렉토리에 개별 gene 또는 genes group 에 대한 survival plot 을 OS, DFS 각각 하나씩 그려 저장한다.

```
# python API to generate survival plots from gepia
# http://gepia2.cancer-pku.cn/#index
# tutorial: http://gepia2.cancer-pku.cn/#api

# click 'Run' or 'Source Script' on the upper right corner
# Check if prompt from R console '>' changed to Python's '>>>'

# install gepia before importing it
# pip3 install gepia
import gepia
import os

# help(gepia)

output_dir = './' #../gepiaResults/survival_COAD_CAF/'
dataset = 'COAD'
sigs = ['SNCA', 'NOVA1', 'ADH1A', 'GALNT13', 'ADH1B', 'PRKCH', 'FGF13', 'TG
FBR3', 'CFD', 'ADORA1', 'ATP8B4', 'PBX1', 'AKR1C3', 'AKR1C2', 'SMPDL3A', 'T
CF21', 'ATOH8', 'GREM2', 'MASP1', 'PPARG', 'FGFR4', 'METTL7A', 'CACNB2', 'S
HC3', 'LIPG', 'CLIC6', 'TRPC6', 'PTGER2', 'SLIT3', 'ITPR3', 'ZNF536', 'IL33
', 'PF4V1', 'ISLR', 'CYTL1', 'COL14A1', 'GALNT15', 'STK32B', 'TMEM119', 'ED
NRB', 'MOXD1', 'IL13RA2', 'S100A4', 'IMPA2', 'PLPP3', 'PTHLH', 'MMP3', 'FBN
2', 'CXCL12', 'RASSF2', 'SHISA3', 'CCL13', 'STEAP1', 'SST', 'TFPI2', 'CCL8
', 'RSPO3']
# sigs = [], ['ADH1A', 'GALNT13', 'ADH1B'], ]    # for multiple genes resul
t

os.makedirs(output_dir, exist_ok=True)

# survival plot
sv=gepia.survival()
sv.showParams()
sv.setOutDir(output_dir)
sv.setParam('dataset',dataset)

for sig in sigs:
    sv.setParam('signature',sig)

    for method in ['os', 'dfs']:
      sv.setParam('methodoption', method)
      sv.query()

# terminate Python console by typing:
# quit or `Esc` key
```

파이썬 콘솔을 빠져나오기 위해서는 Esc 키를 한번 누르거나 콘솔창에 quit 을 타이핑한다. Python console 의 프롬프트 '>>>'에서 R 의'>'로 바뀐 것을 확인할 수 있다.

SurvExpress



TECNOLÓGICO DE MONTERREY. TECSalud. DEL SISTEMA TECNOLÓGICO DE MONTERREY Bioinformatics

SurvExpress
Biomarker validation for cancer gene expression

Citation:
Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaringa R, Alberto Rodriguez-Barrientos, José G. Tamez-Peña, Victor Treviño (2013) *SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis.*
PLoS ONE 8(9): e74250. doi:10.1371/journal.pone.0074250
PMID: 24066126   SurvExpress Tutorial
Funding: ITESM grant CAT220, CONACyT grants 83929 and 140601.
Design by Leopoldo Cuéllar + Jesús Abrego

*** New: Link to tutorial restored (at top, sept/25/2020) ***

*** New and **faster** gene search (since jan/23/2015) ***
Interface v2.0, Database Update = Dec 20, 2020 7:07:36 AM
Tissues = 26, Datasets = 225, Samples = 39325

*** WARNING ***
Dear SurvExpress user,
The database access for SurvExpress has been lost since Oct/2019 and currently out of funds. We are using our free time to help fixing this issue but unfortunately, we don´t have a final date.
We would appreciate if you want to help us filling out this form Thanks!

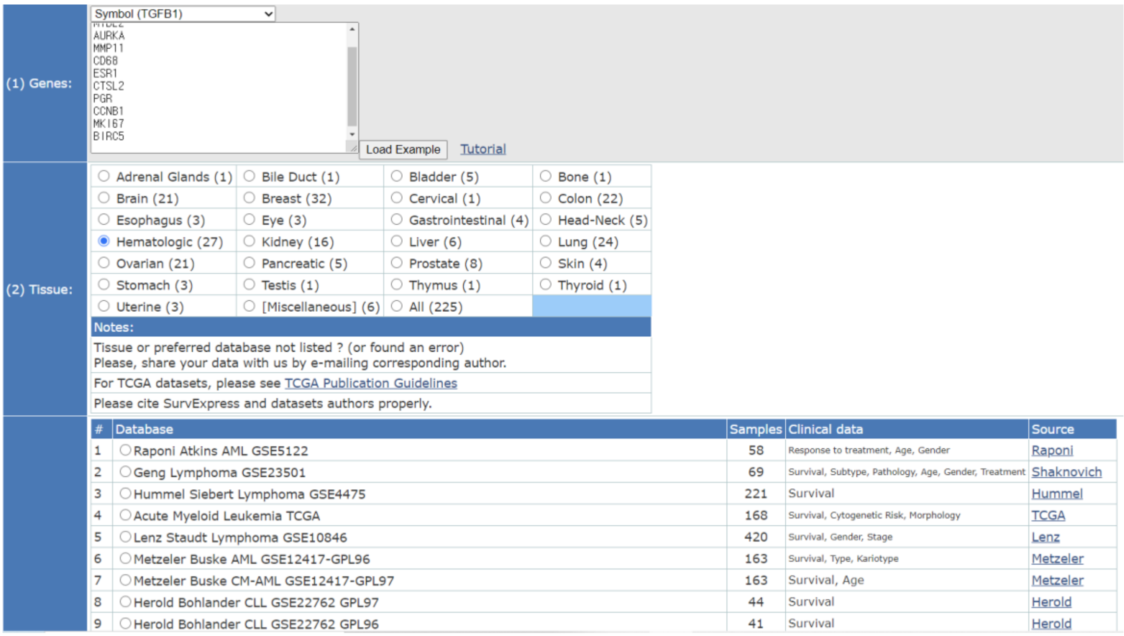
Databases that have been restored (2020):
(785) ACC-TCGA July 2016 Adrenocortical carcinoma
(786) BLCA-TCGA-Bladder Urothelial Carcinoma-July 2016
(787) BRCA-TCGA Breast invasive carcinoma - July 2016
(789) CESC-TCGA Cervical squamous cell carcinoma and endocervical adenocarcinoma July 2016
(791) CHOL-TCGA Cholangiocarcinoma July 2016
(790) CLLE-ES - ICGC Chronic Lymphocytic Leukemia - IgVH +/- June 2016
(792) COAD - TCGA - Colon adenocarcinoma - June 2016
(793) COADREAD - TCGA Colon and Rectum adenocarcinoma June 2016

SurvExpress main

SurvExpress (http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp)는 암종에 따라 TCGA data 뿐만 아니라 GEO data 등 다수의 데이터셋을 reference dataset 으로 두고 있다. 하나의 dataset 에서 원하는 결과가 나오지 않을 경우,

dataset 을 바꿔가며 결과가 나오는지 확인할 수 있다. 서로 다른 dataset 끼리 합치는 기능은 없다.

사용방법은 그림에 나오는 순서대로 설정을 입력한 뒤 `SurvExpress Analysis` 를 클릭하는 것이다.



(1) Genes: Symbol (TGFB1)

AURKA
MMP11
CD68
ESR1
CTSL2
PGR
CCNB1
MKI67
BIRC5

(2) Tissue:

○ Adrenal Glands (1)	○ Bile Duct (1)	○ Bladder (5)	○ Bone (1)
○ Brain (21)	○ Breast (32)	○ Cervical (1)	○ Colon (22)
○ Esophagus (3)	○ Eye (3)	○ Gastrointestinal (4)	○ Head-Neck (5)
● Hematologic (27)	○ Kidney (16)	○ Liver (6)	○ Lung (24)
○ Ovarian (21)	○ Pancreatic (5)	○ Prostate (8)	○ Skin (4)
○ Stomach (3)	○ Testis (1)	○ Thymus (1)	○ Thyroid (1)
○ Uterine (3)	○ [Miscellaneous] (6)	○ All (225)	

Notes:

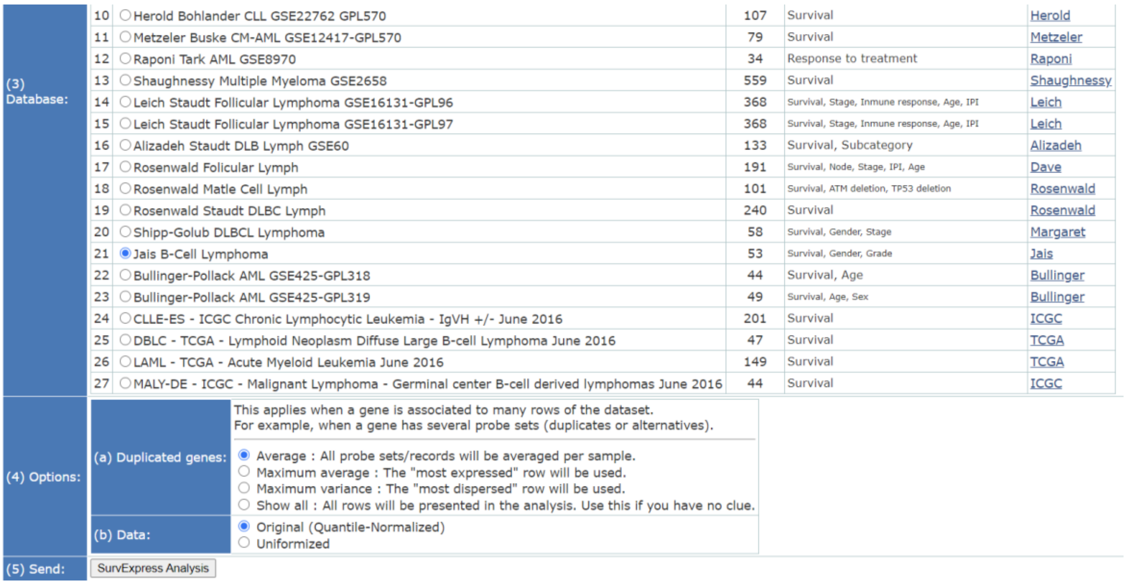
Tissue or preferred database not listed ? (or found an error)
Please, share your data with us by e-mailing corresponding author.

For TCGA datasets, please see TCGA Publication Guidelines

Please cite SurvExpress and datasets authors properly.

#	Database	Samples	Clinical data	Source
1	○ Raponi Atkins AML GSE5122	58	Response to treatment, Age, Gender	Raponi
2	○ Geng Lymphoma GSE23501	69	Survival, Subtype, Pathology, Age, Gender, Treatment	Shaknovich
3	○ Hummel Siebert Lymphoma GSE4475	221	Survival	Hummel
4	○ Acute Myeloid Leukemia TCGA	168	Survival, Cytogenetic Risk, Morphology	TCGA
5	○ Lenz Staudt Lymphoma GSE10846	420	Survival, Gender, Stage	Lenz
6	○ Metzeler Buske AML GSE12417-GPL96	163	Survival, Type, Kariotype	Metzeler
7	○ Metzeler Buske CM-AML GSE12417-GPL97	163	Survival, Age	Metzeler
8	○ Herold Bohlander CLL GSE22762 GPL97	44	Survival	Herold
9	○ Herold Bohlander CLL GSE22762 GPL96	41	Survival	Herold

SurvExpress setting 1



(3) Database:

#	Database	Samples	Clinical data	Source
10	○ Herold Bohlander CLL GSE22762 GPL570	107	Survival	Herold
11	○ Metzeler Buske CM-AML GSE12417-GPL570	79	Survival	Metzeler
12	○ Raponi Tark AML GSE8970	34	Response to treatment	Raponi
13	○ Shaughnessy Multiple Myeloma GSE2658	559	Survival	Shaughnessy
14	○ Leich Staudt Follicular Lymphoma GSE16131-GPL96	368	Survival, Stage, Inmune response, Age, IPI	Leich
15	○ Leich Staudt Follicular Lymphoma GSE16131-GPL97	368	Survival, Stage, Inmune response, Age, IPI	Leich
16	○ Alizadeh Staudt DLB Lymph GSE60	133	Survival, Subcategory	Alizadeh
17	○ Rosenwald Folicular Lymph	191	Survival, Node, Stage, IPI, Age	Dave
18	○ Rosenwald Matle Cell Lymph	101	Survival, ATM deletion, TP53 deletion	Rosenwald
19	○ Rosenwald Staudt DLBC Lymph	240	Survival	Rosenwald
20	○ Shipp-Golub DLBCL Lymphoma	58	Survival, Gender, Stage	Margaret
21	● Jais B-Cell Lymphoma	53	Survival, Gender, Grade	Jais
22	○ Bullinger-Pollack AML GSE425-GPL318	44	Survival, Age	Bullinger
23	○ Bullinger-Pollack AML GSE425-GPL319	49	Survival, Age, Sex	Bullinger
24	○ CLLE-ES - ICGC Chronic Lymphocytic Leukemia - IgVH +/- June 2016	201	Survival	ICGC
25	○ DBLC - TCGA - Lymphoid Neoplasm Diffuse Large B-cell Lymphoma June 2016	47	Survival	TCGA
26	○ LAML - TCGA - Acute Myeloid Leukemia June 2016	149	Survival	TCGA
27	○ MALY-DE - ICGC - Malignant Lymphoma - Germinal center B-cell derived lymphomas June 2016	44	Survival	ICGC

(4) Options:

(a) Duplicated genes:

This applies when a gene is associated to many rows of the dataset.
For example, when a gene has several probe sets (duplicates or alternatives).

● Average : All probe sets/records will be averaged per sample.
○ Maximum average : The "most expressed" row will be used.
○ Maximum variance : The "most dispersed" row will be used.
○ Show all : All rows will be presented in the analysis. Use this if you have no clue.

(b) Data:

● Original (Quantile-Normalized)
○ Unformized

(5) Send:

SurvExpress setting 2

R

geo data 와 tcga data 에 대해 survival plot 을 그린 예제를 참고하자. 두 예제 모두 하나의 gene 에 대해 여러 개의 probe index 가 조회될 경우 평균값을 취했고, median expression level 보다 큰 샘플을 Up-regulated 로, 작은 샘플을 Down-regulated 로 분류하였다.

- geo data (/CAF_validation/CaF_validation_script.R)

해당 script 의 17 번째 줄의 gene_of_interest 를 수정하여 사용한다. 또한, annot_data 의 양식이 통일되어있지 않고 저마다 다르기때문에 `head(annot_data)`를 통해 내용물을 열어보고 필요한 변수(Death, OverallSurvival_months, TumorFreeSurvival_months)를 뽑아 데이터 프레임에 담는 전처리 과정이 필요하다.

```r
source("../functions/geo_data.R")
geo_series_idx <- "gse12945"
gse <- download_gse(geo_series_idx) # geo data 로드
data <- extract_gse(gse, "../geo_Rdata", geo_series_idx) # data 는 exprs_mat, gene_info, annot_data 를 담고있는 리스트
attach(data)

## Extract real clinical data : gse file 마다 다른 포맷일 수도 있음
head(annot_data, 3)
colnames(annot_data) <- gsub(":ch1$", "", colnames(annot_data))  #열 이름에서 :ch1 로 끝나는 부분 삭제

library(dplyr)
annot_data <- annot_data %>% select(Death, OverallSurvival_months, TumorFreeSurvival_months)
annot_data$Death <- as.integer(annot_data$Death)
annot_data$OverallSurvival_months <- as.numeric(annot_data$OverallSurvival_months)

# Survival plot
gene_of_interest <- "GLI2"

library(survival)
library(survminer)
gene_idx <- which(gene_info$`Gene Symbol` == gene_of_interest)
# median expression level whenever multiple probe ids exist
if (length(gene_idx) >1) {
  gene_exprs_level <- colMeans(exprs_mat[gene_idx,])
} else {
  gene_exprs_level <- exprs_mat[gene_idx,]
}
summary(gene_exprs_level)
over_exprs_level <- median(gene_exprs_level)
gp_idx <- ifelse(gene_exprs_level>= over_exprs_level, "up", "down")
surv1 <- survfit(Surv(OverallSurvival_months,Death)~gp_idx, data=annot_dat
a)
ggsurvplot(surv1, data=annot_data, pval=TRUE, risk.table=T, title=gene_of_
interest)
```

- TCGA data (/TCGA_validation/CaF_validation_script.R)

다운받는 과정이 길기 때문에 코드 결과 출력은 생략한다.

```r
## Clinical validation using TCGA data
## reference: https://costalab.ukaachen.de/open_data/Bioinformatics_Analys
is_in_R_2019/BIAR_D3/handout.html
## Hyemin Gu (nicolegu6616@gmail.com)
## ver 1.0 (2020-11-24)

## 1. Search and download TCGA data
library(TCGAbiolinks)
GDCprojects = getGDCprojects()
View(GDCprojects[c("project_id", "name")])  # select a project
TCGAbiolinks:::getProjectSummary("TCGA-OV") # further look on a project

## ===== MODIFY THESE LINE BY YOUR NEED ======
#rm(list=ls()) # if you need to clear up the environment
CancerProject <- "TCGA-OV" # pick a project
# 2 settings to fetch all RNA-seq data
DataCategory <- "Transcriptome Profiling"
DataType <- "Gene Expression Quantification"
ExpStragegy <- "RNA-Seq"
WorkFlowType <- "HTSeq - Counts"


## ===========================================
query = GDCquery(
  project = CancerProject,
  data.category = DataCategory, # parameter enforced by GDCquery
  data.type = DataType,
  experimental.strategy = ExpStragegy,
  workflow.type = WorkFlowType)
GDCdownload(query = query, directory = "../GDCdata")

# 2. Load and obtain exprs_mat, gene_info, annot_data
data = GDCprepare(query, directory = "../GDCdata")
dim(data)
## 3 main functions to access the data: colData(), rowData(), assays()
library(SummarizedExperiment)
colnames(colData(data))
```

```
table(data@colData$vital_status)
table(data@colData$tumor_stage)
table(data@colData$definition)
table(data@colData$tissue_or_organ_of_origin)
table(data@colData$gender)

head(rowData(data))

dim(assay(data))
head(assay(data)[,1:5])

annot_data <- colData(data)
gene_info <- rowData(data)
exprs_mat <- assay(data)

## 3. Survival Analysis
library(survival)
library(survminer)

gene_info$external_gene_name[grep("^CD2", gene_info$external_gene_name)]
## === SPECIFY THESE LINES ===
gene_of_interest <- "CD274"

annot_data$Death <- ifelse(annot_data$vital_status=="Dead", 1, 0)
annot_data$OverallSurvival_months = ifelse(
  annot_data$Death,
  annot_data$days_to_death/30,
  annot_data$days_to_last_follow_up/30)
## ==========================
gene_idx <- which(gene_info$external_gene_name == gene_of_interest)

# median expression level whenever multiple probe ids exist
if (length(gene_idx) >1) {
  gene_exprs_level <- colMeans(exprs_mat[gene_idx,])
} else {
  gene_exprs_level <- exprs_mat[gene_idx,]
}
summary(gene_exprs_level)
over_exprs_level <- median(gene_exprs_level)

gp_idx <- ifelse(gene_exprs_level>= over_exprs_level, "Up", "Down")
surv1 <- survfit(Surv(OverallSurvival_months,Death)~gp_idx, data=annot_dat
a)
ggsurvplot(surv1, data=annot_data, pval=TRUE, risk.table=T, title=gene_of_
interest)
```

Chapter 3: DEA practices

Hyemin Gu

2020-12-18

Table of Contents

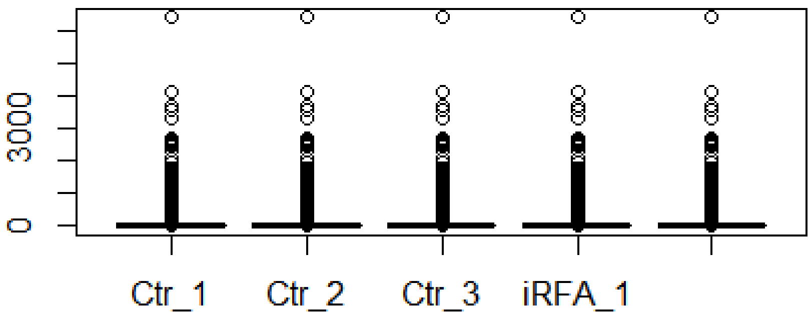
R 로 분석한 실습 예제

DEA on Microarray data

GSE138224 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138224)에 대한 DEA

```
# download excel exprs_mat directly from
# https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138224
# then combine them by columns and save as /geo_exprs_mat/GSE138224.csv
# ensemble gene id -> gene symbol :
https://www.biotools.fr/mouse/ensembl_symbol_converter
rm(list = ls())
```

```
exprs_mat <- read.csv("../geo_data/GSE138224.csv")
head(exprs_mat)

##                   id    Ctr_1    Ctr_2    Ctr_3   iRFA_1   iRFA_2   iRFA_3
## 1 ENSMUSG00000028180 11.19866 10.77352 16.16627 15.70104 15.86787
14.66599
## 2 ENSMUSG00000028182  0.30710  0.22366  0.28851  0.30798  0.35380
0.29713
## 3 ENSMUSG00000028185  0.02041  0.00000  0.00000  0.00000  0.00000
0.00000
## 4 ENSMUSG00000028184 11.25099 11.95978  2.90199  2.34468  2.94021
2.70212
## 5 ENSMUSG00000028187  5.01563  5.19307 13.65286 14.82518 15.28217
13.92727
## 6 ENSMUSG00000028186  0.03412  0.05682  0.00000  0.02069  0.02354
0.00000
##    symbol
## 1  Zranb2
## 2  Lrriq3
## 3 Dnase2b
## 4  Adgrl2
## 5    Rpf1
## 6     Uox

exprs_mat <- exprs_mat[!duplicated(exprs_mat$symbol),]
rownames(exprs_mat) <- exprs_mat$symbol
exprs_mat <- exprs_mat[,-c(1, 8)]

dim(exprs_mat)

## [1] 32041     6

# preprocessing
source("../functions/preprocess_expression_mat.R")
exprs_mat <- as.matrix(exprs_mat, rownames=T)

exprs_mat <- qnormalize(exprs_mat)
```



```
exprs_mat <- qfilter(exprs_mat, 0.25)
min(exprs_mat)

## [1] 0

dim(exprs_mat)

## [1] 24030     6

# dea
source("../functions/dea.R")
ctr <- exprs_mat[, 1:3]
iRFA <- exprs_mat[, 4:6]
res_filt_up <- analyze_DEG(iRFA, ctr, "P.Value < 0.05 & logFC>=1")
head(res_filt_up, 10)

##                logFC    AveExpr         t      P.Value adj.P.Val          B
## Slamf9     7.351666  17.070038 14.879696 4.247157e-05 0.3659394
0.19109663
## Gzmb      10.779201  12.867411 14.495852 4.785238e-05 0.3659394
0.16902226
## Tmem170b   2.143460   4.990642 13.646859 6.300026e-05 0.3659394
0.11437692
## Dnaja1     6.149525  25.069759 13.345288 6.974151e-05 0.3659394
0.09278967
## Prpf38a    2.105224   6.321256 10.874896 1.760756e-04 0.3659394 -
0.14261058
## Inpp4b     2.909138   2.552922 10.485438 2.074231e-04 0.3659394 -
0.19224961
## Snx8       7.521916  16.550237 10.358951 2.190273e-04 0.3659394 -
0.20931866
## Fam122b    2.252108   2.523631  9.629378 3.036122e-04 0.3659394 -
0.31797210
## Eif4a-ps4 22.157794 140.639316  9.340175 3.477713e-04 0.3659394 -
0.36641085
```

```
## Bcl2a1d    5.651290  11.691813  9.276082 3.585831e-04 0.3659394 -
0.37760313

#write.csv(res_filt_up, "overexprs-pval0_01-logFC1.csv")

## heatmap
#install.packages("gplots")
#library(gplots)
#heatmap.2(exprs_mat[rownames(res_filt_up),], scale="row", Rowv = NA, Cowv
= NA, trace = "none", col=greenred(10), density.info = "none")
```

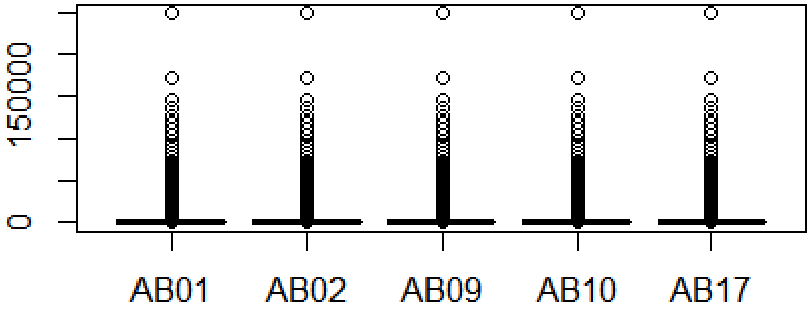
DEA on RNA-Seq count data

GSE117358 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117358)에 대한 DEA

서로 다른 treatment 에 대한 respoder vs. nonresponder 에 대한 common DEGs 확인

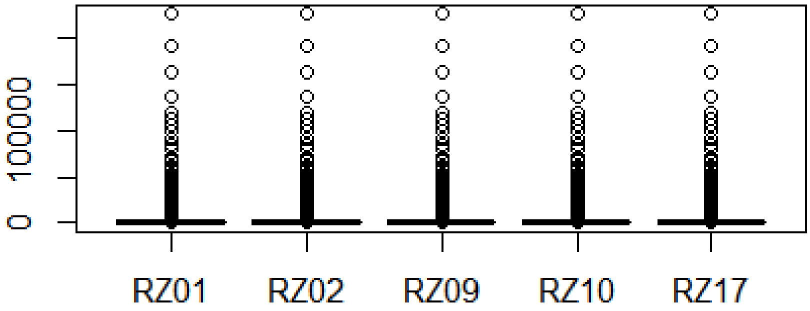
```
rm(list=ls())
gse_serial_no <- "gse117358"
gene_counts <- read.csv("../geo_data/GSE117358_genecounts.csv")
head(gene_counts[,1:10])

##   AB01  AB02 AB09  AB10 AB17 AB18 AB25  AB26 AB33  AB34
## 1 5344  6003 5392  5452 4652 4620 4510  5077 4313  6617
## 2    0     0    0     0    0    0    0     0    0     0
## 3  812   967  841   882  658  434  686   867  568  1247
## 4 8346 14646 6320 16274 5830 5422 4158 13197 4356 13387
## 5   88   142   83   128   70   75   51   107   45   143
## 6    2     3    0     0    0    5    4     1    1     5

# sample can be grouped into (AB responder, AB nonresponder, RZ responder,
RZ nonresponder)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

AB <- gene_counts %>% select(colnames(gene_counts[grep("AB",
colnames(gene_counts))]))
rownames(AB) <- gene_counts$Symbol
AB <- as.matrix(AB)
```

```
RZ <- gene_counts %>% select(colnames(gene_counts[grep("RZ",
colnames(gene_counts))]))
rownames(RZ) <- gene_counts$Symbol
RZ <- as.matrix(RZ)

source("../functions/preprocess_expression_mat.R")
######### DEG analysis on AB
# normalization of genes
AB <- qnormalize(AB)
# quantile filter of genes
AB1_Filt <- qfilter(AB, qnt.cut =  0.25)

source("../functions/dea.R")
AB1_R <- AB1_Filt[, seq(1, ncol(AB1_Filt), by=2)]
AB1_NR <- AB1_Filt[, seq(2, ncol(AB1_Filt), by=2)]
res_filt_up1 <- analyze_DEG_cnt(AB1_R, AB1_NR, "logFC > 1 & FDR < 0.01")
```

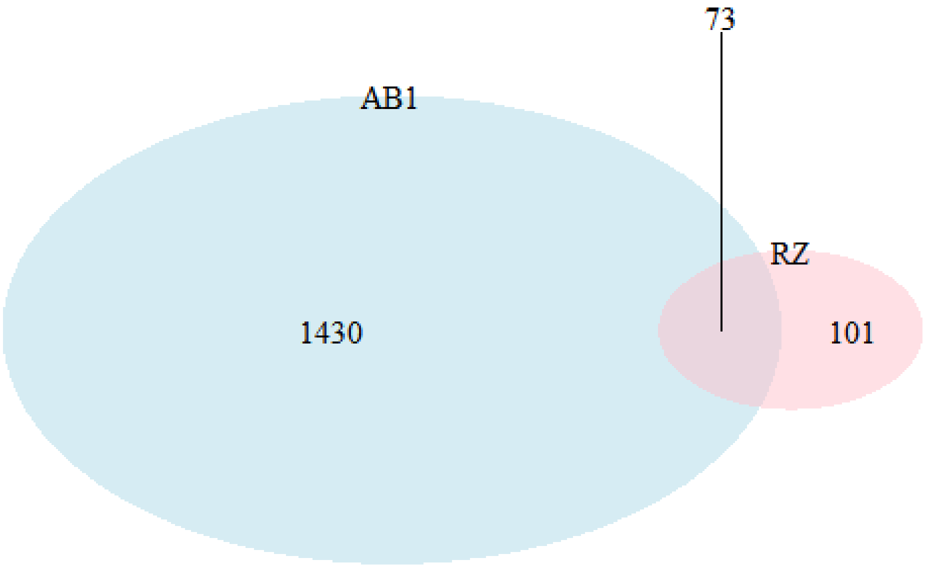


```
## Disp = 0.10439 , BCV = 0.3231

res_filt_down1 <- analyze_DEG_cnt(AB1_R, AB1_NR, "logFC < -1 & FDR <
0.01")

## Disp = 0.10439 , BCV = 0.3231

head(res_filt_up1)

##                logFC    logCPM        PValue          FDR
## Rpl36a-ps3 6.776259 3.1668831 1.269852e-132 3.535267e-128
## Tpt1-ps5   4.647967 3.5673318  4.371627e-89  1.106419e-85
## Rps11-ps3  4.960140 1.3504531  9.041029e-75  8.390075e-72
## Lcn2       3.894966 4.9180481  4.200203e-73  3.340961e-70
## Gm15056    3.755328 5.4481528  3.232379e-69  2.249736e-66
## Rpl27-ps1  4.713970 0.7153889  2.398134e-62  1.236371e-59

head(res_filt_down1)
```

```
##                  logFC   logCPM        PValue           FDR
## Krt77      -8.193637 2.119829 1.272916e-126 1.771900e-122
## Gm8623     -7.211263 2.349354 2.608129e-123 2.420344e-119
## Gm5879     -5.457276 3.661994 1.408424e-110 9.802634e-107
## Rpl31-ps13 -5.751612 2.601681 4.600136e-106 2.561356e-102
## Krtap3-2   -5.113961 3.381023 6.311631e-100  2.928597e-96
## Krt33b     -4.777194 4.420790  2.616557e-96  1.040642e-92

######### DEG analysis on RZ
# normalization of genes
RZ <- qnormalize(RZ)
```

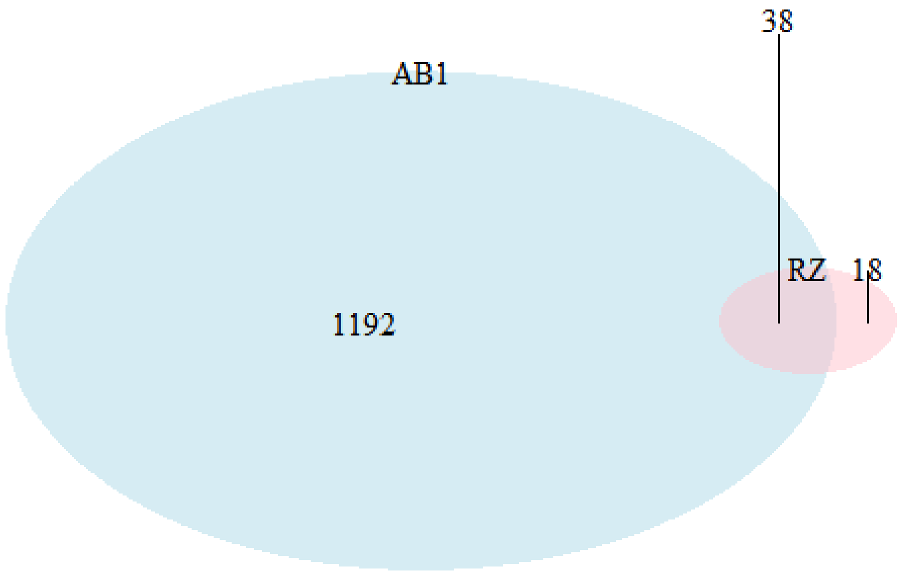


```
# quantile filter of genes
RZ_Filt <- qfilter(RZ, qnt.cut =  0.25)

# Diff.expr.analysis (DEA)
RZ_R <- RZ_Filt[, seq(1, ncol(RZ_Filt), by=2)]
RZ_NR <- RZ_Filt[, seq(2, ncol(RZ_Filt), by=2)]
res_filt_up2 <- analyze_DEG_cnt(RZ_R, RZ_NR, "logFC > 1 & FDR < 0.01")

## Disp = 0.02767 , BCV = 0.1664

res_filt_down2 <- analyze_DEG_cnt(RZ_R, RZ_NR, "logFC < -1 & FDR < 0.01")

## Disp = 0.02767 , BCV = 0.1664

head(res_filt_up2, 10)

##             logFC    logCPM        PValue           FDR
## Gm3756  3.701348  4.7912410 1.803775e-237 1.761266e-233
## Gm5526  3.591216  2.0691955 8.112465e-139 2.640427e-135
## Gm14173 3.660629  1.2493848 6.978950e-109 1.703620e-105
## Gm13192 3.749630  1.1055264 2.230250e-106 5.025440e-103
## Gm13237 2.019772  3.2739056  1.638361e-74  3.428036e-71
## Gm11401 2.462057  1.6153053  2.517539e-71  4.916418e-68
## Gm14536 7.316399 -0.8014681  6.763791e-61  1.100732e-57
## Gm15027 7.316399 -0.8014681  6.763791e-61  1.100732e-57
## Gm13167 1.878762  2.9639339  8.731235e-61  1.346127e-57
## Gm8451  1.741164  3.6465473  7.795332e-59  1.087375e-55

head(res_filt_down2, 10)

##                   logFC    logCPM        PValue           FDR
## Tpt1-ps6      -7.411835 4.6674962  0.000000e+00  0.000000e+00
## Gm7429        -8.458690 2.5263703 1.098597e-306 1.609060e-302
## E030024N20Rik -5.522032 1.9425928 7.039195e-206 5.154978e-202
## Gm5550        -5.956170 1.6077485 1.585137e-191 9.286683e-188
## Ftl2-ps       -2.746508 6.9946779 6.561429e-150 3.203399e-146
## Gm7816        -5.535396 0.9986649 1.150376e-142 4.813997e-139
## Rpl35a-ps2    -8.883831 0.6012592 7.394092e-142 2.707439e-138
## Gm10012       -5.309774 0.8609898 1.212730e-130 3.552451e-127
## Gm7665        -3.800612 1.2438596 1.008371e-113 2.685293e-110
## Gm11808       -2.422366 1.5001987  3.633532e-67  6.652316e-64

### Venn diagram for DEG list
library(VennDiagram)

## Warning: package 'VennDiagram' was built under R version 4.0.3

## Loading required package: grid

## Loading required package: futile.logger

ups <- list(AB1 = rownames(res_filt_up1), RZ = rownames(res_filt_up2))
ups_list <- get.venn.partitions(ups)

grid.newpage()
draw.pairwise.venn(area1 = ups_list$..count..[3]+ups_list$..count..[2],
                  area2 = ups_list$..count..[1]+ups_list$..count..[2],
                  cross.area = ups_list$..count..[2],
                  category = c("AB1", "RZ"),
                  fill = c("light blue", "pink"),
                  lty = "blank",
                  alpha = rep(0.5, 2),
                  cat.pos = c(0,0), # category label position
                  cat.dist = c(0,0)) # category label distance from circle
```



```
## (polygon[GRID.polygon.11], polygon[GRID.polygon.12],
polygon[GRID.polygon.13], polygon[GRID.polygon.14], text[GRID.text.15],
text[GRID.text.16], text[GRID.text.17], lines[GRID.lines.18],
text[GRID.text.19], text[GRID.text.20])
```

```
common_up <- ups_list[1, "..values.."] # common up-regulated genes


downs <- list(AB1 = rownames(res_filt_down1), RZ =
rownames(res_filt_down2))
downs_list <- get.venn.partitions(downs)

grid.newpage()
draw.pairwise.venn(area1 =
downs_list$..count..[3]+downs_list$..count..[2],
                  area2 = downs_list$..count..[1]+downs_list$..count..[2],
                  cross.area = downs_list$..count..[2],
                  category = c("AB1", "RZ"),
                  fill = c("light blue", "pink"),
                  lty = "blank",
                  alpha = rep(0.5, 2),
                  cat.pos = c(0,0), # category label position
                  cat.dist = c(0,0)) # category label distance from circle
```



```
## (polygon[GRID.polygon.21], polygon[GRID.polygon.22],
polygon[GRID.polygon.23], polygon[GRID.polygon.24], text[GRID.text.25],
text[GRID.text.26], lines[GRID.lines.27], text[GRID.text.28],
lines[GRID.lines.29], text[GRID.text.30], text[GRID.text.31])

common_down <- downs_list[1,"..values.."]  # common down-regulated genes
```

Bevacizumab responder vs nonresponder DEA

- 김이준 교수님이 제공해주신 코드를 수정하여 사용

```
#
https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst
/doc/analysis.html

# Install and load libraries
if (length(grep("^BiocManager$", rownames(installed.packages())))<1)
  install.packages("BiocManager")
if (length(grep("^TCGAbiolinks$", rownames(installed.packages())))<1)
  BiocManager::install("TCGAbiolinks")
if (length(grep("^EDASeq$", rownames(installed.packages())))<1)
  BiocManager::install("EDASeq")
if (length(grep("^stringr$", rownames(installed.packages())))<1)
  install.packages("stringr")
if (length(grep("^dplyr$", rownames(installed.packages())))<1)
  install.packages("dplyr")

# load clinical data
if (!file.exists("../GDCdata"))
  dir.create("../GDCdata")

# clinical.BCRtab.all 라는 전체 리스트의 각 항목을 따로따로 저장
nte <- as.data.frame(clinical.BCRtab.all[1])
colnames(nte) <- nte[1,]
nte <- nte[-c(1:2),]
```

```r
follow_up <- as.data.frame(clinical.BCRtab.all[2])
colnames(follow_up) <- follow_up[1,]
follow_up <- follow_up[-c(1:2),]
omf <- as.data.frame(clinical.BCRtab.all[3])
colnames(omf) <- omf[1,]
omf <- omf[-c(1:2),]
patient <- as.data.frame(clinical.BCRtab.all[4])
colnames(patient) <- patient[1,]
patient <- patient[-c(1:2),]
follow_up_nte <- as.data.frame(clinical.BCRtab.all[5])
colnames(follow_up_nte) <- follow_up_nte[1,]
follow_up_nte <- follow_up_nte[-c(1:2),]
drug <- as.data.frame(clinical.BCRtab.all[6])
colnames(drug) <- drug[1,]
drug <- drug[-c(1:2),]
radiation <- as.data.frame(clinical.BCRtab.all[7])
colnames(radiation) <- radiation[1,]
radiation <- radiation[-c(1:2),]


# Bevacizumab 약물처리가 된 데이터 뽑기
# grepl() 함수가 key point
# grepl(포함된_문자열, 조회할_벡터, ignore.case = T 또는 F)
# ignore.case: 대소문자 구분 무시
# bavacizumab 을 치료제로 쓴 사람은 총 31 명이었다...
beva <- subset(drug, grepl("(beva|avastin)", drug_name, ignore.case = T))
# 49 samples found

# 환자 군 분류 Step 1
# 데이터베이스에 labeling 된 response 별로 샘플을 구분
# CR = Complete Response
# PR = Partial Response
# SD = Stable Disease
# PD = Clinical Progressive Disease
# UK = [Not Applicable] or [Unknown]
# 1. response 별로 샘플 구분
unique(beva$measure_of_response)

## [1] "[Not Available]"              "Clinical Progressive Disease"
## [3] "Complete Response"            "[Not Applicable]"
## [5] "Partial Response"             "Stable Disease"
## [7] "[Unknown]"

beva_CR <- subset(beva, measure_of_response=="Complete Response")
beva_PR <- subset(beva, measure_of_response=="Partial Response")
beva_SD <- subset(beva, measure_of_response=="Stable Disease")
beva_PD <- subset(beva, measure_of_response=="Clinical Progressive
Disease")
```

```
beva_UK <- subset(beva, measure_of_response=="[Not Available]" |
measure_of_response=="[Not Applicable]"| measure_of_response==
"[Unknown]")

# CR 3 명, PR 4 명, SD 2 명, PD 12 명, Unknown 28 명
cat(sprintf("Numbers of samples\nCR: %d \nPR: %d \nSD: %d \nPD: %d
\nUK: %d", nrow(beva_CR), nrow(beva_PR), nrow(beva_SD), nrow(beva_PD),
nrow(beva_UK)))
```

```
## Numbers of samples
## CR: 3
## PR: 4
## SD: 2
## PD: 12
## UK: 28
```

```
# 환자 군 분류 Step 2
# 어떻게 구분 기준을 정의하느냐에 따른데, refractory 한지 아닌지에 따라
CR+PR+SD (responder, R) vs. PD (non-responder, NR) 이렇게 나눠서 분석
beva_R <- rbind(beva_CR, beva_PR, beva_SD) # 9 명
beva_R$response <- 1
beva_PD$response <- 0
beva <- rbind(beva_R, beva_PD)

beva <- dplyr::select(beva, bcr_patient_barcode, response)
colnames(beva) <- c("pt_id", "response")
head(beva)
```

```
##            pt_id response
## 171 TCGA-AA-3517        1
## 306 TCGA-AY-A8YK        1
## 308 TCGA-AZ-4308        1
## 218 TCGA-AA-3869        1
## 280 TCGA-AA-A02F        1
## 455 TCGA-F4-6806        1
```

```
# patient id 는 sample barcode 보다 짧다. 앞쪽 번호임...
# paitent id 앞쪽이 동일하면서 normal tissue 가 아닌 sample list 를 뽑아야
함...
# 더 간단한 방법이 있을수도 있는데, 일단 제 방식대로 뽑아보겠습니다.

COADMatrix <- SummarizedExperiment::assay(COADRnaseqSE,"raw_count")

id <- as.data.frame(colnames(COADMatrix))
colnames(id) <- c("sample_id")
library(stringr)
id$pt_id <- str_sub(id$sample_id, 1, 12)
```

```
id2 <- dplyr::inner_join(id, beva, by="pt_id") # 갯수가 13 개로 주는데... 일
부 missing value 가 있는 것 같다...
listSamples <- as.character(id2$sample_id)

TCGAquery_SampleTypes(listSamples, typesample=c("NT")) # normal tissue -->
이건 제거해야...

## [1] "TCGA-A6-2671-11A-01R-A32Z-07" "TCGA-AA-3517-11A-01R-A32Z-07"

TCGAquery_SampleTypes(listSamples, typesample=c("TP")) # tumor primary

## [1] "TCGA-AA-A02K-01A-03R-A32Y-07" "TCGA-NH-A8F7-01A-11R-A41B-07"
## [3] "TCGA-F4-6806-01A-11R-1839-07" "TCGA-NH-A6GB-01A-11R-A37K-07"
## [5] "TCGA-RU-A8FL-01A-11R-A37K-07" "TCGA-NH-A50U-01A-33R-A37K-07"
## [7] "TCGA-NH-A6GA-01A-11R-A37K-07" "TCGA-A6-5664-01A-21R-1839-07"
## [9] "TCGA-AY-A8YK-01A-11R-A41B-07"

TCGAquery_SampleTypes(listSamples, typesample=c("TM")) # tumor metastatic

## [1] "TCGA-NH-A8F7-06A-31R-A41B-07"

id3 <- subset(id2, !(sample_id %in% TCGAquery_SampleTypes(listSamples,
typesample=c("NT"))))
cat(sprintf("Number of samples \nbefore removing normal tissue: %d\nafter
removing normal tissue: %d", nrow(id2), nrow(id3)))

## Number of samples
## before removing normal tissue: 13
## after removing normal tissue: 11

# download RNA seq data
# 1. beva_responder vs. beva_non-resonder 로 분류된 환자 샘플 지정:
listsamples
# 2. 해당 샘플들을 barcode 로 지정하여 RNA sequence data 를 다운
# listSamples : normal tissue 제외
listSamples <- id3$sample_id

# expression matrix 전처리
COADMatrix <- SummarizedExperiment::assay(COADRnaseqSE_beva,"raw_count")

# For gene expression if you need to see a boxplot correlation and AAIC
plot to define outliers you can run
getwd() # 현재 workding directory 아래에 boxplot 이 만들어집니다.

## [1] "G:/내 드라이브/2020TLO/Work/Bioinformatics_study/R-
project/book_ed2"

COADRnaseq_CorOutliers_beva <-
TCGAanalyze_Preprocessing(COADRnaseqSE_beva,
```

```r
filename="./pictures/ch3_1.png")


View(COADRnaseq_CorOutliers_beva) # 보시면 각 sample 별 raw read count 가 보입니다.
head(COADRnaseq_CorOutliers_beva)
```

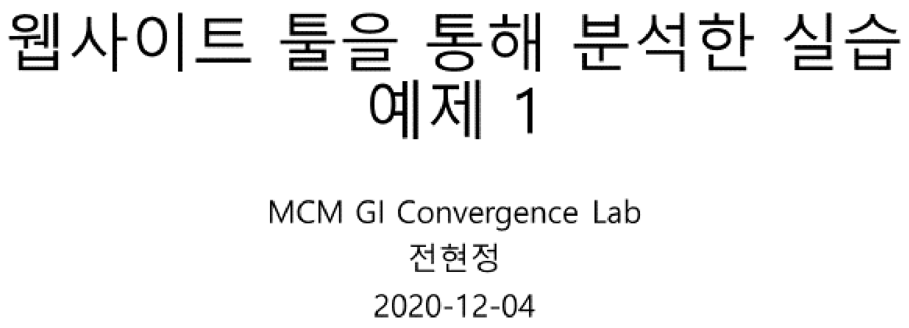
```
##             TCGA-AA-A02K-01A-03R-A32Y-07 TCGA-NH-A8F7-01A-11R-A41B-07
## A1BG|1                             43.50                            9
## A2M|2                            3755.96                         3835
## NAT1|9                            211.00                          276
## NAT2|10                           173.00                          255
## SERPINA3|12                         9.00                           51
## AADAC|13                           12.00                            9
##             TCGA-F4-6806-01A-11R-1839-07 TCGA-NH-A6GB-01A-11R-A37K-07
## A1BG|1                             70.54                        20.82
## A2M|2                            8934.97                      4281.95
## NAT1|9                            571.00                       147.00
## NAT2|10                           497.00                       111.00
## SERPINA3|12                      1787.00                        54.00
## AADAC|13                            6.00                         1.00
##             TCGA-RU-A8FL-01A-11R-A37K-07 TCGA-NH-A50U-01A-33R-A37K-07
## A1BG|1                             19.00                        56.84
## A2M|2                            2166.93                     12821.90
## NAT1|9                            173.00                       417.00
## NAT2|10                            62.00                       222.00
## SERPINA3|12                        20.00                        81.00
## AADAC|13                           31.00                         4.00
##             TCGA-NH-A6GA-01A-11R-A37K-07 TCGA-A6-5664-01A-21R-1839-07
## A1BG|1                             49.70                        46.13
## A2M|2                           43009.94                     31403.87
## NAT1|9                            287.00                       299.00
## NAT2|10                            42.00                       211.00
## SERPINA3|12                        18.00                     11263.00
## AADAC|13                           26.00                        21.00
##             TCGA-AY-A8YK-01A-11R-A41B-07 TCGA-NH-A8F7-06A-31R-A41B-07
## A1BG|1                             34.00                        11.93
## A2M|2                           10105.95                      3943.85
## NAT1|9                            200.00                       165.00
## NAT2|10                           291.00                       171.00
## SERPINA3|12                       934.00                        56.00
## AADAC|13                            8.00                       430.00
```

```r
library(EDASeq)
dataNorm <- TCGAanalyze_Normalization(tabDF = COADRnaseqSE_beva, geneInfo =  geneInfo)
```

```
## I Need about  2.5 seconds for this Complete Normalization Upper
Quantile  [Processing 80k elements /s]

## Step 1 of 4: newSeqExpressionSet ...
## Step 2 of 4: withinLaneNormalization ...

## Step 3 of 4: betweenLaneNormalization ...

## Step 4 of 4: exprs ...

# quantile filter of genes
dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
                               method = "quantile",
                               qnt.cut =  0.25)


# DEG 분석
# bevacizumab responder vs. non-responder 로 그룹 나누고 DEG 분석
# 결과물은 DEG_beva_resp_vs_nonresp.csv 에서도 확인 가능
# groups
beva_R_id <- as.character(subset(id3, response==1)$sample_id)
beva_NR_id <- as.character(subset(id3, response==0)$sample_id)

# Diff.expr.analysis (DEA)
dataDEGs <- TCGAanalyze_DEA(mat1 = dataFilt[,beva_R_id],
                          mat2 = dataFilt[,beva_NR_id],
                          Cond1type = "Responder",
                          Cond2type = "Nonresponder",
                          fdr.cut = 0.01 ,
                          logFC.cut = 1,
                          method = "glmLRT")

## Batch correction skipped since no factors provided

## ----------------------- DEA -------------------------------

## there are Cond1 type Responder in  4 samples

## there are Cond2 type Nonresponder in  7 samples

## there are  14892 features as miRNA or genes

## I Need about  5.5 seconds for this DEA. [Processing 30k elements /s]

## ----------------------- END DEA -------------------------------

# DEGs table with expression values in nonresponder and responder samples
dataDEGsFiltLevel <-
TCGAanalyze_LevelTab(dataDEGs,"Nonresponder","Responder",
dataFilt[,beva_NR_id],dataFilt[,beva_R_id])

head(dataDEGsFiltLevel) # DEGs

##            mRNA     logFC         FDR Nonresponder Responder      Delta
## PPBP      PPBP 11.015466 0.004148567   65797.7143     37.50 724792.465
## LY6E      LY6E  2.860467 0.003376640   14639.7143   1534.50  41876.426
## CRABP2   CRABP2  4.115481 0.009941498    1510.8571     62.00   6217.903
## HAGHL     HAGHL  3.907485 0.001393997    1197.4286     67.50   4678.935
## PACSIN3 PACSIN3  2.942383 0.001663923    1361.0000    130.50   4004.583
## HEPHL1   HEPHL1  5.620872 0.008228477     612.4286     13.75   3442.383

#write.csv(dataDEGsFiltLevel, "DEG_beva_resp_vs_nonresp.csv")
```

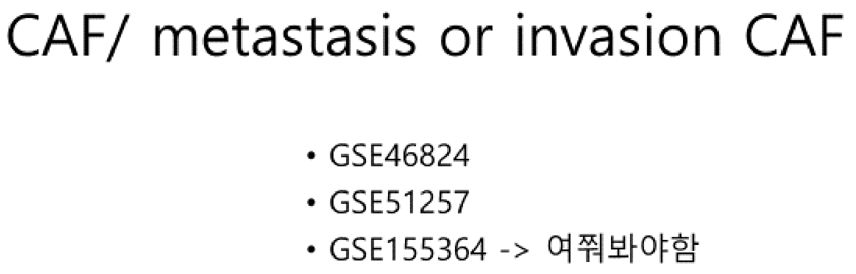
웹사이트 툴을 통해 분석한 실습 예제

CAF/ metastasis or invasion CAF



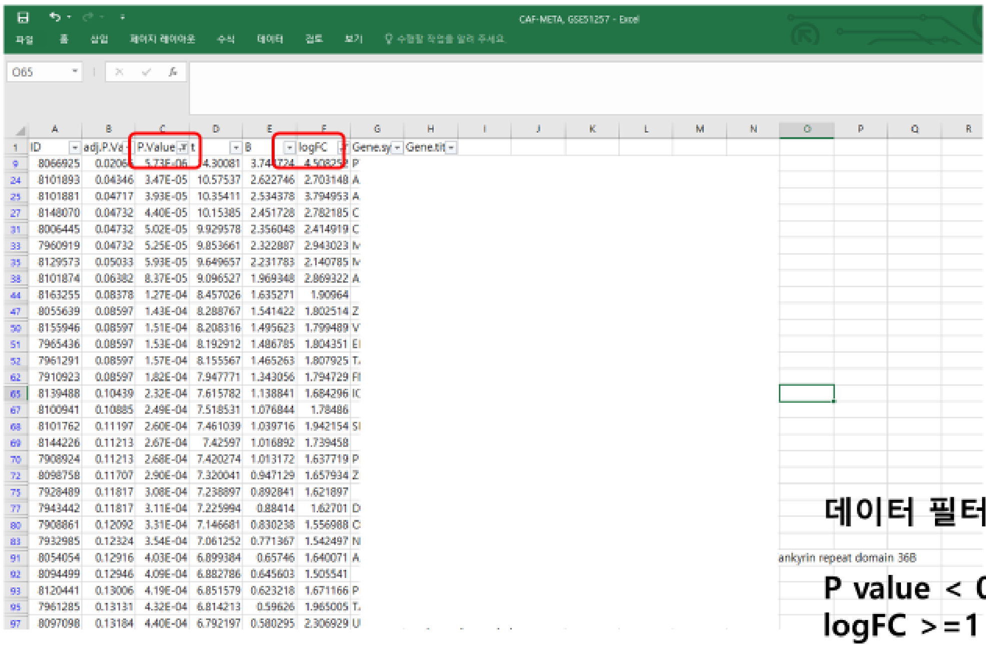
웹사이트 툴을 통해 분석한 실습 예제 1

MCM GI Convergence Lab
전현정
2020-12-04



CAF/ metastasis or invasion CAF

- GSE46824
- GSE51257
- GSE155364 -> 여쭤봐야함

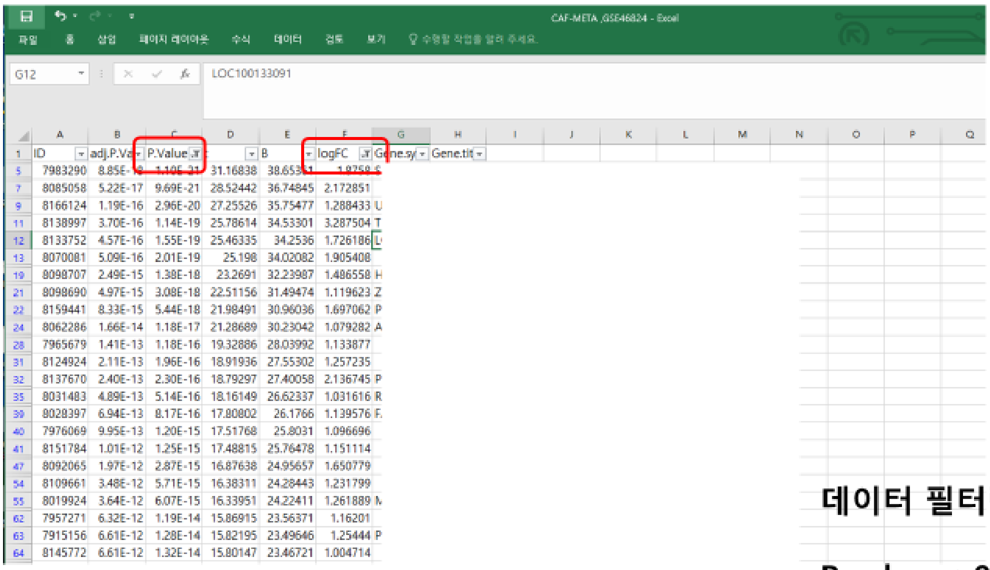


	ID	adj.P.Va	P.Value	t	B	logFC	Gene.sy	Gene.tit
9	8066925	0.0206	5.73E-06	4.30081	3.74[illegible]724	4.50825	P	
24	8101893	0.04346	3.47E-05	10.57537	2.622746	2.703148	A	
25	8101881	0.04717	3.93E-05	10.35411	2.534378	3.794953	A	
27	8148070	0.04732	4.40E-05	10.15385	2.451728	2.782185	C	
31	8006445	0.04732	5.02E-05	9.929578	2.356048	2.414919	C	
33	7960919	0.04732	5.25E-05	9.853661	2.322887	2.943023	M	
35	8129573	0.05033	5.93E-05	9.649657	2.231783	2.140785	M	
38	8101874	0.06382	8.37E-05	9.096527	1.969348	2.869322	A	
44	8163255	0.08378	1.27E-04	8.457026	1.635271	1.90964		
47	8055639	0.08597	1.43E-04	8.288767	1.541422	1.802514	Z	
50	8155946	0.08597	1.51E-04	8.208316	1.495623	1.799489	V	
51	7965436	0.08597	1.53E-04	8.192912	1.486785	1.804351	EI	
52	7961291	0.08597	1.57E-04	8.155567	1.465263	1.807925	T.	
62	7910923	0.08597	1.82E-04	7.947771	1.343056	1.794729	FI	
65	8139488	0.10439	2.32E-04	7.615782	1.138841	1.684296	IC	
67	8100941	0.10885	2.49E-04	7.518531	1.076844	1.78486		
68	8101762	0.11197	2.60E-04	7.461039	1.039716	1.942154	SI	
69	8144226	0.11213	2.67E-04	7.42597	1.016892	1.739458		
70	7908924	0.11213	2.68E-04	7.420274	1.013172	1.637719	P	
72	8098758	0.11707	2.90E-04	7.320041	0.947129	1.657934	Z	
75	7928489	0.11817	3.08E-04	7.238897	0.892841	1.621897		
77	7943442	0.11817	3.11E-04	7.225994	0.88414	1.62701	D	
80	7908861	0.12092	3.31E-04	7.146681	0.830238	1.556988	C	
83	7932985	0.12324	3.54E-04	7.061252	0.771367	1.542497	N	
91	8054054	0.12916	4.03E-04	6.899384	0.65746	1.640071	A	
92	8094499	0.12946	4.09E-04	6.882786	0.645603	1.505541		
93	8120441	0.13006	4.19E-04	6.851579	0.623218	1.671166	P	
95	7961285	0.13131	4.32E-04	6.814213	0.59626	1.965005	T.	
97	8097098	0.13184	4.40E-04	6.792197	0.580295	2.306929	U	

ankyrin repeat domain 36B

데이터 필터링 기준

P value < 0.05
logFC >=1

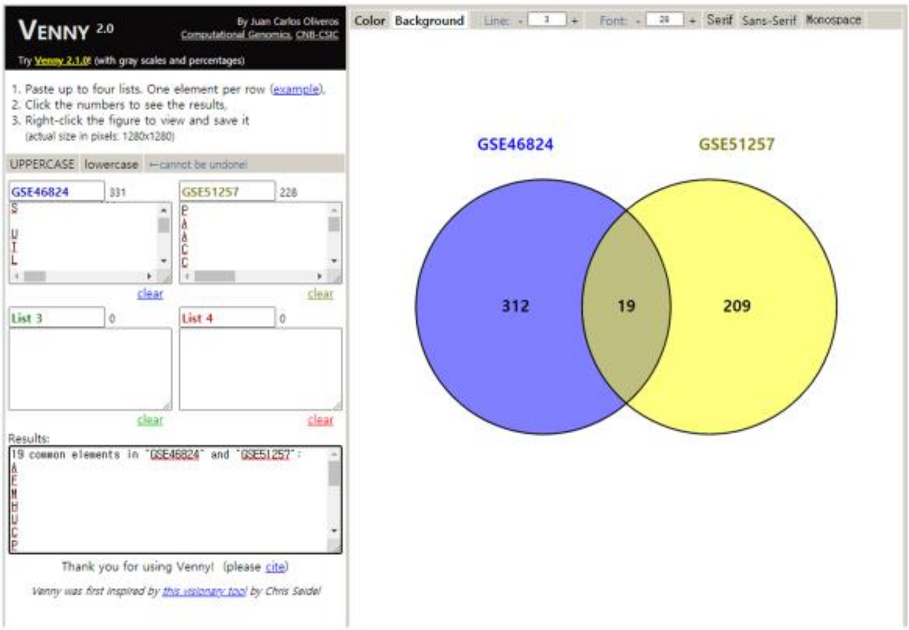


LOC100133091

	ID	adj.P.Va	P.Value	t	B	logFC	Gene.sy	Gene.tit
5	7983290	8.85E-[illegible]	1.10E-21	31.16838	38.653[illegible]1	1.9759	S	
7	8085058	5.22E-17	9.69E-21	28.52442	36.74845	2.172851		
9	8166124	1.19E-16	2.96E-20	27.25526	35.75477	1.288433	U	
11	8138997	3.70E-16	1.14E-19	25.78614	34.53301	3.287504	T	
12	8133752	4.57E-16	1.55E-19	25.46335	34.2536	1.726186	L	
13	8070081	5.09E-16	2.01E-19	25.198	34.02082	1.905408		
19	8098707	2.49E-15	1.38E-18	23.2691	32.23987	1.486558	H	
21	8098690	4.97E-15	3.08E-18	22.51156	31.49474	1.119623	Z	
22	8159441	8.33E-15	5.44E-18	21.98491	30.96036	1.697062	P	
24	8062286	1.66E-14	1.18E-17	21.28689	30.23042	1.079282	A	
28	7965679	1.41E-13	1.18E-16	19.32886	28.03992	1.133877		
31	8124924	2.11E-13	1.96E-16	18.91936	27.55302	1.257235		
32	8137670	2.40E-13	2.30E-16	18.79297	27.40058	2.136745	P	
35	8031483	4.89E-13	5.14E-16	18.16149	26.62337	1.031616	R	
39	8028397	6.94E-13	8.17E-16	17.80802	26.1766	1.139576	F.	
40	7976069	9.95E-13	1.20E-15	17.51768	25.8031	1.096696		
41	8151784	1.01E-12	1.25E-15	17.48815	25.76478	1.151114		
47	8092065	1.97E-12	2.87E-15	16.87638	24.95657	1.650779		
54	8109661	3.48E-12	5.71E-15	16.38311	24.28443	1.231799		
55	8019924	3.64E-12	6.07E-15	16.33951	24.22411	1.261889	N	
62	7957271	6.32E-12	1.19E-14	15.86915	23.56371	1.16201		
63	7915156	6.61E-12	1.28E-14	15.82195	23.49646	1.25444	P	
64	8145772	6.61E-12	1.32E-14	15.80147	23.46721	1.004714		

데이터 필터링 기준

P value < 0.05
logFC >=1



VENNY 2.0

By Juan Carlos Oliveros
Computational Genomics, CNB-CSIC

Try Venny 2.1.0! (with gray scales and percentages)

1. Paste up to four lists. One element per row (example).
2. Click the numbers to see the results.
3. Right-click the figure to view and save it
(actual size in pixels: 1280x1280)

UPPERCASE lowercase ←cannot be undone!

GSE46824 331
S
U
I
L

clear

GSE51257 228
P
A
A
C
C

clear

List 3 0

clear

List 4 0

clear

Results:

```
19 common elements in "GSE46824" and "GSE51257":
A
E
M
H
U
C
P
```

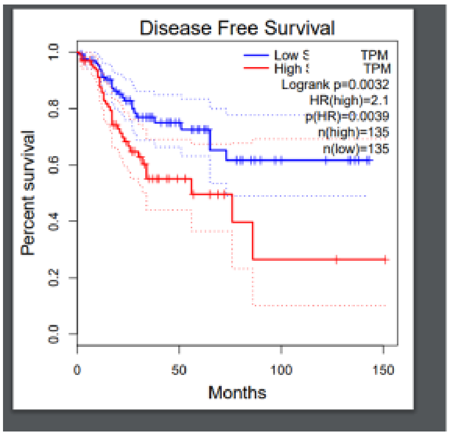
Thank you for using Venny! (please cite)

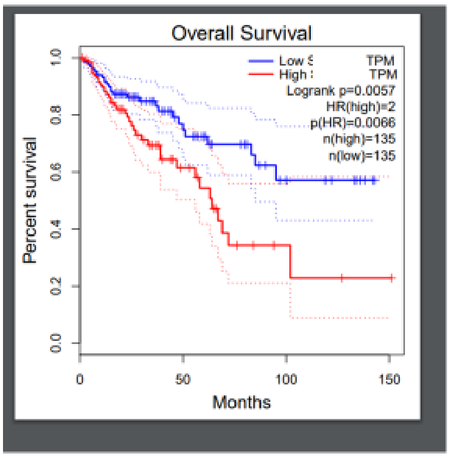
Venny was first inspired by this visionary tool by Chris Seidel

DEGs

- A
- F
- M
- H
- U
- C
- P
- S
- N
- A
- S
- S
- P
- N
- M
- E
- S
- IT
- IL

13. S





CAF/ NF

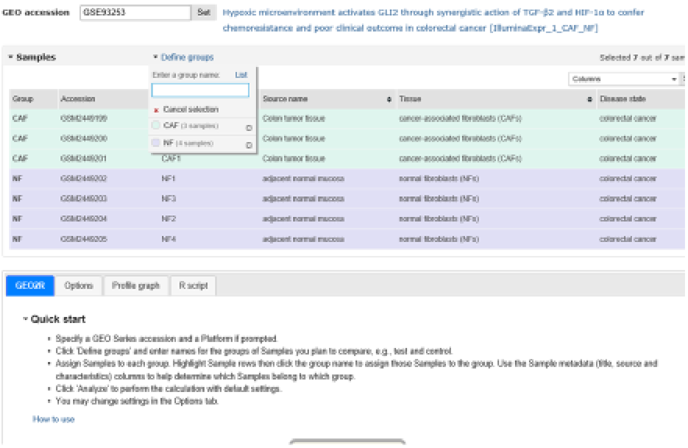
웹사이트 툴을 통해 분석한 실습 예제 2

MCM GI Convergence Lab
전현정
2020-12-08

웹사이트 툴을 통해 분석한 실습 예제 2

MCM GI Convergence Lab
전현정
2020-12-08

CAF/NF로 그룹 설정



GEO accession GSE93253 Set Hypoxic microenvironment activates GLI2 through synergistic action of TGF-β2 and HIF-1α to confer chemoresistance and poor clinical outcome in colorectal cancer [IlluminaExpr_1_CAF_NF]

Group	Accession		Source name	Tissue	Disease state
CAF	GSM2449199		Colon tumor tissue	cancer-associated fibroblasts (CAFs)	colorectal cancer
CAF	GSM2449200		Colon tumor tissue	cancer-associated fibroblasts (CAFs)	colorectal cancer
CAF	GSM2449201	CAF1	Colon tumor tissue	cancer-associated fibroblasts (CAFs)	colorectal cancer
NF	GSM2449202	NF1	adjacent normal mucosa	normal fibroblasts (NFs)	colorectal cancer
NF	GSM2449203	NF3	adjacent normal mucosa	normal fibroblasts (NFs)	colorectal cancer
NF	GSM2449204	NF2	adjacent normal mucosa	normal fibroblasts (NFs)	colorectal cancer
NF	GSM2449205	NF4	adjacent normal mucosa	normal fibroblasts (NFs)	colorectal cancer

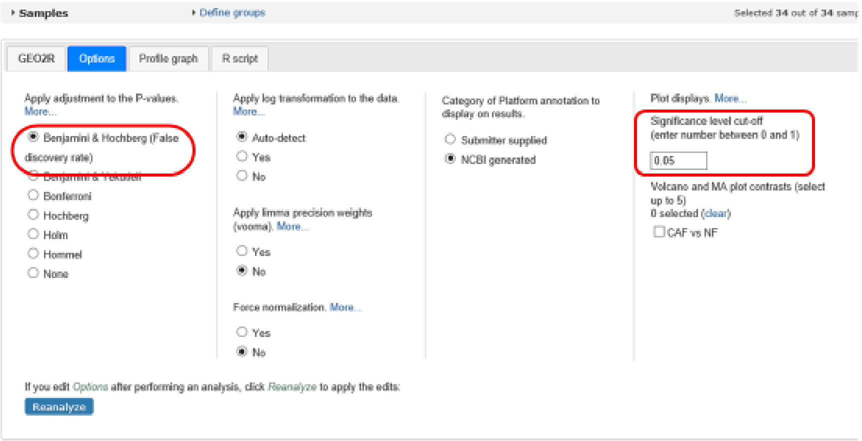
▾ Quick start

- 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, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong to which group.
- Click 'Analyze' to perform the calculation with default settings.
- You may change settings in the Options tab.

How to use

GSE51257
GSE93253
GSE46824
GSE70468-> gene acc로 나옴
GSE155364->geo2r 이 불가능?

GEO2R로 analysis



Apply adjustment to the P-values. More...

- ● Benjamini & Hochberg (False discovery rate)
- ○ Benjamini & Yekutieli
- ○ Bonferroni
- ○ Hochberg
- ○ Holm
- ○ Hommel
- ○ None

Apply log transformation to the data. More...

- ● Auto-detect
- ○ Yes
- ○ No

Apply limma precision weights (vooma). More...

- ○ Yes
- ● No

Force normalization. More...

- ○ Yes
- ● No

Category of Platform annotation to display on results.

- ○ Submitter supplied
- ● NCBI generated

Plot displays. More...

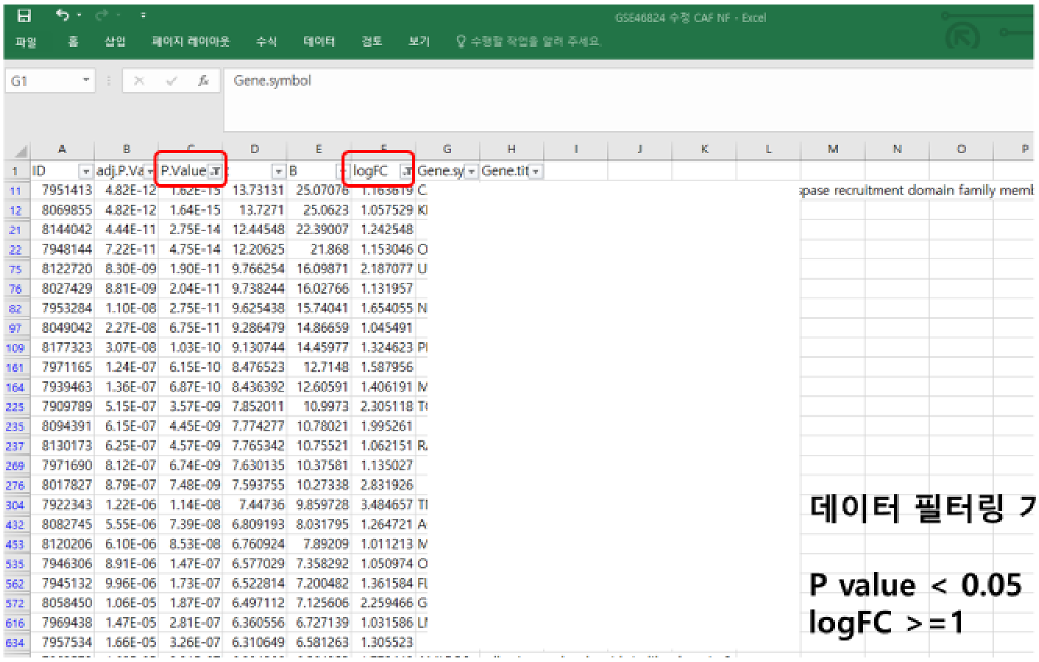
Significance level cut-off (enter number between 0 and 1)

0.05

Volcano and MA plot contrasts (select up to 5)
0 selected (clear)

□ CAF vs NF

If you edit Options after performing an analysis, click Reanalyze to apply the edits:



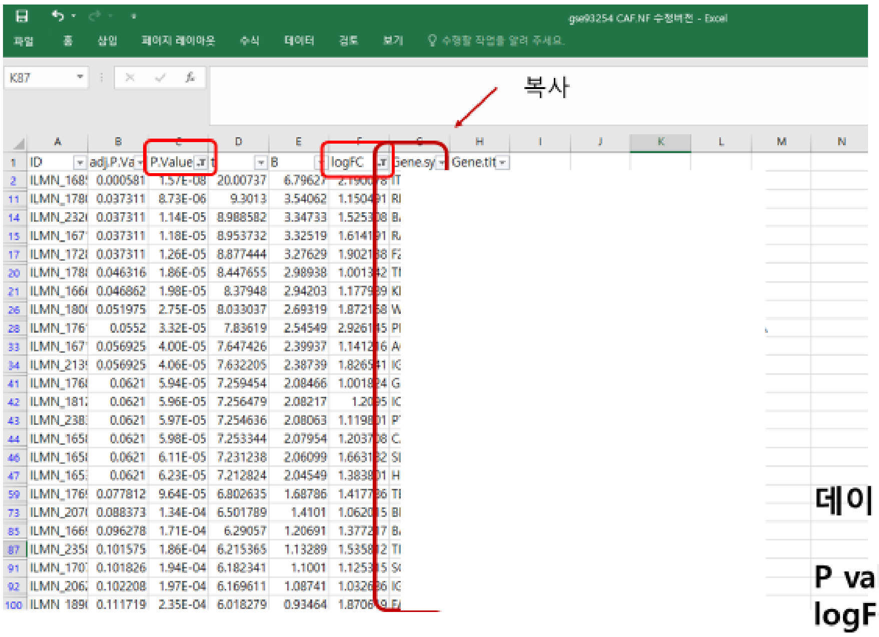
Gene.symbol

1	ID	adj.P.Va	P.Value		B	logFC	Gene.sy	Gene.tit
11	7951413	4.82E-12	1.62E-15	13.73131	25.07076	1.163619	C	
12	8069855	4.82E-12	1.64E-15	13.7271	25.0623	1.057529	Kl	
21	8144042	4.44E-11	2.75E-14	12.44548	22.39007	1.242548		
22	7948144	7.22E-11	4.75E-14	12.20625	21.868	1.153046	O	
75	8122720	8.30E-09	1.90E-11	9.766254	16.09871	2.187077	U	
76	8027429	8.81E-09	2.04E-11	9.738244	16.02766	1.131957		
82	7953284	1.10E-08	2.75E-11	9.625438	15.74041	1.654055	N	
97	8049042	2.27E-08	6.75E-11	9.286479	14.86659	1.045491		
109	8177323	3.07E-08	1.03E-10	9.130744	14.45977	1.324623	Pl	
161	7971165	1.24E-07	6.15E-10	8.476523	12.7148	1.587956		
164	7939463	1.36E-07	6.87E-10	8.436392	12.60591	1.406191	M	
225	7909789	5.15E-07	3.57E-09	7.852011	10.9973	2.305118	T	
235	8094391	6.15E-07	4.45E-09	7.774277	10.78021	1.995261		
237	8130173	6.25E-07	4.57E-09	7.765342	10.75521	1.062151	R	
269	7971690	8.12E-07	6.74E-09	7.630135	10.37581	1.135027		
276	8017827	8.79E-07	7.48E-09	7.593755	10.27338	2.831926		
304	7922343	1.22E-06	1.14E-08	7.44736	9.859728	3.484657	Tl	
432	8082745	5.55E-06	7.39E-08	6.809193	8.031795	1.264721	A	
453	8120206	6.10E-06	8.53E-08	6.760924	7.89209	1.011213	M	
535	7946306	8.91E-06	1.47E-07	6.577029	7.358292	1.050974	O	
562	7945132	9.96E-06	1.73E-07	6.522814	7.200482	1.361584	Fl	
572	8058450	1.06E-05	1.87E-07	6.497112	7.125606	2.259466	G	
616	7969438	1.47E-05	2.81E-07	6.360556	6.727139	1.031586	Ll	
634	7957534	1.66E-05	3.26E-07	6.310649	6.581263	1.305523		

ɔpase recruitment domain family meml

데이터 필터링 기준

P value < 0.05
logFC >=1

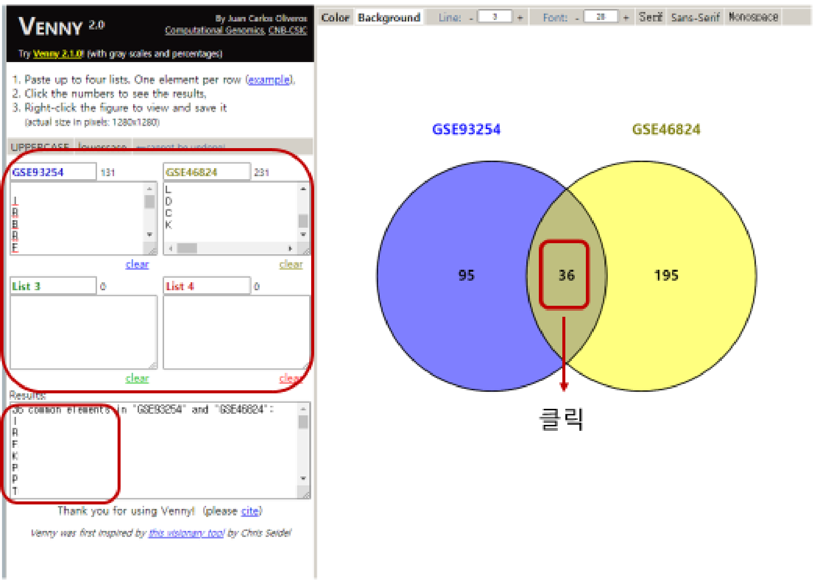


	A	B	C	D	E	F	G	H
1	ID	adj.P.Va	P.Value	t	B	logFC	Gene.sy	Gene.tit
2	ILMN_168	0.000581	1.57E-08	20.00737	6.79627	2.190078	IT	
11	ILMN_178	0.037311	8.73E-06	9.3013	3.54062	1.150491	RI	
14	ILMN_232	0.037311	1.14E-05	8.988582	3.34733	1.525308	B.	
15	ILMN_167	0.037311	1.18E-05	8.953732	3.32519	1.614191	R.	
17	ILMN_172	0.037311	1.26E-05	8.877444	3.27629	1.902188	F.	
20	ILMN_178	0.046316	1.86E-05	8.447655	2.98938	1.001342	TI	
21	ILMN_166	0.046862	1.98E-05	8.37948	2.94203	1.177989	KI	
26	ILMN_180	0.051975	2.75E-05	8.033037	2.69319	1.872168	W	
28	ILMN_176	0.0552	3.32E-05	7.83619	2.54549	2.926185	PI	
33	ILMN_167	0.056925	4.00E-05	7.647426	2.39937	1.141216	A	
34	ILMN_213	0.056925	4.06E-05	7.632205	2.38739	1.826541	IG	
41	ILMN_176	0.0621	5.94E-05	7.259454	2.08466	1.001824	G	
42	ILMN_181	0.0621	5.96E-05	7.256479	2.08217	1.2095	IC	
43	ILMN_238	0.0621	5.97E-05	7.254636	2.08063	1.119801	P'	
44	ILMN_165	0.0621	5.98E-05	7.253344	2.07954	1.203708	C.	
46	ILMN_165	0.0621	6.11E-05	7.231238	2.06099	1.663182	SI	
47	ILMN_165	0.0621	6.23E-05	7.212824	2.04549	1.383801	H	
59	ILMN_176	0.077812	9.64E-05	6.802635	1.68786	1.417786	TI	
73	ILMN_207	0.088373	1.34E-04	6.501789	1.4101	1.062015	BI	
85	ILMN_166	0.096278	1.71E-04	6.29057	1.20691	1.377217	B.	
87	ILMN_235	0.101575	1.86E-04	6.215365	1.13289	1.535812	TI	
91	ILMN_170	0.101826	1.94E-04	6.182341	1.1001	1.125315	S(	
92	ILMN_206	0.102208	1.97E-04	6.169611	1.08741	1.032686	IC	
100	ILMN_189	0.111719	2.35E-04	6.018279	0.93464	1.870619	F.	

데이터 필터링 기준

P value < 0.05
logFC >=1

각 gene symbol
붙여넣기



1. Paste up to four lists. One element per row (example).
2. Click the numbers to see the results.
3. Right-click the figure to view and save it
(actual size in pixels: 1280x1280)

GSE93254 131
I
A
B
A
E

GSE46824 231
L
D
C
K

List 3 0

List 4 0

Results:

```
36 common elements in "GSE93254" and "GSE46824":
I
A
F
K
P
P
T
```

Thank you for using Venny! (please cite)

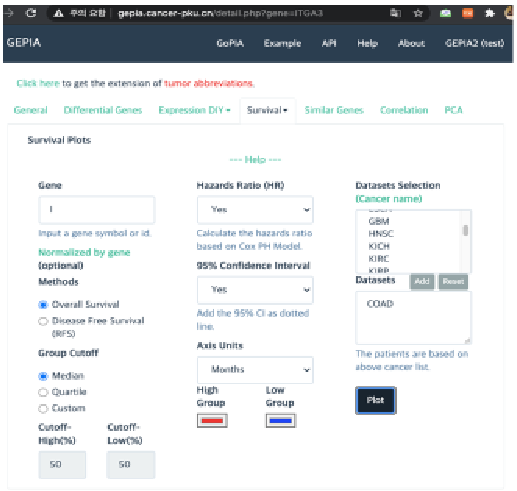
Venny was first inspired by this visionary tool by Chris Seidel

클릭

DEGs

- ITGA3
- RARB
- F2RL2
- KRT17
- PPP1R14A
- PTPRE
- TGFB2
- PLCB4
- EDIL3
- F2R
- TNFSF18
- OXTR
- SULF1
- SORBS2
- PDGFA
- IL6
- SEMA5A
- PKP2
- GRAMD3
- ST6GALNAC5
- DYSF
- SERTAD4
- TNFSF4
- SPINT2
- PDLIM3
- C5orf46
- KLF2
- GRIK2
- C7orf69
- BST1
- KCNMA1
- EFHD1
- CDH13
- ABI3BP
- LEF1
- KRT18

GEPIA



Click here to get the extension of tumor abbreviations.

Survival Plots

Gene: I

Input a gene symbol or id.

Normalized by gene (optional)

Methods

● Overall Survival

○ Disease Free Survival (RFS)

Group Cutoff

● Median

○ Quartile

○ Custom

Cutoff-High(%): 50

Cutoff-Low(%): 50

Hazards Ratio (HR): Yes

Calculate the hazards ratio based on Cox PH Model.

95% Confidence Interval: Yes

Add the 95% CI as dotted line.

Axis Units: Months

High Group 🟥

Low Group 🟦

Datasets Selection (Cancer name)

GBM
HNSC
KICH
KIRC
KIRP

Datasets: COAD

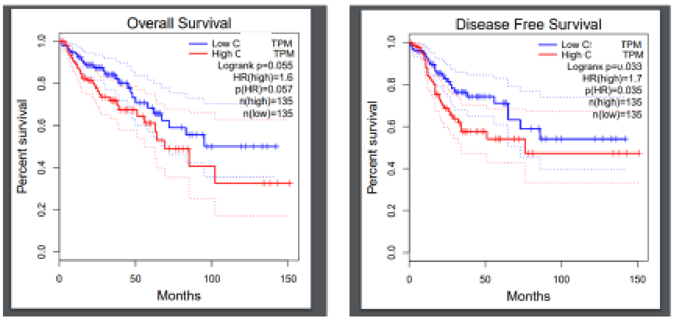
The patients are based on above cancer list.

http://gepia.cancer-pku.cn/

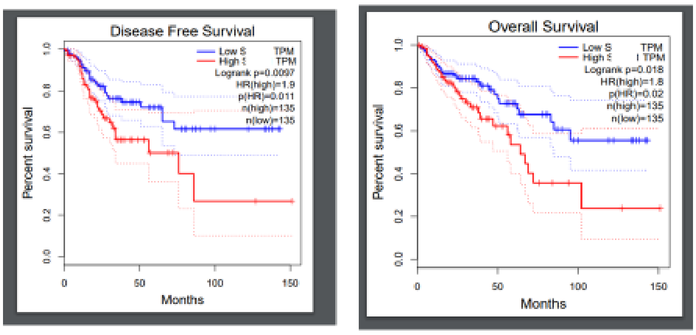
Logrank $p<0.05$

단, C 은 근접하
여기록

C



S



김이준 교수님의 강연 자료

TCGA database 를 활용한 transcriptome 연구 방법론 (김이준 교수님 논문)

- 논문: Kim, YJ., Kim, K., Lee, K.H. et al. Immune expression signatures as candidate prognostic biomarkers of age and gender survival differences in cutaneous melanoma. Sci Rep 10, 12322 (2020). *https://doi.org/10.1038/s41598-020-69082-z*

TCGA database를 활용한 transcriptome 연구 방법론

이화여대부속목동병원
융합의학연구원
김이준
kimyj.ro@gmail.com

2020-10-28

TCGA를 활용한 연구

- 논문 예시
  - Melanoma 에서 성별로 prognosis가 다른 이유를 gene expression 차이에서 설명할 수 있는지 확인...

www.nature.com/scientificreports

SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN **Immune expression signatures as candidate prognostic biomarkers of age and gender survival differences in cutaneous melanoma**

Yi-Jun Kim[1,2], Kyubo Kim[2], Kye Hwa Lee[3], Jiyoung Kim[1] & Wonguen Jung[1]

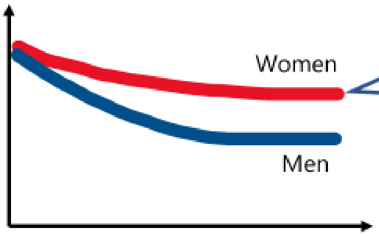
This study aims to investigate the difference of gene expression and its prognostic significance in younger women with melanoma. Significantly upregulated genes in tumors compared to normal skin tissues were extracted. Among these genes, genes that significantly affected survival according to expression level were selected, and pathway annotation was performed. The patient proportion with high/low expression of the most significant pathways was analyzed in each age (< 50, 50–59, ≥ 60) and gender group. Survival was analyzed according to age, gender, and pathways. The most significant pathways that were upregulated in tumor tissues and also had impacts on survival were programmed cell death protein (PD)-1, interferon-γ, and interferon-α/β pathways. In women, the immune signaling rate in patients was higher than men and decreased with age (63.5%, 53.8%, and 47.6%). In men, the decreasing tendency was minimal (47.8%, 50.0%, and 41.6%). In patients aged < 60 years, women had a favorable survival rate than men ($p = 0.055$). Except for patients with high immune signaling, no survival difference was observed between genders ($p = 0.6$). In conclusion, younger female melanoma patients had high immune signaling than older women and men. This immune signaling improved survival of the younger female patients.

Introduction

Cutaneous melanoma

- 1.6% of all cancers
  - ~= brain/nervous system cancer
  - ~= ovarian caner
- In the United States,
  - New cases in 2019
    - 57,220 in men
    - 39,260 in women
  - Death in 2019
    - 4,740 in men
    - 2,490 in women

Sexual disparity in cutaneous melanoma prognosis



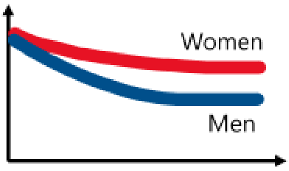
Good prognosis than men
- lower risk of metastasis
- longer survival

Esp. Younger women

Why?

Sex hormones?
Immune system?
Oxidative stress?
Environmental differences
(tanning bed exposure) ?

Genetic analyses so far…



The differences in
- Pigmentation genes,
- Apoptosis genes,
- Reactive oxygen species genes,
- Sex-specific pleiotropic cancer-related genes
- DNA mutational burden difference

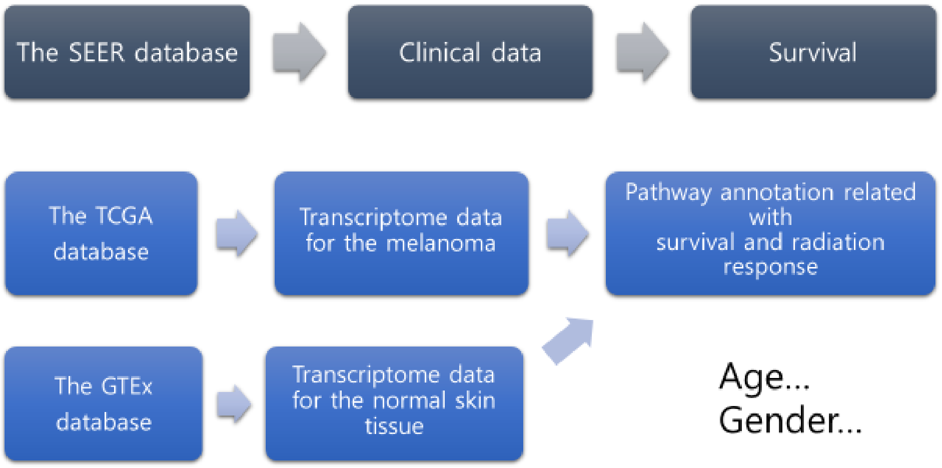
Q1. Age?

Menopausal status?

Q2. Age and gender matter.

✓ Mutational analysis?
✓ SNPs analysis?
✓ Transcriptome analysis!

In this study...



The SEER database

Clinical data

Survival

The TCGA database

Transcriptome data for the melanoma

Pathway annotation related with survival and radiation response

The GTEx database

Transcriptome data for the normal skin tissue

Age...
Gender...

Method

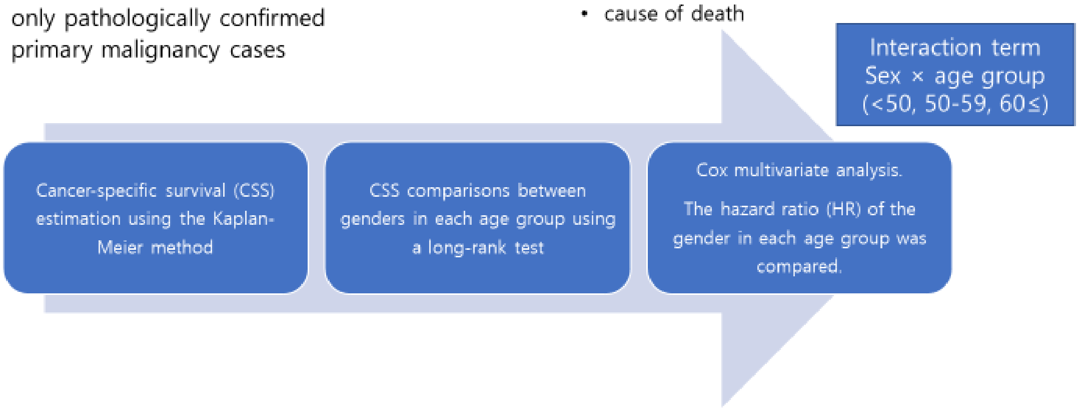
SEER database

- The SEER database is an authoritative cancer registry of the National Cancer Institute (NCI) in the USA. The SEER database collects and registers information on all cancer patients in the United States, approximately 34.6% of the total United States population. SEER collects data on patient demographics, primary tumor sites, tumor morphology, diagnostic stages, primary treatment progress, and tracking of critical conditions.

2.1 Survival comparison between gender and age groups using the SEER database

- Inclusion
  - melanoma of the skin (C440–449)
  - from 1975 to 2015
  - only pathologically confirmed
  - primary malignancy cases

- Exclusion
  - No information on
    - survival time
    - cause of death



Interaction term
Sex × age group
(<50, 50-59, 60≤)

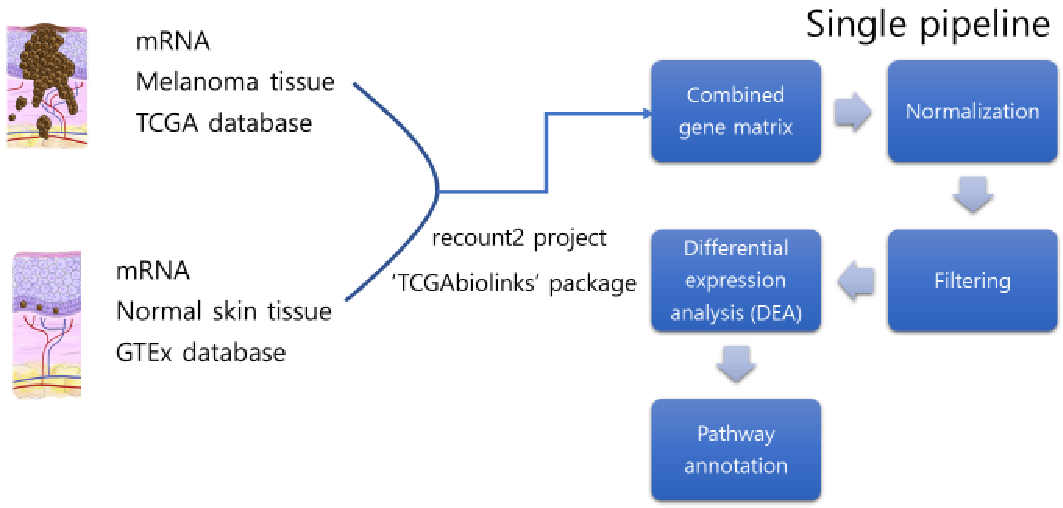
Cancer-specific survival (CSS) estimation using the Kaplan-Meier method

CSS comparisons between genders in each age group using a long-rank test

Cox multivariate analysis.

The hazard ratio (HR) of the gender in each age group was compared.

2.2 mRNA expression data preparation



Single pipeline

mRNA
Melanoma tissue
TCGA database

mRNA
Normal skin tissue
GTEx database

recount2 project
'TCGAbiolinks' package

Combined gene matrix

Normalization

Filtering

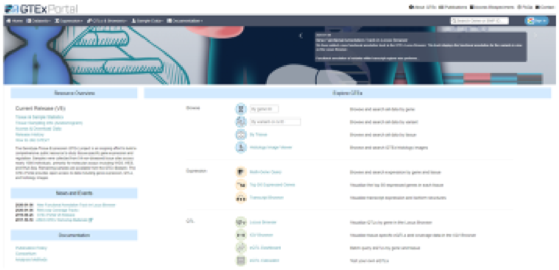
Differential expression analysis (DEA)

Pathway annotation

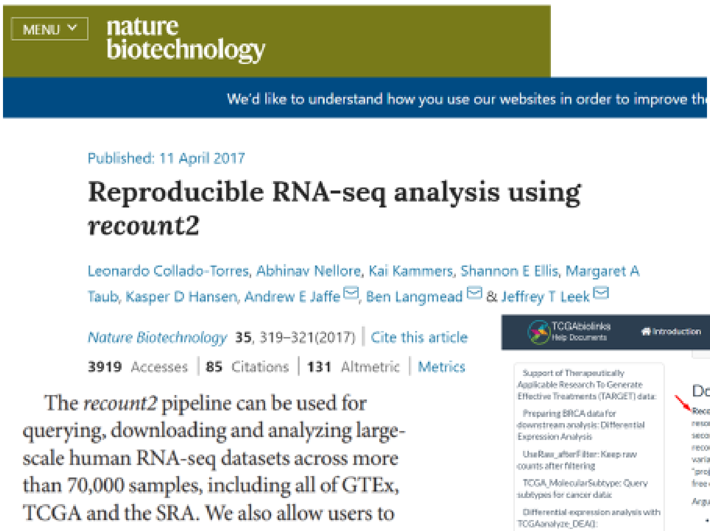
TCGA database

- The Cancer Genome Atlas program (TCGA , RRID: SCR_003193) has collected over 20,000 primary cancer and matched normal samples spanning 33 cancer types carcinomas since 2006 by the National Cancer Institute and the National Human Genome Research Institute. Collection items include genomic, epigenomic, transcriptomic, and proteomic data.

GTEx database (Genotype-Tissue Expression)



- The GTEx Project of the NIH Common Fund has established a resource database and tissue bank in which to study the relationship between genetic variation, gene expression, and molecular phenotypes in multiple tissues. The project includes data on approximately 900 post-donors by 2015.


MENU
nature biotechnology
We'd like to understand how you use our websites in order to improve th

Published: 11 April 2017

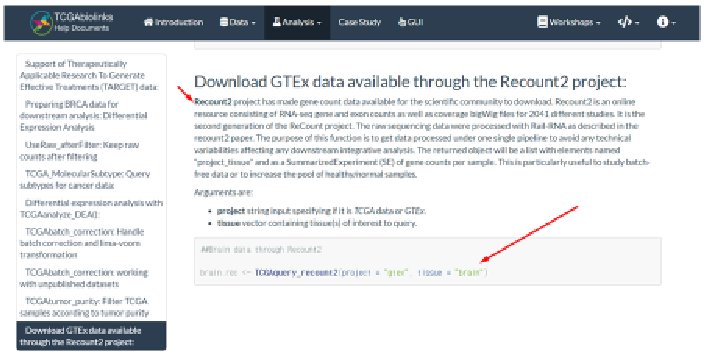
Reproducible RNA-seq analysis using *recount2*

Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead & Jeffrey T Leek

Nature Biotechnology **35**, 319–321(2017) | Cite this article

3919 Accesses | **85** Citations | **131** Altmetric | Metrics

The *recount2* pipeline can be used for querying, downloading and analyzing large-scale human RNA-seq datasets across more than 70,000 samples, including all of GTEx, TCGA and the SRA. We also allow users to


TCGAbiolinks Help Documents
Introduction
Data
Analysis
Case Study
GUI
Workshops
Support of Therapeutically Applicable Research To Generate Effective Treatments (TARGET) data:
Preparing BRCA data for downstream analysis: Differential Expression Analysis
UseRaw_afterFilter: Keep raw counts after filtering
TCGA_MolecularSubtype: Query subtypes for cancer data:
Differential expression analysis with TCGAanalyze_DEA():
TCGAbatch_correction: Handle batch correction and lima-voom transformation
TCGAbatch_correction: working with unpublished datasets
TCGAtumor_purity: Filter TCGA samples according to tumor purity
Download GTEx data available through the Recount2 project:

Download GTEx data available through the Recount2 project:

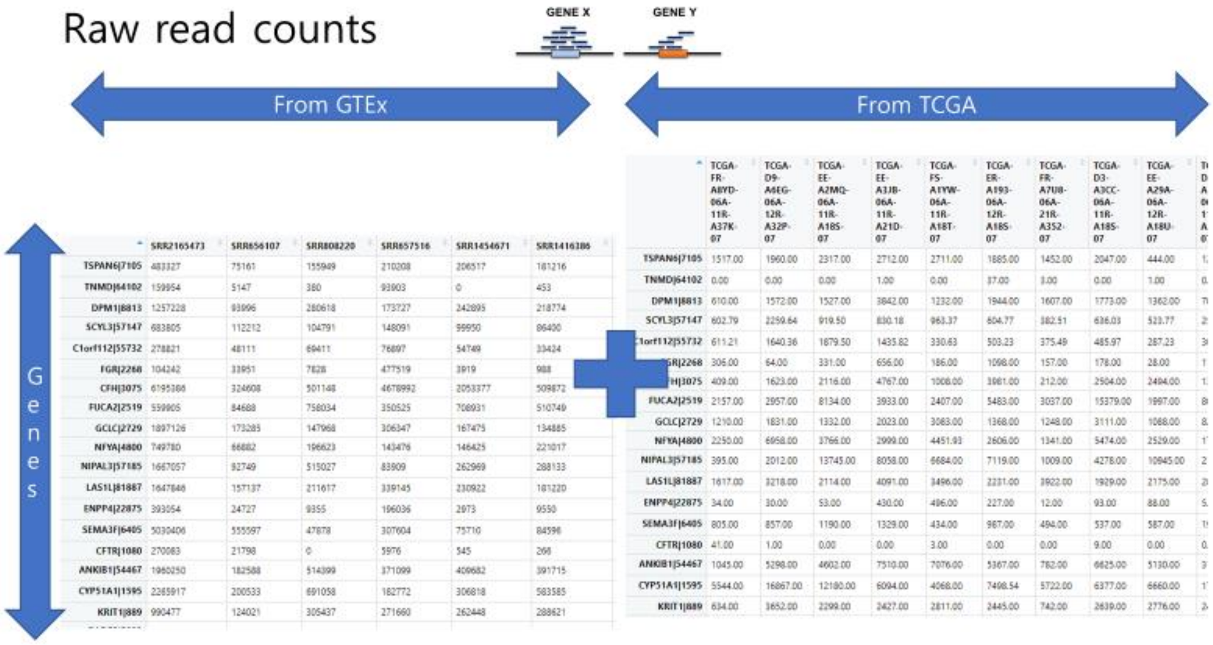
Recount2 project has made gene count data available for the scientific community to download. Recount2 is an online resource consisting of RNA-seq gene and exon counts as well as coverage bigWig files for 2041 different studies. It is the second generation of the ReCount project. The raw sequencing data were processed with Rail-RNA as described in the recount2 paper. The purpose of this function is to get data processed under one single pipeline to avoid any technical variabilities affecting any downstream integrative analysis. The returned object will be a list with elements named "project_tissue" and as a SummarizedExperiment (SE) of gene counts per sample. This is particularly useful to study batch-free data or to increase the pool of healthy/normal samples.

Arguments are:

- **project** string input specifying if it is TCGA data or GTEx.
- **tissue** vector containing tissue(s) of interest to query.

```
#Brain data through Recount2

brain.rec <- TCGAquery_recount2(project = "gtex", tissue = "brain")
```

Raw read counts


GENE X
GENE Y
From GTEx
From TCGA
Genes

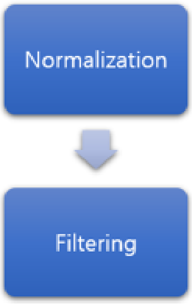
	SRR2165473	SRR656107	SRR808220	SRR657516	SRR1454671	SRR1416386
TSPAN6|7105	483327	75161	155949	210208	206517	181216
TNMD|64102	159954	5147	380	93903	0	453
DPM1|8813	1257228	93996	280618	173727	242895	218774
SCYL3|57147	683805	112212	104791	148091	99950	86400
C1orf112|55732	278821	48111	69411	76897	54749	33424
FGR|2268	104242	33951	7828	477519	3919	988
CFH|3075	6195388	324608	501148	4678992	2053377	509872
FUCA2|2519	559905	84688	758034	350525	708931	510749
GCLC|2729	1897126	173285	147968	306347	167475	134885
NFYA|4800	749780	66882	196623	143476	146425	221017
NIPAL3|57185	1667057	92749	515027	83909	262969	288133
LAS1L|81887	1647848	157137	211617	339145	230922	181220
ENPP4|22875	393054	24727	9355	196036	2973	9550
SEMA3F|6405	5030406	555597	47878	307604	75710	84596
CFTR|1080	270083	21798	0	5976	545	268
ANKIB1|54467	1960250	182588	514399	371099	409682	391715
CYP51A1|1595	2265917	200533	691058	182772	306818	583585
KRIT1|889	990477	124021	305437	271660	262448	288621

	TCGA-FR-A8YD-06A-11R-A37K-07	TCGA-D9-A6EG-06A-12R-A32P-07	TCGA-EE-A2MQ-06A-11R-A18S-07	TCGA-EE-A3JB-06A-11R-A21D-07	TCGA-FS-A1YW-06A-11R-A18T-07	TCGA-ER-A193-06A-12R-A18S-07	TCGA-FR-A7U8-06A-21R-A352-07	TCGA-D3-A3CC-06A-11R-A18S-07	TCGA-EE-A29A-06A-12R-A18U-07	T D A 0 1 A 0
TSPAN6|7105	1517.00	1960.00	2317.00	2712.00	2711.00	1885.00	1452.00	2047.00	444.00	1.
TNMD|64102	0.00	0.00	0.00	1.00	0.00	37.00	3.00	0.00	1.00	0.
DPM1|8813	610.00	1572.00	1527.00	3842.00	1232.00	1944.00	1607.00	1773.00	1362.00	7
SCYL3|57147	602.79	2259.64	919.50	830.18	963.37	604.77	382.51	636.03	523.77	2
C1orf112|55732	611.21	1640.36	1879.50	1435.82	330.63	503.23	375.49	485.97	287.23	3
GR|2268	306.00	64.00	331.00	656.00	186.00	1098.00	157.00	178.00	28.00	1
FH|3075	409.00	1623.00	2116.00	4767.00	1008.00	3981.00	212.00	2504.00	2494.00	1.
FUCA2|2519	2157.00	2957.00	8134.00	3933.00	2407.00	5483.00	3037.00	15379.00	1997.00	8
GCLC|2729	1210.00	1831.00	1332.00	2023.00	3083.00	1368.00	1248.00	3111.00	1068.00	8.
NFYA|4800	2250.00	6958.00	3766.00	2999.00	4451.93	2606.00	1341.00	5474.00	2529.00	1
NIPAL3|57185	395.00	2012.00	13745.00	8058.00	6684.00	7119.00	1009.00	4278.00	10945.00	2
LAS1L|81887	1617.00	3218.00	2114.00	4091.00	3496.00	2231.00	3922.00	1929.00	2175.00	2
ENPP4|22875	34.00	30.00	53.00	430.00	496.00	227.00	12.00	93.00	88.00	5.
SEMA3F|6405	805.00	857.00	1190.00	1329.00	434.00	967.00	494.00	537.00	587.00	1
CFTR|1080	41.00	1.00	0.00	0.00	3.00	0.00	0.00	9.00	0.00	0.
ANKIB1|54467	1045.00	5298.00	4602.00	7510.00	7076.00	5367.00	782.00	6625.00	5130.00	3
CYP51A1|1595	5544.00	16867.00	12180.00	6094.00	4068.00	7498.54	5722.00	6377.00	6660.00	1
KRIT1|889	634.00	3652.00	2299.00	2427.00	2811.00	2445.00	742.00	2639.00	2776.00	2

Normalization

- Read counts need to be properly normalized to accommodate for the following biases and extract meaningful expression estimates:
- Sequencing depth – Higher the sequencing depths, higher the counts
- Gene length - Longer transcripts are expected to generate more reads
- Count distribution

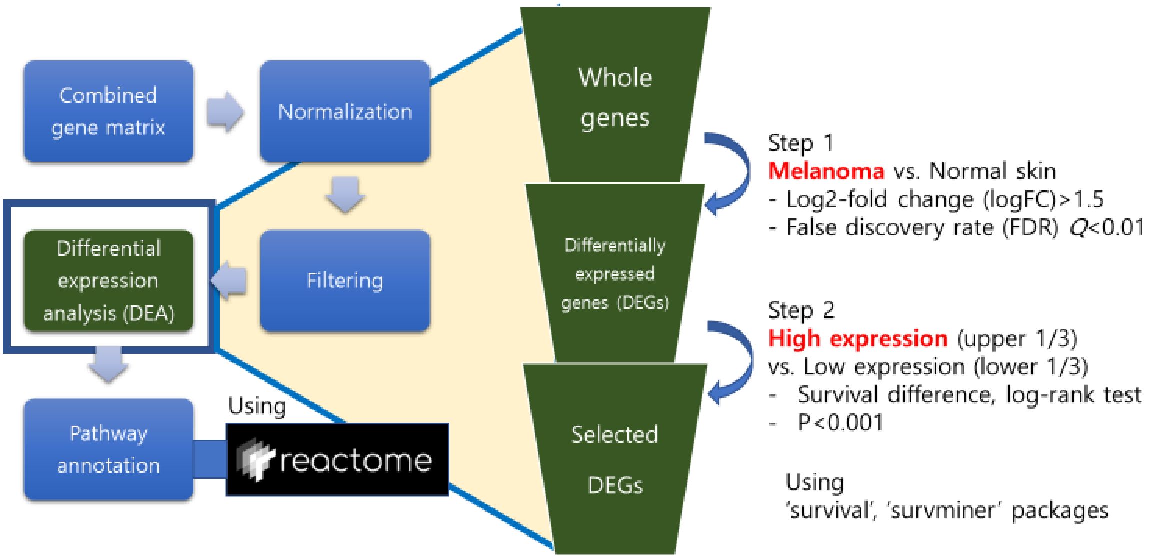
```
dataNorm <- TCGAanalyze_Normalization(tabDF = sumg2, geneInfo = geneInfo)
```



Filtering

```
# quantile filter of genes
dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
                                  method = "quantile",
                                  qnt.cut =  0.25)
```

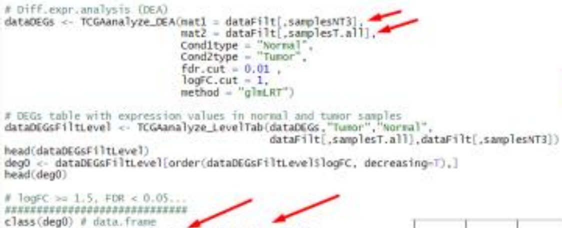
2.3 DEA, survival analysis, and pathway annotation



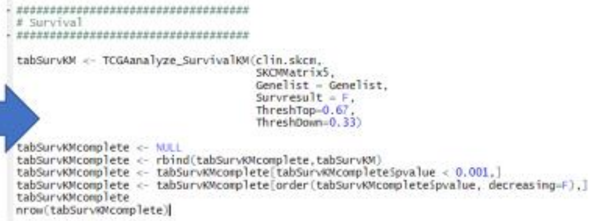
Step 1
Melanoma vs. Normal skin
- Log2-fold change (logFC)>1.5
- False discovery rate (FDR) *Q*<0.01

Step 2
High expression (upper 1/3)
vs. Low expression (lower 1/3)
- Survival difference, log-rank test
- P<0.001

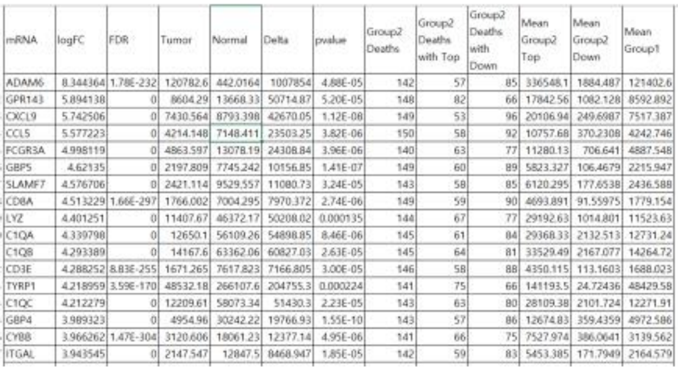
Using
'survival', 'survminer' packages



```
# Diff.expr.analysis (DEA)
dataDEGs <- TCGAanalyze_DEA(mat1 = dataFilt[,samplesNT3],
                            mat2 = dataFilt[,samplesT.all],
                            Cond1type = "Normal",
                            Cond2type = "Tumor",
                            fdr.cut = 0.01 ,
                            logFC.cut = 1,
                            method = "glmLRT")

# DEGs table with expression values in normal and tumor samples
dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGs,"Tumor","Normal",
                                          dataFilt[,samplesT.all],dataFilt[,samplesNT3])
head(dataDEGsFiltLevel)
deg0 <- dataDEGsFiltLevel[order(dataDEGsFiltLevel$logFC, decreasing=T),]
head(deg0)

# logFC >= 1.5, FDR < 0.05...
##############################
class(deg0) # data.frame
deg0.3 <- subset(deg0, logFC >= 1.5 & FDR < 0.05)
```

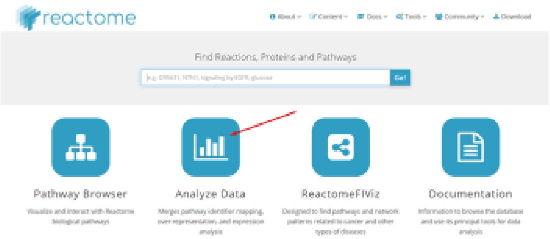


```
#####################################
# Survival
#####################################

tabSurvKM <- TCGAanalyze_SurvivalKM(clin.skcm,
                                    SKCMMatrix5,
                                    Genelist = Genelist,
                                    Survresult = F,
                                    ThreshTop=0.67,
                                    ThreshDown=0.33)

tabSurvKMcomplete <- NULL
tabSurvKMcomplete <- rbind(tabSurvKMcomplete,tabSurvKM)
tabSurvKMcomplete <- tabSurvKMcomplete[tabSurvKMcomplete$pvalue < 0.001,]
tabSurvKMcomplete <- tabSurvKMcomplete[order(tabSurvKMcomplete$pvalue, decreasing=F),]
tabSurvKMcomplete
nrow(tabSurvKMcomplete)
```



mRNA	logFC	FDR	Tumor	Normal	Delta	pvalue	Group2 Deaths	Group2 Deaths with Top	Group2 Deaths with Down	Mean Group2 Top	Mean Group2 Down	Mean Group1
ADAM6	8.344364	1.78E-232	120782.6	442.0164	1007854	4.88E-05	142	57	85	336548.1	1884.487	121402.6
GPR143	5.894138	0	8604.29	13668.33	50714.87	5.20E-05	148	82	66	17842.56	1082.128	8592.892
CXCL9	5.742506	0	7430.564	8793.398	42670.05	1.12E-08	149	53	96	20106.94	249.6987	7517.387
CCL5	5.577223	0	4214.148	7148.411	23503.25	3.82E-06	150	58	92	10757.68	370.2308	4242.746
FCGR3A	4.998119	0	4863.597	13078.19	24308.84	3.96E-06	140	63	77	11280.13	706.641	4887.548
GBP5	4.62135	0	2197.809	7745.242	10156.85	1.41E-07	149	60	89	5823.327	106.4679	2215.947
SLAMF7	4.576706	0	2421.114	9529.557	11080.73	3.24E-05	143	58	85	6120.295	177.6538	2436.588
CD8A	4.513229	1.66E-297	1766.002	7004.295	7970.372	2.74E-06	149	59	90	4693.891	91.55975	1779.154
LYZ	4.401251	0	11407.67	46372.17	50208.02	0.000135	144	67	77	29192.63	1014.801	11523.63
C1QA	4.339798	0	12650.1	56109.26	54898.85	8.46E-06	145	61	84	29368.33	2132.513	12731.24
C1QB	4.293389	0	14167.6	63362.06	60827.03	2.63E-05	145	64	81	33529.49	2167.077	14264.72
CD3E	4.288252	8.83E-255	1671.265	7617.823	7166.805	3.00E-05	146	58	88	4350.115	113.1603	1688.023
TYRP1	4.218959	3.59E-170	48532.18	266107.6	204755.3	0.000224	141	75	66	141193.5	24.72436	48429.58
C1QC	4.212279	0	12209.61	58073.34	51430.3	2.23E-05	143	63	80	28109.38	2101.724	12271.91
GBP4	3.989323	0	4954.96	30242.22	19766.93	1.55E-10	143	57	86	12674.83	359.4359	4972.586
CYBB	3.966262	1.47E-304	3120.606	18061.23	12377.14	4.95E-06	141	66	75	7527.974	386.0641	3139.562
ITGAL	3.943545	0	2147.547	12847.5	8468.947	1.85E-05	142	59	83	5453.385	171.7949	2164.579

Reactome



Find Reactions, Proteins and Pathways

e.g. O95631, NTN1, signaling by EGFR, glucose

Pathway Browser

Visualize and interact with Reactome biological pathways

Analyze Data

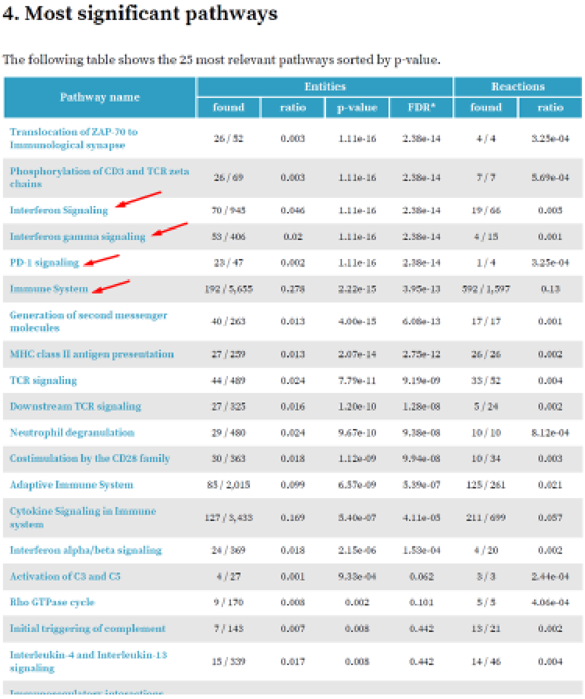
Merges pathway identifier mapping, over-representation, and expression analysis

ReactomeFIViz

Designed to find pathways and network patterns related to cancer and other types of diseases

Documentation

Information to browse the database and use its principal tools for data analysis



4. Most significant pathways

The following table shows the 25 most relevant pathways sorted by p-value.

Pathway name	Entities				Reactions	
	found	ratio	p-value	FDR*	found	ratio
Translocation of ZAP-70 to Immunological synapse	26 / 52	0.003	1.11e-16	2.38e-14	4 / 4	3.25e-04
Phosphorylation of CD3 and TCR zeta chains	26 / 69	0.003	1.11e-16	2.38e-14	7 / 7	5.69e-04
Interferon Signaling	70 / 945	0.046	1.11e-16	2.38e-14	19 / 66	0.005
Interferon gamma signaling	53 / 406	0.02	1.11e-16	2.38e-14	4 / 15	0.001
PD-1 signaling	23 / 47	0.002	1.11e-16	2.38e-14	1 / 4	3.25e-04
Immune System	192 / 5,655	0.278	2.22e-15	3.95e-13	592 / 1,597	0.13
Generation of second messenger molecules	40 / 263	0.013	4.00e-15	6.08e-13	17 / 17	0.001
MHC class II antigen presentation	27 / 259	0.013	2.07e-14	2.75e-12	26 / 26	0.002
TCR signaling	44 / 489	0.024	7.79e-11	9.19e-09	33 / 52	0.004
Downstream TCR signaling	27 / 325	0.016	1.20e-10	1.28e-08	5 / 24	0.002
Neutrophil degranulation	29 / 480	0.024	9.67e-10	9.38e-08	10 / 10	8.12e-04
Costimulation by the CD28 family	30 / 363	0.018	1.12e-09	9.94e-08	10 / 34	0.003
Adaptive Immune System	85 / 2,015	0.099	6.57e-09	5.39e-07	125 / 261	0.021
Cytokine Signaling in Immune system	127 / 3,433	0.169	5.40e-07	4.11e-05	211 / 699	0.057
Interferon alpha/beta signaling	24 / 369	0.018	2.15e-06	1.53e-04	4 / 20	0.002
Activation of C3 and C5	4 / 27	0.001	9.33e-04	0.062	3 / 3	2.44e-04
Rho GTPase cycle	9 / 170	0.008	0.002	0.101	5 / 5	4.06e-04
Initial triggering of complement	7 / 143	0.007	0.008	0.442	13 / 21	0.002
Interleukin-4 and Interleukin-13 signaling	15 / 339	0.017	0.008	0.442	14 / 46	0.004
Immunoregulatory interactions						

2.4 Validation

- Validation set 1
  - Other mRNA database of **melanoma**
  - EMBL–EBI: E-GEOD-65904 (= GEO: GSE65904)
  - 'GEOquery', 'illuminaHumanv4.db' R packages

EMBL-EBI

The home for big data in biology

The European Bioinformatics Institute

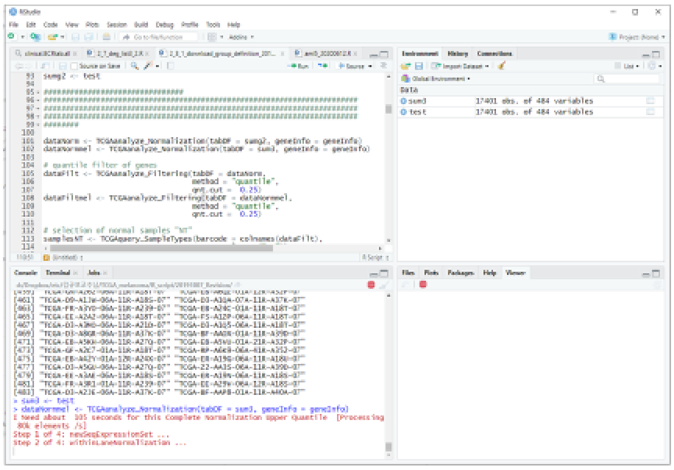
- Validation set 2
  - Other mRNA database of **melanocyte**
  - GEO: 4 projects (SRP022259, SRP039354, SRP057616, and SRP058120)
  - 12 samples
  - 'recount' R packages

GEO

Gene Expression Omnibus

Statistical and computational methods

- R software (version 3.6.1)
- RStudio (version 1.2.1335)



```r
samg2 <- test

################################
#####################################################################
#####################################################################
#####################################################################
########

dataNorm <- TCGAanalyze_Normalization(tabDF = samg2, geneInfo = geneInfo)
dataNormmel <- TCGAanalyze_Normalization(tabDF = sum3, geneInfo = geneInfo)

# quantile filter of genes
dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
                                  method = "quantile",
                                  qnt.cut =  0.25)
dataFiltmel <- TCGAanalyze_Filtering(tabDF = dataNormmel,
                                  method = "quantile",
                                  qnt.cut =  0.25)

# selection of normal samples "NT"
samplesNT <- TCGAquery_SampleTypes(barcode = colnames(dataFilt),
```

```
[461] "TCGA-D9-A1JW-06A-11R-A18S-07" "TCGA-D3-A1Q4-07A-11R-A37K-07"
[463] "TCGA-FR-A3YO-06A-11R-A239-07" "TCGA-EB-A24C-01A-11R-A18T-07"
[465] "TCGA-EE-A2A2-06A-11R-A18T-07" "TCGA-FS-A1ZP-06A-11R-A18T-07"
[467] "TCGA-D3-A3MO-06A-11R-A21D-07" "TCGA-D3-A1Q5-06A-11R-A18T-07"
[469] "TCGA-D3-A8GR-06A-11R-A37K-07" "TCGA-BF-AAOX-01A-11R-A39D-07"
[471] "TCGA-EB-A5KH-06A-11R-A27Q-07" "TCGA-EB-A5VU-01A-21R-A32P-07"
[473] "TCGA-GF-A2C7-01A-11R-A18T-07" "TCGA-BF-A6K9-06A-41R-A352-07"
[475] "TCGA-EB-A42Y-01A-12R-A24X-07" "TCGA-ER-A19G-06A-11R-A18U-07"
[477] "TCGA-D3-A5GU-06A-11R-A27Q-07" "TCGA-Z2-AA3S-06A-11R-A39D-07"
[479] "TCGA-EE-A3AE-06A-11R-A18S-07" "TCGA-ER-A19N-06A-11R-A18S-07"
[481] "TCGA-FR-A3R1-01A-11R-A239-07" "TCGA-EE-A29V-06A-12R-A18S-07"
[483] "TCGA-D3-A2JE-06A-11R-A37K-07" "TCGA-BF-AAP8-01A-11R-A40A-07"
> sum3 <- test
> dataNormmel <- TCGAanalyze_Normalization(tabDF = sum3, geneInfo = geneInfo)
I Need about  105 seconds for this Complete Normalization Upper Quantile  [Processing
 80k elements /s]
Step 1 of 4: newSeqExpressionSet ...
Step 2 of 4: withinLaneNormalization ...
```

Data

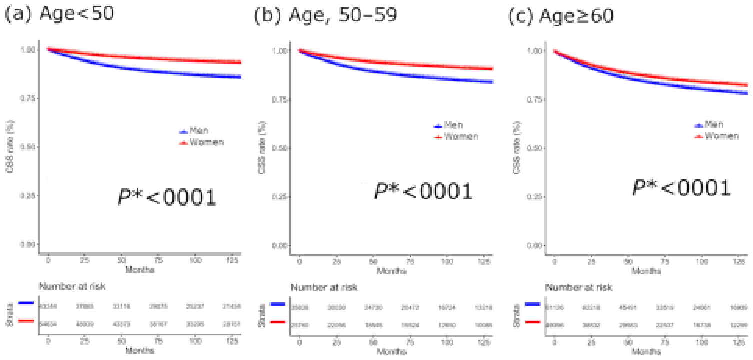
sum3	17401 obs. of 484 variables
test	17401 obs. of 484 variables

Results

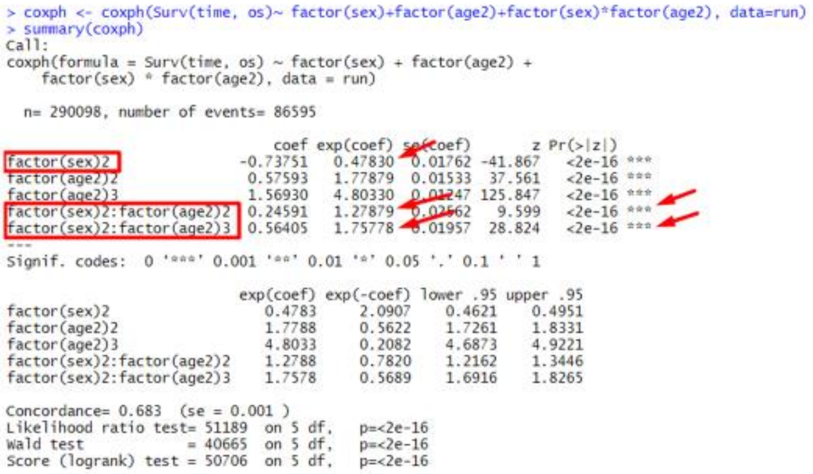
3.1. SEER database analysis

N=290,098

No	exclusion	Contents
371116	16933	C440-449 with histology 8720-8780
293462	77654	sequence number = one primary only or 1st of 2 or more primaries
291797	1665	Pathologically confirmed
291757	40	survival months = unknown 제외
290098	1659	css = missing or unknown 제외

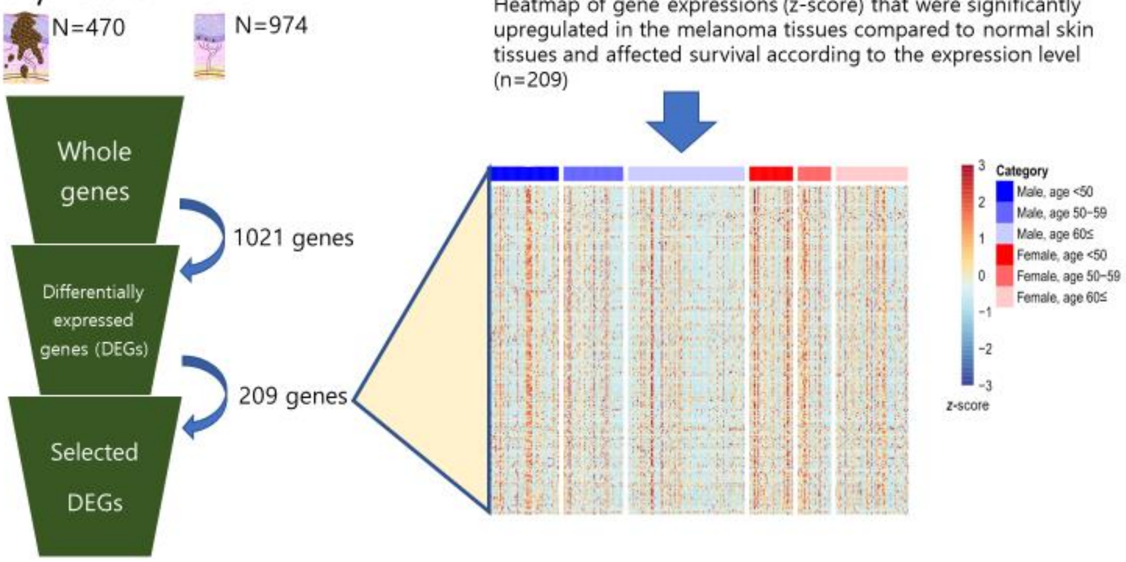

(a) Age<50
(b) Age, 50-59
(c) Age≥60

Sex and the age group, interaction term



```
> coxph <- coxph(Surv(time, os)~ factor(sex)+factor(age2)+factor(sex)*factor(age2), data=run)
> summary(coxph)
Call:
coxph(formula = Surv(time, os) ~ factor(sex) + factor(age2) + 
    factor(sex) * factor(age2), data = run)

  n= 290098, number of events= 86595 

                              coef exp(coef) se(coef)       z Pr(>|z|)    
factor(sex)2              -0.73751   0.47830  0.01762 -41.867   <2e-16 ***
factor(age2)2              0.57593   1.77879  0.01533  37.561   <2e-16 ***
factor(age2)3              1.56930   4.80330  0.01247 125.847   <2e-16 ***
factor(sex)2:factor(age2)2 0.24591   1.27879  0.02562   9.599   <2e-16 ***
factor(sex)2:factor(age2)3 0.56405   1.75778  0.01957  28.824   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                          exp(coef) exp(-coef) lower .95 upper .95
factor(sex)2                 0.4783     2.0907    0.4621    0.4951
factor(age2)2                1.7788     0.5622    1.7261    1.8331
factor(age2)3                4.8033     0.2082    4.6873    4.9221
factor(sex)2:factor(age2)2   1.2788     0.7820    1.2162    1.3446
factor(sex)2:factor(age2)3   1.7578     0.5689    1.6916    1.8265

Concordance= 0.683  (se = 0.001 )
Likelihood ratio test= 51189  on 5 df,   p=<2e-16
Wald test            = 40665  on 5 df,   p=<2e-16
Score (logrank) test = 50706  on 5 df,   p=<2e-16
```

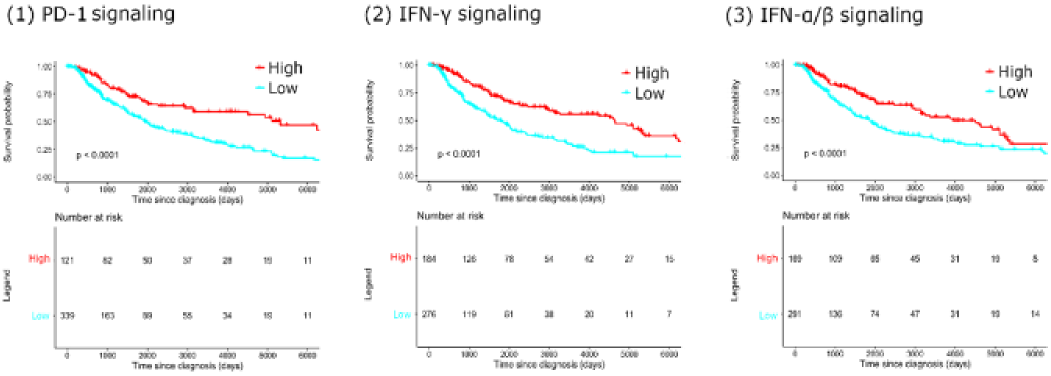
3.2. Genes that significantly impact survival according to the expression level



Heatmap of gene expressions (z-score) that were significantly upregulated in the melanoma tissues compared to normal skin tissues and affected survival according to the expression level (n=209)

Pathway annotation of the 209 genes

➔ Upregulations of **the PD-1, IFN-γ, IFN-α/β pathways** were correlated with favorable survival in melanoma



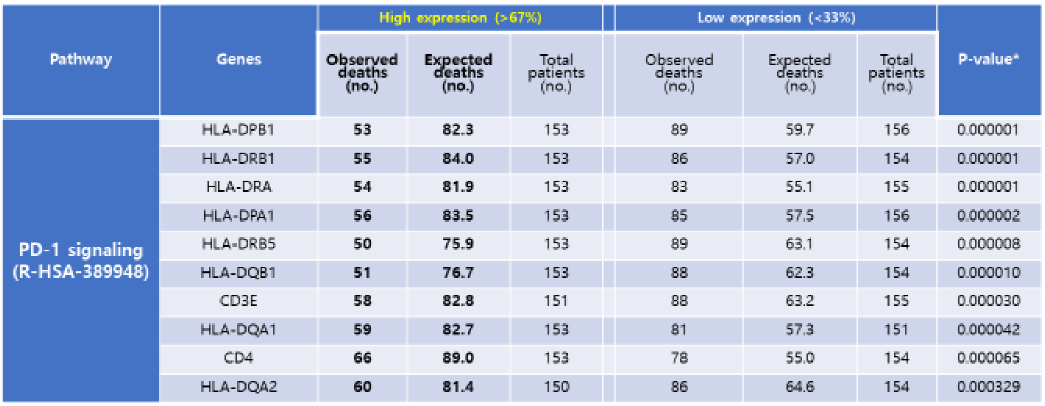
(1) PD-1 signaling

(2) IFN-γ signaling

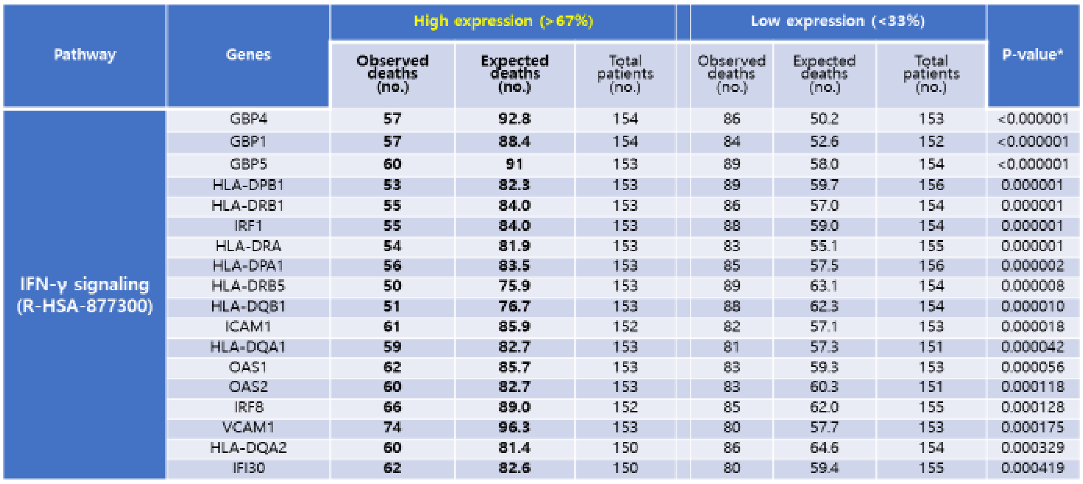
(3) IFN-α/β signaling

* Definition of patients with Upregulated/downregulated signaling
- Upregulated: **1≤ genes** with high expression (top third) of specific signaling.
- Downregulated: **No genes** with high expression of specific signaling.

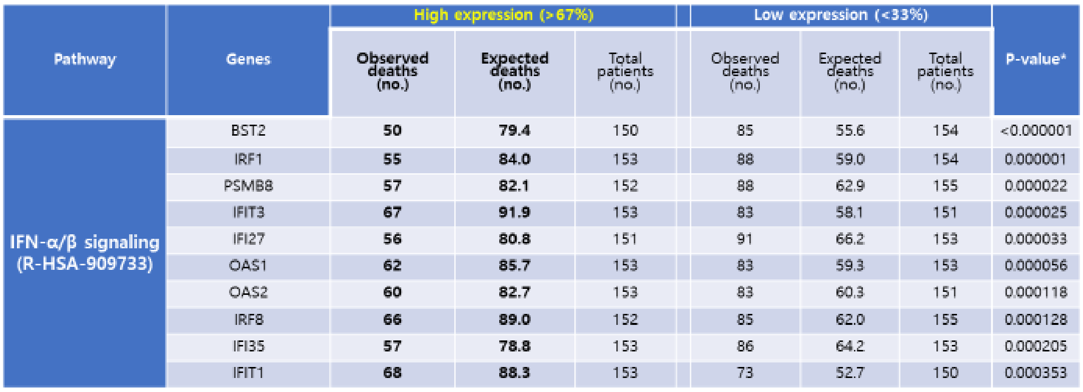
Table 1. Survival comparison according to the expression level of immune pathway-related genes that are upregulated in melanoma tissues than normal skin tissues and affect survival according to the expression level



Pathway	Genes	High expression (>67%)			Low expression (<33%)			P-value*
		Observed deaths (no.)	Expected deaths (no.)	Total patients (no.)	Observed deaths (no.)	Expected deaths (no.)	Total patients (no.)	
PD-1 signaling (R-HSA-389948)	HLA-DPB1	**53**	**82.3**	153	89	59.7	156	0.000001
	HLA-DRB1	**55**	**84.0**	153	86	57.0	154	0.000001
	HLA-DRA	**54**	**81.9**	153	83	55.1	155	0.000001
	HLA-DPA1	**56**	**83.5**	153	85	57.5	156	0.000002
	HLA-DRB5	**50**	**75.9**	153	89	63.1	154	0.000008
	HLA-DQB1	**51**	**76.7**	153	88	62.3	154	0.000010
	CD3E	**58**	**82.8**	151	88	63.2	155	0.000030
	HLA-DQA1	**59**	**82.7**	153	81	57.3	151	0.000042
	CD4	**66**	**89.0**	153	78	55.0	154	0.000065
	HLA-DQA2	**60**	**81.4**	150	86	64.6	154	0.000329

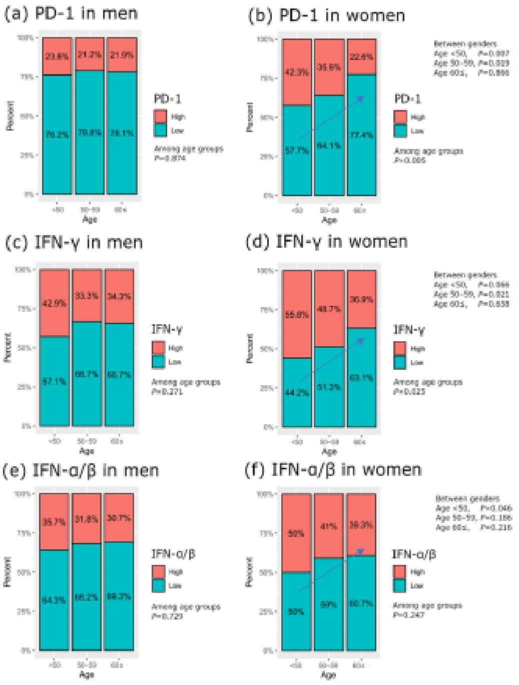


Pathway	Genes	High expression (>67%)			Low expression (<33%)			P-value*
		Observed deaths (no.)	Expected deaths (no.)	Total patients (no.)	Observed deaths (no.)	Expected deaths (no.)	Total patients (no.)	
IFN-γ signaling (R-HSA-877300)	GBP4	**57**	**92.8**	154	86	50.2	153	<0.000001
	GBP1	**57**	**88.4**	154	84	52.6	152	<0.000001
	GBP5	**60**	**91**	153	89	58.0	154	<0.000001
	HLA-DPB1	**53**	**82.3**	153	89	59.7	156	0.000001
	HLA-DRB1	**55**	**84.0**	153	86	57.0	154	0.000001
	IRF1	**55**	**84.0**	153	88	59.0	154	0.000001
	HLA-DRA	**54**	**81.9**	153	83	55.1	155	0.000001
	HLA-DPA1	**56**	**83.5**	153	85	57.5	156	0.000002
	HLA-DRB5	**50**	**75.9**	153	89	63.1	154	0.000008
	HLA-DQB1	**51**	**76.7**	153	88	62.3	154	0.000010
	ICAM1	**61**	**85.9**	152	82	57.1	153	0.000018
	HLA-DQA1	**59**	**82.7**	153	81	57.3	151	0.000042
	OAS1	**62**	**85.7**	153	83	59.3	153	0.000056
	OAS2	**60**	**82.7**	153	83	60.3	151	0.000118
	IRF8	**66**	**89.0**	152	85	62.0	155	0.000128
	VCAM1	**74**	**96.3**	153	80	57.7	153	0.000175
	HLA-DQA2	**60**	**81.4**	150	86	64.6	154	0.000329
	IFI30	**62**	**82.6**	150	80	59.4	155	0.000419



Pathway	Genes	High expression (>67%)			Low expression (<33%)			P-value*
		Observed deaths (no.)	**Expected deaths (no.)**	Total patients (no.)	Observed deaths (no.)	Expected deaths (no.)	Total patients (no.)	
IFN-α/β signaling (R-HSA-909733)	BST2	**50**	**79.4**	150	85	55.6	154	<0.000001
	IRF1	**55**	**84.0**	153	88	59.0	154	0.000001
	PSMB8	**57**	**82.1**	152	88	62.9	155	0.000022
	IFIT3	**67**	**91.9**	153	83	58.1	151	0.000025
	IFI27	**56**	**80.8**	151	91	66.2	153	0.000033
	OAS1	**62**	**85.7**	153	83	59.3	153	0.000056
	OAS2	**60**	**82.7**	153	83	60.3	151	0.000118
	IRF8	**66**	**89.0**	152	85	62.0	155	0.000128
	IFI35	**57**	**78.8**	153	86	64.2	153	0.000205
	IFIT1	**68**	**88.3**	153	73	52.7	150	0.000353

3.3. Immune signaling according to age and gender



(a) PD-1 in men

(b) PD-1 in women

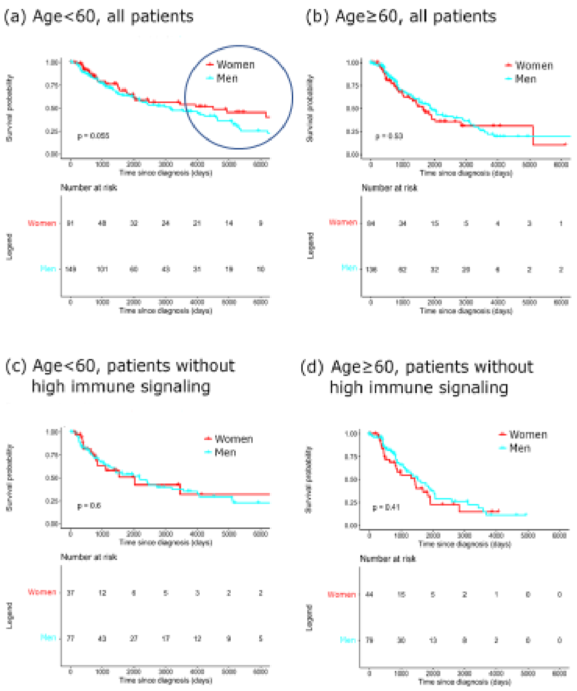
(c) IFN-γ in men

(d) IFN-γ in women

(e) IFN-α/β in men

(f) IFN-α/β in women

3.4. Survival comparison according to age, gender, and immune signaling



(a) Age<60, all patients

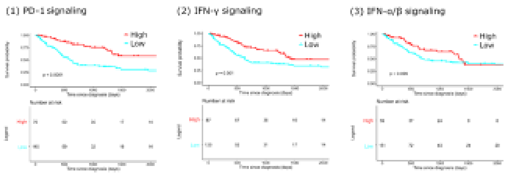
(b) Age≥60, all patients

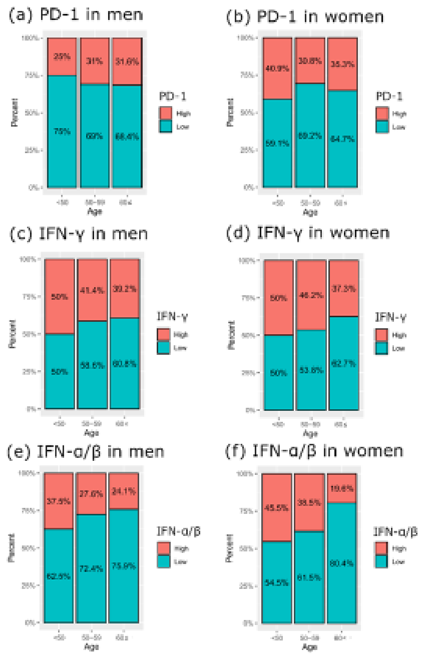
(c) Age<60, patients without high immune signaling

(d) Age≥60, patients without high immune signaling

3.5 Validation

- Similar results.





(a) PD-1 in men

(b) PD-1 in women

(c) IFN-γ in men

(d) IFN-γ in women

(e) IFN-α/β in men

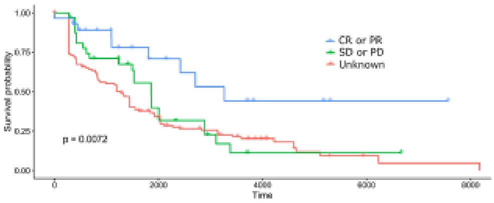
(f) IFN-α/β in women

3.6 Genes that significantly impact response to radiotherapy according to the expression level

- Of the 470 melanoma patients in the TCGA database, 117 patients underwent radiotherapy(Table 2).

Characteristics	Male (N=78)	Female (N=39)	P value
Radiation response			0.103
Complete response	10 (12.8%)	9 (23.1%)	
Partial response	3 (3.8%)	3 (7.7%)	
Stable disease	1 (1.3%)	3 (7.7%)	
Progressive disease	13 (16.7%)	7 (17.9%)	
Unknown	51 (65.4%)	17 (43.6%)	

- Patients with CR or PR showed more favorable survival than patients with SD or PD (Fig. 5).



- Female and younger patients were more likely to have a CR or PR after radiotherapy than those who are male and older patients, respectively (Table 3-4).

Table 3. Response to radiotherapy according to sex in melanoma of the TCGA database.
Abbreviations: CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.
*Pearson's Chi-squared test.

Response to radiotherapy	Age <50 (N=20)	Age 50-59 (N=9)	Age 60≤ (N=19)	P value*
CR or PR	11 (55.0%)	4 (44.4%)	9 (47.4%)	0.834
SD or PD	9 (45.0%)	5 (55.6%)	10 (52.6%)	

Table 4. Response to radiotherapy according to age in melanoma of the TCGA database.
Abbreviations: CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.
*Pearson's Chi-squared test.

Response to radiotherapy	Male (N=27)	Female (N=22)	P value*
CR or PR	13 (48.1%)	12 (54.5%)	0.936
SD or PD	14 (51.9%)	10 (45.5%)	

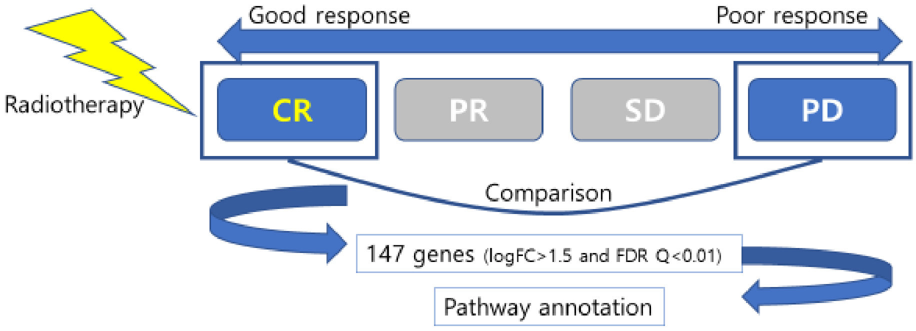


Table 5. Significantly upregulated pathways in melanoma patients with complete response after radiotherapy than those with progressive disease after radiotherapy in the TCGA database. Abbreviation: FDR, false discovery rate.

Pathway identifier	Pathway name	Entities FDR	Submitted entities found
R-HAS-6805567	Keratinization	4.86E-14	KLK5;KRT23;LCE2D;CASP14;LIPM;PI3;KRT6B;CDSN;KRT2;KRT1;KRT79;KRT77;KRT10;LOR;LCE2B;FLG2;LCE2C;KRT17;KRT14;LCE6A;PKP2;DSG1;PKP1;PKP3;DSC1
R-HAS-6809371	Formation of the cornified envelope	4.86E-14	KLK5;KRT23;LCE2D;CASP14;LIPM;PI3;KRT6B;CDSN;KRT2;KRT1;KRT79;KRT77;KRT10;LOR;LCE2B;FLG2;LCE2C;KRT17;KRT14;LCE6A;PKP2;DSG1;PKP1;PKP3;DSC1
R-HAS-6798695	Neutrophil degranulation	0.025264	GRIA2;ARG1;KRT1;HP;MMP9;FLG2;SLPI;SCNN1B;DSG1;PKP1;APOB;DSC1;ELANE;LTF;PLIN5;S100A7

Table 6. Response to radiotherapy according to neutrophil degranulation signals in melanoma patients of the TCGA database. Abbreviations: CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.
*Pearson's Chi-squared test.

Response to radiotherapy	High neutrophil signals (N=6)	Low neutrophil signals (N=55)	P value*
CR or PR	4 (100.0%)	21 (46.7%)	0.128
SD or PD	0 (0.0%)	24 (53.3%)	

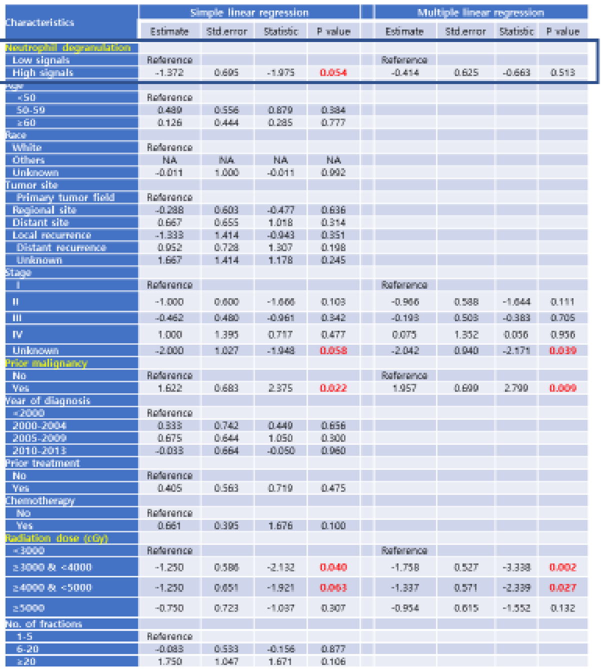
Table 7. Neutrophil degranulation signaling according to sex in melanoma patients of the TCGA database.
*Pearson's Chi-squared test.

Response to radiotherapy	Male (N=78)	Female (N=39)	P value*
High neutrophil signals	10 (12.8%)	6 (15.4%)	0.924
Low neutrophil signals	68 (87.2%)	33 (84.6%)	

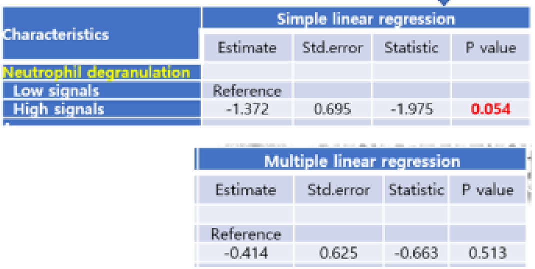
Table 8. Neutrophil degranulation signaling according to age in melanoma patients of the TCGA database.
*Pearson's Chi-squared test.

Response to radiotherapy	<50 years (N=47)	50-59 years (N=19)	≥60 years (N=49)	P value*
High neutrophil signals	4 (8.5%)	1 (5.3%)	9 (18.4%)	0.202
Low neutrophil signals	43 (91.5%)	18 (94.7%)	40 (81.6%)	

Table 9. Simple and multiple linear regressions for the radiation response in melanoma patients in the TCGA database.



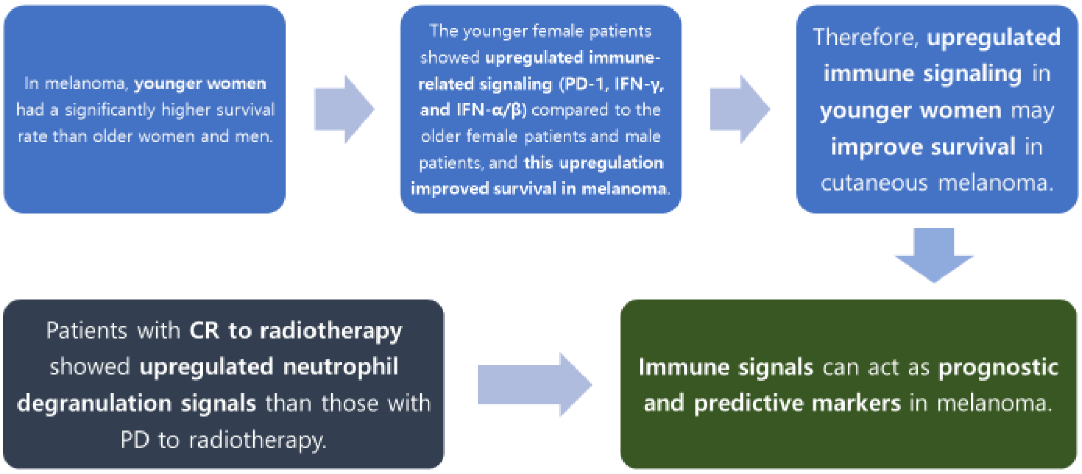
Characteristics	Simple linear regression				Multiple linear regression			
	Estimate	Std.error	Statistic	P value	Estimate	Std.error	Statistic	P value
Neutrophil degranulation								
Low signals	Reference				Reference			
High signals	-1.372	0.695	-1.975	0.054	-0.414	0.625	-0.663	0.513
Age								
<50	Reference							
50-59	0.489	0.556	0.879	0.384				
≥60	0.126	0.444	0.285	0.777				
Race								
White	Reference							
Others	NA	NA	NA	NA				
Unknown	-0.011	1.000	-0.011	0.992				
Tumor site								
Primary tumor field	Reference							
Regional site	-0.288	0.603	-0.477	0.636				
Distant site	0.667	0.655	1.018	0.314				
Local recurrence	-1.333	1.414	-0.943	0.351				
Distant recurrence	0.952	0.728	1.307	0.198				
Unknown	1.667	1.414	1.178	0.245				
Stage								
I	Reference				Reference			
II	-1.000	0.600	-1.666	0.103	-0.966	0.588	-1.644	0.111
III	-0.462	0.480	-0.961	0.342	-0.193	0.503	-0.383	0.705
IV	1.000	1.395	0.717	0.477	0.075	1.352	0.056	0.956
Unknown	-2.000	1.027	-1.948	0.058	-2.042	0.940	-2.171	0.039
Prior malignancy								
No	Reference				Reference			
Yes	1.622	0.683	2.375	0.022	1.957	0.699	2.799	0.009
Year of diagnosis								
<2000	Reference							
2000-2004	0.333	0.742	0.449	0.656				
2005-2009	0.675	0.644	1.050	0.300				
2010-2013	-0.033	0.664	-0.050	0.960				
Prior treatment								
No	Reference							
Yes	0.405	0.563	0.719	0.475				
Chemotherapy								
No	Reference							
Yes	0.661	0.395	1.676	0.100				
Radiation dose (cGy)								
<3000	Reference				Reference			
≥3000 & <4000	-1.250	0.586	-2.132	0.040	-1.758	0.527	-3.338	0.002
≥4000 & <5000	-1.250	0.651	-1.921	0.063	-1.337	0.571	-2.339	0.027
≥5000	-0.750	0.723	-1.037	0.307	-0.954	0.615	-1.552	0.132
No. of fractions								
1-5	Reference							
6-20	-0.083	0.533	-0.156	0.877				
≥20	1.750	1.047	1.671	0.106				



Characteristics	Simple linear regression			
	Estimate	Std.error	Statistic	P value
Neutrophil degranulation				
Low signals	Reference			
High signals	-1.372	0.695	-1.975	0.054

Multiple linear regression			
Estimate	Std.error	Statistic	P value
Reference			
-0.414	0.625	-0.663	0.513

Conclusions



In melanoma, **younger women** had a significantly higher survival rate than older women and men.

The younger female patients showed **upregulated immune-related signaling (PD-1, IFN-γ, and IFN-α/β)** compared to the older female patients and male patients, and **this upregulation improved survival in melanoma**.

Therefore, **upregulated immune signaling** in **younger women** may **improve survival** in cutaneous melanoma.

Patients with **CR to radiotherapy** showed **upregulated neutrophil degranulation signals** than those with PD to radiotherapy.

Immune signals can act as **prognostic and predictive markers** in melanoma.

Thank you for your attention!

논문 예제를 통한 실습

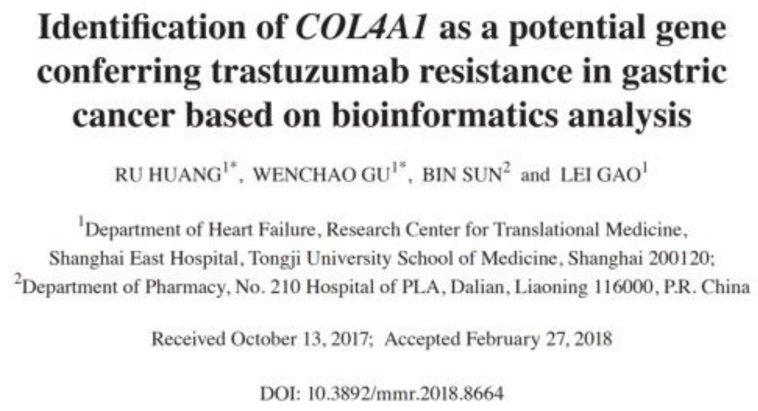
- 논문: Huang R, Gu W, Sun B, Gao L. Identification of COL4A1 as a potential gene conferring trastuzumab resistance in gastric cancer based on bioinformatics analysis. Mol Med Rep. 2018 May;17(5):6387-6396. doi: 10.3892/mmr.2018.8664. Epub 2018 Mar 1. PMID: 29512712; PMCID: PMC5928613.

논문 예제를 통한 실습

이화여대부속목동병원
융합의학연구원
김이준
kimyj.ro@gmail.com

2020-11-04

Identification of COL4A1 as a potential gene conferring trastuzumab resistance in gastric cancer based on bioinformatics analysis



Identification of *COL4A1* as a potential gene conferring trastuzumab resistance in gastric cancer based on bioinformatics analysis

RU HUANG[1*], WENCHAO GU[1*], BIN SUN[2] and LEI GAO[1]

[1]Department of Heart Failure, Research Center for Translational Medicine,
Shanghai East Hospital, Tongji University School of Medicine, Shanghai 200120;
[2]Department of Pharmacy, No. 210 Hospital of PLA, Dalian, Liaoning 116000, P.R. China

Received October 13, 2017; Accepted February 27, 2018

DOI: 10.3892/mmr.2018.8664

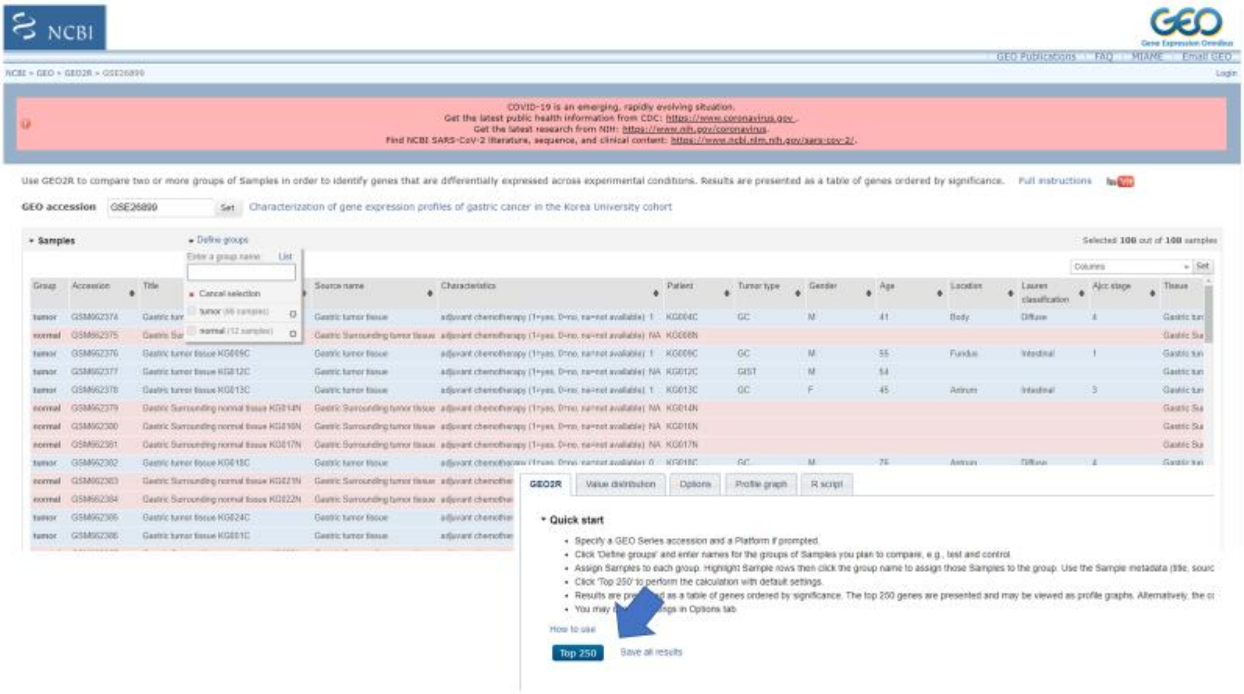
Materials and methods

- Microarray data.
- The gene expression profiles of GSE26899, GSE77346, GSE54129, and GSE65801 were obtained from the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo). In detail, GSE26899 dataset is consisted of 96 clinical gastric tumor tissues and 12 adjacent normal tissues; GSE77346 dataset is consisted of 1 trastuzumab-sensitive cell line and 4 trastuzumab-resistant cell lines (12); GSE54129 includes 111 human gastric cancer tissues and 21 non-cancerous tissues; GSE65801 contains 32 gastric cancer tissues and 32 paired non-cancerous tissues (13).

- Processing of microarray data.

- The raw microarray data files of the datasets downloaded from the GEO website were subsequently analyzed via using the GEO2R (www.ncbi.nlm.nih.gov/geo/geo2r/), an online tool comparing two or more groups of samples in the same experimental setting (14).

- False Discovery Rate (FDR) of P-value adjusted (adj. P) to 0.05 and |logFC|>1 were set as the cut-off criteria.



GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > GEO2R > GSE26899

Login

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. Full instructions

GEO accession: GSE26899 Set — Characterization of gene expression profiles of gastric cancer in the Korea University cohort

Samples — Define groups — Selected 108 out of 108 samples

Enter a group name: List

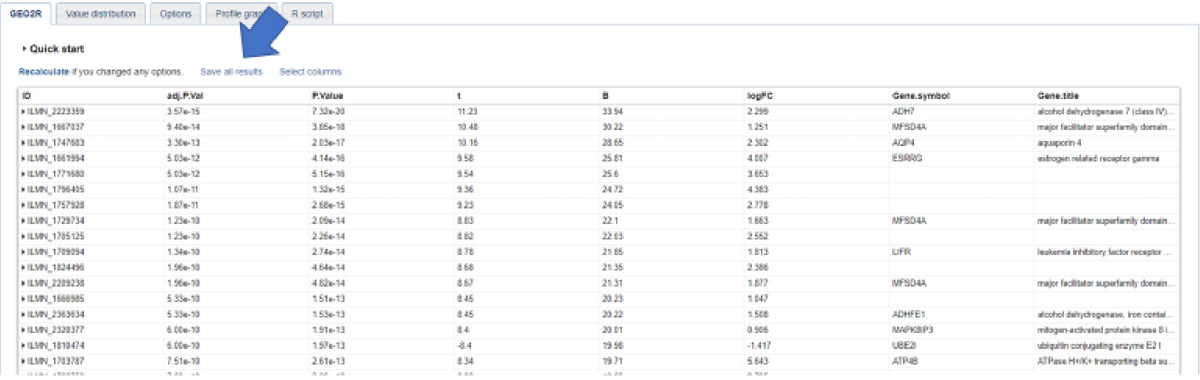
Cancel selection
tumor (96 samples)
normal (12 samples)

Columns — Set

Group	Accession	Title	Source name	Characteristics	Patient	Tumor type	Gender	Age	Location	Lauren classification	Ajcc stage	Tissue
tumor	GSM662374	Gastric tum	Gastric tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): 1	KG004C	GC	M	41	Body	Diffuse	4	Gastric tun
normal	GSM662375	Gastric Su	Gastric Surrounding tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): NA	KG008N							Gastric Su
tumor	GSM662376	Gastric tumor tissue KG009C	Gastric tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): 1	KG009C	GC	M	55	Fundus	Intestinal	1	Gastric tun
tumor	GSM662377	Gastric tumor tissue KG012C	Gastric tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): NA	KG012C	GIST	M	54				Gastric tun
tumor	GSM662378	Gastric tumor tissue KG013C	Gastric tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): 1	KG013C	GC	F	45	Antrum	Intestinal	3	Gastric tun
normal	GSM662379	Gastric Surrounding normal tissue KG014N	Gastric Surrounding tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): NA	KG014N							Gastric Su
normal	GSM662380	Gastric Surrounding normal tissue KG016N	Gastric Surrounding tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): NA	KG016N							Gastric Su
normal	GSM662381	Gastric Surrounding normal tissue KG017N	Gastric Surrounding tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): NA	KG017N							Gastric Su
tumor	GSM662382	Gastric tumor tissue KG018C	Gastric tumor tissue	adjuvant chemotherapy (1=yes, 0=no, na=not available): 0	KG018C	GC	M	75	Antrum	Diffuse	4	Gastric tun
normal	GSM662383	Gastric Surrounding normal tissue KG021N	Gastric Surrounding tumor tissue	adjuvant chemother								
normal	GSM662384	Gastric Surrounding normal tissue KG022N	Gastric Surrounding tumor tissue	adjuvant chemother								
tumor	GSM662385	Gastric tumor tissue KG024C	Gastric tumor tissue	adjuvant chemother								
tumor	GSM662386	Gastric tumor tissue KG001C	Gastric tumor tissue	adjuvant chemother								

Quick start

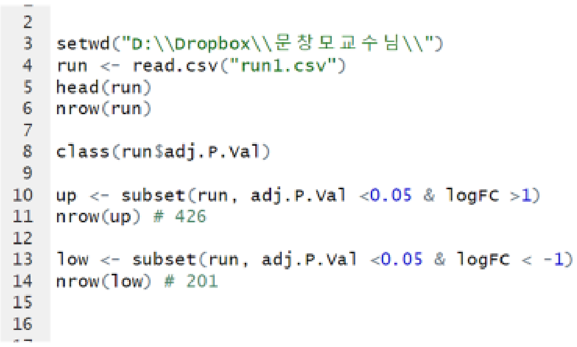
- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, sourc
- Click 'Top 250' to perform the calculation with default settings.
- Results are pre d as a table of genes ordered by significance. The top 250 genes are presented and may be viewed as profile graphs. Alternatively, the c
- You may ngs in Options tab.



▸ Quick start

Recalculate if you changed any options.   Save all results   Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
▸ ILMN_2223359	3.57e-15	7.32e-20	11.23	33.94	2.299	ADH7	alcohol dehydrogenase 7 (class IV)...
▸ ILMN_1667037	9.40e-14	3.85e-18	10.48	30.22	1.251	MFSD4A	major facilitator superfamily domain...
▸ ILMN_1747683	3.30e-13	2.03e-17	10.16	28.65	2.302	AQP4	aquaporin 4
▸ ILMN_1661994	5.03e-12	4.14e-16	9.58	25.81	4.007	ESRRG	estrogen related receptor gamma
▸ ILMN_1771680	5.03e-12	5.15e-16	9.54	25.6	3.653		
▸ ILMN_1796405	1.07e-11	1.32e-15	9.36	24.72	4.383		
▸ ILMN_1757928	1.87e-11	2.68e-15	9.23	24.05	2.778		
▸ ILMN_1729734	1.23e-10	2.09e-14	8.83	22.1	1.663	MFSD4A	major facilitator superfamily domain...
▸ ILMN_1705125	1.23e-10	2.26e-14	8.82	22.03	2.552		
▸ ILMN_1709094	1.34e-10	2.74e-14	8.78	21.85	1.813	LIFR	leukemia inhibitory factor receptor ...
▸ ILMN_1824496	1.96e-10	4.64e-14	8.68	21.35	2.386		
▸ ILMN_2209238	1.96e-10	4.82e-14	8.67	21.31	1.877	MFSD4A	major facilitator superfamily domain...
▸ ILMN_1666985	5.33e-10	1.51e-13	8.45	20.23	1.047		
▸ ILMN_2363634	5.33e-10	1.53e-13	8.45	20.22	1.508	ADHFE1	alcohol dehydrogenase, iron contai...
▸ ILMN_2320377	6.00e-10	1.91e-13	8.4	20.01	0.906	MAPK8IP3	mitogen-activated protein kinase 8 i...
▸ ILMN_1810474	6.00e-10	1.97e-13	-8.4	19.98	-1.417	UBE2I	ubiquitin conjugating enzyme E2 I
▸ ILMN_1703787	7.51e-10	2.61e-13	8.34	19.71	5.643	ATP4B	ATPase H+/K+ transporting beta su...

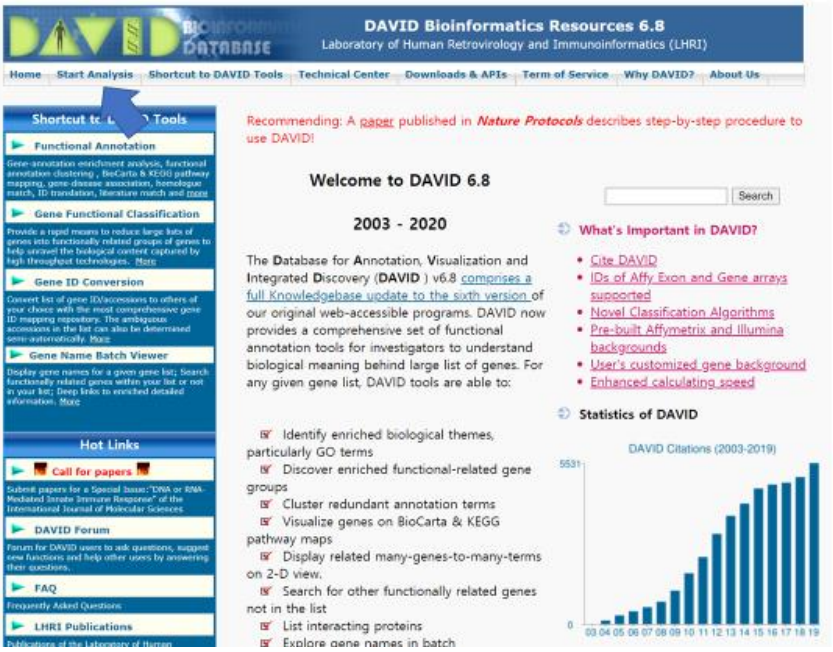


```

setwd("D:\\Dropbox\\문창모교수님\\")
run <- read.csv("run1.csv")
head(run)
nrow(run)

class(run$adj.P.Val)

up <- subset(run, adj.P.Val <0.05 & logFC >1)
nrow(up) # 426

low <- subset(run, adj.P.Val <0.05 & logFC < -1)
nrow(low) # 201


```

- Functional and pathway enrichment analyses.

- Gene ontology (GO) analysis is a commonly used approach for functional studies with three ontologies including biological process, molecular function, and cellular component (15), while Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge base for the systematic study of gene functions (16).

- To study the functional annotations of differentially expressed genes (DEGs), we next employed Database for Annotation, Visualization and Integrated Discovery (DAVID, david.abcc.ncifcrf.gov/,) to process the GO and KEGG analyses of DEGs identified in gastric cancer samples.



DAVID Bioinformatics Resources 6.8

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Hot Links

Call for papers

Submit papers for a Special Issue:"DNA or RNA-Mediated Innate Immune Response" of the International Journal of Molecular Sciences

DAVID Forum

Forum for DAVID users to ask questions, suggest new functions and help other users by answering their questions.

FAQ

Frequently Asked Questions

LHRI Publications

Publications of the Laboratory of Human

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2020

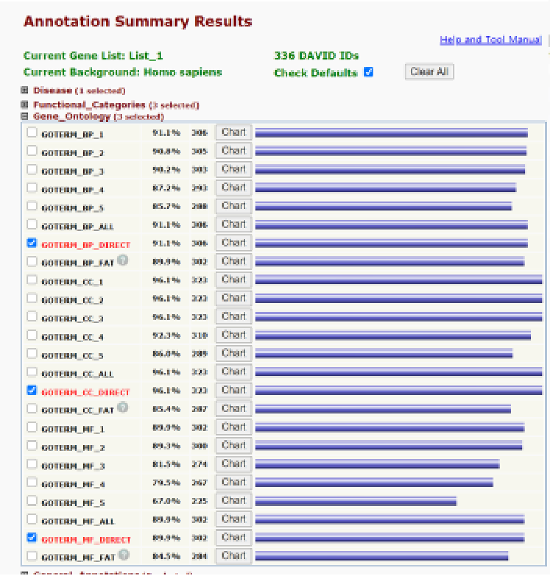
The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch

What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID



Annotation Summary Results

Help and Tool Manual

Current Gene List: List_1 336 DAVID IDs

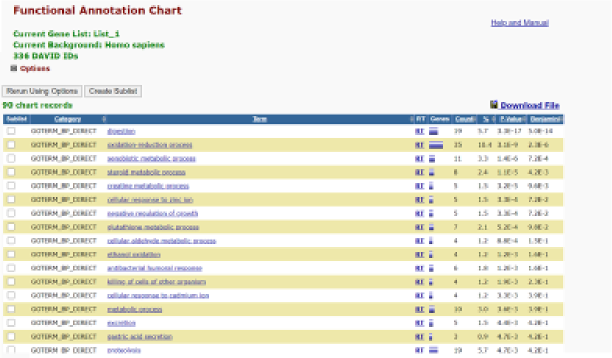
Current Background: Homo sapiens Check Defaults ☑ Clear All

⊞ Disease (1 selected)

⊞ Functional_Categories (3 selected)

⊟ Gene_Ontology (3 selected)

☐	GOTERM_BP_1	91.1%	306	Chart
☐	GOTERM_BP_2	90.8%	305	Chart
☐	GOTERM_BP_3	90.2%	303	Chart
☐	GOTERM_BP_4	87.2%	293	Chart
☐	GOTERM_BP_5	85.7%	288	Chart
☐	GOTERM_BP_ALL	91.1%	306	Chart
☑	GOTERM_BP_DIRECT	91.1%	306	Chart
☐	GOTERM_BP_FAT	89.9%	302	Chart
☐	GOTERM_CC_1	96.1%	323	Chart
☐	GOTERM_CC_2	96.1%	323	Chart
☐	GOTERM_CC_3	96.1%	323	Chart
☐	GOTERM_CC_4	92.3%	310	Chart
☐	GOTERM_CC_5	86.0%	289	Chart
☐	GOTERM_CC_ALL	96.1%	323	Chart
☑	GOTERM_CC_DIRECT	96.1%	323	Chart
☐	GOTERM_CC_FAT	85.4%	287	Chart
☐	GOTERM_MF_1	89.9%	302	Chart
☐	GOTERM_MF_2	89.3%	300	Chart
☐	GOTERM_MF_3	81.5%	274	Chart
☐	GOTERM_MF_4	79.5%	267	Chart
☐	GOTERM_MF_5	67.0%	225	Chart
☐	GOTERM_MF_ALL	89.9%	302	Chart
☑	GOTERM_MF_DIRECT	89.9%	302	Chart
☐	GOTERM_MF_FAT	84.5%	284	Chart



Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
336 DAVID IDs

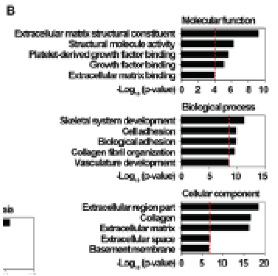
⊞ Options

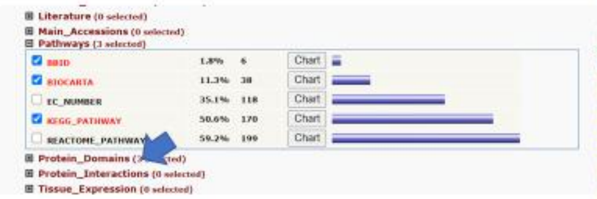
Rerun Using Options | Create Sublist

90 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_BP_DIRECT	digestion	RT		19	5.7	3.3E-17	5.0E-14
☐	GOTERM_BP_DIRECT	oxidation-reduction process	RT		35	10.4	3.1E-9	2.3E-6
☐	GOTERM_BP_DIRECT	xenobiotic metabolic process	RT		11	3.3	1.4E-6	7.2E-4
☐	GOTERM_BP_DIRECT	steroid metabolic process	RT		8	2.4	1.1E-5	4.2E-3
☐	GOTERM_BP_DIRECT	creatine metabolic process	RT		5	1.5	3.2E-5	9.6E-3
☐	GOTERM_BP_DIRECT	cellular response to zinc ion	RT		5	1.5	3.3E-4	7.2E-2
☐	GOTERM_BP_DIRECT	negative regulation of growth	RT		5	1.5	3.3E-4	7.2E-2
☐	GOTERM_BP_DIRECT	glutathione metabolic process	RT		7	2.1	5.2E-4	9.0E-2
☐	GOTERM_BP_DIRECT	cellular aldehyde metabolic process	RT		4	1.2	8.8E-4	1.5E-1
☐	GOTERM_BP_DIRECT	ethanol oxidation	RT		4	1.2	1.2E-3	1.6E-1
☐	GOTERM_BP_DIRECT	antibacterial humoral response	RT		6	1.8	1.2E-3	1.6E-1
☐	GOTERM_BP_DIRECT	killing of cells of other organism	RT		4	1.2	1.9E-3	2.3E-1
☐	GOTERM_BP_DIRECT	cellular response to cadmium ion	RT		4	1.2	3.3E-3	3.9E-1
☐	GOTERM_BP_DIRECT	metabolic process	RT		10	3.0	3.6E-3	3.9E-1
☐	GOTERM_BP_DIRECT	excretion	RT		5	1.5	4.4E-3	4.2E-1
☐	GOTERM_BP_DIRECT	gastric acid secretion	RT		3	0.9	4.7E-3	4.2E-1
☐	GOTERM_BP_DIRECT	proteolysis	RT		19	5.7	4.7E-3	4.2E-1





⊞ Literature (0 selected)

⊞ Main_Accessions (0 selected)

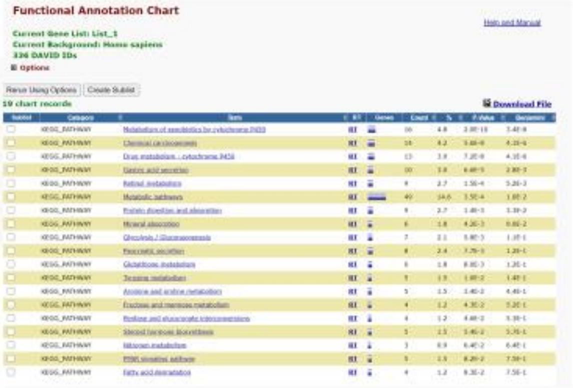
⊟ Pathways (3 selected)

☑	BBID	1.8%	6	Chart
☑	BIOCARTA	11.3%	38	Chart
☐	EC_NUMBER	35.1%	118	Chart
☑	KEGG_PATHWAY	50.6%	170	Chart
☐	REACTOME_PATHWAY	59.2%	199	Chart

⊞ Protein_Domains (3 selected)

⊞ Protein_Interactions (0 selected)

⊞ Tissue_Expression (0 selected)



Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
336 DAVID IDs

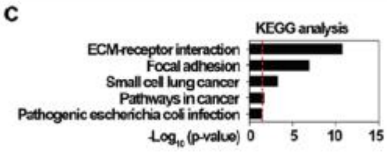
⊞ Options

Rerun Using Options | Create Sublist

19 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		16	4.8	2.0E-11	3.4E-8
☐	KEGG_PATHWAY	Chemical carcinogenesis	RT		14	4.2	5.6E-8	4.2E-6
☐	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT		13	3.9	7.2E-8	4.1E-6
☐	KEGG_PATHWAY	Gastric acid secretion	RT		10	3.0	6.4E-5	2.8E-3
☐	KEGG_PATHWAY	Retinol metabolism	RT		9	2.7	1.5E-4	5.2E-3
☐	KEGG_PATHWAY	Metabolic pathways	RT		49	14.6	3.5E-4	1.0E-2
☐	KEGG_PATHWAY	Protein digestion and absorption	RT		9	2.7	1.4E-3	3.3E-2
☐	KEGG_PATHWAY	Mineral absorption	RT		6	1.8	4.2E-3	8.8E-2
☐	KEGG_PATHWAY	Glycolysis / Gluconeogenesis	RT		7	2.1	5.8E-3	1.1E-1
☐	KEGG_PATHWAY	Pancreatic secretion	RT		8	2.4	7.7E-3	1.2E-1
☐	KEGG_PATHWAY	Glutathione metabolism	RT		6	1.8	8.8E-3	1.2E-1
☐	KEGG_PATHWAY	Tyrosine metabolism	RT		5	1.5	1.0E-2	1.4E-1
☐	KEGG_PATHWAY	Arginine and proline metabolism	RT		5	1.5	1.4E-2	4.4E-1
☐	KEGG_PATHWAY	Fructose and mannose metabolism	RT		4	1.2	4.3E-2	5.2E-1
☐	KEGG_PATHWAY	Pentose and glucuronate interconversions	RT		4	1.2	4.4E-2	5.3E-1
☐	KEGG_PATHWAY	Steroid hormone biosynthesis	RT		5	1.5	5.4E-2	5.7E-1
☐	KEGG_PATHWAY	Nitrogen metabolism	RT		3	0.9	6.4E-2	6.4E-1
☐	KEGG_PATHWAY	PPAR signaling pathway	RT		5	1.5	8.2E-2	7.5E-1
☐	KEGG_PATHWAY	Fatty acid degradation	RT		4	1.2	9.3E-2	7.5E-1

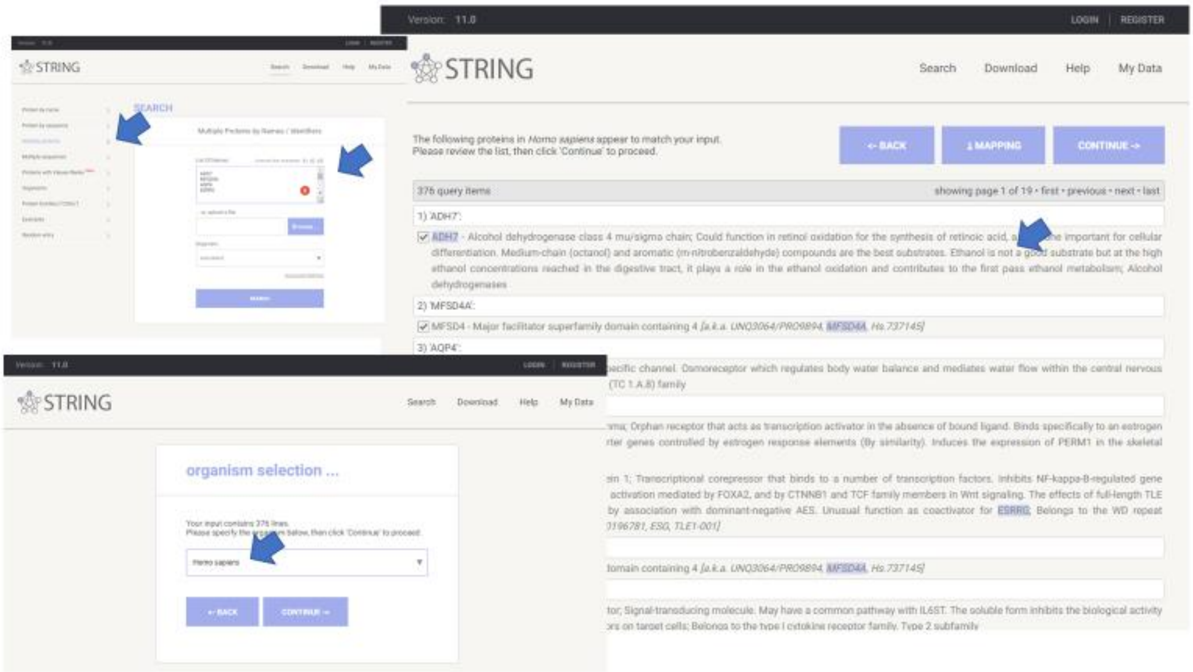


- Protein-protein interaction (PPI) network construction and module analysis.

- The Search Tool for the Retrieval of Interacting Genes (STRING), an online database (string-db.org) designed to evaluate PPI information, covers 9,643,763 proteins from more than 2,000 organisms, which was used to construct the PPI.

- To evaluate the interactive associations of DEGs identified from GSE26899, we mapped these DEGs to the STRING (version 10.5) database. Confidence score >0.4 was selected as significant. PPI networks were constructed by STRING and visualized by Cytoscape.

- Subsequently, the plug-in Molecular Complex Detection (MCODE) was employed to screen the modules of PPI networks in Cytoscape with the threshold set as follows: MCODE scores >10.



The following proteins in *Homo sapiens* appear to match your input.
Please review the list, then click 'Continue' to proceed.

376 query items — showing page 1 of 19 • first • previous • next • last

1) 'ADH7':

☑ ADH7 - Alcohol dehydrogenase class 4 mu/sigma chain; Could function in retinol oxidation for the synthesis of retinoic acid, a [illegible] be important for cellular differentiation. Medium-chain (octanol) and aromatic (m-nitrobenzaldehyde) compounds are the best substrates. Ethanol is not a good substrate but at the high ethanol concentrations reached in the digestive tract, it plays a role in the ethanol oxidation and contributes to the first pass ethanol metabolism; Alcohol dehydrogenases

2) 'MFSD4A':

☑ MFSD4 - Major facilitator superfamily domain containing 4 *[a.k.a. UNQ3064/PRO9894, MFSD4A, Hs.737145]*

3) 'AQP4':

pecific channel. Osmoreceptor which regulates body water balance and mediates water flow within the central nervous (TC 1.A.8) family

ma; Orphan receptor that acts as transcription activator in the absence of bound ligand. Binds specifically to an estrogen rter genes controlled by estrogen response elements (By similarity). Induces the expression of PERM1 in the skeletal

ein 1; Transcriptional corepressor that binds to a number of transcription factors. Inhibits NF-kappa-B-regulated gene activation mediated by FOXA2, and by CTNNB1 and TCF family members in Wnt signaling. The effects of full-length TLE by association with dominant-negative AES. Unusual function as coactivator for ESRRG; Belongs to the WD repeat *0196781, ESG, TLE1-001]*

domain containing 4 *[a.k.a. UNQ3064/PRO9894, MFSD4A, Hs.737145]*

tor; Signal-transducing molecule. May have a common pathway with IL6ST. The soluble form inhibits the biological activity ors on target cells; Belongs to the type I cytokine receptor family. Type 2 subfamily

Your input contains 376 lines.
Please specify the organism below, then click 'Continue' to proceed.

Cytoscape

Cytoscape
Network Data Integration, Analysis, and Visualization in a Box
Introduction
Download 3.8.2

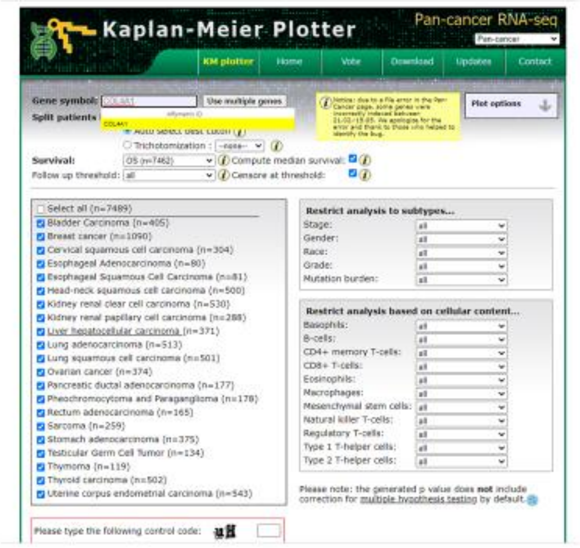
- Cytoscape 내 geneMANIA 있음

- Survival analysis of collagen type IV α1 chain (COL4A1).

- To evaluate the association between COL4A1 level and its clinical outcomes, Kaplan-Meier plotter (KM plotter; www. kmplot.com), an online survival analysis tool, was performed.

- KM plotter is capable of assessing the effect of 54,675 genes on overall survival via using 10,188 cancer samples including 4,142 breast, 1,648 ovarian, 2,437 lung, and 1,065 gastric cancer patients (17).

- Patients with gastric cancer were separated into high- and low-expression groups according to the level of COL4A1, and the overall survival was then analyzed. The hazard ratio (HR) with 95% confidence intervals and log rank P-value were calculated.



Gene symbol: COL4A1

Split patients by: Affymetrix ID: COL4A1

● Auto select best cutoff

○ Trichotomization : --none--

Notice: due to a file error in the Pan Cancer page, some genes were incorrectly indexed between 21-02-15-05. We apologize for the error and thank to those who helped to identify the bug.

Survival: OS (n=7462) Compute median survival: ☑

Follow up threshold: all Censore at threshold: ☑

- ☐ Select all (n=7489)
- ☑ Bladder Carcinoma (n=405)
- ☑ Breast cancer (n=1090)
- ☑ Cervical squamous cell carcinoma (n=304)
- ☑ Esophageal Adenocarcinoma (n=80)
- ☑ Esophageal Squamous Cell Carcinoma (n=81)
- ☑ Head-neck squamous cell carcinoma (n=500)
- ☑ Kidney renal clear cell carcinoma (n=530)
- ☑ Kidney renal papillary cell carcinoma (n=288)
- ☑ Liver hepatocellular carcinoma (n=371)
- ☑ Lung adenocarcinoma (n=513)
- ☑ Lung squamous cell carcinoma (n=501)
- ☑ Ovarian cancer (n=374)
- ☑ Pancreatic ductal adenocarcinoma (n=177)
- ☑ Pheochromocytoma and Paraganglioma (n=178)
- ☑ Rectum adenocarcinoma (n=165)
- ☑ Sarcoma (n=259)
- ☑ Stomach adenocarcinoma (n=375)
- ☑ Testicular Germ Cell Tumor (n=134)
- ☑ Thymoma (n=119)
- ☑ Thyroid carcinoma (n=502)
- ☑ Uterine corpus endometrial carcinoma (n=543)

Restrict analysis to subtypes...

Stage: all
Gender: all
Race: all
Grade: all
Mutation burden: all

Restrict analysis based on cellular content...

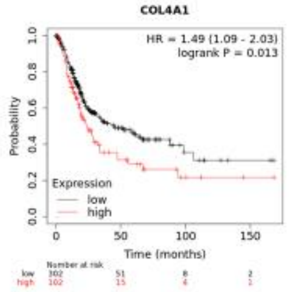
Basophils: all
B-cells: all
CD4+ memory T-cells: all
CD8+ T-cells: all
Eosinophils: all
Macrophages: all
Mesenchymal stem cells: all
Natural killer T-cells: all
Regulatory T-cells: all
Type 1 T-helper cells: all
Type 2 T-helper cells: all

Please note: the generated p value does **not** include correction for multiple hypothesis testing by default.

Please type the following control code: uH

Results

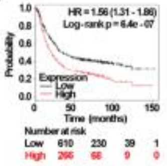
P value: 0.0125
FDR: over 50%



Download plot as a PDF

Download p values vs. cutoff table

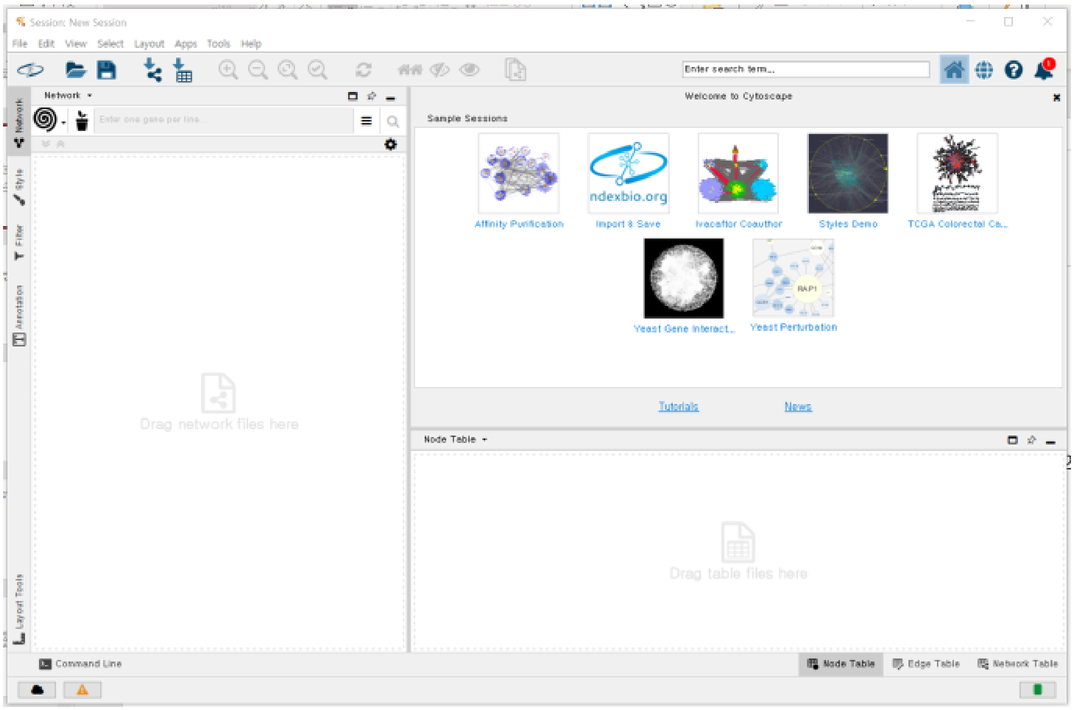
F

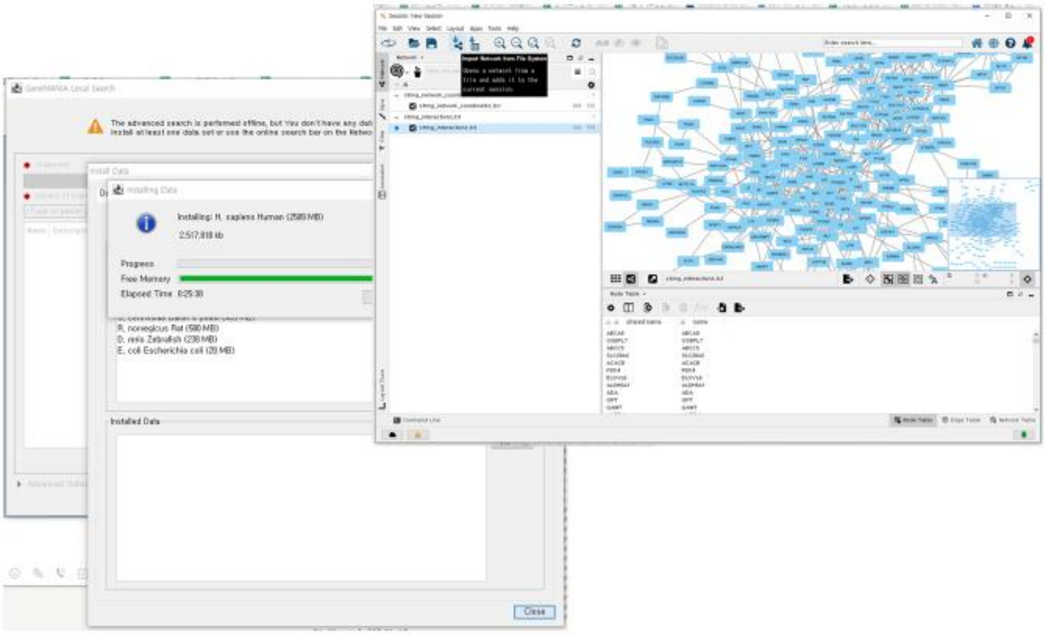


- Analysis of COL4A1 by geneMANIA and coremine.

- GeneMANIA, an online tool (www.genemania.org/), can be used to generate hypotheses of gene function, analyze gene lists, and prioritize genes for functional assays (18).

- After selecting Homo sapiens from the nine optional organisms, COL4A1 was entered into the search bar and the results were then collected. Annotation of biological processes involving COL4A1 was performed by consulting the Coremine Medical online database (www.coremine.com/medical/).





The advanced search is performed offline, but You don't have any dat
install at least one data set or use the online search bar on the Netwo

Installing: H. sapiens Human (2589 MB)

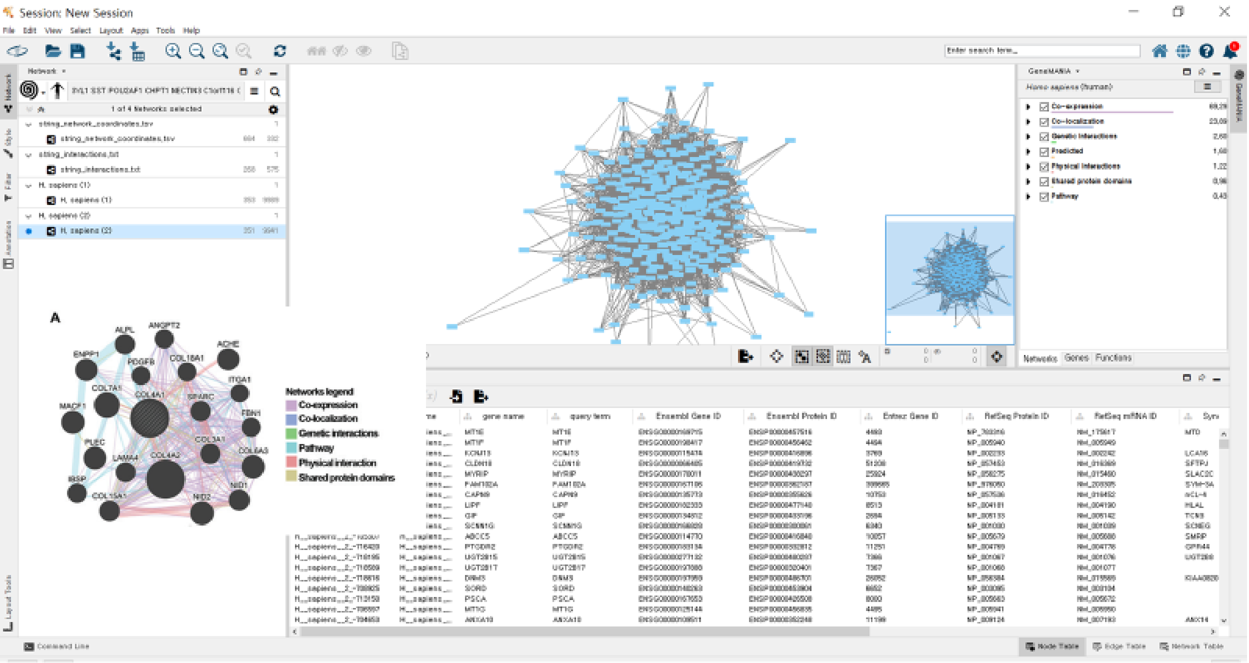
2,517,818 kb

Progress

Free Memory

Elapsed Time 0:25:38

R. norvegicus Rat (580 MB)
D. rerio Zebrafish (238 MB)
E. coli Escherichia coli (20 MB)



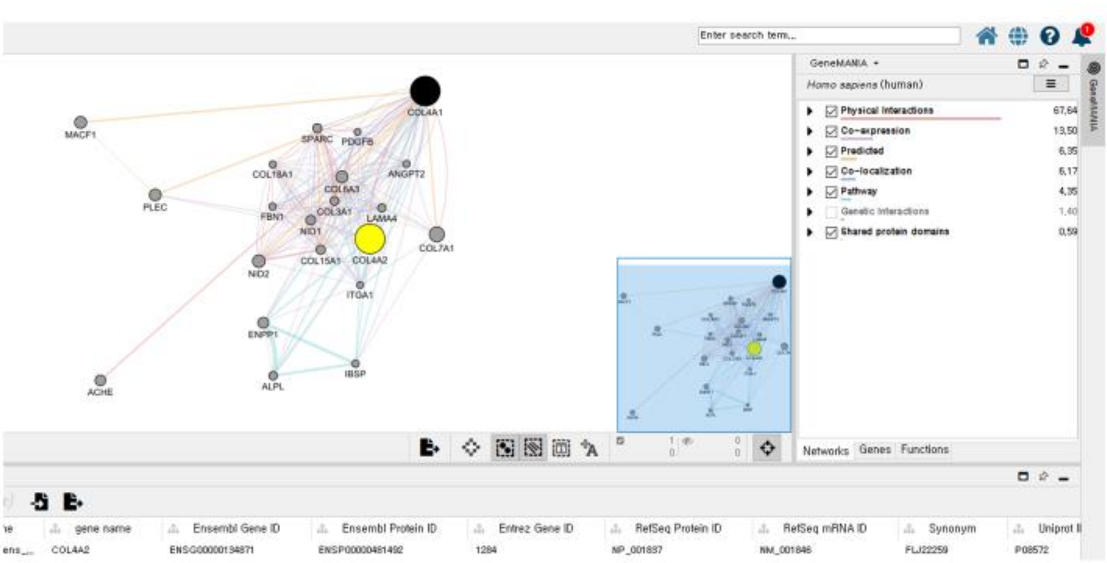
Session: New Session

⌄ string_network_coordinates.tsv		1
string_network_coordinates.tsv	664	332
⌄ string_interactions.txt		1
string_interactions.txt	258	575
⌄ H. sapiens (1)		1
H. sapiens (1)	353	9999
⌄ H. sapiens (2)		1
● H. sapiens (2)	351	3941

▶ ☑ Co-expression	69,28
▶ ☑ Co-localization	23,08
▶ ☑ Genetic Interactions	2,68
▶ ☑ Predicted	1,68
▶ ☑ Physical Interactions	1,22
▶ ☑ Shared protein domains	0,98
▶ ☑ Pathway	0,43

Networks legend
🟪 Co-expression
🟦 Co-localization
🟩 Genetic interactions
🟦 Pathway
🟥 Physical interaction
🟨 Shared protein domains

	ne	gene name	query term	Ensembl Gene ID	Ensembl Protein ID	Entrez Gene ID	RefSeq Protein ID	RefSeq mRNA ID	Syn
	iens_...	MT1E	MT1E	ENSG00000169715	ENSP00000457516	4493	NP_783316	NM_175617	MTD
	iens_...	MT1F	MT1F	ENSG00000198417	ENSP00000456462	4494	NP_005940	NM_005949	
	iens_...	KCNJ15	KCNJ15	ENSG00000115474	ENSP00000416896	3769	NP_002233	NM_002242	LCA16
	iens_...	CLDN18	CLDN18	ENSG00000066405	ENSP00000419732	51208	NP_057453	NM_016369	SFTPJ
	iens_...	MYRIP	MYRIP	ENSG00000170011	ENSP00000400297	25924	NP_056275	NM_015460	SLAC2C
	iens_...	FAM102A	FAM102A	ENSG00000167106	ENSP00000362187	399665	NP_976050	NM_203305	SYM-3A
	iens_...	CAPN9	CAPN9	ENSG00000135773	ENSP00000355826	10753	NP_057536	NM_016452	nCL-4
	iens_...	LIPF	LIPF	ENSG00000182333	ENSP00000477140	8513	NP_004181	NM_004190	HLAL
	iens_...	GIP	GIP	ENSG00000134812	ENSP00000433196	2694	NP_008133	NM_006142	TCN3
	iens_...	SCNN1G	SCNN1G	ENSG00000166828	ENSP00000300061	6340	NP_001030	NM_001039	SCNEG
H__sapiens__2_-T16047	H__sapiens_...	ABCC5	ABCC5	ENSG00000114770	ENSP00000416840	10057	NP_005679	NM_005688	SMRP
H__sapiens__2_-T16429	H__sapiens_...	PTGDR2	PTGDR2	ENSG00000183134	ENSP00000332812	11251	NP_004769	NM_004778	GPR44
H__sapiens__2_-T18195	H__sapiens_...	UGT2B15	UGT2B15	ENSG00000277132	ENSP00000480297	7366	NP_001067	NM_001076	UGT2B8
H__sapiens__2_-T18589	H__sapiens_...	UGT2B17	UGT2B17	ENSG00000197888	ENSP00000320401	7367	NP_001068	NM_001077	
H__sapiens__2_-T18818	H__sapiens_...	DNM3	DNM3	ENSG00000197959	ENSP00000486701	26052	NP_056384	NM_015569	KIAA0820
H__sapiens__2_-T08925	H__sapiens_...	SORD	SORD	ENSG00000140263	ENSP00000453804	6652	NP_003095	NM_003104	
H__sapiens__2_-T13159	H__sapiens_...	PSCA	PSCA	ENSG00000167653	ENSP00000426508	8000	NP_005663	NM_005672	
H__sapiens__2_-T08597	H__sapiens_...	MT1G	MT1G	ENSG00000125144	ENSP00000456835	4495	NP_005941	NM_005950	
H__sapiens__2_-T04859	H__sapiens_...	ANXA10	ANXA10	ENSG00000109511	ENSP00000352248	11199	NP_008124	NM_007193	ANX14



▶ ☑ Physical Interactions	67,64
▶ ☑ Co-expression	13,50
▶ ☑ Predicted	6,35
▶ ☑ Co-localization	6,17
▶ ☑ Pathway	4,35
▶ □ Genetic Interactions	1,40
▶ ☑ Shared protein domains	0,59

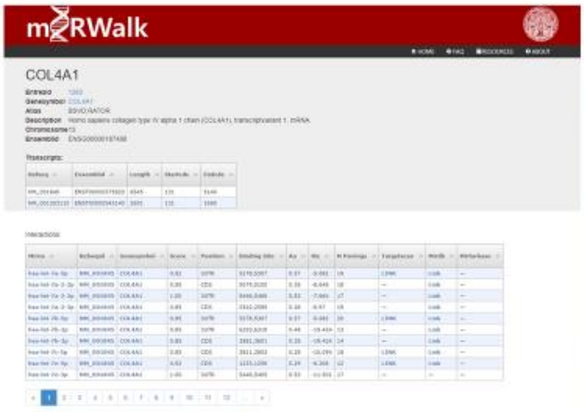
ne	gene name	Ensembl Gene ID	Ensembl Protein ID	Entrez Gene ID	RefSeq Protein ID	RefSeq mRNA ID	Synonym	Uniprot I
ens_...	COL4A2	ENSG00000134871	ENSP00000481492	1284	NP_001837	NM_001846	FLJ22259	P08572

- Prediction of miRNAs.

- To predict the miRNAs targeting the mRNA of COL4A1, miRWalk (version 2.0, zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/), an online platform supplying information about predicted and experimentally validated miRNA-target interactions, was then employed (19).

- Herein, nine prediction programs (miRWalk, miRanda, miRDB, miRNAMap, Pictar2, PITA, RNA22, RNAhybrid and Targetscan) were selected. These predicted miRNAs were then overlapped by at least seven programs, and selected for further analysis.

- Pathway enrichment analysis of these miRNAs was performed by using the DIANA-mirPath web server (snf-515788.vm.okeanos.grnet.gr/index.php?r=mirpath) (20).

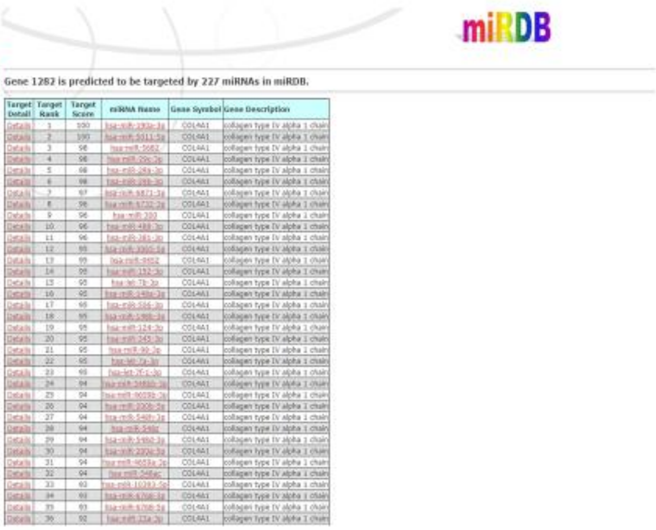


COL4A1

Entrezid 1282
Genesymbol COL4A1
Alias [illegible]
Description Homo sapiens collagen type IV alpha 1 chain (COL4A1), transcript variant 1, mRNA
Chromosome 13
Ensembleid ENSG00000187498

Transcripts:

Interactions



Gene 1282 is predicted to be targeted by 227 miRNAs in miRDB.

Target Detail	Target Rank	Target Score	miRNA Name	Gene Symbol	Gene Description
Details	1	100	hsa-miR-193a-3p	COL4A1	collagen type IV alpha 1 chain
Details	2	100	hsa-miR-5011-5p	COL4A1	collagen type IV alpha 1 chain
Details	3	98	hsa-miR-5682	COL4A1	collagen type IV alpha 1 chain
Details	4	98	hsa-miR-29c-3p	COL4A1	collagen type IV alpha 1 chain
Details	5	98	hsa-miR-29a-3p	COL4A1	collagen type IV alpha 1 chain
Details	6	98	hsa-miR-29b-3p	COL4A1	collagen type IV alpha 1 chain
Details	7	97	hsa-miR-6871-3p	COL4A1	collagen type IV alpha 1 chain
Details	8	96	hsa-miR-6732-3p	COL4A1	collagen type IV alpha 1 chain
Details	9	96	hsa-miR-330	COL4A1	collagen type IV alpha 1 chain
Details	10	96	hsa-miR-488-3p	COL4A1	collagen type IV alpha 1 chain
Details	11	96	hsa-miR-381-3p	COL4A1	collagen type IV alpha 1 chain
Details	12	95	hsa-miR-3065-5p	COL4A1	collagen type IV alpha 1 chain
Details	13	95	hsa-miR-8482	COL4A1	collagen type IV alpha 1 chain
Details	14	95	hsa-miR-152-3p	COL4A1	collagen type IV alpha 1 chain
Details	15	95	hsa-let-7b-3p	COL4A1	collagen type IV alpha 1 chain
Details	16	95	hsa-miR-148a-3p	COL4A1	collagen type IV alpha 1 chain
Details	17	95	hsa-miR-556-3p	COL4A1	collagen type IV alpha 1 chain
Details	18	95	hsa-miR-148b-3p	COL4A1	collagen type IV alpha 1 chain
Details	19	95	hsa-miR-124-3p	COL4A1	collagen type IV alpha 1 chain
Details	20	95	hsa-miR-345-3p	COL4A1	collagen type IV alpha 1 chain
Details	21	95	hsa-miR-98-3p	COL4A1	collagen type IV alpha 1 chain
Details	22	95	hsa-let-7a-3p	COL4A1	collagen type IV alpha 1 chain
Details	23	95	hsa-let-7f-1-3p	COL4A1	collagen type IV alpha 1 chain
Details	24	94	hsa-miR-3680-3p	COL4A1	collagen type IV alpha 1 chain
Details	25	94	hsa-miR-6003b-3p	COL4A1	collagen type IV alpha 1 chain
Details	26	94	hsa-miR-3306-5p	COL4A1	collagen type IV alpha 1 chain
Details	27	94	hsa-miR-548h-3p	COL4A1	collagen type IV alpha 1 chain
Details	28	94	hsa-miR-548z	COL4A1	collagen type IV alpha 1 chain
Details	29	94	hsa-miR-5480-3p	COL4A1	collagen type IV alpha 1 chain
Details	30	94	hsa-miR-3306-5p	COL4A1	collagen type IV alpha 1 chain
Details	31	94	hsa-miR-4653a-3p	COL4A1	collagen type IV alpha 1 chain
Details	32	94	hsa-miR-548ac	COL4A1	collagen type IV alpha 1 chain
Details	33	93	hsa-miR-10393-5p	COL4A1	collagen type IV alpha 1 chain
Details	34	93	hsa-miR-6768-3p	COL4A1	collagen type IV alpha 1 chain
Details	35	93	hsa-miR-6768-5p	COL4A1	collagen type IV alpha 1 chain
Details	36	92	hsa-miR-33a-3p	COL4A1	collagen type IV alpha 1 chain

감사합니다.

Appendix

Hyemin Gu

2020 12 17

Appendix: Classification of entire data in GDC portal by filetypes

1 차 분류: Data Category 2 차 분류: Data Type 3 차 분류: Experiment Stratagy='RNA-Seq'

*표시는 TCGAbiolinks 에서 지원하는 Data Category ### Harmonized database cf) TCGAbiolinks 에서 지원하는 Experiment Stratagy: WXS, RNA-Seq, miRNA-Seq, Genotyping Array

- simple nucleotide variation *
  - Annotated Somatic Mutation
    - Raw Simple Somatic Mutation
    - Masked Annotated Somatic Mutation
    - Aggregated Somatic Mutation
    - Masked Somatic Mutation
- copy number variation *
  - Gene Level Copy Number Scores
  - Copy Number Segment
  - Masked Copy Number Segment
  - Gene Level Copy Number
  - Allele-specific Copy Number Segment
- transcriptome profiling *
  - Gene Expression Quantification
    - RNA-Seq
  - Isoform Expression Quantification
  - miRNA Expression Quantification
  - Splice Junction Quantification
    - RNA-Seq
- sequencing reads *
  - Aligned Reads
    - RNA-Seq
- biospecimen *
  - Slide Image
  - Biospecimen Supplement
- clinical *
  - Clinical Supplement
- dna methylation *

  - Methylation Beta Value
- somatic structural variation
  - Transcript Fusion
  - Structural Rearrangement
- structural variation
  - Transcript Fusion
    - RNA-Seq
- combined nucleotide variation
  - Raw CGI Variant

Legacy archive

cf) TCGAbiolinks 에서 지원하는 Experiment Stratagy: WXS, RNA-Seq, miRNA-Seq, Genotyping Array, DNA-Seq, Methylation array, Protein expression array, WXS,CGH array, VALIDATION, Gene expression array,WGS, MSI-Mono-Dinucleotide Assay, miRNA expression array, Mixed strategies, AMPLICON, Exon array, Total RNA-Seq, Capillary sequencing, Bisulfite-Seq

- Copy number variation *
  - Copy number segmentation
  - Copy number estimate
  - Normalized copy numbers
  - Copy number variation
  - Copy number QC metrics
  - LOH
- Gene expression *
  - Gene expression quantification
    - RNA-Seq
    - Gene expression array
  - miRNA gene quantification
  - miRNA isoform quantification
  - Isoform expression quantification
    - RNA-Seq
  - Exon quantification
    - RNA-Seq
  - Exon junction quantification
    - RNA-Seq
  - rtPCR quantification
  - Gene expression summary
    - Gene expression array
- Raw microarray data *
  - Raw intensities
    - Protein expression array

        - Gene expression array
    - Normalized intensities
        - Gene expression array
    - CGH array QC
    - Intensities
        - Protein expression array
        - Gene expression array
    - Intensities LogRatio
    - Methylation array QC metrics
- Raw sequencing data *
    - Aligned reads
        - RNA-Seq
    - Coverage WIG
    - Unaligned reads
        - RNA-Seq
    - Sequencing tag
        - RNA-Seq
    - Sequencing tag counts
        - RNA-Seq
- Simple nucleotide variation *
    - Simple nucleotide variation
        - RNA-Seq
    - Genotypes
    - Simple somatic mutation
- Clinical *
    - Tissue slide image
    - Diagnostic image
    - Clinical Supplement
    - Pathology report
    - Clinical data
    - Biospecimen data
- DNA methylation *
    - Methylation beta value
    - Bisulfite sequence alignment
    - Methylation percentage
- Biospecimen *
    - Biospecimen Supplement
- Other
    - Microsatellite instability
    - Auxiliary test
    - ABI sequence trace
- Protein expression *

  - Protein expression quantification
    - Protein expression array
- Archive
  - TCGA DCC Archive
- Structural rearrangement
  - Structural variation
    - RNA-Seq
- Processed microarray data
  - Processed intensities
    - Gene expression array