# data_2_GSE16873

Sedreh

5/9/2019

```
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
```

```
# read the dataset into R
library(GEOquery)
```

```
## 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 objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union
```

```
## 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)
```

```
library(limma)
```

```
##
## Attaching package: 'limma'
```

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

```
# library for Human annotation
library(org.Hs.eg.db)
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: stats4
```

```
## Loading required package: IRanges
```

```
## Loading required package: S4Vectors
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, rename
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
```

```
##
## Attaching package: 'IRanges'
```

```
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
```

```
##
## Attaching package: 'AnnotationDbi'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
```

```
# for collapseBy and other functions
source("/home/sedreh/Documents/rnaseq/functions.r")
### load the dataset here
res <- getGEO("GSE16873", AnnotGPL = TRUE)[[1]]
```

```
## Found 1 file(s)
```

```
## GSE16873_series_matrix.txt.gz
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ID_REF = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
## File stored at:
```

```
## /tmp/RtmpeAGKsQ/GPL96.annot.gz
```

```
## Warning: 1176 parsing failures.
##   row           col           expected               actual         file
## 10161 UniGene title 1/0/T/F/TRUE/FALSE Clone HQ0117 PRO0117 literal data
## 10161 UniGene ID    1/0/T/F/TRUE/FALSE Hs.670442            literal data
## 10179 UniGene title 1/0/T/F/TRUE/FALSE Transcribed locus    literal data
## 10179 UniGene ID    1/0/T/F/TRUE/FALSE Hs.621370            literal data
## 10347 UniGene title 1/0/T/F/TRUE/FALSE Transcribed locus    literal data
## ..... ............. .................. ..................... ............
## See problems(...) for more details.
```

```
# GEOquery is working, this is a list of files, I can see all the information
# to access individual list I need to use this format res$data@data
# for example, res@experimentData@title will give us details about the experiment
res@experimentData@title
```

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

```
# this is mouse dataset
res@experimentData@abstract
```

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

```
#  simple ductal hyperplasia (SH) and atypical ductal hyperplasia (ADH) are considerable issues in this pape
r
# This dataset doesn't contain the abstract or experimental information. Let's continue to work on it.
```

```
# every GEO data has these internal identifiers: pData is phenotypeData, fData is featureData
str(experimentData(res))
```

```
## Formal class 'MIAME' [package "Biobase"] with 13 slots
##   ..@ name             : chr ""
##   ..@ lab              : chr ""
##   ..@ contact          : chr ""
##   ..@ title            : chr ""
##   ..@ abstract         : chr ""
##   ..@ url              : chr ""
##   ..@ pubMedIds        : chr ""
##   ..@ samples          : list()
##   ..@ hybridizations   : list()
##   ..@ normControls     : list()
##   ..@ preprocessing    : list()
##   ..@ other            : list()
##   ..@ .__classVersion__:Formal class 'Versions' [package "Biobase"] with 1 slot
##   .. .. ..@ .Data:List of 2
##   .. .. .. ..$ : int [1:3] 1 0 0
##   .. .. .. ..$ : int [1:3] 1 1 0
```

```
str(pData(res))
```

```
## 'data.frame':    40 obs. of  40 variables:
##  $ title                 : Factor w/ 40 levels "226 ADH","226 DCIS",..: 39 40 37 38 15 13 14 21 19 20
...
##  $ geo_accession         : chr  "GSM422873" "GSM422874" "GSM422875" "GSM422876" ...
##  $ status                : Factor w/ 1 level "Public on Sep 08 2009": 1 1 1 1 1 1 1 1 1 1 ...
##  $ submission_date       : Factor w/ 1 level "Jun 29 2009": 1 1 1 1 1 1 1 1 1 1 ...
##  $ last_update_date      : Factor w/ 1 level "Sep 08 2009": 1 1 1 1 1 1 1 1 1 1 ...
##  $ type                  : Factor w/ 1 level "RNA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ channel_count         : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ source_name_ch1       : Factor w/ 4 levels "atypical ductal hyperplasia epithelium",..: 3 4 1 2 3 1 2
3 1 2 ...
##  $ organism_ch1          : Factor w/ 1 level "Homo sapiens": 1 1 1 1 1 1 1 1 1 1 ...
##  $ characteristics_ch1   : Factor w/ 1 level "tissue: laser capture microdissected human breast": 1 1 1
1 1 1 1 1 1 1 ...
##  $ characteristics_ch1.1 : Factor w/ 1 level "cell type: epithelial": 1 1 1 1 1 1 1 1 1 1 ...
##  $ characteristics_ch1.2 : Factor w/ 4 levels "disease state: atypical ductal hyperplasia",..: 3 4 1 2 3
1 2 3 1 2 ...
##  $ characteristics_ch1.3 : Factor w/ 1 level "age range: 48-92 years old": 1 1 1 1 1 1 1 1 1 1 ...
##  $ treatment_protocol_ch1 : Factor w/ 1 level "Samples were microdissected from lightly H&E stained seria
l frozen tissue sections.": 1 1 1 1 1 1 1 1 1 1 ...
##  $ molecule_ch1          : Factor w/ 1 level "total RNA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ extract_protocol_ch1  : Factor w/ 1 level "Following the manufacturer's protocol, total RNA was extra
cted and purified using the Picopure RNA Isolation Ki"| __truncated__: 1 1 1 1 1 1 1 1 1 1 ...
##  $ label_ch1             : Factor w/ 1 level "biotin": 1 1 1 1 1 1 1 1 1 1 ...
##  $ label_protocol_ch1    : Factor w/ 1 level "To convert the RNA to cDNA, the purified total RNA was lin
early amplified for two rounds using the MessageAMP a"| __truncated__: 1 1 1 1 1 1 1 1 1 1 ...
##  $ label_protocol_ch1.1  : Factor w/ 1 level "For second round amplification, cDNA for each sample was i
n vitro transcribed (IVT) to biotin-labeled cRNA with"| __truncated__: 1 1 1 1 1 1 1 1 1 1 ...
##  $ taxid_ch1             : Factor w/ 1 level "9606": 1 1 1 1 1 1 1 1 1 1 ...
##  $ hyb_protocol          : Factor w/ 1 level "10 µg of fragmented cRNA and hybridization controls were h
ybridized to each U133A GeneChip (Affymetrix) for 16 "| __truncated__: 1 1 1 1 1 1 1 1 1 1 ...
##  $ scan_protocol         : Factor w/ 1 level "The stained arrays were scanned using a G2500 Scanner (Agi
lent)": 1 1 1 1 1 1 1 1 1 1 ...
##  $ description           : Factor w/ 40 levels "Gene expression data from case 226 ADH (rehyb)",..: 39 4
0 37 38 15 13 14 21 19 20 ...
##  $ data_processing       : Factor w/ 1 level "Images from the scanned chips were quantified and scaled u
sing Affymetrix Microarray Suite 5.0 (Affymetrix), an"| __truncated__: 1 1 1 1 1 1 1 1 1 1 ...
##  $ platform_id           : Factor w/ 1 level "GPL96": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_name          : Factor w/ 1 level "Lyndsey,A.,Emery": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_laboratory    : Factor w/ 1 level "Carol L. Rosenberg": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_department    : Factor w/ 1 level "Medicine - Hematology/Oncology": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_institute     : Factor w/ 1 level "Boston University Medical Center": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_address       : Factor w/ 1 level "650 Albany Street - EBRC 4th Floor": 1 1 1 1 1 1 1 1 1 1
...
##  $ contact_city          : Factor w/ 1 level "Boston": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_state         : Factor w/ 1 level "MA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_zip/postal_code: Factor w/ 1 level "02118": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contact_country       : Factor w/ 1 level "USA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ supplementary_file    : Factor w/ 40 levels "ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM422nnn/GSM4228
73/suppl/GSM422873.CEL.gz",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ data_row_count        : Factor w/ 1 level "22283": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age range:ch1         : chr  "48-92 years old" "48-92 years old" "48-92 years old" "48-92 years old"
...
##  $ cell type:ch1         : chr  "epithelial" "epithelial" "epithelial" "epithelial" ...
##  $ disease state:ch1     : chr  "histologically normal" "simple ductal hyperplasia" "atypical ductal hyp
erplasia" "ductal carcinoma in situ" ...
##  $ tissue:ch1            : chr  "laser capture microdissected human breast" "laser capture microdissecte
d human breast" "laser capture microdissected human breast" "laser capture microdissected human breast" ...
```

```
head(fData(res))
```

```
##                 ID
## 1007_s_at 1007_s_at
## 1053_at     1053_at
## 117_at       117_at
## 121_at       121_at
## 1255_g_at 1255_g_at
## 1294_at     1294_at
##                                                          Gene title
## 1007_s_at microRNA 4640///discoidin domain receptor tyrosine kinase 1
## 1053_at                                 replication factor C subunit 2
## 117_at                  heat shock protein family A (Hsp70) member 6
## 121_at                                                 paired box 8
## 1255_g_at                               guanylate cyclase activator 1A
## 1294_at   microRNA 5193///ubiquitin like modifier activating enzyme 7
##              Gene symbol          Gene ID UniGene title UniGene symbol
## 1007_s_at MIR4640///DDR1  100616237///780           NA             NA
## 1053_at             RFC2             5982           NA             NA
## 117_at             HSPA6             3310           NA             NA
## 121_at              PAX8             7849           NA             NA
## 1255_g_at         GUCA1A             2978           NA             NA
## 1294_at   MIR5193///UBA7 100847079///7318           NA             NA
##           UniGene ID
## 1007_s_at         NA
## 1053_at           NA
## 117_at            NA
## 121_at            NA
## 1255_g_at         NA
## 1294_at           NA
##                                                                                Nucleotide Title
## 1007_s_at                                 Human receptor tyrosine kinase DDR gene, complete cds
## 1053_at                     Human replication factor C, 40-kDa subunit (A1) mRNA, complete cds
## 117_at                                                  Human heat-shock protein HSP70B' gene
## 121_at                                                                    H.sapiens Pax8 mRNA
## 1255_g_at Homo sapiens guanylate cyclase activating protein (GCAP) gene exons 1-4, complete cds
## 1294_at          Homo sapiens ubiquitin-activating enzyme E1 related protein mRNA, complete cds
##                GI GenBank Accession Platform_CLONEID Platform_ORF
## 1007_s_at 1753221            U48705               NA           NA
## 1053_at   1590810            M87338               NA           NA
## 117_at      35221            X51757               NA           NA
## 121_at      38425            X69699               NA           NA
## 1255_g_at  623404            L36861               NA           NA
## 1294_at    520832            L13852               NA           NA
##           Platform_SPOTID Chromosome location
## 1007_s_at              NA              6p21.3
## 1053_at                NA             7q11.23
## 117_at                 NA                1q23
## 121_at                 NA                2q13
## 1255_g_at              NA              6p21.1
## 1294_at                NA                3p21
##
Chromosome annotation
## 1007_s_at                         Chromosome 6, NC_000006.12 (30890883..30890972)///Chromosome 6, NC_0000
06.12 (30880909..30900156)
## 1053_at                                                                  Chromosome 7, NC_000007.14 (74231
502..74254458, complement)
## 117_at                                                                       Chromosome 1, NC_00000
1.11 (161524540..161526897)
## 121_at                                                                 Chromosome 2, NC_000002.12 (1132159
97..113278950, complement)
## 1255_g_at                                                                    Chromosome 6, NC_0000
06.12 (42155377..42180083)
## 1294_at   Chromosome 3, NC_000003.12 (49806137..49806245, complement)///Chromosome 3, NC_000003.12 (49805
205..49813958, complement)
##
GO:Function
## 1007_s_at
ATP binding///collagen binding///collagen binding///metal ion binding///protein binding///protein tyrosine k
inase collagen receptor activity///transmembrane receptor protein tyrosine kinase activity
## 1053_at
ATP binding///contributes_to DNA clamp loader activity///enzyme binding///protein binding///contributes_to s
ingle-stranded DNA-dependent ATPase activity
## 117_at
```

```
ATP binding///ATPase activity, coupled///enzyme binding///heat shock protein binding///protein binding///unf
olded protein binding
## 121_at    DNA binding///DNA binding///RNA polymerase II core promoter proximal region sequence-specific D
NA binding///RNA polymerase II core promoter sequence-specific DNA binding///protein binding///thyroid-stimu
lating hormone receptor activity///transcription factor activity, sequence-specific DNA binding///transcript
ion regulatory region DNA binding///transcriptional activator activity, RNA polymerase II core promoter prox
imal region sequence-specific binding
## 1255_g_at
calcium ion binding///calcium sensitive guanylate cyclase activator activity///guanylate cyclase regulator a
ctivity
## 1294_at
ATP binding///ISG15 activating enzyme activity///protein binding///ubiquitin activating enzyme activity///ub
iquitin-protein transferase activity///ubiquitin-protein transferase activity
##
GO:Process
## 1007_s_at
branching involved in mammary gland duct morphogenesis///cell adhesion///collagen-activated tyrosine kinase
receptor signaling pathway///collagen-activated tyrosine kinase receptor signaling pathway///ear developmen
t///embryo implantation///extracellular matrix organization///lactation///mammary gland alveolus developmen
t///negative regulation of cell proliferation///organ regeneration///peptidyl-tyrosine autophosphorylatio
n///protein autophosphorylation///regulation of cell growth///regulation of cell-matrix adhesion///regulatio
n of extracellular matrix disassembly///skin development///smooth muscle cell migration///smooth muscle cell
-matrix adhesion///wound healing, spreading of cells
## 1053_at
DNA damage response, detection of DNA damage///DNA replication///error-free translesion synthesis///error-pr
one translesion synthesis///nucleotide-excision repair, DNA gap filling///nucleotide-excision repair, DNA in
cision///nucleotide-excision repair, DNA incision, 5'-to lesion///positive regulation of DNA-directed DNA po
lymerase activity///regulation of signal transduction by p53 class mediator///telomere maintenance via recom
bination///transcription-coupled nucleotide-excision repair///translesion synthesis
## 117_at
NOT cellular heat acclimation///cellular response to heat///cellular response to heat///protein refolding///
response to unfolded protein
## 121_at    anatomical structure morphogenesis///branching involved in ureteric bud morphogenesis///cellula
r response to gonadotropin stimulus///central nervous system development///inner ear morphogenesis///kidney
development///mesenchymal to epithelial transition involved in metanephros morphogenesis///mesonephros devel
opment///metanephric S-shaped body morphogenesis///metanephric comma-shaped body morphogenesis///metanephric
distal convoluted tubule development///metanephric epithelium development///metanephric nephron tubule forma
tion///negative regulation of apoptotic process involved in metanephric collecting duct development///negati
ve regulation of apoptotic process involved in metanephric nephron tubule development///negative regulation
of mesenchymal cell apoptotic process involved in metanephric nephron morphogenesis///negative regulation of
mesenchymal cell apoptotic process involved in metanephros development///otic vesicle development///positive
regulation of branching involved in ureteric bud morphogenesis///positive regulation of mesenchymal to epith
elial transition involved in metanephros morphogenesis///positive regulation of metanephric DCT cell differe
ntiation///positive regulation of thyroid hormone generation///positive regulation of transcription from RNA
polymerase II promoter///positive regulation of transcription, DNA-templated///positive regulation of transc
ription, DNA-templated///pronephric field specification///pronephros development///regulation of apoptotic p
rocess///regulation of metanephric nephron tubule epithelial cell differentiation///regulation of thyroid-st
imulating hormone secretion///sulfur compound metabolic process///thyroid gland development///thyroid gland
development///thyroid-stimulating hormone signaling pathway///transcription from RNA polymerase II promote
r///transcription, DNA-templated///urogenital system development
## 1255_g_at
cellular response to calcium ion///phototransduction///positive regulation of guanylate cyclase activity///r
egulation of rhodopsin mediated signaling pathway///signal transduction///visual perception
## 1294_at
ISG15-protein conjugation///cellular protein modification process///modification-dependent protein catabolic
process///negative regulation of type I interferon production///protein ubiquitination///translesion synthes
is
##
GO:Component
## 1007_s_at basolateral plasma membrane///extracellular exosome///extracellular space///integral component
of plasma membrane///plasma membrane///receptor complex
## 1053_at                                                                                    Ctf18 RFC-like c
omplex///DNA replication factor C complex///nucleoplasm
## 117_at                                          colocalizes_with COP9 signalosome///blood microparticle///
centriole///cytoplasm///cytosol///extracellular exosome
## 121_at
nucleoplasm///nucleoplasm///nucleus
## 1255_g_at                                                                  photoreceptor disc m
embrane///photoreceptor inner segment///plasma membrane
## 1294_at
cytosol///cytosol///nucleoplasm///nucleus
##
GO:Function ID
```

```
## 1007_s_at                          GO:0005524///GO:0005518///GO:0005518///GO:0046872///GO:0005515///GO:0
038062///GO:0004714
## 1053_at                        GO:0005524///contributes_to GO:0003689///GO:0019899///GO:0005515///contri
butes_to GO:0043142
## 117_at                                          GO:0005524///GO:0042623///GO:0019899///GO:0031072///GO:0
005515///GO:0051082
## 121_at    GO:0003677///GO:0003677///GO:0000978///GO:0000979///GO:0005515///GO:0004996///GO:0003700///GO:0
044212///GO:0001077
## 1255_g_at                                                                                GO:0005509///GO:0
008048///GO:0030249
## 1294_at                                         GO:0005524///GO:0019782///GO:0005515///GO:0004839///GO:0
004842///GO:0004842
##
GO:Process ID
## 1007_s_at
GO:0060444///GO:0007155///GO:0038063///GO:0038063///GO:0043583///GO:0007566///GO:0030198///GO:0007595///GO:0
060749///GO:0008285///GO:0031100///GO:0038083///GO:0046777///GO:0001558///GO:0001952///GO:0010715///GO:00435
88///GO:0014909///GO:0061302///GO:0044319
## 1053_at
GO:0042769///GO:0006260///GO:0070987///GO:0042276///GO:0006297///GO:0033683///GO:0006296///GO:1900264///GO:1
901796///GO:0000722///GO:0006283///GO:0019985
## 117_at
NOT GO:0070370///GO:0034605///GO:0034605///GO:0042026///GO:0006986
## 121_at    GO:0009653///GO:0001658///GO:0071371///GO:0007417///GO:0042472///GO:0001822///GO:0003337///GO:0
001823///GO:0072284///GO:0072278///GO:0072221///GO:0072207///GO:0072289///GO:1900215///GO:1900218///GO:00723
05///GO:1900212///GO:0071599///GO:0090190///GO:0072108///GO:2000594///GO:2000611///GO:0045944///GO:004589
3///GO:0045893///GO:0039003///GO:0048793///GO:0042981///GO:0072307///GO:2000612///GO:0006790///GO:0030878///
GO:0030878///GO:0038194///GO:0006366///GO:0006351///GO:0001655
## 1255_g_at
GO:0071277///GO:0007602///GO:0031284///GO:0022400///GO:0007165///GO:0007601
## 1294_at
GO:0032020///GO:0006464///GO:0019941///GO:0032480///GO:0016567///GO:0019985
##                                                                                          GO:Component ID
## 1007_s_at                  GO:0016323///GO:0070062///GO:0005615///GO:0005887///GO:0005886///GO:0043235
## 1053_at                                                                 GO:0031390///GO:0005663///GO:0005654
## 117_at    colocalizes_with GO:0008180///GO:0072562///GO:0005814///GO:0005737///GO:0005829///GO:0070062
## 121_at                                                                  GO:0005654///GO:0005654///GO:0005634
## 1255_g_at                                                               GO:0097381///GO:0001917///GO:0005886
## 1294_at                                                GO:0005829///GO:0005829///GO:0005654///GO:0005634
```

```
# we can see that disease state is our condition in this data
# here with gsub, we are just cleaning the data. each entry begins with the symbols \\+, _
# we have to CLEAN these symbols

condition<- res$`disease state:ch1`
```

```
res$condition <- gsub("\\+", "_", condition)
res$condition
```

```
##  [1] "histologically normal"       "simple ductal hyperplasia"
##  [3] "atypical ductal hyperplasia" "ductal carcinoma in situ"
##  [5] "histologically normal"       "atypical ductal hyperplasia"
##  [7] "ductal carcinoma in situ"    "histologically normal"
##  [9] "atypical ductal hyperplasia" "ductal carcinoma in situ"
## [11] "histologically normal"       "simple ductal hyperplasia"
## [13] "atypical ductal hyperplasia" "ductal carcinoma in situ"
## [15] "histologically normal"       "atypical ductal hyperplasia"
## [17] "ductal carcinoma in situ"    "histologically normal"
## [19] "atypical ductal hyperplasia" "ductal carcinoma in situ"
## [21] "histologically normal"       "atypical ductal hyperplasia"
## [23] "ductal carcinoma in situ"    "histologically normal"
## [25] "atypical ductal hyperplasia" "ductal carcinoma in situ"
## [27] "histologically normal"       "simple ductal hyperplasia"
## [29] "atypical ductal hyperplasia" "ductal carcinoma in situ"
## [31] "histologically normal"       "simple ductal hyperplasia"
## [33] "atypical ductal hyperplasia" "ductal carcinoma in situ"
## [35] "histologically normal"       "atypical ductal hyperplasia"
## [37] "ductal carcinoma in situ"    "histologically normal"
## [39] "atypical ductal hyperplasia" "ductal carcinoma in situ"
```

```
#clean white spaces
res$condition <- c("histologically_normal","simple_ductal_hyperplasia","atypical_ductal_hyperplasia","ductal
_carcinoma_in_situ","histologically_normal","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histol
ogically_normal","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","simple_du
ctal_hyperplasia","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","atypical
_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","atypical_ductal_hyperplasia","ducta
l_carcinoma_in_situ","histologically_normal","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histo
logically_normal","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","simple_d
uctal_hyperplasia","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","simple_
ductal_hyperplasia","atypical_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","atypic
al_ductal_hyperplasia","ductal_carcinoma_in_situ","histologically_normal","atypical_ductal_hyperplasia","duc
tal_carcinoma_in_situ")
res$condition
```

```
##  [1] "histologically_normal"       "simple_ductal_hyperplasia"
##  [3] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
##  [5] "histologically_normal"       "atypical_ductal_hyperplasia"
##  [7] "ductal_carcinoma_in_situ"    "histologically_normal"
##  [9] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
## [11] "histologically_normal"       "simple_ductal_hyperplasia"
## [13] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
## [15] "histologically_normal"       "atypical_ductal_hyperplasia"
## [17] "ductal_carcinoma_in_situ"    "histologically_normal"
## [19] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
## [21] "histologically_normal"       "atypical_ductal_hyperplasia"
## [23] "ductal_carcinoma_in_situ"    "histologically_normal"
## [25] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
## [27] "histologically_normal"       "simple_ductal_hyperplasia"
## [29] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
## [31] "histologically_normal"       "simple_ductal_hyperplasia"
## [33] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
## [35] "histologically_normal"       "atypical_ductal_hyperplasia"
## [37] "ductal_carcinoma_in_situ"    "histologically_normal"
## [39] "atypical_ductal_hyperplasia" "ductal_carcinoma_in_situ"
```

```
# Now we collapse the dataset with genesymbols, similar to what we did in phantasus
res <- collapseBy(res, fData(res)$`Gene symbol`, FUN=median)
res <- res[!grepl("///", rownames(res)), ]
res <- res[rownames(res) != "", ]
```

```
# let's annotate the symbols with the human database entries

fData(res) <- data.frame(row.names = rownames(res))
fData(res)$entrez <- row.names(fData(res))
fData(res)$symbol <- mapIds(org.Hs.eg.db, keys=fData(res)$entrez,
                                  keytype="SYMBOL",column="ENTREZID" )
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
# let's normalize this data

res.qnorm <- res

summary(exprs(res.qnorm))
```

```
##    GSM422873         GSM422874         GSM422875         GSM422876
##  Min.   :   0.10   Min.   :    0.1   Min.   :   0.10   Min.   :    0.10
##  1st Qu.:  23.80   1st Qu.:   28.0   1st Qu.:  28.30   1st Qu.:   30.60
##  Median :  61.85   Median :   77.5   Median :  81.15   Median :   94.15
##  Mean   : 171.92   Mean   :  223.3   Mean   : 254.73   Mean   :  293.53
##  3rd Qu.: 150.35   3rd Qu.:  206.5   3rd Qu.: 235.50   3rd Qu.:  274.07
##  Max.   :9700.40   Max.   :10890.8   Max.   :9974.90   Max.   :10874.10
##    GSM422877         GSM422878         GSM422879         GSM422880
##  Min.   :   0.1   Min.   :   0.0   Min.   :    0.1   Min.   :    0.1
##  1st Qu.:  20.7   1st Qu.:  14.6   1st Qu.:   18.0   1st Qu.:   14.8
##  Median :  66.0   Median :  46.9   Median :   61.2   Median :   46.7
##  Mean   : 215.8   Mean   : 192.2   Mean   :  250.5   Mean   :  190.7
##  3rd Qu.: 196.2   3rd Qu.: 158.3   3rd Qu.:  203.7   3rd Qu.:  153.3
##  Max.   :8736.5   Max.   :9052.4   Max.   :10988.9   Max.   :11150.7
##    GSM422881         GSM422882         GSM422883         GSM422884
##  Min.   :    0.0   Min.   :   0.00   Min.   :   0.1   Min.   :   0.20
##  1st Qu.:   16.0   1st Qu.:   9.70   1st Qu.:  15.1   1st Qu.:  17.30
##  Median :   49.7   Median :  29.30   Median :  38.7   Median :  42.20
##  Mean   :  217.4   Mean   : 124.89   Mean   : 105.2   Mean   :  98.85
##  3rd Qu.:  169.8   3rd Qu.:  91.97   3rd Qu.:  92.2   3rd Qu.:  91.70
##  Max.   :10044.1   Max.   :8542.20   Max.   :9757.2   Max.   :7475.00
##    GSM422885         GSM422886         GSM422887         GSM422888
##  Min.   :   0.1   Min.   :   0.2   Min.   :   0.10   Min.   :   0.1
##  1st Qu.:  19.8   1st Qu.:  17.9   1st Qu.:  17.73   1st Qu.:  13.4
##  Median :  51.3   Median :  44.7   Median :  56.75   Median :  39.4
##  Mean   : 130.6   Mean   : 121.3   Mean   : 236.29   Mean   : 175.4
##  3rd Qu.: 120.6   3rd Qu.: 106.9   3rd Qu.: 214.40   3rd Qu.: 146.4
##  Max.   :7821.9   Max.   :7877.8   Max.   :9292.90   Max.   :8618.7
##    GSM422889         GSM422890         GSM422891         GSM422892
##  Min.   :   0.1   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:  16.6   1st Qu.:  12.3   1st Qu.:  13.0   1st Qu.:  13.8
##  Median :  52.9   Median :  36.4   Median :  38.2   Median :  42.5
##  Mean   : 254.5   Mean   : 149.9   Mean   : 148.0   Mean   : 176.1
##  3rd Qu.: 219.0   3rd Qu.: 123.3   3rd Qu.: 122.3   3rd Qu.: 151.1
##  Max.   :9457.8   Max.   :8259.8   Max.   :7580.6   Max.   :9050.7
##    GSM422893         GSM422894         GSM422895         GSM422896
##  Min.   :    0.2   Min.   :    0.1   Min.   :    0.1   Min.   :    0.0
##  1st Qu.:   17.9   1st Qu.:   17.7   1st Qu.:   17.1   1st Qu.:   21.3
##  Median :   54.2   Median :   58.1   Median :   54.7   Median :   69.5
##  Mean   :  199.7   Mean   :  261.0   Mean   :  260.0   Mean   :  294.3
##  3rd Qu.:  161.3   3rd Qu.:  211.0   3rd Qu.:  206.6   3rd Qu.:  256.9
##  Max.   :10939.4   Max.   :11938.8   Max.   :10417.3   Max.   :12837.3
##    GSM422897         GSM422898         GSM422899         GSM422900
##  Min.   :    0.1   Min.   :    0.1   Min.   :   0.2   Min.   :    0.1
##  1st Qu.:   16.7   1st Qu.:   18.5   1st Qu.:  21.6   1st Qu.:   19.5
##  Median :   53.0   Median :   58.6   Median :  57.0   Median :   53.9
##  Mean   :  229.8   Mean   :  272.3   Mean   : 157.1   Mean   :  159.7
##  3rd Qu.:  184.6   3rd Qu.:  214.8   3rd Qu.: 139.9   3rd Qu.:  139.2
##  Max.   :10791.9   Max.   :11911.4   Max.   :9342.3   Max.   :10150.1
##    GSM422901         GSM422902         GSM422903         GSM422904
##  Min.   :    0.0   Min.   :    0.1   Min.   :   0.10   Min.   :   0.1
##  1st Qu.:   30.4   1st Qu.:   24.2   1st Qu.:   8.40   1st Qu.:   8.2
##  Median :   84.0   Median :   67.0   Median :  24.20   Median :  23.7
##  Mean   :  228.8   Mean   :  201.4   Mean   :  98.84   Mean   : 104.6
##  3rd Qu.:  218.8   3rd Qu.:  181.1   3rd Qu.:  72.40   3rd Qu.:  76.9
##  Max.   :10776.3   Max.   :10159.2   Max.   :6829.00   Max.   :6748.8
##    GSM422905         GSM422906         GSM422907         GSM422908
##  Min.   :   0.10   Min.   :   0.10   Min.   :   0.10   Min.   :    0.20
##  1st Qu.:   5.80   1st Qu.:   9.80   1st Qu.:  16.10   1st Qu.:   16.30
##  Median :  16.10   Median :  28.50   Median :  50.95   Median :   51.55
##  Mean   :  59.54   Mean   : 107.55   Mean   : 204.97   Mean   :  230.47
##  3rd Qu.:  43.50   3rd Qu.:  88.28   3rd Qu.: 177.30   3rd Qu.:  190.50
##  Max.   :4687.10   Max.   :8931.20   Max.   :9576.50   Max.   :10534.10
##    GSM422909         GSM422910         GSM422911         GSM422912
##  Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0.1
##  1st Qu.:  12.0   1st Qu.:   7.10   1st Qu.:   8.5   1st Qu.:   7.9
##  Median :  38.7   Median :  20.60   Median :  24.6   Median :  23.6
##  Mean   : 193.4   Mean   :  76.55   Mean   : 113.9   Mean   : 104.3
##  3rd Qu.: 146.3   3rd Qu.:  59.67   3rd Qu.:  86.0   3rd Qu.:  78.5
##  Max.   :7896.0   Max.   :4754.10   Max.   :5545.2   Max.   :6452.8
```

```
exprs(res.qnorm) <- normalizeBetweenArrays(log2(exprs(res.qnorm)+1), method="quantile")
summary(exprs(res.qnorm))
```

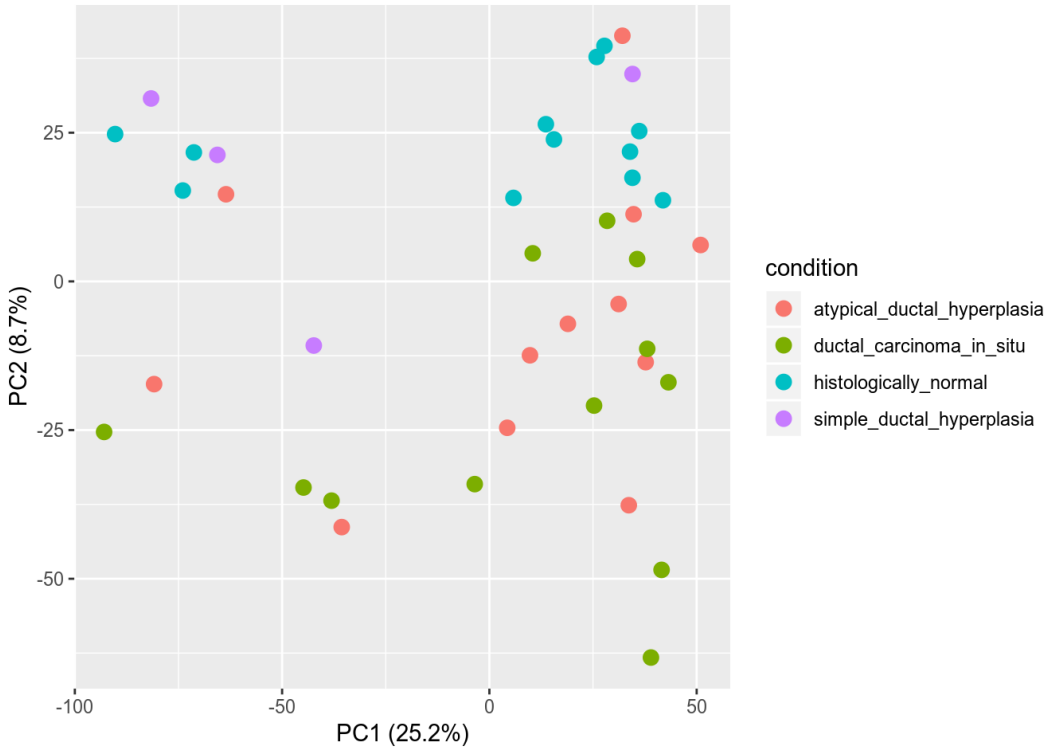
```
##    GSM422873         GSM422874         GSM422875         GSM422876
##  Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154
##  1st Qu.: 4.0462   1st Qu.: 4.0479   1st Qu.: 4.0512   1st Qu.: 4.0487
##  Median : 5.5486   Median : 5.5479   Median : 5.5483   Median : 5.5486
##  Mean   : 5.5693   Mean   : 5.5693   Mean   : 5.5693   Mean   : 5.5693
##  3rd Qu.: 7.1614   3rd Qu.: 7.1606   3rd Qu.: 7.1603   3rd Qu.: 7.1604
##  Max.   :13.1259   Max.   :13.1259   Max.   :13.1259   Max.   :13.1259
##    GSM422877         GSM422878         GSM422879         GSM422880
##  Min.   : 0.163   Min.   : 0.1154   Min.   : 0.239   Min.   : 0.163
##  1st Qu.: 4.048   1st Qu.: 4.0534   1st Qu.: 4.046   1st Qu.: 4.050
##  Median : 5.548   Median : 5.5472   Median : 5.551   Median : 5.549
##  Mean   : 5.569   Mean   : 5.5693   Mean   : 5.569   Mean   : 5.569
##  3rd Qu.: 7.160   3rd Qu.: 7.1600   3rd Qu.: 7.160   3rd Qu.: 7.160
##  Max.   :13.126   Max.   :13.1259   Max.   :13.126   Max.   :13.126
##    GSM422881         GSM422882         GSM422883         GSM422884
##  Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154
##  1st Qu.: 4.0462   1st Qu.: 4.0442   1st Qu.: 4.0484   1st Qu.: 4.0499
##  Median : 5.5477   Median : 5.5474   Median : 5.5506   Median : 5.5488
##  Mean   : 5.5694   Mean   : 5.5693   Mean   : 5.5693   Mean   : 5.5692
##  3rd Qu.: 7.1598   3rd Qu.: 7.1606   3rd Qu.: 7.1600   3rd Qu.: 7.1600
##  Max.   :13.1259   Max.   :13.1259   Max.   :13.1259   Max.   :13.1259
##    GSM422885         GSM422886         GSM422887         GSM422888
##  Min.   : 0.1154   Min.   : 0.163   Min.   : 0.163   Min.   : 0.2105
##  1st Qu.: 4.0467   1st Qu.: 4.047   1st Qu.: 4.048   1st Qu.: 4.0486
##  Median : 5.5479   Median : 5.548   Median : 5.549   Median : 5.5488
##  Mean   : 5.5693   Mean   : 5.569   Mean   : 5.569   Mean   : 5.5693
##  3rd Qu.: 7.1601   3rd Qu.: 7.160   3rd Qu.: 7.160   3rd Qu.: 7.1616
##  Max.   :13.1259   Max.   :13.126   Max.   :13.126   Max.   :13.1259
##    GSM422889         GSM422890         GSM422891         GSM422892
##  Min.   : 0.2105   Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154
##  1st Qu.: 4.0466   1st Qu.: 4.0525   1st Qu.: 4.0492   1st Qu.: 4.0455
##  Median : 5.5489   Median : 5.5509   Median : 5.5506   Median : 5.5488
##  Mean   : 5.5693   Mean   : 5.5694   Mean   : 5.5693   Mean   : 5.5693
##  3rd Qu.: 7.1604   3rd Qu.: 7.1600   3rd Qu.: 7.1605   3rd Qu.: 7.1603
##  Max.   :13.1259   Max.   :13.1259   Max.   :13.1259   Max.   :13.1259
##    GSM422893         GSM422894         GSM422895         GSM422896
##  Min.   : 0.2674   Min.   : 0.163   Min.   : 0.1154   Min.   : 0.1154
##  1st Qu.: 4.0492   1st Qu.: 4.053   1st Qu.: 4.0479   1st Qu.: 4.0524
##  Median : 5.5483   Median : 5.548   Median : 5.5489   Median : 5.5489
##  Mean   : 5.5693   Mean   : 5.569   Mean   : 5.5693   Mean   : 5.5693
##  3rd Qu.: 7.1606   3rd Qu.: 7.160   3rd Qu.: 7.1603   3rd Qu.: 7.1604
##  Max.   :13.1259   Max.   :13.126   Max.   :13.1259   Max.   :13.1259
##    GSM422897         GSM422898         GSM422899         GSM422900
##  Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.1154
##  1st Qu.: 4.0519   1st Qu.: 4.0483   1st Qu.: 4.0483   1st Qu.: 4.0513
##  Median : 5.5488   Median : 5.5494   Median : 5.5488   Median : 5.5486
##  Mean   : 5.5693   Mean   : 5.5693   Mean   : 5.5692   Mean   : 5.5693
##  3rd Qu.: 7.1612   3rd Qu.: 7.1603   3rd Qu.: 7.1600   3rd Qu.: 7.1612
##  Max.   :13.1259   Max.   :13.1259   Max.   :13.1259   Max.   :13.1259
##    GSM422901         GSM422902         GSM422903         GSM422904
##  Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.2105   Min.   : 0.3168
##  1st Qu.: 4.0488   1st Qu.: 4.0495   1st Qu.: 4.0440   1st Qu.: 4.0479
##  Median : 5.5488   Median : 5.5497   Median : 5.5519   Median : 5.5471
##  Mean   : 5.5692   Mean   : 5.5693   Mean   : 5.5693   Mean   : 5.5692
##  3rd Qu.: 7.1603   3rd Qu.: 7.1600   3rd Qu.: 7.1605   3rd Qu.: 7.1612
##  Max.   :13.1259   Max.   :13.1259   Max.   :13.1259   Max.   :13.1259
##    GSM422905         GSM422906         GSM422907         GSM422908
##  Min.   : 0.2674   Min.   : 0.163   Min.   : 0.163   Min.   : 0.2105
##  1st Qu.: 4.0556   1st Qu.: 4.049   1st Qu.: 4.047   1st Qu.: 4.0519
##  Median : 5.5471   Median : 5.552   Median : 5.548   Median : 5.5490
##  Mean   : 5.5694   Mean   : 5.569   Mean   : 5.569   Mean   : 5.5693
##  3rd Qu.: 7.1600   3rd Qu.: 7.160   3rd Qu.: 7.160   3rd Qu.: 7.1603
##  Max.   :13.1259   Max.   :13.126   Max.   :13.126   Max.   :13.1259
##    GSM422909         GSM422910         GSM422911         GSM422912
##  Min.   : 0.1154   Min.   : 0.1154   Min.   : 0.163   Min.   : 0.2674
##  1st Qu.: 4.0466   1st Qu.: 4.0525   1st Qu.: 4.048   1st Qu.: 4.0486
##  Median : 5.5500   Median : 5.5455   Median : 5.546   Median : 5.5488
##  Mean   : 5.5693   Mean   : 5.5693   Mean   : 5.569   Mean   : 5.5693
##  3rd Qu.: 7.1603   3rd Qu.: 7.1608   3rd Qu.: 7.161   3rd Qu.: 7.1600
##  Max.   :13.1259   Max.   :13.1259   Max.   :13.126   Max.   :13.1259
```

```
res.qnorm.top12K <- res.qnorm
# let's get top 12000 entries
res.qnorm.top12K <- res.qnorm.top12K[head(order(apply(exprs(res.qnorm.top12K), 1, mean),
                                                decreasing = TRUE), 12000), ]
```

```
# Now let's look at the dataset
## pdf('pca_dataset2.pdf')

#also we can make PCA plot from our dataset
pcaPlot(res.qnorm.top12K, 1, 2) + aes(color=condition)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```




```
# dev.off()
```

```
# Now we make a design matrix that will be used to make a model for the given data
res.design <- model.matrix(~0+condition, data=pData(res.qnorm.top12K))
res.design
```

```
##           conditionatypical_ductal_hyperplasia
## GSM422873                                    0
## GSM422874                                    0
## GSM422875                                    1
## GSM422876                                    0
## GSM422877                                    0
## GSM422878                                    1
## GSM422879                                    0
## GSM422880                                    0
## GSM422881                                    1
## GSM422882                                    0
## GSM422883                                    0
## GSM422884                                    0
## GSM422885                                    1
## GSM422886                                    0
## GSM422887                                    0
## GSM422888                                    1
## GSM422889                                    0
## GSM422890                                    0
## GSM422891                                    1
## GSM422892                                    0
## GSM422893                                    0
## GSM422894                                    1
## GSM422895                                    0
## GSM422896                                    0
## GSM422897                                    1
## GSM422898                                    0
## GSM422899                                    0
## GSM422900                                    0
## GSM422901                                    1
## GSM422902                                    0
## GSM422903                                    0
## GSM422904                                    0
## GSM422905                                    1
## GSM422906                                    0
## GSM422907                                    0
## GSM422908                                    1
## GSM422909                                    0
## GSM422910                                    0
## GSM422911                                    1
## GSM422912                                    0
##           conditionductal_carcinoma_in_situ conditionhistologically_normal
## GSM422873                                 0                               1
## GSM422874                                 0                               0
## GSM422875                                 0                               0
## GSM422876                                 1                               0
## GSM422877                                 0                               1
## GSM422878                                 0                               0
## GSM422879                                 1                               0
## GSM422880                                 0                               1
## GSM422881                                 0                               0
## GSM422882                                 1                               0
## GSM422883                                 0                               1
## GSM422884                                 0                               0
## GSM422885                                 0                               0
## GSM422886                                 1                               0
## GSM422887                                 0                               1
## GSM422888                                 0                               0
## GSM422889                                 1                               0
## GSM422890                                 0                               1
## GSM422891                                 0                               0
## GSM422892                                 1                               0
## GSM422893                                 0                               1
## GSM422894                                 0                               0
## GSM422895                                 1                               0
## GSM422896                                 0                               1
## GSM422897                                 0                               0
## GSM422898                                 1                               0
## GSM422899                                 0                               1
## GSM422900                                 0                               0
## GSM422901                                 0                               0
## GSM422902                                 1                               0
```

```
## GSM422903                                  0                                1
## GSM422904                                  0                                0
## GSM422905                                  0                                0
## GSM422906                                  1                                0
## GSM422907                                  0                                1
## GSM422908                                  0                                0
## GSM422909                                  1                                0
## GSM422910                                  0                                1
## GSM422911                                  0                                0
## GSM422912                                  1                                0
##           conditionsimple_ductal_hyperplasia
## GSM422873                                   0
## GSM422874                                   1
## GSM422875                                   0
## GSM422876                                   0
## GSM422877                                   0
## GSM422878                                   0
## GSM422879                                   0
## GSM422880                                   0
## GSM422881                                   0
## GSM422882                                   0
## GSM422883                                   0
## GSM422884                                   1
## GSM422885                                   0
## GSM422886                                   0
## GSM422887                                   0
## GSM422888                                   0
## GSM422889                                   0
## GSM422890                                   0
## GSM422891                                   0
## GSM422892                                   0
## GSM422893                                   0
## GSM422894                                   0
## GSM422895                                   0
## GSM422896                                   0
## GSM422897                                   0
## GSM422898                                   0
## GSM422899                                   0
## GSM422900                                   1
## GSM422901                                   0
## GSM422902                                   0
## GSM422903                                   0
## GSM422904                                   1
## GSM422905                                   0
## GSM422906                                   0
## GSM422907                                   0
## GSM422908                                   0
## GSM422909                                   0
## GSM422910                                   0
## GSM422911                                   0
## GSM422912                                   0
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$condition
## [1] "contr.treatment"
```

```
#we have 4 condition:
intermediate <- data.frame (res.design)
colnames(intermediate) <-c("conditionatypical_ductal_hyperplasia","conditionductal_carcinoma_in_situ", "cond
itionhistologically_normal", "conditionsimple_ductal_hyperplasia")

rm(res.design)
res.design <- as.matrix(intermediate)


# based on this matrix we fit our data
fit <- lmFit(res.qnorm.top12K, res.design)

# we will also make a bayisian model for the data called fit2
# this is the tricky part, because we need to choose contrast names which specify the sample groups to compa
re! we need to specify condition of interest and level to compare.

fit2 <- contrasts.fit(fit,makeContrasts(conditionhistologically_normal,conditionatypical_ductal_hyperplasia,
conditionhistologically_normal-conditionductal_carcinoma_in_situ, conditionhistologically_normal-conditionsi
mple_ductal_hyperplasia, levels=res.design))


# View(res_data2.design)
fit2 <- eBayes(fit2)

# now let's do a bonferroni-hochback correction
de <- topTable(fit2, adjust.method="BH", number=Inf)
head(de)
```

```
##        entrez symbol conditionhistologically_normal
## RPL39   RPL39   6170                       12.23252
## RPL37A RPL37A   6168                       12.85915
## RPS4X   RPS4X   6191                       12.54691
## LAMP1   LAMP1   3916                       12.41891
## RPL14   RPL14   9045                       12.29853
## RPS10   RPS10   6204                       12.91701
##        conditionatypical_ductal_hyperplasia
## RPL39                              12.18162
## RPL37A                             12.76278
## RPS4X                              12.51378
## LAMP1                              12.54163
## RPL14                              12.10078
## RPS10                              12.70586
##        conditionhistologically_normal...conditionductal_carcinoma_in_situ
## RPL39                                                          0.04723625
## RPL37A                                                         0.07666757
## RPS4X                                                          0.07553233
## LAMP1                                                         -0.13540110
## RPL14                                                          0.21157569
## RPS10                                                          0.14978246
##        conditionhistologically_normal...conditionsimple_ductal_hyperplasia
## RPL39                                                          0.035772903
## RPL37A                                                        -0.004462746
## RPS4X                                                         -0.058475821
## LAMP1                                                         -0.115227520
## RPL14                                                         -0.134144327
## RPS10                                                          0.152796922
##         AveExpr        F      P.Value    adj.P.Val
## RPL39  12.19950 22515.52 2.780155e-67 3.336186e-63
## RPL37A 12.80769 20465.68 1.925631e-66 9.555901e-63
## RPS4X  12.52016 20249.19 2.388975e-66 9.555901e-63
## LAMP1  12.50787 16379.42 1.760097e-64 4.388052e-61
## RPL14  12.18915 16348.70 1.828355e-64 4.388052e-61
## RPS10  12.79345 15353.59 6.530166e-64 1.306033e-60
```

```
# Here, we have a matrix that contains the enriched genes, we take the top genes and submit to database (msi
gdbr) to get the enriched pathways. We first target the hallmark pathways, which are well studied and then w
e target all the pathways. We try to find out what special pathways are involved in our normal versus condit
ion. This will further give us insight into the comparision.
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:IRanges':
##
##     shift
```

```
## The following objects are masked from 'package:S4Vectors':
##
##     first, second
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
de <- as.data.table(de, keep.rownames=TRUE)
de[entrez == "RPL37A"]
```

```
##        rn entrez symbol conditionhistologically_normal
## 1: RPL37A RPL37A   6168                       12.85915
##    conditionatypical_ductal_hyperplasia
## 1:                              12.76278
##    conditionhistologically_normal...conditionductal_carcinoma_in_situ
## 1:                                                         0.07666757
##    conditionhistologically_normal...conditionsimple_ductal_hyperplasia
## 1:                                                        -0.004462746
##     AveExpr        F      P.Value    adj.P.Val
## 1: 12.80769 20465.68 1.925631e-66 9.555901e-63
```

```
# BiocManager::install('fgsea')
library(fgsea)
```

```
## Loading required package: Rcpp
```

```
library(tibble)

# We use the matrix de to make a new matrix which contains annotated information about the pathways
# Let's make a new matrix de2 which will store information about pathways
de2 <- data.frame(de$entrez, de$P.Value)
colnames(de2) <- c('ENTREZ', 'stat')

# let's get the rank of genes from top differentially expressed to non significant
ranks <- deframe(de2)
head(ranks, 20)
```

```
##        RPL39       RPL37A        RPS4X        LAMP1        RPL14
## 2.780155e-67 1.925631e-66 2.388975e-66 1.760097e-64 1.828355e-64
##        RPS10        RPL41         EIF1        RPL12        RPL27
## 6.530166e-64 2.121528e-63 1.519644e-62 3.541907e-62 1.887323e-61
##        COPB1        HUWE1       HNRNPD          B2M        EEF1D
## 2.032329e-61 2.714707e-61 3.452425e-61 4.620123e-61 6.539432e-61
##       FTH1P5        RBM39        SUMO1         SKP1         RPL7
## 1.147525e-60 1.773340e-60 6.597470e-60 1.013885e-59 1.237329e-59
```

```
# Load the pathways into a named list
# BiocManager::install('msigdbr')
library(msigdbr)

m_df <- msigdbr(species = "Homo sapiens")
# View(m_df)
pathways <- split(m_df$human_gene_symbol, m_df$gs_name)
head(pathways)
```

```
## $AAACCAC_MIR140
##   [1] "ABCC4"     "ACTN4"     "ACVR1"     "ADAM9"     "ADAMTS5"
##   [6] "AGER"      "ANK2"      "API5"      "BACH1"     "BAZ2B"
##  [11] "BCL11A"    "BCL2L2"    "BCL9"      "C15orf29"  "C1orf21"
##  [16] "C3orf58"   "C7orf60"   "CACNA1C"   "CEBPA"     "CHD4"
##  [21] "CIT"       "COL23A1"   "CSK"       "CSNK1G3"   "CTCF"
##  [26] "CUL3"      "DAZL"      "DBNDD2"    "DCUN1D4"   "DDX3X"
##  [31] "DDX3Y"     "DHX57"     "DPP4"      "DSCAM"     "DTNA"
##  [36] "E2F3"      "EHD1"      "EPHB1"     "ERC2"      "ETV3"
##  [41] "EYA2"      "FAM123A"   "FAM175B"   "FAM178A"   "GABARAP"
##  [46] "GALNTL1"   "GDF6"      "GIT1"      "GYS1"      "HDAC4"
##  [51] "HNRNPH3"   "HSPA13"    "IGFBP5"    "KCND2"     "KIAA1370"
##  [56] "LOC440742" "LOXL3"     "LRRC4"     "LRRC8E"    "MAP3K8"
##  [61] "MDGA2"     "MEX3C"     "MGAT1"     "MMD"       "NAV3"
##  [66] "NKIRAS2"   "NR3C1"     "NUTF2"     "OGT"       "OSTM1"
##  [71] "PDGFRA"    "PFN1"      "PHF20L1"   "PHYHIP"    "PITX2"
##  [76] "PPP1CC"    "PRIMA1"    "R3HDM1"    "REEP1"     "RNF19A"
##  [81] "RTKN2"     "SENP1"     "SIAH1"     "SLC25A13"  "SLC38A2"
##  [86] "SLC41A2"   "SLMAP"     "SNX2"      "SOX4"      "SRR"
##  [91] "STAG1"     "STRADB"    "SYT6"      "TAF9B"     "TBX3"
##  [96] "TP53INP2"  "TSHZ1"     "TSPAN2"    "TSSK2"     "TTYH2"
## [101] "UBASH3B"   "USP6"      "VEGFA"     "WHSC1L1"   "WNT1"
## [106] "YES1"      "ZBED4"     "ZBTB10"    "ZNF182"    "ZNF608"
## [111] "ZNF654"
##
## $AAAGACA_MIR511
##   [1] "ABCG8"     "ACE"       "ADAMTSL3"  "ADGRF5"    "ADSS"
##   [6] "AGBL3"     "ALCAM"     "ANKZF1"    "AQP6"      "ARHGEF17"
##  [11] "ATL2"      "ATP2B2"    "ATRX"      "BCL11A"    "BTG1"
##  [16] "BUB3"      "BZRAP1"    "C11orf51"  "C18orf34"  "C1orf21"
##  [21] "C1QL2"     "C21orf59"  "C2orf71"   "C5orf41"   "C6orf106"
##  [26] "C7orf23"   "C7orf42"   "CALM1"     "CAMK2N1"   "CAMTA1"
##  [31] "CAPRIN1"   "CCND1"     "CCNT2"     "CDH2"      "CDK14"
##  [36] "CDK19"     "CELF1"     "CELF6"     "CEP350"    "CLK2"
##  [41] "CLTC"      "CNOT4"     "CORIN"     "CREM"      "CRIM1"
##  [46] "DCTN4"     "DDX3X"     "DDX3Y"     "DEDD"      "DNAJB12"
##  [51] "DNAJC13"   "DSC1"      "DUSP6"     "DYRK1B"    "E2F3"
##  [56] "EDEM3"     "EFR3A"     "EIF2C1"    "EIF2C2"    "EIF2C4"
##  [61] "ELAVL3"    "EMILIN2"   "EML4"      "ENPP1"     "ENPP4"
##  [66] "EPHA4"     "ESRRG"     "EYA1"      "EYA4"      "FAM117A"
##  [71] "FAM60A"    "FGF13"     "FIP1L1"    "FMR1"      "FN1"
##  [76] "FNDC1"     "FNDC5"     "FOXK2"     "FOXN3"     "GAD2"
##  [81] "GEMIN2"    "GFAP"      "GJA1"      "GLRA2"     "GPR116"
##  [86] "HAS2"      "HCN4"      "HLF"       "HLTF"      "HOXA13"
##  [91] "IGF2BP1"   "IGF2BP3"   "KCNE1"     "KCNMA1"    "KHDRBS2"
##  [96] "KIAA1429"  "KLF9"      "KLHL18"    "KLHL24"    "LATS1"
## [101] "LINC00483" "LMCD1"     "LPP"       "LRCH4"     "LUC7L3"
## [106] "MAP3K2"    "MAP4K4"    "MAPK1IP1L" "MBD2"      "MBD6"
## [111] "MDGA2"     "METAP2"    "MIB1"      "MINK1"     "MRPL21"
## [116] "MSTN"      "MTAP"      "MYCBP"     "MYO19"     "NACC1"
## [121] "NEUROD6"   "NHLH2"     "NLK"       "NR4A2"     "NRXN3"
## [126] "NTRK2"     "NXPH1"     "ONECUT2"   "PAX8"      "PCDH10"
## [131] "PCDH17"    "PELI1"     "PHLPP1"    "PIK3R3"    "PMEPA1"
## [136] "POGK"      "POU4F2"    "PPARGC1A"  "PRELP"     "PRPF4B"
## [141] "PSMA1"     "PSMD10"    "QKI"       "RAB22A"    "RAB2A"
## [146] "RBM15B"    "RBM26"     "RECK"      "REV3L"     "RGL1"
## [151] "RHOJ"      "RHOT1"     "RNF19A"    "ROBO2"     "RPS6KB1"
## [156] "RPS6KL1"   "SATB2"     "SCN4B"     "SEMA3F"    "SEMA6D"
## [161] "SEPP1"     "SLC22A17"  "SLC25A26"  "SLC6A6"    "SLITRK1"
## [166] "SMARCE1"   "SOCS2"     "SORCS3"    "SOST"      "SOX12"
## [171] "SPTBN4"    "SPTLC2"    "SRGAP3"    "SS18"      "ST18"
## [176] "SYT11"     "T"         "TAF5"      "THOC5"     "TIAL1"
## [181] "TMEM196"   "TNRC6A"    "TNRC6B"    "TOB1"      "TRAPPC3"
## [186] "TRAPPC8"   "TRIM2"     "TRIM24"    "TXNL1"     "UBE2H"
## [191] "VANGL2"    "VAV3"      "VKORC1L1"  "VMP1"      "WNT16"
## [196] "YTHDF2"    "YY1"       "ZADH2"     "ZCCHC24"   "ZDHHC21"
## [201] "ZNF319"    "ZNF654"    "ZNF706"
##
## $AAAGGAT_MIR501
##   [1] "ACACA"    "ACADSB"   "ADCYAP1"  "ADIPOR2"  "ALS2"     "AMMECR1"
##   [7] "APOLD1"   "ATP6V1H"  "BCL6"     "BCLAF1"   "C8orf82"  "CA6"
##  [13] "CACHD1"   "CAMTA1"   "CCDC140"  "CD164"    "CELF2"    "CELSR2"
```

```
##  [19] "CHODL"    "CLK1"     "CLK2"     "CTDSP1"   "CTDSPL2"  "CUL1"
##  [25] "CUX2"     "DCX"      "DNAJB12"  "ELAVL4"   "ERRFI1"   "FAM179B"
##  [31] "GIF"      "GRAMD4"   "GRB10"    "H2AFX"    "HAS2"     "HES5"
##  [37] "HOXB8"    "JUN"      "KCND2"    "KCNRG"    "KIAA2022" "KIF1C"
##  [43] "KIF2A"    "KLHL14"   "KRR1"     "LARP1"    "LEPROTL1" "LPGAT1"
##  [49] "LPIN1"    "LRRC1"    "MAP2K1"   "MAP3K8"   "MCU"      "MEF2C"
##  [55] "MYB"      "MYCL1"    "MYLK"     "NFASC"    "NFIL3"    "NFIX"
##  [61] "NPR3"     "NR2F2"    "NR4A3"    "PCDH19"   "PDK1"     "PHC1"
##  [67] "PHF16"    "PHF6"     "PIK3AP1"  "PITX2"    "PLP1"     "PLXNB1"
##  [73] "PNN"      "PPP1CB"   "PPP2R5E"  "PPP6R3"   "PRKCE"    "PURA"
##  [79] "QKI"      "RAB22A"   "RABGEF1"  "RASL10B"  "RCN1"     "RDX"
##  [85] "RET"      "RGL1"     "RNF11"    "ROBO2"    "RPGRIP1L" "RSBN1"
##  [91] "SATB2"    "SCN3A"    "SENP3"    "SEPHS1"   "SGPP1"    "SLC25A3"
##  [97] "SLC35B3"  "SLITRK5"  "SMC1A"    "SMEK1"    "SNAP29"   "SOX11"
## [103] "SOX4"     "SPOPL"    "SRR"      "SRSF2"    "SYNC"     "SYNJ1"
## [109] "SYT7"     "TAF5L"    "TAPT1"    "TNNI2"    "TOMM70A"  "TRIM39"
## [115] "UBAP1"    "UBE2Q1"   "UBE4B"    "USP12"    "VDAC2"    "WDFY3"
## [121] "WIPF2"    "WT1-AS"   "ZC3H7A"   "ZIC4"     "ZMYM5"    "ZNF238"
##
## $AAAGGGA_MIR204_MIR211
##   [1] "ADAMTS9"   "ADCY6"    "AKAP1"    "ALPL"     "ANGPT1"   "ANKRD13A"
##   [7] "ANXA11"    "AP1S1"    "AP1S3"    "AP2A2"    "AP3M1"    "APH1A"
##  [13] "ARAP2"     "ARCN1"    "ARGLU1"   "ARHGAP29" "ARL8B"    "ATF2"
##  [19] "ATP2B1"    "AUP1"     "BAZ2A"    "BCL11B"   "BCL2"     "BCL9"
##  [25] "BCL9L"     "BRD4"     "BRPF3"    "BUD31"    "C16orf72" "C17orf48"
##  [31] "C1orf144"  "C21orf63" "CAPRIN1"  "CCNT2"    "CCPG1"    "CDC25B"
##  [37] "CDC42"     "CDH2"     "CELSR3"   "CHD5"     "CHN2"     "CHP"
##  [43] "CLIP1"     "CORO1C"   "COX5A"    "CPD"      "CPNE8"    "CREB5"
##  [49] "CRKL"      "CTDNEP1"  "DAG1"     "DCAF5"    "DCUN1D3"  "DENND5A"
##  [55] "DHH"       "DLG5"     "DMTF1"    "DNAJC13"  "DNM2"     "DTX1"
##  [61] "DVL3"      "DYRK1A"   "EDEM1"    "EEF1E1"   "EFNB3"    "EIF2C4"
##  [67] "ELAVL3"    "ELF2"     "ELL2"     "ELMOD3"   "ELOVL6"   "EPHA7"
##  [73] "EPHB6"     "ESR1"     "ESRRG"    "EZR"      "FAM117B"  "FAM120C"
##  [79] "FAM122B"   "FAM160A2" "FAM175B"  "FARP1"    "FBN2"     "FBXW7"
##  [85] "FJX1"      "FNIP1"    "FRAS1"    "FREM1"    "FRY"      "GABRB3"
##  [91] "GAPVD1"    "GGA2"     "GLIS3"    "GPM6A"    "GRM1"     "HIC2"
##  [97] "HMGA2"     "HOOK3"    "HOXC8"    "HS2ST1"   "IGF2R"    "ING4"
## [103] "ITPR1"     "JPH3"     "KCNA3"    "KCTD1"    "KDM2A"    "KHDRBS1"
## [109] "KHDRBS3"   "KITLG"    "KLF12"    "KLHL13"   "LATS1"    "LRRC8D"
## [115] "MALL"      "MAML3"    "MAP1LC3B" "MAP3K3"   "MBNL1"    "MED13L"
## [121] "METAP1"    "MIR600HG" "MLL"      "MLLT3"    "MMGT1"    "MON2"
## [127] "MRPL35"    "MRPL52"   "MYO10"    "NAA15"    "NBEA"     "NCOA7"
## [133] "NEUROG1"   "NOVA1"    "NPTX1"    "NR3C1"    "NR4A2"    "NRBF2"
## [139] "NTRK2"     "P4HB"     "PCDH9"    "PHF13"    "PID1"     "PLAG1"
## [145] "POU3F2"    "PPARGC1A" "PPP3R1"   "PRDM2"    "PRPF38B"  "PRRX1"
## [151] "RAB10"     "RAB14"    "RAB1A"    "RAP2C"    "REEP1"    "RERE"
## [157] "RHOBTB3"   "RHOT1"    "RICTOR"   "RPS6KA3"  "RPS6KA5"  "RPS6KC1"
## [163] "RSPO3"     "RTKN2"    "RUNX2"    "SATB2"    "SCRT2"    "SEC24D"
## [169] "SEC61A2"   "SERINC3"  "SETD8"    "SF3B1"    "SGCZ"     "SGIP1"
## [175] "SHC1"      "SIN3A"    "SIRT1"    "SLC17A7"  "SLC22A2"  "SLC37A3"
## [181] "SLITRK4"   "SLTM"     "SMOC1"    "SOCS6"    "SOX11"    "SOX4"
## [187] "SPOP"      "SPRED1"   "SPRYD7"   "SSRP1"    "ST7"      "STXBP5"
## [193] "SUMO2"     "SUMO4"    "TAF5"     "TCF12"    "TCF7L1"   "TGFBR2"
## [199] "TMEM30A"   "TMOD3"    "TNRC6B"   "TP53INP1" "TRIAP1"   "TRIP12"
## [205] "TRPC5"     "TTYH1"    "UBE2R2"   "UHRF2"    "USP6"     "WEE1"
## [211] "WNT4"      "WSB1"     "XRN1"     "YTHDF3"   "YWHAG"    "ZCCHC14"
## [217] "ZCCHC24"   "ZDHHC17"  "ZFC3H1"   "ZFP91"    "ZFYVE20"  "ZNF282"
## [223] "ZNF335"    "ZNF423"
##
## $AAANWWTGC_UNKNOWN
##   [1] "ACTB"      "ADHFE1"    "AFF4"      "ANK2"      "ANK3"
##   [6] "APP"       "ASPA"      "ATOH7"     "ATP1B1"    "ATP2B4"
##  [11] "ATXN7L1"   "BCL11A"    "BCL6"      "BNC2"      "C11orf87"
##  [16] "C17orf85"  "CACNA1D"   "CACNG3"    "CALM1"     "CD14"
##  [21] "CDC42EP3"  "CDC42EP5"  "CDH13"     "CDK2AP1"   "CEPT1"
##  [26] "CHD2"      "CITED2"    "CNTFR"     "DAB1"      "DCAF11"
##  [31] "DCHS2"     "DDIT3"     "DIS3L"     "DLG2"      "DLGAP4"
##  [36] "DMD"       "DNAJB5"    "DPYSL5"    "DRD3"      "DSCAM"
##  [41] "DSEL"      "DSTN"      "DTX3L"     "DUSP1"     "DYNC1I2"
##  [46] "EBF1"      "EFNA5"     "EGFLAM"    "EIF4EBP2"  "ELAVL4"
##  [51] "ELF4"      "EPHA7"     "EPHB2"     "ESR1"      "FBXW7"
##  [56] "FGF7"      "FGFR2"     "FLJ45983"  "FN1"       "FOXN3"
##  [61] "FOXP1"     "FOXP2"     "FTHL17"    "FZD7"      "GANAB"
```

```
##  [66] "GATA3"     "GLRA2"     "GPC3"      "GPC6"      "GPR21"
##  [71] "GPRIN3"    "GRHL3"     "GRIN2B"    "GTF2E2"    "HEPACAM"
##  [76] "HHEX"      "HOXA2"     "HOXA3"     "HOXB2"     "HOXB6"
##  [81] "HOXC4"     "IGF2BP1"   "INHBA"     "ITM2C"     "KANK1"
##  [86] "KCNJ13"    "KLF12"     "KLF14"     "KRTAP8-1"  "LEAP2"
##  [91] "LECT1"     "LIPG"      "LOC148872" "LOX"       "LOXL4"
##  [96] "LRRC3B"    "LRRN1"     "LSAMP"     "LUC7L3"    "MAML3"
## [101] "MAN2A2"    "MAP3K4"    "MAPK3"     "MBNL1"     "MEF2C"
## [106] "MEIS1"     "MGLL"      "MID1"      "MLLT6"     "MMP3"
## [111] "MPZL3"     "MRPL24"    "MRPS18B"   "MYCL1"     "MYH2"
## [116] "MYLK"      "NEK6"      "NEUROG1"   "NFE2L2"    "NNAT"
## [121] "NR2F2"     "NRAS"      "NTN1"      "NTRK3"     "OLFM1"
## [126] "OLIG2"     "OMG"       "OTX2"      "PATZ1"     "PAX1"
## [131] "PAX6"      "PCSK1"     "PCTP"      "PDGFRB"    "PHF15"
## [136] "PHOX2B"    "PHTF1"     "PIK3R3"    "POU2F1"    "POU4F1"
## [141] "PPARGC1A"  "PPFIA2"    "PPP1R10"   "PPP2R2A"   "PPP3CC"
## [146] "PRDM16"    "PRIMA1"    "PRKRIR"    "PRPF4B"    "RAB10"
## [151] "RBMX"      "RORA"      "RRS1"      "RSPO2"     "S100PBP"
## [156] "SALL3"     "SAMD12"    "SATB2"     "SEMA6C"    "SESN2"
## [161] "SFRP2"     "SGCD"      "SHC3"      "SIX5"      "SKIL"
## [166] "SKP2"      "SLMAP"     "SNCAIP"    "SNX25"     "SORT1"
## [171] "SOX13"     "SOX4"      "SOX5"      "SPAG9"     "SPARCL1"
## [176] "SSBP3"     "STEAP2"    "TBC1D8B"   "TFAP4"     "TFDP2"
## [181] "TGIF1"     "THBS2"     "TLE4"      "TLK1"      "TLX3"
## [186] "TRAM1"     "TRPM3"     "TSC22D4"   "ZFPM1"     "ZHX3"
## [191] "ZNF462"    "ZNF827"    "ZW10"
##
## $AAAYRNCTG_UNKNOWN
##   [1] "ABT1"      "ACVR1"     "ADAM12"    "ADD3"      "AGGF1"
##   [6] "ANKRD12"   "ANKRD28"   "AP4S1"     "APBB2"     "APOBR"
##  [11] "AQP2"      "ARHGAP44"  "ARID1A"    "ARID4A"    "ARPC2"
##  [16] "ARSG"      "ARX"       "ASB4"      "ASPH"      "ATOH8"
##  [21] "ATP1A2"    "ATP5L"     "ATPIF1"    "AXDND1"    "B4GALT6"
##  [26] "BAI3"      "BAMBI"     "BCL2L1"    "BCL9"      "BMPR1B"
##  [31] "BMX"       "BRSK2"     "BTBD3"     "BUB3"      "C11orf84"
##  [36] "C11orf92"  "C12orf65"  "C13orf30"  "C14orf1"   "C15orf26"
##  [41] "C17orf28"  "C20orf197" "C3orf19"   "C6orf138"  "CA3"
##  [46] "CACNA2D3"  "CACNB2"    "CAPN1"     "CAPZA1"    "CASQ2"
##  [51] "CBX2"      "CCNJ"      "CCNY"      "CDC23"     "CDH2"
##  [56] "CER1"      "CHRM1"     "CITED2"    "CLDN5"     "CLTC"
##  [61] "CMKLR1"    "CNTLN"     "CNTN1"     "COCH"      "COL12A1"
##  [66] "COL1A2"    "COL4A5"    "COL4A6"    "COLEC10"   "CRAT"
##  [71] "CRH"       "CRKL"      "CRYGD"     "CRYGS"     "CSNK1A1"
##  [76] "CSRNP3"    "CSTF3"     "CYBRD1"    "DAAM1"     "DBNDD2"
##  [81] "DCAKD"     "DDAH2"     "DDX4"      "DEF6"      "DENND4A"
##  [86] "DGKB"      "DHH"       "DHRS4"     "DHRS4L2"   "DIDO1"
##  [91] "DMD"       "DMRT1"     "DNAJA2"    "DNAJB3"    "DNAJB4"
##  [96] "DSCAML1"   "DUSP4"     "DYNC1I1"   "DYRK1A"    "EDA"
## [101] "EFNA1"     "EGFLAM"    "EIF5"      "EMX2"      "EPC1"
## [106] "EPHA7"     "ERBB4"     "ERRFI1"    "ESRP2"     "ESRRB"
## [111] "ESRRG"     "EYA1"      "FAM49A"    "FAM83F"    "FCER1A"
## [116] "FGD4"      "FGF10"     "FGF12"     "FGFR1"     "FGFR1OP2"
## [121] "FIZ1"      "FKRP"      "FMNL3"     "FNDC9"     "FOXA1"
## [126] "FOXG1"     "FOXO4"     "FOXP2"     "FSIP2"     "FST"
## [131] "GABRA3"    "GDNF"      "GFI1"      "GGNBP2"    "GJB4"
## [136] "GLDN"      "GNAQ"      "GPR85"     "GPRC5D"    "GRIN2B"
## [141] "H3F3A"     "HDAC8"     "HESX1"     "HEXIM2"    "HGF"
## [146] "HIC2"      "HIP1R"     "HN1"       "HOXA10"    "HOXA5"
## [151] "HOXB8"     "HPSE2"     "HSD3B7"    "ICAM4"     "ID1"
## [156] "IGF1"      "IL1RAPL1"  "INHBC"     "IP6K2"     "ITGA10"
## [161] "ITGA8"     "JPH1"      "KANK2"     "KCNIP2"    "KCNK5"
## [166] "KCNN3"     "KCNQ1DN"   "KIAA0182"  "KITLG"     "KLF5"
## [171] "KLHDC10"   "KLHL20"    "KLHL3"     "LARS2"     "LENG9"
## [176] "LHFP"      "LHX9"      "LMO7"      "LOC151534" "LRP5"
## [181] "LRRC4"     "LRRN4CL"   "LTBP1"     "MAML1"     "MANF"
## [186] "MAP2"      "MAP3K5"    "MAP6"      "MEIS1"     "MGAT1"
## [191] "MGAT4A"    "MID1"      "MLL"       "MOAP1"     "MPP6"
## [196] "MPPED2"    "MRPL13"    "MTA2"      "MTBP"      "MYF6"
## [201] "MYH1"      "MYH10"     "MYO18A"    "NAGLU"     "NAPB"
## [206] "NAV2"      "NAV3"      "NCDN"      "NDNF"      "NDST4"
## [211] "NDUFS4"    "NEK1"      "NEK2"      "NFATC4"    "NFYB"
## [216] "NMI"       "NMT1"      "NR2F1"     "NRG1"      "NTRK2"
## [221] "NUP54"     "NXPH4"     "OMA1"      "OMG"       "OR2L13"
```

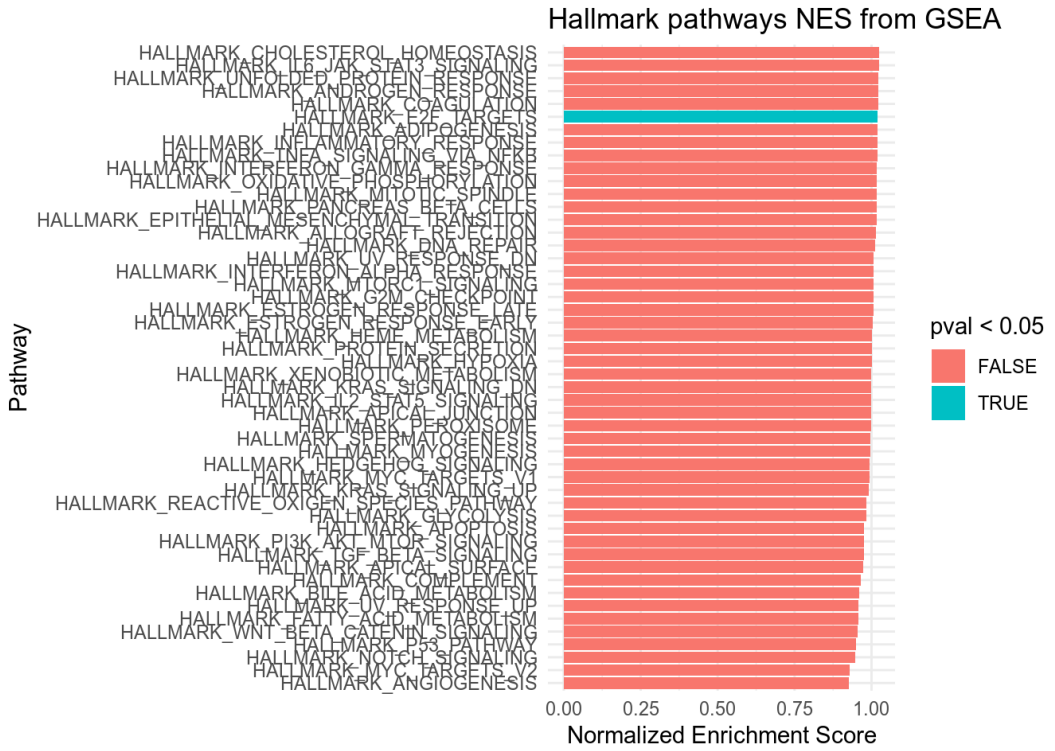
```
## [226] "OTX2"      "PACRG"     "PAPD5"     "PARK2"     "PART1"
## [231] "PCDH17"    "PCDH18"    "PCF11"     "PCYT1B"    "PDGFB"
## [236] "PDGFRA"    "PDLIM2"    "PDS5B"     "PDZRN4"    "PFN2"
## [241] "PHC2"      "PHEX"      "PHF1"      "PHF15"     "PHF6"
## [246] "PHOX2B"    "PLAGL2"    "PLEC"      "PLEKHM1"   "PLP2"
## [251] "PMCH"      "PMCHL1"    "PODXL2"    "POFUT1"    "POU2AF1"
## [256] "POU4F1"    "PPAP2B"    "PPP1R9B"   "PPP2R3A"   "PPP2R4"
## [261] "PPP2R5E"   "PPP3CA"    "PRELP"     "PRKCG"     "PRKCQ"
## [266] "PROK2"     "PTH1R"     "PXN"       "R3HDM1"    "RAB30"
## [271] "RAB5B"     "RAB5C"     "RAPGEF4"   "RBMS3"     "RGS17"
## [276] "RNF146"    "ROBO4"     "ROR1"      "RPLP0"     "RTN1"
## [281] "RUFY3"     "S1PR2"     "SCN3B"     "SCN5A"     "SCN8A"
## [286] "SCOC"      "SDCBP"     "SEMA6D"    "SEPT7"     "SESN3"
## [291] "SGCD"      "SH2D6"     "SHC3"      "SHCBP1L"   "SIPA1"
## [296] "SIRPA"     "SLC26A6"   "SLC4A1"    "SLC6A1"    "SMARCA2"
## [301] "SNX9"      "SORBS2"    "SOX12"     "SOX21"     "SOX30"
## [306] "SOX5"      "SPOCK2"    "SPTLC2"    "SRGAP2"    "SRSF8"
## [311] "SSBP2"     "ST7L"      "STAC3"     "STAG1"     "STAG2"
## [316] "STC2"      "STRN3"     "STRN4"     "TAS1R2"    "TEF"
## [321] "TFAP4"     "TFDP2"     "TM2D3"     "TMEM182"   "TMEM27"
## [326] "TMEM69"    "TMSB4X"    "TMSB4XP1"  "TMSL3"     "TMSL6"
## [331] "TNFAIP8"   "TNS1"      "TNXB"      "TP53INP2"  "TRDN"
## [336] "TREML1"    "TRIM28"    "TRIM68"    "TRIM8"     "TRIML1"
## [341] "TRPS1"     "TSC22D3"   "TSPAN7"    "TSPY26P"   "TSSK3"
## [346] "TTC17"     "TUSC2"     "UBE2W"     "UBXN10"    "USP1"
## [351] "VDR"       "VIP"       "VKORC1L1"  "VWA5A"     "WBP1"
## [356] "WNT2B"     "WT1"       "WT1-AS"    "XRCC1"     "ZADH2"
## [361] "ZBTB11"    "ZFP91"     "ZFPM2"     "ZIC1"      "ZIC4"
## [366] "ZMAT3"     "ZNF238"    "ZNF296"    "ZNF503"    "ZNF521"
## [371] "ZNF524"    "ZNF654"    "ZNF687"    "ZNF710"
```

```
# filter the list to include only hallmark pathways

library(data.table)

pathways.hallmark <- m_df[m_df$gs_name %like% "HALLMARK_", ]
pathways.hallmark <- split(pathways.hallmark$human_gene_symbol, pathways.hallmark$gs_name)

# Show the first few pathways, and within those, show only the first few genes.
pathways.hallmark %>%
  head() %>%
  lapply(head)
```

```
## $HALLMARK_ADIPOGENESIS
## [1] "ABCA1" "ABCB8" "ACAA2" "ACADL" "ACADM" "ACADS"
##
## $HALLMARK_ALLOGRAFT_REJECTION
## [1] "AARS"   "ABCE1"  "ABI1"   "ACHE"   "ACVR2A" "AKT1"
##
## $HALLMARK_ANDROGEN_RESPONSE
## [1] "ABCC4"   "ABHD2"   "ACSL3"   "ACTN1"   "ADAMTS1" "ADRM1"
##
## $HALLMARK_ANGIOGENESIS
## [1] "APOH"   "APP"    "CCND2"  "COL3A1" "COL5A2" "CXCL6"
##
## $HALLMARK_APICAL_JUNCTION
## [1] "ACTA1" "ACTB"  "ACTC1" "ACTG1" "ACTG2" "ACTN1"
##
## $HALLMARK_APICAL_SURFACE
## [1] "ADAM10"  "ADIPOR2" "AFAP1L2" "AIM1"    "AKAP7"   "APP"
```

```
# running the fgsea algorithm on hallmark.pathways

fgseaRes <- fgsea(pathways=pathways.hallmark, stats=ranks, nperm=1000)

fgseaResTidy <- fgseaRes %>%
  as_tibble() %>%
  arrange(desc(NES))  #ggploting for halmark pathways
```

```
# ggplotting for hallmark pathways
library(ggplot2)
# pdf("fgseaResTidy2.pdf", width = 10, height = 10)

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) +
  geom_col(aes(fill=pval<0.05)) +
  coord_flip() +
  labs(x="Pathway", y="Normalized Enrichment Score",
       title="Hallmark pathways NES from GSEA") +
  theme_minimal()
```



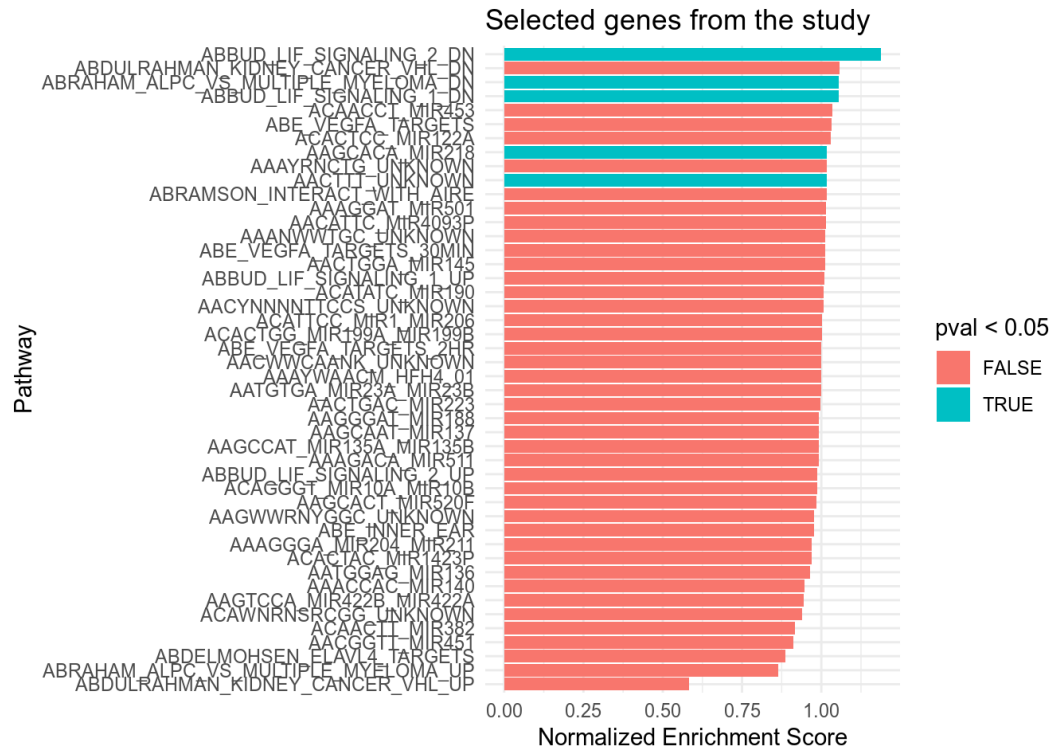

```
## dev.off()
```

```
# We have plotted all the significant patways in the hallmark pathways as 'blue'
# We can see that:
    # HALLMARK_E2F_TARGETS
# pathway is activated!

# Let's look at all pathways involving the following genes that they mentioned in the paper
# ACTG2,ADAMTS1,CAPN6,CAV1,CAV2,CCND2,COL14A1,COL15A1,EGF,EGFR,FGF1,FGF2,FGFR2,FIGF,FN1,FYN,IGF1,ITGA10,LAMA
2,LAMA3,LAMB1,LAMB3,LAMC2,MME,MYLK,NCAM1,PAK3,PDGFA,PDGFD,PDGFRA,PIK3R1,PIK3R3,PIP5K1B,PPP1R12B,RELN,SPP1,TH
BS1,TIAM1,TNN,TNXB,VCAM1,VEGFA

# We are going to search the entire pathway list for any pathway that contains these genes, this can be done
by subsetting and appending to a new dataframe of pathways.

#_-----------------------------------------------

# let's make a list of all pathways fgseares.all
fgseaRes.all <- fgsea(pathways=pathways, stats=ranks, nperm=1000)

item <- data.frame('ACTG2','ADAMTS1','CAPN6','CAV1','CAV2','CCND2','COL14A1','COL15A1','EGF','EGFR','FGF1',
'FGF2','FGFR2','FIGF','FN1','FYN','IGF1','ITGA10','LAMA2','LAMA3','LAMB1','LAMB3','LAMC2','MME','MYLK','NCAM
1','PAK3','PDGFA','PDGFD','PDGFRA','PIK3R1','PIK3R3','PIP5K1B','PPP1R12B','RELN','SPP1','THBS1','TIAM1','TN
N','TNXB','VCAM1','VEGFA')

item<- t(item)
rownames(item) <- NULL

entry <- function(){

  x<- for (i in item){
    print(de[entrez == i])

  }

  return(x)

}

# searching for the genes in pathway and appending the rownumbers
# sink('numbers.txt')
#
# options(max.print=2000)
#
# for(i in item){
#   print(grep(i, fgseaRes.all$leadingEdge))
# }
#
# sink()

# we have to do a lot of cleaning of the data before importing it as csv
# getting only unique values from all numbers, because one gene may overlap with other, we only want the uni
que row numbers
numbers <- read.delim("~/Documents/rnaseq/data2/numbers.txt", header=FALSE, comment.char="#")

unique_vals <- data.frame(as.integer(unique(unlist(numbers))))

colnames(unique_vals) <- c('row_number')

# View(unique_vals)

pathways.final <- subset(fgseaRes.all, rownames(fgseaRes.all) %in% unique_vals$row_number)

#View(pathways.final)

#_-----------------------------------------------


# Show the first few pathways, and within those, show only the first few genes.
pathways.final %>%
  head() %>%
  lapply(head)
```

```
## $pathway
## [1] "AAACCAC_MIR140"        "AAAGACA_MIR511"        "AAAGGAT_MIR501"
## [4] "AAAGGGA_MIR204_MIR211" "AAANWWTGC_UNKNOWN"     "AAAYRNCTG_UNKNOWN"
##
## $pval
## [1] 0.97802198 0.77722278 0.17582418 0.97702298 0.17482517 0.05394605
##
## $padj
## [1] 1.0000000 1.0000000 0.6624186 1.0000000 0.6618753 0.6371385
##
## $ES
## [1] 0.9231347 0.9709258 0.9928014 0.9514715 0.9918829 0.9993489
##
## $NES
## [1] 0.9461955 0.9909595 1.0158418 0.9706567 1.0129512 1.0174617
##
## $nMoreExtreme
## [1] 978 777 175 977 174  53
##
## $size
## [1]  80 155  97 167 139 257
##
## $leadingEdge
## $leadingEdge[[1]]
## [1] "PITX2"   "CACNA1C" "TSPAN2"  "WNT1"
##
## $leadingEdge[[2]]
## [1] "EPHA4"  "PMEPA1" "ACE"
##
## $leadingEdge[[3]]
## [1] "TNNI2" "PITX2"
##
## $leadingEdge[[4]]
## [1] "MRPL35"  "WNT4"    "JPH3"    "HOXC8"   "EPHA7"   "NEUROG1" "FAM120C"
##
## $leadingEdge[[5]]
## [1] "DRD3"  "INHBA"
##
## $leadingEdge[[6]]
## [1] "WT1"  "ZIC1"
```

```
final <- data.frame(pathways.final)
# running the fgsea algorithm on final pathways
# Let's look at the plot
```

```
# ggplotting for final pathways
library(ggplot2)

ggplot(final, aes(reorder(pathway, NES), NES)) +
  geom_col(aes(fill=pval<0.05)) +
  coord_flip() +
  labs(x="Pathway", y="Normalized Enrichment Score",
       title="Selected genes from the study") +
  theme_minimal()
```



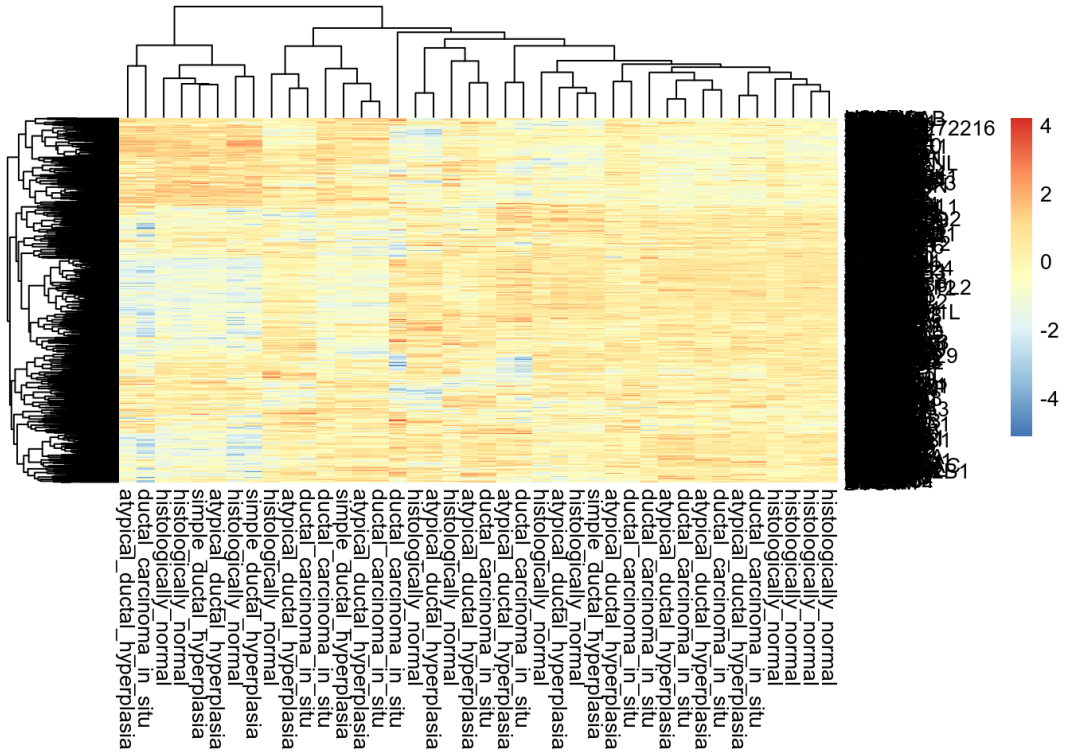

```
# install.packages('DT')
library(DT)
# Show in a table for all pathways
fgseaResTidy %>%
  dplyr::select(-leadingEdge, -ES, -nMoreExtreme) %>%
  arrange(padj) %>%
  DT::datatable()
```

Show 10 entries Search:

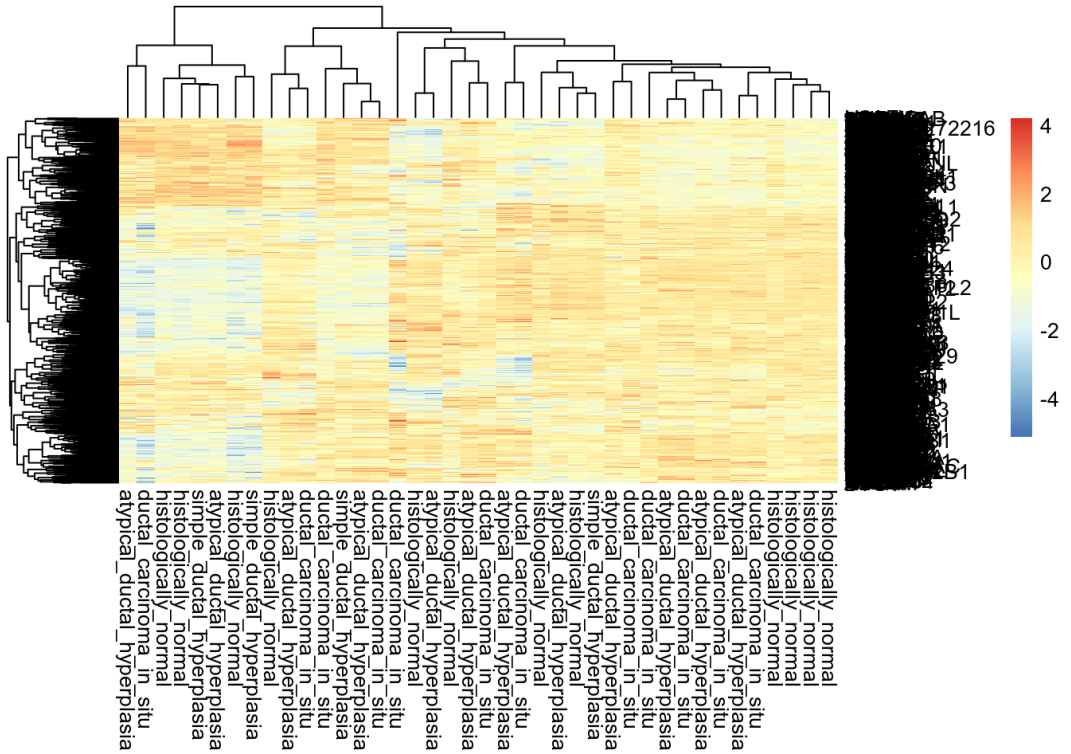
| | pathway | pval | padj | NES | size |
|---|---|---|---|---|---|
| 1 | HALLMARK_IL6_JAK_STAT3_SIGNALING | 0.11988011988012 | 0.749250749250749 | 1.02360270260286 | 84 |
| 2 | HALLMARK_UNFOLDED_PROTEIN_RESPONSE | 0.0969030969030969 | 0.749250749250749 | 1.02286366488764 | 105 |
| 3 | HALLMARK_ANDROGEN_RESPONSE | 0.111888111888112 | 0.749250749250749 | 1.02270179664728 | 96 |
| 4 | HALLMARK_COAGULATION | 0.0809190809190809 | 0.749250749250749 | 1.02197825962398 | 125 |
| 5 | HALLMARK_E2F_TARGETS | 0.017982017982018 | 0.749250749250749 | 1.02053068243033 | 185 |
| 6 | HALLMARK_ADIPOGENESIS | 0.143856143856144 | 0.749250749250749 | 1.01858827678428 | 170 |
| 7 | HALLMARK_INFLAMMATORY_RESPONSE | 0.135864135864136 | 0.749250749250749 | 1.01844500970978 | 185 |
| 8 | HALLMARK_TNFA_SIGNALING_VIA_NFKB | 0.131868131868132 | 0.749250749250749 | 1.01828402986031 | 191 |
| 9 | HALLMARK_INTERFERON_GAMMA_RESPONSE | 0.153846153846154 | 0.749250749250749 | 1.01790852730123 | 176 |
| 10 | HALLMARK_OXIDATIVE_PHOSPHORYLATION | 0.153846153846154 | 0.749250749250749 | 1.01780765899996 | 192 |

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

```
# heatmap
library(pheatmap)

#scale rows
xt <-t(as.matrix(res.qnorm.top12K)) # this is a matrix of normalised 12k genes
# let's get a heatmap of 1000 genes suggested in the study
xts <-scale(xt)
xtst <-t(xts)
xtst <- na.omit(xtst)
colnames(xtst) <- res$condition

#only grab top 1048 by p-value
h<-head(xtst, n = 1048L)

#set layout options - adjust if labels get cut off
#pdf("heatmap.pdf",width=10, height=100)

#draw heatmap allowing larger margins and adjusting row label font size
# pdf('heatmap.pdf', width = 20, height = 200)

pheatmap(h)
```

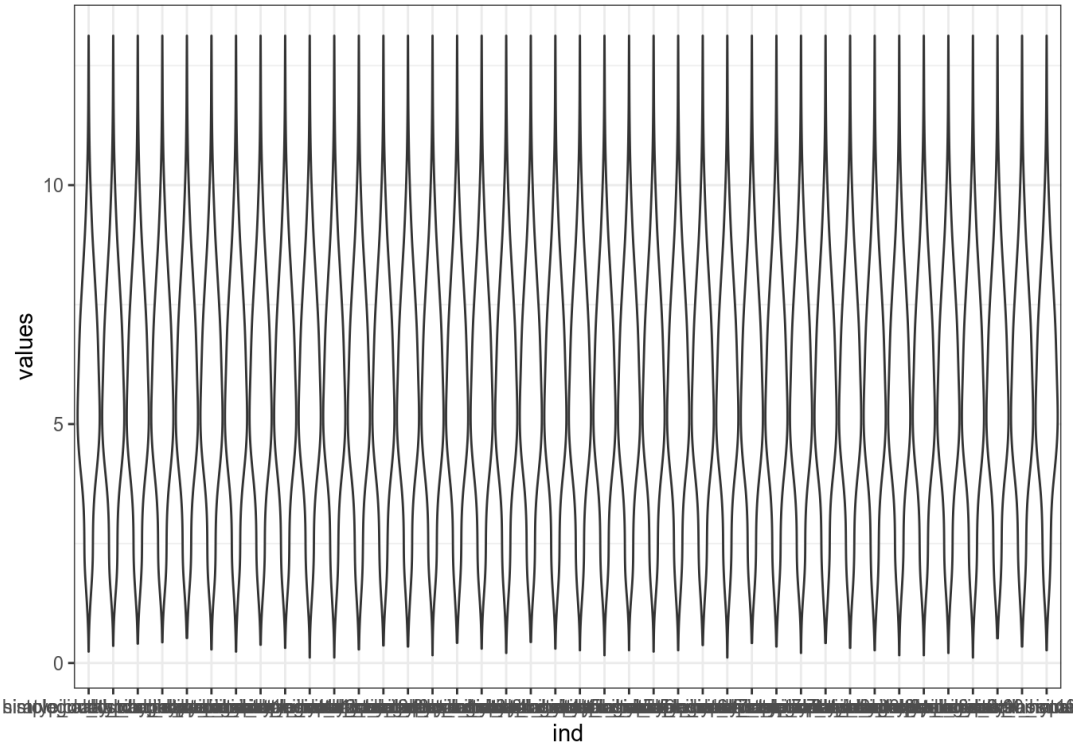



```
#output plot to file
 # dev.off()

 # let's get a smaller heatmap
 # pdf('heatmap_small.pdf', width = 20, height = 50)

pheatmap(h)
```




```
#output plot to file
 # dev.off()
```

```
# let's make a boxplot of the data

# install.packages('devtools')
library(devtools)
# devtools::install_github("sinhrks/ggfortify")
library(ggfortify)

#pdf('box_dataset.pdf', width = 50)

gt <- t(xt) # taking xt from the heatmap and transposing it
colnames(gt)<- res$condition # now giving it labels from condition
ggplot(stack(data.frame(gt)), aes(x = ind, y = values)) +
  geom_violin() + theme_bw()
```



```
#dev.off()
```