happy path implementation

2023-03-25

Happy Path implementation for grnaeR

This is an happy path implementation for our designed functionality Find_DEG and gene_enrichment_visualization.

First, we test on the gene visualization function using the example data geneList from the DOSE library. Please note other related dependencies should also be installed to run the package.

Make sure you have had installed the package and example data

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
    BiocManager::install("DOSE")
```

```
## Bioconductor version 3.16 (BiocManager 1.30.19), R 4.2.2 (2022-10-31)
```

```
## Warning: package(s) not installed when version(s) same as or greater than
current; use
##   `force = TRUE` to re-install: 'DOSE'
```

```
## Old packages: 'AnnotationDbi', 'BiocManager', 'BiocParallel', 'blob', 'boo
t',
##   'broom', 'class', 'cli', 'codetools', 'commonmark', 'dbplyr', 'dplyr',
##   'dtplyr', 'fastmap', 'foreign', 'gh', 'googledrive', 'googlesheets4',
##   'gtable', 'haven', 'hms', 'htmltools', 'htmlwidgets', 'httpuv', 'httr',
##   'Matrix', 'mgcv', 'modelr', 'nlme', 'openssl', 'pillar', 'ps',
##   'RcppArmadillo', 'RSQLite', 'S4Vectors', 'sourcetools', 'spatial',
##   'survival', 'tibble', 'tinytex', 'utf8', 'vctrs', 'xfun', 'XML'
```

```
# Notice: the package should reside under current working directory for insta
llation
install.packages('grnaeR_0.1.0.tar.gz',repos = NULL)
```

Step 1:Happy PATH: Gene_enrichment_visualization

Load the example data geneList

```
library(DOSE)
```

```
##
```

```
## DOSE v3.24.2  For help: https://yulab-smu.top/biomedical-knowledge-mining-
book/
##
## If you use DOSE in published research, please cite:
```

```
## Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an R/Biocon
ductor package for Disease Ontology Semantic and Enrichment analysis. Bioinfo
rmatics 2015, 31(4):608-609

data(geneList)
library('grnaeR')

## Loading required package: tidyverse

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors

library(enrichplot)
library(ggpubr)

##
## Attaching package: 'ggpubr'
##
## The following object is masked from 'package:enrichplot':
##
##     color_palette

library('org.Hs.eg.db')

## Loading required package: AnnotationDbi
## Loading required package: stats4
## Loading required package: BiocGenerics
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
##
## 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, aperm, 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.max, which.min
##
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
##
## Loading required package: IRanges
## Loading required package: S4Vectors
##
## Attaching package: 'S4Vectors'
##
## The following objects are masked from 'package:lubridate':
##
##     second, second<-
##
## The following objects are masked from 'package:dplyr':
##
##     first, rename
##
## The following object is masked from 'package:tidyr':
##
##     expand
##
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
##
##
## Attaching package: 'IRanges'
##
## The following object is masked from 'package:lubridate':
##
##     %within%
##
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
```

```
##
## The following object is masked from 'package:purrr':
##
##     reduce
##
##
## Attaching package: 'AnnotationDbi'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(DESeq2)
```

```
## Loading required package: GenomicRanges
## Loading required package: GenomeInfoDb
## Loading required package: SummarizedExperiment
## Loading required package: MatrixGenerics
## Loading required package: matrixStats
##
## Attaching package: 'matrixStats'
##
## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians
##
## The following object is masked from 'package:dplyr':
##
##     count
##
##
## Attaching package: 'MatrixGenerics'
##
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
##
## The following object is masked from 'package:Biobase':
##
##     rowMedians
```

```
library('RColorBrewer')
library("pheatmap")
```

The package conflict still exist, effort should be paid to solve the problem

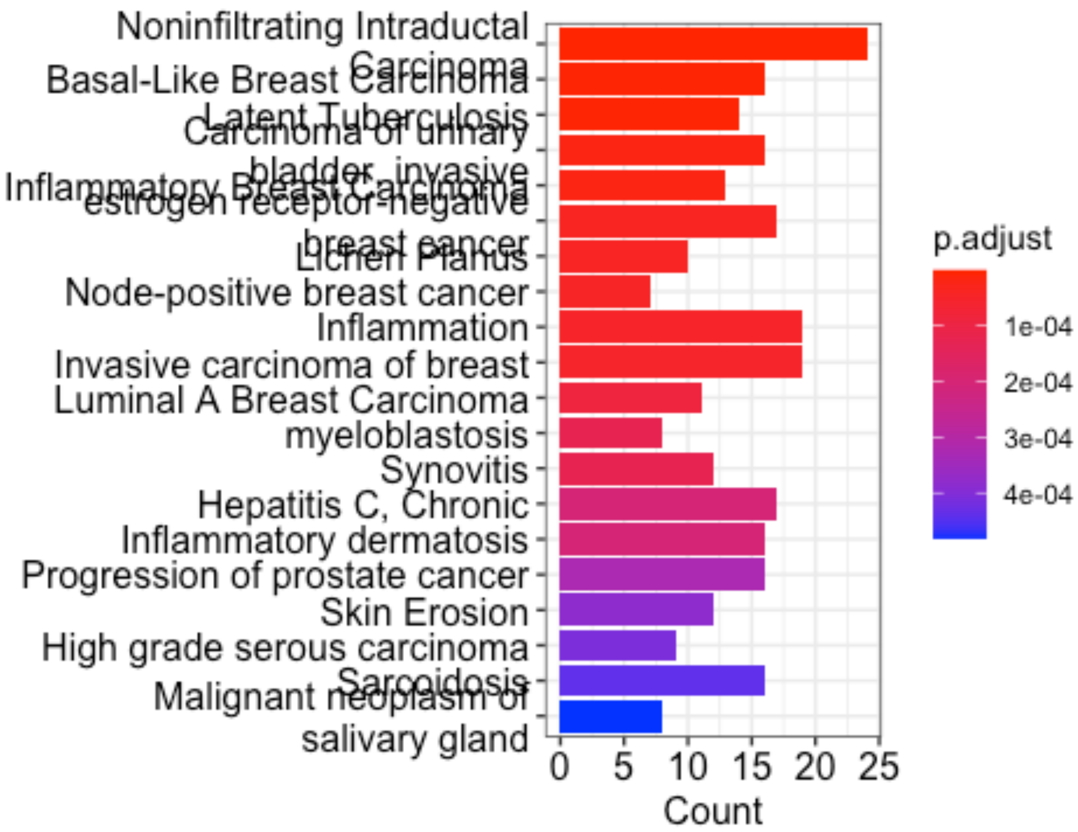
Here, we call the filter_genelist function to select interested genes with abolute value greater than assigned threshold and convert to large enrichResult

```
edo <- grnaeR::filter_genelist(geneList,standard_fc = 2)
```

```
## [1] "enrichResult object generated"
```

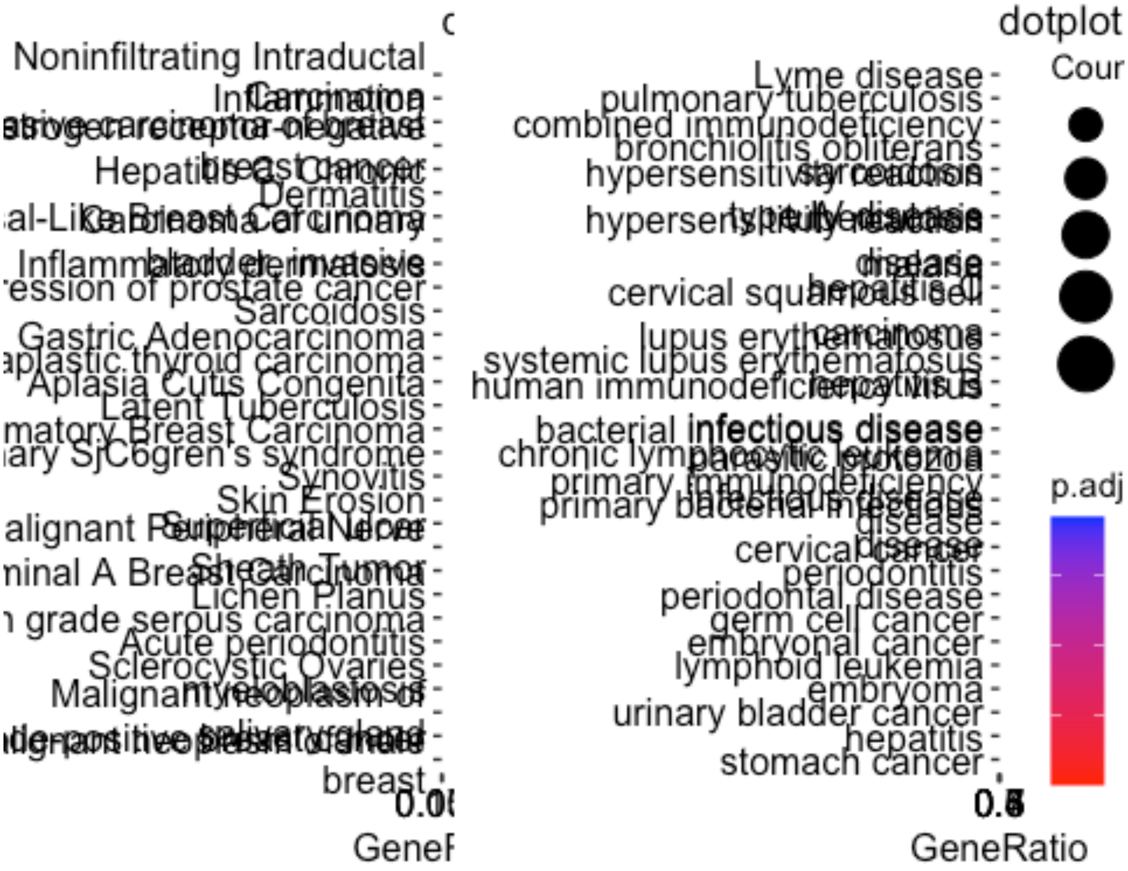
Then, we try to visualize the example data in the format of barplot, dotplot and gene_network

```
barplot <- grnaeR::show_barplot(edo,showCategory_num = 20)
barplot
```



```
dotplot<-grnaeR::show_dotplot(edo,showCategory_num=30)
```

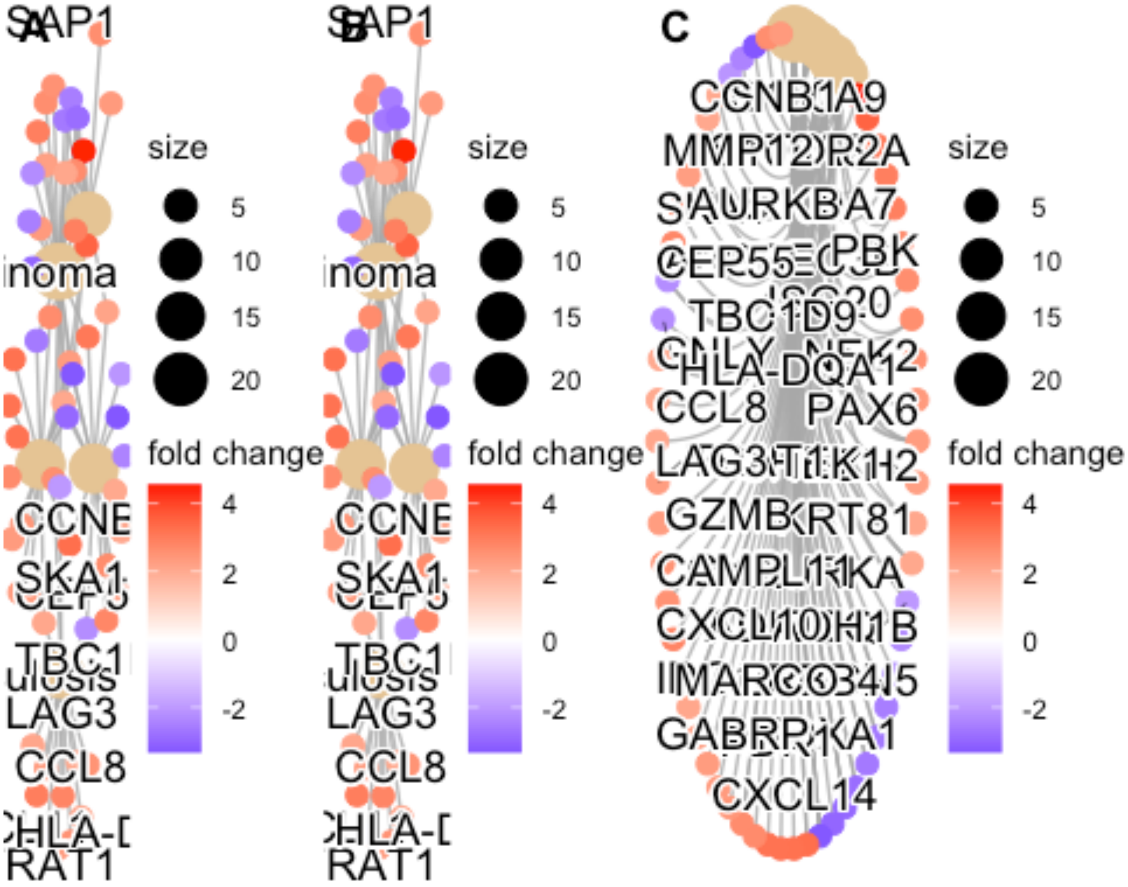
```
## preparing geneSet collections...

## GSEA analysis...

## Warning in fgseaMultilevel(pathways = pathways, stats = stats, minSize =
## minSize, : For some pathways, in reality P-values are less than 1e-10. You
can
## set the `eps` argument to zero for better estimation.

## leading edge analysis...

## done...

dotplot
```



```
gene_network <- grnaeR::develop_Gene_Network(edo,geneList)

## Scale for size is already present.
## Adding another scale for size, which will replace the existing scale.
## Scale for size is already present.
## Adding another scale for size, which will replace the existing scale.
## Scale for size is already present.
## Adding another scale for size, which will replace the existing scale.

gene_network
```

```
## Warning: ggrepel: 50 unlabeled data points (too many overlaps). Consider i
ncreasing max.overlaps
## ggrepel: 50 unlabeled data points (too many overlaps). Consider increasing
max.overlaps
```

```
## Warning: ggrepel: 24 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



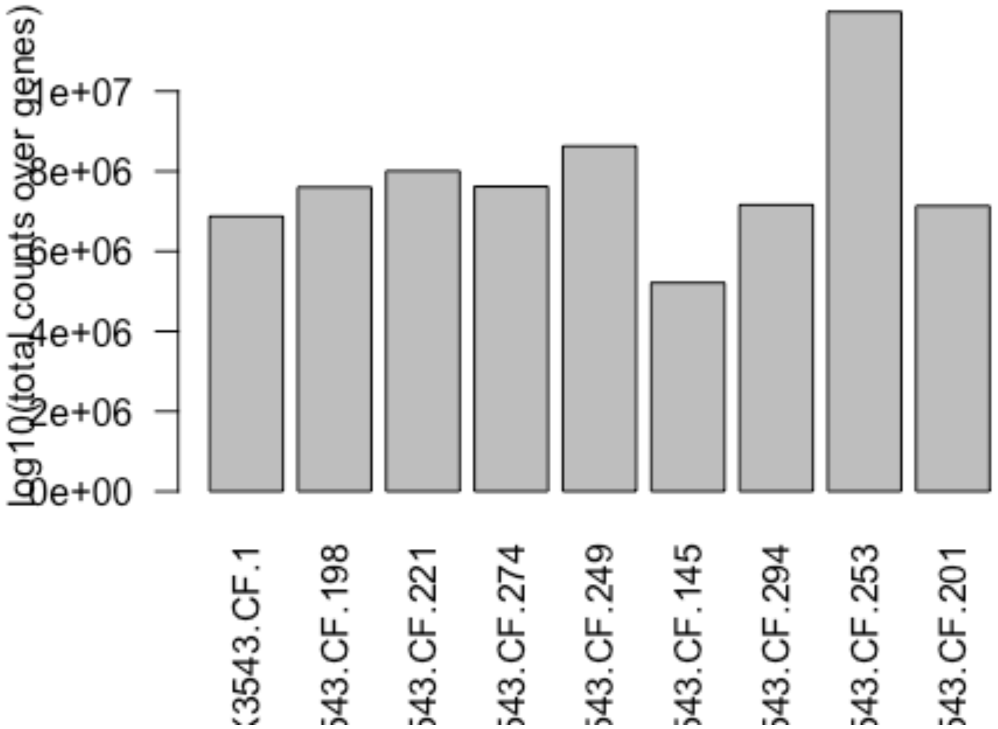
STEP2: Happy PATH Find_DEG

load the example data CRS_34v0

```
# working directory
dir = getwd()
# file The path of rnaseq raw count
file = '/Users/jesi/Documents/CRS_34v0.txt'
readcount = grnaeR::load_data(dir,file)
```

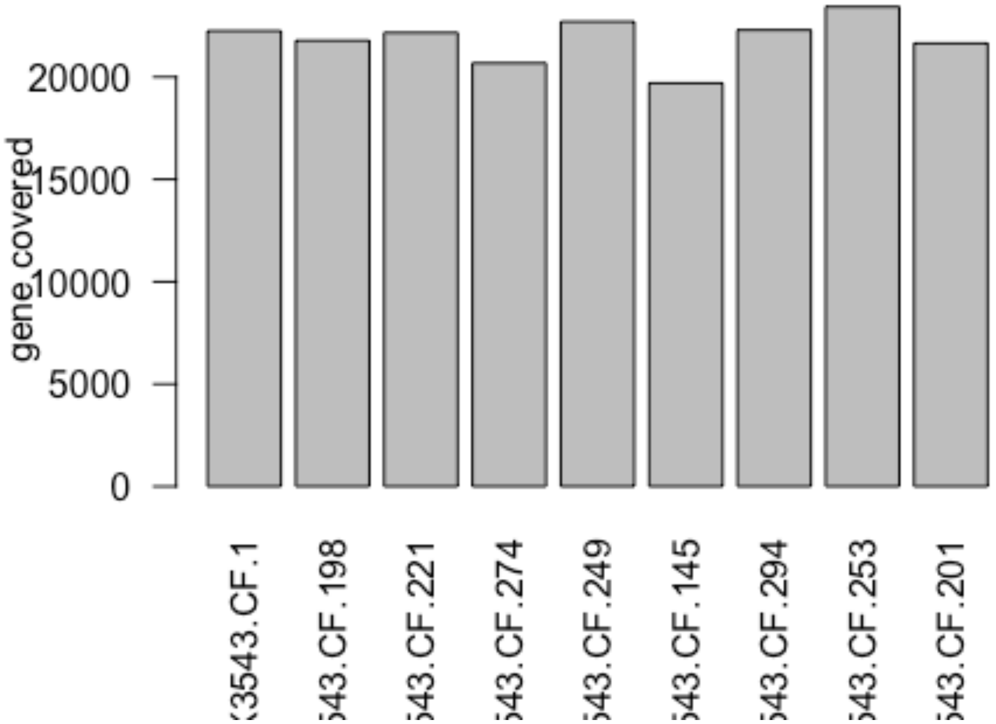
Perform quality control: check the total coverage

```
total.cov = grnaeR:: check_totalcov_quality(readcount)
```



quality control: check the number of genes being covered

```
gene.cov = grnaeR:: check_genecovered_quality(readcount)
```



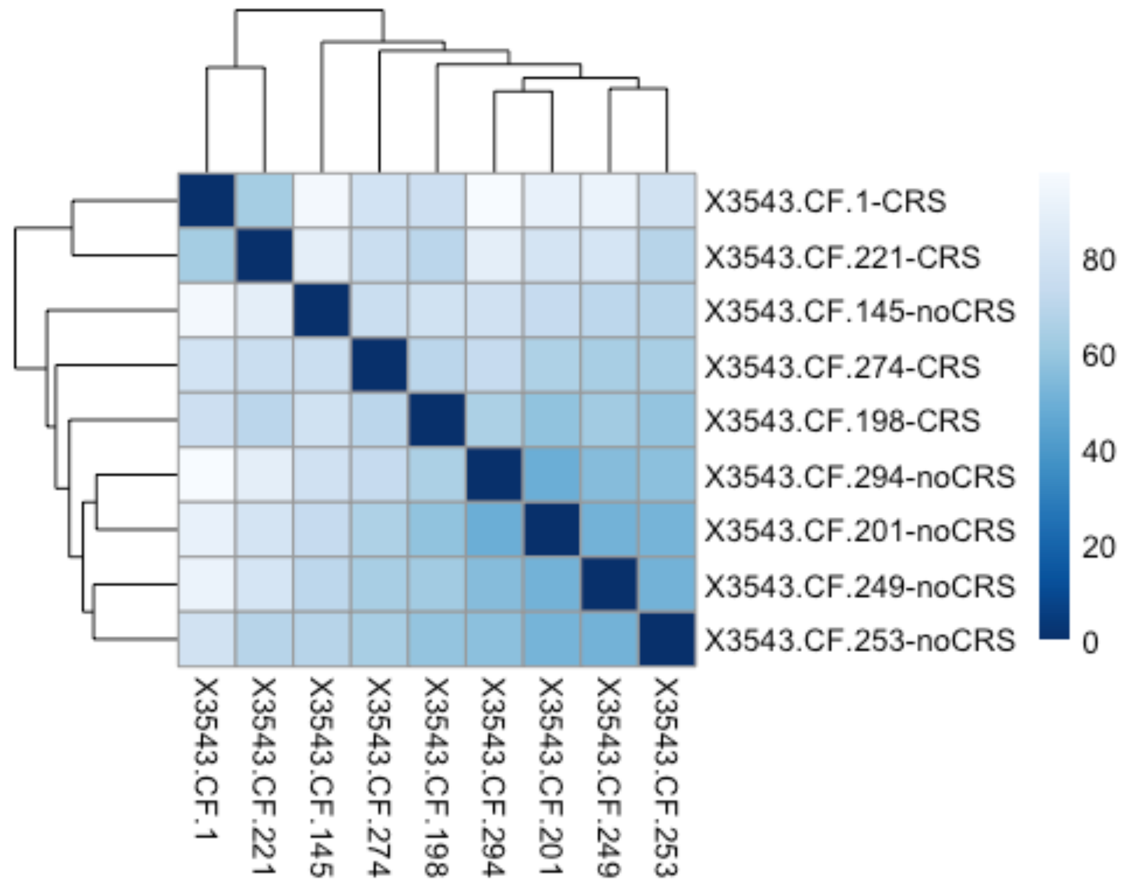
```
grnaeR::calculate_RPKM(readcount)

## [1] "the exonlength should be contained in provided file to calculate RPKM
"
```

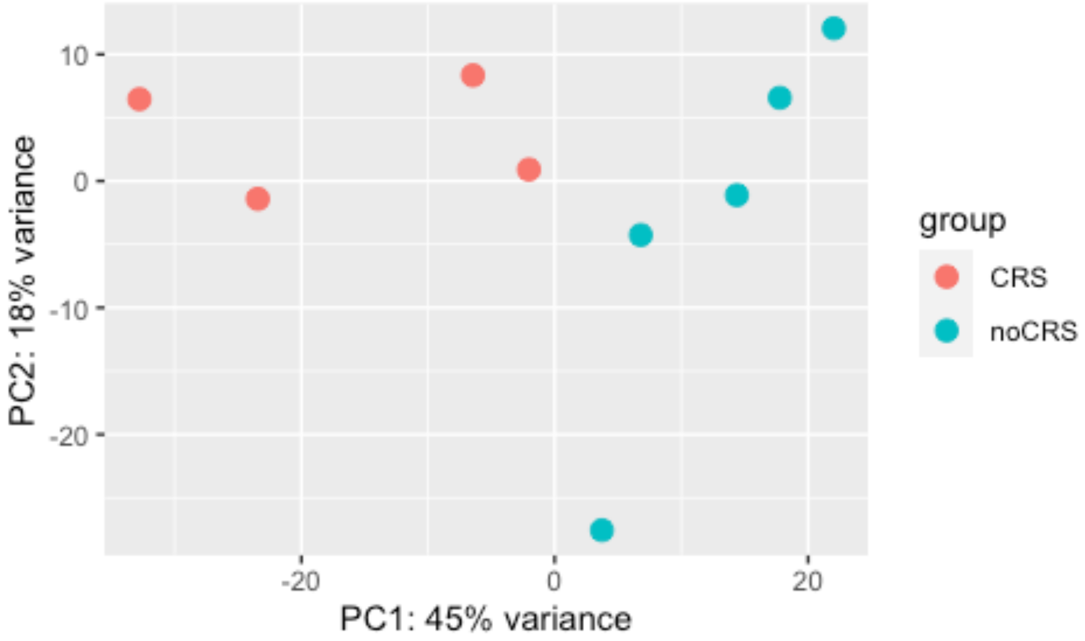
Next, we need to specifies configure and types used for the design of our data. Since the example data contains 4 patient samples develop CRS and 5 samples without CRS, we set the condition vector to describe the CRS condition, while the type_vector represents the individual sample.

```
type_vector = c(colnames(readcount))
condition_vector = c(c(rep('CRS',4)),c(rep('noCRS',5)))
```

load the data into DESeq2 object and normalized the dataset, further check the similarity between samples

```
dds = grnaeR::load_data_for_DESeq2(file,condition_vector,type_vector)
normalized_dds = normalize_dataset(dds)
check_sample_distance(normalized_dds)
```


X3543.CF.1-CRS
X3543.CF.221-CRS
X3543.CF.145-noCRS
X3543.CF.274-CRS
X3543.CF.198-CRS
X3543.CF.294-noCRS
X3543.CF.201-noCRS
X3543.CF.249-noCRS
X3543.CF.253-noCRS
X3543.CF.1
X3543.CF.221
X3543.CF.145
X3543.CF.274
X3543.CF.198
X3543.CF.294
X3543.CF.201
X3543.CF.249
X3543.CF.253
80
60
40
20
0



obtain the dataframe of the differentially expressed genes

```
select_DEGs = select_DEG(dds = dds,filter_thresh = 0,log2_fc = log(1.5,2), pa
djust = 0.05)

## [1] "filtering 2455 genes with low counts"

## using pre-existing size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

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

select_DEGs

##                         name    baseMean log2FoldChange     lfcSE      stat
## ENSG00000029534         ANK1  869.220107      -2.498666 0.4792864 -3.992819
## ENSG00000070182      SPTB  541.114023      -2.437783 0.4640356 -3.992842
## ENSG00000073464     CLCN4   66.277236      -2.345809 0.4164819 -4.227907
## ENSG00000086506      HBQ1   24.800945      -2.739773 0.5324346 -4.047090
## ENSG00000103942    HOMER2  259.512260      -2.815284 0.5103725 -4.369987
## ENSG00000117400       MPL   59.966231      -2.600807 0.3603960 -5.593415
## ENSG00000117560     FASLG   73.148087       1.986889 0.3560559  3.937376
## ENSG00000119508     NR4A3   86.218139      -4.667013 0.6940295 -5.881667
## ENSG00000123689      G0S2  320.776174      -5.363862 1.2263622 -3.896809
## ENSG00000124491     F13A1 8458.774527      -2.213657 0.4115042 -3.957906
## ENSG00000132613     MTSS2   60.875567      -2.666934 0.5102822 -4.080040
## ENSG00000133069     TMCC2  215.788863      -2.528897 0.4812075 -4.039700
## ENSG00000136842     TMOD1   49.874397      -3.165629 0.6248289 -4.130197
## ENSG00000140416      TPM1 1059.146658      -2.565856 0.4267550 -4.641757
## ENSG00000143995     MEIS1  194.884120      -2.144377 0.3947467 -3.950418
## ENSG00000145335      SNCA  826.232985      -2.508172 0.3910070 -4.918605
## ENSG00000156206   CFAP161   28.385474      -3.076096 0.5764862 -4.321237
## ENSG00000161513      FDXR  115.164295      -3.412218 0.6015795 -4.699720
## ENSG00000163736      PPBP 7795.616104      -2.457837 0.4658269 -4.020537
## ENSG00000184319  RPL23AP7   88.630827      -2.041034 0.2770360 -5.255893
## ENSG00000189060      H1-0  305.427716      -1.538643 0.2117097 -4.504660
## ENSG00000196565      HBG2  645.719741      -4.340141 0.9326037 -4.026554
## ENSG00000205639    MFSD2B   69.528876      -2.570594 0.4660148 -4.260876
## ENSG00000211829      <NA>  107.899418       2.842184 0.5381599  4.194333
## ENSG00000213931      HBE1   13.211798      -6.881100 1.3379859 -4.705683
## ENSG00000214076   CPSF1P1    7.591993      -6.680109 1.4626757 -4.167121
## ENSG00000223855  PDGFA-DT  121.656606      -2.612231 0.5195349 -3.902083
## ENSG00000228463 RPL23AP21   87.191066      -2.189608 0.3640058 -4.408295
## ENSG00000236397      <NA>   60.751566      -2.779214 0.5436198 -4.036371
## ENSG00000237541  HLA-DQA1   20.411057      -8.103383 1.0601802 -7.091643
## ENSG00000240356      <NA>  163.938451      -3.011172 0.5400743 -4.492363
## ENSG00000240583      AQP1   29.070822      -3.588255 0.7578724 -3.962794
## ENSG00000274602   PI4KAP1   96.551222      -2.024702 0.3038521 -4.738288
## ENSG00000276107      <NA>   79.844163      -5.714999 1.2655401 -4.053634
##                       pvalue         padj
## ENSG00000029534 6.529236e-05 3.999856e-02
## ENSG00000070182 6.528607e-05 3.999856e-02
## ENSG00000073464 2.358756e-05 2.528734e-02
## ENSG00000086506 5.185840e-05 3.830762e-02
## ENSG00000103942 1.242539e-05 1.639482e-02
## ENSG00000117400 2.226462e-08 1.273017e-04
## ENSG00000117560 8.237749e-05 4.415691e-02
## ENSG00000119508 4.061547e-09 3.483386e-05
## ENSG00000123689 9.746843e-05 4.917282e-02
## ENSG00000124491 7.560983e-05 4.316744e-02
## ENSG00000132613 4.502790e-05 3.830762e-02
## ENSG00000133069 5.351959e-05 3.830762e-02
## ENSG00000136842 3.624528e-05 3.272186e-02
## ENSG00000140416 3.454587e-06 6.584059e-03
## ENSG00000143995 7.801496e-05 4.316744e-02
## ENSG00000145335 8.716306e-07 2.990216e-03
## ENSG00000156206 1.551569e-05 1.901004e-02
## ENSG00000161513 2.605183e-06 5.585838e-03
## ENSG00000163736 5.806554e-05 3.830762e-02
## ENSG00000184319 1.473079e-07 6.316930e-04
## ENSG00000189060 6.647923e-06 1.098370e-02
## ENSG00000196565 5.660033e-05 3.830762e-02
## ENSG00000205639 2.036270e-05 2.328542e-02
## ENSG00000211829 2.736752e-05 2.761383e-02
## ENSG00000213931 2.530173e-06 5.585838e-03
## ENSG00000214076 3.084706e-05 2.939554e-02
## ENSG00000223855 9.536837e-05 4.917282e-02
## ENSG00000228463 1.041874e-05 1.489273e-02
## ENSG00000236397 5.428442e-05 3.830762e-02
## ENSG00000237541 1.325288e-12 2.273266e-08
## ENSG00000240356 7.043710e-06 1.098370e-02
## ENSG00000240583 7.407757e-05 4.316744e-02
## ENSG00000274602 2.155309e-06 5.585838e-03
## ENSG00000276107 5.042806e-05 3.830762e-02
```