cancer-specificity-analysis

2023-04-19

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.5
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Some function Reuqires the in-house package "protools"
if (!require("protools", quietly = TRUE))
   devtools::install_github("https://github.com/FDUguchunhui/protools")
library('protools')
```

import data

nasf_primary_cell_2.csv: all primary cell data filtered with at least 2 spectral count and then normalized to NSAF cell-line-raw-data/nasf_tce_2.csv: all cell line data filtered with at least 2 spectral count and then normalized to NSAF IPAS_annotation.csv: contain disease and subtype information for above samples

```
# misspelling in the original file name
primary_cell_nsaf <- read.csv('primary-cell/nasf_primary_cell_2.csv', row.names = 1)
primary_cell_nsaf <- as.matrix(primary_cell_nsaf)
tce_nsaf <- read.csv('cell-line-raw-data/nasf_tce_2.csv', row.names = 1)
tce_nsaf <- as.matrix(tce_nsaf)
# read annotation file
IPAS_annotation <- read_csv('support-data/IPAS_annotation.csv')
```

```
## Rows: 545 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): ipas, disease, subtype
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

create a helper function for creating annotation file for the expression matrix

```
create_annoataion <- function(data, annotation_table) {
  annotation_tbl <- annotation_table %>% select(ipas, disease, subtype) %>% filter(ipas %in% colnames(da
  annotation_df <- data.frame(disease=annotation_tbl$disease, subtype=annotation_tbl$subtype)
  rownames(annotation_df) <- annotation_tbl$ipas
  return(annotation_df)
}

primary_cell_annotation <- create_annoataion(primary_cell_nsaf, annotation_table = IPAS_annotation)
primary_cell_annotation$type <- paste(primary_cell_annotation$disease, primary_cell_annotation$subtype)
primary_cell_annotation <- primary_cell_annotation['type']
colnames(primary_cell_annotation) <- 'disease'
table(primary_cell_annotation$disease)
```

```
##
##  Gastric Ascites     Leukemia ALL     Leukemia AML Leukemia Control
##               35               60               77                2
##     Leukemia MDS  Ovarian Ascites
##               37               33
```

create annotation dataframe for cell-line TCE data

```
tce_annotation <- create_annoataion(tce_nsaf, annotation_table = IPAS_annotation)
tce_annotation$type <- paste(tce_annotation$disease, tce_annotation$subtype)
tce_annotation <- tce_annotation[, c('disease', 'type')]
colnames(tce_annotation) <- c('cancer', 'subtype')
# check distribution of each cancer and subtype
table(tce_annotation$cancer)
```

```
##
##        Breast         Colon       Gastric        Glioma      Leukemia
##            39            10             9            10            14
##     LungAdeno   Lymphoblast      Melanoma Neuroblastoma       Ovarian
##            60             2             5             2            14
##    Pancreatic      Prostate          SCLC
##            18             5            27
```

```
table(tce_annotation$subtype)
```

```
##
##             Breast Basal              Breast HER2        Breast LuminalA/B
##                        1                         6                         6
##              Breast TNBC              Colon CMS1              Colon CMS3
##                       26                         3                         1
##              Colon CMS4           Gastric Adeno          Gastric Ascites
##                        6                         7                         2
```

```
##      Glioma Glioblastoma       Glioma Mesenchymal                Glioma NA
##                        4                        4                        2
##             Leukemia AML            Leukemia CML     LungAdeno Epithelial
##                       13                        1                       15
##    LungAdeno Mesenchymal            LungAdeno NA          Lymphoblast NA
##                        9                       36                        2
##      Melanoma Metastatic             Melanoma NA         Neuroblastoma NA
##                        1                        4                        2
##            Ovarian Adeno        Ovarian Carcinoma             Ovarian Xeno
##                       11                        1                        2
## Pancreatic ExocrineAdeno           Pancreatic NA          Pancreatic PDAC
##                       14                        3                        1
##           Prostate Adeno       Prostate Carcinoma  SCLC ASCL1-SLFN11(high)
##                        2                        3                        4
##   SCLC ASCL1-SLFN11(low)                  SCLC NA       SCLC Neurendocrine
##                        1                        2                       10
##             SCLC NEUROD1               SCLC NonNE              SCLC POU2F3
##                        3                        2                        2
##           SCLC VariantNE
##                        3
```

import missing protein

```
missing_proteins <- readxl::read_xlsx('Supplementary file 2 identified missing protein details.xlsx', sh
unique_67_accession  <- missing_proteins %>% filter(TPM > 0) %>% select(accession) %>% pull() %>% unique
```

create a varaible to denote cancer-testis for the identified MPs

```
cancer_testis_gene <- c('WFDC11','SPATA21','C19orf67','CCNYL2','TRAV9-2','TRAV6','TMEM105','NCBP2L','TM
missing_proteins_annotation <- missing_proteins %>% filter(TPM > 0) %>%  select(accession, gene_symbol)
missing_proteins_annotation <- missing_proteins_annotation %>% mutate(cancer_testis=if_else(gene_symbol
missing_proteins_annotation <- missing_proteins_annotation %>% arrange(cancer_testis)
```

extract the IPAS to keep

For primary cell, only use following cancers: “Leukemia ALL”, “Leukemia AML”, “Leukemia MDS”, “Ovarian Ascites”, “Gastric Ascites”

For cell-line TCE, only use the following cancers/subtypes “Breast HER2”, “Breast LuminalA/B”, “Breast TNBC”, “Gastric Ascites”, “Gastric Adeno”, “Leukemia AML”, “LungAdeno Mesenchymal”, “LungAdeno Epithelial”, “LungAdeno NA”, “Ovarian Adeno”, “Pancreatic ExocrineAdeno”, “SCLC Neurendocrine”

The reason for not using all cancer/subtype is because some subtypes only have few samples, and including them will make heatmap plot color hard to read

```
extract_sample_id <- function(x, disease_col, disease) {
  rownames(x$annotation[x$annotation[[disease_col]] %in% disease, ,drop=FALSE])
```

```
}

# diseases_keep <- c('Colon', 'Glioma', 'Leukemia', 'Lymphoblast', 'Melanoma', 'Neuroblastoma', 'Prosta
diseases_keep <- c("Leukemia ALL", "Leukemia AML", "Leukemia MDS", "Ovarian Ascites", "Gastric Ascites"
SpC_list_primary_cell <- SpC_List(primary_cell_nsaf, annotation = primary_cell_annotation, proteins_fil
```

```
## Number of rows before filtering: 50743
```

```
## Number of rows after filtering: 67
```

```
##
```

```
replicates_for_keep <- extract_sample_id(SpC_list_primary_cell, disease_col = 'disease', disease = disea
```

```
# diseases_keep <- c('Colon', 'Glioma', 'Leukemia', 'Lymphoblast', 'Melanoma', 'Neuroblastoma', 'Prosta
diseases_keep_tce <- c("Breast HER2", "Breast LuminalA/B", "Breast TNBC", "Gastric Ascites", "Gastric A
SpC_list_tce <- SpC_List(tce_nsaf, annotation = tce_annotation, proteins_filter = unique_67_accession)
```

```
## Number of rows before filtering: 55221
```

```
## Number of rows after filtering: 61
```

```
##
```

```
replicates_for_keep_tce <- extract_sample_id(SpC_list_tce, disease_col = 'subtype', disease = diseases_
```

extract the IPAS to keep for each cancer/subtype

This information is later used to reorder samples in matrix for better plotting heatmap

```
IPAS <- mapply(FUN=extract_sample_id, disease=c("Leukemia ALL", "Leukemia AML", "Leukemia MDS", "Ovaria
# SpC_lists <- mapply(FUN=SpC_List, replicates_keep=IPAS, MoreArgs = list(df=primary_cell_nasf_gene_sym
IPAS_tce <- mapply(FUN=extract_sample_id, disease= c("Breast HER2", "Breast LuminalA/B", "Breast TNBC",
```

reorder samples to make samples of the same cancer/subtype next to each other

```
primary_cell_nasf <- primary_cell_nsaf[, unlist(IPAS)]
tce_nsaf <- tce_nsaf[, unlist(IPAS_tce)]
```

Create the final data object used for both primaru cell and cell-line TCE

SpC_List is a convenient customized object for simplifying analysis pipeline See ?SpC_List for more details

```
all_cancer_primary_cell <- SpC_List(df=primary_cell_nasf, annotation=primary_cell_annotation, proteins_
```

```
## Number of rows before filtering: 50743
```

```
## Number of rows after filtering: 67
```

```
##
```

```
all_cancer_tce <- SpC_List(df=tce_nsaf, annotation=tce_annotation, proteins_filter = unique_67_accessio
```

```
## Number of rows before filtering: 55221
```

```
## Number of rows after filtering: 61
```

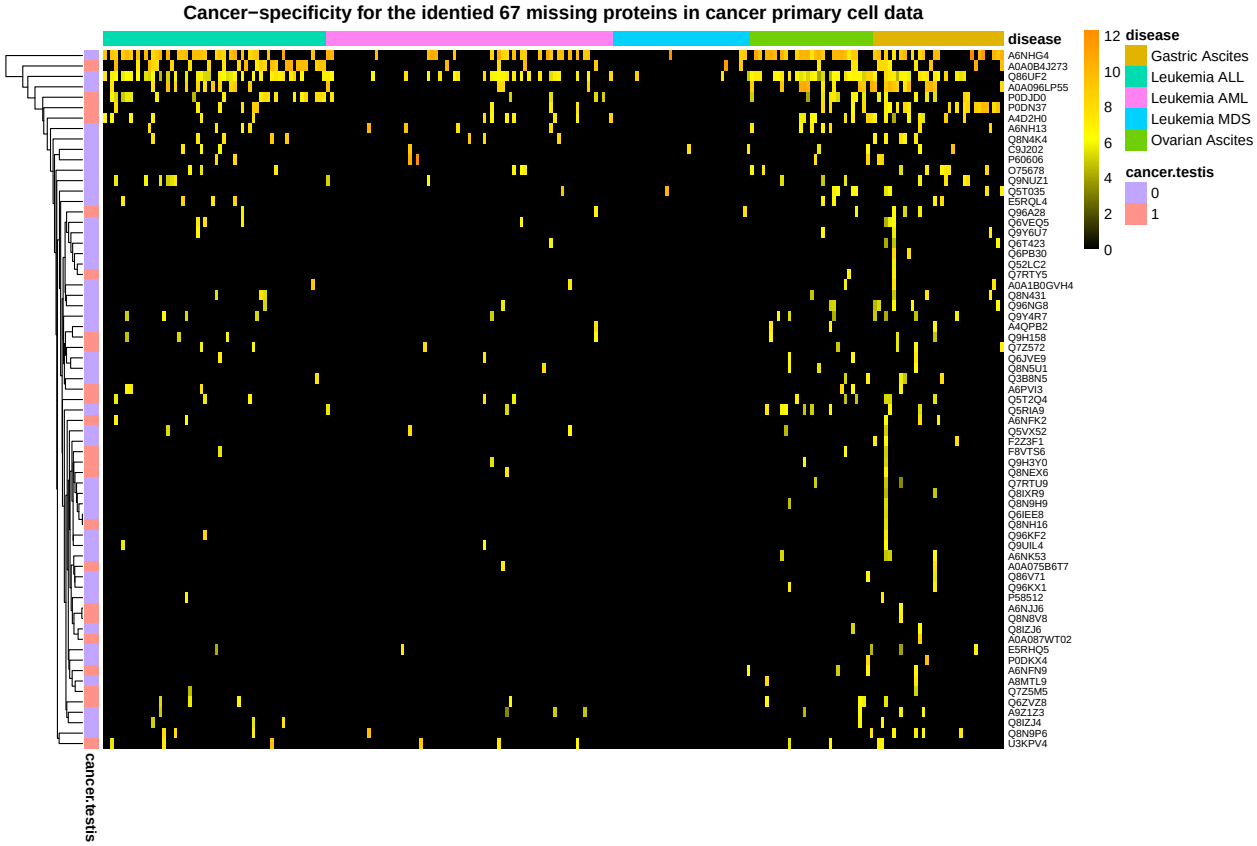
```
##
```

create cancer-testis row annotation for primary cell

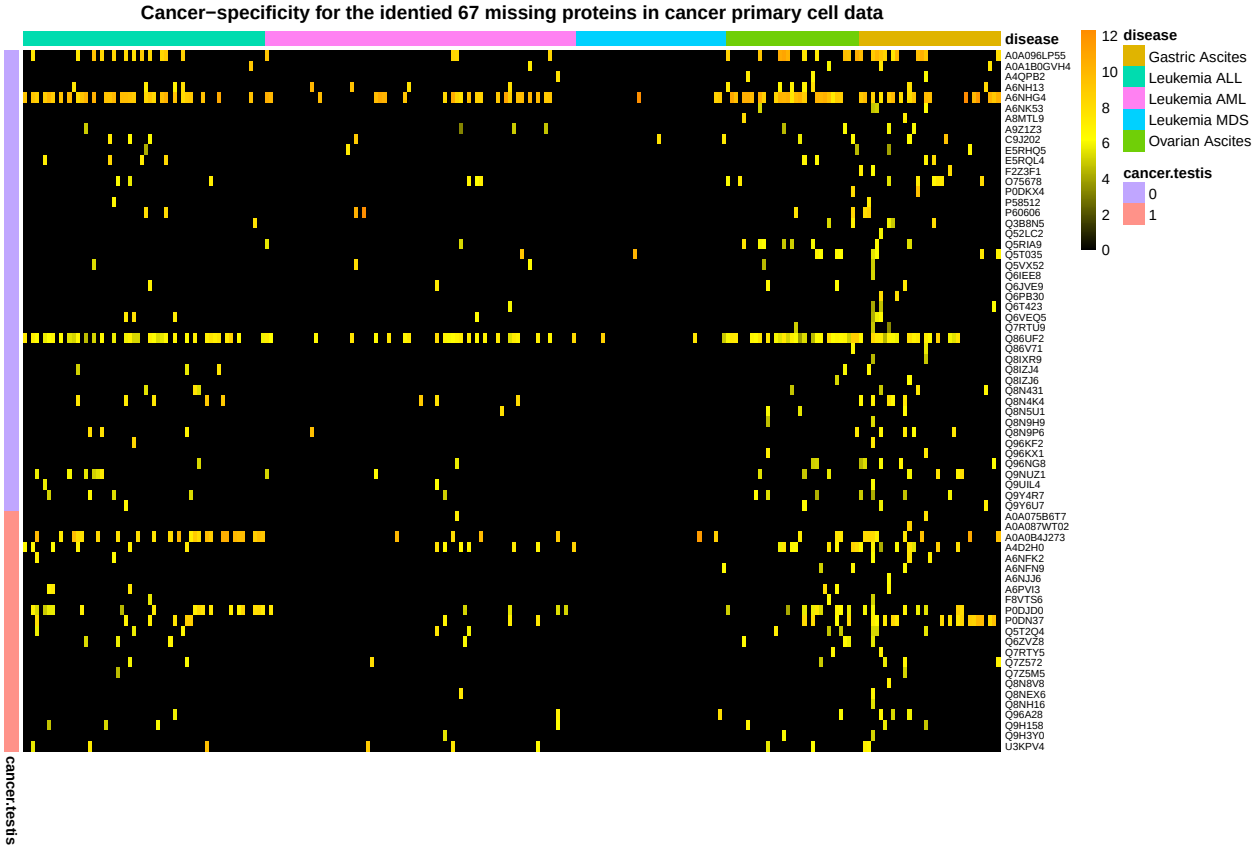
```
all_cancer_primary_cell$matrix <- all_cancer_primary_cell$matrix[missing_proteins_annotation$accession,

row_annotation <- missing_proteins_annotation[, c("accession", "cancer_testis")]
row_annotation <- data.frame(`cancer testis`=row_annotation$cancer_testis, row.names = row_annotation$a
row_annotation$cancer.testis <- as.factor(row_annotation$cancer.testis)
```

```
plot_heat_map(all_cancer_primary_cell,
         annotation_row = row_annotation,
         cluster_rows = T,
         cluster_cols = F,
         show_rownames = T,
         fontsize_row=7,
              main='Cancer-specificity for the identied 67 missing proteins in cancer primary cell data
         show_colnames = F,)
```




```
plot_heat_map(all_cancer_primary_cell,
          annotation_row = row_annotation,
          cluster_rows = F,
          cluster_cols = F,
          show_rownames = T,
          fontsize_row=7,
               main='Cancer-specificity for the identied 67 missing proteins in cancer primary cell data
          show_colnames = F,)
```

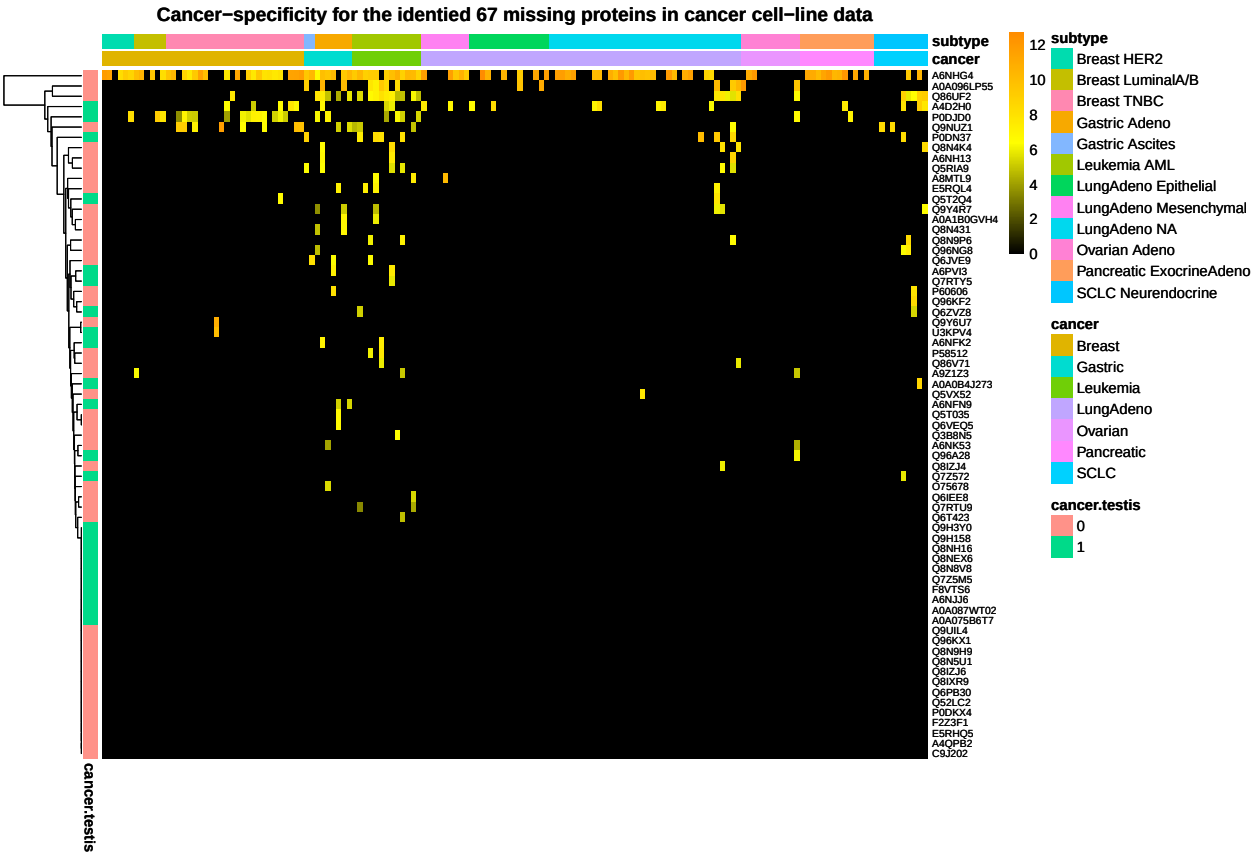


reorder to make caqncer-testis perotein to bottom part

```
# find those MPs that were empty in data
empty_MPs <- missing_proteins_annotation$accession[which(!(missing_proteins_annotation$accession %in% r
additional_matrix <- matrix(0, nrow = length(empty_MPs), ncol = ncol(all_cancer_tce$matrix))
rownames(additional_matrix) <- empty_MPs
all_cancer_tce$matrix <- rbind(all_cancer_tce$matrix, additional_matrix)
all_cancer_tce$matrix <- all_cancer_tce$matrix[missing_proteins_annotation$accession, ]
```

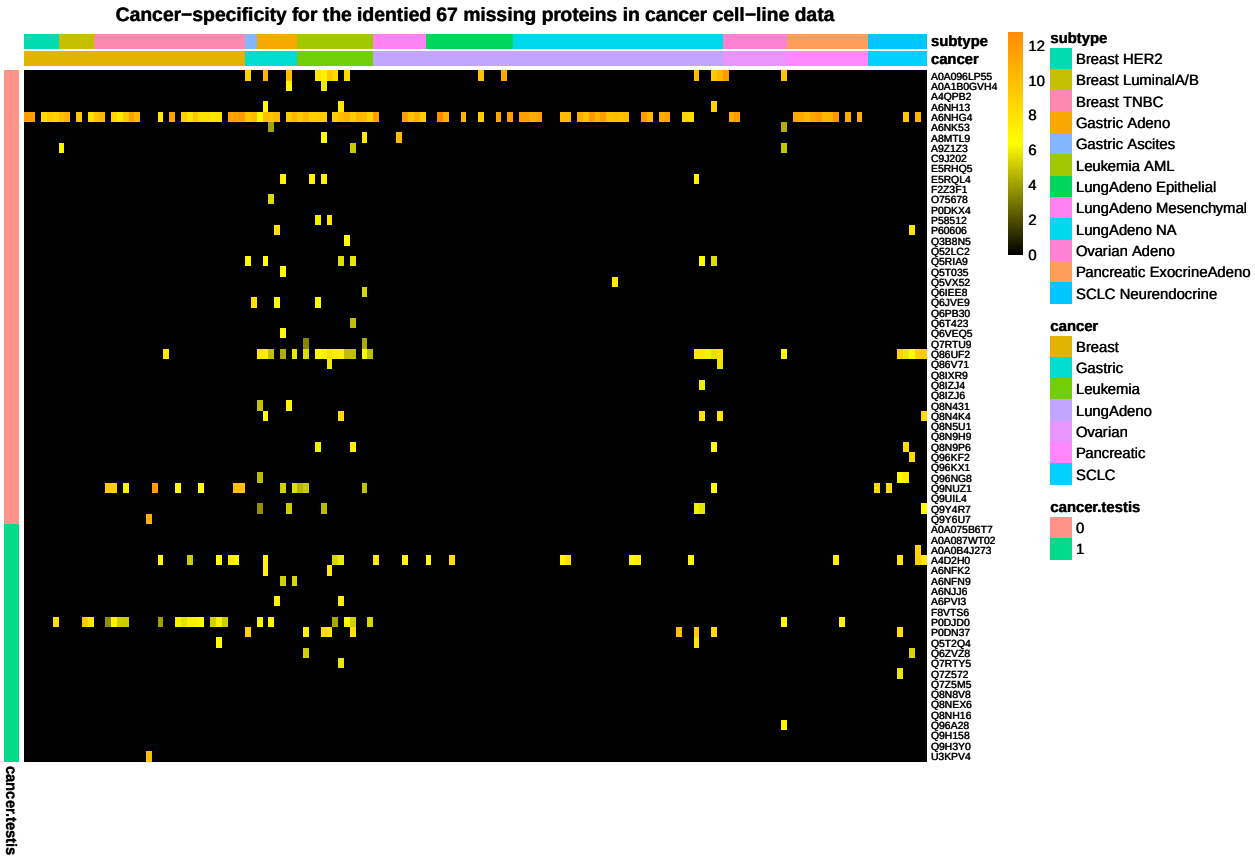
```
row_annotation <- missing_proteins_annotation[, c("accession", "cancer_testis")]
row_annotation <- data.frame(`cancer testis`=row_annotation$cancer_testis, row.names = row_annotation$a
row_annotation$cancer.testis <- as.factor(row_annotation$cancer.testis)
```

```
(cancer_specificity_cell_line_row_clustered <- plot_heat_map(all_cancer_tce,
         annotation_row = row_annotation,
         cluster_rows = T,
         cluster_cols = F,
         show_rownames = T,
         fontsize_row=7,
              main='Cancer-specificity for the identied 67 missing proteins in cancer cell-line data',
         show_colnames = F))
```

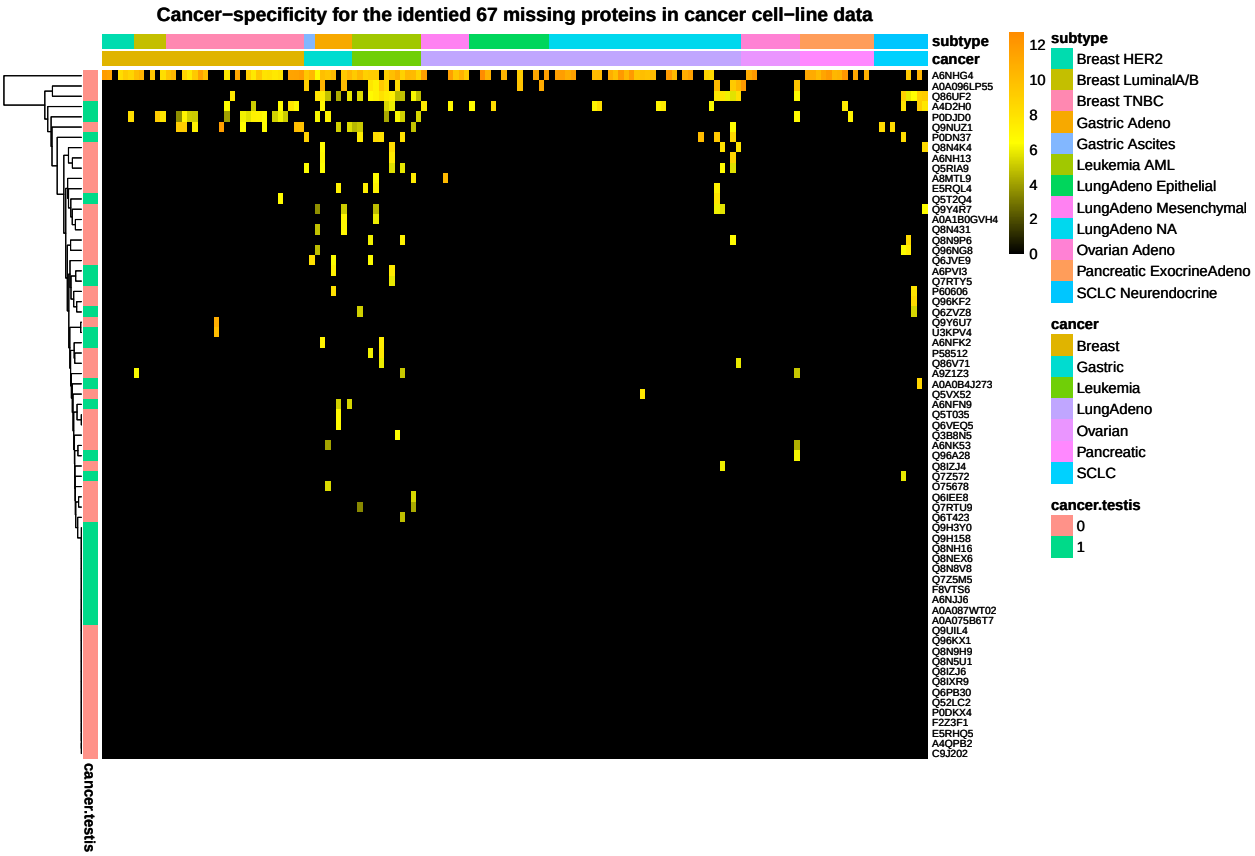



```
# ggsave('~/Downloads/cancer_specificity_primary_cell_row_clustered.png', cancer_specificity_primary_ce
#        width = 1020, height=735, units = 'px', dpi=80)
```

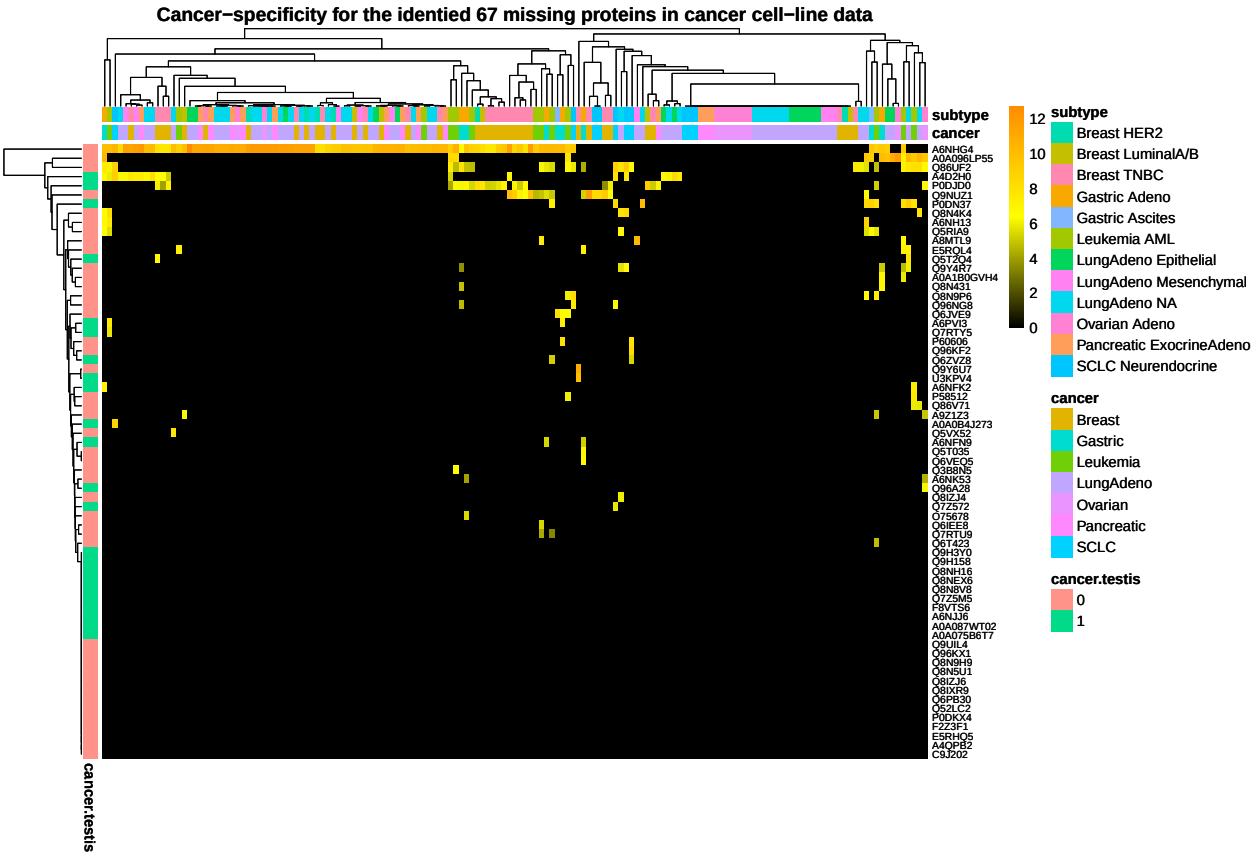
```
(cancer_specificity_cell_line <- plot_heat_map(all_cancer_tce,
         annotation_row = row_annotation,
         cluster_rows = F,
         cluster_cols = F,
         show_rownames = T,
         fontsize_row=7,
              main='Cancer-specificity for the identied 67 missing proteins in cancer cell-line data',
         show_colnames = F))
```




```
(cancer_specificity_cell_line_col_clustered <- plot_heat_map(all_cancer_tce,
         annotation_row = row_annotation,
         cluster_rows = T,
         cluster_cols = F,
         show_rownames = T,
         fontsize_row=7,
              main='Cancer-specificity for the identied 67 missing proteins in cancer cell-line data',
         show_colnames = F))
```




```
(cancer_specificity_cell_line_col_clustered <- plot_heat_map(all_cancer_tce,
         annotation_row = row_annotation,
         cluster_rows = T,
         cluster_cols = T,
         show_rownames = T,
         fontsize_row=7,
              main='Cancer-specificity for the identied 67 missing proteins in cancer cell-line data',
         show_colnames = F))
```


Cancer−specificity for the identied 67 missing proteins in cancer cell−line data
subtype
cancer
cancer.testis
A6NHG4
A0A096LP55
Q86UF2
A4D2H0
P0DJD0
Q9NUZ1
P0DN37
Q8N4K4
A6NH13
Q5RIA9
A8MTL9
E5RQL4
Q5T2Q4
Q9Y4R7
A0A1B0GVH4
Q8N431
Q8N9P6
Q96NG8
Q6JVE9
A6PVI3
Q7RTV5
P60606
Q96KF2
Q6ZVZ8
Q9Y6U7
U3KPV4
A6NFK2
P58512
Q86V71
A9Z1Z3
A0A0B4J273
Q5VX52
A6NFN9
Q5T035
Q6VEQ5
Q3B8N5
A6NK53
Q96A28
Q8IZJ4
Q7Z572
O75678
Q6IEE8
Q7RTU9
Q6T423
Q9H3Y0
Q9H158
Q8NH16
Q8NEX6
Q6N8V8
Q7Z5M5
F8VTS6
A6NJJ6
A0A087WT02
A0A075B6T7
Q9UIL4
Q96KX1
Q8N9H9
Q8N5U1
Q8IZJ6
Q8IXR9
Q6PB30
Q52LC2
P0DKX4
F2Z3F1
E5RHQ5
A4QPB2
C9J202
12
10
8
6
4
2
0
subtype
Breast HER2
Breast LuminalA/B
Breast TNBC
Gastric Adeno
Gastric Ascites
Leukemia AML
LungAdeno Epithelial
LungAdeno Mesenchymal
LungAdeno NA
Ovarian Adeno
Pancreatic ExocrineAdeno
SCLC Neurendocrine
cancer
Breast
Gastric
Leukemia
LungAdeno
Ovarian
Pancreatic
SCLC
cancer.testis
0
1