

# 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 Requires 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(d
  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
```

import missing protein

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

extract the IPAS to keep

```
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 = dise

# 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", "Ovarian
# SpC_lists <- mapply(FUN=SpC_List, replicates_keep=IPAS, MoreArgs = list(df=primary_cell_nsf_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_nsf <- primary_cell_nsaf[, unlist(IPAS)]
tce_nsaf <- tce_nsaf[, unlist(IPAS_tce)]

```

## Create the final data object used for both primary 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_accession
```

```
## 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,
```

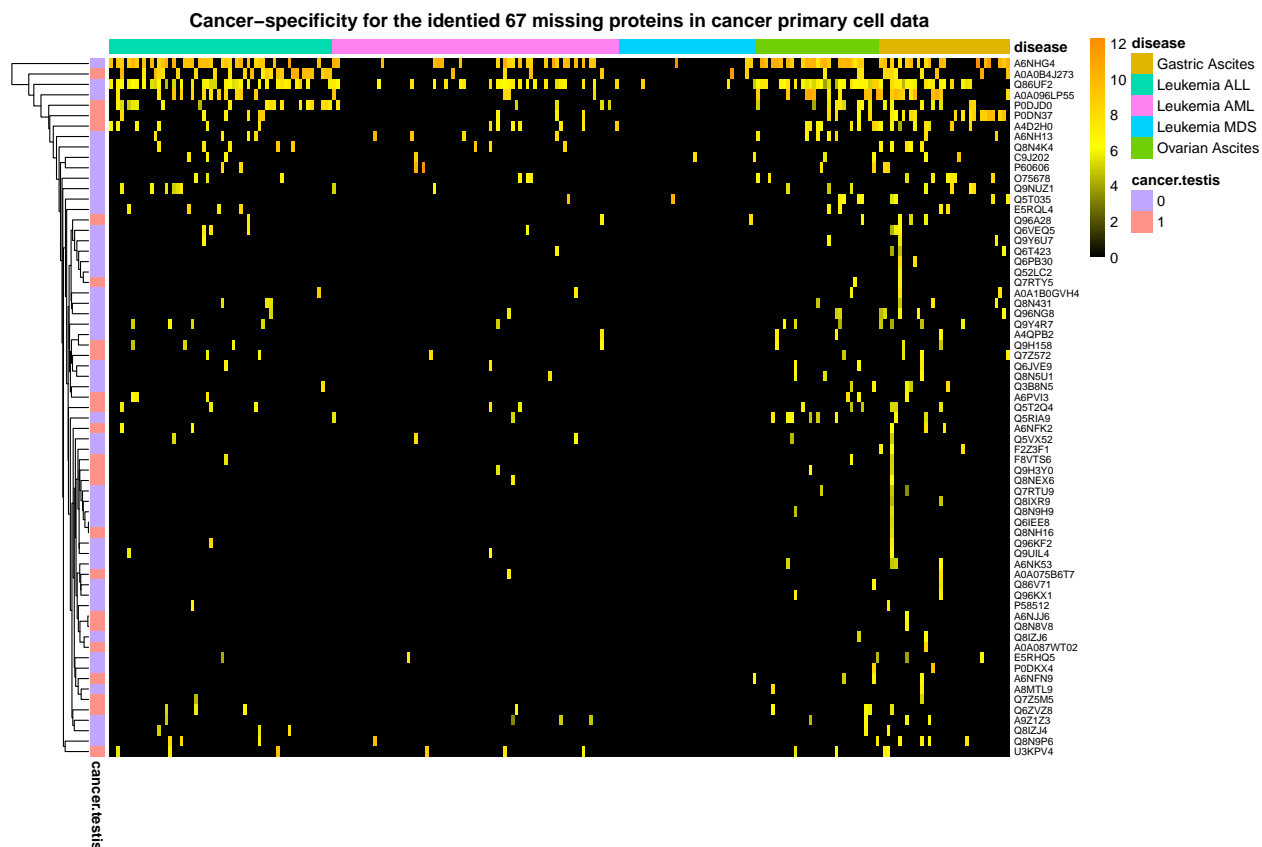
```
  cluster_cols = F,
```

```
  show_rownames = T,
```

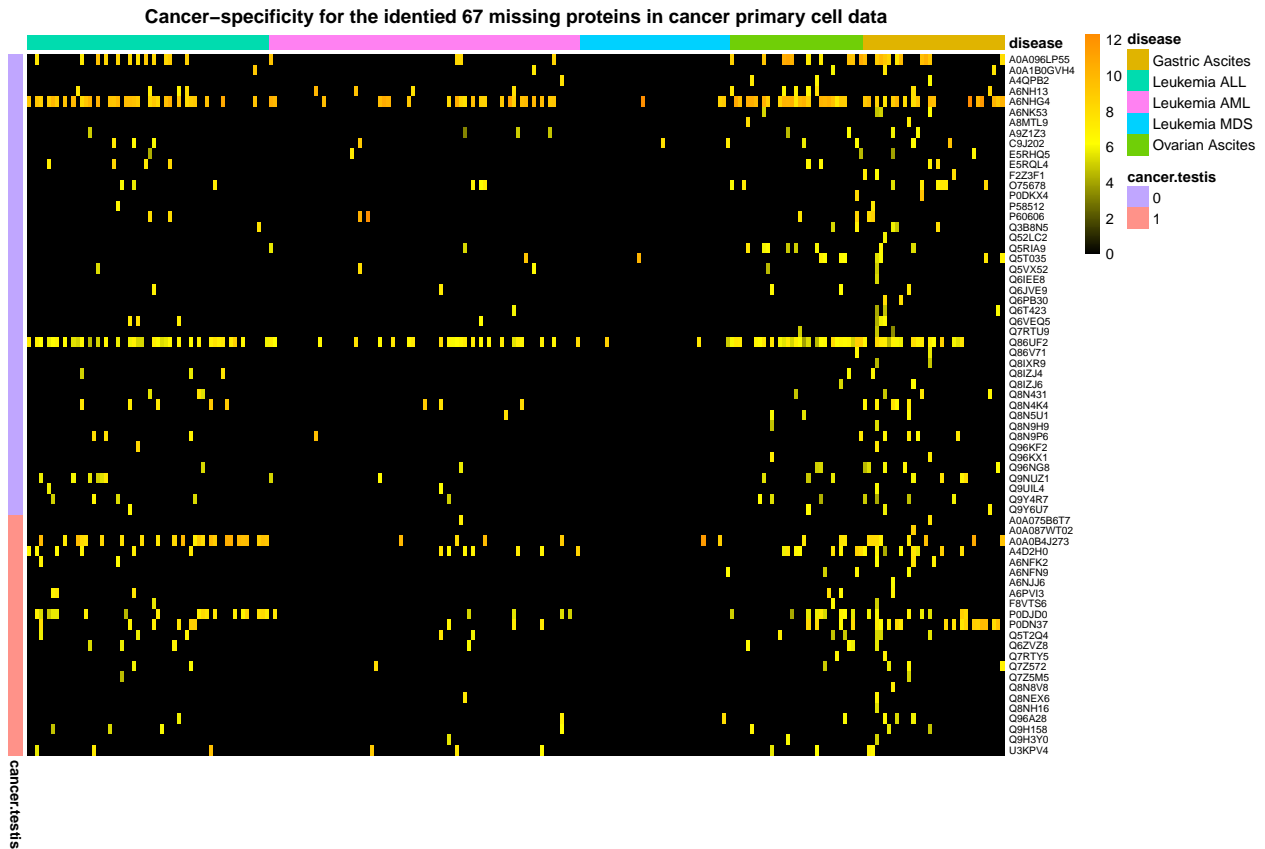
```
  fontsize_row=7,
```

```
  main='Cancer-specificity for the identified 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 identified 67 missing proteins in cancer primary cell data',
              show_colnames = F,)
```

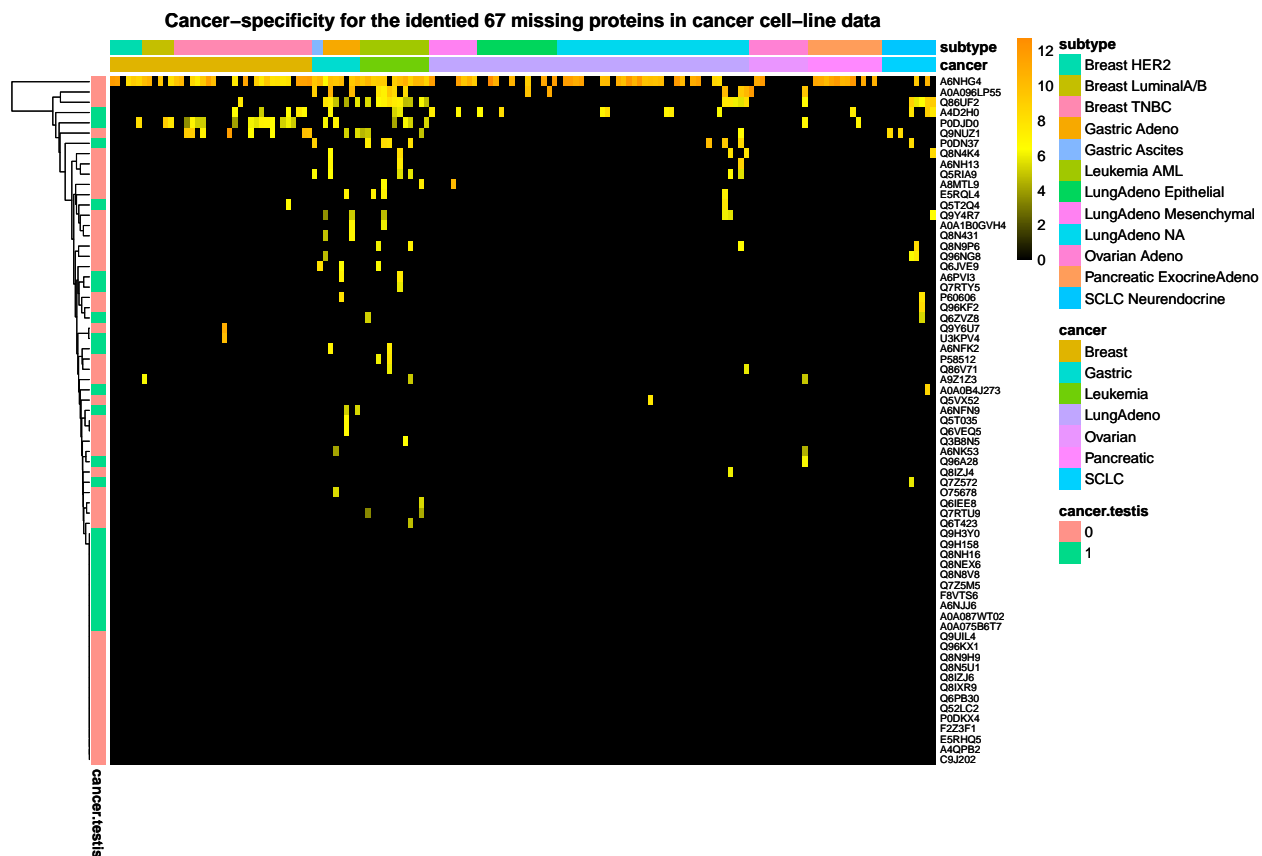


reorder to make caqncr-testis perotein to bottom part

```
# find those MPs that were empty in data
empty_MPs <- missing_proteins_annotation$accession[which(!(missing_proteins_annotation$accession %in% rownames(all_cancer_tce$matrix)))]
additional_matrix <- matrix(0, nrow = length(empty_MPs), ncol = ncol(all_cancer_tce$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$cancer_testis, row.names = row_annotation$accession)
row_annotation$cancer.testis <- as.factor(row_annotation$cancer.testis)
```

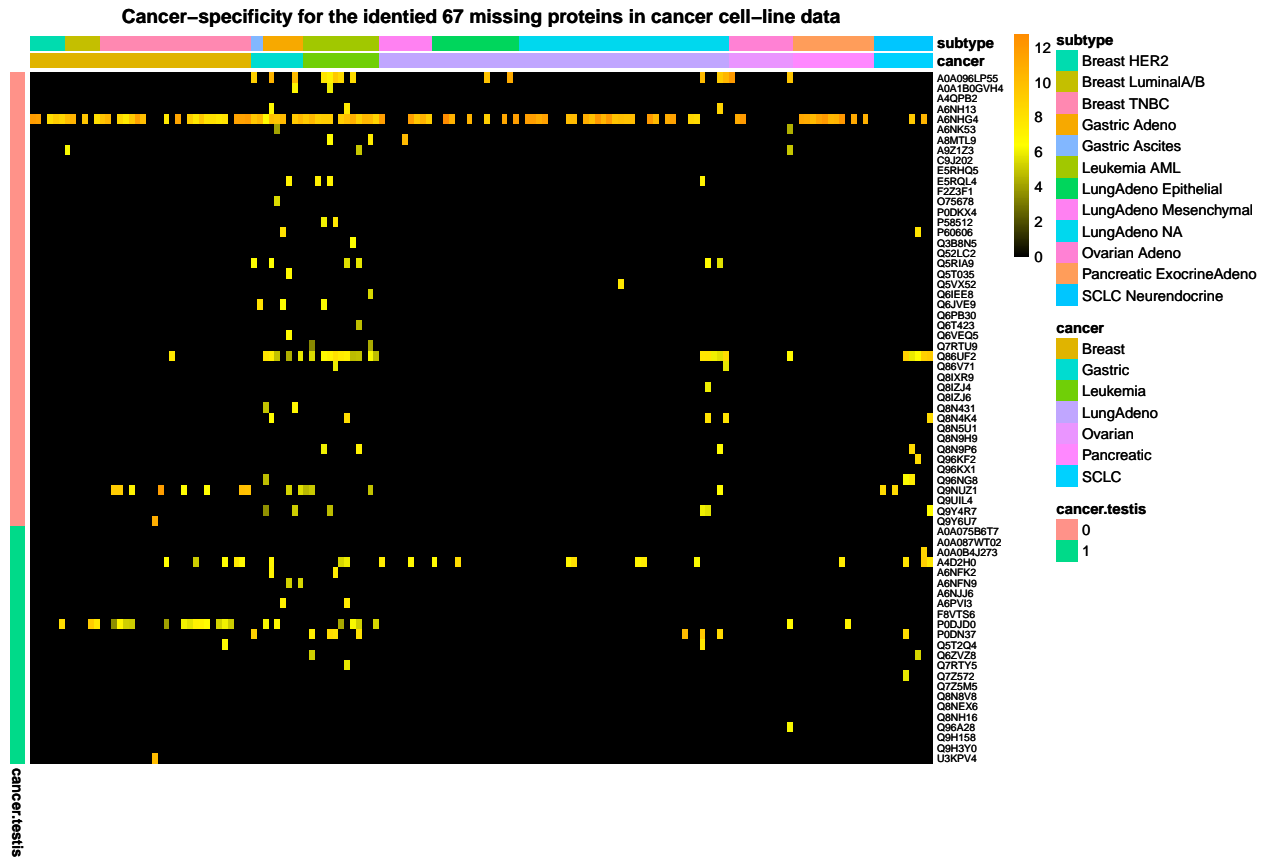
```
(cancer_specificity_cell_line_row_clustered <- plot_heat_map(all_cancer_tce,
  annotation_row = row_annotation,
  cluster_rows = T,
  cluster_cols = F,
  show_rownames = T,
  fontsize_row=7,
  main='Cancer-specificity for the identified 67 missing proteins in cancer cell-line data',
  show_colnames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



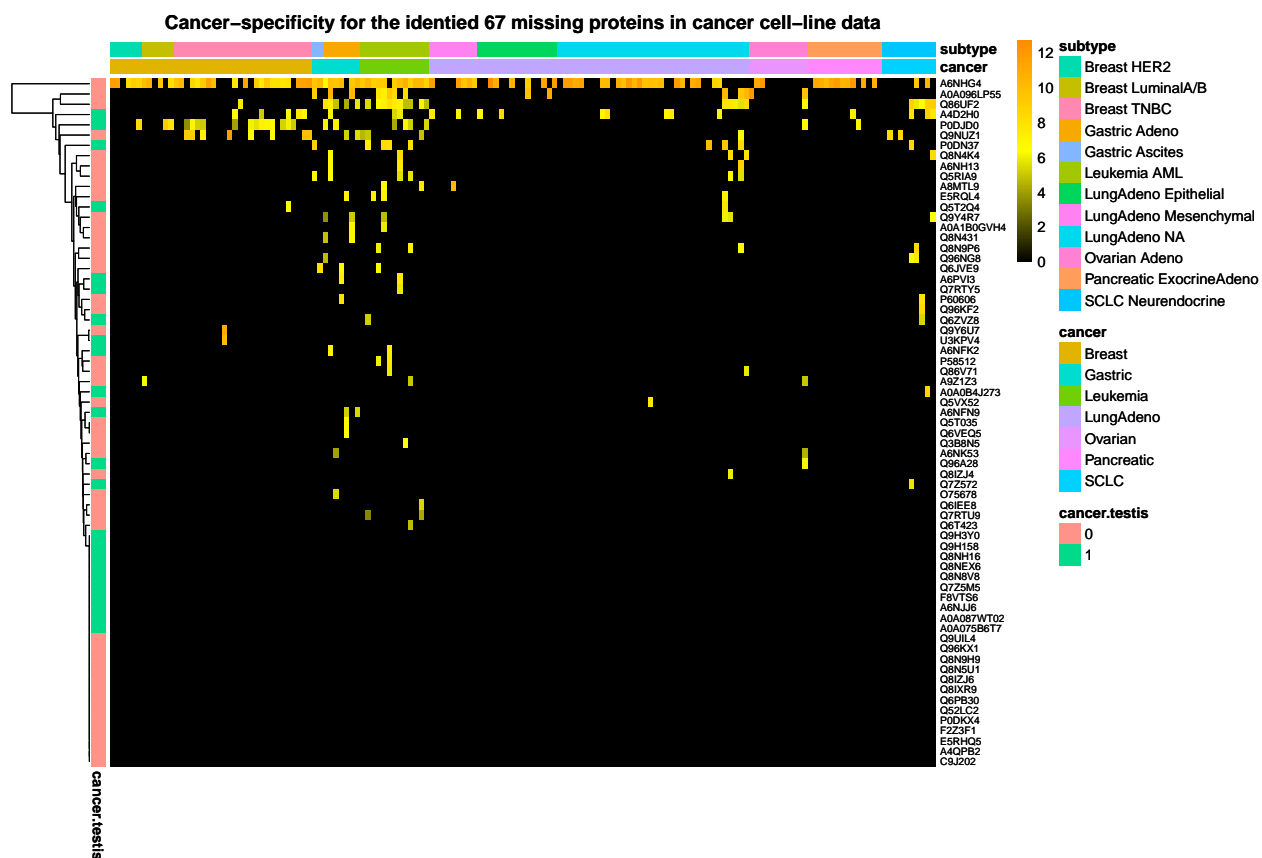
```
# ggsave('~Downloads/cancer_specificity_primary_cell_row_clustered.png', cancer_specificity_primary_cell_row_clustered.png,
#         width = 1020, height=735, units = 'px', dpi=80)
```

```
(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 identified 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 identified 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 identified 67 missing proteins in cancer cell-line data',
  show_colnames = F))
```

