Data Analysis

Jieqi Tu

12/10/2020

Import data

```r
HPV.data = readxl::read_excel("TCGA_HPV_mutation_Summary_042720 copy.xlsx")
```

First, we want extract genes: CD3e, CD3d, CD4, CD8a, CD8b, Foxp3, IFNg, IL2, PD1, PDL1, TP53, MYC, EGFR, ERBB2, ERBB3, EGF, NRG1 (Heregulin1/Neuregulin1), IL-6, TGFb1, FGF1, FGF2.

Extract genes

```r
HPV.data1 =
  HPV.data %>%
  filter(str_detect(name, c("CD3E")))
HPV.data2 =
  HPV.data %>%
  filter(str_detect(name, c("CD3D")))
HPV.data3 =
  HPV.data %>%
  filter(str_detect(name, c("CD4")))
HPV.data4 =
  HPV.data %>%
  filter(str_detect(name, c("CD8A")))
HPV.data5 =
  HPV.data %>%
  filter(str_detect(name, c("CD8B")))
HPV.data6 =
  HPV.data %>%
  filter(str_detect(name, c("FOXP3")))
HPV.data7 =
  HPV.data %>%
  filter(str_detect(name, c("IFNG")))
HPV.data8 =
  HPV.data %>%
  filter(str_detect(gene_name, c("COL8431")))
HPV.data9 =
  HPV.data %>%
  filter(str_detect(name, c("PD1"))) # not sure about the gene "PD1"
HPV.data10 =
  HPV.data %>%
  filter(str_detect(name, c("PDL"))) # not sure about the gene "PDL1"
HPV.data11 =
  HPV.data %>%
  filter(str_detect(name, c("TP53")))
HPV.data12 =
  HPV.data %>%
  filter(str_detect(name, c("MYC")))
HPV.data13 =
  HPV.data %>%
  filter(str_detect(name, c("EGFR")))
HPV.data14 =
  HPV.data %>%
  filter(str_detect(name, c("ERBB2")))
HPV.data15 =
  HPV.data %>%
  filter(str_detect(name, c("ERBB3")))
HPV.data16 =
  HPV.data %>%
  filter(str_detect(gene_name, c("COL5428")))
HPV.data17 =
  HPV.data %>%
  filter(str_detect(name, c("NRG1")))
HPV.data18 =
  HPV.data %>%
  filter(str_detect(name, c("IL6")))
HPV.data19 =
  HPV.data %>%
  filter(str_detect(name, c("TGFB1")))
HPV.data20 =
  HPV.data %>%
  filter(str_detect(gene_name, c("COL6387")))
HPV.data21 =
  HPV.data %>%
  filter(str_detect(gene_name, c("COL6392")))

# Write the results as an excel file
library(openxlsx)
write.xlsx(list(HPV.data1, HPV.data2, HPV.data3, HPV.data4, HPV.data5, HPV.data6, HPV.data7,
                HPV.data8, HPV.data9, HPV.data10, HPV.data11, HPV.data12, HPV.data13, HPV.data14,
                HPV.data15, HPV.data16, HPV.data17, HPV.data18, HPV.data19, HPV.data20, HPV.data21), "E
```

Genes are categorized in different sheets. Since we are not sure which genes are the targeted ones, we want reviewers to decide/filter what the correct genes should be.

Then we want to check whether there are other genes with p-values less than 0.05.

Check genes with p-value < 0.05

```
# Select genes with p-values < 0.05
HPV.data.pval =
  HPV.data %>%
  filter(p_value <= 0.05)
```

```r
nrow(HPV.data.pval)
```

```
## [1] 3609
```

```r
# Write the results as an excel file
write.xlsx(HPV.data.pval, "Genes_smallP.xlsx")
```

Since there are 3609 genes having p-values less than 0.05. We cannot make a scatter point plot with all the gene names added on.