Code

Finding the patterns of phosphorylation of NSCLC pathway-related genes by mutational profiles

Prepare for exploring data

First, prepare the packages necessary for the project.

Hide

```
library(tidyverse)
library(readxl)
library(dplyr)
library(reshape2)
```

Then, load and manipulate the data to be used.

Hide

```
d_4a<- read_excel('1-s2.0-S0092867420307431-mmc4.xlsx', sheet= 2, skip = 2, col_names = T)

d_5d<- read_excel('1-s2.0-S0092867420307431-mmc5.xlsx', sheet= 5, col_names = T)
d_5d[is.na(d_5d)]<- "0"
colnames(d_5d)<- c("Uniprot", "Gene", "KW_p-value", "Rank Median G1", "Rank Median G4", "Rank M
edian G3", "Rank Median twmix", "Rank Median G5",  "Rank Median G2", "Median T/N Ratio G1", "Me
dian T/N Ratio G4", "Median T/N Ratio G3", "Median T/N Ratio twmix", "Median T/N Ratio G5", "Me
dian T/N Ratio G2", "Enriched pathway", "Aryl hydrocarbon receptor signalling", "Chemical carci
nogenesis", "Drug metabolism - cytochrome P450", "NSCLC", "ErbB signaling pathway",  "MAPK sign
aling pathway")

d_5c<- read_excel('1-s2.0-S0092867420307431-mmc5.xlsx', sheet= 4, skip = 1, col_names = T)
```

```
New names:
* `` -> ...2
* `` -> ...3
* mRNA -> mRNA...4
* Protein -> Protein...5
* Phospho -> Phospho...6
* ...
```

Hide

```
colnames(d_5c)<- c("Carcinogen_enriched_pathway", "number", "-", "PAHs_mRNA", "PAHs_protein",
"PAHs_phospho", "NitroPAHs_mRNA", "NitroPAHs_protein", "NitroPAHs_phospho", "Mixed_mRNA", "Mixe
d_protein", "Mixed_phospho", "Nitrosamine_mRNA", "Nitrosamin_protein", "Nitrosamine_phospho",
"p-value_mRNA", "p-value_protein", 'p-value_phos')

d_2d<- read_excel('1-s2.0-S0092867420307431-mmc2.xlsx', sheet= 5, col_names = T)
colnames(d_2d)<- unlist(c("Accession", "Gene", "total", "count_log2T/N>0.5", "(%)_log2T/N>0.5",
"count_log2T/N<-0.5", "(%)_log2T/N<-0.5", "count_-0.5≤log2T/N≤0.5", "(%)_-0.5≤log2T/N≤0.5",
"missing protein count","missing protein count(%,/89)"))
d_2e<- read_excel('1-s2.0-S0092867420307431-mmc2.xlsx', sheet= 6, col_names = T)
```

Find the NSCLC pathway related genes

Find common genes between table S2D and table S5D. Table S2D suggests genes which are related to NSCLC pathway and table S5D gives us the median values of log2 tumor/normal ratio by five carcinogen groups.

Hide

```
common<- intersect(d_2d$Gene, d_5d$Gene)
common
```

```
 [1] "BAD"    "CDK4"   "CDK6"   "EGFR"   "ERBB2"  "MAP2K2" "PLCG1"  "PRKCA"  "RB1"    "SOS1"
```

Before exploring the data, let's first find out which pathway each gene is associate with and check overall enriched pathways.

Hide

```
d_5d %>% filter(Gene %in% common) %>%
  select(Gene, `Enriched pathway`)
```

Gene <chr>
SOS1
CDK4
CDK6
BAD
PLCG1
MAP2K2
ERBB2
EGFR
RB1
PRKCA

1-10 of 10 rows | 1-1 of 2 columns

Hide

```
 intersect(d_5d %>%filter(Gene %in% common) %>%  select(`Enriched pathway`) %>% str_split(., "WW;|WW,|WWWn") %>% unlist(), d_5c$Carcinogen_enriched_pathway)
```

```
Warning in stri_split_regex(string, pattern, n = n, simplify = simplify,  :
  argument is not an atomic vector; coercing
```

```
 [1] "ErbB signaling pathway"                    "Hepatitis C"                               "MAPK signaling pathway"                    "Natural killer cell mediated cytotoxicity"
 [5] "Non-small cell lung cancer"                "Regulation of actin cytoskeleton"          "T cell receptor signaling pathway"         "Epstein-Barr virus infection"
 [9] "Metabolic pathways"                        "Phosphatidylinositol signaling system"     "Toll-like receptor signaling pathway"
```

Through this, it can be seen that the 10 NSCLC-related genes are particularly related to the above 11 pathways.

Now, manipulate the table S5D so that you can easily see the log2 T/N ratio and Median rank for each gene according to mutational profiles.

Hide

```
d_5d_rank<- d_5d %>% filter(NSCLC == "v" & Gene %in% common) %>%
  melt(., id.vars= c("Gene")) %>%
  filter(variable %in% c('Rank Median G1', 'Rank Median G2', 'Rank Median G3', 'Rank Median G4', 'Rank Median G5', 'Rank Median twmix')) %>%
  mutate(group= case_when(variable == "Rank Median G1" ~ "G1",
                          variable == "Rank Median G2" ~ "G2",
                          variable == "Rank Median G3" ~ "G3",
                          variable == "Rank Median G4" ~ "G4",
                          variable == "Rank Median G5" ~ "G5",
                          variable == "Rank Median twmix" ~ "twmix")) %>%
  select(Gene, value, group)
colnames(d_5d_rank)<- unlist(c("Gene", "Rank", "Group"))

d_5d_ratio<- d_5d %>% filter(NSCLC == "v" & Gene %in% common) %>%
  melt(., id.vars= c("Gene")) %>%
  filter(variable %in% c('Median T/N Ratio G1', 'Median T/N Ratio G2', 'Median T/N Ratio G3', 'Median T/N Ratio G4', 'Median T/N Ratio G5', 'Median T/N Ratio twmix')) %>%
  mutate(group= case_when(variable == "Median T/N Ratio G1" ~ "G1",
                          variable == "Median T/N Ratio G2" ~ "G2",
                          variable == "Median T/N Ratio G3" ~ "G3",
                          variable == "Median T/N Ratio G4" ~ "G4",
                          variable == "Median T/N Ratio G5" ~ "G5",
                          variable == "Median T/N Ratio twmix" ~ "twmix" )) %>%
  select(Gene, value, group)
colnames(d_5d_ratio)<- unlist(c("Gene", "T/N_ratio", "Group"))

d_5d_merge<- merge(d_5d_rank, d_5d_ratio, by= c("Gene", "Group"))
d_5d_merge
```

Gene	Group	Rank	T/N_ratio
<chr>	<chr>	<chr>	<chr>
BAD	G1	6	-0.1755957095

Gene <chr>	Group <chr>	Rank <chr>	T/N_ratio <chr>
BAD	G2	2	0.2021475385
BAD	G3	1	0.250158289
BAD	G4	3	0.041708743
BAD	G5	4	0.033069077
BAD	twmix	5	-0.0974589805
CDK4	G1	6	-0.0209002715
CDK4	G2	3	0.22841123
CDK4	G3	4	0.09639055
CDK4	G4	2	0.370303858

1-10 of 60 rows Previous **1** 2 3 4 5 6 Next

Hide

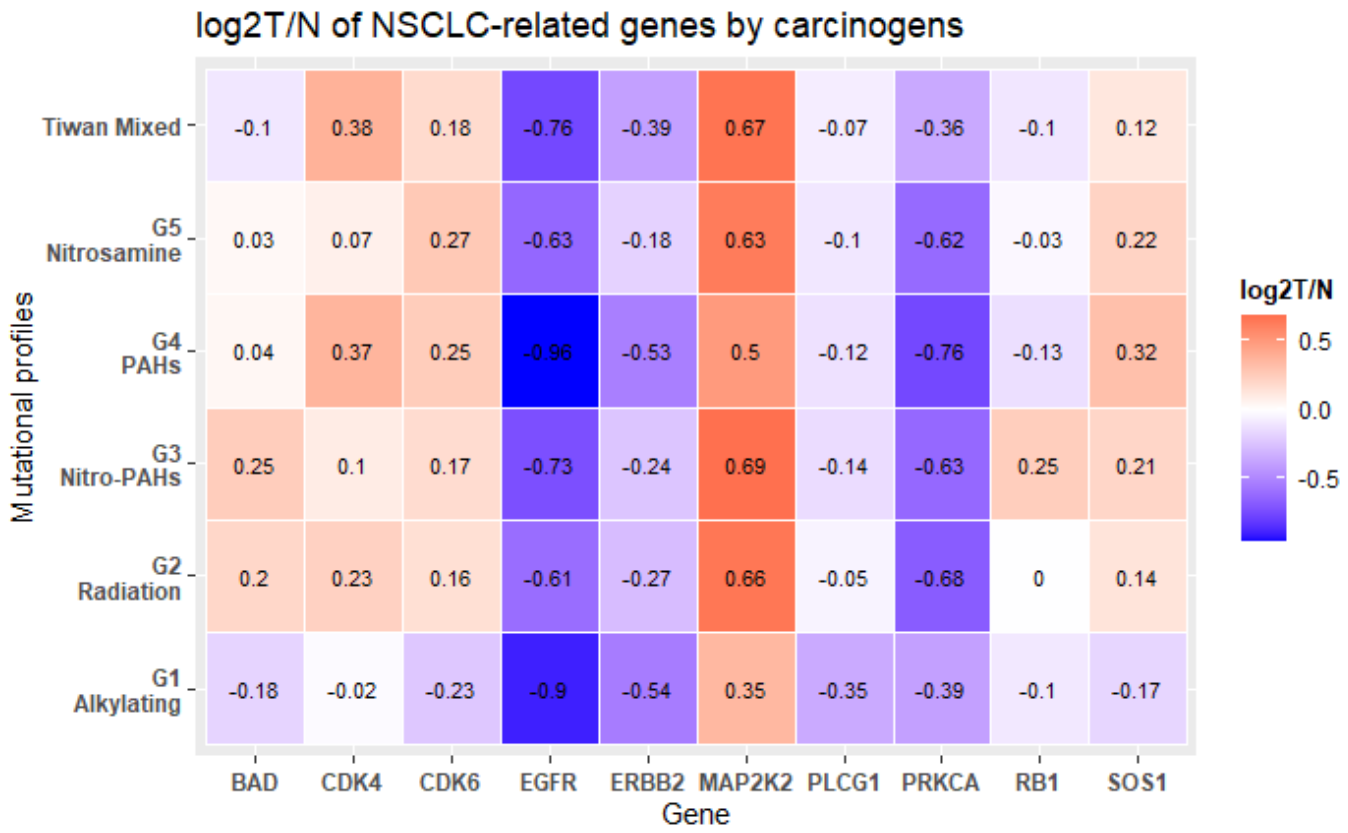
```
NA
```

Expression of NSCLC-pathway related genes

Using the processed data, plot the T/N ratio of 10 NSCLC-related genes for each mutational profile.

Hide

```
d_5d_merge %>%
  mutate(`T/N_ratio`= as.numeric(`T/N_ratio`)) %>%
  ggplot(aes(Gene, Group, fill= `T/N_ratio`))+
  geom_tile(color= "white")+
  geom_text(aes(label= round(`T/N_ratio`, 2)), color= "black", size=3)+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0)+
  labs(fill= "log2T/N", y= "Mutational profiles", title= "log2T/N of NSCLC-related genes by car
cinogens")+
  theme(legend.title = element_text(size= 10, face= "bold"),
        axis.text = element_text(face = "bold"))+
  scale_y_discrete(labels= c("twmix"= "Tiwan Mixed", "G1"= "G1WnAlkylating", "G2"= "G2WnRadiati
on", "G3"= "G3WnNitro-PAHs", "G4"= "G4WnPAHs", "G5"= "G5WnNitrosamine"))
```



Through the graph, it can be seen that the EGFR, ERBB2, and PRKCA genes are remarkably downregulated, and the MAP2K2 gene is upregulated. Considering that MAP2K2 is involved in the metabolic and signaling pathways of cancer cells, this result is reasonable. In addition, as the expression of EGFR and ERBB2, which are involved in adherens junctions, is reduced, it can be considered in association with EMT.

Phosphrylation patten of NSCLC-pathway related genes

Now, let's check what pattern of mutation occurred in the above genes using table S2E. As you can see from the results below, there is only information on 5 genes, so let's look at those genes. Also, sample sizes are too small so note that we cannot generalize the result.

Hide

```
d_2e %>% filter(`Gene name` %in% common) %>% select(`Gene name`) %>% table()
```

```
.
  BAD  EGFR MAP2K2 PLCG1   RB1
    4     8      1     3     6
```

And let's see the results in our final portfolio.. To be continued…