APOBEC mutational signature

1. Introduction

1.1 Background

Lung cancer is commonly known to develop easily among smokers or people exposed to cigarettes. However, lung cancer began to develop among women who had never smoked especially in East Asia. The most surprising thing is that early-onset patients of non-smoking women who developed lung cancer before the age of 60 showed similar patterns to smoking men. This appears to be the influence of environmental factors such as air pollution. We need to note that the APOBEC signature of these patients showed differences from others(Chen et al.,2020). Therefore, I intend to visualize the data by focusing on the APOBEC mutation signature.

1.2 Introduce the topic

The topic what I'm going to visualize is **Visualizing APOBEC mutational signature at the phosphoproteome levels using kinase enrichment**. In order to look at the APOBEC signature of early-onset women, we will look at the APOBEC mutational signature at the phosphoproteome level. Phosphorylation plays an important role, such as functioning of proteins, complex formation, degradation of proteins, and regulation of cell signaling networks. Knowing which kinase is associated with APOBEC signature will help develop targeted therapies at the phosphoproteome level. According to the research, kinase enrichment analysis identified AurB, CK2, CDK1, and CDK2 as the top-ranking activated kinases in the APOBEC-high female group(Weidner et al.,2014). And the activation of CDK1, CDK2, and AurB offers actionable intervention candidates for female patients with high APOBEC signature (Lin et al., 2018; Maslyk et al., 2017; Mross et al., 2016).

I plotted the bar graph using information of substrates' type and enrichment of kinase from the regulated phosphosites between APOBEC high and low groups in the given data-sets. Since paper's(Chen et al.,2020) graph is for groups with APOBEC-high, we can only determine which kinase is concentrated a lot in which substrate. However, can this be seen as a unique feature of APOBEC-high group? In order to conclude that this phenomenon has a significant result in APOBEC-high, I thought it should be compared with the APOBEC-low group, and I would like to graph both at the same time to compare the enrichment by group.

2. Explore the data

2.1 Unpacking data

```
library(dplyr)
```

```
##
##         : 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(ggplot2)
```

```
enriched_kinase<-read_excel("pathway_with_APOBEC.xlsx", sheet=3)
```

2.2 Processing the data

First, let's look what kind of data is in enriched_kinase data set.

```
head(enriched_kinase)
```

```
## # A tibble: 6 x 5
##   Kinase Substrate   `APOBEC high` `APOBEC low` `P-value`
##   <chr>  <chr>               <dbl>        <dbl>     <dbl>
## 1 CDK1   ANAPC1_S688         0.180        0.593  0.0194
## 2 CDK1   USP24_S2047        -0.119        0.142  0.00898
## 3 CDK1   TFPT_S180          -0.627        0.272  0.00207
## 4 CDK1   EEF1D_S133         -0.445        0.381  0.000386
## 5 CDK1   U2AF2_S79          -1.07         0.637  0.000794
## 6 CDK1   EIF4G2_T508         0.747        0.132  0.00246
```

It has 5 columns. Type of kinase, type of substrate, enrichment of kinase in APOBEC high and low patients, and p-value. Then let's take a closer look at kinase's data. How many kinds of kinases are there?

```
table(enriched_kinase$Kinase)
```

```
##
##      AurB AurB, CK2      CDK1 CDK1,CDK2      CDK2       CK2
##         5         2        18         2        12        12
```

There are 4 kinds of kinase! Oops! But two types of kinases are related to one substrate, so two pieces of information are located in one row. In this case, the graph cannot be drawn neatly. Let's manipulate the data frame so that only one type of information can fit in a row. After making a copy of a row containing two pieces of information, change the row name of the original data to the first of the two pieces of information, and also change the row name of the copy data to the name of the second data. Finally, combine the two data frames.

```
# Make a copy of a row containing two pieces of information
enriched_kinase1<-enriched_kinase %>% filter(Kinase=="CDK1,CDK2"|Kinase=="AurB, CK2")

# Change the row name of the copy data
enriched_kinase1$Kinase<- ifelse(enriched_kinase1$Kinase=="CDK1,CDK2","CDK2","CK2")

# Change the row name of the original data
enriched_kinase$Kinase<-ifelse(enriched_kinase$Kinase=="CDK1,CDK2","CDK1",
                               enriched_kinase$Kinase)
enriched_kinase$Kinase<-ifelse(enriched_kinase$Kinase=="AurB, CK2","AurB",
                               enriched_kinase$Kinase)

# Combine the two data frames
enriched_kinase_total<-bind_rows(enriched_kinase,enriched_kinase1)
```

Let's see if we have 55 objects!

```
enriched_kinase_total
```

```
## # A tibble: 55 x 5
##    Kinase Substrate          `APOBEC high` `APOBEC low` `P-value`
##    <chr>  <chr>                      <dbl>        <dbl>     <dbl>
##  1 CDK1   ANAPC1_S688                0.180       0.593    0.0194
##  2 CDK1   USP24_S2047               -0.119       0.142    0.00898
##  3 CDK1   TFPT_S180                 -0.627       0.272    0.00207
##  4 CDK1   EEF1D_S133                -0.445       0.381    0.000386
##  5 CDK1   U2AF2_S79                 -1.07        0.637    0.000794
##  6 CDK1   EIF4G2_T508                0.747       0.132    0.00246
##  7 CDK1   PDS5BS_S1358;T1370         1.34        0.706    0.0149
##  8 CDK1   SRRM2_T1413                0.585      -0.00836  0.00722
##  9 CDK1   SRRM2_T866                 0.670       0.305    0.0163
## 10 CDK1   LMNB1_S23                  0.195      -0.252    0.0233
## # ... with 45 more rows
```

Well done! Let's take a closer look at kinase's data one more time. Are the two information separated differently?

```
table(enriched_kinase_total$Kinase)
```

```
##
## AurB CDK1 CDK2  CK2
##    7   20   14   14
```

Perfect! Let's move on to the next step. The current data frame is wider. APOBEC-high and low group enrichments are located in different columns, so it is complicated to express them as a single graph. Let's combine the two groups into one column by transforming the data frame longer using the `pivot_longer` function.

```
# Make wider dataframe to longer dataframe

enriched_kinase2<- enriched_kinase_total%>% select(-`P-value`)%>%
  pivot_longer(!c('Substrate','Kinase'), names_to = 'APOBEC_signature',
               values_to = 'enrichment')
```

Let's see how the data frame has changed.

```
head(enriched_kinase2)
```

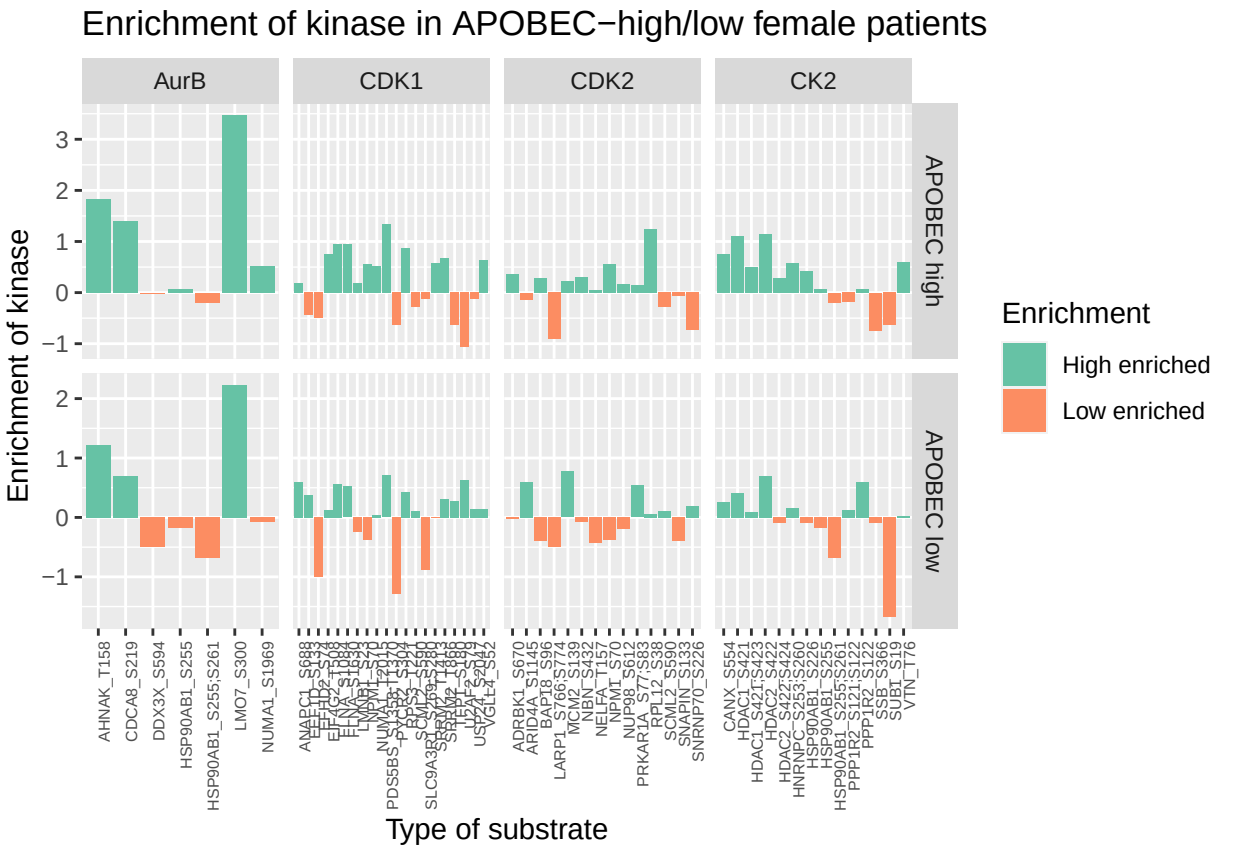
```
## # A tibble: 6 x 4
##   Kinase Substrate   APOBEC_signature enrichment
##   <chr>  <chr>       <chr>                <dbl>
## 1 CDK1   ANAPC1_S688 APOBEC high           0.180
## 2 CDK1   ANAPC1_S688 APOBEC low            0.593
## 3 CDK1   USP24_S2047 APOBEC high          -0.119
## 4 CDK1   USP24_S2047 APOBEC low            0.142
## 5 CDK1   TFPT_S180   APOBEC high          -0.627
## 6 CDK1   TFPT_S180   APOBEC low            0.272
```

Now, let's visualize with this data.

3. Data visualization

I will draw a bar graph to compare the enrichment for each Kinase according to the group.

```
enriched_kinase2 %>%
  mutate(Enrichment=ifelse(`enrichment`<0, "Low enriched","High enriched")) %>%
  ggplot(aes(Substrate,enrichment,fill=Enrichment))+
  geom_bar(stat = "identity")+
  facet_grid(APOBEC_signature~Kinase, scales="free") +
  theme(axis.text.x=element_text(angle=90, hjust=1,size=6)) +
  ggtitle("Enrichment of kinase in APOBEC-high/low female patients") +
  scale_fill_brewer(palette="Set2") +
  xlab("Type of substrate")+
  ylab("Enrichment of kinase")
```



A group is displayed at once using the `facet_grid` function for comparison by group. In order to determine the degree of enrichment, values were classified by color based on 0. Using the color of 'carrot', the high enriched one is shown in green and the low enriched one is shown in orange.

4. Discussion

According to the graph, we can compare the enrichment of kinase of APOBEC-high female patient groups to APOBEC-low group. As mentioned above, it can be seen that kinase is highly enriched in patients with APOBEC-high compared to patients with APOBEC-low. The graph showed that the concentration of these kinases was related to APOBEC mutation, which could cause lung cancer. Through this, as mentioned earlier, it is expected that knowing which kinase is associated with APOBEC signature will help develop targeted therapies at the phosphoproteome level.

Despite trying to draw the graph neatly and intuitively, there was a limitation. In order to make it easier to compare by group, scales should be adjusted to each other, but when scales are adjusted, the value becomes very small, making it difficult to recognize the graph. Therefore, `scale="free"` was used, which changed the scale of enrichment by group. In this process, distortion of the graph would have occurred. If this is supplemented, better results will be produced.