Correlation between Ber Protein and APOBEC3 families.

2020250151 lee seung yub

1.instruction

APOBEC3 enzymes are single-stranded DNA cytidine deaminases that can induce mutagenesis by catalyzing the conversion of cytidine to promutagenic uridine on single-stranded DNA. APOBEC signature mutations are common in several cancer types including breast, head/neck, cervical, bladder, and lung. This signature is comprised of C-to-T transitions and C-to-G transversions within TC(A/T) trinucleotide motifs resulting in TT(A/T) and TG(A/T) mutations. APOBEC-catalyzed C-to-U lesions must escape multiple DNA repair processes before becoming immortalized as mutations in tumor genomes. Most DNA uracil lesions are counteracted by base excision repair (BER), which excises the uracil base and choreographs its faithful replacement by another cytosine.

2.visualize this

2-1 setup

first ,load the packages

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
library(readxl)
library(ggplot2)
library(ggridges)
library(patchwork)
```

```
setwd("C:/Users/USER/Desktop/biostat/bsms222_151_lee")
```

```
getwd()
```

```
## [1] "C:/Users/USER/Desktop/biostat/bsms222_151_lee"
```

and load the excel files

```
read_excel('1-s2.0-S0092867420307431-mmc1..xlsx',sheet=2,na='NA')->d1
read_excel('1-s2.0-S0092867420307431-mmc1..xlsx',sheet=6,na='NA')->d2
```

change the row and colume of d2 to matchd1 and d2

```
d3<-as.data.frame(t(d2))
colnames(d3)<-d3[2,]
d3<-d3[-c(1,2,3), ]
```

delete the rows which have no data

```
data1<-d1[-c(91,82,78,79,73,74,64,43,42,37,34,31,3,2),]
```

check the type of base excision repair proteins

```
d2%>%filter(Gene%in%c('MBD4', 'APEX1' ,'FEN1', 'POLD2', 'HAT1', 'MCM2', 'MCM4', 'SMC2', 'NCAPD2',
'NCAPG', 'TOP2A', 'PRIM1' ,'PMS1', 'MLH1', 'MSH2' ,'MSH6'))
```

Accession <chr>
O14929
O95347
P11388
P27695
P33991
P39748
P40692
P43246
P49005
P49736

1-10 of 16 rows | 1-1 of 92 columns Previous **1** 2 Next

merge the needed parts of d3 and data1 and make the final data

```
data2<-data1%>%mutate(APOBEC3A=d3$APOBEC3A,APOBEC3B=d3$APOBEC3B,APOBEC3C=d3$APOBEC3C,APOBEC3D=d3$APOBEC3D,APOBEC3F=d3$APOBEC3F,APOBEC3G=d3$APOBEC3G,MBD4=d3$MBD4, APEX1=d3$APEX1, FEN1=d3$FEN1, POLD2=d3$POLD2, HAT1=d3$HAT1, MCM2=d3$MCM2 ,MCM4=d3$MCM4, SMC2=d3$SMC2, NCAPD2=d3$NCAPD2,NCAPG=d3$NCAPG, TOP2A=d3$TOP2A, PRIM1=d3$PRIM1, PMS1=d3$PMS1, MLH1=d3$MLH1, MSH2=d3$MSH2, MSH6=d3$MSH6 )
```

As there are too many ber protein genes, only use MBD4,APEX1,FEN1 and POLD2 which mentioned in the article

change the characters into numeric class

```
data2$APOBEC3A<-as.numeric(data2$APOBEC3A)
data2$APOBEC3B<-as.numeric(data2$APOBEC3B)
data2$APOBEC3C<-as.numeric(data2$APOBEC3C)
data2$APOBEC3D<-as.numeric(data2$APOBEC3D)
data2$APOBEC3F<-as.numeric(data2$APOBEC3F)
data2$APOBEC3G<-as.numeric(data2$APOBEC3G)
data2$MBD4<-as.numeric(data2$MBD4)
data2$APEX1<-as.numeric(data2$APEX1)
data2$FEN1<-as.numeric(data2$FEN1)
data2$POLD2<-as.numeric(data2$POLD2)
```

2-2 plotting

```
g1<-data2%>%filter(!is.na(APOBEC3A) & APOBEC3A!='NA')%>%
    filter(APOBEC3A>0)%>%
    ggplot()+
    geom_point(aes(APOBEC3A,MBD4,col='red'),alpha=0.3)+
    geom_smooth(aes(APOBEC3A,MBD4,col='red'),method='lm')+
    geom_point(aes(APOBEC3A,APEX1,col='yellow'),alpha=0.3)+
    geom_smooth(aes(APOBEC3A,APEX1,col='yellow'),method='lm')+
    geom_point(aes(APOBEC3A,FEN1,col='green'),alpha=0.3)+
    geom_smooth(aes(APOBEC3A,FEN1,col='green'),method='lm')+
    geom_point(aes(APOBEC3A,POLD2,col='blue'),alpha=0.3)+
    geom_smooth(aes(APOBEC3A,POLD2,col='blue'),method='lm')+
    scale_color_discrete(name='Berprotein',labels=c('MBD4(red)','APEX1(green)',
                                        'FEN1(blue)','POLD2(purple)'))+
   #add the y=x line to see the correlation clearly
   geom_abline(slope=1)+
   ylab('Berprotein')+ggtitle('Correlation between Ber Portein and APOBEC3 family')
```

repeat the process

```
g2<-data2%>%filter(!is.na(APOBEC3B) & APOBEC3B!='NA')%>%
    filter(APOBEC3B>0)%>%
    ggplot()+
    geom_point(aes(APOBEC3B,MBD4,col='red'),alpha=0.3)+
    geom_smooth(aes(APOBEC3B,MBD4,col='red'),method='lm')+
    geom_point(aes(APOBEC3B,APEX1,col='yellow'),alpha=0.3)+
    geom_smooth(aes(APOBEC3B,APEX1,col='yellow'),method='lm')+
    geom_point(aes(APOBEC3B,FEN1,col='green'),alpha=0.3)+
    geom_smooth(aes(APOBEC3B,FEN1,col='green'),method='lm')+
    geom_point(aes(APOBEC3B,POLD2,col='blue'),alpha=0.3)+
    geom_smooth(aes(APOBEC3B,POLD2,col='blue'),method='lm')+
    scale_color_discrete(name='Berprotein',labels=c('MBD4(red)','APEX1(green)',
                                        'FEN1(blue)','POLD2(purple)'))+
    geom_abline(slope=1)+
  ylab('Berprotein')+ggtitle('Correlation between Ber Portein and APOBEC3 family')
```

```
g3<-data2%>%filter(!is.na(APOBEC3C) & APOBEC3C!='NA')%>%
    filter(APOBEC3C>0)%>%
    ggplot()+
    geom_point(aes(APOBEC3C,MBD4,col='red'),alpha=0.3)+
    geom_smooth(aes(APOBEC3C,MBD4,col='red'),method='lm')+
    geom_point(aes(APOBEC3C,APEX1,col='yellow'),alpha=0.3)+
    geom_smooth(aes(APOBEC3C,APEX1,col='yellow'),method='lm')+
    geom_point(aes(APOBEC3C,FEN1,col='green'),alpha=0.3)+
    geom_smooth(aes(APOBEC3C,FEN1,col='green'),method='lm')+
    geom_point(aes(APOBEC3C,POLD2,col='blue'),alpha=0.3)+
    geom_smooth(aes(APOBEC3C,POLD2,col='blue'),method='lm')+
    scale_color_discrete(name='Berprotein',labels=c('MBD4(red)','APEX1(green)',
                                        'FEN1(blue)','POLD2(purple)'))+
    geom_abline(slope=1)+
  ylab('Berprotein')+ggtitle('Correlation between Ber Portein and APOBEC3 family')
```

```
g4<-data2%>%filter(!is.na(APOBEC3D) & APOBEC3D!='NA')%>%
    filter(APOBEC3D>0)%>%
    ggplot()+
    geom_point(aes(APOBEC3D,MBD4,col='red'),alpha=0.3)+
    geom_smooth(aes(APOBEC3D,MBD4,col='red'),method='lm')+
    geom_point(aes(APOBEC3D,APEX1,col='yellow'),alpha=0.3)+
    geom_smooth(aes(APOBEC3D,APEX1,col='yellow'),method='lm')+
    geom_point(aes(APOBEC3D,FEN1,col='green'),alpha=0.3)+
    geom_smooth(aes(APOBEC3D,FEN1,col='green'),method='lm')+
    geom_point(aes(APOBEC3D,POLD2,col='blue'),alpha=0.3)+
    geom_smooth(aes(APOBEC3D,POLD2,col='blue'),method='lm')+
    scale_color_discrete(name='Berprotein',labels=c('MBD4(red)','APEX1(green)',
                                        'FEN1(blue)','POLD2(purple)'))+
    geom_abline(slope=1)+
  ylab('Berprotein')+ggtitle('Correlation between Ber Portein and APOBEC3 family')
```

```
g5<-data2%>%filter(!is.na(APOBEC3F) & APOBEC3F!='NA')%>%
    filter(APOBEC3F>0)%>%
    ggplot()+
    geom_point(aes(APOBEC3F,MBD4,col='red'),alpha=0.3)+
    geom_smooth(aes(APOBEC3F,MBD4,col='red'),method='lm')+
    geom_point(aes(APOBEC3F,APEX1,col='yellow'),alpha=0.3)+
    geom_smooth(aes(APOBEC3F,APEX1,col='yellow'),method='lm')+
    geom_point(aes(APOBEC3F,FEN1,col='green'),alpha=0.3)+
    geom_smooth(aes(APOBEC3F,FEN1,col='green'),method='lm')+
    geom_point(aes(APOBEC3F,POLD2,col='blue'),alpha=0.3)+
    geom_smooth(aes(APOBEC3F,POLD2,col='blue'),method='lm')+
    scale_color_discrete(name='Berprotein',labels=c('MBD4(red)','APEX1(green)',
                                        'FEN1(blue)','POLD2(purple)'))+
    geom_abline(slope=1)+
  ylab('Berprotein')+ggtitle('Correlation between Ber Portein and APOBEC3 family')
```

```
g6<-data2%>%filter(!is.na(APOBEC3G) & APOBEC3G!='NA')%>%
    filter(APOBEC3G>0)%>%
    ggplot()+
    geom_point(aes(APOBEC3G,MBD4,col='red'),alpha=0.3)+
    geom_smooth(aes(APOBEC3G,MBD4,col='red'),method='lm')+
    geom_point(aes(APOBEC3G,APEX1,col='yellow'),alpha=0.3)+
    geom_smooth(aes(APOBEC3G,APEX1,col='yellow'),method='lm')+
    geom_point(aes(APOBEC3G,FEN1,col='green'),alpha=0.3)+
    geom_smooth(aes(APOBEC3G,FEN1,col='green'),method='lm')+
    geom_point(aes(APOBEC3G,POLD2,col='blue'),alpha=0.3)+
    geom_smooth(aes(APOBEC3G,POLD2,col='blue'),method='lm')+
    scale_color_discrete(name='Berprotein',labels=c('MBD4(red)','APEX1(green)',
                                        'FEN1(blue)','POLD2(purple)'))+
    geom_abline(slope=1)+
  ylab('Berprotein')+ggtitle('Correlation between Ber Portein and APOBEC3 family')
```

2-3 result

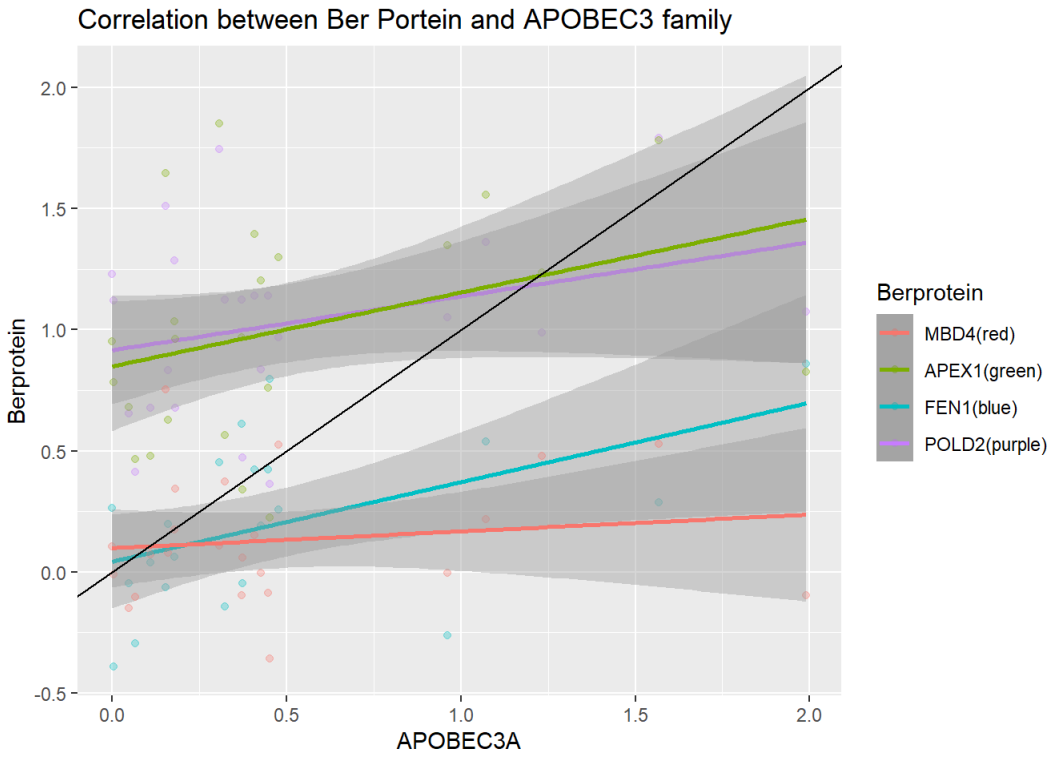
```
g1
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

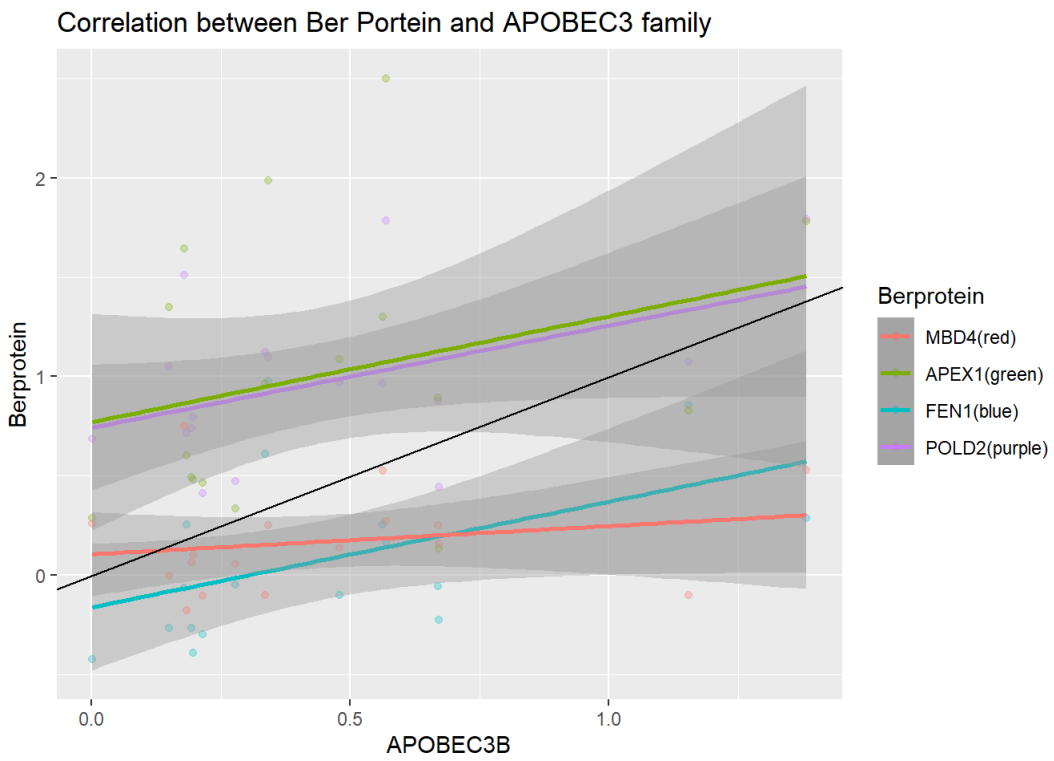
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
g2
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



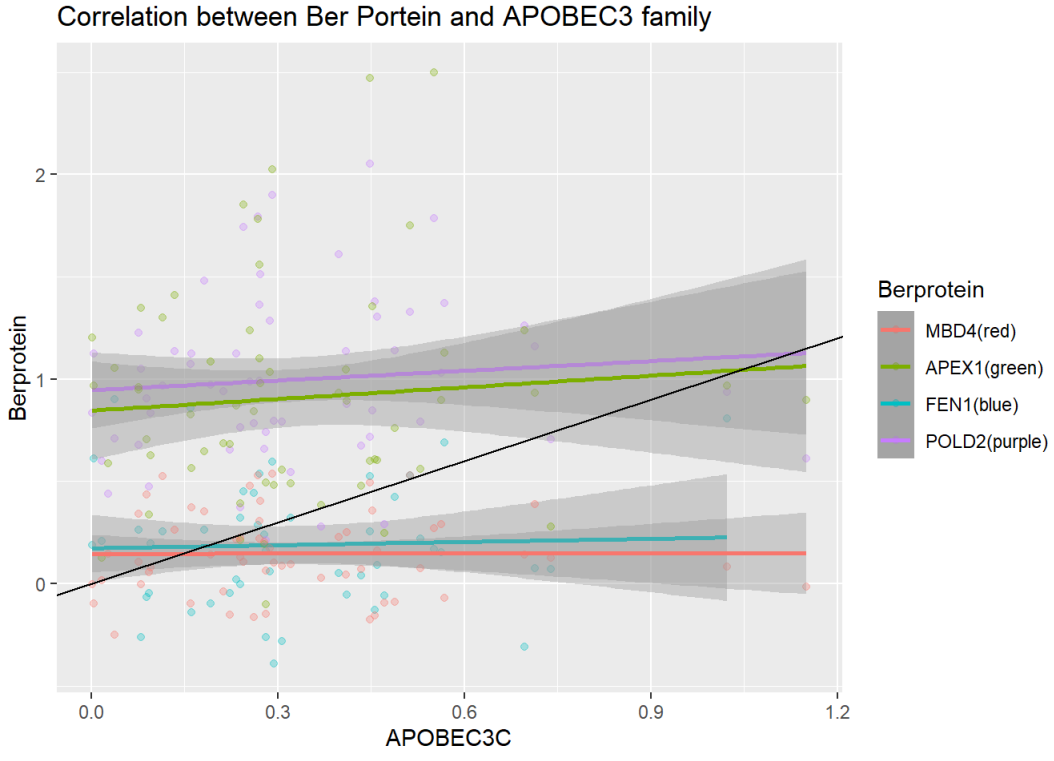

```
g3
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 11 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



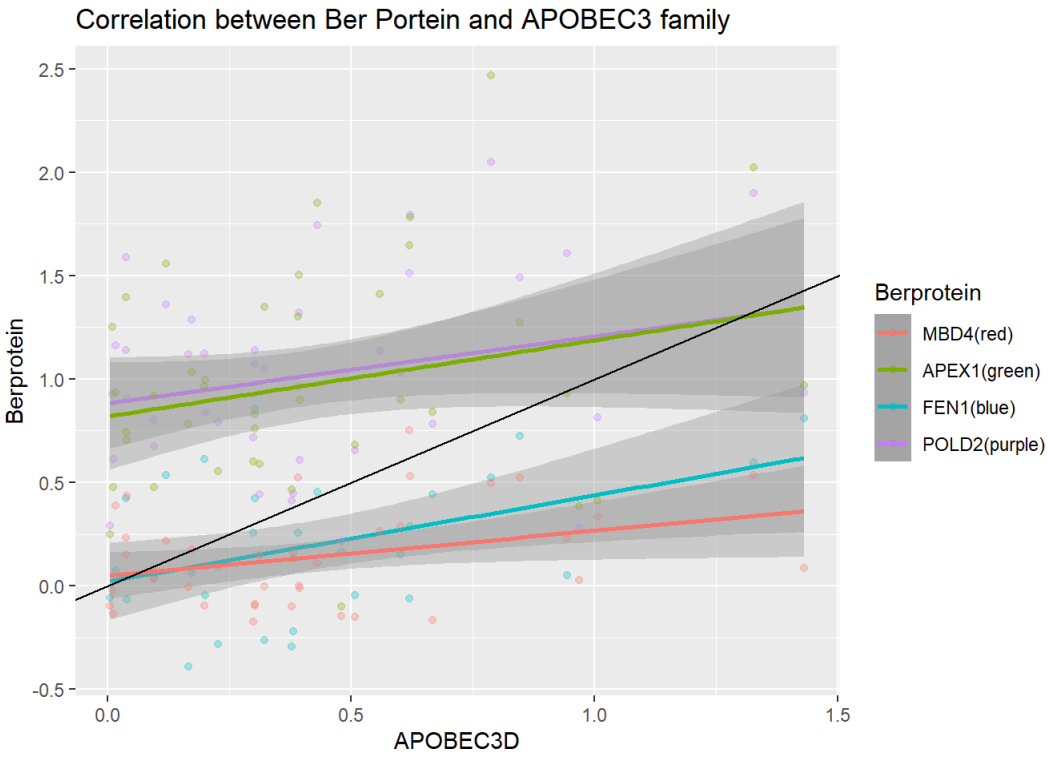

```
g4
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 10 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



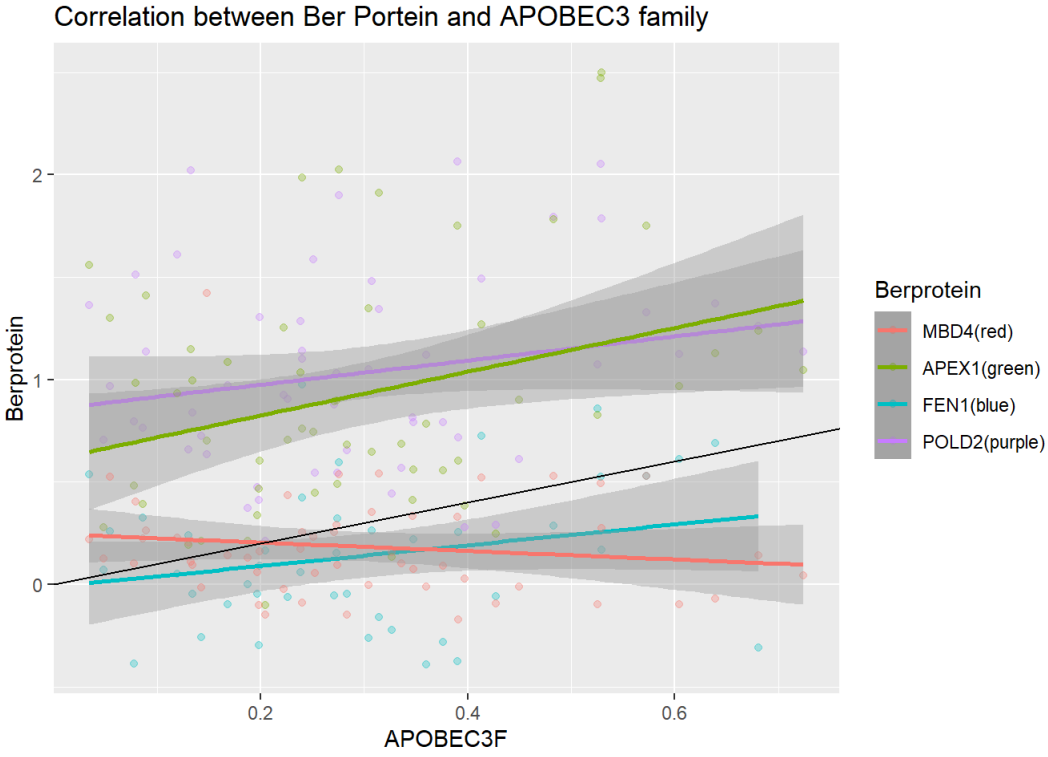

```
g5
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```



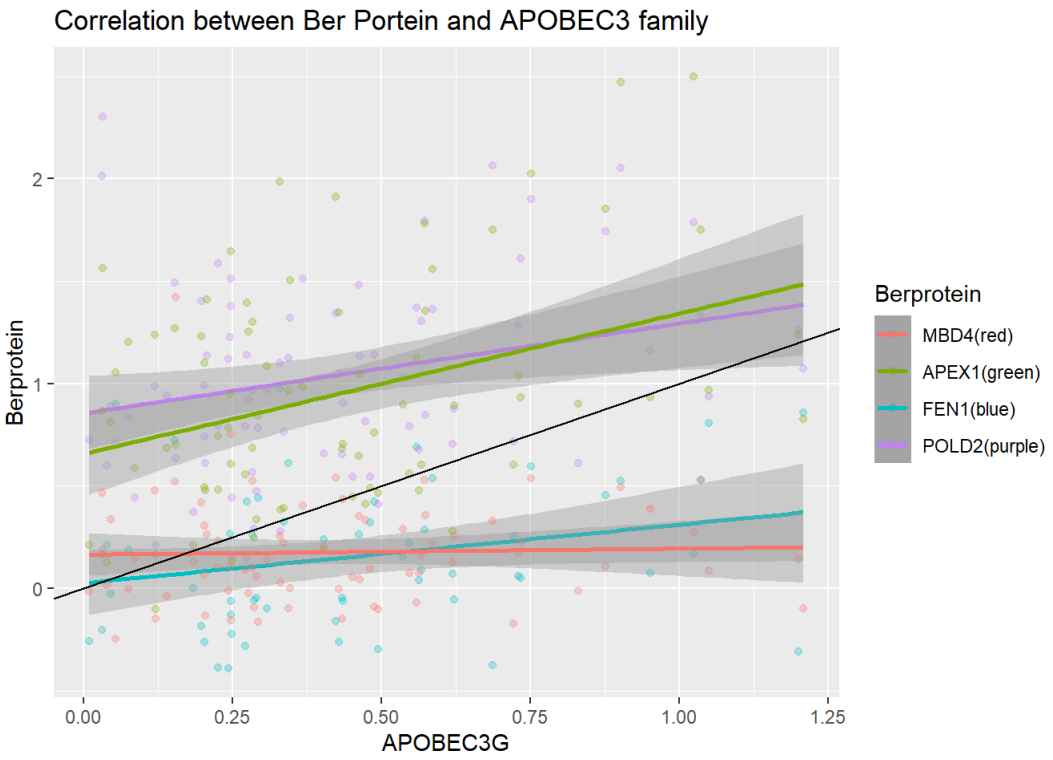
```
g6
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 18 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 18 rows containing missing values (geom_point).
```




```
(g1+g2)/(g3+g4)+(g6+g5)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 11 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 10 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 18 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

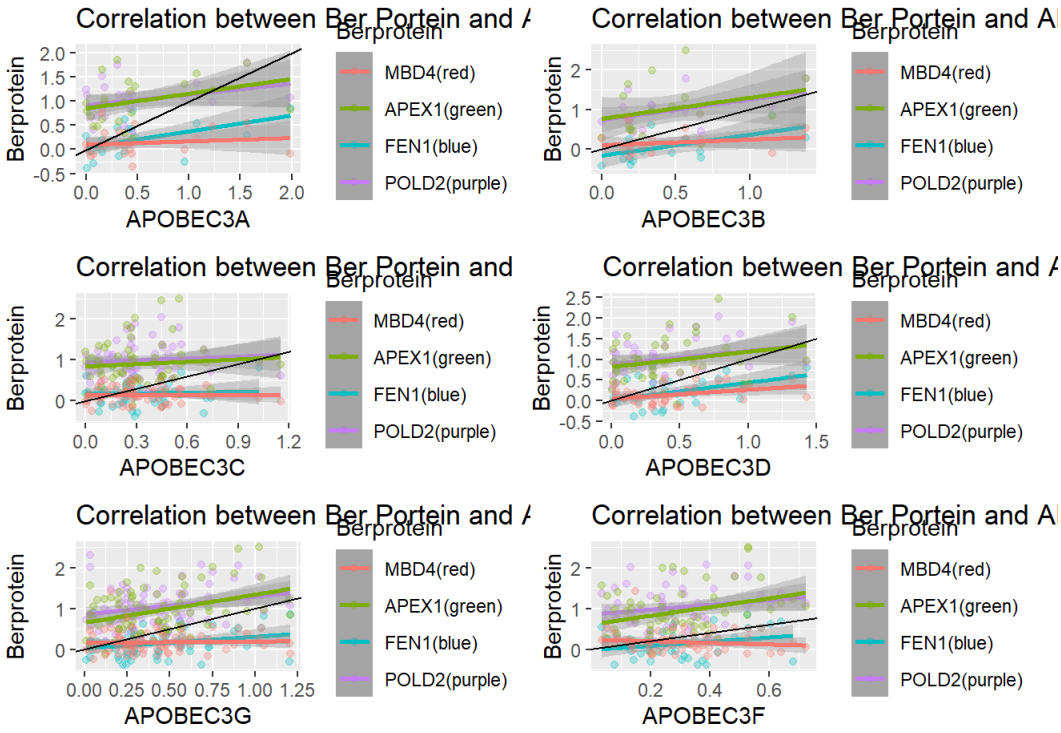
```
## Warning: Removed 18 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```



Except APOBec3C, there is positive correlation between APOBEC3 families and Berprotein.