Correlation between protein expression level of DNA replication, repair pathway and TP53 mutation in female nonsmoker lung cancer patient

Contents

1 Introduction

Lung cancer is very common and one of the leading cause of death in worldwide. There is lots of genetic, environmental factors affecting to the lung cancer and smoking is famous one that we know. But the nonsmokers can get the lung cancer. In Taiwanese population, nonsmoker patients are predominant (53%), especially among females (93%). It means nonsmoker females in East Asia can get the lung cancer more easily. There's more interesting sentences in our original paper(Chen et al., 2020). Let's read some of them. Researchers found that DNA replication, DNA repair upregulated in lung cancer through pathway enrichment analysis using protein Log2 T/N value. They also found positive association of TP53 mutation and higher phosphorylation of DNA repair protein. Thanks to the above consequences, I got a questions like these, 'Is there upregulation of DNA replication, DNA repair proteins in nonsmoker female?', 'Is there a correlation between TP53 mutation and DNA replication, DNA repair proteins in nonsmoker female?'. For these questions, I made 2 visualization figures. Figure 1 includes distribution of DNA replication, DNA repair proteins Log2 T/N value in nonsmoker females. For comparing, Figure 1 also has distribution of DNA replication, DNA repair proteins Log2 T/N value of other three groups(nonsmoker male, ex-smoker male, current-smoker male), divided by gender and smoking status, in taiwan cohort. Figure 2 includes the heatmap of Log2 T/N value of DNA replication, DNA repair proteins with TP53 mutation or not in nonsmoker female. Genes used in figure 1, 2 selected from 15 genes(HAT1, SMC2, NCAPD2, NCAPG, TOP2A, PRIM1, MBD4, APEX1, FEN1, POLD2, PMS1, MLH1, MSH2, MSH6, MCM2), used in analysis of correlation between APOBEC signature and DNA replication, DNA repair pathway, and additional 3 MCM families(MCM3, MCM6, MCM7).

Before start, let's look at the terms used in figure 1, 2.

1. Genes related to DNA replication

> HAT1 : HAT1 (Histone Acetyltransferase 1) is a Protein Coding gene. Among its related pathways are Chromatin organization and IL-2 Pathway. Gene Ontology (GO) annotations related to this gene include histone acetyltransferase activity and H4 histone acetyltransferase activity.

From [GeneCards]https://www.genecards.org/cgi-bin/carddisp.pl?gene=HAT1

> MCM family : MCM(minichromosome maintenance) complex has a role in both the initiation and the elongation phases of eukaryotic DNA replication, specifically the formation and elongation of the replication fork. MCM is a component of the pre-replication complex, which is a component of the licensing factor. MCM is a hexamer of six related polypeptides (MCM2 through MCM7) that form a ring structure.

From HGNC

> SMC2 : SMC2 (Structural Maintenance Of Chromosomes 2) is a Protein Coding gene. Diseases associated with SMC2 include Pleural Empyema and Progeroid Syndrome. Among its related pathways are Cell Cycle, Mitotic and Cell cycle_Chromosome condensation in prometaphase. Gene Ontology (GO) annotations related to this gene include protein heterodimerization activity. An important paralog of this gene is SMC3.

From GeneCards

> NCAPD2 : NCAPD2 (Non-SMC Condensin I Complex Subunit D2) is a Protein Coding gene. Diseases associated with NCAPD2 include Microcephaly 21, Primary, Autosomal Recessive and Microcephaly. Among its related pathways are Cell Cycle, Mitotic and Cell cycle_Chromosome condensation in prometaphase. Gene Ontology (GO) annotations related to this gene include binding and histone binding. An important paralog of this gene is NCAPD3.

From GeneCards

> TOP2A : This gene encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This nuclear enzyme is involved in processes such as chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication. It catalyzes the transient breaking and rejoining of two strands of duplex DNA which allows the strands to pass through one another, thus altering the topology of DNA.

From GeneCards

> PRIM1 : The replication of DNA in eukaryotic cells is carried out by a complex chromosomal replication apparatus, in which DNA polymerase alpha and primase are two key enzymatic components. Primase, which is a heterodimer of a small subunit and a large subunit, synthesizes small RNA primers for the Okazaki fragments made during discontinuous DNA replication. The protein encoded by this gene is the small, 49 kDa primase subunit.

From GeneCards

2. Genes related to DNA repair

MBD4 : The protein encoded by this gene is a member of a family of nuclear proteins related by the presence of a methyl-CpG binding domain (MBD). These proteins are capable of binding specifically to methylated DNA, and some members can also repress transcription from methylated gene promoters. This protein contains an MBD domain at the N-terminus that functions both in binding to methylated DNA and in protein interactions and a C-terminal mismatch-specific glycosylase domain that is involved in DNA repair.

From GeneCards

APEX1 : The APEX gene encodes the major AP endonuclease in human cells. It encodes the APEX endonuclease, a DNA repair enzyme with apurinic/apyrimidinic (AP) activity. Such AP activity sites occur frequently in DNA molecules by spontaneous hydrolysis, by DNA damaging agents or by DNA glycosylases that remove specific abnormal bases. The AP sites are the most frequent pre-mutagenic lesions that can prevent normal DNA replication.

From GeneCards

FEN1 : FEN1 (Flap Structure-Specific Endonuclease 1) is a Protein Coding gene. Diseases associated with FEN1 include Werner Syndrome and Vitelliform Macular Dystrophy. Among its related pathways are Cell Cycle, Mitotic and Homology Directed Repair. Gene Ontology (GO) annotations related to this gene include magnesium ion binding and damaged DNA binding. An important paralog of this gene is GEN1.

From GeneCards

POLD2 : This gene encodes the 50-kDa catalytic subunit of DNA polymerase delta. DNA polymerase delta possesses both polymerase and 3' to 5' exonuclease activity and plays a critical role in DNA replication and repair. The encoded protein is required for the stimulation of DNA polymerase delta activity by the processivity cofactor proliferating cell nuclear antigen (PCNA).

From GeneCards

MLH1 : The protein encoded by this gene can heterodimerize with mismatch repair endonuclease PMS2 to form MutL alpha, part of the DNA mismatch repair system. When MutL alpha is bound by MutS beta and some accessory proteins, the PMS2 subunit of MutL alpha introduces a single-strand break near DNA mismatches, providing an entry point for exonuclease degradation.

From GeneCards

MSH2 : The MSH2 gene provides instructions for making a protein that plays an essential role in repairing DNA. This protein helps fix errors that are made when DNA is copied (DNA replication) in preparation for cell division. The MSH2 protein joins with one of two other proteins, MSH6 or MSH3 (each produced from a different gene), to form a two-protein complex called a dimer. This complex identifies locations on the DNA where errors have been made during DNA replication. Another group of proteins, the MLH1-PMS2 dimer, then binds to the MSH2 dimer and repairs the errors by removing the mismatched DNA and replicating a new segment. The MSH2 gene is one of a set of genes known as the mismatch repair (MMR) genes.

From MedlinePlus

MSH6 : The MSH6 gene provides instructions for making a protein that plays an essential role in repairing DNA. This protein helps fix errors that are made when DNA is copied (DNA replication) in preparation for cell division. The MSH6 protein joins with another protein called MSH2 (produced from the MSH2 gene) to form a two-protein complex called a dimer. This complex identifies locations on the DNA where errors have been made during DNA replication. Additional proteins, including another dimer called the MLH1-PMS2 dimer, then repair the errors by removing the mismatched DNA and replicating a new segment. The MSH6 gene is a member of a set of genes known as the mismatch repair (MMR) genes.

From MedlinePlus

3. TP53

TP53 : This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome.

From GeneCards

4. Smoking status

nonsmoker : nonsmoker is someone who has not smoked more than 100 cigarettes in their lifetime and does not currently smoke.

From MINISTRY OF HEATH

ex-smoker : The term ex-smoker refers to an individual who has given up (i.e., quit) cigarette and/or tobacco smoking. Ex-smokers were previous current smokers, but are no longer smoking.

From Springer Link

current-smoker : Percent of adults who reported they have smoked at least 100 cigarettes in their entire life and that they now smoke some days or every day.

From ldchealth

2 Data wrangling

2.1 importing dataset

Load all packages that we need in this assignment.

```
# Load packages that we need
library(tidyverse)
library(readxl)
library(dplyr)
library(ggplot2)
library(ggridges)
library(janitor)
library(RColorBrewer)
library(ggside)
library(cowplot)
```

Let's import data using read_excel function! We can use supplementary data provided by our original paper(Chen et al., 2020).

```
# Import main data
main_data <-
  read_excel("F:/bsms222_141_Jang/portfolio/1-s2.0-S0092867420307431-mmc1.xlsx",
                         sheet = "Table S1A_clinical_103patient",
                         col_names = TRUE,
                         na = "NA")

# Import protein expression level data
protein_expression <-
  read_excel("F:/bsms222_141_Jang/portfolio/1-s2.0-S0092867420307431-mmc1.xlsx",
                         sheet = "Table S1E_ProteomeLog2TN",
                         col_names = TRUE,
                         na = "NA")

# Import Somatic mutation profile data
somatic_mutation <-
  read_excel("F:/bsms222_141_Jang/portfolio/1-s2.0-S0092867420307431-mmc1.xlsx",
                         sheet = "Table S1C_SNV",
                         col_names = TRUE,
                         skip = 1,
                         na = "NA")
```

2.2 Manipulating dataset

Let's filter the DNA repair, DNA replication protein from **protein_expression** dataset and TP53 somatic mutation data from **somatic_mutation** dataset.

```
# Filter DNA repair, replication protein from dataset.
DNA_rep <- protein_expression %>%
  filter(Gene %in% c("MCM2", "MCM3", "MCM6", "MCM7", "MBD4", "APEX1", "FEN1", "POLD2",
                     "HAT1", "SMC2", "NCAPD2", "NCAPG", "TOP2A", "PRIM1", "PMS1" ,"MLH1",
                     "MSH2", "MSH6")) %>%
  t() %>%
  as.data.frame() %>%
  mutate(ID = colnames(protein_expression)) %>%
  row_to_names(row_number = 2) %>%
  rename(ID = Gene) %>%
  slice(-(1)) %>%
  mutate(across(c(where(is.character), -ID), as.numeric))

# Filter TP53 somatic mutation from dataset.
TP53 <- somatic_mutation %>%
  filter(Gene == "TP53") %>%
  t() %>%
  as.data.frame() %>%
  mutate(ID = colnames(somatic_mutation)) %>%
  row_to_names(row_number = 1) %>%
  slice(-(1:2)) %>%
  rename(ID = Gene)
```

For data visualization, we have to merge three dataset(main_data, DNA_rep, TP53).

```r
# Merge three datasets!

merged_data <-
  merge(main_data, TP53, all = TRUE) %>%
  merge(., DNA_rep, all = TRUE)  %>%
  mutate(TP53 = ifelse(
    TP53 %in% c("stopgain",
                "frameshift_deletion",
                "nonsynonymous_SNV",
                "nonframeshift_insertion",
               "frameshift_deletion,nonsynonymous_SNV"), "Yes", "No"))
```

There's 18 columns including Log2 T/N values. For data visualization, we need to bind the Log2 T/N values and their names in each column using pivot_longer function.

```r
# Use pivot_longer function
data <- merged_data %>%
  pivot_longer(cols = -(1:10),
               names_to = "Gene",
               values_to = "Log2TN")
```

3 Data visualization

3.1 Figure 1

As mentioned in introduction, our original paper analyzed protein expression value(Log2 T/N value) of DNA replication, DNA repair pathway. But that analysis is based on whole patients level. So I divided Taiwan cohort patients in 4 groups(nonsmoker female, nonsmoker male, ex-smoker male, current-smoker male) by smoking status and gender. And compared nonsmoker female and other 3 groups. For this comparing, I made 4 ridgeline plots using geom_density_ridges function in ggridges package and modified plots by median Log2 T/N value. And then combined them using plot_grid function in cowplot package.

```r
p1 <- data %>%
  filter(Gender == "Female" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = Log2TN, y = reorder(Gene, Log2TN, FUN = median), fill = stat(x))) +
  geom_density_ridges_gradient(bandwidth = 0.1) +
  scale_fill_gradient(limits = c(-0.5,2), low = "Blue1", high = "Red1") +
  xlim(-0.5,2) +
  labs(title = "Female Nonsmoker",
       x = "Log2 T/N value",
       y = "DNA replication, repair proteins",
       fill = "Log2 \n T/N") +
  theme_ridges() +
  theme(plot.title = element_text(size = 10),
        axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8),
        legend.key.height = unit(0.5, 'cm'),
        legend.key.width = unit(0.5, 'cm'),
        legend.title = element_text(size = 10, hjust = -0.2),
        axis.text.y = element_text(size = 5.5, vjust = 0.4)) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed")
```

```r
p2 <- data %>%
  filter(Gender == "Male" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = Log2TN, y = reorder(Gene, Log2TN, FUN = median), fill = stat(x))) +
  geom_density_ridges_gradient(bandwidth = 0.1) +
  scale_fill_gradient(limits = c(-0.5,2), low = "Blue1", high = "Red1") +
  xlim(-0.5,2) +
  labs(title = "Male Nonsmoker",
       x = "Log2 T/N value",
       y = "DNA replication, repair proteins",
       fill = "Log2 \n T/N") +
  theme_ridges() +
  theme(plot.title = element_text(size = 10),
        axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8),
        legend.key.height = unit(0.5, 'cm'),
        legend.key.width = unit(0.5, 'cm'),
        legend.title = element_text(size = 10, hjust = -0.2),
        axis.text.y = element_text(size = 5.5, vjust = 0.4)) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed")
```

```r
p3 <- data %>%
  filter(Gender == "Male" & `Smoking Status` == "Ex-smoker" & !is.na(Log2TN)) %>%
  ggplot(aes(x = Log2TN, y = reorder(Gene, Log2TN, FUN = median), fill = stat(x))) +
  geom_density_ridges_gradient(bandwidth = 0.1) +
  scale_fill_gradient(limits = c(-0.5,2), low = "Blue1", high = "Red1") +
  xlim(-0.5,2) +
  labs(title = "Male Ex-smoker",
       x = "Log2 T/N value",
       y = "DNA replication, repair proteins",
       fill = "Log2 \n T/N") +
  theme_ridges() +
  theme(plot.title = element_text(size = 10),
        axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8),
        legend.key.height = unit(0.5, 'cm'),
        legend.key.width = unit(0.5, 'cm'),
        legend.title = element_text(size = 10, hjust = -0.2),
        axis.text.y = element_text(size = 5.5, vjust = 0.4)) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed")
```

```r
p4 <- data %>%
  filter(Gender == "Male" & `Smoking Status` == "Current_Smoker" & !is.na(Log2TN)) %>%
  ggplot(aes(x = Log2TN, y = reorder(Gene, Log2TN, FUN = median), fill = stat(x))) +
  geom_density_ridges_gradient(bandwidth = 0.1) +
  scale_fill_gradient(limits = c(-0.5,2), low = "Blue1", high = "Red1") +
  xlim(-0.5,2) +
  labs(title = "Male Current-smoker",
       x = "Log2 T/N value",
       y = "DNA replication, repair proteins",
       fill = "Log2 \n T/N") +
  theme_ridges() +
  theme(plot.title = element_text(size = 10),
        axis.title.x = element_text(size = 8),
```

```
        axis.title.y = element_text(size = 8),
        legend.key.height = unit(0.5, 'cm'),
        legend.key.width = unit(0.5, 'cm'),
        legend.title = element_text(size = 10, hjust = -0.2),
        axis.text.y = element_text(size = 5.5, vjust = 0.4)) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed")
```

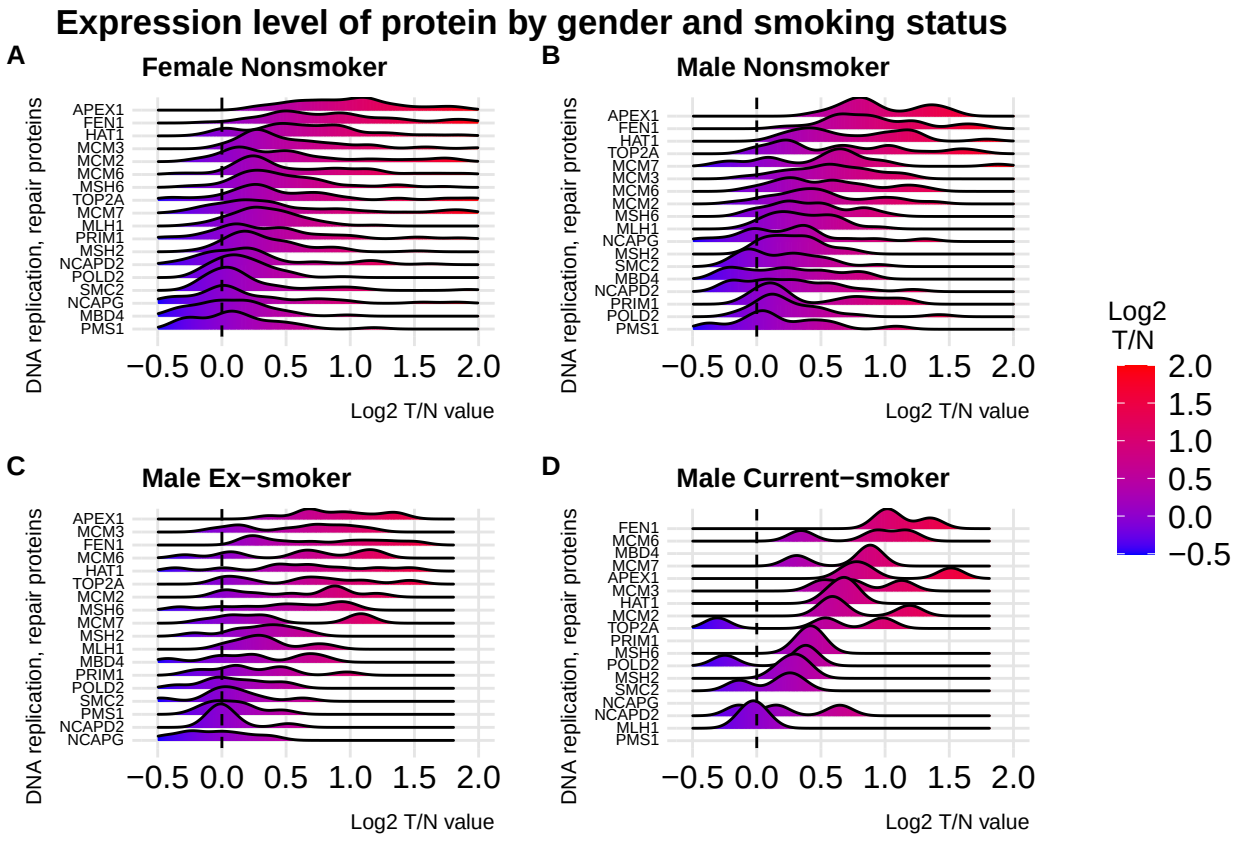
```
title <- ggdraw() +
  draw_label("Expression level of protein by gender and smoking status",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

p <- plot_grid(p1 + theme(legend.position="none"),
               p2 + theme(legend.position="none"),
               p3 + theme(legend.position="none"),
               p4 + theme(legend.position="none"),
               labels = c("A", "B", "C", "D"),
               label_size = 10)

pp <- plot_grid(title, p, ncol=1, rel_heights=c(0.1, 2))

legend <- get_legend(p1 + theme(legend.box.margin = margin(0, 0, 0, 12)))

plot_grid(pp, legend, rel_widths = c(3, 0.5))
```



As a result, all 4 groups have overall upregulation of DNA replication, DNA repair proteins. And there's almost no difference in nonsmoker male and ex-smoker male and current-smoker male has more upregulation pattern. current-smoker male > ex-smoker male, nonsmoker male > nonsmoker female is the arrangement of expression level in order. Then, why these 4 groups have an overall upregulation? Let's look at the original paper for suggestion, there's 2 sentences like 'Pathway analysis in female patients revealed a number of proteins involved in DNA repair and replication more abundant in the tumors with high APOBEC signature', 'Higher expression of base excision repair (BER) proteins in APOBEC-high females, including MBD4, APEX1, FEN1, and POLD2, implicates a role of BER in counteracting APOBEC-induced mutagenesis.'. Look at the first sentence. By reading original paper, we can know that high APOBEC signature has more mutations than low APOBEC signature, so it can mean there's correlation between increasing of mutations and DNA replication, repair proteins upregulation. Generally upregulation of abnormal DNA replication protein can induce genomic instability by producing mutations. It means upregulation of abnormal DNA replication in cancer patients can induce increasing of mutations of high APOBEC signature. By second sentence, I can know that increased mutations can be repaired by DNA repair proteins upregulation. Because of above suggestions I made a one hypothesis that upregulation of abnormal DNA replication proteins induces increasing of mutations and these mutations repaired by upregulation of DNA repair proteins. By applying this hypothesis, overall upregulation of DNA replication and DNA repair proteins in 4 groups can be explained. And let's think about why current-smoker male has more upregulation than other 3 groups. Generally, the smoking can make more mutations and cause genetic instability and induce the cancer. So, more upregulation pattern of current-smoking male than other 3 groups is because of increasing of mutations and genomic instabilities. And I found that the overall slight upregulation of DNA replication, DNA repair proteins in nonsmoker male patients than nonsmoker female patients.

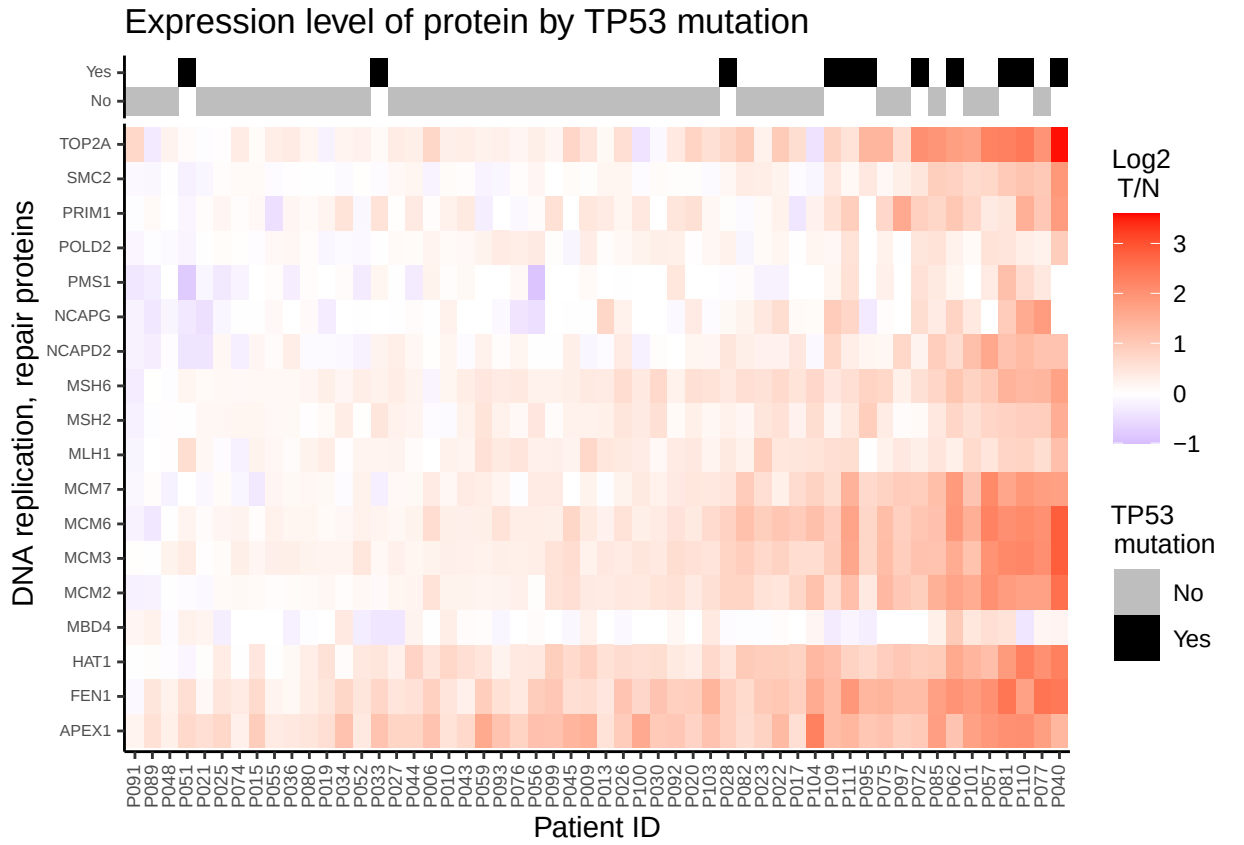
In short, from the original paper sentences, I made the hypothesis that upregulation of abnormal DNA replication proteins can induce increasing of mutations and upregulation of DNA repair proteins for restoring mutations. According to this hypothesis, all 4 group cancer patients have upregulation of DNA replication, DNA repair proteins and especially current-smoker has more upregulation because of increasing of mutations induced by smoking. My hypothesis is just an suggestion based on original paper and general knowledge but quite fit with Figure 1 consequence.

3.2 Figure 2

To visualize expression level of DNA replication, DNA repair protein in nonsmoker female, I used geom_tile function to make heatmap arranged by median Log2 T/N value from ggplot2 package and made a side bar using ggside package to express whether TP53 mutation is or not.

```
data %>%
  filter(Gender == "Female" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 7, vjust = 0.4),
        axis.text.y = element_text(size = 6, vjust = 0.4),
        legend.title = element_text(size = 10, hjust = -0.1)) +
  scale_fill_gradient2(midpoint=0,limit=c(-1, 3.6),
                       low="blue", mid="white",
                       high="red", na.value="white") +
  geom_xsidetile(aes(y = TP53, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Expression level of protein by TP53 mutation",
       x = "Patient ID",
       y = "DNA replication, repair proteins",
```

```
        xfill = "TP53 \nmutation",
        fill = "Log2 \n T/N")
```



As a result, DNA replication and DNA repair protein Log2 T/N value of nonsmoker female increased like we saw in Figure 1. And I found that DNA replication and DNA repair protein more upregulated with TP53 mutation. I think my hyphthesis described in Figure 1 consequence also can be applied in Figure 2 consequence. In normal state, TP53 can regulate Cell cycle, DNA damage protein so it can block cancer proliferation. But if there's mutation in TP53, it loses cell cycle regulation ability and abnormal DNA replication increased. It can induce increasing of mutation and accelerate cancer proliferation. And increased mutation can be repaired by upregulation of DNA repair protein. As I mentioned in introduction, according to original paper, TP53 mutation is correlated with high phosphorylation of DNA repair mechasnism. Generally, high phosphorylation of DNA repair can activate DNA repair mechanism. It means TP53 mutation can activate DNA repair.

Taken together, TP53 mutation can upregulate abnormal DNA replication protein by losing cell cycle regulation ability and DNA repair protein upregulated to repair mutations.

4 Discussion

In Figure 1, nonsmoker female has upregulation of DNA replication, repair protein but other 3 groups, nonsmoker male, ex-smoker male, current-smoker male, have more upregulation pattern than nonsmoker female. And we can know that smoking can amplify DNA replication and DNA repair protein.

In Figure 2, we can also see upregulation of DNA replication, repair protein in nonsmoker female. And TP53 mutation can induce more upregulation of DNA replication by losing cell cycle control ability and

upregulation of DNA repair protein. Description of original paper that TP53 mutation associated with high phosphorylation of DNA repair protein underpins the consequence.

I wrote a few things that I think could have been better.

- Log2 T/N value of the dataset aquired from individual level, not cell level. So I can get expression level of DNA replication and DNA repair protein just in patient level, not in cancer cell level.
- Current- smoker, ex-smoker male has a few sample so I can't analyze exactly.
- I don't know which environmental, genetic factors related to upregulation of DNA replication protein.
- The number of Log2 T/N value of Patients with TP53 mutation is a few so I can't take it in plot.

If there's a dataset that can supplement above problems, better consequences will come out.

5 Reference

Yi-Ju Chen, Theodoros I.Roumeliotis, Ya-Hsuan Chang, Ching-Tai Chen, Chia-Li Han, Miao-Hsia Lin, ...& Yu-Ju Chen. (2020). Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell.