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

Contents

Dae Hee Jang

2018250141, School of Biosystem and Biomedical Science, College of Health Science, Korea University

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 pathway in nonsmoker female and other groups?', 'Is there a correlation between TP53 mutation and DNA replication, DNA repair pathway in nonsmoker female and other groups?'. For these questions, I made 2 visualization figures. Figure 1 visualized protein expression level of DNA replication, DNA repair pathways in Taiwan cohort nonsmoker females using ridgeline plot and heatmap. For comparing, figure 1 also visualized protein expression level of DNA replication, DNA repair pathways in other three groups(nonsmoker male, ex-smoker male, current-smoker male), divided by gender and smoking status, in Taiwan cohort using ridgeline plot, heatmap. Figure 2 has same format used in figure 1. But figure 2 used RNA expression level other than figure 1. Some peer reviewers wondered why there isn't ex-smoker and current smoker data of female. But there isn't ex-smoker or current smoker female because only nonsmoker female included in supplementary data of original paper. Genes used in this portfolio 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 in original paper(Chen et al., 2020), and this portfolio also used additional 3 genes of MCM families(MCM3, MCM6, MCM7). Some peer reviewers wanted to analyze phosphorylation. So I tried to visualize phosphorylation level of DNA replication and DNA repair protein but there's few observations in supplementary data of original paper so I can't visualize phosphorylation.

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, (https://www.genenames.org/data/genegroup/#!/group/1085)

> 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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=SMC2)

> 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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=NCAPD2)

> 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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TOP2A)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=PRIM1)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=MBD4)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=APEX1)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=FEN1)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=POLD2)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=MLH1)

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, (https://medlineplus.gov/genetics/gene/msh2/)

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, (https://medlineplus.gov/genetics/gene/msh6/)

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, (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TP53)

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, (https://www.health.govt.nz/our-work/preventative-health-wellness/tobacco-control/tobacco-control-information-practitioners/definitions-smoking-status)

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, (https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-1005-9_313)

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, (https://ldchealth.org/177/Current-Smoker)

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 RNA expression level data
RNA_expression <-
  read_excel("F:/bsms222_141_Jang/portfolio/1-s2.0-S0092867420307431-mmc1.xlsx",
                        sheet = "Table S1D_transcriptome_log2TN",
                        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 RNA expression level of DNA repair, DNA replication genes from **RNA_expression** dataset and the DNA repair, protein expression level of DNA replication genes from **Protein_expression** dataset and TP53 somatic mutation data from **somatic_mutation** dataset.

```r
# Filter genes related to DNA repair, replication pathway from RNA_expression dataset.
RNA <-
  RNA_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(RNA_expression)) %>%
  row_to_names(row_number = 1) %>%
  rename(ID = gene) %>%
  slice(-(1:2)) %>%
  mutate(across(c(where(is.character), -ID), as.numeric))

# Filter genes related to DNA repair, replication pathway from Protein_expression dataset.
protein <-
  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 make two merged data for RNA, protein expression level. Let's make two merged data(merged_data1 and merged_data2) for each RNA, protein expression level.

```r
# Merge three datasets for RNA expression level visualization!

merged_data1 <-
  merge(main_data, TP53, all = TRUE) %>%
  merge(., RNA, by.y = "ID") %>%
  mutate(TP53 = ifelse(
    TP53 %in% c("stopgain",
                "frameshift_deletion",
                "nonsynonymous_SNV",
                "nonframeshift_insertion",
              "frameshift_deletion,nonsynonymous_SNV"), "Yes", "No"))

# Merge three datasets for protein expression level visualization!

merged_data2 <-
  merge(main_data, TP53, all = TRUE) %>%
  merge(., protein, by.y = "ID") %>%
  mutate(TP53 = ifelse(
    TP53 %in% c("stopgain",
                "frameshift_deletion",
                "nonsynonymous_SNV",
                "nonframeshift_insertion",
              "frameshift_deletion,nonsynonymous_SNV"), "Yes", "No"))
```

There's 18 columns in merged_data1 and merged_data2 datasets 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.

```
# Use pivot_longer function.
RNA_data <-
  merged_data1 %>%
  pivot_longer(cols = -(1:10),
               names_to = "Gene",
               values_to = "Log2TN")

Protein_data <-
  merged_data2 %>%
  pivot_longer(cols = -(1:10),
               names_to = "Gene",
               values_to = "Log2TN")
```

To distinguish genes included in DNA replication, repair pathway, let's make a 'pathway' column to divide genes in two groups(DNA replication, DNA repair) using as.factor function.

```
# Make DNA replication, DNA repair related genes in RNA_data dataset as factor.
RNA_data <-
  RNA_data %>%
  mutate(pathway = as.factor(ifelse(
    Gene %in% c("HAT1", "MCM2", "MCM3", "MCM6", "MCM7", "SMC2", "NCAPD2", "TOP2A", "PRIM1"),
    "DNA replication", "DNA repair")))

# Make DNA replication, DNA repair related genes in Protein_data dataset as factor.
Protein_data <-
  Protein_data %>%
  mutate(pathway = as.factor(ifelse(
    Gene %in% c("HAT1", "MCM2", "MCM3", "MCM6", "MCM7", "SMC2", "NCAPD2", "TOP2A", "PRIM1"),
    "DNA replication", "DNA repair")))
```

3 Data visualization

3.1 Figure 1

As mentioned in introduction, figure 1 expresses protein expression level of genes related to DNA replication, DNA repair pathway in nonsmoker female and other groups using ridgeline plot and heatmap. For this, 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 heatmap using geom_tile function in ggplot package and modified these plots by median Log2 T/N value. In heatmap, I made a side bar using geom_xsidetile function in ggside package to express whether TP53 mutation is or not. And then combined them using plot_grid function in cowplot package.

p_A1~A4 are ridgeline plot.

```
p_A1 <-
  Protein_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 = "Nonsmoker Female",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")

p_A2 <-
  Protein_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 = "Nonsmoker Male",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")

p_A3 <-
  Protein_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 = "Ex-smoker Male",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")

p_A4 <-
  Protein_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 = "Current-smoker Male",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")
```

p_B1~B4 are heatmap.

```
p_B1 <-
  Protein_data %>%
  filter(Gender == "Female" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Nonsmoker Female",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")

p_B2 <-
  Protein_data %>%
  filter(Gender == "Male" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Nonsmoker Male",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")

p_B3 <-
  Protein_data %>%
  filter(Gender == "Male" & `Smoking Status` == "Ex-smoker" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Ex-smoker Male",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")

p_B4 <-
  Protein_data %>%
  filter(Gender == "Male" & `Smoking Status` == "Current_Smoker" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Current-smoker Male",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")
```

Let's combine two type plots using plot_grid function and save these plots using ggsave function for better visualization.

```
title_1_1 <-
  ggdraw() +
  draw_label("Figure 1 - Expression level of protein by gender and smoking status",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

title_1_2 <-
  ggdraw() +
  draw_label("ridgeline plot",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

title_1_3 <-
  ggdraw() +
  draw_label("heatmap",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

p_A <- plot_grid(p_A1 + theme(legend.position="none"),
               p_A2 + theme(legend.position="none"),
               p_A3 + theme(legend.position="none"),
               p_A4 + theme(legend.position="none"),
               labels = c("A", "B", "C", "D"),
               label_size = 10)

p_B <- plot_grid(p_B1 + theme(legend.position="none"),
               p_B2 + theme(legend.position="none"),
               p_B3 + theme(legend.position="none"),
               p_B4 + theme(legend.position="none"),
               labels = c("E", "F", "G", "H"),
               label_size = 10)

pp_1 <- plot_grid(title_1_2, p_A, ncol=1, rel_heights=c(0.1, 2))

pp_A <- plot_grid(title_1_1, pp_1, ncol=1, rel_heights=c(0.1, 2))

pp_B <- plot_grid(title_1_3, p_B, ncol=1, rel_heights=c(0.1, 2))

legend_A <- get_legend(p_A1 + theme(legend.box.margin = margin(0, 0, 0, 12)))

legend_B <- get_legend(p_B1 + theme(legend.box.margin = margin(0, 0, 0, 12)))

fig1_1 <- plot_grid(pp_A, legend_A, rel_widths = c(3, 0.5))

fig1_2 <- plot_grid(pp_B, legend_B, rel_widths = c(3, 0.5))

fig1 <- plot_grid(fig1_1, fig1_2, nrow = 2, rel_widths = c(3, 0.5))

ggsave("Figure 1.pdf", fig1, width = 12, height = 15)
```

(Figure 1 at the end)

In ridgeline plot(A, B, C, D), all 4 groups(nonsmoker female, nonsmoker male, ex-smoker male, current-smoker male) have overall upregulation of protein expression level of DNA replication, DNA repair pathway. but no significant difference was found between these groups. 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.

In heatmap(E, F, G, H), nonsmoker female, nonsmoker male has tendency that more 'clear' upregulation of RNA expression level of DNA replication and DNA repair pathway revealed with TP53 mutation and ex-smoker also has this tendency. But current-smoker male wasn't found this tendency because these groups have few number of patients. I think my hypothesis described in above ridgeline plot consequence also can be applied in this heatmap 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 mechanism. Generally, high phosphorylation of DNA repair can activate DNA repair mechanism. It means TP53 mutation can activate DNA repair.

Taken together, 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 groups have upregulation of protein expression level of DNA replication, DNA repair pathway and TP53 mutation can upregulate abnormal DNA replication protein by losing cell cycle regulation ability and DNA repair protein upregulated to repair mutations. My hypothesis is just an suggestion based on original paper and general knowledge but quite fit with Figure 1 consequence.

3.2 Figure 2

As mentioned in introduction and like figure 1, figure 2 expresses RNA expression level of genes related to DNA replication, DNA repair pathways in nonsmoker female and other groups using ridgeline plot and heatmap. For this, 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 heatmap using geom_tile function in ggplot package and modified these plots by median Log2 T/N value. In heatmap, I made a side bar using geom_xsidetile function in ggside package to express whether TP53 mutation is or not. And then combined them using plot_grid function in cowplot package.

p_C1~C4 are ridgeline plot.

```
p_C1 <-
  RNA_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 = "Nonsmoker Female",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")

p_C2 <-
  RNA_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 = "Nonsmoker Male",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")

p_C3 <-
  RNA_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 = "Ex-smoker Male",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")

p_C4 <-
  RNA_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 = "Current-smoker Male",
       x = "Log2 T/N value",
       y = "DNA replication, repair genes",
       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 = 12),
        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") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y")
```

p_D1~D4 are heatmap.

```
p_D1 <-
  RNA_data %>%
  filter(Gender == "Female" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Nonsmoker Female",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")

p_D2 <-
  RNA_data %>%
  filter(Gender == "Male" & `Smoking Status` == "Nonsmoke" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Nonsmoker Male",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")

p_D3 <-
  RNA_data %>%
  filter(Gender == "Male" & `Smoking Status` == "Ex-smoker" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Ex-smoker Male",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")

p_D4 <-
  RNA_data %>%
  filter(Gender == "Male" & `Smoking Status` == "Current_Smoker" & !is.na(Log2TN)) %>%
  ggplot(aes(x = reorder(ID, Log2TN, FUN = median), y = Gene, fill = Log2TN)) +
  geom_tile() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5),
        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 = 0, xfill = TP53)) +
  scale_xfill_manual(values = c("Grey", "Black")) +
  labs(title = "Current-smoker Male",
       x = "Patient ID",
       y = "DNA replication, repair genes",
       xfill = "TP53 \nmutation",
       fill = "Log2 \n T/N") +
  facet_grid(pathway~., switch="x", scales = "free_y", space = "free_y") +
  ggside(collapse = "x")
```

Let's combine two type plots using plot_grid function and save these plots using ggsave function for better visualization.

```r
title_2_1 <-
  ggdraw() +
  draw_label("Figure 2 - Expression level of RNA by gender and smoking status",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

title_2_2 <-
  ggdraw() +
  draw_label("ridgeline plot",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

title_2_3 <-
  ggdraw() +
  draw_label("heatmap",
             fontface = 'bold',
             size = 13,
             vjust = 0.8)

p_C <- plot_grid(p_C1 + theme(legend.position="none"),
               p_C2 + theme(legend.position="none"),
               p_C3 + theme(legend.position="none"),
               p_C4 + theme(legend.position="none"),
               labels = c("A", "B", "C", "D"),
               label_size = 10)

p_D <- plot_grid(p_D1 + theme(legend.position="none"),
               p_D2 + theme(legend.position="none"),
               p_D3 + theme(legend.position="none"),
               p_D4 + theme(legend.position="none"),
               labels = c("E", "F", "G", "H"),
               label_size = 10)

pp_2 <- plot_grid(title_2_2, p_C, ncol=1, rel_heights=c(0.1, 2))

pp_C <- plot_grid(title_2_1, pp_2, ncol=1, rel_heights=c(0.1, 2))

pp_D <- plot_grid(title_2_3, p_D, ncol=1, rel_heights=c(0.1, 2))

legend_C <- get_legend(p_C1 + theme(legend.box.margin = margin(0, 0, 0, 12)))

legend_D <- get_legend(p_D1 + theme(legend.box.margin = margin(0, 0, 0, 12)))

fig2_1 <- plot_grid(pp_C, legend_C, rel_widths = c(3, 0.5))

fig2_2 <- plot_grid(pp_D, legend_D, rel_widths = c(3, 0.5))

fig2 <- plot_grid(fig2_1, fig2_2, nrow = 2, rel_widths = c(3, 0.5))

ggsave("Figure 2.pdf", fig2, width = 12, height = 15)
```

(Figure 2 at the end)

In ridgeline plot(A, B, C, D), all 4 groups(nonsmoker female, nonsmoker male, ex-smoker male, current-smoker male) upregulated RNA expression level of DNA replication and DNA repair pathway but no significant difference was found between these groups. RNA expression level shows more wide-spreading distributtion pattern than protein expression level.

In heatmap(E, F, G, H), nonsmoker female, nonsmoker male has tendency that more upregulation of RNA expression level of DNA replication and DNA repair pathway revealed with TP53 mutation. But ex-smoker, current-smoker male weren't found this tendency because these groups have few number of patients. Genes like POLD2, MLH1, MBD4 shows unclear association with TP53 mutation.

4 Discussion

I visualized ridgeline, heatmap plots to show distribution and association with TP53 mutation in DNA replication, repair pathway.

In figure 1, all 4 groups has 'clear' upregulation of protein expression level of DNA replication, DNA repair pathway. I made a hypothesis that abnormal DNA replication proteins can induce increasing of mutations and upregulation of DNA repair proteins for restoring mutations 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.

In figure 2, we can also see all 4 groups has upregulation of RNA expression level of DNA replication, DNA repair pathway. But in heatmap, some genes like POLD2, MLH1, MBD4 show unclear association with TP53 mutation.

Taken together, RNA, protein expression level of DNA replication, DNA repair pathway show upregulation in ridgeline plot. But RNA expression level shows more wide-spreading distribution pattern of DNA replication, DNA repair pathway than protein expression level in ridgeline plot and protein expression level of DNA replication and DNA repair pathway in heatmap shows clear association than RNA expression level.

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.
- I wanted to show phosphorylation level of DNA replication, DNA repair protein but there's few data in supplementary data to use.

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.

Figure 1 − Expression level of protein by gender and smoking status

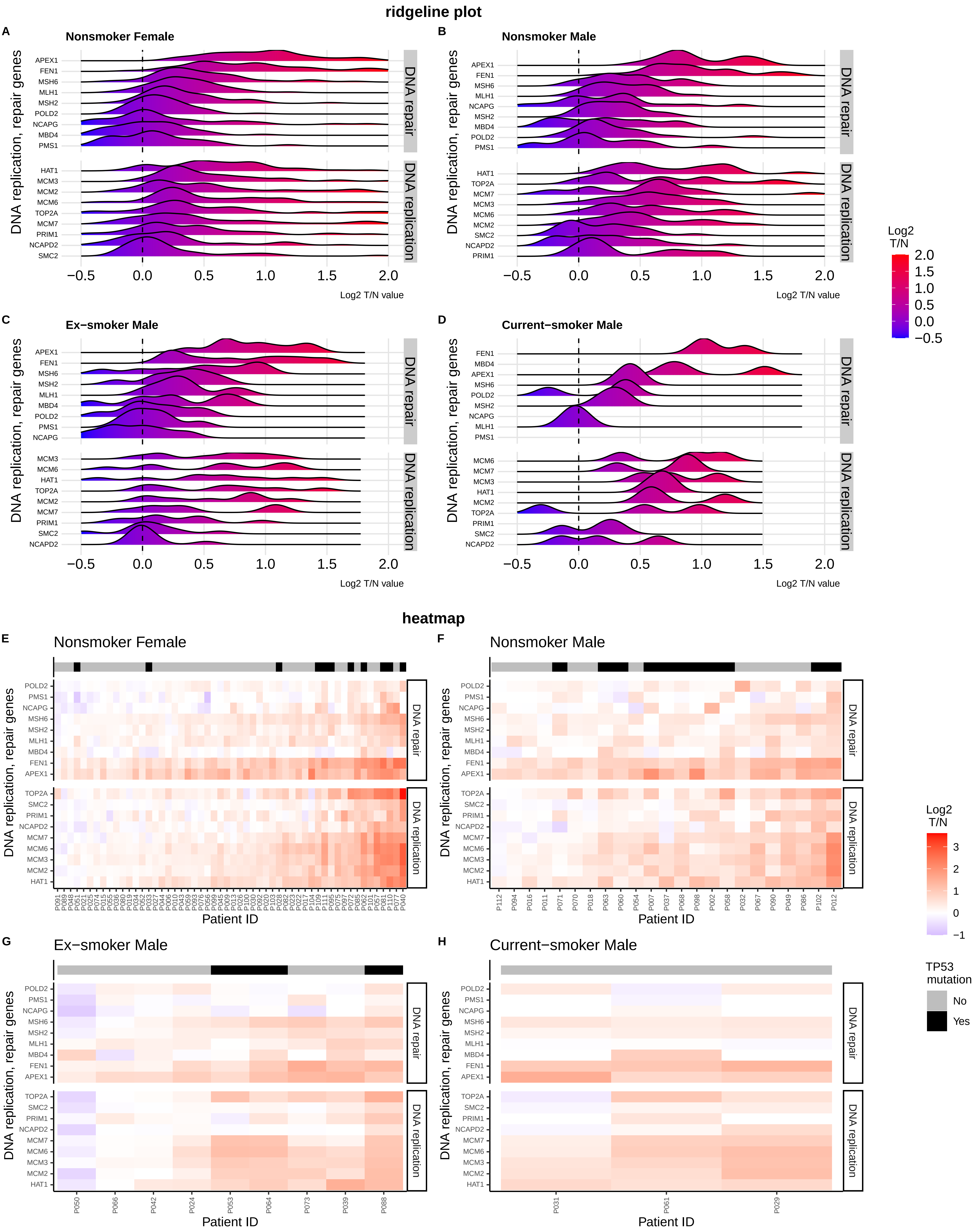
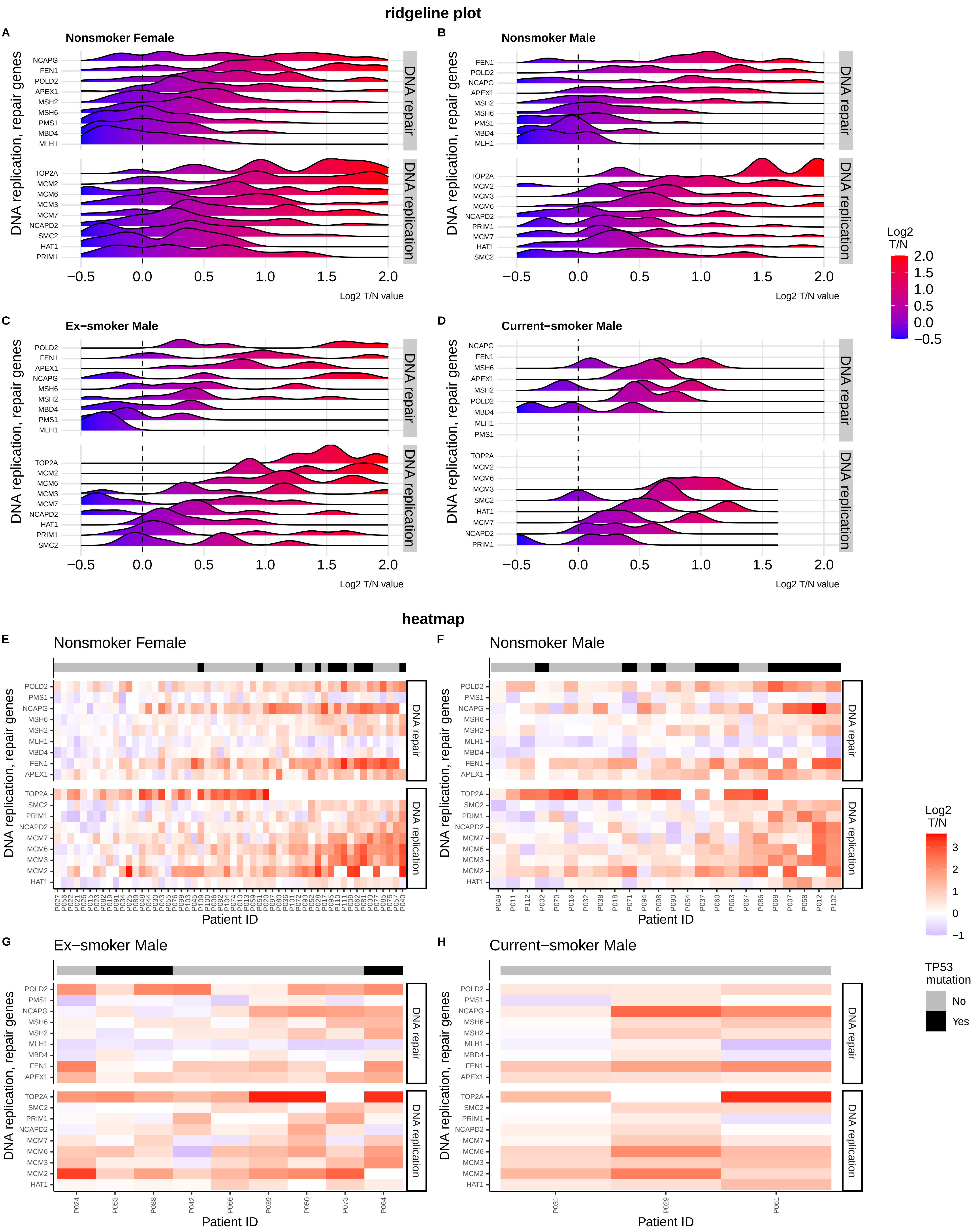


Figure 2 − Expression level of RNA by gender and smoking status


ridgeline plot
A
Nonsmoker Female
NCAPG
FEN1
POLD2
APEX1
MSH2
MSH6
PMS1
MBD4
MLH1
TOP2A
MCM2
MCM6
MCM3
MCM7
NCAPD2
SMC2
HAT1
PRIM1
DNA repair
DNA replication
DNA replication, repair genes
Log2 T/N value
B
Nonsmoker Male
FEN1
POLD2
NCAPG
APEX1
MSH2
MSH6
PMS1
MBD4
MLH1
TOP2A
MCM2
MCM3
MCM6
NCAPD2
PRIM1
MCM7
HAT1
SMC2
DNA repair
DNA replication
DNA replication, repair genes
Log2 T/N value
Log2
T/N
C
Ex−smoker Male
POLD2
FEN1
APEX1
NCAPG
MSH6
MSH2
MBD4
PMS1
MLH1
TOP2A
MCM2
MCM6
MCM3
MCM7
NCAPD2
HAT1
PRIM1
SMC2
DNA repair
DNA replication
DNA replication, repair genes
Log2 T/N value
D
Current−smoker Male
NCAPG
FEN1
MSH6
APEX1
MSH2
POLD2
MBD4
MLH1
PMS1
TOP2A
MCM2
MCM6
MCM3
SMC2
HAT1
MCM7
NCAPD2
PRIM1
DNA repair
DNA replication
DNA replication, repair genes
Log2 T/N value
heatmap
E
Nonsmoker Female
POLD2
PMS1
NCAPG
MSH6
MSH2
MLH1
MBD4
FEN1
APEX1
TOP2A
SMC2
PRIM1
NCAPD2
MCM7
MCM6
MCM3
MCM2
HAT1
DNA repair
DNA replication
DNA replication, repair genes
P027 P056 P022 P021 P026 P015 P023 P082 P019 P091 P034 P025 P089 P048 P044 P030 P043 P055 P076 P099 P103 P045 P109 P100 P006 P092 P104 P074 P010 P013 P059 P051 P020 P097 P080 P036 P101 P072 P093 P052 P028 P017 P095 P110 P111 P009 P062 P081 P033 P077 P085 P075 P057 P040
Patient ID
F
Nonsmoker Male
POLD2
PMS1
NCAPG
MSH6
MSH2
MLH1
MBD4
FEN1
APEX1
TOP2A
SMC2
PRIM1
NCAPD2
MCM7
MCM6
MCM3
MCM2
HAT1
DNA repair
DNA replication
DNA replication, repair genes
P049 P011 P112 P002 P070 P016 P032 P038 P018 P071 P094 P098 P090 P054 P037 P060 P063 P067 P086 P068 P007 P058 P012 P102
Patient ID
Log2
T/N
G
Ex−smoker Male
POLD2
PMS1
NCAPG
MSH6
MSH2
MLH1
MBD4
FEN1
APEX1
TOP2A
SMC2
PRIM1
NCAPD2
MCM7
MCM6
MCM3
MCM2
HAT1
DNA repair
DNA replication
DNA replication, repair genes
P024 P053 P088 P042 P066 P039 P050 P073 P064
Patient ID
H
Current−smoker Male
POLD2
PMS1
NCAPG
MSH6
MSH2
MLH1
MBD4
FEN1
APEX1
TOP2A
SMC2
PRIM1
NCAPD2
MCM7
MCM6
MCM3
MCM2
HAT1
DNA repair
DNA replication
DNA replication, repair genes
P031 P029 P061
Patient ID
TP53
mutation
No
Yes