Portfolio

Signaling pathways caused by exogenous carcinogens in vitro associated with oncogenesis

2019250140 Jeon in a

1. Introduction

1. 1. Background

Lung cancer remains the most common malignancy and the leading cause of cancer mortality worldwide. Lung cancer is known to be mainly caused by direct exposure to cigarettes. But LUAD (lung adenocarcinoma) in East Asia, especially in Taiwan, is characterized by a high rate of never-smokers, early onset, and predominant EGFR mutations (Chen et al., 2020). According to these analyses, APOBEC mutational signatures are frequently observed in younger females and enviromental carcinogen-like mutational signatures highly occur in older females. In addition, early onset is a distinct feature of LUAD in East Asia, especially among never-smokers. From this, we can think that LUAD is associated with genetic and environmental factors, especially in this paper, which analyzed that air pollution in Taiwan correlates with LUAD incidence in never-smokers. The carcinogen signals presented in the paper include (1) Nitrosamine-like, (2) Nitro-PAH, (3) radiation, (4) Alkylating agents, and (5) PAHs. Among these five carcinogen, I will focus on Nitrosamine, Nitro-PAH, and PAHs. Among the many components in tobacco smoke and outdoor and indoor air pollution are polycyclic aromatic hydrocarbons (PAHs), which are considered to be the most important carcinogens in these complex mixtures. Metabolism of PAHs leads to the formation of the active carcinogens. These reactive metabolites produce DNA adducts, resulting in DNA mutations, alteration of gene expression profiles, and tumorigenesis(Moorthy et al., 2015). Nitrosamines are formed by a reaction between nitrates or nitrites and certain amines. Nitrosamines and/or their precursors can be found in diverse consumer products such as processed meats, alcoholic beverages, cosmetics, cigarette smoke and also be formed in the mouth or stomach if the food contains nitrosamine precursors. Nitrosamines are considered to be strong carcinogens that may produce cancer in diverse organs and tissues including lung, brain, liver, kidney, bladder, stomach, esophagus, and nasal sinus(H. Robles, 2014). Nitrated polycyclic aromatic hydrocarbons (Nitro-PAHs) are derivatives of PAHs with at least one nitro-functional group (-NO2) on the aromatic ring. Nitro-PAHs are mainly generated by incomplete combustion and pyrolysis of fossil fuels and biomass. Nitro-PAHs are direct-acting mutagens and carcinogens. The mechanisms underlying some of these toxicological effects of nitro-PAHs include DNA damage, DNA adduct formation, aryl hydrocarbon receptor activation, changes in gene and protein expression, cell cycle alternations, increased levels of reactive oxygen species and pro-inflamation. Inhalation, oral ingestion and dermal contact are the main routes of nitro-PAH intake from the environment by humans and animals(Benjamin, 2017).

1. 2. Data Visualization Topic

The topic for data visualization is to plot the correlation between environmental carcinogen and enriched pathway. Particularly, focusing on the pathway that has a significant correlation with carcinogen, I will examine whether the carcinogens presented in the paper have a direct relationship with mutational metabolism.

2. Exploring Data

2. 1. Unboxing Dataset

Before drawing the plot, I loaded the packages needed to create a portfolio by using 'library()'.

```
library(readxl)
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(ggplot2)
library(dplyr)
library(corrplot)
```

```
## corrplot 0.91 loaded
```

```
library(grid)
library(ggside)
```

```
## Registered S3 method overwritten by 'ggside':
##   method from
##   +.gg   ggplot2
```

Since the supplementary table is MS Excel format, the file must be loaded using 'readxl' package. I stored the file in a folder called 'chen2020', so the sheet was loaded using the following code.

```
readxl::excel_sheets('Chen2020/1-s2.0-S0092867420307431-mmc5.xlsx')
```

```
## [1] "Description"                   "S5A.APOBEC_repair_mRNAprotein"
## [3] "S5B.Kinase-substrate pairs"    "S5C_carcinogen 1Dpath_Fig5E"
## [5] "S5D_related to Fig5F"
```

Since the sheet to be used in this file is [4] "S5C_carcinogen 1Dpath_Fig5E", the data set was loaded using 'sheet'.

```
d <- read_excel('Chen2020/1-s2.0-S0092867420307431-mmc5.xlsx', sheet = 4)
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...5
## * `` -> ...6
## * ...
```

2. 2. Manipulating Data Frame

After checking the data using 'head()', let's write 'colnames()' to check whether the column names are proper or not.

```
head(d)
```

```
## # A tibble: 6 x 18
##   ...1    ...2 ...3  PAHs   ...5   ...6  NitroPAHs ...8  ...9  Mixed ...11 ...12
##   <chr>  <dbl> <lgl> <chr>  <chr>  <chr> <chr>     <chr> <chr> <chr> <chr> <chr>
## 1 Carci~    NA NA    mRNA   Prote~ Phos~ mRNA      Prot~ Phos~ mRNA  Prot~ Phos~
## 2 Chemi~     1 NA    -0.31~ 0.124~ -0.4~ -0.14544~ -3.4~ -0.4~ 4.10~ 7.96~ -2.4~
## 3 Drug ~     2 NA    -0.15~ 0.201~ -0.4~ 4.894531~ 1.77~ -0.8~ -9.7~ 0.17~ -9.0~
## 4 Vitam~     3 NA    -0.16~ 0.114~ -0.6~ 0.247891~ 4.67~ -0.1~ -0.5~ 5.31~ 5.68~
## 5 Stero~     4 NA    -0.40~ 4.677~ NA    -0.37392~ -6.3~ NA    -8.5~ 3.31~ NA
## 6 Tyros~     5 NA    0.645~ 1.021~ 8.23~ -0.16271~ -2.4~ 4.89~ -0.1~ 0.11~ -0.1~
## # ... with 6 more variables: Nitrosamine <chr>, ...14 <chr>, ...15 <chr>,
## #   Comparison of 6 carcinogen group (p-value) <chr>, ...17 <chr>, ...18 <chr>
```

```
colnames(d)
```

```
##  [1] "...1"
##  [2] "...2"
##  [3] "...3"
##  [4] "PAHs"
##  [5] "...5"
##  [6] "...6"
##  [7] "NitroPAHs"
##  [8] "...8"
##  [9] "...9"
## [10] "Mixed"
## [11] "...11"
## [12] "...12"
## [13] "Nitrosamine"
## [14] "...14"
## [15] "...15"
## [16] "Comparison of 6 carcinogen group (p-value)"
## [17] "...17"
## [18] "...18"
```

Since the column names are not organized, I set them as I wanted. In addition, row 1 and column 2, 3 are unnecessary, so I deleted them. I changed the columns except for the first column to numeric for facilitate processing, and rearranged the columns into a 'type'.

```
d <- d[c(2:54), c(1, 4:18)]

d <- d %>%
  rename(pathway = ...1,
         PAHs_mRNA = PAHs,
         PAHs_prot = ...5,
         PAHs_phos = ...6,
         NitroPAHs_mRNA = NitroPAHs,
```

```
        NitroPAHs_prot = ...8,
        NitroPAHs_phos = ...9,
        Mixed_mRNA = Mixed,
        Mixed_prot = ...11,
        Mixed_phos = ...12,
        Nitrosamine_mRNA = Nitrosamine,
        Nitrosamine_prot = ...14,
        Nitrosamine_phos = ...15,
        mRNA = "Comparison of 6 carcinogen group (p-value)",
        prot = ...17,
        phos = ...18) %>%
  mutate(PAHs_mRNA = as.numeric(PAHs_mRNA),
        PAHs_prot = as.numeric(PAHs_prot),
        PAHs_phos = as.numeric(PAHs_phos),
        NitroPAHs_mRNA = as.numeric(NitroPAHs_mRNA),
        NitroPAHs_prot = as.numeric(NitroPAHs_prot),
        NitroPAHs_phos = as.numeric(NitroPAHs_phos),
        Mixed_mRNA = as.numeric(Mixed_mRNA),
        Mixed_prot = as.numeric(Mixed_prot),
        Mixed_phos = as.numeric(Mixed_phos),
        Nitrosamine_mRNA = as.numeric(Nitrosamine_mRNA),
        Nitrosamine_prot = as.numeric(Nitrosamine_prot),
        Nitrosamine_phos = as.numeric(Nitrosamine_phos),
        mRNA = as.numeric(mRNA),
        prot = as.numeric(prot),
        phos = as.numeric(phos)) %>%
  gather(key = "type", value = "log", -c("pathway", "mRNA", "prot", "phos")) %>%
  gather("mRNA", "prot", "phos", key = "p_value", value = "P")
```

```
## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA

## Warning in mask$eval_all_mutate(quo):          NA
```

Let's check again to see whether it is changed properly.

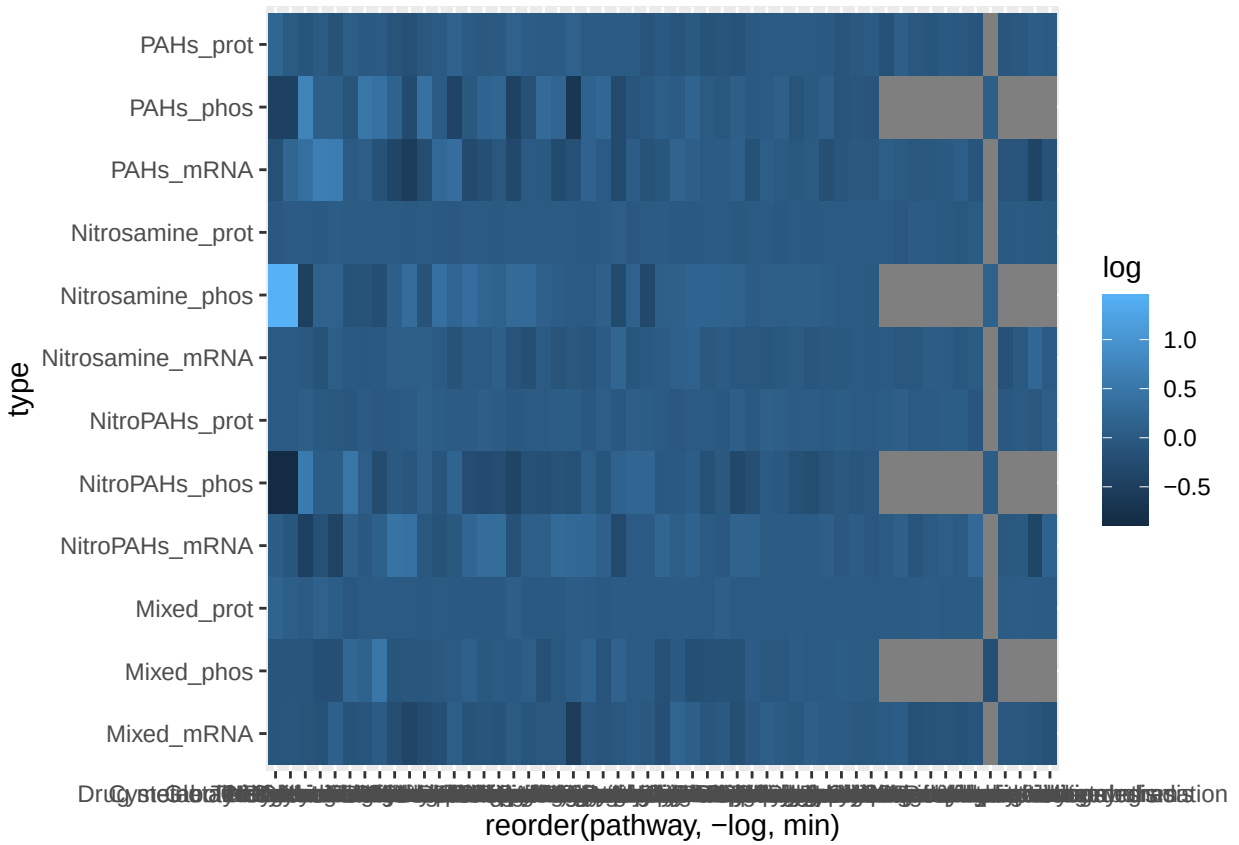
```
head(d)
```

```
## # A tibble: 6 x 5
##   pathway                           type         log p_value      P
##   <chr>                             <chr>      <dbl> <chr>    <dbl>
## 1 Chemical carcinogenesis           PAHs_mRNA -0.311 mRNA    0.231
## 2 Drug metabolism - cytochrome P450 PAHs_mRNA -0.158 mRNA    0.256
## 3 Vitamin B6 metabolism             PAHs_mRNA -0.162 mRNA    0.660
## 4 Steroid hormone biosynthesis      PAHs_mRNA -0.408 mRNA    0.286
## 5 Tyrosine metabolism               PAHs_mRNA  0.646 mRNA    0.674
## 6 Renin-angiotensin system          PAHs_mRNA  0.146 mRNA    0.0645
```

3. Data Visualization

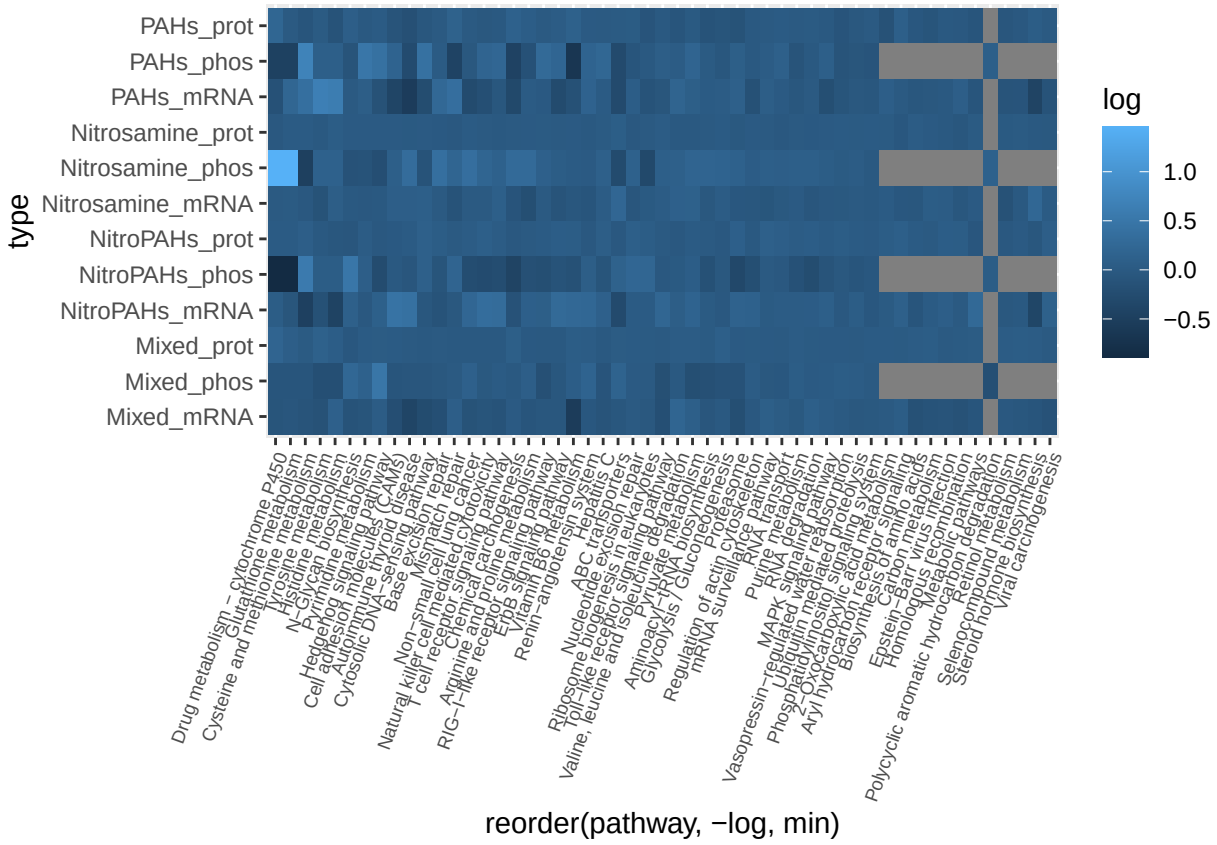
At first, I will draw Heatmap using the values of 'relative log2T/N', classified into mRNA, protein, and phosphate to see whether each carcinogens have significant effects on the enriched pathway. First, set the x-axis to 'pathway' and y-axis to 'log2T/N' values. Then write the code using 'geom_tile()' because I will draw a heatmap. At this time, the degree of log value will be compared, so write 'aes(fill=log)' in 'geom_tile().

```
d %>% ggplot(aes(x = reorder(pathway, -log, min), y = type)) + geom_tile(aes(fill = log))
```



Since 'pathway' letters on the x-axis overlap, let's adjust the angle and size so that the letters don't overlap.

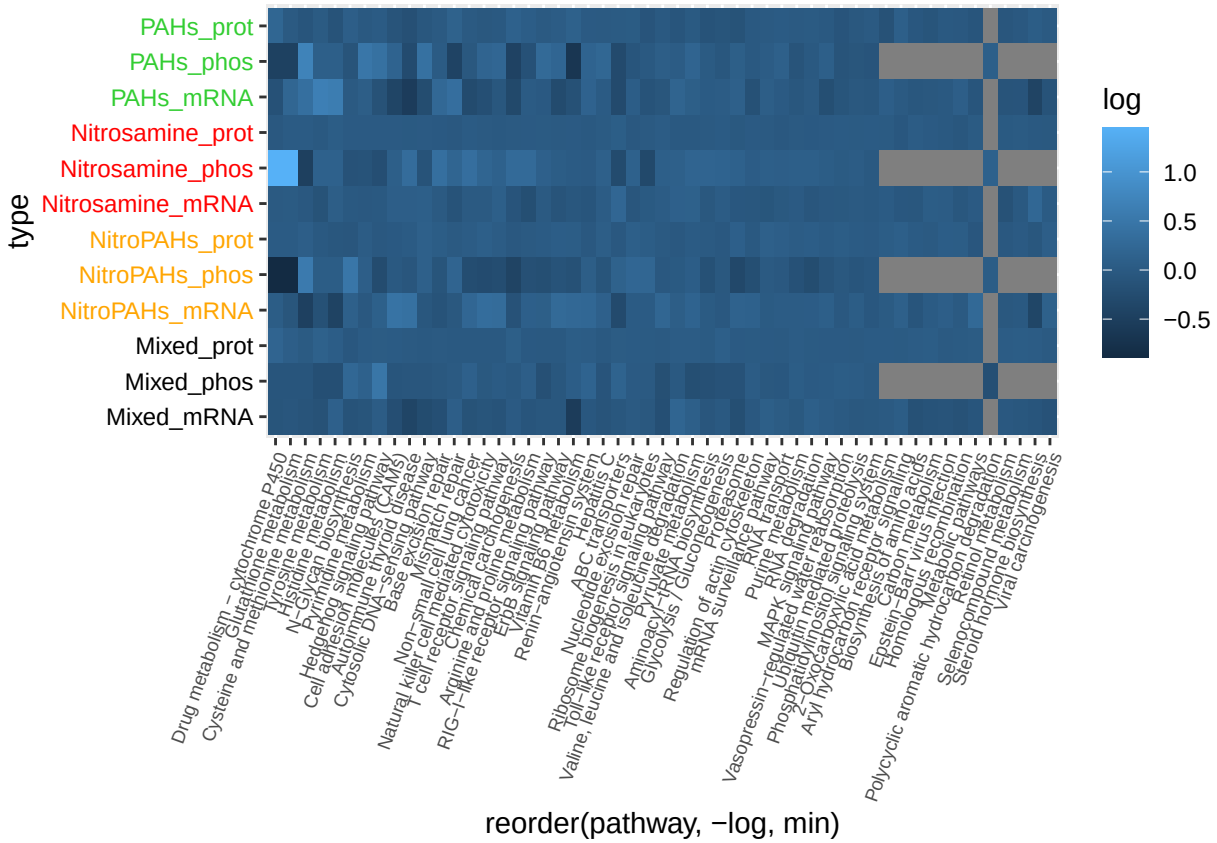
```
d %>% ggplot(aes(x = reorder(pathway, -log, min), y = type)) + geom_tile(aes(fill = log)) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = 7))
```



Assign the color for good visibility. So that the carcinogen on the y-axis can be easily seen for each type. Also adjust the font size.

```
d %>% ggplot(aes(x = reorder(pathway, -log, min), y = type)) + geom_tile(aes(fill = log)) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = 7),
        axis.text.y = element_text(hjust = 1, size = 9, colour = c("black", "black", "black", "orange",
```

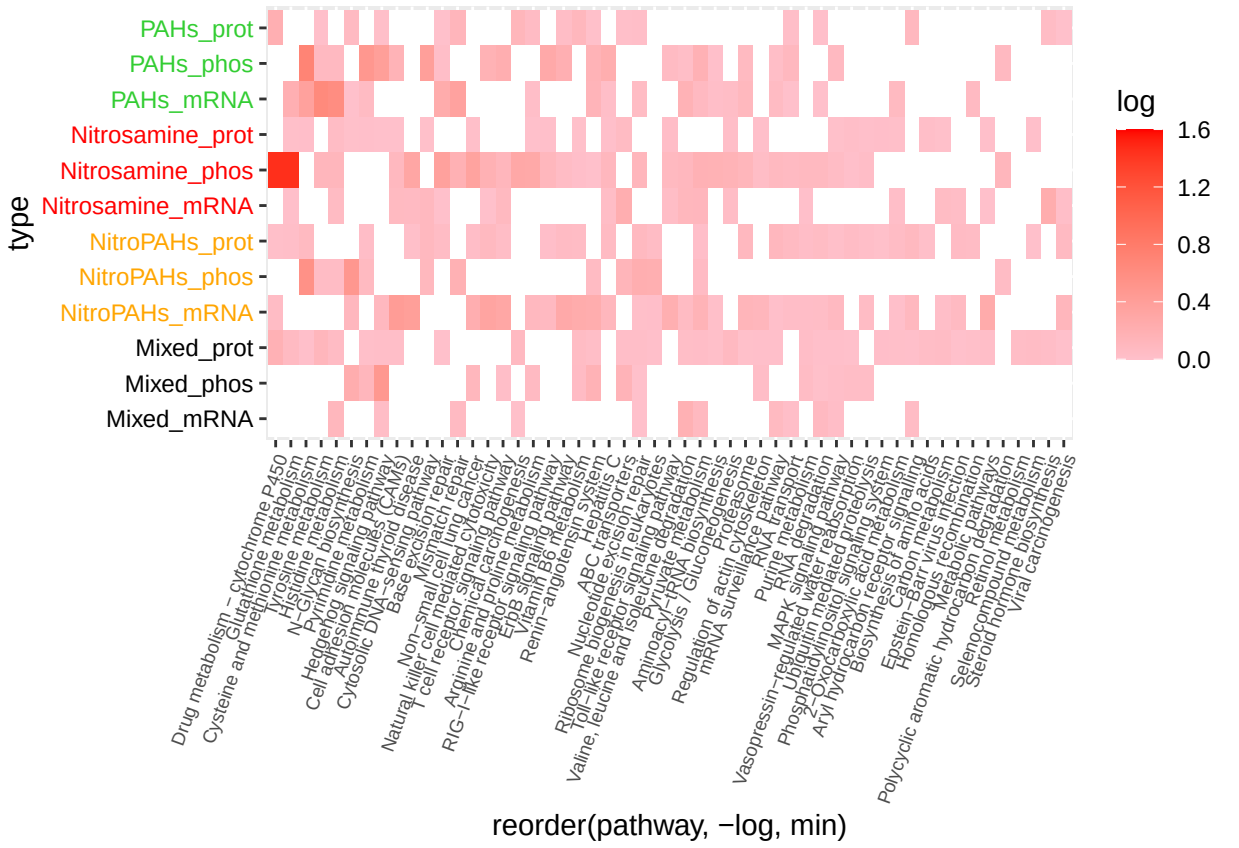
```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```



Change the color of heatmap to see the values easily, so the low-value was set to pink and the high-value to red. In this case, when the log value is 0, the value of Tumor and NAT is the same, and when the log value is negative, the NAT value is higher than Tumor. So I thought log value that below 0 are unnecessary informations to confirm the relationship between carcinogen and pathway. I set only positive cases to the range using 'limits'. Values below 0 are marked in white and the missing value 'NA' was also marked in white.

```
d %>% ggplot(aes(x = reorder(pathway, -log, min), y = type)) + geom_tile(aes(fill = log)) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = 7),
        axis.text.y = element_text(hjust = 1, size = 9, colour = c("black", "black", "black", "orange",
  scale_fill_gradient2(midpoint = 0, mid = "pink", high = "red", limits = c(0, 1.6), na.value = "white"
```

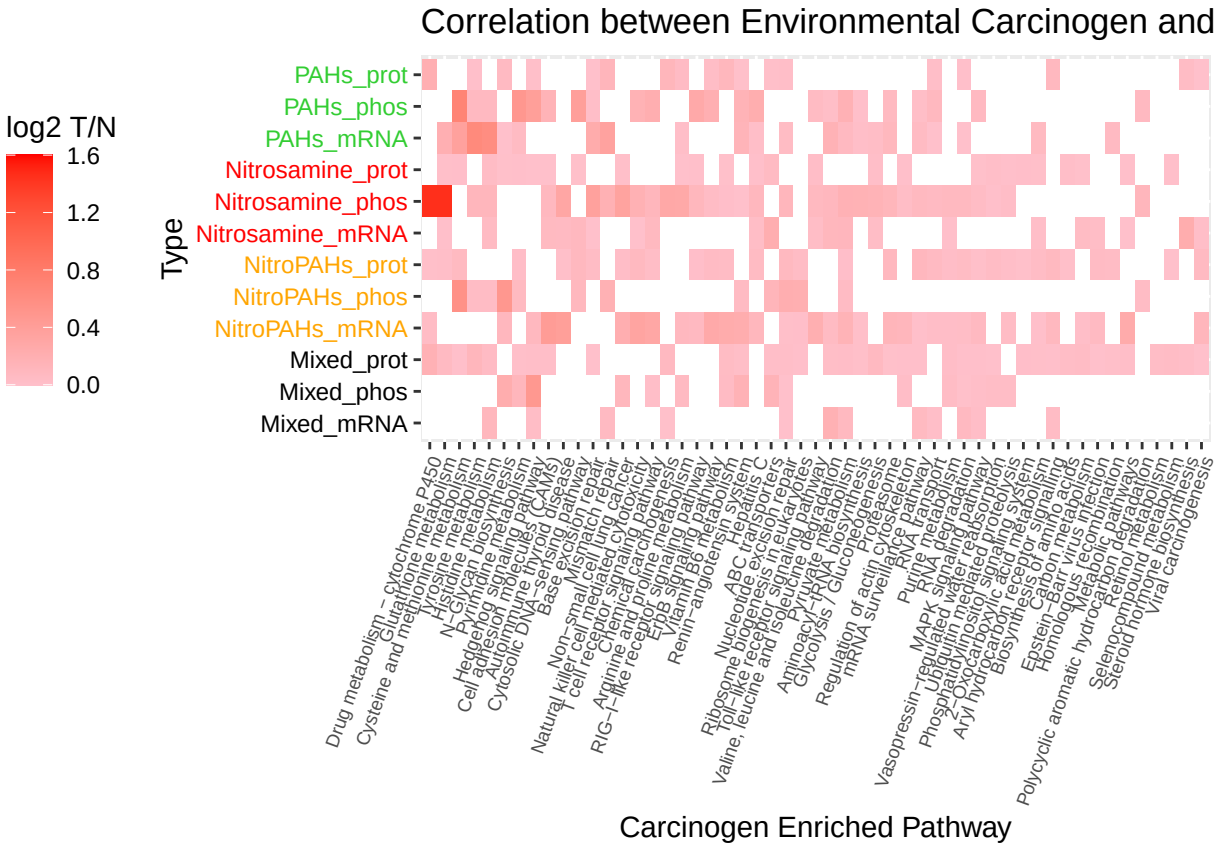
```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```



Finally, name the title of the plot, the x- and y-axes, and the legend. Also change the position of legend to left.

```
d %>% ggplot(aes(x = reorder(pathway, -log, min), y = type)) + geom_tile(aes(fill = log)) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = 7),
        axis.text.y = element_text(hjust = 1, size = 9, colour = c("black", "black", "black", "orange",
  scale_fill_gradient2(midpoint = 0, mid = "pink", high = "red", limits = c(0, 1.6), na.value = "white"
  labs(title = "Correlation between Environmental Carcinogen and Enriched Pathway",
       cex.main = 8,
       x = "Carcinogen Enriched Pathway",
       y = "Type",
       fill = "log2 T/N") +
  theme(legend.position="left")
```

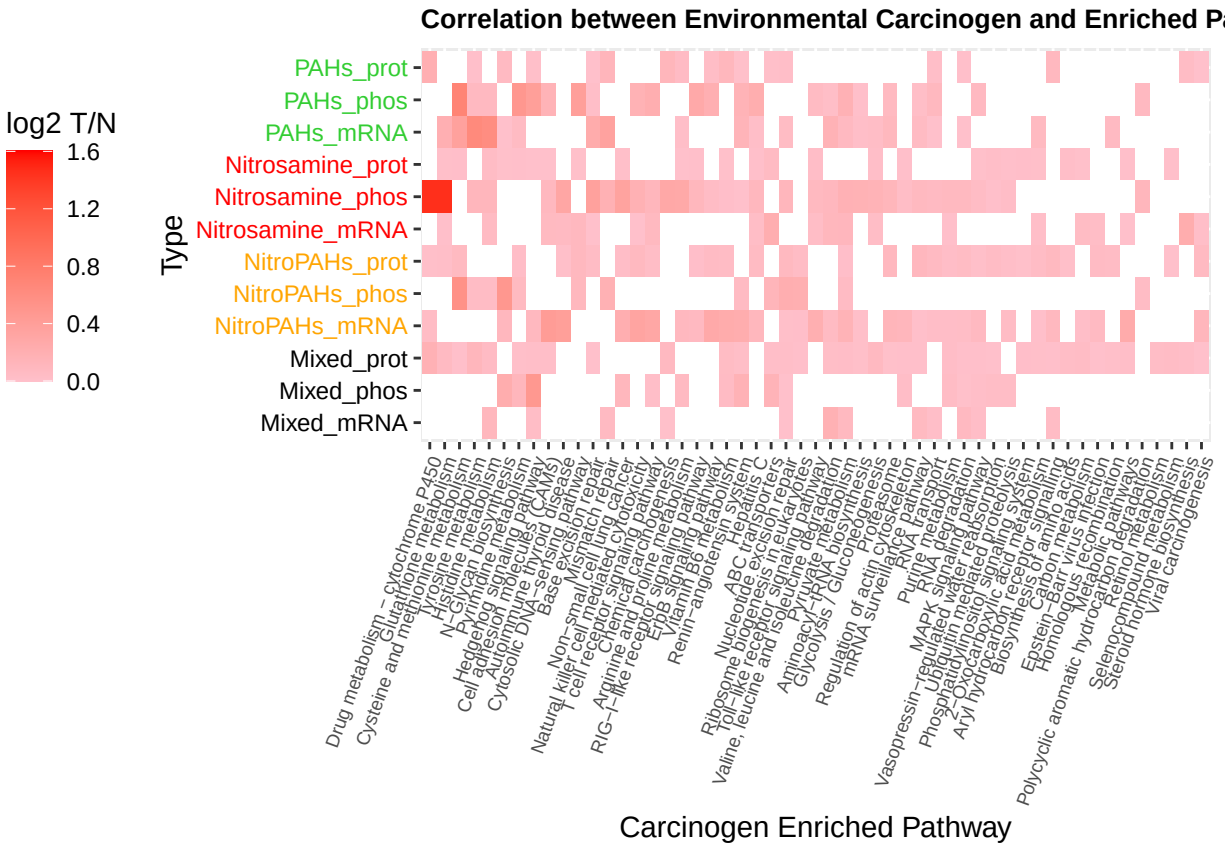
```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```



Since the title doesn't appear completely, adjust the size of the title.

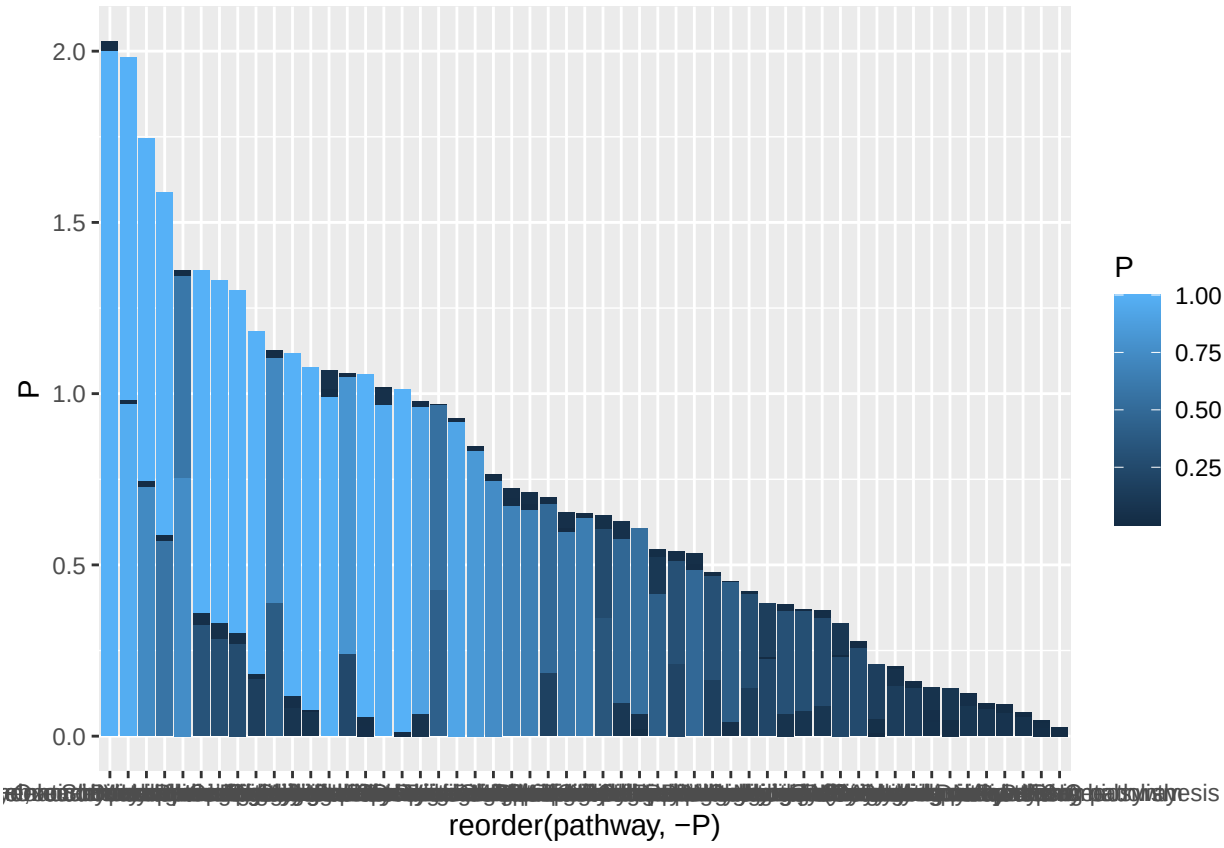
```
d %>% ggplot(aes(x = reorder(pathway, -log, min), y = type)) + geom_tile(aes(fill = log)) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = 7),
        axis.text.y = element_text(hjust = 1, size = 9, colour = c("black", "black", "black", "orange",
  scale_fill_gradient2(midpoint = 0, mid = "pink", high = "red", limits = c(0, 1.6), na.value = "white"
  labs(title = "Correlation between Environmental Carcinogen and Enriched Pathway",
       cex.main = 8,
       x = "Carcinogen Enriched Pathway",
       y = "Type",
       fill = "log2 T/N") +
  theme(legend.position="left",
        plot.title = element_text(size=10, face="bold"))
```

```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```



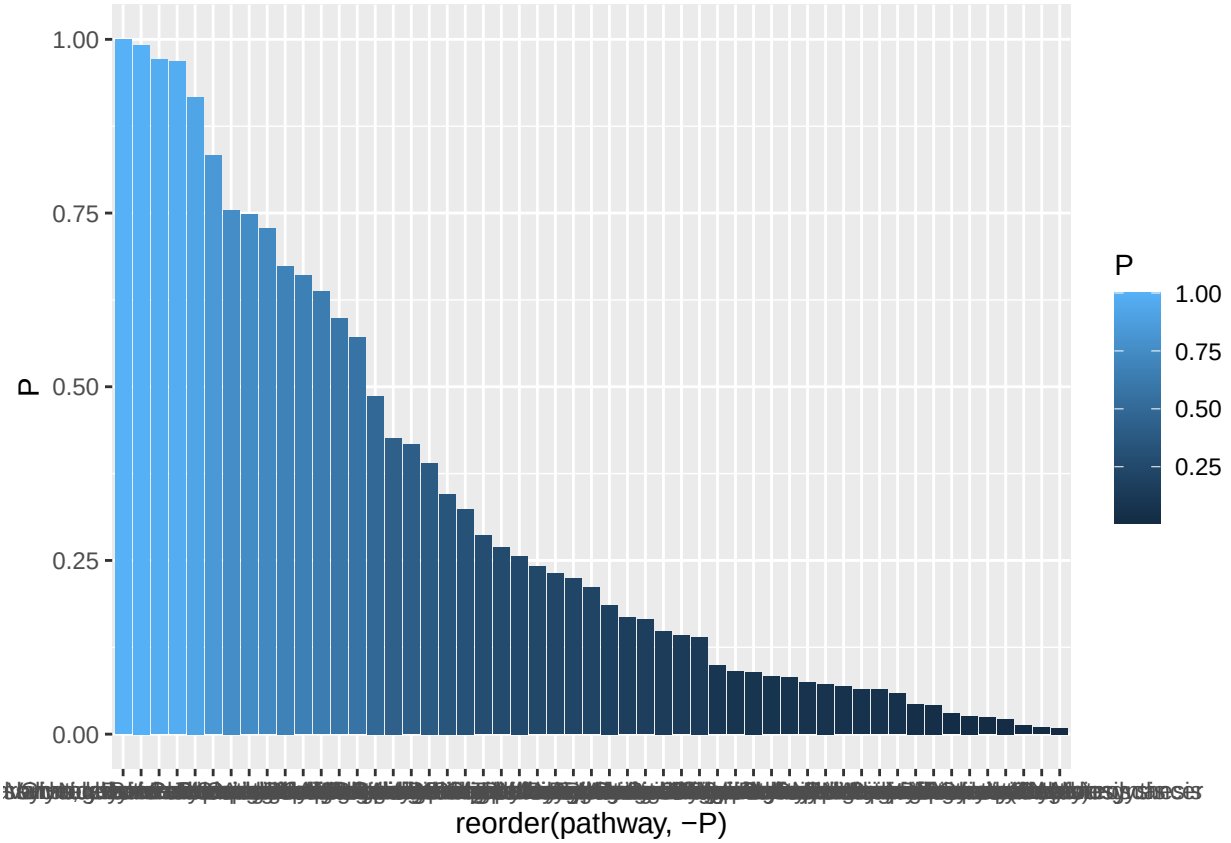
Next, I will draw a plot to compare p-values to see if the hypothesis that carcinogen affects to the enriched pathway is significant. First of all, in data 'd', columns were regathered for each type(PAHs, NitroPAHs, Nitrosamine, Mixed), so the pathways overlapped. Therefore, using 'filter()', one type was selected. Draw a barplot by setting the 'pathway' as the x-axis and 'P' as the y-axis. Sort according to the p-value to increase readability.

```
d %>% filter(type == "PAHs_mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P)) +
  geom_bar(stat = 'identity')
```



mRNA, protein, and phosphate are all present in p_value, so the data are superimposed and expressed. Let's draw a plot by designating only mRNA separately to create a proper plot first.

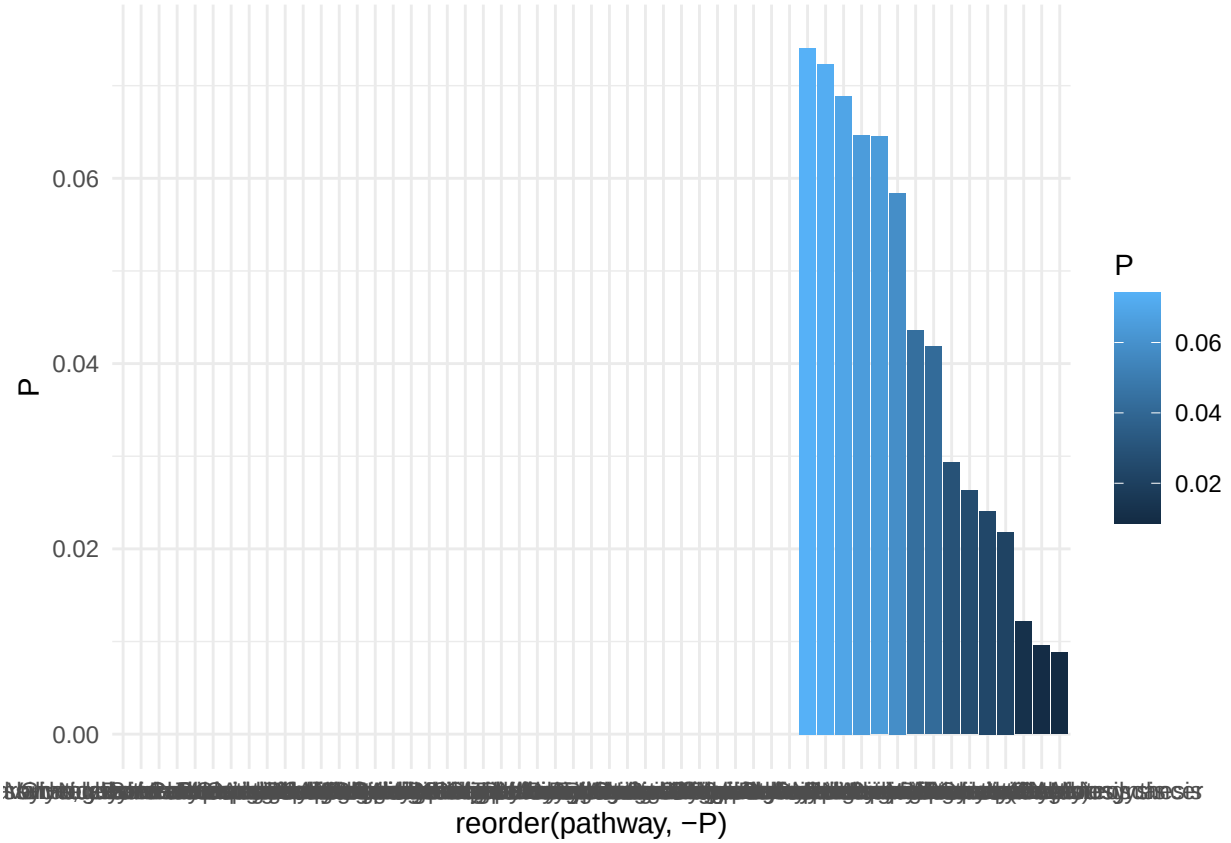
```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P)
  geom_bar(stat = 'identity')
```



Set theme as 'minimal' to eliminate the background color, and specify the range of p-value as 0 to 0.75.

```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P))
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075)
```

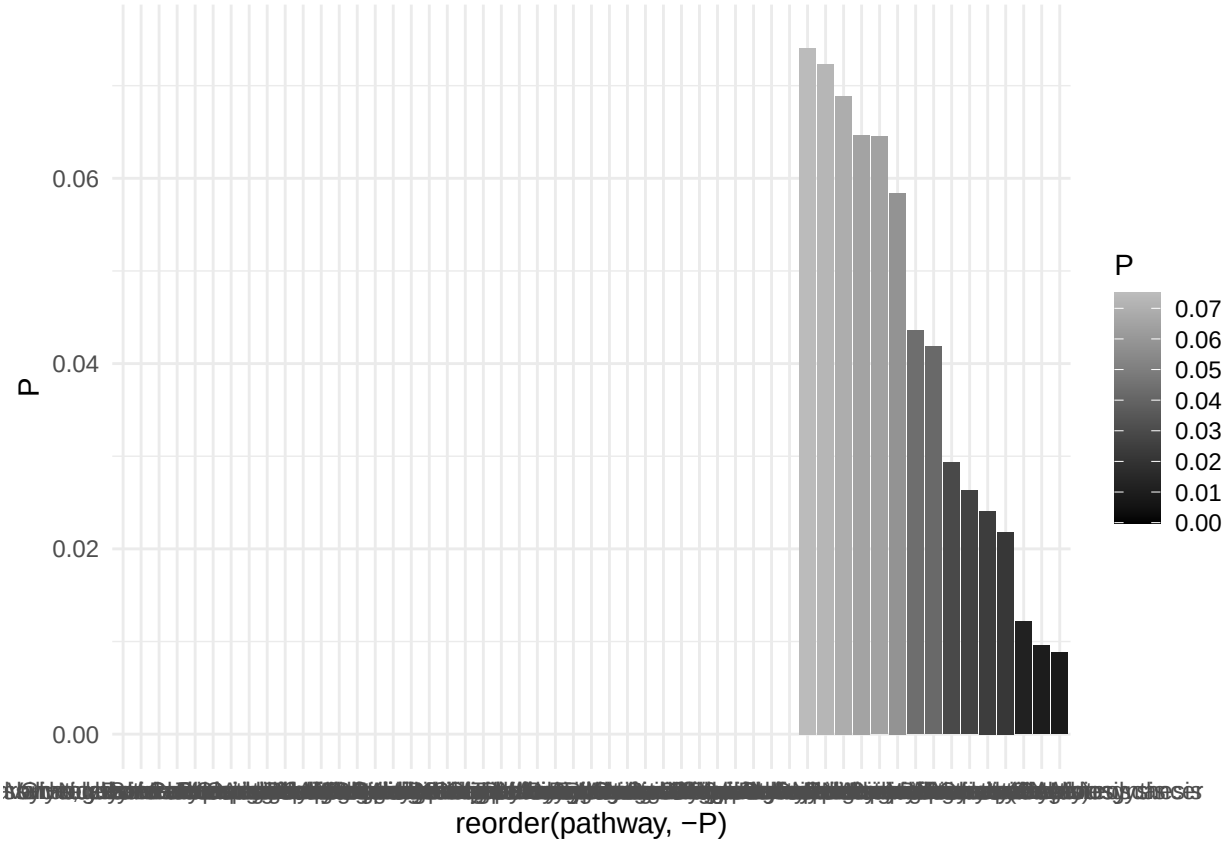
```
## Warning: Removed 38 rows containing missing values (position_stack).
```



Change the color of barplot to see the values easily. In this case, the lower p-value is, the more significant it is, so the low-value was set to black and the high-value to grey. The missing value 'NA' was marked in white.

```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P)
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075) +
  scale_fill_gradient2(midpoint = 0.05, low = "black", mid = "grey50", high = "white", limits = c(0, 0.
```

```
## Warning: Removed 38 rows containing missing values (position_stack).
```

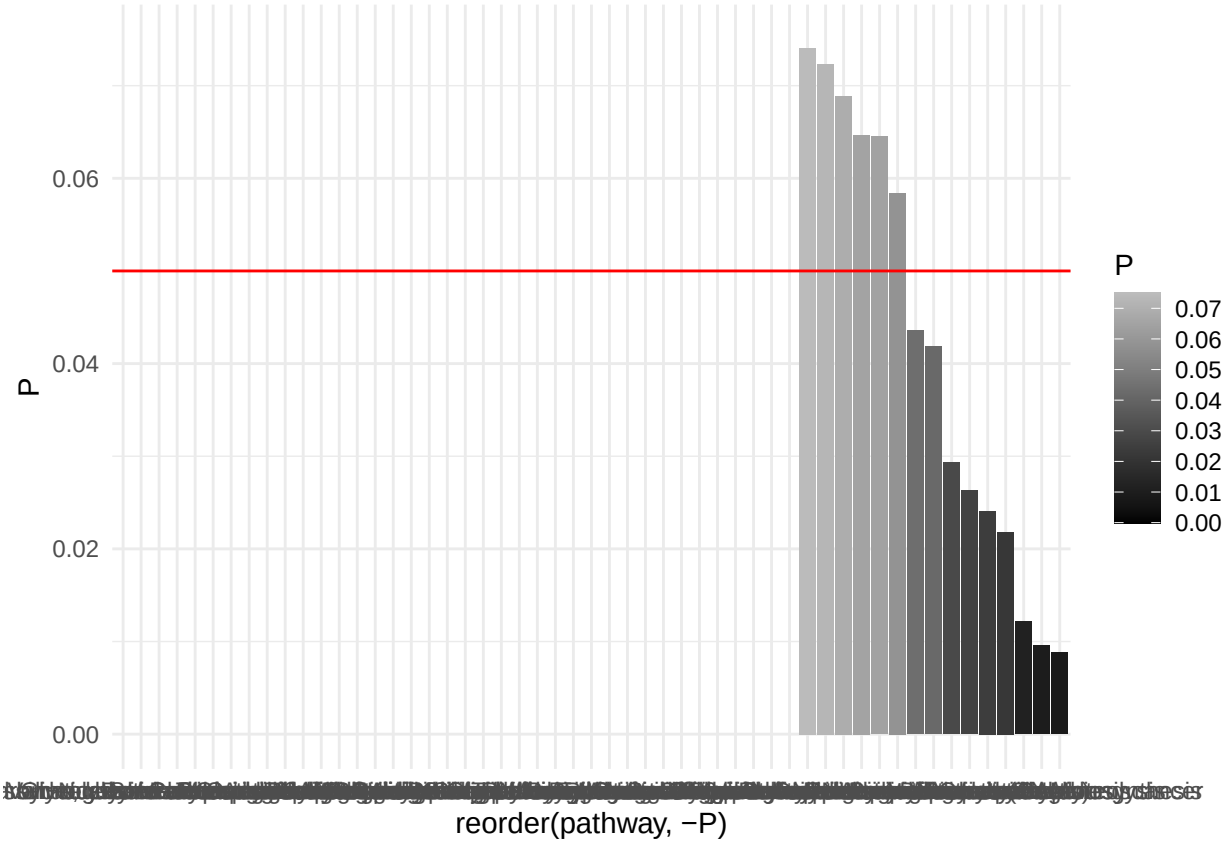


As our statistical hypothesis will, by definition, state some property of the distribution, the null hypothesis is the default hypothesis under which that property does not exist. This hypothesis might specify the probability distribution of X precisely, or it might only specify that it belongs to some class of distributions. The p-value is used in the context of null hypothesis testing in order to quantify the statistical significance of a result, the result being the observed value of the chosen statistic T. The lower the p-value is, the lower the probability of getting that result if the null hypothesis were true. A result is said to be statistically significant if it allows us to reject the null hypothesis. All other things being equal, smaller p-values are taken as stronger evidence against the null hypothesis.

For typical analysis, using the standard = 0.05 cutoff, the null hypothesis is rejected when p <= 0.05 and not rejected when p > 0.05. So, I marked p = 0.05 as a reference point for significance with a red line.

```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P))
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075) +
  scale_fill_gradient2(midpoint = 0.05, low = "black", mid = "grey50", high = "white", limits = c(0, 0.0
  geom_hline(yintercept = 0.05, col = "red")
```

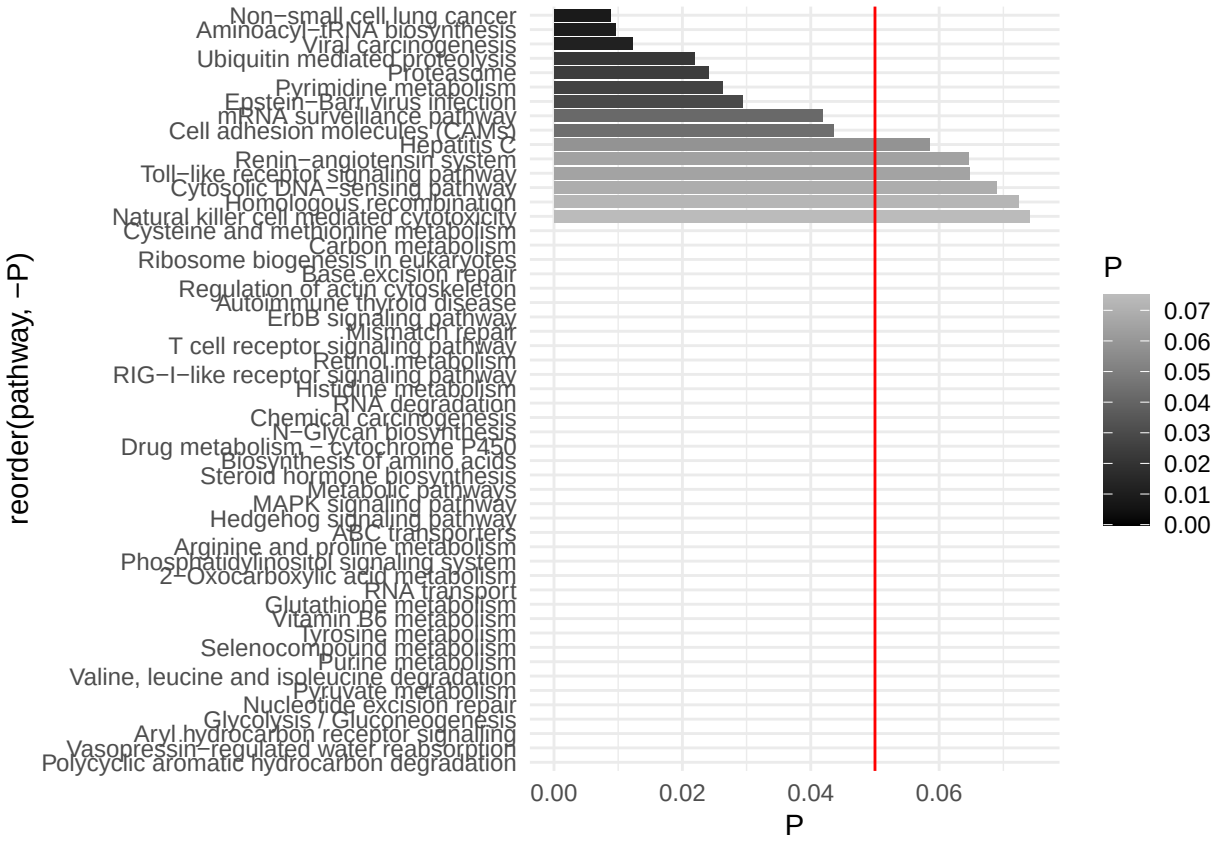
```
## Warning: Removed 38 rows containing missing values (position_stack).
```



To see the plot easliy, switch the positions of x-axis and y-axis.

```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P))
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075) +
  scale_fill_gradient2(midpoint = 0.05, low = "black", mid = "grey50", high = "white", limits = c(0, 0.0
  geom_hline(yintercept = 0.05, col = "red") +
  coord_flip()
```

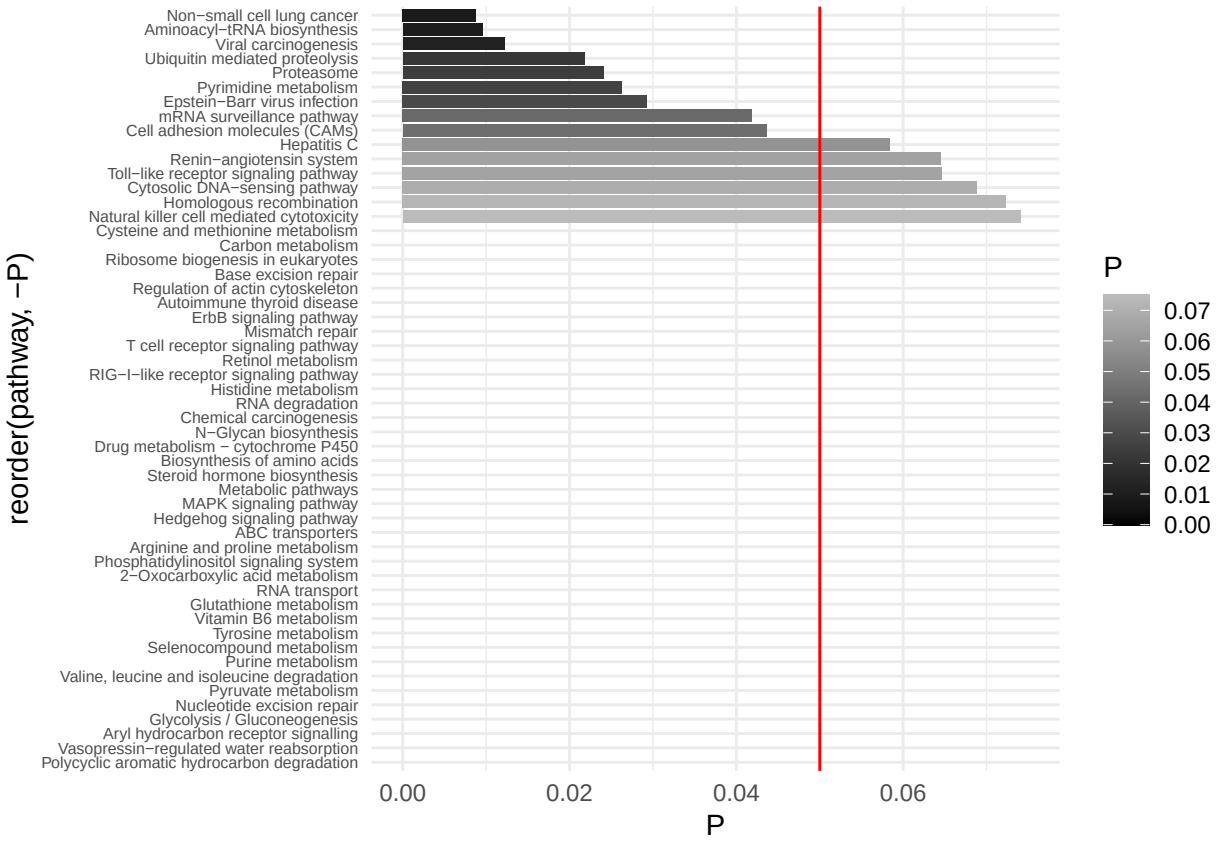
```
## Warning: Removed 38 rows containing missing values (position_stack).
```



Since 'pathway' letters on the y-axis overlap, let's adjust the size so that the letters don't overlap.

```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P))
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075) +
  scale_fill_gradient2(midpoint = 0.05, low = "black", mid = "grey50", high = "white", limits = c(0, 0.0
  geom_hline(yintercept = 0.05, col = "red") +
  coord_flip() +
  theme(axis.text.y = element_text(hjust = 1, size = 6))
```

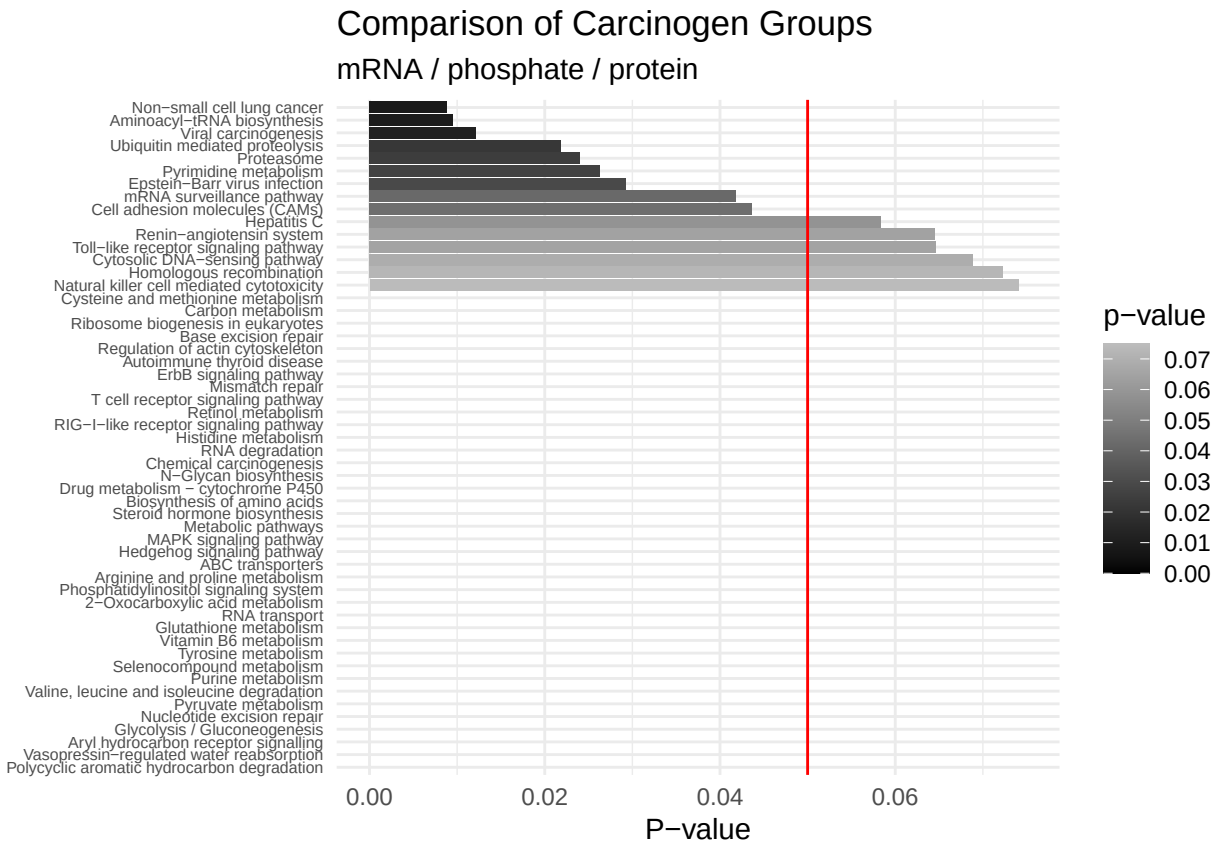
```
## Warning: Removed 38 rows containing missing values (position_stack).
```



Name the title of the plot, the x- and y-axes, and the legend. Also change the position of legend to right.

```
d %>% filter(type == "PAHs_mRNA" & p_value == "mRNA") %>% ggplot(aes(reorder(pathway, -P), P, fill = P))
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075) +
  scale_fill_gradient2(midpoint = 0.05, low = "black", mid = "grey50", high = "white", limits = c(0, 0.0
  geom_hline(yintercept = 0.05, col = "red") +
  coord_flip() +
  theme(axis.text.y = element_text(hjust = 1, size = 6)) +
  labs(title = "Comparison of Carcinogen Groups",
       subtitle = "mRNA / phosphate / protein",
       y = "P-value",
       x = NULL,
       fill = "p-value") +
  theme(legend.position="right")
```

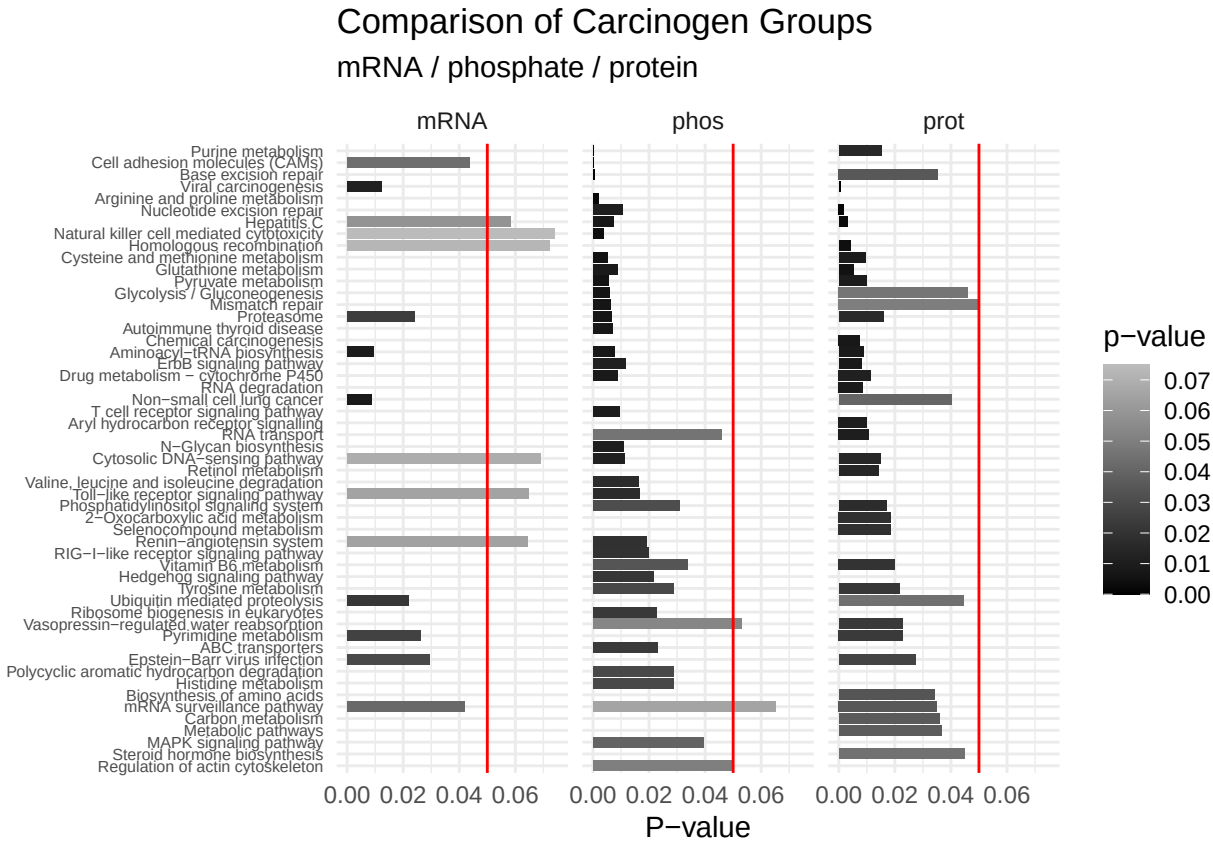
```
## Warning: Removed 38 rows containing missing values (position_stack).
```



Finally, using facet_wrap, present all p-values of mRNA, phosphate, and protein.

```
d %>% filter(type == "PAHs_mRNA") %>% ggplot(aes(reorder(pathway, -P, max), P, fill = P)) +
  geom_bar(stat = 'identity') +
  theme_minimal() +
  ylim(0, 0.075) +
  scale_fill_gradient2(midpoint = 0.05, low = "black", mid = "grey50", high = "white", limits = c(0, 0.0
  geom_hline(yintercept = 0.05, col = "red") +
  coord_flip() +
  theme(axis.text.y = element_text(hjust = 1, size = 6)) +
  labs(title = "Comparison of Carcinogen Groups",
       subtitle = "mRNA / phosphate / protein",
       y = "P-value",
       x = NULL,
       fill = "p-value") +
  theme(legend.position="right") +
  facet_wrap(~p_value)
```

```
## Warning: Removed 71 rows containing missing values (position_stack).
```



4. Discussion

Figure 1 shows that PAHs and nitrosamine have a high mutational signature ratio in phosphate and NitroPAHs have a high proteinic mutation ratio. In addition, in Figure 2, similar to Figure 1, it was found that phosphate and protein affect pathway as carcinogen. According to the Figure 1, tumors harboring PAH or nitro-PAH signatures showed significant enrichment for pathways associated with metabolism and detoxification of chemical carcinogens, including the AHR and Cytochrome P450 pathways, known to contribute to carcinogenesis by PAH. The nitro-PAH and nitrosamines-like groups were dominated by DNA repair, ERBB/MAPK pathway, and TLR/RIG-1 T-cell signaling, which potentially link to the tumor initiation, cell proliferation, EMT malignant progression, and immune modulation in early carcinogenesis(Chen et al., 2020). Through this plotting, it was possible to determine that various carcinogens (PAHs, nitroPAHs, nitrosamine, etc.) absorbed into the body through smoking or air pollution affect to metabolism. Especially, the fact that experimental group used in data is never-smoker makes us to think environmental pollution is main carcinogen in TW cohort.

5. Reference

Chen et al. (2020), Cell, Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression Moorthy et al. (2015), Toxicological Sciences, Polycyclic Aromatic Hydrocarbons: From Metabolism to Lung Cancer H.Robles (2014), Encyclopedia of Toxicology Benjamin A. Musa Bandowe et al. (2017), Science of The Total Environment, Nitrated polycyclic aromatic hydrocarbons (nitro-PAHs) in the environment – A review