Characteristics of Ahr downstream genes at transcriptomic/proteomic level

I. Introduction

Ahr signaling pathway is ~~~~(뭔지, 뭐 포함돼있는지, 어떤 작용하는지, 어떻게 작동/활성화되는지, 논문에서는 어떻게 다뤄졌는지) 논문에서 깊게 다뤄지지 않았는데, anti-apoptosis를 통해 tumorigenesis에 기여할 수 있다는 부분을 보고 흥미가 생겨서 data visualization의 주제로 선정했다

@@ warning 끌까? 어떻게 끄지 @@ sandstone/simplex/

II. Visualization

II-1 Loading/Manipulating data

To see the characteristics of Ahr downstream genes among patients, comparing cohort from the paper with other cohort would be helpful. So I used tables from Chen et al.(2020), which has informations about Taiwan cohort(hencforth TW), and Gillette et al.(2020), which cover geographically diverse population(henceforth g; general). The tables I used include patients' clinical data(TW-S1A, g-S1A), SNV informations(TW-S1C, g-S1A), gene expressions at transcriptomic level(TW-S1D, g-S2D), and gene expressions at proteomic level(TW-S1E, g-S3A).

```
#load packages I need
library(tibble)
library(dplyr)
```

```
##
## 다음의 패키지를 부착합니다: 'dplyr'
```

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

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

```
library(ggplot2)
library(ggridges)
library(dslabs)
library(readxl)
library(cowplot)
library(reshape2)
library(RColorBrewer)
library(tidyverse)
```

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

```
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

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

```
library(ggside)
```

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

```
library(janitor)
```

```
##
## 다음의 패키지를 부착합니다: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggpubr)
```

```
##
## 다음의 패키지를 부착합니다: 'ggpubr'
```

```
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

```
#load data

S1A <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//1-s2.0-S0092867420307431
-mmc1.xlsx", sheet = "Table S1A_clinical_103patient"))

S1D <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//1-s2.0-S0092867420307431
-mmc1.xlsx", sheet = "Table S1D_transcriptome_log2TN"))

S1E <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//1-s2.0-S0092867420307431
-mmc1.xlsx", sheet = "Table S1E_ProteomeLog2TN", na = "NA"))

S1C <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//1-s2.0-S0092867420307431
-mmc1.xlsx", sheet = "Table S1C_SNV", na = "NA"))
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `256` -> `256...6`
## * `86` -> `86...8`
## * `64` -> `64...11`
## * ...
```

```
gS1A <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//mmc1.xlsx", sheet = "An
notions_S1A", na = "NA"))

gS2D <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//mmc2.xlsx", sheet = "Ta
ble S2D", na = "NA"))
```

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ...
```

```
gS3A <- as.data.frame(read_excel("C:/Users/user/Desktop/github/bsms222_158_mun//mmc3.xlsx", sheet = "Ta
ble S3A", na = "NA"))
```

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ...
```

```
#Make data tidy & usable

S1A <- S1A %>%
  rename("variable" = ID)

S1C <- S1C %>%
  row_to_names(row_number = 1) %>%
  subset(select = -c(`COUNT`, `%`)) %>%
  pivot_longer(!Gene,
               names_to = "variable",
               values_to = "SNV")

S1D <- subset(S1D, select = -c(ensembl_gene_id, Median)) %>%
  pivot_longer(!gene,
               names_to = "variable",
               values_to = "value")

S1E <- subset(S1E, select = -c(Accession, Protein)) %>%
  pivot_longer(!Gene,
               names_to = "variable",
               values_to = "value") %>%
  rename("gene" = "Gene")

gS1A <- gS1A %>%
  rename("variable" = "Participant")

gS2D <- gS2D %>%
  subset(select = -c(`...2`, `...3`, `...4`, `...5`, `...6`)) %>%
  row_to_names(row_number = 2, remove_rows_above = TRUE) %>%
  slice(-c(1:68)) %>%
  pivot_longer(!id,
               names_to = "variable",
               values_to = "log2value") %>%
  rename("gene" = "id")

gS3A <- gS3A %>%
  subset(select = -c(`...2`, `...4`, `#1.3`, `...5`, `...6`, `...7`, `...8`, `...9`, `...10`, `...11`,
`...12`, `...13`, `...14`, `...15`, `...16`, `...17`)) %>%
  row_to_names(row_number = 2, remove_rows_above = TRUE) %>%
  slice(-c(1:16)) %>%
  pivot_longer(!geneSymbol,
               names_to = "variable",
               values_to = "log2value") %>%
  rename("gene" = "geneSymbol")
```

II-2 Figure 1

Figure 1 will be density_ridges plot for expression of Ahr dowstream genes at transcriptomic/proteomic level.

```
#Make dataframe with patients' clinical data and Ahr downstream genes for Ahr downstream genes
de <-S1E %>%
  filter(gene %in% c("CYP1A1", "CYP1B1", "GSTA1", "CDA"))
de <- merge(S1A, de, by = "variable") %>%
  mutate(level = "Proteome")
dd <- S1D %>%
  filter(gene %in% c("CYP1A1", "CYP1B1", "GSTA1", "CDA"))
dd <- merge(S1A, dd, by = "variable") %>%
  mutate(level = "Transcriptome")
TWDownstream_DE <- rbind(dd, de)

#Draw plot for total(TW)

TWDownstream_DE$level <- factor(TWDownstream_DE$level, levels=c("Transcriptome", "Proteome"))
TWDownstream_DE$gene <- factor(TWDownstream_DE$gene, levels=c("CYP1A1", "CYP1B1", "CDA", "GSTA1"))

Total_downstream <- TWDownstream_DE %>%
  filter(gene %in% c("CYP1A1", "CYP1B1", "GSTA1", "CDA")) %>%
  ggplot(aes(value, gene, fill = stat(x))) +
  geom_density_ridges_gradient() +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_point(aes(col = stat(x))) +
  scale_color_gradient2(high = "red", mid = "gray90", low = "blue4") +
  guides(color = "none", fill = "none") +
  scale_x_continuous(limits = c(-8, 8)) +
  theme_ridges() +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  labs(x = 'Log2T/N',
       y = 'Ahr downstream genes',
       fill = 'Log2 \n T/N',
       title = "",
       subtitle = "TW cohort - total") +
  theme(axis.text.x = element_text(size = 9, face  = "bold"),
        axis.text.y = element_text(size = 13),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        axis.title.x = element_text(size = 11, hjust = 0.5),
        axis.title.y = element_text(size = 17, hjust = 0.5),
        plot.subtitle = element_text(color = "gray40", face = "italic")) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  facet_wrap(~level)

#Draw plot with category(TW)

Categorized_downstream <- TWDownstream_DE %>%
  mutate(class = case_when(
    Gender == "Female" & `Smoking Status` == "Nonsmoke" & Age <= 60 ~ "A",
    Gender == "Female" & `Smoking Status` == "Nonsmoke" & Age >= 60 ~ "B",
    Gender == "Male" & `Smoking Status` %in% c("Current_Smoker", "Ex-smoker") ~ "C",
    Gender == "Male" & `Smoking Status` == "Nonsmoke" ~ "D"
  )) %>%
  filter(gene %in% c("CYP1A1", "CYP1B1", "GSTA1", "CDA")) %>%
  ggplot(aes(value, gene, fill = stat(x))) +
  geom_density_ridges_gradient() +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_point(aes(col = stat(x))) +
  scale_color_gradient2(high = "red", mid = "gray90", low = "blue4") +
  guides(color = "none") +
  scale_x_continuous(limits = c(-8, 8)) +
  theme_ridges() +
  labs(x = 'Log2T/N',
       y = '',
       fill = 'Log2 \n T/N',
       title = "",
       subtitle = "TW cohort - categorized by patients",
       caption = "A = Young(<=60) Nonsmoker Female / B = Old(>60) Nonsmoker Female \n C = Ex/Current Sm
oker Male / D = Nonsmoker Male") +
  theme(axis.text.x = element_text(size = 9, face  = "bold"),
        axis.text.y = element_text(size = 13),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        axis.title.x = element_text(size = 11, hjust = 0.5),
        axis.title.y = element_text(size = 17, vjust = 0.5),
        plot.caption = element_text(size = 10, face = "italic", hjust = 0),
        plot.subtitle = element_text(color = "gray40", face = "italic")) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  facet_grid(level~class)

#Bring together and add title by plot_grid

plot_row <- plot_grid(Total_downstream, Categorized_downstream, labels = c("A", "B"))
```

```
## Picking joint bandwidth of 0.714
```

```
## Picking joint bandwidth of 0.311
```

```
## Warning: Removed 111 rows containing missing values (geom_point).
```

```
## Picking joint bandwidth of 0.908
```

```
## Picking joint bandwidth of 0.96
```

```
## Picking joint bandwidth of 1.18
```

```
## Picking joint bandwidth of 0.778
```

```
## Picking joint bandwidth of 0.2
```

```
## Picking joint bandwidth of 0.355
```

```
## Picking joint bandwidth of 0.346
```

```
## Picking joint bandwidth of 0.313
```

```
## Warning: Removed 111 rows containing missing values (geom_point).
```

```
title <- ggdraw() +
  draw_label(
    "Figure 1 - Expression of Ahr downstream genes at transcriptomic/proteomic level",
    fontface = "bold",
    x = 0,
    hjust = 0,
    size = 20
  ) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.margin = margin(0, 0, 0, 7))
AHR_downstream <- plot_grid(title, plot_row, ncol = 1, rel_heights = c(0.1,1))

#Calculating Log2T/N

DST <- gS2D %>%
  filter(gene %in% c("CYP1A1", "CYP1B1", "GSTA1", "CDA")) %>%
  rename("Sample.ID" = "variable")
DST <- merge(gS1A, DST, by = "Sample.ID") %>%
  mutate(level = "Transcriptome")

DSP <- gS3A %>%
  filter(gene %in% c("CYP1A1", "CYP1B1", "GSTA1", "CDA")) %>%
  rename("Sample.ID" = "variable")
DSP <- merge(gS1A, DSP, by = "Sample.ID") %>%
  mutate(level = "Proteome")

gDownstream <- rbind(DST, DSP) %>%
  subset(select = c(Sample.ID, variable, Type, Smoking.Status, Age, Gender, Stage, Ethnicity, TP53.muta
tion, EGFR.mutation, gene, log2value, level)) %>%
  slice(-c(1673:1688, 829:844))

NAT <- gDownstream %>%
  filter(Type == "NAT") %>%
  subset(select = c("Sample.ID", "variable", "gene", "log2value", "level"))
Tumor <- gDownstream %>%
  filter(Type == "Tumor", !(variable %in% c("C3L-01862", "C3N-00294", "C3N-01074", "C3N-01842", "C3N-02
422")))
gAHRdownstream <- merge(Tumor, NAT, by = c("variable", "gene", "level")) %>%
  rename("Tumor.Sample.ID" = "Sample.ID.x", "NAT.Sample.ID" = "Sample.ID.y", "Tumor.value" = "log2valu
e.x", "NAT.value" = "log2value.y")

gAHRdownstream$Tumor.value <- as.numeric(gAHRdownstream$Tumor.value)
gAHRdownstream$NAT.value <- as.numeric(gAHRdownstream$NAT.value)

gAHRdownstream$Tumor.value <- 2^(gAHRdownstream$Tumor.value)
gAHRdownstream$NAT.value <- 2^(gAHRdownstream$NAT.value)

gAHRdownstream <- gAHRdownstream %>%
    mutate(logvalue = log2(Tumor.value / NAT.value))

#Draw plot for total(g)

gAHRdownstream$level <- factor(gAHRdownstream$level, levels=c("Transcriptome", "Proteome"))
gAHRdownstream$gene <- factor(gAHRdownstream$gene, levels=c("CYP1A1", "CYP1B1", "CDA", "GSTA1"))

g_total <- gAHRdownstream %>%
  ggplot(aes(logvalue, gene, fill = stat(x))) +
  geom_density_ridges_gradient() +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_point(aes(col = stat(x))) +
  scale_color_gradient2(high = "red", mid = "gray90", low = "blue4") +
  guides(color = "none", fill = "none") +
  scale_x_continuous(limits = c(-8, 8)) +
  theme_ridges() +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  labs(x = 'Log2T/N',
       y = 'Ahr downstream genes',
       fill = 'Log2 \n T/N',
       title = "",
       subtitle = "General cohort - total") +
  theme(axis.text.x = element_text(size = 9, face  = "bold"),
        axis.text.y = element_text(size = 13),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        axis.title.x = element_text(size = 11, hjust = 0.5),
        axis.title.y = element_text(size = 17, hjust = 0.5),
        plot.subtitle = element_text(color = "gray40", face = "italic")) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  facet_wrap(~level)

#Draw plot with category(g)

g_categorized <- gAHRdownstream %>%
  mutate(class = case_when(
    Gender == "female" & Smoking.Status == "non-smoker" & Age <= 60 ~ "A",
    Gender == "female" & Smoking.Status == "non-smoker" & Age >= 60 ~ "B",
    Gender == "female" & Smoking.Status == "smoker" ~ "C",
    Gender == "male" & Smoking.Status == "smoker" ~ "D"
  )) %>%
  filter(!is.na(class)) %>%
  ggplot(aes(logvalue, gene, fill = stat(x))) +
  geom_density_ridges_gradient() +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_point(aes(col = stat(x))) +
  scale_color_gradient2(high = "red", mid = "gray90", low = "blue4") +
  guides(color = "none") +
  scale_x_continuous(limits = c(-8, 8)) +
  theme_ridges() +
  labs(x = 'Log2T/N',
       y = '',
       fill = 'Log2 \n T/N',
       title = "",
       subtitle = "General cohort - categorized by patients",
       caption = "A = Young(<=60) Nonsmoker Female / B = Old(>60) Nonsmoker Female \n  C = Ex/Current S
moker Female / D = Ex/Current smoker Male") +
  theme(axis.text.x = element_text(size = 9, face  = "bold"),
        axis.text.y = element_text(size = 13),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        axis.title.x = element_text(size = 11, hjust = 0.5),
        axis.title.y = element_text(size = 17, vjust = 0.5),
        plot.caption = element_text(size = 10, face = "italic", hjust = 0),
        plot.subtitle = element_text(color = "gray40", face = "italic")) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  facet_grid(level~class)

#Bring together(g)

g_AHR_downstream <- plot_grid(g_total, g_categorized, labels = c("C", "D"))
```

```
## Picking joint bandwidth of 0.746
```

```
## Picking joint bandwidth of 0.983
```

```
## Warning: Removed 124 rows containing missing values (geom_point).
```

```
## Picking joint bandwidth of 0.805
```

```
## Picking joint bandwidth of 0.745
```

```
## Picking joint bandwidth of 0.721
```

```
## Picking joint bandwidth of 0.991
```

```
## Picking joint bandwidth of 1.23
```

```
## Picking joint bandwidth of 1.05
```

```
## Picking joint bandwidth of 1.24
```

```
## Picking joint bandwidth of 1.31
```

```
## Warning: Removed 94 rows containing missing values (geom_point).
```

```
#Bring all 4 plots together

expression_TW_general <- plot_grid(AHR_downstream, g_AHR_downstream, ncol = 1)
```

이렇게 하니까 ~~~ bimodal~~ 한게 나타났다 그래서 그부분만 빼서 나타내봤다

```
#density plot for GSTA1(Asian vs Non-Asian)

dens <- gAHRdownstream %>%
  mutate(region = case_when(
    Ethnicity %in% c("asian", "han") ~ "Asian",
    is.na(Ethnicity) ~ "NA",
    TRUE ~ "Non-Asian"
  )) %>%
  filter(region %in% c("Non-Asian", "Asian")) %>%
  filter(level == "Proteome", gene == "GSTA1") %>%
  ggplot(aes(logvalue, gene, fill = stat(x))) +
  geom_density_ridges_gradient() +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_point(aes(col = stat(x))) +
  scale_color_gradient2(high = "red", mid = "gray90", low = "blue4") +
  guides(color = "none") +
  scale_x_continuous(limits = c(-8, 8)) +
  theme_ridges() +
  labs(x = 'Log2T/N',
       y = '',
       fill = 'Log2 \n T/N',
       title = "",
       subtitle = "GSTA1 expression(General cohort, proteomic level)") +
  theme(axis.text.x = element_text(size = 9, face  = "bold"),
        axis.text.y = element_text(size = 13),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        axis.title.x = element_text(size = 11, hjust = 0.5),
        axis.title.y = element_text(size = 17, vjust = 0.5),
        plot.subtitle = element_text(color = "gray40", face = "italic")) +
  geom_vline(xintercept = 0, col = "black", linetype = "dashed") +
  facet_grid(Gender~region)

#Draw boxplot for GSTA1 expression of Asian/Non-Asian males

box <- gAHRdownstream %>%
  mutate(region = case_when(
    Ethnicity %in% c("asian", "han") ~ "Asian",
    is.na(Ethnicity) ~ "NA",
    TRUE ~ "Non-Asian"
  )) %>%
  filter(region %in% c("Non-Asian", "Asian"), Gender == "male") %>%
  filter(level == "Proteome", gene == "GSTA1") %>%
  ggplot(aes(x = region, y = logvalue)) +
  geom_boxplot(aes(fill = region), width = 0.3, alpha = 0.7) +
  scale_fill_manual(values = c("dodgerblue4", "firebrick4")) +
  geom_jitter(aes(color = region),alpha = 0.4) +
  scale_color_manual(values = c("dodgerblue4", "firebrick4")) +
  theme_classic() +
  theme(legend.position = "none",
        axis.text.x = element_text(face = "bold", size = 12),
        axis.title.y = element_text(face = "bold"),
        axis.title.x = element_blank()) +
  geom_hline(yintercept = 0, col = "black", linetype = "dashed") +
  labs(x = "Region",
       y = "Log2T/N")

#Bring together

GSTA1 <- plot_grid(dens, box, ncol = 2, labels = c("E"))
```

```
## Picking joint bandwidth of 1.84
```

```
## Picking joint bandwidth of 2.24
```

```
## Picking joint bandwidth of 1.06
```

```
## Picking joint bandwidth of 2.09
```

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

```
#Bring all the plots for Ahr downstream expression together

fig1 <- plot_grid(expression_TW_general, GSTA1, ncol = 1, rel_heights = c(4, 1))
ggsave("Figure1.pdf", fig1, width = 12, height = 16)
```

> Figure 1 is about ~~~. Plot A and C shows the relative expression of Ahr downstream genes of TW cohort and g cohort, respectively. Plot B and D shows the relative expression of Ahr downstream genes of TW cohort and g cohort, respectively, categorized by patients' age/smoking status/gender. I used scale_fill_gradient_2 so that 눈에 잘띄는 red~blue 쓸 수 있게. Also, I drew a vertical line at x=0 by geom_vline to make it easier to see whether the expression이 높은/낮은 경향을 보이는지. geom_point on x axis is to helps the viewers to see the 분포/밀집도 easily, and can show values that did not draw a density plot because of small number. Plot E shows the expression of GSTA1 gene, comparing Asian with Non-Asian, male with female. The boxplot helps to see the 경향성 clearly.

II-3 Figure 2

Figure 2 will be heatmaps for Ahr dowmstream genes' expression, with patiens' age/smoking stats/stage/TP53 mutation/EGFR mutation informations.

```
#Merge TP53 mutation information to previous dataframe

tp53 <- S1C %>%
  mutate(TP53 = case_when(
    is.na(SNV) ~ "TP53_WT",
    TRUE ~ "TP53_mutated"
  )) %>%
  subset(select = -c(SNV)) %>%
  rename("gene" = "Gene")
TWDownstream_DE <- merge(TWDownstream_DE, tp53, by = c("variable", "gene"))

#Draw heatmap for TW cohort

TWtile <- TWDownstream_DE %>%
  mutate(Smoking_Status = "Smoking_Status",
         EGFR_mutation = "EGFR_mutation",
         TP53_mutation = ">TP53_mutation",
         stage = "Stage",
         age = "Patient_age") %>%
  mutate(smoke = case_when(
    `Smoking Status` %in% c("Ex-smoker", "Current_Smoker") ~ ".smoker",
    TRUE ~ ".non-smoker"
  )) %>%
  mutate(EGFR = case_when(
    EGFR_Status %in% c("WT", "others") ~ "EGFR_WT/others",
    TRUE ~ "EGFR_L858R/Del19")) %>%
  mutate(Stage = case_when(
    Stage == "IA" ~ "=IA",
    Stage == "IB" ~ "=IB",
    TRUE ~ ">=II"
  )) %>%
  mutate(Age = case_when(
    Age <=60 ~ "Age_younger",
    TRUE ~ "Age_older"
  )) %>%
  mutate(variable = reorder(variable, value, median)) %>%
  ggplot(aes(x = variable, y = gene)) +
  geom_tile(aes(fill = value)) +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_xsidetile(aes(y = Smoking_Status, xfill = smoke)) +
  geom_xsidetile(aes(y = EGFR_mutation, xfill = EGFR)) +
  geom_xsidetile(aes(y = TP53_mutation, xfill = TP53)) +
  geom_xsidetile(aes(y = age, xfill = Age)) +
  geom_xsidetile(aes(y = stage, xfill = Stage)) +
  scale_xfill_manual(values = c("darkorange", "darkolivegreen4", "darkslategray1", "deepskyblue2", "dee
pskyblue4", "wheat3", "wheat1", "mediumorchid4", "mediumorchid1", "black", "floralwhite")) +
  scale_xsidey_discrete(labels = NULL) +
  theme_light() +
  facet_wrap(~level) +
  labs(title = "Figure 2 - Correlation between EGFR/TP53 mutation and Ahr downstream gene expression",
       y = "Ahr downstream genes",
       subtitle = "TW cohort") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        strip.background = element_rect(fill = "gray88"),
        axis.title.x = element_blank(),
        legend.position = "none",
        strip.text = element_text(size = 15, color = "black"),
        plot.subtitle = element_text(color = "gray40", face = "italic"))

#Draw heatmap for g cohort

gtile <- gAHRdownstream %>%
  mutate(Smoking_Status = "Smoking_Status",
         EGFR_mutation = "EGFR_mutation",
         TP53_mutation = ">TP53_mutation",
         stage = "Stage",
         age = "Patient_age") %>%
  mutate(smoke = case_when(
    Smoking.Status == "smoker" ~ ".smoker",
    Smoking.Status == "non-smoker" ~ ".non-smoker"
  )) %>%
  mutate(EGFR = case_when(
    is.na(EGFR.mutation) ~ "EGFR_WT",
    TRUE ~ "EGFR_mutated")) %>%
  mutate(TP53 = case_when(
    is.na(TP53.mutation) ~ "TP53_WT",
    TRUE ~ "TP53_mutated"
  )) %>%
  mutate(Stage = case_when(
    Stage == "1A" ~ "=IA",
    Stage == "1B" ~ "=IB",
    Stage == "1" & is.na(Stage) ~ "NA",
    TRUE ~ ">=II"
  )) %>%
  mutate(Age = case_when(
    Age <= 60 ~ "Age_younger",
    TRUE ~ "Age_older"
  )) %>%
  mutate(variable = reorder(variable, logvalue, median)) %>%
  ggplot(aes(x = variable, y = gene)) +
  geom_tile(aes(fill = logvalue)) +
  scale_fill_gradient2(high = "red", mid = "white", low = "blue") +
  geom_xsidetile(aes(y = Smoking_Status, xfill = smoke)) +
  geom_xsidetile(aes(y = EGFR_mutation, xfill = EGFR)) +
  geom_xsidetile(aes(y = TP53_mutation, xfill = TP53)) +
  geom_xsidetile(aes(y = age, xfill = Age)) +
  geom_xsidetile(aes(y = stage, xfill = Stage)) +
  scale_xfill_manual(values = c("darkorange", "darkolivegreen4", "darkslategray1", "deepskyblue2", "dee
pskyblue4", "wheat3", "wheat1", "mediumorchid4", "mediumorchid1", "black", "floralwhite")) +
  scale_xsidey_discrete(labels = NULL) +
  theme_light() +
  facet_wrap(~level) +
  labs(y = "Ahr downstream genes",
       subtitle = "General cohort") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        legend.position = "none",
        strip.background = element_blank(),
        strip.text = element_blank(),
        axis.title.x = element_blank(),
        plot.subtitle = element_text(color = "gray40", face = "italic"))

#Make legends

legend_C <- get_legend(
  TWDownstream_DE %>%
    mutate(variable = reorder(variable, value, median)) %>%
    mutate(Smoke = case_when(
      `Smoking Status` %in% c("Ex-smoker", "Current_Smoker") ~ ".smoker",
      TRUE ~ ".non-smoker"
      )) %>%
    mutate(Smoking_Status = "Smoking_Status") %>%
    ggplot() +
    geom_tile(aes(x = variable, y = Smoking_Status, fill = Smoke)) +
    scale_fill_manual(values = c("darkorange", "darkolivegreen4")))

legend_E <- get_legend(
  TWDownstream_DE %>%
    mutate(EGFR_mutation = "EGFR_mutation") %>%
    mutate(EGFR = case_when(
    EGFR_Status %in% c("WT", "others") ~ "EGFR_WT/others",
    TRUE ~ "EGFR_L858R/Del19")) %>%
    mutate(variable = reorder(variable, value, median)) %>%
    ggplot() +
    geom_tile(aes(x = variable, y = EGFR_mutation, fill = EGFR)) +
    scale_fill_manual(values = c("mediumorchid4", "mediumorchid1")))

legend_T <- get_legend(
  TWDownstream_DE %>%
    mutate(TP53_mutation = ">TP53_mutation") %>%
    mutate(variable = reorder(variable, value, median)) %>%
    ggplot() +
    geom_tile(aes(x = variable, y = TP53_mutation, fill = TP53)) +
      scale_fill_manual(values = c("black", "floralwhite")))

legend_S <- get_legend(
  TWDownstream_DE %>%
    mutate(Stage = case_when(
    Stage == "IA" ~ "=IA",
    Stage == "IB" ~ "=IB",
    TRUE ~ ">=II"
  )) %>%
    mutate(stage = "Stage") %>%
    ggplot() +
    geom_tile(aes(x = variable, y = stage, fill = Stage)) +
    scale_fill_manual(values = c("darkslategray1", "deepskyblue2", "deepskyblue4")))

legend_A <- get_legend(
  TWDownstream_DE %>%
    mutate(Age = case_when(
    Age <= 60 ~ "Age_younger",
    TRUE ~ "Age_older"
  )) %>%
    mutate(age = "Patient_age") %>%
    ggplot() +
    geom_tile(aes(x = variable, y = age, fill = Age)) +
    scale_fill_manual(values = c("wheat3", "wheat1")))


legends <- plot_grid(legend_S, legend_C, legend_A, legend_E, legend_T, ncol = 1)


#Bring plots together and put legends inside

TWgtile <- plot_grid(TWtile, gtile, ncol = 1)

fig2 <- plot_grid(TWgtile, legends, byrow = TRUE, rel_widths = c(6.6,1))

ggsave("Figure2.pdf", fig2, width = 12, height = 10)
```

> Figure 2 is heatmap for Ahr downstream gene expression of TW and g cohort at transcriptomic/proteomic level. The heatmap is drawn with geom_tile. Stage, smoking status, age, EGFR mutation, TP53 mutation informations are included by geom_xsidetile, and each variables are displayed by color. Large grey 구간 is NAs

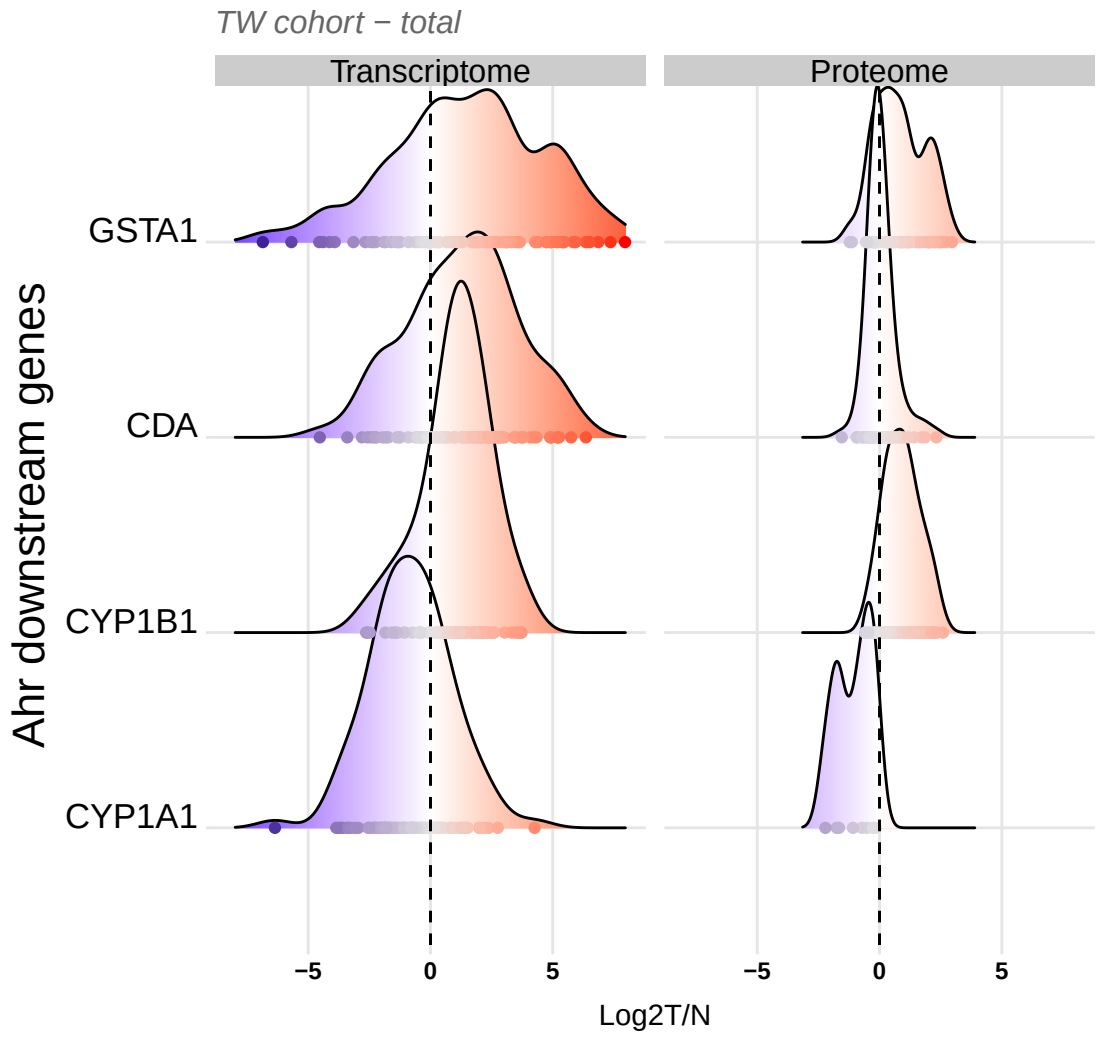
III Consclusion & Discission

III-1 Figure 1

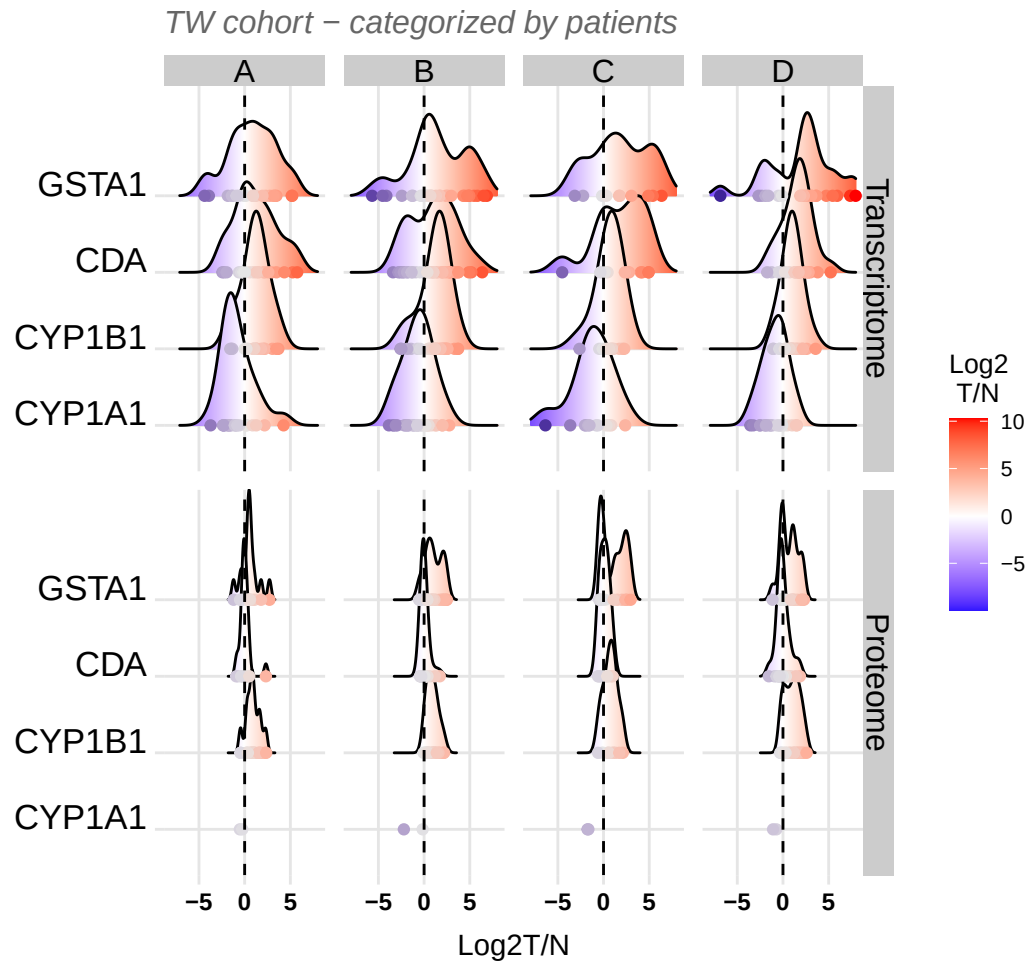
III-2 Figure 2

Figure 1 − Expression of Ahr downstream genes at transcriptomic/proteomic level

A

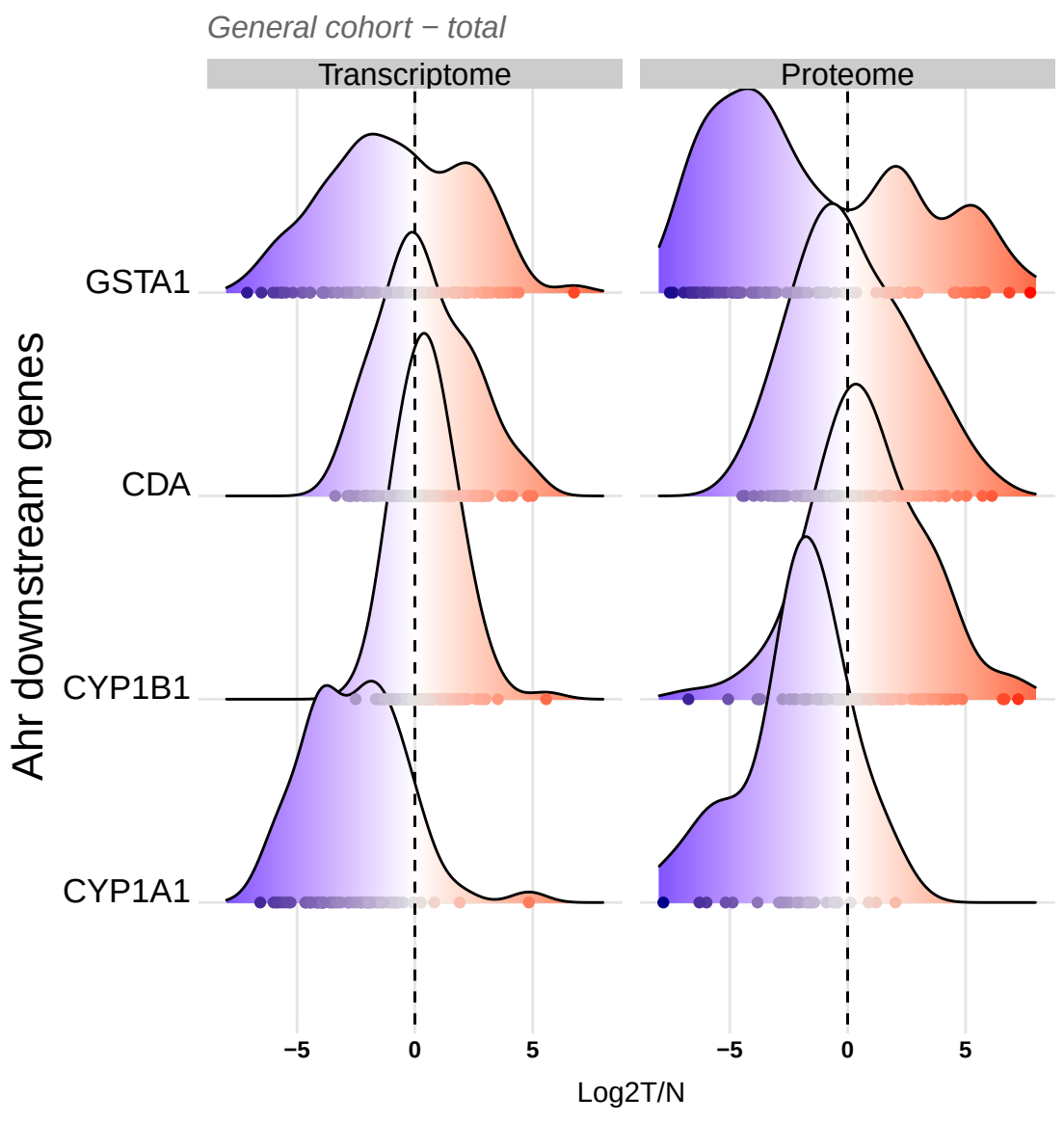


B

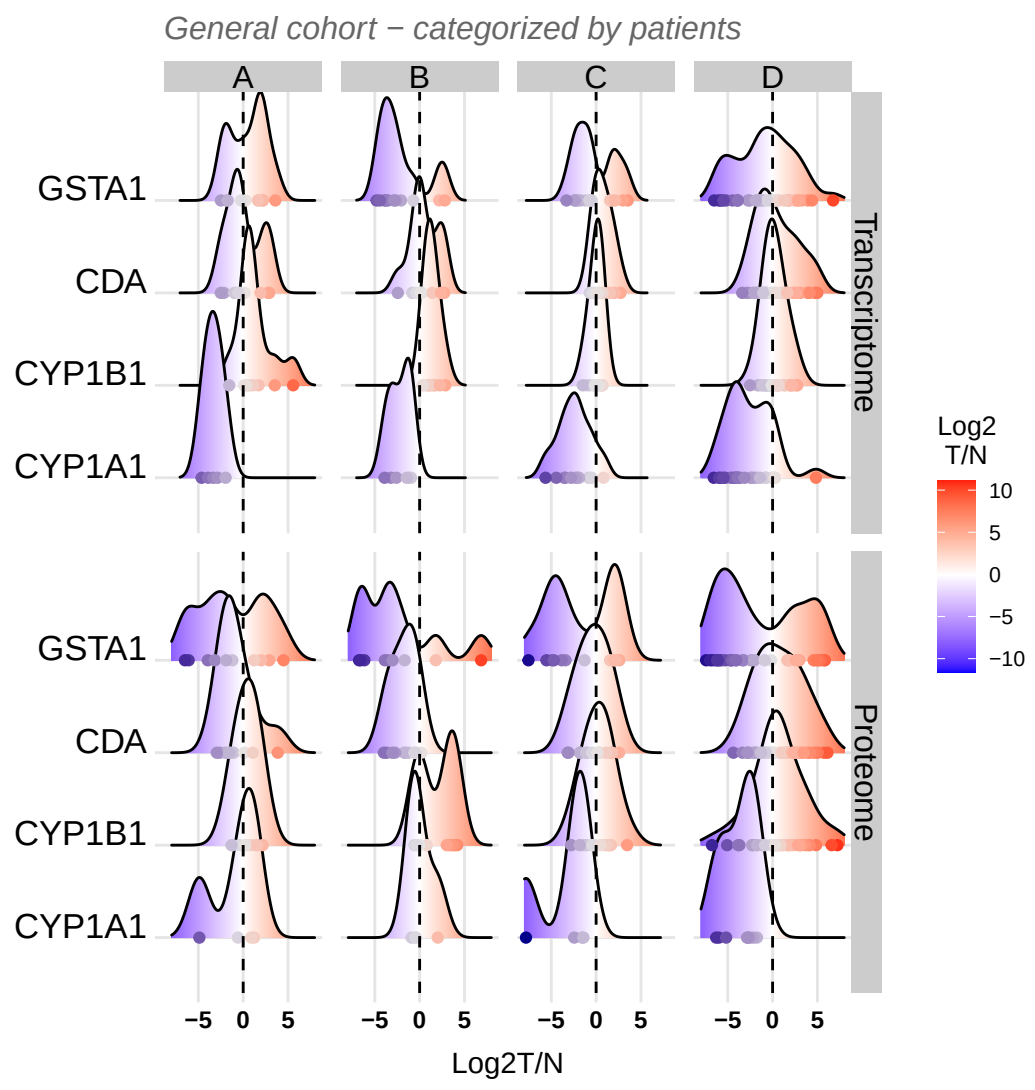


A = Young(<=60) Nonsmoker Female / B = Old(>60) Nonsmoker Female
C = Ex/Current Smoker Male / D = Nonsmoker Male

C

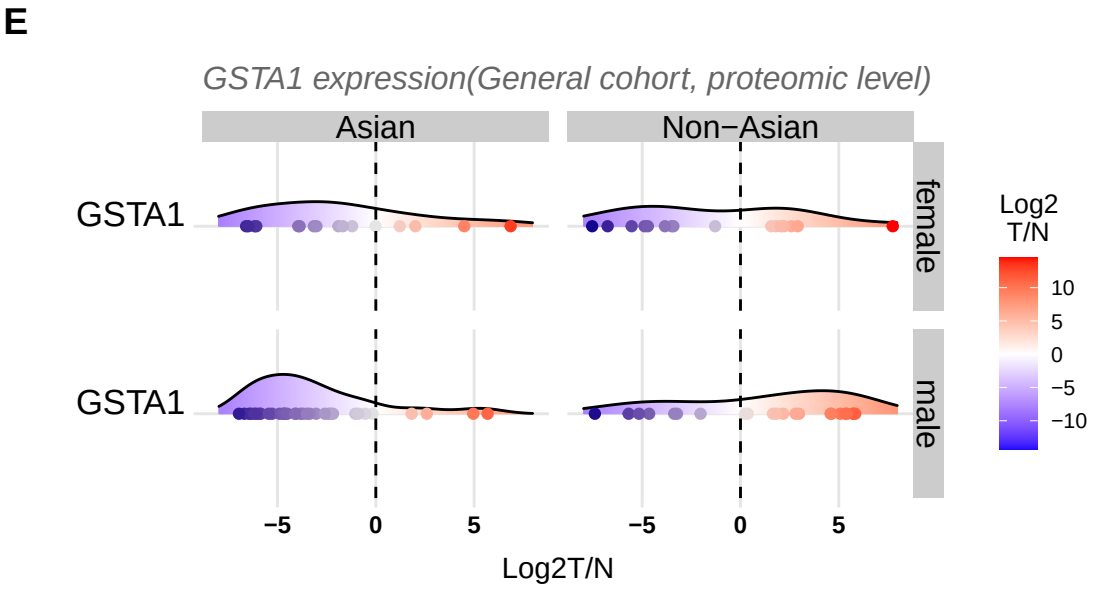


D



A = Young(<=60) Nonsmoker Female / B = Old(>60) Nonsmoker Female
C = Ex/Current Smoker Female / D = Ex/Current smoker Male

E



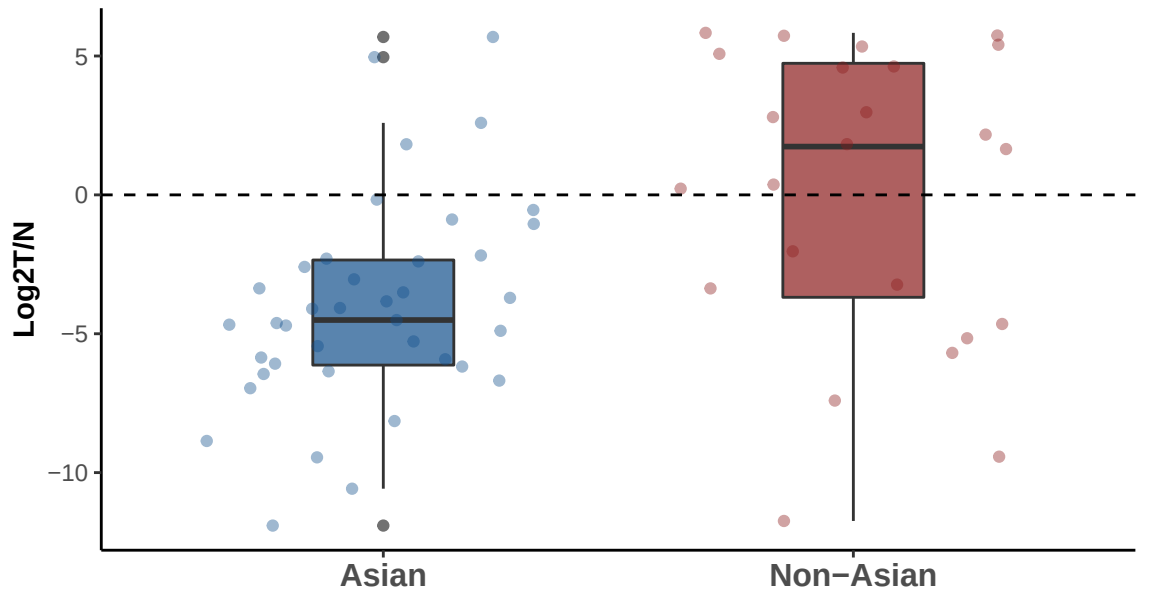


Figure 2 − Correlation between EGFR/TP53 mutation and Ahr downstream gene expression