Assignment1_2017250105_CHOI

Correlation between EGFR mutation and expression of proliferation and antiapoptosis pathway.

Seunghwan Choi, 2017250105

Contents

- Introduction
- EGFR and EGFR activating mutation
- Candidate genes related to pathway involved in tumorigenesis
- Discussion1 - Figure1
- Discussion2 - Figure2
- References
- Code
  - Data reconstruction processes
  - Data Visualizations

Introduction

Lung cancer is one of major cancer and known as leading causes of cancer death in 2020(reported by WHO). This paper, Chen et al (2020), points out that the lung cancer in East Asia have very distinct features from other region where lung carcinoma is mostly found at smoking patient not never-smoking. In East Asia, non-smoker, early onset female patients account for quite large portion of lung cancer cases and these issues remain significant health problem now.

In this paper, whole exome sequencing (WES), RNA-seq, proteomics and phosphoproteomics data were collected from patient-matched tumor and NAT (normal tissue adjacent to the tumor) from 103 treatment-naïve patients from Taiwan. From this multi-omics data, proteogenomic characteristic of East Asian patients (TW cohort) could be clearly explained.

EGFR and EGFR activating mutation

EGFR mutations were frequently found in TW cohort (85%) meanwhile in TCGA cohort, they don't seem to have that high frequency at EGFR. Before getting into main topic, I should refer to the EGFR shortly. EGFR is a type I receptor tyrosine kinase (RTK) and along with its ligands, EGFR is involved in the regulation of multiple cellular pathways which include cellular processes such as cell cycle, cell migration, cell survival, as well as cell proliferation (2). There are two EGFR activating mutation: L858R and exon19del. These two are classical activating mutation comprising the vast majority of EGFR mutation and frequently found in lung cancer patients(3).

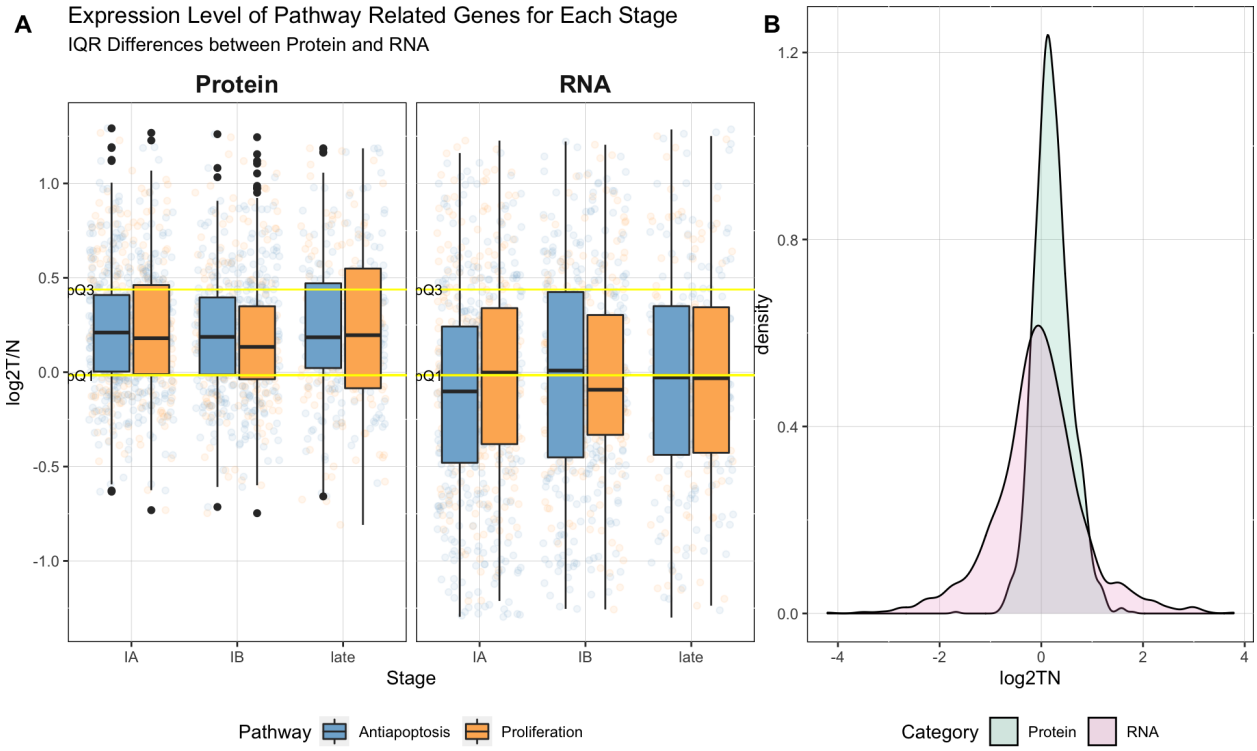
EGFR is comprised of four domains: extracellular EGF-binding domain, a hydrophobic transmembrane domain, and cytoplasmic tyrosine kinase domain and C-terminal phosphorylation domain (2). And EGFR activating mutation, L858R(at exon21) and exon19deletion are mutations in the kinase domain and known to destabilize the inactive conformation of the receptor and further stabilize active conformation leading increased receptor dimerization and activity compared to WT (2).

Candidate genes related to pathway involved in tumorigenesis

I would like to identify a correlation between EGFR mutation and EGFR related pathways involved in tumorigenesis, such as proliferation and anti-apoptosis pathway. Candidate genes related to these two pathways were selected in reference to Figure2F in this paper and you can see in Table below. Additionally, I tried to identify a correlation between cancer stage and expression level of candidate genes each related to proliferation or anti-apoptosis pathway.

Pathway	Candidate gene
Proliferation	ARAF, BRAF, RAF1, MAP2K2, MAP2K1, MAPK1, MAPK3
Anti-apoptosis	PIK3CA, PIK3CB, PIK3CD, PIK3R1, PDPk1, AKT3, AKT2, AKT1 STAT3, STAT5A, STAT5B, JAK3

Discussion1 - Figure1



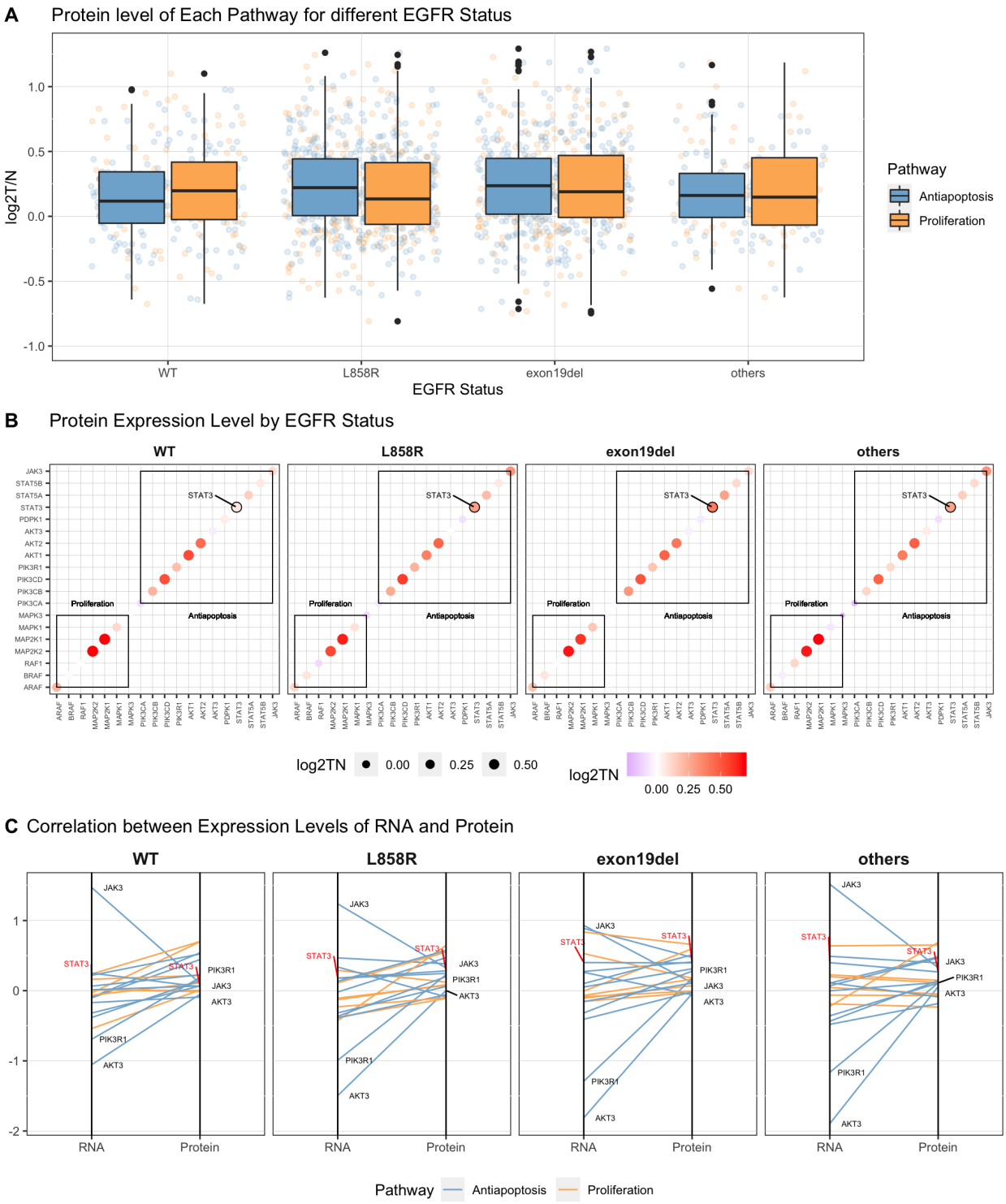
First, expression level of candidate gene for each stage were visualized in Figure 1A. It seems that there are no significant differences between expression level of each pathway across different stages and no significant difference between these two pathways as well. I guess, based on Figure1, these two pathways don't seem to be related to cancer progression. According to this paper, it was reported that proteins that function in cell-to-cell communication, signaling presented negative regulation trend during cancer progression in contrast to proteins in glycolysis, DNA replication (1). Back to the figure1, since the expression levels of each gene are not shown here, I cannot exclude the possibility of wrong generalization that two pathways don't seem

to be related to cancer progression. I should remember that each gene used to be differently and exquisitely regulated even though they are involved in same pathway.

By the way, the most interesting part here is that distribution of expression level is quite different between protein and RNA. Expression levels of RNA are similar to that of NAT meanwhile expression levels of protein tend to increase in tumor tissues. Also, interquartile region (IQR) of protein seems twice smaller than that of RNA. This IQR difference means expression of protein is more sensitively and exquisitely regulated and more consistent about extracellular signals (Figure1B).

From this data visualization, in tumorigenesis, existence and importance of post-translational regulation is easily delivered. This suggests that visualization is powerful method to deliver the intrinsic information in big data.

Discussion2- Figure2



Protein level of each pathway for different EGFR status was analyzed (Figure2A). In EGFR activating mutation, not significant though, expression levels of the anti-apoptosis seem to surpass expression levels of proliferation compared to EGFR wildtype. So, I tried additional visualization to identify which gene is

upregulated in EGFR activating mutation.

It is discovered that STAT3, one of the candidate genes for anti-apoptosis pathway, was activated in patients with EGFR activating mutation (Figure2B). This is related to the other research suggesting STAT3 as important targets for cancer treatment(4). Also, this figure shows highly activated MAP2K regardless of EGFR status.

After I found that STAT3 is upregulated at protein level, I wonder whether expression level of RNA is activated as well. This wondering makes me to do another visualization to identify a correlation between RNA and protein expression level for each candidate genes (Figure2C). Figure 5 shows that Protein expression level is regulated as I expected. However, increase in RNA expression level of STAT3 was similar to the increase in protein expression level. In other words, STAT3 is upregulated at both RNA and Protein level in EGFR activating mutation.
Also, there are some genes that is quite distinct from others in figure5. RNA expression levels of JAK3, PIK3R1 and AKT3 are significantly decreased in EGFR activating mutation meanwhile protein level is quite maintained. And these genes are all related to anti-apoptosis pathway.
JAK is known as major activators of signal transducer and activator of transcription (STAT) proteins. JAK–STAT3 signaling is crucial for cancer development in both tumor cells and the tumor microenvironment (4). I am still wondering what makes these genes downregulated at RNA level compared to protein level. Only thing I can say here is that these genes are highly regulated at protein level.

References

- (1) Chen YJ, Roumeliotis TI, Chang YH, et al. Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell. 2020;182(1):226-244.e17. doi:10.1016/j.cell.2020.06.012
- (2) Rajaram P, Chandra P, Ticku S, Pallavi BK, Rudresh KB, Mansabdar P. Epidermal growth factor receptor: Role in human cancer. Indian J Dent Res. 2017;28(6):687-694. doi:10.4103/ijdr.IJDR_534_16
- (3) Harrison PT, Vyse S, Huang PH. Rare epidermal growth factor receptor (EGFR) mutations in non-small cell lung cancer. Semin Cancer Biol. 2020;61:167-179. doi:10.1016/j.semcancer.2019.09.015
- (4) Yu H, Lee H, Herrmann A, Buettner R, Jove R. Revisiting STAT3 signalling in cancer: new and unexpected biological functions. Nat Rev Cancer. 2014;14(11):736-746. doi:10.1038/nrc3818

Code

Data reconstruction process

```
# DATA From supplementary Excel file. (Table S1)

Patient <- readxl::read_xlsx("/Users/choeseunghwan/Documents/GitHub/bsms222_105_choi/portfolio/Table S1
    sheet = 2)
save(Patient, file = "Patient.Rda")

RNA <- readxl::read_xlsx("/Users/choeseunghwan/Documents/GitHub/bsms222_105_choi/portfolio/Table S1_rela
    sheet = 5)
save(RNA, file = "RNA.Rda")

Protein <- read.csv("/Users/choeseunghwan/Documents/GitHub/bsms222_105_choi/portfolio/TEStCSVfile.csv")
save(Protein, file = "Protein.Rda")  # from same excel data (Table S1, sheet=6)
load("Protein.Rda")
```

```
# Pathway

Proliferation <- c("ARAF", "BRAF", "RAF1",
    "MAP2K2", "MAP2K1", "MAPK1", "MAPK3")

Antiapoptosis <- c("PIK3CA", "PIK3CB", "PIK3CD",
    "PIK3R1", "AKT1", "AKT2", "AKT3", "PDPK1",
    "STAT3", "STAT5A", "STAT5B", "JAK3")
```

```
## Transposing original data

# RNA data -> RNAtrp
RNAtrp <- t(RNA)
RNAtrp <- as.data.frame(RNAtrp)
colnames(RNAtrp) <- RNA$gene
RNAtrp <- RNAtrp[-c(1:3), ]
RNAtrp <- tibble::rownames_to_column(RNAtrp,
    "PatientID")
RNAtrp <- RNAtrp %>%
    mutate_at(-1, as.numeric)

# Protein data -> Proteintrp There is
# some unknown gene in original Protein
# data. so I delete it.
Protein <- Protein %>%
    filter(Gene != "")
Proteintrp <- t(Protein)
Proteintrp <- as.data.frame(Proteintrp)
colnames(Proteintrp) <- Protein$Gene
Proteintrp <- Proteintrp[-c(1:3), ]
Proteintrp <- tibble::rownames_to_column(Proteintrp,
    "PatientID")
Proteintrp <- Proteintrp %>%
    mutate_at(-1, as.numeric)
```

```r
# Use this dataset

load("Protein.Rda")
load("RNA.Rda")
load("Proteintrp.Rda")
load("RNAtrp.Rda")
load("PatientRNA.Rda")
load("PatientProtein.Rda")
```

```r
# Data merging

# PatientRNA

PatientRNA <- merge(Patient %>%
    select(ID, Gender, Age, `Smoking Status`,
        Stage, EGFR_Status), RNAtrp, by.x = "ID",
    by.y = "PatientID")
PatientRNA <- PatientRNA %>%
    mutate(Staging = case_when(Stage == "IA" ~
        "IA", Stage == "IB" ~ "IB", TRUE ~
        "late"))
PatientRNA <- PatientRNA[, c(1:5, ncol(PatientRNA),
    6:(ncol(PatientRNA) - 1))]
PatientRNA <- PatientRNA %>%
    mutate(Staging = factor(Staging, levels = c("IA",
        "IB", "late"))) %>%
    filter(EGFR_Status != "L858R.exon19del") %>%
    mutate(EGFR_Status = factor(EGFR_Status,
        levels = c("WT", "L858R", "exon19del",
            "others"))) %>%
    mutate(Gender = factor(Gender, levels = c("Male",
        "Female"))) %>%
    mutate(`Smoking Status` = factor(`Smoking Status`))

save(PatientRNA, file = "PatientRNA.Rda")

# PatientProtein

PatientProtein <- merge(Patient %>%
    select(ID, Gender, Age, `Smoking Status`,
        Stage, EGFR_Status), Proteintrp,
    by.x = "ID", by.y = "PatientID")
save(PatientProtein, file = "PatientProtein.Rda")
PatientProtein <- PatientProtein %>%
    mutate(Staging = case_when(Stage == "IA" ~
        "IA", Stage == "IB" ~ "IB", TRUE ~
        "late"))
PatientProtein <- PatientProtein[, c(1:5,
    ncol(PatientProtein), 6:(ncol(PatientProtein) -
        1))]
PatientProtein <- PatientProtein %>%
    mutate(Staging = factor(Staging, levels = c("IA",
        "IB", "late"))) %>%
```

```
    filter(EGFR_Status != "L858R.exon19del") %>%
    mutate(EGFR_Status = factor(EGFR_Status,
        levels = c("WT", "L858R", "exon19del",
            "others"))) %>%
    mutate(Gender = factor(Gender, levels = c("Male",
        "Female"))) %>%
    mutate(`Smoking Status` = factor(`Smoking Status`))

save(PatientProtein, file = "PatientProtein.Rda")
```

```
# RNA

longer_RNA_P <- PatientRNA %>%
    select(Staging, all_of(Proliferation)) %>%
    pivot_longer(2:ncol(.), names_to = "Gene",
        values_to = "log2TN")

longer_RNA_A <- PatientRNA %>%
    select(Staging, all_of(Antiapoptosis)) %>%
    pivot_longer(2:ncol(.), names_to = "Gene",
        values_to = "log2TN")

longer_RNA_P <- longer_RNA_P %>%
    mutate(Pathway = "Proliferation")
longer_RNA_A <- longer_RNA_A %>%
    mutate(Pathway = "Antiapoptosis")

longer_RNA <- rbind(longer_RNA_P, longer_RNA_A)
```

```
# Protein

longer_Protein_P <- PatientProtein %>%
    select(Staging, all_of(Proliferation)) %>%
    pivot_longer(2:ncol(.), names_to = "Gene",
        values_to = "log2TN")

longer_Protein_A <- PatientProtein %>%
    select(Staging, all_of(Antiapoptosis)) %>%
    pivot_longer(2:ncol(.), names_to = "Gene",
        values_to = "log2TN")

#
longer_Protein_P <- longer_Protein_P %>%
    mutate(Pathway = "Proliferation")
longer_Protein_A <- longer_Protein_A %>%
    mutate(Pathway = "Antiapoptosis")

longer_Protein <- rbind(longer_Protein_P,
    longer_Protein_A)
```

```
# EGFR RNA
EGFR_longer_RNA_P <- PatientRNA %>%
    select(Staging, EGFR_Status, all_of(Proliferation)) %>%
```

```
    pivot_longer(3:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Pathway = "Proliferation")


EGFR_longer_RNA_A <- PatientRNA %>%
    select(Staging, EGFR_Status, all_of(Antiapoptosis)) %>%
    pivot_longer(3:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Pathway = "Antiapoptosis")
```

Visualization

```
# Figure1A

Figure1 <- rbind(longer_RNA %>%
    mutate(Category = "RNA"), longer_Protein %>%
    mutate(Category = "Protein")) %>%
    ggplot(aes(Staging, log2TN)) + geom_jitter(aes(color = Pathway),
    alpha = 0.1) + geom_boxplot(aes(fill = Pathway)) +
    facet_wrap(Category ~ .) + ylim(-1.3,
    1.3) + ggtitle("Expression Level of Pathway Related Genes for Each Stage") +
    scale_fill_manual(values = c("#80B1D3",
        "#FDB462")) + scale_color_manual(values = c("#80B1D3",
    "#FDB462")) + geom_rect(xmin = 0, xmax = 4,
    ymin = quantile(longer_Protein$log2TN,
        0.25), ymax = quantile(longer_Protein$log2TN,
        0.75), fill = NA, color = "yellow",
    size = 0.4) + geom_text(data = data.frame(y = c(quantile(longer_Protein$log2TN,
    0.25), quantile(longer_Protein$log2TN,
    0.75)), x = c(0.52, 0.52), q = c("pQ1",
    "pQ3")), aes(x = x, y = y, label = q),
    size = 3) + theme(legend.position = "bottom",
    strip.background = element_rect(fill = NA),
    strip.text.x = element_text(face = "bold",
        size = 14), panel.background = element_rect(fill = "white",
        color = "black", linetype = "solid"),
    panel.grid.major = element_line(color = "grey",
        size = 0.1), plot.margin = unit(c(0,
        0, 0, 0), "cm")) + labs(subtitle = "IQR Differences between Protein and RNA",
    x = "Stage", y = "log2T/N")
```

```
# Figure1B

Figure2 <- rbind(longer_RNA %>%
    mutate(Category = "RNA"), longer_Protein %>%
    mutate(Category = "Protein")) %>%
    ggplot() + geom_density(aes(x = log2TN,
    fill = Category), alpha = 0.2) + theme(strip.background = element_rect(fill = NA),
    panel.background = element_rect(fill = "white",
        color = "black", linetype = "solid"),
```

```
    panel.grid.major = element_line(color = "grey",
        size = 0.1), plot.margin = unit(c(0,
        0, 0, 0), "cm"), legend.position = "bottom") +
    scale_fill_manual(values = c("#66C2A5",
        "#E78AC3"))
```

```
# Figure2A

# Protein
EGFR_longer_Protein_P <- PatientProtein %>%
    select(Staging, EGFR_Status, all_of(Proliferation)) %>%
    pivot_longer(3:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Pathway = "Proliferation")

EGFR_longer_Protein_A <- PatientProtein %>%
    select(Staging, EGFR_Status, all_of(Antiapoptosis)) %>%
    pivot_longer(3:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Pathway = "Antiapoptosis")

Figure3 <- rbind(EGFR_longer_Protein_P, EGFR_longer_Protein_A) %>%
    ggplot(aes(EGFR_Status, log2TN)) + geom_jitter(alpha = 0.2,
    aes(color = Pathway)) + geom_boxplot(aes(fill = Pathway)) +
    ylim(-1, 1.3) + ggtitle("Protein level of Each Pathway for different EGFR Status") +
    scale_fill_manual(values = c("#80B1D3",
        "#FDB462")) + scale_color_manual(values = c("#80B1D3",
    "#FDB462")) + theme(strip.background = element_rect(fill = NA),
    strip.text.x = element_text(face = "bold",
        size = 14), panel.background = element_rect(fill = "white",
        color = "black", linetype = "solid"),
    panel.grid.major = element_line(color = "grey",
        size = 0.1)) + labs(y = "log2T/N",
    x = "EGFR Status")
```

```
# Figure2B
IDEGFR_longer_Protein_P <- PatientProtein %>%
    select(ID, Staging, EGFR_Status, all_of(Proliferation)) %>%
    pivot_longer(4:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Pathway = "Proliferation")

IDEGFR_longer_Protein_A <- PatientProtein %>%
    select(ID, Staging, EGFR_Status, all_of(Antiapoptosis)) %>%
    pivot_longer(4:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Pathway = "Antiapoptosis")

Figure4 <- rbind(IDEGFR_longer_Protein_P,
    IDEGFR_longer_Protein_A) %>%
    group_by(EGFR_Status, Pathway, Gene) %>%
    summarize(log2TN = median(log2TN)) %>%
    mutate(Gene = factor(Gene, levels = c("ARAF",
```

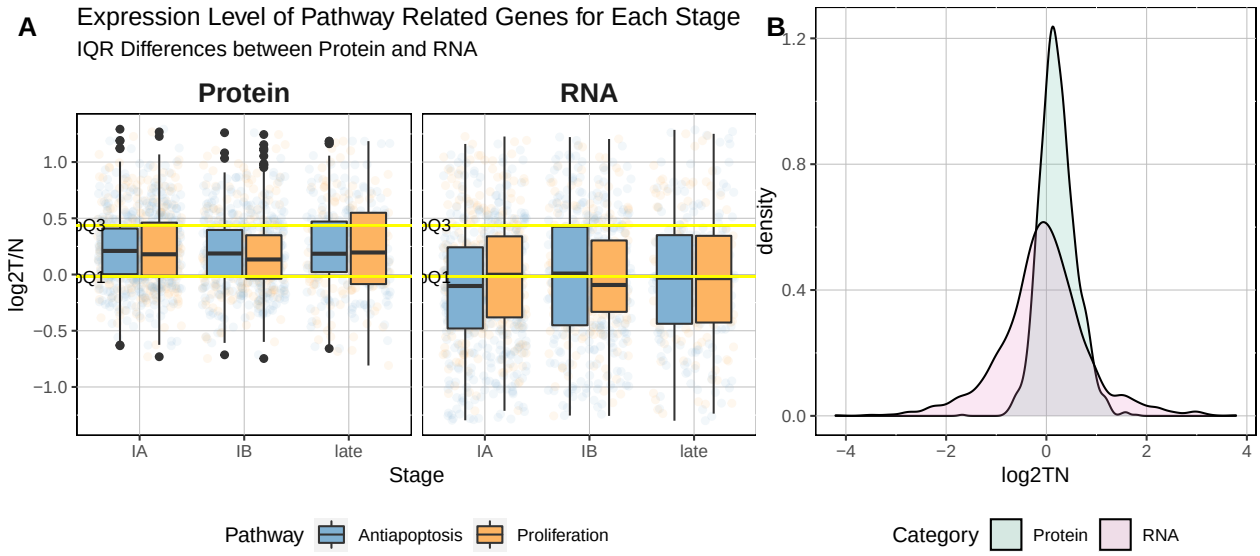
```
        "BRAF", "RAF1", "MAP2K2", "MAP2K1",
        "MAPK1", "MAPK3", "PIK3CA", "PIK3CB",
        "PIK3CD", "PIK3R1", "AKT1", "AKT2",
        "AKT3", "PDPK1", "STAT3", "STAT5A",
        "STAT5B", "JAK3"))) %>%
    ggplot() + geom_point(aes(Gene, Gene,
    size = log2TN, color = log2TN)) + geom_rect(xmin = "ARAF",
    xmax = "MAPK3", ymin = "ARAF", ymax = "MAPK3",
    fill = NA, color = "black", size = 0.2,
    linetype = "solid") + geom_rect(xmin = "PIK3CA",
    xmax = "JAK3", ymin = "PIK3CA", ymax = "JAK3",
    fill = NA, color = "black", size = 0.2,
    linetype = "solid") + geom_text(aes(x = "MAP2K2",
    y = "PIK3CA"), label = "Proliferation",
    size = 2) + geom_text(aes(x = "AKT3",
    y = "MAPK3"), label = "Antiapoptosis",
    size = 2) + scale_color_gradient2(low = "Purple",
    high = "Red", mid = "white", midpoint = 0) +
    facet_grid(. ~ EGFR_Status) + theme(legend.position = "bottom",
    axis.text.x = element_text(size = 5,
        angle = 90, hjust = 1), axis.text.y = element_text(size = 5,
        angle = 0, hjust = 1), strip.background = element_rect(fill = NA),
    strip.text.x = element_text(face = "bold",
        size = 10), panel.background = element_rect(fill = "white",
        color = "black", linetype = "solid"),
    panel.grid.major = element_line(color = "grey",
        size = 0.1), aspect.ratio = 1, legend.key.size = unit(6,
        "mm"), axis.title = element_blank()) +
    scale_size_continuous(range = c(1, 3)) +
    geom_text_repel(data = data.frame(x = c("STAT3"),
        y = c("STAT3")), aes(x, y, label = "STAT3"),
        nudge_x = -3, nudge_y = 1, size = 2) +
    geom_point(data = data.frame(x = c("STAT3"),
        y = c("STAT3")), aes(x, y), shape = 1,
        size = 3) + labs(title = "Protein Expression Level by EGFR Status")
```

```
# Figure2C

Figure5 <- rbind(PatientRNA %>%
    select(ID, EGFR_Status, Staging, all_of(c(Proliferation,
        Antiapoptosis))) %>%
    pivot_longer(4:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Category = "RNA"), PatientProtein %>%
    select(ID, EGFR_Status, Staging, all_of(c(Proliferation,
        Antiapoptosis))) %>%
    pivot_longer(4:ncol(.), names_to = "Gene",
        values_to = "log2TN") %>%
    mutate(Category = "Protein")) %>%
    mutate(Category = factor(Category, levels = c("RNA",
        "Protein"))) %>%
    group_by(EGFR_Status, Gene, Category) %>%
    summarize(log2TN = median(log2TN)) %>%
```

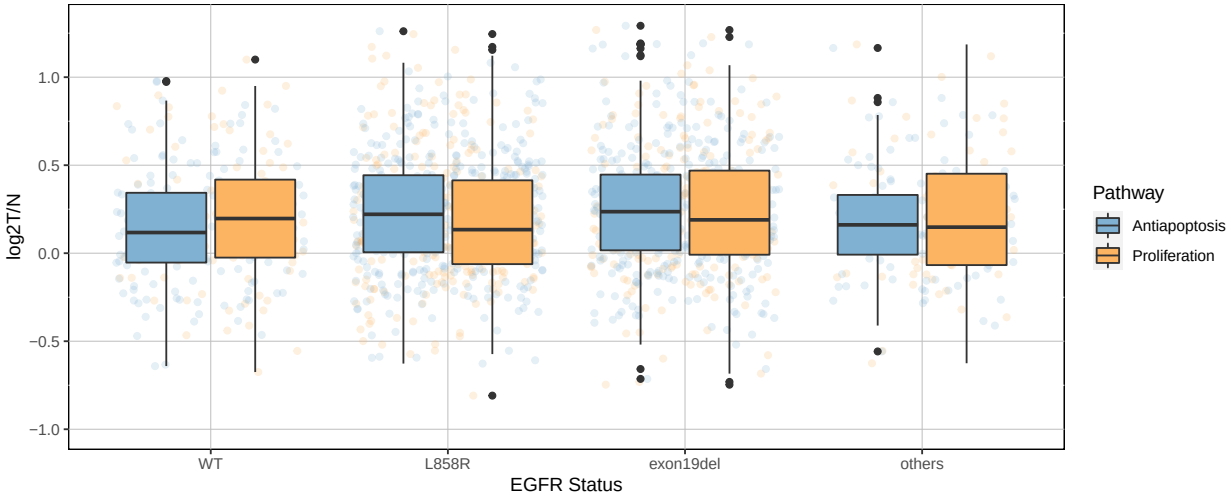
```
    mutate(Pathway = ifelse(Gene %in% all_of(Proliferation),
        "Proliferation", "Antiapoptosis")) %>%
    ggplot(aes(Category, log2TN)) + geom_line(aes(group = Gene,
    color = Pathway)) + facet_grid(. ~ EGFR_Status) +
    scale_fill_manual(values = c("#80B1D3",
        "#FDB462")) + scale_color_manual(values = c("#80B1D3",
    "#FDB462")) + theme(strip.background = element_rect(fill = NA),
    strip.text.x = element_text(face = "bold",
        size = 12), panel.background = element_rect(fill = "white",
        color = "black", linetype = "solid"),
    panel.grid.major = element_line(color = "grey",
        size = 0.1), legend.key.size = unit(6,
        "mm"), legend.position = "bottom",
    axis.title = element_blank()) + geom_text_repel(aes(label = ifelse(Gene %in%
    c("JAK3", "AKT3", "PIK3R1"), Gene, "")),
    size = 2, nudge_x = 0.2) + geom_text_repel(aes(label = ifelse(Gene ==
    "STAT3", Gene, "")), size = 2, nudge_y = 0.3,
    nudge_x = -0.1, color = "Red") + geom_vline(xintercept = c("RNA",
    "Protein")) + labs(title = "Correlation between Expression Levels of RNA and Protein")
```

```
# Figure1
plot_grid(Figure1, Figure2, labels = c("A",
    "B"), rel_widths = c(3, 2))
```

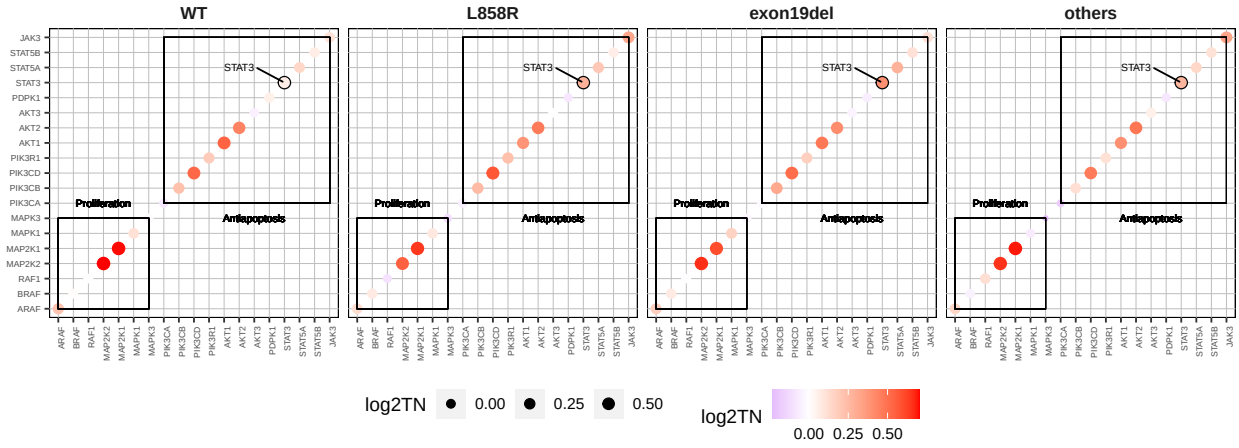



```
# Figure2
plot_grid(Figure3, Figure4, Figure5, labels = c("A",
    "B", "C"), nrow = 3)
```

A Protein level of Each Pathway for different EGFR Status



B Protein Expression Level by EGFR Status



C Correlation between Expression Levels of RNA and Protein