Relationship between tumor purity and immune subtypes in lung cancer

2018250103 Kim Yeonjae

1. Introduction

Using TW LUAD data, I would like to examine the expression patterns of immune-related genes according to tumor purity. Tumor purity is defined as the proportion of cancer cells in the tumor tissue, which reflects the characteristics of TME. The tumor microenvironment is a complex milieu consisting of factors that promote growth and inhibit it, as well as nutrients, chemokines, and very importantly, other non-cancerous cell types. These cells include fibroblasts, immune cells, endothelial cells and normal epithelial cells. These non-cancerous components of the tumor may play an important role in cancer biology. Among the various roles, I would like to consider what indicator of tumor purity can be used, especially by examining the relationship with the immune system. Yoshihara et al have developed ESTIMATE algorithm for assessment of the presence of stromal cells and the infiltration of immune cells in tumor samples using gene expression data and further calculating tumor purity. I will use data of Taiwanese lung cancer patient's tumor purity which is calculated by ESTIMATE algorithm. We will look at the expression of mRNA in immune-related genes in groups divided based on tumor purity. The genes to look at are as follows.

Immune cell =T-reg(CD3,CD4,CD25,FoxP3), M2(CD163)

Immune checkpoint = PD-1(PDCD1) ,PD-L1(CD274), LAG3, TIGIT

Immune surpressive pathway = IL6-STAT5(IL6, STAT5), IL2-JAK-STAT5(STAT5A, STAT5B, IL2RA, IL2RG)

2. Data Visualization

2.1 Data Visualization plan

1. Exploring data
2. Grouping patients by Tumor purity level
3. Select the gene
4. Visualizing data
5. Combine plots

2.2.1 Load data

```
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   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
library(readxl)
library(ggplot2)
```

Load "Tumor purity table". Both tumor tissue and normal tissue of patients' tumor purity is calculated. So I removed normal tissue to avoid patient code overlap.

```
# Load Tumor purity table
p = read_excel('portfolio/1-s2.0-S0092867420307431-mmc7.xlsx', sheet=4)
pT <- p[1:103,] # Remove normal tissue
```

I will divide them into 4 groups according to quantile values.

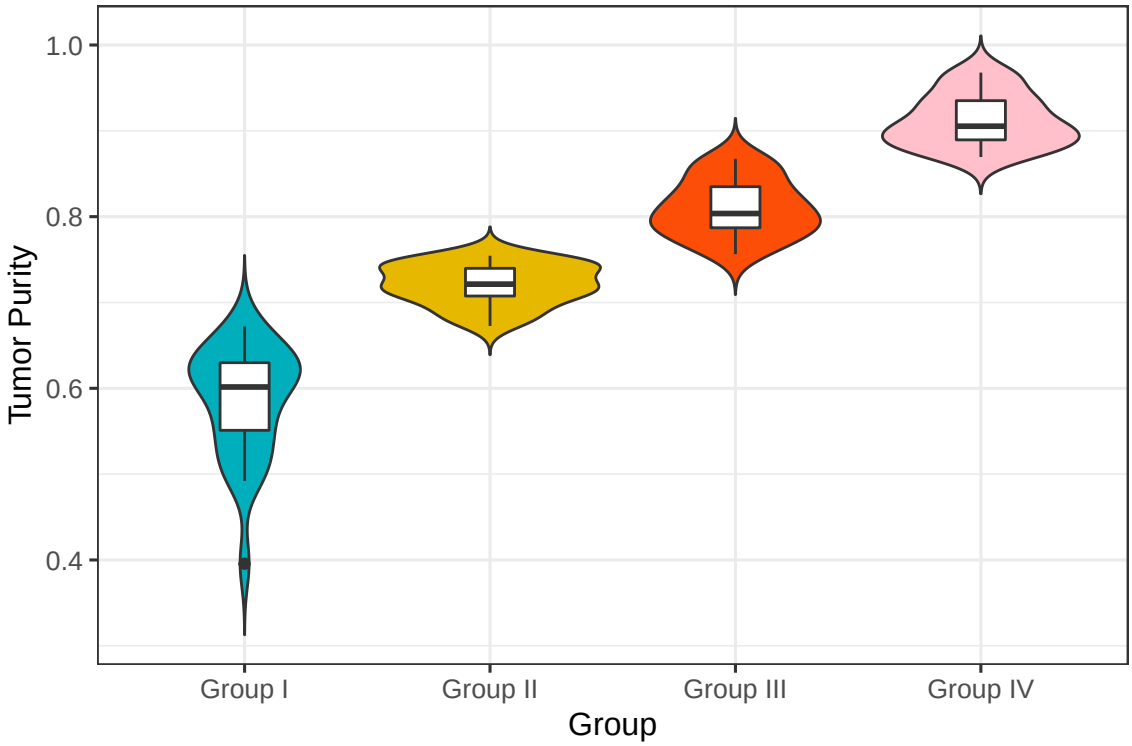
```
quantile(pT$`Tumor Purity`, c(0, 0.25, 0.5, 0.75, 1))
```

```
##        0%       25%       50%       75%      100%
## 0.3955870 0.6723151 0.7544549 0.8684486 0.9678307
```

```
pT$tp_level = ifelse(pT$`Tumor Purity` < 0.6723151, "Group I",
                     ifelse(pT$`Tumor Purity` < 0.7544549, "Group II",
                            ifelse(pT$`Tumor Purity` < 0.8684486, "Group III", "Group IV")))
```

Let's try to make boxplot to see the distribution of groups in this samlple.

```
ggplot(pT, aes(x = tp_level, y = `Tumor Purity`)) +
  geom_violin(aes(fill = tp_level), trim = FALSE) +
  geom_boxplot(width = 0.2)+
  theme_bw(base_size = 12) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800", "#FC4E07","pink"))+
  theme(legend.position = "none")+
  xlab("Group")
```



<Figure.1>

And then, Load transcriptome log2TN table. This table contain TW LUAD patient's mRNA expression results.

```
p5 = read_excel('portfolio/1-s2.0-S0092867420307431-mmc1.xlsx', sheet=5)
```

2.2.2 Grouping patients by Tumor purity level

New median values are defined by extracted patients, divided according to purity.

```
## Group I
pg1 <- pT %>% filter(tp_level == "Group I")
pg1$patient_code <- sapply(strsplit(as.character(pg1$`Patient.tissue sample`),"[.]"), "[", 1)
# Some patients doesn't have log2TN value, so need to remove them
pg1 <- pg1[!pg1$`Patient.tissue sample` %in% c("P004.T", "P046.T","P084.T"), ]
pg1.patient_code <- pg1 %>% pull(patient_code)

p5_group1 <- p5 %>% select(gene, all_of(pg1.patient_code)) # Select gene and log2TN value of pg1 patien
p5_group1$Median <- apply(p5_group1[,2:24],1,median) # Calculate Median log2TN value

## Group II
pg2 <- pT %>% filter(tp_level == "Group II")
pg2$patient_code <- sapply(strsplit(as.character(pg2$`Patient.tissue sample`),"[.]"), "[", 1)
pg2 <- pg2[!pg2$`Patient.tissue sample` %in% c("P035.T", "P047.T","P083.T"), ]
pg2.patient_code <- pg2 %>% pull(patient_code)
pg2.patient_code
```

```
##  [1] "P070" "P088" "P099" "P101" "P072" "P085" "P095" "P097" "P016" "P058"
```

```
## [11] "P027" "P081" "P021" "P056" "P055" "P019" "P023" "P015" "P028" "P045"
## [21] "P049" "P051" "P068"
```

```
p5_group2 <- p5 %>% select(gene, all_of(pg2.patient_code))
p5_group2$Median <- apply(p5_group2[,2:24],1,median)

## Group III
pg3 <- pT %>% filter(tp_level == "Group III")
pg3$patient_code <- sapply(strsplit(as.character(pg3$`Patient.tissue sample`),"[.]"), "[", 1)
pg3 <- pg3[!pg3$`Patient.tissue sample` %in% c("P087.T", "P096.T","P078.T","P079.T"), ]
pg3.patient_code <- pg3 %>% pull(patient_code)

p5_group3 <- p5 %>% select(gene, all_of(pg3.patient_code))
p5_group3$Median <- apply(p5_group3[,2:22],1,median)

## Group IV
pg4 <- pT %>% filter(tp_level == "Group IV")
pg4$patient_code <- sapply(strsplit(as.character(pg4$`Patient.tissue sample`),"[.]"), "[", 1)
pg4 <- pg4[!pg4$`Patient.tissue sample` %in% c("P005.T", "P069.T","P041.T"), ]
pg4.patient_code <- pg4 %>% pull(patient_code)
pg4.patient_code
```

```
##  [1] "P007" "P012" "P086" "P009" "P089" "P091" "P061" "P102" "P103" "P090"
## [11] "P092" "P098" "P024" "P032" "P018" "P064" "P042" "P040" "P054" "P080"
## [21] "P110" "P111" "P112"
```

```
p5_group4 <- p5 %>% select(gene, all_of(pg4.patient_code))
p5_group4$Median <- apply(p5_group4[,2:24],1,median)
```

2.2.3 Select the gene we want to see.

Pro-tumorigenic inflammation M2 macrophages(CD163), T-regs(CD3D, CD3E, CD3G, CD4, FoxP3)

Immune Checkpoints PD-1(PDCD1) ,PD-L1(CD274), LAG3, TIGIT

Immune surpressive pathway IL6-STAT5 (IL6, STAT5), IL2-JAK-STAT5(STAT5A, STAT5B, IL2RA, IL2RG)

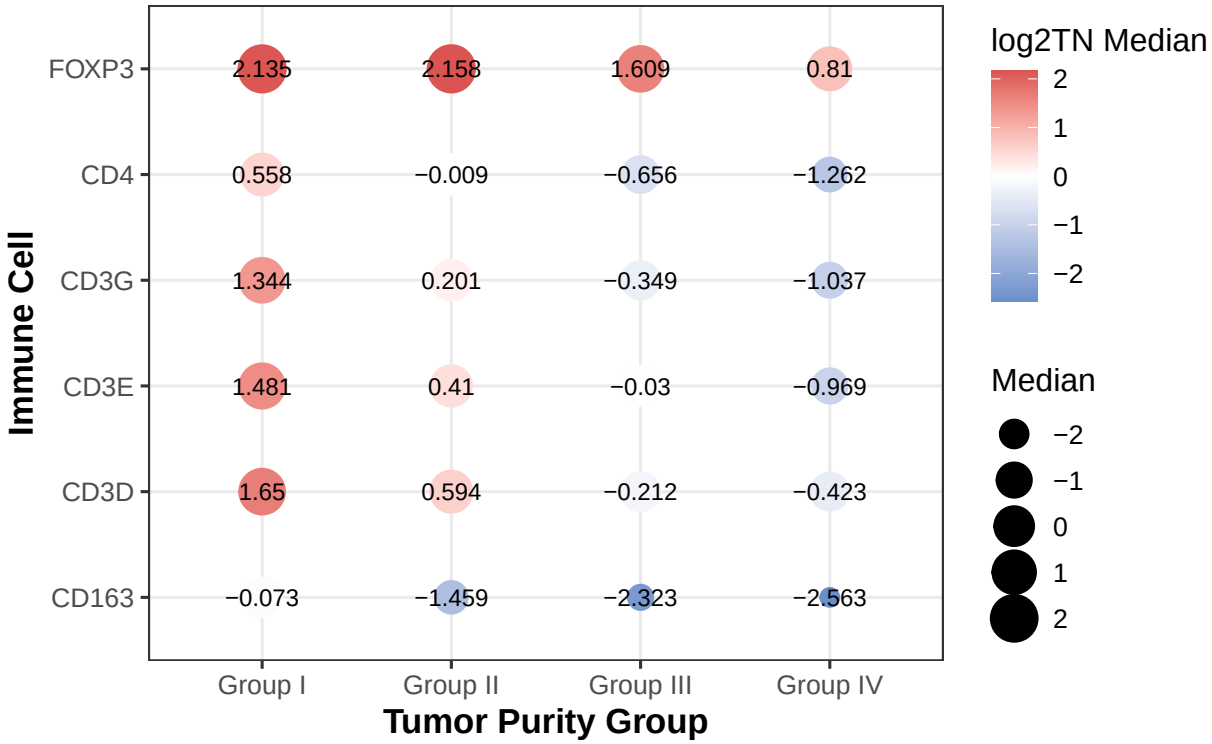
2.2.4 Visualize geom_point

Pro-tumorigenic inflammation

M2 macrophages(CD163), T-regs(CD3D, CD3E, CD3G, CD4, FoxP3)

```
## Visualize geom_point

### Immune cell
immune_cell <- c("CD3D","CD3E","CD3G","CD4", "FOXP3","CD163")
p5_group1.c <- p5_group1 %>% filter(gene %in% immune_cell) %>% select(gene, Median) %>% mutate(group =
p5_group2.c <- p5_group2 %>% filter(gene %in% immune_cell) %>% select(gene, Median) %>% mutate(group =
p5_group3.c <- p5_group3 %>% filter(gene %in% immune_cell) %>% select(gene, Median) %>% mutate(group =
p5_group4.c <- p5_group4 %>% filter(gene %in% immune_cell) %>% select(gene, Median) %>% mutate(group =

p5_total.c = rbind.data.frame(p5_group1.c, p5_group2.c, p5_group3.c, p5_group4.c)
```

```
p5i =  ggplot(data = p5_total.c, aes(x = group, y = gene, col = Median)) +
  geom_point(aes(size = Median), fill = "#1b98e0", stroke = 2) +
  geom_text(data= p5_total.c, mapping = aes(x = group, y = gene, label = round(Median, digits = 3)),
            color = "black", size = 3.3) +
  theme_bw(base_size = 12) +
  scale_color_gradient2(name ="log2TN Median",low="#6A90CA", mid = 'white', high="#CD2836") +
  labs(title = '',
       subtitle = '',
       x = 'Tumor Purity Group', y = 'Immune Cell')+
  theme(axis.title = element_text(face = "bold", size = 13))
p5i
```



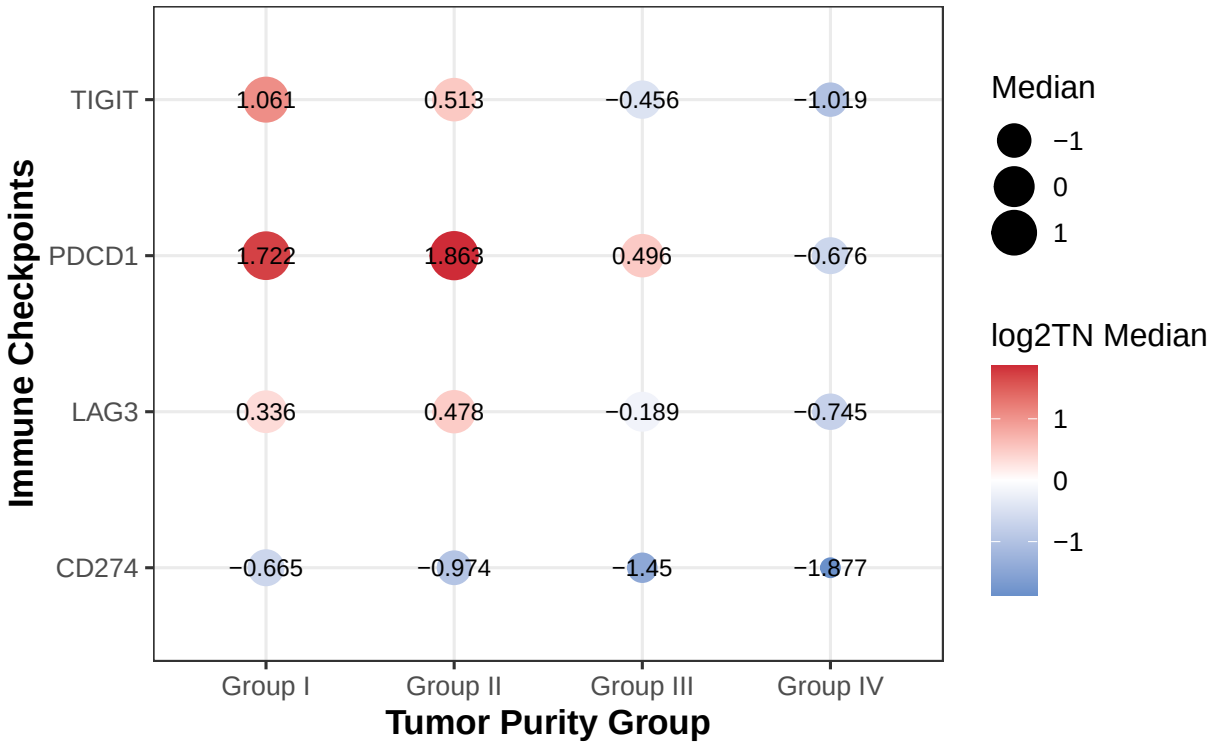
<Figure.2 Immune cell>

Immune Checkpoints

PD-1(PDCD1) ,PD-L1(CD274), LAG3, TIGIT

```
immune_checkpoints <- c("PDCD1", "CD274", "LAG3", "TIGIT")

p5_group1.cp <- p5_group1 %>% filter(gene %in% immune_checkpoints) %>% select(gene, Median) %>% mutate(
p5_group2.cp <- p5_group2 %>% filter(gene %in% immune_checkpoints) %>% select(gene, Median) %>% mutate(
p5_group3.cp <- p5_group3 %>% filter(gene %in% immune_checkpoints) %>% select(gene, Median) %>% mutate(
p5_group4.cp <- p5_group4 %>% filter(gene %in% immune_checkpoints) %>% select(gene, Median) %>% mutate(

p5_total.cp = rbind.data.frame(p5_group1.cp, p5_group2.cp, p5_group3.cp, p5_group4.cp)
```

```r
p5c = ggplot(data = p5_total.cp, aes(x = group, y = gene, col = Median)) +
  geom_point(aes(size = Median), fill = "#1b98e0", stroke = 2) +
  geom_text(data= p5_total.cp, mapping = aes(x = group, y = gene, label = round(Median, digits = 3)), c
  theme_bw(base_size = 12) +
  scale_color_gradient2(name ="log2TN Median",low="#6A90CA", mid = 'white', high="#CD2836") +
  labs(title = '',
       subtitle = '',
       x = 'Tumor Purity Group', y = 'Immune Checkpoints') +
  theme(axis.title = element_text(face = "bold", size = 13))

p5c
```



<Figure.3 Immune Checkpoints>

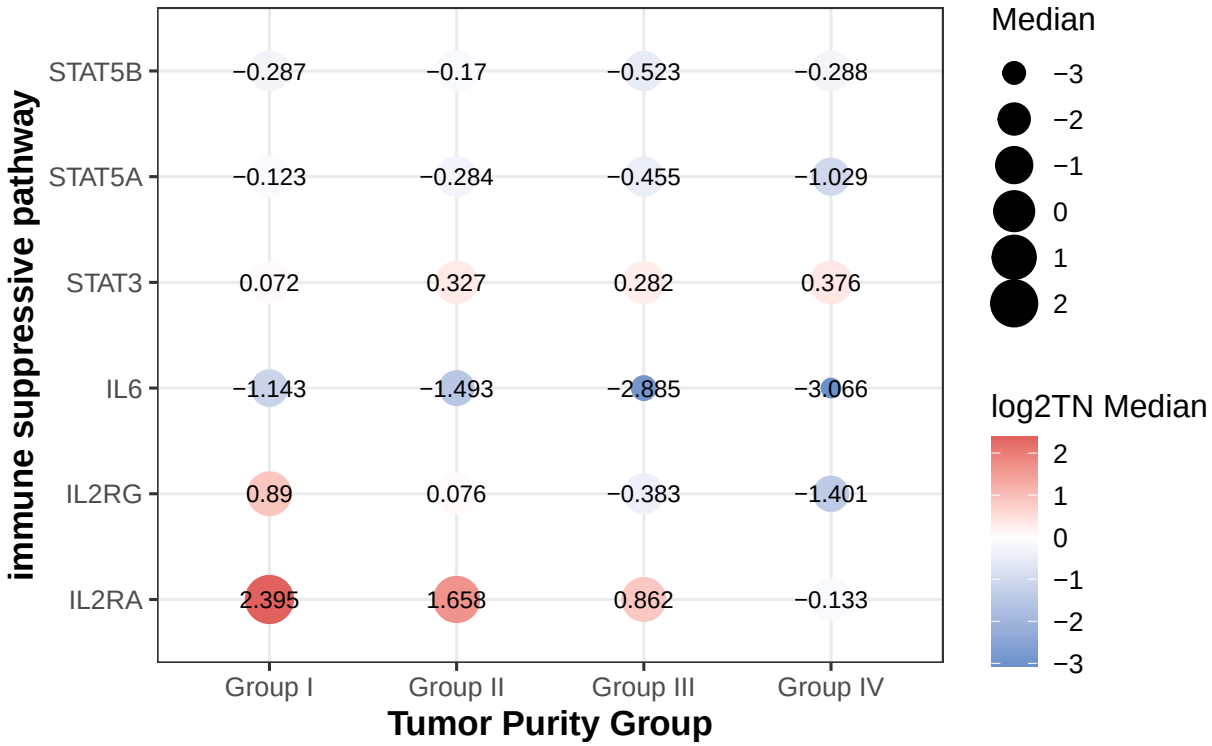
Immune surpressive pathway

Gene List IL6-STAT5 (IL6, STAT5), IL2-JAK-STAT5(STAT5A, STAT5B, IL2RA,IL2RG)

```r
### Immune suppressive pathway
immune_suppressive_pathway <- c("IL6", "STAT3", "STAT5A", "STAT5B", "IL2RA","IL2RG")



p5_group1.sp <- p5_group1 %>% filter(gene %in% immune_suppressive_pathway) %>% select(gene, Median) %>%
p5_group2.sp <- p5_group2 %>% filter(gene %in% immune_suppressive_pathway) %>% select(gene, Median) %>%
p5_group3.sp <- p5_group3 %>% filter(gene %in% immune_suppressive_pathway) %>% select(gene, Median) %>%
```

```
p5_group4.sp <- p5_group4 %>% filter(gene %in% immune_suppressive_pathway) %>% select(gene, Median) %>%

p5_total.sp = rbind.data.frame(p5_group1.sp, p5_group2.sp, p5_group3.sp, p5_group4.sp)
```

```
p5s = ggplot(data = p5_total.sp, aes(x = group, y = gene, col = Median)) +
  geom_point(aes(size = Median), fill = "#1b98e0", stroke = 2) +
  geom_text(data= p5_total.sp, mapping = aes(x = group, y = gene, label = round(Median, digits = 3)),
            color = "black", size = 3.3) +
  theme_bw(base_size = 12) +
  scale_color_gradient2(name ="log2TN Median",low="#6A90CA", mid = 'white', high="#CD2836") +
  labs(title = '',
       subtitle = '',
       x = 'Tumor Purity Group', y = 'immune suppressive pathway')+
  theme(axis.title = element_text(face = "bold", size = 13))

p5s
```

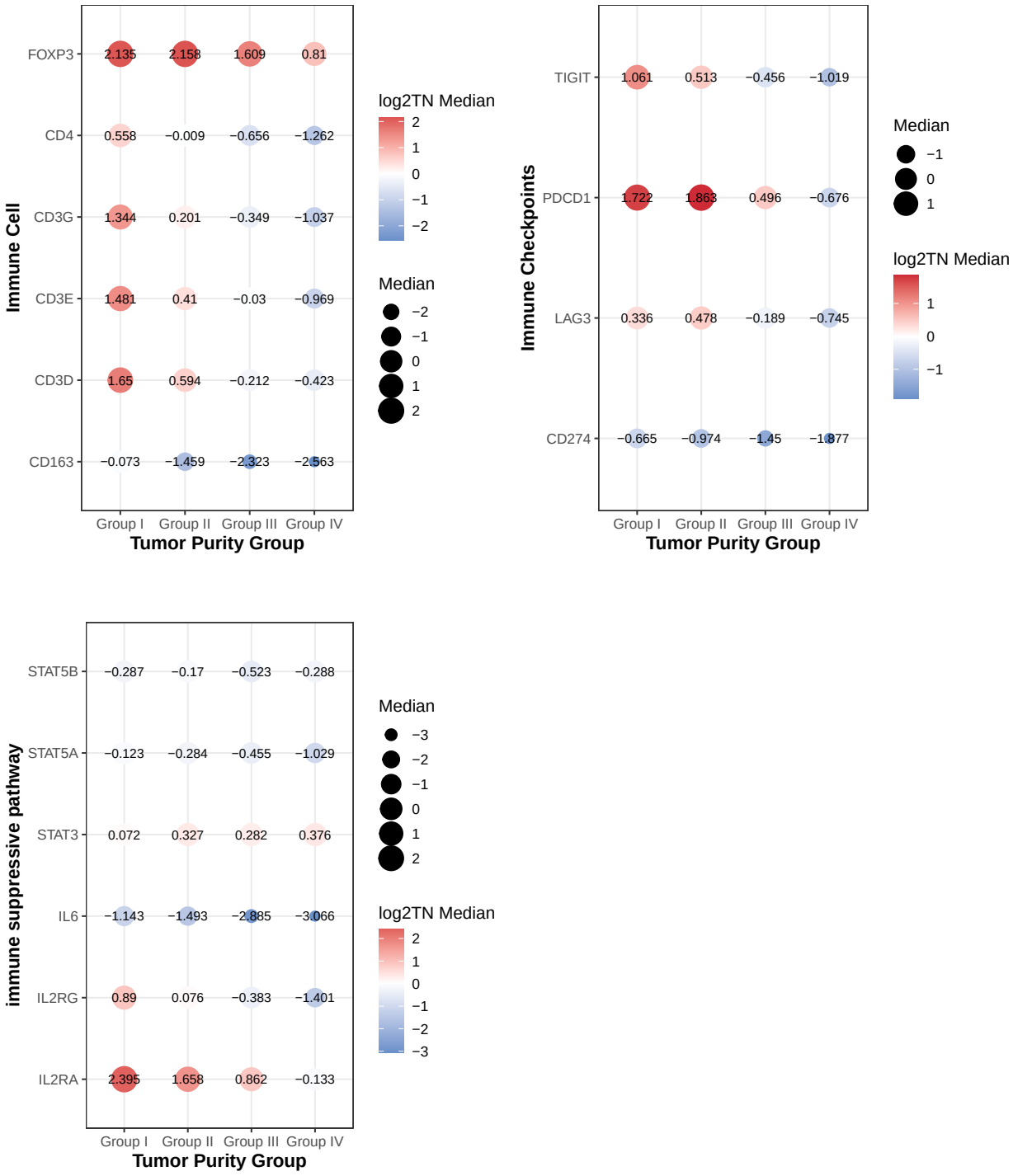


<Figure.4 Immune surpressive pathway>

2.2.5 Combine graphs together using cowplot

```
library(grid)
library(ggplot2)
library(cowplot)
```

```
plot <- plot_grid(p5i,  p5c ,p5s)
plot
```



<Figure. 5>

```
ggsave('plot.relationship_betweeen_tumor_purity_and_immune_subtypes.pdf', plot, width = 12, height = 10)
```

2.2.6 Checking the tendency by dividing the expression amount of the "PDCD" gene by patient stage.

Since "pdcd" is considered to have the largest difference in expression amount according to tumor purity, this study intend to confirm the tendency of the patient by stage.

Load data

```
p2 = read_excel('portfolio/1-s2.0-S0092867420307431-mmc1.xlsx', sheet=2)
```

```
purity_T <- p[1:103,]
purity_T$ID <- sapply(strsplit(as.character(purity_T$`Patient.tissue sample`),"[.]"), "[", 1)
```

```
p2_stage <- p2 %>% select(ID, Stage)
```

```
Stage_purity <- merge (p2_stage,purity_T, by = "ID" , all = F)
```

```
p5_pdcd <- p5 %>% filter(gene == "PDCD1")
```

```
p5_pdcd_t <- data.frame(t(p5_pdcd))
```

```
p5_pdcd_t$ID <- rownames(p5_pdcd_t)
```

```
Stage_purity_PDCD <- merge(p5_pdcd_t,Stage_purity, by = "ID" , all = F)
```

```
Stage_purity_PDCD %>% group_by(Stage) %>% count(Stage)
```

```
## # A tibble: 7 x 2
## # Groups:   Stage [7]
##   Stage     n
##   <chr> <int>
## 1 IA       41
## 2 IB       31
## 3 IIA       5
## 4 IIB       1
## 5 IIIA      7
## 6 IIIB      1
## 7 IV        4
```

```
colnames(Stage_purity_PDCD)[2]<-"PDCD"
```

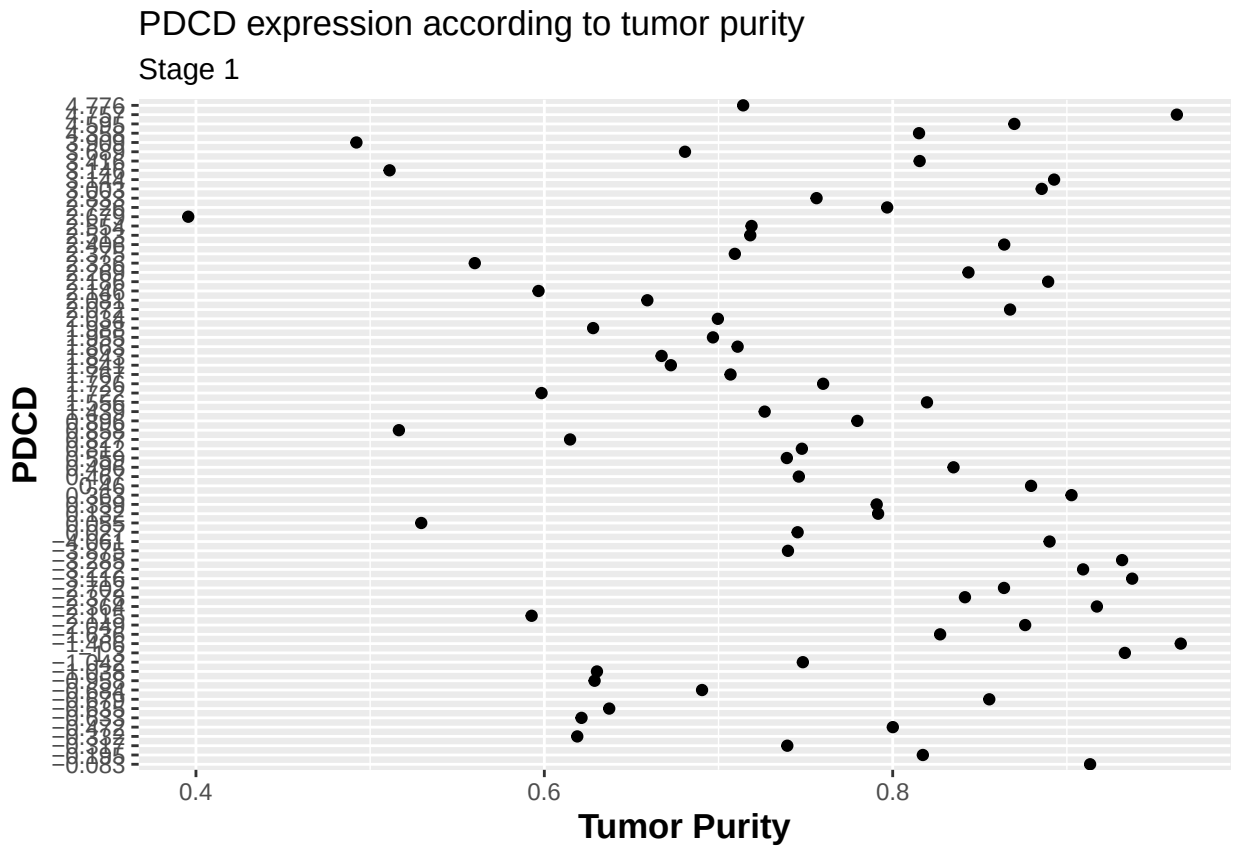
```
stage1 <- c("IA","IB" )
stage2 <- c("IIA","IIB" )
stage3 <- c("IIIA","IIIB" )
stage4 <- c("IV" )
stage_1 <- Stage_purity_PDCD %>% filter(Stage %in% stage1)
stage_2 <- Stage_purity_PDCD %>% filter(Stage %in% stage2)
stage_3 <- Stage_purity_PDCD %>% filter(Stage %in% stage3)
stage_4 <- Stage_purity_PDCD %>% filter(Stage %in% stage4)
```

```
colnames(Stage_purity_PDCD)
```

```
## [1] "ID"                    "PDCD"                  "Stage"
## [4] "Patient.tissue sample" "Stromal Score"         "Immune Score"
## [7] "ESTIMATE Score"        "Tumor Purity"
```

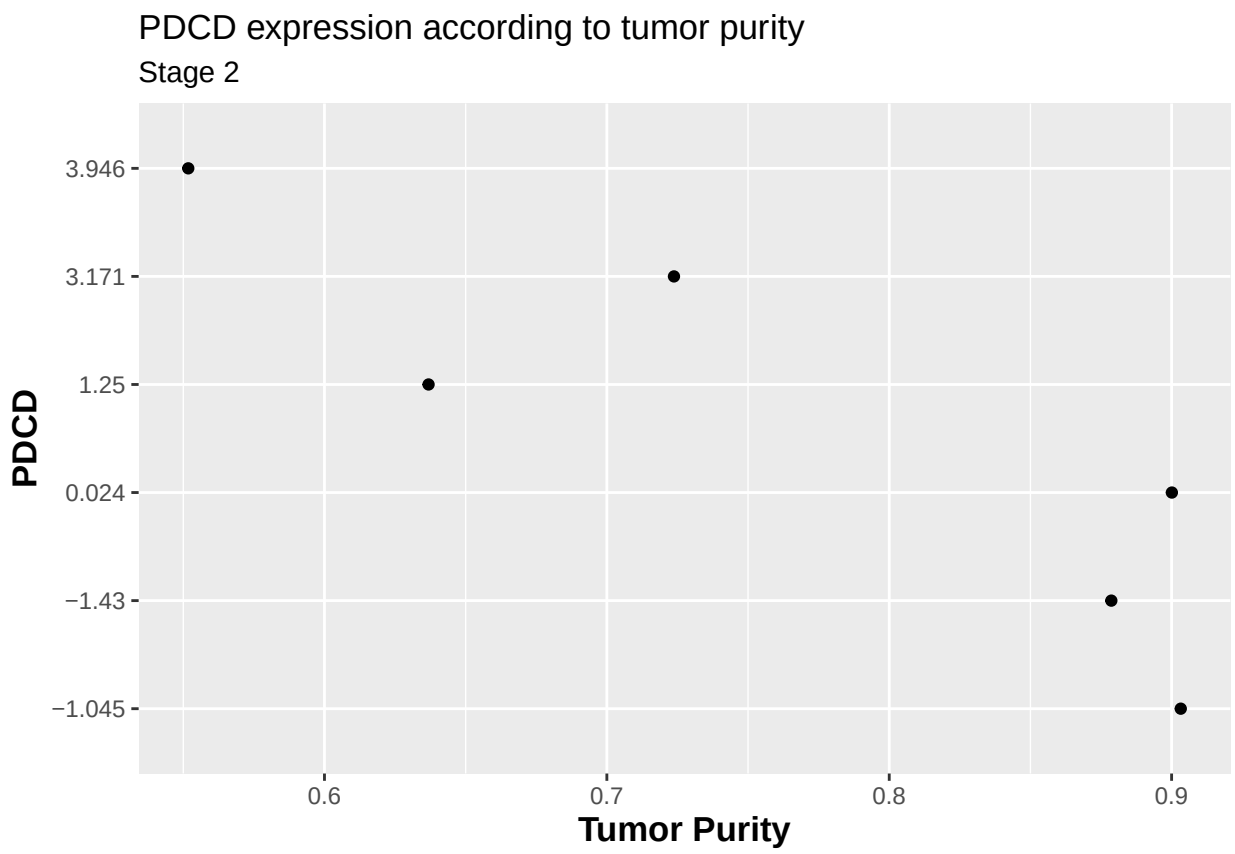
```
stage1_scatter <-ggplot(stage_1, aes(x=`Tumor Purity`, y=PDCD))

s1 <- stage1_scatter+geom_point()+stat_smooth(method=lm)+ geom_smooth()+
  labs(title = 'PDCD expression according to tumor purity',
       subtitle = 'Stage 1')+
  theme(axis.title = element_text(face = "bold", size = 13))
s1
```

```
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



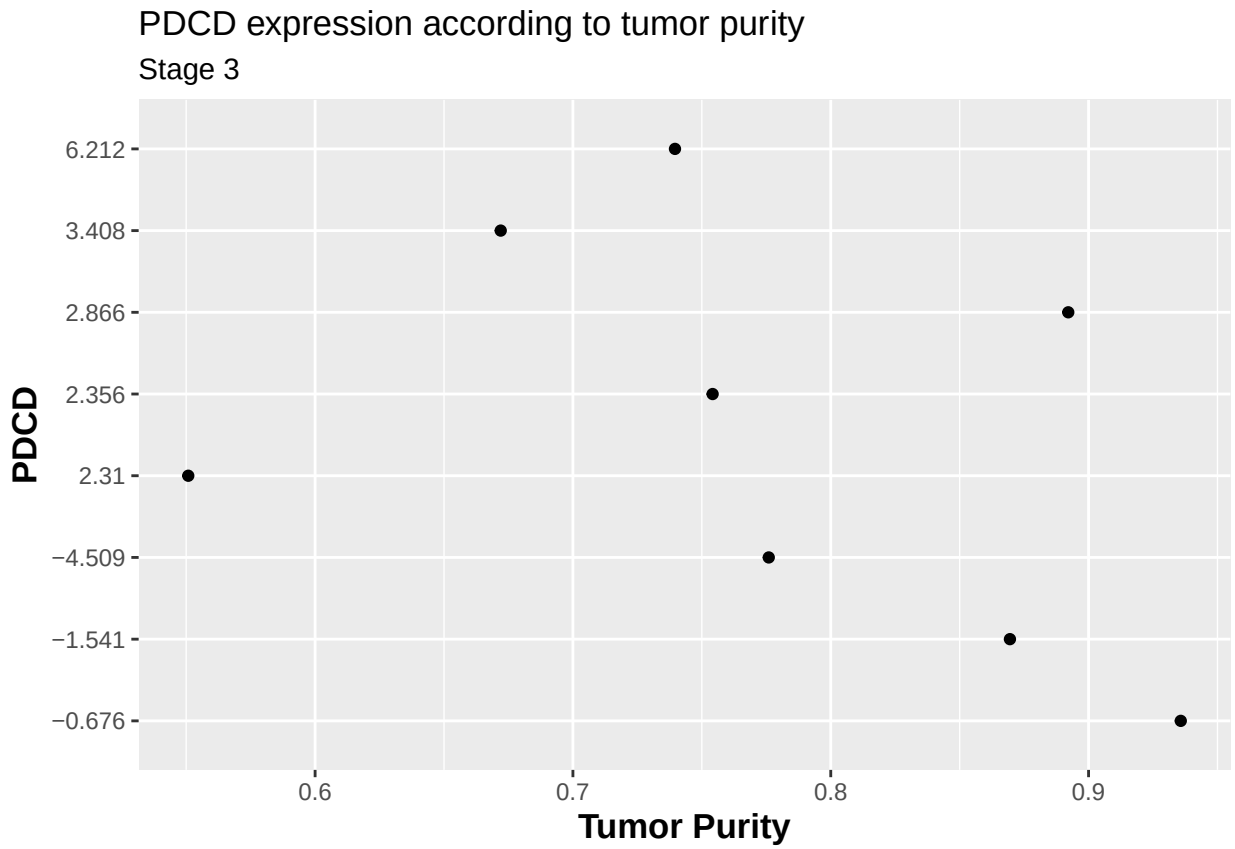

```
stage2_scatter <-ggplot(stage_2, aes(x=`Tumor Purity`, y=PDCD))

s2 <- stage2_scatter+geom_point()+stat_smooth(method=lm) +labs(title = 'PDCD expression according to tu
       subtitle = 'Stage 2')+
  theme(axis.title = element_text(face = "bold", size = 13))
s2
```

```
## `geom_smooth()` using formula 'y ~ x'
```



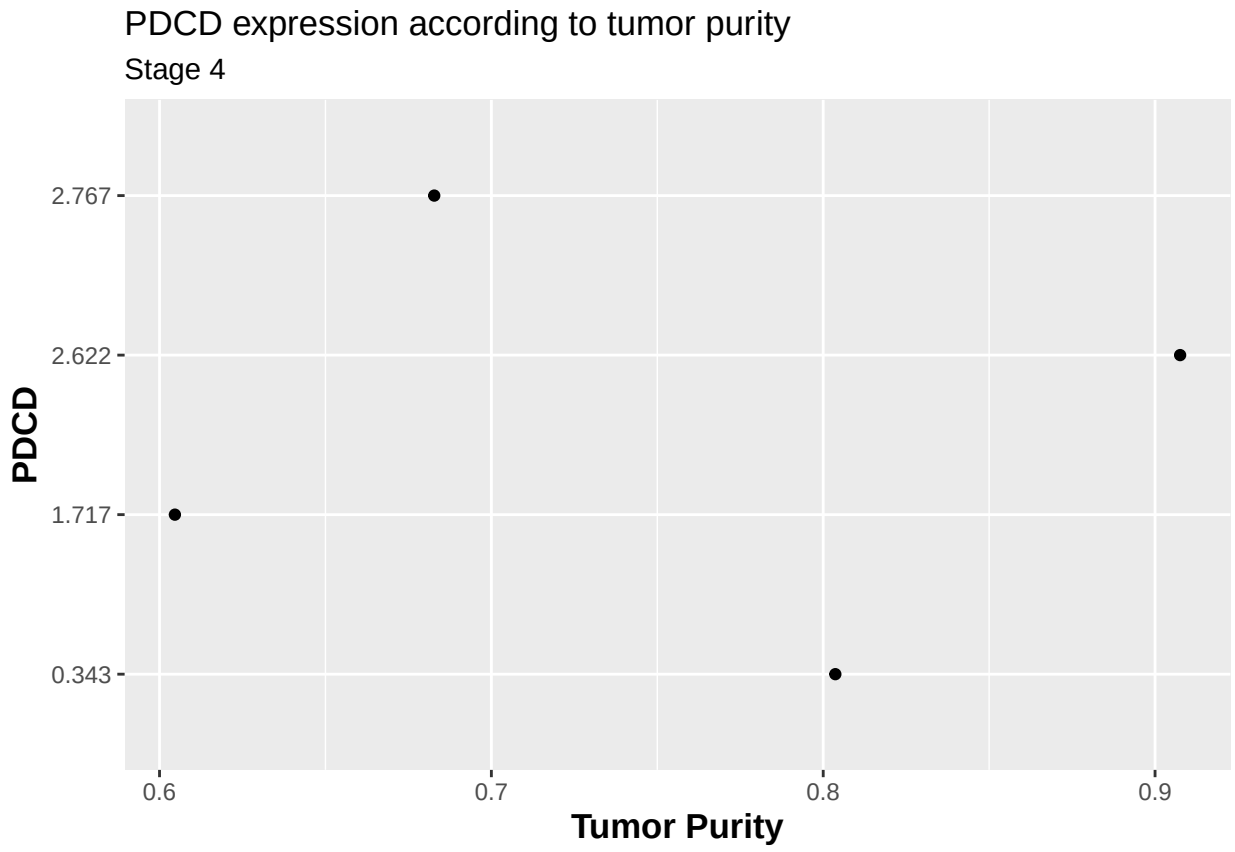

```
stage3_scatter <-ggplot(stage_3, aes(x=`Tumor Purity`, y=PDCD))

s3 <- stage3_scatter+geom_point()+stat_smooth(method=lm)+
  labs(title = 'PDCD expression according to tumor purity',
       subtitle = 'Stage 3')+
  theme(axis.title = element_text(face = "bold", size = 13))
s3
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
stage4_scatter <-ggplot(stage_4, aes(x=`Tumor Purity`, y=PDCD))

s4 <- stage4_scatter+geom_point()+stat_smooth(method=lm) +
  labs(title = 'PDCD expression according to tumor purity',
       subtitle = 'Stage 4')+
  theme(axis.title = element_text(face = "bold", size = 13))
s4
```

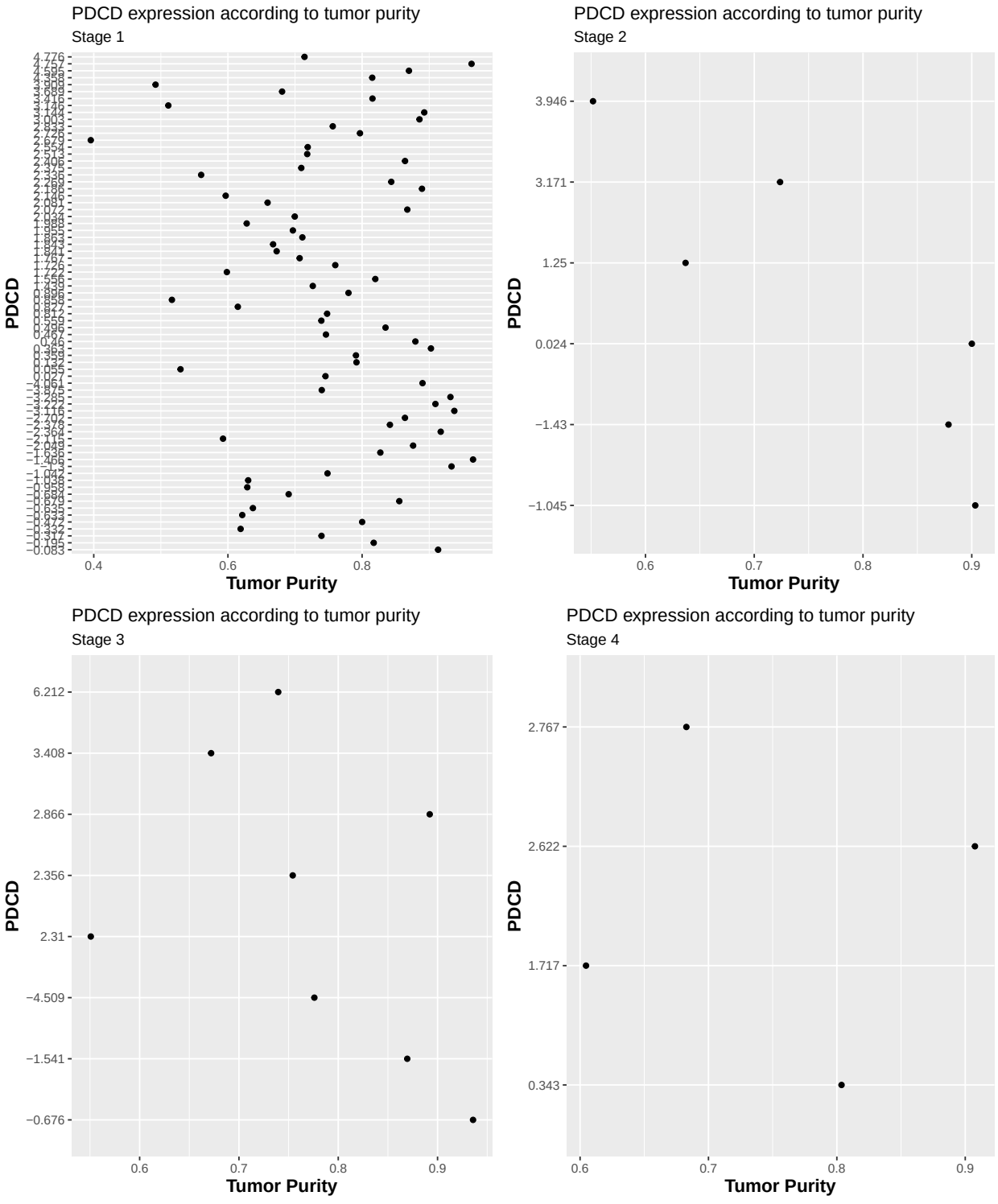
```
## `geom_smooth()` using formula 'y ~ x'
```




```
plot2 <- plot_grid(s1, s2, s3, s4)
```

```
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
plot2
```



In this cohort, it was confirmed that the distribution of patients by stage was uneven and focused on stage 1. Therefore, it seems unreasonable to check whether there is a tendency that varies depending on tumor purity for each stage type.Therefore, it is better to collect additional patient samples or combine other databases with their cohorts to find out the tendency.

3. Result

As a result of visualization, it can be seen that the lower the tumor purity, the greater the expression of the immune system. I initially divided TW LUAD cohort into four groups based on tumor purity,setting low value to group I. And made a boxplot to show the sample size of each group used in the analysis. After that, I filtered the patients to confirm the expression of the gene.

When identifying immune cell genes, T-regs and M2 macrophages marker gene was chosen. Because they are related to pro-tumorigenic inflammation. M2 class have been associated with wound healing and tissue repair, while M1 is associated with mounting defense against infectious agents. Immune checkpoints are regulators of the immune system. These pathways are crucial for immune-tolerance, which prevents the immune system from attacking cells indiscriminately. However, cancers can protect themselves from attack by stimulating immune checkpoint targets. As for the IL6-JAK-STAT3 signaling, prior studies showed it involved in tumor growth, metastasis, and the immune escaping. Based on these facts, the gene of interest was selected and the correlation between tumor purity and immune subtype was confirmed.In all three categories(Immune cell,Immune checkpoint,Immune suppressive pathway), it was confirmed that the lower the tumor purity, the higher the gene expression level.

4 Discussion

To induce an effective immune response that can fight infectious diseases or cancer, cells or active molecules in the innate immune system and the adaptive immune system must interact in harmony. When both immune system cells play a proper role in the very early stages of cancer, cancer cells are removed and returned to their normal state. However, if cancer cells grow too fast for the immune system to handle, the number of surviving cancer cells increases. In this case, cancer cells continue to multiply by deceiving the immune system or taming immune cells not to attack themselves. In most cases, the stage in which cancer is found is difficult for the immune system to easily remove cancer due to the formation of the tumor microenvironment.

In particular, cancer cells protect themselves from immune cells by using adaptive immune response characteristics to their advantage. Lymphocytes begin to be activated from the moment they are activated. That is, the activated T lymphocytes begin to express inhibitory receptors called PD-1 and the degree of expression increases as the stimulation continues. Cancer cells avoid immune responses by expressing a large amount of PD-L1, a ligand that specifically binds to PD-1, and when activated T lymphocytes try to kill themselves, they combine with PD-1 to send a degenerative signal to deactivate cytotoxic T lymphocytes. Recent studies have demonstrated that PD-1 is called an immune checkpoint, and suppression of it can revive inhibited anticancer immune responses. T-reg is known to suppress anti-tumor immune response by impairing cell-mediated immune responses to tumors and further promoting disease progression. The existence of a higher T-reg level may weaken the effect of immunotherapy. Therefore, in order to treat patients with low tumor purity, immunotherapy and other chemotherapy should be combined or different treatments should be used. Originally, macrophages were known to play a role in directly killing cancer cells. But recent studies have shown that tumor-associated macrophages(TAM) are rather infiltrated in a tumor. It helps the proliferation of cancer cells and helps cancer cells in the metastatic stage easily pass through the blood vessel wall. That means the existence of higher T-regs and M2 macrophages may weaken the effect of immunotherapy. Therefore,anti-Tregs and anti-M2 macrophages combined with anti-PD-1/PD-L1 therapy may be a better choice for low-purity group patients.

According to the analysis results, we could find out the relationship between tumor purity and immune subtypes that are involved in immune escaping. In a word, our study revealed that tumor purity plays an important role in prediction of the immune state in lung cancer. Low purity in lung cancer was associated with immune systems, which indicated that low-purity lung cancer patients may benefit more from immunotherapy. Further investigations need to be performed on tumor purity to get a better comprehension of TME and make a better clinical decision.

5 Reference

1. Chen, Y. J., Roumeliotis, T. I., Chang, Y. H., Chen, C. T., Han, C. L., Lin, M. H., ... & Chen, Y. J. (2020). Proteogenomics of non-smoking lung cancer in East Asia delineates molecular signatures of pathogenesis and progression. Cell, 182(1), 226-244.
2. Gong, Z, Zhang, J, Guo, W. Tumor purity as a prognosis and immunotherapy relevant feature in gastric cancer. Cancer Med. 2020; 9: 9052- 9063.
3. Germano, G., Frapolli, R., Belgiovine, C., Anselmo, A., Pesce, S., Liguori, M., ... & Allavena, P. (2013). Role of macrophage targeting in the antitumor activity of trabectedin. Cancer cell, 23(2), 249-262.
4. Weinberg, R. A. (2013). The biology of cancer. Garland science. 5.Martinez, F. O., & Gordon, S. (2014). The M1 and M2 paradigm of macrophage activation: time for reassessment. F1000prime reports, 6.