First TCGA analysis

Ariel Hippen

2023-01-11

Contents

Having run the TCGA RNA-seq data through BayesPrism, this notebook compares the samples' cell type composition with their subtype annotations from the Way pipeline and the patients' survival status/time.

```
suppressPackageStartupMessages({
    library(data.table)
    library(SingleCellExperiment)
    library(dplyr)
    library(yaml)
    library(stringr)
    library(ggplot2)
    library(survival)
    library(ggfortify)
})

params <- read_yaml("../../config.yml")
data_path <- params$data_path
local_data_path <- params$local_data_path
plot_path <- params$plot_path
```

```
tcga <- fread(paste(local_data_path, "deconvolution_output",
                    "TCGA_default_bayesprism_results.tsv", sep = "/"))
tcga_melt <- melt(tcga)
```

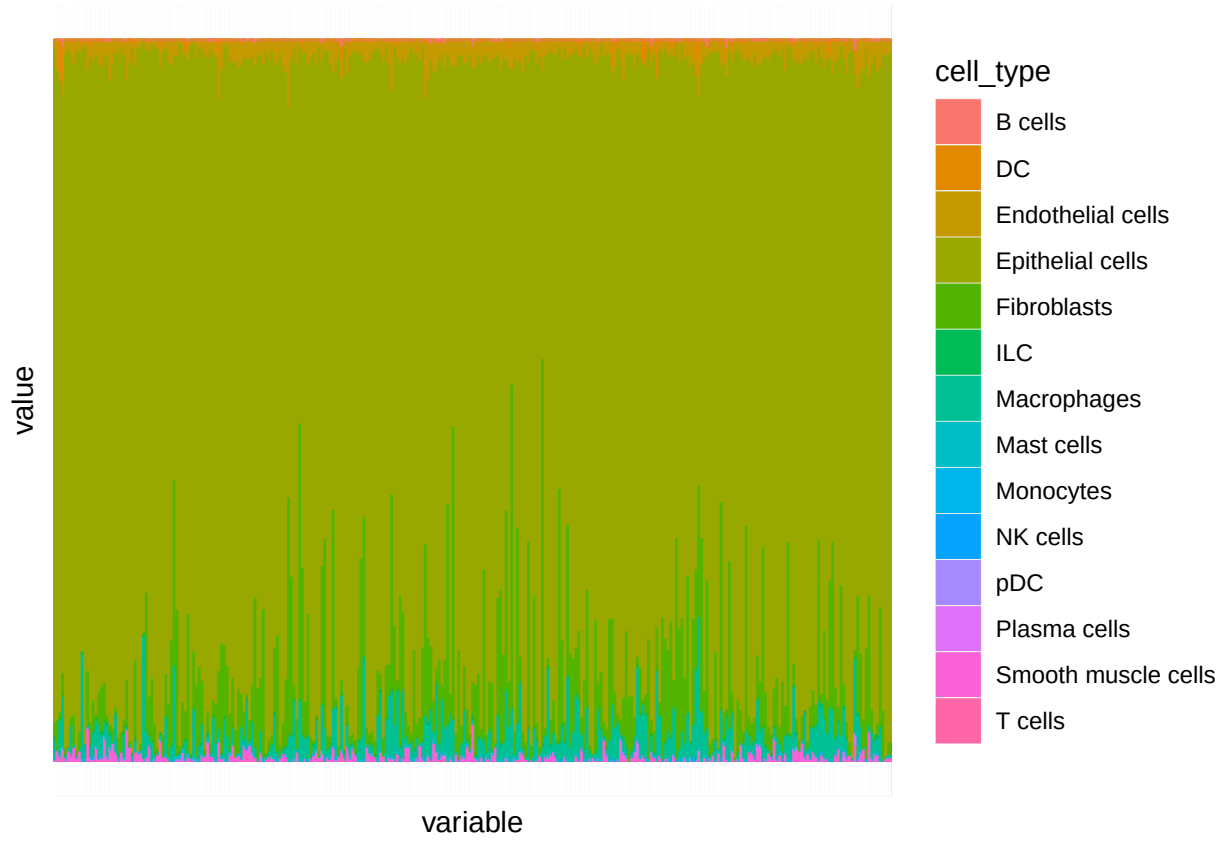
```
## Warning in melt.data.table(tcga): id.vars and measure.vars are internally
## guessed when both are 'NULL'. All non-numeric/integer/logical type columns are
## considered id.vars, which in this case are columns [cell_type, ...]. Consider
## providing at least one of 'id' or 'measure' vars in future.
```

Cell composition

```
g <- ggplot(tcga_melt, mapping = aes(x=variable, y=value, fill=cell_type, color=cell_type)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x=element_blank(), #remove x axis labels
        axis.ticks.x=element_blank(), #remove x axis ticks
        axis.text.y=element_blank(),  #remove y axis labels
        axis.ticks.y=element_blank()) #remove y axis ticks

plotfile <- paste(plot_path, "evaluation_plots",
                  "TCGA_proportion_barchart.png", sep = "/")
png(filename = plotfile, width = 1200)
g
dev.off()
```

```
## pdf
##   2
```

```
g
```




```
# Switch so cell types are columns and samples are rows for easier analysis
cell_types <- tcga$cell_type

tcga$cell_type <- NULL
tcga_t <- t(as.matrix(tcga))
colnames(tcga_t) <- cell_types
tcga_t <- as.data.frame(tcga_t)
```

Cell composition by survival

```
# Load survival data
tcga_survival <- fread(paste(local_data_path, "TCGA", "TCGA_OV_survival.tsv",
                             sep = "/"))
tcga_patients <- str_extract(colnames(tcga), "TCGA-\\w\\w-\\w\\w\\w\\w")
tcga_survival <- subset(tcga_survival, tcga_survival$bcr_patient_barcode %in%
                           tcga_patients)
```

```
# Combine survival data with %
tcga_t$bcr_patient_barcode <- tcga_patients
tcga_master <- full_join(tcga_survival, tcga_t)
```

```
## Joining, by = "bcr_patient_barcode"
```

```
tcga_master$Immune <- tcga_master$Macrophages + tcga_master$Monocytes + tcga_master$`Plasma cells` +
  tcga_master$DC + tcga_master$`NK cells` + tcga_master$pDC + tcga_master$`B cells` + tcga_master$ILC +
  tcga_master$`Mast cells`
```

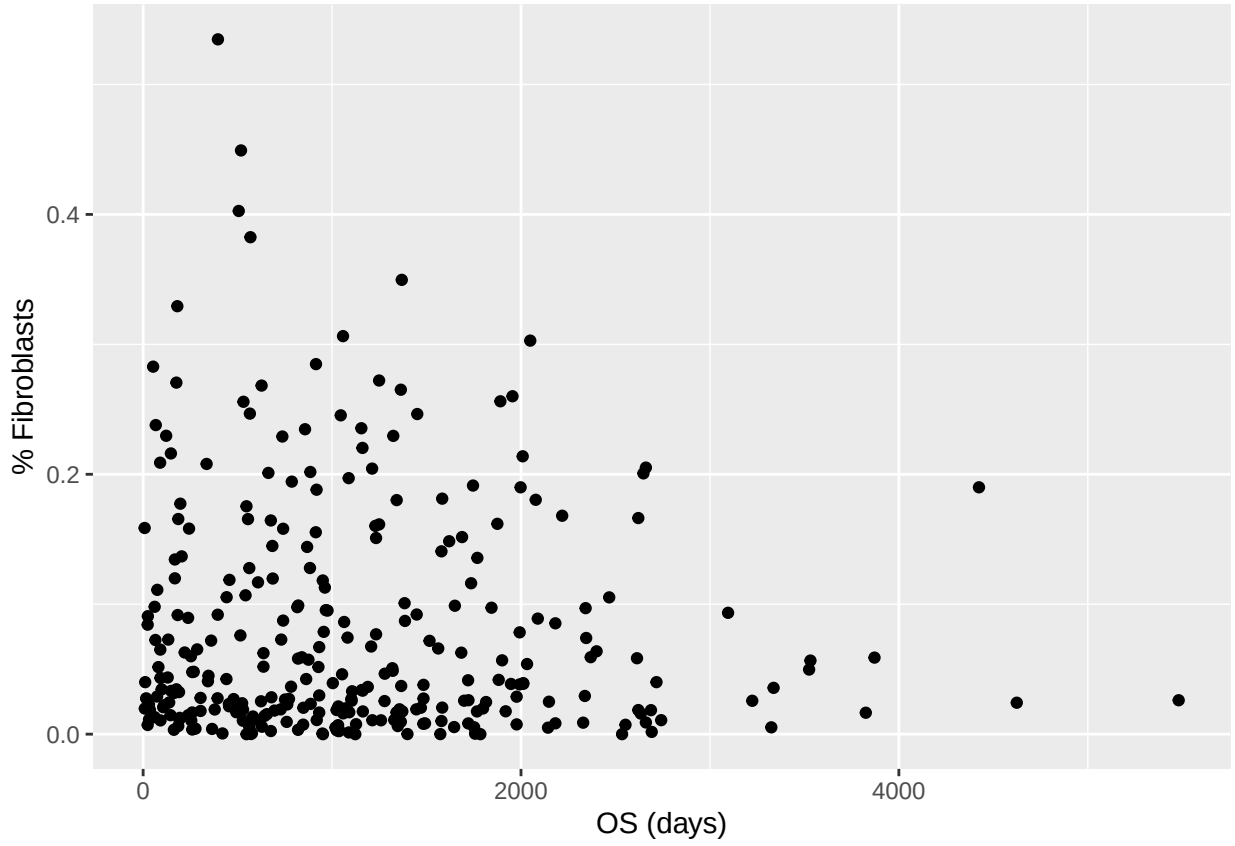
```
g <- ggplot(tcga_master, mapping = aes(x=tcga_master$OS.time, y=tcga_master$Fibroblasts)) +
  geom_point() + xlab("OS (days)") + ylab("% Fibroblasts")
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_survival_by_fibroblasts.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
## pdf
##   2
```

```
g
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



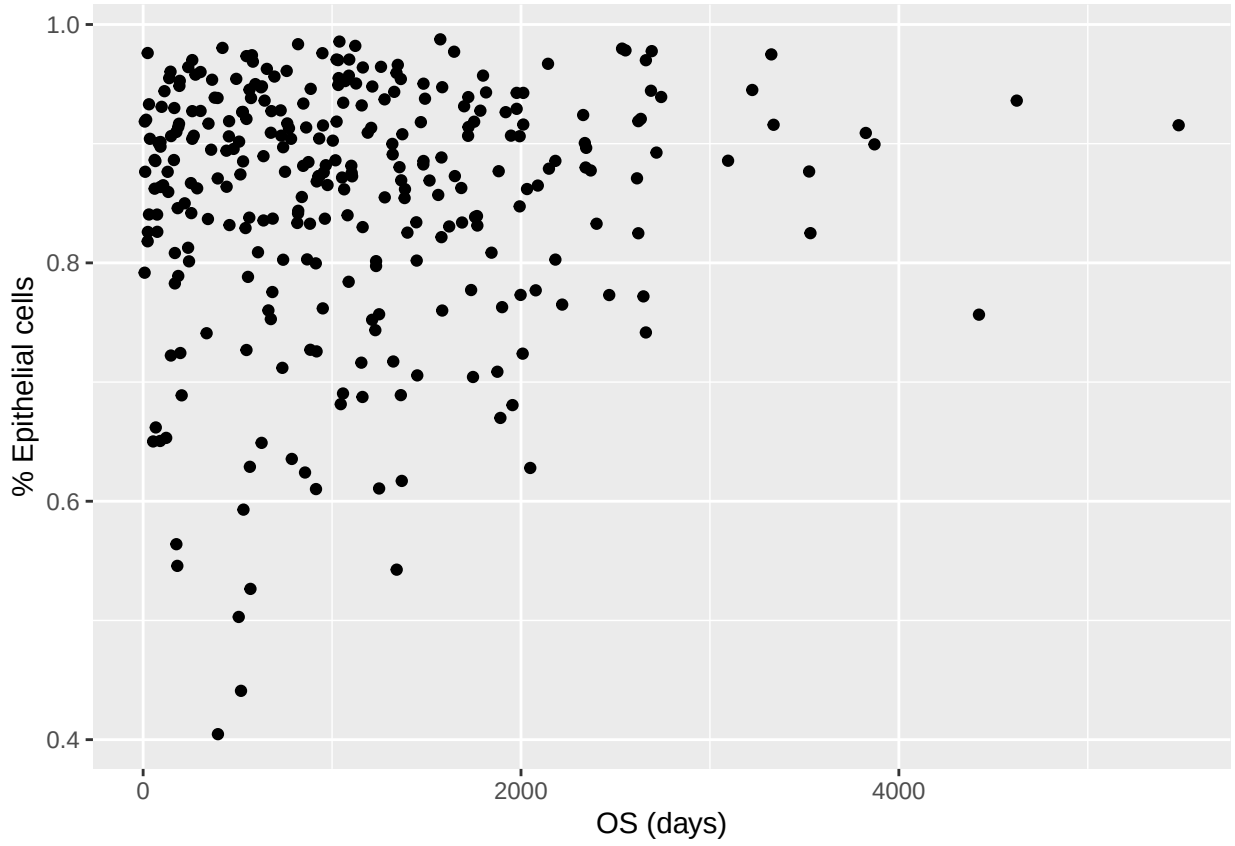

```
g <- ggplot(tcga_master, mapping = aes(x=tcga_master$OS.time, y=tcga_master$`Epithelial cells`)) +
  geom_point() + xlab("OS (days)") + ylab("% Epithelial cells")
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_survival_by_epithelial.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
## pdf
##   2
```

```
g
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



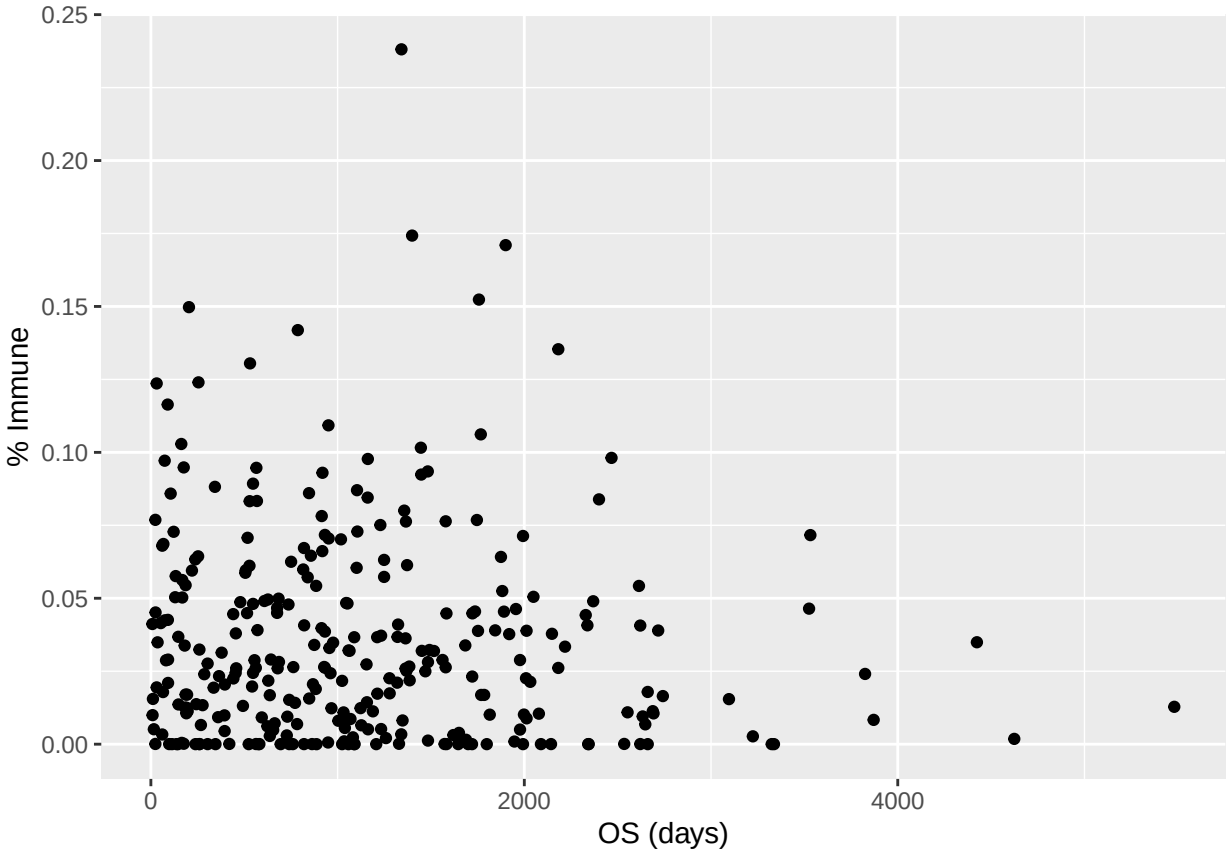

```
g <- ggplot(tcga_master, mapping = aes(x=tcga_master$OS.time, y=tcga_master$Immune)) +
  geom_point() + xlab("OS (days)") + ylab("% Immune")
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_survival_by_immune.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
## pdf
##   2
```

```
g
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



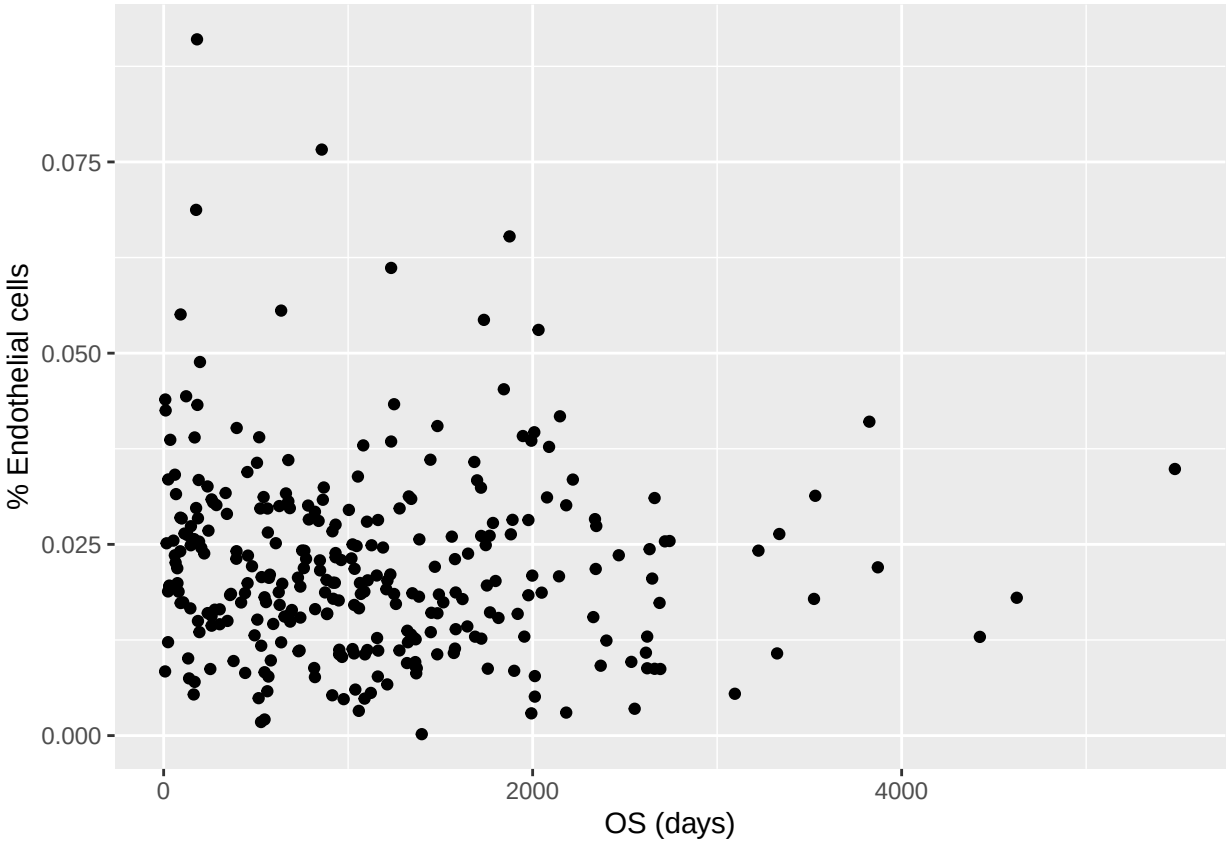

```
g <- ggplot(tcga_master, mapping = aes(x=tcga_master$OS.time, y=tcga_master$`Endothelial cells`)) +
  geom_point() + xlab("OS (days)") + ylab("% Endothelial cells")
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_survival_by_endothelial_cells.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
## pdf
##   2
```

```
g
```

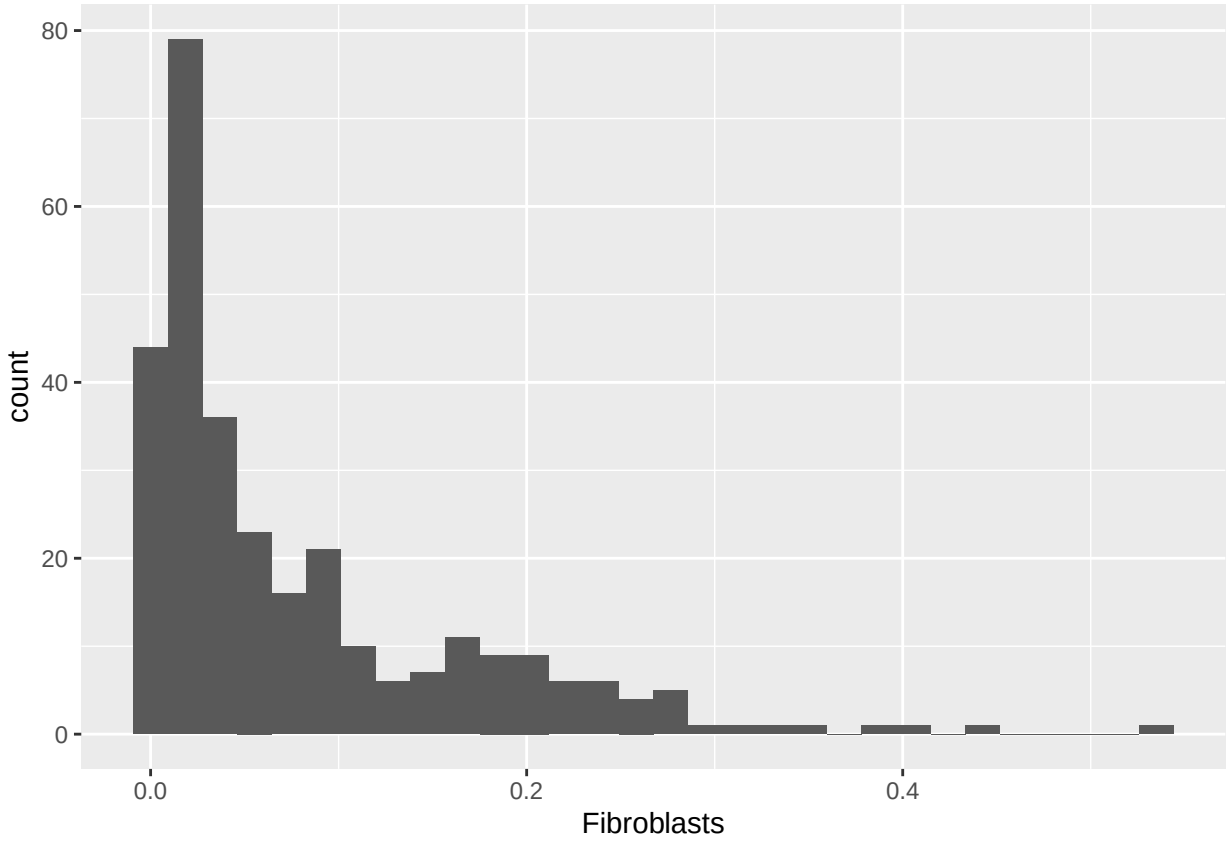
```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Kaplan Meier curves

```
# Put the samples into quartiles based on fibroblast content
ggplot(tcga_master, mapping = aes(x=Fibroblasts)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

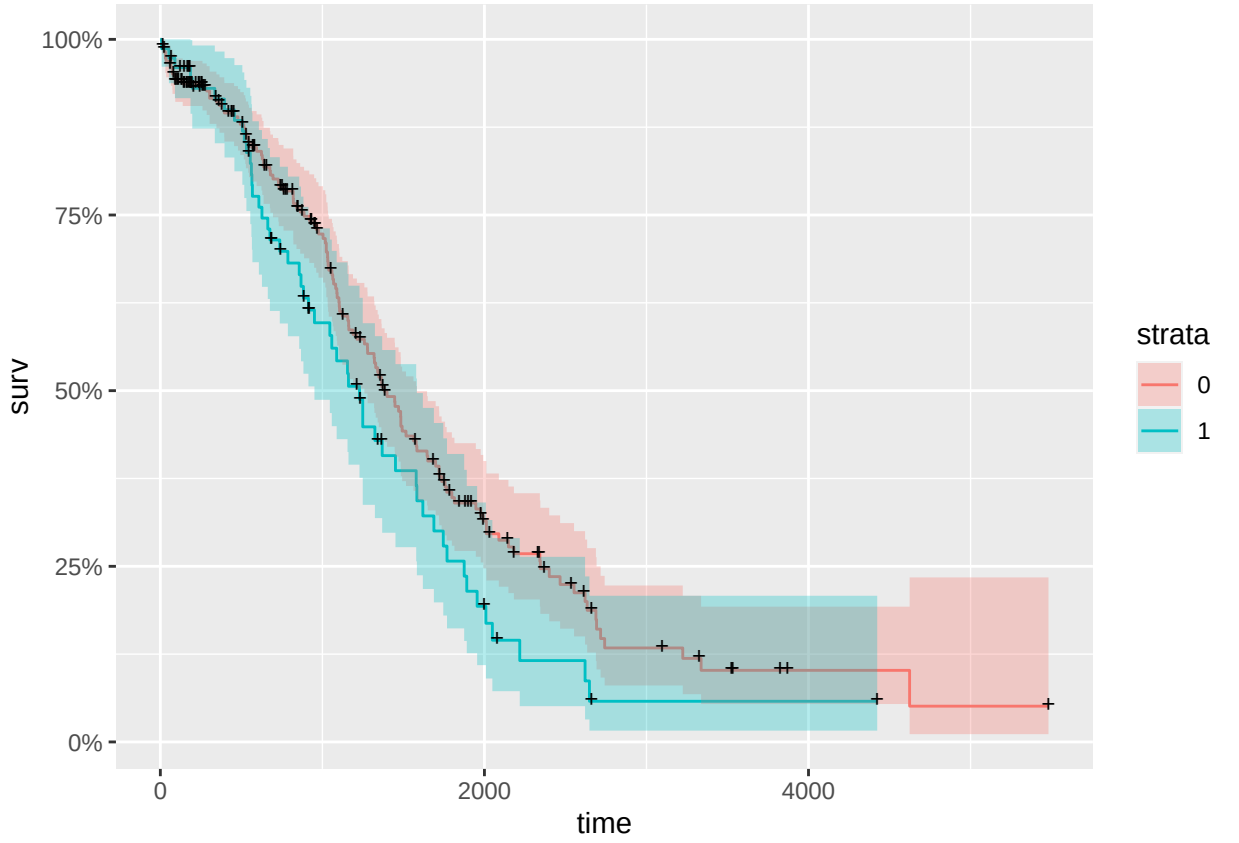



```
quantiles <- quantile(tcga_master$Fibroblasts)
q1 <- quantiles[2]
q3 <- quantiles[4]
tcga_master$high_fibro <- ifelse(tcga_master$Fibroblasts > q3, 1, 0)
```

```
# Get Kaplan-Meier curves
km <- Surv(time = tcga_master$OS.time,  event = tcga_master$OS)
km_treatment<-survfit(km~high_fibro,data=tcga_master,type='kaplan-meier',conf.type='log')

plotfile <- paste(plot_path, "evaluation_plots", "TCGA_KaplanMeier_fibroblasts.png", sep = "/")
png(filename = plotfile)
autoplot(km_treatment)
dev.off()
```

```
## pdf
##   2
```

```
autoplot(km_treatment)
```



Subtypes

```
# Get subtype annotations
cluster_file <- paste(local_data_path, "cluster_assignments", "FullClusterMembership.csv", sep = "/")
cluster_list <- fread(cluster_file)

cluster_list$V1 <- gsub("\\.", "-", cluster_list$V1)
setnames(cluster_list, "V1", "ID")

tcga_t$ID <- str_extract(rownames(tcga_t), "TCGA-\\w\\w-\\w\\w\\w\\w")

tcga_t <- left_join(tcga_t, cluster_list)
```

```
## Joining, by = "ID"
```

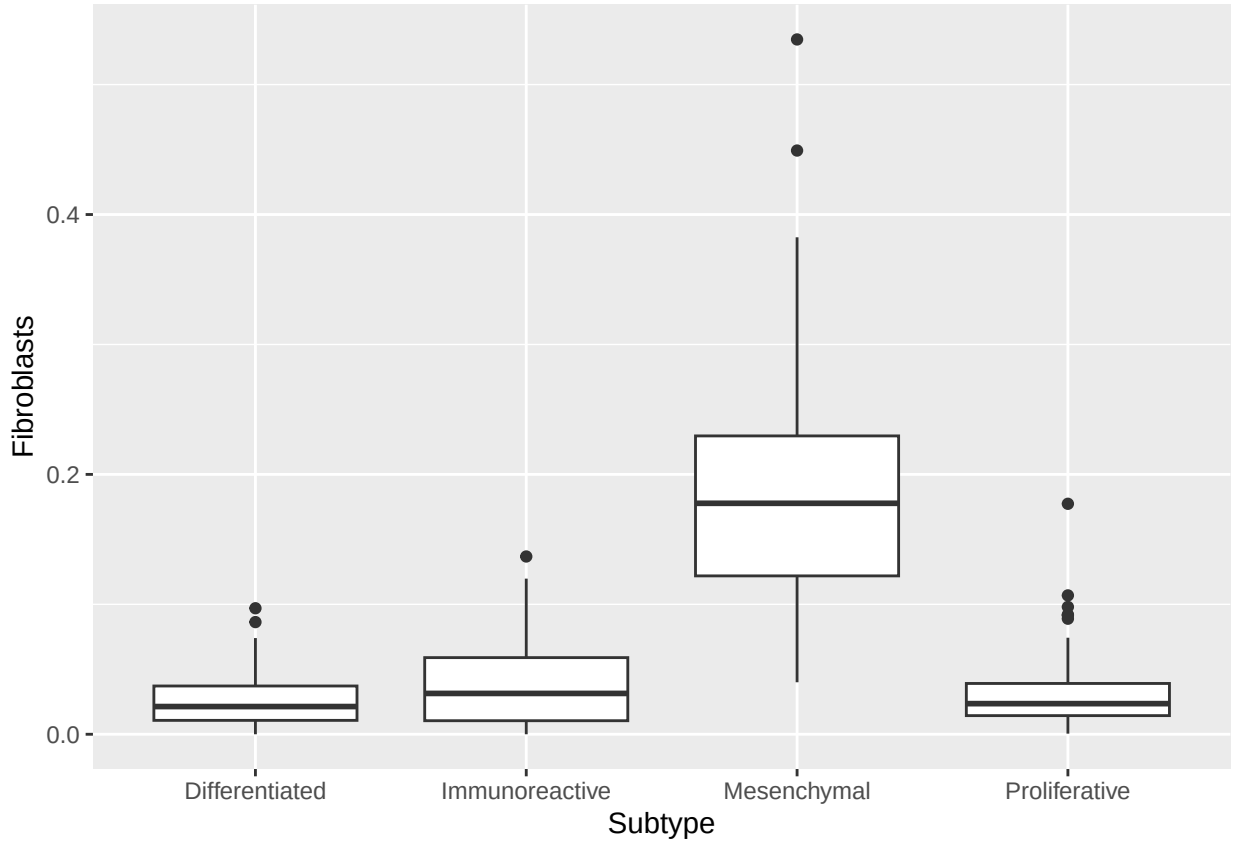
```
tcga_t$Subtype <- recode(tcga_t$ClusterK4_kmeans,
                         "1" = "Mesenchymal",
                         "2" = "Proliferative",
                         "3" = "Immunoreactive",
                         "4" = "Differentiated")

tcga_t$Immune <- tcga_t$`T cells` + tcga_t$Macrophages + tcga_t$Monocytes + tcga_t$`Plasma cells` +
  tcga_t$DC + tcga_t$`NK cells` + tcga_t$pDC + tcga_t$`B cells` + tcga_t$ILC + tcga_t$`Mast cells`
```

```
# Get rid of samples that don't have a subtype label
tcga_t <- subset(tcga_t, !is.na(tcga_t$Subtype))
```

```
# Compare cell type proportions of subtypes
g <- ggplot(tcga_t, mapping = aes(x=Subtype, y=Fibroblasts)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_fibroblasts_by_subtype.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

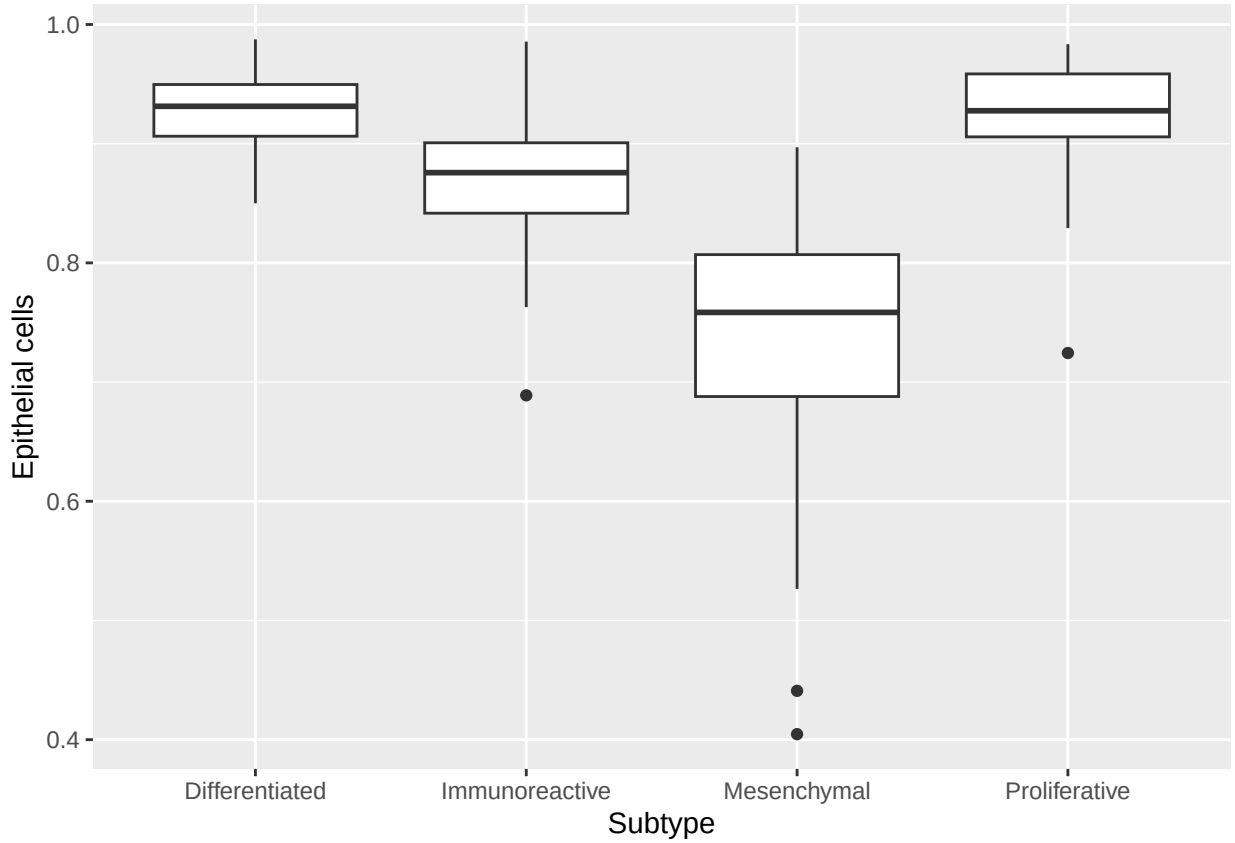
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=Subtype, y=`Epithelial cells`)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_epithelial_by_subtype.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

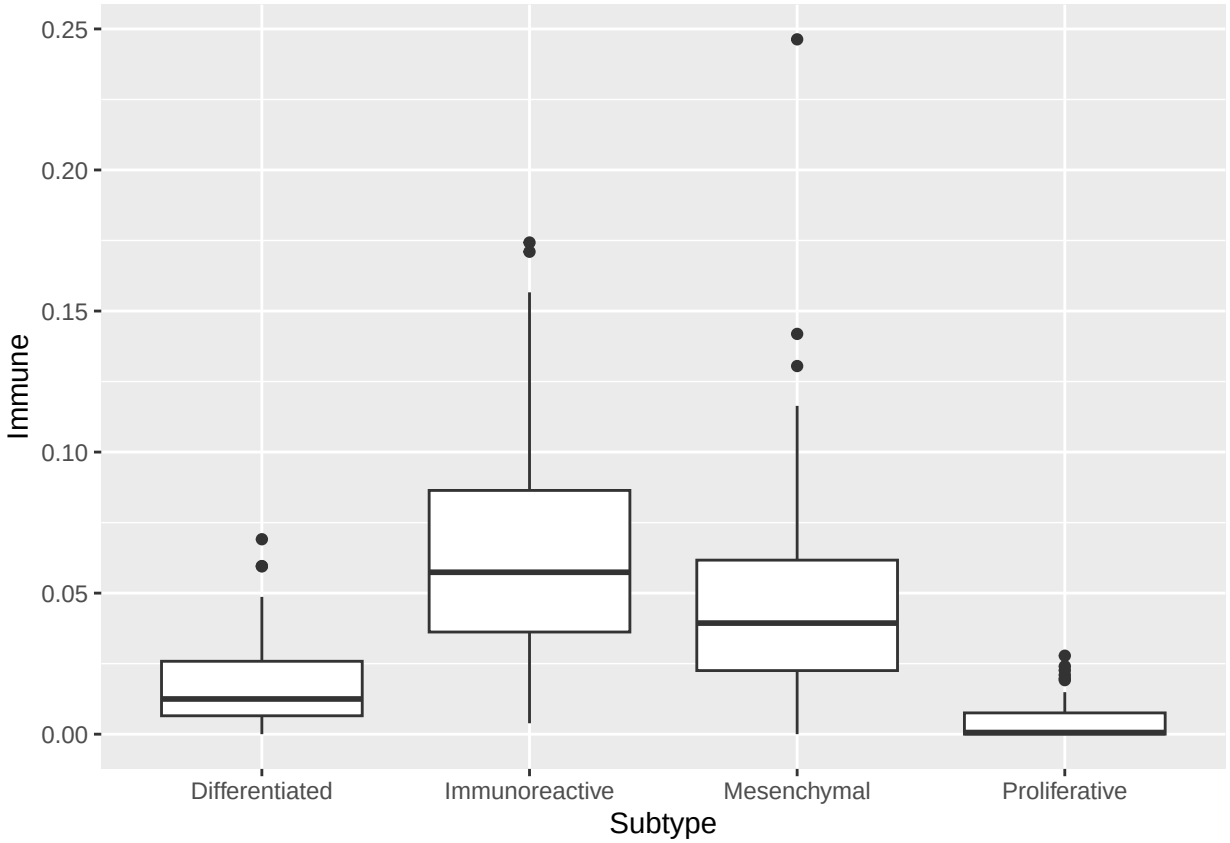
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=Subtype, y=Immune)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_immune_by_subtype.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

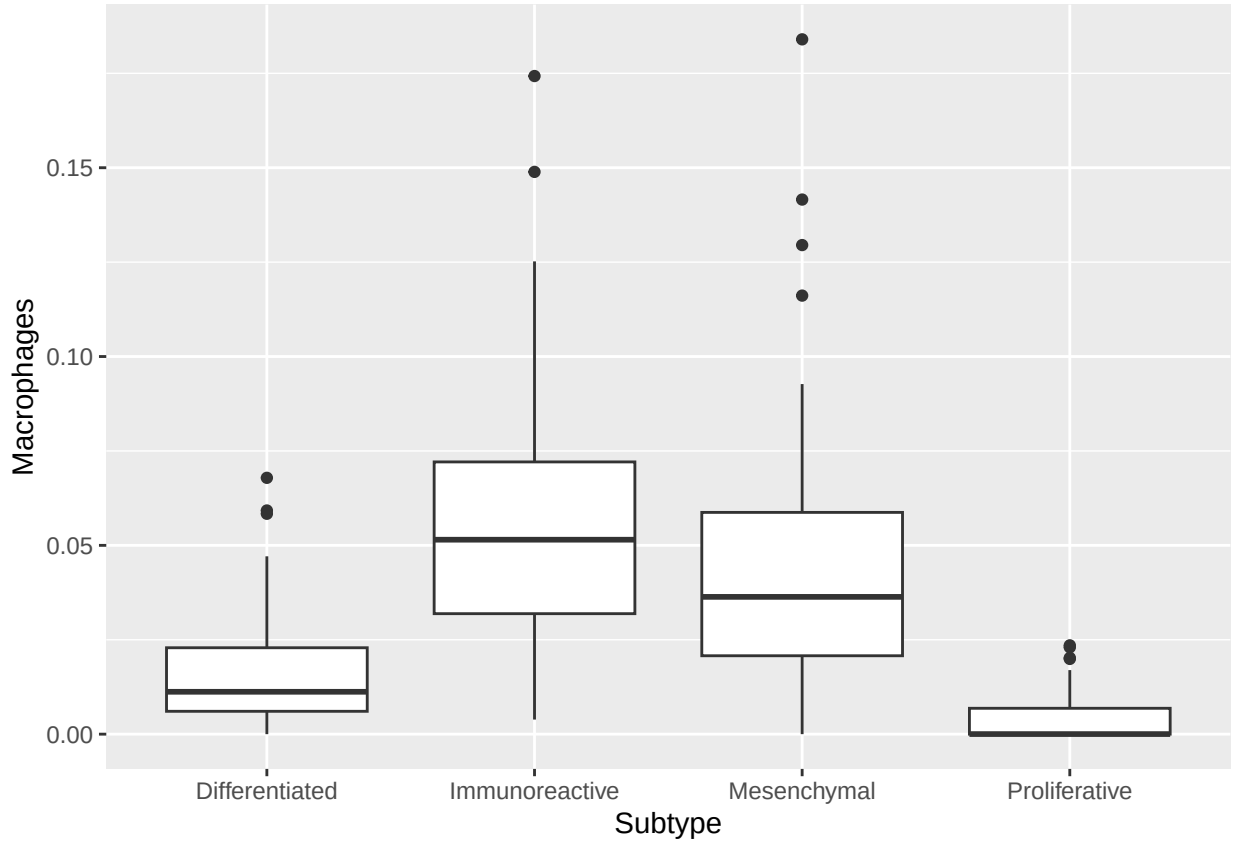
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=Subtype, y=Macrophages)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_macrophages_by_subtype.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

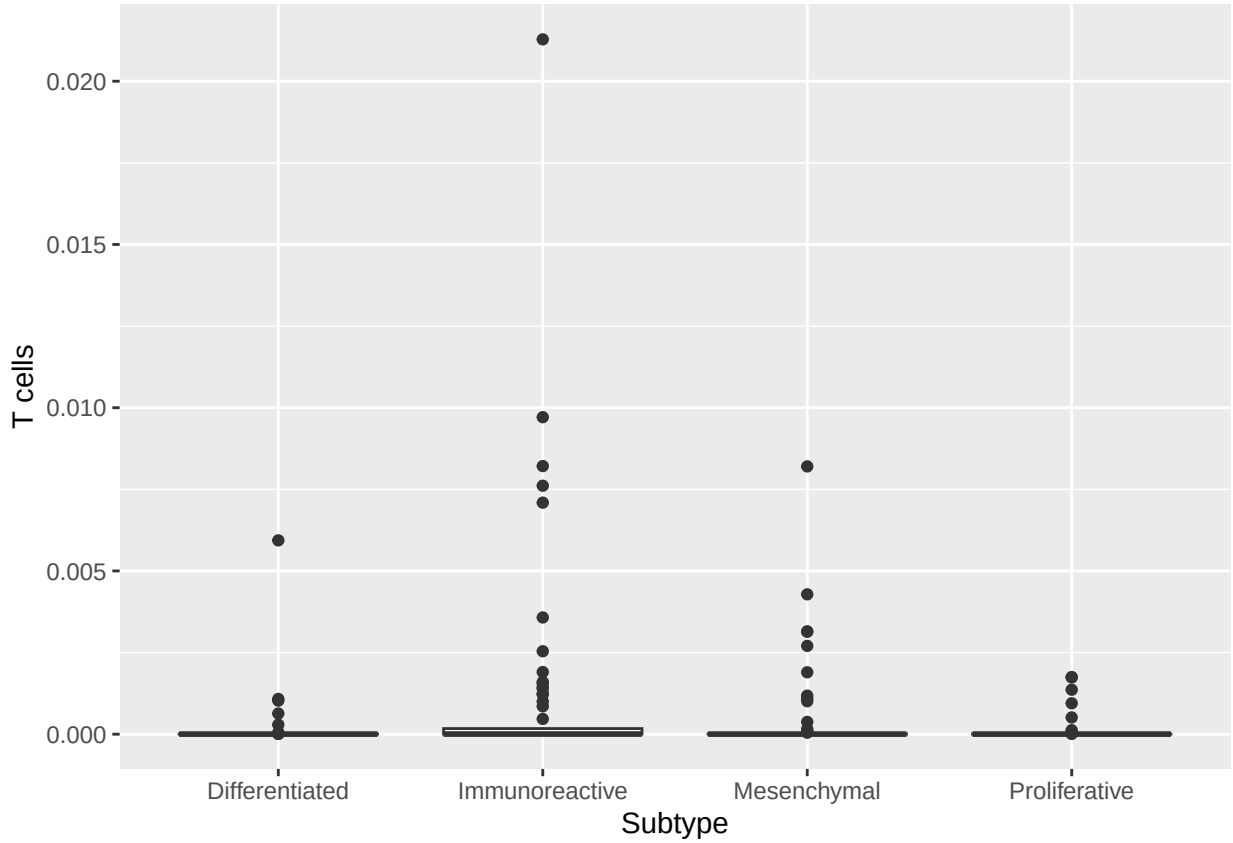
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=Subtype, y=`T cells`)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_tcells_by_subtype.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

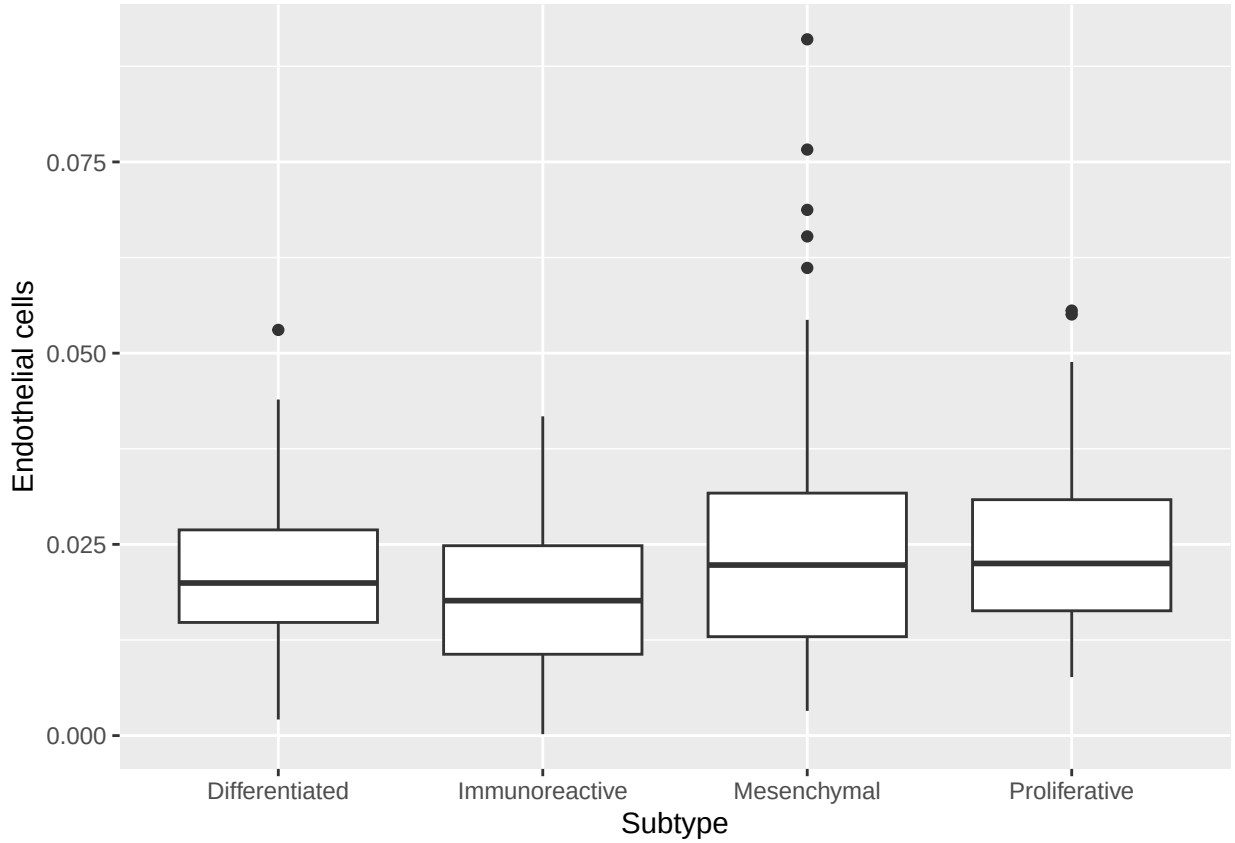
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=Subtype, y=`Endothelial cells`)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_endothelial_by_subtype.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

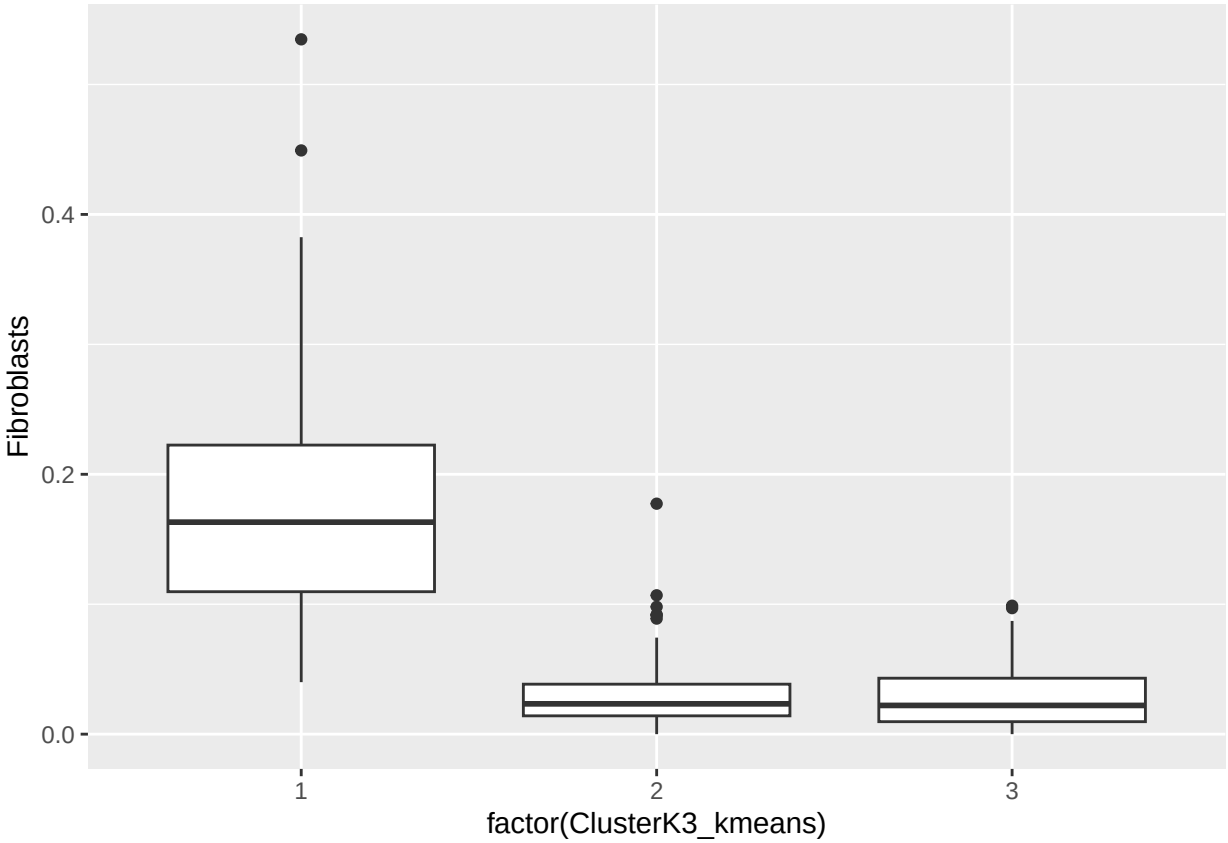
```
g
```




```
# Compare cell type proportions of subtypes for k=3
g <- ggplot(tcga_t, mapping = aes(x=factor(ClusterK3_kmeans), y=Fibroblasts)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_fibroblasts_by_subtype_k3.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

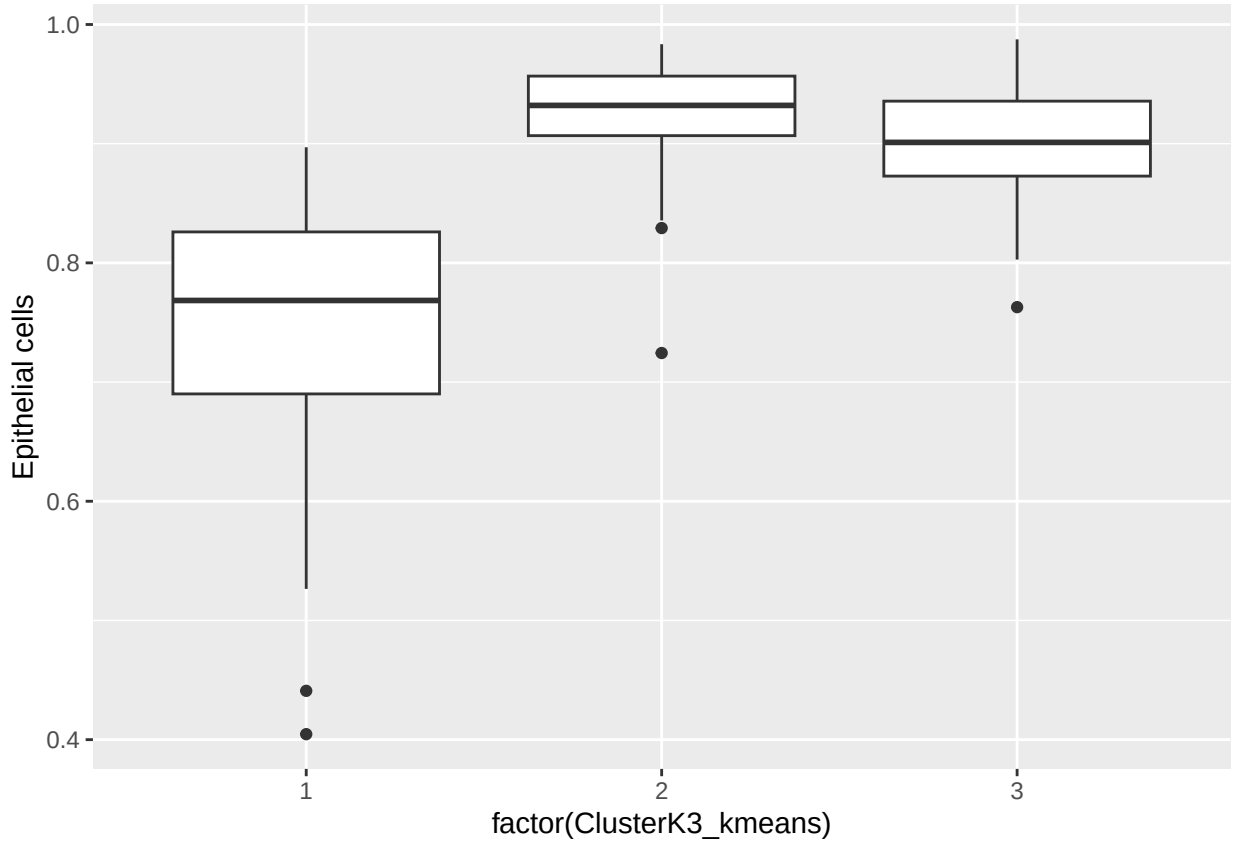
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=factor(ClusterK3_kmeans), y=`Epithelial cells`)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_epithelial_by_subtype_k3.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

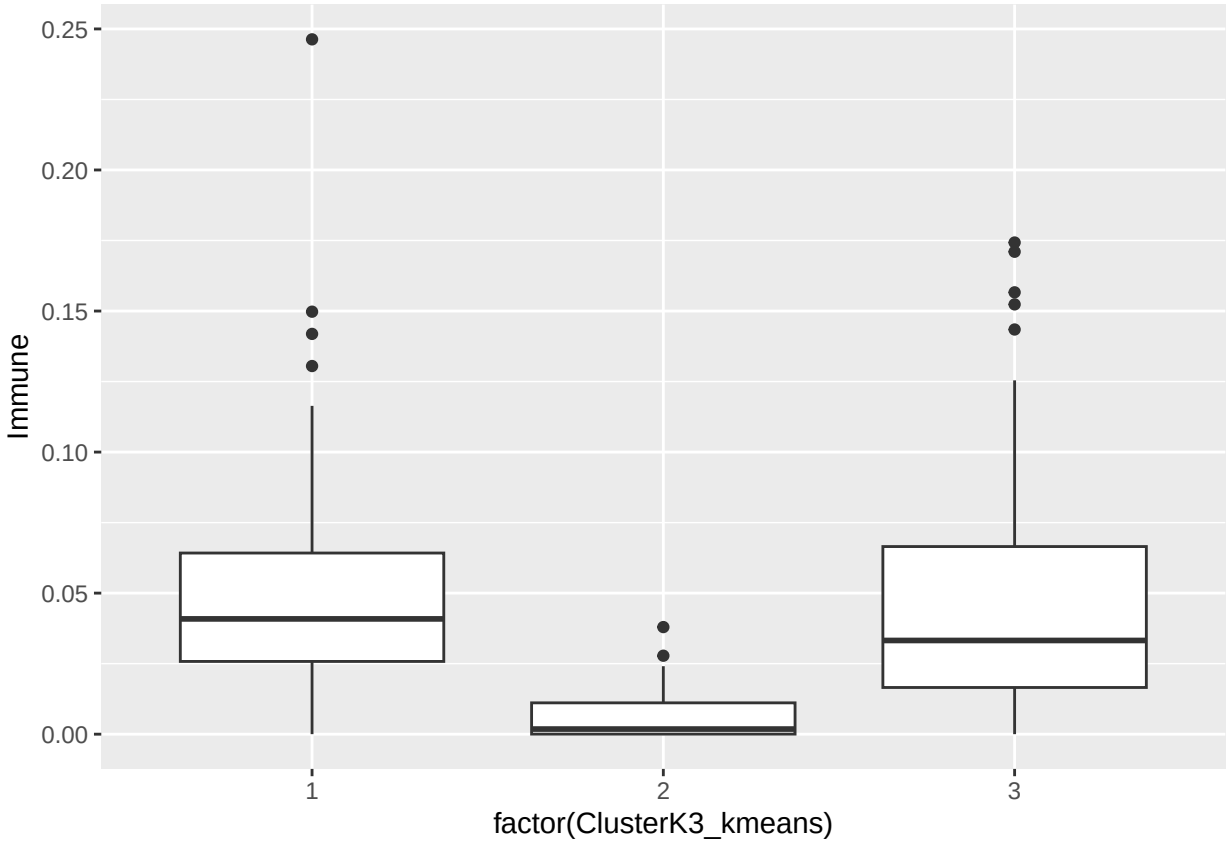
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=factor(ClusterK3_kmeans), y=Immune)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_immune_by_subtype_k3.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

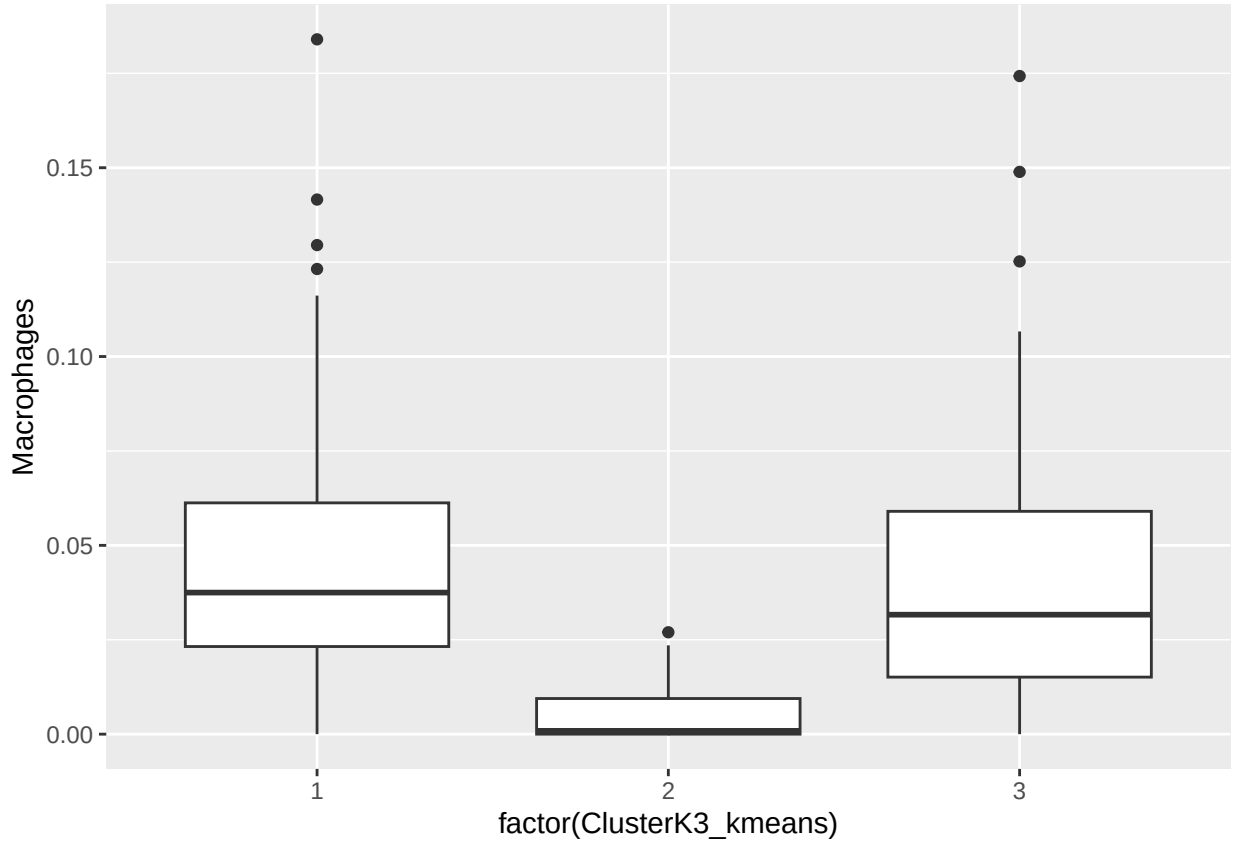
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=factor(ClusterK3_kmeans), y=Macrophages)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_macrophages_by_subtype_k3.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

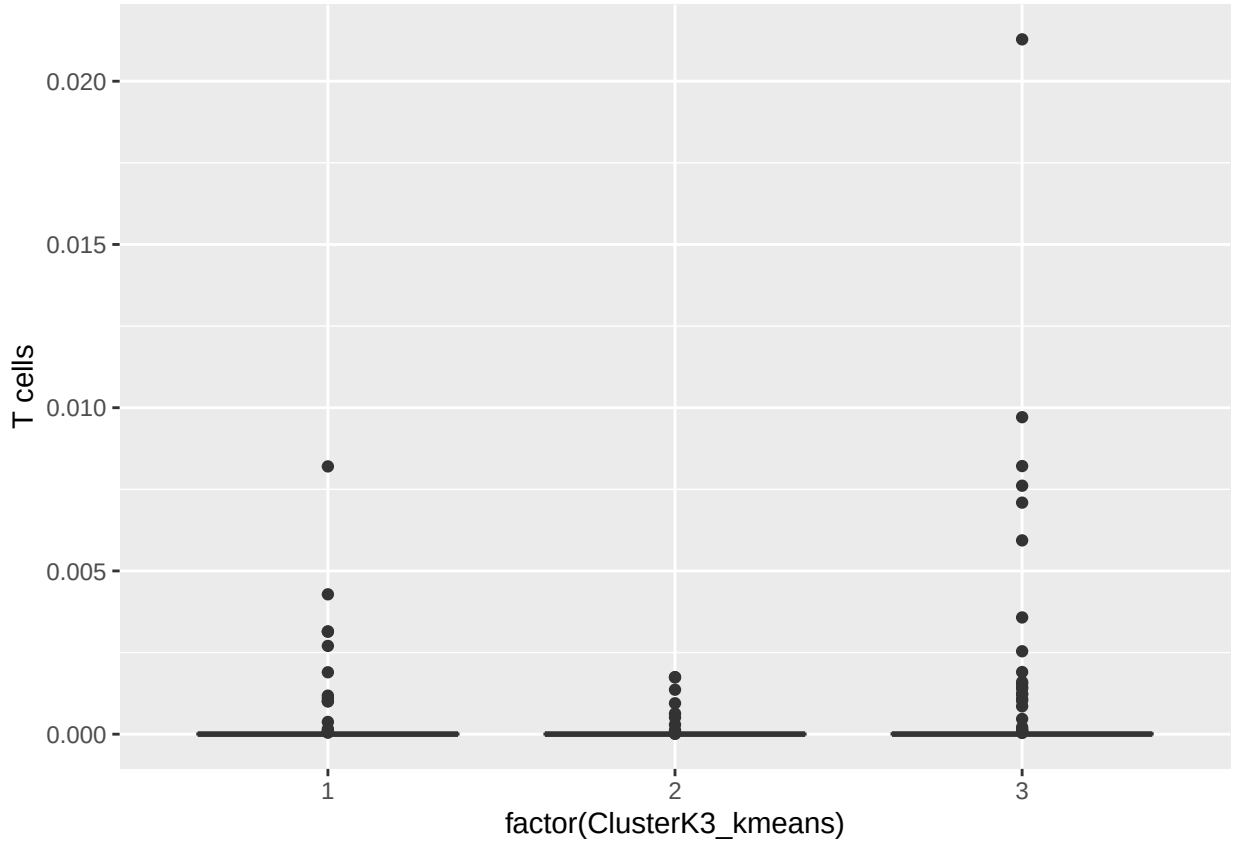
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=factor(ClusterK3_kmeans), y=`T cells`)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_tcells_by_subtype_k3.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

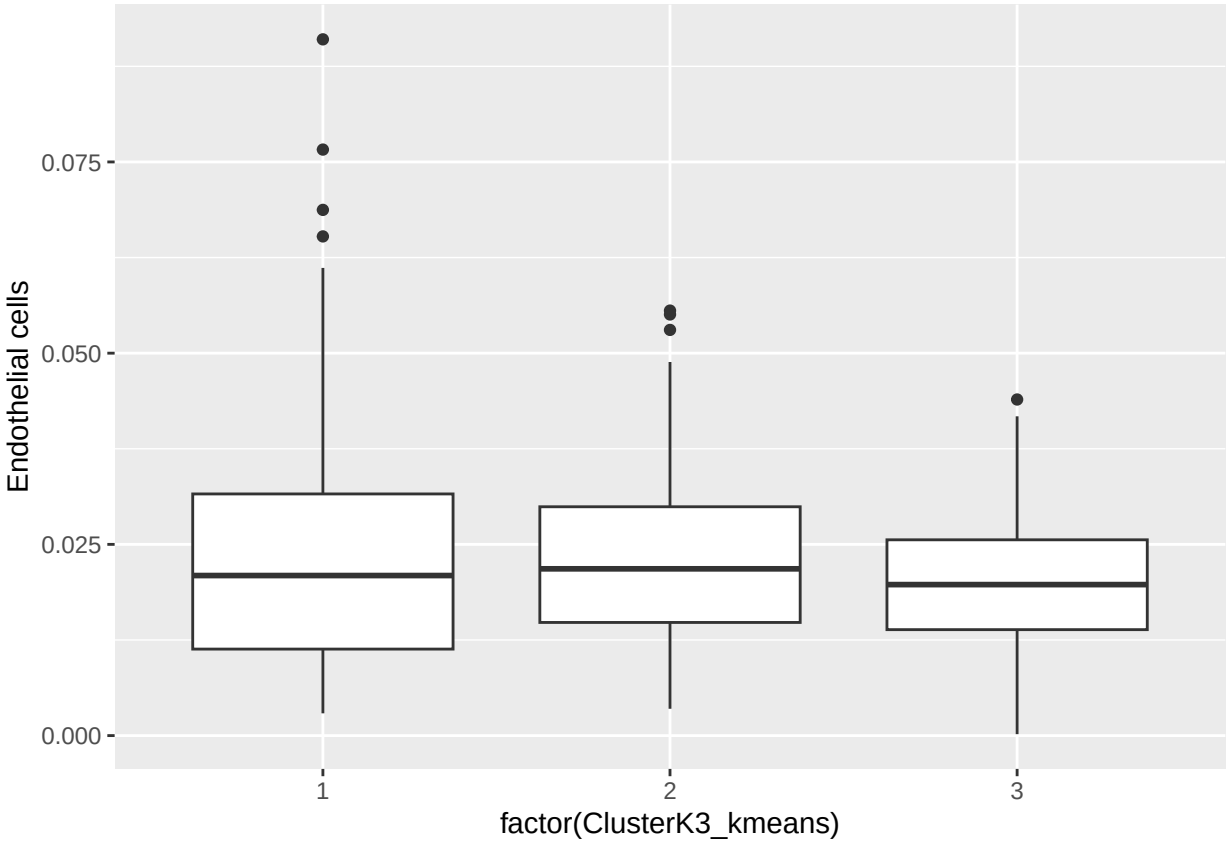
```
g
```




```
g <- ggplot(tcga_t, mapping = aes(x=factor(ClusterK3_kmeans), y=`Endothelial cells`)) + geom_boxplot()
plotfile <- paste(plot_path, "evaluation_plots", "TCGA_endothelial_by_subtype_k3.png", sep = "/")
png(filename = plotfile); g; dev.off()
```

```
## pdf
##   2
```

```
g
```


Endothelial cells
0.075
0.050
0.025
0.000
1
2
3
factor(ClusterK3_kmeans)