Chapter 6 Analysis example

6.1 ALL dataset

- BT 그룹별 유전자들의 발현 분포를 boxplot 이용해서 비교

```
library(tidyr)
library(dplyr)
library(tibble)
library(Biobase)
library(ALL)
library(hgu95av2.db)
library(ggplot2)

data(ALL)
## data
ex_data <- exprs(ALL)[1:30,]
ph_data <- pData(ALL)[,c("cod", "sex", "BT")]

## remove missing | duplicated genes
ph_data <- ph_data[complete.cases(ph_data),]
feature_names <- rownames(ex_data)
gene_names <- unlist(as.list(hgu95av2SYMBOL[feature_names]))
idx <- which(is.na(gene_names) | duplicated(gene_names))
ex_data <- as.data.frame(ex_data[-idx,])
rownames(ex_data) <- gene_names[-idx]
ex_data[1:3,1:3]


ex_data_mlt <- ex_data %>%
  rownames_to_column() %>%
  pivot_longer(-rowname) %>%
  mutate(bt = ph_data[name,"BT"])

### boxplot
ex_data_mlt %>%
  group_by(rowname) %>%
  ggplot(aes(x=bt, y=value, group=bt)) +
  facet_wrap(~rowname, ncol=9, scales="free") +
  geom_boxplot() +
  theme(
    axis.text.x = element_text(angle = 90, size=8, hjust = 1, vjust=0.5)
  )

## NA
ph_data$BT
tmp <- ex_data %>%
  rownames_to_column() %>%
  pivot_longer(-rowname)
ph_data[tmp$name,]$BT


ex_data_mlt <- ex_data %>%
  rownames_to_column() %>%
  pivot_longer(-rowname) %>%
  mutate(bt = ph_data[name,"BT"]) %>%
  drop_na()
ex_data_mlt %>% complete.cases()
```

- BT 그룹별 유전자 발현 평균 비교

```r
### 평균 비교
ex_summary <- ex_data_mlt %>%
  group_by(bt, rowname) %>%
  summarize(m=mean(value))


ggplot(ex_summary, aes(x=bt, y=m, group=bt)) +
  facet_wrap(~rowname, ncol=9, scales="free") +
  geom_bar(stat="identity") +
  theme(
    axis.text.x = element_text(angle = 90, size=8, hjust = 1, vjust=0.5)
  )

### scale
ggplot(ex_summary, aes(x=bt, y=m, group=bt)) +
  facet_wrap(~rowname, ncol=9) +
  geom_bar(stat="identity") +
  theme(
    axis.text.x = element_text(angle = 90, size=8, hjust = 1, vjust=0.5)
  )



### testing (t-test)
ex_data_mlt <- ex_data %>%
  rownames_to_column() %>%
  pivot_longer(-rowname) %>%
  mutate(bt = ph_data[name,"sex"])

ex_data_mlt %>% head
t.test(value~bt, data=ex_data_mlt)

test_results <- ex_data_mlt %>%
  group_by(rowname) %>%
  summarize(
    tval=t.test(value~bt)$statistic,
    pval=t.test(value~bt)$p.value,
    )



#### test all
data(ALL)
## data
ex_data <- exprs(ALL)
ph_data <- pData(ALL)[,c("cod", "sex", "BT")]

## remove missing | duplicated genes
ph_data <- ph_data[complete.cases(ph_data),]
```

```
feature_names <- rownames(ex_data)
gene_names <- unlist(as.list(hgu95av2SYMBOL[feature_names]))
idx <- which(is.na(gene_names) | duplicated(gene_names))
ex_data <- as.data.frame(ex_data[-idx,])
rownames(ex_data) <- gene_names[-idx]
dim(ex_data)

ex_data_mlt <- ex_data %>%
  rownames_to_column() %>%
  pivot_longer(-rowname) %>%
  mutate(bt = ph_data[name,"sex"]) %>%
  drop_na()

test_results <- ex_data_mlt %>%
  group_by(rowname) %>%
  summarize(
    tval=t.test(value~bt)$statistic,
    pval=t.test(value~bt)$p.value,
  )

dim(test_results)
head(test_results)
sig_results <- test_results %>%
  filter(pval<0.01)

sel_genes <- ex_data_mlt %>%
  filter(rowname %in% sig_results$rowname)

sel_genes %>%
  group_by(rowname) %>%
  ggplot(aes(x=bt, y=value, group=bt)) +
  facet_wrap(~rowname, ncol=9, scales="free") +
  geom_boxplot() +
  theme(
    axis.text.x = element_text(angle = 90, size=8, hjust = 1, vjust=0.5)
  )
```

- BT 그룹별 유전자 발현 anova 테스트

```
### testing (anova)
lm(data=ex_data_mlt, formula = value~bt)
```

6.2 Heatmap with sequence data

```
### Let's do it together!
```