Lab11: Genome informatics and high throughput sequencing

Duy An Le (PID:A16400411)

Read in file:

```
geneexp <- read.table("geneexpression.txt")
head(geneexp)
```

```
   sample geno      exp
1 HG00367  A/G 28.96038
2 NA20768  A/G 20.24449
3 HG00361  A/A 31.32628
4 HG00135  A/A 34.11169
5 NA18870  G/G 18.25141
6 NA11993  A/A 32.89721
```

> Q13: Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
aa = geneexp %>% filter(geno=="A/A")
ag = geneexp %>% filter(geno=="A/G")
gg = geneexp %>% filter(geno=="G/G")

aaplot <- boxplot(aa$exp, plot=F)
agplot <- boxplot(ag$exp, plot=F)
ggplot <- boxplot(gg$exp, plot=F)
```

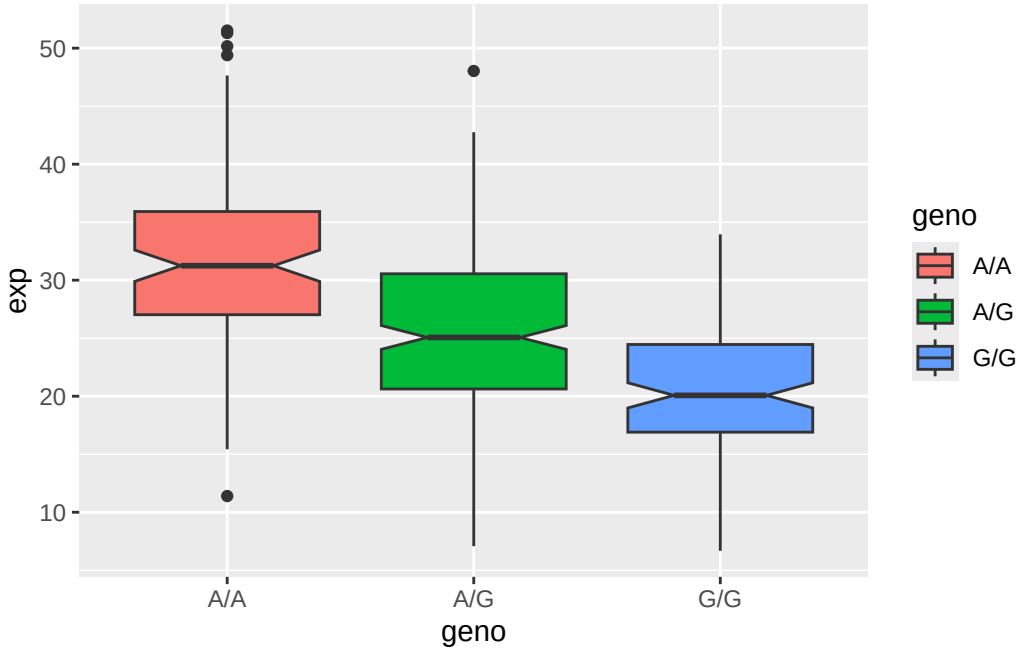
The sample size of A/A genotype is 108 and the median expression level is 31.248475. The sample size of the A/G genotype is 233 and the median expression level is 25.06486. The sample size of the G/G genotype is 121 and the median expression level is 20.07363.

> Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

```
library(ggplot2)
```

Make a boxplot with ggplot

```
ggplot(geneexp) + aes(geno, exp, fill=geno) + geom_boxplot(notch=T)
```



It seems that the relative expression level of ORMDL3 in A/A genotypes is higher than that of the G/G genotypes. Having the SNP seems to affect the expression of ORMDL3 and may be correlated with having asthma.