HW 12: Population Scale Analysis

Amanda Wilpitz | A17463962

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

```r
data <- read.table("rs8067378_ENSG00000172057.6.txt")
head(data)
```

```
   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
```

```r
table(data$geno)
```

```

A/A A/G G/G 
108 233 121 
```

```r
lapply(split(data$exp, data$geno), summary)
```

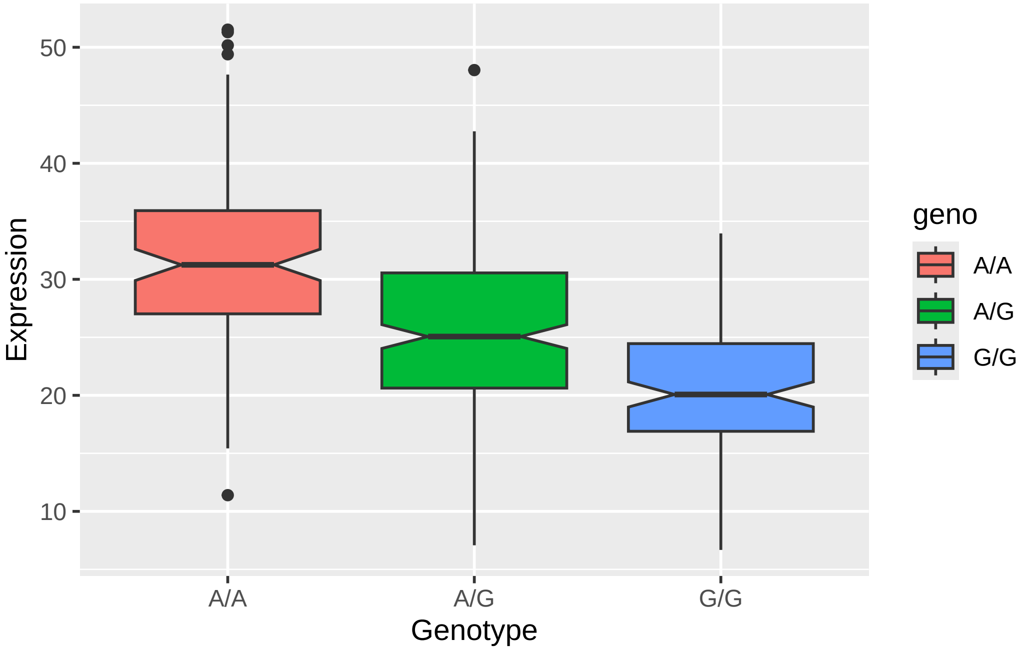
```
$`A/A`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  11.40   27.02   31.25   31.82   35.92   51.52 

$`A/G`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  7.075  20.626  25.065  25.397  30.552  48.034 
```

```
$`G/G`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.675  16.903  20.074  20.594  24.457  33.956 
```

> 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)
```

```
ggplot(data) +
  aes(geno, exp, fill = geno) +
  geom_boxplot(notch=T) +
  labs(x = "Genotype", y = "Expression")
```



I can infer that A/A has higher expression levels and has a difference of about 10 between A/A and G/G.

Since the expression levels differ across the genotypes enough to visually see, the SNP could effect the expression of ORMDL3.