HW Class 11 Pt.2 (Population analysis) [Extra Credit BoxPlot]

Jimmi

Population Analysis

> 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.

Let's read data and store it into a dataframe.

```
gene <- read.table(file='rs8067378_ENSG00000172057.6.txt', header=TRUE)
head(gene)
```

```
   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
```

Identifying how many samples have a specific genotype.

```
genotypes = table(gene$geno)
genotypes
```

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

We can find useful data with the `summary()` function and specifically get the median by using the built in R `median()` function.

```
summary( gene$exp[gene$geno == "A/A"] )
```

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

```
summary( gene$exp[gene$geno == "A/G"] )
```

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

```
summary( gene$exp[gene$geno == "G/G"] )
```

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

```
median(gene$exp[gene$geno == "A/A"])
```

```
[1] 31.24847
```

```
median(gene$exp[gene$geno == "A/G"])
```

```
[1] 25.06486
```

```
median(gene$exp[gene$geno == "G/G"])
```

```
[1] 20.07363
```

The sample size for A/A is 108 with an 31.248475 median expression level.

The sample size for A/G is 233 with an 25.06486 median expression level.

The sample size for G/G is 121 with an 20.07363 median expression level.

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?

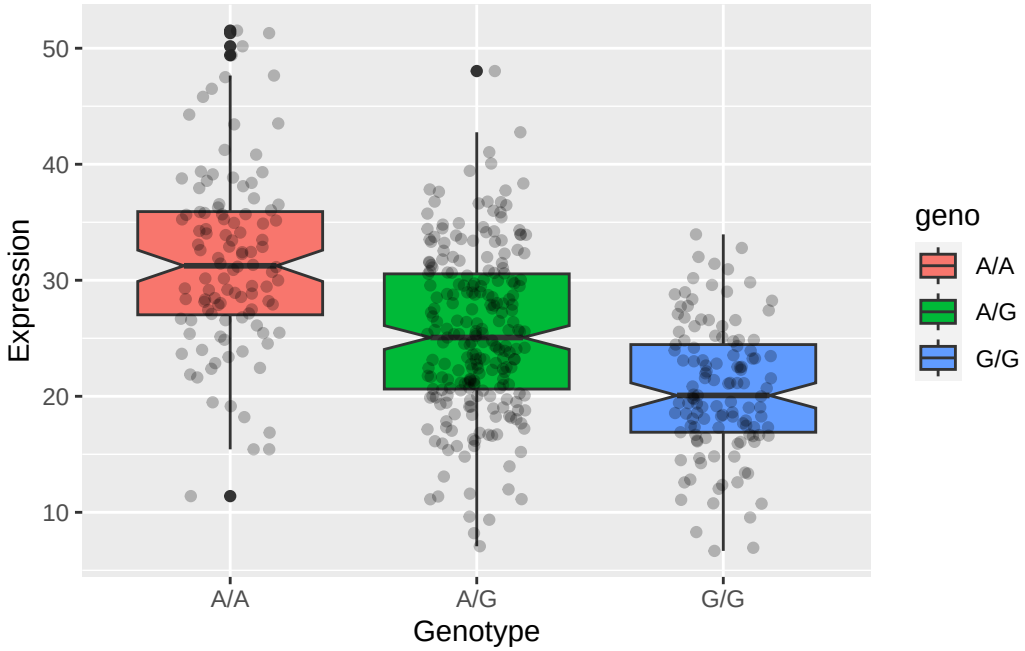
Turning genotypes are factors before loading up ggplot to make a boxplot.

```
gene$geno = as.factor(gene$geno)
head(gene)
```

```
   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
```

```
library(ggplot2)

ggplot(gene, aes(x=geno, y=exp,fill=geno)) +
  geom_boxplot(notch=TRUE) +
  geom_jitter(alpha=0.25, fill="black",width = 0.2) +
  theme(legend.position="right",
        plot.title = element_text(size=11)) +
  labs(x="Genotype", y = "Expression")
```



We can infer that A/A is incurs the most expression of ORMDL3 and dominant while G/G seen in underexpressed ORMDL3 and recessive. The SNP loci appears to have an effect on the ORMDL3 gene, however, there are obvious variable ranges between each genotypes. Therefor, without a statistical significance test it is not certain.