Class 11 hw

AUTHOR
Lucy Wang

Section 4: Population Scale 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.

Read in the data:

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

```
   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
```

How many samples do we have - check the number of rows:

```
nrow(expr)
```

```
[1] 462
```

```
#Genotype info
table(expr$geno)
```

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

```
aa <- expr$exp[expr$geno == "A/A"]
ag <- expr$exp[expr$geno == "A/G"]
gg <- expr$exp[expr$geno == "G/G"]
summary(aa)
```

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

```
summary(ag)
```

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

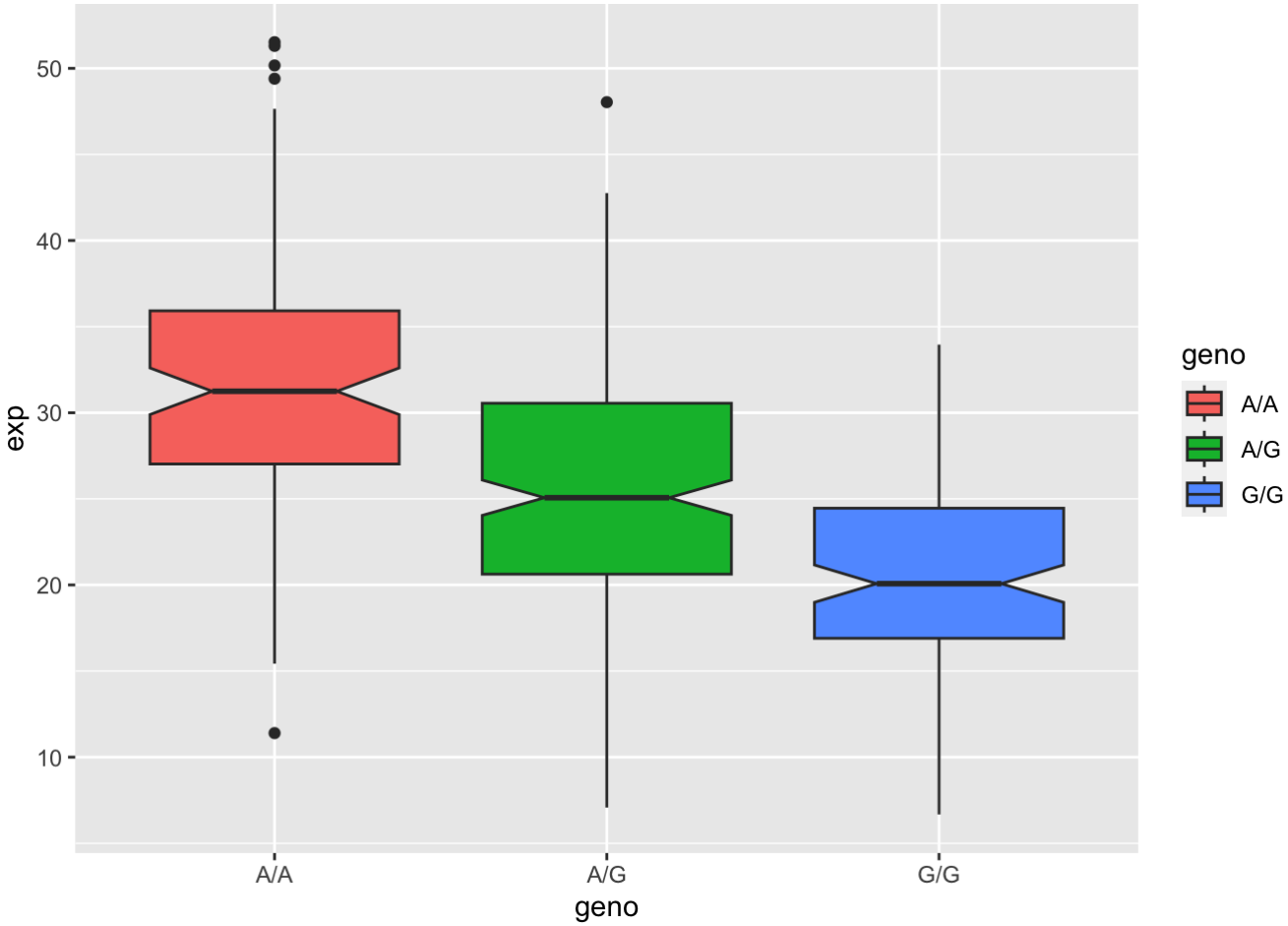
```
summary(gg)
```

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

```
library(ggplot2)
```

make a boxplot w/ ggplot

```
#use `notch` get a more clear indication
ggplot(expr) + aes(geno, exp, fill=geno) +
  geom_boxplot(notch = TRUE)
```



> 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?

A/A expression is higher than G/G expression, with almost no overlap between their central 50% (the box). It can be inferred that this SNP affect the expression of ORMDL3 from the data.

THE END