# class11

Katie Mostoller A17259578

## Section 1. proption of G/G in a population

Read the csv file

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mxl)
```

```
  Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
1                  NA19648 (F)                       A|A ALL, AMR, MXL      -
2                  NA19649 (M)                       G|G ALL, AMR, MXL      -
3                  NA19651 (F)                       A|A ALL, AMR, MXL      -
4                  NA19652 (M)                       G|G ALL, AMR, MXL      -
5                  NA19654 (F)                       G|G ALL, AMR, MXL      -
6                  NA19655 (M)                       A|G ALL, AMR, MXL      -
  Mother
1      -
2      -
3      -
4      -
5      -
6      -
```

```
table(mxl$Genotype..forward.strand.)
```

```

A|A A|G G|A G|G 
 22  21  12   9 
```

```
table(mxl$Genotype..forward.strand.)/nrow(mxl) *100
```

```
    A|A     A|G     G|A     G|G 
34.3750 32.8125 18.7500 14.0625 
```

### Section 4: population analysis

```
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
```

```
nrow(expr)
```

```
[1] 462
```

> 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

```
table(expr$geno)
```

```

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

```
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
```

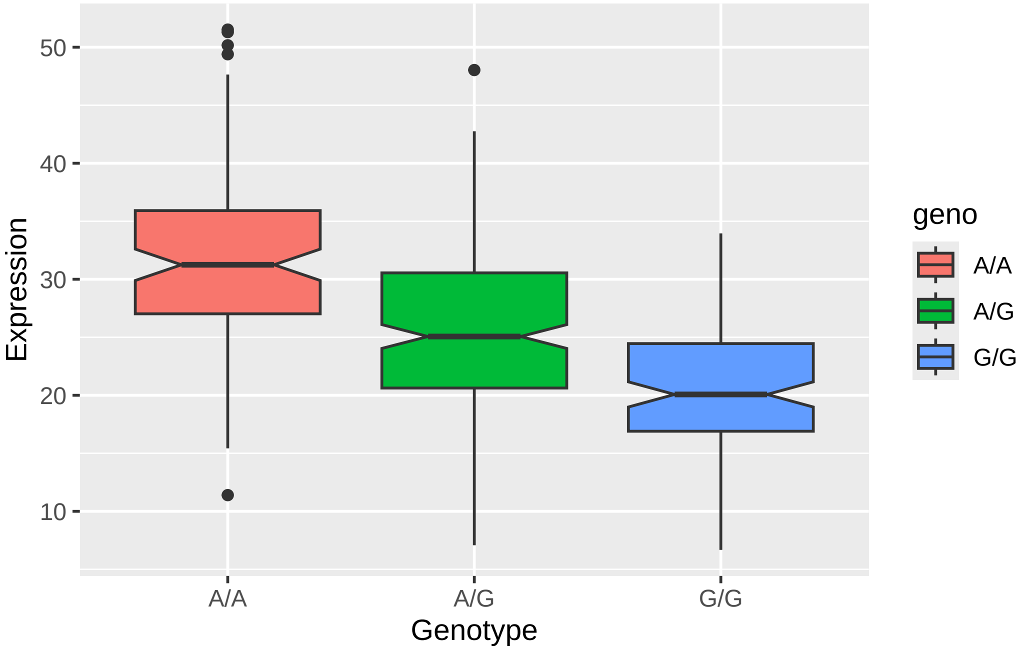
```
summary_stats <- expr %>%
  group_by(geno) %>%
  summarise(
    Sample_Size = n(),
    Median_Expression = median(exp))

summary_stats
```

```
# A tibble: 3 x 3
  geno  Sample_Size Median_Expression
  <chr>       <int>             <dbl>
1 A/A           108              31.2
2 A/G           233              25.1
3 G/G           121              20.1
```

```
library(ggplot2)
```

Q14. Make a boxplot!

```
ggplot(expr) + aes(x = geno, y = exp, fill = geno) + geom_boxplot(notch = TRUE) + xlab("Geno
```


50
40
30
20
10
Expression
A/A
A/G
G/G
Genotype
geno
A/A
A/G
G/G