Lab 11

Rosa Chavez (PID: A16487039)

Section 1

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (2).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
```

Extra Credit

How many samples do we have?

```
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
```

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

```
nrow(expr)
```

```
[1] 462
```

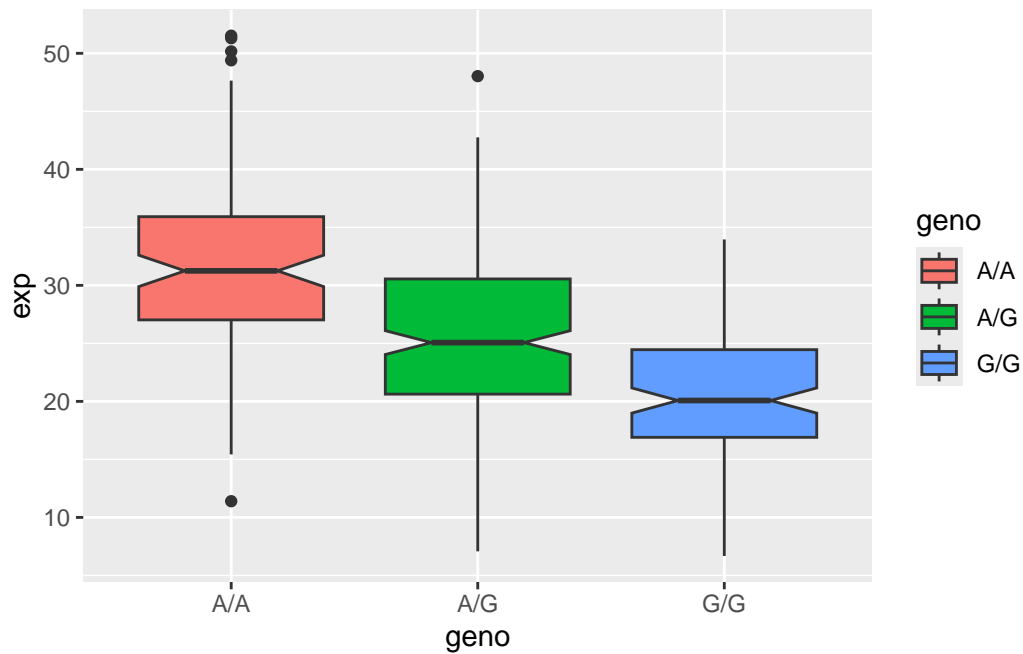What is the sample size for each of the individual genotypes?

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

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

```
library(ggplot2)
```

```
boxplot_data <-  ggplot(expr) + aes(geno, exp, fill = geno) +
  geom_boxplot(notch = TRUE)

boxplot_data
```


What are the median for each genotype?

```
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
```

```
ggplot_build_data <- ggplot_build(boxplot_data)$data[[1]]

median <- ggplot_build_data %>%
group_by(x) %>%
summarize(median = middle)

median
```

```
# A tibble: 3 x 2
  x           median
  <mppd_dsc>  <dbl>
1 1             31.2
2 2             25.1
3 3             20.1
```