class12

Ani A16647613

Section 1: Proporation of G/G in a population

Downloaded a CSV file from the Ensemble from the lab directions

Now we will read the csv file

```r
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).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      -
```

```r
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
```

14.06% G/G

Now we are gonna look at population GBR (Great Britain)

```
gbr <- read.csv("gbr.csv")
```

Find proportion of G/G

```
round(table(gbr$Genotype..forward.strand.) / nrow(gbr) *100, 2)
```

```
  A|A   A|G   G|A   G|G
25.27 18.68 26.37 29.67
```

This varient of childhood asthma is more frequent in GBR than MXL population, so noew we can look into this further:

##Section 4: Population Scale Analysis

Answering the following questions:

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

Q14. Generate a boxplot with a box per genotype, what would you inger from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3? How many samples do we have?

```
expr <- read.table("project12part2.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
```

From the table function below, we can see the distribution amongst the different genotypes to
see their sample sizes.
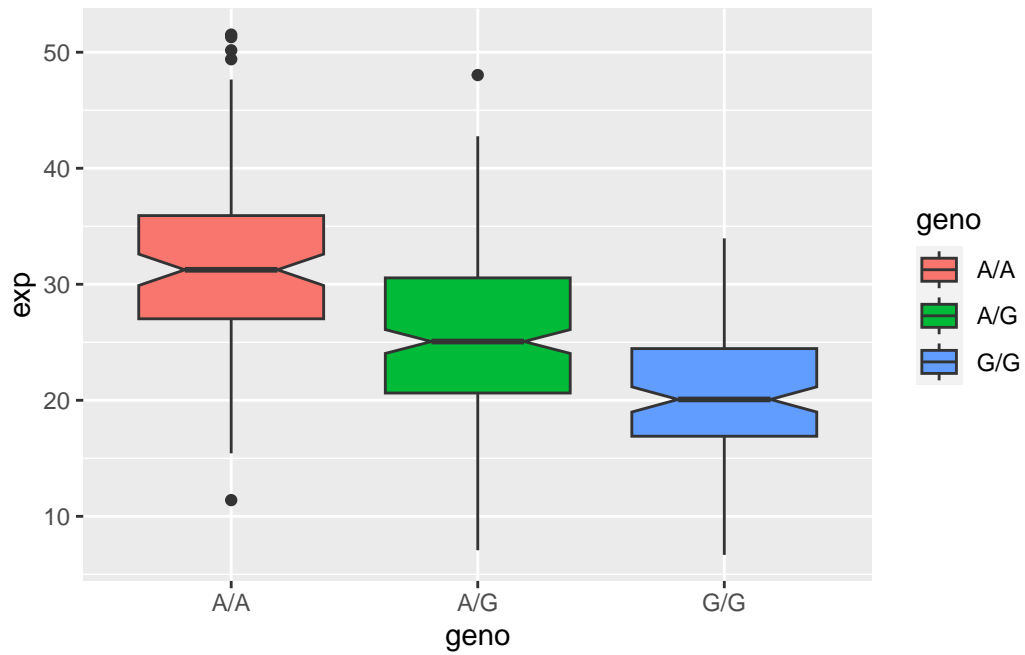
```
nrow(expr)
```

```
[1] 462
```

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

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

```
library(ggplot2)
```

Make a boxggpplot

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

Above, we can interpret that having a G/G shows reduced expression of the gene. We can also interpret that A/A shows high expression of the gene.