Lab 14 Genome Informatics with Homework

##Introduction to Genome Informatics Lab #Section 1: Identify genetic variants of interest

Q5: Q5: What proportion of the Mexican Ancestry in Los Angeles sample population (MXL) are homozygous for the asthma associated SNP (G|G)?

Downloaded CSV file from Ensemnble < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core
39895595;v=rs8067378;vdb=variation;vf=105535077

Read CSV file and analyze the sample:

```
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      -
```

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

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

A5:

```
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% are homozygous for asthma within MXL population sample in LA

Q6. Back on the ENSEMBLE page, use the "search for a sample" field above to find the particular sample HG00109. This is a male from the GBR population group. What is the genotype for this sample?

Now let's look at diff population, specifically GBR, "Great Britain"

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.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
```

A: 29.67% are homozygous for asthma within GBR population sample in Great Britain

#Section 4: Population Scale Analysis How many samples doe 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
```

Total individuals?

```
nrow(expr)
```

```
[1] 462
```

How many of each genotype?

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

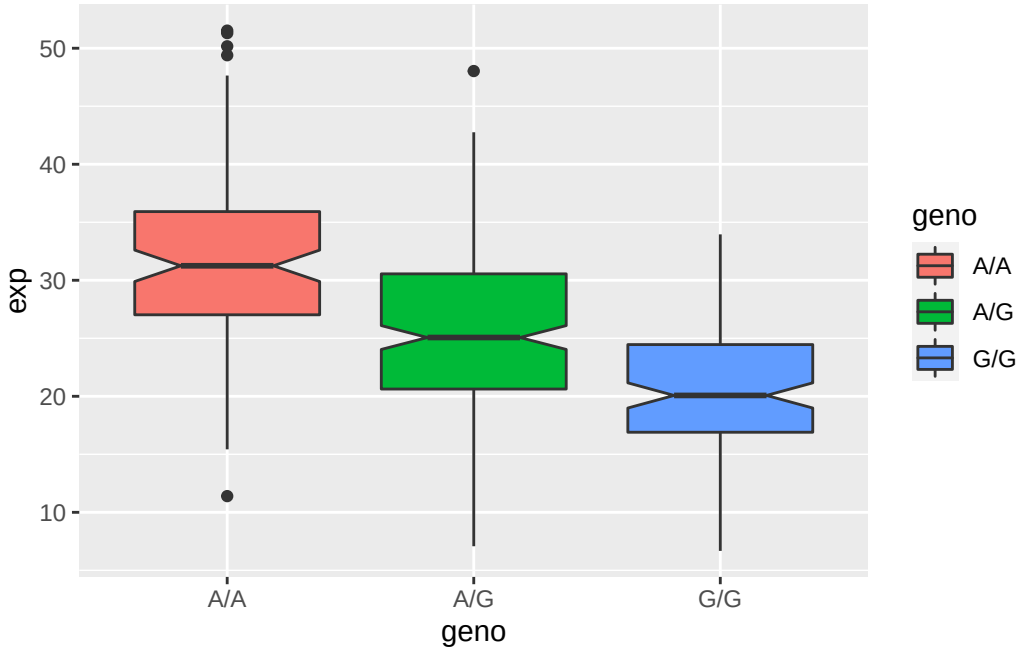
```
A/A A/G G/G
108 233 121
```

How do we deplay these results?

```
library(ggplot2)
```

let's make a boxplot with this data:

```
ggplot(expr) + aes(x=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?

A14: According to the boxplot above, epression of ORMDL3 is highest when individuals experess homologous alleles A|A, and lowest for G|G. The SNP does effect the expression of ORDML3