HW Class 12 Pt.2: Population Analysis

Ruth Barnes: A16747659

Class 12 Pt. 1: RNASeq Galaxy

Download and read a CSV from ENSEMBLE:

```
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                  NA19651 (F)                       A|A ALL, AMR, MXL      -
3                  NA19658 (M)                       A|A ALL, AMR, MXL      -
4                  NA19663 (F)                       A|A ALL, AMR, MXL      -
5                  NA19669 (F)                       A|A ALL, AMR, MXL      -
6                  NA19670 (M)                       A|A ALL, AMR, MXL      -
  Mother
1      -
2      -
3      -
4      -
5      -
6      -
```

How many of each genotype are there?

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

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

Proportion or percent of total for each genotype:

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

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

This variant that is associated with childhood asthma is more frequent in the GBR population than the MKL population.

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

```
  Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
1                  HG00096 (M)                       A|A ALL, EUR, GBR      -
2                  HG00100 (F)                       A|A ALL, EUR, GBR      -
3                  HG00101 (M)                       A|A ALL, EUR, GBR      -
4                  HG00102 (F)                       A|A ALL, EUR, GBR      -
5                  HG00105 (M)                       A|A ALL, EUR, GBR      -
6                  HG00108 (M)                       A|A ALL, EUR, GBR      -
  Mother
1      -
2      -
3      -
4      -
5      -
6      -
```

Let's now dig into this further.

Class 12 Pt. 2: Population Analysis

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale.

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.

```
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. How many samples do we have?

```
nrow(expr)
```

```
[1] 462
```

Sample size of genotypes and summary of expr:

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

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

```
summary(expr)
```

```
    sample              geno                exp
 Length:462         Length:462         Min.   : 6.675
 Class :character   Class :character   1st Qu.:20.004
 Mode  :character   Mode  :character   Median :25.116
                                       Mean   :25.640
                                       3rd Qu.:30.779
                                       Max.   :51.518
```

Q13. Let's examine the three genotypes to find the median expression levels for each:

```
inden <- expr$geno == "G/G"
summary(expr[inden, "exp"])
```

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

```
inden <- expr$geno == "A/A"
summary(expr[inden, "exp"])
```

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

```
inden <- expr$geno == "A/G"
summary(expr[inden, "exp"])
```

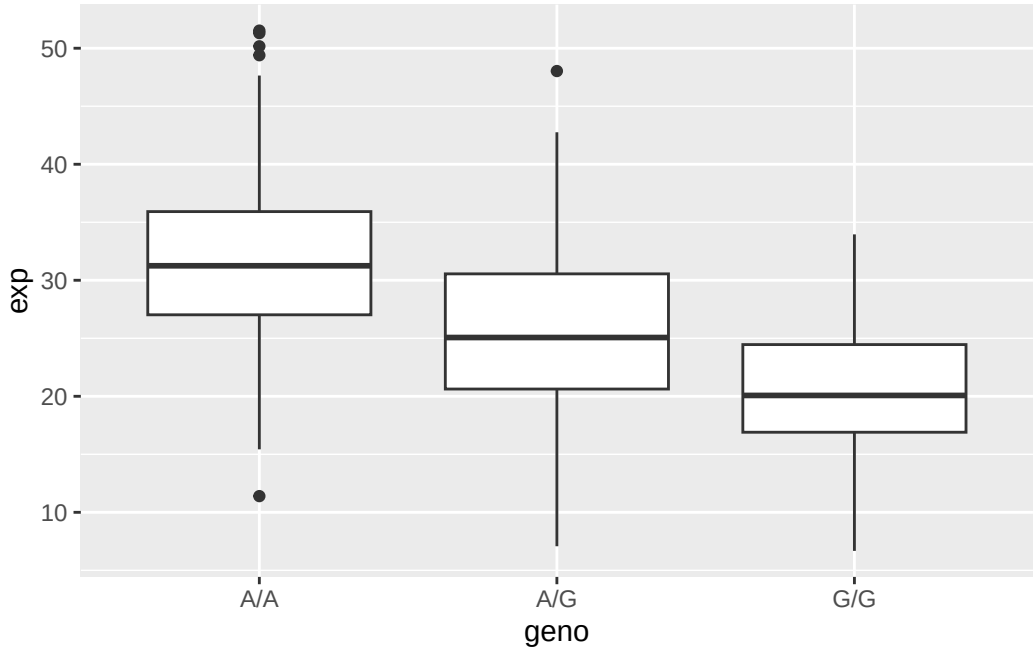
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  7.075  20.626  25.065  25.397  30.552  48.034 
```

Q14: Generate a boxplot with a box per genotype

Let's make a boxplot:

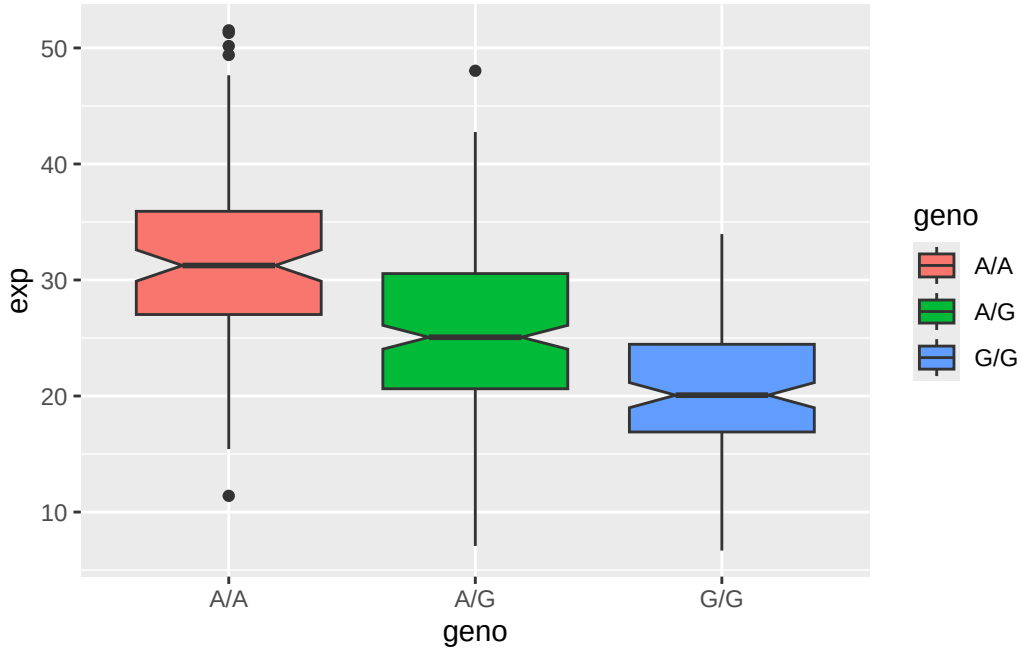
```
library(ggplot2)
```

```
ggplot(expr) + aes(geno, exp) +
  geom_boxplot()
```



Let's make the boxplot nicer and easier to understand:

```
ggplot(expr) + aes(geno, exp, fill = geno) +
  geom_boxplot(notch = T)
```



Q14. 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?

The relative expression values between A/A and G/G show that the A/A genotype has the highest expression of ORMDL3, while G/G has the lowest expression of ORMDL3. This trend suggests that the SNP does affect the expression of ORMDL3, where the G allele is associated with lower expression levels.