Class 11: HW Population Analysis

AUTHOR
Hyeseung (Frankie) Son PID: A16025601

Section 1: Proportion of G/G in a population

Downloaded a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39894946-39895247;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel>

Here we read a csv file to determine the allele frequency.

```
mxl <- read.csv("/Users/frank/Downloads/class12/373531-SampleGenotypes-Homo_sapiens_Variation_Sam

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
```

In the MXL population, the G|G homozygous for childhood asthma is 14%.

Now let's look at a different population. I picked GBR.

```
gbr <- read.csv("/Users/frank/Downloads/class12/373522-SampleGenotypes-Homo_sapiens_Variation_Sam
```

Find the proportion of G|G.

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

```
A|A A|G G|A G|G
 25  19  26  30
```

The proportion of G|G in this population is 30%, so childhood asthma is more frequent in GBR than MXL.

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.

So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

> 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. Hint: The read.table(), summary() and boxplot() functions will likely be useful here.

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
```

```
nrow(expr)
```

```
[1] 462
```

There are 462 individuals in this sample.

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

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

The sample size for each genotype is: 108 individuals with A|A, 233 with A|G, 121 with G|G.

```
geno.AA <- expr[expr$geno == "A/A", "exp"]
geno.AG <- expr[expr$geno == "A/G", "exp"]
geno.GG <- expr[expr$geno == "G/G", "exp"]

median.AA <- median(geno.AA)
median.AG<- median(geno.AG)
median.GG <- median(geno.GG)

median.AA
```

```
[1] 31.24847
```

```
median.AG
```

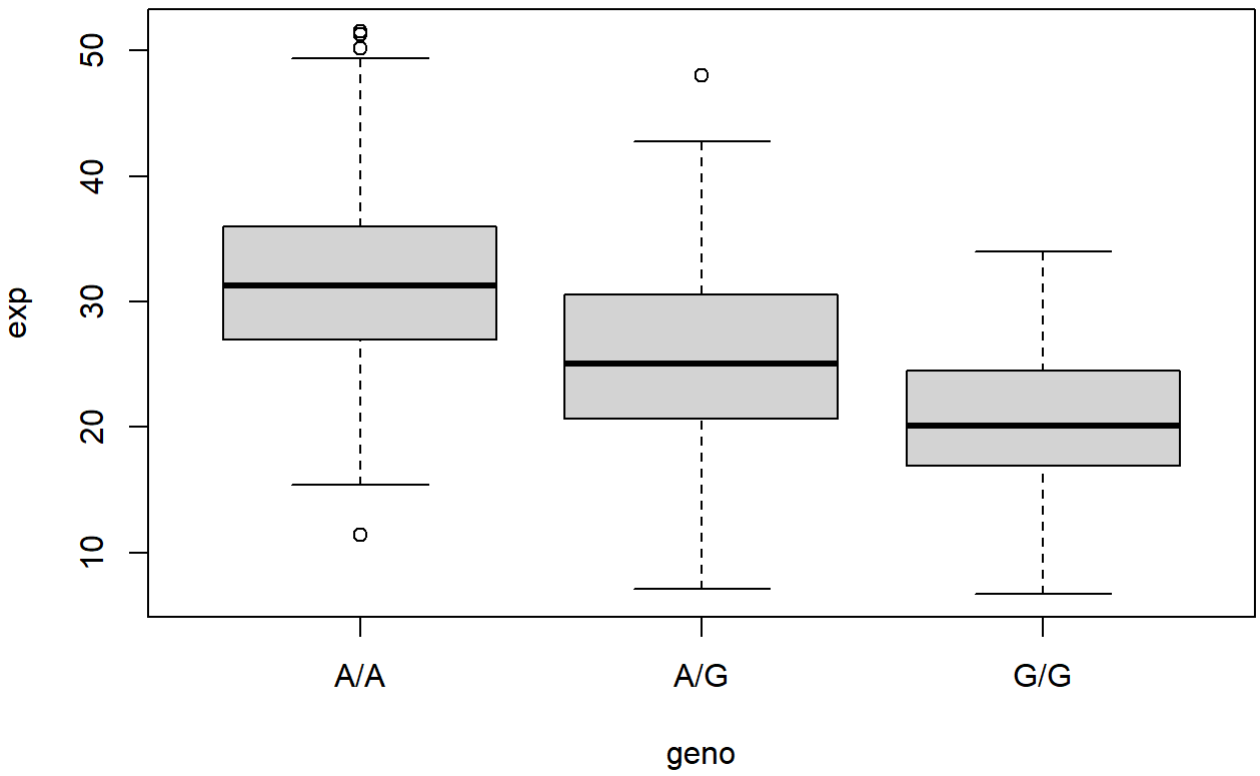
```
[1] 25.06486
```

```
median.GG
```

```
[1] 20.07363
```

A more simpler method may be:

```
median <- boxplot(exp ~ geno, data=expr)
```




```
summary.stats <- summary(median$stats)

summary.stats
```

```
      V1              V2              V3
Min.   :15.43   Min.   : 7.075   Min.   : 6.675
1st Qu.:26.95   1st Qu.:20.626   1st Qu.:16.903
Median :31.25   Median :25.065   Median :20.074
Mean   :31.80   Mean   :25.215   Mean   :20.413
3rd Qu.:35.96   3rd Qu.:30.552   3rd Qu.:24.457
Max.   :49.40   Max.   :42.757   Max.   :33.956
```

The median expression levels for each genotype are as follows: A/A: 31.2, A/G: 25.1, G/G: 20.1.

> 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?

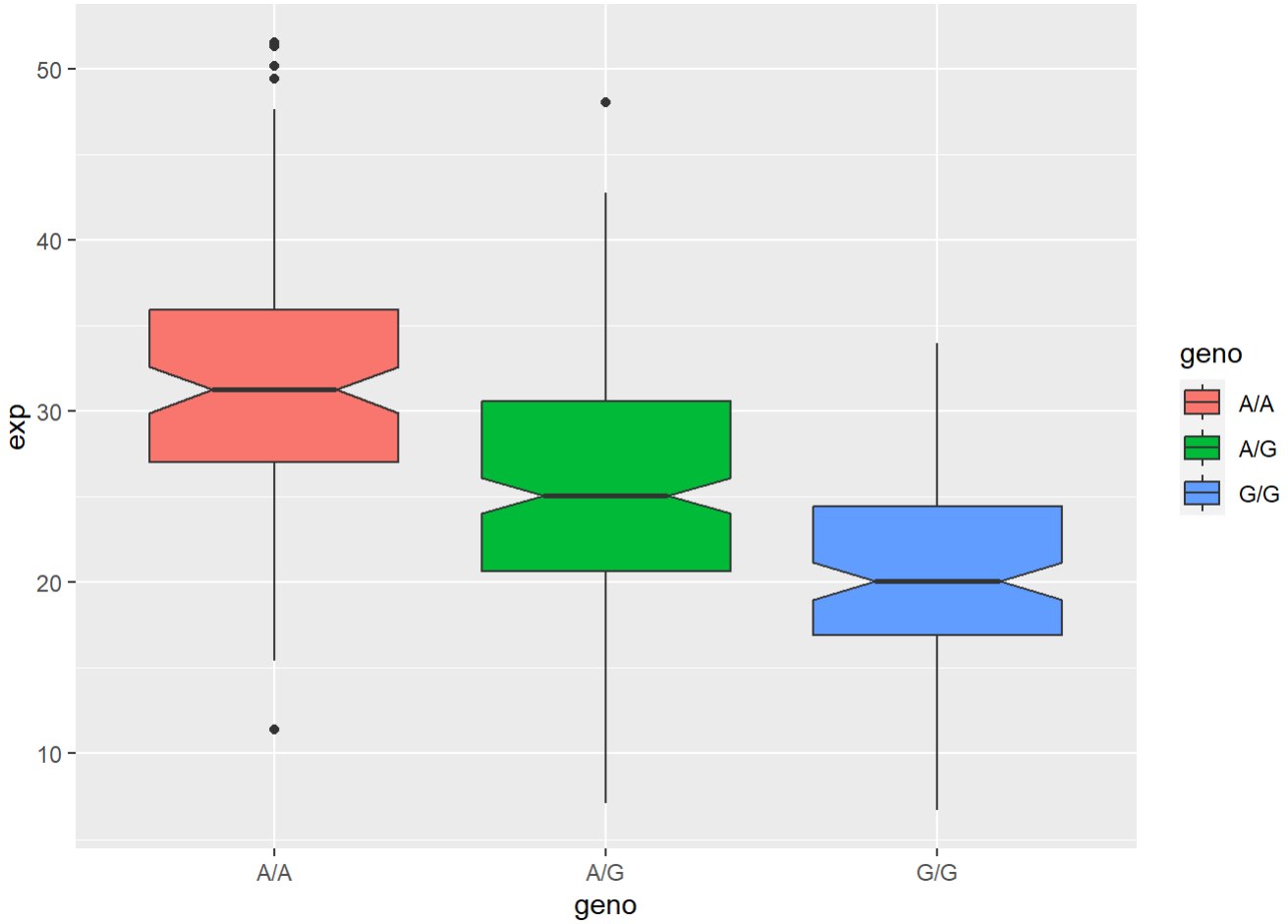
```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

Let's make a boxplot.

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

boxplot.expr
```



Having a G/G genotype is associated with having a decreased expression of the ORMDL3 gene, while A/A genotype is associated with having increased expression of the ORMDL3 gene.

The A/G genotype has ORMDL3 gene expression levels in between that of the A/A and G/G genotypes' expression levels, and overlaps with both the expression levels of the A/A and G/G boxplots, so the heterozygous genotype may not be all that different from either homozygous genotype. We find that overall, having a a higher percentage of G alleles (asthma-related SNPs) in your genotype are associated with decreased levels of ORMLD3 gene expression.