HW class10 Pt.2

Seong Tae Gwon

2/21/2022

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensemble https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39894970-39895222;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel

Here, we read this CSV file.

```
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                 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
```

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

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

Find proportion of G|G.

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

```
##
##      A|A      A|G      G|A      G|G
## 25.27473 18.68132 26.37363 29.67033
```

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

Let's now dig into this further.

Note: Section 2. Initial RNA-Seq analysis and Section 3. Mapping RNA-Seq reads to genome on lab PDF submission

Section 4. Population Scale Analysis

```
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
```

How many samples do we have?

```
# Total sample size
nrow(expr)
```

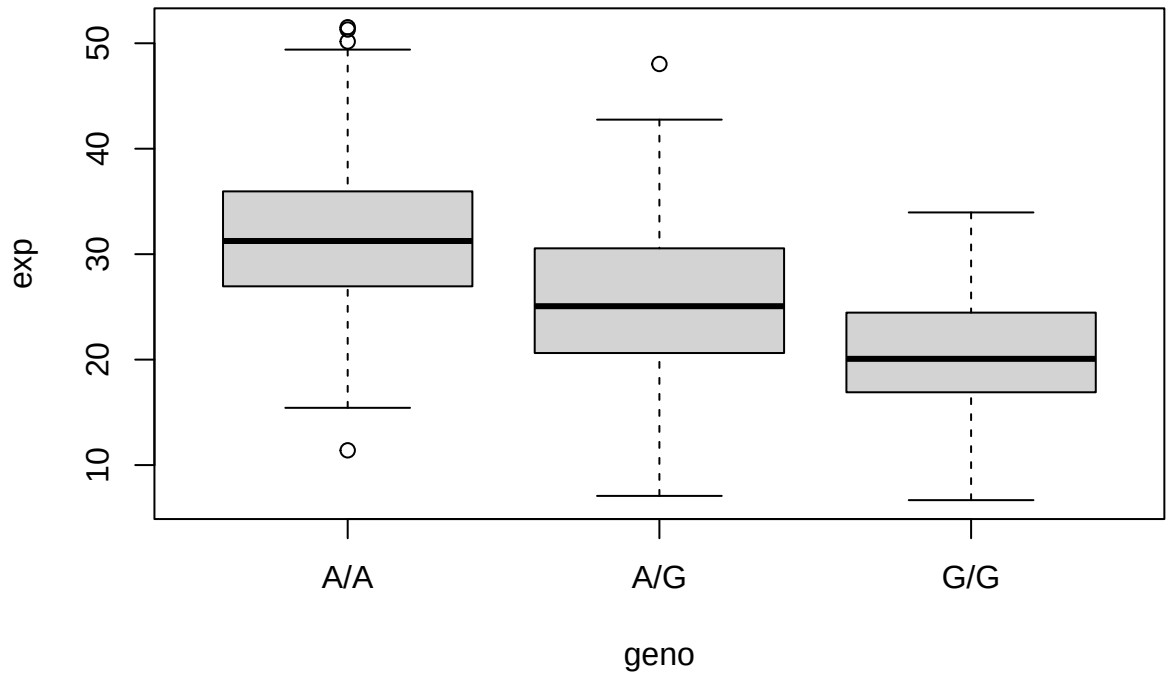
```
## [1] 462
```

> 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.

```
# Sample size for each genotype
table(expr$geno)
```

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

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




```
# Third row contains mdeian values for each genotype
bp$stats
```

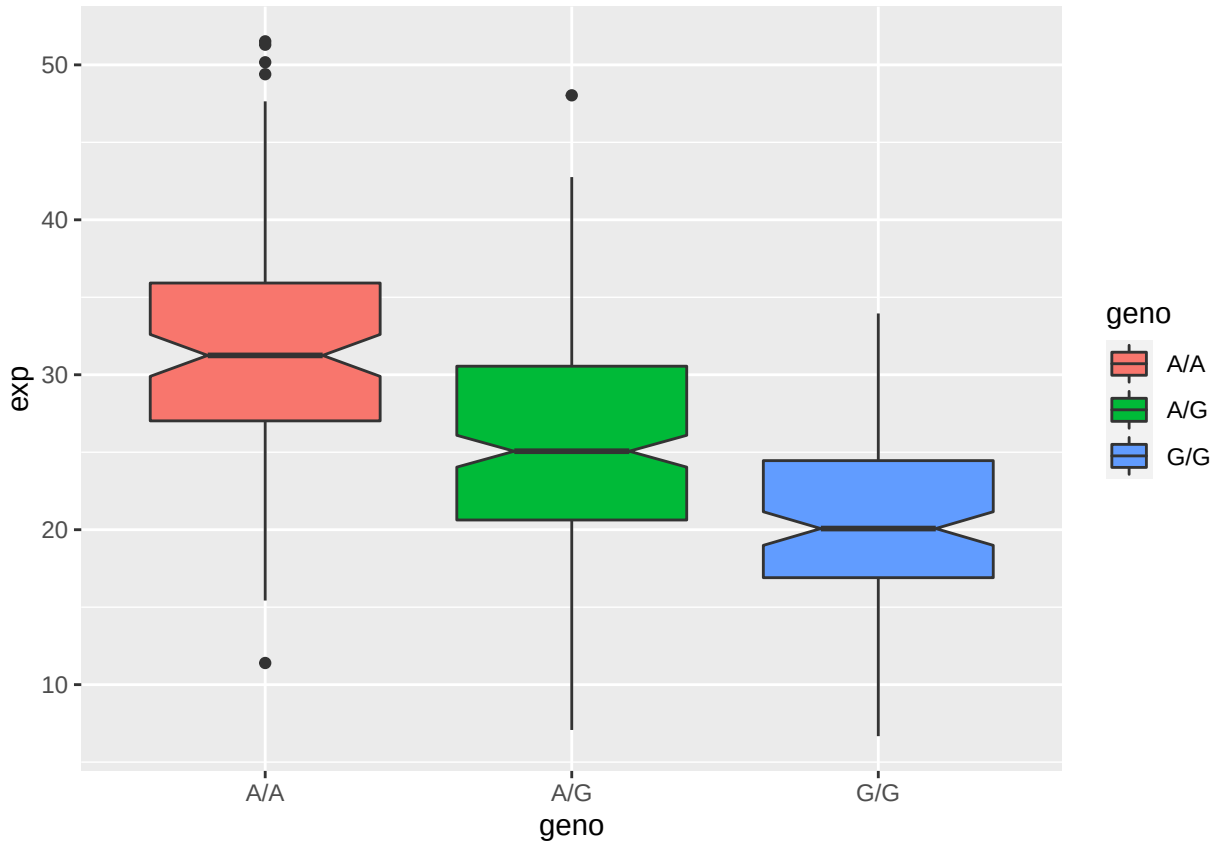
```
##          [,1]     [,2]     [,3]
## [1,] 15.42908  7.07505  6.67482
## [2,] 26.95022 20.62572 16.90256
## [3,] 31.24847 25.06486 20.07363
## [4,] 35.95503 30.55183 24.45672
## [5,] 49.39612 42.75662 33.95602
```

Answer: Genotype A|A has sample size of 108 and median value of 31.25. Genotype A|G has sample size of 233 and median value of 25.06. Genotype G|G has sample size of 121 and median value of 20.07.

> 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)

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



Answer: Homozygous A|A genotype has higher relative expression than that of G|G (i.e. A|A is up-regulated and G|G is down-regulated). Hence, this indicate that the SNP affect the expression of ORMDL3.