

Class10

Samuel Do (PID:A15803613)

2/17/2022

```
#read genotype file\  
mx1 <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")  
mx1
```

##	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	-
## 7	NA19657 (F)	A G	ALL, AMR, MXL	-
## 8	NA19658 (M)	A A	ALL, AMR, MXL	-
## 9	NA19661 (M)	A G	ALL, AMR, MXL	-
## 10	NA19663 (F)	A A	ALL, AMR, MXL	-
## 11	NA19664 (M)	G A	ALL, AMR, MXL	-
## 12	NA19669 (F)	A A	ALL, AMR, MXL	-
## 13	NA19670 (M)	A A	ALL, AMR, MXL	-
## 14	NA19676 (M)	G G	ALL, AMR, MXL	-
## 15	NA19678 (F)	A A	ALL, AMR, MXL	-
## 16	NA19679 (M)	A G	ALL, AMR, MXL	-
## 17	NA19681 (F)	A G	ALL, AMR, MXL	-
## 18	NA19682 (M)	A G	ALL, AMR, MXL	-
## 19	NA19684 (F)	A G	ALL, AMR, MXL	-
## 20	NA19716 (F)	G A	ALL, AMR, MXL	-
## 21	NA19717 (M)	A G	ALL, AMR, MXL	-
## 22	NA19719 (F)	G G	ALL, AMR, MXL	-
## 23	NA19720 (M)	G G	ALL, AMR, MXL	-
## 24	NA19722 (F)	G A	ALL, AMR, MXL	-
## 25	NA19723 (M)	G G	ALL, AMR, MXL	-
## 26	NA19725 (F)	A G	ALL, AMR, MXL	-
## 27	NA19726 (M)	A A	ALL, AMR, MXL	-
## 28	NA19728 (F)	A A	ALL, AMR, MXL	-
## 29	NA19729 (M)	A G	ALL, AMR, MXL	-
## 30	NA19731 (F)	A A	ALL, AMR, MXL	-
## 31	NA19732 (M)	A G	ALL, AMR, MXL	-
## 32	NA19734 (F)	G A	ALL, AMR, MXL	-
## 33	NA19735 (M)	G G	ALL, AMR, MXL	-
## 34	NA19740 (F)	A A	ALL, AMR, MXL	-
## 35	NA19741 (M)	A A	ALL, AMR, MXL	-
## 36	NA19746 (F)	A A	ALL, AMR, MXL	-
## 37	NA19747 (M)	G A	ALL, AMR, MXL	-
## 38	NA19749 (F)	A G	ALL, AMR, MXL	-
## 39	NA19750 (M)	A G	ALL, AMR, MXL	-
## 40	NA19752 (F)	A G	ALL, AMR, MXL	-
## 41	NA19755 (F)	A A	ALL, AMR, MXL	-
## 42	NA19756 (M)	G A	ALL, AMR, MXL	-
## 43	NA19758 (F)	A G	ALL, AMR, MXL	-
## 44	NA19759 (M)	G A	ALL, AMR, MXL	-
## 45	NA19761 (F)	G A	ALL, AMR, MXL	-
## 46	NA19762 (M)	A A	ALL, AMR, MXL	-
## 47	NA19764 (F)	A A	ALL, AMR, MXL	-
## 48	NA19770 (F)	A G	ALL, AMR, MXL	-
## 49	NA19771 (M)	A A	ALL, AMR, MXL	-
## 50	NA19773 (F)	A A	ALL, AMR, MXL	-
## 51	NA19774 (M)	A G	ALL, AMR, MXL	-

## 52	NA19776 (F)	A G ALL, AMR, MXL	-
## 53	NA19777 (M)	A A ALL, AMR, MXL	-
## 54	NA19779 (F)	G A ALL, AMR, MXL	-
## 55	NA19780 (M)	A A ALL, AMR, MXL	-
## 56	NA19782 (F)	G A ALL, AMR, MXL	-
## 57	NA19783 (M)	A G ALL, AMR, MXL	-
## 58	NA19785 (F)	A A ALL, AMR, MXL	-
## 59	NA19786 (M)	G A ALL, AMR, MXL	-
## 60	NA19788 (F)	A G ALL, AMR, MXL	-
## 61	NA19789 (M)	G G ALL, AMR, MXL	-
## 62	NA19792 (M)	A A ALL, AMR, MXL	-
## 63	NA19794 (F)	G A ALL, AMR, MXL	-
## 64	NA19795 (M)	A G ALL, AMR, MXL	-
##	Mother		
## 1	-		
## 2	-		
## 3	-		
## 4	-		
## 5	-		
## 6	-		
## 7	-		
## 8	-		
## 9	-		
## 10	-		
## 11	-		
## 12	-		
## 13	-		
## 14	-		
## 15	-		
## 16	-		
## 17	-		
## 18	-		
## 19	-		
## 20	-		
## 21	-		
## 22	-		
## 23	-		
## 24	-		
## 25	-		
## 26	-		
## 27	-		
## 28	-		
## 29	-		
## 30	-		
## 31	-		
## 32	-		
## 33	-		
## 34	-		
## 35	-		
## 36	-		
## 37	-		
## 38	-		

```
## 39      -
## 40      -
## 41      -
## 42      -
## 43      -
## 44      -
## 45      -
## 46      -
## 47      -
## 48      -
## 49      -
## 50      -
## 51      -
## 52      -
## 53      -
## 54      -
## 55      -
## 56      -
## 57      -
## 58      -
## 59      -
## 60      -
## 61      -
## 62      -
## 63      -
## 64      -
```

```
table(mx1$Genotype..forward.strand.)/nrow(mx1)
```

```
##
##      A|A      A|G      G|A      G|G
## 0.343750 0.328125 0.187500 0.140625
```

Section 4: Population Scale Analysis

[Q13] 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
```

```
nrow(expr)
```

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

```
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
```

```
aggregate(x=expr$exp,  
          by=list(expr$geno),  
          FUN=median)
```

```
##   Group.1      x  
## 1    A/A 31.24847  
## 2    A/G 25.06486  
## 3    G/G 20.07363
```

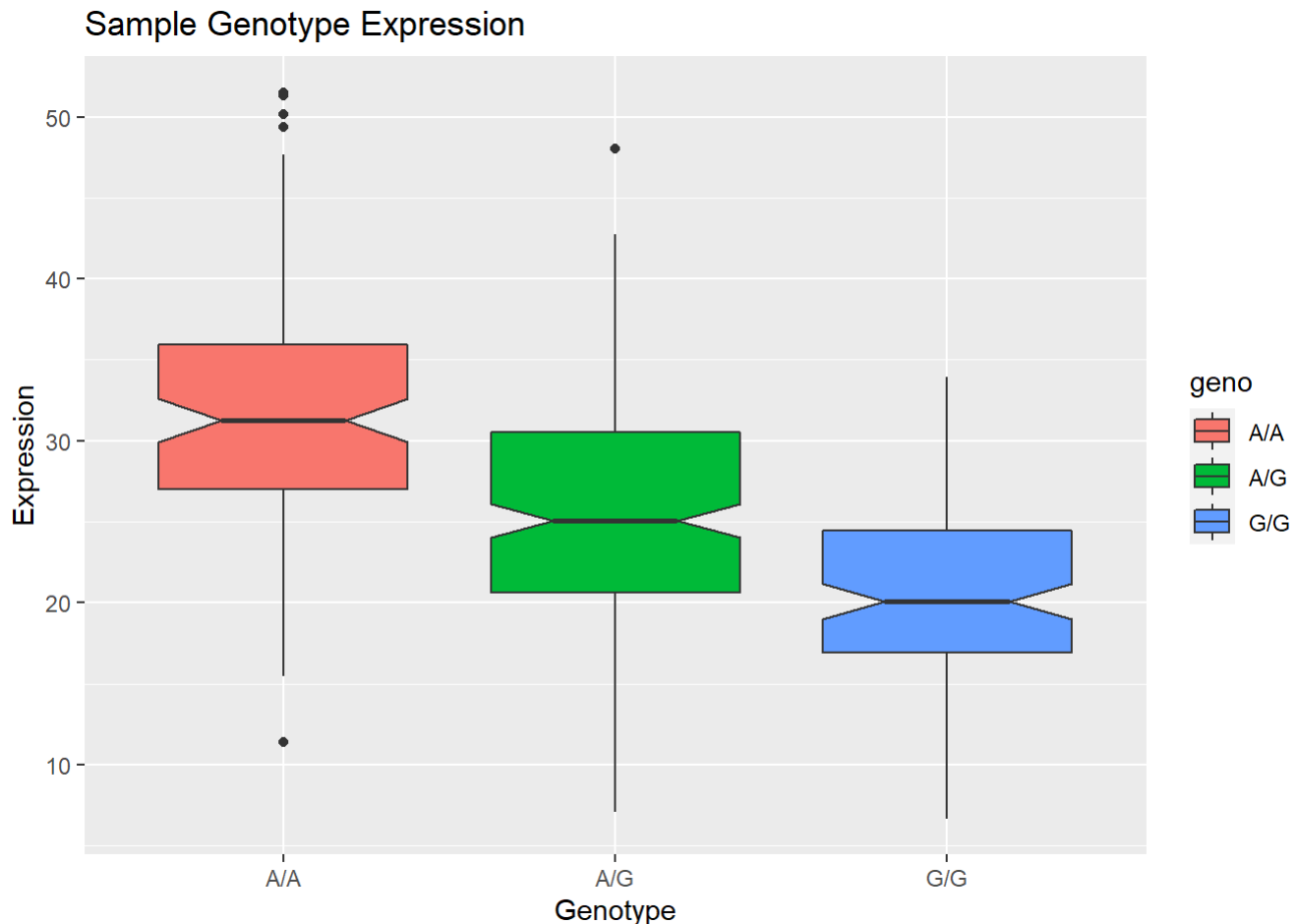
The sample size for each genotype is as follows: There are 108 homozygous A/A samples, 233 heterozygous A/G samples, and 121 homozygous G/G samples. There are a total of 462 samples in this dataset.

The median expression levels for each genotype is as follows: The median expression level of A/A is 31.248, the median expression

level of A/G is 25.065, and the median expression level of G/G is 20.074.

[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 ORM3L3?

```
library(ggplot2)
ggplot(expr, aes(geno, exp, fill=geno)) + geom_boxplot(notch=TRUE) + xlab("Genotype") + ylab("Expression") + labs(title="Sample Genotype Expression")
```



Because the first and third quartiles of the A/A and G/G samples do not overlap, there is no major overlap between the G/G samples and A/A samples' expression levels. Therefore, there is a significant difference in expression levels between the A/A and G/G samples,

which means that the SNP does effect expression of ORMDL3.