class12 hw population scale analysis

amy (pid A16962111)

population scale analysis

Q13. Read this file (https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057.6.t 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.csv")
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
```

The sample size for each genotype is as follows: A/A has 108 samples, A/G has 233 samples, and G/G has 121 samples.

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

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

The median expression level is 31 for A/A, 25 for A/G, and 20 for G/G.

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
expr %>%
  group_by(geno) %>%
  summarize(median(exp))
```

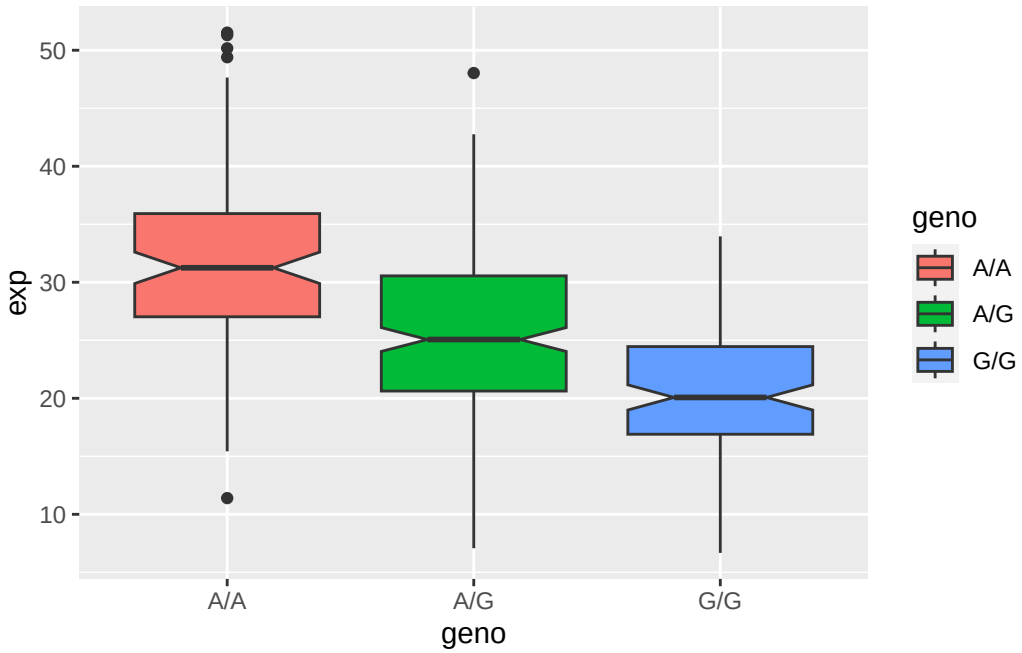
```
# A tibble: 3 x 2
  geno  `median(exp)`
  <chr>         <dbl>
1 A/A            31.2
2 A/G            25.1
3 G/G            20.1
```

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?

From the boxplot below, we can infer that the relative expression between A/A and G/G is significantly different because there is no overlap between the boxes. Since G is the ancestral allele, we can say that the SNP (allele A) increases expression of ORMDL3.

```
library(ggplot2)

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


50
40
30
20
10
exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G