Lab Class11 Pt.1 (RNASeq Galaxy)

Moises Gonzalez (A17579866)

5/21/23

Section 1. Identify genetic variants of interest

Download CSV file and read

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).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
```

Q1: What are those 4 candidate SNPs?

(rs12936231, rs8067378, rs9303277, and rs7216389)

Q2: What three genes do these variants overlap or effect?

rs8067378, rs9303277, rs12936231

Q3: What is the location of rs8067378 and what are the different alleles for rs8067378?

Location: 17:39,894,595-39,895,595, ACG

Q4: Name at least 3 downstream genes for rs8067378?

ENSG00000172057, ENSG00000073605, ENSG00000167914

Q5: What proportion of the Mexican Ancestry in Los Angeles sample population (MXL) are homozygous for the asthma associated SNP (G|G)?

```
mean(mxl$Genotype..forward.strand. == "G|G")
```

```
[1] 0.140625
```

Q6. Back on the ENSEMBLE page, use the "search for a sample" field above to find the particular sample HG00109. This is a male from the GBR population group. What is the genotype for this sample?

G|G

Section 2: Initial RNA-Seq analysis

Q7: How many sequences are there in the first file? What is the file size and format of the data? Make sure the format is **fastqsanger** here!

3,863 sequences and the file has a size of 775 KB and is a fastq format.

Q8: What is the GC content and sequence length of the second fastq file?

GC: 54 sequence length: 50-75

Q9: How about per base sequence quality? Does any base have a mean quality score below 20?

The per base sequence quality is about 35-36 and the base has no mean quality below 20.

Section 3: Mapping RNA-Seq reads to genome

Q10: Where are most the accepted hits located?

Between 38,060,000 and 38,080,000

Q11: Following Q10, is there any interesting gene around that area?

Not sure

Q12: Cufflinks again produces multiple output files that you can inspect from your right-hand- side galaxy history. From the "gene expression" output, what is the FPKM for the ORMDL3 gene? What are the other genes with above zero FPKM values?

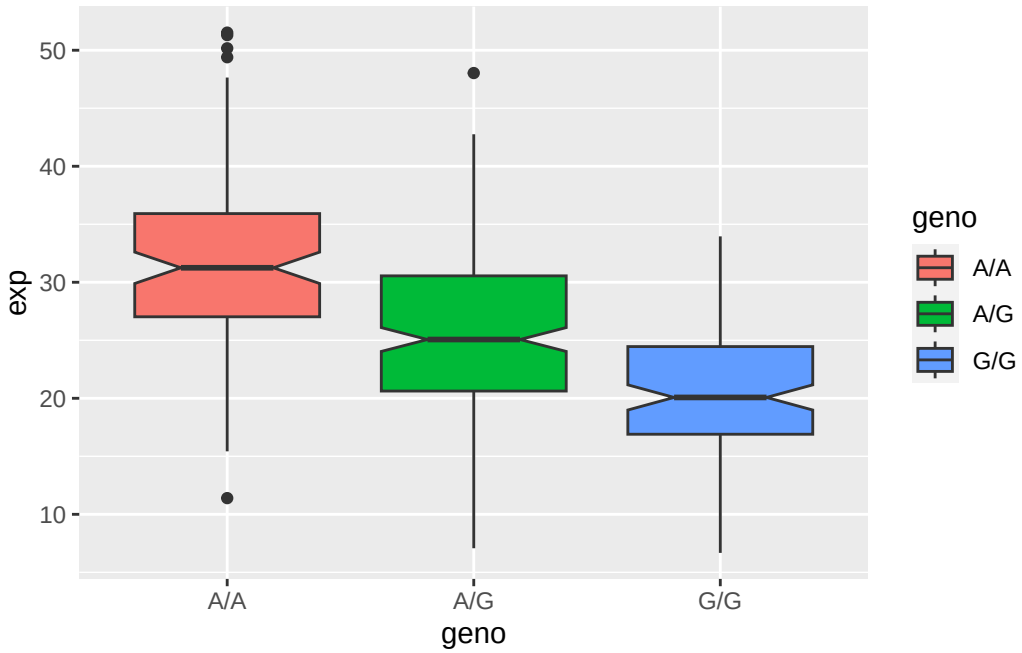
Section 4: Population Scale Analysis

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.

```
library(ggplot2)
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
```

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


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