Lab 12: Genome Informatics

Rahul Nedunuri (PID: A16297840)

Section 1: Proportion of G|G in a population

Q1: What are those 4 candidate SNPs? rs12936231,rs8067378,rs9303277, rs7216389

Q2: What three genes do these variants overlap or effect? ZPBP2, IKZF3, GSDMB

Q3: What is the location of rs8067378 and what are the different alleles for rs8067378? location: Chromosome.17:39895095.(forward strand) alleles: A/C/G | Ancestral:G | Highest population MAF:0.50

Q4: Name at least 3 downstream genes for rs8067378? GSDMA, CSF3, RARA

Let's gather the data of the MXL SNPs for SNP rs8067378

Downloaded 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
```

Q5: What proportion of the Mexican Ancestry in LA sample population (MXL) are homozygous for the asthma associated SNP (G|G)? 14.0625% are homozygous for G|G

Let's look at a diff population. We picked GBR.

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

```
  Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
1                 HG00096 (M)                         A|A ALL, EUR, GBR      -
2                 HG00097 (F)                         G|A ALL, EUR, GBR      -
3                 HG00099 (F)                         G|G ALL, EUR, GBR      -
4                 HG00100 (F)                         A|A ALL, EUR, GBR      -
5                 HG00101 (M)                         A|A ALL, EUR, GBR      -
6                 HG00102 (F)                         A|A ALL, EUR, GBR      -
  Mother
1      -
2      -
3      -
4      -
5      -
6      -
```

Find portion of the population with 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
```

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?

```
gbr$Genotype..forward.strand.[grepl('HG00109', gbr$Sample..Male.Female.Unknown.)]
```

```
[1] "G|G"
```

The genotype for this sample is G|G.

The variant associated with childhood asthma is more common in GBR population than MXL

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

Q8: What is the GC content and sequence length of the second fastq file? 54% GC 50-75 sequence length

Q9: How about per base sequence quality? Does any base have a mean quality score below 20? Trimming is not needed, all bases have a mean quality > 20.

Section 3: Mapping RNA-Seq reads to genome

Q10: Where are most the accepted hits located? chr17:38,150,000

Q11: Following Q10, is there any interesting gene around that area? PSMD3

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? ORMDL3: 136853 GSDMA: 133.634 GSDMB: 26366.3 ZPBP2: 4613.49 PSMD3: 299021

Section 4: Population Scale Analysis

Let's read this file: https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057.6.tx

```
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
```

Let's make the boxplot

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.

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

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

Sample sizes of each genotype A|A: 108 A|G: 233 G|G: 121

```
paste('A|A median expression', median(expr$exp[grepl('A/A', expr$geno)]))
```

```
[1] "A|A median expression 31.248475"
```

```
paste('A|G median expression', median(expr$exp[grepl('A/G', expr$geno)]))
```

```
[1] "A|G median expression 25.06486"
```

```
paste('G|G median expression', median(expr$exp[grepl('G/G', expr$geno)]))
```

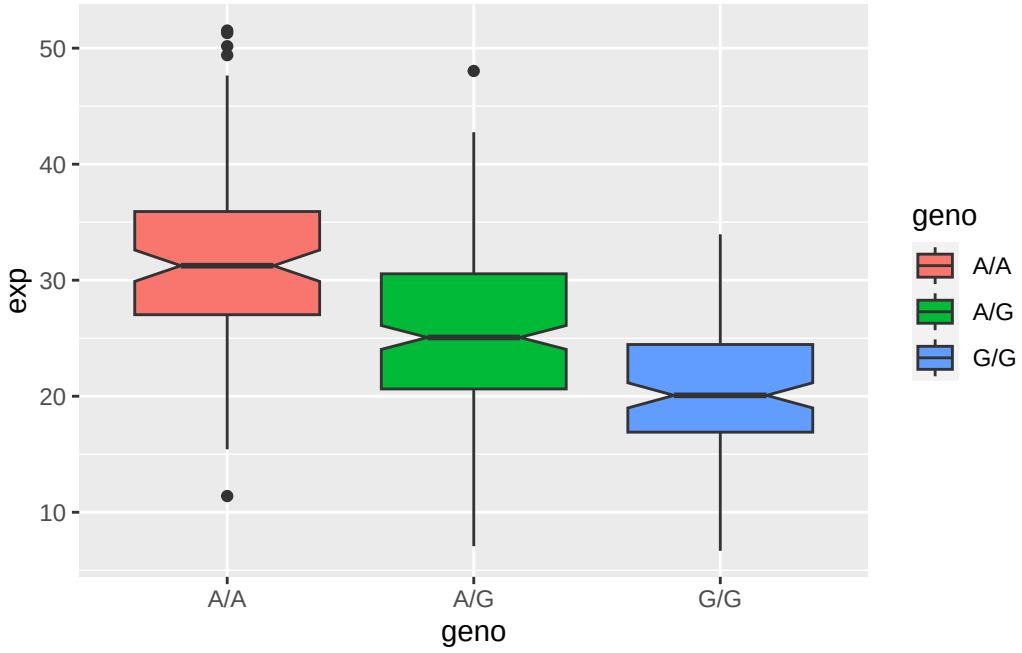
```
[1] "G|G median expression 20.07363"
```

A|A median expression 31.248475 A|G median expression 25.06486 G|G median expression 20.07363

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?

```
#install.packages("ggplot2")
library(ggplot2)

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



It appears that the SNP decreases expression of ORMDL3 in general, although this difference in relative expression level doesn't appear to be statistically significant as the interquartile ranges overlap with the medians of the expression data for each genotype.