Class 12: Introduction to Genome Informatics

Ashley Allen (PID: A14633373)

Table of contents

Section 1: Identify genetic variants of interest

Q1. What are those 4 candidate SNPs?

s12936231, rs8067378, rs9303277, and rs7216389

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

ZPBP2, GSDMB, and ORMDL3

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

location: Chromosome 17:39895095 different alleles: A/C/G, Ancestral: G, Highest population MAF: 0.49

Q4. Name at least 3 downstream genes for rs8067378?

GSDMB-205, ORMDL3-201, LRRC3C-201

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

Here we read in the csv file obtained from ensembl:

```r
mxl <- read.csv("mxl_asthma.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      -
```

```r
table(mxl$Genotype..forward.strand.)
```

```
A|A A|G G|A G|G
 22  21  12   9
```

Proportion of G|G:

```r
round(table(mxl$Genotype..forward.strand.) / nrow(mxl) * 100, 2)
```

```
  A|A   A|G   G|A   G|G
34.38 32.81 18.75 14.06
```

Comparing MXL and GBR:

```r
gbr <- read.csv("gbr_asthma.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      -
```

```
round(table(gbr$Genotype..forward.strand.) / nrow(gbr) * 100, 2)
```

```
  A|A   A|G   G|A   G|G
25.27 18.68 26.37 29.67
```

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!

3863 sequences 741.9 KB fastqsanger

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

54% 3863 seguences

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

The sequence quality is good, there are no mean quality scores below 20.

Section 3: Mapping RNA-Seq reads to genome

Q10. Where are most the accepted hits located?

PSMD3

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

gasdermin A (GSDMA), which from a quick google search says "a protein that forms pores in cell membranes, causing cell death" and are important in host defense.

Q12. Cufflinks again produces multiple output files that you can inspect from your right-handside 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 FPKM: 136853

Above zero FPKM: ZPBP2: 4613.49 GSDMB: 26366.3 GSDMA: 133.634 PSMD3: 299021

Section 4: Population Scale Analysis

Reading in our data:

```
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
```

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.

```
nrow(expr)
```

```
[1] 462
```

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

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

```
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_expr = median(exp))
```

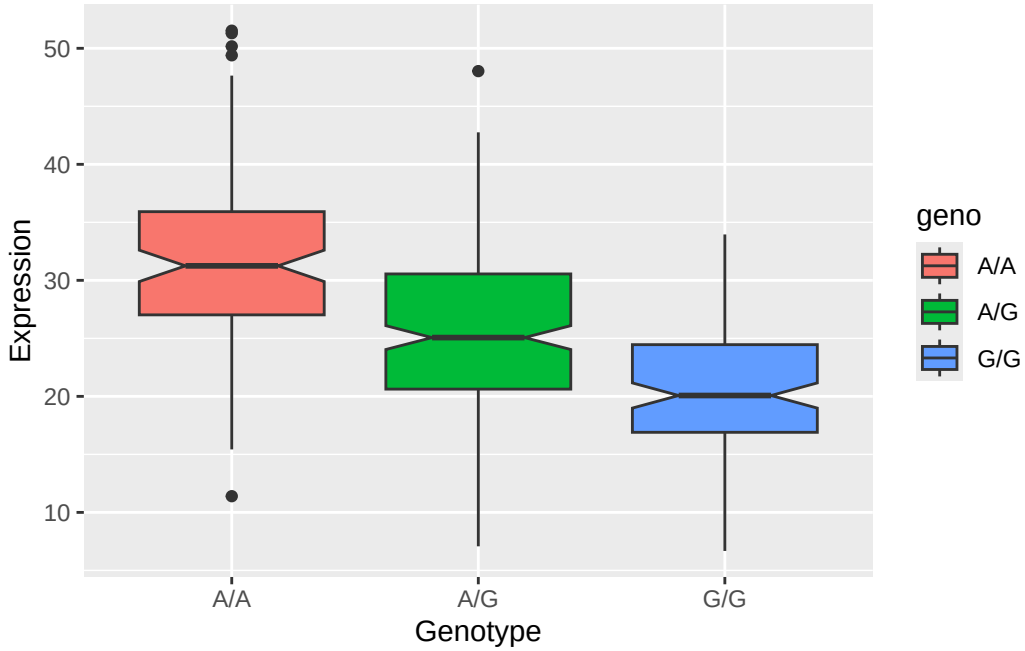
```
# A tibble: 3 x 2
  geno  median_expr
  <chr>       <dbl>
1 A/A          31.2
2 A/G          25.1
3 G/G          20.1
```

> 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?

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
ggplot(expr) + aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE) +
  xlab("Genotype") +
  ylab("Expression")
```



From our box plot you could infer that A|A has a higher expression level overall compared to G|G. The SNP could cause a higher expression level in ORMDL3. We can infer that from our plot as well, since we know the ancestral genotype is G rather than A, and ORMDL3 is associated with that SNP.