Class 11: Genome informatics

Pagna Hout

5/10/23

Section 1: Identify gene of interest

Q1: What are those 4 candidate SNPs?

rs12936231, rs8067378, rs9303277, and 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?

Alleles: A/C/G, Ancestral: G, MAF: 0.43 (G)

Location: Chromosome 17:39895095 (forward strand)

Q4: Name at least 3 downstream genes for rs8067378?

GRB7, IKZF3, MIEN1

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

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

table(mxl$Genotype..forward.strand.) / nrow(mxl) * 100
```

```
    A|A     A|G     G|A     G|G 
34.3750 32.8125 18.7500 14.0625 
```

14% are homozygous for the asthma associated SNP (G|G).

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?

The genotype for this sample is 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?

There are 3863 sequences in the first file.

The file size is 741.9 KB. The format is fastqsanger.

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

The GC content is 54%. The sequence of the second fastq file is 50-75.

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

None of the base have a mean quality score below 20. Trimming is not needed for the data set.

Section 3: Mapping RNA-Seq reads to genome

Q10: Where are most the accepted hits located?

PSMD3, ORMDL3, GSDMB

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

PSMD3, ORMDL3, GSDMB

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?

The FPKM for the ORMDL3 gene is 136853.

GSDMA, PSMD3, GSDMB, and ZPBP2 are genes with FPKM values above 0.

Section 4: Population Scale Analysis [HOMEWORK]

```
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.

```
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
```

```
gene_summary <- expr %>%
  group_by(geno) %>%
  summarize(Sample_Size = n(), Median_exp = median(exp))

print(gene_summary)
```

```
# A tibble: 3 x 3
  geno  Sample_Size Median_exp
  <chr>       <int>      <dbl>
1 A/A           108       31.2
2 A/G           233       25.1
3 G/G           121       20.1
```

There are 108 A|A, 233 A|G, and 121 G|G.

The median expression level for A|A is 31.25.

The median expression level for A|G is 25.06.

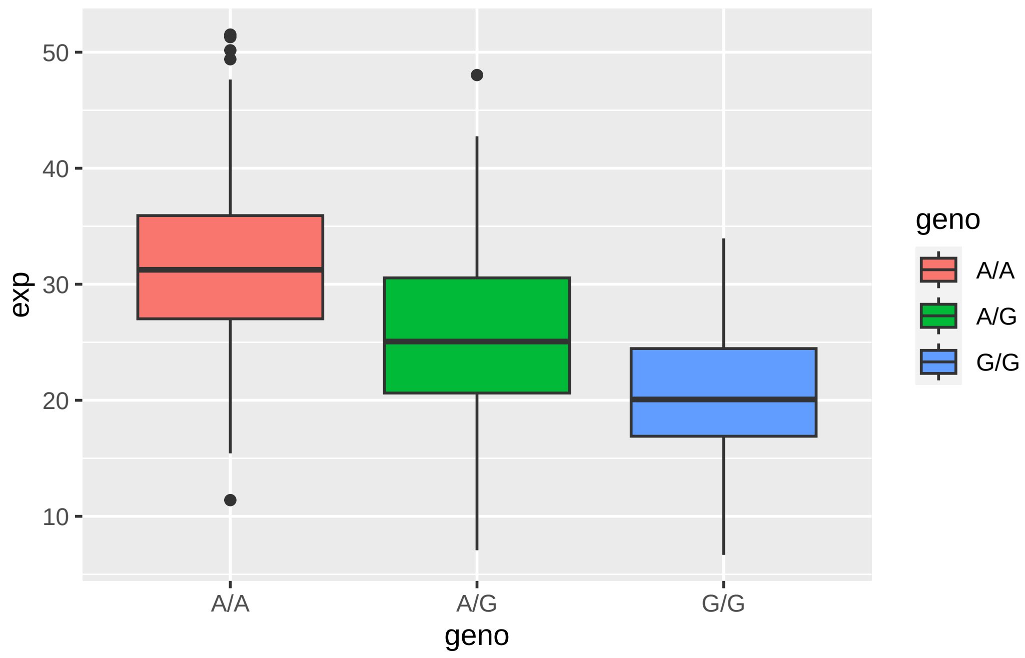
The median expression level for G|G is 20.07.

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)
```

Boxplot:

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



The G|G genotype is associated with having a reduced expression of the gene compared to the A|A genotype.