Class12 Structural Bioinformatics (pt2. Focus on new AlphaFold2

AUTHOR
Erin McTavish PID: A17300519

> Q1. What are those 4 candidate SNPs?

rs12936231, rs8067378, rs9303277, rs7216389

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

GSDMB (Gasdermin B): encodes a protein responsible for tissue repair, cell death, and immune responses

ZPBP2 (Zona Pellucida Binding Protein 2): Ensures proper sperm development

IKZF3 (Ikaros Family Zinc Finger 3): High concentration in immune tissues, and esnures B and T cell development.

> Q3: What is the location of rs8067378 and what are the different alleles for rs8067378? [HINT, alleles and location are listed at the top of the Ensemble page as chromosome number and position. You may search in a genome browser to find this information]

A/C/G|Ancestral: G|Highest population MAF: 0.50 Chromosome 17:39895095

> Q4: Name at least 3 downstream genes for rs8067378?

PSMD3

GSDMA

ORMDL3

Section 1. Proportion of G/G in a population

Downloaded CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39894595-39895595;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel

Here we read this CSV file

```
mxl <- data <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.c
                stringsAsFactors = FALSE)
head(mxl)
```

```
  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      -
```

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).csv",
                stringsAsFactors = FALSE)
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      -
```

> Q5: What proportion of the Mexican Ancestry in Los Angeles sample population (MXL) are homozygous for the asthma associated SNP (G|G)? [HINT: You can filter the displayed genotypes by entering the population code MXL. Then either count those of interest or download a CVS file for this population and use excel or the R functions read.csv(), and table() to answer this question]

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

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

14.0625%

> 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

```
# Check where R is currently looking
getwd()
```

```
[1] "/Users/erin/Documents/BIMM 143/class12/class 12"
```

```
# Show what R can see in that folder
list.files()
```

```
[1] "373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv"
[2] "373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).csv"
[3] "class 12.Rproj"
[4] "class12pt1.qmd"
[5] "class12pt1.rmarkdown"
[6] "rs8067378_ENSG00000172057.6.txt"
```

```
# Now read the NEW csv directly from your class folder
mxl <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv",
                stringsAsFactors = FALSE)

head(mxl)
```

```
  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      -
```

Now let's look at a different population. I picked the GBR.

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

Find proportion of G|G

```
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 
```

29.67%

This variant that is associated with childhood asthma is more frequent in the GBR population than the MKL population.

Lets dig into this further.

> 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! [HINT, you can check the fastq format wiki for more information]

3,863 sequences File size 0.74 MB Format: FASTQ

> Q8: What is the GC content and sequence length of the second fastq file? [HINT, you may check “Basic Statistics”]

GC content: 54.14% Sequence length: ~74.7 base pairs

> Q9: How about per base sequence quality? Does any base have a mean quality score below 20? [HINT, blue line is the mean quality score and for this exercise, assume a median quality score of below 20 to be unusable. Given this criterion, is trimming needed for the dataset?]

No base has a mean quality score below 20, so no trimming is needed.

> Q10: Where are most the accepted hits located? [HINT, you can view the SAM version of your accepted hits file in galaxy and also use the UCSC Genome Browser via following the galaxy provided link and focusing on particular regions as described above]

Around IKZF3, GSDMB, and ORMDL3.

> Q11: Following Q10, is there any interesting gene around that area? [HINT, you can find genes around accepted hits in either the UCSC Genome Browser or IGV - depending on which browser you prefer]

There is a gene cluster that overlaps with ORMDL3. This is interesting, particularly because this gene was noted in previous questions for being associated with asthma.

> 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?

FPKM for the ORMDL3: 128189 Other genes with above zero FPKM values: PSMD3, GSDMB, ZPBP2

> 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. Hint: The read.table(), summary() and boxplot() functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the boxplot() function to an R object and examining this object. There is also the medium() and summary() function that you can use to check your understanding. How many sample do we have?

```
 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
```

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

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

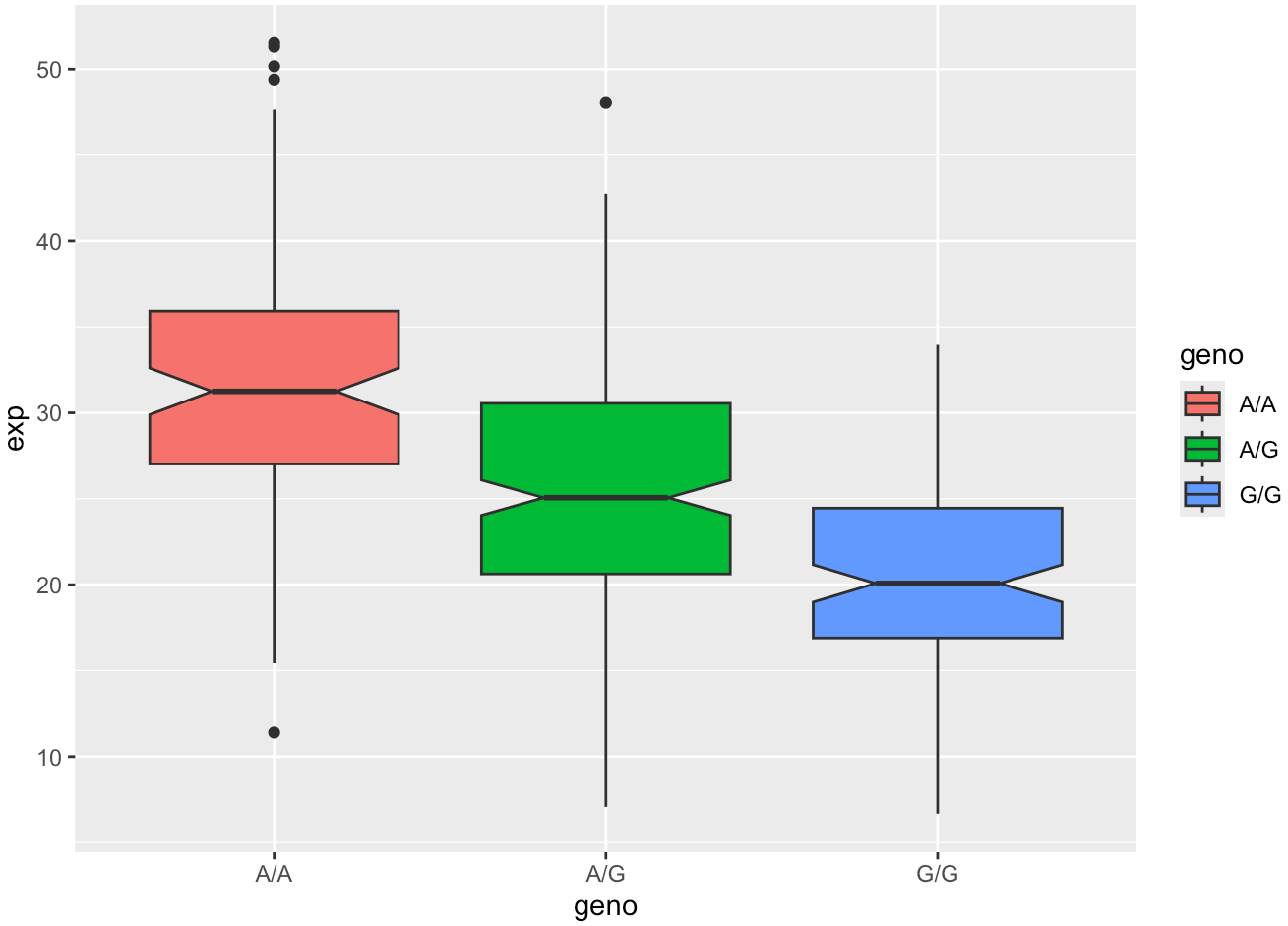
```
summary(expr)
```

```
    sample              geno                exp
 Length:462         Length:462         Min.   : 6.675
 Class :character   Class :character   1st Qu.:20.004
 Mode  :character   Mode  :character   Median :25.116
                                       Mean   :25.640
                                       3rd Qu.:30.779
                                       Max.   :51.518
```

462 samples in the dataset. A|G: 233 A|A: 108 Median ORMDL3 expression levels were highest ~31 in A|A group. Intermediate expression levels in A|G: ~25 Lowest expression levels in G|G ~20.

> Q14: Generate a boxplot with a box per genotype, what could you infer from the rel- ative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3? Hint:Anexampleboxplotisprovided overleaf – yours does not need to be as polished as this one.

```
library(ggplot2)
ggplot(expr) +
  aes(geno, exp, fill = geno) +
  geom_boxplot(notch = TRUE)
```



The boxplot shows that ORMDL3 expression differs depending on genotype. Individuals with the A/A genotype have the highest median expression, those with A/G display intermediate levels, and individuals with G/G have the lowest expression. This trend indicates that the rs8067378 SNP influences ORMDL3 expression in a dose-dependent manner, where carrying the G allele is associated with reduced gene expression.

```
SampleGenotypes <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs80673
                            stringsAsFactors = FALSE)
```