Class 11

Emmanuel Robles

Section 1. Proportion of G|G in a population.

Q1: What are those 4 candidate SNPs?

(rs12936231, rs8067378, rs9303277, and rs7216389)

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

rs930… IKZF3 rs721… GSDMB rs806… none

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

Chromosome 17:39895095 (forward strand)

Q4: Name at least 3 downstream genes for rs8067378?

ORMDL3 GSDMB GSDMA

Downloaded csv file from Ensamble. Now we'll read the csv file.

```
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
```

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

14% of the sample population are homozygous for the asthma SNP.

Now let's look at a different population. 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
```

> 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

This variant that is associated with childhood asthma is more frequent in the bgr population than the MXL population. Let's dig further.

Section 2

> 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

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

53%

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

All the mean quality scores for the bases are at 30 or higher. No.

Section 3

Q10: Where are most the accepted hits located?

Most of his are located around ~38,060,000 and ~38,150,000

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

For the first we see ORMDL3 an for the one furthest away we see 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?

128189