12: Genome informatics

Mason Lew (PID: A17533139

Section 1: Identify genetic variants of interest

Q1. What are those 4 candidate SNPs?

The 4 candidate SNPs are: rs12936231, rs8067378, rs9303277, and rs7216389

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

ZPBP2

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

Location: Chromosome 17:39895095 Alleles: A/C/G | Ancestral: G | MAF: 0.49

Q4. Name at least 3 downstream genes for rs8067378?

GSDMA, CSF3, GSDMB, LRRC3C

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

Reading the CSV and tabling the data:

```
MXL <- read.csv("MXL data.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
```

Find the porportion of G|G:

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

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

##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

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

%GC: 54 Sequence length: 50-75

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

No, no base has a mean minimum quality score below 20

Section 3: Mapping RNA-Seq reads to genome

Q10. Where are most the accepted hits located?

The most accepted hits are located between 38,150,000 and 38,160,000

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-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 ORMDL3: 128189 Other genes with above zero FPKM values: ZPBP2, GSDMB, and PSMD3