class17

Ryan Bench

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensemble

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.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.) / nrow(mxl) * 100
```

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

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

```
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
```

Q1: What are those 4 candidate SNPs?

The 4 SNPs are rs12936231, rs8067378, rs9303277, and rs7216389.

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

ZPBP2, IKZF3, and GSDMB

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

Chromosome 17:39872867 is the location, the different alleles listed are A/C/G.

Q4: Name at least 3 downstream genes for rs8067378?

GSDMB-219, GSDMB-205, and GSDMB-201

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 samples in the dataset are homozygous for 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?

This sample is homozygous.

> 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!

There are 3,863 sequences in the first file. The file size is 741.9 KB, and the format is fastqsanger.

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

The GC content is 54%, and the sequence length is 50-75.

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

Everything looks to be of usable quality, trimming is not necessary in this case.

> Q10: Where are most the accepted hits located?

The most accepted hits located are PSMD3, ORMDL3, and IKZF3.

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

One gene that may be of interest is GSDMB.

> 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?

The FPKM value is 128189 for the ORMDL3. GSDMB, GSDMA, PSMD3 and ZPBP2 have above 0 FPKM values.