Class 11 Lab: RNASeq Galaxy

Jessica Diaz-Vigil

2023-05-21

Section 1: Identify Genetic Variants of Interest

Q1: What are those 4 candidate SNPs? 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?

A/C/G

Ancestral: G|MAF: 0.43 (G)

Highest population MAF: 0.50

Chromosome 17: 39,895,045-39,895,145 (forward strand)

Q4: Name at least 3 downstream genes for rs8067378?

GDSMA, CASC3, WIPF2

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

```r
mxl <- read.csv("MXL.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      -
```

```r
table(mxl$Genotype..forward.strand.)
```

```
##
## A|A A|G G|A G|G
##  22  21  12   9
```

```r
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% of the people with Mexican ancestry in Los Angeles 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?

3,863 sequences

format fastqsanger

741.9 KB

Q8: What is the GC content and sequence length of the second fastq file? The GC content is 54%

The sequence length is 50-75

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

There are no bases with a mean quality score under 20, so no trimming will be required

Section 3: Mapping RNA-Seq Reads to Genome

Q10: Where are most the accepted hits located?

Most of the accepted hits are in between 38,050,000 and 38,100,000

(Viewed SAM in UCSC Genome Browser)

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

GSDMB is around that area as well as ORMDL3

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?

136853

The other genes are GSDMA, GSDMB and ZPBP2