Class 11: Genome Informatics

Heidi Nam

Identifying genetic variants of interest

Q1: What are those 4 candidate SNPs?

- The 4 candidate SNPs that are correlated with childhood asthma are rs12936231, rs8067378 , rs9303277 , rs7216389

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

- rs12936231 affects ZPBP2, rs9303277 affects IKZF3, and rs7216389 affects GSDMB.

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

- location of rs8067378: chromosome 17:39895095
- alleles: A,C,G

Q4: Name at least 3 downstream genes for rs8067378?

- GSDMA, PSMD3, THRA

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

```
rs806.data <- read.csv("rs8067378.csv")
table(rs806.data$Genotype..forward.strand.)
```

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

```
9 / (9 + 12 + 21 + 22)
```

```
[1] 0.140625
```

For the rs8067378 SNP, 14.1% of the sample population 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 is also G|G for this sample HG00109.

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!

- There are 3863 sequences with file size of 741.9KB and format of fastqsanger.

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

- GC content of 53 and sequence length of 50-75

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

- There is no base with a median quality score below 20.

Mapping RNA-seq reads to genome

Q10: Where are most the accepted hits located?

- chr17:38007296-38170000

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

- IKZF3, GSDMB, and ORMDL3

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?

- ORMDL3 has a FPKM value of 136853, and genes with above zero FPKM values included GSDMA, GSDMB, ZPBP2