Class 12- Introduction to Genome Informatics

Renee Zuhars (PID: A17329856)

Table of contents

Section 1: Identify Genetic Variants of Interest

There are a number of gene variants associated with childhood asthma. A study from Verlaan et al. (2009) shows that 4 candidate SNPs demonstrate significant evidence for association.

Question 1

What are those 4 candidate SNPs?

After a bit of digging, I was able to find the full paper here: cell.com, but still used OMIM to find the following:

The four candidate SNPs are rs12936231, rs8067378, rs9303277, and rs7216389.

Question 2

> What three genes do these variants overlap or affect?

rs12936231 affects the gene ZPBP2. rs9303277 affects the gene IKZF3. rs7216389 affects the gene GSDMB. rs8067378 has not yet been proven to affect any variants.

Question 3

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

The location of rs8067378 is on Chromosome 17:39,895095 (forward strand), and the different alleles are A/C/G, with the ancestral allele being G.

Question 4

> Name at least 3 downstream genes for rs8067378?

Some downstream genes include ZPBP2, GSDMA, LRRC3C, and PSMD3.

Question 5

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

```
# Save in project directory
MXL.data <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")

# input data, store as MXL.df
MXL.df <- data.frame(MXL.data, row.names = 1)

head(MXL.df)
```

```
            Genotype..forward.strand. Population.s. Father Mother
NA19648 (F)                       A|A ALL, AMR, MXL      -      -
NA19649 (M)                       G|G ALL, AMR, MXL      -      -
NA19651 (F)                       A|A ALL, AMR, MXL      -      -
NA19652 (M)                       G|G ALL, AMR, MXL      -      -
NA19654 (F)                       G|G ALL, AMR, MXL      -      -
NA19655 (M)                       A|G ALL, AMR, MXL      -      -
```

```
# save genotype as a vector
genotype <- factor(MXL.df$Genotype..forward.strand.)

# to find proportion homozygous for G/G:
table(genotype)
```

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

Out of the 64 observations, 9 are homozygous for the G/G SNP. This means that about 14 percent of the Mexican Ancestry in Los Angeles sample population is homozygous for G/G.

Question 6

> 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 HG00109 is G/G.

Section 2: Initial RNA-seq analysis

The FASTQ files related to HG00109 were downloaded into this project directory.

Question 7

> How many sequences are there in the first file? What is the file size and format of the data?

In both HG00109_1.fastq and HG00109_2.fastq, there are 3,863 sequences. The file size for HG00109_1 is 741.9 KB, and the file size for HG00109_2 is 741.2 KB. The format of the data is fastqsanger.

Some quality control was run on the data (both files) using the FastQC tool. The outputs included a webpage and a raw data file for both.

Question 8

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

When looking at the webpage for HG00109_2, the GC content (%GC) is 54%. The sequence length is 50-75 bp.

Question 9

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

When looking at the per base sequence quality on the webpage for HG00109_2, no bases appear to have a mean quality score below 20, as none of them are in the red range. There are a couple in the yellow range, but the majority are in the green.

Section 3: Mapping RNA-Seq reads to genome

The Tophat tool within Galaxy was used to map RNA-seq reads to the hg19 build of the Human reference genome. Our data was listed as paired end, and HG00109_1 was placed in the first (forward read) box while HG00109_2 was placed in the second (reverse read) box.

Out of the results, we are focusing only on the alignment of the accepted_hits data. The accepted_hits data was then converted from BAM format to SAM format using the BAM -> SAM conversion tool within Galaxy.

From here, the 'display at UCSC main' link was clicked, and the location was set to chr17:38007296-38170000.

Question 10

> Where are most of the accepted hits located?

Most of the accepted hits are located between chr17: 38,140,000-38,160,000 and chr17: 38,070,000-38,090,000.

Question 11

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

The areas where most of the accepted hits are located also include the genes PSMD3, GSDMA, GSDMB, and ORMDL3.

For the final step, the file genes.chr17.gtf was downloaded from the class website, and then run through Galaxy's 'cufflinks' tool.

Question 12

> 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?

The FPKM for the ORMDL3 gene is 136853. The other genes with above zero FPKM values include ZPBP2, GSDMB, GSDMA, and PSMD3.