Genome Informatics

Matt Hashimoto

12/6/2021

Identify Genetic Variants of Interest

Q1 “What are those 4 candidate SNPs?”

The four candidate SNPs investigated in the Verlaan et al. paper about childhood asthma are rs12936231, rs8067378, rs9303277, and rs7216389.

Q2 “What three genes do these variants overlap or effect?”

Based on the ENSEMBL pages for the SNPs listed above, the three genes they affect are ZPBP2, IKZF3, and GSDMB.

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

rs8067378 is located at position 17:39895095, meaning it is on chromosome 17. The different alleles are A, C, and G.

Q4 “Name at least 3 downstream genes for rs8067378?”

Three downstream genes are LRRC3C, PSMD3, and CSF3.

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

```r
# Load data from .csv file and check
mxlname <- "373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv"
mxldata <- read.csv(mxlname)
table(mxldata)
```

```
## , , Population.s. = ALL, AMR, MXL, Father = -, Mother = -
##
##                              Genotype..forward.strand.
## Sample..Male.Female.Unknown. A|A A|G G|A G|G
##                  NA19648 (F)   1   0   0   0
##                  NA19649 (M)   0   0   0   1
##                  NA19651 (F)   1   0   0   0
##                  NA19652 (M)   0   0   0   1
##                  NA19654 (F)   0   0   0   1
##                  NA19655 (M)   0   1   0   0
##                  NA19657 (F)   0   1   0   0
##                    NA19658 (M)   1   0   0   0
##                    NA19661 (M)   0   1   0   0
##                    NA19663 (F)   1   0   0   0
##                    NA19664 (M)   0   0   1   0
##                    NA19669 (F)   1   0   0   0
##                    NA19670 (M)   1   0   0   0
##                    NA19676 (M)   0   0   0   1
##                    NA19678 (F)   1   0   0   0
##                    NA19679 (M)   0   1   0   0
##                    NA19681 (F)   0   1   0   0
##                    NA19682 (M)   0   1   0   0
##                    NA19684 (F)   0   1   0   0
##                    NA19716 (F)   0   0   1   0
##                    NA19717 (M)   0   1   0   0
##                    NA19719 (F)   0   0   0   1
##                    NA19720 (M)   0   0   0   1
##                    NA19722 (F)   0   0   1   0
##                    NA19723 (M)   0   0   0   1
##                    NA19725 (F)   0   1   0   0
##                    NA19726 (M)   1   0   0   0
##                    NA19728 (F)   1   0   0   0
##                    NA19729 (M)   0   1   0   0
##                    NA19731 (F)   1   0   0   0
##                    NA19732 (M)   0   1   0   0
##                    NA19734 (F)   0   0   1   0
##                    NA19735 (M)   0   0   0   1
##                    NA19740 (F)   1   0   0   0
##                    NA19741 (M)   1   0   0   0
##                    NA19746 (F)   1   0   0   0
##                    NA19747 (M)   0   0   1   0
##                    NA19749 (F)   0   1   0   0
##                    NA19750 (M)   0   1   0   0
##                    NA19752 (F)   0   1   0   0
##                    NA19755 (F)   1   0   0   0
##                    NA19756 (M)   0   0   1   0
##                    NA19758 (F)   0   1   0   0
##                    NA19759 (M)   0   0   1   0
##                    NA19761 (F)   0   0   1   0
##                    NA19762 (M)   1   0   0   0
##                    NA19764 (F)   1   0   0   0
##                    NA19770 (F)   0   1   0   0
##                    NA19771 (M)   1   0   0   0
##                    NA19773 (F)   1   0   0   0
##                    NA19774 (M)   0   1   0   0
##                    NA19776 (F)   0   1   0   0
##                    NA19777 (M)   1   0   0   0
##                    NA19779 (F)   0   0   1   0
##                    NA19780 (M)   1   0   0   0
##                    NA19782 (F)   0   0   1   0
##                    NA19783 (M)   0   1   0   0
##                    NA19785 (F)   1   0   0   0
##                    NA19786 (M)   0   0   1   0
##                    NA19788 (F)   0   1   0   0
##                    NA19789 (M)   0   0   0   1
##                    NA19792 (M)   1   0   0   0
##                    NA19794 (F)   0   0   1   0
##                    NA19795 (M)   0   1   0   0
```

```
# Count G|G column of table and divide by total count
colSums(table(mxldata))[4] / sum(colSums(table(mxldata)))
```

```
## [1] 0.140625
```

Thus, 14.06% of those with Mexican Ancestry in the Los Angeles sample population are homozygous for the asthma-associated SNP.

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.

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 is only one sequence in the first file. The file is 741.9 KB and is in the fastqsanger format.

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

The percent 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?"

No base has a mean quality score below 20, and thus no trimming is required.

Mapping RNA-Seq Reads to Genome

Q10 "Where are most the accepted hits located?"

The vast majority of hits are located on Chromosome 17, from 38.05 Mbp to 38.1 Mbp. There are also smaller regions of note upstream and downstream.

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

GSDMB and ORMDL3 are located near the center, where most of the accepted hits are clustered. Upstream are IKZF3 and ZPBP2, while downstream is PSMD3.

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 for the ORMDL3 gene is 136853. Other genes with FPKM values > 0 are PSMD3, GSDMA, GSDMB, and ZPBP2.