W8_Genomics_Lab

Vera Sophia Beliaev

2/17/2022

Section 1: Identify genetic variants of interest

> Q1: What are those 4 candidate SNPs? rs12936231, rs8067378, rs9303277, and rs7216389
>
> Q2: What three genes do these variants overlap or effect? ZPBP2, GSDMB, and ORMDL3
>
> Q3: What is the location of rs8067378 and what are the different alleles for rs8067378? The alleles (different since recorded video) are A/C/G. The location is Chromosome 17:39895095.
>
> Q4: Name at least 3 downstream genes for rs8067378? LRRC3C, MSL1, ND1D1
>
> Q5: What proportion of the Mexican Ancestry in Los Angeles sample population (MXL) are homozygous for the asthma associated SNP (G|G)? 0.140625 or ~14%

Downloaded a CSV file from Ensembl

```
#Read csv file, can type in first few chrs in file name and tab to autocomplete
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).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      -
```

```
mxl$Genotype..forward.strand.
```

```
##  [1] "A|A" "G|G" "A|A" "G|G" "G|G" "A|G" "A|G" "A|A" "A|G" "A|A" "G|A" "A|A"
## [13] "A|A" "G|G" "A|A" "A|G" "A|G" "A|G" "A|G" "G|A" "A|G" "G|G" "G|G" "G|A"
## [25] "G|G" "A|G" "A|A" "A|A" "A|G" "A|A" "A|G" "G|A" "G|G" "A|A" "A|A" "A|A"
## [37] "G|A" "A|G" "A|G" "A|G" "A|A" "G|A" "A|G" "G|A" "G|A" "A|A" "A|A" "A|G"
## [49] "A|A" "A|A" "A|G" "A|G" "A|A" "G|A" "A|A" "G|A" "A|G" "A|A" "G|A" "A|G"
## [61] "G|G" "A|A" "G|A" "A|G"
```

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

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

```
#divide by nrow because number of rows in table is the number of people total
table(mxl$Genotype..forward.strand.) / nrow(mxl) * 100
```

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

> 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? G|G

#Section 2: Initial RNA-Seq analysis USing Galaxy for NGS analyses

> 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! The first file has 3,863 sequences and a file size of 741.9 kb The format is fastqsanger where the first line is @a descriptor, the second line are the bases, the third a +, and the fourth quality scores for each base.

Quality Control with Fast QC

> Q8: What is the GC content and sequence length of the second fastq file? There is 54% GC content and a sequence length of 50-75.

> Q9: How about per base sequence quality? Does any base have a mean quality score below 20? The per base sequence quality is good and there is no bases with mean quality scores below 20.

Section 3: Mapping RNA-Seq reads to genome

using Bowtie

> Q10: Where are most the accepted hits located? The greatest number of hits map over PSMD3 and ORMDL3

> Q11: Following Q10, is there any interesting gene around that area? PSMD3 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? The FPKM for the ORMDL3 gene is 136853. Other genes with above zero FPKM values are GSDMB, GSDMA, ZPBP2, and PSMD3.

Section 4: Population Analysis

> Q13: Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```
expr <-  read.table("rs8067378_ENSG00000172057.6.txt")
head(expr)
```

```
##    sample geno      exp
## 1 HG00367  A/G 28.96038
## 2 NA20768  A/G 20.24449
## 3 HG00361  A/A 31.32628
## 4 HG00135  A/A 34.11169
## 5 NA18870  G/G 18.25141
## 6 NA11993  A/A 32.89721
```

```
#number of samples
nrow(expr)
```

```
## [1] 462
```

```
#Make a table counting each genotype
table(expr$geno)
```

```
##
## A/A A/G G/G
## 108 233 121
```

```
summary(expr)
```

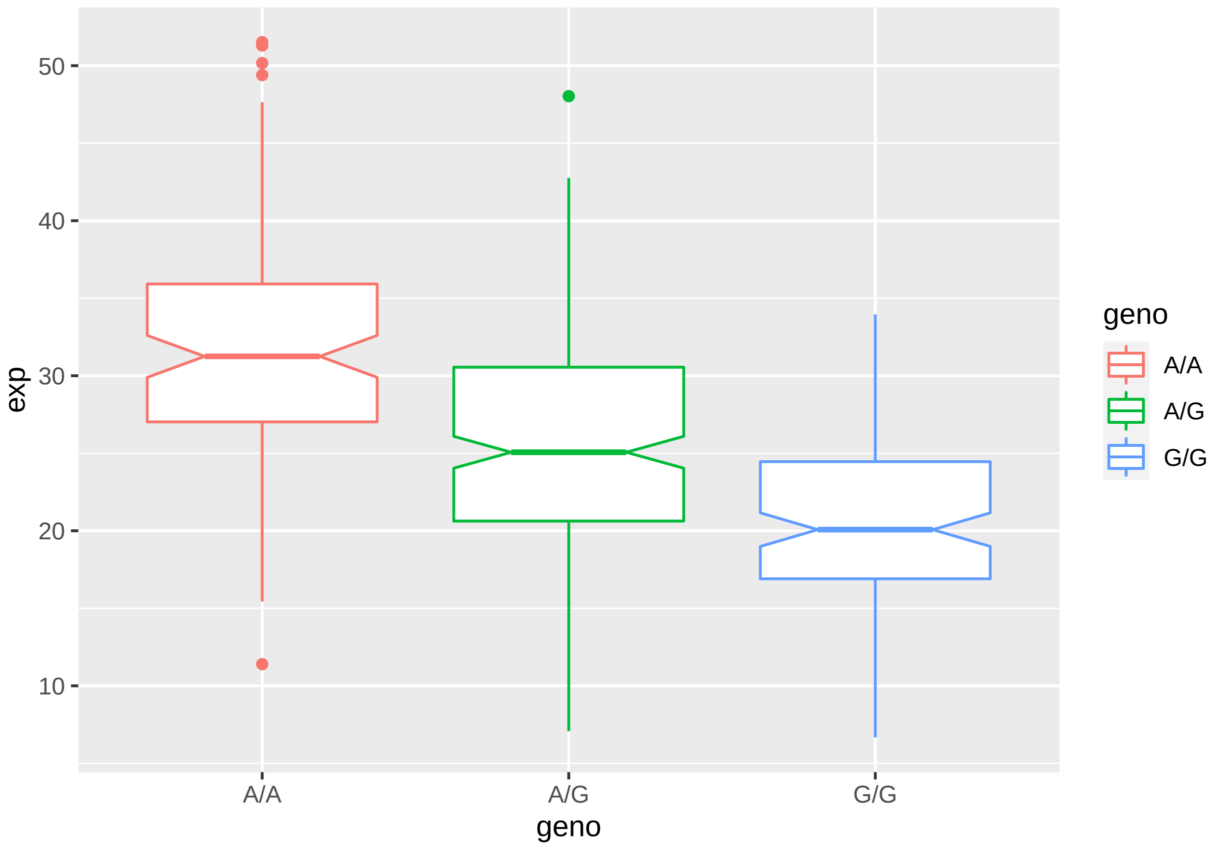
```
##     sample              geno                exp
##  Length:462         Length:462         Min.   : 6.675
##  Class :character   Class :character   1st Qu.:20.004
##  Mode  :character   Mode  :character   Median :25.116
##                                        Mean   :25.640
##                                        3rd Qu.:30.779
##                                        Max.   :51.518
```

> Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3? The A/A genotype has much higher expression than the G/G genotype as evidenced by the boxplot. Having an A/A genotype corresponds to high expression of ORMDL3 while having a G/G genotype corresponds to low expression of ORMDL3.

```
library(ggplot2)
```

Make a boxplot

```
ggplot(expr) + aes(geno, exp, col= geno) + geom_boxplot(notch=TRUE)
```


exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G