Class 17: Genome Informatics (Q13/Q14)

Wade Ingersoll (PID: 69038080)

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core
41432946;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel

Here we read this CSV file

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

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

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

```
table(mxl$Genotype..forward.strand.) / nrow(mxl) * 100
```

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

Now let's look at a different population. I picked the GBR.

```
gbr <- read.csv("GBR_373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

Find proportion of G|G

```
round( table(gbr$Genotype..forward.strand.) / nrow(gbr) * 100, 2 )
```

```
  A|A   A|G   G|A   G|G
25.27 18.68 26.37 29.67
```

This variant that is associated with childhood asthma is more frequent in the GBR population than the MKL population.

Let's now dig into this further.

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

How many samples do we have?

```
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
```

```
nrow(expr)
```

```
[1] 462
```

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.

Answer: (See proceeding code for calculations):

```
> Sample size for A/A = 108 (median expression: 31.24847)
> Sample size for A/G = 233 (median expression: 25.06486)
> Sample size for G/G = 121 (median expression: 20.07363)
```

```
table(expr$geno)
```

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

```
# Calculate median expression for each genotype
median(expr$exp[expr$geno == "A/A"])
```

```
[1] 31.24847
```

```
median(expr$exp[expr$geno == "A/G"])
```

```
[1] 25.06486
```

```
median(expr$exp[expr$geno == "G/G"])
```

```
[1] 20.07363
```

```
library(ggplot2)
```

Let's make a boxplot

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?

Answer: Based on the boxplot shown below, it appears the SNP *does* affect ORMDL3 expression. I would infer that Gs negatively impact ORMDL3 expression since A/A shows the highest median expression whereas A/G and G/G genotypes show lower median expression.

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