Class.10.md

Mirte Ciz Marieke Kuijpers

18/02/2022

Asthma associated SNP genotypes from the 1000 Genomes Project

We have downloaded data about one of the Asthma associated SNPs from Verlaan et. al. 2009 in the MXL population of the 1000 Genomes project.

```
# Read in csv file
mxl <- read.csv("373531-SampleGenotypes-
Homo_sapiens_Variation_Sample_rs8067378.csv", header = T)
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      -
```

We can now determine the frequency of different alleles in the MXL population.

```
# Make binary presence absence table
mxl.t <- table(mxl)

# Calculate frequencies
mxl.f <- (colSums(mxl.t)/nrow(mxl))*100

mxl.f
```

```
## , , Father = -, Mother = -
##
##                           Population.s.
## Genotype..forward.strand. ALL, AMR, MXL
##                       A|A       34.3750
##                       A|G       32.8125
##                       G|A       18.7500
##                       G|G       14.0625
```

OR

```
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 compare for a different population.

```
# Load file
gbr <- read.csv("373522-SampleGenotypes-
Homo_sapiens_Variation_Sample_rs8067378.csv", header = T)

# Check file
head(gbr)
```

```
##   Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s.
Father
## 1                  HG00096 (M)                       A|A ALL, EUR, GBR
-
## 2                  HG00097 (F)                       G|A ALL, EUR, GBR
-
## 3                  HG00099 (F)                       G|G ALL, EUR, GBR
-
## 4                  HG00100 (F)                       A|A ALL, EUR, GBR
-
## 5                  HG00101 (M)                       A|A ALL, EUR, GBR
-
## 6                  HG00102 (F)                       A|A ALL, EUR, GBR
-
##   Mother
## 1      -
## 2      -
## 3      -
## 4      -
## 5      -
## 6      -
```

```
# Get frequencies
table(gbr$Genotype..forward.strand.)/nrow(gbr)*100
```

```
##
##      A|A      A|G      G|A      G|G
## 25.27473 18.68132 26.37363 29.67033
```

This shows that these two populations have very different genotype frequncies for this Asthma implicated SNP. (N.B. the implication is just a correlation, and correlation does not = conservation, but these results are intersting.)

Section 2

```
# Set up
library("ShortRead")

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: BiocParallel

## Loading required package: Biostrings

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit

## Loading required package: Rsamtools

## Loading required package: GenomicRanges

## Loading required package: GenomicAlignments

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
```

```
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians

# Read in Fastq files
HG1 <- readFastq("HG00109_1.fastq")
HG2 <- readFastq("HG00109_2.fastq")
```

Homework section

We have data for the gene expression of the gene associated with the rs8067378 SNP linked to Asthma for ~230 samples.

```
# Read in data to an r object
dat <- read.table("rs8067378_ENSG00000172057.6.txt", row.names = 1, header =
TRUE)

# Inspect data
str(dat)

## 'data.frame':    462 obs. of  3 variables:
##  $ sample: chr  "HG00367" "NA20768" "HG00361" "HG00135" ...
##  $ geno  : chr  "A/G" "A/G" "A/A" "A/A" ...
##  $ exp   : num  29 20.2 31.3 34.1 18.3 ...

summary(dat)

##     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
```

To determine the frequencies of each genotype we can use the following code:

```
# For now, ignore gene expression
gen <- dat[,-3]

# Make binary presence absence table
```

```
geno.t <- table(gen)

head(geno.t)

##          geno
## sample    A/A A/G G/G
##   HG00096   1   0   0
##   HG00097   0   1   0
##   HG00099   0   0   1
##   HG00100   1   0   0
##   HG00101   1   0   0
##   HG00102   1   0   0

# Sum across the columns to get frequencies
geno.f <- (colSums(geno.t)/nrow(dat))*100

geno.f

##      A/A      A/G      G/G
## 23.37662 50.43290 26.19048
```

To find the median expression for each of these genotypes we can group the dataset by genotype using the `dplyr` package.

```
# Load the dplyr package
library("dplyr")

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:ShortRead':
##
##     id

## The following objects are masked from 'package:GenomicAlignments':
##
##     first, last

## The following object is masked from 'package:Biobase':
##
##     combine

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following object is masked from 'package:XVector':
##
##     slice

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Group the data set by genotype and find the median

dat %>%
  filter(geno == "A/G") %>%
  summary()
```

```
##     sample              geno                exp
##  Length:233         Length:233         Min.   : 7.075
##  Class :character   Class :character   1st Qu.:20.626
##  Mode  :character   Mode  :character   Median :25.065
##                                        Mean   :25.397
##                                        3rd Qu.:30.552
##                                        Max.   :48.034
```

```
dat %>%
  filter(geno == "G/G") %>%
  summary()
```

```
##     sample              geno                exp
##  Length:121         Length:121         Min.   : 6.675
```

```
##  Class :character   Class :character   1st Qu.:16.903
##  Mode  :character   Mode  :character   Median :20.074
##                                        Mean   :20.594
##                                        3rd Qu.:24.457
##                                        Max.   :33.956
```

```
dat %>%
  filter(geno == "A/A") %>%
  summary()
```

```
##     sample              geno                exp
##  Length:108         Length:108         Min.   :11.40
##  Class :character   Class :character   1st Qu.:27.02
##  Mode  :character   Mode  :character   Median :31.25
##                                        Mean   :31.82
##                                        3rd Qu.:35.92
##                                        Max.   :51.52
```

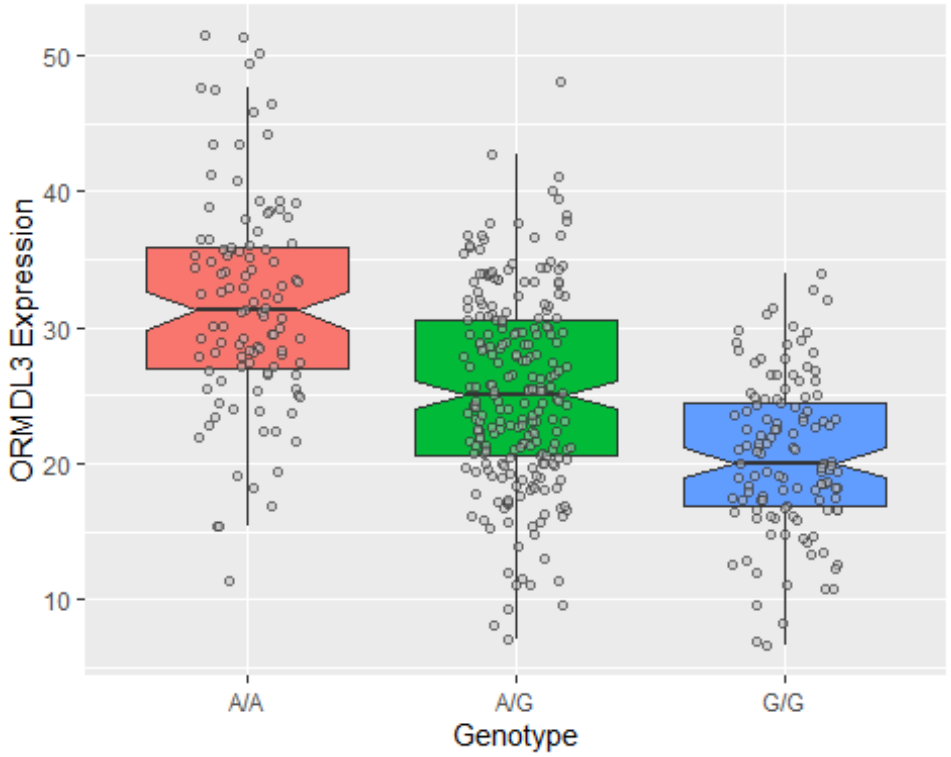
So the expression median for the "A/G" = 25.065, "G/G" = 20.074 and "A/A" = 31.25.

We can also plot this data.

```
# Set up
library("ggplot2")

# Plot

ggplot(dat, aes(x = geno, y = exp, fill = geno)) +
  geom_boxplot(notch = TRUE, outlier.shape = NA, show.legend = FALSE) +
  geom_jitter(alpha = 0.5, shape=21, position=position_jitter(0.2), fill =
"grey") +
  labs(x = "Genotype", y = "ORMDL3 Expression")
```



This plot suggests that the A SNP variant causes greater expression of ORMDL3. Furthermore, this effect appears to be additive rather tna dominant, with ORMDL3 expression for A/A > A/G > G/G.