Wk6: Genome Informatics

Yiyu

2025-02-14

Section 1. Proportion of G/G population

Downloaded a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39895035-39895162;v=rs8067378;vdb=variation;vf=959672880#373531__tablePanel

Here we read the CSV file

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

```
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
```

#Section 4: Population Scale Analysis [HOMEWORK] 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
```

```r
nrow(expr)
```

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

> Q13. Determine sample size for each genotype and their corresponding median expression levels for each of these genotypes.

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

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

Sample size for A/A is 108, A/G is 233, G/G is 121.

```r
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'

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

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

```r
median <- expr |>
  group_by(geno) |>
  summarize(median_expression = median(exp))
print(median)
```

```
## # A tibble: 3 x 2
##   geno  median_expression
##   <chr>             <dbl>
## 1 A/A                31.2
## 2 A/G                25.1
## 3 G/G                20.1
```

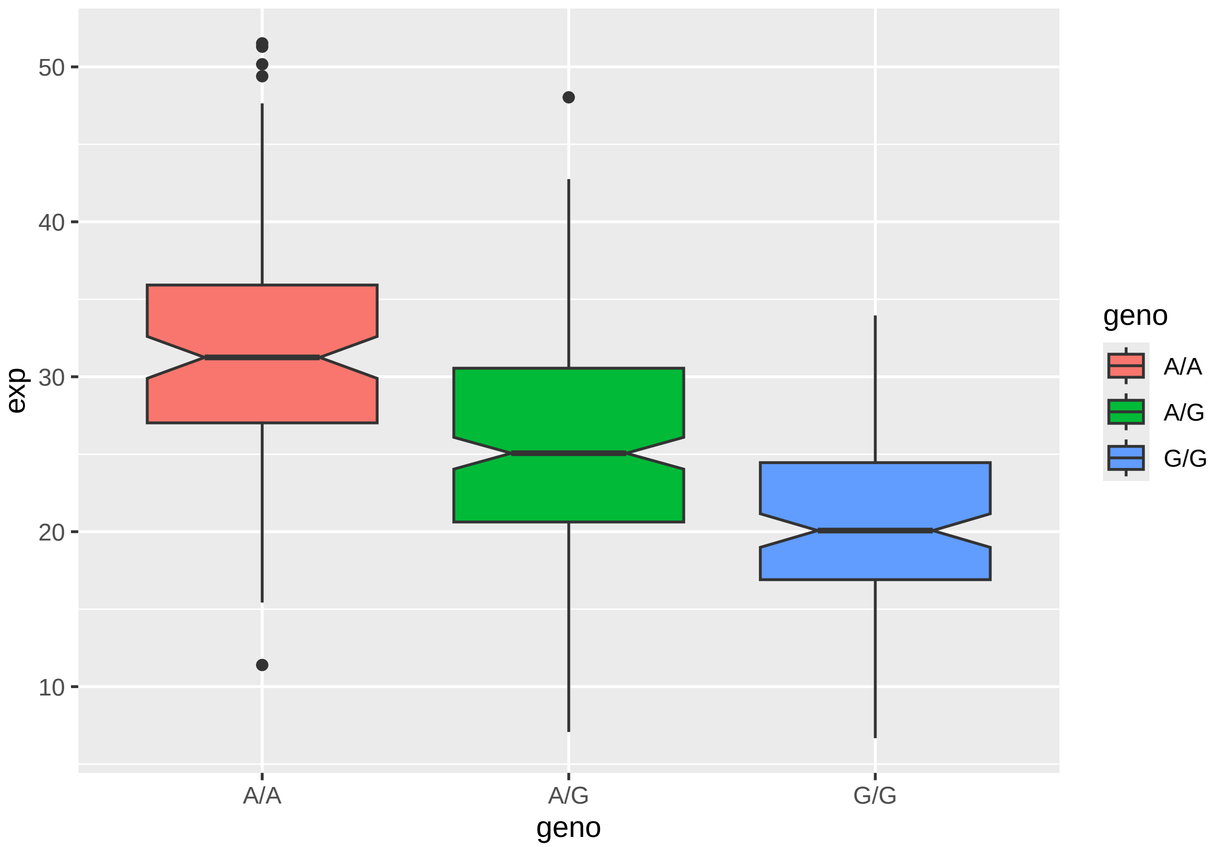
Can use **dplyr** package to print median expression for each genotype using the `group_by` function.

```r
library(ggplot2)
```

> Q14. Generate a boxplot wiht 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 affect the expression of ORMDL3?

Let's make a boxplot with our data

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



Yes, the box plot indicates that having G/G correlates to reduced expression of ORMDL3 gene, while A/A correlates to higher expression.