Class 10: Genome Informatics

Taylor F. (A59010460)

2/18/2022

Counting Entries in a CSV

```
#We first need to read in the .csv file
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")

#We can then look at the SNP genotype in each of the observations
#mxl$Genotype..forward.strand.

#Then we have to get a summary of each genotype as a percentage by dividing total instances of each var
table(mxl$Genotype..forward.strand.) / nrow(mxl)
```

```
##
##      A|A      A|G      G|A      G|G
## 0.343750 0.328125 0.187500 0.140625
```

```
#Let's compare the MXL values to the GBR dataset
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")

table(gbr$Genotype..forward.strand.) / nrow(gbr)
```

```
##
##       A|A       A|G       G|A       G|G
## 0.2527473 0.1868132 0.2637363 0.2967033
```

RNA-Seq Genotyping Results: What Does it all Mean?

```
#We need to read in the appropriate .csv file
x <- read.table("rs8067378_ENSG00000172057.6.txt")
head(x)
```

```
##    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
```

How many different genotypes do we have?

```
table(x$geno)
```

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

```
#Search through 'x' for the rows that contain G/G
x[x$geno == "G/G",]
```

```
##      sample geno      exp
## 5   NA18870  G/G 18.25141
## 9   HG00327  G/G 17.67473
## 17  NA12546  G/G 18.55622
## 20  NA18488  G/G 23.10383
## 23  NA19214  G/G 30.94554
## 28  HG00112  G/G 21.14387
## 29  NA20518  G/G 18.39547
## 31  NA19119  G/G 12.02809
## 32  HG00247  G/G 17.44761
## 35  NA20758  G/G 29.82254
## 41  NA12249  G/G 23.01983
## 46  HG00320  G/G 13.42470
## 47  NA11843  G/G 22.65437
## 49  NA20588  G/G 11.07445
## 50  NA20510  G/G 28.35841
## 56  HG00118  G/G 28.79371
## 57  NA18520  G/G 27.08956
## 61  NA12234  G/G 16.11138
## 72  NA19152  G/G 26.61928
## 73  NA20761  G/G 30.18323
## 77  NA18923  G/G 19.40790
## 79  HG00238  G/G 19.52301
## 85  NA12058  G/G 26.56808
## 89  HG00129  G/G 17.34076
## 92  HG00183  G/G 10.74263
## 93  HG00109  G/G 16.66051
## 104 NA18517  G/G 29.01720
## 105 NA20801  G/G 20.69333
## 106 NA20529  G/G 21.15677
## 109 HG00349  G/G 18.58691
## 110 HG00234  G/G 19.04962
## 111 NA19248  G/G 22.81974
## 114 NA12813  G/G 32.01142
## 115 NA20537  G/G 21.12823
## 117 HG00332  G/G 18.61268
## 118 HG00152  G/G 19.37093
## 119 NA20783  G/G 31.42162
## 128 HG00185  G/G 16.67764
## 132 NA20531  G/G 19.08659
## 135 HG00277  G/G 21.55001
## 140 HG00336  G/G  8.29591
## 143 NA20581  G/G 12.58869
## 150 NA20538  G/G 17.34109
## 153 NA20814  G/G 28.23642
## 156 NA19171  G/G 19.99979
## 159 HG00141  G/G 25.55413
## 163 NA19190  G/G 24.45672
## 166 NA10851  G/G 23.53572
## 170 HG00116  G/G 22.48273
## 171 NA12272  G/G 14.66862
## 172 NA19096  G/G 33.95602
## 175 NA19236  G/G 18.26466
## 178 HG00345  G/G 16.06661
## 190 HG00156  G/G 17.32504
## 193 HG00282  G/G 19.14766
## 194 HG00343  G/G 12.57599
## 195 HG00139  G/G 22.28749
## 199 HG00232  G/G 17.29261
## 201 HG00122  G/G 24.18141
## 207 NA19149  G/G 16.07627
## 211 HG00189  G/G 14.80495
## 218 HG00126  G/G 23.46573
## 224 HG00265  G/G 28.97074
## 225 HG00378  G/G 27.78837
## 232 NA20796  G/G 23.92355
## 233 NA12399  G/G  9.55902
## 239 HG00099  G/G 12.35836
## 241 NA19114  G/G 22.53910
## 247 NA19210  G/G 21.98118
## 250 HG00276  G/G 16.40569
## 253 HG00181  G/G 25.21931
## 254 HG00346  G/G 24.32857
## 259 HG00142  G/G 19.42882
## 261 HG00315  G/G 26.56993
## 267 HG00250  G/G 13.34557
## 268 NA20769  G/G 16.60507
## 271 NA19144  G/G 24.85165
## 272 NA12815  G/G 21.56943
## 280 NA19175  G/G 23.95528
## 283 NA18519  G/G 16.18962
## 285 NA20535  G/G 22.53720
## 287 HG00260  G/G 26.04123
## 288 HG00372  G/G  6.67482
## 292 HG00261  G/G 20.07363
## 293 HG00273  G/G 19.76527
## 299 HG00358  G/G 18.50772
## 307 NA19121  G/G 20.14146
## 308 NA20515  G/G 18.07151
## 314 NA10847  G/G  6.94390
## 316 NA12400  G/G 22.14277
## 319 HG00342  G/G 14.23742
## 330 HG00136  G/G 19.85388
## 340 NA20765  G/G 27.73467
## 344 NA18502  G/G 19.02064
## 351 NA20772  G/G 14.49816
## 355 HG00257  G/G 26.78940
## 356 NA18486  G/G 20.84709
## 357 HG00188  G/G 10.77316
## 361 HG00280  G/G 12.82128
## 362 HG00308  G/G 16.90256
## 364 NA18910  G/G 29.60045
## 369 HG00281  G/G 14.81945
## 373 NA12275  G/G 17.46326
## 375 HG00351  G/G 23.26922
## 376 HG00186  G/G 21.39806
## 378 HG00275  G/G 18.06320
## 379 HG00325  G/G 15.91528
## 380 NA19118  G/G 24.80823
## 381 HG00124  G/G 26.04514
## 383 HG02215  G/G 18.28089
## 385 HG00134  G/G 23.24907
## 391 NA11931  G/G 17.91118
## 393 HG00120  G/G 21.09502
## 421 NA20582  G/G 24.74366
## 428 NA12889  G/G 27.40521
## 435 NA12006  G/G 24.85772
## 436 NA19108  G/G 23.08482
## 446 NA07346  G/G 16.56929
## 454 HG00154  G/G 16.69044
## 457 HG00233  G/G 25.08880
## 458 HG00131  G/G 32.78519
```

```
#To get the expression values for the G/G genotypes
x[x$geno == "G/G","exp"]
```

```
##   [1] 18.25141 17.67473 18.55622 23.10383 30.94554 21.14387 18.39547 12.02809
##   [9] 17.44761 29.82254 23.01983 13.42470 22.65437 11.07445 28.35841 28.79371
##  [17] 27.08956 16.11138 26.61928 30.18323 19.40790 19.52301 26.56808 17.34076
##  [25] 10.74263 16.66051 29.01720 20.69333 21.15677 18.58691 19.04962 22.81974
##  [33] 32.01142 21.12823 18.61268 19.37093 31.42162 16.67764 19.08659 21.55001
##  [41]  8.29591 12.58869 17.34109 28.23642 19.99979 25.55413 24.45672 23.53572
##  [49] 22.48273 14.66862 33.95602 18.26466 16.06661 17.32504 19.14766 12.57599
##  [57] 22.28749 17.29261 24.18141 16.07627 14.80495 23.46573 28.97074 27.78837
##  [65] 23.92355  9.55902 12.35836 22.53910 21.98118 16.40569 25.21931 24.32857
##  [73] 19.42882 26.56993 13.34557 16.60507 24.85165 21.56943 23.95528 16.18962
##  [81] 22.53720 26.04123  6.67482 20.07363 19.76527 18.50772 20.14146 18.07151
##  [89]  6.94390 22.14277 14.23742 19.85388 27.73467 19.02064 14.49816 26.78940
##  [97] 20.84709 10.77316 12.82128 16.90256 29.60045 14.81945 17.46326 23.26922
## [105] 21.39806 18.06320 15.91528 24.80823 26.04514 18.28089 23.24907 17.91118
## [113] 21.09502 24.74366 27.40521 24.85772 23.08482 16.56929 16.69044 25.08880
## [121] 32.78519
```

```
summary(x[x$geno == "G/G","exp"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.675  16.903  20.074  20.594  24.457  33.956
```

```
#Summaries of expressions for every genotype
summary(x[x$geno == "G/G","exp"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.675  16.903  20.074  20.594  24.457  33.956
```

```
summary(x[x$geno == "A/A","exp"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.40   27.02   31.25   31.82   35.92   51.52
```

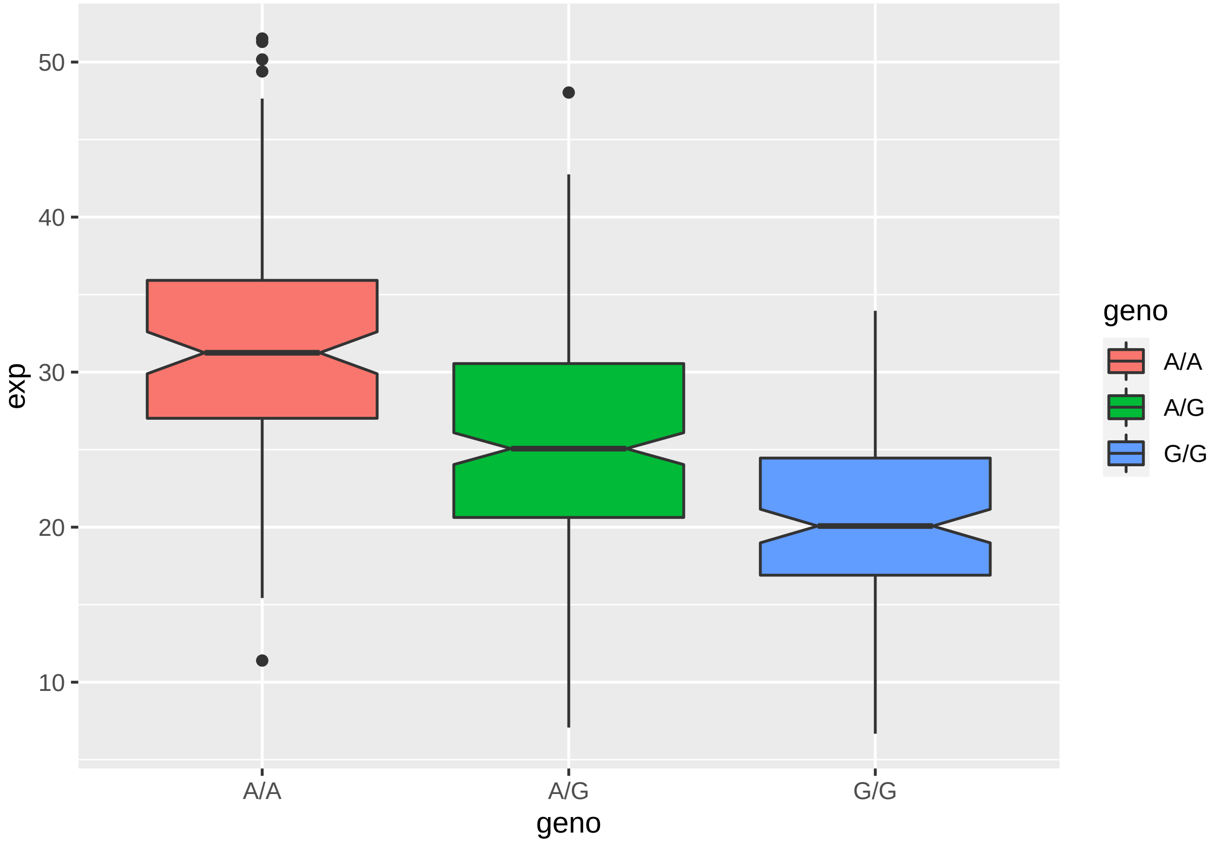
```
summary(x[x$geno == "A/G","exp"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.075  20.626  25.065  25.397  30.552  48.034
```

Now let's create a graphical summary of this information

```
library(ggplot2)

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


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