Class10

Morgan Farrell

2/18/2022

#Genotype data from 1000 Genomes Project

Download MXL dataset from https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39894595-39895595;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel

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

Find the frequency of G|G homozygous gentoypes in the dataset

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

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

Download the GBR dataset

```
gbr <- read.csv("GBR-SampleGenotypes.csv")
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      -
```

Same calculation of G|G with the GBR dataset

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

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

Inital RNA-seq Analysis

Question 13

Using read.csv first

```
tbl <- read.csv("rs8067378_ENSG00000172057.6.csv", header= TRUE, sep= "")
```

Second method using read.table

```
x <- read.table("rs8067378_ENSG00000172057.6.csv")
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 of each genoytpe

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

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

Find where all the G|G genotypes are

```
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
```

Summary statistics

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

```
##     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
```

Check the summary of the other genoytpes

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

```
##     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
```

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

```
##     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
```

Question 14

make a box plot

```
library(ggplot2)
```

```
## Warning in register(): Can't find generic 'scale_type' in package ggplot2 to
## register S3 method.
```

```
ggplot(x, aes(geno,exp, fill=geno))+
  geom_boxplot(notch=TRUE) +
  geom_jitter(width=0.2, size= 0.5)
```