Class_12_RNA_seq_021624

Johann Tailor

Reading the sample summary file into R:

```r
expression_file <- read.table("https://bioboot.github.io/bggn213_W24/class-material/rs8067

#Finding the summary of the file loaded:
summary(expression_file)
```

```
    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
```

Identifying how many patients are there in each genotype category:

```r
All_AG <- grep("A/G", expression_file$geno)
All_GG <- grep("G/G", expression_file$geno)
All_AA <- grep("A/A", expression_file$geno)

All_AG
```

```
  [1]   1   2   7  10  11  12  14  19  21  22  25  26  33  34  36  38  39  40
 [19]  42  43  44  45  51  52  54  58  59  60  63  64  65  66  68  69  70  71
 [37]  74  75  78  80  81  83  84  86  87  88  90  91  94  95  98 100 101 103
 [55] 107 108 113 116 120 123 124 125 126 129 130 131 134 136 137 138 139 141
 [73] 142 144 145 146 147 148 149 151 152 157 158 160 161 164 165 168 173 176
 [91] 181 183 184 185 186 189 191 192 196 197 200 203 204 206 208 209 210 212
```

```
[109] 213 216 219 220 221 222 223 231 234 236 237 238 244 249 251 252 255 256
[127] 258 260 262 273 274 276 277 279 281 282 284 286 289 291 294 295 296 297
[145] 300 301 302 303 309 310 311 312 313 315 317 318 320 321 323 324 325 326
[163] 327 328 329 331 333 334 335 336 338 339 342 345 347 349 350 352 353 358
[181] 363 365 366 367 372 377 387 388 392 394 395 396 397 398 399 402 403 404
[199] 406 407 410 411 413 414 415 416 417 418 419 422 423 425 427 429 430 431
[217] 432 434 437 438 439 440 441 444 445 447 449 450 452 453 455 460 462
```

```
#A/G_patients <- row.names(df[All_AG, ])
```

Another way:

Trying codes to figure it out how to get row.names from the grepED files.

```
matching_AG <- rownames(expression_file)[grep("A/G", expression_file$geno)]
matching_GG <- rownames(expression_file)[grep("G/G", expression_file$geno)]
matching_AA <- rownames(expression_file)[grep("A/A", expression_file$geno)]

matching_GG
```

```
  [1] "5"   "9"   "17"  "20"  "23"  "28"  "29"  "31"  "32"  "35"  "41"  "46" 
 [13] "47"  "49"  "50"  "56"  "57"  "61"  "72"  "73"  "77"  "79"  "85"  "89" 
 [25] "92"  "93"  "104" "105" "106" "109" "110" "111" "114" "115" "117" "118"
 [37] "119" "128" "132" "135" "140" "143" "150" "153" "156" "159" "163" "166"
 [49] "170" "171" "172" "175" "178" "190" "193" "194" "195" "199" "201" "207"
 [61] "211" "218" "224" "225" "232" "233" "239" "241" "247" "250" "253" "254"
 [73] "259" "261" "267" "268" "271" "272" "280" "283" "285" "287" "288" "292"
 [85] "293" "299" "307" "308" "314" "316" "319" "330" "340" "344" "351" "355"
 [97] "356" "357" "361" "362" "364" "369" "373" "375" "376" "378" "379" "380"
[109] "381" "383" "385" "391" "393" "421" "428" "435" "436" "446" "454" "457"
[121] "458"
```

```
matching_AA
```

```
  [1] "3"   "4"   "6"   "8"   "13"  "15"  "16"  "18"  "24"  "27"  "30"  "37" 
 [13] "48"  "53"  "55"  "62"  "67"  "76"  "82"  "96"  "97"  "99"  "102" "112"
 [25] "121" "122" "127" "133" "154" "155" "162" "167" "169" "174" "177" "179"
 [37] "180" "182" "187" "188" "198" "202" "205" "214" "215" "217" "226" "227"
 [49] "228" "229" "230" "235" "240" "242" "243" "245" "246" "248" "257" "263"
 [61] "264" "265" "266" "269" "270" "275" "278" "290" "298" "304" "305" "306"
```

```
[73] "322" "332" "337" "341" "343" "346" "348" "354" "359" "360" "368" "370"
[85] "371" "374" "382" "384" "386" "389" "390" "400" "401" "405" "408" "409"
[97] "412" "420" "424" "426" "433" "442" "443" "448" "451" "456" "459" "461"
```

```
matching_AG
```

```
  [1] "1"   "2"   "7"   "10"  "11"  "12"  "14"  "19"  "21"  "22"  "25"  "26" 
 [13] "33"  "34"  "36"  "38"  "39"  "40"  "42"  "43"  "44"  "45"  "51"  "52" 
 [25] "54"  "58"  "59"  "60"  "63"  "64"  "65"  "66"  "68"  "69"  "70"  "71" 
 [37] "74"  "75"  "78"  "80"  "81"  "83"  "84"  "86"  "87"  "88"  "90"  "91" 
 [49] "94"  "95"  "98"  "100" "101" "103" "107" "108" "113" "116" "120" "123"
 [61] "124" "125" "126" "129" "130" "131" "134" "136" "137" "138" "139" "141"
 [73] "142" "144" "145" "146" "147" "148" "149" "151" "152" "157" "158" "160"
 [85] "161" "164" "165" "168" "173" "176" "181" "183" "184" "185" "186" "189"
 [97] "191" "192" "196" "197" "200" "203" "204" "206" "208" "209" "210" "212"
[109] "213" "216" "219" "220" "221" "222" "223" "231" "234" "236" "237" "238"
[121] "244" "249" "251" "252" "255" "256" "258" "260" "262" "273" "274" "276"
[133] "277" "279" "281" "282" "284" "286" "289" "291" "294" "295" "296" "297"
[145] "300" "301" "302" "303" "309" "310" "311" "312" "313" "315" "317" "318"
[157] "320" "321" "323" "324" "325" "326" "327" "328" "329" "331" "333" "334"
[169] "335" "336" "338" "339" "342" "345" "347" "349" "350" "352" "353" "358"
[181] "363" "365" "366" "367" "372" "377" "387" "388" "392" "394" "395" "396"
[193] "397" "398" "399" "402" "403" "404" "406" "407" "410" "411" "413" "414"
[205] "415" "416" "417" "418" "419" "422" "423" "425" "427" "429" "430" "431"
[217] "432" "434" "437" "438" "439" "440" "441" "444" "445" "447" "449" "450"
[229] "452" "453" "455" "460" "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.

Here are the median for each genotype:

```
summary(expression_file[expression_file[,2] == "A/A",3] )
```

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

```
summary(expression_file[expression_file[,2] == "G/G",3] )
```

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

```
summary(expression_file[expression_file[,2] == "A/G",3] )
```

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

ANSWER:

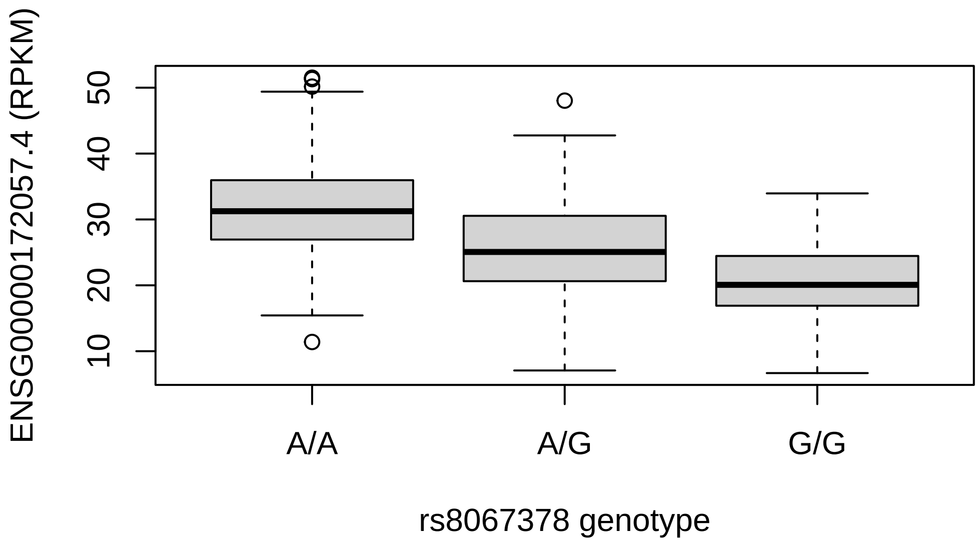
Median for each genotype is as follows

A/A : 31.25 G/G : 20.074 A/G : 25.065

> 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?

Let's generate a box plot to compare the means of the three genotypes and their expression of ORMDL3 gene:

```
box_plot_genotypes <- boxplot(exp~geno, data=expression_file, xlab="rs8067378 genotype", y
```



```
box_plot_genotypes
```

```
$stats
         [,1]     [,2]     [,3]
[1,] 15.42908  7.07505  6.67482
[2,] 26.95022 20.62572 16.90256
[3,] 31.24847 25.06486 20.07363
[4,] 35.95503 30.55183 24.45672
[5,] 49.39612 42.75662 33.95602

$n
[1] 108 233 121

$conf
         [,1]     [,2]     [,3]
[1,] 29.87942 24.03742 18.98858
[2,] 32.61753 26.09230 21.15868

$out
[1] 51.51787 50.16704 51.30170 11.39643 48.03410

$group
[1] 1 1 1 1 2

$names
[1] "A/A" "A/G" "G/G"
```

ANSWER:

I do think the SNP G|G renders overall less gene expression of ORMDL3 compared to A|A; the mean of A|A is 31.82 and G|G is 20.5, respectively.