Assignment-2

Γιώργος Γεραμούτσος, csd3927

Libraries

```
library(matrixStats)
library(limma)
library(gplots)
```

Μέρος Α - Ανάλυση Μικροστοιχείων

Επεξεργασία Δεδομένων

```
gds_ds = readLines("GDS3709.soft")
gds_cl = gds_ds[!grepl("^[!^#]", gds_ds)]
writeLines(gds_cl,"GDS3709.soft")

myDt = read.table("GDS3709.soft", sep="\t", header=T, na.strings="null")
head(myDt, n=6)
```

```
##      ID_REF IDENTIFIER GSM447401 GSM447411 GSM447413 GSM447415 GSM447416
## 1 1007_s_at    MIR4640  1124.000  1196.000   982.800  1075.000  1114.000
## 2   1053_at       RFC2   203.300   181.500   229.600   160.400   209.500
## 3    117_at      HSPA6    53.400    55.600    49.040    39.100    51.940
## 4    121_at       PAX8   245.600   209.400   252.900   223.300   186.500
## 5 1255_g_at     GUCA1A     8.228     8.131     8.994     7.576     8.383
## 6   1294_at    MIR5193   157.300   149.800   131.300   127.700   154.100
```

Δημιουργια των factors

```
gender = factor(rep(c('f', 'm'), each = 40))
gender = gender[2:80]
smoking = factor(rep(rep(c('s', 'ns'), each = 20), 2))
smoking = smoking[2:80]

expr = myDt[,3:ncol(myDt)]
head(expr)
```

```
##   GSM447401 GSM447411 GSM447413 GSM447415 GSM447416 GSM447425 GSM447430
## 1  1124.000  1196.000   982.800  1075.000  1114.000  1302.000  1279.000
## 2   203.300   181.500   229.600   160.400   209.500   191.900   180.600
## 3    53.400    55.600    49.040    39.100    51.940    69.800    52.440
## 4   245.600   209.400   252.900   223.300   186.500   155.100   277.300
## 5     8.228     8.131     8.994     7.576     8.383     8.768     8.679
## 6   157.300   149.800   131.300   127.700   154.100   158.200   147.200
##   GSM447435 GSM447440 GSM447444 GSM447448 GSM447449 GSM447450 GSM447452
## 1  1210.000  1199.000  1172.000  1245.000  1307.000  1263.000  1282.000
## 2   204.300   173.500   249.800   213.300   191.800   202.700   187.000
## 3    51.250    47.070   109.600    48.100    43.840    49.290    59.640
## 4   189.200   260.200   225.800   211.000   220.600   212.400   178.600
## 5     7.806     7.795     8.418     7.862     7.667     9.543     9.772
## 6   161.200   142.800   138.600   109.600   154.400   174.900   169.500
## and more ...
```

Ερώτηση 1

```
design_1 = model.matrix(~ 0 + gender*smoking)
colnames(design_1) = c("female", "male", "smoking", "male_smoking")

fit_1 = lmFit(expr, design_1, intercept=T)
#fit_1
```

- **1.1 - επίδραση φύλου**

```
contrasts_11 = makeContrasts(female-male, levels=design_1)

fit_11 = contrasts.fit(fit_1,contrasts_11)
#fit_11

fit_11 = eBayes(fit_11)
#fit_11

tb_11 = topTable(fit_11, n=Inf)
head(tb_11, n=10)
```

```
##            logFC    AveExpr         t      P.Value    adj.P.Val         B
## 16147  -170.8905  125.27684 -9.077815 9.886918e-14 5.405672e-09 10.553016
## 23518   305.3888  191.36570  8.578709 8.887561e-13 1.323790e-08  9.518388
## 11358 -2667.0730 1903.62709 -8.560697 9.621229e-13 1.323790e-08  9.480457
## 31009   658.1400  425.56608  8.559202 9.684790e-13 1.323790e-08  9.477306
## 37747  -117.4413   85.39362 -8.007311 1.099457e-11 1.050127e-07  8.296610
## 33848  2278.0030 1443.55747  7.996614 1.152403e-11 1.050127e-07  8.273394
## 16510   -79.8790   70.56278 -7.687915 4.469852e-11 3.491274e-07  7.598532
## 13858  -111.1106   86.18557 -7.651029 5.254212e-11 3.590925e-07  7.517296
```

```
## 20513    -9.1133   13.74244 -7.588891 6.897906e-11 4.190478e-07  7.380176
## 13857  -164.5102  122.85527 -7.562567 7.740435e-11 4.232083e-07  7.321988
```

- **1.2 - επίδραση καπνίσματος**

```
contrasts_12 = makeContrasts(smoking, levels=design_1)

fit_12 = contrasts.fit(fit_1, contrasts_12)
#fit_12

fit_12 = eBayes(fit_12)
#fit_12

tb_12 = topTable(fit_12, n=Inf)
head(tb_12, n=10)
```

```
##           logFC   AveExpr        t      P.Value    adj.P.Val         B
## 36708 104.25595 138.52975 6.928823 1.220454e-09 0.0000667283 10.614618
## 11885 273.55253 211.12380 6.250361 2.217693e-08 0.0005223870  8.192117
## 29122  28.44658  49.10924 6.189474 2.866321e-08 0.0005223870  7.976944
## 38609  39.29587  87.34291 5.822495 1.324264e-07 0.0018101036  6.691074
## 16573 343.39605 582.22532 5.596728 3.343663e-07 0.0028416422  5.911200
## 13980 359.18500 607.39241 5.578536 3.600738e-07 0.0028416422  5.848787
## 11884  50.82387  45.42494 5.573619 3.673498e-07 0.0028416422  5.831929
## 15541 240.62474 429.07848 5.543141 4.157867e-07 0.0028416422  5.727550
## 11886  93.26905  65.19885 5.428380 6.613444e-07 0.0037090011  5.336301
## 37349  34.71913  86.03557 5.422068 6.783724e-07 0.0037090011  5.314865
```

Ερώτηση 2

```
design_2 = model.matrix(~ 0 + gender+smoking)

colnames(design_2) = c("female", "male", "smoking")

fit_2 = lmFit(expr, design_2, intercept=T)
#fit_2
```

- **2.1**

```
contrasts_21 = makeContrasts(female-male, levels=design_2)
contrasts_21
```

```
##          Contrasts
## Levels    female - male
##   female              1
##   male               -1
##   smoking             0
```

```
fit_21 = contrasts.fit(fit_2, contrasts_21)
#fit_21

fit_21 = eBayes(fit_21)
#fit_21

tb_21 = topTable(fit_21, n=Inf)
head(tb_21, n=10)
```

```
##              logFC    AveExpr         t      P.Value    adj.P.Val        B
## 16147  -168.812110  125.27684 -12.68127 1.670634e-20 5.349121e-16 29.68518
## 11358 -2790.748987 1903.62709 -12.64290 1.956697e-20 5.349121e-16 29.57238
## 31009   672.599117  425.56608  12.36609 6.150566e-20 1.120941e-15 28.75114
## 13857  -189.458827  122.85527 -12.10697 1.811345e-19 2.475882e-15 27.97046
## 23518   298.884077  191.36570  11.86982 4.900215e-19 4.637628e-15 27.24590
## 13858  -122.842575   86.18557 -11.86083 5.089304e-19 4.637628e-15 27.21823
## 20513    -9.937403   13.74244 -11.63110 1.342739e-18 1.048775e-14 26.50697
## 16510   -82.364623   70.56278 -11.20342 8.293826e-18 5.143349e-14 25.15963
## 37747  -116.008617   85.39362 -11.18581 8.943253e-18 5.143349e-14 25.10350
## 33848  2250.830727 1443.55747  11.17399 9.407131e-18 5.143349e-14 25.06584
```

- **2.2**

```
contrasts_22 = makeContrasts("smoking", levels=design_2)
contrasts_22
```

```
##          Contrasts
## Levels    smoking
##   female        0
##   male          0
##   smoking       1
```

```
fit_22 = contrasts.fit(fit_2, contrasts_22)
#fit_22

fit_22 = eBayes(fit_22)
#fit_22

tb_22 = topTable(fit_22, n=Inf)
head(tb_22, n=10)
```

```
##            logFC    AveExpr        t      P.Value    adj.P.Val         B
## 36708   91.20237  138.52975 8.600944 7.318853e-13 4.001583e-08 17.869145
## 38609   36.71535   87.34291 7.780183 2.769455e-11 7.530571e-07 14.665652
## 29122   24.85955   49.10924 7.689495 4.132001e-11 7.530571e-07 14.311334
## 11885  222.41465  211.12380 7.155318 4.312740e-10 5.658587e-06 12.229046
## 37349   31.87156   86.03557 7.113581 5.174748e-10 5.658587e-06 12.066920
## 11886   80.83565   65.19885 6.695668 3.173051e-09 2.891442e-05 10.451036
## 15197   42.40659   49.40544 6.413110 1.066616e-08 8.331031e-05  9.368711
## 15071 1616.27922 3660.62025 6.273696 1.930472e-08 1.272879e-04  8.838606
## 15311  264.98513  894.35316 6.254379 2.095274e-08 1.272879e-04  8.765388
## 29713   98.50650  165.20063 6.069341 4.575549e-08 2.499695e-04  8.067091
```

Ερώτηση 3

Δοκιμή με διαφορετικό τρόπο

```
library(GEOquery)
```

Φορτώνουμε το dataset και κρατάμε στο phen τα χαρακτηριστικά των δειγμάτων από το eset

```
gds = getGEO("GDS3709", GSEMatrix=TRUE)
eset = GDS2eSet(gds, do.log2=TRUE)

phen = pData(eset)
phen
```

```
##              sample gender           agent
## GSM447401 GSM447401 female cigarette smoke
## GSM447411 GSM447411 female cigarette smoke
## GSM447413 GSM447413 female cigarette smoke
## GSM447415 GSM447415 female cigarette smoke
## GSM447416 GSM447416 female cigarette smoke
## GSM447425 GSM447425 female cigarette smoke
## ....
## GSM447400 GSM447400 female         control
## GSM447402 GSM447402 female         control
## GSM447403 GSM447403 female         control
## GSM447405 GSM447405 female         control
## GSM447418 GSM447418 female         control
## GSM447422 GSM447422 female         control
## ....
## GSM447404 GSM447404   male cigarette smoke
## GSM447406 GSM447406   male cigarette smoke
## GSM447407 GSM447407   male cigarette smoke
## GSM447409 GSM447409   male cigarette smoke
## GSM447412 GSM447412   male cigarette smoke
## GSM447426 GSM447426   male cigarette smoke
## ....
## GSM447398 GSM447398   male         control
```

```
## GSM447399 GSM447399   male         control
## GSM447408 GSM447408   male         control
## GSM447410 GSM447410   male         control
## GSM447414 GSM447414   male         control
## GSM447417 GSM447417   male         control
## .....
##                                                                                 description
## GSM447401             Value for GSM447401: Smoker female study #107; src: Smoker female buccal mucosa
## GSM447411              Value for GSM447411: Smoker female study #20; src: Smoker female buccal mucosa
## .....
```

Δημιουργία των factors και αποθήκευση στις λίστες gender & smoking τα στοιχεία εκείνα του dataset που κάνουνε match τα gender & agent

```
gender = factor(phen$gender, levels = c("female","male"))
gender
```

```
##  [1] female female female female female female female female female female
## [11] female female female female female female female female female female
## [21] female female female female female female female female female female
## [31] female female female female female female female female female male
## [41] male   male   male   male   male   male   male   male   male   male
## [51] male   male   male   male   male   male   male   male   male   male
## [61] male   male   male   male   male   male   male   male   male   male
## [71] male   male   male   male   male   male   male   male   male
## Levels: female male
```

```
smoking = factor(phen$agent, levels = c("cigarette smoke", "control"))
smoking
```

```
##  [1] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
##  [5] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
##  [9] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
## [13] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
## [17] cigarette smoke cigarette smoke cigarette smoke control
## [21] control         control         control         control
## [25] control         control         control         control
## [29] control         control         control         control
## [33] control         control         control         control
## [37] control         control         control         cigarette smoke
## [41] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
## [45] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
## [49] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
## [53] cigarette smoke cigarette smoke cigarette smoke cigarette smoke
## [57] cigarette smoke cigarette smoke cigarette smoke control
## [61] control         control         control         control
## [65] control         control         control         control
## [69] control         control         control         control
## [73] control         control         control         control
## [77] control         control         control
## Levels: cigarette smoke control
```

- **3.1**

Κρατάμε στις δύο λίστες τα indexes των καπνιστών και μη-καπνιστών και εφαρμόζουμε το t-test. Από αυτά κρατάμε μόνο όσα p-values είναι < 0.05 και τυπώνονται ενδεικτικά μερικά.

```
smoker_ind = which(phen$agent == "cigarette smoke")
non_smoker_in = which(phen$agent == "control")

ttest_smoking = apply(expr, 1, function(x) t.test(x[smoker_ind], x[non_smoker_in])$p.value)
significant_genes_smoking = which(ttest_smoking < 0.05)
# significant_genes_smoking


for (i in head(significant_genes_smoking, 20)) {
  result = paste("Gene position:", i, "p-value:", ttest_smoking[i])
  cat(result, "\n")
}
```

```
## Gene position: 6 p-value: 0.0343909301236234
## Gene position: 9 p-value: 0.000607615703685403
## Gene position: 14 p-value: 0.007810223220614
## Gene position: 23 p-value: 0.0327207483557484
## Gene position: 30 p-value: 0.00651755485459916
## Gene position: 37 p-value: 0.0107220806269349
## Gene position: 47 p-value: 0.00618988783129787
## Gene position: 51 p-value: 0.00954081344081713
## Gene position: 55 p-value: 0.00319779946530702
## Gene position: 59 p-value: 0.00340279350645569
## Gene position: 64 p-value: 0.0291664130755789
## Gene position: 74 p-value: 0.0192797248152838
## Gene position: 97 p-value: 0.00889531421209028
## Gene position: 99 p-value: 0.0270604760749846
## Gene position: 100 p-value: 0.0412184781145362
## Gene position: 113 p-value: 0.0328441876871098
## Gene position: 114 p-value: 0.00704794876747234
## Gene position: 174 p-value: 0.0172609670574429
## Gene position: 176 p-value: 0.039979236805241
## Gene position: 188 p-value: 0.00411490165761968
```

- **3.2**

Αντίστοιχα η ίδια διαδικασία για τα φύλλα

```
female_ind = which(phen$gender == "female")
male_ind = which(phen$gender == "male")


ttest_gender = apply(expr, 1, function(x) t.test(x[female_ind], x[male_ind])$p.value)

significant_genes_gender = which(ttest_gender < 0.05)

for (i in head(significant_genes_gender, 20)) {
  result = paste("Gene position:", i, "p-value:", ttest_smoking[i])
  cat(result, "\n")
}
```

```
## Gene position: 13 p-value: 0.33866064677971
## Gene position: 23 p-value: 0.0327207483557484
## Gene position: 36 p-value: 0.60196513902111
## Gene position: 39 p-value: 0.227977755721823
## Gene position: 41 p-value: 0.340622085825681
## Gene position: 85 p-value: 0.331129908330557
## Gene position: 87 p-value: 0.411159920263638
## Gene position: 98 p-value: 0.130351684960388
## Gene position: 121 p-value: 0.805772411851807
## Gene position: 131 p-value: 0.196350584835517
## Gene position: 139 p-value: 0.757620435825385
## Gene position: 195 p-value: 0.285176446648799
## Gene position: 201 p-value: 0.0621648907571032
## Gene position: 219 p-value: 0.519545033311186
## Gene position: 226 p-value: 0.0659081265797984
## Gene position: 282 p-value: 0.0495800371914624
## Gene position: 291 p-value: 0.625761757313732
## Gene position: 303 p-value: 0.978908751915445
## Gene position: 309 p-value: 0.715080296806899
## Gene position: 316 p-value: 0.684540019615368
```

Μέρος Β

Επεξεργασία αρχείου

```
# Processing the file
processFile = function(filepath){

  con = file(filepath, "r")
  on.exit(close(con))
  seqs = list()

  while(TRUE){

    line = readLines(con, n=1)

    if(length(line)==0){
      break
    }

    isNewSeq = length(grep(">", line, ignore.case=TRUE, perl=TRUE)) > 0

    if(isNewSeq){
      motiv = ""
      name = gsub(">([^_]*_[^_ ]*).*", x=line, replacement="\\1", perl=TRUE)
    }else{
      motiv = paste(motiv, line, sep="")
      seqs[[name]] = motiv
    }
  }
  return(seqs)
}
```

α) Καταμέτρηση μετρώντας πολλές φορές ένα μοτίβο αν υπάρχει πάνω από 1 φορά σε κάποια περιοχή

- έφτιαξα ένα demo αρχείο για να ελέγξω τον κώδικα αλλά και για πιο γρήγορα

```
# all patterns list
patterns_a = list()

# exporting the patterns for question 1
getPatternsA = function(ptr_string, base_length=6){

  start_pos = 1
  end_pos = nchar(ptr_string) - base_length + 1

  v = strsplit(ptr_string, split="")[[1]]

  for(i in start_pos:end_pos){

    motiv = paste(v[i:(i+base_length-1)], collapse="")

    if(!is.null(patterns_a[[motiv]])){
      patterns_a[[motiv]] = patterns_a[[motiv]] + 1

    } else{
      patterns_a[[motiv]] = 1
    }
  }

  return(patterns_a)
}

human_file_a = processFile("test_file.fa")
# human_file_a = processFile("upstream1000_human.fa")
# chimp_file_a = processFile("upstream1000_chimpanzee.fa")
# mouse_file_a = processFile("upstream1000_mouse.fa")
# rat_file_a = processFile("upstream1000_rat.fa")

cat("Demo File: -test_file.fa-\n\n")
cat("Names and sequences:\n\n")
for(i in seq_along(human_file_a)){
  cat(names(human_file_a)[i], human_file_a[[i]], "\n")
  cat("\n")
}
cat("Total counts for each pattern:\n\n")
for(i in names(human_file_a)){
  print(getPatternsA(human_file_a[[i]]))
}
```

```
## Demo File: -test_file.fa-

## Names and sequences:

## NM_000299 cttttacttttatttccatcaaagtaaatacttaaaaaaaaaaaaaaaacactcagctcctgttacacaccaaattcactgatgtgggctc

## NM_001299 cacgccgaaagatttccatcaaagtaaatacttaaaaaaaaaaaaaaaacactcagctcctgttacacaccaaattcactgatgtgggcca

## NM_002289 cacgccgaaagatttccatcaaagtaaatacttaaaaacgccgagaaacactcagctcctgttacacaccaaattcactgatgtgggcca

## NM_003119 cacgccgaaagatttccatcaaagtaaatacttaaaaacgccgagaaacactcagctcctgttacacaccaaattcactgatgtgggcca

## Total counts for each pattern:

## $ctttta
## [1] 2
##
## $ttttac
## [1] 1
##
## $tttact
## [1] 1
##
## $ttactt
## [1] 1
##
## $tacttt
## [1] 1
##
## $actttt
## [1] 1
##
## $cacgcc
## [1] 3
##
## $ttttat
## [1] 1
##
## $tttatt
## [1] 1
##
## $ttattt
## [1] 1
##
## $tatttc
## [1] 1
##
## $aaaaaa
## [1] 10
##
## $atttcc
## [1] 1
##
```

```
## $tttcca
## [1] 1
## and more .....
```

α) Καταμέτρηση μετρώντας μόνο 1 φορά το μοτίβο

```
#unique patterns list
patterns_b = list()

# exporting all patterns
getPatternsB = function(ptr_string, base_length=6){

  start_pos = 1
  end_pos = nchar(ptr_string) - base_length + 1

  v = strsplit(ptr_string, split="")[[1]]

  # patterns_a = list()

  for(i in start_pos:end_pos){

    motiv = paste(v[i:(i+base_length-1)], collapse="")

    if(!is.null(patterns_b[[motiv]])){
      patterns_b[[motiv]] =  1

    } else{
      patterns_b[[motiv]] = 1
    }
  }

  return(patterns_b)
}

human_file_b = processFile("test_file.fa")
# human_file_b = processFile("upstream1000_human.fa")
# chimp_file_b = processFile("upstream1000_chimpanzee.fa")
# mouse_file_b = processFile("upstream1000_mouse.fa")
# rat_file_b = processFile("upstream1000_rat.fa")

cat("Demo File: -test_file.fa-\n\n")
cat("Names and sequences:\n\n")
for(i in seq_along(human_file_b)){
  cat(names(human_file_b)[i], human_file_b[[i]], "\n")
  cat("\n")
 }
 cat("Total counts for each pattern:\n\n")

for(i in names(human_file_a)){
  print(getPatternsB(human_file_b[[i]]))
}
```

```
## Demo File: -test_file.fa-

## Names and sequences:

## NM_000299 cttttacttttatttccatcaaagtaaatacttaaaaaaaaaaaaaaaacactcagctcctgttacacaccaaattcactgatgtgggctc
##
## NM_001299 cacgccgaaagatttccatcaaagtaaatacttaaaaaaaaaaaaaaaacactcagctcctgttacacaccaaattcactgatgtgggcca
##
## NM_002289 cacgccgaaagatttccatcaaagtaaatacttaaaaacgccgagaaacactcagctcctgttacacaccaaattcactgatgtgggcca
##
## NM_003119 cacgccgaaagatttccatcaaagtaaatacttaaaaacgccgagaaacactcagctcctgttacacaccaaattcactgatgtgggcca

## Total counts for each pattern

## $ctttta
## [1] 1
##
## $ttttac
## [1] 1
##
## $tttact
## [1] 1
##
## $ttactt
## [1] 1
##
## $tacttt
## [1] 1
##
## $actttt
## [1] 1
##
## $ttttat
## [1] 1
##
## $tttatt
## [1] 1
##
## $ttattt
## [1] 1
##
## $tatttc
## [1] 1
##
## and more .....
##
```