R Boot Camp Problem Set

Carly Bobak

August 10, 2020

Establishing reliable biomarkers for assessing and validating clinical diagnosis at early prodromal stages of Parkinson’s disease is crucial for developing therapies to slow or halt disease progression. This data set uses whole blood gene expression profiling from over 500 individuals where we will attempt to find a gene signature. This repository contains the gene expression profiles collected in the GENEPARK consortium. The main study sought a classifier for IPD. These data contain 233 healthy controls, 205 IPD patients, and 48 patients with other neurodegenerative diseases (NDD). Other samples are available in the data and can be used for additional analyses. The largest class of these additional samples are 22 samples from genetic unaffected controls and 41 genetic PD patients.

Note: the original study which uploaded this data to NIH Geo is not yet published.

Data Wrangling

Let’s start by loading in our data sets. Download these from the sharepoint site, and make a new folder for R bootcamp. We’ll switch to this directory here.

Note that we have both a phenotype file, as well as a file which includes the normalized and log transformed expression values. We can use the read.csv function to load in these files.

```
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
```

```
library(stringr)


pheno<-read.csv("parkPheno.csv")
expr<-read.csv("simulatedData.csv")
```

We should start by summarizing both these files. Try the following functions: head(), and View(). Note that while the dimensions on our phenotype file are reasonable, we have 552 columns in our expression file. Just summarize the first 10 columns of this file.

```r
## Enter your own code here

head(pheno)
```

```
##   geo_accession submission_date last_update_date type      tissue     organism
## 1    GSM2631171     May 17 2017      May 20 2017  RNA Whole blood Homo sapiens
## 2    GSM2631309     May 17 2017      May 20 2017  RNA Whole blood Homo sapiens
## 3    GSM2631219     May 17 2017      May 20 2017  RNA Whole blood Homo sapiens
## 4    GSM2630775     May 17 2017      May 20 2017  RNA Whole blood Homo sapiens
## 5    GSM2631147     May 17 2017      May 20 2017  RNA Whole blood Homo sapiens
## 6    GSM2630853     May 17 2017      May 20 2017  RNA Whole blood Homo sapiens
##   subject_id disease_label     sex mutated_pd_genes age_at_exam age_at_symptoms
## 1      L2899   ATYPICAL_PD    Male             NONE          NA              53
## 2      L2872   ATYPICAL_PD    Male             NONE          NA              64
## 3      L2131   ATYPICAL_PD    Male             NONE          NA              NA
## 4      L2573           CBD  Female             NONE          NA              60
## 5      L2697           CBD  Female             NONE          NA              66
## 6      L3031       CONTROL    Male             NONE          NA              41
##   updrs updrs_ii updrs_iii_score_on updrs_iii_score_off updrs_iv hoehn_yahr_on
## 1     1        4                 19                   0        0             2
## 2     0        0                  0                   0        0             9
## 3     0        0                  0                   0        0             0
## 4     0        0                  0                   0        0             9
## 5     0        0                 30                   0        0             9
## 6     0        0                  1                   0        0             8
##   hoehn_yahr_off moca_score
## 1              0         21
## 2              0          0
## 3              0          0
## 4              0          0
## 5              0          0
## 6              0         30
```

```r
View(pheno)
summary(expr[1:20668,1:10, drop=FALSE])
```

```
##        X             GeneName       GSM2631171         GSM2631309
##  Min.   :    1   A1BG    :    1   Min.   :-5.223788   Min.   :-6.09018
##  1st Qu.: 5168   A1BG-AS1:    1   1st Qu.:-0.960423   1st Qu.:-0.92906
##  Median :10334   A1CF    :    1   Median :-0.004842   Median : 0.01385
##  Mean   :10334   A2M     :    1   Mean   :-0.009648   Mean   : 0.01249
##  3rd Qu.:15501   A2M-AS1 :    1   3rd Qu.: 0.953228   3rd Qu.: 0.95912
##  Max.   :20668   (Other) :20662   Max.   : 5.766301   Max.   : 5.66627
##                  NA's    :    1
##    GSM2631219         GSM2630775          GSM2631147         GSM2630853
##  Min.   :-6.39097   Min.   :-5.206869   Min.   :-5.27578   Min.   :-6.115736
##  1st Qu.:-0.97337   1st Qu.:-0.981831   1st Qu.:-0.96379   1st Qu.:-0.944666
##  Median :-0.01097   Median : 0.001772   Median : 0.01906   Median :-0.007942
##  Mean   :-0.00354   Mean   :-0.000010   Mean   : 0.00298   Mean   :-0.009892
##  3rd Qu.: 0.95324   3rd Qu.: 0.971013   3rd Qu.: 0.98545   3rd Qu.: 0.945826
##  Max.   : 6.56118   Max.   : 5.275719   Max.   : 5.18612   Max.   : 5.570111
##
```

```
##    GSM2630769          GSM2631196
##  Min.   :-5.608142   Min.   :-6.303044
##  1st Qu.:-0.968002   1st Qu.:-0.970730
##  Median :-0.001583   Median :-0.004689
##  Mean   : 0.014813   Mean   :-0.006484
##  3rd Qu.: 0.987677   3rd Qu.: 0.977216
##  Max.   : 5.591597   Max.   : 5.434250
##
```

Try summarizing the phenotype data:

```
## Enter your own code here

summary(pheno)
```

```
##     geo_accession    submission_date    last_update_date  type
##  GSM2630758:  1   May 17 2017:550    May 20 2017:550     RNA:550
##  GSM2630759:  1
##  GSM2630760:  1
##  GSM2630761:  1
##  GSM2630762:  1
##  GSM2630763:  1
##  (Other)   :544
##         tissue            organism     subject_id              disease_label
##  Whole blood:550   Homo sapiens:550    B25   :  1    CONTROL           :233
##                                        B27   :  1    IPD               :205
##                                        B28   :  1    GPD               : 41
##                                        B29   :  1    GENETIC_UNAFFECTED: 22
##                                        B32   :  1    HD                : 19
##                                        B36   :  1    MSA               :  8
##                                       (Other):544   (Other)            : 22
##       sex     mutated_pd_genes  age_at_exam    age_at_symptoms
##         : 45    NONE  :428      Min.   :30.00   Min.   :10.00
##   Female:281          : 48      1st Qu.:54.75   1st Qu.:45.00
##   Male  :224    PARKIN: 22      Median :61.00   Median :55.00
##                 PINK1 : 21      Mean   :60.56   Mean   :53.61
##                 NMF   : 12      3rd Qu.:68.25   3rd Qu.:64.00
##                 LRRK2 : 11      Max.   :82.00   Max.   :78.00
##                (Other):  8      NA’s   :266     NA’s   :325
##      updrs           updrs_ii      updrs_iii_score_on updrs_iii_score_off
##  Min.   : 0.000   Min.   : 0.000   0      :198        0      :381
##  1st Qu.: 0.000   1st Qu.: 0.000          :152               :108
##  Median : 0.000   Median : 0.000   1      : 13        1      : 11
##  Mean   : 1.171   Mean   : 4.593   2      : 12        2      : 10
##  3rd Qu.: 2.000   3rd Qu.: 7.000   15     :  9        17     :  4
##  Max.   :36.000   Max.   :35.000   18     :  8        19     :  3
##  NA’s   :122      NA’s   :123      (Other):158        (Other): 33
##     updrs_iv      hoehn_yahr_on hoehn_yahr_off   moca_score
##  Min.   : 0.000   0      :164   0      :406    0      :191
##  1st Qu.: 0.000          :148          :101    30     : 95
##  Median : 0.000   8      : 57    ND    :  8    29     : 76
##  Mean   : 1.236   1      : 43   1      :  8    28     : 47
##  3rd Qu.: 1.000   2      : 37   2      :  6    26     : 30
##  Max.   :14.000   3      : 31   4      :  6   27     : 24
##  NA's   :118      (Other): 70   (Other): 15   (Other): 87
```

We make the following observations.

1. We have some unnecessary data in this file. We aren't interested in the submission and last update date. We can reduce the dimensions of this file so it handles nicer from now on.
2. We have a LOT of missing data. You'll learn how to handle this in some of your biostats classes! For now, we'll run what analyses we can given the data we have.
3. Some of our scores have been read in as character values (and they should be numbers). If you investigate this further, you'll find that some values have been recorded as "ND", which we'll assume means "no data". We will need to record these as NA values in R.

Our next step is to address item one. We will reduce the dimensions of our pheno data frame to include only that information that we're interested in modelling. We can exclude the dates, type (as it's all RNA), tissue (all whole blood), organism (all homo sapiens), and subject ID (we will be using geo_accession as our unique indicator). As well, we will exclude mutated_pd_genes, as we indend to define our own gene signature later this week.

Subset your pheno data frame to include columns 1,8,9,11:20.

```
## Enter your own code here

library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------

## v ggplot2 3.3.0     v readr   1.3.1
## v tibble  3.0.1     v purrr   0.3.4
## v tidyr   1.1.0     v forcats 0.5.0

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## -- Conflicts -----------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Enter your own code here FIX
pheno <- pheno %>% select(1,8,9,11:20)
pheno
```

```
##     geo_accession       disease_label     sex age_at_exam age_at_symptoms updrs
## 1      GSM2631171         ATYPICAL_PD    Male          NA              53     1
## 2      GSM2631309         ATYPICAL_PD    Male          NA              64     0
## 3      GSM2631219         ATYPICAL_PD    Male          NA              NA     0
## 4      GSM2630775                 CBD  Female          NA              60     0
## 5      GSM2631147                 CBD  Female          NA              66     0
## 6      GSM2630853             CONTROL    Male          NA              41     0
## 7      GSM2630769             CONTROL  Female          NA              43     0
## 8      GSM2631196             CONTROL    Male          40              NA    NA
## 9      GSM2631194             CONTROL    Male          42              NA    NA
## 10     GSM2631197             CONTROL    Male          43              NA    NA
## 11     GSM2631195             CONTROL    Male          49              NA    NA
## 12     GSM2631198             CONTROL    Male          40              NA    NA
## 13     GSM2631306             CONTROL                  NA              NA     0
## 14     GSM2631162             CONTROL                  NA              NA     0
## 15     GSM2631172             CONTROL                  NA              NA     0
## 16     GSM2631241             CONTROL  Female          NA              NA     0
## 17     GSM2631252             CONTROL  Female          NA              NA     0
## 18     GSM2630927             CONTROL  Female          NA              NA     0
## 19     GSM2630928             CONTROL  Female          NA              NA     0
## 20     GSM2631227             CONTROL  Female          NA              NA     0
## 21     GSM2631231             CONTROL  Female          NA              NA     0
## 22     GSM2631235             CONTROL  Female          NA              NA     0
## 23     GSM2631236             CONTROL  Female          NA              NA     0
## 24     GSM2631238             CONTROL  Female          NA              NA     0
## 25     GSM2631239             CONTROL  Female          NA              NA     0
## 26     GSM2631243             CONTROL  Female          NA              NA     0
## 27     GSM2630771             CONTROL  Female          NA              NA     0
## 28     GSM2630783             CONTROL  Female          NA              NA     0
## 29     GSM2630830             CONTROL  Female          NA              NA     0
## 30     GSM2630857             CONTROL  Female          NA              NA     0
## 31     GSM2630868             CONTROL  Female          NA              NA     0
## 32     GSM2630818             CONTROL  Female          NA              NA     0
## 33     GSM2630907             CONTROL  Female          NA              NA     0
## 34     GSM2630909             CONTROL  Female          NA              NA     0
## 35     GSM2630916             CONTROL  Female          NA              NA     0
## 36     GSM2630923             CONTROL  Female          NA              NA     0
## 37     GSM2630925             CONTROL  Female          NA              NA     0
## 38     GSM2630929             CONTROL  Female          NA              NA     0
## 39     GSM2630930             CONTROL  Female          NA              NA     0
## 40     GSM2630932             CONTROL  Female          NA              NA     0
## 41     GSM2631221             CONTROL    Male          NA              NA     0
## 42     GSM2631230             CONTROL    Male          NA              NA     0
## 43     GSM2631232             CONTROL    Male          NA              NA     0
## 44     GSM2631234             CONTROL    Male          NA              NA     0
## 45     GSM2631237             CONTROL    Male          NA              NA     0
## 46     GSM2631240             CONTROL    Male          NA              NA     0
## 47     GSM2631242             CONTROL    Male          NA              NA     0
## 48     GSM2631248             CONTROL    Male          NA              NA     0
## 49     GSM2630899             CONTROL    Male          NA              NA     0
## 50     GSM2630905             CONTROL    Male          NA              NA     0
## 51     GSM2630906             CONTROL    Male          NA              NA     0
## 52     GSM2630917             CONTROL    Male          NA              NA     0
## 53     GSM2630922             CONTROL    Male          NA              NA     0
## 54     GSM2630924             CONTROL    Male          NA              NA     0
## 55     GSM2631298             CONTROL                  NA              NA     0
## 56     GSM2631300             CONTROL                  NA              NA     0
## 57     GSM2631301             CONTROL                  NA              NA     0
## 58     GSM2631304             CONTROL                  NA              NA     0
## 491    GSM2630762                 IPD  Female          NA              NA     0
## 492    GSM2630817                 IPD  Female          NA              NA     0
## 493    GSM2630826                 IPD  Female          NA              NA     0
## 494    GSM2630895                 IPD  Female          NA              NA     0
## 495    GSM2630915                 IPD  Female          NA              NA     0
## 496    GSM2630838                 IPD  Female          NA              NA     0
## 497    GSM2630900                 IPD  Female          NA              NA     0
## 498    GSM2631224                 IPD    Male          NA              NA     0
## 499    GSM2631244                 IPD    Male          NA              NA     0
## 500    GSM2631245                 IPD    Male          NA              NA     0
## 501    GSM2631246                 IPD    Male          NA              NA     0
## 502    GSM2631247                 IPD    Male          NA              NA     0
## 503    GSM2630794                 IPD    Male          NA              NA     0
## 504    GSM2630870                 IPD    Male          NA              NA     0
## 505    GSM2630890                 IPD    Male          NA              NA     0
## 506    GSM2630901                 IPD    Male          NA              NA     0
## 507    GSM2630902                 IPD    Male          NA              NA     0
## 508    GSM2630904                 IPD    Male          NA              NA     0
## 509    GSM2630764                 IPD    Male          NA              NA     0
## 510    GSM2630781                 IPD    Male          NA              NA     0
## 511    GSM2630791                 IPD    Male          NA              NA     0
## 512    GSM2630862                 IPD    Male          NA              NA     0
## 513    GSM2630898                 IPD    Male          NA              NA     0
## 514    GSM2630903                 IPD    Male          NA              NA     0
## 515    GSM2630931                 IPD    Male          NA              NA     0
## 516    GSM2631150                 IPD                  NA              NA     0
## 517    GSM2631151                 IPD                  NA              NA     0
## 518    GSM2631155                 IPD                  NA              NA     0
## 519    GSM2631156                 IPD                  NA              NA     0
## 520    GSM2631157                 IPD                  NA              NA     0
## 521    GSM2631160                 IPD                  NA              NA     0
## 522    GSM2631169                 IPD                  NA              NA     0
## 523    GSM2631154                 IPD                  NA              NA     0
## 524    GSM2631164                 IPD                  NA              NA     0
## 525    GSM2631166                 IPD                  NA              NA     0
## 526    GSM2631168                 IPD                  NA              NA     0
## 527    GSM2631170                 IPD                  NA              NA     0
## 528    GSM2631178                 IPD                  NA              NA     0
## 529    GSM2631019                 IPD  Female          60              NA     3
## 530    GSM2630886                 IPD  Female          NA              NA     0
## 531    GSM2630765                 IPD    Male          NA              NA     0
## 532    GSM2630831                 MSA    Male          NA              51     0
## 533    GSM2631315                 MSA    Male          66              60     0
## 534    GSM2630814                 MSA  Female          NA              64     1
## 535    GSM2630776                 MSA    Male          NA              68     0
## 536    GSM2631204                 MSA  Female          75              71     5
## 537    GSM2630943                 MSA    Male          76              72     6
## 538    GSM2631199                 MSA  Female          81              74     1
## 539    GSM2631299                 MSA                  NA              NA     0
## 540    GSM2631223         PD_DEMENTIA    Male          NA              69     1
## 541    GSM2631233         PD_DEMENTIA    Male          NA              NA     0
## 542    GSM2631313                 PSP    Male          63              52    NA
## 543    GSM2630808                 PSP  Female          NA              53     0
## 544    GSM2631314                 PSP    Male          76              66    36
## 545    GSM2630836                 PSP  Female          NA              72     2
## 546    GSM2631216                 PSP  Female          NA              72     0
## 547    GSM2630889                 PSP                  NA              NA     0
## 548    GSM2631174                 PSP                  NA              NA     0
## 549    GSM2630894                 PSP    Male          NA              NA     0
## 550    GSM2630792   Vascular dementia                  NA              NA     0
##     updrs_ii updrs_iii_score_on updrs_iii_score_off updrs_iv hoehn_yahr_on
## 1          4                 19                   0        0             2
## 2          0                  0                   0        0             9
## 3          0                  0                   0        0             0
## 4          0                  0                   0        0             9
## 5          0                 30                   0        0             9
## 6          0                  1                   0        0             8
## 7          0                  0                   0        0             8
## 8         NA                                      0        0             0
## 9         NA                                      0        0             0
## 10        NA                                      0        0             0
## 11        NA                                      0        0             0
## 12        NA                                      0       NA
## 13         0                  0                   0        0             0
## 14         0                  0                   0        0             0
## 15         0                  0                   0        0             0
## 16         0                  0                   0        0             0
## 17         0                  0                   0        0             0
## 18         0                  0                   0        0             0
## 19         0                  0                   0        0             0
## 20         0                  0                   0        0             0
## 21         0                  0                   0        0             0
## 22         0                  0                   0        0             0
## 23         0                  0                   0        0             0
## 24         0                  0                   0        0             0
## 25         0                  0                   0        0             0
## 26         0                  0                   0        0             0
## 27         0                  0                   0        0             0
## 28         0                  0                   0        0             0
## 29         0                  0                   0        0             0
## 30         0                  0                   0        0             0
## 31         0                  0                   0        0             0
## 32         0                  0                   0        0             0
## 33         0                  0                   0        0             0
## 34         0                  0                   0        0             0
## 35         0                  0                   0        0             0
## 36         0                  0                   0        0             0
## 37         0                  0                   0        0             0
## 38         0                  0                   0        0             0
## 39         0                  0                   0        0             0
## 40         0                  0                   0        0             0
## 41         0                  0                   0        0             0
## 42         0                  0                   0        0             0
## 43         0                  0                   0        0             0
## 44         0                  0                   0        0             0
## 45         0                  0                   0        0             0
## 46         0                  0                   0        0             0
## 47         0                  0                   0        0             0
## 534       16                 35                   0        6             3
## 535        0                 46                   0        0             9
## 536       28                                     64        1
## 537       25                 43                            0             4
## 538       26                                     55        1
## 539        0                  0                   0        0             0
## 540        7                 26                   0        0             3
## 541        0                  2                   0        0             8
## 542       NA                                     26       NA
## 543        0                 37                   0        0             9
## 544       NA                 75                            0             5
## 545       15                 31                   0        0             3
## 546        0                  0                   0        0             9
## 547        0                  0                   0        0             0
## 548        0                  0                   0        0             0
## 549        0                  2                   0        1             8
## 550        0                  0                   0        0             0
##     hoehn_yahr_off moca_score
## 1                0         21
## 2                0          0
## 3                0          0
## 4                0          0
## 5                0          0
## 6                0         30
## 7                0          0
## 8                0          0
## 9                0          0
## 10               0          0
## 11               0          0
## 12               0         30
## 13               0          0
## 14               0          0
## 15               0          0
## 16               0          0
## 17               0          0
## 18               0          0
## 19               0          0
## 20               0          0
## 21               0          0
## 22               0          0
## 23               0          0
## 24               0          0
## 25               0          0
## 26               0          0
## 27               0          0
## 28               0          0
## 29               0          0
## 30               0          0
## 31               0          0
## 32               0          0
## 33               0          0
## 34               0          0
## 35               0          0
## 36               0          0
## 523                  0          0
## 524                  0          0
## 525                  0          0
## 526                  0          0
## 527                  0          0
## 528                  0          0
## 529                  2         26
## 530                  0         30
## 531                  0         30
## 532                  0          0
## 533                  4
## 534                  0         23
## 535                  0          0
## 536                  5
## 537                            24
## 538                  5         27
## 539                  0          0
## 540                  0         20
## 541                  0         29
## 542                2.5
## 543                  0          0
## 544
## 545                  0          0
## 546                  0          0
## 547                  0          0
## 548                  0          0
## 549                  0         29
## 550                  0          0
```

Next we need to correct the columns which contain "ND". You can use the "which" function to find the index of of the matrices which are "ND", and then set these to NA. Set columns 8,9,11,12,13 to numeric values using the "as.numeric" function inside a "sapply" loop. Run a summary of the data frame again.

```
index<-which(pheno == " ND",arr.ind = T)
pheno[index]<-NA
j<-c(8,9,11,12,13)
pheno[,j]<-sapply(unlist(pheno[,j]),as.numeric)
summary(pheno)
```

```
##     geo_accession            disease_label      sex        age_at_exam
##  GSM2630758:  1    CONTROL           :233          : 45   Min.   :30.00
##  GSM2630759:  1    IPD               :205    Female:281   1st Qu.:54.75
##  GSM2630760:  1    GPD               : 41    Male  :224   Median :61.00
##  GSM2630761:  1    GENETIC_UNAFFECTED: 22                 Mean   :60.56
##  GSM2630762:  1    HD                : 19                 3rd Qu.:68.25
##  GSM2630763:  1    MSA               :  8                 Max.   :82.00
##  (Other)   :544   (Other)            : 22                 NA's   :266
##  age_at_symptoms     updrs           updrs_ii      updrs_iii_score_on
##  Min.   :10.00   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.:45.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :55.00   Median : 0.000   Median : 0.000   Median : 3.000
##  Mean   :53.61   Mean   : 1.171   Mean   : 4.593   Mean   : 9.593
##  3rd Qu.:64.00   3rd Qu.: 2.000   3rd Qu.: 7.000   3rd Qu.:12.250
##  Max.   :78.00   Max.   :36.000   Max.   :35.000   Max.   :53.000
##  NA's   :325     NA's   :122      NA's   :123      NA's   :2
##  updrs_iii_score_off    updrs_iv      hoehn_yahr_on   hoehn_yahr_off
##  Min.   : 1.000      Min.   : 0.000   Min.   : 1.00   Min.   : 1.00
##  1st Qu.: 3.000      1st Qu.: 0.000   1st Qu.: 1.00   1st Qu.: 3.00
##  Median : 3.000      Median : 0.000   Median : 3.00   Median : 3.00
##  Mean   : 4.829      Mean   : 1.236   Mean   :15.63   Mean   : 4.45
##  3rd Qu.: 3.000      3rd Qu.: 1.000   3rd Qu.:26.00   3rd Qu.: 3.00
##  Max.   :58.000      Max.   :14.000   Max.   :60.00   Max.   :60.00
##                      NA's   :118      NA's   :10      NA's   :8
##    moca_score
##  Min.   : 1.00
##  1st Qu.: 3.00
##  Median :22.00
##  Mean   :16.09
##  3rd Qu.:25.00
##  Max.   :27.00
##  NA's   :9
```

Let's look at a summary of the first 10 columns of expression data set.

```
## Enter your own code here

summary(expr[1:20668,1:10, drop=FALSE])
```

```
##        X              GeneName       GSM2631171          GSM2631309
##  Min.   :    1   A1BG    :    1   Min.   :-5.223788   Min.   :-6.09018
##  1st Qu.: 5168   A1BG-AS1:    1   1st Qu.:-0.960423   1st Qu.:-0.92906
##  Median :10334   A1CF    :    1   Median :-0.004842   Median : 0.01385
##  Mean   :10334   A2M     :    1   Mean   :-0.009648   Mean   : 0.01249
##  3rd Qu.:15501   A2M-AS1 :    1   3rd Qu.: 0.953228   3rd Qu.: 0.95912
##  Max.   :20668   (Other) :20662   Max.   : 5.766301   Max.   : 5.66627
##                  NA's    :    1
##    GSM2631219         GSM2630775          GSM2631147          GSM2630853
##  Min.   :-6.39097   Min.   :-5.206869   Min.   :-5.27578   Min.   :-6.115736
##  1st Qu.:-0.97337   1st Qu.:-0.981831   1st Qu.:-0.96379   1st Qu.:-0.944666
##  Median :-0.01097   Median : 0.001772   Median : 0.01906   Median :-0.007942
##  Mean   :-0.00354   Mean   :-0.000010   Mean   : 0.00298   Mean   :-0.009892
##  3rd Qu.: 0.95324   3rd Qu.: 0.971013   3rd Qu.: 0.98545   3rd Qu.: 0.945826
##  Max.   : 6.56118   Max.   : 5.275719   Max.   : 5.18612   Max.   : 5.570111
##
##    GSM2630769          GSM2631196
##  Min.   :-5.608142   Min.   :-6.303044
##  1st Qu.:-0.968002   1st Qu.:-0.970730
##  Median :-0.001583   Median :-0.004689
##  Mean   : 0.014813   Mean   :-0.006484
##  3rd Qu.: 0.987677   3rd Qu.: 0.977216
##  Max.   : 5.591597   Max.   : 5.434250
##
```

We don't need the X1 variable - this is just remaining row labels in the csv file. Let's remove this variable.

```
## Enter your own code here

expr <- expr %>%
  select(-X)
```

We don't see any evidence of missing values in our summary, but we should check all of the columns (excluding the ProbeID and GeneName). You can check this with the “anyNA”" function.

```
## Enter your own code here
expr_na_CheckPrep = subset(expr, select = -c(GeneName)) # temp dropping these for N/A check
anyNA(expr_na_CheckPrep, recursive = FALSE)
```

```
## [1] FALSE
```

```
#anyNA(expr)
```

Let's identify how big this problem is, and where it occurs.

```
which(is.na(expr),arr.ind = T)
```

```
##         row col
## [1,] 20668   1
```

So one of our gene names is NA! This isn't useful, so let's remove this row.

```
## Enter your own code here

expr <- expr[-20668,]
```

We should see if the unique identifiers in our two data sets match. Check for a perfect match using the “identical” function.

```
identical(colnames(expr[,-1]),as.character(pheno[,1]))
```

```
## [1] TRUE
```

Question: why is the '-1' necessary here? Answer below!

The -1 is necessary, because exclude GeneNames

So that we don't lose any work, let's clean up our workspace to include only our cleaned expression and pheno data sets, which we can reload later.

Exploratory Data Analysis

In this section we are going to explore some of the data we have, and maybe develop a diagnostic signature for Parkinson's disease.

First, load in your data from yesterday.

Let's re-examine our pheno data set with the summary function again.

```
## Enter your own code here

summary(pheno)
```

```
##     geo_accession             disease_label      sex       age_at_exam
##  GSM2630758:  1    CONTROL           :233           : 45   Min.   :30.00
##  GSM2630759:  1    IPD               :205    Female:281   1st Qu.:54.75
##  GSM2630760:  1    GPD               : 41    Male  :224   Median :61.00
##  GSM2630761:  1    GENETIC_UNAFFECTED: 22                 Mean   :60.56
##  GSM2630762:  1    HD                : 19                 3rd Qu.:68.25
##  GSM2630763:  1    MSA               :  8                 Max.   :82.00
##  (Other)   :544   (Other)           : 22                 NA's   :266
##  age_at_symptoms     updrs           updrs_ii      updrs_iii_score_on
##  Min.   :10.00   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.:45.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :55.00   Median : 0.000   Median : 0.000   Median : 3.000
##  Mean   :53.61   Mean   : 1.171   Mean   : 4.593   Mean   : 9.593
##  3rd Qu.:64.00   3rd Qu.: 2.000   3rd Qu.: 7.000   3rd Qu.:12.250
##  Max.   :78.00   Max.   :36.000   Max.   :35.000   Max.   :53.000
##  NA's   :325     NA's   :122      NA's   :123      NA's   :2
##  updrs_iii_score_off    updrs_iv      hoehn_yahr_on    hoehn_yahr_off
##  Min.   : 1.000      Min.   : 0.000   Min.   : 1.00   Min.   : 1.00
##  1st Qu.: 3.000      1st Qu.: 0.000   1st Qu.: 1.00   1st Qu.: 3.00
##  Median : 3.000      Median : 0.000   Median : 3.00   Median : 3.00
##  Mean   : 4.829      Mean   : 1.236   Mean   :15.63   Mean   : 4.45
##  3rd Qu.: 3.000      3rd Qu.: 1.000   3rd Qu.:26.00   3rd Qu.: 3.00
##  Max.   :58.000      Max.   :14.000   Max.   :60.00   Max.   :60.00
##                      NA's   :118      NA's   :10      NA's   :8
##    moca_score
##  Min.   : 1.00
##  1st Qu.: 3.00
##  Median :22.00
##  Mean   :16.09
##  3rd Qu.:25.00
##  Max.   :27.00
##  NA's   :9
```

We need to further delve into our disease label in order to simplify some of this analysis. Attach your pheno data frame using the attach function, and then summarize the disease label vector.

```
## Enter your own code here

attach(pheno)

summary(disease_label)
```

```
##         ATYPICAL_PD                 CBD             CONTROL                 DRD
##                   3                   2                 233                   3
##            DRD-DYT5  GENETIC_UNAFFECTED                 GPD                  HD
##                   3                  22                  41                  19
##                 IPD                 MSA         PD_DEMENTIA                 PSP
##                 205                   8                   2                   8
##   Vascular dementia
##                   1
```

Here we have the counts of all the diseases in our data set. If you look at the actual excel file (not the csv), I've put in a dictionary for these acronyms if you're curious. Here, our controls and our genetic unaffected are both considered to be healthy controls. Any label which contains PD is some subset of Parkison's Disease, and the other labels represent other neurological disorders. We need to make a variable which records a 1 for our cases, and a 0 for our controls. Here, since we are interested in a signature that distinguishes PD from our other disease, the other diseases are technically part of the control set.

Try to set your case control vector using the grep function to find the indicies which contain "PD". At the end, sum your case vector to check that it worked. Make another variable of the words "case" and "control"

```
## Enter your own code here
pheno <- pheno %>%
  mutate(case = if_else(str_detect(disease_label, "PD"), 1, 0))
#Case Set
case <-  pheno %>%
  mutate(case = if_else(str_detect(disease_label, "PD"), 1, 0)) %>%
  subset(case == 1)
#Control
control <-  pheno %>%
  mutate(case = if_else(str_detect(disease_label, "PD"), 1, 0)) %>%
  subset(case == 0)
```

We need to find differentially expressed genes. You'll learn more about this later. For now, feel free to use some of my code. Start by downloading the limma package

```
## If using Windows, first go to https://cran.rstudio.com/bin/windows/Rtools/ and install the appropria
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

#uncomment to load limma for first run through

BiocManager::install("limma")
library(limma)
```

We will use the following code. Please add comments to every line to tell me what its doing!

```
## Subset our data for a training and test set
set.seed(2) #two random numbers generated for simulation
prob<-runif(ncol(expr)-2) #generates random deviates of the uniform distribution for expr
k<-which(prob>=0.3333333) #stores indexes that have a prob > 1/3
#Subsets out the GeneName column
eset<-expr[,2:ncol(expr)] # Explanation Below
eset<-eset[,k] #Stores the rows that meet the threshold of  prob > 1/3
rownames(eset)<-expr[,1] #take the row names from expr dataframe and put them in eset dataframe
design <- model.matrix(~0+as.factor(pheno$case[k])) #creating a design matrix and getting the independe
fit <- eBayes(lmFit(eset,design)) #fitting the model with the parkinsons labels
topTable(fit, coef=2) # 2 coef is the optimal coef
```

```
##               logFC   AveExpr         t       P.Value     adj.P.Val        B
## EXOC3L4     3.142306  1.471867  29.51102 4.199361e-149 4.759985e-145 329.8083
## FAM132A    -3.159535 -1.461491 -29.50594 4.606362e-149 4.759985e-145 329.7162
## MDM2        3.127701  1.426391  29.45147 1.239815e-148 8.541085e-145 328.7308
## CCR3       -3.145072 -1.547047 -29.41096 2.588618e-148 1.337474e-144 327.9981
## MYO9A      -3.172079 -1.571136 -29.27526 3.042541e-147 1.257604e-143 325.5456
## GADD45GIP1 -3.086818 -1.391784 -29.09131 8.553065e-146 2.946103e-142 322.2251
## ANXA2      -3.106348 -1.517455 -29.04409 2.011963e-145 5.940178e-142 321.3737
## CCNJ       -3.123721 -1.454289 -28.99841 4.602051e-145 1.188882e-141 320.5503
## EMC6        3.070413  1.590585  28.97850 6.599845e-145 1.515544e-141 320.1914
## GEMIN4      3.100428  1.427101  28.89106 3.211859e-144 6.637949e-141 318.6165
```

```
results<-topTable(fit, coef=2, number=Inf) # showing the inferential stats
```

Here, we have our gene names, our log fold change for expression, average expression, t statistic, pvalue, adjusted pvalue (for multiple testing!!), and the log odds of differential expression.

Next, we select those genes that have adjusted p-values below 0.001. Again, add comments to every line to describe what the code is doing.

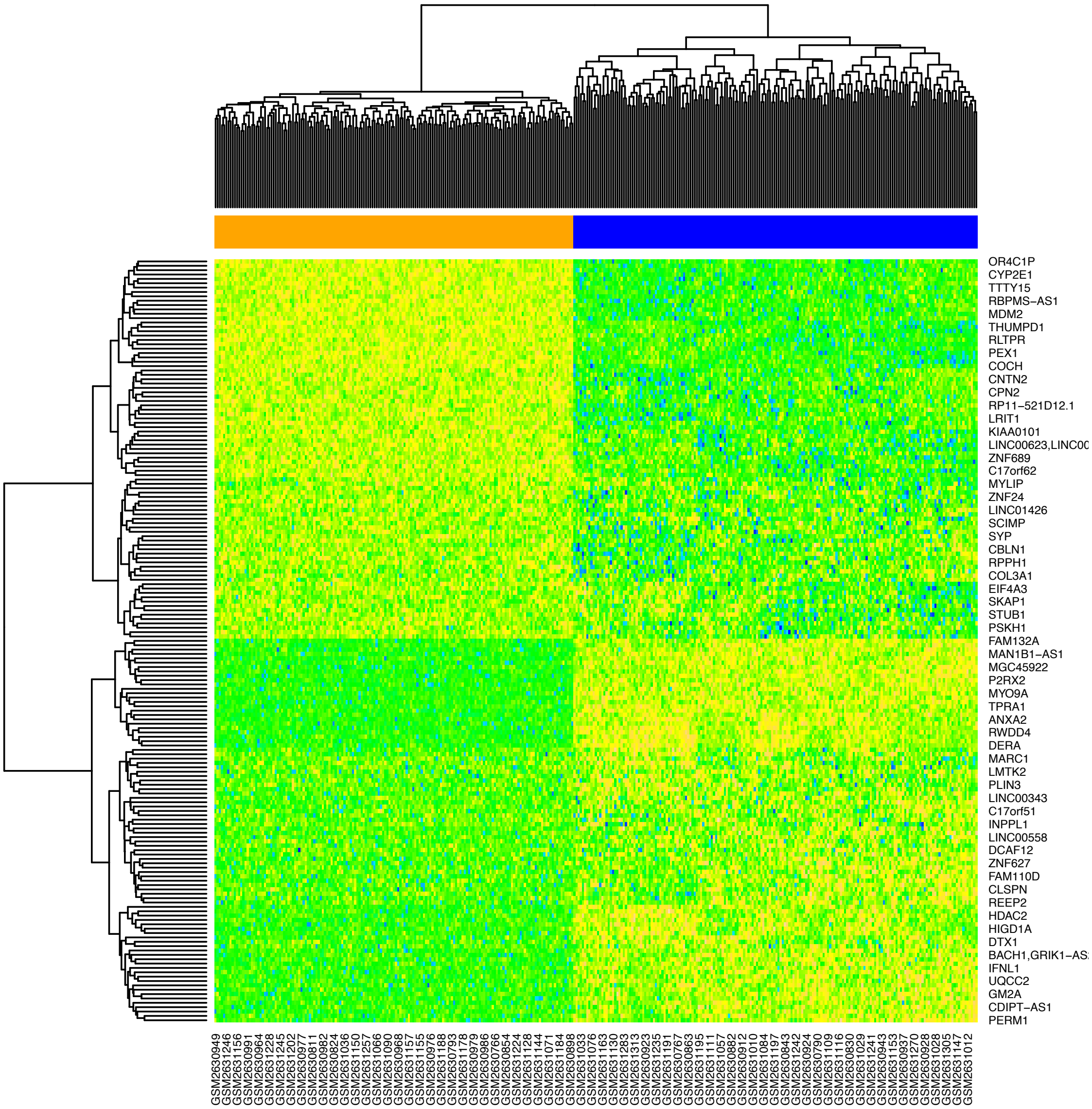
```
selected  <- row.names(results)[p.adjust(results$P.Value, method="fdr")<0.001]# gets the characters tha
direction <- sign(results$logFC) ## generates vector of numbers based on if the logFC has positve or ne
esetSel <- eset[selected, ] #storing the occurences of <.001 into esetSel
nrow(esetSel) # how many occurences of <.001
```

```
## [1] 175
```

Okay! So we're now looking at just 175 probes!

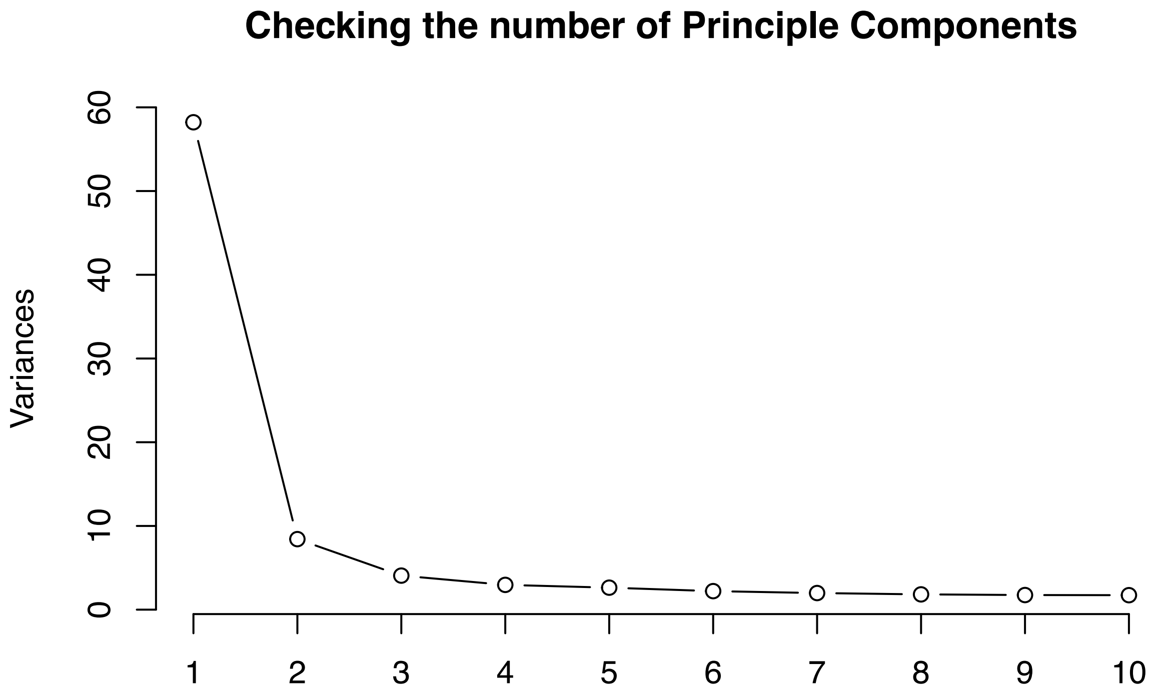
We are going to make a heat map here. I've provided the code, but try changing colours, labels, etc. to make it your own. You can try typing '?heatmap' into the console to see the help page and provide more ideas for what you'd like to change!

```
patientcolors <-ifelse(pheno$case[k]==1,"orange","blue")
heatmap(as.matrix(esetSel), col=topo.colors(100), ColSideColors=patientcolors, distfun = function(x) di
```


OR4C1P
CYP2E1
TTTY15
RBPMS−AS1
MDM2
THUMPD1
RLTPR
PEX1
COCH
CNTN2
CPN2
RP11−521D12.1
LRIT1
KIAA0101
LINC00623,LINC00
ZNF689
C17orf62
MYLIP
ZNF24
LINC01426
SCIMP
SYP
CBLN1
RPPH1
COL3A1
EIF4A3
SKAP1
STUB1
PSKH1
FAM132A
MAN1B1−AS1
MGC45922
P2RX2
MYO9A
TPRA1
ANXA2
RWDD4
DERA
MARC1
LMTK2
PLIN3
LINC00343
C17orf51
INPPL1
LINC00558
DCAF12
ZNF627
FAM110D
CLSPN
REEP2
HDAC2
HIGD1A
DTX1
BACH1,GRIK1−AS
IFNL1
UQCC2
GM2A
CDIPT−AS1
PERM1
GSM2630949
GSM2631246
GSM2631156
GSM2630991
GSM2630964
GSM2631228
GSM2631245
GSM2631202
GSM2630977
GSM2630811
GSM2630982
GSM2630824
GSM2631036
GSM2631150
GSM2631257
GSM2631066
GSM2631090
GSM2630968
GSM2631157
GSM2631155
GSM2630976
GSM2631188
GSM2630793
GSM2631178
GSM2630979
GSM2630986
GSM2630766
GSM2630854
GSM2631224
GSM2631128
GSM2631144
GSM2631071
GSM2631184
GSM2630898
GSM2631033
GSM2631076
GSM2631163
GSM2631130
GSM2631283
GSM2631313
GSM2630923
GSM2631235
GSM2631191
GSM2630767
GSM2630863
GSM2631195
GSM2631111
GSM2631057
GSM2630882
GSM2630912
GSM2631010
GSM2631084
GSM2631197
GSM2630843
GSM2631242
GSM2630924
GSM2630790
GSM2631109
GSM2631116
GSM2630830
GSM2631029
GSM2631241
GSM2630943
GSM2631153
GSM2630937
GSM2631270
GSM2630928
GSM2631028
GSM2631305
GSM2631147
GSM2631012

Notice the annotation bar along the top. This indicates PD vs not PD samples. This heat map is an example of a ‘non-supervised method’ - where we didn’t feed the labelled data to the algorithm. Instead, it is just clustering similar samples together. Because all of our PD samples cluster away from the non-PD samples, we are relatively certian we’ve picked good biomarkers! We should also check a PCA plot.

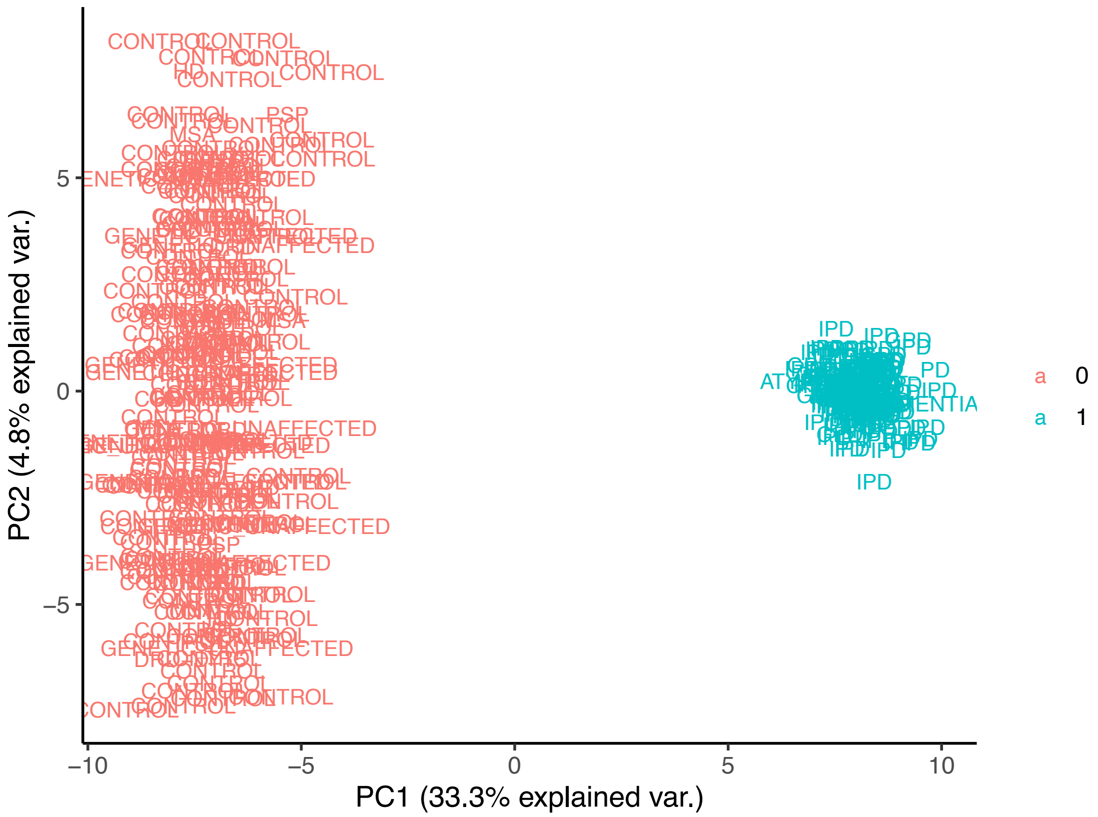
```
pc<-prcomp(t(esetSel),center=T,scale=T)
plot(pc,type="l",main="Checking the number of Principle Components")
```



Again, I’ve provided code for you here. Change it to something you like better!

```
#install.packages("devtools")
library(devtools)

#install.packages("ggpubr")
library(ggpubr)

#install_github("vqv/ggbiplot")
#library(ggbiplot)
source("ggbiplot.R")
g <- ggbiplot(pc, obs.scale = 1, var.scale = 1,
              groups = as.factor(pheno$case[k]), ellipse = F,
              circle = F, labels=pheno$disease_label[k],var.axes = F)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
```

```
g <- g + theme_classic2()
print(g)
```


PC2 (4.8% explained var.)
PC1 (33.3% explained var.)
5
0
−5
−10
−5
0
5
10
CONTROL
HD
PSP
MSA
GENETIC_UNAFFECTED
DRUG
IPD
PD
ATYPICAL
DEMENTIA
a 0
a 1

We have separation! Notice the obvious differences between cases and controls.

Make a variable which only contains the differential gene names and call it diffGenes AND print out all of these gene names using one line of code. The parentheses around the full line of code do this!

```
(diffGenes<-selected)
```

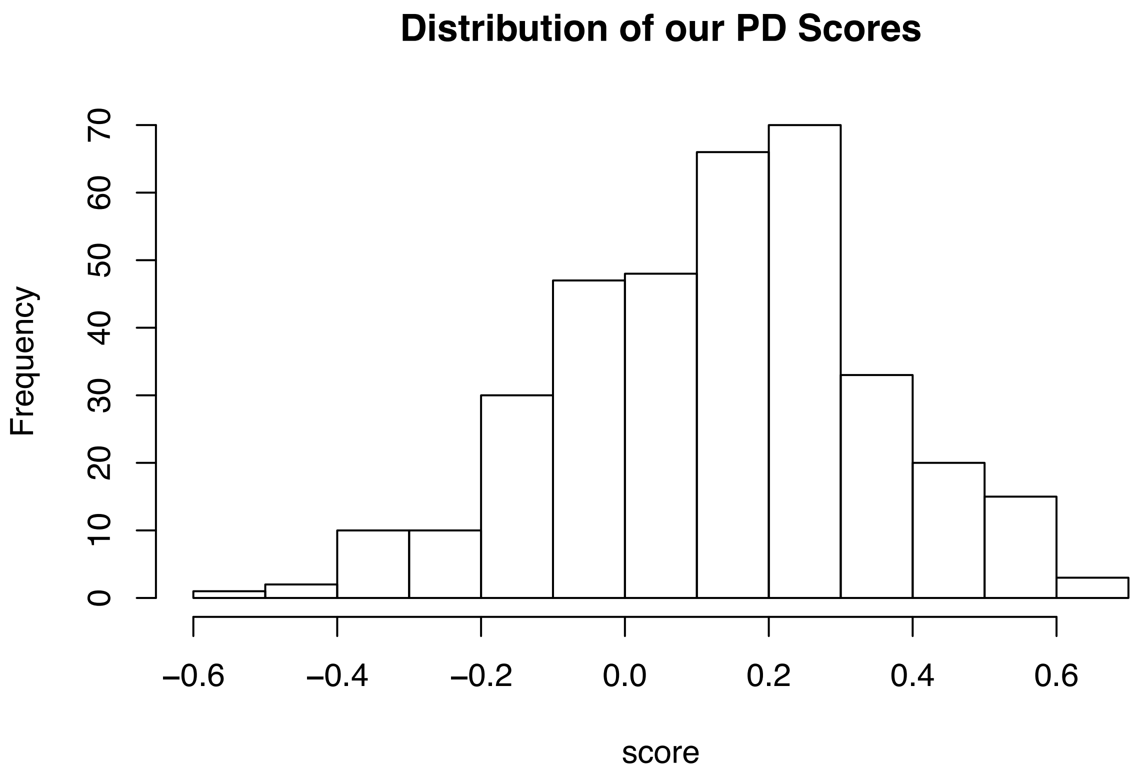
```
##   [1] "EXOC3L4"
##   [2] "FAM132A"
##   [3] "MDM2"
##   [4] "CCR3"
##   [5] "MYO9A"
##   [6] "GADD45GIP1"
##   [7] "ANXA2"
##   [8] "CCNJ"
##   [9] "EMC6"
##  [10] "GEMIN4"
##  [11] "PPM1K"
##  [12] "TTL"
##  [13] "DEFB103A,DEFB103B"
##  [14] "STX3"
##  [15] "DERA"
##  [16] "HIST1H2BO"
##  [17] "RGS4"
##  [18] "MAN1B1-AS1"
##  [19] "PSG5"
##  [20] "OR4C1P"
##  [21] "ABHD12B,MIR4454"
##  [22] "COCH"
##  [23] "RWDD4"
##  [24] "FAR1"
##  [25] "PEX1"
##  [26] "THUMPD1"
##  [27] "CTB-31O20.9"
##  [28] "GCLC"
##  [29] "SEC16B"
##  [30] "CYP2E1"
##  [31] "EGLN1"
##  [32] "PRKAG2"
##  [33] "NPC2"
##  [34] "TPRA1"
##  [35] "SLC26A4"
##  [36] "XAB2"
##  [37] "C10orf88"
##  [38] "MGC45922"
##  [39] "P2RX2"
##  [40] "AARS"
##  [41] "RNF157"
##  [42] "PSMC2"
##  [43] "RBPMS-AS1"
##  [44] "PCDHGB8P"
##  [45] "CD207"
##  [46] "RLTPR"
##  [47] "TTTY15"
##  [48] "TRAF5"
##  [49] "RP11-245J9.5"
##  [50] "KCNA10"
##  [51] "UQCC2"
##  [52] "RP11-324J3.1"
##  [53] "GABRA1"
##  [54] "CPN2"
##  [55] "MAP6"
##  [56] "C17orf62"
##  [57] "TRAFD1"
##  [58] "HIPK4"
##  [59] "GM2A"
##  [60] "N6AMT1"
##  [61] "RABGAP1L"
##  [62] "ANP32A"
##  [63] "ROBO2"
##  [64] "TOPORS-AS1"
##  [65] "STEAP2"
##  [66] "RNF167"
##  [67] "HDAC2"
##  [68] "ETFB"
##  [69] "RP11-521D12.1"
##  [70] "PHKA1"
##  [71] "TNS4"
##  [72] "EIF4A2"
##  [73] "ZNF689"
##  [74] "BACH1,GRIK1-AS2"
##  [75] "LRIT1"
##  [76] "KBTBD8"
##  [77] "B3GAT2"
##  [78] "DNM2"
##  [79] "DDIAS"
##  [80] "C2CD3"
##  [81] "CNTN2"
##  [82] "AP1S3"
##  [83] "CDIPT-AS1"
##  [84] "HIGD1A"
##  [85] "KIAA0101"
##  [86] "PERM1"
##  [87] "IFNL1"
##  [88] "CYP4Z1"
##  [89] "R3HDM4"
##  [90] "HMGCLL1"
##  [91] "RBM41"
##  [92] "RP11-108P20.4"
##  [93] "ARL6"
##  [94] "LINC00623,LINC00869,LINC01138,LOC103091866"
##  [95] "LINC00865"
##  [96] "ASMTL-AS1"
##  [97] "CASP14"
##  [98] "OR5J2"
##  [99] "DDX60L"
## [100] "ZDHHC24"
## [101] "MUC20"
## [102] "SYNPO"
## [103] "LAIR2"
## [104] "UCP3"
## [105] "REEP2"
## [106] "HDAC10"
## [107] "CBLN1"
## [108] "AP2M1"
## [109] "FOXN3-AS2"
## [110] "SYP"
## [111] "PPP6R2"
## [112] "CDH26"
## [113] "RPPH1"
## [114] "NT5DC3"
## [115] "ZNF627"
## [116] "STUB1"
## [117] "DTX1"
## [118] "CCDC136"
## [119] "FAM169A"
## [120] "LINC00558"
## [121] "CLCA2"
## [122] "GINM1"
## [123] "GHRHR"
## [124] "PKD2L2"
## [125] "RP11-742B18.1"
## [126] "LPGAT1"
## [127] "EIF4A3"
## [128] "CTD-2033C11.1"
## [129] "LEF1"
## [130] "LMTK2"
## [131] "A1BG"
## [132] "LINC00343"
## [133] "FAM110D"
## [134] "ADORA3"
## [135] "DKC1,MIR664B,SNORA56"
## [136] "BOP1,MIR7112"
## [137] "SCIMP"
## [138] "MAB21L1,MIR548F5"
## [139] "ZNF883"
## [140] "ZC3H14"
## [141] "PADI4"
## [142] "CLSPN"
## [143] "ZNF24"
## [144] "PLIN3"
## [145] "AURKC"
## [146] "RP11-320N7.2"
## [147] "FAM99B"
## [148] "LPCAT4"
## [149] "MPV17L"
## [150] "CD22"
## [151] "NEK11"
## [152] "MARC1"
## [153] "NR3C1"
## [154] "USO1"
## [155] "GJD4"
## [156] "RP11-21L23.2"
## [157] "LINC01426"
## [158] "STAT1"
## [159] "IGLC1,IGLV3-10,IGLV3-10"
## [160] "MRPL15"
## [161] "INPPL1"
## [162] "C17orf51"
## [163] "DCAF12"
## [164] "LINC00337"
## [165] "CYFIP2"
## [166] "LINC00927"
## [167] "ALK"
## [168] "SSX2,SSX2B,SSX3"
## [169] "ROCK2"
## [170] "MAGEC3"
## [171] "PSKH1"
## [172] "SKAP1"
## [173] "COL3A1"
## [174] "MYLIP"
## [175] "RP11-613M5.1"
```

To use these genes as a classifier, we will need to define a score function. Our score will be the sum of the average expression for the upregulated (positive) genes and the average for the down regulated (negative) genes. Here, I've written you a function which will do this. Please enter it and make comments to show you understand what its doing.

```
PDscore<-function(x,g,v,s){
  #x expression values for a sample
  #g all the genes
  #v the diffGenes
  #s is the sign of the logFC

  i<-which(g%in%v)   #Subset for diffGenes within the entire list of all the genes
  x<-x[i] # stores the value at the ith index into x (for expression values for a sample)
  s<-s[i] # stores the value at the ith index into s (for sign of LogFC)
  #Create vectors for genes with positive and negative momentum
  p<-c()
  n<-c()
  for(i in 1:length(x)){ # loop through the entire expression values
    if(s[i]>0){ #if the LogFC sign is positive than append it to the list p for positive
      p<-append(p,(x[i]))
    }
    else if(s[i]<0){ ## if the LogFC sign is negative than append it to the list n for negative
      n<-append(n,(x[i]))
    }
  }

  #If neither positive nor negative set to 0
  if(is.null(p)){p[1]=0}
  if(is.null(n)){n[1]=0}

  # the "score" is the differential of the mean of positive and negative
  score<-mean(p)-mean(n)
  return(score)
}
```

Now we can apply our function to our expression set to define a score for each patient. Comment what this is doing and why each step is necessary!

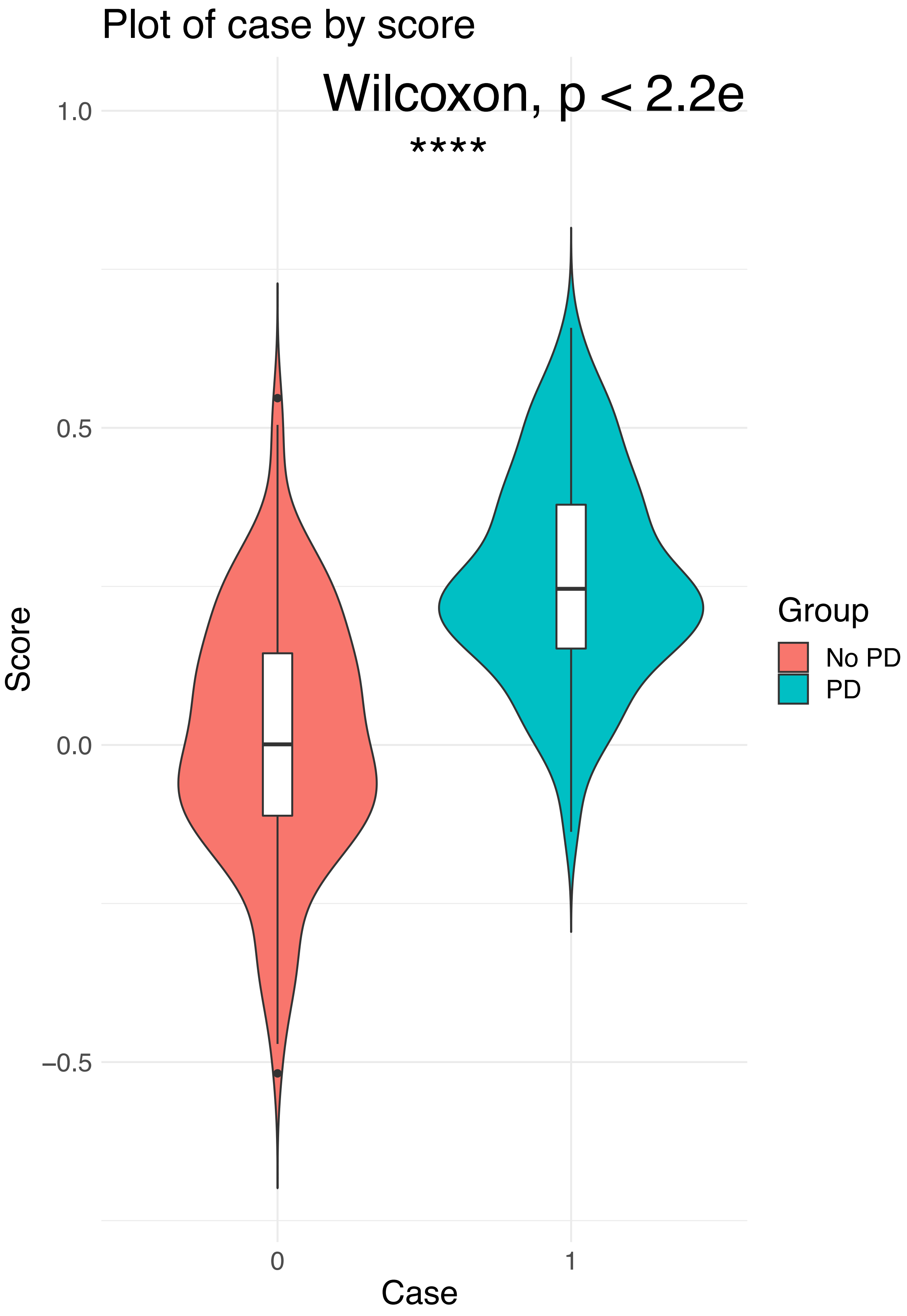
```
#Create vector
score<-c()


# create a vector of all genes in expr. Necessary for future steps to use gene stats to predict case ou
allGenes<-as.character(expr[as.character(expr$GeneName)%in%rownames(results),1])

#Apply our function to calculate scores
for(i in 1:ncol(eset)){
  score[i]<-PDscore(eset[,i],allGenes,diffGenes,direction)
}

#Generates histogram
hist(score,main="Distribution of our PD Scores")
```



Now we'll use ggplot to make and interpret a violin plot of our score. I've provided some code to do this, but try to change labels, colours, etc. to make it your own.

```
df<-data.frame(cbind(pheno$case[k],score))
```

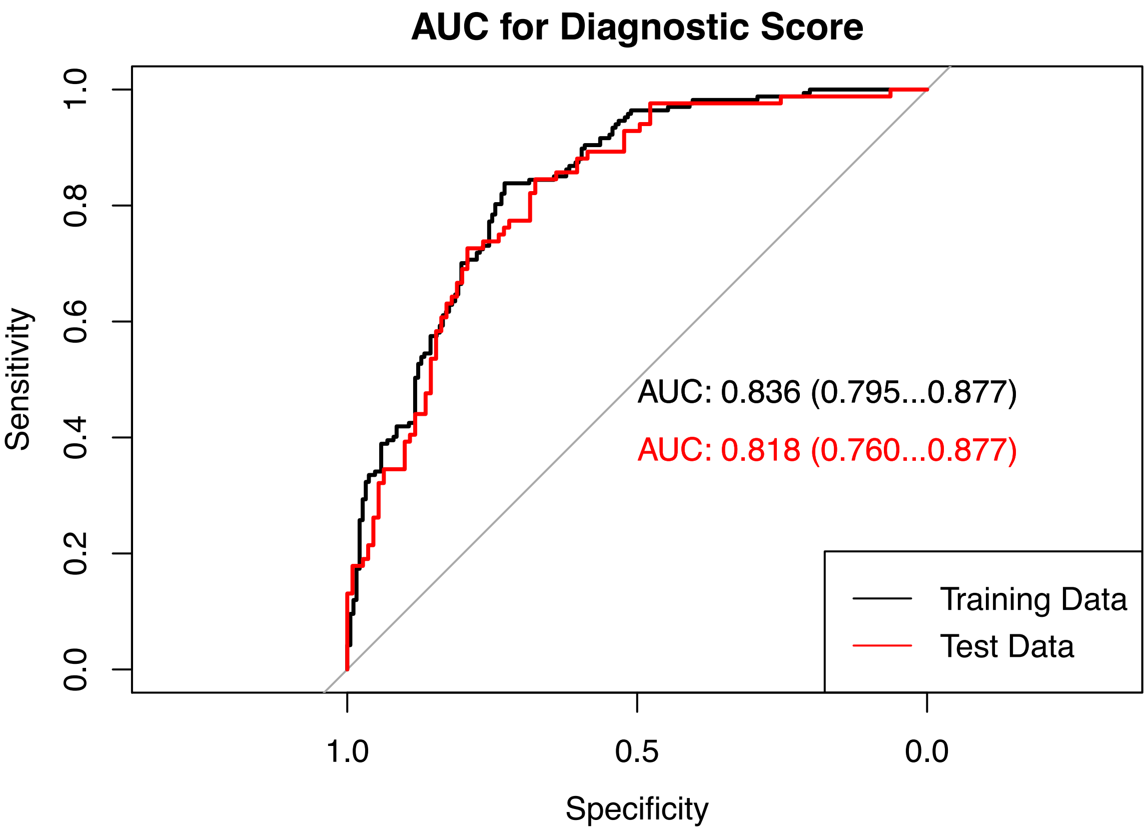
```
dp <- ggplot(df, aes(x=as.factor(pheno$case[k]), y=score, fill=as.factor(pheno$case[k]))) +
  geom_violin(trim=FALSE)+
  geom_boxplot(width=0.1, fill="white")+
  labs(title="Plot of case by score",x="Case ", y = "Score")+
  stat_compare_means(label.x = 1.5, label.y = 1, size=10)+
  stat_compare_means(aes(label = ..p.signif..),
                        label.x = 1.5, label.y = 0.9, size =10) + theme_minimal() +
  scale_fill_discrete(name = "Group", labels = c("No PD", "PD")) +
  theme(text = element_text(size = 18))

dp
```


Plot of case by score
Wilcoxon, p < 2.2e
****
1.0
0.5
0.0
−0.5
Score
0
1
Case
Group
No PD
PD

This shows not only the boxplot of our data, but also the distribution of our data points around the boxplot! As before, we can see that we don't have significant separation for our score, although we can see that the cases are trending to have a higher score. With more time and data cleaning we may be able to find something here!

Let's make an ROC plot, first with our training data, and then with our test data. As before, play with the plot options to make something you like! Note, there are MANY packages to build ROC plots, this one is just simple. Feel free to play with other packages to make publication ready plots if you'd like!

```
#install.packages("verification")
#install.packages("pROC")
library("pROC")
testEset<-expr[,2:ncol(expr)]
testEset<-testEset[,-k]
newScore<-apply(testEset,2,FUN=PDscore,allGenes,diffGenes,direction)
plot.roc(pheno$case[k]~score, data=df,legacy.axes=F,print.auc=T, ci=T, main="AUC for Diagnostic Score")
plot.roc(pheno$case[-k]~newScore,data=data.frame(cbind(pheno$case[-k],newScore)),add=T,print.auc=T, ci=
legend("bottomright",c("Training Data","Test Data"),lty=c(1,1),col=c("black","red"))
```



Notice that our score does better with our training data - this is expected! This is why we need to split our data, to avoid problems with over-fitting. These scores are better than random (the grey line), but we'd like to see an AUC as close to 1 as possible. Let's See if we can do better!

Statistics!

We can run a t-test to see if our score is significantly different between cases and controls. Try using the t.test function in R.

```
allScore<-c(score,newScore)
mergeCase<-c(pheno$case[k],pheno$case[-k]) ## to preserve order

## Do the t.test here

t.test(allScore ~ mergeCase)
```

```
##
##  Welch Two Sample t-test
##
## data:  allScore by mergeCase
## t = -15.842, df = 548, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2743704 -0.2138353
## sample estimates:
## mean in group 0 mean in group 1
##     0.005519575     0.249622456
```

The mean scores for our cases and controls are close, but they are significantly different with an extremely small p-value of 2.787e-13. This highlights a classical statistical fallacy - while small p-values are great, they are often meaningless without a large enough effect size. Here, we have achieved significance due to the large sample size of our study, hence our study is adequately powered.

We could also run a simple regression to examine the impact of the score on the log odds of being a case.

```
smallModel<-glm(pheno$case[k]~score, family=binomial)
summary(smallModel)
```

```
##
## Call:
## glm(formula = pheno$case[k] ~ score, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5109  -0.7565  -0.2132   0.8773   2.1995
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2373     0.1862  -6.646 3.02e-11 ***
## score         7.9472     0.8874   8.955  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 490.89  on 354  degrees of freedom
## Residual deviance: 350.94  on 353  degrees of freedom
## AIC: 354.94
##
## Number of Fisher Scoring iterations: 5
```

Summarize this output!

Again, we conclude that the score is a statistically significant indicator of the odds of having PD. Let's build a larger model which examines other phenotype variables.

First, build a data frame which includes all the model data we're interested in. Start with the age variables in your pheno set, and then use the cbind() function to add on our scores and the binary case vector. Print a summary of the model data.

```
## Enter your own code here

modelData <- data.frame(cbind(age_at_exam, age_at_symptoms, allScore, mergeCase ))
modelData
```

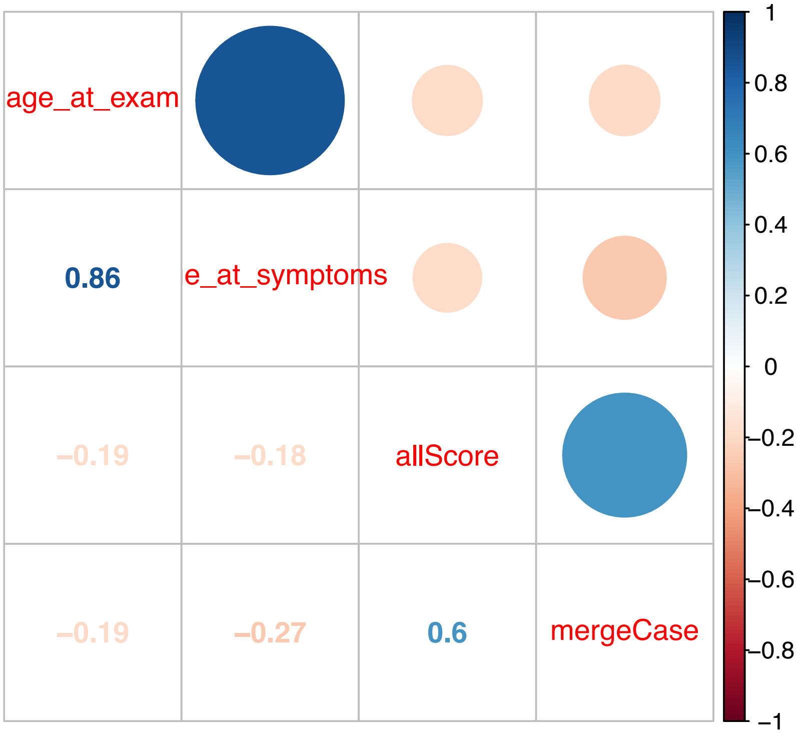
```
##            age_at_exam age_at_symptoms      allScore mergeCase
## X                   NA              53 -0.1347098260         1
## X.1                 NA              64  0.1521432297         1
## X.2                 NA              NA  0.1034828276         0
## X.3                 NA              60 -0.0596159956         0
## X.4                 NA              66  0.2008197232         0
## X.5                 NA              41  0.1159874423         0
## X.6                 NA              43 -0.1716217278         0
## X.7                 40              NA  0.3145752497         0
## X.8                 42              NA -0.0783260787         0
## X.9                 43              NA  0.3160259096         0
## X.10                49              NA  0.2077043555         0
## X.11                40              NA  0.1054308479         0
## X.12                NA              NA  0.0388705862         0
## X.13                NA              NA -0.1137639583         0
## X.14                NA              NA -0.3133639163         0
## X.15                NA              NA -0.3253562144         0
## X.16                NA              NA  0.2621020054         0
## X.17                NA              NA -0.0677631764         0
## X.18                NA              NA -0.1049747583         0
## X.19                NA              NA  0.3126505620         0
## X.20                NA              NA  0.1711543562         0
## X.21                NA              NA  0.1299140878         0
## X.22                NA              NA -0.1625402453         0
## X.23                NA              NA -0.0001492302         0
## X.24                NA              NA  0.1440582062         0
## X.25                NA              NA  0.2269270956         0
## X.26                NA              NA  0.2834992804         0
## X.27                NA              NA -0.0734041713         0
## X.28                NA              NA  0.2290789771         0
## X.29                NA              NA -0.2087020865         0
## X.30                NA              NA -0.3854779012         0
## X.31                NA              NA -0.0848262372         0
## X.32                NA              NA -0.0960736661         0
## X.33                NA              NA  0.2615004252         0
## X.34                NA              NA  0.1353291115         0
## X.35                NA              NA -0.2326583465         0
## X.36                NA              NA -0.2099056367          0
## X.37                NA              NA  0.2205223646          0
## X.38                NA              NA -0.0605756069          0
## X.39                NA              NA -0.0207499227          0
## X.40                NA              NA -0.0142608639          0
## X.41                NA              NA -0.0021409175          0
## X.42                NA              NA  0.2607879538          0
## X.43                NA              NA  0.0834096624          0
## X.44                NA              NA  0.0686908765          0
## X.45                NA              NA  0.1759604057          0
## X.46                NA              NA -0.4334295525          0
## X.47                NA              NA -0.0825491905          0
## X.48                NA              NA -0.0973115985          0
## X.49                NA              NA  0.2189929066          0
## X.50                NA              NA  0.5046861931          0
## X.51                NA              NA -0.2715498546          0
## X.52                NA              NA -0.1108084334          0
## X.53                NA              NA -0.0175931857          0
## X.54                NA              NA  0.1213403250          0
## X.55                NA              NA  0.1760353223          0
## X.56                NA              NA -0.1352559821          0
## X.57                NA              NA -0.0863252329          0
## X.58                NA              NA  0.2269926177          0
## X.59                NA              NA -0.0904286923          0
## X.60                NA              NA  0.0073166342          0
## X.61                NA              NA  0.1191124525          0
## X.62                NA              NA  0.1460380005          0
## X.63                NA              NA -0.2058909418          0
## X.64                NA              NA  0.2144559649          0
## X.65                NA              NA  0.1374674165          0
## X.66                NA              NA -0.1305490226          0
## X.67                NA              NA -0.0502469710          0
## X.68                NA              NA  0.4213617050          0
## X.69                NA              NA -0.0160066746          0
## X.70                NA              NA  0.0731164951          0
## X.71                NA              NA -0.1353828390          0
## X.72                NA              NA -0.1952449189          0
## X.73                NA              NA  0.0020716362          0
## X.74                NA              NA  0.1601396132          0
## X.75                NA              NA -0.1962461422          0
## X.76                NA              NA  0.1259466914          0
## X.77                NA              NA -0.0337047700          0
## X.78                NA              NA  0.0475631840          0
## X.79                68              NA -0.0691503670          0
## X.80                82              NA  0.1639790493          0
## X.81                63              NA -0.3291049779          0
## X.82                70              NA  0.0235955686          0
## X.83                49              NA -0.0664110238          0
## X.84                58              NA -0.3697818244          0
## X.85                65              NA -0.3608025295          0
## X.86                78              NA -0.3191115091          0
## X.87                42              NA -0.3108080449          0
## X.88                67              NA  0.0961788529          0
## X.89                69              NA -0.0021726334          0
## GSM2630927          49              46 -0.0522823732         0
## GSM2631227          NA              46 -0.1467890637         0
## GSM2631238          NA              46  0.3225404641         0
## GSM2630771          50              46 -0.3682175723         0
## GSM2630857          49              47 -0.1639329943         0
## GSM2630868          52              47  0.4116132618         0
## GSM2630818          50              47 -0.0608316878         0
## GSM2630929          54              47 -0.1272356228         0
## GSM2630932          51              47 -0.0845866511         0
## GSM2631230          53              49  0.0032341971         0
## GSM2631232          57              49  0.3899492115         0
## GSM2631234          69              50  0.1273560475         0
## GSM2630906          53              50  0.3872403761         0
## GSM2630917          58              50  0.3127298280         0
## GSM2631298          59              50  0.0257254438         0
## GSM2631158          52              51 -0.0597313257         0
## GSM2631167          60              51 -0.0120072012         0
## GSM2631175          63              51 -0.2229541259         0
## GSM2631176          NA              51  0.0556446776         0
## GSM2631177          58              51  0.0281937628         0
## GSM2630777          NA              52 -0.3812652966         0
## GSM2630788          58              52 -0.0931504744         0
## GSM2631173          59              53 -0.0207580150         0
## GSM2630787          54              53 -0.3740003573         0
## GSM2631100          61              53 -0.1098869820         0
## GSM2631118          NA              54  0.0182209788         0
## GSM2631039          60              54  0.0342952062         0
## GSM2631024          64              54 -0.1815996949         0
## GSM2631275          NA              55  0.3135896765         0
## GSM2631061          NA              55  0.2616320367         0
## GSM2631027          NA              55 -0.0554127774         0
## GSM2630980          NA              55  0.3108372458         0
## GSM2631105          56              55 -0.2570759631         0
## GSM2631032          NA              55  0.1758353429         0
## GSM2631082          58              55  0.0491639967         0
## GSM2631131          NA              55 -0.0824422320         0
## GSM2631047          60              55 -0.0422465024         0
## GSM2631180          65              55 -0.1377294518         0
## GSM2631038          NA              56  0.0999457968         0
## GSM2631059          59              56 -0.1383420739         0
## GSM2631085          59              56  0.3021180006         0
## GSM2630761          59              56 -0.0748953939         0
## GSM2631011          64              56 -0.0903771134         0
## GSM2631137          67              57  0.0538060543         0
## GSM2631277          73              57  0.0602106297         0
## GSM2631072          61              57 -0.0939756419         0
## GSM2631026          62              57  0.0998868259         0
## GSM2631284          62              57  0.0630271069         0
## GSM2631274          71              58  0.1136231806         0
## GSM2630774          71              58 -0.1640405477         0
## GSM2630993          63              58 -0.2129823439         0
## GSM2631112          67              58  0.0362726719         0
## GSM2631142          60              58  0.0647821034         0
## GSM2630984          71              58  0.1349552299         0
## GSM2631083          67              58  0.1358762580         0
## GSM2631068          NA              59 -0.0590313000         0
## GSM2631272          64              59  0.2241232311         0
## GSM2631123          65              59 -0.2695422786         0
## GSM2630855          71              60 -0.2321818066         0
## GSM2630758          64              60 -0.1726227631         0
## GSM2630866          80              60 -0.2094973202         0
## GSM2630840          63              60  0.0192708477         0
## GSM2630778          62              60 -0.1123960834         0
## GSM2630876          63              60  0.1347894708         0
## GSM2630965          NA              60 -0.3182871966         0
## GSM2630958          61              60  0.0202881196         0
## GSM2631044          77              60 -0.2164530381         0
## GSM2631080          66              60 -0.3462555919         0
## GSM2631179          63              61  0.2444754976         0
## GSM2631181          74              61 -0.0535447428         0
## GSM2631132          65              61 -0.0323790327         0
## GSM2631122          66              62 -0.0444032947         0
## GSM2631058          70              62  0.1472488193         0
## GSM2631081          64              62  0.3407157285         0
## GSM2631140          NA              62  0.0864484831         0
## GSM2631045          66              62  0.0073262075         0
## GSM2630812          67              63  0.1645513140         0
## GSM2630816          66              63 -0.1324290947         0
## GSM2630873          63              63 -0.0638843470         0
## GSM2630874          78              63  0.1110271307         0
## GSM2630803          NA              64  0.2043994103         0
## GSM2630772          NA              64  0.1289805806         0
## GSM2630875          NA              64 -0.1921792326         0
## GSM2631303          70              64  0.2524339302         0
## GSM2631222          69              64  0.1626319108         0
## GSM2630935          77              64 -0.4060597178         0
## GSM2631250          NA              64  0.0967163831         0
## GSM2630936          NA              65  0.1594038858         0
## GSM2630926          NA              65 -0.3816749088         0
## GSM2630833          67              65  0.0294311589         0
## GSM2630920          74              65  0.0255902903         0
## GSM2630844          70              65 -0.6118184477         0
## GSM2631217          67              65 -0.1690363581         0
## GSM2631226          69              65  0.4014600534         1
## GSM2631193          67              65  0.1813764093         1
## GSM2631282          NA              66 -0.0246015592         1
## GSM2631063          69              66  0.3264019816         1
## GSM2631141          71              66  0.0420676727         1
## GSM2630946          NA              67  0.0386019162         1
## GSM2630809          71              67  0.3809984358         1
## GSM2631292          72              67  0.2501502113         1
## GSM2630887          71              67  0.0775886379         1
## GSM2631251          70              67  0.2012195917         1
## GSM2630782          71              67  0.1620374371         1
## GSM2630858          69              67  0.4540467256         1
## GSM2630859          70              68  0.4829596283         1
## GSM2630852          NA              68  0.5219211804         1
## GSM2631139          NA              68 -0.1669816927         0
## GSM2630952          69              68 -0.1193681318         0
## GSM2631086          78              68  0.3593611172         0
## GSM2630962          69              68  0.0757851007         0
## GSM2631034          NA              69 -0.0362659932         0
## GSM2631092          72              69 -0.1461183888         0
## GSM2631286          74              69  0.2439021957         0
## GSM2631113          75              70  0.1200285112         0
## GSM2631093          71              70 -0.0703832811         0
## GSM2630989          70              70  0.4284944864         1
## GSM2631127          73              70  0.2018891315         1
## GSM2631258          76              72  0.3873551052         1
## GSM2630802          74              72  0.0704288303         1
## GSM2630983          NA              72  0.3162031378         1
## GSM2630914          75              73  0.2071276910         1
## GSM2630997          NA              74  0.2971352645         1
## GSM2630921          NA              75  0.1123782133         1
## GSM2630992          NA              76  0.3446099937         1
## GSM2630823          NA              76  0.2637098838         1
## GSM2631050          78              78  0.4122471406         1
## GSM2630994          NA              78  0.1968760655         1
## GSM2631021          NA              NA  0.1856891642         1
## GSM2630825          NA              NA -0.1368355207         1
## GSM2630821          NA              NA  0.3325432561         1
## GSM2631297          NA              NA  0.2087770486         1
## GSM2631114          NA              NA  0.4229274243         1
## GSM2631117          NA              NA  0.1731362552         1
## GSM2631260          NA              NA  0.1675814527         1
## GSM2631096          NA              NA  0.2129642554         1
## GSM2631003          NA              NA  0.2228625058         1
## GSM2631095          NA              NA  0.0146353863         1
## GSM2630940          NA              NA  0.2491197581         1
## GSM2630805          NA              NA  0.2363934829         1
## GSM2631009          NA              NA  0.0128259099         1
## GSM2630944          NA              NA  0.0724033832         1
## GSM2630947          NA              NA -0.3523216244         1
## GSM2630892          NA              NA  0.4874447599         1
## GSM2630973          NA              NA  0.5383556380         1
## GSM2630828          NA              NA  0.3688085064         1
## GSM2631192          NA              NA  0.6262424973         1
## GSM2630822          NA              NA  0.0592759754         1
## GSM2630969          NA              NA  0.5034897321         1
## GSM2631296          NA              NA  0.1295880360         1
## GSM2630956          NA              NA  0.0785098765         1
## GSM2630959          NA              NA  0.0802919119         1
## GSM2631291          NA              NA  0.0116655885         1
## GSM2630981          NA              NA  0.0802672895         1
## GSM2631067          NA              NA  0.1490238129         1
## GSM2631107          NA              NA  0.2600287801         1
## GSM2631186          NA              NA  0.3400259568         1
## GSM2630942          NA              NA  0.2002346464         1
## GSM2630867          NA              NA  0.1083356794         1
## GSM2631266          NA              NA  0.3190021207         1
## GSM2630763          NA              NA  0.3409741983         1
## GSM2631267          NA              NA  0.4056896209         1
## GSM2631004          NA              NA  0.3515907890         1
## GSM2631104          NA              NA  0.4513433797         1
## GSM2630966          NA              NA  0.0324209020         1
## GSM2631269          NA              NA  0.1468274640         1
## GSM2631013          NA              NA  0.1451585074         1
## GSM2630995          NA              NA  0.1642121476         1
## GSM2631106          60              NA  0.2759626014         1
## GSM2631187          NA              NA  0.4110372889         1
## GSM2631261          NA              NA -0.0208964163         1
## GSM2631078          NA              51  0.3397995067         1
## GSM2631060          66              60  0.2339563384         1
## GSM2630762          NA              64  0.3560714020         1
## GSM2630838          NA              68  0.3017586700         1
## GSM2630900          75              71 -0.0246343182         1
## GSM2631244          76              72  0.2478615815         1
## GSM2630870          81              74  0.1617287673         1
## GSM2630862          NA              NA  0.2109527968         1
## GSM2631169          NA              69  0.1448082916         1
## GSM2631168          NA              NA  0.2337505989         1
## GSM2631170          63              52 -0.0018886159         1
## GSM2631019          NA              53  0.1181109146         1
## GSM2630765          76              66  0.1547808579         1
## GSM2630814          NA              72 -0.3945095738         0
## GSM2631199          NA              72 -0.0681534488         0
## GSM2631233          NA              NA  0.3345370758         1
## GSM2631216          NA              NA -0.1650202314         0
## GSM2630889          NA              NA -0.0844095755         0
## GSM2630792          NA              NA  0.3939263236         0
```

```
summary(modelData)
```

```
##   age_at_exam    age_at_symptoms    allScore         mergeCase
##  Min.   :30.00   Min.   :10.00   Min.   :-0.6118   Min.   :0.0000
##  1st Qu.:54.75   1st Qu.:45.00   1st Qu.:-0.0334   1st Qu.:0.0000
##  Median :61.00   Median :55.00   Median : 0.1298   Median :0.0000
##  Mean   :60.56   Mean   :53.61   Mean   : 0.1169   Mean   :0.4564
##  3rd Qu.:68.25   3rd Qu.:64.00   3rd Qu.: 0.2598   3rd Qu.:1.0000
##  Max.   :82.00   Max.   :78.00   Max.   : 0.6577   Max.   :1.0000
##  NA's   :266     NA's   :325
```

We should examine the correlations in our data set. You can do this quickly by building a correlation plot matrix.

```
install.packages("corrplot")
library(corrplot)
M<-cor(na.omit(modelData))
corrplot.mixed(M)
```


age_at_exam
e_at_symptoms
allScore
mergeCase
0.86
−0.19
−0.18
−0.19
−0.27
0.6
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

How would you interpret this output? Write a few sentences below!

- age_at_exam and age_at_symptoms is highly correlated to eachother
- mergeCase and allScore are also very correlated

Let’s build our first model. Here, we consider the case as our dependent variable, and the others as our explanatory variables.

```
model1<-glm(mergeCase~.,family=binomial,data=modelData)
```

This error is important - it represetns that our model is drastically overfit. We can easily fix this using the BayesGLM model from the arm package

```
#install.packages("arm")
library(arm)
model1<-bayesglm(mergeCase~.,family=binomial,data=modelData)
summary(model1)
```

```
##
## Call:
## bayesglm(formula = mergeCase ~ ., family = binomial, data = modelData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1132  -0.6876  -0.2873   0.7161   2.1547
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.16206    1.23382  -0.131   0.8955
## age_at_exam      0.04603    0.03531   1.304   0.1923
## age_at_symptoms -0.07472    0.03125  -2.391   0.0168 *
## allScore         8.25921    1.33364   6.193  5.9e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 215.96  on 157  degrees of freedom
## Residual deviance: 138.16  on 154  degrees of freedom
##   (392 observations deleted due to missingness)
## AIC: 146.16
##
## Number of Fisher Scoring iterations: 9
```

We cannot use the step function for bayes glm. We will iteratively remove variables with the highest p-values, and then rerun the model.

Try this on your own, removing one by one and checking the output to find the next largest p-value. OR if you feel up to the challenge, write your own function to automate this process for you! There are bonus points available ;)

```r
## Enter your own code here
model<-bayesglm(mergeCase~.,family=binomial,data=modelData)
model2 <- step(model)
```

```
## Start:  AIC=146.16
## mergeCase ~ age_at_exam + age_at_symptoms + allScore
##
##                    Df Deviance    AIC
## <none>                  138.16 146.16
## - age_at_exam      1   140.56 146.56
## - age_at_symptoms  1   145.36 151.36
## - allScore         1   203.64 209.64
```

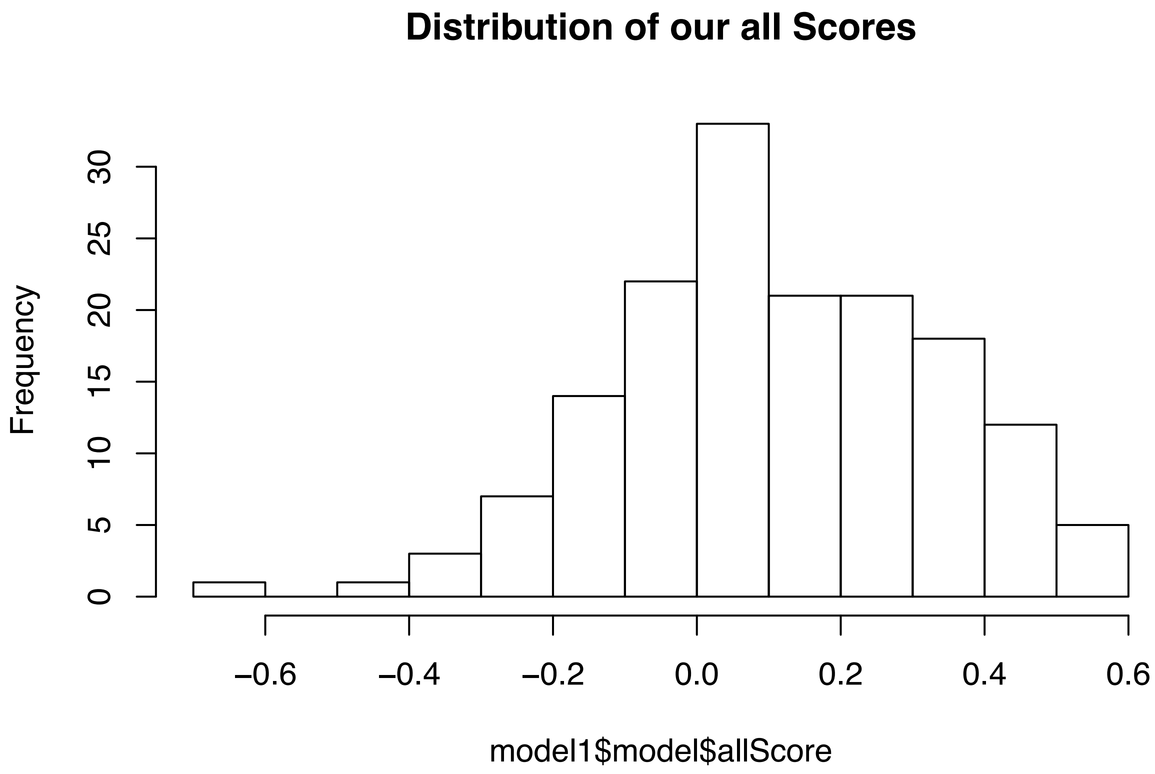
```r
summary(model2)
```

```
##
## Call:
## bayesglm(formula = mergeCase ~ age_at_exam + age_at_symptoms +
##     allScore, family = binomial, data = modelData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1132  -0.6876  -0.2873   0.7161   2.1547
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.16206    1.23382  -0.131   0.8955
## age_at_exam      0.04603    0.03531   1.304   0.1923
## age_at_symptoms -0.07472    0.03125  -2.391   0.0168 *
## allScore         8.25921    1.33364   6.193  5.9e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 215.96  on 157  degrees of freedom
## Residual deviance: 138.16  on 154  degrees of freedom
##   (392 observations deleted due to missingness)
## AIC: 146.16
##
## Number of Fisher Scoring iterations: 9
```

This is our final model! Notice that our largest effect size is controlled by our genetic score. At a first glance, we might assume this means that the score has the largest effect on the model. However, if we recall how to interpret our coefficients, the estimated effect size is the change in log odds of being a case for a 1 unit increase in our score. Think about the score distribution: the range of our scores is fairly small. In contrast, the range of the updrs scores varies from 0 to 36. Keep in mind the scale of our data when interpretting these models!

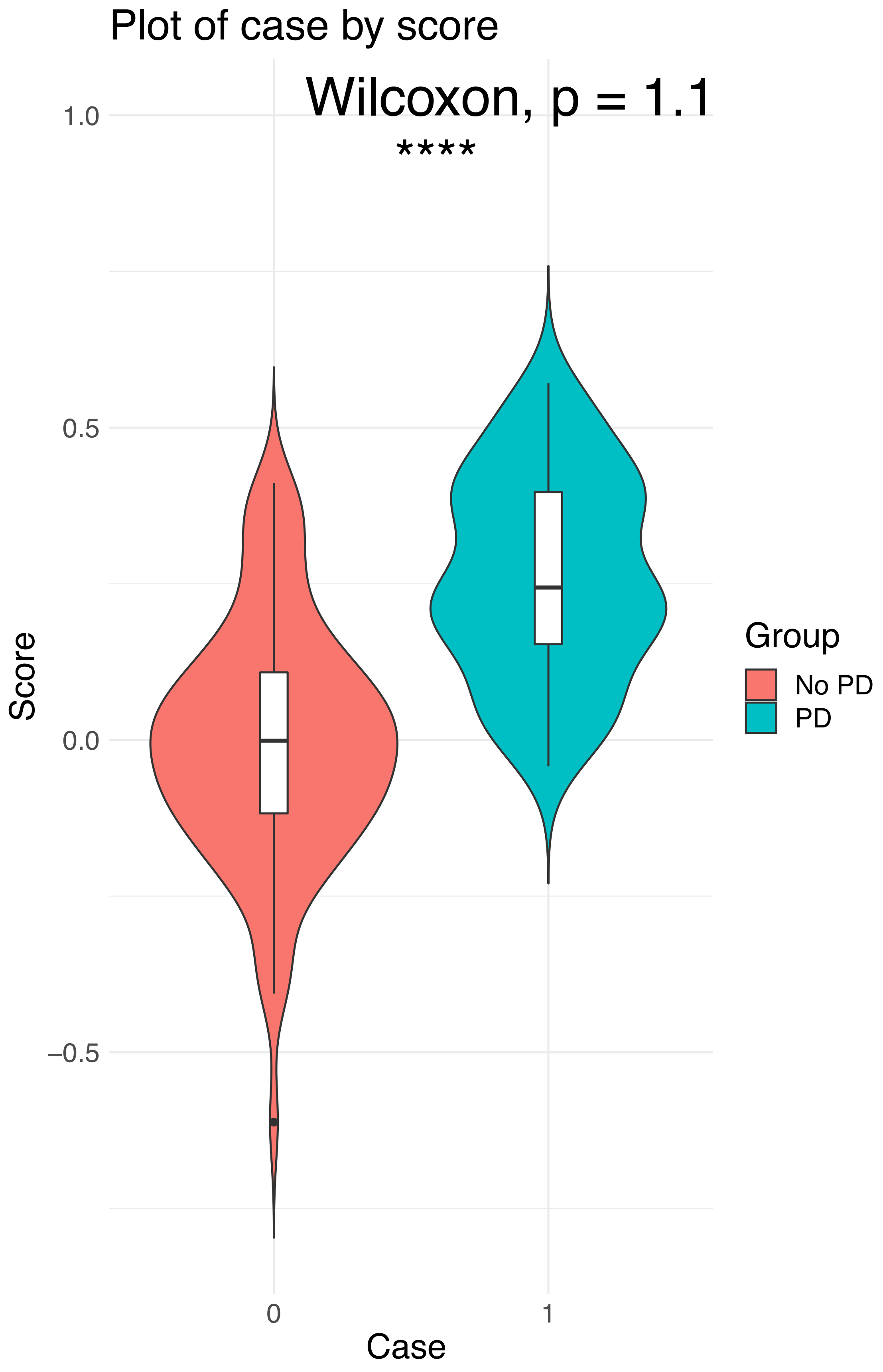
Let's predict the probability of having a case given our model. Make a histogram of the score from this model.

```
## Enter your own code here
hist(model1$model$allScore,main="Distribution of our all Scores")
```



Like before, we'll build a violin plot to compare the output of our regression model. See if you can adapt the violin plot code from before to do this now.

```
## Enter your own code here

df<-data.frame(cbind(model1$model$mergeCase,model1$model$allScore))

dp <- ggplot(df, aes(x=as.factor(model1$model$mergeCase), y=model1$model$allScore, fill=as.factor(model
  geom_violin(trim=FALSE)+
  geom_boxplot(width=0.1, fill="white")+
  labs(title="Plot of case by score",x="Case ", y = "Score")+
  stat_compare_means(label.x = 1.5, label.y = 1, size=10)+
  stat_compare_means(aes(label = ..p.signif..),
                     label.x = 1.5, label.y = 0.9, size =10) + theme_minimal() +
  scale_fill_discrete(name = "Group", labels = c("No PD", "PD")) +
  theme(text = element_text(size = 18))

dp
```


Plot of case by score
Wilcoxon, p = 1.1
****
1.0
0.5
0.0
−0.5
Score
0
1
Case
Group
No PD
PD

Now we're starting to see a clearer separation of scores! It's clear that by including the established tests to pre-screen patients for PD and other neurological diseases we have improved overall performance. While this may be an obvious conclusion, it is worth noting that the context with which our diagnostic signature would be used would be on patients already exhibiting potential PD symptoms. Clearly this needs a little more work, but for a first pass at assessing raw data, it's not bad!

Again, we can examine ROC curves. I've done some of the set up to get the data in the right format. Use the ROC code above to then build your own plot!

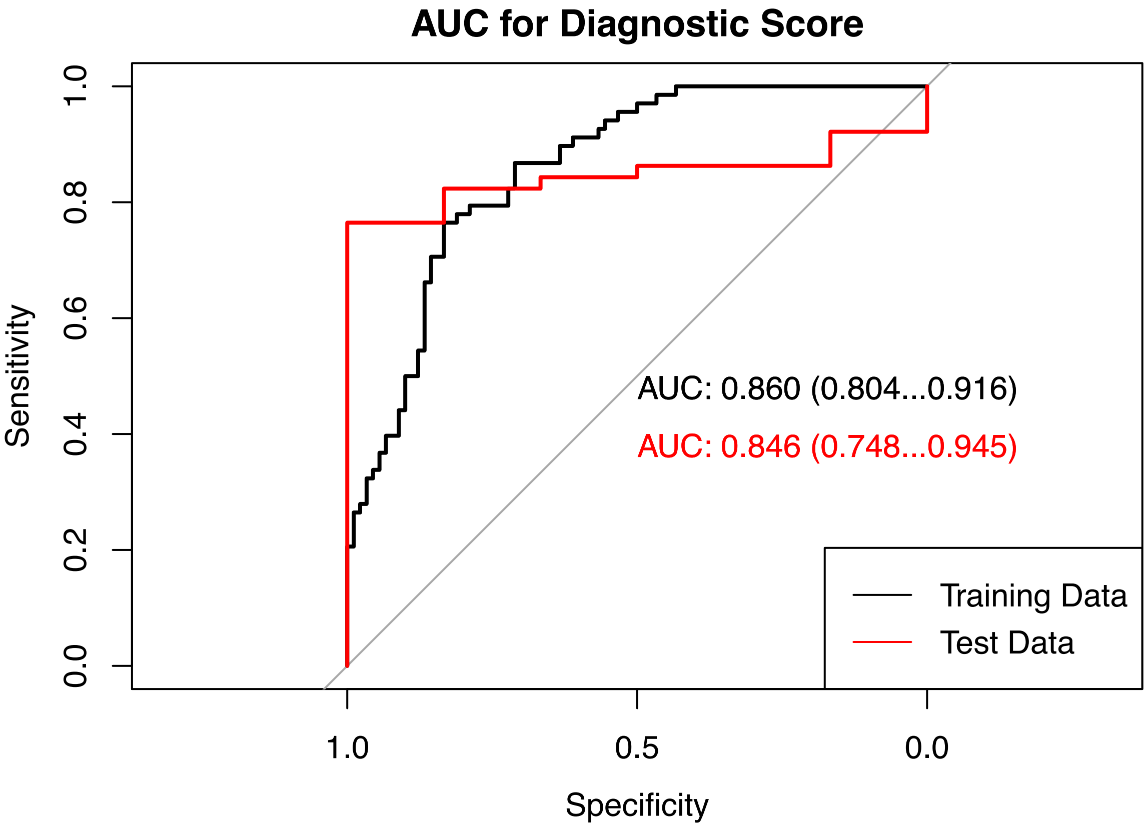
```
library("pROC")
nd<-cbind(pheno[-k,4:ncol(pheno)],newScore)
colnames(nd)<-c(colnames(modelData[1:ncol(modelData)-1]),"score")
newMScore<-predict(model2,newdata=nd)

## Enter your own code here
plot.roc(model1$model$mergeCase ~ model1$model$allScore, data=df, legacyaxes=F,print.auc=T, ci=T, main=
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
plot.roc(pheno$case[-k] ~ newMScore, data=data.frame(cbind(pheno$case[-k],newMScore)), add=T,print.auc=
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
legend("bottomright",c("Training Data","Test Data"),lty=c(1,1),col=c("black","red"))
```



Here, we have a notable increase in AUC, particularly for our training data. Our test data shows an overal improvement as well, although with a large confidence interval. There are clearly some data points in here which are abnormal - and perhaps worth investigating.

##Congratulations, you have finished the R Bootcamp Assignment!