Kallisto V3 Sex Prediction

```
knitr::opts_knit$set(root.dir = "D:/PBTA")

setwd("D:/PBTA")

ge <- readRDS("pbta-gene-expression-kallisto_v3.rds")

colnames(ge)[1:10]

##  [1] "transcript_id" "gene_id"       "BS_ZF6BSFNF"   "BS_H97S5SQN"
##  [5] "BS_0ZA67BBC"   "BS_W7MFJZ5A"   "BS_68KX6A42"   "BS_NB9XXBW6"
##  [9] "BS_1N7MQZGR"   "BS_2JP7RBMB"

ge[1:10,1:10]

##                    transcript_id                    gene_id BS_ZF6BSFNF
## 1     ENST00000000233.9_ARF5-201    ENSG00000004059.10_ARF5  109.299000
## 2     ENST00000000412.7_M6PR-201     ENSG00000003056.7_M6PR   23.189400
## 3   ENST00000000442.10_ESRRA-201   ENSG00000173153.13_ESRRA   12.912400
## 4    ENST00000001008.5_FKBP4-201    ENSG00000004478.7_FKBP4   33.004500
## 5  ENST00000001146.6_CYP26B1-201  ENSG00000003137.8_CYP26B1    3.761880
## 6  ENST00000002125.8_NDUFAF7-201 ENSG00000003509.15_NDUFAF7    3.407500
## 7   ENST00000002165.10_FUCA2-201   ENSG00000001036.13_FUCA2   19.489400
## 8  ENST00000002501.10_DBNDD1-201  ENSG00000003249.13_DBNDD1   18.193000
## 9   ENST00000002596.5_HS3ST1-201   ENSG00000002587.9_HS3ST1    4.468570
## 10  ENST00000002829.7_SEMA3F-201  ENSG00000001617.11_SEMA3F    0.233314
##    BS_H97S5SQN BS_0ZA67BBC BS_W7MFJZ5A BS_68KX6A42 BS_NB9XXBW6 BS_1N7MQZGR
## 1  124.8580000   96.839400  131.467000  180.688000  106.569000  112.266000
## 2   27.3794000   20.560800   47.567300   44.889600   29.834300   40.906400
## 3   12.0704000   14.063200   18.476000   13.663400    8.237950   16.658100
## 4   41.7035000   32.766300   25.777400   31.139200   54.034800   39.243200
## 5    6.6264000    3.398600    1.295910    2.471260    3.131460    1.501020
## 6    7.4102400    6.587250    5.526510    7.139690    6.241450    5.135780
## 7   27.4604000   22.131100   30.323600   33.936500   35.699900   30.217600
## 8   11.0535000    5.930340   74.715400    9.661620   15.913700   16.774900
## 9    8.9291800    5.147840    8.784990   12.606900   12.124100    5.722520
## 10   0.0731159    0.405104    0.239548    0.206973    0.160525    0.507516
##    BS_2JP7RBMB
## 1    118.26300
## 2     15.39210
## 3     11.26210
## 4     37.35410
## 5      2.82956
## 6      6.05726
## 7     16.31480
## 8     10.97500
```

```
## 9     11.04320
## 10     2.54941

is(ge)

## [1] "data.frame" "list"       "oldClass"   "vector"

hist <- read.delim("pbta-histologies_v3.tsv", header=TRUE, sep="\t",
stringsAsFactors = FALSE)
str(hist)

## 'data.frame':    2835 obs. of  24 variables:
##  $ Kids_First_Participant_ID: chr  "PT_00G007DM" "PT_00G007DM"
"PT_00G007DM" "PT_00G007DM" ...
##  $ Kids_First_Biospecimen_ID: chr  "BS_K07KNTFY" "BS_CTEM6SYF"
"BS_AQMKA8NC" "BS_QWNBZ9RJ" ...
##  $ experimental_strategy    : chr  "WGS" "WGS" "WGS" "RNA-Seq" ...
##  $ sample_type              : chr  "Tumor" "Normal" "Tumor" "Tumor" ...
##  $ composition              : chr  "Solid Tissue" "Peripheral Whole Blood"
"Solid Tissue" "Solid Tissue" ...
##  $ tumor_descriptor         : chr  "Initial CNS Tumor" "" "Recurrence" ""
...
##  $ primary_site             : chr  "Cerebellum/Posterior Fossa" ""
"Cerebellum/Posterior Fossa" "Cerebellum/Posterior Fossa" ...
##  $ reported_gender          : chr  "Male" "Male" "Male" "Male" ...
##  $ race                     : chr  "White" "White" "White" "White" ...
##  $ ethnicity                : chr  "Not Hispanic or Latino" "Not Hispanic
or Latino" "Not Hispanic or Latino" "Not Hispanic or Latino" ...
##  $ age_at_diagnosis         : int  464 NA 3178 464 3178 NA 546 546 NA 2223
...
##  $ sample_id                : chr  "7316-272" NA "7316-2577" "7316-272"
...
##  $ aliquot_id               : chr  "588001" NA "601267" "588287" ...
##  $ disease_type_old         : chr  "Medulloblastoma" NA "Other"
"Medulloblastoma" ...
##  $ disease_type_new         : chr  "Medulloblastoma" NA "Ependymoblastoma"
"Medulloblastoma" ...
##  $ short_histology          : chr  "Medulloblastoma" NA "CNS Embryonal
tumor" "Medulloblastoma" ...
##  $ broad_histology          : chr  "Embryonal tumor" NA "Embryonal tumor"
"Embryonal tumor" ...
##  $ molecular_subtype        : chr  "" "" "" "Group3" ...
##  $ broad_composition        : chr  "tumor" "non-tumor" "tumor" "tumor" ...
##  $ Notes                    : chr  "" "" "changed from other to
Ependymoblastoma per CBTTC all sheet" "Subtype based on prediction" ...
##  $ germline_sex_estimate    : chr  "Male" "Male" "Male" "Male" ...
##  $ RNA_library              : chr  "" "" "" "stranded" ...
##  $ OS_days                  : int  783 783 783 783 783 680 680 680 663 663
...
##  $ OS_status                : chr  "LIVING" "LIVING" "LIVING" "LIVING" ...
```

```
I <- intersect(colnames(ge), hist$Kids_First_Biospecimen_ID)
hist[which(hist$reported_gender != hist$germline_sex_estimate),
c("Kids_First_Biospecimen_ID", "reported_gender",

"germline_sex_estimate")]
```

```
##      Kids_First_Biospecimen_ID reported_gender germline_sex_estimate
## 181                BS_3X8GAQ2M            Male                Female
## 182                BS_9X9E86DY            Male                Female
## 183                BS_E94QRJDW            Male                Female
## 768                BS_5KFAEMQJ                                  Male
## 769                BS_6S7T92KS                                  Male
## 770                BS_W9AT0MR7                                  Male
## 911                BS_0GDJQANN          Female               Unknown
## 983                BS_8CDVTS20          Female                  Male
## 984                BS_70FFPA51          Female                  Male
## 985                BS_BA0M71QZ          Female                  Male
## 1105               BS_9YPJANGX            Male                Female
## 1106               BS_AF6A572P            Male                Female
## 1107               BS_HB03GSHF            Male                Female
## 1199               BS_WGK8ZMVA                                  Male
## 1200               BS_EJ1H9PZY                                  Male
## 1201               BS_5PYJQBEA                                  Male
## 1877               BS_HJ99CJ2R          Female                  Male
## 1878               BS_VJ2S5DVS          Female                  Male
## 1879               BS_9GJHMA3J          Female                  Male
```

```
hist[768, "reported_gender"]
```

```
## [1] ""
```

```
#sum(is.na.data.frame(ge))
#result = 0

#sum(ge == "")
#result = 0

sum(hist[, "reported_gender"] == "")
```

```
## [1] 6
```

```
#result = 6
missing_reported_gender_samples <- hist[which(hist[, "reported_gender"] ==
""), "Kids_First_Biospecimen_ID"]

sum(hist[, "germline_sex_estimate"] == "")
```

```
## [1] NA
```

```
#result = 0
```

```
library(stats)

mads <- apply(ge[ , 3:1030], 1, function(x) mad(x, high=TRUE))

summary(mads)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     0.00     0.09     2.06     0.29 82081.93

mads_lo_hi <- sort(mads, index.return=TRUE)
tail(mads_lo_hi$x, 10)

##  [1]  5793.133  5847.189  5847.189  6643.227  8184.619  8292.686 21848.335
##  [8] 32403.706 61040.125 82081.925

tail(mads_lo_hi$ix, 10)

##  [1]  19028 160967 185084 178390 114100  38806 192290 196113 187627  94026
```

Let's try a model with the top 25% highest mads. Start a df with just the rows of ge corresponding to the top 25% highest mads. ge_top. transpopse ge_top[, 3:1030]. df. Set colnames(df) to transcript IDs. Set rownames(df) to sample names.

```
tail_percent = 0.75
ge_top <- ge[mads_lo_hi$ix[floor(tail_percent*length(mads)):length(mads)], ]
df <- t(ge_top[, 3:1030])
dim(df)

## [1]  1028 50102

colnames(df) <- ge_top[, 1]
head(colnames(df))

## [1] "ENST00000618025.4_AGAP6-203"     "ENST00000619710.4_PARVB-210"
## [3] "ENST00000622441.1_TMED10-207"    "ENST00000311915.12_C12orf66-201"
## [5] "ENST00000366778.5_COQ8A-202"     "ENST00000622656.1_PCDHGC3-207"

rownames(df) <- colnames(ge_top[, 3:1030])
head(rownames(df))

## [1] "BS_ZF6BSFNF" "BS_H97S5SQN" "BS_0ZA67BBC" "BS_W7MFJZ5A" "BS_68KX6A42"
## [6] "BS_NB9XXBW6"
```

Use reported_gender column from hist as the response. Eliminate rows of df corresponding to samples with missing reported_gender. Extract reported_gender values from reported_gender column of hist.

```
df <- df[!rownames(df) %in% missing_reported_gender_samples, ]
dim(df)

## [1]  1026 50102
```

```
reported_gender <- hist[hist$Kids_First_Biospecimen_ID %in% rownames(df),
c("Kids_First_Biospecimen_ID", "reported_gender")]
dim(reported_gender)

## [1] 1027    2
```

reported_gender is 1027 x 2, and shoule be 1026 by 2. Find duplicated row! Remove duplicated row.

Check sequence of rownames(df) and reported_gender[, 1]

```
x <- which(hist$Kids_First_Biospecimen_ID %in% rownames(df))
y <- duplicated(hist$Kids_First_Biospecimen_ID[x])
sum(y)

## [1] 1

which(y == TRUE)

## [1] 176

x[176]

## [1] 473

hist$Kids_First_Biospecimen_ID[473]

## [1] "BS_6DCSD5Y6"

which(hist$Kids_First_Biospecimen_ID =="BS_6DCSD5Y6")

## [1] 472 473

hist$Kids_First_Biospecimen_ID[472:473]

## [1] "BS_6DCSD5Y6" "BS_6DCSD5Y6"

which(reported_gender[, 1] == "BS_6DCSD5Y6")

## [1] 175 176

reported_gender <- reported_gender[-175,]

match_index <- sapply(rownames(df), function(x) which(reported_gender[, 1] ==
x))
#match_index <- unlist(match_index)
head(rownames(df[, ]))

## [1] "BS_ZF6BSFNF" "BS_H97S5SQN" "BS_0ZA67BBC" "BS_W7MFJZ5A" "BS_68KX6A42"
## [6] "BS_NB9XXBW6"

head(reported_gender[match_index, 1])
```

```
## [1] "BS_ZF6BSFNF" "BS_H97S5SQN" "BS_0ZA67BBC" "BS_W7MFJZ5A" "BS_68KX6A42"
## [6] "BS_NB9XXBW6"

reported_gender_response <- reported_gender[match_index, 2]
```

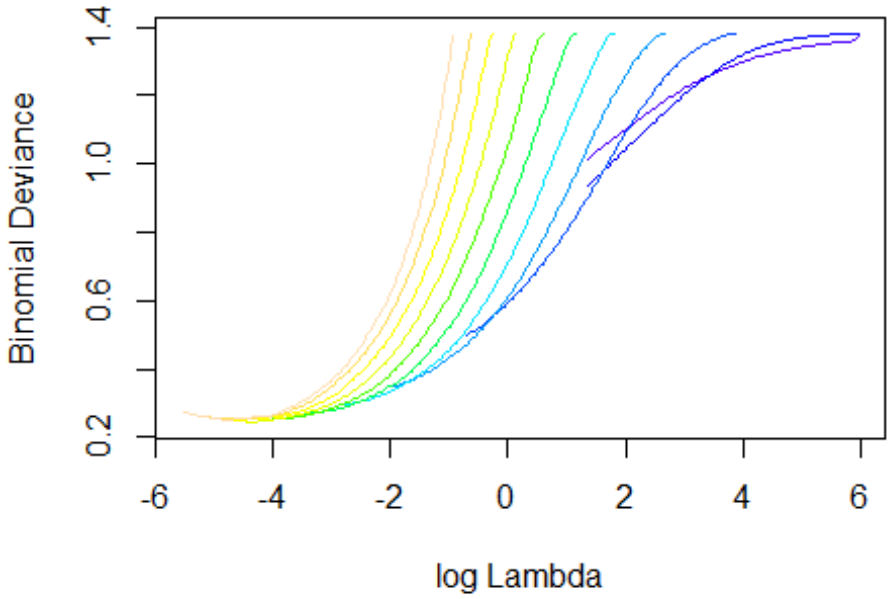
Check values in reported_gender_response.

```
table(reported_gender_response)

## reported_gender_response
##        Female          Male Not Available
##           470           551             5

reported_gender_response_hold <- reported_gender_response
reported_gender_response <- reported_gender_response[reported_gender_response
!= "Not Available"]
df <- df[reported_gender_response_hold != "Not Available", ]
```

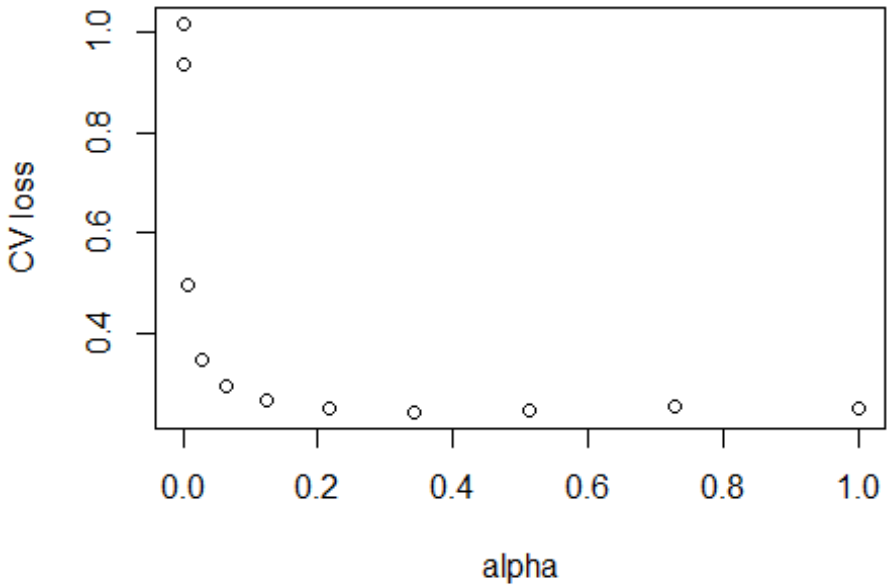
Build predictive model

```
library(glmnet)

## Warning: package 'glmnet' was built under R version 3.6.1

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-18

library(glmnetUtils)

## Warning: package 'glmnetUtils' was built under R version 3.6.1

##
## Attaching package: 'glmnetUtils'

## The following objects are masked from 'package:glmnet':
##
##     cv.glmnet, glmnet

train_percent <- 0.70
train_set <- sample(1:nrow(df), floor(train_percent*nrow(df)))
test_set <- setdiff(1:nrow(df), train_set)


ptm <- proc.time()
sex.cva <- cva.glmnet(df[train_set, ], reported_gender_response[train_set],
standardize=TRUE,
                      alpha = seq(0, 1, len = 11)^3, family="binomial")
proc.time() - ptm

##    user  system elapsed
## 1424.36   28.30 1459.58
```

```
plot(sex.cva)
```




```
minlossplot(sex.cva, cv.type="min")
```



```r
best_alpha_values <- sapply(sex.cva$modlist, function(x) min(x$cvm))
best_alpha_index <- which(best_alpha_values == min(best_alpha_values))
best_alpha_index
```

```
## [1] 8
```

```r
sex.cva$alpha[best_alpha_index]
```

```
## [1] 0.343
```

```r
best_fit <- sex.cva$modlist[[best_alpha_index]]
non_zero_features <- which(coef(best_fit, s =best_fit$lambda.min) != 0)
non_zero_features
```

```
##   [1]     1    80   212   272   486   491   574   971  1127  1294  1455
##  [12]  1956  2172  2332  2505  3076  3225  3939  4315  4433  4600  5473
##  [23]  5730  5740  6059  6225  6327  6513  6578  6698  6964  6985  7018
##  [34]  7063  7520  8242  8348  8591  8925  9176  9291  9412  9444  9767
##  [45]  9775 10294 10307 10548 10576 11006 11064 11271 11402 11658 12169
##  [56] 12484 12553 12639 12740 12857 12965 13097 13377 13711 13885 14448
##  [67] 14681 14720 14820 15020 15028 15180 15312 15729 16180 16417 16722
##  [78] 16961 17492 17631 17906 18221 18567 18734 18946 19051 19625 19749
##  [89] 19861 20003 20078 20171 20645 20908 21195 21358 21394 21717 21763
## [100] 22057 22242 22868 23214 23229 23315 23850 24291 24320 24689 24831
## [111] 24911 25138 25206 25530 25675 26026 26066 26107 26170 26371 26657
## [122] 26813 28252 28321 28385 28414 28454 28992 29089 29683 29684 29692
## [133] 29695 29783 29798 30366 30391 30594 30679 30732 31082 31366 31413
## [144] 31438 31446 31735 31874 31950 32258 32484 32490 32491 33808 34548
## [155] 34731 34870 35107 35396 36048 36531 36556 37178 37785 37986 38057
## [166] 38095 38261 38660 38749 39489 39769 39871 40048 40319 40469 41244
## [177] 41898 42740 43016 43128 43446 44347 45073 46762 47470 47609 48043
## [188] 48120 48321 48603 48845 49382 49475 49597 49604 49704 49914
```

```r
coef(best_fit, s=best_fit$lambda.min)[non_zero_features]
```

```
##   [1] -1.326429e+00  1.769810e-01 -2.648804e-02  1.575575e-02 -7.120632e-
03
##   [6]  4.332870e-03 -3.382337e-01  3.605991e-02  2.853135e-02 -1.361507e-
02
##  [11] -3.974114e-03 -5.578607e-02  7.440275e-01  5.548425e-02  1.999993e-
02
##  [16]  7.236112e-01  1.042827e-04  2.260844e-03 -4.119263e-03  2.569131e-
02
##  [21] -1.751178e-02  2.096583e-02  4.711296e-02  7.238293e-02 -4.056419e-
03
##  [26]  1.663763e-01  1.134825e-01 -3.493600e-01 -4.473584e-02  1.174995e-
01
##  [31] -1.302024e-01  8.135716e-03 -1.544153e-02  2.919632e-03 -1.857127e-
01
##  [36]  5.919073e-03  3.100159e-02 -2.449340e-02  8.112750e-03 -2.150408e-
03
```

```
##  [41] -2.179651e-03  1.699990e-02  1.692678e-03 -8.776507e-02  3.356741e-
01
##  [46]  1.753979e-02  9.727909e-01 -6.770068e-04 -4.536443e-02  2.412795e-
01
##  [51] -1.325086e-02  2.983582e-02  6.721438e-03  1.529252e-01  4.473603e-
03
##  [56]  5.472328e-02  3.465273e-03 -1.909214e-01  1.075191e-01 -3.036203e-
02
##  [61]  5.098318e-01 -5.902445e-02  1.202772e-02  6.874560e-02 -6.808256e-
05
##  [66]  9.332329e-04 -1.074307e-03  4.201348e-01  1.625237e-03  6.405066e-
02
##  [71]  6.871970e-02  1.064768e-03 -7.259769e-02  1.526945e-02  8.457059e-
02
##  [76]  3.748639e-02  1.433007e-02  4.723783e-02  1.423747e-01  8.655072e-
04
##  [81] -1.055955e-02 -4.442677e-02  7.349697e-02 -4.174199e-02 -3.042276e-
02
##  [86]  2.092883e-03 -1.991239e-01  6.282735e-03  4.582002e-04  9.074638e-
02
##  [91] -5.643055e-03  2.312609e-03 -7.430933e-02  2.190789e-02  3.741915e-
02
##  [96] -4.405910e-03  4.486601e-02  1.885338e-01 -1.705173e-03 -2.603998e-
01
## [101]  2.261279e-02  5.840200e-04  4.418720e-02 -1.140835e-01 -4.088778e-
02
## [106]  5.028387e-01 -3.162483e-01 -9.404800e-02  1.846097e-03  4.030649e-
03
## [111] -8.443456e-04 -3.320966e-01 -7.404352e-03 -1.870132e-02  1.349339e-
02
## [116]  5.087160e-02  2.394962e-02  8.136520e-03  1.921950e-02 -7.767639e-
02
## [121]  3.530238e-02  1.594078e-02 -3.255933e-02  1.100327e-03  5.390311e-
03
## [126]  1.694944e-02 -3.388324e-03  1.948103e-02  1.107446e-02 -6.257867e-
03
## [131]  1.683471e-02 -1.841799e-02  4.502421e-02  1.405173e-03  1.130694e-
02
## [136]  8.771071e-03 -1.517453e-02  2.101171e-02  8.660615e-03 -4.163081e-
02
## [141] -3.971428e-03 -6.819695e-03 -6.021585e-03 -6.313388e-02 -9.967325e-
03
## [146]  9.526462e-03  1.712468e-03  4.164589e-02  3.783758e-03  7.681532e-
03
## [151]  1.823444e-02  2.411423e-02  1.031931e-02 -6.202074e-03 -9.113554e-
02
## [156]  2.700069e-02  3.693120e-03  5.002459e-02  2.358431e-02 -9.649503e-
03
## [161]  1.142738e-03 -2.282465e-03 -1.173931e-02  1.183234e-02 -3.096587e-
03
```

```
## [166]  3.602908e-03 -4.196288e-02 -4.893907e-02 -1.632167e-02 -4.521758e-
04
## [171]  4.251941e-04 -9.929364e-03  3.303243e-01 -2.346901e-03  2.130912e-
03
## [176] -1.781038e-03  6.916535e-03 -3.317337e-02  4.639991e-03  1.484494e-
02
## [181] -9.432643e-03  8.440369e-03 -4.719038e-03 -2.734570e-03  4.389493e-
05
## [186]  3.021306e-03  3.526754e-03  3.652616e-03  7.525278e-03  5.847855e-
03
## [191] -8.945592e-04 -3.730636e-03  1.649570e-05  1.320241e-03  7.610267e-
06
## [196]  8.765788e-05  7.214162e-05
```

```
colnames(df)[non_zero_features]
```

```
##   [1] "ENST00000618025.4_AGAP6-203"
##   [2] "ENST00000475993.1_ABCF1-205"
##   [3] "ENST00000396331.5_LILRB1-205"
##   [4] "ENST00000422178.1_AC016745.1-201"
##   [5] "ENST00000254286.8_ACTR10-201"
##   [6] "ENST00000608346.1_AC083798.2-203"
##   [7] "ENST00000592705.5_DNMT1-228"
##   [8] "ENST00000580716.5_GPS1-212"
##   [9] "ENST00000554507.5_KTN1-214"
##  [10] "ENST00000477238.1_SEPT5-213"
##  [11] "ENST00000565225.1_PIGBOS1-202"
##  [12] "ENST00000595677.5_FCGRT-207"
##  [13] "ENST00000358647.4_GIMAP5-201"
##  [14] "ENST00000484683.1_RPRD1B-205"
##  [15] "ENST00000557120.5_ARG2-203"
##  [16] "ENST00000574217.1_PRPF8-210"
##  [17] "ENST00000546627.1_C12orf76-202"
##  [18] "ENST00000591326.5_MOB3A-206"
##  [19] "ENST00000490521.1_COMTD1-205"
##  [20] "ENST00000489153.1_SRC-209"
##  [21] "ENST00000337612.9_BCL2L13-202"
##  [22] "ENST00000417450.5_GTDC1-209"
##  [23] "ENST00000477925.5_SRPK2-210"
##  [24] "ENST00000522968.1_UBXN8-205"
##  [25] "ENST00000492487.5_SMYD3-216"
##  [26] "ENST00000599650.1_TIMM44-209"
##  [27] "ENST00000555206.5_ZFP37-203"
##  [28] "ENST00000389243.9_POU6F1-202"
##  [29] "ENST00000406116.7_RPS6KB1-203"
##  [30] "ENST00000311595.13_ENGASE-202"
##  [31] "ENST00000562400.1_FAM192A-204"
##  [32] "ENST00000587627.1_PLEKHH3-203"
##  [33] "ENST00000504298.1_BACH1-IT1-201"
##  [34] "ENST00000637674.1_AC116565.2-201"
```

```
##  [35] "ENST00000442406.5_EIF4G1-222"
##  [36] "ENST00000618947.1_TPT1-AS1-221"
##  [37] "ENST00000538083.1_SOX5-211"
##  [38] "ENST00000429907.5_RPE-205"
##  [39] "ENST00000427372.5_FAM133B-202"
##  [40] "ENST00000523721.2_PEX16-203"
##  [41] "ENST00000468797.1_AC021205.1-201"
##  [42] "ENST00000468508.1_SLC19A1-208"
##  [43] "ENST00000261520.8_ICE2-201"
##  [44] "ENST00000429663.5_AFG3L1P-209"
##  [45] "ENST00000594602.5_ZNF28-207"
##  [46] "ENST00000534225.1_IL18-206"
##  [47] "ENST00000573466.5_MAPK7-215"
##  [48] "ENST00000568731.1_ADCY7-211"
##  [49] "ENST00000359308.8_XRCC6-201"
##  [50] "ENST00000586495.1_TRIM47-203"
##  [51] "ENST00000555648.1_SNX6-205"
##  [52] "ENST00000486000.2_SLC2A10-202"
##  [53] "ENST00000560343.1_IGF1R-214"
##  [54] "ENST00000515332.5_SGMS2-209"
##  [55] "ENST00000490846.5_ANKRD13C-204"
##  [56] "ENST00000437805.5_SAMD9L-204"
##  [57] "ENST00000448644.2_VNN3-209"
##  [58] "ENST00000438261.5_SNRPG-204"
##  [59] "ENST00000488770.1_MCUR1-202"
##  [60] "ENST00000477897.1_PSMA5-202"
##  [61] "ENST00000429808.5_LINC00426-205"
##  [62] "ENST00000508532.5_ATP2C1-214"
##  [63] "ENST00000464308.1_POLR2M-204"
##  [64] "ENST00000483047.5_SEPT10-212"
##  [65] "ENST00000537232.5_TEX9-202"
##  [66] "ENST00000576925.3_BSG-210"
##  [67] "ENST00000476796.2_ZNF33B-204"
##  [68] "ENST00000354323.2_HRCT1-201"
##  [69] "ENST00000406949.5_PCNX4-204"
##  [70] "ENST00000548309.1_CRIP2-205"
##  [71] "ENST00000395656.6_PRPSAP2-202"
##  [72] "ENST00000506131.1_MARCH6-206"
##  [73] "ENST00000476826.5_MRPL18-202"
##  [74] "ENST00000557421.5_ATP5S-211"
##  [75] "ENST00000396056.6_ZNF35-202"
##  [76] "ENST00000606812.5_CCDC120-207"
##  [77] "ENST00000369985.8_MYO6-204"
##  [78] "ENST00000228862.3_DUSP16-201"
##  [79] "ENST00000491513.5_UBE2G2-209"
##  [80] "ENST00000421275.1_CELSR3-AS1-201"
##  [81] "ENST00000473513.5_FGD4-204"
##  [82] "ENST00000450730.5_PRKACB-214"
##  [83] "ENST00000561080.5_SLC12A6-219"
##  [84] "ENST00000460898.5_HAUS7-205"
```

```
##  [85] "ENST00000554339.5_DNAL1-205"
##  [86] "ENST00000405454.1_FAM102B-202"
##  [87] "ENST00000396385.3_MPV17L-202"
##  [88] "ENST00000396501.8_MPDU1-203"
##  [89] "ENST00000337432.8_RAD51C-201"
##  [90] "ENST00000498704.6_IFT22-209"
##  [91] "ENST00000558801.1_DNM1P51-202"
##  [92] "ENST00000462913.1_SFXN4-205"
##  [93] "ENST00000618976.1_AL163051.2-201"
##  [94] "ENST00000553935.5_KHNYN-202"
##  [95] "ENST00000468989.1_CDK5RAP2-207"
##  [96] "ENST00000513789.1_ZFYVE16-213"
##  [97] "ENST00000392244.7_SUMO1-201"
##  [98] "ENST00000623113.1_C19orf12-209"
##  [99] "ENST00000551731.1_ARSA-206"
## [100] "ENST00000362916.1_RNA5SP188-201"
## [101] "ENST00000620612.5_NBPF26-208"
## [102] "ENST00000345941.2_LPAR6-201"
## [103] "ENST00000384551.1_Y_RNA.517-201"
## [104] "ENST00000520841.1_SLU7-205"
## [105] "ENST00000422452.2_TENM1-202"
## [106] "ENST00000490628.1_GOLGA2-210"
## [107] "ENST00000315684.12_CTC1-201"
## [108] "ENST00000572329.5_BAIAP2-209"
## [109] "ENST00000260505.12_TTLL7-201"
## [110] "ENST00000430328.6_RIF1-204"
## [111] "ENST00000392290.5_MAIP1-202"
## [112] "ENST00000381329.5_WDR37-203"
## [113] "ENST00000595889.1_PNPLA6-213"
## [114] "ENST00000355898.5_ZNF507-202"
## [115] "ENST00000623867.1_AC007405.1-204"
## [116] "ENST00000593635.1_ZNF730-201"
## [117] "ENST00000446894.5_ZMYND8-210"
## [118] "ENST00000418917.6_TFG-202"
## [119] "ENST00000535549.5_ETS1-208"
## [120] "ENST00000512302.1_DROSHA-218"
## [121] "ENST00000603575.1_SRGAP2-205"
## [122] "ENST00000349606.4_MYLIP-201"
## [123] "ENST00000439976.5_LAMB1-205"
## [124] "ENST00000576407.1_PRPF8-213"
## [125] "ENST00000409421.5_MIER3-205"
## [126] "ENST00000458407.1_FAM201B-201"
## [127] "ENST00000438998.6_PSMG4-206"
## [128] "ENST00000599699.2_SSBP4-206"
## [129] "ENST00000477391.6_DBN1-207"
## [130] "ENST00000268184.10_CRTC3-201"
## [131] "ENST00000448745.5_CPSF7-206"
## [132] "ENST00000582236.1_MIR3194-201"
## [133] "ENST00000611986.1_PLCD3-209"
## [134] "ENST00000594436.5_MIA-203"
```

```
## [135] "ENST00000471218.5_PTPRE-207"
## [136] "ENST00000341068.7_ANAPC1-201"
## [137] "ENST00000488575.1_HLA-DPB1-216"
## [138] "ENST00000489390.1_SHOC2-205"
## [139] "ENST00000374597.3_STARD8-202"
## [140] "ENST00000488799.5_ADD3-215"
## [141] "ENST00000524356.1_PCM1-223"
## [142] "ENST00000522434.5_TM2D2-209"
## [143] "ENST00000527622.5_CTTN-211"
## [144] "ENST00000264658.10_FBXL20-201"
## [145] "ENST00000606892.1_TRAF3IP2-AS1-211"
## [146] "ENST00000614273.1_CYP1B1-207"
## [147] "ENST00000427358.3_FAHD1-203"
## [148] "ENST00000369098.3_C1orf54-201"
## [149] "ENST00000311277.8_MAP9-201"
## [150] "ENST00000576214.2_ACTG1-213"
## [151] "ENST00000264741.9_ITGA9-201"
## [152] "ENST00000360127.6_C17orf97-201"
## [153] "ENST00000360911.7_ZMYND8-205"
## [154] "ENST00000391952.7_CLASRP-202"
## [155] "ENST00000411028.1_RNU6-272P-201"
## [156] "ENST00000557320.5_PAPOLA-223"
## [157] "ENST00000365249.1_RNU6-302P-201"
## [158] "ENST00000373747.7_PUM1-204"
## [159] "ENST00000485960.6_TBC1D15-210"
## [160] "ENST00000376049.4_AIF1-202"
## [161] "ENST00000302609.7_ZNF25-201"
## [162] "ENST00000399777.1_BCL2L13-204"
## [163] "ENST00000439637.5_ADAM10-204"
## [164] "ENST00000395548.6_AP3S1-202"
## [165] "ENST00000263773.9_FNBP4-201"
## [166] "ENST00000446231.6_SMG1-202"
## [167] "ENST00000442216.1_CSNK1E-210"
## [168] "ENST00000344417.9_C7orf26-201"
## [169] "ENST00000547630.1_C12orf49-203"
## [170] "ENST00000447833.1_PDS5B-202"
## [171] "ENST00000526324.5_RSF1-205"
## [172] "ENST00000378802.4_CYP4V2-201"
## [173] "ENST00000329630.9_ZNF775-201"
## [174] "ENST00000582492.1_MIR4480-201"
## [175] "ENST00000476993.1_VWA1-203"
## [176] "ENST00000359208.6_DNTTIP2-201"
## [177] "ENST00000574331.5_SRRM2-218"
## [178] "ENST00000241600.9_MRPS2-201"
## [179] "ENST00000330137.11_SKA2-201"
## [180] "ENST00000224949.8_PITRM1-201"
## [181] "ENST00000355190.7_C6orf89-201"
## [182] "ENST00000502690.5_UBE2D3-213"
## [183] "ENST00000235382.6_RGS2-201"
## [184] "ENST00000629269.2_TBCB-214"
```

```
## [185] "ENST00000420698.5_NCOR2-207"
## [186] "ENST00000481027.5_LRRC75A-AS1-209"
## [187] "ENST00000490035.6_LSAMP-205"
## [188] "ENST00000402696.7_TF-201"
## [189] "ENST00000285930.8_AKR1B1-201"
## [190] "ENST00000270776.12_PGD-201"
## [191] "ENST00000248342.8_EIF3K-201"
## [192] "ENST00000286301.7_CSF1R-201"
## [193] "ENST00000597629.1_ZFP36-203"
## [194] "ENST00000523339.1_NPM1-210"
## [195] "ENST00000576209.5_ACTG1-212"
## [196] "ENST00000451270.6_ANXA2-204"
## [197] "ENST00000230050.3_RPS12-201"
```

```
p <- predict(best_fit, newx = df[test_set, ], type = "class", s
=best_fit$lambda.min)
length(which(p == reported_gender_response[test_set]))/length(p)
```

```
## [1] 0.9674267
```

```
which(p == reported_gender_response[test_set])
```

```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  29  30  31  32  33  34  35
##  [35]  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52
##  [52]  53  54  55  56  57  58  59  60  61  62  63  65  66  67  68  69  70
##  [69]  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87
##  [86]  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104
## [103] 105 106 107 108 109 110 111 112 113 114 115 116 118 119 120 121 122
## [120] 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139
## [137] 140 141 142 143 144 145 147 148 149 150 151 152 153 154 155 156 157
## [154] 158 159 160 161 162 163 164 165 167 168 169 170 171 172 173 174 175
## [171] 176 177 178 179 180 181 182 183 184 185 186 188 189 190 191 192 193
## [188] 194 195 196 197 198 199 200 201 202 204 205 206 207 208 209 210 211
## [205] 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228
## [222] 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245
## [239] 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 262 263
## [256] 264 265 266 267 269 270 271 272 273 274 275 276 277 279 280 281 282
## [273] 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299
## [290] 300 301 302 303 304 305 306 307
```

```
which(p != reported_gender_response[test_set])
```

```
##  [1]  28  64 117 146 166 187 203 261 268 278
```