Can Gene Expression Data Identify Patients With Inflammatory Bowel Disease?

A Consulting Project for Math 6627 (3/3)

Benjamin Smith

28 March 2022

Contents

Introduction

(Quoted from SSC website.)

Inflammatory bowel disease (IBD), which is comprised of the two disease entities of Crohn's disease (CD) and ulcerative colitis (UC), is an incurable gastrointestinal illness that results in chronic inflammation. IBD greatly affects patients' quality of life. Approximately 1.5 million people have IBD in the United States and Canada, where the rates are among the highest in the world.

There are currently no biomarkers for IBD, which could help to identify better treatments and individualize patient care. Such biomarkers could also be used to facilitate the development of clinical trials involving new medications. Recently, genome-wide association studies (GWAS) have significantly advanced our understanding about the importance of genetic susceptibility in IBD. Studies have identified a total of 201 IBD loci (Liu et al. 2015). However, these loci have yielded only a handful of candidate genes which often have small contributory effect in IBD.

The aim of this case study is to construct classifiers for IBD using global gene expression data based on these candidate genes. The research questions are:

1. Can data features (i.e., variables or probesets or genes) be used to cluster individuals into three biological groups (i.e., healthy individuals, CD patients, UC patients)?
2. Can data features (i.e., variables, probesets or genes) predict the disease state of individuals from three biological groups (i.e., healthy individuals, CD patients, UC patients)?

The Data

The data available consists of two datasets:

(Quoted from SSC website.)

1. Global gene expression data: Burczynski et al. (2006) generated genome-wide gene expression profiles for 41 healthy individuals (note that the processed data includes only 41 individuals although the original study included 42 individuals), 59 CD patients, and 26 UC patients using Affymetrix HG-U133A human GeneChip array. The GeneChip include approximate 22,000 probesets (each gene may have multiple probesets). The expression level of each probeset in each individual was quantified using MAS 5.0 software (we downloaded the processed data from ArrayExpress: E-GEOD-3365).
2. IBD candidate genes: IBD candidate genes implicated in the 201 IBD associated loci were evaluated using GRAIL (Gene Relationships across Implicated Loci) and DAPPLE (Disease Association Protein-Protein Link Evaluator) software tools. A total of 225 unique genes (see Supplementary Table 9 of Liu et al. 2015) were identified, and 185 of these 225 genes are on the Affymetrix HG-U133A human GeneChip array. These 185 candidate genes include 309 probesets.

Due to lack of domain knowledge and the ambiguity of the dataset, only the first data set will be used for this analysis. After the global gene expression data shaped properly for the analysis, there is no missing data present. With this taken care of, lets proceed with answering the two research questions.

Analysis

Clustering Individuals Into Three Biological Groups

On the SSC website it mentions that Liu, van Sommeren, Huang, et al found through association analyses which identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. To explain that in terms for that people with a non-genetics background - the result found by Liu, van Sommeren, Huang, et al identifies 38 specific locations on an subjects chromosomes which highlight shared genetic risk across populations.

As is the case with the assignments in this practicum, the ability to thoroughly analyze the data under time constraint presents its challenges. An approach to cluster the data into three groups can be done through principal component analysis. Interestingly enough, ~ 80% of the variance in the data can be explained by 38 principal components as shown in the R output below:

To view the effectiveness of using PCA to cluster individuals into three biological groups, a pca regression is preformed with 10-fold cross validation. From figures 1 and 2 the MSE produced from PCA regression using 38 components produces a MSE of around 0.43 and the spread between predicted and measured groups is quite large. From this it can be seen that the 38 principal components are useful for explaining the variance in the data, however as a method for classifying individuals it is quite poor.

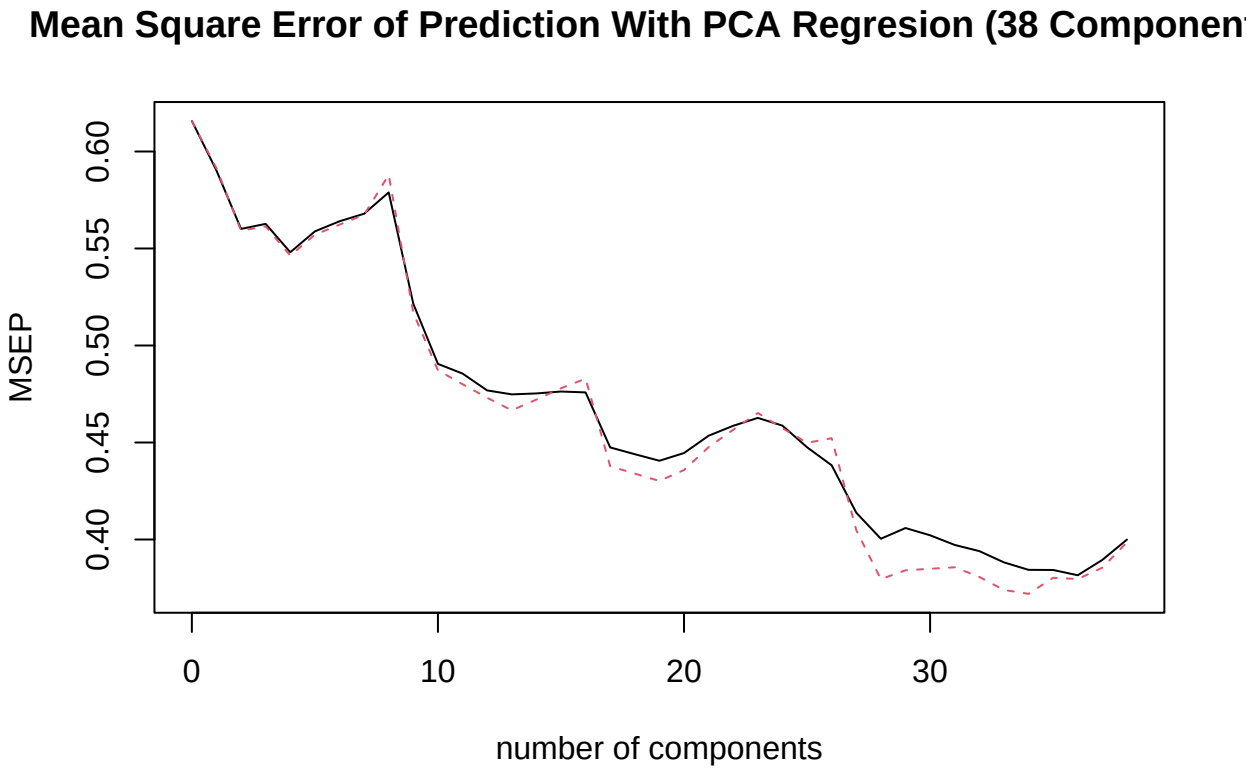

Figure 1: Mean Square Error of Prediction With PCA Regresion (38 Components)

While this method did prove to be a successful method for clustering data, it does demonstrate (and possibly confirms) the finding by Liu, van Sommeren, Huang, et al identifies loci which are susceptible to receiving the diseases in question by the general population, but are not useful as predictors in of themselves.

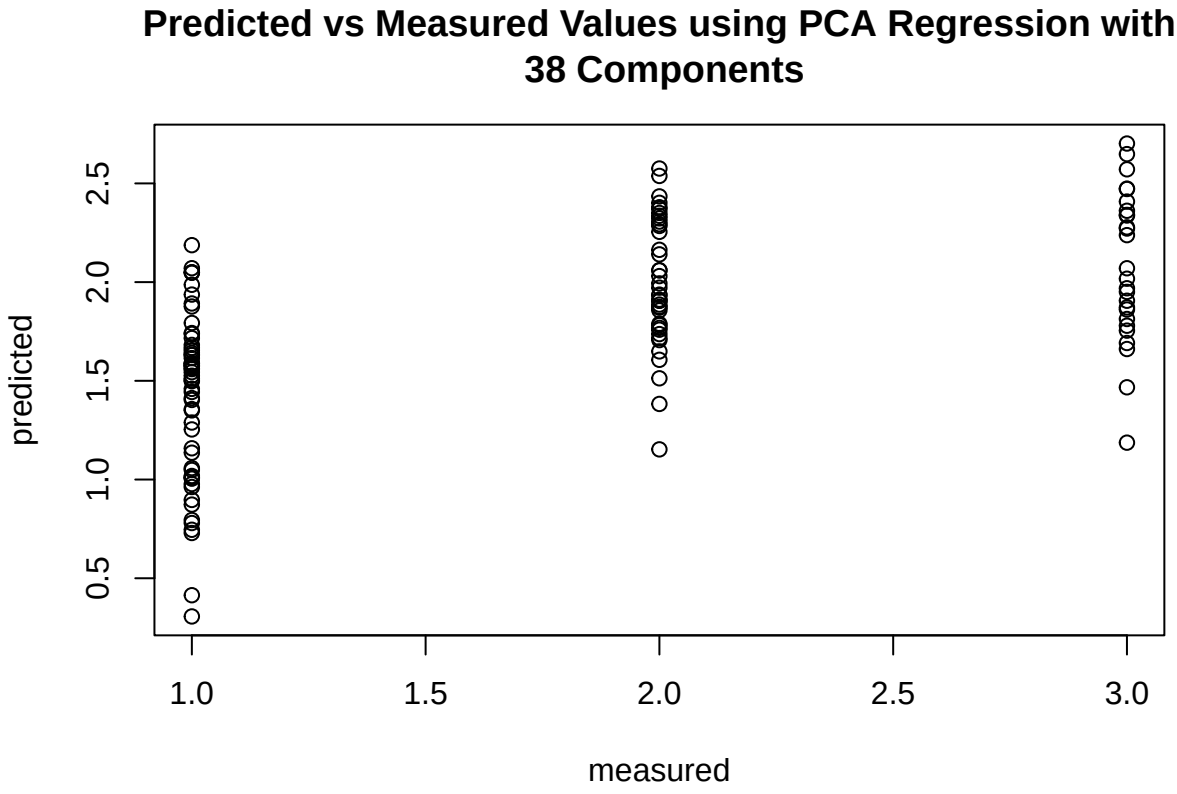

Figure 2: Predicted vs Measured Values using PCA Regression with 38 Components

What Data Features Can Predict The Disease State From Three Biological Groups

With a focus on prediction, the use of multinomial regression with 10-fold cross validation is employed with a traditional machine learning approach where the data is split into training and testing sets with a 75%/25% training-testing split in the data. Figure 3 shows that the multinomial deviance is minimized when $\lambda_{Group} = 0.003367685$. When modeling the testing data, the model managed to classify %76.47 of subjects correctly. This can be seen in the confusion matrix heatmap in figure 4.

As far as features is concerned, table 1 the sparse estimates from the model produces that there are 60 probesets which are significant predictors for classifying an individual as Normal, having Crohn's disease or Ulcerative Collitis. Table 2 shows that when looking Crohn's and UC alone, there are a total of 44 probesets. Interestingly enough the number of sparse estimates is only 6 more than that of the principal components which explain ~80% of the variance in the dataset.

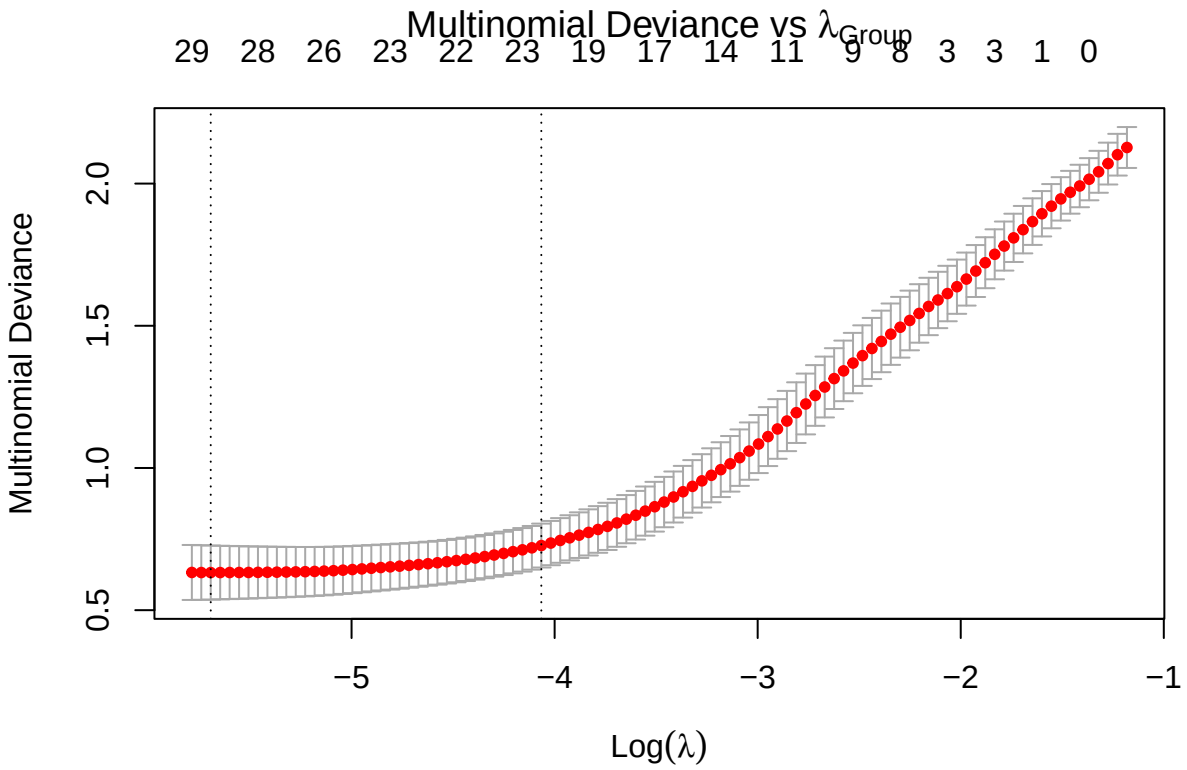

Figure 3: Multinomial Deviance vs λ_{Group}

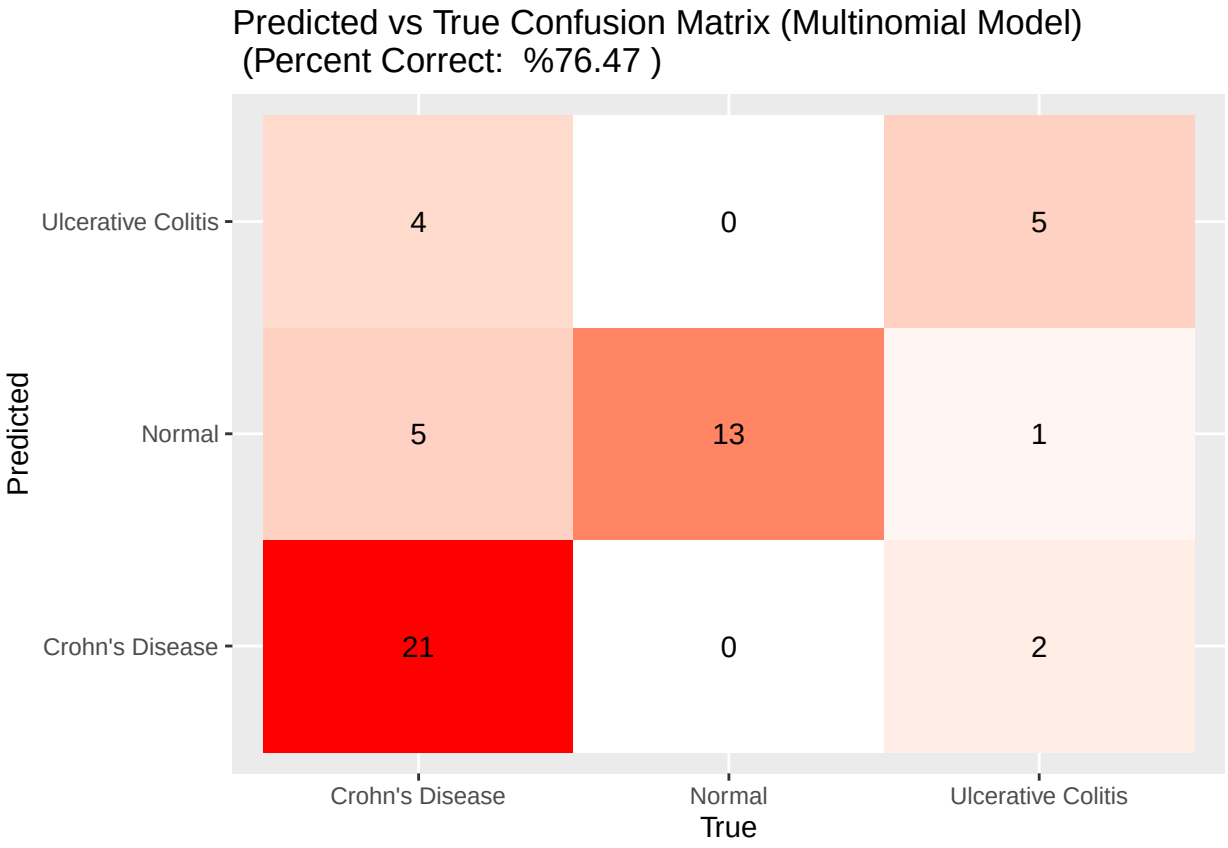

Figure 4: Predicted vs True Confusion Matrix (Multinomial Model)

Table 1: Sparse Estimates for Multinomial Classification of Groups

	Crohn's Disease	Normal	Ulcerative Colitis
Age	-0.0097589	0.0000000	0.0000000
200930_s_at	0.0000000	0.0000000	-0.0448403
201783_s_at	0.0000000	0.0017043	0.0000000
202531_at	0.0000000	0.0000000	-0.0009233
202716_at	0.0000000	0.0000000	0.0200865
203236_s_at	-0.0104848	0.0000000	0.0006280
203320_at	0.0000000	0.0030645	0.0000000
204057_at	0.0000000	0.0067957	0.0000000
204213_at	-0.0035302	0.0000000	0.0000000
204417_at	0.0000000	0.0000000	0.0000171
204785_x_at	0.0019350	0.0000000	0.0000000
204902_s_at	0.0000000	0.0000000	-0.0211848
204906_at	0.0283967	0.0000000	0.0000000
205098_at	0.0000000	-0.0033583	0.0003681
205266_at	0.0349097	0.0000000	0.0000000
206035_at	0.0000000	0.0320196	0.0000000
206390_x_at	0.0018512	0.0000000	0.0000000
206828_at	0.0000000	0.0169354	0.0000000
207072_at	0.0000000	0.0018425	-0.0003017
207526_s_at	0.0000000	0.0000000	0.0587881
207535_s_at	0.0000000	0.0042939	0.0000000
207844_at	-0.0355289	0.0000000	0.0000000
208010_s_at	0.0000000	0.0000000	-0.0138359
208038_at	0.0085615	0.0000000	0.0000000
208304_at	0.0000000	0.0000000	0.0227305
208621_s_at	0.0000000	0.0000000	-0.0084096
209544_at	0.0172310	0.0000000	0.0000000
209545_s_at	0.0000000	0.0000000	-0.0024219
209575_at	0.0000000	-0.0072154	0.0000000
209664_x_at	0.0000000	0.0000000	0.0173999
209782_s_at	0.0000000	0.0137451	0.0000000
209967_s_at	0.0000000	0.0000000	0.0063665
210133_at	0.0274815	0.0000000	0.0000000
210422_x_at	0.0000000	0.0000000	0.0016145
211153_s_at	0.0263287	0.0000000	0.0000000
211639_x_at	0.0000000	0.0000000	0.0333122
211856_x_at	-0.0120901	0.0000000	0.0000000
212501_at	0.0000000	0.0000000	0.0002571
212587_s_at	0.0000000	0.0000000	-0.0018010
212588_at	0.0000000	-0.0004283	0.0000000
212668_at	0.0000000	0.0426836	0.0000000
212912_at	0.0000000	-0.0154505	0.0000000
214228_x_at	0.0000000	0.0000000	-0.0038817
214467_at	0.0076350	0.0000000	0.0000000
215050_x_at	0.0000000	0.0000000	-0.0208561
215346_at	-0.0019466	0.0000000	0.0000000
216734_s_at	0.0000000	0.0000000	-0.0782045
216889_s_at	0.0000000	0.0204305	0.0000000
216986_s_at	0.0000000	-0.0429206	0.0000000
216987_at	-0.0172659	0.0000000	0.0000000

	Crohn's Disease	Normal	Ulcerative Colitis
217473_x_at	0.0019883	0.0000000	0.0000000
217689_at	0.0000000	0.0000000	-0.0067085
218648_at	0.0000000	0.0000000	0.0026076
219209_at	0.0000000	0.0158086	-0.0117014
220704_at	0.0000000	0.0596789	0.0000000
221092_at	0.0000000	0.0000000	-0.0486084
221111_at	0.0000000	0.0000000	-0.0634202
221331_x_at	-0.0876114	0.0138713	0.0000000
221690_s_at	0.0000000	0.0058926	0.0000000
222292_at	0.0000000	-0.0247665	0.0000000

Table 2: Sparse Estimates of Crohn's disease and Ulcerative Collitis Predictors

	Crohn's Disease	Ulcerative Colitis
Age	-0.0097589	0.0000000
200930_s_at	0.0000000	-0.0448403
202531_at	0.0000000	-0.0009233
202716_at	0.0000000	0.0200865
203236_s_at	-0.0104848	0.0006280
204213_at	-0.0035302	0.0000000
204417_at	0.0000000	0.0000171
204785_x_at	0.0019350	0.0000000
204902_s_at	0.0000000	-0.0211848
204906_at	0.0283967	0.0000000
205098_at	0.0000000	0.0003681
205266_at	0.0349097	0.0000000
206390_x_at	0.0018512	0.0000000
207072_at	0.0000000	-0.0003017
207526_s_at	0.0000000	0.0587881
207844_at	-0.0355289	0.0000000
208010_s_at	0.0000000	-0.0138359
208038_at	0.0085615	0.0000000
208304_at	0.0000000	0.0227305
208621_s_at	0.0000000	-0.0084096
209544_at	0.0172310	0.0000000
209545_s_at	0.0000000	-0.0024219
209664_x_at	0.0000000	0.0173999
209967_s_at	0.0000000	0.0063665
210133_at	0.0274815	0.0000000
210422_x_at	0.0000000	0.0016145
211153_s_at	0.0263287	0.0000000
211639_x_at	0.0000000	0.0333122
211856_x_at	-0.0120901	0.0000000
212501_at	0.0000000	0.0002571
212587_s_at	0.0000000	-0.0018010
214228_x_at	0.0000000	-0.0038817
214467_at	0.0076350	0.0000000
215050_x_at	0.0000000	-0.0208561
215346_at	-0.0019466	0.0000000
216734_s_at	0.0000000	-0.0782045

	Crohn's Disease	Ulcerative Colitis
216987_at	-0.0172659	0.0000000
217473_x_at	0.0019883	0.0000000
217689_at	0.0000000	-0.0067085
218648_at	0.0000000	0.0026076
219209_at	0.0000000	-0.0117014
221092_at	0.0000000	-0.0486084
221111_at	0.0000000	-0.0634202
221331_x_at	-0.0876114	0.0000000

Conclusion

The challenge of clustering and prediction of individuals into biological groups are two unique questions. In terms of clustering, the use of principal component analysis did manage to explain the variance an possibly highlights the same result noted by Liu, van Sommeren, Huang H, et al of there being 38 locations which highlight shared genetic risk across populations but does not help or explain (in the context of this analysis) why.

When focusing on prediction of individuals into biological groups, there are 60 probesets which serve to be useful sparse predictors. When focusing on individuals suffering from Crohn's and UC, that number is reduced to 44- just 6 more than the number of principal compontents which explain ~80% of the variance in the data.

Further research would involve exploring the relationship between the principal components and the sparse estimates and if any commonalities are shared between the two results.

References

1. Statistical Society of Canada, Can Gene Expression Data Identify Patients With Inflammatory Bowel Disease? (2017). https://ssc.ca/en/meeting/annual/2017/case-study-2
2. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 47(9):979-86 (2015).
3. Ron LP, Natalie CT, Krystyna AZ, et al. Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. Michael E Burczynski, J Mol Diagn 8(1):51-61 (2006).
4. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA. 99(10):6562-6 (2002).
5. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 99(2):147-57 (2007).
6. Cross-Validated, How to use R prcomp results for prediction? https://stats.stackexchange.com/questions/72839/how-to-use-r-prcomp-results-for-prediction

Code Appendix

```
library(tidyverse)
library(reshape2)
library(readxl)
library(nlme)
library(glmnet)

dt1 <- read_excel("ConsultingData_3.xlsx", sheet = 1)
dt2 <- read_excel("ConsultingData_3.xlsx", sheet = 2)

dtt2 <- model.matrix(~Gene.Symbol, dt2)


dtt <- dt1 %>%
    t() %>%
    as_tibble() %>%
    setNames(dt1 %>%
        t() %>%
        as_tibble() %>%
        slice(1:2) %>%
        as.list() %>%
        lapply(function(x) x[!is.na(x)]) %>%
        unname() %>%
        unlist()) %>%
    slice(c(3:n()))

dtt <- dtt %>%
    mutate(across(names(dtt)[-c(1, 3, 4)], ~as.numeric(.x)),
        Group = ifelse(Group == "Ulcerative" | Group == "Ulcerative Colitis",
            "Ulcerative Colitis", Group), Ethnicity = ifelse(Ethnicity ==
            "cacuasian" | Group == "caucasian", "caucasian",
            Ethnicity))



library(factoextra)
pca <- prcomp(data.matrix(dtt[, -1]), center = TRUE, scale = TRUE)

# We can use the first 38 principal components
pca_summary <- summary(pca)

# This agrees with the paper results
pca_summary$importance[, 1:38]


library(pls)
library(caret)
set.seed(1)

pcr_model <- pcr(Group ~ ., data = data.matrix(dtt) %>%
    as.data.frame.matrix(), scale = TRUE, validation = "CV",
    ncomp = 38)
```

```r
validationplot(pcr_model, val.type = c("MSEP"), main = "Mean Square Error of Prediction With PCA Regres
```

```r
predplot(pcr_model, main = "Predicted vs Measured Values using PCA Regression with\n 38 Components")
```

```r
set.seed(1)
dtt2 <- data.matrix(dtt)
y <- dtt2[, 1]
X <- dtt2[, -1]

train = sample(seq(length(y)), 75, replace = FALSE)
# 10 -fold crossvalidation
cv_model <- cv.glmnet(X, y, family = "multinomial", alpha = 1)
best_lambda <- cv_model$lambda.min
```

```r
plot(cv_model, main = expression("Multinomial Deviance vs " *
    lambda[Group] * ""))
```

```r
library(plyr)
train_model <- glmnet(X[train, ], y[train], family = "multinomial",
    alpha = 1, lambda = best_lambda)

confusion.glmnet(train_model, newx = X[-train, ], newy = y[-train]) %>%
    melt() %>%
    mutate(Predicted = revalue(as.character(Predicted), c(`1` = "Crohn's Disease",
        `2` = "Normal", `3` = "Ulcerative Colitis")), True = revalue(as.character(True),
        c(`1` = "Crohn's Disease", `2` = "Normal", `3` = "Ulcerative Colitis"))) %>%
    ggplot(mapping = aes(x = True, y = Predicted, fill = value,
        label = value)) + geom_tile() + geom_text() + scale_fill_gradient(low = "white",
    high = "red") + ggtitle("Predicted vs True Confusion Matrix (Multinomial Model)\n (Percent Correct:
    theme(legend.position = "none")
```

```r
sparseEstimates <- data.frame(`Crohn's Disease` = train_model$beta$`1`[,
    ], Normal = train_model$beta$`2`[, ], `Ulcerative Colitis` = train_model$beta$`3`[,
    ])

colnames(sparseEstimates) <- c("Crohn's Disease", "Normal", "Ulcerative Colitis")

sparseEstimates %>%
    filter(!(`Crohn's Disease` == 0 & Normal == 0 & `Ulcerative Colitis` ==
        0)) %>%
    knitr::kable(caption = "Sparse Estimates for Multinomial Classification of Groups")
```

```
sparseEstimates %>%
    select(!(Normal)) %>%
    filter(!(`Crohn's Disease` == 0 & `Ulcerative Colitis` ==
        0)) %>%
    knitr::kable(caption = "Sparse Estimates of Crohn's disease and Ulcerative Collitis Predictors")
```