8-12-2016

Identifying Inflammatory Bowel Disease Patients in TCGA Database

Regina Chang

IDENTIFYING INFLAMMATORY BOWEL DISEASE PATIENTS IN TCGA DATABASE

by

REGINA CHANG

Under the Direction of Yi Jiang, PhD

ABSTRACT

Chronic inflammation increases the risk of developing cancer. We aim to investigate the molecular pathway of inflammation induced cancer by comparing gene expression in colorectal (CRC) tumors of patients with inflammatory bowel disease (IBD) to sporatic colorectal tumors. Since mRNA microarray data of IBD induced CRC is not readily available, we attempt to isolate IBD patients in a public database based on their gene expression signatures.

INDEX WORDS: TCGA, microarray, Cancer, Inflammation, CRC, Colon Rectal Cancer, IBD.

IDENTIFYING INFLAMMATORY BOWEL DISEASE PATIENTS IN TCGA DATABASE

by

REGINA CHANG

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2016

IDENTIFYING INFLAMMATORY BOWEL DISEASE PATIENTS IN TCGA DATABASE

by

REGINA CHANG

Committee Chair: Yi Jiang

Committee: Gengsheng Qin

Remus Oşan

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2016

ACKNOWLEDGEMENTS

I would like to first thank my thesis advisor, Dr. Yi Jiang, without whom this thesis could not have been completed. She is knowledgeable and pleasant to work with, always encouraging and supporting me to explore new ideas.

I also would like to thank Dr. Yichuan Zhao, Dr. Remus Oşan, and Dr. Gengsheng Qin for all their help in many different forms during my graduate study in this department, either academically or materially.

Finally, thanks to Phuongan A. Dam from the Voit lab at Georgia Tech for guiding me through the day to day details of Biostatistic analysis and R coding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- TCGA - The Cancer Genome Atlas
- IBD - Inflammatory Bowel Disease
- CRC - Colorectal Cancer
- UC - Ulcerative Colitis
- CD - Crohn's Disease
- GEO - Gene Expression Omnibus

Chapter 1

INTRODUCTION

1.1 Cancer-Related Inflammation

There is a strong link between cancer and chronic inflammation [1]. We propose an approach to understanding the role of chronic inflammation in cancer by comparing the gene expression of one particular cancer that can be subdivided into two groups: sporadic and inflammation-driven. The sporadic subtype will be cancer patients who have no prior history of an inflammatory disease. The inflammation-driven subtype will be cancer patients who have a pre-existing inflammatory disease associated with increased risk of developing that paticular subtype of cancer.

Inflammatory bowel disease (IBD), inclusive of ulcerative colitis (UC) and Crohn's disease (CD), is associated with an increased risk for developing colorectal cancer (CRC) [2]. Even though IBD patients account for less than 2% of CRC cases, the risk of developing either a cancer precursor or cancer of the colon over 30 years increases from 2% to 18% for those with chronic inflammation [2].

1.2 Description of Data

The data was collected from two databases. The databases were chosen based on their open access and large data sets. The first is from Gene Expression Omnibus (GEO) at (http://www.ncbi.nlm.nih.gov/geo/) [3]. Data consisted of gene expression analysis of colon biopsies of CRC (n=15) and IBD (n=14) patients along with healthy normal controls (n=8) using high-density oligonucleotide microarray (series accession number GSE4183) [4]. Genome-wide gene expression profile was evaluated by HGU133 Plus 2.0 microarrays which measured the expression of 54,675 genes. According to the methods section of the research paper, pre-processing quality control and normalization of the data were preformed prior to

publishing on GEO. We hereafter refer to this data set as GEO data.

The second data set is from The Cancer Genome Atlas Research Network (http://cancergenome.nih.gov The Biospecimen Core Resource did not exclude colon or rectal cases for IBD, nor do the standard clinical forms allow mention of IBD. Thus, we can not use the clinical data to filter for IBD patients in the TCGA database. Some papers have used the TCGA data under the assumption that all samples originate from sporadic cases of colorectal cancer [9]. The R package RTCGA was used to download mRNA data from colon adenocarcinoma (COAD) samples consisting of 172 patients and 17,815 genes.

1.3 Purpose of the Thesis

The purpose of this study is two-fold. First, repeat and validate the classification procedure described in the *Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature* paper published in 2008 using their publicly available data. Second, apply the analysis on TCGA data to classify any IBD patients present in the database. This study aims to objectively identify any IBD patients in a large database based on gene expression when traditional histological diagnosis is unavailable.

Chapter 2

METHODOLOGIES

2.1 Preliminary Data Analysis

The R code used in this analysis can be found in Appendix B. TCGA data was uploaded to R using RTCGA.data packages [5]. The matrix contained expression levels of 17,815 genes of 172 patients. GSE4183 series matrix was uploaded on to R using the GEOquery package provided by GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/) [6]. Preliminary analysis of expression levels were conducted using GEO2R code provided by the website. KEGG analysis was done using the Pathview package [7]. TCGA and GEO data sets were cross-referenced for gene names contained in both data-sets. Unfortunately, none of the genes found to be significant (p ¡ 0.01) in the GEO2R analysis are present in the TCGA gene name list. The remaining analysis was limited to only genes profiled in both TCGA and GEO.

2.2 PAM Classification

Prediction analysis for Microarrays (PAM) uses soft thresholding to produce a shrunken centroid, which allows the selection of genes with high predictive potential [8]. Both data sets were normalized before training the classifier. The classifier was trained on the GEO set with only Normal (n=8) and IBD (n=15) patient data. The trainer was cross-validated, and the cross-validated errors were plotted. A threshold of 3 was chosen to minimize non zero errors. Cross-validated class probabilities by class, class centroids, and gene plot of most significant genes were computed. TCGA data (n=172) were then classified using the trainer. For plot aesthetics, only samples that changed in classification as threshold changed were shown. All other samples were classified as IBD regardless of threshold.

Chapter 3

RESULTS AND DISCUSSION

3.1 Preliminary Analysis of GEO data

Preliminary analysis of GEO data was done to gauge the expression profiles of "treatment" (IBD) versus "control" (normal). Further detail on the top 50 differentially expressed genes in IBD compared to normal can be found in Appendix A. Figure 3.1 is a heatmap of the difference in expression for each IBD sample. We can see a mix of upregulated and down regulated genes. No gene is uniformly up or down regulated across all samples. Figure 3.2 shows the KEGG pathway for IBD, entry hsa05321, of one IBD sample with up or down regulation of displayed genes colored red or green, respectively.

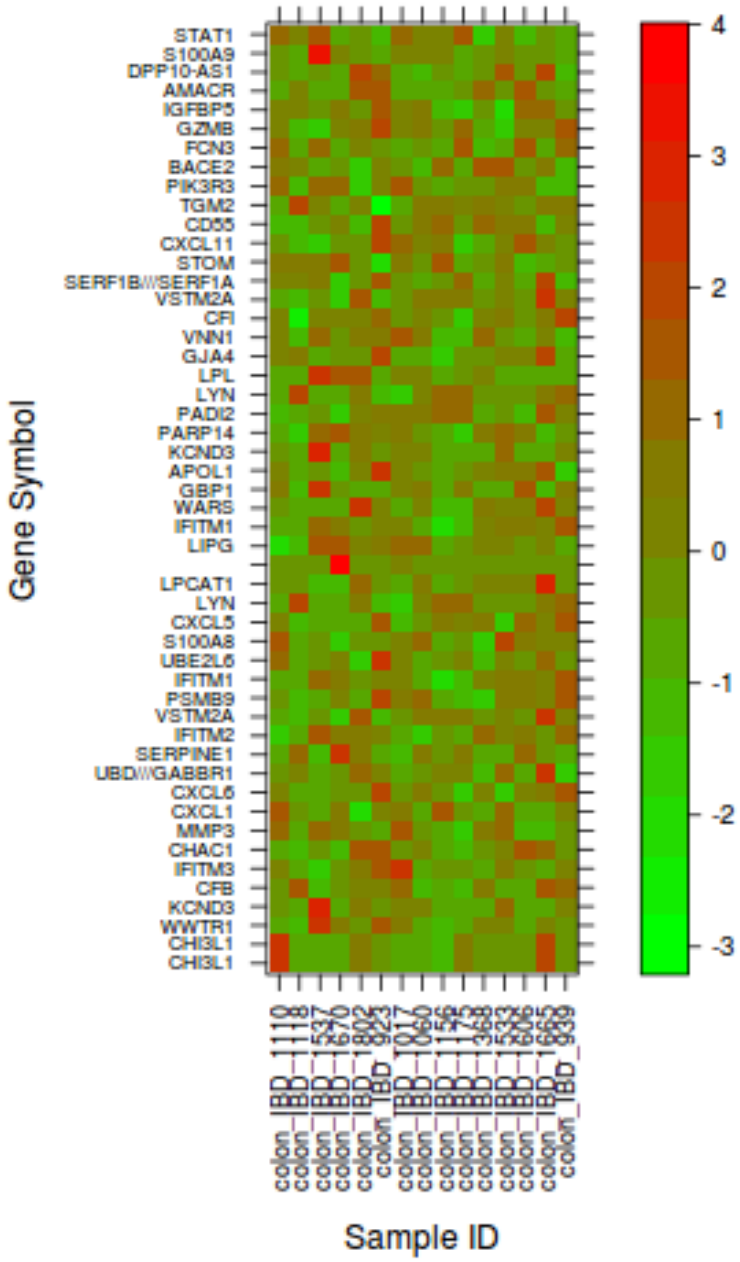

Figure (3.1) Heatmap of difference in expression of IBD patients compared to normal. Top 50 genes shown.

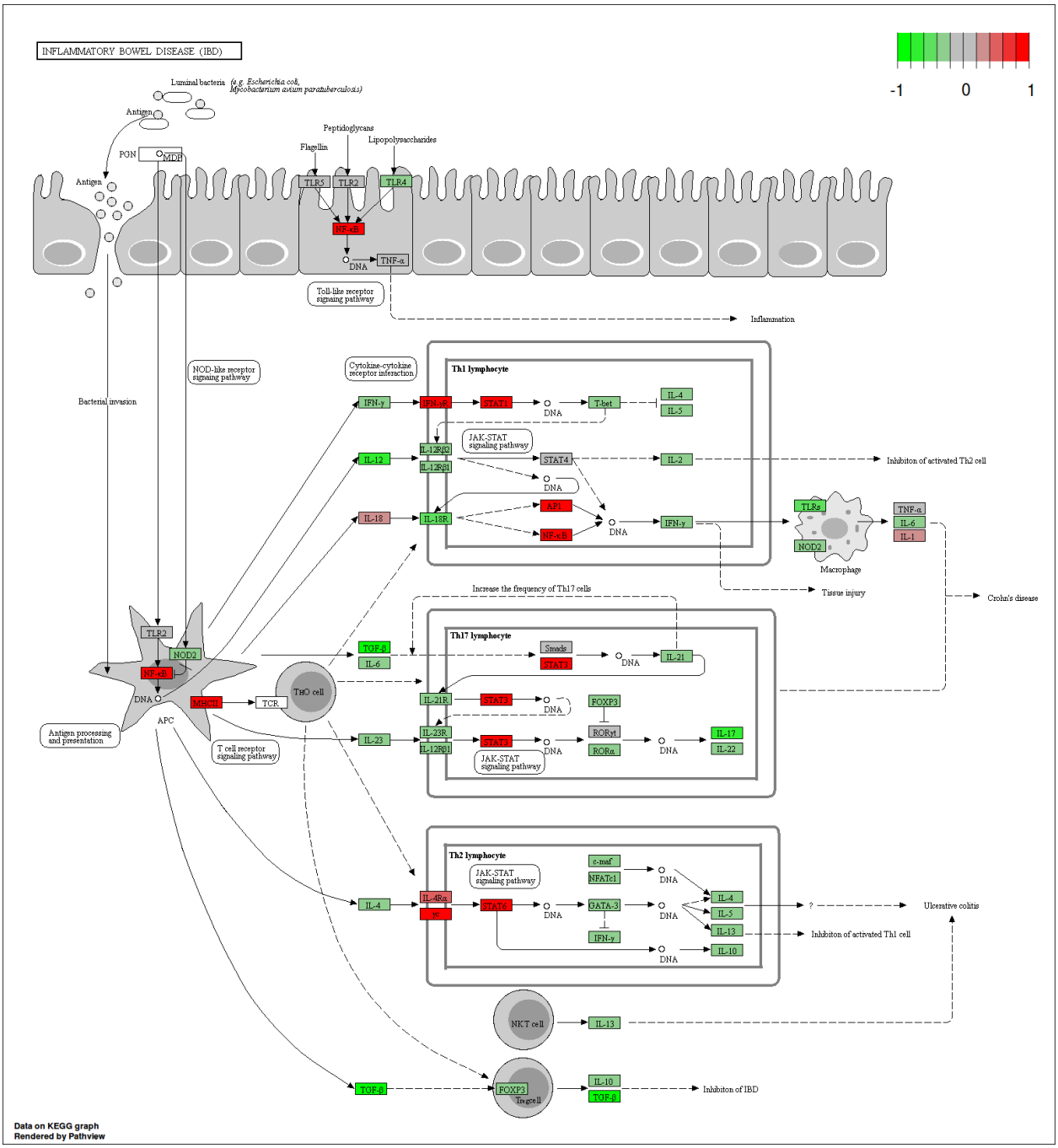

Figure (3.2) KEGG pathway hsa05321 of patient colon_IBD_939.

3.2 PAM Classification

The GEO dataset consisting of 14,554 genes of 53 columns was used to train the classifier, shown in table 3.1. As threshold increases, the number of genes used in the classifier decreases and the number of normal samples misclassified as IBD are shown under errors.

The cross-validated error curves the nearest shrunken centroid classifier are shown in figure 3.3. The error bars shown are confidence interval for misclassifcation (for example misclassifying 1 IBD sample out of 15 will result in an overall error rate of 0.043). Optimal threshold will be around 3.1 since it minimizes both number of genes used in the classifying and misclassification error. The classier has a bias towards misclassifying samples as IBD until threshold passes 3.1, then misclasification drastically increases towards normal samples.

Cross-validated class probabilities for each sample (n=23) are shown in figure 3.4. The x axis indicates sample number with the first 15 samples catagorized under IBD (red) and the last 8 categorized under normal (green). Each sample has two different color dots to indicate the probability of being categorized as either IBD or normal, y axis. The points are mirrored along the 0.5 probability line because there are only two categories for classification.

Table 3.2 and figure 3.5 shows the 25 genes used to classify the the samples. IBD/NOR-score are the raw gene expression for genes that survive the specified threshold of 3.1. The y axis on figure 3.5 is the raw gene expression value for each sample along the x axis. Red circles indicate IBD samples and green circles indicate normal samples. Plot is stratified by class. All genes except TMEM67A and STIP1 have higher expression in normal samples.

Finally, TCGA samples (n=172) were fed into the classifier at varying thresholds (min of 2.4 and max of 3.3). Majority of samples were classified as IBD (not shown) for all thresholds. The samples that changed classification as a function of threshold are shown in figure 3.6. Each row (y axis) indicates a different sample from the TCGA database. Each column is the classification of that sample as either normal (blue) or IBD (purple) based on the threshold (x axis). Most samples were reclassified from normal to IBD with the exception of two samples, which were reclassified twice.

	threshold	genes	errors
1	0.000	14554	1
2	0.125	13005	1
3	0.250	11455	1
4	0.374	9952	1
5	0.499	8547	1
6	0.624	7299	1
7	0.749	6168	1
8	0.874	5106	1
9	0.999	4174	1
10	1.123	3352	1
11	1.248	2694	1
12	1.373	2140	1
13	1.498	1656	1
14	1.623	1237	1
15	1.747	930	1
16	1.872	709	1
17	1.997	544	1
18	2.122	409	1
19	2.247	292	1
20	2.372	192	1
21	2.496	136	1
22	2.621	98	1
23	2.746	70	0
24	2.871	39	0
25	2.996	25	0
26	3.121	21	0
27	3.245	12	1
28	3.370	10	6
29	3.495	4	8
30	3.620	0	8

Table (3.1) PAM classifier output.

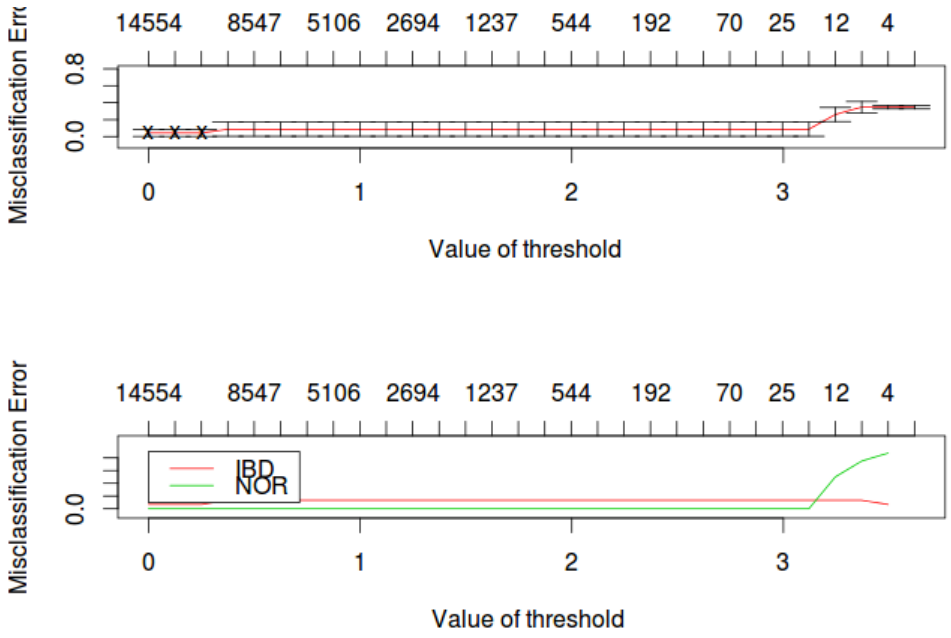

Figure (3.3) Cross-validated error curves of PAM for classification of IBD vs Normal.

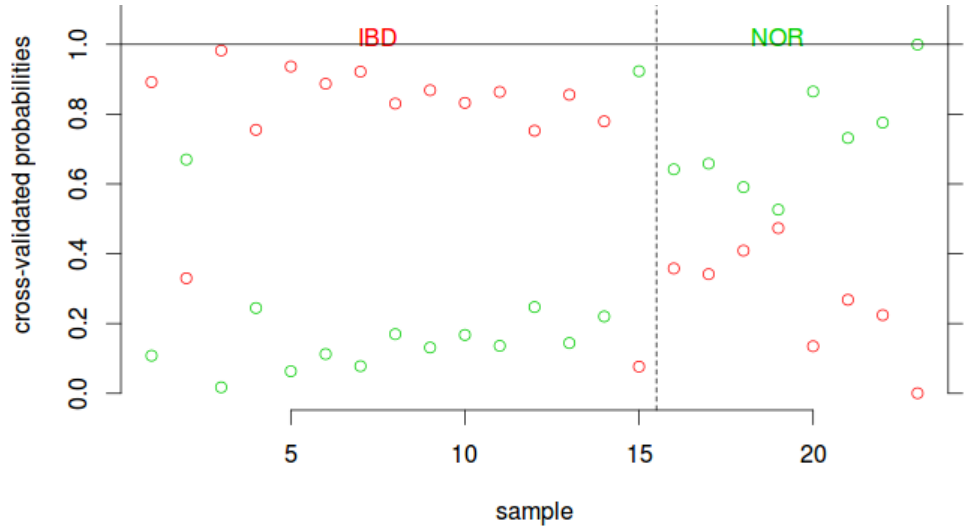

Figure (3.4) Cross-validated class probabilities of PAM for classification of IBD vs Normal.

	Gene Name	IBD-score	NOR-score
1	SLC6A12	-0.0944	0.177
2	PLA2G12B	-0.0942	0.1766
3	PAQR9	-0.0895	0.1678
4	EGFLAM	-0.0785	0.1472
5	YIPF1	-0.0695	0.1304
6	PADI2	-0.0675	0.1265
7	PIGL	-0.0663	0.1242
8	OLFM2	-0.0642	0.1204
9	KRT27	-0.0632	0.1185
10	C9orf69	-0.0629	0.118
11	CASP6	-0.0458	0.086
12	POU4F1	-0.0439	0.0823
13	FNDC4	-0.0324	0.0607
14	JPH3	-0.029	0.0544
15	C17orf59	-0.0259	0.0486
16	PHOSPHO1	-0.0238	0.0446
17	MASTL	-0.023	0.0431
18	DYNLT1	-0.0228	0.0427
19	PRDX6	-0.0209	0.0392
20	TMEM87A	0.0208	-0.039
21	POM121	-0.0203	0.038
22	YIPF2	-0.0183	0.0343
23	SLC5A7	-0.0172	0.0323
24	TACC3	-0.0089	0.0167
25	STIP1	0.0075	-0.0141

Table (3.2) PAM classification list of significant genes.

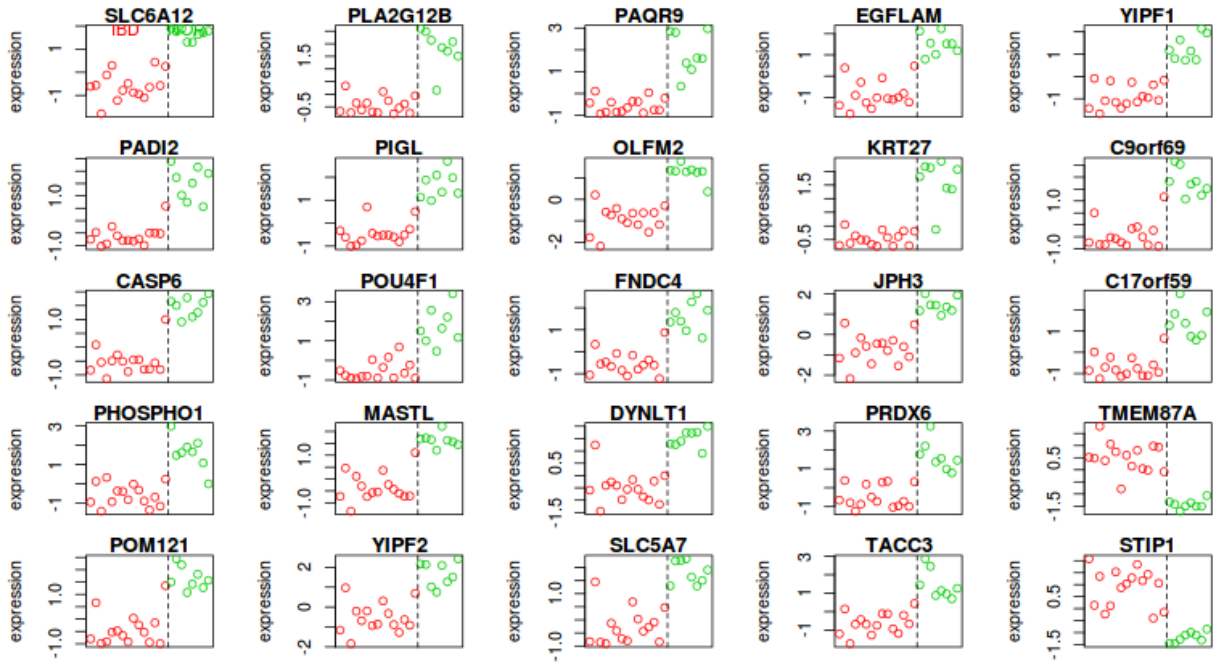

Figure (3.5) A gene plot of the most significant genes. Red indicates IBD classification and Green indicates Normal classification.

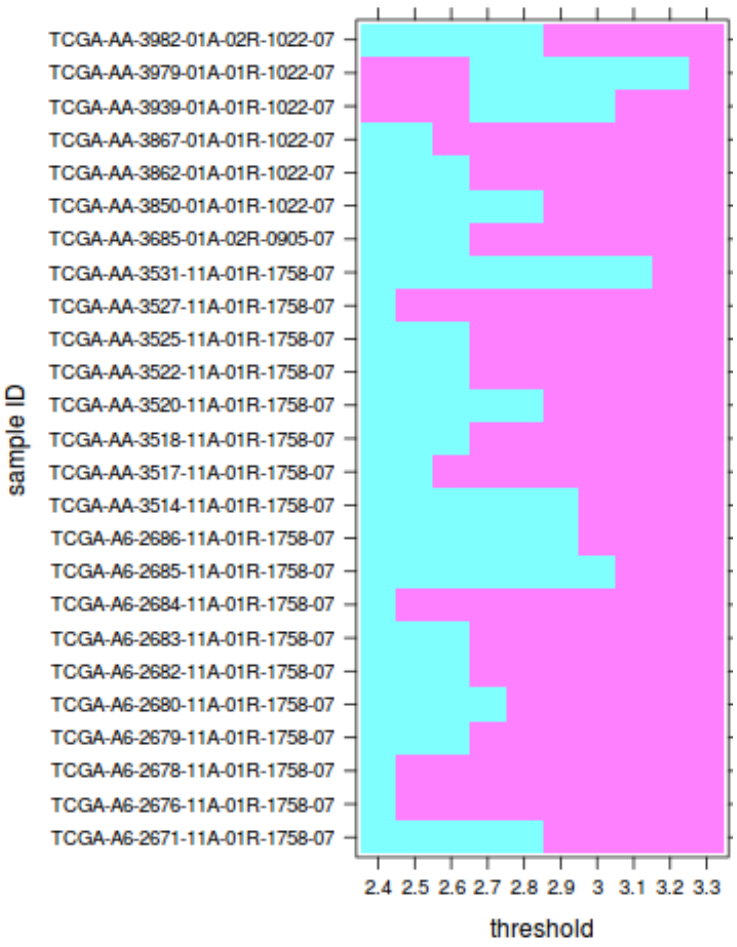

Figure (3.6) Heatmap of TCGA samples that changed sorting values with threshold. A value of 2 (blue) indicates Normal and a value of 1 (purple) indicates IBD.

Chapter 4

CONCLUSION

High throughput gene analysis databases such as TCGA and GEO are complicated and require years of bioinformatics expertise to master. Even the open source tools used to streamline the analysis of such databases can be hard to operate without proper training by an expert. There are many variables that go into genetics data and not all those are immediately visible even after careful consideration of the data available.

Initial analysis of the GEO data do not reveal any consistent expression patterns among all samples in the top 50 genes. Non of the genes are consistently up regulated or down regulated among all samples, posing a difficult task to track samples based on their signature expression differences. Pathway analysis has shown that not all genes present in the IBD pathway hsa05321 are significantly deferentially expressed compared to healthy normal subjects.

The PAM classifier has identified 25 genes used to sort between IBD and normal samples at a optimal threshold of 3.1. Surprisingly, these genes are not in the top 250 significantly differentially expressed genes. The results of the PAM classification of TCGA data indicate that a majority of samples can be classified as IBD. This may be because expression patterns of CRC are more similar to IBD than normal tissue samples.

Possible errors made in this analysis include matching gene names between the two databases and normalizing expression values. It is unclear how the TCGA samples were normalized, and thus difficult to repeat the normalization procedure on GEO samples. If normalization for TCGA samples were from same patient tissue samples of non-disease tissue, or a housekeeping gene not present in the GEO array, then it would be impossible to match the GEO samples to TCGA. This will result in errors with the sorting algorithm since the two samples are not consistent. Further direction in categorizing TCGA data may include

Principal Component Analysis or Discriminant Analysis.

REFERENCES

[1] Mantovani A, Allavena P, Sica A, Balkwill F. (July 2008). Cancer-related inflammation. *Nature* 454 (7203):436-44.

[2] Triantafillidis JK, Nasioulas, G, Kosmidis, PA (Jul 2009). Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies. *Anticancer Research* 22 (20):4794-801.

[3] Edgar R, Domrachev M, Lash AE (Jan 2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository *Nucleic Acids Res* 30(1):207-10

[4] Galamb O, Gyrffy B, Sipos F, Spisk S et al (2008). Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. *Dis Markers* 25(1):1-16.

[5] Chodor W (2015). RTCGA.mRNA: mRNA datasets from The Cancer Genome Atlas Project. R package version 1.0.2.

[6] Sean Davis and Paul S. Meltzer (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23(14): 1846-1847

[7] Luo, Weijun, Brouwer and Cory (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics, 29(14), pp. 1830-1831.

[8] R.Tibshirani, T. Hastie, B. Barasimhan and G. Chu (2002). Diagnosis of multiple cancer types oby shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 99, 6567-6572.

[9] Robles AI, Traverso G, Zhang M, Roberts NJ, Khan MA, Joseph C, Lauwers GY, Selaru FM, Popoli M, Pittman ME, Ke X, Hruban RH, Meltzer SJ, Kinzler KW, Vogelstein B, Harris CC, Papadopoulos N. (April 2016). Whole-Exome Sequencing Analyses of Inflammatory Bowel Disease-Associated Colorectal Cancers. *Gastroenterology* 150(4):931-43.

Appendix A

TOP 50 GENES

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
209396_s_at	2.56e-10	4.69e-15	-17.586	22.526	-6.966	CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39)
209395_at	2.74e-09	1e-13	-15.275	20.132	-6.270	CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39)
202134_s_at	2.82e-07	1.74e-11	-11.947	15.799	-2.322	WWTR1	WW domain containing transcription regulator 1
215014_at	2.82e-07	2.06e-11	-11.847	15.649	-3.018	KCND3	potassium voltage-gated channel, Shal-related subfamily, member 3
202357_s_at	6.12e-07	6.33e-11	-11.206	14.662	-2.712	CFB	complement factor B
212203_x_at	6.12e-07	7.5e-11	-11.111	14.511	-1.768	IFITM3	interferon induced transmembrane protein 3
219270_at	6.12e-07	7.83e-11	-11.087	14.472	-3.331	CHAC1	ChaC, cation transport regulator homolog 1 (E. coli)
205828_at	1.1e-06	1.73e-10	-10.651	13.764	-5.202	MMP3	matrix metallopeptidase 3 (stromelysin 1, progelatinase)
204470_at	1.1e-06	1.85e-10	-10.614	13.703	-4.298	CXCL1	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
206336_at	1.1e-06	2.02e-10	-10.568	13.626	-4.338	CXCL6	chemokine (C-X-C motif) ligand 6
205890_s_at	1.34e-06	2.7e-10	-10.412	13.365	-3.671	UBD/// GABBR1	ubiquitin D/// gamma-aminobutyric acid (GABA) B receptor, 1
202628_s_at	1.36e-06	2.99e-10	-10.358	13.273	-4.080	SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
201315_x_at	1.43e-06	3.66e-10	-10.250	13.090	-1.960	IFITM2	interferon induced transmembrane protein 2
1554530_at	1.43e-06	3.67e-10	10.249	13.088	3.928	VSTM2A	V-set and transmembrane domain containing 2A
204279_at	1.43e-06	3.93e-10	-10.213	13.026	-1.988	PSMB9	proteasome (prosome, macropain) subunit, beta type, 9
214022_s_at	1.56e-06	4.58e-10	-10.133	12.889	-1.393	IFITM1	interferon induced transmembrane protein 1
201649_at	1.61e-06	5e-10	-10.087	12.809	-1.701	UBE2L6	ubiquitin-conjugating enzyme E2L 6
202917_s_at	2.17e-06	7.14e-10	-9.902	12.486	-5.078	S100A8	S100 calcium binding protein A8
214974_x_at	2.64e-06	9.18e-10	-9.773	12.258	-5.488	CXCL5	chemokine (C-X-C motif) ligand 5
202625_at	7.18e-06	2.63e-09	-9.244	11.295	-1.506	LYN	LYN proto-oncogene, Src family tyrosine kinase
201818_at	7.38e-06	2.83e-09	-9.206	11.225	-3.138	LPCAT1	lysophosphatidylcholine acyltransferase 1
231078_at	9.91e-06	3.99e-09	-9.039	10.911	-2.914		
219181_at	1.02e-05	4.28e-09	-9.004	10.846	-1.400	LIPG	lipase, endothelial
201601_x_at	1.22e-05	5.35e-09	-8.895	10.639	-1.760	IFITM1	interferon induced transmembrane protein 1
200629_at	1.35e-05	6.36e-09	-8.812	10.480	-2.208	WARS	tryptophanyl-tRNA synthetase
231577_s_at	1.35e-05	6.4e-09	-8.809	10.474	-2.629	GBP1	guanylate binding protein 1, interferon-inducible
209546_s_at	1.46e-05	7.25e-09	-8.749	10.359	-2.748	APOL1	apolipoprotein L, 1
213832_at	1.46e-05	7.49e-09	-8.734	10.329	-2.064	KCND3	potassium voltage-gated channel, Shal-related subfamily, member 3
224701_at	1.86e-05	9.89e-09	-8.601	10.071	-1.571	PARP14	poly (ADP-ribose) polymerase family, member 14
1554385_a_at	2.14e-05	1.18e-08	8.519	9.910	1.844	PADI2	peptidyl arginine deiminase, type II
202626_s_at	2.16e-05	1.23e-08	-8.499	9.872	-1.486	LYN	LYN proto-oncogene, Src family tyrosine kinase
203548_s_at	2.2e-05	1.29e-08	-8.477	9.828	-3.370	LPL	lipoprotein lipase
40687_at	2.26e-05	1.37e-08	-8.447	9.768	-2.496	GJA4	gap junction protein, alpha 4, 37kDa
205844_at	2.26e-05	1.45e-08	-8.420	9.716	-4.317	VNN1	vanin 1
203854_at	2.26e-05	1.45e-08	-8.420	9.716	-3.026	CFI	complement factor I
236308_at	2.29e-05	1.51e-08	8.401	9.678	2.039	VSTM2A	V-set and transmembrane domain containing 2A
223539_s_at	2.68e-05	1.81e-08	8.315	9.506	1.563	SERF1B/// SERF1A	small EDRK-rich factor 1B (centromeric)/// small EDRK-rich factor 1A (telomeric)
201061_s_at	2.68e-05	1.91e-08	-8.291	9.458	-1.663	STOM	stomatin
211122_s_at	2.68e-05	1.91e-08	-8.291	9.458	-3.865	CXCL11	chemokine (C-X-C motif) ligand 11
201925_s_at	2.81e-05	2.06e-08	-8.257	9.389	-2.917	CD55	CD55 molecule, decay accelerating factor for complement (Cromer blood group)
201042_at	2.86e-05	2.14e-08	-8.237	9.351	-1.877	TGM2	transglutaminase 2
202743_at	3.01e-05	2.32e-08	-8.202	9.279	-1.390	PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)
227051_at	3.22e-05	2.54e-08	-8.160	9.194	-2.123	BACE2	beta-site APP-cleaving enzyme 2
205866_at	3.46e-05	2.84e-08	-8.107	9.089	-2.726	FCN3	ficolin (collagen/fibrinogen domain containing) 3
210164_at	3.46e-05	2.85e-08	-8.106	9.085	-3.013	GZMB	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)
211959_at	3.54e-05	2.98e-08	-8.085	9.043	-2.464	IGFBP5	insulin-like growth factor binding protein 5
209425_at	3.62e-05	3.13e-08	8.063	8.998	1.655	AMACR	alpha-methylacyl-CoA racemase
236351_at	3.62e-05	3.3e-08	8.039	8.949	2.915	DPP10-AS1	DPP10 antisense RNA 1
203535_at	3.62e-05	3.3e-08	-8.038	8.948	-4.701	S100A9	S100 calcium binding protein A9
AFFX-HUMISGF3A/M97935_3_at	3.62e-05	3.31e-08	-8.037	8.945	-1.304	STAT1	signal transducer and activator of transcription 1, 91kDa

Table (A.1) Top 50 genes differentially expressed in IBD compared to normal.

Appendix B

R CODE

```
#### Load TCGA Data ####
#https://rtcga.github.io/RTCGA/
# source("https://bioconductor.org/biocLite.R")
# biocLite("RTCGA.mRNA")
library('RTCGA.mRNA')
COAD.data=COAD.mRNA

#### Load GEO Data ####
#http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4183
#Download GSE4183_series_matrix1.txt this is a modified version of the Series Matrix Fil
datadir <- "/home/gina/Downloads/" #change this to your directory
GSE4183.data <- read.table(file.path(datadir, "GSE4183_series_matrix1.txt"), fill = TRUE
colnames(GSE4183.data)=c("colon_normal_1024", "colon_normal_1081","colon_normal_1114","c
                         "colon_adenoma_1115","colon_adenoma_1138","colon_adenoma_1141",
                         "colon_CRC_1146","colon_CRC_1158","colon_CRC_1293","colon_CRC_1
                         "colon_IBD_1110","colon_IBD_1118","colon_IBD_1537","colon_IBD_1
#But oh no! The gene names are probe-IDs of Affymetrix Human Genome U133 Plus 2.0 Array
#How do we get the correct gene names to match this expression data with TCGA?
#Don your hazmat suits, cause we bout to crawl thru some shit

#### Alternative Method of Loading GEO Data: GEO2R####
#Make sure your R is version 3.0 or higher
#load required libraries from Bioconductor (This may require tears and more understandin
```

```
source("http://bioconductor.org/biocLite.R")
biocLite("GEOquery")
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.36.0, limma 3.26.8
# R scripts generated  Mon May 23 19:13:44 EDT 2016

# load series and platform data from GEO

gset <- getGEO("GSE4183", GSEMatrix =TRUE) #don't panic this will take a while
if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
gsms <- "11111111XXXXXXXXXXXXXXXXXXXXXXXXXXXXX000000000000000"
sml <- c()
for (i in 1:nchar(gsms)) { sml[i] <- substr(gsms,i,i) }

# eliminate samples marked as "X"
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]

# log2 transform
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
```

```
  (qx[6]-qx[1] > 50 && qx[2] > 0) ||
  (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
exprs(gset) <- log2(ex) }

# set up the data and proceed with analysis
sml <- paste("G", sml, sep="")    # set group names
fl <- as.factor(sml)
gset$description <- fl
design <- model.matrix(~ description + 0, gset)
colnames(design) <- levels(fl)
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(G1-G0, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
#tT.all <- topTable(fit2, adjust="fdr", sort.by="B", number=3082)

# load NCBI platform annotation
gpl <- annotation(gset)
platf <- getGEO(gpl, AnnotGPL=TRUE)
ncbifd <- data.frame(attr(dataTable(platf), "table"))

# replace original platform annotation
tT <- tT[setdiff(colnames(tT), setdiff(fvarLabels(gset), "ID"))]
tT <- merge(tT, ncbifd, by="ID")
tT <- tT[order(tT$P.Value), ]  # restore correct order
```

```
tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene
write.table(tT, file=stdout(), row.names=F, sep="\t")

#NOW we can name GEO data
rownames(GSE4183.data)=make.names(ncbifd$Gene.symbol, unique=TRUE)
#Make GEO data the same format as TCGA
GEO.data=data.frame(t(GSE4183.data))

##### KEGG pathway analysis #####
source("http://bioconductor.org/biocLite.R")
biocLite("pathview")
biocLite(c("Rgraphviz", "png", "KEGGgraph", "org.Hs.eg.db"))

library(pathview)
#name genes as KEGG names
rownames(ex)=ncbifd$Gene.ID
#use IBD pathway
#http://www.genome.jp/dbget-bin/www_bget?pathway+hsa05321
pv.out <- pathview(gene.data = scale(ex[, 23]), pathway.id = "05321",
                   species = "hsa", out.suffix = "gse4183", kegg.native = T)
#look for a file called hsa05321.gse4183.png

### Create a heatmap of top 50 expressed genes ###
heatmap.data=GEO.data[c(1:8,39:53),tT$Gene.symbol[1:50]]
m=as.vector(apply(heatmap.data[1:8,],2,mean))
heatmap.normalized=as.matrix(heatmap.data[-c(1:8),])-matrix(t(replicate(15,m)),nrow=15,n
colnames(heatmap.normalized)=tT$Gene.symbol[1:50]
library(lattice)
```

```
col.l <- colorRampPalette(c('green', 'red'))(30)
levelplot(scale(heatmap.normalized),xlab="Sample ID",ylab="Gene Symbol",
          col.regions=col.l, scales=list(x=list(rot=90, cex=0.75),y=list(cex=0.6)))

#### Determining Overlap b/w TCGA and GEO gene expression ####
genenames.TCGA=names(COAD.data[-1])
genenames.GEO=ncbifd$Gene.symbol
#Oh, btw, there's another problem with GEO data (actually there's a lot of problems with
sum(duplicated(genenames.GEO))
#TCGA does not have duplicated gene names

#Find which genes are present in both TCGA and GEO w/o duplicates (using first instance
samegenes.idx=match(genenames.TCGA,genenames.GEO)
sum(is.na(samegenes.idx)) #some TCGA genes not in GEO
samegenes.names=genenames.GEO[samegenes.idx]
samegenes.names=samegenes.names[!is.na(samegenes.names)]
length(samegenes.names) #number of genes to be used in our analysis

samegenes.idx=match(names(COAD.data),names(GEO.data))
sum(is.na(samegenes.idx)) #some TCGA genes not in GEO
genenames.GEO=names(GEO.data)
samegenes.names=genenames.GEO[samegenes.idx]
samegenes.names=samegenes.names[!is.na(samegenes.names)]
length(samegenes.names) #number of genes to be used in our analysis

#Truncate TCGA and GEO data to only include same genes
#COAD.short=COAD.data[as.character(samegenes.names)]
b=match(samegenes.names,names(COAD.data))
```

```
COAD.short=COAD.data[b]
d=match(samegenes.names,names(GEO.data))
GEO.short=GEO.data[d]

#Normalize both data sets
COAD.norm=as.data.frame(scale(COAD.short))
GEO.norm=as.data.frame(scale(GEO.short))

#### PAMR analysis ####
#http://statweb.stanford.edu/~tibs/PAM/Rdist/doc/readme.html
library('pamr')
##PAM expects the data in an object (class=list) with
##components x (the expression matrix of genes by samples) and
##y, a vector of class labels.

x=as.matrix(GEO.norm)
x=unname(x, force = TRUE)
#Only look into Normal and IBD patients
GEO.pamr=list(x=t(x[c(1:8,39:53),]),
              y=c(rep('NOR',8),rep('IBD',15)),
              geneid=samegenes.names,
              samplelables=c(rownames(GEO.norm)[1:8],rownames(GEO.norm)[39:53]))
TCGA.pamr=list(x=unname(t(as.matrix(COAD.norm)),force = TRUE),
               y=rep('NOR',172), #assume TCGA samples are all normal?
               geneid=samegenes.names,
               samplelables=COAD.data$bcr_patient_barcode)

## Train the classifier
```

```
GEO.train <- pamr.train(GEO.pamr)

## Type name of object to see the results
GEO.train

## Cross-validate the classifier
GEO.results<- pamr.cv(GEO.train, GEO.pamr)
GEO.results #min nonzeros and errors?

## Plot the cross-validated error curves
pamr.plotcv(GEO.results)

## Compute the confusion matrix for a particular model
threshold=3.1
pamr.confusion(GEO.results, threshold)

## Plot the cross-validated class probabilities by class
pamr.plotcvprob(GEO.results, GEO.data, threshold)

## Plot the class centroids
pamr.plotcen(GEO.train, GEO.data, threshold)

## Make a gene plot of the most significant genes
pamr.geneplot(GEO.train, GEO.pamr, threshold)


# Estimate false discovery rates and plot them
fdr.obj<- pamr.fdr(GEO.train, GEO.pamr)
```

```
pamr.plotfdr(fdr.obj) #NULL

## List the significant genes
x=pamr.listgenes(GEO.train, GEO.pamr, threshold)
xtable(x)
##  prediction information, from a nearest shrunken centroid fit
pamr.predict(GEO.train, TCGA.pamr$x, threshold=3.1)
prob=pamr.predict(GEO.train, TCGA.pamr$x, threshold=2.5, type="posterior")

## A function to classify samples, allowing for an indeterminate (doubt) category
pamr.indeterminate(prob,mingap=.75)

x=matrix(nrow=10,ncol=172)
for (i in 1:10){
  x[i,]=pamr.predict(GEO.train, TCGA.pamr$x, threshold=(2.3+i*0.1))
  #1 is IBD and 2 is NOR
}
rownames(x) <- 2.3+0.1*c(1:10)
colnames(x)=t(COAD.data[1])
remov=0
for (i in 1:172){
  if(all(x[,i]==rep(1,10))){remov=append(remov,i)}
}
x=x[,-remov]
library(lattice)
levelplot(x)
```

Appendix C

PAMR CLASSIFICATION OF TCGA

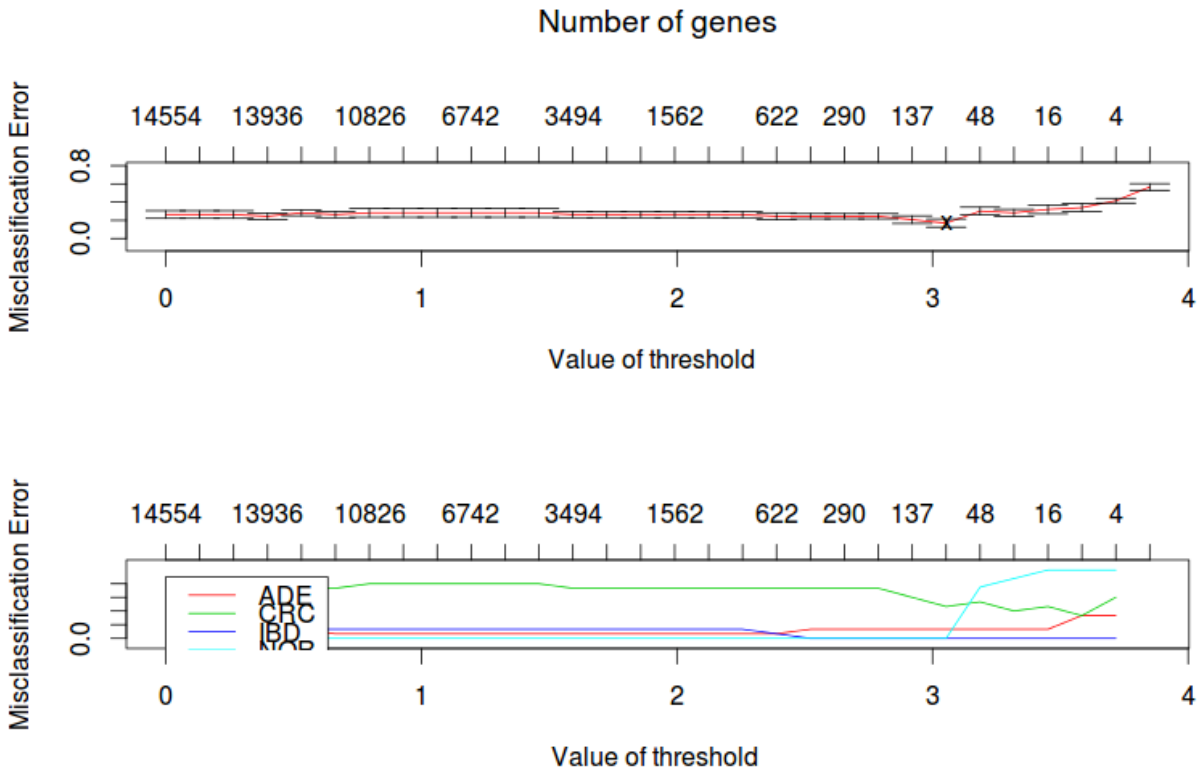

Figure (C.1) Plot the cross-validated error curves with all 4 categories present in GEO data.

```
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.3)
  [1] IBD IBD IBD IBD IBD ADE IBD CRC CRC CRC CRC CRC IBD CRC ADE CRC IBD
 [18] IBD CRC IBD IBD IBD CRC CRC ADE CRC IBD CRC CRC CRC ADE CRC IBD CRC
 [35] CRC IBD CRC IBD CRC ADE ADE CRC ADE CRC CRC ADE IBD CRC CRC IBD IBD
 [52] IBD ADE IBD IBD ADE ADE ADE IBD IBD CRC IBD IBD ADE CRC ADE IBD IBD
 [69] ADE ADE ADE IBD IBD ADE CRC ADE ADE CRC IBD ADE IBD CRC IBD ADE ADE
 [86] CRC ADE IBD ADE IBD IBD ADE CRC ADE ADE ADE IBD IBD ADE ADE ADE IBD
```

```
[103] CRC ADE IBD IBD IBD IBD IBD IBD IBD ADE IBD IBD IBD CRC IBD IBD IBD
[120] IBD IBD ADE IBD ADE ADE ADE CRC ADE ADE IBD IBD CRC ADE IBD IBD ADE
[137] ADE IBD ADE CRC ADE IBD ADE CRC ADE ADE IBD ADE ADE IBD ADE ADE ADE
[154] ADE CRC ADE IBD ADE ADE CRC ADE ADE ADE ADE CRC CRC ADE ADE IBD ADE
[171] CRC ADE
Levels: ADE CRC IBD NOR
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.4)
  [1] IBD CRC IBD IBD IBD ADE IBD CRC IBD CRC CRC CRC IBD CRC CRC CRC IBD
 [18] IBD CRC IBD IBD IBD CRC CRC ADE IBD IBD CRC CRC CRC ADE CRC IBD CRC
 [35] CRC IBD CRC IBD CRC ADE ADE CRC ADE CRC CRC ADE IBD CRC CRC IBD ADE
 [52] IBD ADE ADE IBD ADE ADE ADE IBD IBD ADE IBD IBD ADE CRC ADE IBD IBD
 [69] ADE ADE ADE IBD IBD ADE ADE ADE ADE CRC IBD ADE IBD CRC IBD ADE ADE
 [86] CRC ADE IBD IBD IBD IBD ADE CRC ADE ADE ADE IBD IBD ADE ADE ADE ADE
[103] IBD ADE IBD IBD IBD IBD IBD IBD IBD ADE IBD IBD IBD CRC IBD IBD IBD
[120] IBD ADE ADE IBD ADE ADE ADE CRC ADE ADE ADE IBD CRC IBD IBD IBD ADE
[137] ADE IBD ADE CRC ADE CRC ADE CRC ADE ADE IBD ADE ADE IBD ADE ADE ADE
[154] ADE CRC ADE IBD ADE ADE CRC ADE ADE ADE ADE ADE ADE ADE ADE IBD ADE
[171] CRC ADE
Levels: ADE CRC IBD NOR
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.5)
  [1] IBD CRC IBD IBD IBD ADE IBD ADE IBD CRC IBD IBD IBD CRC CRC CRC IBD
 [18] IBD CRC IBD IBD IBD CRC CRC ADE IBD CRC CRC CRC CRC ADE CRC IBD CRC
 [35] CRC IBD CRC IBD ADE ADE ADE CRC ADE ADE CRC ADE IBD IBD CRC IBD ADE
 [52] IBD ADE ADE IBD ADE ADE ADE IBD ADE ADE IBD IBD ADE CRC ADE IBD IBD
 [69] ADE ADE ADE IBD IBD ADE IBD ADE ADE CRC IBD CRC IBD CRC IBD ADE ADE
 [86] CRC ADE IBD ADE IBD IBD ADE IBD ADE ADE ADE IBD ADE ADE ADE ADE ADE
[103] IBD ADE IBD IBD IBD IBD IBD IBD IBD ADE IBD IBD IBD CRC IBD IBD IBD
[120] IBD ADE ADE IBD ADE ADE ADE ADE ADE ADE ADE IBD ADE IBD IBD IBD ADE
```

```
[137] ADE IBD ADE CRC ADE CRC ADE CRC ADE ADE IBD ADE ADE IBD ADE ADE ADE
[154] ADE ADE ADE IBD ADE ADE ADE IBD ADE ADE ADE ADE ADE ADE ADE IBD ADE
[171] CRC CRC
Levels: ADE CRC IBD NOR
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.6)
  [1] IBD CRC IBD IBD CRC IBD IBD ADE IBD CRC IBD IBD IBD CRC CRC CRC IBD
 [18] IBD CRC IBD IBD IBD CRC CRC ADE IBD CRC CRC CRC CRC ADE CRC IBD CRC
 [35] CRC IBD CRC IBD ADE ADE ADE IBD ADE ADE CRC ADE IBD IBD CRC IBD ADE
 [52] IBD ADE ADE IBD ADE ADE ADE IBD ADE ADE IBD IBD ADE CRC ADE IBD IBD
 [69] IBD ADE ADE IBD IBD ADE IBD IBD ADE CRC IBD CRC IBD CRC IBD ADE IBD
 [86] CRC ADE IBD ADE IBD IBD ADE IBD ADE ADE ADE IBD ADE ADE ADE ADE ADE
[103] IBD ADE IBD IBD IBD IBD IBD IBD IBD ADE IBD IBD IBD CRC IBD IBD IBD
[120] IBD ADE ADE IBD ADE IBD IBD ADE IBD ADE ADE IBD ADE IBD IBD IBD ADE
[137] ADE IBD ADE CRC ADE CRC ADE CRC ADE ADE IBD ADE ADE IBD ADE ADE ADE
[154] ADE ADE ADE IBD ADE ADE ADE IBD ADE ADE ADE ADE ADE ADE ADE IBD ADE
[171] CRC ADE
Levels: ADE CRC IBD NOR
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.7)
  [1] IBD CRC CRC IBD CRC IBD IBD IBD IBD CRC IBD IBD IBD IBD IBD CRC IBD
 [18] ADE CRC IBD IBD CRC CRC CRC ADE IBD CRC CRC IBD IBD ADE IBD IBD CRC
 [35] CRC IBD CRC CRC ADE ADE IBD IBD IBD ADE IBD ADE IBD IBD CRC IBD ADE
 [52] IBD IBD ADE CRC ADE ADE ADE IBD ADE ADE IBD IBD ADE CRC ADE IBD IBD
 [69] IBD ADE ADE IBD IBD IBD IBD IBD IBD CRC CRC CRC CRC CRC IBD ADE IBD
 [86] CRC ADE IBD ADE IBD IBD ADE IBD ADE ADE ADE IBD ADE ADE ADE ADE IBD
[103] IBD ADE IBD ADE IBD IBD IBD IBD IBD ADE IBD IBD IBD CRC IBD IBD CRC
[120] IBD ADE ADE CRC ADE IBD IBD ADE IBD ADE ADE ADE ADE IBD CRC IBD ADE
[137] ADE IBD ADE CRC ADE CRC ADE CRC ADE ADE IBD ADE ADE CRC ADE ADE ADE
[154] ADE ADE ADE IBD ADE ADE ADE ADE ADE ADE ADE ADE IBD ADE ADE IBD ADE
```

```
[171] CRC ADE
Levels: ADE CRC IBD NOR
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.8)
  [1] IBD ADE ADE ADE ADE IBD IBD ADE IBD ADE IBD IBD IBD IBD IBD ADE IBD
 [18] ADE ADE IBD ADE IBD ADE IBD IBD IBD ADE ADE IBD IBD IBD IBD IBD ADE
 [35] IBD IBD IBD ADE ADE ADE ADE IBD IBD ADE IBD ADE IBD ADE ADE IBD ADE
 [52] IBD IBD ADE ADE ADE IBD ADE IBD ADE IBD IBD IBD IBD ADE IBD IBD IBD
 [69] IBD IBD ADE ADE IBD IBD IBD IBD IBD ADE ADE ADE ADE ADE IBD ADE IBD
 [86] ADE IBD IBD ADE IBD IBD ADE IBD IBD ADE IBD IBD ADE IBD IBD ADE IBD
[103] IBD IBD ADE ADE IBD IBD IBD IBD IBD IBD IBD IBD IBD ADE IBD IBD ADE
[120] IBD IBD IBD ADE IBD IBD IBD IBD IBD ADE ADE ADE ADE IBD IBD ADE IBD
[137] ADE IBD ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE IBD IBD ADE IBD ADE
[154] ADE ADE ADE IBD ADE ADE ADE IBD ADE IBD ADE ADE IBD IBD ADE IBD ADE
[171] ADE IBD
Levels: ADE CRC IBD NOR
> pamr.predict(GEO.train2, TCGA.pamr$x, threshold=3.9)
  [1] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
 [18] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
 [35] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
 [52] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
 [69] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
 [86] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
[103] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
[120] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
[137] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
[154] ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE ADE
[171] ADE ADE
Levels: ADE CRC IBD NOR
```

threshold	ADE	CRC	IBD
3.3	38	42	65
3.4	44	38	63
3.5	50	30	65
3.6	44	29	72
3.7	40	33	72
3.8	51	94	0
3.9	145	0	0

Table (C.1) PAM classifier output. No samples classified as Normal.