Differential Gene Expression Analysis of Ulcerative Colitis and Crohn's Disease Samples and Predictive Modeling for Disease Diagnosis

Patricia Mason
December 2020

Background

Ulcerative Colitis and Crohn's Disease are inflammatory bowels diseases that cause gastrointestinal inflammation and tissue damage. Both are caused by a combination of inappropriate immune response to normal gut flora and environmental factors. Diagnosis is often made with biopsies from colonoscopies, which are invasive and pose an element of risk, as do all surgeries. Blood samples, in contrast, are easy to obtain, less expensive, and minimally invasive.

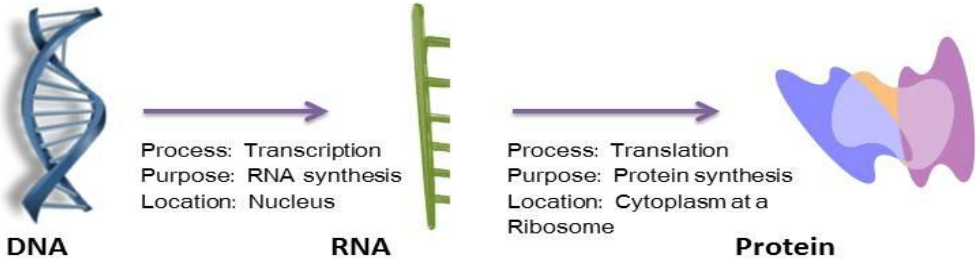
Problem Statement:

- Is there a standard set of genes that distinguish the Peripheral Blood Monocyte Cells (PBMCs) between Ulcerative Colitis, Crohn's Disease, and Normal patients.

- Based on the differentially expressed genes, Is it possible to create a model to predict the origin of the PBMCs.

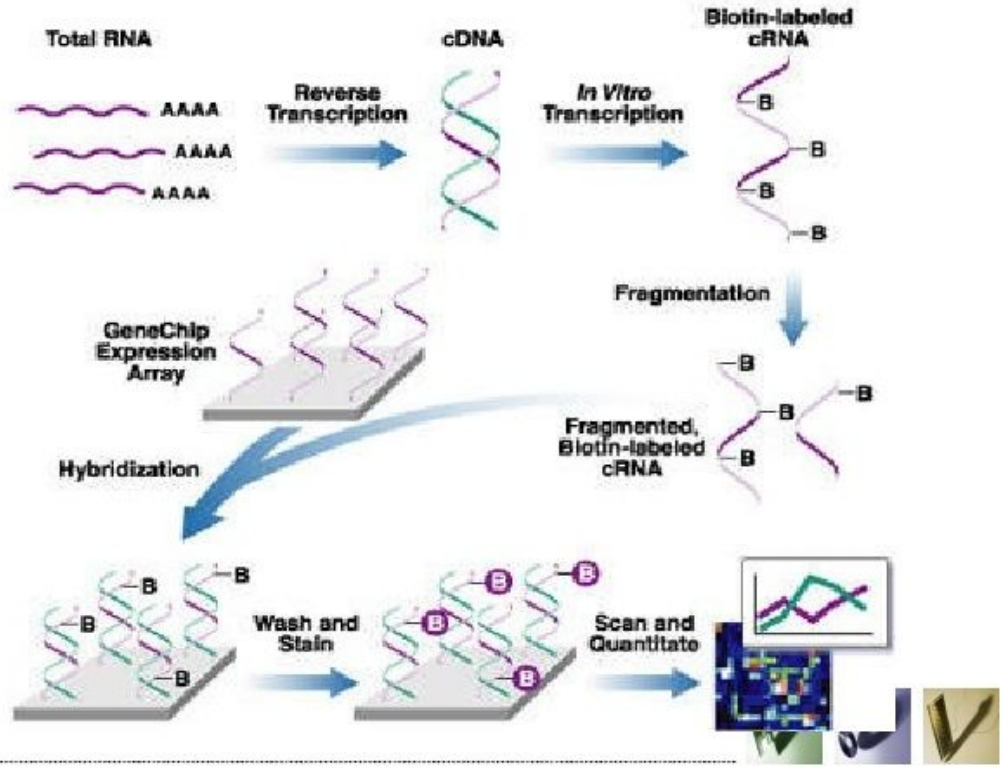
- Is it possible to replicate the analysis performed by Bioconductor/limma library: a package that has been an open-sourced project for 20 years.

The Central Dogma



DNA contains the original codes for making the proteins that living cells need. mRNA is a copy of a gene located on the DNA molecule. mRNA will leave the nucleus of the cell and the ribosome will read its coding sequences and put the appropriate amino acids together.

Affymetrix GeneChip experiment



Microarray (slideshare.net)

Comparison of Packages

Bioconductor (limma library)

Linear Regression

Empirical Bayes

UMAP

Genomatix

PCA

NCBI Resources How To Sign in to NCBI

NCBI
National Center for Biotechnology Information

All Databases | Search

COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: https://www.coronavirus.gov

Get the latest research information from NIH: https://www.nih.gov/coronavirus

Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

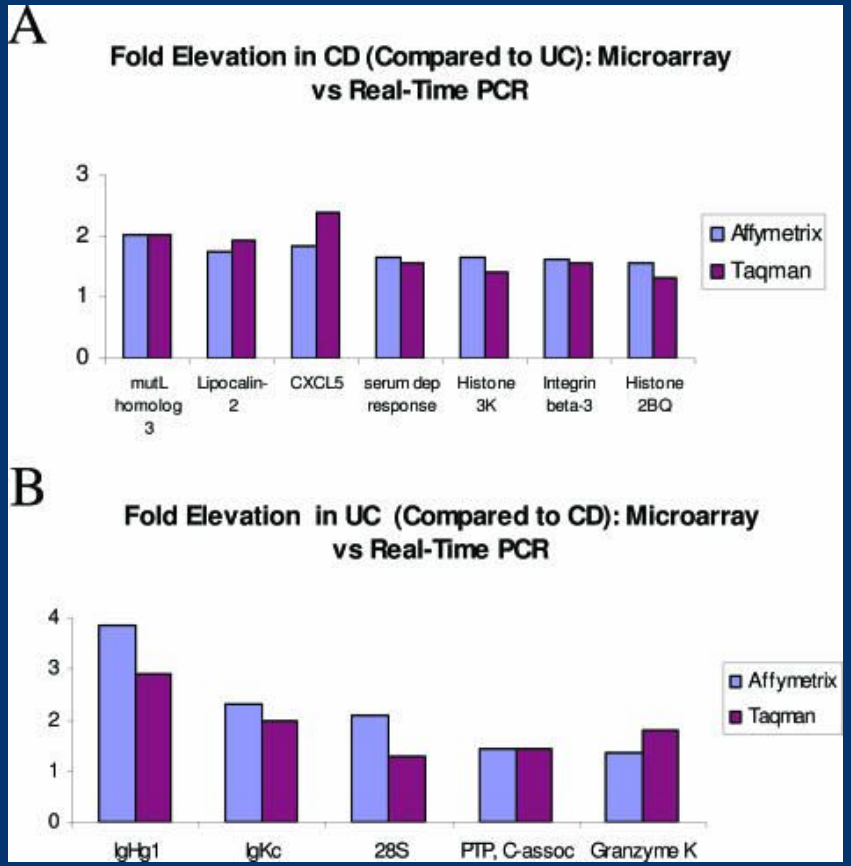
NCBI Virus: Test drive our new SARS-CoV-2 interactive data dashboard!
03 Dec 2020
Are you looking for SARS-CoV-2 sequence data? Look no further! The

December 9 Webinar: Using BLAST+ in Docker and on the cloud
30 Nov 2020
Join us on December 9, 2020 to learn about containerized BLAST+ in Docker

Read assembly and Annotation Pipeline Tool (RAPT) is available for use and testing
24 Nov 2020
We are excited to launch a beta version

More...

Verification of Results with TaqMan PCR



Data Source:

Burczynski, Michael E et al. “Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells.” *The Journal of molecular diagnostics : JMD* vol. 8,1 (2006): 51-61. doi:10.2353/jmoldx.2006.050079

Dataset contains microarray signals of
22,283 human genes
assayed against the Peripheral Blood Monocyte Cells (PBMCs) in
127 samples:
26 Ulcerative Colitis
59 Crohn’s Disease
96 Normal

Data Cleaning and Processing

Data downloaded in SOFT format (Simple Omnibus Format in Text)

→

Separate meta-data (clinical and demographic) from raw data into two different csv files

→

Create lists of accession numbers that correlated with sample types

→

Calculate log2 of the fluorescence values

→

Transpose the dataframe such that the genes are the features (columns and the samples are the rows

→

Scale with StandardScaler for PCA Analysis

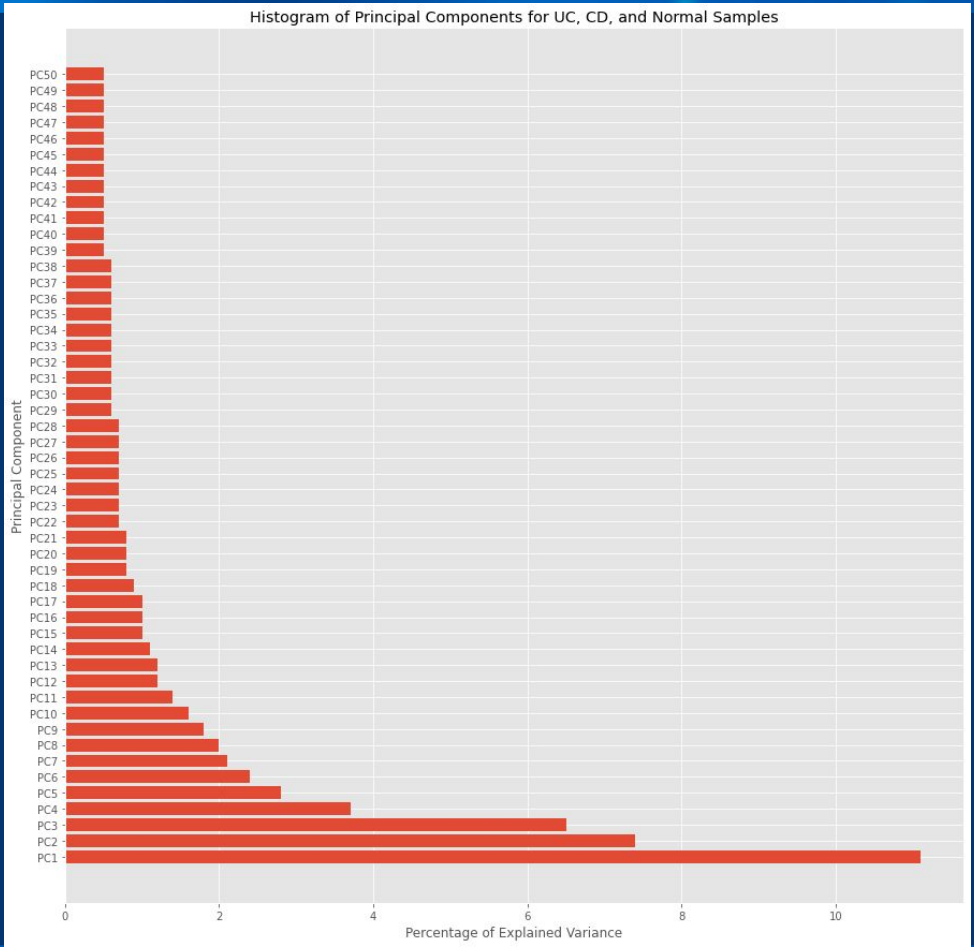
→

Add category column with UC, CD, NM for modeling

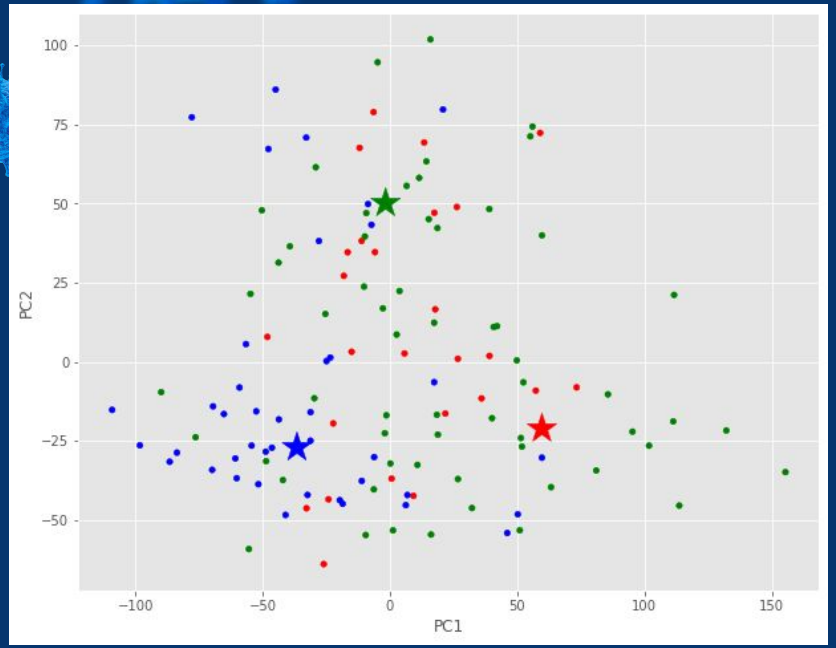
PCA Analysis of all three sample types: Ulcerative Colitis, Crohn's Disease, and Normal

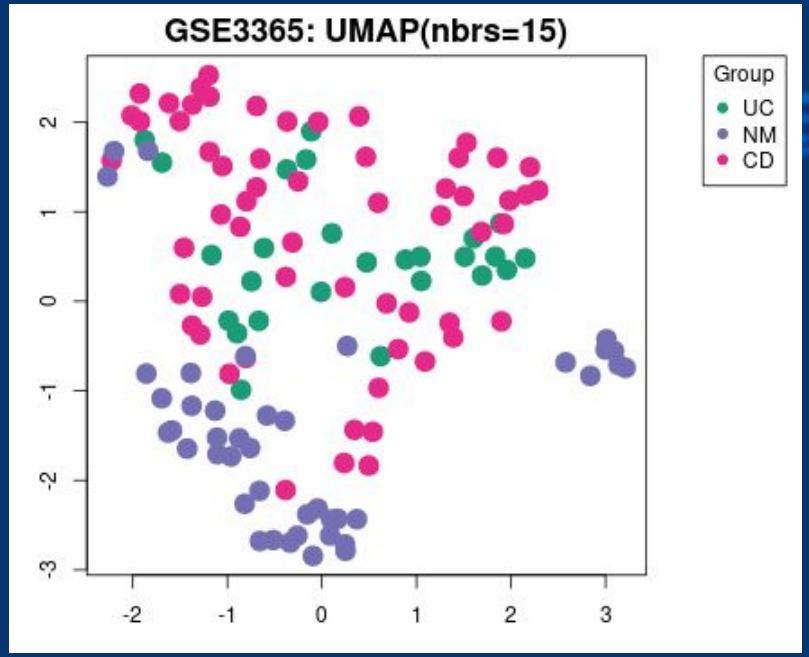
Cumulative Explained Variance - 50 Components (Genes)
0.70

Cumulative Explained Variance - 100 Components (Genes)
0.91


Histogram of Principal Components for UC, CD, and Normal Samples
Principal Component
PC50
PC49
PC48
PC47
PC46
PC45
PC44
PC43
PC42
PC41
PC40
PC39
PC38
PC37
PC36
PC35
PC34
PC33
PC32
PC31
PC30
PC29
PC28
PC27
PC26
PC25
PC24
PC23
PC22
PC21
PC20
PC19
PC18
PC17
PC16
PC15
PC14
PC13
PC12
PC11
PC10
PC9
PC8
PC7
PC6
PC5
PC4
PC3
PC2
PC1
0
2
4
6
8
10
Percentage of Explained Variance

K-Means Clustering and UMAP of Ulcerative Colitis, Crohn's Disease, and Normal Samples



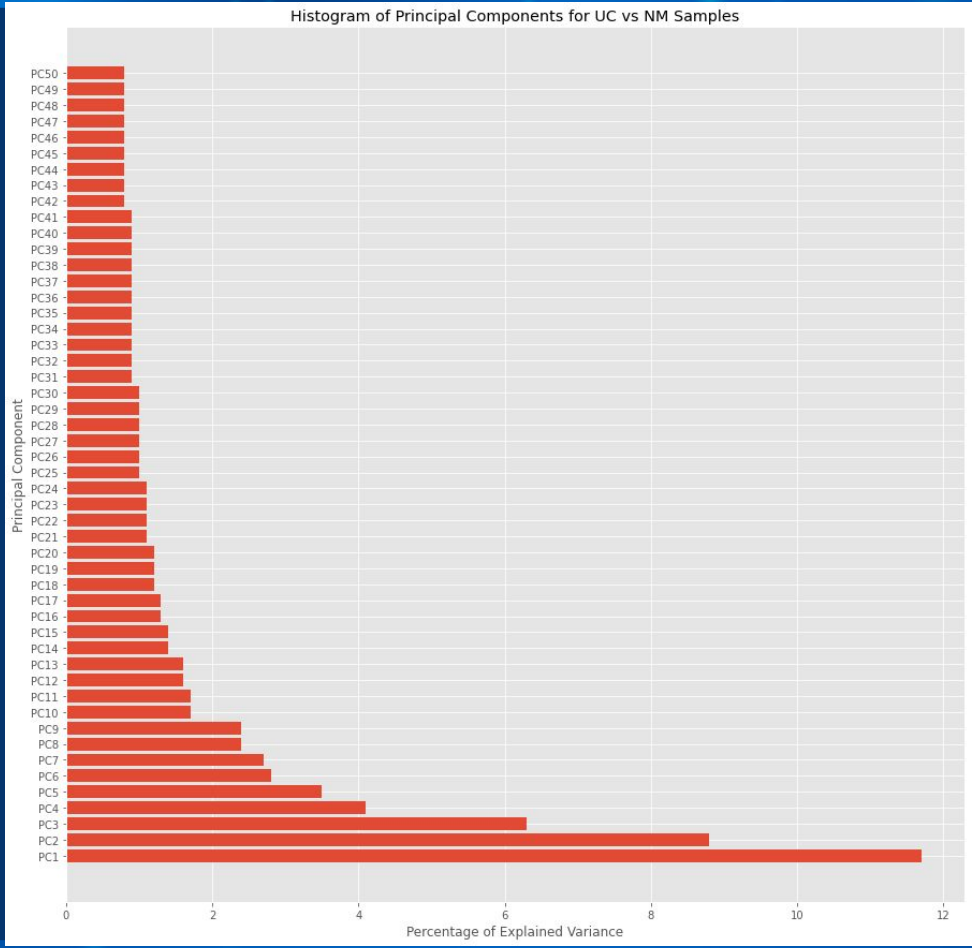


Silhouette scores
UC = 0.41
CD = 0.39
NM = 0.52

Top Ten Differentially Expressed Genes - Ulcerative Colitis, Crohn's Disease, and Normal Samples for PCA and Bioconductor

PCA Differential/UC, CD, NM	Bioconductor Differential/UC, CD, NM
solute carrier family 4 member 3	histone cluster 1, H2ac
stearoyl-CoA desaturase	histone cluster 1, H2bk
alkaline phosphatase, intestinal	brain abundant membrane attached signal protein 1
HAUS augmin like complex subunit 7///three pri...	progesterone receptor membrane component 1
ATPase H+ transporting V0 subunit a1	histone cluster 2, H2be
bone morphogenetic protein 1	folate receptor 1
Kruppel like factor 1	serpin family B member 2
TBC1 domain family member 10B	monocyte to macrophage differentiation associated
homeobox B5	amyloid beta precursor protein
crystallin beta B3	transmembrane protein 158 (gene/pseudogene)

N=0

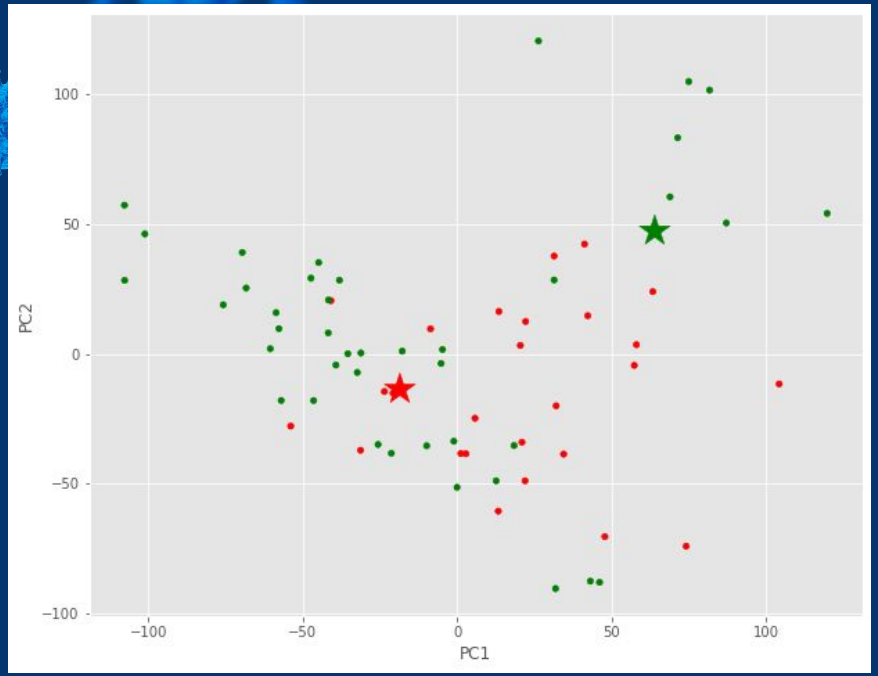

Histogram of Principal Components for UC vs NM Samples
Principal Component
PC50
PC49
PC48
PC47
PC46
PC45
PC44
PC43
PC42
PC41
PC40
PC39
PC38
PC37
PC36
PC35
PC34
PC33
PC32
PC31
PC30
PC29
PC28
PC27
PC26
PC25
PC24
PC23
PC22
PC21
PC20
PC19
PC18
PC17
PC16
PC15
PC14
PC13
PC12
PC11
PC10
PC9
PC8
PC7
PC6
PC5
PC4
PC3
PC2
PC1
0
2
4
6
8
10
12
Percentage of Explained Variance

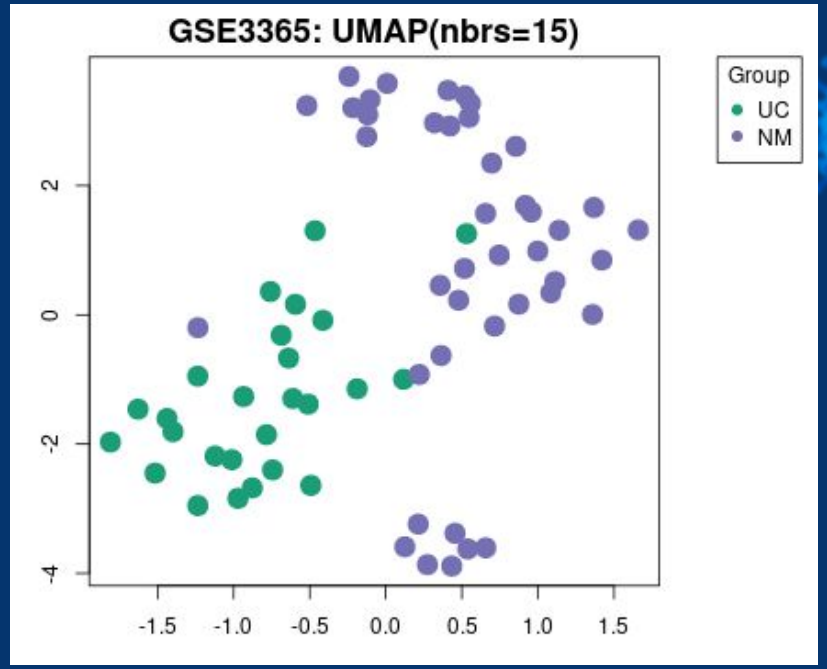
Top Ten Differentially Expressed Genes - Ulcerative Colitis and Normal Samples for PCA and Bioconductor

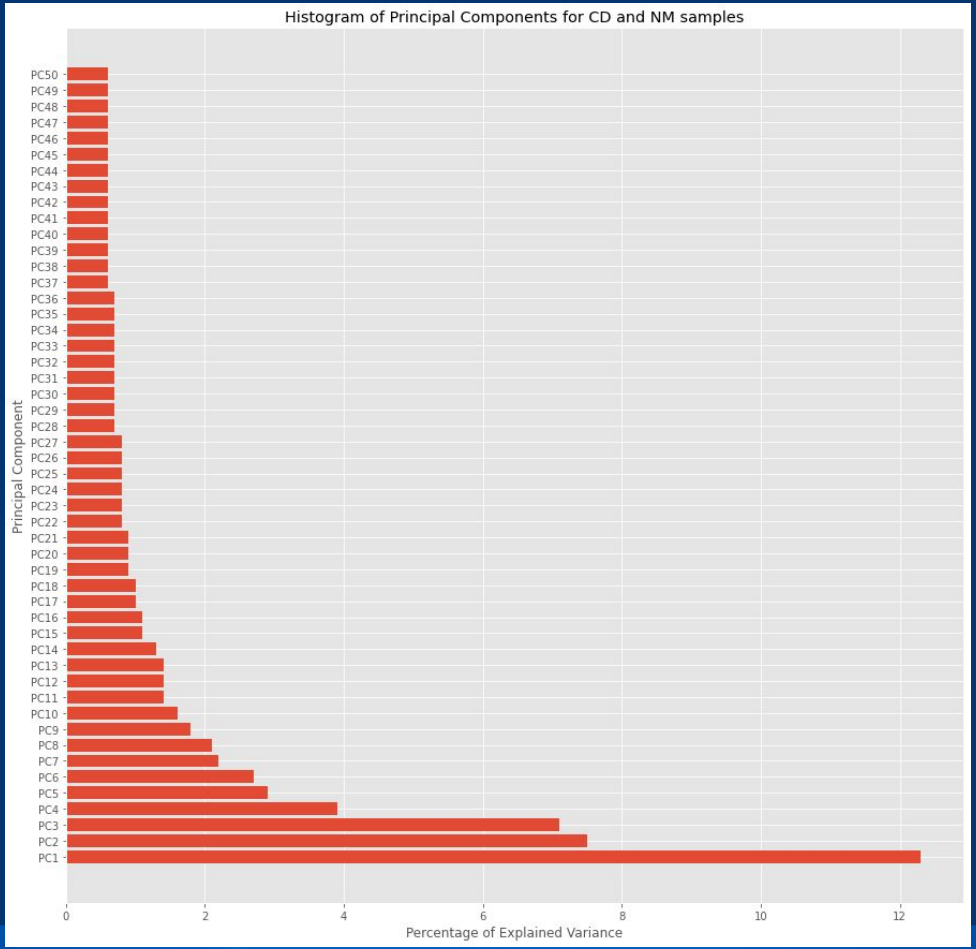
PCA Differential/UC vs NM	Bioconductor Differential/UC vs NM
ubiquitin conjugating enzyme E2 L3	folate receptor 1
NMD3 ribosome export adaptor	brain abundant membrane attached signal protein 1
proteasome subunit alpha 3	Affy Gene
cold shock domain containing E1	high mobility group box 1
small ubiquitin-like modifier 1	S100 calcium binding protein A11
basic leucine zipper and W2 domains 1	microRNA 8071-2///microRNA 8071-1///immunoglob...
poly(A) polymerase alpha	ARP5 actin-related protein 5 homolog
KRR1, small subunit processome component homolog	lysine demethylase 2A
transcription elongation factor B subunit 1	solute carrier family 22 member 4
heat shock protein family A (Hsp70) member 13	immunoglobulin kappa locus///immunoglobulin ka...

N =0

K-Means Clustering and UMAP(Bioconductor) of Ulcerative Colitis vs. Normal Samples





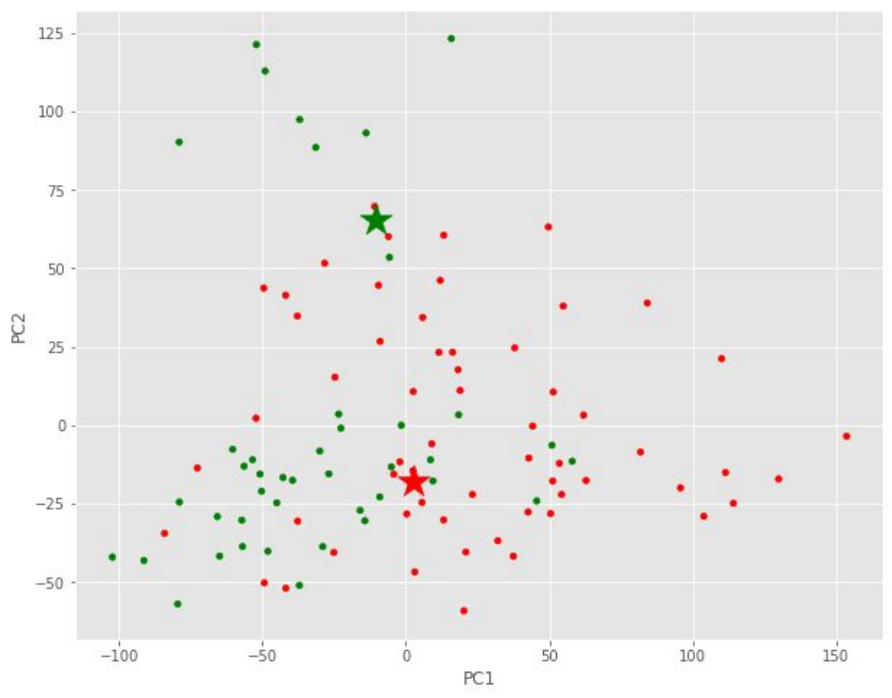

Histogram of Principal Components for CD and NM samples
Principal Component
PC50
PC49
PC48
PC47
PC46
PC45
PC44
PC43
PC42
PC41
PC40
PC39
PC38
PC37
PC36
PC35
PC34
PC33
PC32
PC31
PC30
PC29
PC28
PC27
PC26
PC25
PC24
PC23
PC22
PC21
PC20
PC19
PC18
PC17
PC16
PC15
PC14
PC13
PC12
PC11
PC10
PC9
PC8
PC7
PC6
PC5
PC4
PC3
PC2
PC1
0
2
4
6
8
10
12
Percentage of Explained Variance

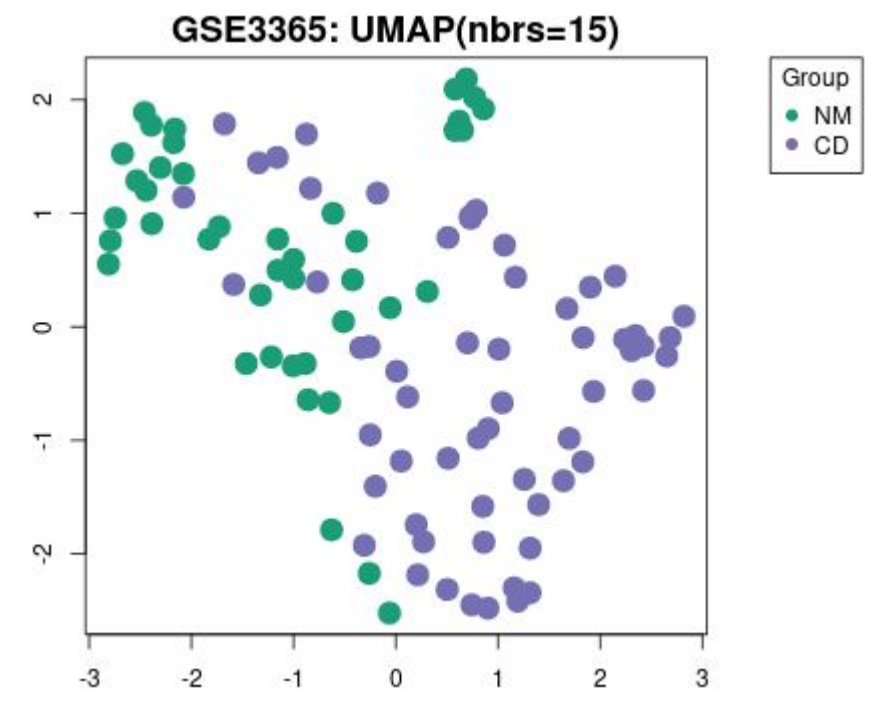
Top Ten Differentially Expressed Genes - Crohn's Disease and Normal Samples for PCA and Bioconductor

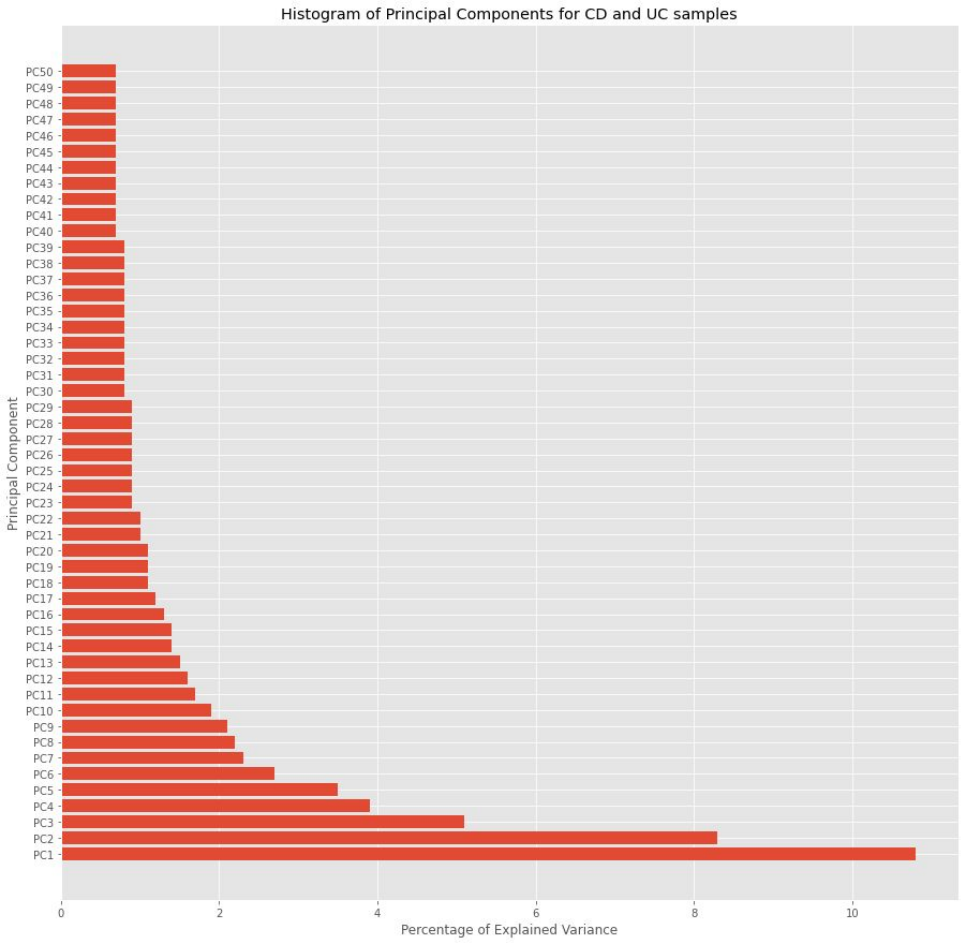
PCA Differential/CD vs NM	Bioconductor Differential/CD vs NM
solute carrier family 4 member 3	histone cluster 1, H2ac
stearoyl-CoA desaturase	histone cluster 1, H2bk
bone morphogenetic protein 1	progesterone receptor membrane component 1
alkaline phosphatase, intestinal	serpin family B member 2
crystallin beta B3	histone cluster 2, H2be
mucin 3B, cell surface associated///mucin 3A, ...	monocyte to macrophage differentiation associated
HAUS augmin like complex subunit 7///three pri...	brain abundant membrane attached signal protein 1
NEDD4 binding protein 1	transmembrane protein 158 (gene/pseudogene)
collagen type XI alpha 2 chain	histone cluster 1, H2bd
homeobox B5	amyloid beta precursor protein

N = 0

K-Means Clustering and UMAP(Bioconductor) of Crohn's Disease vs. Normal Samples





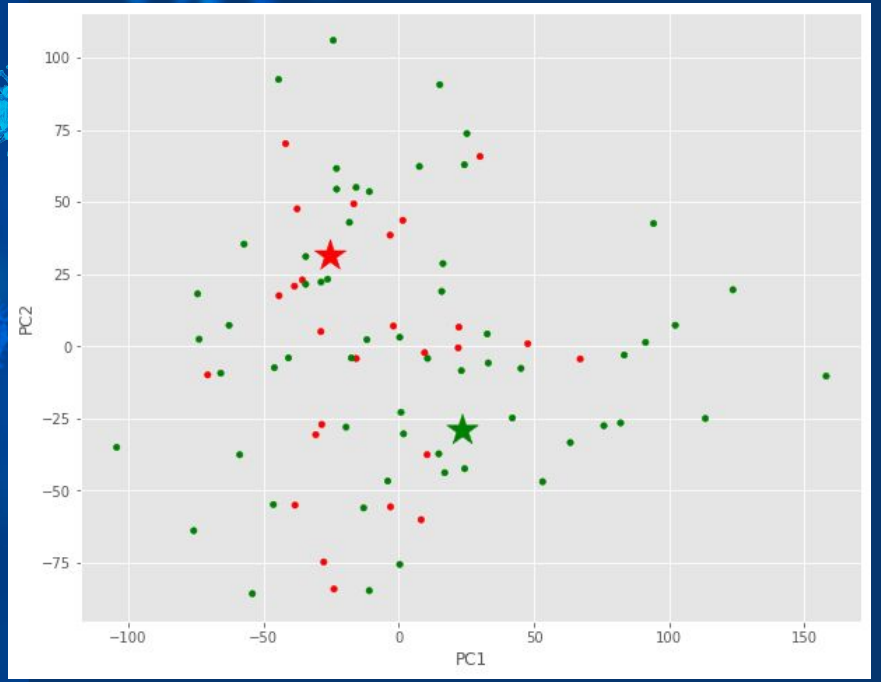

Histogram of Principal Components for CD and UC samples
Principal Component
PC50
PC49
PC48
PC47
PC46
PC45
PC44
PC43
PC42
PC41
PC40
PC39
PC38
PC37
PC36
PC35
PC34
PC33
PC32
PC31
PC30
PC29
PC28
PC27
PC26
PC25
PC24
PC23
PC22
PC21
PC20
PC19
PC18
PC17
PC16
PC15
PC14
PC13
PC12
PC11
PC10
PC9
PC8
PC7
PC6
PC5
PC4
PC3
PC2
PC1
0
2
4
6
8
10
Percentage of Explained Variance

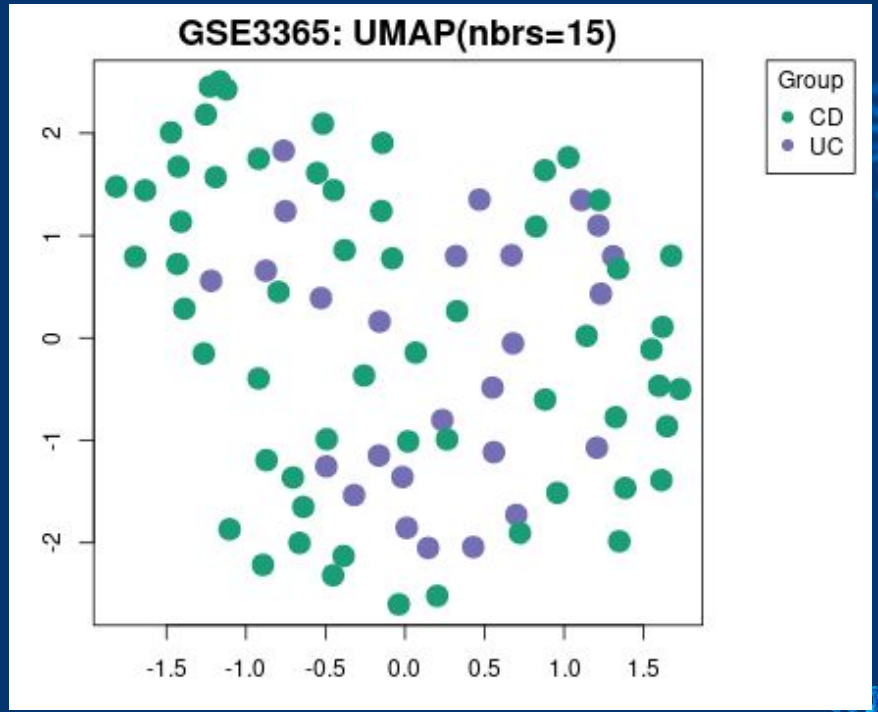
Top Ten Differentially Expressed Genes - Crohn's Disease and Ulcerative Colitis Samples for PCA and Bioconductor

PCA Differential/CD vs UC	Bioconductor Differential/CD vs UC
solute carrier family 4 member 3	microRNA 8071-2///microRNA 8071-1///immunoglob...
Rap guanine nucleotide exchange factor 3	folate receptor 1
crystallin beta B3	immunoglobulin heavy constant gamma 1 (G1m mar...
alkaline phosphatase, intestinal	Affy Gene
forkhead box N3	mutL homolog 3
bone morphogenetic protein 1	immunoglobulin kappa locus///immunoglobulin ka...
HAUS augmin like complex subunit 7///three pri...	immunoglobulin kappa locus///immunoglobulin ka...
homeobox B5	U2 small nuclear RNA auxiliary factor 2
WD repeat domain 55	immunoglobulin kappa locus///immunoglobulin ka...
NEDD4 binding protein 1	serum deprivation response

N = 0

K-Means Clustering and UMAP (Bioconductor - Crohn's Disease vs Ulcerative Colitis Samples





UC, CD, & NM Sample Classification Models

Model	Train	Test	Best Parameters
kNN with PCA	1.0	0.72	Metric:Euclidean n-neighbors: 1
kNN all genes	0.75	0.59	Metric: Manhattan n-neighbors: 21
Random Forest with PCA	1.0	0.65	Max-Depth: None n-estimators: 150
Random Forest All genes	1.0	0.70	Max-Depth: None n-estimators: 150

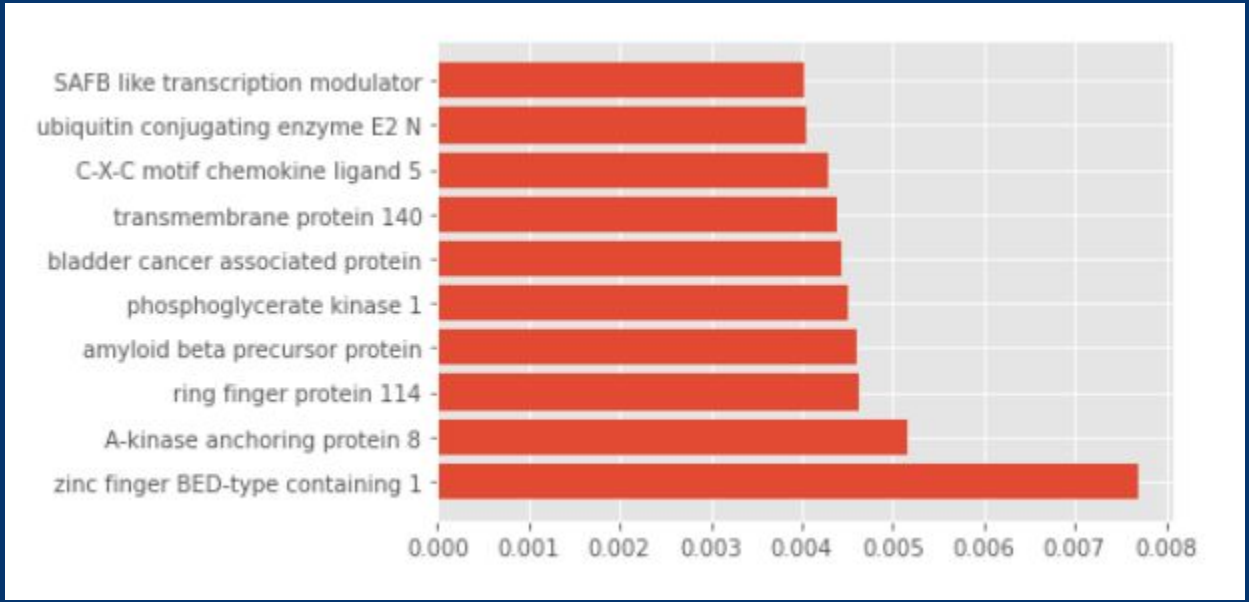
PCA = 100 Components

Binary Classification

Logistic Regression (default)	Baseline	Train	Test
UC vs NM with PCA	0.62	1.0	1.0
UC vs NM All Genes	0.62	1.0	1.0
CD vs NM with PCA	0.58	1.0	0.88
CD vs NM All Genes	0.58	1.0	0.92

PCA = 50 Components

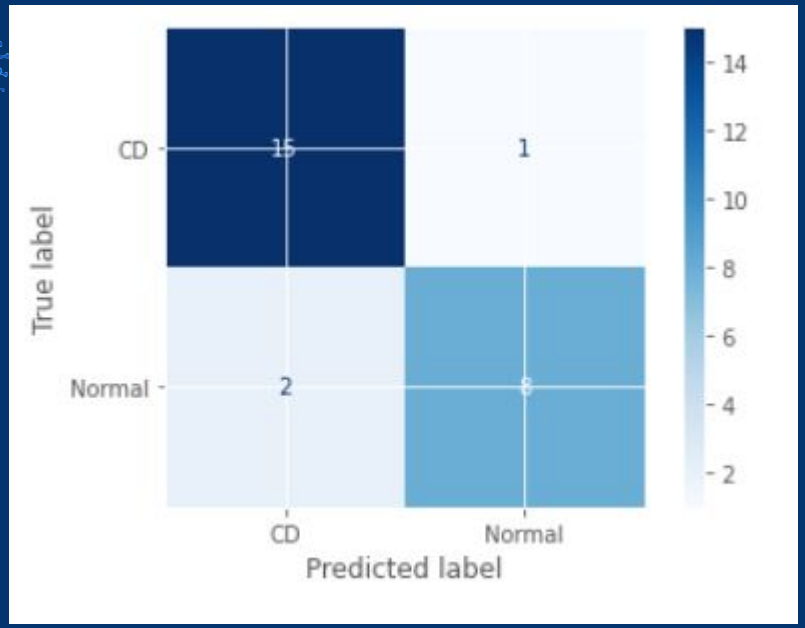
Top 10 Genes Differentially Expressed Between Ulcerative Colitis, Crohn's Disease and Normal Samples Random Forest



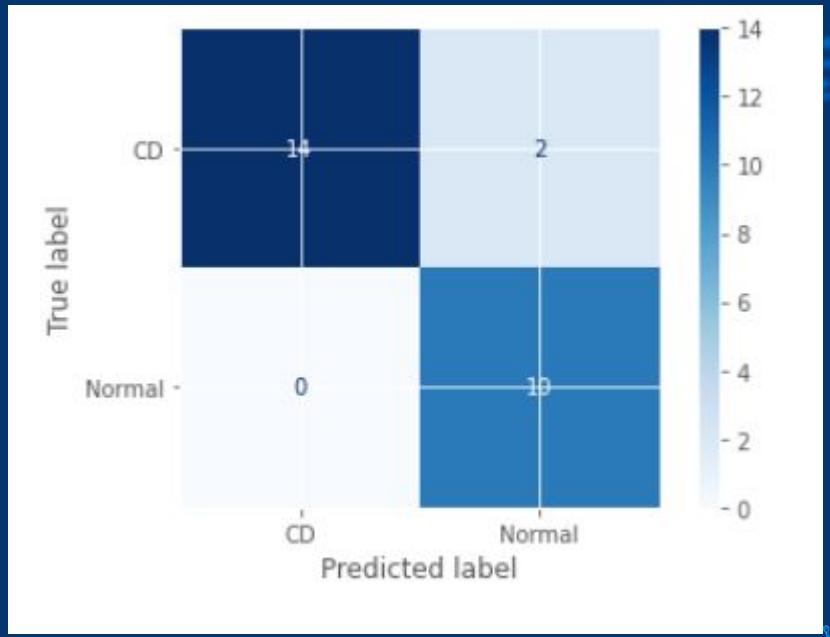
Top Ten Differentially Expressed Genes - Ulcerative Colitis, Crohn's Disease, Normal Samples for Random Forest and Bioconductor

RF Differential/UC,CD,NM	Bioconductor Differential/UC,CD,NM
zinc finger BED-type containing 1	histone cluster 1, H2ac
A-kinase anchoring protein 8	histone cluster 1, H2bk
ring finger protein 114	brain abundant membrane attached signal protein 1
amyloid beta precursor protein	progesterone receptor membrane component 1
phosphoglycerate kinase 1	histone cluster 2, H2be
bladder cancer associated protein	folate receptor 1
transmembrane protein 140	serpin family B member 2
C-X-C motif chemokine ligand 5	monocyte to macrophage differentiation associated
ubiquitin conjugating enzyme E2 N	amyloid beta precursor protein
SAFB like transcription modulator	transmembrane protein 158 (gene/pseudogene)

Ulcerative Colitis vs Normal Samples

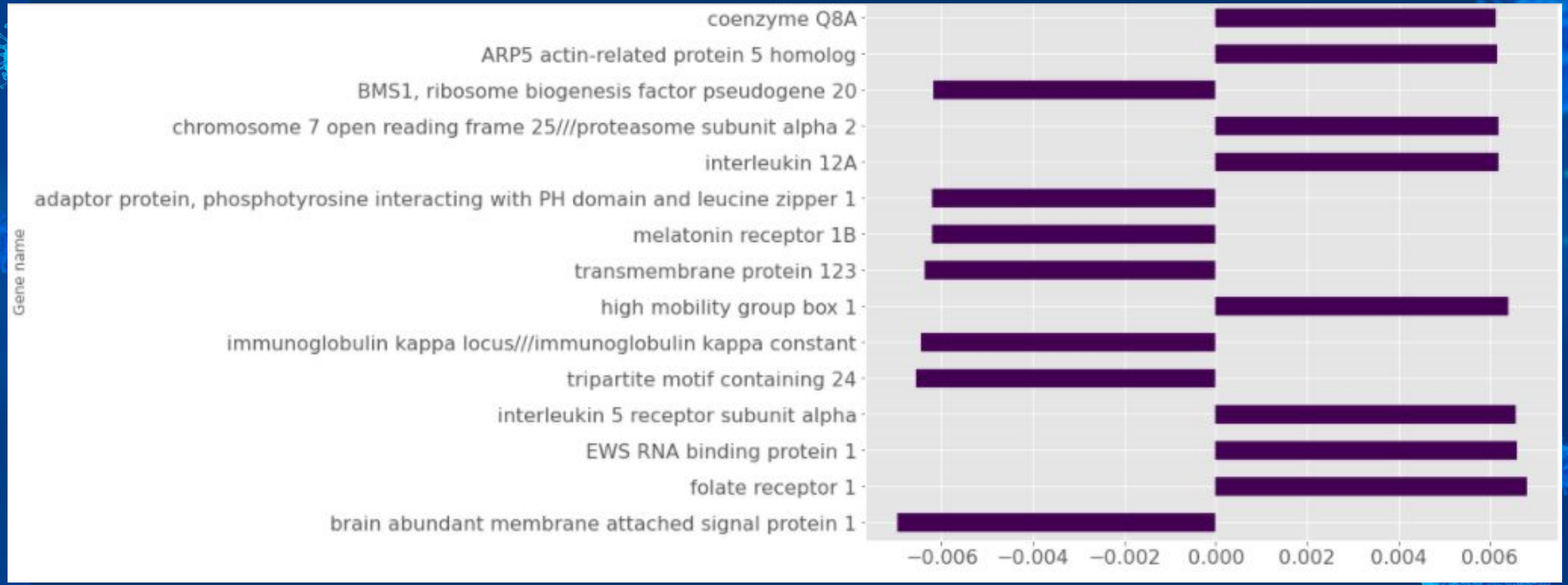


PCA (50) and Linear Regression



Linear Regression

Top 15 Genes Differentially Expressed Between Ulcerative Colitis and Normal Samples
Logistic Regression

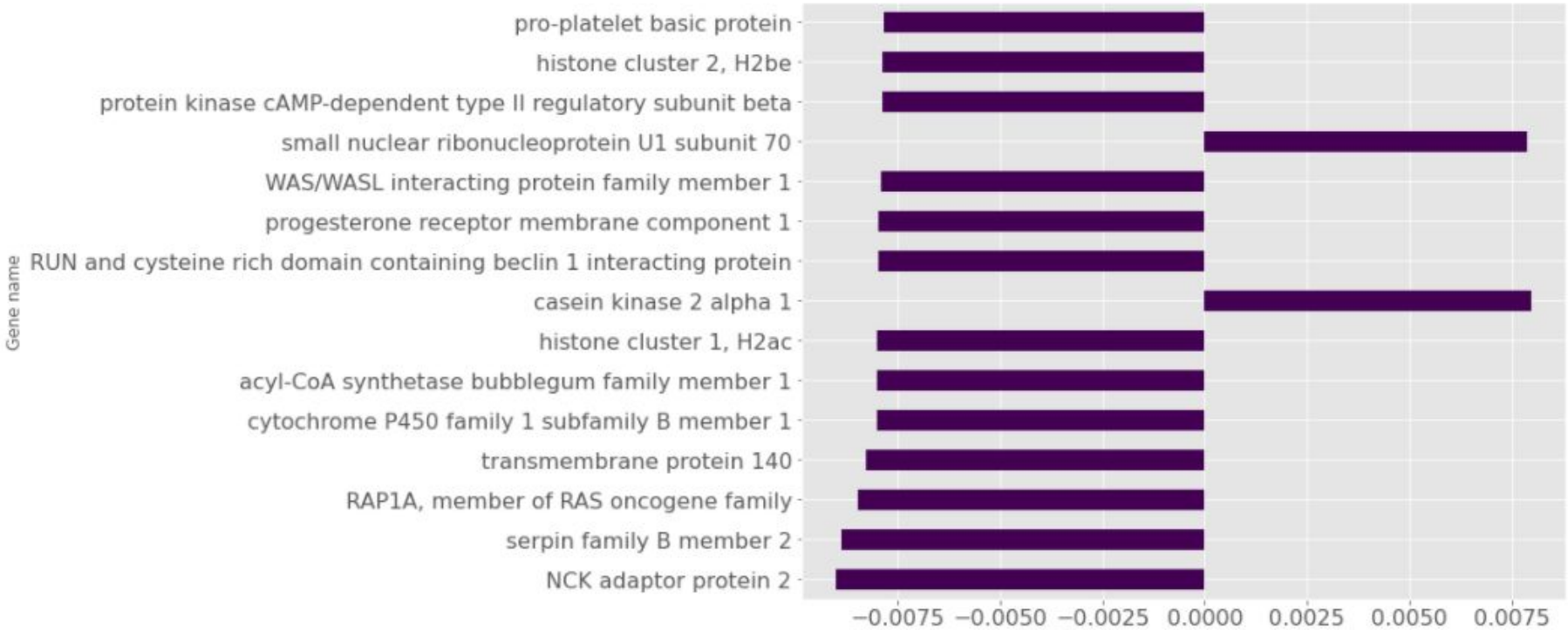


Top Ten Differentially Expressed Genes - Ulcerative Colitis vs Normal Samples for Linear Regression and Bioconductor

LR Differential/UC vs NM	Bioconductor Differential/UC vs NM
brain abundant membrane attached signal protein 1	folate receptor 1
folate receptor 1	brain abundant membrane attached signal protein 1
EWS RNA binding protein 1	Affy Gene
interleukin 5 receptor subunit alpha	high mobility group box 1
tripartite motif containing 24	S100 calcium binding protein A11
immunoglobulin kappa locus///immunoglobulin ka...	microRNA 8071-2///microRNA 8071-1///immunoglob...
high mobility group box 1	ARP5 actin-related protein 5 homolog
transmembrane protein 123	lysine demethylase 2A
melatonin receptor 1B	solute carrier family 22 member 4
adaptor protein, phosphotyrosine interacting w...	immunoglobulin kappa locus///immunoglobulin ka...

N = 4

Top 15 Genes Differentially Expressed Between CD and NM
Logistic Regression



Top Ten Differentially Expressed Genes - Crohn's Disease vs Normal Samples for Linear Regression and Bioconductor

LR Differential/CD vs NM	Bioconductor Differential/CD vs NM
NCK adaptor protein 2	histone cluster 1, H2ac
serpin family B member 2	histone cluster 1, H2bk
RAP1A, member of RAS oncogene family	progesterone receptor membrane component 1
transmembrane protein 140	serpin family B member 2
cytochrome P450 family 1 subfamily B member 1	histone cluster 2, H2be
acyl-CoA synthetase bubblegum family member 1	monocyte to macrophage differentiation associated
histone cluster 1, H2ac	brain abundant membrane attached signal protein 1
casein kinase 2 alpha 1	transmembrane protein 158 (gene/pseudogene)
RUN and cysteine rich domain containing beclin...	histone cluster 1, H2bd
progesterone receptor membrane component 1	amyloid beta precursor protein

N = 3

Conclusions:

- PCA does not produce any of the same genes as Bioconductor in any of the four comparisons that were performed-that is not to say the genes that PCA chooses as principal components are not important in the disease model

- PCA did not produce clear distinction of clusters in the K-means model, in spite of the fact that the number of clusters were known in advance. At the the same time, UMAP was not able to cluster the samples well either.

- Both kNN and Random Forest performed better on train and test sets with all genes as opposed to models based on PCA components

- Logistic Regression classification with and without PCA produced perfect scores for classifying ulcerative colitis samples. The model that used all the genes had 4 genes in common with the Bioconductor analysis.

- Logistic Regression classification with and without PCA produced perfect scores for train sets and 0.88 and 0.92,respectively, for each analysis of Crohn's disease samples in the test set. The model that used all the genes had 3 genes in common with the Bioconductor analysis.

- Differential gene expression analysis can be a multi-pronged approach using several methods, including PCA, PCA + logistic regression, logistic regression solo on top of the Bioconductor analysis

Further Study:

- Increase Sample Size: Blood draws are often easily obtained from voluntary blood from donation
- TaqMan PCR verification of differentially expressed genes identified by various models in this study
- Use other microarray data to compare/contrast the different models used in this study
- Compare the ulcerative colitis and Crohn's disease samples through additional modeling

Thank you!

Global Instructors: Dan, Noelle, and Riley

Local Instructors: Caroline, Kai, and Heather

Outcomes: Rachel

Support: Naida

Guest Speakers

Fellow Students in the East Coast Cohort