**Transcriptomic Triumph: Pioneering Faster Clinical Diagnosis for Ulcerative Colitis**

20.440 - Project Report 1

*Saravanan Thangarajan*

**Introduction**:

Ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD) [[1]](https://www.zotero.org/google-docs/?iuEiax) . UC is a disease that is the result of a combination of factors including genetics, environmental, diet, and immune system function. Crohn's and other IBDs can occur from inflammation at any part of the gastrointestinal tract as opposed to UC which is limited to the colon [[2]](https://www.zotero.org/google-docs/?87vcty). As a result of this, diffuse mucosal inflammation and mixture of complex inflammatory mediators mucosal ulcerations develop [[3]](https://www.zotero.org/google-docs/?LgZQIu). This pathology is best observed through histology in areas where there is increase in the neutrophil counts in the lamina propria and crypts and are deemed micro-abscesses [[3]](https://www.zotero.org/google-docs/?tW3emB).

The immune pathology of ulcerative colitis (UC) is characterized by a complex interplay between the host's immune system and various environmental factors. UC begins with an abnormal immune response to the intestinal microbiota in genetically predisposed individuals. This can be triggered or worsened by environmental and dietary factors. The body's immune system mistakenly attacks harmless substances in the colon, releasing various inflammatory mediators. These include cytokines such as tumor necrosis factor-alpha (TNF-α), interleukins (e.g., IL-1, IL-6, IL-13), and other substances that promote inflammation [[4]](https://www.zotero.org/google-docs/?mlZAGc). Based on the literature, in Ulcerative Colitis (UC), specific genes are more likely to be significant than others. Studies have identified genes like CCR7, CXCL10, CXCL9, CXCL8, CCL19, CTLA4, CCR1, CD69, CD163, IL7R, PECAM1, TLR8, and TLR2 as hub genes associated with immune-infiltrated cells in UC [[5]](https://www.zotero.org/google-docs/?xNC7Gu).

Biological networks and pathways in UC involve immune response pathways such as Toll-like receptor signaling, IL-17 signaling, and chemokine signaling. These pathways are enriched with differentially expressed genes (DEGs) that contribute to the inflammatory processes seen in UC. Biological networks and pathways in UC involve immune response pathways such as Toll-like receptor signaling, IL-17 signaling, and chemokine signaling [[6]](https://www.zotero.org/google-docs/?JjaNhE). These pathways are enriched with differentially expressed genes that contribute to the inflammatory processes seen in UC. Certain genes can serve as prognosis markers in UC. For instance, the identified hub genes associated with immune-infiltrated cells play a role in disease progression and can serve as markers for the diagnosis and treatment of UC [[5]](https://www.zotero.org/google-docs/?1a2doC). These genes provide insights into the molecular mechanisms involved in UC and can aid in validating drug targets.

A biomarker has yet to be discovered to diagnose ulcerative colitis. The current method is to combine imaging studies, medical history and symptoms. This can be a lengthy process and delay the time for treatment. It is estimated that around 1 million people have UC in the United States. Just since the start of the year there have been over a thousand papers that have been published using the search term “ulcerative colitis”. The question remains that despite there being significant research groups focused on UC why is there still not a streamlined diagnostic or detectable biomarker available for clinicians? Xiong et al aimed to use visceral fat distribution from CT images to distinguish between Crohn's Disease and UC but fat distribution is heavily influenced by other factors and can confound results [[7]](https://www.zotero.org/google-docs/?H0E1Fr). An additional subset of studies have focused on investigating treatment options and efficacy of drugs [[8]](https://www.zotero.org/google-docs/?Vbek3K)[[9]](https://www.zotero.org/google-docs/?nlVXkN). Some possible biomarkers that have been investigated is anti-integrin alpha-v-beta-6 autoantibodies however they lack continued prevalence in UC patients. Only 30.8% of the UC cohort tested had detectable levels of this autoantibody [[10]](https://www.zotero.org/google-docs/?4imzhr)***.***

Additionally, studies with larger cohorts that provide more predictive power lack an algorithm that takes into account medical history to make a final prediction. Serum histone H4 was observed to have a significant decrease in UC patients compared to healthy individuals but once a cut off value was established (via ROC curve generation) the final diagnostic accuracy was 78.2% [[11]](https://www.zotero.org/google-docs/?bTqYr6). The question remains that if more advanced machine learning algorithms were applied to their study could they see better accuracy with using serum histone H4 as a potential biomarker. This would be one of the few projects that would be able to use transcriptomic data and advanced machine learning algorithms to better diagnostic UC in patients.

Noble et al published a study in 2008 and performed a genome-wide expression analysis of 98 patients (67 with ulcerative colitis and 31 control subjects) [[12]](https://www.zotero.org/google-docs/?sKkraT). A total of 215 biopsies were taken (up to 5 anatomical locations per patient) and a dataset containing RNA microarray results was analyzed. The group performed unsupervised hierarchical clustering and then once genes of interest were determined focused on the biological implications and using the genes to understand the pathology of UC.

Although RNA microarray data has seen decreased popularity compared to its higher throughput counterpart of RNAseq the Noble study provides a rich and interesting dataset to uncover novel findings. One important factor when maintaining clinical translation is ease of sample collection and the robustness of the testing process to ensure accurate results. Having multiple anatomical location samples from a single patient can determine if there is an ideal biopsy location for diagnosis.

Additionally, an advantage of using a slightly dated dataset is that the analysis and experiment was done prior to the advance and accessibility of machine learning methods and algorithms. The computational modeling ceases after discovering DEGs and provides an opportunity to make decision trees and consider multiple DEGs and parameters for classifying UC patients and controls. This will also provide an opportunity to grow the test data set by using more recent publicly available datasets.

Our team proposes to perform a more robust machine learning analysis on the data to further investigate the diagnostic potential. Nobel claims differential expression of SAA1, DEFA5, DEFA6, MMP3, and MMP7 was observed between UC and healthy control biopsies using hierarchical clustering [[12]](https://www.zotero.org/google-docs/?aYkYvY). **Our hypothesis is to reanalyze the RNA microanalysis novel, to discover DEGs, that can be used in a combinatorial fashion (ML methods) that might be beneficial in the clinical setting.**

**Methods:**

Nobel et al generated a patient cohort of 108 (n=67 UC) patients and during endoscopy generated a 202 sample set with four different biopsy locations. These locations and number of samples include the terminal ileum (n=10), ascending colon (n=50), descending colon (n=58), and sigmoid colon (n=84). Total RNA extraction was performed using the micro total RNA isolation kit and then was purified using the RNeasy Mini Kit (Qiagen). The Agilent Whole Human Genome microarrays were used to measure the expression of 33,296 genes. The microarray images were analyzed and normalex using the Stratagene Universal Human Reference and the log intensities of each sample was deposited to Gene Expression Omnibus (GEO) with the accession code of GSE11223.

Taking the appropriate files from GEO and cleaning the expression and patient information data hierarchical clustering was performed. Samples and genes were clustered separately (horizontal and vertical respectively) and via euclidean clustering measurements. To determine the effectiveness of the clustering the silhouette coefficient, Calinski-Harabasz score, Davies-Bouldin score, Adjusted Rand Index (ARI) and accuracy were calculated when needed.

The differential gene expression analysis was performed by calculating the average expression value for the healthy and UC samples and the fold change was calculated by dividing the UC average by the healthy average. T-tests were performed to determine if there was a significant difference between the UC and healthy samples’ expression. The p-values were adjusted using the Bonferroni-correction for multiple hypothesis testing. Volcano plots were generated and display the log2(fold change (UC/healthy) and -log(adjusted p-val). The same process was repeated for each biopsy sample. The adjusted p-val had to be <0.05 and the absolute value of the log2(fold-change) had to be greater than three. The fold change cut off was an arbitrary selection but was optimized to generate a gene of interest (GOI) list of a workable value for the purpose of the analysis.

Random forest decision trees are a common type of supervised machine learning and were created using the sklearn

**Results:**

Using the already normalized gene expression data hierarchical clustering was performed to see if full RNA micro-array data analysis successfully clustered the samples. To determine the success of the clustering different sample parameters were taken into account. These parameters included disease status (‘healthy’ or ‘UC’), by biopsy type (sigmoid colon, descending colon, terminal ileum, and ascending colon), and by paired disease status and biopsy location (14 combinations).

To quantify and compare the clustering the silhouette coefficient, calinski-harabasz score, davies bouldin score were calculated. The tabulated results are presented in Table 1.

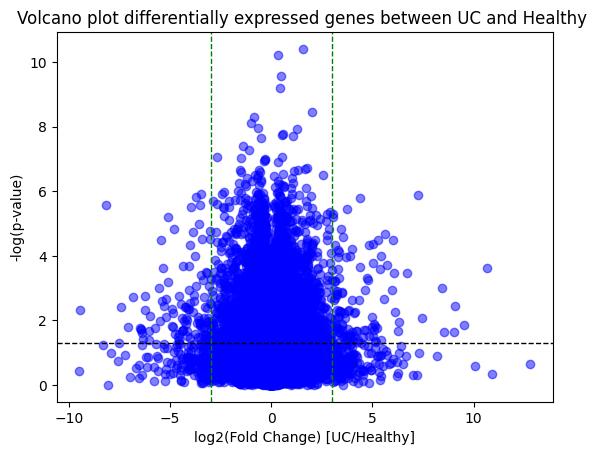
**Table 1:** Silhouette coefficients, Calinkski-Harabasz score, and Davies Bouldin score for the varying numbers of clusters for a given parameter (n= number of clusters).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Disease Status (n=2) | Biopsy Type (n=4) | Disease & Biopsy (n=14) |
| Silhouette Coefficient | 0.271 | 0.0942 | 0.0751 |
| Calinkski-Harabasz Score | 43.8 | 26.7 | 11.8 |
| Davies Bouldin Score | 1.66 | 2.32 | 2.23 |

Overall the effectiveness of clustering based on all gene’s expression was very poor with silhouette coefficients that were very far from 1. This was especially present when clustering by biopsy type with a silhouette coefficient of almost zero. This indicates that there is little dis-similarity between the two clusters and not a distinct separation between the types of biopsy. This proves that using RNAseq results as a diagnostic tool are ineffective at predicting the tissue type of the sample. This further motivates the need to investigate for a single gene expression paired with patient history to provide a clearer depiction of the disease present.

The Calinkski-Harabasz score is most effective when comparing the results of different clustering solutions across the same dataset. \_\_\_\_\_\_. For this reason we found it most beneficial to continue our analysis based on disease status in an attempt to generate a robust machine learning model to diagnose patients.

A differential gene expression analysis was performed. The p-values were adjusted using a Bonferonni-correction and the absolute log2(Fold Change) threshold was set to three. This was an arbitrary cut off established to maintain a workable list of genes. The volcano plot is below in Figure 1.



**Figure 1:** Volcano plot of all genes. Green lines represent the log2(FC) threshold of +/-3 and the black horizontal line is equivalent to the adjusted p-values of 0.05.

There are 628 genes that exceed the threshold for significance and fold change between all UC and healthy samples. The resulting genes of interest (GOI) were then sorted by the absolute fold change values. The top 4 genes that demonstrate the most extreme and statistically significant fold change are S100P, MGC23985 (now known as C5orf46), MGRN1, and BSCL2.

To narrow the search for one candidate gene a similar differential analysis was performed for each of the four biopsy types. The same thresholds and Bonferonni-corrections for multiple hypothesis testing was applied. The following total number of differentially expressed genes between UC and healthy patients for each sample location is sigmoid colon (87), descending colon (57), terminal ileum (12), and ascending colon (49).

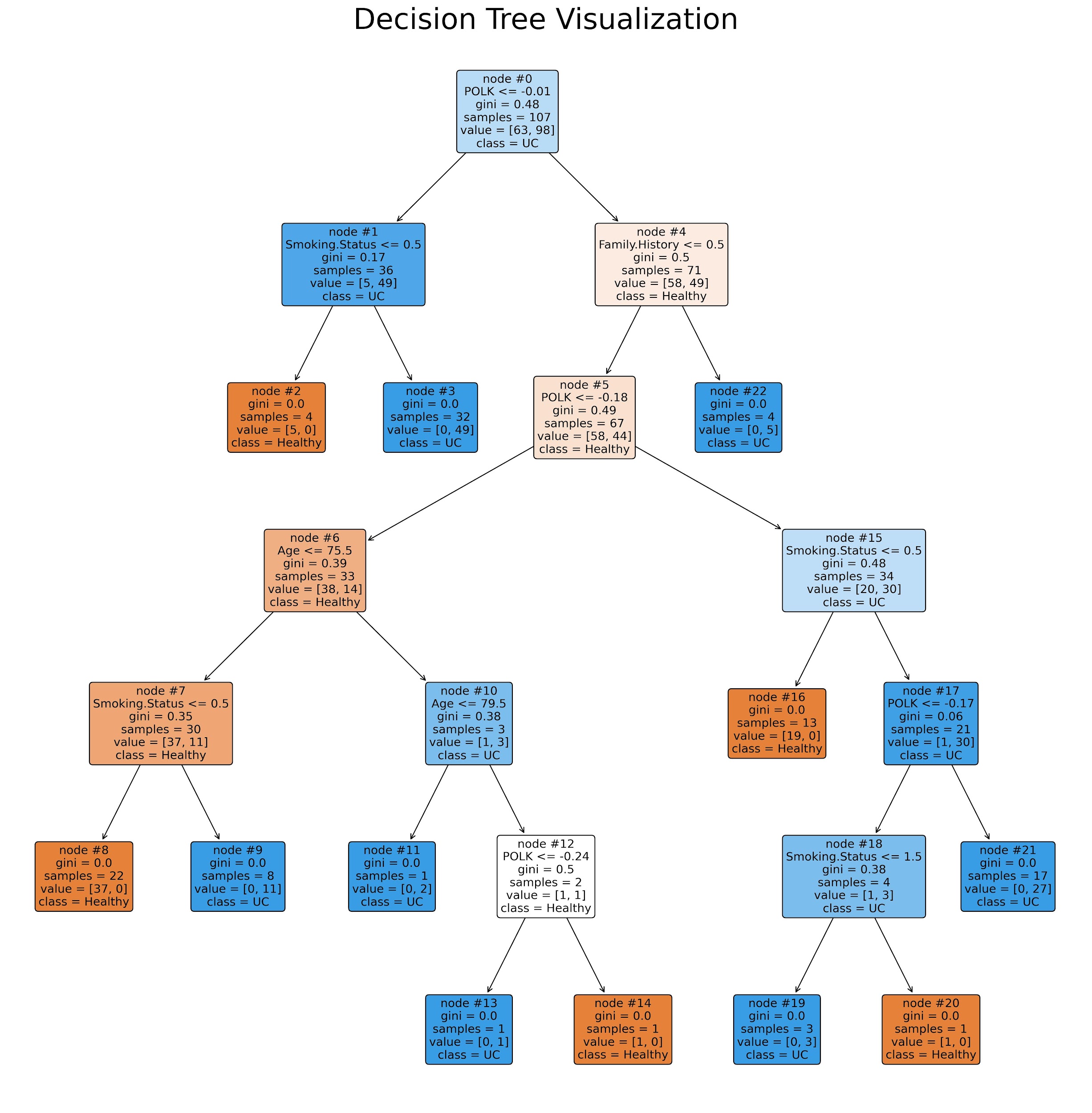


**Figure 2:** Visual depiction of overlap of statistically significant differentially expressed genes by sample type.

Due to variation in the number of samples per biopsy this could result in the adjusted p-values falling outside of the threshold of significance and removed from the GOI per biopsy. To account for this all of the gene were combined into a final GOI of six genes. Each gene expression was appointed as a distinct feature and the samples were classified as UC or healthy for all samples regardless of sample location. The random forest tree classifier is represented in Figure S5. The decision tree is 80.5% accurate with a sensitivity of 96.3% and specificity of 50.0%. The high sensitivity of the tree

Special attention was given to the first few nodes of the decision tree. The features with the largest gini values were BAIAP3 (gini = 0.47) and POLK (gini = 0.49)(Fig S5). To continue narrowing our search for a candidate gene of interest to design a diagnostic test these two genes were deemed the top candidates. A supplementary analysis to determine the co-occurance of expression between healthy and UC was performed (Fig S6). There was little observed correlation between the expression of the two genes therefore are not interchangeable as a sole biomarker for diagnosis. This was somewhat predicted since there was not full overlap between the sample types and significant gene expression but worth confirming.

Next to fully integrate both patient history and single gene expression and hopefully improve the performance of a decision tree the following parameters from the patient information was used as classification features: age (at time of publication), ethnicity, family history, smoking status, with POLK and BAIAP3 expression. The classification tree is below in Figure 3. The final accuracy, sensitivity, and specificity all came to be 100%.



**Figure 3:** Random forest classifier using normalized expression values for the following features (age, ethnicity, family history of UC, smoking status, and POLK and BAIAP3 expression). The feature name, threshold, number of samples, and classification are in each node. Orange nodes represent a healthy classification and blue represent a predicted UC diagnosis.

**Discussion**:

There have been fewer studies connecting BAIAP3 to colorectal health. However, Lui et al in 2019 found eight genes (one being BAIAP3) with promoter hypermethylation and decreased expression in sessile serrated adenomas (SSA) compared to SSAs with dysplasia (SSADs). SSAs is a type of polyp found in 20-30% of colorectal cancers and once dysplasia develops leads to rapid progression of the cancer. Although encouraging to see a connection of BAIAP3 expression and colon health this study limited their expression data between cancer types and did not include non-disease individuals in their sample cohort.

* Insert lit review of connection between POLK and colon inflammation. Has very little tissue specificity which is good for when we are looking at colon locations out of the scope of the Nobel et al study

With the conclusion that POLK and BAIAP3 would be appropriate gene expression biomarkers a sanity check was performed to validate the findings. POLK has disease associations with some cancer types such as prostate and endometrial but has a less clear implication with

* + Papers are in google drive folder

To further validate the model the next steps would be to find additional studies that performed RNAseq (preferable over RNA microarray) and determine if POLK and BAIAP3 display significant expression differences between UC and healthy patients from a variety of biopsy locations. Once the model has been further validated the following wet lab validation should occur. A multiplex RT-qPCR should be designed for POLK, BAIAP3, an endogenous control (usually a housekeeping gene such as ACTB or HPRT), and an exogenous control (a randomly generated sequence not found in the human genome). Once the appropriate RNA tissue extraction protocols have been developed and optimized the RT-qPCR should be performed on a sample set of patients with a variety of features and characteristics.

New thresholds should be applied to the RT-qPCR results since the units will differ from an exploratory RNA microarray analysis such as the one performed here and gene expression data (typically ΔΔCts). Either a decision tree classifier can be regenerated or receiver operator characteristic (ROC) curves can be generated to determine the most appropriate threshold to maximize the sensitivity.

This tool would help improve UC diagnostics since sample collection could be done during a routine outpatient procedure such as a colonoscopy and quickly and inexpensively processed and performed. Additional routes of investigation would be to see if POLK and BAIAP3 protein levels are detectable and differentiable in stool samples and therefore could be included as a part of other ‘at-home-collection-kits’ that screen for colon cancer (ie ExactSciences Cologuard).

Methods:

* Combine and clean microarray data
  + Looked at age distribution of UC and healthy samples
    - Both normally distributed but they’re means are statistically significantly different (t-test)
    - Need to be cautious for confounding age effects of DEGs between UC and healthy
* Cluster normalized reads by sample
* Evaluate clustering: Silhouette coefficient, Calinkski-harabasz index, davies bouldin index
  + By disease (n=2)
  + By biopsy/sample type (n=4)
  + By disease, biopsy and sample type (n=14)
* Silhouette score of n=2 was the highest (0.27054)
  + Calculate the adjusted rand index (ARI = - 0.0216)
  + Calculate the accuracy (56.4%)
* Determine differentially expressed genes between UC and healthy samples
  + Separate microarray data frame into Normal and UC
  + Calculate p-values between Normal and UC for each gene (Bonferonni adjusted p-vals)
  + Calculate the fold change values for Normal and UC
* Generate a volcano plot of log2(FC) vs log10(p-val) of all genes
  + Filter gene list with parameters of p<0.05 and |FC| >3
    - |FC| >3 was arbitrary and produced a manageable sized genes of interest (GOI) list for the purpose of our analysis
* Final GOI list contains 628 genes
  + ***Perform GSEA on gene list?***
  + ***See how it relates to biology and pathway analysis of UC?***
* **Regenerate heat map with the 628 GOI (This should probably be the figure for this section. Going to be easier to interpret than the entire heatmap since that doesn’t provide much insight)**
* **See if we can recapitulate Nobel clustering (split into 3? Left and right colon or something like that?) A way to validate or invalidate their results.**
* **Separate data into biopsy location** 
  + For a biopsy location determine differentially expressed genes between UC and healthy (calculate fold change and adjusted p-vlaues)
  + Generate volcano plot with equivalent cut off values for p and FC
  + Generate GOI list per biopsy
  + How much overlap is there in the GOIs of biopsy location?
  + This is going to be the key conclusion for if we can translate to clinic
* Using patient info and expression of 1-2 GOI generate a random forest tree
  + Idea would be that the input expression level of 1-2 GOI would go in and the ML will generate a diagnosis based on the patients age, sex, date of diagnosis etc…

**References**:

[[1] J. Huang, F. Wang, and X. Tang, “Uncovering the shared molecule and mechanism between ulcerative colitis and atherosclerosis: an integrative genomic analysis,” *Front Immunol*, vol. 14, p. 1219457, Aug. 2023, doi: 10.3389/fimmu.2023.1219457.](https://www.zotero.org/google-docs/?QqccSC)

[[2] I. Cleynen *et al.*, “Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: a genetic association study,” *Lancet*, vol. 387, no. 10014, pp. 156–167, Jan. 2016, doi: 10.1016/S0140-6736(15)00465-1.](https://www.zotero.org/google-docs/?QqccSC)

[[3] “Unravelling the pathogenesis of inflammatory bowel disease | Nature.” Accessed: Mar. 14, 2024. [Online]. Available: https://www.nature.com/articles/nature06005](https://www.zotero.org/google-docs/?QqccSC)

[[4] V. Gross, T. Andus, H. G. Leser, M. Roth, and J. Schölmerich, “Inflammatory mediators in chronic inflammatory bowel diseases,” *Klin Wochenschr*, vol. 69, no. 21–23, pp. 981–987, Dec. 1991, doi: 10.1007/BF01645143.](https://www.zotero.org/google-docs/?QqccSC)

[[5] “Identification of hub genes and immune infiltration in ulcerative colitis using bioinformatics | Scientific Reports.” Accessed: Mar. 15, 2024. [Online]. Available: https://www.nature.com/articles/s41598-023-33292-y](https://www.zotero.org/google-docs/?QqccSC)

[[6] W. Shi, R. Zou, M. Yang, J. Wen, and R. Lai, “Analysis of Genes Involved in Ulcerative Colitis Activity and Tumorigenesis Through Systematic Mining of Gene Co-expression Networks,” *Front. Physiol.*, vol. 10, May 2019, doi: 10.3389/fphys.2019.00662.](https://www.zotero.org/google-docs/?QqccSC)

[[7] Z. Xiong *et al.*, “Radiological biomarkers reflecting visceral fat distribution help distinguish inflammatory bowel disease subtypes: a multicenter cross-sectional study,” *Insights Imaging*, vol. 15, no. 1, p. 70, Mar. 2024, doi: 10.1186/s13244-024-01640-9.](https://www.zotero.org/google-docs/?QqccSC)

[[8] K. M. Nguyen and E. K. Wright, “Vancomycin and Ustekinumab Combination Therapy in Acute Ulcerative Colitis,” *ACG Case Rep J*, vol. 11, no. 3, p. e01302, Mar. 2024, doi: 10.14309/crj.0000000000001302.](https://www.zotero.org/google-docs/?QqccSC)

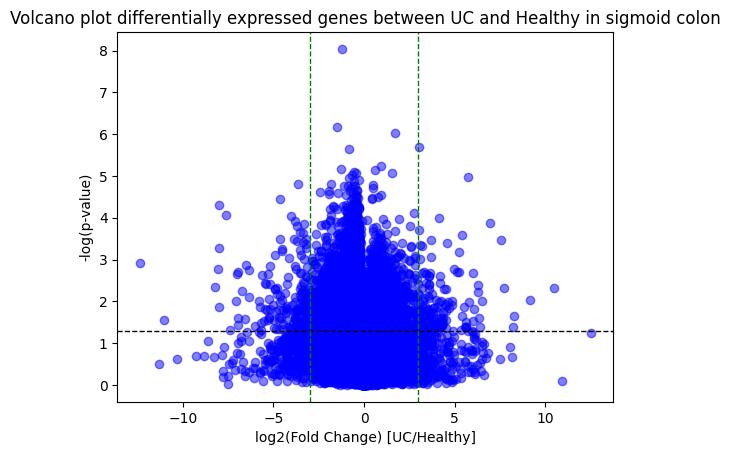
[[9] I. R. L. Andriolo, L. Venzon, and L. M. da Silva, “Perspectives About Ascorbic Acid to Treat Inflammatory Bowel Diseases,” *Drug Res (Stuttg)*, Mar. 2024, doi: 10.1055/a-2263-1388.](https://www.zotero.org/google-docs/?QqccSC)

[[10] M. Yokode *et al.*, “Anti-integrin αvβ6 autoantibodies are a potential biomarker for ulcerative colitis-like immune checkpoint inhibitor-induced colitis,” *Br J Cancer*, Mar. 2024, doi: 10.1038/s41416-024-02647-1.](https://www.zotero.org/google-docs/?QqccSC)

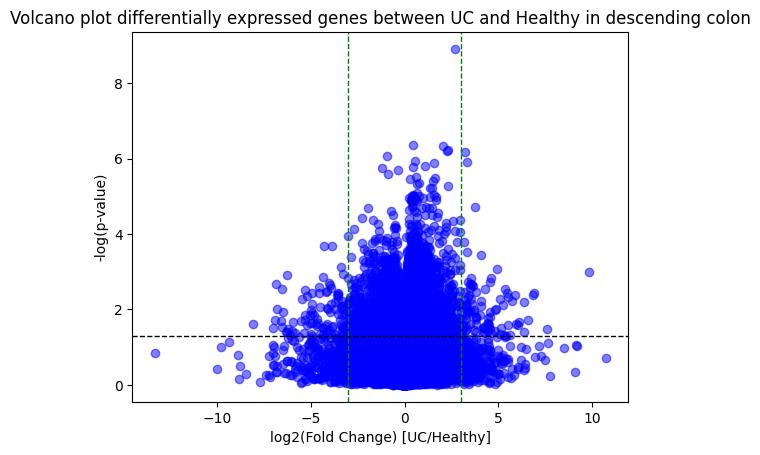
[[11] İ. Küçük, F. Özçelik, Y. Yazgan, İ. Yılmaz, M. Kaplan, and İ. Yıldırım, “Can Serum Histone H4 Level Be a Biomarker in Ulcerative Colitis?,” *Turk J Gastroenterol*, vol. 35, no. 1, pp. 4–10, Jan. 2024, doi: 10.5152/tjg.2024.22385.](https://www.zotero.org/google-docs/?QqccSC)

[[12] C. L. Noble *et al.*, “Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis,” *Gut*, vol. 57, no. 10, pp. 1398–1405, Oct. 2008, doi: 10.1136/gut.2008.148395.](https://www.zotero.org/google-docs/?QqccSC)

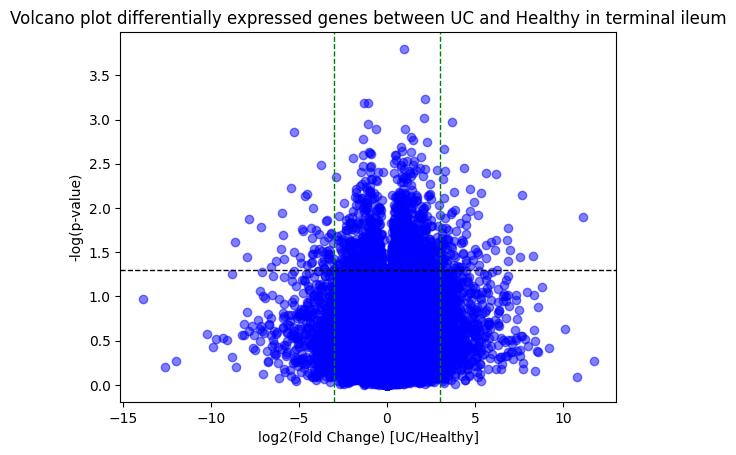
**Supplemental Figures:**

****

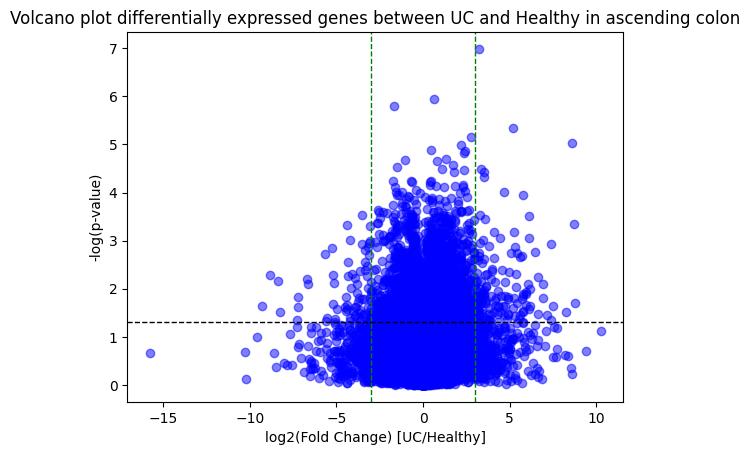
**Figure S1:** Volcano plot of differentially expressed genes within all sigmoid colon samples



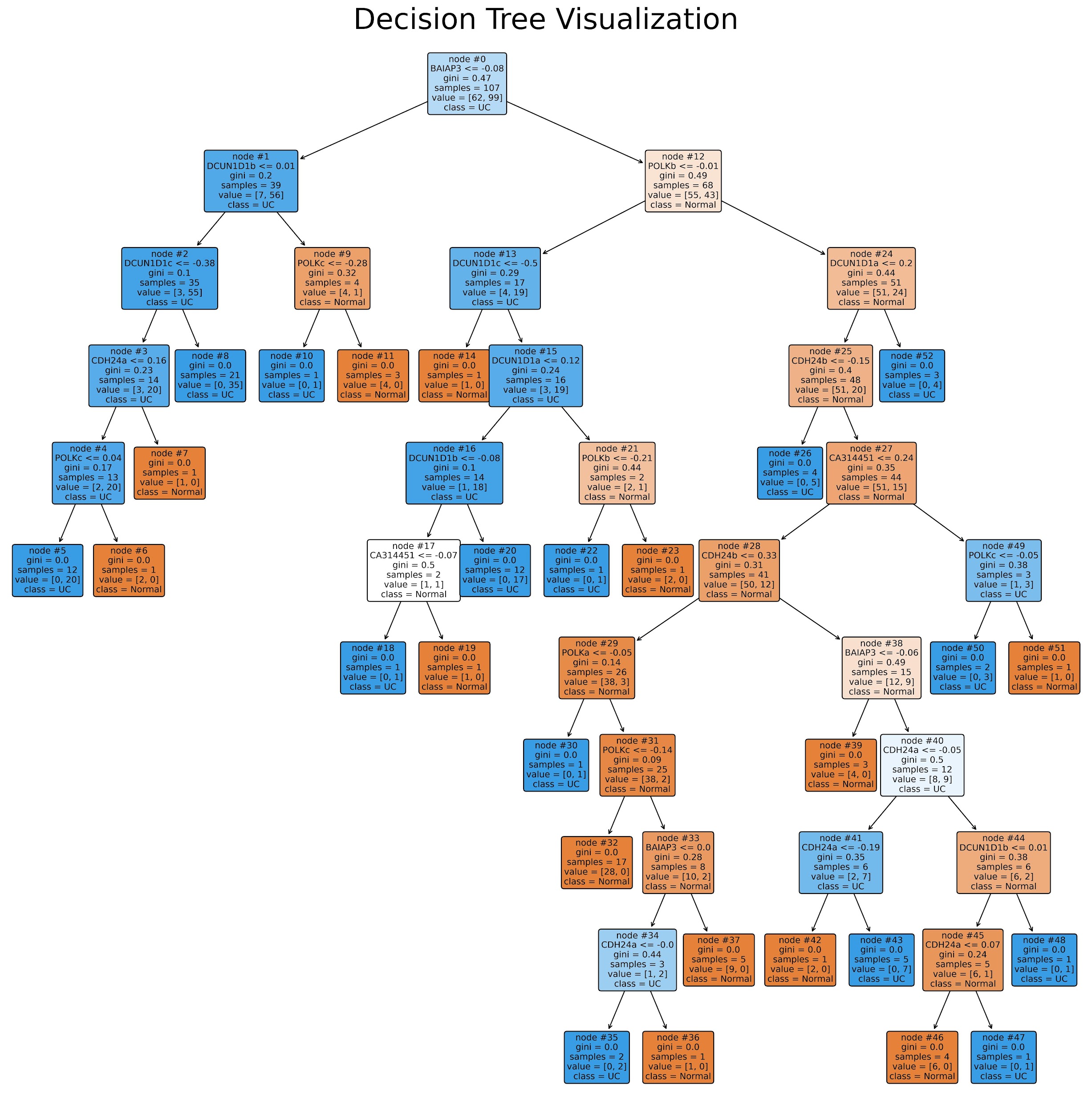
**Figure S2:** Volcano plot of differentially expressed genes within all descending colon samples



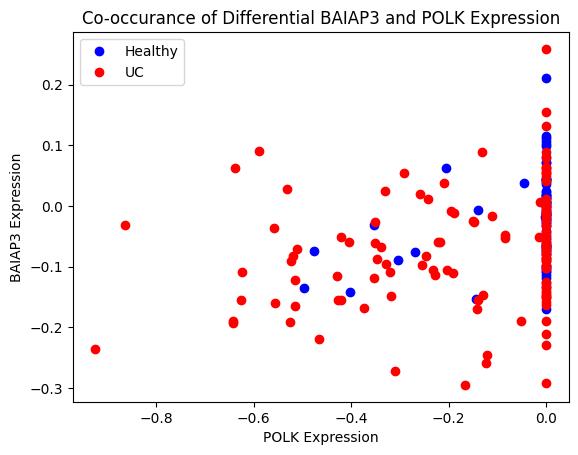
**Figure S3:** Volcano plot of differentially expressed genes within all terminal ileum samples



**Figure S4:** Volcano plot of differentially expressed genes within all ascending colon samples



**Figure S5:** Random forest classifier using normalized expression values for the six genes of interest (HCLS1, CDH24, CA314451, DCUN1D1, POLK, and BAIAP3). The feature name, threshold, number of samples, and classification are in each node. Orange nodes represent a normal classification and blue represent a predicted UC diagnosis.

****

**Figure S6:** Correlation in expression between BAIAP3 and POLK expression across all samples colored by disease status. Healthy samples are blue and UC samples are red.