

Differential Gene Expression Analysis of Ulcerative Colitis Tissue Types
Chandershekhar Shori, Steven Kim
BCH 339N--Dr. Cenik
7 May 2020

Abstract:

Inflammatory bowel disease (IBD) has plagued many individuals across the world by manifesting itself primarily in the forms of Crohn's disease and ulcerative colitis. Currently these potentially fatal illnesses have no cure, and to be diagnosed with these illnesses means to battle IBD for the rest of your lifetime. Additionally, not much is known about which tissue specifically causes this IBD phenotype. However, research has shown that the key to ameliorate these illnesses may lie in differential gene expression. In this investigation, we looked at oligonucleotide microarray data from 202 ulcerative colitis patients in order to determine the differential gene expression among whether the individual has ulcerative colitis, region of the colon, and the interaction between both of these variables. We identified 4 genes that have significant interaction between ulcerative colitis phenotype and the sigmoid region.

Background:

Inflammatory bowel disease (IBD) is a chronic autoimmune illness that causes inflammation throughout one's gastrointestinal tract and has been known to be the leading cause of colon cancer. Symptoms include diarrhea, constipation, stomach pain, and rectal bleeding. There are two primary types of IBD: ulcerative colitis and Crohn's disease. Ulcerative colitis causes inflammation that is limited to the colon, while Crohn's disease causes inflammation that is spread out through the entire gastrointestinal tract. Currently there is no cure for these severe illnesses, but both of these illnesses seem to be intermittently linked due to the fact that both cause inflammation in various parts of the digestive symptom and manifest through similar symptoms. It has been hypothesized that certain genes may lead to a genetic predisposition for IBD that may make it more likely for a patient to develop this illness in their lifetime. The reason

being is that mutations in genes can lead to upregulation and downregulation of the production of certain proteins that lead to a potentially exacerbated response in times when immune action is not needed. This exacerbated response is what is known as an autoimmune response as the body begins to hurt itself by causing excessive inflammation when the body is at rest. In the long run, if this response is not contained, it can lead to permanent scarring in the intestines. This scarring can decrease nutrient and water absorption, which can lead to abnormal growth as the body is not able to absorb the enough resources to provide to its cells. Furthermore, in extreme cases, the inability to absorb nutrients can lead to other bodily organs to malfunction or stop working entirely, which can result in the death of an individual. That is why it is necessary to find a way to contain these illnesses in order to help people in the long run, and the solution to this may lie in gene expression.

Data:

For this investigation, a dataset from the Gene Expression Omnibus (GEO) database will be used (GDS 3268). This dataset contains information from 202 individuals, with 135 of those individuals having ulcerative colitis (UC) in different locations of the colon. The dataset contains information about how expression of different genes varies for patients who have UC affecting different parts of their colon. The oligonucleotide microarray data was normalized to the log2 ratio. The source of the microarray is from Agilent Technologies and used to identify genes (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL1708>).

Goals:

The goal of this study is to contribute knowledge to the field of IBD by showing which genes show a significant role in the development of ulcerative colitis and vary in their expression in different parts of the colon.

Methods:

First, import the GEO file using the “GEOquery” package. Then, we shall remove all the genes that do not contain values in all the columns and are denoted by “NA.” The reason being is that these genes represent incomplete cases when we are in need of gene expression data for that gene for all 202 patients that are being investigated in this study. If these “NA” values were kept, then an error is observed in the code.

After confirming that the data was normalized and in log2 ratio, a linear model with $\beta_0 = 0$ was created using the limma package. Data was inputted as a 2x2 factorial matrix with condition normal and condition ulcerative colitis as one variable and sigmoidal/non-sigmoidal tissue as the other. For the purposes of this model, any region that was not sigmoidal (ascending colon, descending colon, terminal ileum) were grouped into one. As such, there were 4 levels to this linear model. Contrasts were made to compare variables to each other as well as the interaction, the effect of the condition phenotype on the regional expression. We are most particularly interested in the interaction as this indicates that the ulcerative condition phenotype will impact specifically the sigmoidal region. The contrast matrix corresponds with coefficients of the data. A value's coefficient corresponds with which gene has a higher fold expression relative to the contrasted variable. This is shown in Table 1.

Table 1: Coefficients for each contrast, UC stands for ulcerative colitis, wt stands for the normal phenotype relative to ulcerative colitis, sigmoid means tissue samples taken from the sigmoidal tissue/region, and other means non-sigmoidal tissue.

Levels	Contrasts				
	condition_sigmoid	condition_other	tissue_wt	tissue_UC	interaction
UC.other	0	-1	0	1	-1
UC.sigmoid	-1	0	0	-1	1
wt.other	0	1	1	0	1
wt.sigmoid	1	0	-1	0	-1

With this model, t-statistics and f-statistics were calculated between each of these contrasts. We set a threshold for differentially expressed genes ($p < 0.001$) and looked at genes with the most significance, or lowest p-value as well as larger coefficients. These differentially expressed genes are given as log2 fold changes.

Preliminary Analysis:

Data is shown to be normalized among all trials. This is shown in Figure 1. Since the data is normalized and no uneven distributions exist, a linear model is created and corresponding tests were calculated. The significant genes from the contrast are shown below in Table 2.

Figure 1: Distribution of normalized gene density from GEO database record GDS3268

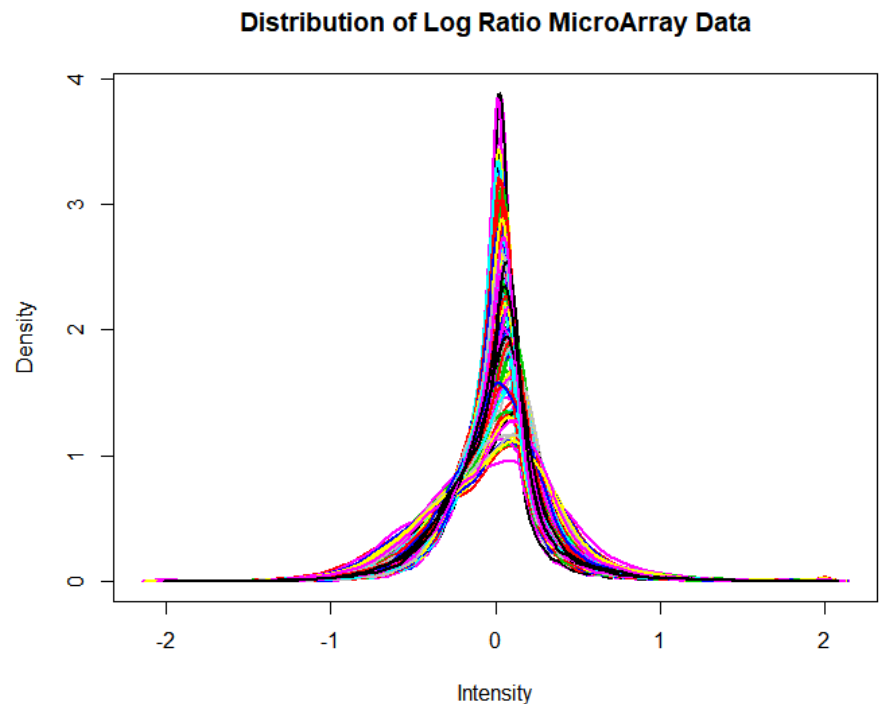


Table 2: Classified expression of genes relative to each contrast. 369/178 genes were up/down regulated in normal sigmoid tissue compared to ulcerative colitis. 520/309 genes were up/down in the normal sigmoid tissue compared to ulcerative colitis. 1337/821 genes were down/up regulated in normal sigmoid tissue compared to normal non-sigmoidal tissue. 1655/661 genes were down-regulated in the ulcerative colitis sigmoid tissue compared to ulcerative colitis non-sigmoidal tissue. 28/25 genes show an increase/decrease in gene expression when changing from normal phenotype to ulcerative colitis or vice versa.

	condition_sigmoid	condition_other	tissue_wt	tissue_UC	interaction
-1	178	309	821	661	28
0	30407	30125	28796	28638	30901
1	369	520	1337	1655	25

Results:

Looking at the first contrast, variable phenotype in sigmoidal tissue, a volcano plot was created mapping the log fold change along the x axis and the significance along the y-axis. The top 4 genes were highlighted based on highest significance, indicated in Figure 2. Those with high absolute value log-fold change were also looked into.

Figure 2: Expression fold change for disease phenotype with sigmoidal tissue. Top 4 significant genes were highlighted and labeled.

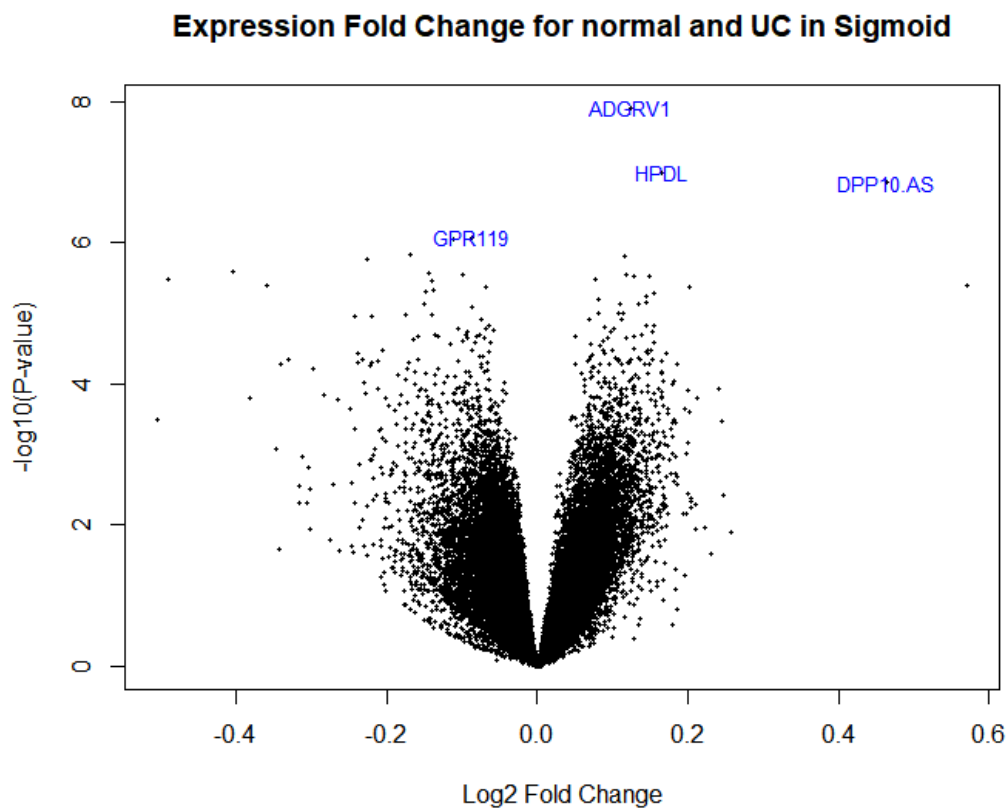
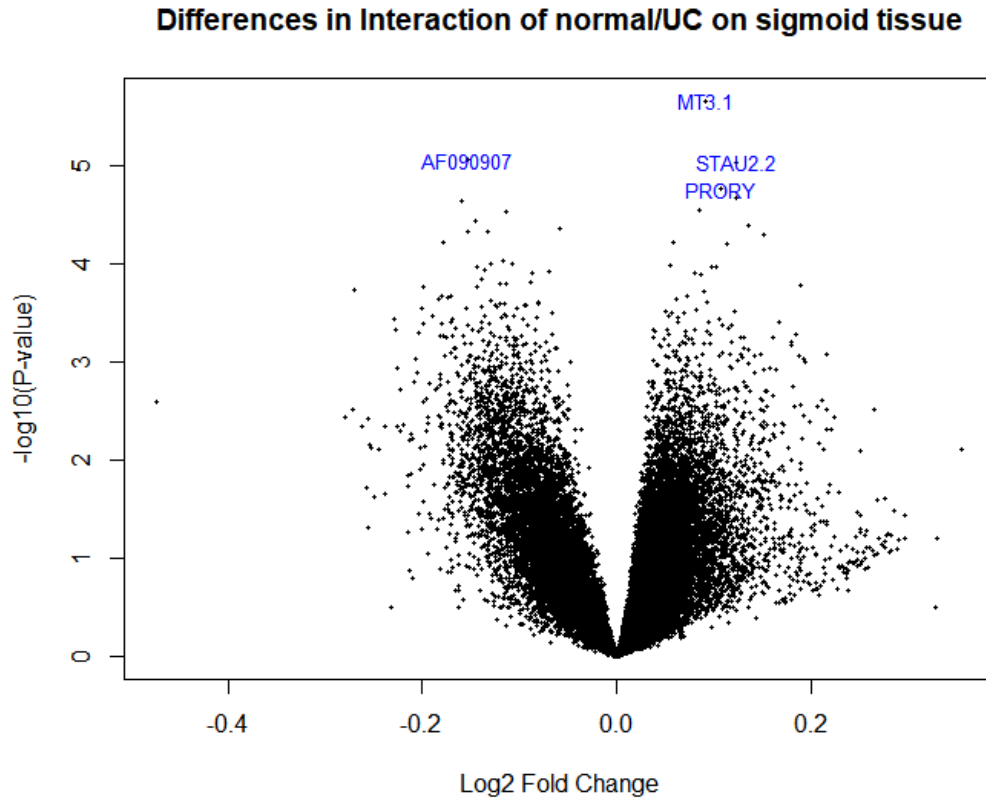


Figure 3: A volcano plot of the interaction of the disease phenotype on the regional expression. Same as with the contrast for condition with sigmoid tissue, the top 4 significant genes are highlighted in the plot. It is worth noting that the p-values for interaction is lower compared to other contrasts statistics.



With this interaction data, a filter was made for those genes with a p-value<0.001 and a threshold log fold change greater than 0.2, these genes are shown in Table 3 below.

Table 3: Genes that had significant and high LFC of interaction with phenotype and region. Two genes are unknown in function (A_24_P76362 and A_24_P144543).

	gene	LFC_conditi on_sigmoid	LFC_conditi on_other	LFC_tissue_ wt	LFC_tissue_ UC	LFC_interac tion	adj.P.Val
1	CXCL8	-0.49142	-0.13599	0.019351	-0.33607	0.355423	2.82E-06
2	A_24_P763 62	0.024418	-0.20449	-0.2906	-0.06169	-0.22891	3.89E-08
3	FMOD	0.048445	-0.15416	-0.28304	-0.08044	-0.2026	0.000193
4	DUOX2.1	-0.40503	-0.18887	0.058714	-0.15745	0.216165	1.08E-05
5	DPP10.AS1	0.464421	0.214602	-0.0899	0.159915	-0.24982	7.99E-07
6	TCN1	-0.23333	-0.01767	0.051143	-0.16452	0.215658	0.000294
7	FAM111B	0.18543	-0.03451	-0.30565	-0.08571	-0.21994	7.86E-05
8	A_24_P144 543	0.028169	0.236376	0.092302	-0.11591	0.208208	0.000286
9	UGT2A3	0.571	0.097747	-0.20716	0.266097	-0.47325	0.000233
10	CXCL1	-0.50452	-0.27757	-0.1366	-0.36355	0.226944	2.17E-05

Conclusions:

In this investigation, 4 important genes were found in the sigmoid colon that showed significant difference in expression when comparing the sigmoid colon of an ulcerative colitis patient to one of a normal individual. These genes were: HPDL, ADGRV10, GPR110, and DPP110.AS1. The reason these genes are important is because they were found significant in other papers where the ulcerative colitis patients were compared to normal patients. This helps reaffirm the results obtained in our investigation are valid as they are shown to be true by looking at previous studies. Thus, these genes are essentially controls in our project as they serve as a reference point to show that our analysis was done correctly. It is worth noting that these genes are typically expressed higher in the sigmoid colon in comparison to other parts of the

colon, regardless of whether the patient has ulcerative colitis or not (data not shown). Once this was observed, we can now proceed to look at some other genes that are shown to have a significant interaction effect, which are shown in Table 3.

Although 10 genes were shown to have interactions in Table 3, specifically 4 genes are of tremendous interest in regards to our study as they are directly related to inflammatory processes related to ulcerative colitis. These genes are: UGT2A3, CXCL1, CXCL8, and DPP10.AS1. UGT2A3, which is known to play a significant role in metabolism--specifically of drugs, is of interest because alteration to the expression of this gene can detrimentally affect the extent to which drugs are metabolized. By looking at Table 3, it is observed that the UGT2A3 gene has a positive value for LFC_condition_sigmoid and LFC_condition_other, which means that it is downregulated in ulcerative colitis tissues relative to normal patients. This implies that ulcerative colitis patients are not able to effectively break down drugs effectively, which can lead to the buildup of toxins in the intestines. These toxins could alter cellular function in the intestines leading to inflammation being triggered by the immune response, which ultimately leads to the ulcerative colitis phenotype in patients.

Next, we move onto the CXCL1 and CXCL8 genes as they are directly involved in the inflammatory response by helping macrophages release chemokines that trigger inflammatory processes. By looking at Table 3, it is interesting to see that we have significantly more negative values for LFC_condition_sigmoid in comparison to LFC_condition_other. This means that the CXCL1 and CXCL8 genes are upregulated in ulcerative colitis patients, and leads to an overexpression of the chemokines that result in an exacerbated inflammatory response. Looking at the disease itself, this makes sense with what has been stated in previous literature as excessive unnecessary inflammation is common in many cases of ulcerative colitis.

Now, we move on to discussing DPP10.AS1, which is a gene that has a connection to the development of colon cancer. Given that LFC_condition_sigmoid has positive value for this gene, it is implied that this gene is downregulated in ulcerative colitis patients in comparison to normal patients. What further makes this interesting is that since this gene is linked to colon cancer, it is likely that is a tumor suppressor gene whose function is lowered leading to the manifestation of ulcerative colitis, which ultimately progresses to cancer as IBD is one of the leading causes of colon cancer world-wide. Lastly, it is worth noting that A_24_P144543 and A_24_P76362 also possess significant results as these genes appear to very downregulated in ulcerative colitis patients in comparison to normal patients given the fact that these genes have positive values in the LFC_condition_sigmoid columns. Currently much knowledge of these genes is currently unknown, but future investigations of these genes may provide further insight to the functions of A_24_P144543 and A_24_P76362 and their connection to ulcerative colitis.

A limitation to this data is how the factorial matrix was created. Because the tissue types were only sigmoid tissue compared to other tissue, significantly expressed genes in a region such as the ascending colon relative to the sigmoid region cannot be analyzed. Future models could implement different base tissue such as setting the sigmoid region to ascending colon in this model. Additionally, the condition for ulcerative colitis is identified as a dummy variable. Someone who has more severe symptoms such as more frequent bowel movements is treated the same in this model as someone with less. As such, variable expressivity isn't accounted for in the linear model created. Future remedies for this seem difficult, as a term such as number of movements is subject to error.

For future investigations, we must take other factors into account that can drastically affect our results in future investigations. One of these factors is taking other kinds of IBD into

account, such as Crohn's disease, in order to better generalize our results to all kinds of IBD. The reason being is that ulcerative colitis and Crohn's disease are different to an extent to which parts of the gastrointestinal tract they affect, and as a result both IBDs must be looked at in order to identify generalizable results that apply to both IBDs. Additionally, age of diagnosis must be taken into account for future investigations. The reason being is that IBD does not manifest at birth, but rather typically appears randomly in people's lifetimes. This implies that for cases of IBD have variable penetrance in regards to age of manifestation, which may be due to the result of how mutated a particular gene may be. Lastly, For example, more mutations may lead to earlier penetrance of the IBD phenotype, while less mutations may lead to later penetrance of this phenotype. Lastly, environmental factors, such as stress, diet, place of living, and etc, must also be taken into account as previous literature has shown that IBD is not necessarily the product of gene mutations, but may be rather a product of gene-environment interaction. Thus, in the end, this investigation only brings us a small step closer to a cure as there is much we still need to discover in regards to inflammatory bowel disease.

Contributions:

Conceptualization-- Chandershekhar was the one who primarily undertook the endeavor of producing the ideas, goals, and gathering background information of this investigation.

Data Curation-- Steven and Chandershekhar both assisted in helping prepare this data for the purposes of this investigation and for later use.

Formal Analysis-- Steven and Chandershekhar both undertook the endeavor of producing the code necessary for analysis of this investigation.

Methodology-- Steven and Chandershekhar both assisted in creating the methods needed to proceed with this investigation.

Funding Acquisition-- There really was no extra funding needed to carry out this investigation. All resources were provided by the University of Texas at Austin and Dr. Cenik's teaching team.

Investigation-- The investigation was supervised and conducted by both Steven and Chandershekhar.

Project Administration-- The investigation was managed and coordinated by both Steven and Chandershekhar.

Resources-- Chandershekhar and Steven gathered their resources using the Gene Expression Omnibus (GEO) database, and attained much of their other resources from Dr. Cenik's class and teaching team.

Software-- Chandershekhar and Steven attained much of their software from R packages online, and Dr. Cenik's teaching team. The software used for this investigation was R.

Visualization-- Chandershekhar and Steven both assisted in creating the data and preparing it for visualization through the powerpoint presentation.

Supervision-- Chandershekhar and Steven received supervision from Dr. Cenik's teaching team as they helped provide the necessary mentorship to help make this project a success. They did this by providing the research team with constructive critiques to help guide their project onto the right path.

Validation-- Chandershekhar and Steven both helped verify the replicability of the results produced in this investigation.

Writing -original draft-- Chandershekhar and Steven both assisted in preparing the initial draft and presentation for this investigation.

Writing - review & editing-- Chandershekhar and Steven both assisted in finalizing and proofing the final overall presentation and paper.

References:

Mackenzie, Peter I., et al. "The Novel UDP Glycosyltransferase 3A2: Cloning, Catalytic Properties, and Tissue Distribution." *Molecular Pharmacology*, vol. 79, no. 3, 2010, pp. 472–478., doi:10.1124/mol.110.069336.

Mirza, Aashiq H, et al. "Transcriptomic Landscape of LncRNAs in Inflammatory Bowel Disease." *Genome Medicine*, vol. 7, no. 1, 2015, doi:10.1186/s13073-015-0162-2.

Noble, C L, et al. "Regional Variation in Gene Expression in the Healthy Colon Is Dysregulated in Ulcerative Colitis." *Gut*, vol. 57, no. 10, 2008, pp. 1398–1405., doi:10.1136/gut.2008.148395.

Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015).limma powers differential expression analyses for RNA-sequencing and microarray studies.*Nucleic Acids Research*43 (7), e47.

Phipson, B, Lee, S, Majewski, IJ, Alexander, WS, and Smyth, GK (2016). Robust Hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression.*Annals of Applied Statistics* 10(2), 946–963

Law, CW, Chen, Y, Shi, W, and Smyth, GK (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.*Genome Biology*15, R29

Code:

```
library(rjson)
library(BiocManager)
library(GEOquery)
library(pheatmap)
library(DESeq2)
library(limma)
library(tidyverse)
# updateR()

#Setting up for limma, the expression matrix, the feature data, and the phenotype data
BiocManager::install("limma")

gds858 <- getGEO('GDS3268', destdir=".") #expression
agilent_source = getGEO("GPL1708", destdir=".") #feature matrix of the assay, includes genes
and descriptions
tab = Table(gds858)
tab_gds858 = tab %>% select(3:length(tab[1,]))
gds858_logcounts = as.matrix(tab_gds858)
rownames(gds858_logcounts)=make.names(tab[,2],unique=TRUE)
rownames(gds858_logcounts)
gds858_logcounts=na.omit(gds858_logcounts)

agilentmatrix=data.frame(agilent_source@dataTable@table)
rownames(agilentmatrix)=make.names(agilentmatrix$SPOT_ID,unique=TRUE)
# agilent = agilentmatrix %>% select(GENE_SYMBOL,GENE_NAME,DESCRIPTION)
# reorder_agilent <- match(rownames(agilentmatrix),rownames(gds858_logcounts))
# reorder_agilent = na.omit(reorder_agilent)
# identical(rownames(agilentmatrix),rownames(gds858_logcounts))
# agilentmatrix=na.omit(agilentmatrix)

# agilentmatrix = agilentmatrix %>% filter(rownames(agilentmatrix) %in%
rownames(gds858_logcounts))
gds858_metadata= gds858@dataTable@columns
gds858_metadata=column_to_rownames(gds858_metadata,"sample") #phenotype

#This is adjusting the phenotype data to match the expression data
reorder_idx <- match(rownames(gds858_metadata),colnames(gds858_logcounts))
prepared_gds858 = gds858_metadata[reorder_idx,]
prepared_gds858 = data.frame(prepared_gds858)
```


#column names were adjusted to have better labels, as well as group the other sigmoid regions into one

```
prepared_gds858 = prepared_gds858 %>% mutate(sigmoid=
  case_when(
    tissue=="sigmoid colon"~"sigmoid",
    TRUE ~ "other"
  ))
prepared_gds858 = prepared_gds858 %>% mutate(disease.state=
  case_when(
    disease.state=="ulcerative colitis"~"UC",
    TRUE ~ "wt"
  ))
rownames(prepared_gds858)=colnames(gds858_logcounts)
```

```
# dim(gds858_logcounts)
prepared_gds858 = sapply(prepared_gds858,as.character)
# sapply(prepared_gds858,class)
prepared_gds858=data.frame(prepared_gds858)
rownames(prepared_gds858)=colnames(gds858_logcounts)
```

#There was trouble adding the `agilent_microarray_matrix` onto the feature data for the `ExpressionSet` object, so if gene names were different than what was in the limma model ID!=gene, then it was checked manually.

```
eset =ExpressionSet(gds858_logcounts,phenoData=AnnotatedDataFrame(prepared_gds858))
plotDensities(eset,legend=FALSE,main="Distribution of Log Ratio MicroArray Data")
```

```
#model matrix was created with the 4 parts in the model
group <- with(pData(eset), trimws(paste(disease.state, sigmoid, sep = ".")))
group <- factor(group)
group
design <- model.matrix(~0 + group)
colnames(design) <- levels(group)
colSums(design)
```

```
#Contrasts were made among the two factors as well as the interaction
cm <- makeContrasts(condition_sigmoid = wt.sigmoid - UC.sigmoid,
  condition_other = wt.other - UC.other,
  tissue_wt = wt.other - wt.sigmoid,
```

```

tissue_UC = UC.other - UC.sigmoid,
interaction = (wt.other - wt.sigmoid) - (UC.other - UC.sigmoid),
levels = design)
fit <- lmFit(eset, design)

fit2 <- contrasts.fit(fit, contrasts = cm)

# Calculate the t-statistics for the contrasts
fit2 <- eBayes(fit2)

# Summarize results
results <- decideTests(fit2)
results <- classifyTestsF(fit2, p.value=0.0001)
# vennDiagram(results,include=c("up"))
#understanding the diagram by looking at contrasts,
summary(results)
cm

stats <- topTable(fit2, number = nrow(fit2), sort.by = "none",adjust.method='BH' )
# stats %>% order_by(adj.P.Val)
hist(stats[, "P.Value"])
#Histogram was plotted to make sure the p values distributions reflect typical limma models
#Left skewed is normal

volcanoplot(fit2, highlight = 4, names =
rownames(fit2$p.value),coef="interaction",main="Differences in Interaction of normal/UC on
sigmoid tissue")
volcanoplot(fit2, highlight = 4, names =
rownames(fit2$p.value),coef="condition_sigmoid",main="Expression Fold Change for normal
and UC in Sigmoid")

#looked at gene expression in other tissue, omitted from paper
volcanoplot(fit2, highlight = 4, names =
rownames(fit2$p.value),coef="condition_other",main="Expression Fold Change for normal and
UC in non-sigmoid")

#These two volcano plots show which genes are expressed more in which tissue.
#For the purposes of the paper, they were excluded, but there are high significance values
#as well as nice plots.

```

```
volcanoplot(fit2, highlight = 4, names =  
rownames(fit2$p.value),coef="tissue_wt",main="Expression Fold Change for regions in WT")  
volcanoplot(fit2, highlight = 4, names =  
rownames(fit2$p.value),coef="tissue_UC",main="Expression Fold Change for regions in UC")
```

```
stats = rownames_to_column(stats,var="gene")
```

```
#analysis was done on the interaction term. Here I filtered genes that had p value<0.001,  
#and these genes had a coefficient higher than 0.2.
```

```
interaction_stats = stats %>% filter(adj.P.Val<0.001,abs(interaction)>0.2)
```

```
interaction_stats
```

```
#CXCL1 and CXCL8 are chemokines, which is interesting.
```

```
#UGT3A2 is a detoxifying protein that changes lipophilic things to liquid. This may explain  
damage,
```

```
#though unknown for now.
```

```
# transformed_stats %>% mutate(average=average())
```

```
write.csv(interaction_stats,"finish_data2.csv")
```

```
write.csv(stats,"whole_stats.csv")
```

```
# cor_normalized = cor(gds858_rawcounts_,y=NULL, use = "everything", method =  
"spearman")
```

```
# pheatmap(cor_normalized,na.rm=TRUE)
```

```
# dds_all <- DESeqDataSetFromMatrix(countData = gds858_rawcounts_,
```

```
#           colData = prepared_gds858,
```

```
#           design = ~condition + region)
```

```
# class(dds_all)
```

```
# class(deseq_colon)
```

```
# deseq_colon = DESeq(dds_all)
```

```
# keep <- rowMeans(counts(deseq_colon)) >= 500
```

```

# deseq_colon = deseq_colon[keep,]
# deseq_colon$region= relevel(deseq_colon$region,ref="sigmoid colon")
# deseq_colon$condition= relevel(deseq_colon$condition,ref="normal")
#
# #outlier detection
# BiocManager::install("apeglm")
# resLFC <- lfcShrink(deseq_colon,
# coef="condition_ulcerative.colitis_vs_normal",type="apeglm")
# plotDispEsts()
# plotDispEsts(resLFC)
# plotMA(resLFC, ylim=c(-2,2))
# plotMA(dds_results,ylim=c(-2,2))
#
# #vst was chosen for log transform and normalization
# vsd_all <- vst(deseq_colon, blind = TRUE)
# plotPCA(vsd_all, intgroup=c("condition", "region"))
# plotPCA(vsd_all, intgroup=c("condition"))
# plotPCA(vsd_all, intgroup=c("region"))
# # resultsNames(deseq_colon)
# vsst_colon = vst(deseq_colon, blind=TRUE)
# dds_results = results(deseq_colon,contrast=c("condition","ulcerative.colitis","normal"),
# # alpha=0.05, lfcThreshold=0.32)
# dds_results <- lfcShrink(deseq_colon,
# # contrast=c("condition","ulcerative.colitis","normal"),
# # res=dds_results)
# dds_data_frame = data.frame(dds_results)
# dds_final <- subset(dds_data_frame, pvalue < 0.05)
# write.csv(dds_final,"DESeq_gene_results")
# vst_ordered = vsd_all[order(vsd_all$pvalue),]
#
#
# dds_condition_only = DESeqDataSetFromMatrix(countData = gds858_rawcounts_,
# # colData = prepared_gds858,
# # design = ~condition)
# dds_condition_only$condition= relevel(dds_condition_only$condition,ref="normal")
# dds_condition_only = DESeq(dds_condition_only)
# plotDispEsts(dds_condition_only, main = "condition only DispEsts")
# dds_region_res = vst(dds_region_only,blind=TRUE)
# plotPCA(dds_region_res, intgroup="region")
#
# dds_region_only = DESeqDataSetFromMatrix(countData = gds858_rawcounts_,
# # colData = prepared_gds858,
# # design = ~region)
# dds_region_only$region= relevel(dds_region_only$region,ref="sigmoid colon")

```

```

# dds_region_only = DESeq(dds_region_only)
# plotDispEsts(dds_region_only,main="region only DispEsts")
# dds_region_res = vst(dds_region_only,blind=TRUE)
# plotPCA(dds_region_res, intgroup="region")
#
#
# library("RColorBrewer")
# library("pheatmap")
# sampleDists <- dist(t(assay(vsd_all)))
# sampleDistMatrix <- as.matrix(sampleDists)
# rownames(sampleDistMatrix) <- paste(vsd_all$condition, vsd_all$region)
# colnames(sampleDistMatrix) <- NULL
# colors <- colorRampPalette( rev(brewer.pal(9, "Blues"))) (255)
# window(width=12,height=10)
# pheatmap(sampleDistMatrix,
#   clustering_distance_rows=sampleDists,
#   clustering_distance_cols=sampleDists,
#   col=colors,show_rownames = F)
#
#

```