Cancer-Associated Dysregulated Genes in IBD and Crohn's Disease Patients

1st Runa Cheng
FAES BIOF509 FALL
NIH
Bethesda, USA
runa.cheng@nih.gov

2nd Christina Park
FAES Teaching Assistance
NIH
Bethesda, USA
parkyc@nih.gov

3rd James Anibal
FAES Course Instructor
NIH
Bethesda, USA
james.anibal@nih.gov

***Abstract*—Inflammatory bowels disease (IBD), including Crohn's disease, is known to associate with increased cancer risk, particularly colorectal cancer. Using the E-GEOD-100833 dataset from ArrayExpress database, which contained microarray data from healthy controls and IBD/Crohn's disease patients, the gene expression profiles were dimensionally reduced with UMAP and clustered with density-based DBSCAN. Although no cancer-associated genes were found uniquely relate to any patient subset identified with this dataset, shared dysregulated proto-ocogene and tumor-suppressor genes among different clusters were identified.**

***Index Terms*—IBD, Crohn's disease, cancer, microarray data**

I. Introduction

The prevalence of inflammatory bowel disease increased significantly since the late 20th century (CDC, 2020). Both Crohn's disease and ulcerative colitis were under this category and they were not well understood. It was estimated that the annual incidence of Crohn's disease was about 3-20 cases per 100,000 (Feuerserin and Cheiferz, 2017). It was highly heterogenous and could manifest in any region of the gastrointestinal tract. Both environmental and genetics factors could trigger the disease (Chowers, 2013). The differences in disease susceptibility and prognosis among patients could be partly explained by the HLA genotyping result, which revealed differing HLA allele distribution between Crohn's disease subgroups (Lombardi et al., 2001).

In addition to diarrhea, blood in stool, abdominal pain, and gastric inflammation, Crohn's disease patients were also prone to developing cancer ("Crohn's disease", 2020). A study reported that the prevalence of colorectal cancer in patients with Crohn's disease was about twenty times that of the overall population (Weedon et al, 1973). In addition, rare cancers such as anorectal fistulae adenocarcinoma were also associated with Crohn's disease, with poor prognosis and a higher fatality rate compared to normal cancer patients (Kodama et al., 2018). (See project progress report.)

II. Methods

A. Data Pre-Processing

E-GEOD-100833 contained data from 338 individuals, including 261 patients with either general IBD or Crohn's disease. The microarray used contained 54715 probes, which were coded and should be matched to genes to analyze the dataset. The gene expression profile should be standardized. It was unlikely that a dataset with healthy controls and diseased individuals would follow a Gaussian distribution, so a min-max scaling would be more appropriate. Different types of samples were included in the original data, including blood and biopsy specimens from normal or inflamed regions. There were more blood samples than any other types. Because expression data can differ based on the sample type, so only the blood samples were analyzed.

III. Dimensional Reduction with UMAP

The dataset was too large to directly be passed to the clustering algorithm, so UMAP was applied. The dimensionality reduction algorithm was implemented based on non-linear methods so the structure of the original dataset would be preserved better than its linear counterpart (i.e. PCA). However, the algorithm also has its own disadvantages. Its runtime was significantly longer. In addition, the stochastic method implemented would render different results during each run, holding all the data and parameters constant.

There are three main parameters, n-neighbors, min-dist, n-components. n-neighbors would determine how many data points the program look at in the local neighborhood. min-dist determined how close the points were allowed to be. Both parameters would determine whether the local or the global structure would be better preserved. n-components determined the resulting number of dimensions after UMAP.

IV. Density-Based Clustering with DBSCAN

Density-based clustering was suitable for gene expression profiles. DBSCAN was applied to the dimensionally reduced data after UMAP. Two parameters were passed in – eps and min-samples. The former represent the max distance between two data point for them to be considered in the same neighborhood. Min-samples determined the number of points close by for one point to be considered as the centroid.

V. Identifying Dysregulated Gene with Percentile Ranking

Each of the 54715 genes was ranked based on percentile with the Pandas built-in ranking function (df.rank(pct=True)).

Genes expressed below the fifth percentile or above the ninety-fifth percentile across all samples within at least one cluster were identified.

VI. Results

A. Exploratory Analysis with Normal Control

Eight clusters were identified with the complete dataset, including the normal controls (Fig.1).

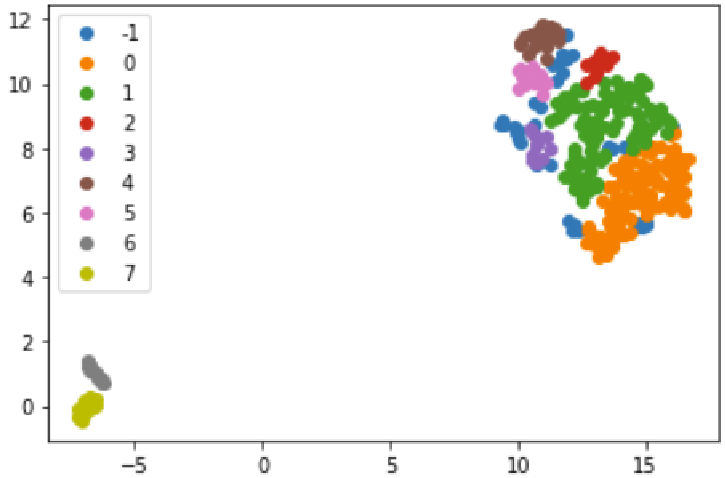


Fig. 1. DBSCAN of Gene Expression Profile Data with Normal Control

The most distinct clusters emerged from the normal control compared to the diseased population (Fig.2). It was interesting to note that there seemed to be two populations of normal controls, although the normal controls closer the diseased population was still distinct. The gene expression profile of IBD patients was not separable from that of Crohn's disease. Crohn's disease was closely associated with general IBD and often categorized under the broader term, which was reflected by the loose clusters.

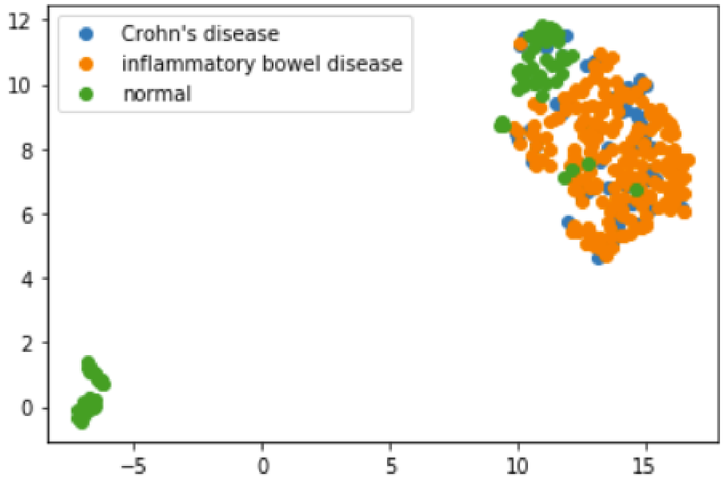


Fig. 2. UMAP of Gene Expression Profile Data by Disease Category

B. DBSCAN Clustering with Patient Data

To focus on the patient data, the normal control samples were taken out. 7 clusters were identified (Fig.3). However, the clusters were loosely formed and the data seemed moderately randomly distributed. The clustering was not related to whether the patients had general IBD or Crohn's disease (Fig.4).

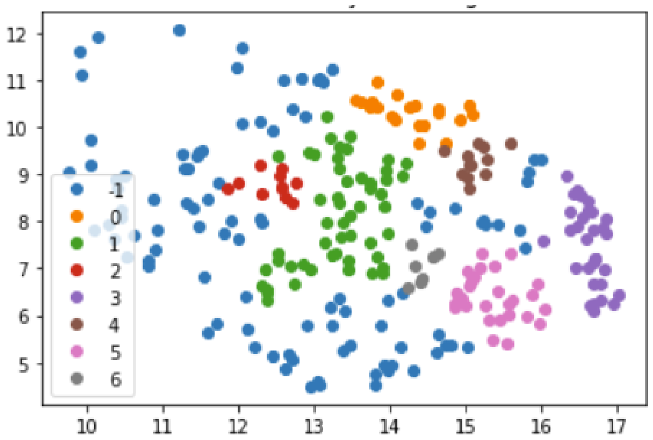


Fig. 3. DBSCAN of Gene Expression Profile Data of IBD and Crohn's Patients Only

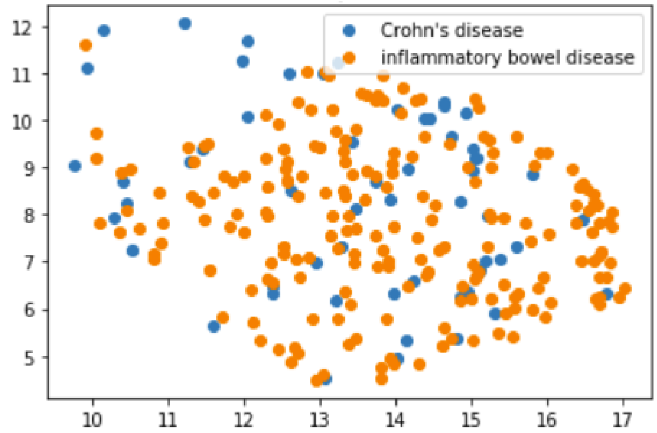


Fig. 4. DBSCAN of Gene Expression Profile Data of IBD and Crohn's Patients Only by Disease

Furthermore, the distribution did not appeared to be associated with sex (Fig.5).

C. Dysregulated Genes Associated with Cancer

The gene expression profile was standardized with the min-max method and genes whose expression levels were in the lowest or the top 5 percent for all samplesin at least one cluster, which was categorized as dysregulated, was identified (Fig.6).

There were a total of 30 genes that showed up in this screening. 13 of the hits were known to associate with colorectal and gastric cancers. All 13 were dysregulated in all clusters

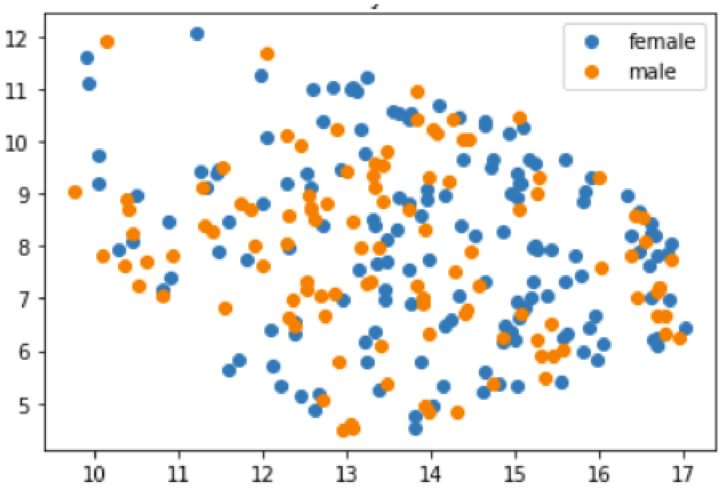
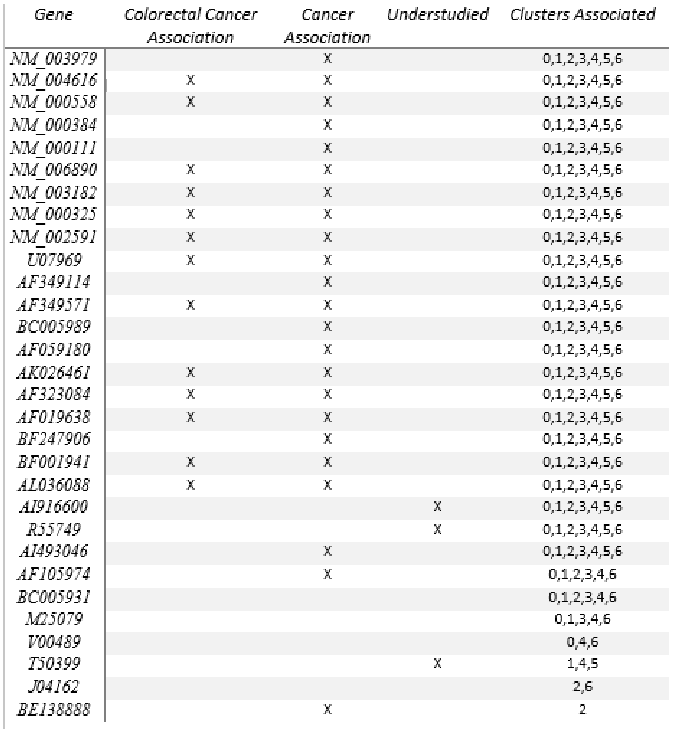


Fig. 5. DBSCAN of Gene Expression Profile Data of IBD and Crohn's Patients Only by Gender

identified. 23 hits in total were known to either be proto-oncogenes or tumor-suppressor genes. Of the 5 remaining genes, 2 were not well studied and 3 were associated with immune responses and inflammation. More detailed information on these genes and references can be found in 'gene and cancer association.xlsx' in this repository.



Gene	Colorectal Cancer Association	Cancer Association	Understudied	Clusters Associated
NM_003979		x		0,1,2,3,4,5,6
NM_004616	x	x		0,1,2,3,4,5,6
NM_000558	x	x		0,1,2,3,4,5,6
NM_000384		x		0,1,2,3,4,5,6
NM_000111		x		0,1,2,3,4,5,6
NM_006890	x	x		0,1,2,3,4,5,6
NM_003182	x	x		0,1,2,3,4,5,6
NM_000325	x	x		0,1,2,3,4,5,6
NM_002591	x	x		0,1,2,3,4,5,6
U07969	x	x		0,1,2,3,4,5,6
AF349114		x		0,1,2,3,4,5,6
AF349571	x	x		0,1,2,3,4,5,6
BC005989		x		0,1,2,3,4,5,6
AF059180		x		0,1,2,3,4,5,6
AK026461	x	x		0,1,2,3,4,5,6
AF323084	x	x		0,1,2,3,4,5,6
AF019638	x	x		0,1,2,3,4,5,6
BF247906		x		0,1,2,3,4,5,6
BF001941	x	x		0,1,2,3,4,5,6
AL036088	x	x		0,1,2,3,4,5,6
AI916600			x	0,1,2,3,4,5,6
R55749			x	0,1,2,3,4,5,6
AI493046		x		0,1,2,3,4,5,6
AF105974		x		0,1,2,3,4,6
BC005931				0,1,2,3,4,6
M25079				0,1,3,4,6
V00489				0,4,6
T50399			x	1,4,5
J04162				2,6
BE138888		x		2

Fig. 6. Dysregulated Genes Identified in DBSCAN Clusters

Discussion

The goal of the study was to identify subgroups of IBD and Crohn's patients who would be susceptible to developing cancer, particularly those located in the gastrointestinal area. The 1000IBD dataset contained clinical, microbiomic, and gene expression data from a large number of patients and was an integral part of the original project design. However, the data request required an extensive approval process with funded ongoing projects so it could not be obtained. I then turned to the E-GEOD-100833 from ArrayExpress which contained microarray data from a sizable population with IBD and Crohn's disease.

The data dimensionality was successfully reduced and then passed to the DBSCAN algorithm for density clustering. The normal controls' expression profiles were distinct from those from the diseased individuals' and two populations could be identified, one more similar to the diseased population.

Looking at the data only comprising of the general IBD and Crohn's disease patients, loose clusters were identified, although the distribution of the data points appeared to be random. The distribution was not related to sex or the disease category.

Unfortunately, I could not obtain a dataset which categorized the patients based on their IBD/Crohn's disease as well as cancer status. However, I ranked the expression of each gene from all the samples to identify individuals whose expression profiles were different from others. Dysregulated genes were defined as genes whose expression levels were above the top fifth percentile, or below the bottom fifth percentile. Genes that are dystregulated in all individuals within a cluster were identified.

A total of 30 genes were found and the set contained 13 genes were known to associate with colorectal and gastric cancers; they were prevalent showed up in all 7 clusters. Given that the clusters were loosely defined and showed no distinct pattern, it was not surprising no significant hits were uniquely associated with only one cluster. 23 identified genes were known to either be proto-oncogenes or tumor-suppressor gene. It appeared that IBD/Crohn's disease were associated with cancer, particularly colorectal cancers.

A larger dataset that examined both cancer and IBD/Crohn's disease would be necessary to continue the study. To better understand the pathology of inflammation-association cancer would be important for the disease management.

Acknowledgment

I appreciate all the patience and support from my TA Christina and lecturer James.

References

[1] Data and Statistics. (2020, August 11). Retrieved December 04, 2020, from https://www.cdc.gov/ibd/data-statistics.htm

[2] "Crohn Disease: Epidemiology, Diagnosis, and Management." Mayo Clinic Proceedings , 7 June 2017, www.mayoclinicproceedings.org/article/S0025-6196(17)30313-0/fulltext.

[3] Chowers, Yehuda. "Taking Crohn's Disease Personally." Rambam Maimonides Medical Journal, Rambam Health Care Campus, 30 Apr. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC3678831/.

[4] Lombardi, Maria Luisa, et al. "Crohn Disease: Susceptibility and Disease Heterogeneity Revealed by HLA Genotyping." Human Immunology, Elsevier, 18 June 2001,

[5] “Crohn’s Disease.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 13 Oct. 2020, www.mayoclinic.org/diseases-conditions/crohns-disease/symptoms-causes/syc- 20353304..

[6] Weedon DD;Shorter RG;Ilstrup DM;Huizenga KA;Taylor WF; “Crohn’s Disease and Cancer.” The New England Journal of Medicine, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/4754948/.

[7] Kodama, Makoto, et al. “Adenocarcinoma within Anorectal Fistulae: Different Clinicopathological Characteristics between Crohn’s Disease-Associated Type and the Usual Type.” Nature News, Nature Publishing Group, 11 Sept. 2018, www.nature.com/articles/s41379-018-0105-8.