1 Sample collection

The datasets for this paper were retrieved from the NCBI-GEO (Gene Expression Omnibus) a free public database of microarray or gene profile, and we obtained the gene expression profile of GSE6731-platform [HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array for healthy, UC and CD samples of the dataset. High throughput sequencing data profiling of gene expression is included in the data. The present data for GSE6731 includes out of these 36 cases, 4 are controlled cases, while 32 are samples with adult inflammatory bowel disease (IBD). Majority of cases have Crohn's disease i.e., 19 while 9 have ulcerative colitis. Although the age group has not been considered for the analysis for this paper, the number of cases for each dataset is summarised below in Table 1. The research data was collected and conducted at the Johns Hopkins University School of Medicine,specifically in the Department of Medicine, Baltimore, USA. The study was submitted on January 12, 2007, and last updated on December 13, 2018.

Data set Accession ID	**Type of the data (file format)**	**Technology type (if any)**	**Number of cases**	**Reference**
GSE6731	Microarray(CEL files)	in situ oligonucleotide	36 samples (34 considered)	https://pubmed.ncbi.nlm.nih.gov/17262812/

Table 1: The following table consists of dataset used for a genome wide sequencing analysis, each containing control/non-inflammatory and inflammatory samples, where GSE6731 is taken from Affymetrix Human Genome U95 Version 2 Array, and dataset GSE9452 and GSE16879 taken from U133 Plus 2.0 Array

4.1.1 GSE6731

The following dataset has been procured from NCBI GEO website where it matched the desired criteria of IBD conditions dataset. In order to perform Genome-wide sequencing of genes, cel files from the Affymetrix array are required. The expression data shows the variations in Crohn's and ulcerative colitis's genome-wide gene expression from endoscopic pinch biopsies. The research study compared active and inactive areas of UC and CD with infectious colitis and healthy controls. A total of 36 samples underwent unsupervised classification, revealing

distinctive gene expression patterns among affected and unaffected IBD tissues, non-IBD colitis, and normal controls.

RAW FILE FORMAT	**SAMPLES NUMBERS**	**CONDITIONS**	**SHORTLISTED CONDITIONS**
celfiles	36	Ulcerative colitis-9	Ulcerative colitis-9
		Crohn's disease-19	Crohn's disease-19
		Normal-4	Normal-4
		Indeterminate colitis -2	
		INF(bacterial infectious colitis)-2	

Table 2: the GEO dataset GSE6731 has .CEL format taken from Affymetrix Human Genome U95 Version 2 Array where the shortlisted conditions are normal, UC and CD which are in total 34 out of the 36 samples present in the original dataset.