4. RESULTS

4.1 Sample collection

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

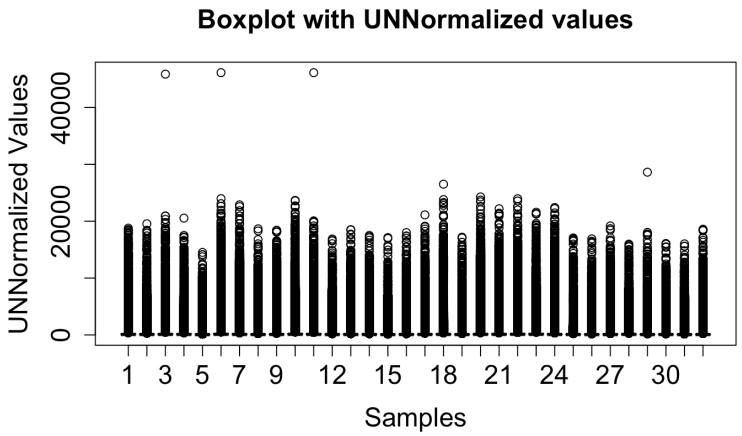
RAW FILE FORMAT	SAMPLES NUMBERS	CONDITIONS	SHORTLISTED CONDITIONS
celfiles	36	Ulcerative colitis-9	Ulcerative colitis-9
		Crohn's disease-19	Crohn's disease-19
		Normal-4	Normal-4
		Indeterminate colitis -2	
		INF(bacterial infectious colitis)-2	

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.

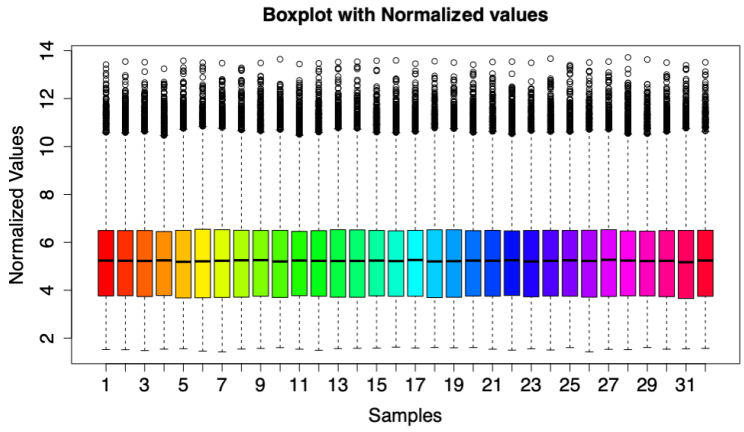
4.3 Normalisation of expressed values

Readaffy function reads the cel files into the R environment and prepares it for normalisation, *exprs* and *rma* are both functions used for normalising the data. Successfully extracting the

expression values of the samples, now QC must be performed on the normalised data to anomalies and outliers. To detect the accurate normalisation of the dataset, Figure 3 shows a *boxplot* plotted to visualise the normalisation of the expression values from the samples.



(a)



(b)

Figure 3: a Boxplot illustration to showcase data variance and normalisation in the samples of the dataset GSE6731. Samples are listed on the x-axis where 32 patient data is listed and the expression values are represented on the y-axis (a) A boxplot showing expression values of the data with unnormalized configuration, (b) Boxplot depicting normalised expression values after quality control of the data. The colour palette can be changed according to the user's choice(note: here the palate has been set as "rainbow" to give a visually appealing aesthetic touch to the box plot).

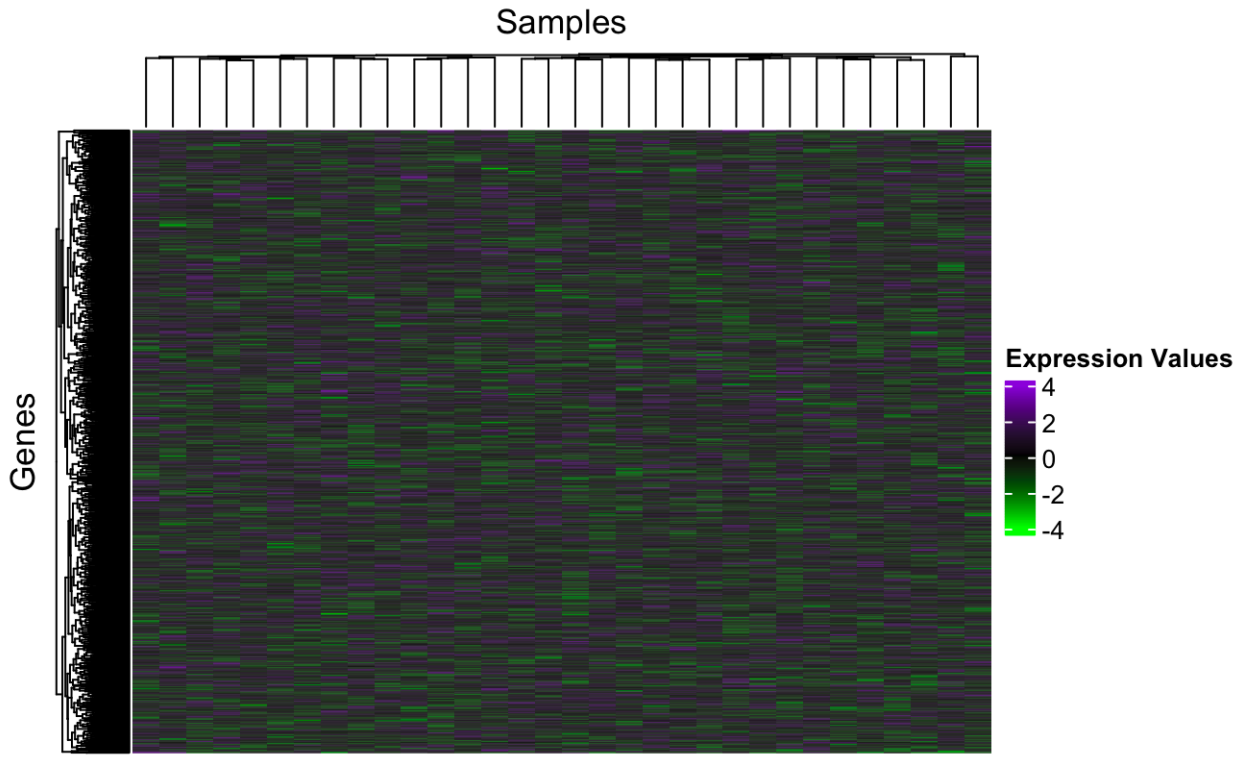


Figure 4: A heatmap constructed with the help of ComplexHeatmap package in R version 4.3.2, in the above graph, columns consist of the 32 samples and the associated regulated genes are somewhere represented in the rows. Due to the large number of expressed values here the map shows significant boxes overlapping with mild to high variance in the values. The **Violet** colour represents the Upregulated genes whereas the downregulated genes are shown in the **green** colour. The value ranges from +6 to -6 in the plot.

4.4 Principal Component Analysis

In the following dataset, a PCA is done using the expression values of Inflammatory Bowel Disease dataset to demonstrate any possible sample clustering within and variation between cohorts. It shows the correlation or variability of the samples from each other and gives a visual characterization to differentiate the normal vs diseased samples. In order to generate the plot, *ggfortify library* has been used in the R pipeline, further the pc data has to be generated where each sample contains the individual pc score using the expression values. The function *autopilot* has been used to generate a pc plot for and has been shown in the figure 5. The Principal Component 1 (PC1) is the first principal component extracted from the data which captures the direction of maximum variance in the data whereas Principal Component 2 (PC2) is the second principal component, orthogonal (perpendicular) to PC1, capturing the next highest amount of variance that is uncorrelated with PC1. The percentage numbers for each

primary component reflect how much of the total variance is explained by that component. For example, if PC1 explains 55.59% of the variance, this means that when the data points are projected onto PC1, this single component accounts for 55.59% of the total variability in the data. In a similar manner if PC2 represents 11% of the variation, it means that PC2 accounts for an additional 11% of overall variability in the data set that PC1 does not explain. It is interpreted that "55.59%" for PC1 and "11.18%" for PC2 indicate that PC1 is the dominating component, accounting for the majority of the variation in the data. PC2 with a percentage that is lower (11.18%), detects less variation, but it provides additional information not described by PC1. When reading a PCA graphic, the axes (PC1 and PC2) do not correspond to the original variables in the dataset.

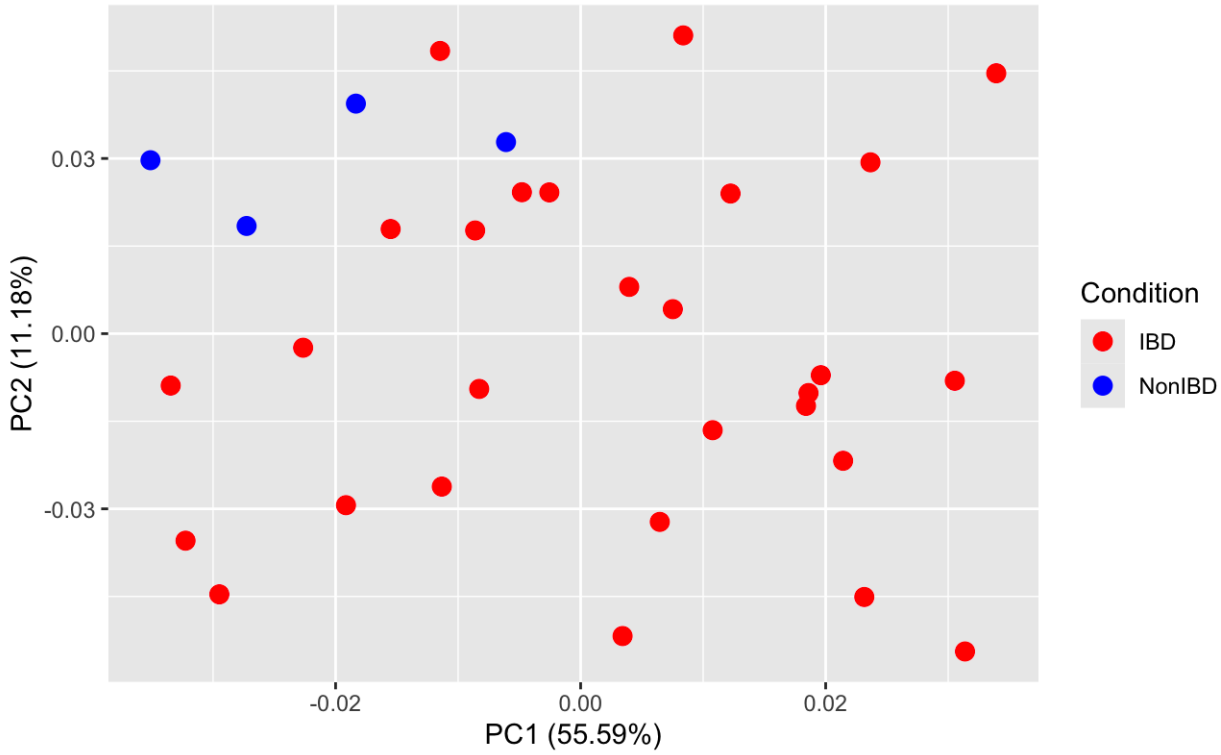
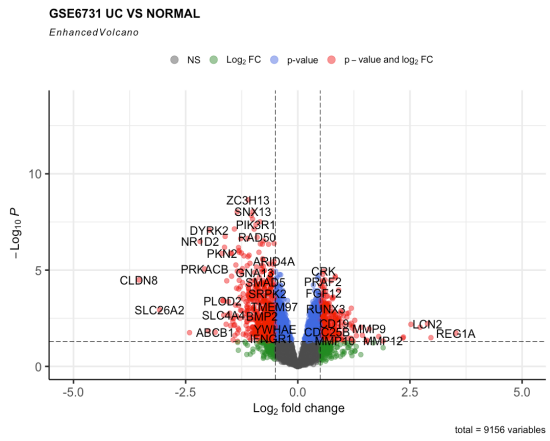


Figure 5: Principal component analysis performed on the IBD dataset GSE6731. The points here represent the samples and are coloured according to the subject cohort. Results are plotted according to the PC1 and PC2 scores from the pc data matrix generated using ggfortify library, with the percent variation explained by the respective axis where the PCA plot is demonstrating variation between IBD and non-IBD samples.The PC1 score on x axis demonstrates a 55.59% showing the maximum variance and the orthogonal PC2 score is 11.18% capturing the next highest amount of variance.

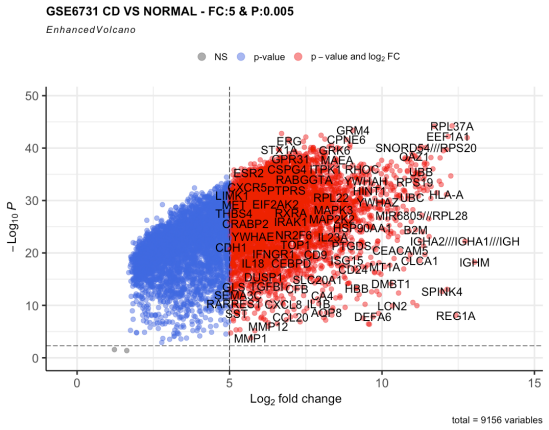
4.5 Differential Genes Expression analysis

S.no	gene.symbol	logfc	AvgExpr	t	P.value	adj.P.val	B
1	RABGGTA	0.112717432	7.300086	1.04211036	3.051403e-01	0.522784076	-5.7643117
2	MAPK3	0.011684912	8.176038	0.06287469	9.502559e-01	0.976666137	-6.2943555
3	TIE1	0.316040658	3.985100	2.71566402	1.055796e-02	0.082077746	-2.9996039
4	CYP2C19	0.083767482	4.166605	0.54176689	5.917174e-01	0.759267407	-6.1507259
5	CXCR5	0.200358888	5.404175	2.21942411	3.364000e-02	0.150200659	-4.0179155

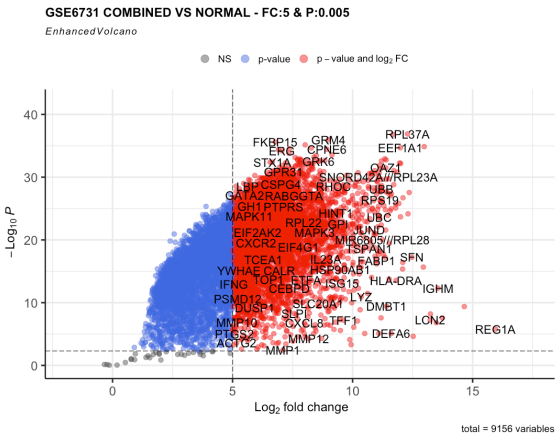
Table 3: A demonstration of the example of how a retrieved set of resulted genes with important variable columns such as *logFC* and P value looks like, the genelist contains upregulated genes responsible for a particular condition in this case Combined UC and CD vs normal.



(a)



(b)

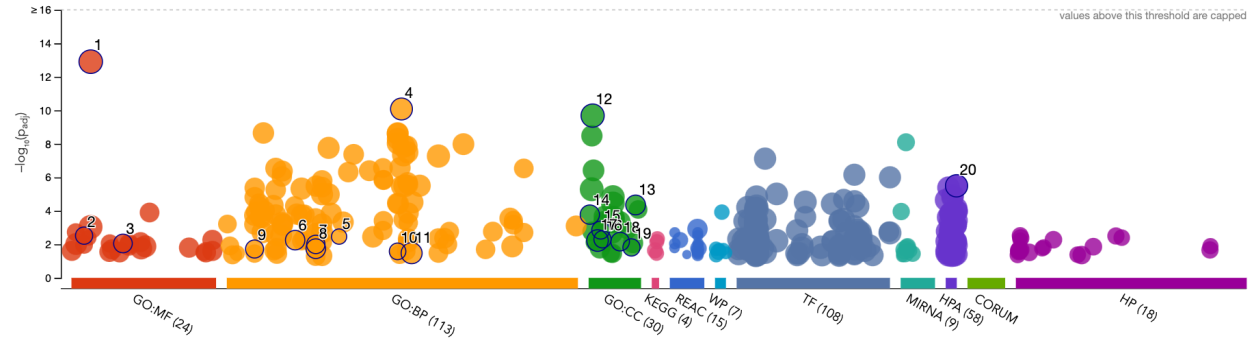


(c)

Figure 6: Differentially expressed genes volcano plot on the condition demonstrating upregulated genes and downregulated genes (a)Ulcerative colitis vs normal, The following data classifies 286 upregulated genes and 1,548 downregulated genes, where fc value was >0.5 and P value set to be < 0.05 (b) Crohn's Disease vs normal with upregulated genes of number 5,100 where fc was set to be >5 and P value < 0.005 (c) combined IBD vs normal genelist, with significant 5,071 upregulated genes where fc was set to be >5 and P value < 0.005. The red spot indicates upregulation, blue indicates downregulation, grey points are non-significant or neutral points that fall between the thresholds for upregulation and downregulation whereas green indicates no significant change.

4.6 Biological Interpretation

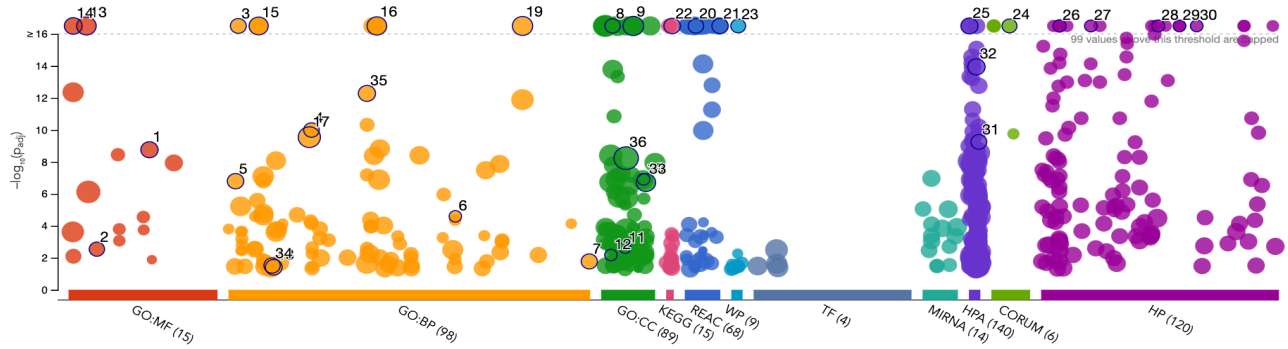
4.6.1 Genes of ulcerative colitis vs normal:



ID	Source	Term ID	Term Name	p_{adj} (query_...
1	GO:MF	GO:0005515	protein binding	1.239×10^{-13}
2	GO:MF	GO:0004674	protein serine/threonine kinase activity	3.018×10^{-3}
3	GO:MF	GO:0030674	protein-macromolecule adaptor activity	8.852×10^{-3}
4	GO:BP	GO:0048856	anatomical structure development	8.274×10^{-11}
5	GO:BP	GO:0034101	erythrocyte homeostasis	3.446×10^{-3}
6	GO:BP	GO:0016477	cell migration	5.615×10^{-3}
7	GO:BP	GO:0030029	actin filament-based process	1.010×10^{-2}
8	GO:BP	GO:0030030	cell projection organization	1.815×10^{-2}
9	GO:BP	GO:0006338	chromatin remodeling	1.897×10^{-2}
10	GO:BP	GO:0048511	rhythmic process	2.623×10^{-2}
11	GO:BP	GO:0051641	cellular localization	3.369×10^{-2}
12	GO:CC	GO:0005737	cytoplasm	2.052×10^{-10}
13	GO:CC	GO:1902494	catalytic complex	4.452×10^{-5}
14	GO:CC	GO:0000785	chromatin	1.677×10^{-4}
15	GO:CC	GO:0030425	dendrite	1.406×10^{-3}
16	GO:CC	GO:0031252	cell leading edge	4.992×10^{-3}
17	GO:CC	GO:0012505	endomembrane system	5.935×10^{-3}
18	GO:CC	GO:0070161	anchoring junction	6.910×10^{-3}
19	GO:CC	GO:0098978	glutamatergic synapse	1.533×10^{-2}
20	HPA	HPA:0590051	thyroid gland; glandular cells[≥Low]	3.191×10^{-6}

Figure 7: The above chart shows the description of the functional enrichment of the genes for Ulcerative colitis condition of IBD samples. Each colour signifies a group where the similar genes are grouped according to their functions, the red colour is for molecular function- protein binding, orange indicates biological process i.e., metabolic or cellular process, green represents cell components like chromosome or cytoplasm, pink, indigo and light blue for signalling pathways, grey colour shows factor binding site motifs, cyan for miRna site, purple shows HPA and maroon for the Hypothetical proteins.

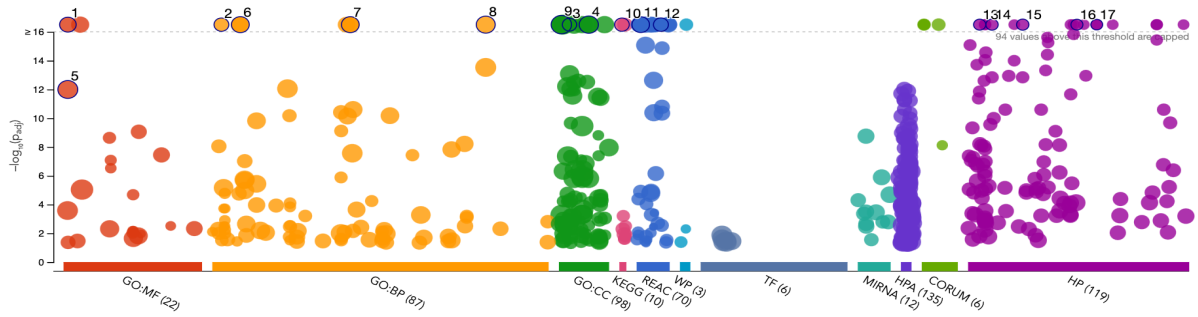
4.6.2 Condition Crohn's disease s normal:



ID	Source	Term ID	Term Name	p_{adj} (query_...
1	GO:MF	GO:0045296	cadherin binding	1.754×10^{-9}
2	GO:MF	GO:0009055	electron transfer activity	2.875×10^{-3}
3	GO:BP	GO:0002181	cytoplasmic translation	6.513×10^{-57}
4	GO:BP	GO:0019730	antimicrobial humoral response	1.024×10^{-10}
5	GO:BP	GO:0001906	cell killing	1.609×10^{-7}
6	GO:BP	GO:0071294	cellular response to zinc ion	2.552×10^{-5}
7	GO:BP	GO:2001242	regulation of intrinsic apoptotic signaling path...	1.681×10^{-2}
8	GO:CC	GO:0022626	cytosolic ribosome	3.012×10^{-69}
9	GO:CC	GO:0070062	extracellular exosome	4.308×10^{-40}
10	GO:CC	GO:0098553	lumenal side of endoplasmic reticulum membr...	1.224×10^{-7}
11	GO:CC	GO:0042589	zymogen granule membrane	2.371×10^{-3}
12	GO:CC	GO:0016327	apicolateral plasma membrane	6.670×10^{-3}
13	GO:MF	GO:0005198	structural molecule activity	6.408×10^{-28}
14	GO:MF	GO:0003735	structural constituent of ribosome	1.952×10^{-35}
15	GO:BP	GO:0006518	peptide metabolic process	3.252×10^{-34}
16	GO:BP	GO:0043603	amide metabolic process	5.900×10^{-30}
17	GO:BP	GO:0019538	protein metabolic process	2.858×10^{-10}
18	GO:BP	GO:0009199	ribonucleoside triphosphate metabolic process	3.909×10^{-2}
19	GO:BP	GO:1901566	organonitrogen compound biosynthetic process	1.067×10^{-22}
20	REAC	REAC:R-HSA-7...	Eukaryotic Translation Initiation	1.875×10^{-55}
21	REAC	REAC:R-HSA-7...	rRNA processing	1.971×10^{-43}
22	KEGG	KEGG:05171	Coronavirus disease - COVID-19	7.438×10^{-38}
23	WP	WP:WP477	Cytoplasmic ribosomal proteins	1.963×10^{-57}
24	CORUM	CORUM:3055	Nop56p-associated pre-rRNA complex	8.573×10^{-26}
25	HPA	HPA:0030892	appendix; goblet cells[≥Medium]	8.639×10^{-17}
26	HP	HP:0001896	Reticulocytopenia	1.751×10^{-17}
27	HP	HP:0005532	Macrocytic dyserythropoietic anemia	7.537×10^{-20}
28	HP	HP:0012410	Pure red cell aplasia	3.071×10^{-19}
29	HP	HP:0030270	Elevated red cell adenosine deaminase activity	1.581×10^{-20}
30	HP	HP:0031640	Abnormal radial artery morphology	1.581×10^{-20}
31	HPA	HPA:0481353	small intestine; paneth cells[High]	5.622×10^{-10}
32	HPA	HPA:0400892	rectum; goblet cells[≥Medium]	1.144×10^{-14}
33	GO:CC	GO:0099503	secretory vesicle	1.951×10^{-7}
34	GO:BP	GO:0009060	aerobic respiration	3.042×10^{-2}
35	GO:BP	GO:0042254	ribosome biogenesis	5.241×10^{-13}
36	GO:CC	GO:0043226	organelle	5.861×10^{-9}

Figure 8: This chart shows the description of the functional enrichment of the genes for Crohn's disease condition of IBD samples. Each colour signifies a group where the similar genes are grouped according to their functions, the red colour is for molecular function- protein binding, orange indicates biological process i.e., metabolic or cellular process, green represents cell components like chromosome or cytoplasm, pink, indigo and light blue for signalling pathways, cyan for miRna site, purple shows HPA and maroon for the Hypothetical proteins.

4.6.3 Condition combined IBD vs normal:



ID	Source	Term ID	Term Name	p_{adj} (query_...
1	GO:MF	GO:0003735	structural constituent of ribosome	1.434×10^{-33}
2	GO:BP	GO:0002181	cytoplasmic translation	1.254×10^{-55}
3	GO:CC	GO:0022626	cytosolic ribosome	9.433×10^{-66}
4	GO:CC	GO:0070062	extracellular exosome	1.471×10^{-42}
5	GO:MF	GO:0003723	RNA binding	1.015×10^{-12}
6	GO:BP	GO:0006518	peptide metabolic process	4.194×10^{-36}
7	GO:BP	GO:0043604	amide biosynthetic process	2.715×10^{-33}
8	GO:BP	GO:1901566	organonitrogen compound biosynthetic process	1.960×10^{-23}
9	GO:CC	GO:0005576	extracellular region	1.044×10^{-29}
10	KEGG	KEGG:03010	Ribosome	5.813×10^{-45}
11	REAC	REAC:R-HSA-2...	Cellular responses to stress	9.477×10^{-33}
12	REAC	REAC:R-HSA-9...	Regulation of expression of SLITs and ROBOs	3.066×10^{-50}
13	HP	HP:0001227	Abnormality of the thenar eminence	8.503×10^{-18}
14	HP	HP:0002669	Osteosarcoma	2.411×10^{-17}
15	HP	HP:0006758	Malignant genitourinary tract tumor	3.949×10^{-17}
16	HP	HP:0012410	Pure red cell aplasia	4.235×10^{-19}
17	HP	HP:0030270	Elevated red cell adenosine deaminase activity	2.181×10^{-20}

version	e111_eg58_p18_30541362
date	22/04/2024, 03:44:35
organism	hsapiens

g:Profiler

Figure 9: This chart shows the description of the functional enrichment of the genes for IBD condition of IBD samples. Each colour signifies a group where the similar genes are grouped according to their functions, the red colour is for molecular function- protein binding, orange indicates biological process i.e., metabolic or cellular process, green represents cell components like chromosome or cytoplasm, pink, indigo and light blue for signalling pathways, cyan for miRna site, purple shows HPA and maroon for the Hypothetical proteins.