



SAPIENZA
UNIVERSITÀ DI ROMA

GENE EXPRESSION DATA ANALYSIS

CROHN'S DISEASE

Bioinformatics course - A.Y. 2022/2023

MSc in Computer Science

Faculty of Information Engineering, Informatics and Statistics
Sapienza University of Rome

Ballanti Chiara

ballanti.1844613@studenti.uniroma1.it



01

Introduction

Crohn's disease

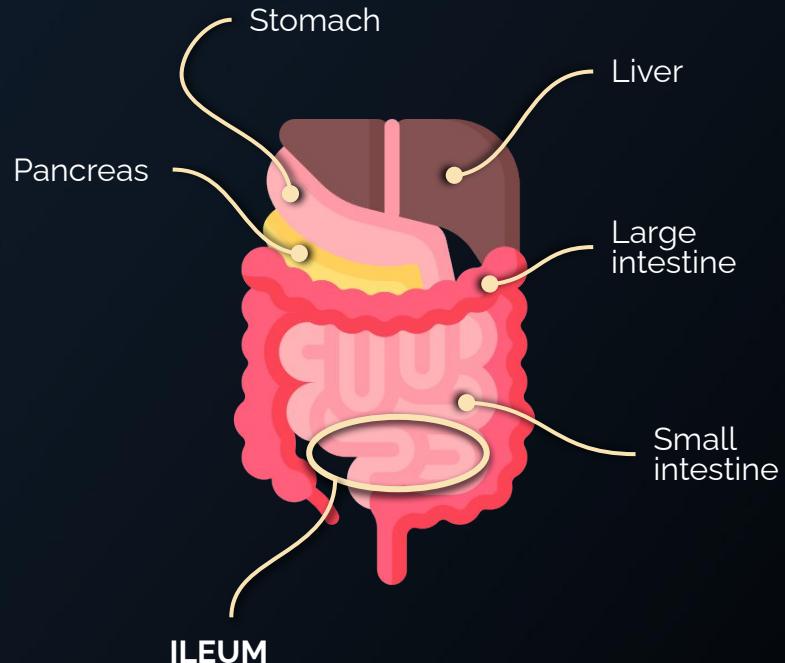
Crohn's disease

Crohn's disease is a chronic inflammatory bowel disease (**IBD**) characterized by inflammation of the gastrointestinal tract.

It is a complex disorder with a multifactorial etiology, involving genetic, environmental, and immunological factors.

Crohn's disease affects millions of people worldwide, causing significant morbidity and reducing their quality of life. It is a chronic condition that requires lifelong management and often leads to serious complications.

Understanding the genetic basis of Crohn's disease is crucial for unraveling its pathogenesis, developing targeted therapies, and improving patient outcomes.





02

Dataset

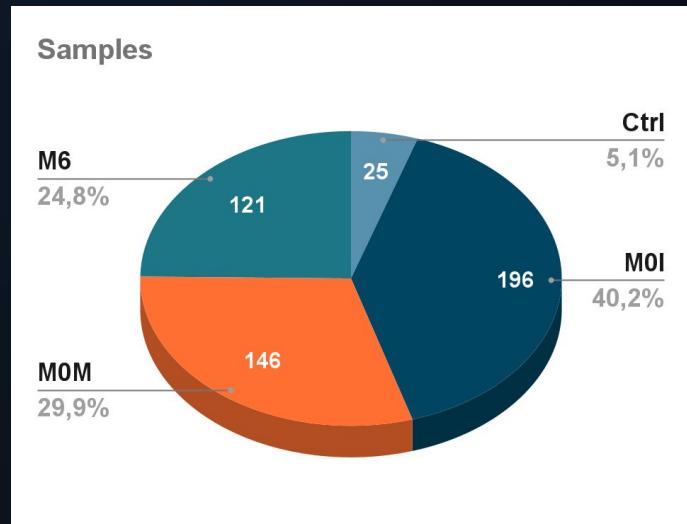
Gene Expression Omnibus
(GEO) series "[GSE186582](#)"

Dataset

The diagnosis of the disease and its monitoring are done by taking biopsies carried out during colonoscopies and gastroscopies.

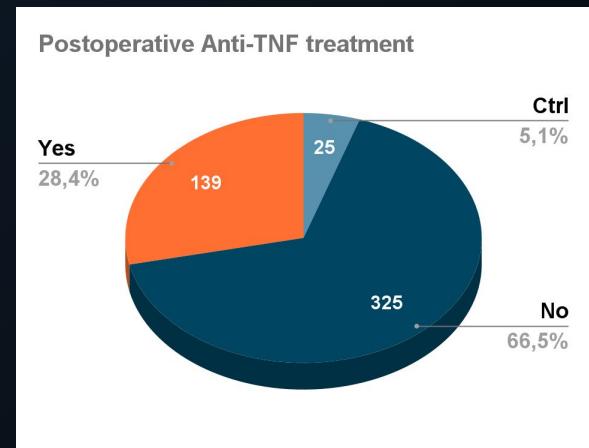
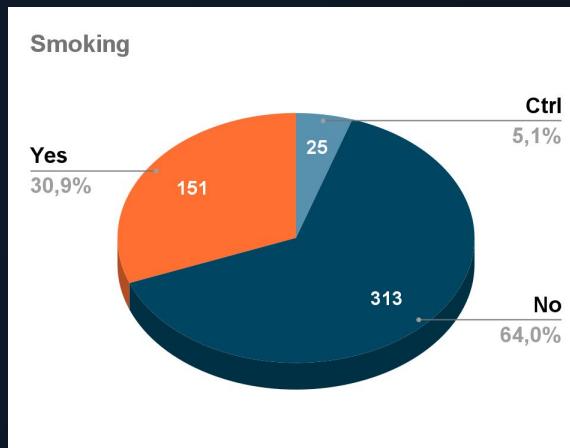
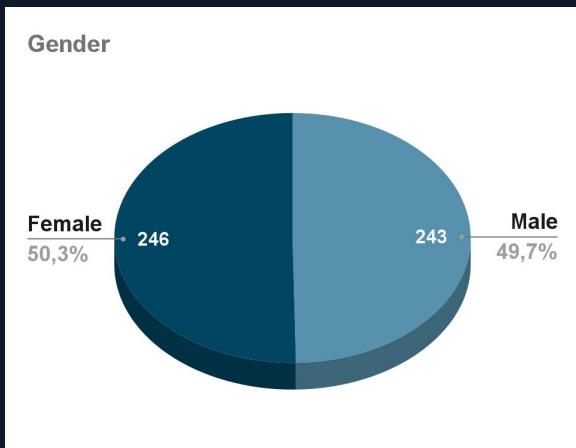
489 total samples:

- 196 (**M0I**): inflamed ileum;
- 147 (**MoM**): ileal margin at time of surgery;
- 121 (**M6**): post-operative endoscopy six months later;
- 25 (**Ctrl**): ileal non-IBD control biopsies from patients who underwent ileocecal resection for colonic tumour with a healthy ileum.



Dataset

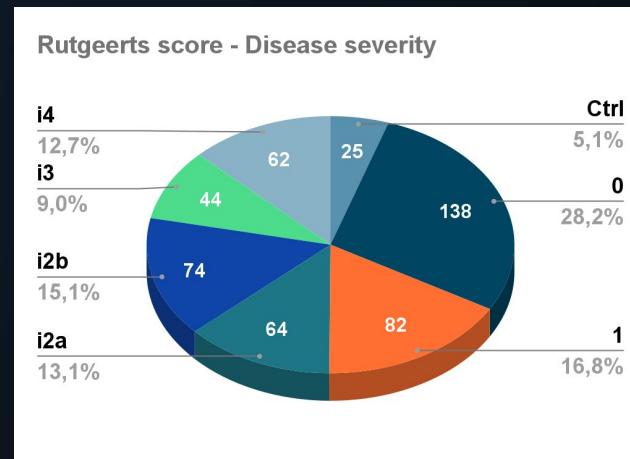
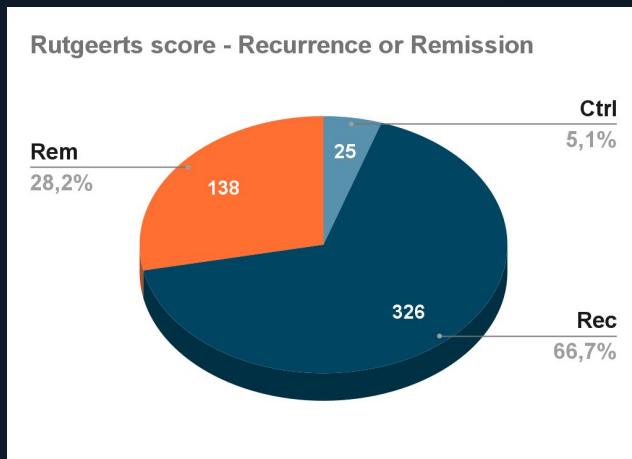
Clinical features provided by metadata:



Dataset

Clinical features provided by metadata:

- **Rutgeerts' endoscopic postoperative recurrence classification score:** it indicates the level of recurrence and disease severity.



Dataset

Starting from the information provided by the GEO dataset, two datasets have been extracted:

- **Control** samples: **25** patients
 - "Ctrl"
- **Case** samples: **31** patients
 - "Mol"
 - Postoperative Anti-TNF treatment: No
 - Rutgeerts' score: Recurrence and (i4 or i3)
 - Gender and Smoking: ignored

The datasets show the expression of each recorded gene (rows) for each patient (columns).

The **goal** of the study is to analyze the expression of 20.061 genes over 56 patients using RNA sequencing data in order to identify genes involved in this disease.



03

Differential expression analysis

— Aim —

Find a list of genes that are differential expressed between **case** and **control** samples and characterize them.

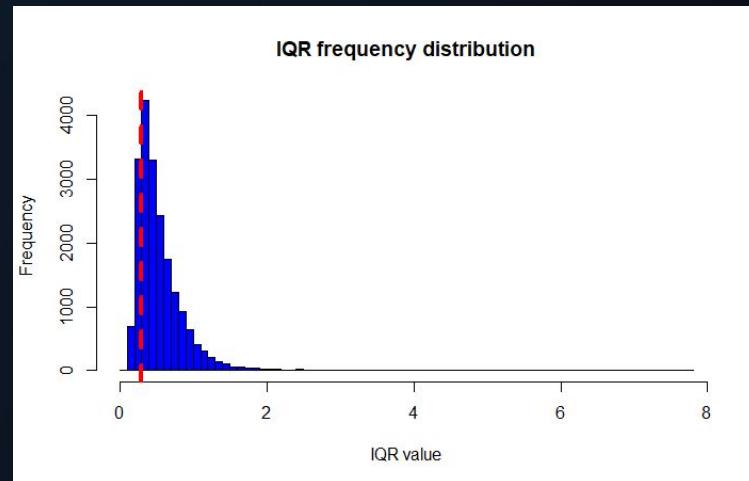
1. Pre-processing

1. Remove the genes whose mean of the expression data is equal to zero in both control and case samples.

1. Pre-processing

1. Remove the genes whose mean of the expression data is equal to zero in both control and case samples.
2. Removing genes with low variation ($IQR \leq 0.2$)

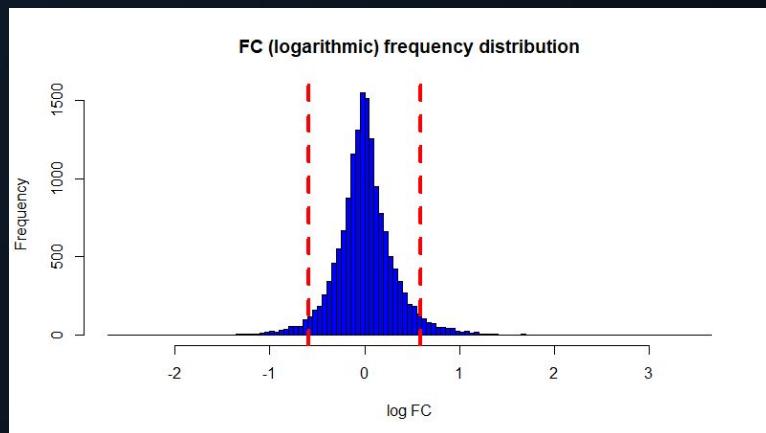
From 20.061 to **16.048** genes.



2. Filtering

1. Remove the genes according to the Fold Change ($|\log FC| \leq 2.0$)

From 16.048 to 856 genes.



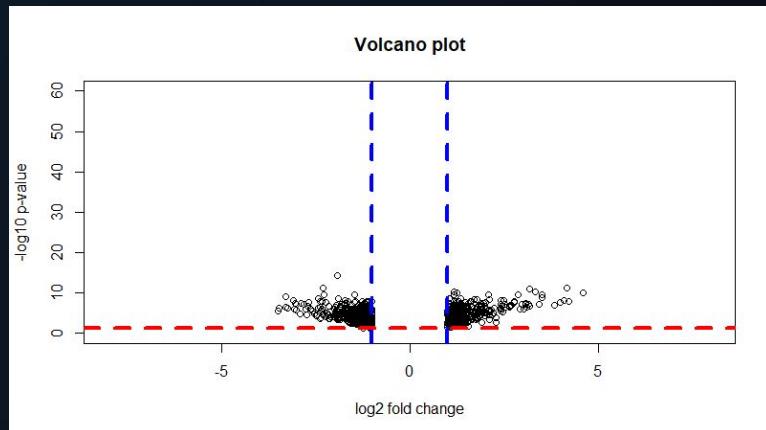
2. Filtering

1. Remove the genes according to the Fold Change ($|\log\text{FC}| \leq 2.0$)

From 16.048 to 856 genes.

2. Remove the genes according to the adjusted p-values and the significance level ($\text{adj_p-val} > 0.05$)

From 856 to **824** genes.

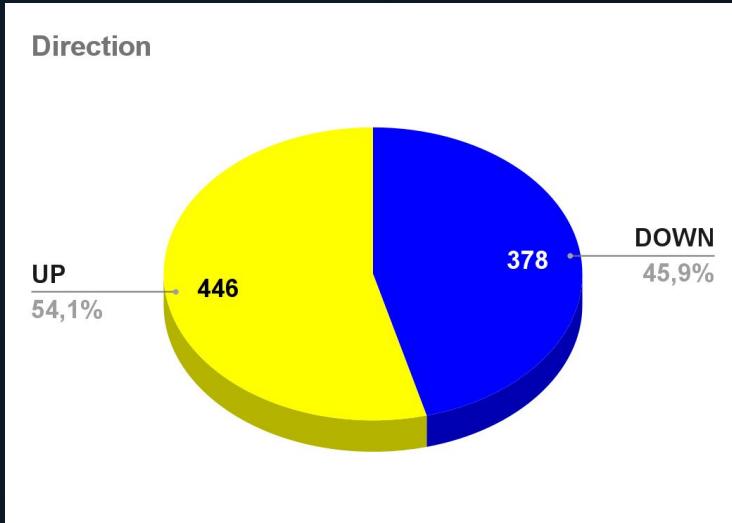


824 DEGs

Genes with a statistically significant change in expression between control and case conditions

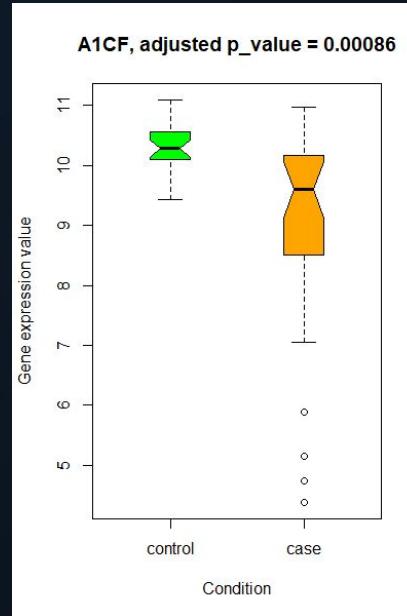
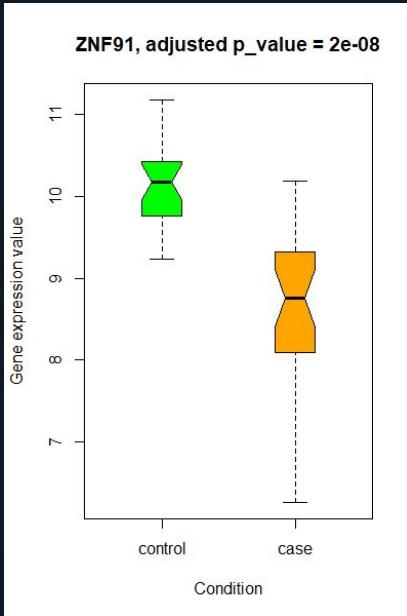
Differential expression analysis

Results: A total of **824 DEGs**, i.e., those genes with a statistically significant change in expression between the control and case conditions, were extracted.



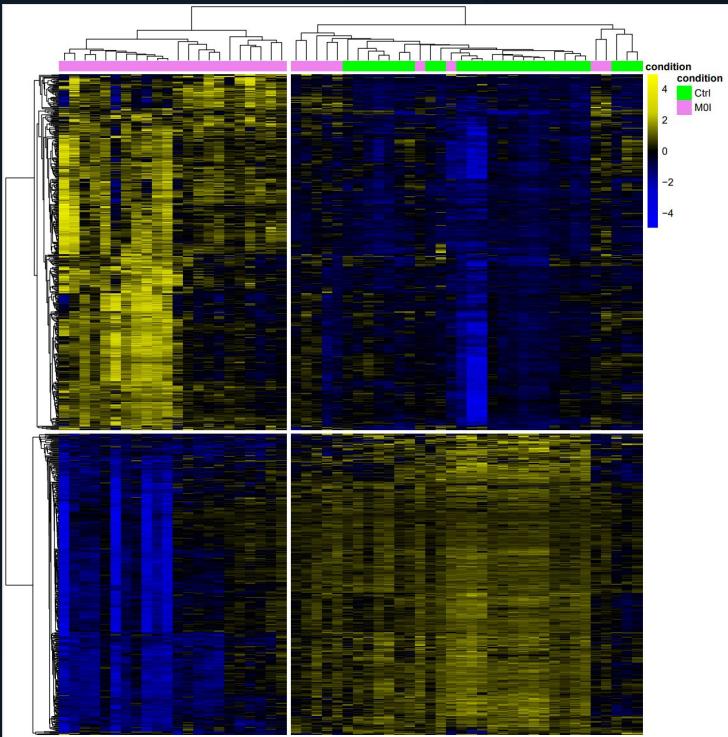
Differential expression analysis

Most **down**-regulated differential expressed gene.



Most **up**-regulated differential expressed gene.

Differential expression analysis





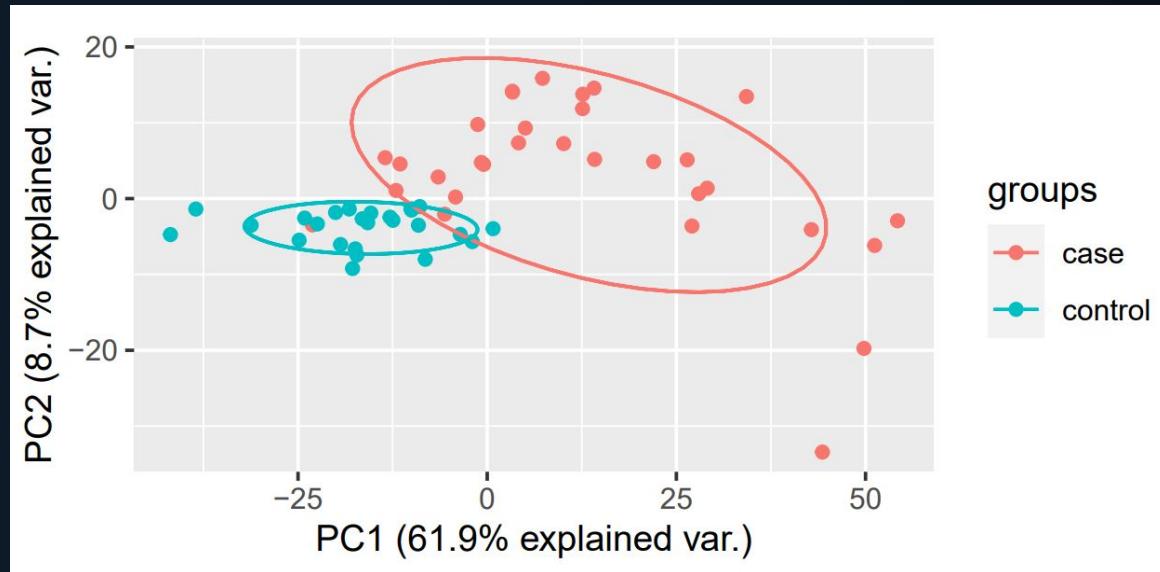
04

PCA

Principal Component Analysis

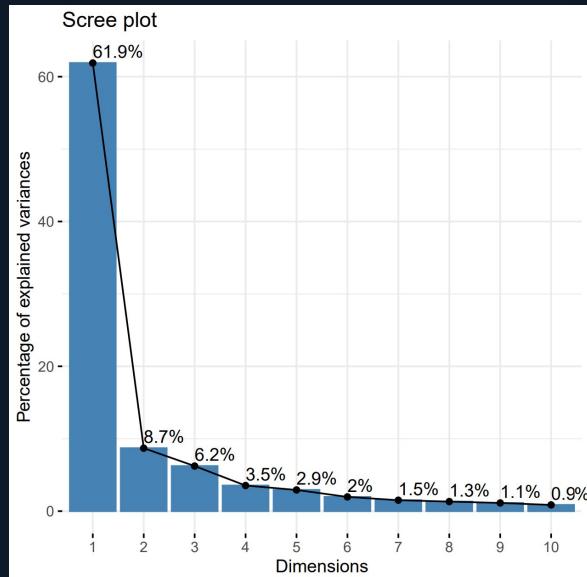
Principal Component Analysis

Score plot: provides a visual representation of the distribution of samples in the principal component space, highlighting patterns, clusters, similarities, and differences among samples based on their gene expression profiles.

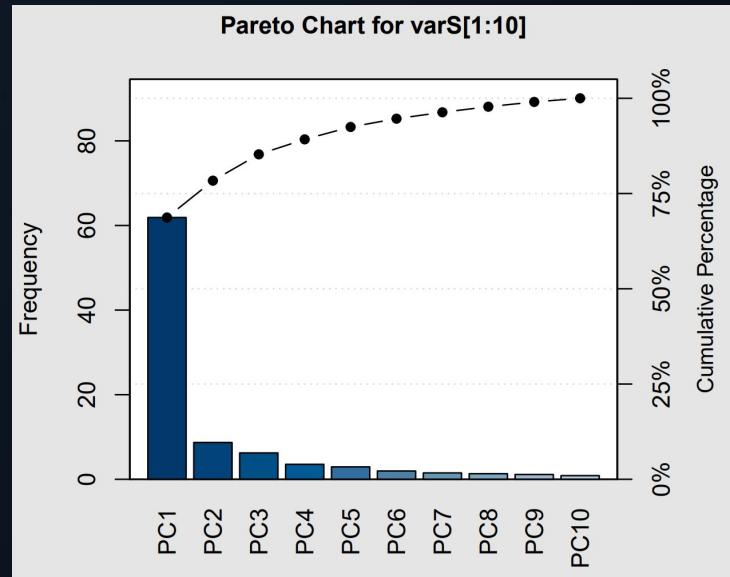


Principal Component Analysis

Scree plot: visual representation of the eigenvalues or variances explained by each principal component.

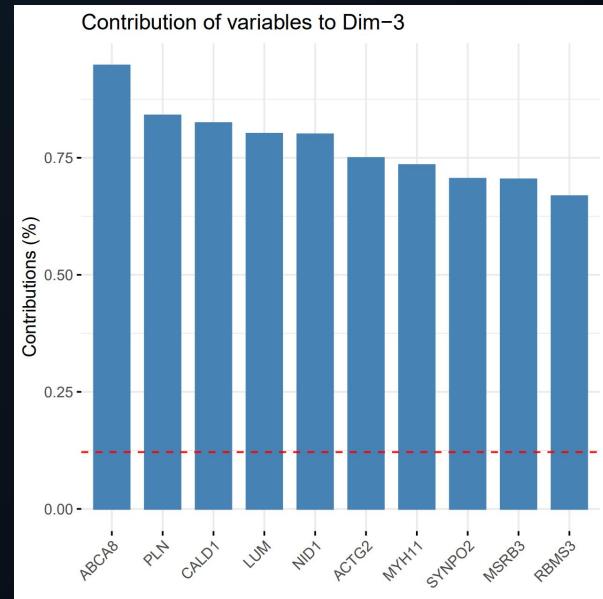
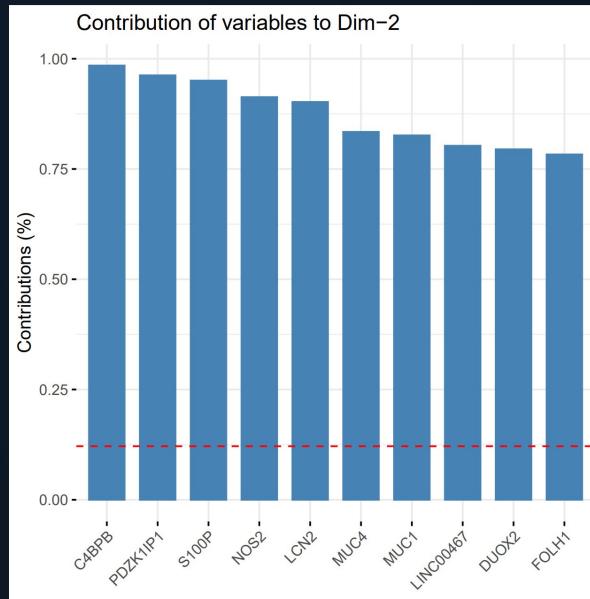
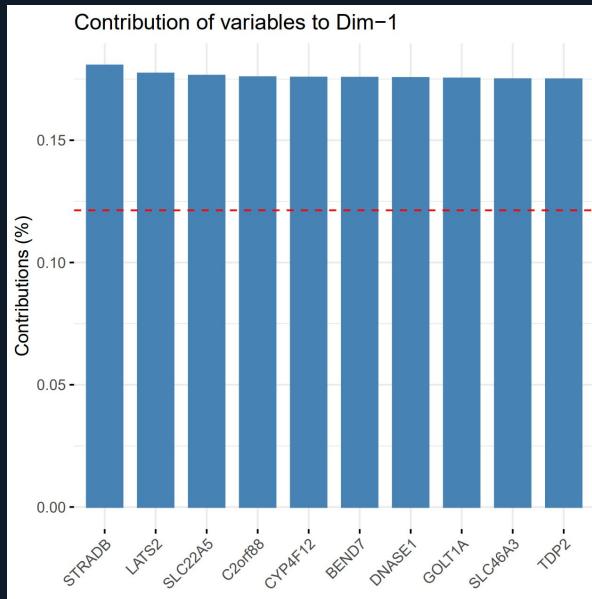


Pareto chart: visual representation of the cumulative variance explained by each principal component.



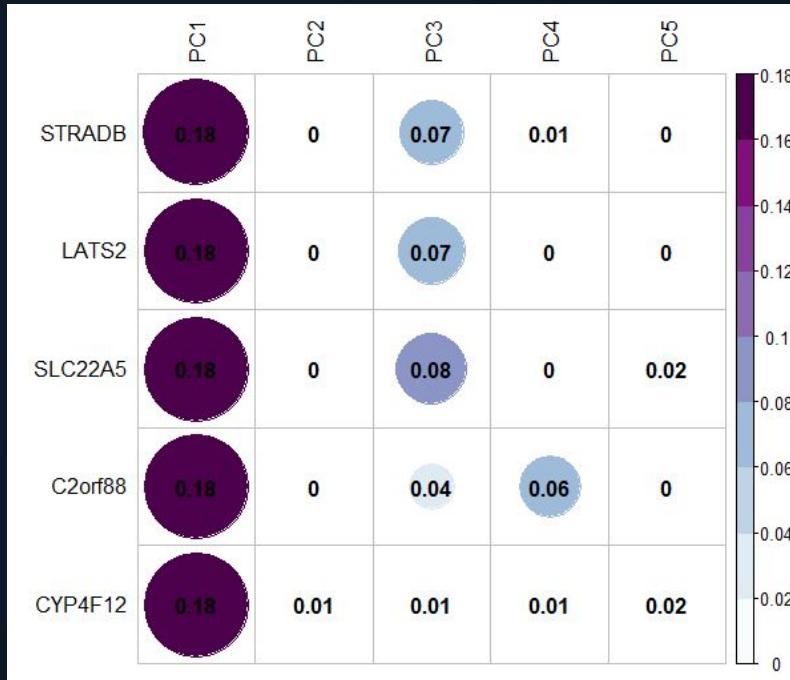
Principal Component Analysis

Visual representation of the contribution of genes to each principal component (PC).



Principal Component Analysis

Visual representation of the contribution of genes to each principal component (PC).





05

Functional enrichment analysis

Tool: EnrichR

Databases



Transcription

- TRANSFAC and JASPAR PWMs



Ontologies

- GO Cellular Component (2023)
- GO Molecular Function (2023)
- GO Biological Process (2023)



Pathways

- Reactome (2022)
- KEGG Human (2021)



Diseases / Drugs

- DisGeNET
- GWAS Catalog (2023)

Up-regulated
genes

IBD and Arthritis

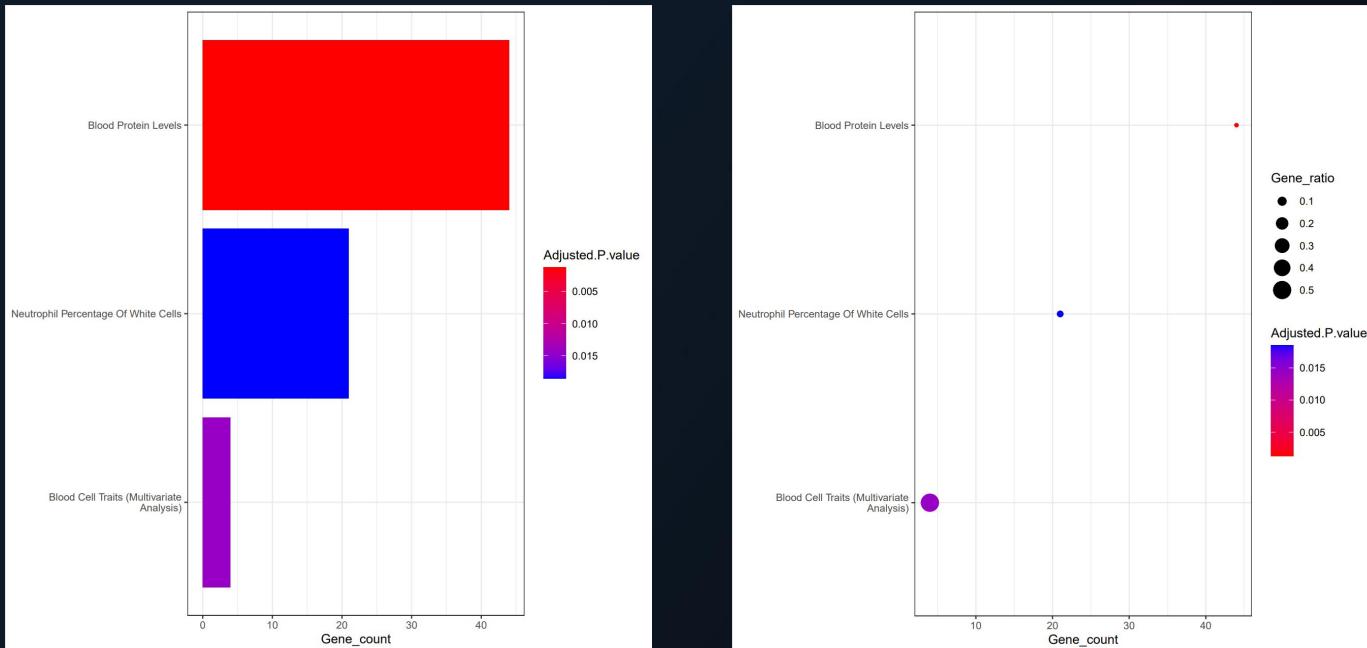
There is a strong connection between inflammatory bowel disease (IBD) and arthritis. In fact, arthritis is one of the most common extra-intestinal manifestations of IBD.

Patients with IBD have a greater risk of developing RA [\[1\]](#).

More in general, Immune-mediated inflammatory diseases (IMIDs) share similar genetic factors, environmental precipitants, and pathophysiological mechanisms. This study demonstrates that the risk of developing a second IMID is significantly higher for individuals who have already experienced a first IMID [\[2\]](#) [\[3\]](#).

GWAS Catalog (2023)

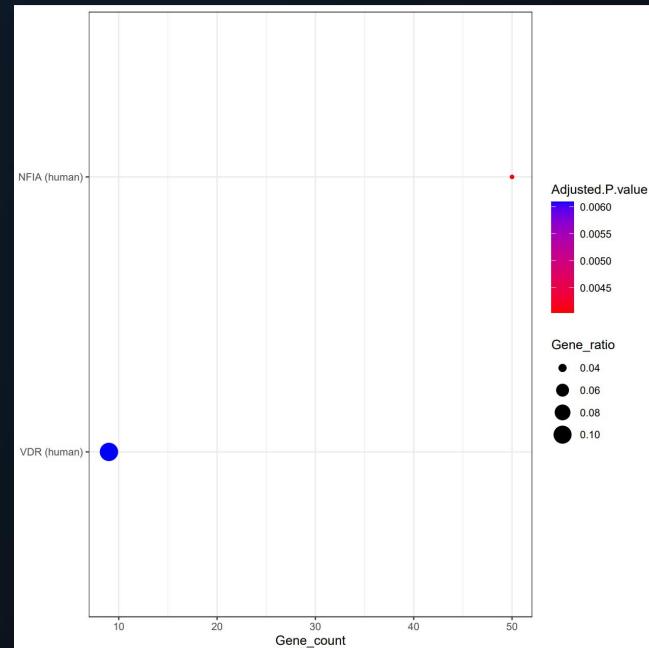
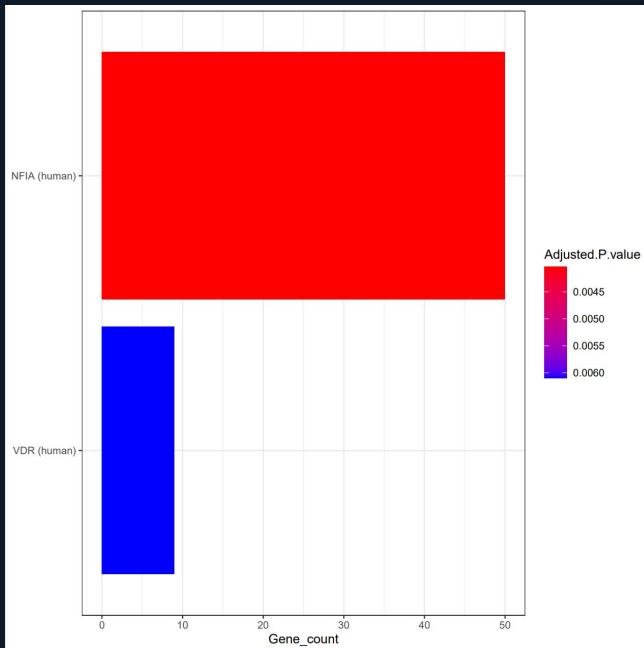
It is a collection of results from GWAS, studies that analyze genetic variations across the genome in large populations to identify associations between genetic markers and traits or diseases.



Down-regulated
genes

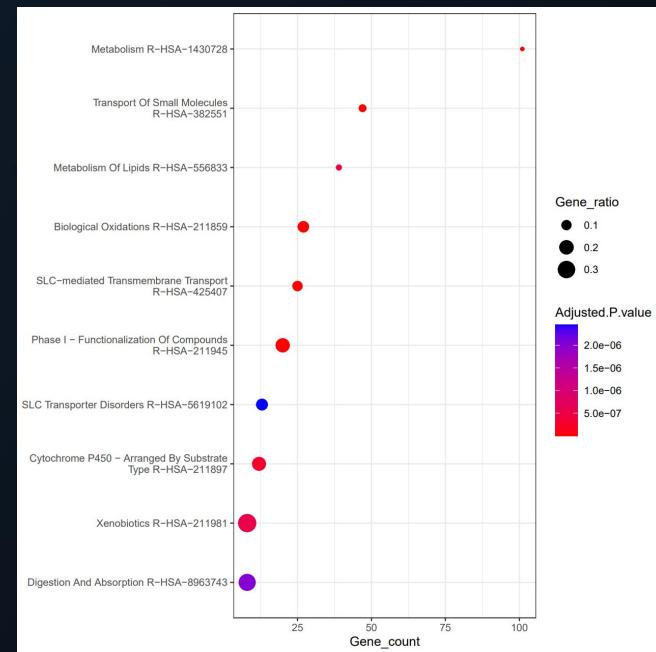
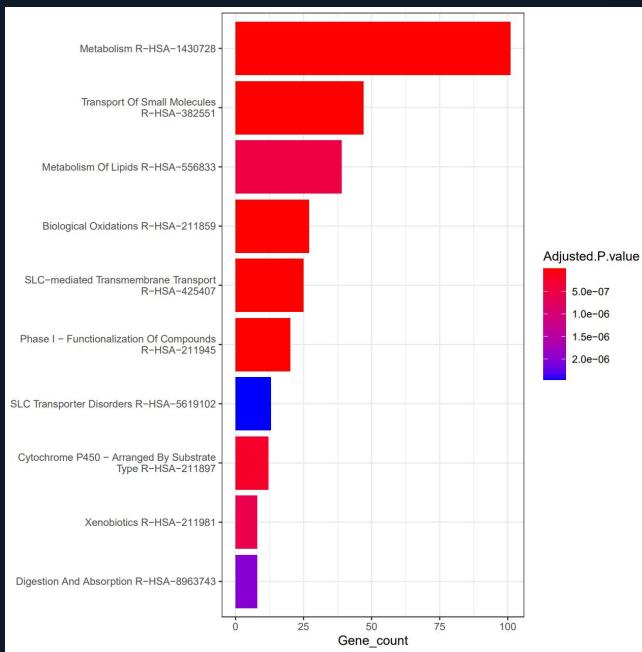
TRANSFAC and JASPAR PWMs

2 databases that provide curated and experimentally validated PWMs that allow for the identification of potential binding sites in DNA sequences.



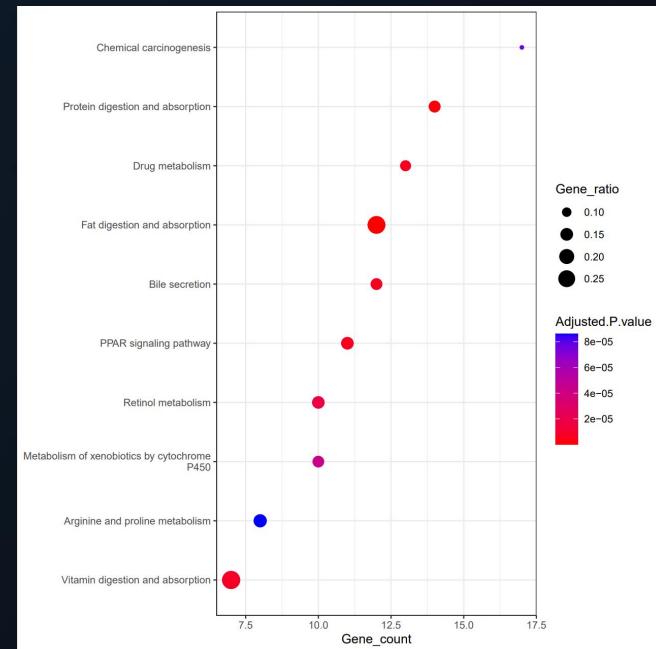
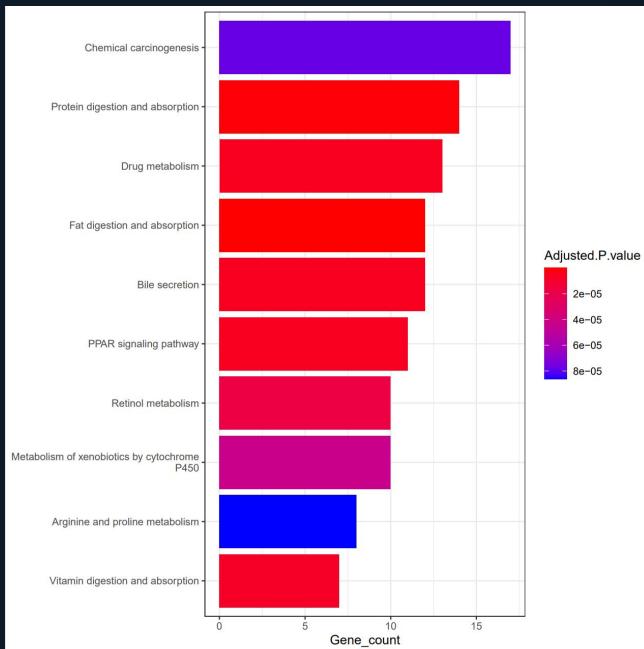
Reactome (2022)

It provides an extensive collection of molecular-level information about biological processes and pathways. It covers a wide range of biological areas.



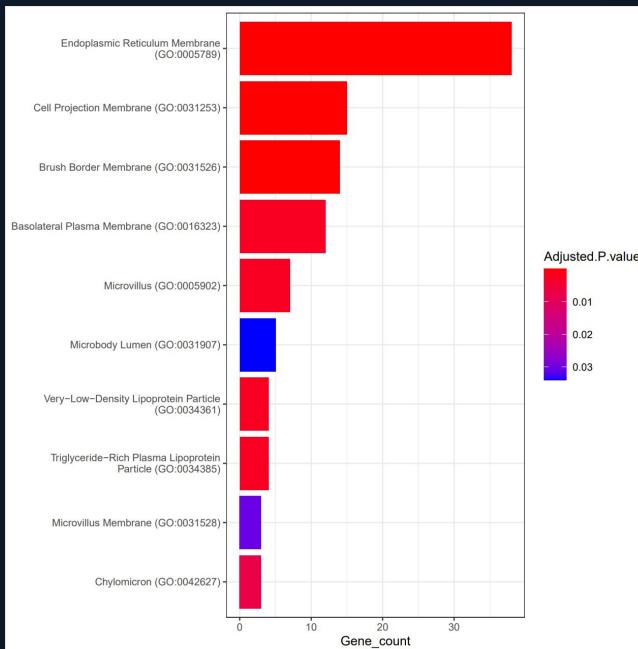
KEGG Human (2022)

It provides a collection of curated and annotated biological pathways specific to humans for studying molecular interactions, cellular processes, and disease-related pathways.

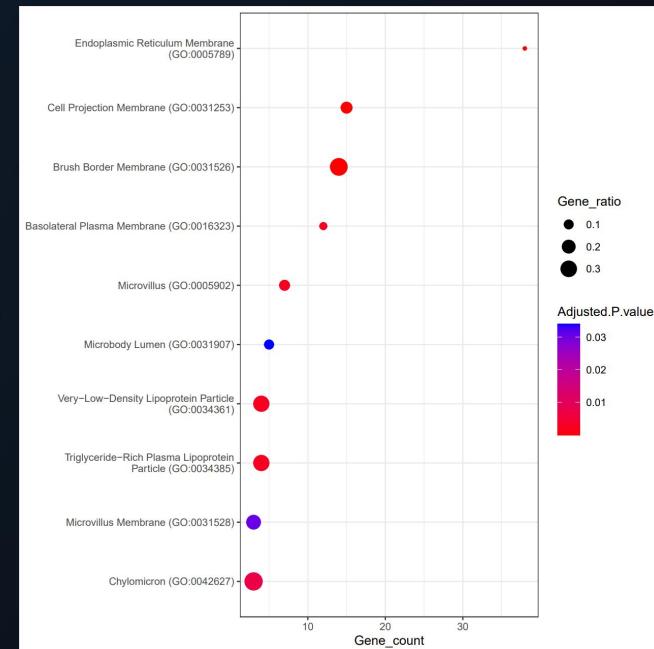


GO Cellular Component (2023)

It is a structured vocabulary within the Gene Ontology that describes the cellular localization or subcellular location of gene products.

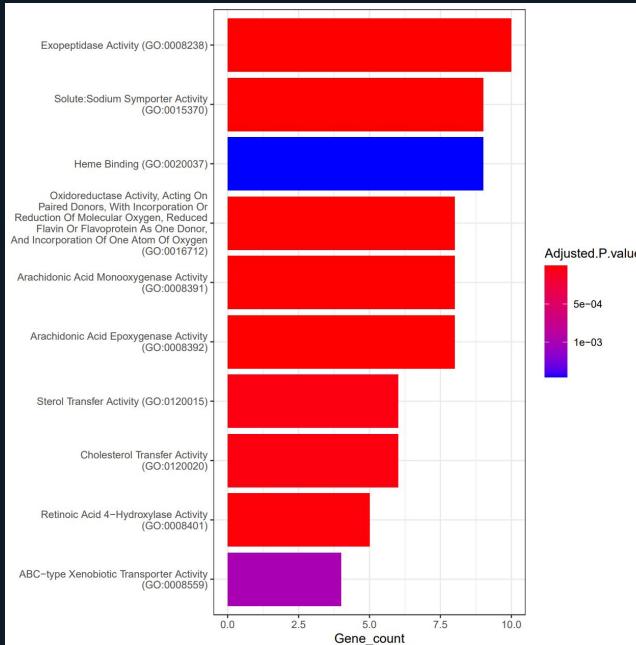


Down-regulated
DEGs

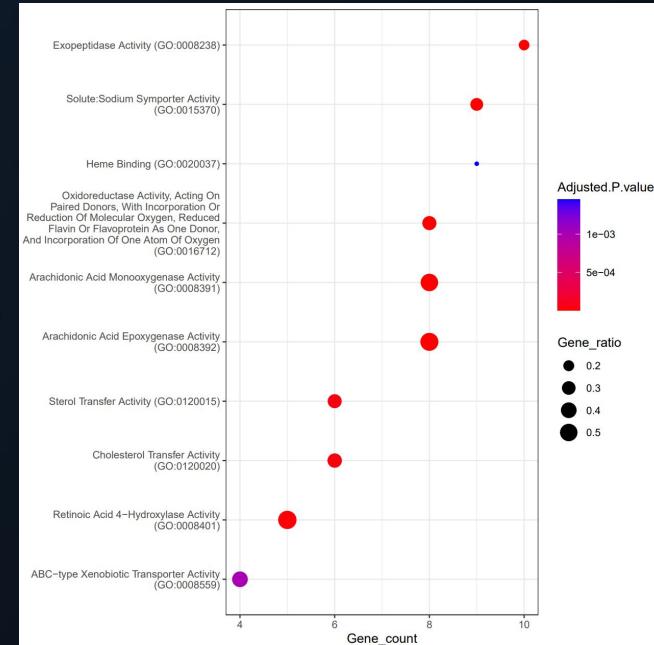


GO Molecular Function (2023)

It is a structured vocabulary within the Gene Ontology that describes the specific biochemical activities or functions performed by gene products.

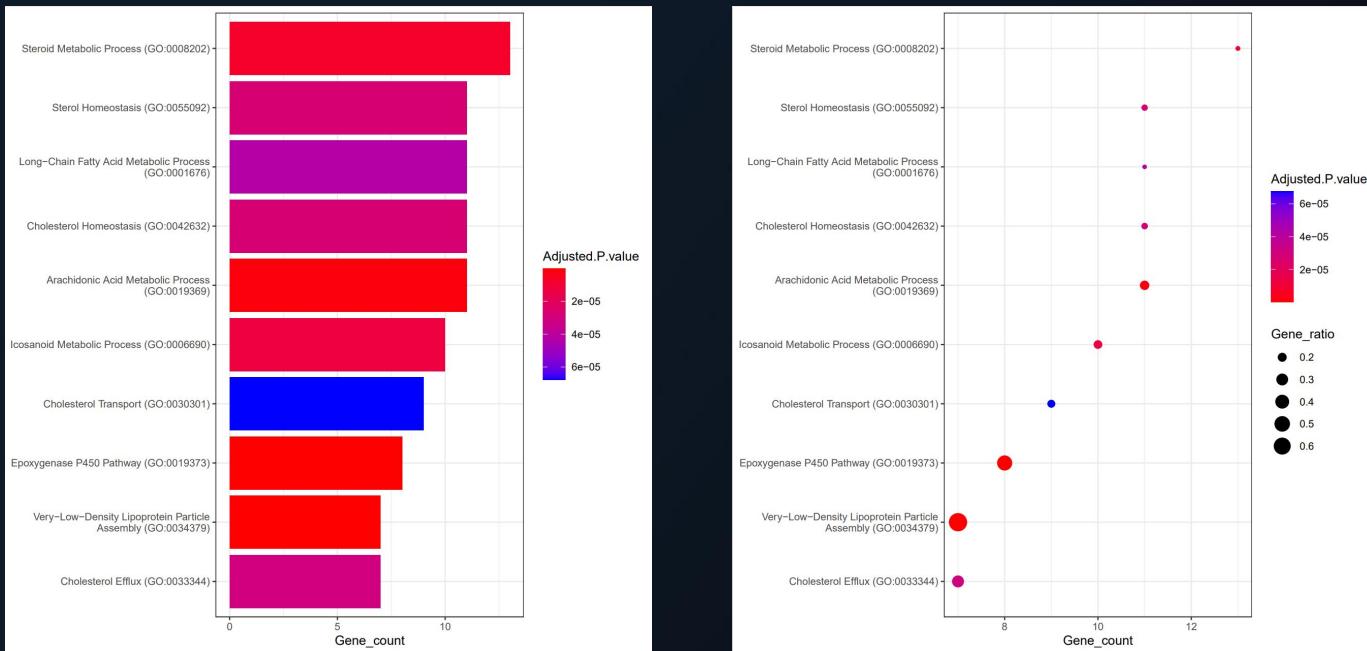


Down-regulated
DEGs



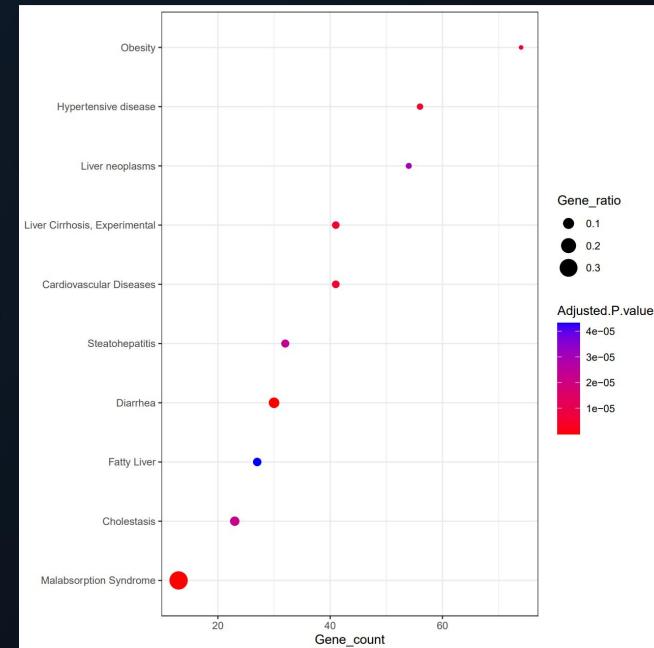
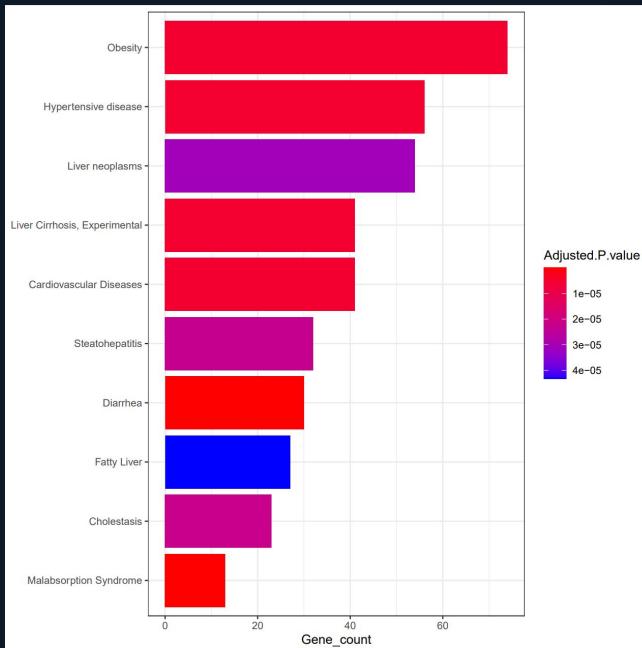
GO Biological Process (2023)

It is a structured vocabulary within the Gene Ontology that describes the series of events or interactions among gene products that lead to a specific biological outcome or result.



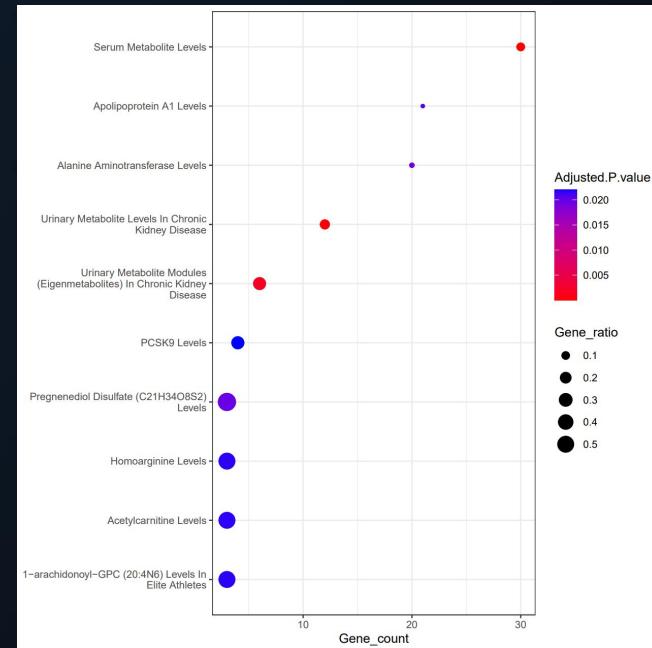
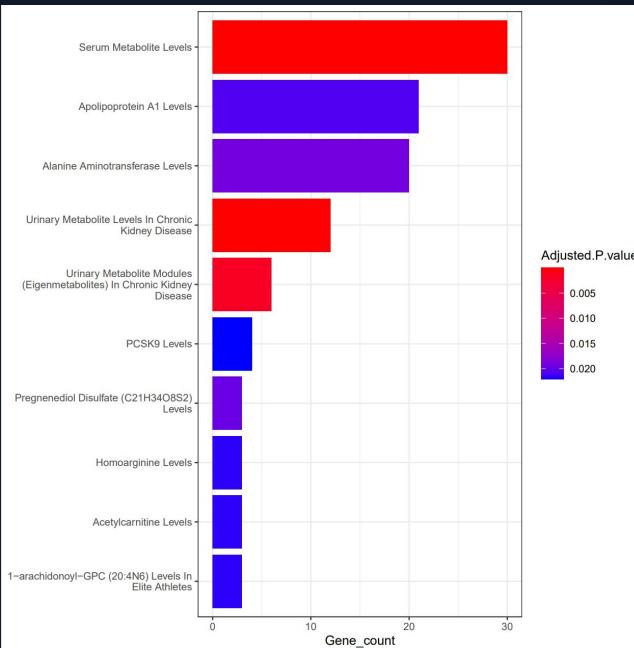
DisGeNET

It focuses on the integration and analysis of gene-disease associations. It provides a wealth of information about the relationship between genetic variants, genes, and human diseases.



GWAS Catalog (2023)

It is a collection of results from GWAS, studies that analyze genetic variations across the genome in large populations to identify associations between genetic markers and traits or diseases.





06

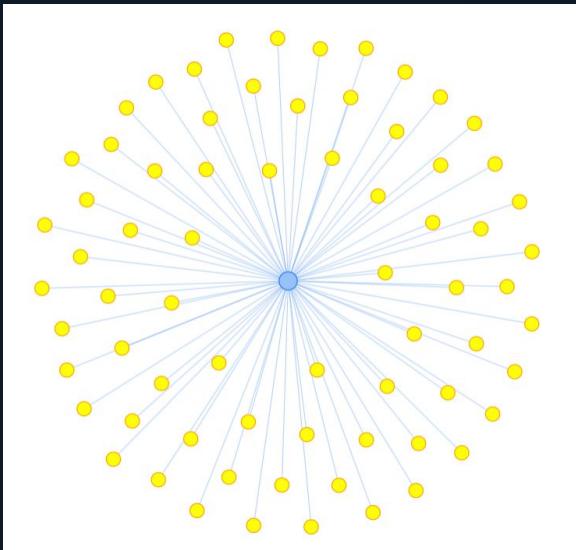
Enrichment analysis of miRNA targets

Tool: MIENTURNET

miRTarBase

It is a comprehensive database that focuses on experimentally validated microRNA-target interactions (MTIs).

Network microRNA/Genes



miRNA: **hsa-miR-124-3p**

Degree: 70

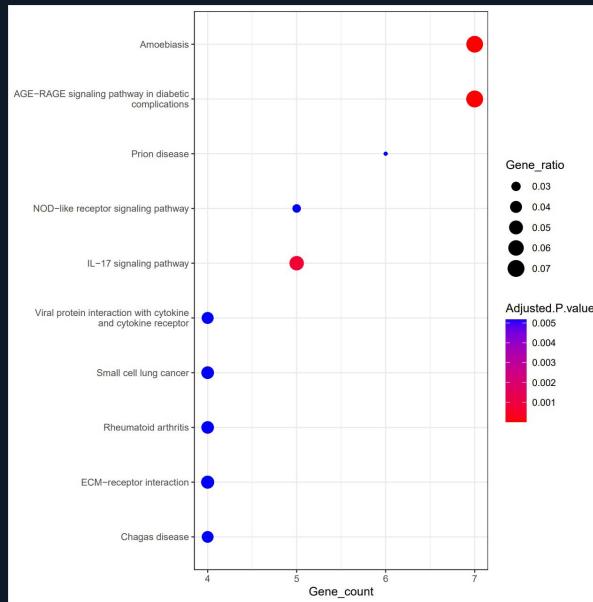
Robustness:

- Weak to targeted attack
- Strong to random attack

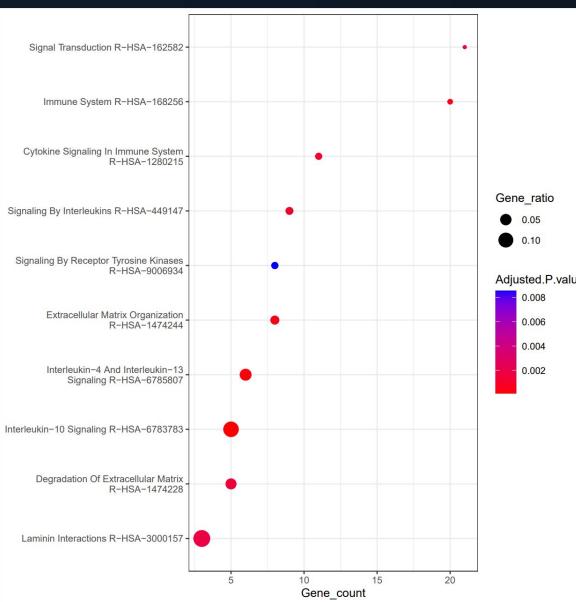
miRTarBase

Functional enrichment analysis of genes targeted by **hsa-miR-124-3p**

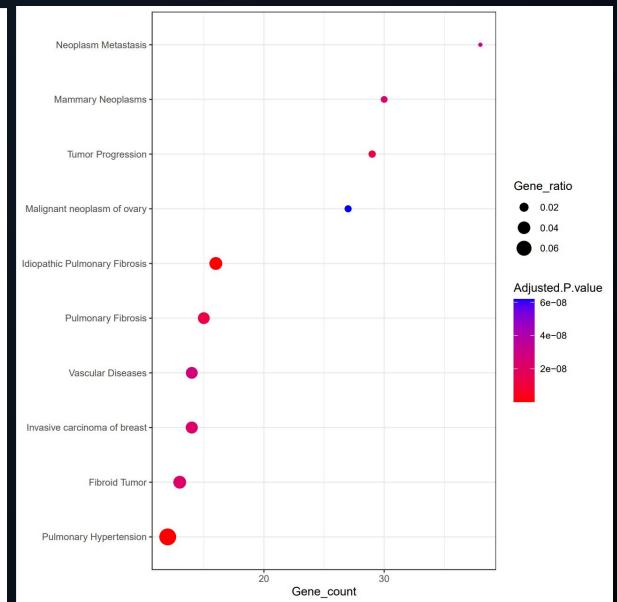
KEGG Human (2022)



Reactome (2022)



DisGeNET



hsa-miR-124

1. Suppresses Aryl hydrocarbon receptor (AhR) expression and increases pro-inflammatory cytokine production [4] [5].
2. Higher probability of damaging effects the intestinal barrier and promote inflammatory bowel disease development in elderly populations [6].

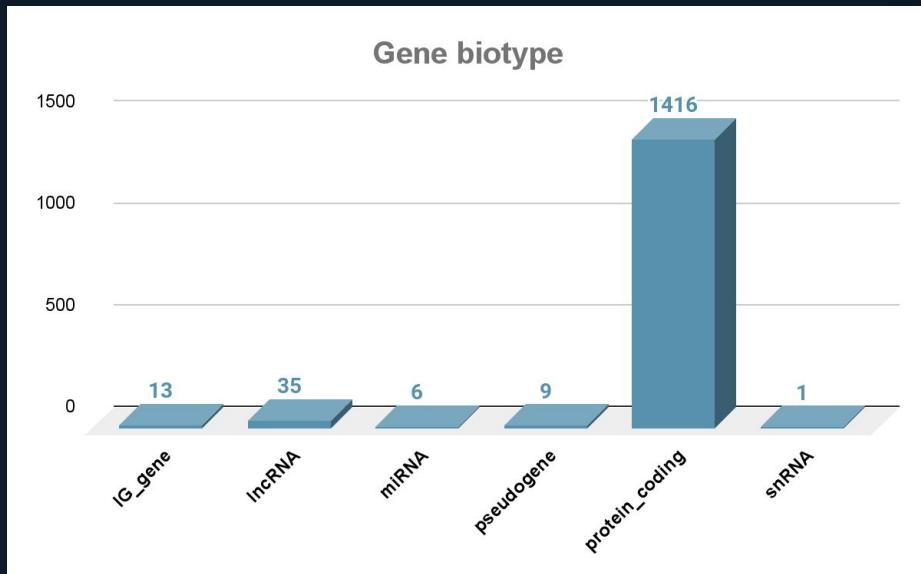


07

Long non-coding RNA analysis

Query on Biomart

Genes biotype



IG_gene: immunoglobulin gene

lncRNA: long non-coding RNA

miRNA: microRNA

snRNA: small nuclear RNA

CARMN

CARMN: Cardiac Mesoderm Enhancer-Associated lncRNA

Visceral smooth muscle cells (SMCs) are an integral component of the gastrointestinal (GI) tract that regulates GI motility.

CARMN is a major regulator of GI motility and viability through actions in visceral SMCs that serve to maintain contractile function and intercellular connectivity.

Loss of function of CARMN may contribute to human visceral myopathy [7]



08

Transcription factors analysis

Query on Biomart

Transcription factors analysis

66 TF

Dysregulated genes that can be
transcription factors.

ATF3

ATF3 is up-regulated in Crohn's disease patients.

Activating transcription factor 3 (ATF3) modulates intestinal epithelial cell (IEC) apoptosis under **stress** via regulating the p53 pathway [8].

Increased IEC apoptosis greatly destroys the intestinal mucosal integrity and contributes to the development of important clinical complications (e.g. colitis).

p53 protein: encoded by the TP53 gene, which is a tumor suppressor gene.

References

- [1] Chen, Y., Chen, L., Xing, C., Deng, G., Zeng, F., Xie, T., ... & Yang, H. (2020). The risk of rheumatoid arthritis among patients with inflammatory bowel disease: a systematic review and meta-analysis. *BMC gastroenterology*, 20, 1-11.
- [2] Halling, M. L., Kjeldsen, J., Knudsen, T., Nielsen, J., & Hansen, L. K. (2017). Patients with inflammatory bowel disease have increased risk of autoimmune and inflammatory diseases. *World journal of gastroenterology*, 23(33), 6137.
- [3] Aletaha, D., Epstein, A. J., Skup, M., Zueger, P., Garg, V., & Panaccione, R. (2019). Risk of developing additional immune-mediated manifestations: a retrospective matched cohort study. *Advances in therapy*, 36, 1672-1683.

References

- [4] Jung, H., Kim, J. S., Lee, K. H., Tizaoui, K., Terrazzino, S., Cargnini, S., ... & Kronbichler, A. (2021). Roles of microRNAs in inflammatory bowel disease. *International journal of biological sciences*, 17(8), 2112.
- [5] Zhao, Y., Ma, T., Chen, W., Chen, Y., Li, M., Ren, L., ... & Shi, R. (2016). MicroRNA-124 promotes intestinal inflammation by targeting aryl hydrocarbon receptor in Crohn's disease. *Journal of Crohn's and Colitis*, 10(6), 703-712.
- [6] Sun, T. Y., Li, Y. Q., Zhao, F. Q., Sun, H. M., Gao, Y., Wu, B., ... & Zhou, D. S. (2022). MiR-1-3p and MiR-124-3p synergistically damage the intestinal barrier in the ageing colon. *Journal of Crohn's and Colitis*, 16(4), 656-667.

References

- [7] He, X., Dong, K., Shen, J., Hu, G., Mintz, J. D., Atawia, R. T., ... & Zhou, J. (2023). The Long Noncoding RNA Cardiac Mesoderm Enhancer-Associated Noncoding RNA (Carmn) Is a Critical Regulator of Gastrointestinal Smooth Muscle Contractile Function and Motility. *Gastroenterology*.
- [8] Gu, L., Ge, Z., Wang, Y., Shen, M., & Zhao, P. (2018). Activating transcription factor 3 promotes intestinal epithelial cell apoptosis in Crohn's disease. *Pathology-Research and Practice*, 214(6), 862-870.

Thanks for the
attention!
