In silico Identification of Transcription based Biomarkers for Early Diagnosis of Inflammatory Bowel Disease and Predicting Progression to Colorectal Cancer

by

Farhat Khan

(1509034)

Thesis

Submitted in fulfilment of the requirements for the degree

Philosophiae Doctor

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor:

Prof Mandeep Kaur

March 2022

DECLARATION

I, **Farhat Khan (1509034),** declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

_____

(Signature of candidate)

7TH SEPTEMBER 2022

ABSTRACT

Inflammatory Bowel Disease (IBD) is a complex intestinal inflammation disorder that negatively impacts the quality life of patients. Diagnosis is confirmed through colonoscopy and in paediatrics this procedure is traumatic which highlights the need of less-invasive screening methods for clinical use. Circumstances is additionally entangled by the way that IBD patients have 20-fold increased risk of developing colorectal cancer (CRC). Due to advancements in computer technologies and development of new methods, bioinformatics has become an essential component of biology research. We hypothesized that utilizing bioinformatics methods to identify transcription-based biomarkers can be used for early diagnosis and screening of IBD, and for predicting the danger of its development to CRC in these patients. Transcription factors (TFs) are regulatory proteins that bind primarily to the promoter region of the target genes and control some of the key fundamental processes in a cell. We aim to identify key transcriptional regulators of IBD, and CRC associated genes by generating and understanding the transcription regulatory networks through our proposed methodology. To our knowledge there is currently no IBD genes database created so far. We first aimed at developing manually curated database of 289 IBD genes that are experimentally validated. Furthermore, to achieve other aims, we used various computational tools. To identify TFs and mapping of transcription factor binding sites (TFBSs) to mammalian matrices were achieved by OPOSSUM and JASPAR respectively. Using Cytoscape, we constructed largest transcriptional regulatory networks of top ranked TFs that were mapped to CRC genes. The pathway analysis was done by KEGG, and gene ontology was conducted using DAVID. Majority of the pathways identified were involved in the processes, known to play potential role in IBD. The top upregulated pathways in IBD identified were cytokine, immune response, autophagy, WNT signalling, transmembrane signalling. Further testing of expression of these key regulators in using publicly available published data were done using ONCOMINE database. We identified serological biomarkers of CRC and combined them with transcription-based biomarkers of IBD (CEA + TIMP1 + CA724 + RUNX1) to predict early diagnosis of IBD progressing to CRC. This project will form the basis to develop a kit in future for use in pathology laboratories and clinical settings.

DEDICATION

This thesis is dedicated to my:

Mother Late Sayeeda Khanam, who gave me this dream.

Grandmother Late Sanjeeda Khanam, who helped me to safeguard it.

Finally, my beloved cat Late Mona (2012 - 2022), who accompanied me throughout my journey of PhD.

RESEARCH OUTPUT

1. Publication

Publication forming part of PhD Thesis

Khan, F., Radovanovic, A., Gojobori, T., and Kaur, M (2021). IBDDB: a manually curated and text mining enhanced database of genes involved in inflammatory bowel disease. Database, Volume 2021, 2021, baab022, https://doi.org/10.1093/database/baab022

2. Poster Presentation

Farhat Khan. Understanding the molecular basis of Inflammatory Bowel Disease using computational methods. MBRT (Molecular Biosciences Research Thrust) PG Research Day. WITS Adler Museum, Faculty of Health Sciences Building, University of Witwatersrand. **28th November 2019**

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor Prof Mandeep Kaur for the continuous support of my Ph.D. study and related research, for her patience, motivation, and immense knowledge. Her expertise was invaluable in formulating the research questions and methodology. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Her guidance helped me throughout the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D. study.

My sincere thanks also go to Prof. Yasien Sayed, Dr. Vanessa Meyer, and Dr. Angela Botes for their guidance.

I would like to acknowledge my colleagues and friends from my laboratory for the stimulating discussions, for the help and support and for all the fun we have had in the last four years.

In addition, I would like to thank my family especially, my husband Mr. Meraj H Khan and his brother Mr. Sartaj H Khan and my brother Kamran Khan, for their wise counsel and sympathetic ear. Thank you for being there for me always.

TABLE OF CONTENT

DECLARATION ii
ABSTRACT iii
DEDICATION iv
RESEARCH OUTPUT v
ACKNOWLEDGEMENTS vi
LIST OF FIGURES ix
LIST OF TABLES x
LIST OF ABBREVIATIONS xi
1 CHAPTER ONE - INTRODUCTION 1
1.1 Hypotheses 1
1.2 Research Aims 2
1.3 Research Objectives 2
2 CHAPTER TWO - Literature Review 3
2.1 Inflammatory bowel disease 3
2.2 Aetiology of IBD 4
2.2.1 Environmental factors 4
2.2.2 Epidemiology 7
2.2.3 Genetic factors 9
2.2.4 Human gut microbiome in IBD 11
2.3 Molecular mechanism of IBD 12
2.3.1 Immunological factors 12
2.4 Biomarkers 15
2.4.1 Classification of Biomarkers 16
2.4.2 Biomarkers in IBD: Clinical Utility 19
2.5 Biomarkers of IBD 19
2.5.1 Serological Biomarkers 19
2.5.2 Faecal Biomarkers 21
2.6 Colorectal cancer 22
2.6.1 Epidemiology of CRC 22
2.6.2 Pathology of CRC 22
2.6.3 Aetiology of CRC 26
2.7 Molecular mechanisms of CRC 27
2.7.1 Pathogenesis in CRC 27
2.8 Biomarkers of CRC 31
2.9 IBD associated CRC 34
2.10 Databases and their importance 34
2.10.1 Types of databases 36
2.10.2 Existing CRC genes and TFs database 36
2.10.3 TFs Database 37
3 CHAPTER THREE – METHODOLOGY 38
3.1 Research design 38
3.2 Database Creation 39
3.2.1 Text mining 39
3.2.2 Curation of IBD genes and data mining 39
3.3 Identification of the biomarkers of IBD genes and CRC genes, and their *in-silico* validation 41
3.3.1 Promoter sets 43
3.3.2 Transcription Factor Binding Sites (TFBSs) 43
3.3.3 Mapping of TFBSs 44
3.3.4 Ranking of TFs 44

3.3.5 Network construction 45
3.3.6 Functional annotation 45
3.3.7 *In silico* validation of transcription factors 45
3.3.8 *In silico* validation of biomarkers 46
3.3.9 Gene expression profiling and Survival analysis 46
3.3.10 Statistical evaluation of biomarkers 46
4..... CHAPTER FOUR – RESULTS 56
4.1 AIM 1 56
4.1.1 Creation of a curated database of IBD genes 56
4.2 AIM 2 65
4.2.1 Identification of Biomarkers of IBD Genes 65
4.3 AIM 3 88
4.3.1 To identify biomarkers which can serve as potential diagnostic measure to predict early risk of CRC in IBD patients. 88
5..... CHAPTER FIVE – DISCUSSION 103
5.1.1 Creation of manually curated IBD genes database 104
5.1.2 Identification of Biomarkers of IBD and their validation 105
5.1.3 To identify biomarkers which can serve as potential diagnostic measure to predict early risk of CRC in IBD patients. 109
5.1.4 Limitations of this study 111
5.1.5 Strengths of this study 111
6..... CHAPTER SIX- CONCLUSION AND FUTURE WORK 113
7..... REFERENCES 114
8..... APPENDIX 151
8.1 APPENDIX A 151
8.2 APPENDIX B 155
8.2.1 Scripting Language R 155
8.2.2 Scripting Language Python 157
8.3 APPENDIX C 160
8.3.1 IBDDB: User manual guide of IBD database 160

LIST OF FIGURES

Figure 2. 1 Anatomy of the human intestine. 4
Figure 2. 2 Global Incidence of CD and UC was reported in the developed nations. 8
Figure 2. 3 Pathophysiology of IBD represents the immune response and inflammatory pathway of IBD. 14
Figure 2. 4 Molecular mechanism involved in the pathogenesis of IBD. 15
Figure 2. 5 The chart represents different categories involved in application of biomarkers (Pierce et al., 2012). 17
Figure 2. 6 This figure demonstrate the multiple epigenetic and genetic pathways that are involved in the development of CRC 28
Figure 2. 7 This figure demonstrates the DBMS design and features. 35

Figure 3. 1 Schematic representation of the methodology used in the research plan of this study. 38
Figure 3. 2 ROC curve illustrates the sensitivity and 1 – specificity plotted on a graph across varied cut- 53

Figure 4. 1 This figure illustrates the different paradigm available under Explore tab of the IBD database (IBDDB) along with visualize and user guide. 57
Figure 4. 2 This figure illustrates the tables containing various column names under Explore tab 'IBD Database'. 58
Figure 4. 3 This figure illustrates the Knowledgebase under Explore tab. 59
Figure 4. 4 This figure illustrated the Biomedical entities under Explore tab. 60
Figure 4. 5 The figure illustrates the text mining approach in IBD database. 61
Figure 4. 6 The hypothesis generated by the hypothesis explorer showing that PG may be linked to lipid metabolism and IBD (Khan et al., 2021). 62
Figure 4. 7 This figure demonstrates the step-by-step network construction to identify the relationship between disease PG and its connection with the pathway lipid metabolism and IBD. 63
Figure 4. 8 This figure illustrates the exportable data from IBD database in various formats. 64
Figure 4. 9 Figure illustrates the TSS location of core promoters. 66
Figure 4. 10 This figure depicts the TFBS of RUNX1 mapped to the JASPAR matrix ID MA0002.1 mammalian matrix model like PFM. 69
Figure 4. 11 This is the matrix profile of TF RUNX1 from the core redundant vertebrate collection, displayed as PWMs scores in the RUNX1 70
Figure 4. 12 This is the outcome from PWMScan. 70
Figure 4. 13 Reconstruction of TRN model of IBD genes. 72
Figure 4. 14 This figure illustrates the GO annotation and KEGG Pathway analysis using R package: ClusterProfiler. 74
Figure 4. 15 Box-whisker plot shows relative transcription of biomarkers in sub-groups of COAD-READ patients, stratified based on gene expression level in CRC tumor vs Normal 81
Figure 4. 16 ROC curve analysis of post-test probability for positive and negative test in individual IBD biomarkers. 83
Figure 4. 17 High expression of combination biomarkers was associated with **(A)** worse OS, hazard ratio (HR)=1.7, log rank P= 0.087 and **(B)** worse DFS, HR= 1.4, log rank P=0.32. 86
Figure 4. 18 Heat map illustrates the expression values of 4 biomarkers in the patients' samples in IBD. 95
Figure 4. 19 Box-whisker plot shows relative expression of biomarkers in sub-groups of COAD-READ patients, stratified based on gene expression level in CRC tumor vs Normal 96
Figure 4. 20 ROC curve analysis of post-test probability for positive and negative test in individual IBD to CRC biomarkers. 98
Figure 4. 21 The combination of 5 biomarkers' expression was associated with **(A)** worse OS, hazard ratio (HR) = 1.3, log rank P = 0.2 and **(B)** worse DFS, HR = 1.6, log rank P = 0.028. 102
Figure 4. 22 The combination of 4 biomarkers' expression was associated with **(A)** worse OS, hazard ratio (HR) = 1.8, log rank P = 0.076 and **(B)** worse DFS, HR = 2.1, log rank P = 0.014. 102

LIST OF TABLES

Table 2. 1 AJCC and TNM staging classification of CRC (Kheirelseid et al., 2013). 26

Table 3. 1 Population of database columns, along with the references and URL. 40
Table 3. 2 List of bioinformatics tools used in the analysis, along with the references and URL. 42
Table 3. 3 This table illustrates the input values for the diagnostic test evaluation in 2x2 contingency matrix table. 51
Table 3. 4 Rule of Thumb: Correspondence between Accuracy, LR+ and LR- and Area under the Receiver Operating Characteristics Curve ($AUCROC$) and the Diagnostic Value of a Biomarker. 54

Table 4. 1 An excerpt from the datasets of IBD genes downloaded from IBDDB database in Excel format. 66
Table 4. 2 This table shows the TFs regulating IBD genes by number of gene hits and TFBS hits based on Fisher score as the outcome by OPOSSUM 3. 67
Table 4. 3 This table shows the analysis of IBD genes outcome through OPOSSUM 3. 68
Table 4. 4 This representative table shows ranking based on Fisher exact test score of some of the TFs out of the complete set of 289 IBD genes. 71
Table 4. 5 This table represents the diagnostic values of biomarkers for clinical utility based on Sensitivity, Specificity, PPV, NPV, Pre-test odds, Post-test odds and LR of IBD biomarkers at 95% CI. 77
Table 4. 6 This table represents the diagnostic capabilities of biomarkers based on Accuracy, LR and AUCROC in predicting risk of IBD at 95% CI. 78
Table 4. 7 Expression of IBD biomarkers in patients' samples. 80
Table 4. 8 Evaluation and combinations of biomarkers of IBD. 85
Table 4. 9 This table represents the diagnostic values of biomarkers for clinical utility based on Sensitivity, Specificity, PPV, NPV, Pre-test odds, Post-test odds and LR of IBD to CRC biomarkers at 95% CI. 90
Table 4. 10 This table represents the diagnostic capabilities of biomarkers based on Accuracy, LRs and AUCROC in predicting risk of IBD to CRC progression at 95% CI. 91
Table 4. 11 Expression of IBD to CRC biomarkers in patients' samples. The expression of 5 biomarkers was compared in IBD and CRC patients' samples at the given concentration unit, predicting progression from IBD to CRC in the early stages of CRC. mRNA Expression was measured from blood serum and platelets using Log 2-Fold change. 94
Table 4. 12 Evaluation and combinations of biomarkers from IBD-CRC. 100

LIST OF ABBREVIATIONS

APC	Adenomatous polyposis coli
API	Application programming Interface
ASCA	Anti-Saccharomyces cerevisiae antibodies
AUC	Area under curve
AJCC	The American Joint committee on Cancer
BRAF	B-rapidly accelerated fibrosarcoma
CD	Crohn's disease
CRC	Colorectal Cancer
CEA	Carcinoembryonic antigen
CA724	Cancer antigen 724
CA199	Cancer antigen 199
CRP	C-reactive protein
CBRC	Computational Bioscience Research Center
CIN	Chromosomal Instability
CIMP	CpG island methylator phenotype
CI	Confidence Interval
COAD	Colon Adenocarcinoma
CSV	Comma separated values
DNA	Deoxyribonucleic Acid
DEGs	Differentially expressed genes
DBMS	A database management system
DES	Dragon Knowledge Exploration System
DNMTs	DNA Methyltransferases
DAVID	Database for Annotation, Visualization, and Integrated Discovery
ESR	Erythrocyte sedimentation rate
EPD	Eukaryotic Promoter Database
ELF5	E74 like ETS transcription factor 5
FEV	FEV transcription factor, ETS family member
FC	Fold-Change
FP	False positive
FPR	False positive rate
FN	False negative
F-Score	Fisher's exact test score
GI tract	Gastrointestinal tract
GWAS	Genome wide association Study
GTEx	Genotype Tissue Expression
GO	Gene Ontology
HTS	High-Throughput Sequencing technology
HD	High-Definition
HGNC	HUGO Gene Nomenclature Committee
IBD	Inflammatory Bowel Disease
IBDDB	Inflammatory Bowel Disease Database
iFOBT	Immunochemical Faecal Occult Blood Test

IHC	Immunohistochemistry
IECs	Intestinal epithelial cells
KRAS	Ki-ras2 kirsten rat sarcoma
Kb	Kilobytes
Kbp	Kilo base pair
kDa	Kilodaltons (molecular weights)
IU/mL	International units per milliliter
JSON	JavaScript Object Notation
KAUST	King Abdullah University of Science and Technology
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KSA	Kingdom of Saudi Arabia
KB	Knowledgebase
LR	Likelihood ratio
LOH	Loss of heterozygosity
mg/L	Milligram per liter
MESH	Medical Subject Headings
MMP	Matrix Metalloproteinases
mRNA	Messenger Ribonucleic Acid
MCC	Mucinous colorectal carcinoma
MC	medullary carcinoma
MSI	Microsatellite instability
nt	Nucleotides
ng/mL	Nanogram per milliliter
NFKB1	NF-kappaB related factors
NSAIDs	Non-steroidal anti-inflammatory drugs
NK cells	Natural killer cells
NUM	Number of sample's populations
NPV	Negative predictive value
OR	Odd Ratio
ORI	Over-Representation Index
PPV	Positive predictive value
Post-test Probability	Posterior test probability
Pre-test probability	Prior test probability
pANCA	Perinuclear Antineutrophil Cytoplasmic Antibodies
PMNs	Polymorphonuclear leukocytes
PIK3CA	Phosphatidylinositol-3-kinase
PMID	PubMed identifier
PMCID	PubMed Central identifier
PNG	Portable Network Graphics format
PG	Pyoderma Gangrenosum
P53	Tumor protein TP53
SNPs	Single nucleotide polymorphisms
PSSM	Position-specific scoring matrix
PPI	Protein-Protein Interaction
PHP	Hypertext Preprocessor
PWMs	Position weight matrix
PFMs	Position frequency matrices

qPCR	Quantitative polymerase chain reaction
RBC	Red blood cell
RUNX1	Runt-Related Transcription Factor 1
RELA	RELA proto-oncogene, NF-kB subunit
RNA-Seq	RNA sequencing
ROC	Receiver operating characteristic curve
RNA	Ribonucleic acid
READ	Rectum Adenocarcinoma
RT-PCR	Real time reverse transcription polymerase chain reaction
SMAD4	Mothers against decapentaplegic homolog 4
SPIB	Spi -B transcription factor
SPI1	Spi-1 proto-oncogene
STARD	Standards for Reporting of Diagnostic Accuracy
SQLite	Structured Query Language (Light weight)
TSS	Transcription Start Site
TIMP1	Tissue Inhibitor Metalloproteinase
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TRN	Transcriptional Regulatory Networks
TPM+1	Transcripts Per Million
TPR	True positive rate
TCGA	The Cancer Genome Atlas
TNM	Tumor, Nodes, Metastases
UICC	Union for International Cancer Control
URL	Uniform Resource Locator
WHO	World Health Organization
Wnt Signaling path-way	Wingless related integration site
XML	extensible markup language

1 CHAPTER ONE - INTRODUCTION

Inflammatory bowel disease (IBD) is a complicated disorder of gastrointestinal (GI) tract. Ulcerative colitis (UC) and Crohn's disease (CD) are two well-known phenotypes of IBD (Hampe et al., 1999). The characteristics and various sites of the inflammatory lesions in the GI tract differ between CD and UC, although both exhibit symptoms such as stomachache, severe pain, loss of weight, gastrointestinal bleeding, and anemic. In advanced cases, surgery can remove damaged parts of intestines, but often requires colostomy. IBD affected about 1.29 million individuals residing in the United States alone in 2020 (Ungaro et al., 2022) and about 1.3 million people approximately in Europe (Zhao et al., 2021) with increasing incidence rate worldwide (Alatab et al., 2020). The frequency of occurrence rate of IBD are highest in developed countries and there is a scarcity of data on the frequency and severity rate of IBD in third world nations (Kaplan and Windsor, 2021). In South Africa, a few reports have emerged reporting on CD (Basson et al., 2014). In a study conducted on IBD patients during COVID-19 pandemic stating the difficulties in consulting with IBD patients from very little to none face-to-face consultation, the only way to consult patients by gastroenterologists was telephone via video calling and found only 2.8% patients responded to call in South Africa (Lees et al., 2020). More than 200 susceptibilities loci of IBD have been discovered in the studies carried out during the last decade (Afrasiabi et al., 2020). The study has confirmed the overall prevalence of IBD as an early-onset colorectal cancer (CRC) between period 2013-2018 to be 18.9/100 000 per population (Glover et al., 2019). In a recent study it was predicted that UC individuals were anticipated to be at a high chance of developing CRC in the course of time (Olén et al., 2020). It is stated that IBD risk is influenced by genetic factors, including *IBD5* locus (human 5q31), that harbours that *IRF1* gene, can be a cause for IBD-associated CRC (Jeyakumar et al., 2019). Despite the accumulating evidence of genetic involvement in IBD, we are still far from diagnosing and prognosing IBD in a minimally-invasive way. There is an urgent need for developing non-invasive methods for timely diagnosis of IBD and we propose to develop a database of transcription factors (TFs) as biomarkers for diagnosis and possibly prognosis of IBD.

1.1 Hypotheses

Currently, bioinformatics development is critical in South Africa in order to contribute to the changing environment of biological research. The proposed study will concentrate on the development and implementation of computational approaches to be evaluated for the diagnosis/prognosis

of IBD. The project is important in three ways: a) it will be the first computational analysis of transcription regulation of genes implicated in IBD; b) it will discover biomarkers for IBD diagnosis and maybe screening; and c) it will identify biomarkers to predict and monitor IBD progression to CRC.

In recent years, the number of studies on IBD genetics has increased (Luo et al., 2018), and several genes have been implicated in the pathogenesis of IBD, implying that IBD is a polygenic disease (Ramos and Papadakis, 2019). In this thesis, I have attempted to computationally find transcription factors (TFs) that regulate these genes, as well as transcriptional regulatory networks (TRNs) within which these genes are expected to act in IBD, which could lead to the discovery of diagnostic biomarkers or potentially therapeutic targets for this disease.

1.2 Research Aims

1. To create a curated database of IBD genes.
2. To identify the transcription factor based biomarkers of IBD.
3. To identify biomarkers which can serve as potential diagnostic measure to predict early risk of CRC in IBD patients.

1.3 Research Objectives

1. To develop a manually curated database of IBD genes.
2. To develop and refine existing methods to identify transcription regulators of disease-associated genes and their utilization as biomarkers.
3. Computationally identify biomarkers for early and less invasive diagnosis and screening of IBD.
4. Generate information to understand the molecular mechanisms underlying the aetiology of IBD.
5. Delineating the TRNs common between IBD and CRC.
6. *In silico* validation of predicted biomarkers using publicly available data for IBD and CRC patient samples.

2 CHAPTER TWO - Literature Review

2.1 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a gastrointestinal (GI) tract disease that is a chronic ailment (Rothfuss et al., 2006). It is an idiopathic disorder with two subtypes ulcerative colitis (UC) and Crohn's disease (CD). These two subtypes vary in the location of the organs (Matsuoka et al., 2018). UC is a disease limited to the colon or small intestine in the mucosal layer, while CD is a perianal disease that can impact any part of GI tract, from mouth to rectum (Alzahrani et al., 2020). Symptoms in IBD patients comprised of rectal bleeding, severe cramping, weight loss, ulcers, frequent diarrhoea, severe abdominal pain, depression and anxiety (Marrie et al., 2021). Loss of quality of life in IBD patients leads to severe depression and anxiety, often when left untreated (Sweeney et al., 2018). In clinical aspects, UC and CD are associated with intestinal and extraintestinal manifestations (Marotto et al., 2020). UC is limited to the colonic mucosa, mixed with blood and stool, but extends continuously from the rectum (De Maio et al., 2022). On the contrary, CD is often delayed in diagnosis, as its symptoms rely on the location of the disease GI tract extent and with severe pain level (Hartwig et al., 2021).

In IBD patients the diagnosis is conducted by Colonoscopy, which is a procedure performed with a long flexible scope, attached is a high-definition camera, an irrigator, and a light (Fig 2.1 A & B). It investigates inflammation, chronic diarrhoea, in the intestines of IBD patients, while polyps in CRC patients (Borsotti et al., 2021). In paediatric patients, the side effects of colonoscopy are much higher than adults, as they have to undergo adverse effects of anaesthesia during the procedure, followed by infection and bleeding as the potential risk of colonoscopy (Kucharzik et al., 2020). Even though colonoscopy is a difficult procedure but prior to colonoscopy preparation like bowel preparation for paediatric patients is also difficult, as children have to consume 4 litre of polyethylene glycol to cleanse their bowel system a day before the procedure (Kesavelu Sr, 2020). In recent years the incidences of IBD has increased in young children at aged 4 years and IBD was considered to be the second leading diagnosis in infants and young children who underwent colonoscopy (Nambu et al., 2019).

A B

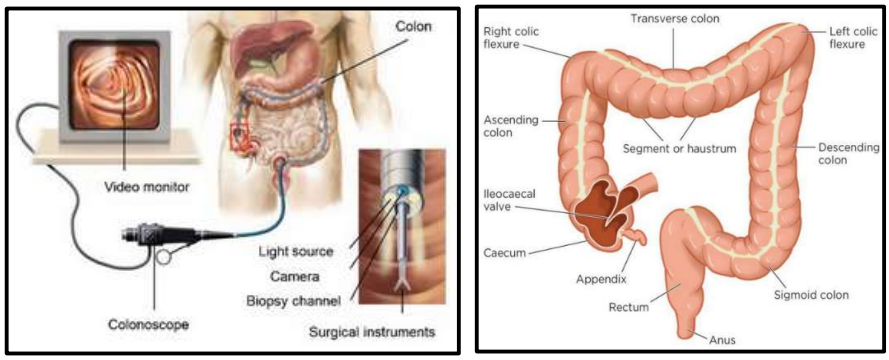


Figure 2. 1 Anatomy of the human intestine. **A)** Represents the view of colonoscopy procedure inside the large intestine. Colonoscope wire has devices attached to its tip with light, HD camera and irrigator (Ahmed et al., 2019). **B)** Represents the different sections of the colon (Nigam et al., 2019)

Although IBD is a disease with no established origin or cause, many studies had suggested IBD's aetiology may be associated with microbial, immunologic, nutritional, environmental and genetic correlations (Ahlawat et al., 2021). Extraintestinal manifestation complications and other autoimmune diseases are connected to systemic diseases of IBD (Garber and Regueiro, 2019). It is suggested that immune response disorder is responsible for extraintestinal manifestation in IBD patients, which is further classified into two forms-Reactive manifestations that reflected pathogenic mechanisms common among IBD skin lesions like pyoderma, arthritis (Greuter et al., 2021) and other is autoimmune diseases (Veloso et al., 1996).

2.2 Aetiology of IBD

2.2.1 Environmental factors

It is true, the vast knowledge provided by the Genome-wide association studies (GWAS) studies dissecting the genetic components and linkage mapping is involved in developing and progressing to IBD. The high incidence of IBD worldwide, suggests that other factors like environmental factors are involved as well. Many environmental factors have been investigated in IBD, but the foremost dominant factors are discussed here such as smoking, appendectomy, dietary factors.

2.2.1.1 Smoking

One of the major environmental factors contributing to IBD is smoking. Smoking causes the pernicious influence on the clinical assessment of the disease. Mucosal damage, changes in gut irrigation, and impaired mucosal immune response are among the mechanisms proposed to explain smoking's role in

IBD (Berkowitz et al., 2018). On the contrary smoking is also associated with UC in non-smokers or former smokers (Salih et al., 2018). The meta-analysis study determined the smoking status of current smokers, former smokers, and non-smokers by comparing smokers to non-smokers and evaluating odd ratios of less than 1.0, indicating a protective association between current smoking and UC, implying more or less 40% lower risk of UC diagnosis (Mahid et al., 2006). Former smokers, in contrast to present smokers, generated an additional risk of UC with odd ratios of 1.8 (Mahid et al., 2006). A study conducted on environmental and demographic differences of late-onset of CD and discovered that ex-smokers and patients with anaemia are at a higher risk of developing CD (Moon et al., 2020). No significant findings were generated in former and never smoker in CD patients with increased risk of two-fold (Mahid et al., 2006).

It is reported that in cigarette smoking, the culprit nicotine as a metabolite is doubtlessly the fundamental driver for the negative consequences causing IBD inflammation (Langsted and Nordestgaard, 2019), even though reviews analysed that either effects of nicotine and tobacco alone neglected to reproduce a similar extent of connection seen by smoking, proposing that different parts of tobacco smoke or chewing of tobacco may be significant (Zou et al., 2021). Many studies have demonstrated the negative influence of passive smoking in children with decreased risk of UC (Sandier et al., 1992; To et al., 2016) and an increased risk of developing CD in children (Braun et al., 2020). Elevated levels of IL-8 were discovered in the colonic mucosa of smoking patients, indicating changes in the immune system because of altered cytokine levels (Allais et al., 2020). However, smokers with CD had lower amounts of IL-8, whereas UC patients had lower levels of both IL-1 and IL-8, suggesting that these lower levels may protect patients (Sher et al., 1999).

2.2.1.2 Appendectomy

One of the other environmental factors that cause IBD is appendectomy, which is a surgical procedure to remove the appendix (Loureiro and Barbosa, 2019). Human appendix is considered as a vestigial organ (Arjomand Fard et al., 2022) in human body because its functions were considered negligible and was thought to be first used by our herbivore ancestors to digest cellulose (Heuthorst et al., 2021). The appendix is a portion of the gastrointestinal system that evolved from the caecum (Arnold et al., 2019) and is related with the lymphoid tissue system, as well as Peyer's patches and tonsils (Koutroubakis et al., 2002). Appendicitis is the inflamed appendix filled with pus and they lack the clear display of signs and symptoms (Butler, 1981) but some of the symptoms are anorexia, urinary frequency, diarrhoea, vomiting and generalize malaise (Jones et al., 2021). The major sign of acute appendicitis is the severe pain in the right side of lower quadrant, abdominal stiffness, resettling of

pain from periumbilical region to the right side of lower quadrant (Meyer et al., 2021; Paulson et al., 2003).

It is important to understand the appendicitis and appendectomy in the light of IBD, and how it is related. Considering the risk in patients developing UC prior appendectomy or after is found to be influenced by age. In many studies it has been reported that there is a low possibility of occurring UC in patients who underwent appendectomy (Rolli-Derkinderen et al., 2020). In a population cohort study of 210,000 patients carried out by Andersson et al, it was investigated that patients underwent appendectomy for appendicitis surgery and mesenteric lymphadenitis before 50 years of age had a much-decreased rate of UC than those with nonspecific abdominal pain without appendectomy (Andersson et al., 2001). The drawback discovered in the same cohort study about an increased risk of CD is only apparent in people who have had an appendectomy at the age of more or less 20 years, nevertheless likelihood of CD was much lowered as the patients were operated earlier than or at the age of ten years (Andersson et al., 2003). On the contrary, some studies investigated the risk involving the occurrence of UC and CD in patients with appendectomy (Chung et al., 2021).

2.2.1.3 Dietary factors

The new data suggests that increased incidence of IBD is arising in the western countries and approximately 0.2% European population suffers from IBD (Zhao et al., 2021). It has been found that adopting the western lifestyle of fast-food, high fatty acid diets, sugar, alcohol, dairy consumption on daily basis can also be the contributing factor in increased risk of IBD among the current food patten in US adults (Han et al., 2020) and in Asian population (Cui and Yuan, 2018). A high intake of dietary fibers like fruits and vegetables reported a low risk of CD and UC in a study (Milajerdi et al., 2021). Meanwhile, in UC patients, high consumption of fats, polyunsaturated fats, omega-6 fatty acids and meats plays the key contributing factor, alongside saturated fats together contributed for higher risk of IBD, particularly in UC populations (Kobayashi et al., 2021). Obesity, on the other hand, has been linked to the aetiology of IBD by inducing mucosal barrier dysfunction, bacterial translocation, dysbacteriosis, and adipocyte activation, with estimates ranging from 20 to 50% of IBD patients being obese (Singh et al., 2017).

Several studies had demonstrated that vitamin D deficiency causes IBD in newly diagnosed patients and that it play a significant role in the innate immune responses (White, 2018) of inflammation that can cause risk of developing 38% in CD and possibly 31% in UC patients (Parizadeh et al., 2019). The

study reported that specific carbohydrate diet is a popular diet for individuals with IBD, and an unsubstantiated data showed it may be useful treating CD and UC. Modification of the bacterial microbiome is thought to be the key mechanism via which specific carbohydrate diet works, as this diet IBD patients are found to have different microbiome in comparison to the control IBD patients (Kakodkar et al., 2013).

2.2.2 Epidemiology

IBD disorder (CD and UC), has been viewed as an issue in developed metropolitan social communities and credited to a great extent to developed countries' lifestyle (Forbes et al., 2017). The highest recorded prevalence was reported in European countries (UC 505/100,000 in Norway; CD 322/100 000 in Germany) and in North America (UC 286/100 000 in USA) (Ng et al., 2017) and Canada 54.6/100 000 (Benchimol et al., 2019). The incidence of IBD is increasing and has turned into a global burden, however Asian countries have recently reported greater incidence rates (Wei et al., 2021). Earlier reports from West Germany rural county Tubigen (FRG) (from 1970 - 1984) suggested high incidences and prevalence cases in CD (4 new cases per 100000/year) in the year 1970 and prevalence of (80 per 100000) respectively followed by high incidence cases of UC (30 per 100000) both in the least populated regions of the county (Daiss et al., 1989). In North America, Oceania, and in many European countries, the prevalence of IBD approached 0.3% (Ng et al., 2017). On the contrary, the causes that contribute to the emergence of IBD are still unknown, these related disorders have long been seen as a problem in developed countries, with the Western way of life playing a significant role in their aetiology (Kaplan and Windsor, 2021). Considering the above facts IBD is becoming more common in underdeveloped countries (Zuo et al., 2018), and it is increasingly being regarded as a global disease according to the incidence rates are mentioned in (Fig 2.2) (M'koma, 2013). Extraintestinal manifestations incidences in IBD patients have ranged from 10% to 50% recently worldwide (Adam et al., 2020).

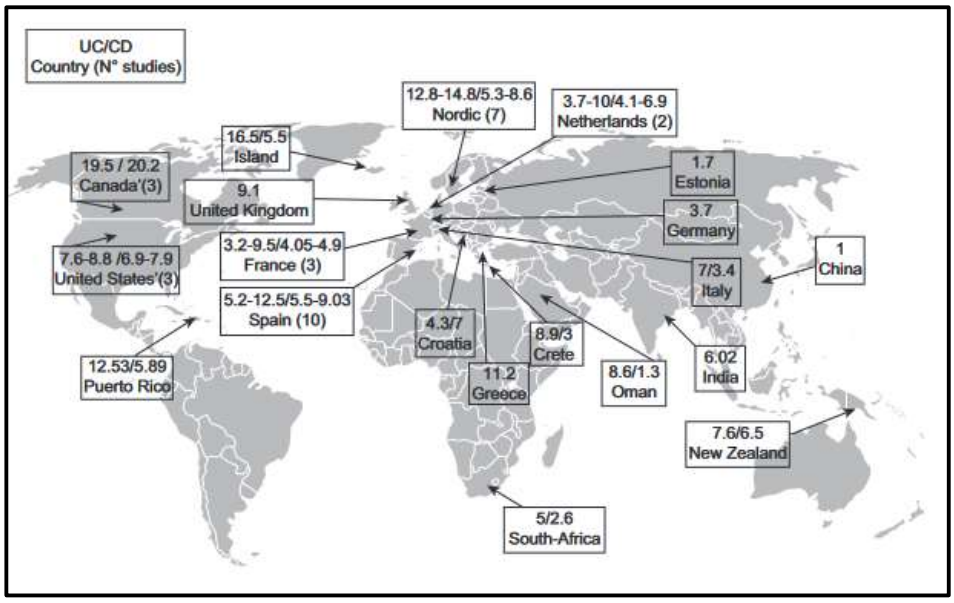


Figure 2. 2 Global Incidence of CD and UC was reported in the developed nations (M'koma, 2013).

It is observed that incidence and prevalence in developing countries are very limited. According to the epidemiological studies on paediatric-onset IBD (PIBD), there is a rapid rise in the incidence and clinical characteristics of PIBD in a South-East Asian cohort in Singapore (Ong et al., 2018). It has been reported that the incidence rate of UC increased from 2015-2015 by 70.7% in urban population of Iran (Sharara et al., 2018). The incidence rate of CD ranges between 1.46 and 1.66 per 100 000 people and 2.33 per 100 000 people for UC in middle eastern countries like Kingdom of Saudi Arabia (KSA) due to westernization lifestyles and smoking (Mosli et al., 2021). Whereas, the incidence of IBD is also reported to be increasing in the Turkish population from 0.99/100000 and 0.45/100 000 in UC and CD (2004-2005) to 4.87/100 000 and 2.09/100 000 in UC and CD patients (2011-2013) (Can et al., 2019).

2.2.2.1 The Epidemiology of IBD in Africa

A study documented the incidence of CD in the Western Cape, South Africa, between 2011 and 2013 based on racial categories signifying 18% white, 78% coloured, and 4% black population, of whom the coloured individuals were diagnosed with CD condition and exposure to second-hand cigarette smoking (Basson et al., 2014). Study was conducted on the incidence of patients with CD, in the Gastrointestinal Clinic from 1975-1980 among racial groups, were 72% white, 37% coloured and 1% black population. Concerning the aetiology of IBD not being completely understood, this published study contemplated the role of hygiene in the rural areas, investigating exposure to parasite helminth may protect against the development of both CD and UC in Western Cape (Chu et al., 2013). The rate of

incidence and the prevalence in IBD is gradually increasing in the Sub-Saharan African black population, but still, due to their traditional lifestyle and not much exposure to the environmental factors, UC prevalence is found rare in them (Agoda-Koussema et al., 2012; Segal et al., 1980). In a preliminary analysis of a large IBD registry of more than 3000 South African patients showed an exponential increase in IBD over the past 70 years in all ethnic group (Watermeyer et al., 2020).

It is reported in a latest study conducted in West Africa, Togo, the data obtained were from the diagnostic procedure of Colonoscopy, investigated lower gastrointestinal bleeding in patients with 31.1% haemorrhoids, 18.9% were diverticulosis, 8.3% CRC and 30.7% of cases by UC, mainly in patients of age 47 years and were male predominance (Bagny et al., 2021). In one study from Algeria, from Constantine hospital, annual incidence cases of CD and UC were 5.87/100000 and 3.29/100000 respectively were reported and stated that incidence of UC has increased during the last five years from 2.76/100000 in 2003 to 5.12/100000 in 2007 (Hammada et al., 2011). The relative frequency of UC diagnosis when compared with Crohn's in Johannesburg was proven by reports from Cairo, where data from a cohort of IBD patients diagnosed between 1995 and 2009 reported that the ratio of UC to Crohn's was 6:1 (Hodges and Kelly, 2020). According to the Global burden of disease, the total number of deaths caused by IBD in North Africa and the Middle East remained the super-region with the lowest number of deaths, with 872 (781-999) in 2017, considering the number of deaths in females reported was 372 (317-432) and in males 500 (436-605) (Alatab et al., 2020).

It has been determined that the rising incidences of IBD in developed nations is due to a change in the western lifestyle of living standard. Environmental factors like consuming processed food with less dietary nutrients like fibers and smoking cigarettes contributed to the health degradation by compromising the immune system. Gradually, the incidence and prevalence are seen to be rising in Asian, African, and middle eastern countries as well.

2.2.3 Genetic factors

GWAS seek to identify genotype-phenotype associations by looking for disparities in the allele frequency of genetic variants among individuals who are genetically similar but differ phenotypically (Uffelmann et al., 2021). In the past few years, large scale GWAS have made the most important contributions in dissecting the genetic component involved in the development and progression of IBD (Verstockt et al., 2018). To date, GWAS has identified >200 risk loci for disease IBD (Jostins et al., 2012; Momozawa et al., 2018). It is considered that CD and UC is a complex polygenic disease with several genes and influenced with environmental factors, contributed to developing IBD (Yamamoto-

Furusho, 2007). Since epidemiology studies of IBD indicated the strong influence of environmental factors, their influence is confirmed by concordance in the study of identical twins with 50% for CD and 10% for UC (Tysk et al., 1988). This Swedish study from 1988 demonstrated that monozygotic twins had a proband concordance rate of 58.3 % for CD and 6.3 % for UC (Halfvarson et al., 2003; Tysk et al., 1988). These first-degree relatives' twin studies with a high proband concordance rate suggested that genetics played an important role in the pathophysiology of IBD (Halfvarson, 2019).

Linkage mapping is a proven powerful tool to map genetic location containing IBD susceptibility genes (Brant et al., 2017). Using linkage mapping analysis, the susceptibility locus (*IBD1*) Nucleotide-binding oligomerization domain-containing protein 2*, NOD2* on chromosome 16 was mapped to CD in this study (Hugot et al., 2001; Mirkov et al., 2017). Mutation in this gene is associated with CD (Kaczmarek-Ryś et al., 2021). Inflammatory Bowel Disease1, *IBD1* (Caspase Recruitment Domain 15, *CARD15* or *NOD2*) is an innate immune receptor responsible for recognizing bacterial molecules within the intestine, is associated with CD pathogenesis (Schaefer et al., 2018). This marked the history not only for establishing aetiology, pathogenicity and immune response disease but also for depicting the positional identification of diseased genes for complex disease (Brant and Shugart, 2004). In 1999, Inflammatory Bowel Disease5 (*IBD5*) gene on chromosomal segment 5q31 were discovered to be linked to CD for the first time (Ma et al., 1999). Statistical fine mapping of the *IBD5* gene revealed a 250-kb haplotype cytokine gene cluster linked to CD and its variations, the organic cation carriers Organic Cation Transporter1, *OCTN1* Solute Carrier family 22 member 4 (*SLC22A4*) and Organic Cation Transporter2, *OCTN2* Solute Carrier family 22 member 5 (*SLC22A5*) (Noble et al., 2005; Palmieri et al., 2006; Park and Jeen, 2019). There has been a conflicting report associating *IBD5* susceptibility loci variants *OCTN* gene associated with UC (León et al., 2018). Although hereditary factors have long been assumed to be a major hazard element for the development of IBD, new research shows that they are not, the contribution of a single variant is small, as the biggest genetic effects have been shown for interleukin-23 receptor (*IL23R)*, *NOD2*, and human leukocyte antigen (*HLA)*, with chances ranging from 1.4 to 3 (Annese, 2020).

2.2.3.1 High-throughput sequencing techniques

The completion of the human genome project, led scientists made considerable progress in comprehending and interpreting their own genetic information (Green et al., 2015; Sakaki, 2019). In the recent advances of technologies representing 'omics' like genomics, transcriptomics and proteomics using microarray, next-generation sequencing (NGS) (Schramm et al., 2013), as well as enhanced bioinformatics and statistical techniques, have sparked the interest in finding biomarkers for complex disorders

such neurological diseases and cancer (Yang et al., 2020). Simultaneously, technologies with ever-improving transcript identification ability, particularly at the higher molecular basis, drastically boost efficiency. Transcriptomics, in general, makes use of high-throughput sequencing (HTS) technology to examine at the heterogeneity of a cell's transcriptome at the integral level over time and space. The development of gene expression microarray technology, for example, has aided in transcriptomic research at HTS. (Schena et al., 1995) and RNA sequencing (RNA-Seq) technology (Wang et al., 2009). HTS like microarray and RNA-Seq are explained below.

1. The first study on IBD using microarray technique was carried out in analysing gene expression from UC patients in comparison to control patients but not CD (Dieckgraefe et al., 2000). A nylon-based chip with only 588 genes and expressed sequence tags was used in the second published investigation including CD patients. It discovered 31 genes that were up regulated in CD compared to UC, including matrix metalloproteinase (MMP) genes, as well as genes implicated in apoptosis and cell adhesion (Uthoff et al., 2001).
2. RNA-Seq technique has become an essential method for transcriptome-wide investigation of differential gene expression (DEGs) and mRNA splicing throughout the last decade. On the other hand, RNA-Seq has progressed in tandem with next-generation sequencing technology (Stark et al., 2019). Identification of long reads of sequence is now possible with the third-generation sequencing platform (Wang, 2021). Normal mucosa, non-inflamed CD mucosa, and inflamed CD mucosa were distinguished by genome-wide transcriptome differences using RNA-Seq. CXCL1 levels in the gut and blood significantly increased with CD activity implying that it might be employed as a possible CD biomarker (Hong et al., 2017). In another study, RNA-Seq technique revealed in the treatment of naïve UC in comparison to control. They provided evidence for a gender-dependent pathogenesis in UC which may be useful in personalised treatment strategies for UC patients (De Zoysa et al., 2017; Taman et al., 2017).

2.2.4 Human gut microbiome in IBD

Human gut microbiome comprised of vast variety of bacterial species, viruses, fungi, and phage's inhabiting the GI tract. In addition to the gut acquiring the largest microbiome compartment in the body, there are few other organs comprising the microbiome like oral cavity, skin, lungs, vagina and nose (Vangoitsenhoven and Cresci, 2020). These microbes exhibit mutual symbiotic and commensal relationship with the host and featured as the good bacteria by maintaining homeostasis (Lamichhane et al., 2018). Irrespective of the gut microbiota maintaining the homeostasis, immune system regulation and pathogen regulation, they may disrupt the homeostasis causing dysbiosis, which can lead to

inflammation (Øyri et al., 2015). The gut location that harbours large number of bacteria is colon and ileum and the four main microbial taxa colonizing these areas are Bacteroidetes, Actinobacteria, Firmicutes and Proteobacteria (Singhvi et al., 2020).

In a study previously reported the reduction in commensal bacteria phyla firmicutes and Bacteroidetes in the two sub types of IBD (Nkosi et al., 2022). Human gut is a home to vast microbial ecosystem. However, it has been found that gut microbial flora consists of 100 trillion microorganism population (Chi et al., 2021). In a study carried out in paediatric CD patients, it was reported that in comparison to controls, microbial diversity was lower in CD paediatric patients, however, in patients who did not attain long-term remission, it was considerably lower (Dunn et al., 2016). As it is stated that genetic predisposition played a role in the development of IBD (Kaur and Goggolidou, 2020). Nevertheless, the rapid increment in the number of people affected suggests that nutrition, as an environmental component, and the intestinal microbiota that it influenced are involved in the aetiology of IBD (Nishida et al., 2018). Non-bacterial organisms like viruses and fungi, should be considered in IBD pathogenesis in addition to bacterial dysbiosis. When compared to healthy controls, IBD patients have been found to have fungal dysbiosis, suggesting that the gut environments of CD patients favour fungus over bacteria (Sokol et al., 2017).

2.3 Molecular mechanism of IBD

2.3.1 Immunological factors

The identification and subsequent investigation of a large array of intestinal inflammation through the study of mouse models has been a significant achievement in the research of IBD (Salem et al., 2022). It has been established from studies that antigens from bacteria and food were always present in the gastrointestinal system, requiring gut equilibrium to be maintained in the absence of intestinal inflammation by minimizing excessive immune response to adverse antigens (Lee et al., 2018). Innate immune response has been highlighted in the studies of IBD pathophysiology since the revelation of GWAS susceptibility gene *NOD2* (Iida et al., 2018). Although intestinal epithelial cells (IECs) function as a wall to inhibit pathogen invasion and antigen influx (Mancini et al., 2021). Mucins and trefoil factors are considered to be the key protective components of the mucus layer that covers the surface of the intestinal lumen produced by IECs (Gerritsen, 2015; Hisamatsu et al., 2013). It is found that healthy mucosa is discovered to be good at keeping back faecal microorganisms, preventing residing microflora from coming into direct contact with the epithelium but due to local inflammatory alterations, this mucosal cleavage function is disrupted in patients with IBD (Sarrabayrouse et al., 2020).

2.3.1.1 Cytokines

Various IBD susceptibility loci comprising genes that encode cytokines and related proteins in cytokine signalling have indeed been established through GWAS (Verstockt et al., 2018). Cytokines govern extraintestinal manifestation, systemic effects in IBD, in addition to driving intestinal inflammation and related symptoms such as diarrhoea (Marafini et al., 2019) (Fig 2.4). The modulation of gut immune system relies completely on cytokines functions, and are produced by monocytes, lymphocytes (T cells e.g., Th1 and Th2), macrophages, granulocytes, IECs, endothelial cells and fibroblasts (Ramos and Papadakis, 2019). Innate immune system respond through macrophages and monocytes cells by releasing proinflammatory cytokines like interleukins such as IL-1, IL-6, IL-8, IL-12 and tumour-necrosis factor (TNF-α) and then the cytokine environments then directs the development of T and B lymphocyte-mediated adaptive immunity (Papadakis and Targan, 2000). According to the compiled research review from mice studies, IBD inflammation occurs either due to the abundance of effector T-cell responses (Th1 and Th2) or by the lack of regulatory responses to the colonic mucosal antigens (Lai et al., 2019).

2.3.1.2 Pathogenesis in IBD

The epithelial layer is the first line of defense for immunological responses in the gut lumen. Transmembrane mucins cover the apical surface of epithelial cells, and goblet cells produce gel-forming mucin to make mucus. The mucinous layer protects the interior of the membrane bound compound of the GI tract. If there is a leaking wall in the mucosa caused by the lack of mucin, then bacteria can encounter epithelial cells. There are two layers in the colon, the outer layer which is called mucinous layer that serves as a home for gut flora and the inner layer which is bacteria resistant. Bacterial enzyme protease has the ability to degrade mucin which can result in penetration of germs into the epithelial layer making it a faulty mucinous layer (Hoskins, 2018). Several types of antigen-presenting cells (dendritic cells and macrophages), T-helper, regulatory T cells, natural killer cells play a crucial role in the pathogenesis of UC (Kałużna et al., 2022). Cytokines networks indicate an important part of signalling for cell communication. Pro-inflammatory cytokines play a crucial role in up-regulation and anti-inflammatory cytokines play in down-regulation of pathogenesis progression (Tatiya-Aphiradee et al., 2019) (Fig 2.3).

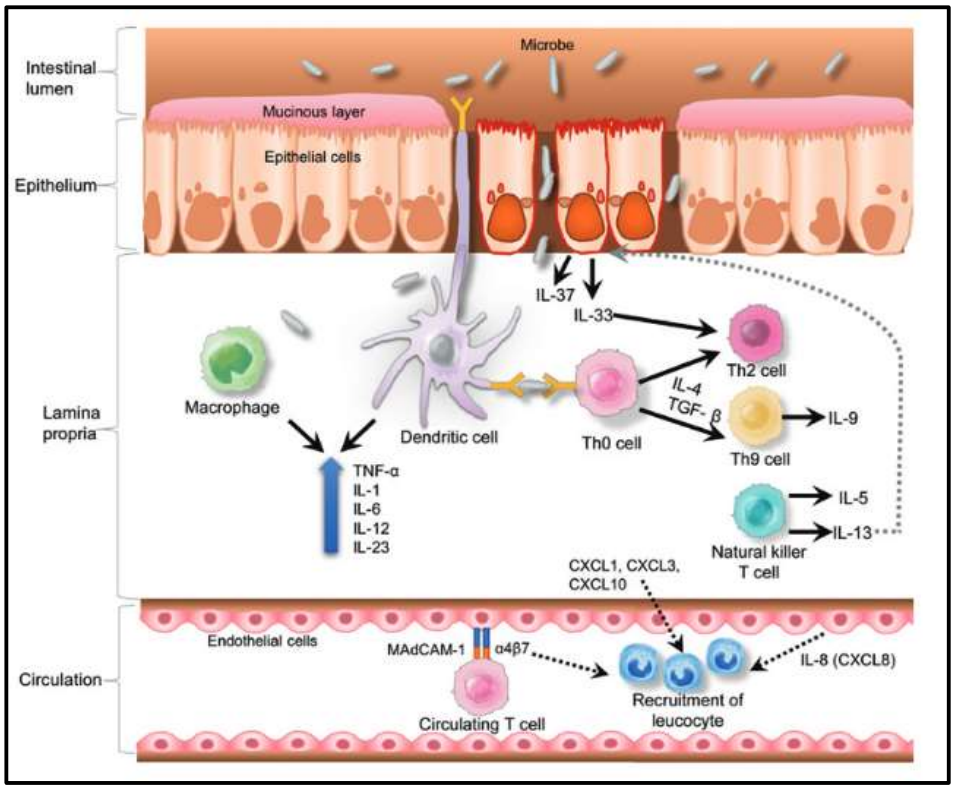


Figure 2. 3 Pathophysiology of IBD represents the immune response and inflammatory pathway of IBD. The pathogenesis of stricture formation in IBD is a complex process with wide variety of clinical, genetic and serologic risk factors. In this diagram it explains epithelial layer is the first line of defense of immune response in gut lumen consisting of transmembrane mucins and goblet cells. The dysregulated immune response to gut microbiota by rupturing the epithelial barrier by becoming permeable resulted in the phenomena leading to an increase in luminal antigen uptake altering the innate and acquired host immune responses (Tatiya-Aphiradee et al., 2019).

2.3.1.3 Wnt signalling pathway

It has been discovered that wingless related integration site or Wingless/integrated pathway's (Wnt pathway) mammalian homologues implies a significant role in stem cell development, embryonic development, and cancer (Nguyen et al., 2018). For two reasons, the significance of the WNT pathway in the pathophysiology of IBD is potentially intriguing in UC patients, first, it have a higher risk of colon cancer (Hogezand et al., 2002) and second, Wnt pathway is critical for the generation and maintenance of GI tract stem cells and colon health (Takahashi, 2020; Van De Wetering et al., 2002). There are two types of Wnt signalling pathways: 1) β-catenin-dependent signalling (canonical pathway) and β -catenin-independent signalling (non-canonical Wnt pathways including planar polarity and Wnt/Ca2+ pathways) (Flores-Hernández et al., 2020). IBD patients have a higher risk of colon

cancer, it's worth noting that mutations that cause constitutive activation of the canonical Wnt pathway are responsible for the majority of colon cancer cases (Zhao et al., 2022).

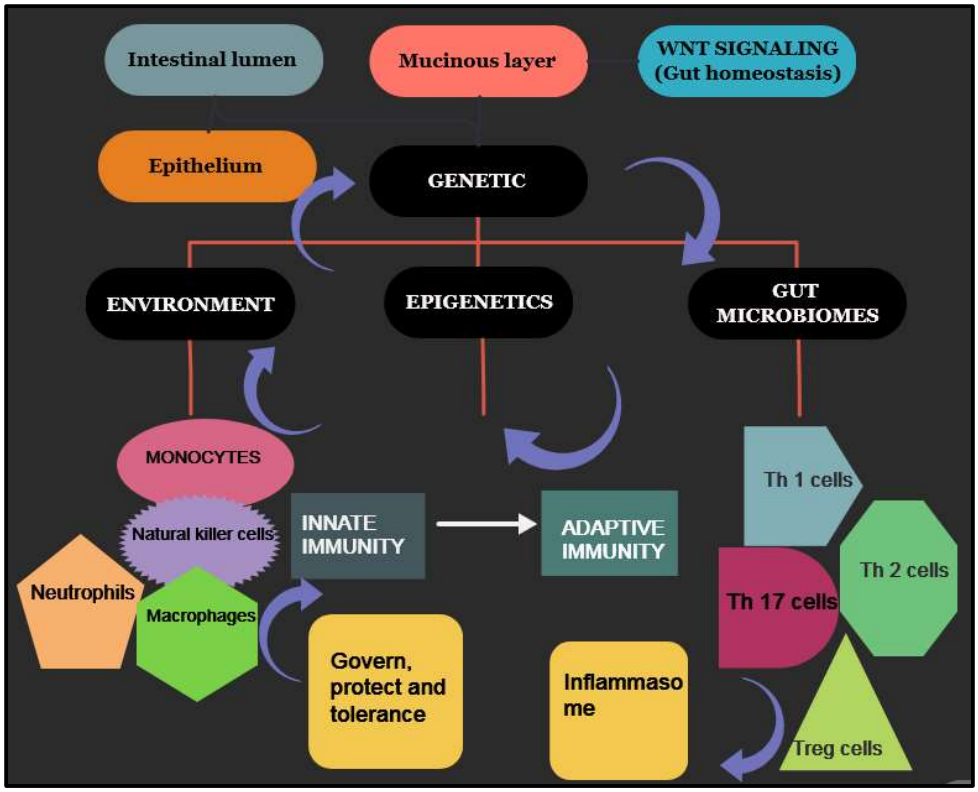


Figure 2. 4 Molecular mechanism involved in the pathogenesis of IBD. Disruption of intestinal barrier due to the genetic, epigenetic and environmental factors and gut microbiomes. This leads to abnormal immune responses in the intestine, responsible for imbalancing the Th1/Th2/Th17/Treg cells and increases the cytotoxicity (NK cells) and increased activation of dendritic cells causing inflammasome.

2.4 Biomarkers

The medical consultation is just the beginning of a much longer process and the sickness process begins with personal acknowledgment of altered body sensations and concluded with the person being classified as sick by relatives or oneself (Kleinman et al., 1978; Leventhal et al., 2020). With the current growth of technology in human genome project sequencing has aided in the advancement of medical research paths and the development of diagnostic instruments and methodologies (Lunshof et al., 2022). It is claimed that a disease begins when the best marker available to define it is identified and it is necessary to indicate clinical symptoms that can aid in the identification of any advancement of an illness discovered at its early stage can be treated (Kraus et al., 2011). Biomarkers (biological markers) are defined as the biological indicators of health. A biomarker is a measurable trait that can be used to assess normal biological mechanisms, biochemical processes, or therapeutic reactions to a

treatment intervention (Dalle-Donne et al., 2006; Zhou et al., 2018). Vital indicators such as pulse rate have traditionally been used by health care practitioners to monitor health status and analyse effectively disease react to therapeutic interventions (Qian et al., 2018). Eventually verified and confirmed biomarkers are statistically detectable indications of physiological or pathologic processes (Depledge, 2020), perform a vital role in disease diagnosis, disease progression prognosis, and therapy response monitoring (Sesay et al., 2018). The National Academy of Sciences in Washington, DC, USA has advocated for the development of a 'New Taxonomy of Disease' to characterize and identify diseases based on the underlying molecular aetiology rather than conventional clinical symptoms to advance understanding of pathophysiologic processes and knowledgebase data of human health (Council, 2011; Kraus, 2018).

2.4.1 Classification of Biomarkers

Biomarkers can be categorised according to a variety of factors (Xu et al., 2020). A biomarker is a precise indicator of an individual's tendency to develop a disorder (Califf, 2018). A biomarker should be applicable to asymptomatic people in order to characterize their risk of developing a disease, implying a target for intervention (Torres et al., 2020). Biomarker is a catch-all term for a wide range of diagnostic and therapeutic phenomena. Biomarkers have been effectively classified into distinct categories, along with predictive and prognostic tests (Fig 2.5). Classification of biomarkers include:

2.4.1.1 Biomarkers based on their diagnostic properties (Carlomagno et al., 2017):

- **Diagnostic biomarker:** allows the early detection of the cancer in a non-invasive way and thus the secondary prevention of the cancer (Link and Kupcinskas, 2018).
- **Predictive biomarker:** allows predicting the response of the patient to a targeted therapy and so defining subpopulations of patients that are likely going to benefit from a specific therapy (Camidge et al., 2019).
- **Prognostic biomarker:** is a clinical or biological characteristic that provides information on the likely course of the disease; it gives information about the outcome of the patient (Moreno et al., 2021).
- **Therapeutic biomarker:** is generally a protein that could be used as target for a therapy (Holmes et al., 2021).

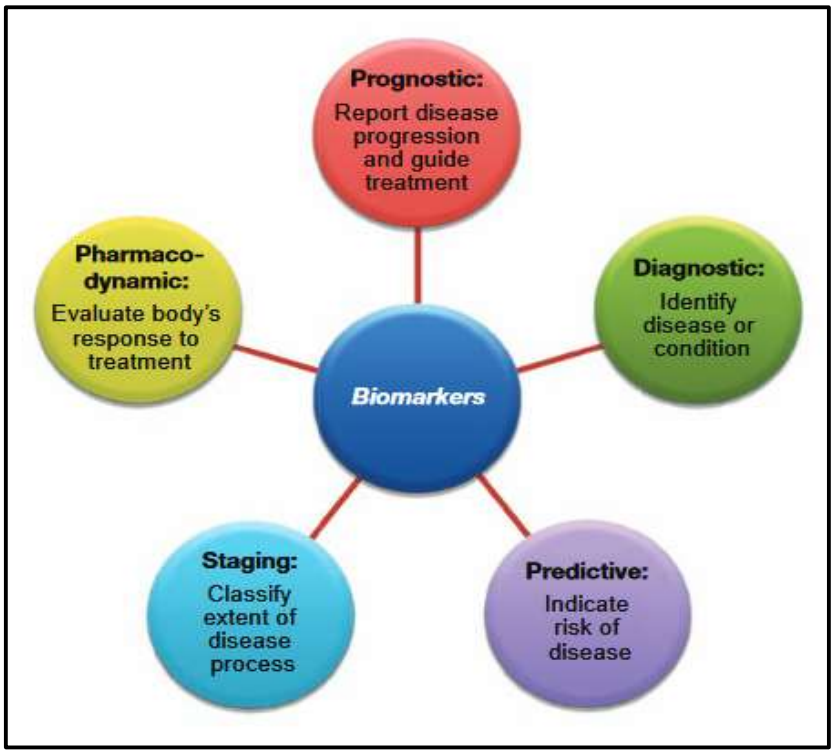


Figure 2. 5 The chart represents different categories involved in application of biomarkers (Pierce et al., 2012).

2.4.1.2 Clinical trial settings-based biomarkers (Kroll, 2008):

- **Disease biomarker:** associated with clinical outcome or disease measurement.
- **Competence biomarker:** biomarker that reflects a beneficial effect of a given treatment (Hampel et al., 2010).
- **Toxicity biomarker:** biomarker that intercepts a drug's toxic effects in an *in vitro* or *in vivo* system (Gadhave and Kokare, 2019).
- **Target biomarker:** biomarker that reports interaction of the drug with its target (Oskooei et al., 2019).
- **Staging biomarker:** biomarker that distinguishes between different stages of chronic disorder (McGorry et al., 2014).
- **Surrogate end-point biomarker:** biomarker that is regarded as a valid substitute for a clinical outcomes measure (Group, 2016).
- **Pharmacodynamics biomarker:** a biomarker that assesses a patient's response to a medical condition or a clinical intervention (García-Gutiérrez et al., 2020).

2.4.1.3 Ideal biomarker characteristics (Bennett and Devarajan, 2011):

- Biomarkers should be non-invasive, easy to measure, inexpensive, and generate results quickly
- Biomarkers should indeed be derived from readily available sources, such as body fluid like urine and blood serum.
- Biomarkers should have a high sensitivity, allowing early diagnosis and thus no interference in values among both affected patients and controls.
- Biomarkers ought to have a high specificity, to be significantly up-regulated (or down-regulated) explicitly in diseased samples and undeterred by confounding factors.
- Biomarker levels should respond quickly to therapy.
- Biomarker levels should assist in risk prediction and must predict prognostic relative value of tangible results; and
- Biomarkers should be physiologically plausible as well as provide understanding of the underlying illness mechanism.

Biomarkers are particularly significant for early detection since they can detect the outcome of interactions between genetic susceptibility and the environment (Srinivas et al., 2001). While not all biomarkers will be able to match all the desired features, they should at least meet a few them (Sobsey et al., 2020). On receiver-operating characteristic curves (ROC), biomarkers must demonstrate robust biomarker qualities (Bantis et al., 2021).

2.4.1.4 Clinical assessment of biomarkers

Clinical evaluation has been the benchmark by which all biomarkers are assessed, as clinical outcomes are the most significant factors in the formulation of disease-specific medicines. When used as a biomarker to predict illness prevalence, progression, and response to treatment, it is considered the gold standard of diagnosis (Kelly et al., 2016).

A further important concept is whether the biomarker's efficacy should be compared with that of clinically useful biomarkers. Even though, setting such as testing standards for example sensitivity and specificity values that potentially have therapeutic utility, is a frequently overlooked aspect of research design (Pepe et al., 2016). Furthermore, for the biomarker to provide accurate and efficient diagnosis of specific diseases as well as respond to therapeutic interventions, it must be aided by proper techniques and statistics, preventing any errors. As a result, biomarker evaluation should be supplemented and assisted via biostatistical techniques.

In 2003, the guidelines and principles for diagnostic tests were established by Standards for Reporting of Diagnostic Accuracy committee (STARD). The main aim of STARD was to test how accurate the test is by improvising and revising the previously existed methods (Bossuyt et al., 2003). It is a quality measure of investigating in clinical settings to provide the best available clinical care facilities to populations. Furthermore, there has been an improvement made to STARD guidelines (Sounderajah et al., 2020). The STARD guidelines table is shown as Table 8.1 in Appendix.

2.4.2 Biomarkers in IBD: Clinical Utility

Biomarkers are quantitative indications of a biological condition that are frequently utilized as disease indicators in clinical care. Although not one biomarker has proven to be robust enough to aid in the diagnosis of IBD and the distinction of UC from CD (Rogler and Biedermann, 2015). Clinical trials, colonoscopy examination, histology, radiologic, and antigen detection testing are all elements to consider in the guidelines according to a second European evidence-based agreement on CD diagnosis and management (Van Assche et al., 2010). Mucosal healing is perhaps the most valid end-point used to determine risk of disease complications and hardly any end-points have been fully validated as surrogates of risk of disease complications (Dulai et al., 2019). However, it is found that serum and faecal biomarkers like c-reactive protein, calprotectin and lactoferrin are the only non-invasive form of diagnosis in IBD (Kopylov et al., 2014).

2.5 Biomarkers of IBD

2.5.1 Serological Biomarkers

The scientific study of serum and other bodily fluids is known as serology (Lee et al., 2019a). Typically, the term refers to the identification of antibodies in serum for diagnostic purposes. Antibodies of this type are usually formed in response to an infection or foreign proteins (Katsanos et al., 2017). Serological markers have long been used to assess disease activity in a non-invasive manner. These biomarkers, on the other hand, are inexpensive, simple to perform, and non-invasive, making them an appealing potential tool for assessing disease activity in a daily practice setting (Miranda-García et al., 2016).

1. **C-reactive protein**

C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) are two commonly utilized clinical surrogative biomarkers for disease activity in CD and UC and other inflammatory illnesses that indicate overall inflammation surrogative biomarkers (Chen et al., 2019). In 1930, CRP was

first identified under different name (Sheriff, 2022; Tillett and Francis Jr, 1930). Tillet and Francis named Fraction-C at the Rockefeller Institute, USA after serum obtained at regular intervals from sick patients with pneumonia or recovered from the illness was diluted with c-polysaccharide and discovered it to be precipitated and later Fraction-C was changed to CRP (Tillett and Francis Jr, 1930). CRP is produced by the liver in response to cytokines, and both IL-1 and IL-6 boost CRP production through improving translation and transcription. In active CD, monocytes release more IL-1 and IL-6 than in UC (Aschenbrenner et al., 2021; Mazlam and Hodgson, 1994). However, increased level of CRP (>45 mg/L) in IBD patients indicates patients need to have colectomy with a high degree of certainty, indicating severe continuing and unpredictable inflammation in the GI tract (Vermeire et al., 2004). However, it was discovered in another study that CRP levels in UC were higher than usual but significantly lower than those in CD (Rubin et al., 2021), implying that serum CRP measurement could assist in distinguishing the two conditions, particularly in severely affected people (Fagan et al., 1982)

2. Erythrocyte sedimentation rate (ESR)

Another broadly adopted biomarker in the diagnosis of IBD is ESR, which is reported to be higher in active CD than in dormant UC and CD (Liu et al., 2013). The ESR is the rate at which red blood cells (RBC) travel via plasma per hour. ESR is a quick but rudimentary way to determine the general typical phase response (Iskandar and Ciorba, 2012). Pro-sedimentation factors, such as fibrinogen, cause RBCs to clump together and settle faster whenever there is inflammation present (Ranjbar et al., 2022). Demographic factors and lifestyle factors including disease symptoms hypertension, dyscrasias, anaemia and maternity are among the factors that influence the ESR (Alende-Castro et al., 2019; Mendoza and Abreu, 2009).

3. pANCA And ASCA biomarker

Perinuclear antineutrophil cytoplasmic antibodies (pANCA) and Anti–*Saccharomyces cerevisiae* antibodies (ASCA) are the recent serum-based biomarkers, help to differentiate between UC and CD (Smids et al., 2017). According to Joosens et al, it is reported that in UC patients 63% were pANCA positive and ASCA negative and in CD patients 80% were found to be pANCA negative and ASCA positive (Joossens et al., 2002). However, it was discovered in a study that among CD patients with ASCA positive and pANCA negative, their diagnostic test had a sensitivity of 55% and a specificity of 92%, making it ineffective for clinical use (Reese et al., 2006).

2.5.2 Faecal Biomarkers

The premise behind faecal biomarkers is that faeces come into direct contact with the intestinal mucosa, it should include specific indicators of mucosal pathology (Mangifesta et al., 2018). As a result, such indicators should be able to accurately indicate the presence and severity of intestinal inflammation (Sutherland et al., 2008). Calprotectin and lactoferrin are faecal inflammatory biomarkers that are an undeviated outcome of intestinal inflammation and can be used to detect and diagnose UC and CD (Kopylov et al., 2014).

1. Calprotectin

Serological markers have long been used to assess disease activity in a non-invasive manner. These biomarkers, on the other hand, are inexpensive, simple to perform, and non-invasive, making them an appealing potential tool for assessing disease activity in a daily practice setting (Steinbakk et al., 1990). While monitoring the extent and severity of inflammation is a significant step. But it is not always possible to follow up on alteration of symptoms, as the only gold standard for examination of GI tract is endoscopic evaluation but it is an invasive form of diagnosis and is expensive (Røseth et al., 1992). Although the level of faecal calprotectin decreases with time, clinicians advised UC patients to take stool samples of their first bowel excretion every two days and to note the consistency and time of the faeces, however its contribution in continuous monitoring while enhancing its deep-rooted and extended consequences has still not been thoroughly investigated (Bressler et al., 2015). However in one recent study a practical guideline was established stating to minimize the dosage variability and monitoring of faecal calprotectin in patients with IBD (D'Amico et al., 2021).

2. Lactoferrin

Lactoferrin is an iron-binding glycoprotein with a molecular weight of 76 kDa (Adlerova et al., 2008). It is found in polymorphonuclear neutrophils' (PMNs) secondary granules and is secreted by colonic mucosa membranes (Legrand et al., 2005). Lactoferrin can be found in human bodily fluids such as tears, bile, and milk. Because lactoferrin serves as a bactericidal agent in the innate immune system, it degranulates secondary granules throughout the inflammatory process (Baveye et al., 1999; Burton and Roitt, 2016). However, one of the principal drawback of faecal lactoferrin is that, as like calprotectin, it might elevate its level after using non-steroidal anti-inflammatory drugs (NSAID's) most likely due to the intestinal obstruction caused, therefore it is criticised by

the clinicians to not use any non-steroidal medicine before the faecal test for IBD (Gisbert et al., 2009; Patil et al., 2019).

Beside the assessment and monitoring the therapy interventions. It has been found that other faecal tests like leukocyte radioactive labelling is not worth than calprotectin and lactoferrin. As it has been reported that faecal calprotectin biomarker level >50 mg/l can predict clinical relapse in IBD patients after a few months in the individuals with clinically dormant IBD with over 80% sensitivity in patients with clinically quiescent IBD (Angriman et al., 2007). Even though faecal biomarkers are inexpensive and non-invasive, still they have some limitations like the test's clinical relevance that is limited by the necessity for gut washout to accurately quantify faecal proteins. The obvious solution to the problem is to use proteins that are stable in the faeces (Angriman et al., 2007). Also, a proper monitoring guideline to monitor the severity of disease and follow up is required.

2.6 Colorectal cancer

2.6.1 Epidemiology of CRC

CRC is the third common cancer and the fourth leading cause of cancer mortality in the world, approximately one to two million new cases diagnosed each year (Mármol et al., 2017). Chronic inflammation in IBD patients predisposes to the development of IBD-associated CRC. As it can cause mutagenesis, and the relapsing-remitting nature of this inflammation, combined with epithelial regeneration, may exert selective pressure that accelerates carcinogenesis (Porter et al., 2021). CRC is also known as adenocarcinoma as it arises from epithelial cell lining of the large intestine, when certain epithelial cells acquire a series of genetic alterations that provide them with a selective advantage (Ewing et al., 2014). Taking age into consideration, it is more common in older men above 50 years of age than in women, as the cause was found to be environmental risk than genetics (Keum and Giovannucci, 2019). According to GLOBOCAN 2018, the highest incidence are found in the developed nations like European countries, New Zealand, Hungary with 70.6 per 100000 in males and Norway with 30 per 100000 in females, as well in some Asian countries but in African nations and South Asian the incidences were found to be low (Rawla et al., 2019). In a recent report from GLOBOCAN 2020, estimated 1.93 million new CRC cases diagnosed and 0.94 million CRC-related deaths globally. US and China were estimated to have the highest CRC cases in the coming next two decades (Sung et al., 2021; Xi and Xu, 2021).

2.6.2 Pathology of CRC

The knowledge of molecular mechanism and the role of molecular testing in the management of CRC can aid in the pathological reporting of disease diagnosis and treatment. The majority of CRCs develop

from precursor lesions such as adenoma and progress to adenocarcinoma (Harada and Morlote, 2020). Pathological stages of CRC can aid in the prediction of disease severity and can accurately aid in identifying the 5-year survival rate, as well as patients' remission, relapse, and recurrence condition (Win et al., 2015).

2.6.2.1 Histology of CRC

The World Health Organization's (WHO) proposition of histological classification of CRC which is internationally accepted for consistency and uniformity in pathological reporting (Hamilton and Aaltonen, 2000). For the first time, many tumour forms are classified based on molecular phenotype along with their histological characteristics; nonetheless, histopathological categorization remains the gold standard for diagnosis in the majority of cases (Nagtegaal et al., 2020).

1. **Adenocarcinoma**

The most prevalent kind of tumour is adenocarcinoma > 90%. Most are poorly differentiated and lack distinctive histological markers, while CRC tumours frequently have cribriform patterns involving central necrosis, which is relevant if a metastasis tumour is discovered without a CRC source. Although surrounding mucosal dysplasia may be seen, the invasive tumour frequently eliminates any existing polyp through which it may have developed (Treanor and Quirke, 2007).

2. **Mucinous adenocarcinoma**

Mucinous colorectal carcinoma (MCC) is a form of CRC marked by an abundance of extracellular mucus. In order to make this diagnosis, at least 50% volume of the tumour must be mucinous. Microsatellite Instability (MSI) is linked to mucinous adenocarcinomas (Debunne and Ceelen, 2013). Poor prognosis, advanced stage upon diagnosis, aggressive activity, and a high rate of metastasis are all traits they share (Kanazawa et al., 2002; Shivji et al., 2020).

3. **Signet-ring cell carcinoma**

Signet-ring carcinoma is a kind of mucin-secreting adenocarcinoma in which the component cells maintain an abundance of intracytoplasmic mucin (Jiang et al., 2011), pushing their nuclei to one side while giving the cells their distinctive histologic appearance (Almagro, 1983). It involves in more advanced stages but presentation at a younger age, a lower rate of curative resection, a greater

risk of local lymphatic dissemination and peritoneal metastasis, more aggressive biological behaviour, and a worse prognosis than ordinary colorectal adenocarcinoma (Liang et al., 2018).

4. Adenosquamous carcinoma

Squamous carcinoma and adenocarcinoma characteristics are present in these rare tumours, either as discrete sections within the tumour or admixed. Adenosquamous carcinoma is a cancer that contains both glandular and squamous components in its histology. Both components are aggressive and have the ability to metastasize (Cagir et al., 1999).

5. Medullary carcinoma

An extremely rare type of CRC is medullary carcinoma (MC) Its clinicopathologic findings are still unclear. CRC with MC is a relatively unusual form. Its clinicopathologic findings are still being studied. Sheets of malignant cells with vesicular nuclei, prominent nucleoli, and dense eosinophilic cytoplasm with substantial intra-epithelial lymphocyte infiltration characterize MC (Pyo et al., 2016). In comparison to certain other poorly differentiated and undifferentiated colorectal carcinomas, it is almost often associated with MSI and has a favourable prognosis (Zenger et al., 2019).

2.6.2.2 Tumour Grading and Staging in CRC

Several grading criteria for CRC have been proposed in the literature. However, the most frequently accepted and regularly used standard for grading is set based on degree of differentiation (Ueno et al., 2012). Tumour grade is determined by the tumour, node metastases (TNM) classification of tumour differentiation, with grade 1 indicating well-differentiated tumours, grade 2 indicating moderately differentiated tumours, grade 3 indicating poorly differentiated tumours, and grade 4 indicating undifferentiated tumours. WHO classifies tumour grade as least differentiated, with well and moderately differentiated adenocarcinoma being considered low grade, and both poorly differentiated and undifferentiated adenocarcinoma being classified high grade (Hamilton and Aaltonen, 2000). Although the basic approach to adenocarcinoma of the colon and rectum is same in the WHO 2019 system, there has been a modest adjustment to criteria for grading and more subtypes have been introduced. The percentage of gland formation within the tumour was the sole basis of grading in the previous 4th edition WHO and the current 8th edition American Joint Committee on Cancer (AJCC) that is < 50% for high grade and > 50% for low grade (Ahadi et al., 2021).

Since the publication of the Duke's classification schemes, the degree of tumour differentiation, as measured primarily by architectural features, has been the representative criterion of tumour grade used for CRC (Salibasic et al., 2019). Cuthbert Esquire Duke (1890 - 1977) categorized them as stage A, B, or C tumours, with stage A tumours confined to the rectal wall, stage B tumours breaching extra-rectal tissues, and stage C tumours having metastasized to lymph nodes (Dukes, 1932). Duke later followed up on the survival rate and his method proven to be effective in identifying individuals who would benefit from postoperative chemotherapy in Duke's C stage and also B but found lowest survival rate in stage D (distant metastases) (Haq et al., 2009)

The TNM staging approach determines the extent of cancer based on anatomical information regarding the size and extent of the primary tumour (T), regional lymph-node status (N), and distant metastases (M), and groups patients with similar prognosis (Wittekind et al., 2019). The International Union for Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC) collaborate to maintain the system, which results in the TNM Classification of Malignant Tumours and the AJCC Cancer Staging Manual being published on a regular basis and simultaneously (Obrocea et al., 2011) (Table 2.1).

Table 2. 1 AJCC and TNM staging classification of CRC (Kheirelseid et al., 2013).

AJCC stage	TNM Stage	TNM stage characteristics for CRC
		Tx: Primary tumour cannot be evaluated
		T0: No proof of primary tumour
Stage 0	Tis N0 M0	Tis: Tumour confined to mucosa; cancer *in situ*
Stage I	T1 N0 M0	T1: Tumour invades submucosa
Stage I	T2 N0 M0	T2: Tumour invades muscularis propria
Stage II-A	T3 N0 M0	T3: Tumour invades subserosa or beyond (without other organs involved).
Stage II-B	T4 N0 M0	T4: Tumour invades adjacent organs or perforated the visceral peritoneum
Stage III-A	T1-2 N1 M0	N1: Metastasis to 1 to 3 regional lymph nodes. T1 or T2.
Stage III-B	T3-4 N1 M0	N1: Metastasis to 1 to 3 regional lymph nodes. T3 or T4.
Stage III-C	Any T, N2 M0	N2: Metastasis to 4 or more regional lymph nodes. Any T.
Stage IV	Any T, any N, M1	M1: Distant metastases present. Any T, Any N.

2.6.3 Aetiology of CRC

CRC usually manifests itself in one of three ways: sporadic, hereditary, or familial. The patient is stated to have a family history of CRC if a first-degree relative is diagnosed with CRC or colon polyps before the age of 60, or if any member of that family is diagnosed at any age, however the data is conflicting (Smith and Samadder, 2018). The transport, ionic strength, oxygen access, nutrition supply, host secretion (such as gastrointestinal enzymes and bile juice), mucus membranes, and immune response linkages, all affect bacterial populations in different parts of the GI tract stabilization (Kvakova et al., 2021; Liu et al., 2019). However some diets can impact into dysbiosis of gut health and can lead to certain illness like obesity (Flint et al., 2012). With short chain fatty acids like acetate is enhanced by food intake and is produced by acetogenic bacteria is the product of fermentation, results in obesity (Louis et al., 2014).

2.7 Molecular mechanisms of CRC

CRC is caused by agglomeration of epigenetic and genetic modifications that lead the normal colonic mucosal epithelium to convert into colon adenocarcinoma (Shirmohamadi et al., 2020). CRC is a multifaceted disease with numerous molecular subgroups (Gaiani et al., 2021). Four key points of cancer pathogenesis have been established because of the study of the molecular genesis of colon cancer. The first principle is that the genetic and epigenetic modifications that drive colon cancer formation encourage cancer formation by providing a clonal growth advantage to the cells that acquire them. So basically, colorectal carcinogenesis genetic model was proffered by Fearon et al, in 1990 in which he explained the DNA alterations that led to the clinical progression of CRC and in the next principle is stated that cancer manifests itself in a sequential process at both molecular and the morphologic levels (Fearon and Vogelstein, 1990). Third principle is that genomic instability is a crucial biological stage in the development of cancer. The fourth point is when hereditary cancer syndromes typically correspond to germ line variations of major genetic abnormalities whose somatic manifestations cause sporadic colon cancers to appear (Kinzler and Vogelstein, 1996).

2.7.1 Pathogenesis in CRC

Proportionately three distinct molecular carcinogenesis or any genetic instability for CRC have been identified in cancer genetics. The chromosomal instability (CIN), which is seen in 70% of patients with sporadic CRC, is the most frequent genetic phenomena (Pino and Chung, 2010). It is characterized by an increased rate of chromosome unrealized losses and gains, replication stress, mitotic check point errors resulting in karyotypic diversity among cells (Lengauer et al., 1997; Patil et al., 2022). The MSI is the second mechanism of genomic instability, contributing for about 15% of the cases. The loss of DNA mismatch repair (MMR) activity causes microsatellite instability, which is a hypermutable trait (Wu et al., 2020a). The third and last epigenetic alteration is CpG island methylation (CIMP), which is more associated with serrated neoplasia (Fig 2.6) (De Palma et al., 2019). More than 80% carcinomas arise along the serrated pathway developing from the precursor lesion known as the sessile serrated adenoma (Overman et al., 2018). CIMP's molecular mechanism is unknown, however it is seen in nearly all tumours, when *MHL1* (Kouzminova et al., 2010) a MMR gene involved in fixing base-to-base mismatches and insertion-deletion loops, is methylated abnormally (Chang et al., 2018).

2.7.1.1 Chromosomal instability and key genes involved in CRC

Majority of malignancies have lost or gained chromosomes, according to karyotypic studies, and molecular investigations suggesting that karyotypic data actually underestimates the full magnitude of

such changes. Heterozygosity losses (LOH), or the loss of a maternal or paternal allele in a tumour, are common, and they are frequently accompanied by a gain of the opposing allele (Lengauer et al., 1998). CIN is found in 60 – 70 % of Sporadic CRC. CIN causes an unequal number of chromosomes (aneuploidy), sub-chromosomal genomic amplifications, and a frequent LOH (Pino and Chung, 2010). Several genes are involved in CIN pathways. The mechanism of CRC involved important gene players associated with it and epigenetics.

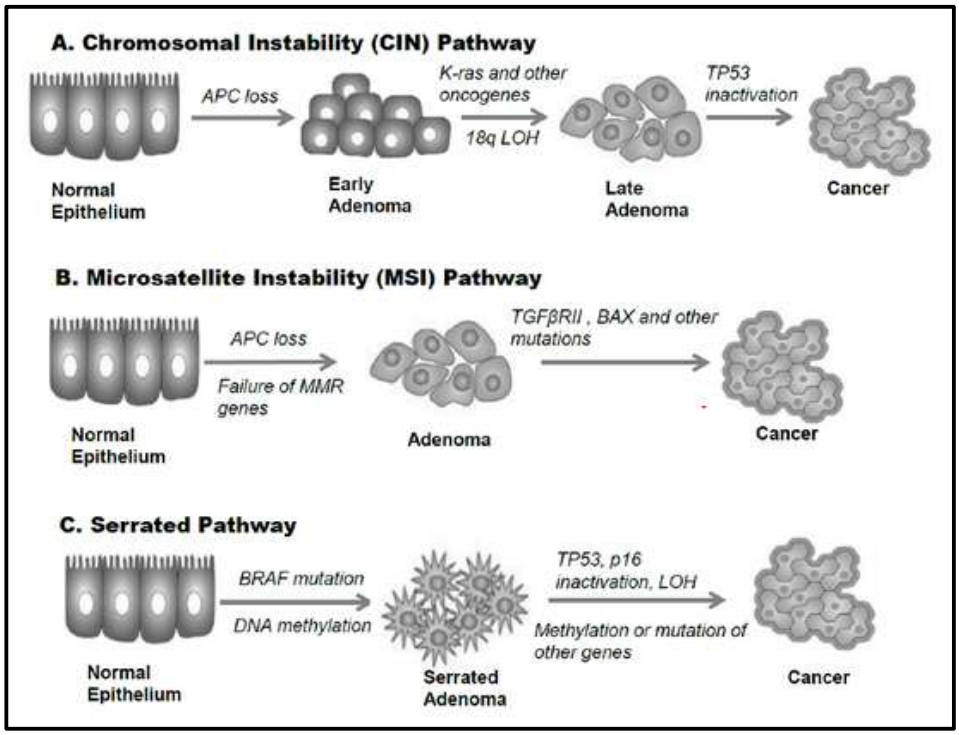


Figure 2. 6 This figure demonstrate the multiple epigenetic and genetic pathways that are involved in the development of CRC (Mundade et al., 2014). The hallmark of most classical (canonical) pathway is CIN that refers to a high load of alterations of whole chromosomes or large portion of chromosomes duplications or depletions (Jung et al., 2020). MSI, involves changes in the number of short repeated sequences termed as microsatellites, spread out across the genome caused by the deficiency in DNA MMR (Wu et al., 2020a). DNA methylation is one of the most ubiquitous epigenetic modifications regulating gene expression. epigenetics defines those of the CIMP subgroup, which are characterized by a high level of CpG island hypermethylation at the promoters of several tumour-suppressor genes (Zhang et al., 2021).

2.7.1.2 Key genes involved in genetic mutations in CRC

1. **Adenomatous polyposis coli (*APC*)**

Adenomatous polyposis coli (*APC*) is an important tumour suppressor gene. The gene has been found to be mutated not just in most colon cancers, but also in several other malignancies, such as

liver tumours. Familial adenomatous polyposis (FAP) is caused by germ-line mutations in the *APC* gene, and it is characterized by many polyps in the intestines (Aoki and Taketo, 2007). In a study conducted by Powell et al, discovered that mutations in the *APC* gene were identified in 60% of adenoma and carcinomas, and that this mutation could be anticipated at an early stage of tumour and remained intact as the tumour advanced from normal to aggressive stages (Powell et al., 1992). In one genetic study, findings showed that *APC* gene alteration to be the key step in the development of intestinal cancers in mice, resulting in an autosomal dominant tendency to produce spontaneous colonic and intestinal tumours (Fodde et al., 1994). By boosting β-catenin levels, mutations in the *APC* activate the Wnt signaling pathway. T-cell factor (TCF) TFs help β-catenin translocate to the nucleus and increase the transcription of different oncogenes (Mann et al., 1999). Gastrointestinal cancers have high amounts of β-catenin (Bakker et al., 2013).

2. Ki-ras2 Kirsten rat sarcoma oncogene (*KRAS*)

Ki-ras2 Kirsten rat sarcoma oncogene homolog (*KRAS*) is a protein kinase mitogen-activated protein kinase (MAPK) pathway member. Cell proliferation, differentiation, senescence, cell survival, cell adhesion and apoptosis are all controlled by the sarcoma virus RAS/ rapidly accelerated fibrosarcoma (RAF)/MAPK system (Khan et al., 2019; Lu et al., 2019). Harvey rat sarcoma virus (HRAS), neuroblastoma RAS viral oncogene homolog (NRAS), and *KRAS* are oncogenes that belong to the RAS family. *KRAS* is the most frequently mutated member of the RAS family in CRC, with mutations seen in 40% of sporadic CRCs. *KRAS* is not thought to be a germline mutation. It is a tiny protein that transmits epidermal growth factor receptor (EGFR) family signals (Armaghany et al., 2012). Several transduction pathways, including RAS/RAF, RAS/PI3-K, and RAS/RAL, promote distinct RAS functions when RAS is activated by certain stimuli and Guanine Exchange Factors (Castagnola and Giaretti, 2005). *KRAS* has been found to be mutated in roughly 30% of colorectal adenomas and 30% to 50% of colorectal cancers (Liu et al., 2011). Although the extent to which colorectal tumours are dependent on *KRAS* is unknown, it is one of the most relevant therapeutic targets for CRC due to the high rate of mutation in this carcinoma (Porru et al., 2018).

3. B- rapidly accelerated fibrosarcoma (*BRAF*)

The RAS-RAF-MAP kinase pathway governs cellular responses to growth signals, and B- rapidly accelerated fibrosarcoma (BRAF) is a proto-oncogene member of the RAF family of serine/threonine kinases (Raskov et al., 2020). BRAF activating mutations have recently been discovered in

roughly 10% of sporadic CRC cases and are uncommon in familial Lynch syndrome CRC (Armaghany et al., 2012). The MAPK pathway is important for tumour cell proliferation and survival. In advanced CRC, data suggests that BRAF oncogene mutations apart from associated with poor prognosis, but is ineffective when treated with anti-epidermal growth factor receptor antibodies (Sanz-Garcia et al., 2017).

4. Phosphatidylinositol-3-kinase (*PIK3CA*)

In nearly 25% of colorectal tumours, PIK3CA, the catalytic subunit of phosphatidylinositol-3-kinase (PI3K), has been somatically altered (Samuels et al., 2004). The p110 catalytic subunit of PI3K, which is one of the most important kinases in PI3K/Akt/ mammalian target of rapamycin (mTOR) signalling, is encoded by *PIK3CA*. However, the effects of *PIK3CA* mutations on first-line chemotherapy treatment are unknown (Wang et al., 2018).

5. Mothers against decapentaplegic homolog 4 (*SMAD4*)

Mothers against decapentaplegic homolog 4 (*SMAD4*) mutations are found sporadically in 2.1–20.0% of CRC tumours. As a common downstream regulator and tumour suppressor gene, *SMAD4* (located on band 18q21) plays a critical function (Sarshekeh et al., 2017). SMADS are a family of proteins that act as intracellular signalling effectors that releases polypeptides belonging to the TGF superfamily (Derynck et al., 1998). Loss of *SMAD4* activity was found to be an independent predictive factor for lower recurrence-free and overall survival in CRC patients, especially those with metastasized cancer, according to studies (Miyaki et al., 1999).

6. Tumour protein gene (*TP53*)

Tumour protein *TP53* gene is frequently mutated in CRC and is an important gene in cell cycle and apoptosis regulation (Lane and Benchimol; Vogelstein et al., 1988). Before a cell commits to the process of DNA replication, the p53 protein initiates G1 cell-cycle arrest and allows DNA repair (Armaghany et al., 2012; Liebl and Hofmann, 2021). The *TP53* mutation is anticipated to develop during the transition from adenoma to carcinoma (Zhou et al., 2017a) and has been discovered in 65-75% of CRC patients (Fearon and Vogelstein, 1990).

2.7.1.3 Epigenetics of CRC

Epigenetics refers to hereditary modifications in gene expression which are regulated primarily by DNA methylation and histone modifications and do not occur as a result of changes in the DNA sequence (Holliday, 1987). It's critical to understand that DNA methylation is a natural process for cells to regulate gene expression in the mammalian genome. The process of DNA methylation involve DNA methyltransferases (DNMTs) by adding a methyl group at 5' position of cytosine forming 5-methylcytosine, a normal nucleotide in DNA (Lao and Grady, 2011). Methylation can take place in both gene promoters and non-promoter areas, known as CpG shores, that are either 2 kbp distant from CpG islands and therefore are regulatory domains of gene activity (Irizarry et al., 2009).

Albeit the fact that gene mutations play a role in cancer development, the significance of epigenetic changes has been contentious until recently (Chen et al., 2020). Epigenetic mutation in carcinoma was first discovered by Feinberg and Vogelstein. They discovered that when colon tumours were compared to normal colons, there was a significant worldwide reduction of 5'-methylcytosine content (Feinberg and Vogelstein, 1983). This global hypomethylation is frequently detected in the early stages of the CRC process, and it is age-dependent (Feinberg, 2004; Suzuki et al., 2006). Many forms of neoplasms exhibit a hypermethylated DNA methylation pattern of typically unmethylated CpG islands in gene promoters (Feinberg, 2004). Indeed, genome-wide investigations of cancer epigenomes have discovered that 1-10% of CpG islands are abnormally methylated, implying that a large proportion of gene promoters are hypermethylated in the average malignancy (Costello et al., 2000; Pfeifer and Rauch, 2009). Epigenetic instability in CRC expresses itself in a variety of ways, including hypermethylation of CpG island-containing gene promoters and genome-wide DNA hypomethylation (Lao and Grady, 2011). Toyota and Issa discovered a significant frequency of methylation of genes in certain CRCs and proposed that these tumours had their own molecular aetiology, named CIMP malignancies (Toyota and Issa, 1999). CIMP tumours make up about 20% of CRCs, and they can now be identified by an array of CIMP marker genes with more than 55% methylation genes. (Lao and Grady, 2011).

2.8 Biomarkers of CRC

Advances in routine screening, laparoscopic methods, adjuvant chemotherapy, and surveillance programs have improved the 5-year survival rate during the last two decades (Ewongwo et al., 2020). CRC is a diverse disease at the molecular level. Differences in illness development, survival, and chemotherapeutic drug response result from this diversity (Newton et al., 2012). This diversity manifests itself in disparities with illness development, survival rate, and chemotherapeutic drug response

(O'Donnell and Dolan, 2009). Although the diagnosis like sigmoidoscopy and colonoscopy are invasive procedures that carry a high risk of comorbidity and necessitate time-consuming pretesting procedures, resulting in a low participation rate (Pellino et al., 2018). Currently the non-invasive CRC biomarkers in use are faecal biomarkers, protein biomarkers and serum biomarkers. Some of the stable serum biomarkers currently in use are mentioned below.

1. Carcinoembryonic antigen (CEA)

During Mid 1960's, almost six decades ago, carcinoembryonic antigen (CEA) was discovered by Gold and Freeman in human colonic tissues (Gold and Freedman, 1965). The same group of research scientists discovered the levels of CEA antigen from serum, taken from the patients with CRC by radioimmunoassay (Thomson et al., 1969). CEA belongs to the CEACAM family. To yield CEA and its preparation process in the earlier decades was lengthy and difficult especially during purification process as the recovery of CEA was substantially low (Krupey et al., 1972). Since several scientist made contributions to the chemical composition of CEA and adopted standard CEA preparation, it was found that CEA is a glycoprotein with 60% make up of carbohydrate weighed together 200 kDa (Westwood et al., 1974; Zhao et al., 2018). Oncofoetal protein is present in amniotic fluid during the embryo development of a foetus, but it is also identified in cancer patients. CEA is the first oncofoetal protein to be discovered (Zaidi et al., 2017). After birth, its production is inhibited, and healthy individuals have very low to no CEA levels (Campos-da-Paz et al., 2018; Sarandakou et al., 2007). The first soluble biomarker for CRC was CEA in serum, and it is now the only soluble biomarker recommended for CRC surveillance (Nielsen et al., 2011). Altogether the sensitivity ranged from 40% to 60 percent. Sensitivity increased significantly with cancer stage, extending from 8% for Dukes A to 89 percent for Dukes D (Hundt et al., 2007). CEA values less than 5 ng/mL were considered normal, whereas those equal to or greater than 5 ng/mL were considered elevated and detected in cancer (Lakemeyer et al., 2021). Although CEA levels are raised in a considerable number of people with preclinical CRC, it is not alone an effective screening tool as its sensitivity is very low and cannot detect illness in immunosuppressed people (Thomas et al., 2015).

2. Tissue inhibitor of metalloproteinase-1 (TIMP1)

A novel biomarker that can be used for early detection and for selection of advent therapy, prognosis, diagnosis and monitoring efficacy of treatment is TIMP1 (Møller Sørensen et al., 2008). The glycoprotein Tissue inhibitor of metalloproteinase-1 (TIMP1) is implicated in cell survival and

cancer (Meng et al., 2018). It belongs to the natural MMP inhibitor family, which has four members, includes TIMP1 (Hornebeck et al., 2005). It weighs about 28.5 kDa (Carmichael et al., 1986). Flat spindle shaped large cells (fibroblast) found in several human tissues like tendons, skin, cartilages are all expressed by TIMP1 (Welgus and Stricklin, 1983). Supplementary chemotherapy is recommended for patients with stage III CRC who are likely to be cured by surgery; nevertheless, it is not encouraged for persons with stage II CRC because overdosing could result in recurrence (Benson et al., 2004). TIMP1 is a potential biomarker that can be exploited for early identification, advent therapy selection, prognosis, and treatment efficacy monitoring.

3. Cancer antigen-724 (CA724)

Cancer antigen-724 or carbohydrate antigen (CA724) is a glycoprotein found on the surface of many cancer cells, including those with hepatocellular carcinoma, ovarian cancer, gastric cancer, and CRC. The diagnostic effectiveness of serum CA724 for CRC diagnosis has been investigated in several studies (Gao et al., 2018; Yanqing et al., 2018). Cancer related antigens' level like CEA, CA724 and CA199 that are elevated have also been linked to a terminal condition (Filella et al., 1992). CA724 and CEA were found to be potential prognostic markers, although Dukes' stage and rectal tumour site were also significant. The serum marker CA724 was found to be high in accordance with metastasis stage 1.8 kU/L (range 0.02–729 kU/L) (Louhimo et al., 2002). It is shown to be elevated in a number of human adenocarcinomas with reported sensitivities as high as 50% and overall specificity above 95% (Mariampillai et al., 2017).

4. Carbohydrate antigen-199 (CA199)

Cancer antigen or Carbohydrate antigen (CA199) is a carbohydrate antigen or cancer antigen that aids in the invasion and spread of cancer cells. E-selectin and P-selectin leukocytes are cell-surface receptors found on endothelial cells adhered when associated with CA199 (Berg et al., 1992). The overall sensitivity and specificity ranged from 18% - 65% and >90% at the threshold of 37 U/mL in most of the experimental investigations (Hundt et al., 2007).

The creation of disease-specific biomarkers is of major therapeutic interest because of the comparative high degree of uncertainty of currently available biomarkers in distinguishing across IBD phenotypes or predicting prognosis and relapse rates. Several studies into novel serum-based antibody or protein

biomarkers for IBD are currently being implemented. Also taking into consideration the ideal biomarker characteristics like non-invasive form of diagnosis and cost-efficient. This study aims to predict early diagnosis of IBD and predicting progression to CRC through computational analysis.

2.9 IBD associated CRC

It is reported that patients with chronic IBD have been found to have a greater chance of getting CRC. There has been a substantial body of evidence relating UC with CRC since the 1920s, when the association was first established (Bargen, 1994). To understand the pathogenesis of CRC in relation to inflammation and pathways between the two diseases at molecular basis is followed by 4 steps: i) Genetic mutation, ii) inflammatory mediators in mucosa like proinflammatory cytokines, COX-2, and chemokines, iii) epithelial cells receptors alterations, iv) factors affecting the gut microbiota (Azer, 2013).

Patients with long-term severe colitis are also more likely to acquire CRC (Rutter, 2011). UC patients have a risk of 5.7 times that of the general population, according to population-based research (Ekbom et al., 1990). CRC is the primary cause of IBD related death, accounting for one out of every six UC patients (Gyde et al., 1982). The most important risk factor for CRC is the anatomical extent of colonic inflammation was first identified in 1990 by Ekbom and colleagues (Ekbom et al., 1990). According to a meta-analysis published in 2001 by Eaden et al, the cumulative CRC risk for a patient with UC is 1.6% after 10 years, 8.3% after 20 years, and 18.4% after 30 years (Eaden et al., 2001). It is reported that in UC patients, the problem of UC-associated CRC rates remained still challenging (Zhou et al., 2019).

2.10 Databases and their importance

With the availability of databases, which allow to store and retrieve files as well as provide with easily accessible data and information at our fingertips, it is necessary to be familiar with and well informed about relevant features and applications, as they have made information in files stored very easily accessible. Therefore, the database is an organized collection of structured information or data typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS) (Fig 2.7). The term ‘datum’ comes from the Latin word ‘datum’, which means ‘provided’. The plural of datum is data. A datum is a logical statement with a label and a value that has been recorded in its most basic form. Even though the phrases data and information are similar

in the dictionary, there is a significant difference in computer jargon since data does not become information until it has been processed and structured (Berrington, 2007). A database is self-describing. This summary is known as metadata, or data about data. The database is linked because it contains both the relationships between data items and the data items themselves (Berg et al., 2013). There are different types of databases existed today and they are described here:

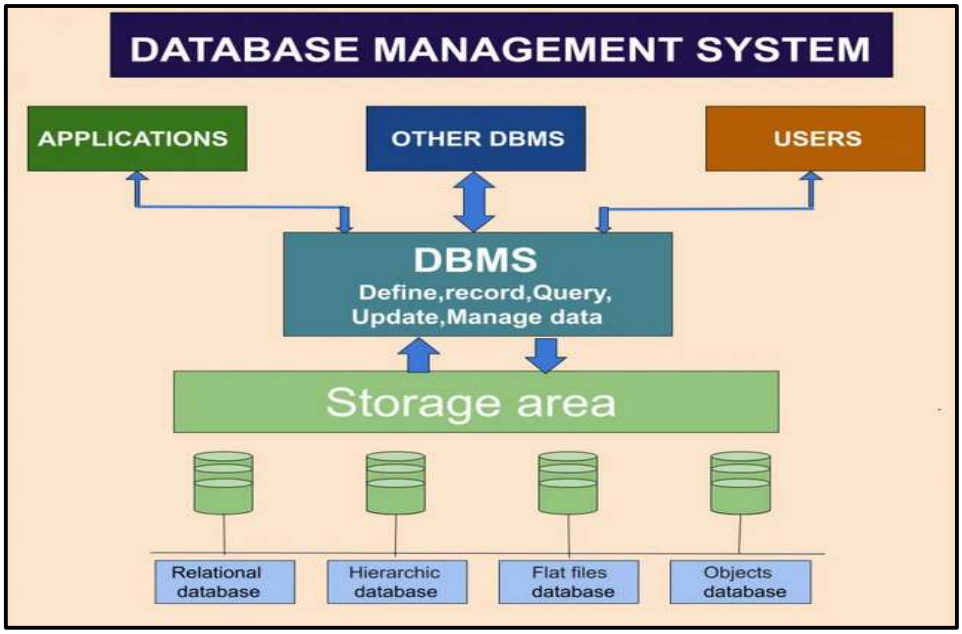


Figure 2. 7 This figure demonstrates the DBMS design and features. The features represent data retrieval, storage, security, integrity and data management. The factors consist of size complexity, types and performance requirement of data. Types of DBMS are hierarchical, network, object-oriented databases and the most popular type of database is relational database. Adapted from https://www.erp-information.com/database-management-system.html**.**

2.10.1 Types of databases

1. Relational database

A relational database management system (RDBMS) is a type of database that stores and provides access to data components that are linked to one another. The relational model is the cornerstone of relational databases (Chen and Lee, 2019). It is a straightforward and apparent way of representing data in tables. In a RDBMS, each row of a table is a record with a unique identifier known as the key. The data's attributes are kept in the table's columns, and each record normally contains a value for each attribute, making it easy to link data points together contains a value for each attribute, making it simple to construct links between data points (Codd, 2007).

- **Object oriented database:** Objects are used to represent data in an object-oriented database, just as they are in object-oriented programming (Kim, 1990).
- **Distributed database:** It is made up of two or more files that are stored in various locations. The database could be spread across various machines in the same physical area or across multiple networks (Rothnie et al., 1980).

  These databases are currently in use today and there are also several other databases. Below are some databases specifically designed for scientific, financial, etcetera use:
- **Open-source databases:** An open-source database system is one whose source code is available to the public; these databases might be SQL or NoSQL (Chen and Xie, 2008).
- **Cloud databases:** A cloud database is a structured or unstructured collection of data stored on a private, commercial, or blended cloud computing platform. Traditional and database as a service are the two sorts of cloud database models (DBaaS). Administrative and maintenance chores are handled by a service provider with DBaaS (Arora and Gupta, 2012).

2.10.2 Existing CRC genes and TFs database

There are few other databases of CRC, called as Colorectal cancer biomarker database (CBD) (Zhang et al., 2018). It was published in 2018 and can provide to users' multiple individual biomarker information, location of tumours, stages, etc. But presently only CoReCG database (Agarwal et al., 2016) for CRC genes are available for users. Currently there are present few other databases of TFs like FANTOM (Lizio et al., 2016), TRANSFAC (Wingender, 2008) and Enrichr (Kuleshov et al., 2016). The following section covers the existing database of CRC that is CoReCG and Enrichr database for TFs.

2.10.2.1 CoReCG Database

The existing database for CRC gene is the CoReCG database, which is a manually curated database of CRC-related genes acquired from the published literature and is accessible at https://lms.snu.edu.in/corecg/. This database was publicly accessible in 2016. It comprises of 2056 genes and their associated data (Agarwal et al., 2016). This database was accessed for this study in the year 2021. CoReCG has compilation of data from 2486 peer-reviewed publications. In contrast, the following cancer and other databases were combined through for evidence of a relationship between colon cancer and genes. This included experimental method applied in the study to validate contribution of genes in CRC, availability of enzyme-drug target for 600 CRC genes, expression changes, renal cancer gene database (Ramana, 2012), CanGeneBase (Kumar et al., 2009), fold change, staging, sample size, gene variants, location of samples collection, post translational modification, various gene IDs, genes annotation, pathway, string protein-protein interaction (PPI) (Franceschini et al., 2012), drug target etc, all retrieved from various sources and databases and verified through PubMed ID.

2.10.3 TFs Database

2.10.3.1 Enrichr Database

Enrichr https://amp.pharm.mssm.edu/Enrichr/enrich# is a web-based interdisciplinary application that incorporates a number of novel genomic libraries, embedded with new approach to rank enriched concepts and provide cohesive and interactive graphical representation of the results and it can be integrated into programming language using key features (Chen et al., 2013; Xie et al., 2021). It evaluates the concepts from gene libraries based on ranking by applying Fisher exact test. It comprises of 35 gene-set libraries either derived from other databases or developed. This gene-set library is divided into six class – transcription, diseases, drugs, pathways and ontologies, cell types and miscellaneous. The transcription class consists of six gene-set libraries – ChEA, position weight matrices (PWMs) from TRANSFAC https://genexplain.com/transfac/ and JASPAR https://jaspar2018.genereg.net/, PWMs from UCSC genome browser https://genome.ucsc.edu/, TFs target retrieved from ENCODE https://www.encodeproject.org/, gene sets associated from histone modification and microRNAs. Many studies have been undertaken to investigate the regulatory interaction using Enrichr between TFs and microRNAs in CRC to better understand the genesis of the disease. (Hossain et al., 2021; Kadkhoda et al., 2021; Peng et al., 2020).

3 CHAPTER THREE – METHODOLOGY

3.1 Research design

The research plan used in this study is designated to provide a pertinent framework of methodology used to accomplish its aims. This project facilitates a novel study for the development of curated IBD database and the identification of potential biomarkers for the early diagnosis of IBD, while predicting progression to CRC. The flow diagram of methodology is illustrated in (Fig 3.1).

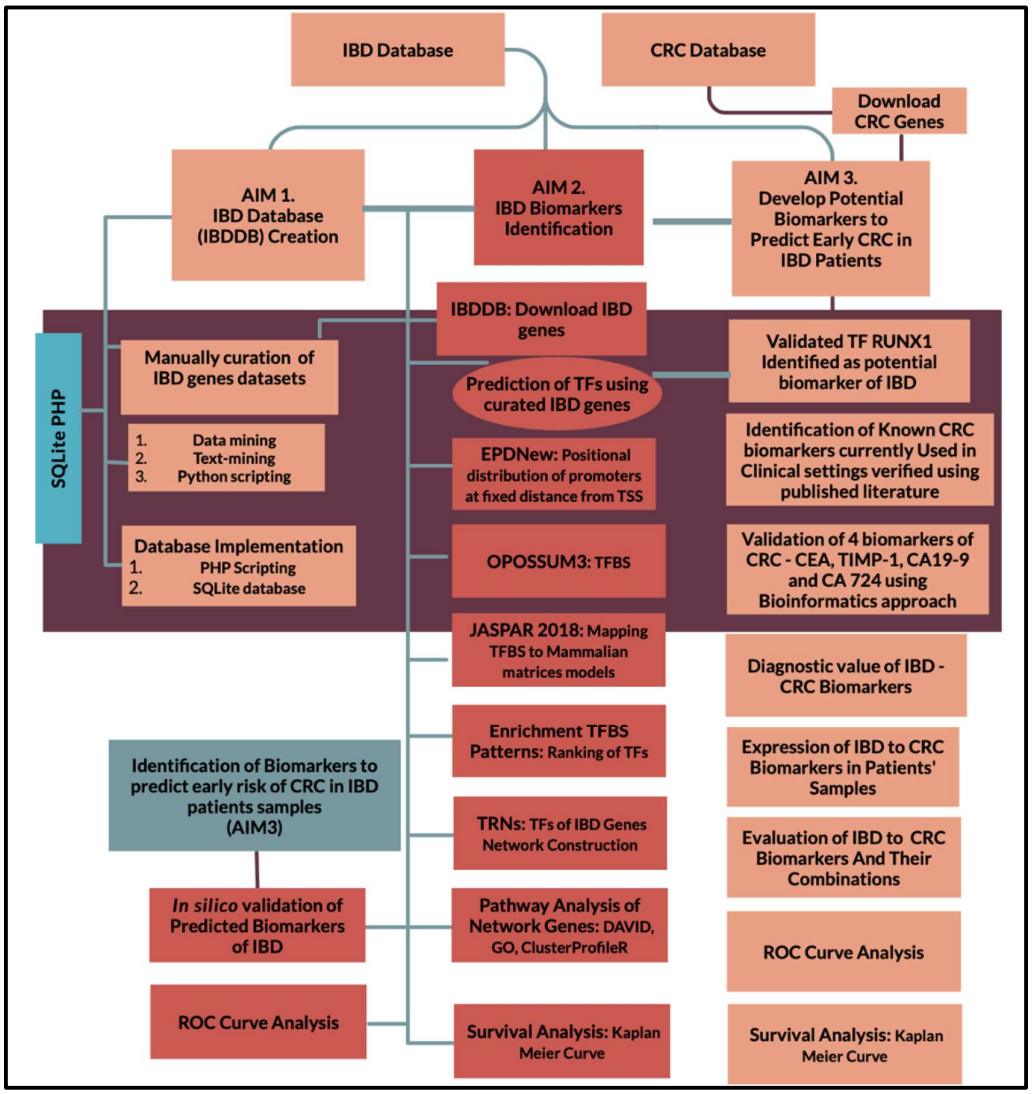


Figure 3. 1 Schematic representation of the methodology used in the research plan of this study.

3.2 Database Creation

We manually curated 289 disease genes linked to IBD. PubMed was used to retrieve IBD genes https://pubmed.ncbi.nlm.nih.gov/, that were found to be experimentally confirmed in the literature. The dataset columns and rows of 289 IBD genes were sorted using python scripting language (Appendix B section 8.1.2). From January 2019 to December 2020, I executed systematic literature search using search queries such as ((“inflammatory bowel disease” OR “IBD”)) OR (“UC” OR “ulcerative colitis”)) OR ((“CD” OR “crohn’s disease OR crohn disease”)) AND ((human OR mice OR mouse OR rat)). PubMed was searched for all published cohort records on solely experimentally validated IBD genes. The criteria for exclusion of genes from the study were those that were not experimentally verified under laboratory settings. We also administered the published literature search using well known Dragon Knowledge Exploration System (DES) knowledgebase (Maqungo et al., 2010; Salhi et al., 2016) organized by our collaborator late Prof Vladimir B. Bajic from the Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Saudi Arabia http://www.cbrc.kaust.edu.sa/des_evolution/.

3.2.1 Text mining

Text mining is the automatic extraction of information from a variety of written materials by a computer to discover new, previously unknown information. Text mining differs from traditional data mining in that patterns are derived from natural language text rather than structured databases of facts in text mining (Hearst, 2003). The amount of published biomedical research, and hence the underlying biomedical knowledgebase, is growing at a rapid pace. Text mining and knowledge extraction are two technologies that might help academics deal with this information overload (Cohen and Hersh, 2005). The text mining module is based on 3 tiers (layers) Data tier, logic tier and presentation tier. Text-mining was done using DES (Salhi et al., 2017) and PolySearch 2.0 (Liu et al., 2015).

3.2.2 Curation of IBD genes and data mining

Data mining followed the curation of IBD genes. Several external databases were employed to populate the IBD database with diverse information connected with IBD genes. It was done manually using some bioinformatics tools and external databases. These tools are listed in (Table 3.1) along with their references and URL.

Table 3. 1 Population of database columns, along with the references and URL.

Database column name	Source URL	References
HGNC_gene_symbol	www.genenames.org	(Stelzer et al., 2016b)
HGNC_gene_name	www.genenames.org	
Chromosome_location	https://amp.pharm.mssm.edu/Enrichr/enr# http://hgdownload.cse.ucsc.edu/downloads.html	(Chen et al., 2013)
Snp_variants	https://varelect.genecards.org/	(Stelzer et al., 2016a)
IBD_phenotypes	www.genenames.org	(Tweedie et al., 2021)
Previous_name	geneanalytics.genecards.org	
Omim_id	www.genecards.org	
Omim_diseases	https://amp.pharm.mssm.edu/Enrichr/enrich#	(Chen et al., 2013)
Uniprotkb_id	www.genecards.org	(Tweedie et al., 2021)
Entrez_id	https://david.ncifcrf.gov/	(Dennis et al., 2003)
HGNC_id	www.genecards.org; www.genenames.org	(Stelzer et al., 2016b)
Ensembl_id	www.genecards.org	
Refseq_DNA_sequence	http://genecards.weizmann.ac.il/geneloc	
Pmid	https://pubmed.ncbi.nlm.nih.gov/	(Fiorini et al., 2017)
Pmcid	https://www.ncbi.nlm.nih.gov/pmc	(Fiorini et al., 2017)
Experimental_evidences	www.genecards.org	(Stelzer et al., 2016b)
Literature_study_subject	https://pubmed.ncbi.nlm.nih.gov/	(Fiorini et al., 2017)
Up-down Regulation	https://pubmed.ncbi.nlm.nih.gov/	
Inflamed_sites	www.genecards.org	(Stelzer et al., 2016a)
Tissues_samples	www.genecards.org; https://pubmed.ncbi.nlm.nih.gov/	(Fiorini et al., 2017)
Cell_lines	https://pubmed.ncbi.nlm.nih.gov/	(Fiorini et al., 2017)
Biological_process	https://david.ncifcrf.gov/	(Dennis et al., 2003)
Cellular_process	https://david.ncifcrf.gov/	
Molecular_function	https://david.ncifcrf.gov/	
Kegg_pathways	https://david.ncifcrf.gov/	
CONTINUED		

Database column name	Source URL	References
Reactome_pathways	https://amp.pharm.mssm.edu/Enrichr/enr#	(Chen et al., 2013; Yoo et al., 2015)
DGIdb_interactiontypes	http://www.dgidb.org	
Dsigdb	tanlab.ucdenver.edu/DsigDB/DSigDBv1.0/	
TFs_transfac	https://amp.pharm.mssm.edu/Enrichr/enrich#	
TFs_chea	https://amp.pharm.mssm.edu/Enrichr/enrich#	
TFs_encode	https://amp.pharm.mssm.edu/Enrichr/enrich#	
TFs_OPOSSUM	https://amp.pharm.mssm.edu/Enrichr/enrich#	

3.3 Identification of the biomarkers of IBD genes and CRC genes, and their *in-silico* validation

To understand the molecular mechanisms of gene regulation can help researchers identify essential genes implicated in IBD that can be used as prognostic and suggestive biomarkers. Although we know that gene expression is linked to a variety of diseases and that TFs regulate many genes, this could lead to the discovery of new biomarkers of IBD that could be valuable in our research. As a result, rather than looking at individual gene expression, we looked for TF-based biomarkers. We began the bioinformatics analysis at first by identifying the IBD genes that were proven to be experimentally verified using PubMed literature and then constructed the database of IBD genes. We then downloaded these genes from the database and with the help of various bioinformatics tools and TFs prediction and mapping analysis was done. At first, we analyzed the promoter sets for the IBD genes. Then these promoters were screened for transcription factor binding sites (TFBS) to provide a link between genes and TFs that regulates them. TFBS of the promoter sequences were then mapped to mammalian matrix models position frequency matrices (PFMs). Further enrichment of these mapped TFBS were positioned as ranks based on their over-representation index (ORI) and this abridged the gap between the expression of genes and their regulatory TFs. These top ranked TFs of IBD clustered genes based on their promoters helped in the creation of largest TRNs of TFs of IBD biomarkers. The pathway analysis and gene ontology of these IBD genes helped in understanding the aetiology underlying the molecular mechanism of IBD. The various bioinformatics tools used in this analysis along with their URL and

data collection date are listed in (Table 3.2). We predicted progression of CRC from IBD by evaluating the known biomarkers of CRC currently used in clinical settings and then combined with the predicted biomarkers of IBD. Combinations of the individual tests from IBD and CRC biomarkers together with their diagnostic value was quite appealing. We even scrutinized the expression value of these biomarkers using gene expression atlas (Papatheodorou et al., 2019). In this thesis we will be discussing the applications of bioinformatics for the identification and evaluation of biomarkers of IBD and the biomarkers of CRC progressing from IBD.

Table 3. 2 List of bioinformatics tools used in the analysis, along with the references and URL.

Bioinformatics tools	Description	Source URL	References
EPDnew database	Promoter sets and location	https://epd.epfl.ch//index.php	(Dreos et al., 2015)
OPOSSUM 3	TFBS	http://OPOSSUM.cis-reg.ca/OPOSSUM3/	(Ho Sui et al., 2005)
CORECG	CRC database of 2056 genes	https://lms.snu.edu.in/corecg/	(Agarwal et al., 2016)
JASPAR	Matrix ID and Position Frequency Matrices (PFMs)	https://jaspar2018.genereg.net/	(Khan et al., 2018)
Position Weight Matrices	Matrix ID PWMs	https://ccg.epfl.ch/pwmscan/	(Ambrosini et al., 2018)
CYTOSCAPE	Network construction	https://cytoscape.org/	(Shannon et al., 2003)
Cluster Profiler	Functional and Pathway analysis gene ontology	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html	(Yu et al., 2012)
GEPIA2	Expression of biomarkers and Survival analysis	http://gepia2.cancer-pku.cn/#index	(Tang et al., 2019)
CONTINUED			

CONTINUED			
Bioinformatics tools	**Description**	**Source URL**	**References**
ONCOMINE	*In silico* validation of biomarkers of IBD	https://www.oncomine.org/resource/login.html	(Rhodes et al., 2004)
EBMTools	ROC curve	https://ebm-tools.knowledgetranslation.net/calculator/diagnostic/	(Shurtz and Foster, 2011)
Gene Expression Atlas	Expression of biomarkers	https://www.ebi.ac.uk/gxa/home	(Papatheodorou et al., 2019)

3.3.1 Promoter sets

The experimentally validated 289 IBD genes were downloaded from IBDDB database and were analyzed in Eukaryotic promoter database (EPDnew) http://epd.vital-it.ch. EPD is an out-of-date database of eukaryotic POL II promoters, also known as transcription start sites (TSS) or regions, that have been experimentally characterized. EPDnew is a database of promoter lists for different organisms. that complements EPD's older corpus of hand compiled promoter information (Dreos *et al.*, 2015). After accessing EPDnew, selected select/download tab, here selected 'refine selection' of *Homo sapiens* promoters were selected from -1000 to +200 (upstream and downstream from the TSS as FASTA format parameters that were downloaded. Further selected OProf tool https://ccg.epfl.ch/ssa/oprof.php, under signal search analysis server (this tool generates motif occurrence profile around TSS positions and also generates the plot displaying the occurrence frequency of TATA boxes between -100 to +100 relative to the TSS).

3.3.2 Transcription Factor Binding Sites (TFBSs)

The bioinformatics approach used to predict binding site in a transcription factor is OPOSSUM version 3.0 http://opossum.cisreg.ca/opossum3/ to perform single-site analysis. This application identified over-representative TFs and its TFBS. The parameters selected were [-1000, +200] upstream and downstream regions in nucleotide. The putative TFs selection was based on one-tailed fisher exact probability (score ≥ 7) score (Fisher, 1922). Fisher's exact test is a prominent statistical significance

test for determining the importance of overlap between two sets of data (Fisher, 1922). OPOSSUM integrates a pre-computed database of conserved TFBSs in human and mouse promoters with statistical approaches to find sites that are over-represented in a group of co-expressed genes. Three transcript profiling investigations revealed that the algorithm successfully identified mediating TFs in control sets of tissue-specific genes and sets of co-expressed genes (Ho Sui et al., 2005; Milanesi et al., 2021). The co-ordinated binding of trans-acting TFs to innumerable binding sites in the region which includes a gene's TSS, as well as to enhancer's locations that initiate gene activation from distal locations, was required for transcriptional regulation of gene expression. Position specific scoring matrices (PSSM) are commonly used to represent TF binding specificities to their homologous DNA binding motifs.

3.3.3 Mapping of TFBSs

By binding to gene promoter regions, TFs control gene expression. TFs binds in two ways either to another TF or to the DNA binding domain. The PWM, also known as the PSSM, score > 85 % for predicted binding sites for vertebrates in JASPAR database contains the log-odds or log-probability weights for calculating the binding affinity score and is one of the most widely used models (Boeva, 2016). Here OPOSSUM 3 web-based database retrieved information to map TFBS to all mammalian matrices models like PWM and PFMs from JASPAR 2018 database (Khan et al., 2018).

3.3.4 Ranking of TFs

Eukaryotic cells respond to growth, development, and environmental cues in large part by regulating the expression of specific sets of genes (Izdebska et al., 2018). One way for measuring or ranking the significance of TFs in this activity is the ratio of the genes of interest hits, first by TF to the total number of targets in the sequence (Hestand et al., 2008). Fisher's exact test is a popular statistical significance test for determining the importance of overlap between two sets. It's an exact test in which the significance of the deviance from the null hypothesis is measured rather than predicted using Chi-squared analogue (Fisher, 1922). OPOSSUM 3.0 examines the significance of an over-represented group of genes of interest in the direct targets of TF using Fisher's exact test (Ho Sui et al., 2005; Milanesi et al., 2021). OPOSSUM-3 is a web-accessible software for identification of over-represented TFBS and families in either DNA sequences of genes or sequences generated from high-throughput methods as ChIP-Seq. it works on two complementary scoring methods 1) Z-scores is based on binomial distribution that measures change in relative number of TFBS motifs in the foreground gene sets

compared to background genes. 2) Fisher scores is based on one-tailed fisher exact probability assessing the number of genes with TFBS motifs in the foreground vs background gene sets. Here we opted for fisher exact probability test which should be greater or equal to 7 (Ho Sui et al., 2005).

3.3.5 Network construction

TRN's can help us better understand the molecular underpinnings of development and cellular reprogramming, leading to more effective ways for generating different cell types for regenerative therapies (He and Tan, 2016). The network construction of IBD genes was performed using Cytoscape version 3.8.2 (Shannon et al., 2003). The plug-in used in Cytoscape was ReNE, to determine the hub of IBD genes controlled by the TFs. ReNE is a plugin program that automatically enriches a standard gene-based regulatory network with more detailed transcriptional, post-transcriptional, and translational data, culminating in a network that more accurately reflects biological regulatory mechanisms (Politano et al., 2014). ReNE plug-in was installed, and a tabulated data was uploaded as a .txt file format in Cytoscape and the source and target were selected as nodes and edges constructing the data into a regulatory network.

3.3.6 Functional annotation

The putative verified IBD genes were classified based on their annotated roles in IBD using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003) and Kyoto Encyclopaedia of Genes and Genomes (KEGG), while extracted using R package Cluster profiler. HUGO Gene Nomenclature Committee (HGNC) verified gene symbols were uploaded (Bruford et al., 2021). The official gene symbol was chosen as the unique identification, and the submit list button was chosen for the gene list. From the output, functional annotation clustering was chosen, and distinct cancer-related processes were chosen from the various clusters. These biochemical processes were compared to the genes that had been annotated.

3.3.7 *In silico* validation of transcription factors

The expression of seven putative TFs potentially regulating IBD genes were accessed using Gene Expression Atlas database (Papatheodorou et al., 2019), queried by typing HGNC gene name of TFs in the gene box, Homo sapiens was selected as the organism and condition IBD was selected for the up/down expression. For the expression of known biomarkers of CRC, condition selected was CRC

disease. Here the expression level was shown as heat map on the basis of log 2-fold change expression level as up and down-regulation in IBD, UC and CD patients' samples.

3.3.8 *In silico* validation of biomarkers

In silico approach was used to identify the validation of biomarkers implicated in the early onset of disease or progression of CRC from IBD disease through mining of biological databases like ONCOMINE https://www.oncomine.org/resource/login.html. ONCOMINE is a web based data-mining cancer microarray database (Rhodes et al., 2004). Here in the concept page, selected 'Analysis type' as Cancer vs Normal analysis, and under drop down menu from' Cancer type' selected Colorectal cancer. And under various author's published literature name-based CRC, adenomas or colon and rectal cancer vs normal, evaluated the TP, FP, TN, and FN diagnostic values for the various IBD TFs and known cancer biomarkers' values under the genes box provided along with the other parameters.

3.3.9 Gene expression profiling and Survival analysis

For assessing gene expression between normal and tumor tissues in box plots, the Cancer Genome Atlas Program (TCGA) Colon Adenocarcinoma (COAD) https://portal.gdc.cancer.gov/projects/TCGA-COAD and Rectal Adenocarcinoma (READ) datasets https://portal.gdc.cancer.gov/projects/TCGA-READ were compared with the TCGA normal tissue database and the Genotype-Tissue Expression database (GTEx). The box whisker plots were created through GEPIA 2 (Gene Expression Profiling Interactive Analysis) an expression DIY tool http://gepia2.cancer-pku.cn/#analysis. Initially the expression data log2(TPM+1) were transformed and the median of tumor and normal tissue was designated as log2FC. One-way ANOVA method was used for differential analysis *p-value Cutoff ≤ 0.01.

The Overall survival (OS) and Disease-Free Survival (DFS) plots were generated using COAD-READ datasets with survival plot tool by GEPIA2 web tool. Gene signature was selected for comparison of biomarkers. The Mantel-Cox test was used to divide the high-expression (50%) and low-expression (50%) in cohorts, along with the hazard ratio included and 95 % confidence interval in the analysis.

3.3.10 Statistical evaluation of biomarkers

In this thesis the application of statistical principles is implemented through various diagnostic tests in predicting the correct accuracy of the biomarkers. The first stage in comparing diagnostic tests to one another or to a previously specified 'gold standard' test is to create the study protocol, which involves

deciding on the study design (Baveja and Aggarwal, 2017). Receiver operating characteristic curve (ROC) analysis and statistical assessment of diagnosis are crucial for evaluating the success of screening treatments and statistical assessors, as well as assisting in the prediction of models and algorithms (Zou et al., 2011). The validity and accuracy of diagnostic tests were determined by several parameters when combined. Sensitivity and specificity are the two examples of such variables (Swift et al., 2020). The quantitative analysis of sensitivity and specificity tests are based on a threshold value above or below whereby the test is positive. The lesser the specificity, the greater the sensitivity, and conversely.

To comprehend the utility of clinical assays, certain terms should be understood first.

1. **True Positive:** When the test reports a positive result for a person who has a disease. This result is referred as true positive (TP).
2. **False Positive:** When the test reports a positive result for a person who is disease free. This result is referred as false positive (FP).
3. **True Negative:** When the test reports a negative result for a person who is disease free. This result is referred as true negative (TN).
4. **False Negative:** When the test reports a negative result for a person who actually have a disease. This result is referred as false negative (FN).

The various parameters of statistical assessment that are significant for all the biomarkers evaluations are discussed below.

3.3.10.1 ***Sensitivity:*** It is a clinical test's ability to correctly identify patients with a disease. The sensitivity of a screening test can be defined in a variety of ways, such as the diagnostic test's ability to identify a true positive, being based on the true positive rate (TPR), representing a test's ability to accurately classify all people who have a condition, or 100 percent of all people with an affliction of involvement by those who test positive on the test (Trevethan, 2017).

Formula: Sensitivity =

$$\frac{True\ Positive}{True\ Positive + False\ Negative}$$

3.3.10.2 Specificity: It is a clinical test's ability to correctly identify patients without any disease. The sensitivity of a screening test can be defined as the diagnostic test with a specificity of 100 percent accurately detects all patients who are without any condition but a test with a specificity of 80% accurately reports the patients without disease that is test negative or TN, leaving to 20% patients without any disease wrongly diagnosed as test positive or FP (Lalkhen and McCluskey, 2008).

Formula: Specificity =

$$\frac{True\ Negative}{True\ Negative + False\ Positive}$$

3.3.10.3 Positive predictive value: When a clinical test indicates that the patient has the disease of interest, it is the probability that the disease is present or that the test is positive. The operating parameters of an assay are defined by sensitivity and specificity; however, it is the predictive value (positive or negative) of the analysis that is generally of diagnostic interest to clinicians and patients (Saah and Hoover, 1997).

Formula: Positive predictive value =

$$\frac{True\ Positive}{True\ Positive + False\ Positive}$$

3.3.10.4 Negative predictive value: It is defined as the probability that the disease is not present when the test is negative (Saah and Hoover, 1997).

Formula: Negative predictive value =

$$\frac{True\ Negative}{True\ Negative + False\ Negative}$$

3.3.10.5 ***Likelihood ratio:*** The likelihood ratio (LR) is the chance of a given test result among persons who have an illness divided by the probability of the same test result among people who do not have the disease (Grimes and Schulz, 2005). The LR is used to determine the efficacy of a diagnostic test and to aid in the choice of the best diagnostic tests and its sequence. LR has a significant advantage over sensitivity and specificity in that it is less likely to alter with disorder prevalence and may be used to aggregate the findings of multiple diagnostic tests to assess the likelihood of a target disorder being diagnosed. (Hayden and Brown, 1999). Tests having a Positive LR more than 10 (or a Negative LR less than 0.1) are normally considered to have the ability to change clinical decision made by experts.

- **Positive Likelihood ratio:** The Positive Likelihood Ratio (LR+) is the ratio of the probability of a positive test result in the presence of disease to the probability of a positive test result in the absence of disease (Shreffler and Huecker, 2020). More pragmatically, LR+ Ranges from 1 – infinity.

Formula: Positive Likelihood Ratio =

$$LR+ = \frac{Sensitivity}{1 - Specificity}$$

- **Negative Likelihood ratio**: The Negative Likelihood Ratio (LR-) is defined as the ratio of the chance of a negative test outcome when the disease is present versus probability of a negative test outcome when the disease is absent (Shreffler and Huecker, 2020).

Formula: Negative Likelihood Ratio =

$$LR- = \frac{1 - Sensitivity}{Specificity}$$

3.3.10.6 Bayesian Theorem: The Bayes' Theorem is a probability theorem that can be used to explain how a new piece of information affects the chance that a hypothesis is true (Ingelfinger, 1994).

- **Pre-test probability**: Pre-test probability is equal to the prevalence of disease in the population if a specific population of patients is being evaluated (Agoritsas et al., 2011).

Formula: Pre-test Probability =

$$LR+ = \frac{Sensitivity}{1 - Specificity} \qquad LR- = \frac{1 - Sensitivity}{Specificity}$$

Pre-test probability =

$$\frac{1 - Sensitivity}{Sensitivity}$$

- **Post-test probability**: Posterior test Probability is the likelihood that the patient may develop an illness once the test findings have been obtained (Agoritsas et al., 2011). There are two post-test probability, positive for population that has disease and negative post-test probability for ones that don't have disease.

Formula: Post-test probability

Positive Post-test Probability =

= (Pre-test Probability) x (Positive Likelihood ratio)

Negative Post-test Probability =

= (Pre-test Probability) x (Negative Likelihood ratio)

3.3.10.7 Accuracy of diagnostic test: It measures the ability of a test to detect a condition when it is present or absent. It should be noted that the prevalence of disease impacts on the accuracy of the test (Šimundić, 2009).

Formula: Accuracy of a diagnostic test =

$$= \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negataive + True\ Negative}$$

3.3.10.8 Prevalence: The percentage of positive outcomes obtained from the administration of a clinical diagnosis to a random selection of persons taken from the population of interest is the conventional estimate of prevalence. The prevalence of an illness refers to the percentage of people in a population who have the ailment in question (Diggle, 2011).

Formula: Prevalence of a diagnostic test =

$$= \frac{True\ Positive + False\ Positive}{True\ Positive + False\ Positive + False\ Negataive + True\ Negative}$$

3.3.10.9 Youden Index: The Youden index is the difference between the test's diagnostic performance and the strongest overall performance often referred to as guilt, which is defined as a utility loss due to ambiguity about the true condition (Šimundić, 2009).

Formula: Youden Index =

$$= Sensitivity + (Specificity - 1)$$

3.3.10.10 ***Contingency 2 X 2 matrix table:*** The presence or absence of a disease must be established by an impartial gold standard before sensitivity and specificity may be calculated (columns in the 2 X 2 table). After the data has been collated in a matrix table (Table 3.3), the sensitivity represents the patient population with positive test results among those who are diseased. The specificity relates to the proportion of people who test negative but are not diseased (Fischer et al., 2003).

Table 3. 3 This table illustrates the input values for the diagnostic test evaluation in 2x2 contingency matrix table.

	Disease	**Non-Disease**	**Total**
Test Positive (+)	**TP** *True Positive*	**FP** *False Positive*	= Total Positive Patients
Test Negative (-)	**FN** *False Negative*	**TN** *True Negative*	= Total Negative Patients
Total	Total Disease Patients =	Total Non-Disease Patients =	

3.3.10.11 Receiver operating characteristics curve (ROC_{AUC})

The receiver operating characteristics curve (ROC) was created for the first time in 1950 (Markowitz and Swets, 1967). It was first used in the detection of signal from noise in radar during 1966 (Green and Swets, 1966). The partial area under the ROC, which provides a nonparametric estimator based on an interpretation of the partial area under the curve (AUC), is a measure of diagnostic test accuracy (Dodd and Pepe, 2003). It was first used in the detection of signal from noise in radar during 1966 (Green and Swets, 1966). The concept of ROC curve has lately been applied to several therapeutic domains that rely largely on monitoring and diagnostic tests. ROC curve is a valuable method for evaluating diagnostic test performance and more broadly, the correctness of a statistical model (e.g., linear regression, principal component analysis) that categorizes people into two groups that is diseased and healthy. The unique performance of ROC curve analysis is a visualization tool for illustrating the accuracy of a medical diagnostic test (Zou et al., 2007). ROC curve graphically represents the sensitivity also known as TPR and 1 – specificity as false positive rate (FPR) of a biomarker diagnostic test on y – axis and on x – axis, respectively (Fig 3.2). A ROC curve that rests on the diagonal line reflects the effectiveness of a diagnostic test that is no better than a probability, i.e., a test that provides positive or negative results that are unrelated to the underlying disease status. The slope of a ROC curve at every juncture is equivalent to the ratio of both the two functionals that are sensitivity and 1 – specificity characterizing the distribution of variables in the diseased and control populations, respectively (Hajian-Tilaki, 2013).

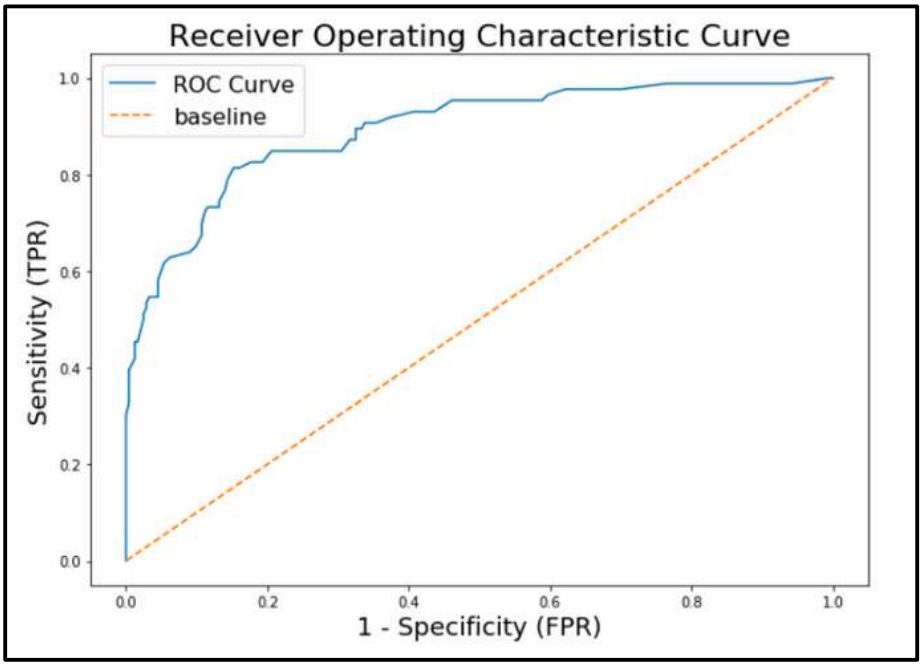


Figure 3. 2 ROC curve illustrates the sensitivity and 1 – specificity plotted on a graph across varied cut-offs. Adapted from https://towardsdatascience.com/roc-curve-in-machine-learning-fea29b14d133.

3.3.10.12 *Thumb rule for diagnostic evaluation value*

Tests with a positive LR greater than 10 (or a negative LR less than 0.1) are normally thought to have the potential to change clinical decisions, according to experts. Although using precise interpretation rules may be enticing as shown in (Table 3.4), scientists must first assess the clinical circumstance to determine what level of increased is clinically important for better treatment planning. It is stated that without a strong LR, no change in prediction is conceivable. Firstly, when an LR gives no additional information (e.g., LR =1.0), as one might assume the pre-test probability matches the post-test probability. Secondly, it is noted as one might assume that pre-test probability might greatly influenced by using a very high predictive biomarker (e.g., LR = 10). Pre-test probabilities that are extremely high (or extremely low) result in smaller adjusted expectations than those that are less extreme (Ray et al., 2010).

Table 3. 4 Rule of Thumb: Correspondence between Accuracy, LR+ and LR- and Area under the Receiver Operating Characteristics Curve (AUC_{ROC}) and the Diagnostic Value of a Biomarker.

	Accuracy	**LR+**	**LR-**	**$\mathbf{AUC_{ROC}}$**
Excellent diagnostic value	> 0.90	>10	< 0.1	> 0.90
Good diagnostic value	0.75 – 0.90	5 – 10	0.1– 0.2	0.75 – 0.90
Poor diagnostic value	0.50 – 0.75	1 – 5	0.2– 1	0.50 – 0.75
No diagnostic value	0.50	1	1	0.50

3.3.10.13 Combinations of multiple biomarkers

In everyday clinical practice, however, we can integrate the results of two independent tests to improve the accuracy of the diagnosis. The diagnostic methods SNOUT and SPIN are related to specificity and sensitivity and these acronyms SNOUT and SPIN stand for "Sensitive test if Negative rules OUT the disease" and "Specific test when Positive rules IN the disease", respectively (Baeyens et al., 2019). Here we have used the formula to combine several independent tests of biomarkers to improve the accuracy of their predictive diagnostic values, by improving low sensitivity and specificity test values by increasing them (Parikh et al., 2008).

Formula: combination of multiple diagnostic test =

For Sensitivity Test =

$$1 - (1 - SEN_{BIOMARKER\ 1})\ X\ (1 - SEN_{BIOMARKER\ 2})\ X\ (1 - SEN_{BIOMARKER\ 3})$$

For Specificity Test =

$$1 - (1 - SPEC_{BIOMARKER\ 1})\ X\ (1 - SPEC_{BIOMARKER\ 2})\ X\ (1 - SPEC_{BIOMARKER\ 3})$$

3.3.10.14 Confidence Interval

Biomedical research is typically conducted on samples rather than entire populations. A sample statistic and sample estimates of error can be used to gain a good understanding of the population parameter as a range of values rather than a single value. This is the confidence interval (CI), which is calculated based on a desired level of confidence (Hazra, 2017). CI also can be used to evaluate if the means of two samples are statistically different. It is predicted that 95% confidence interval is the most prevalent

in biomedical research, a CI can be constructed for any confidence level. Whereas, a 99% CI will be wider than a 95% CI for the same sample (Choudhary and Garg, 2013).

4 CHAPTER FOUR – RESULTS

The main objective of this study to Identify transcription regulators of IBD-associated genes and I used these TFs as biomarkers for the diagnosis and screening of early onset of IBD disease by utilising computational approaches. I also identified the biomarkers of CRC and combined them with IBD biomarkers to predict progression of CRC from IBD. For this it was important to understand the molecular mechanisms underlying the aetiology of IBD. Therefore, this led me to develop the database of manually curated IBD genes, as till date there is no database available for IBD genes. For the curation of genes and associated concepts, I used bioinformatics analysis for the data mining and for biomarkers predictions.

4.1 AIM 1

4.1.1 Creation of a curated database of IBD genes

The literature survey revealed that not a single comprehensive database existed that could be used to mine IBD related genes, therefore, it was a necessity to develop an IBD database that is informative, has usable format and user-friendly. This led to my first aim of the project. I manually curated datasets of 289 IBD genes which were proven to be experimentally validated under laboratory settings and were confirmed in human or mice samples. PubMed (PMID) and PubMed Central (PMCID) was used to search published literature on IBD-linked genes that were validated experimentally. The criteria for the exclusion of genes were based on genes that were not analyzed under laboratory settings or those genes which were recognized solely using computational analysis. Here I have explained the IBD database along with the screenshots taken from its various features. IBD Database (IBDDB) https://www.cbrc.kaust.edu.sa/ibd/ (Khan et al., 2021), lands a user at the front page that contains several tabs like home button, explore tab, visualize tab, about button on the top-left corner and user manual guide on the top-right corner (Fig 4.1).

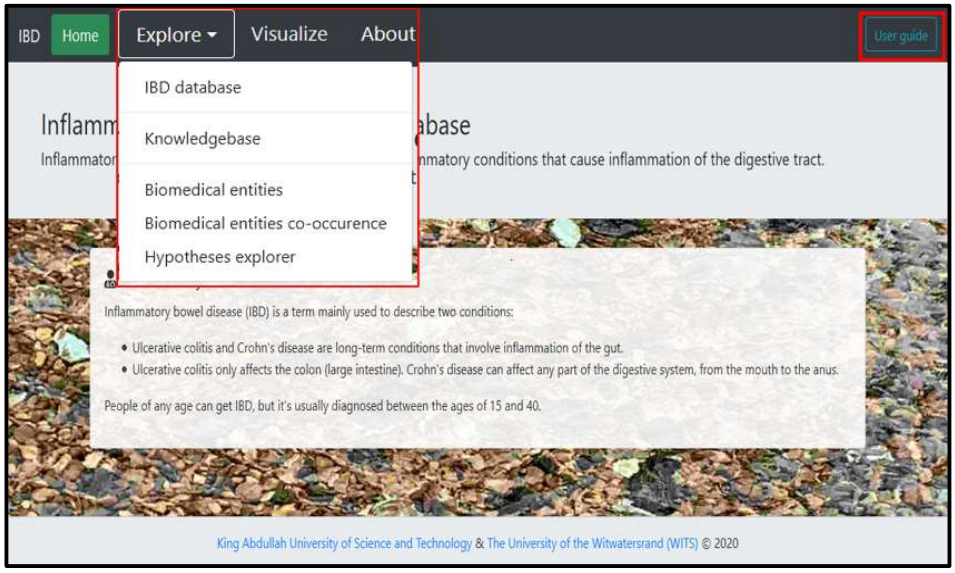


Figure 4. 1 This figure illustrates the different paradigm available under Explore tab of the IBD database (IBDDB) along with visualize and user guide.

Different types of explorable modalities are available under dropdown menu of 'Explore tab' like IBD database, Knowledgebase, Biomedical entities, Biomedical entities co-occurrence and Hypotheses explorer, which are explained in the following sections.

4.1.1.1 IBD database

The IBD database menu under the explore tab displays the manually curated 289 IBD-linked genes that were experimentally verified and were tested experimentally in human and animals' samples in wet laboratory settings using low-throughput investigations like western blot, ELISA, RT-qPCR, immunohistochemistry, etc. Tissue microarray genes and high-throughput RNA-Seq data were included if they were authenticated using any of the low-throughput approaches indicated above and demonstrated to have significantly different expression from healthy samples. The database includes contains single nucleotide polymorphisms (SNPs) and variations of IBD-associated genes. HGNC verified the symbols used for the genes (Tweedie et al., 2021). We then augmented the data by precompiling the data- mining facts and information linked to IBD cited as "concepts". There are 34 columns of concepts comprising various genomic IDs along with TFs from different databases, drug interactions, chromosomal locations, gene ontologies, pathways, fold change, drug signature database (Dsigdb) (Yoo et al., 2015). In the (Fig 4.2) the first table demonstrate the various concepts embedded in the database. Users can press tabs on any of the concept and it will be displayed in the table below. The second table illustrated the HGNC gene symbol, gene name, IBD phenotypes, PubMed ID, experimental evidence

which shows low through-put techniques and study organisms like humans, mice study. It also shows the show/hide tab and download formats, copy, or print tab.

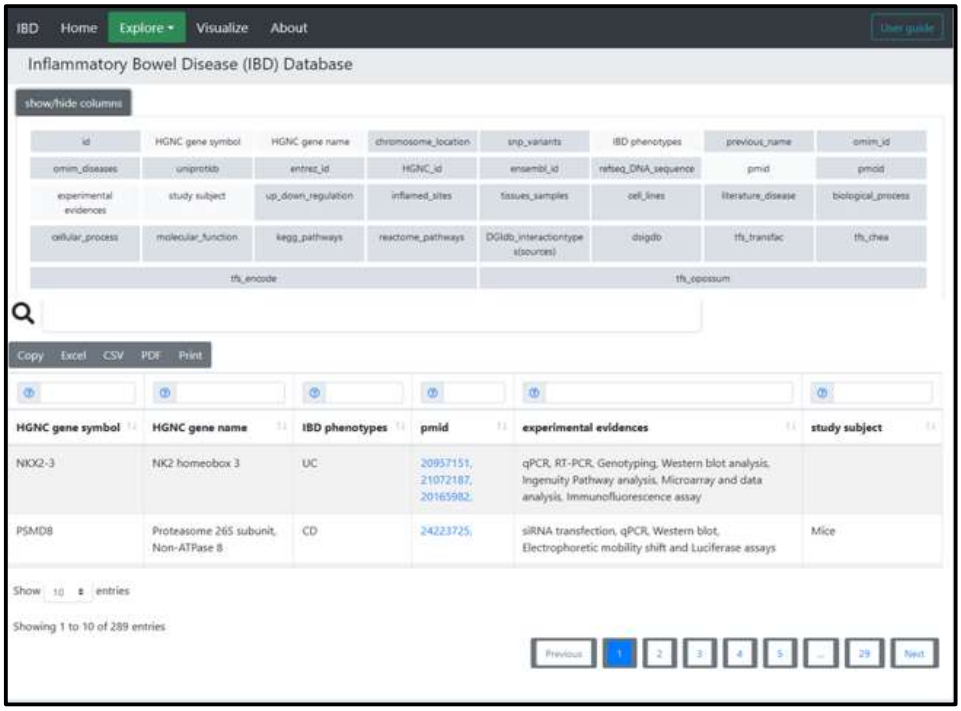


Figure 4. 2 This figure illustrates the tables containing various column names under Explore tab 'IBD Database'.

4.1.1.2 Knowledgebase

The existing database systems in general, as well as the literature on the subject, provide the basis for the knowledge base (KB) for development. In order to identify efficient and sufficient evaluation methods, existing literature on the system must be researched. IBDDB delivers merged data-mined information from a variety of external data sources, as well as text mined information using the KAUST University's Dragon Exploration System (DES) system. Two major components of the DES system are dictionaries and PubMed records. Dictionaries are controlled vocabularies with concepts from certain fields. DES creates a KB index by indexing titles and abstracts retrieved from PubMed against dictionaries. This process makes it possible to highlight concepts in sentences as well as count occurrences and co-occurrences (Fig 4.3). The indexed documents are highlighted on the left-side mentioned in the PubMed literature on the right-side. DES also generated 400 IBD related genes, of which 208 were manually curated and confirmed, while the remaining 81 genes were reviewed using PubMed literature. We also used the well-known text mining application PolySearch 2.0 to see if we were able to locate a sufficient number of relevant genes using DES (Liu et al., 2015). We, therefore, identified

112 genes using PolySearch 2.0, however only 15 of them were manually verified using literature search. Altogether, we managed to reach 289 IBD related genes that were experimentally confirmed using literature search from PubMed.

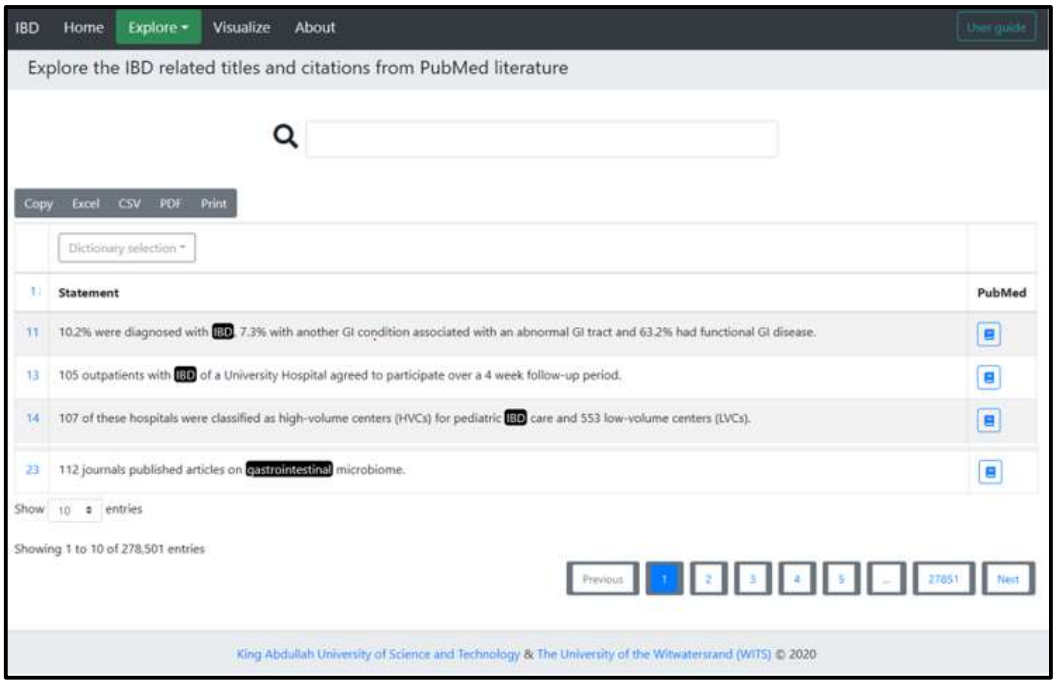


Figure 4. 3 This figure illustrates the Knowledgebase under Explore tab.

4.1.1.3 Biomedical Entities

Text mining was carried out on 106 333 abstracts pertinent to IBD that were retrieved from PubMed. The indexing engine used 11 dictionaries to index 2 22 726 biomedical terms. As a result of the approach 11 728 terms and about 3 000 000 co-occurring terms were generated. The terminology was standardized using a lexicon of synonyms and homophones or homographs. The various types of information in the database can be viewed via medical entities identified as a subject-specific set of documents in reference to their PubMed literature. These entities are terms, assembled into classified dictionaries (Fig 4.4). The concepts accommodated in the KB are the terms organized in the dictionaries, for example 'Disease' is a dictionary and its medical entity can be IBD, UC, CD, etc. The list of 11 dictionaries that were used, together with the number of terms in brackets are Biological action (94), Biological processes (1112), Cellular component (276), ChEBI (76 566), Diseases (15 878), Dietary supplements (2707), Toxins (8407), Metabolites and enzymes (61 127), Drugs (13 582), Path-

ways (3717), Human genes (42 061). Although different versions of the same concept can be discovered, the concepts are normalized to offer non-repetitive content. Hence dictionaries were formatted to erase duplicates, missing data and finally, parsed the indexed data into standardization consistency.

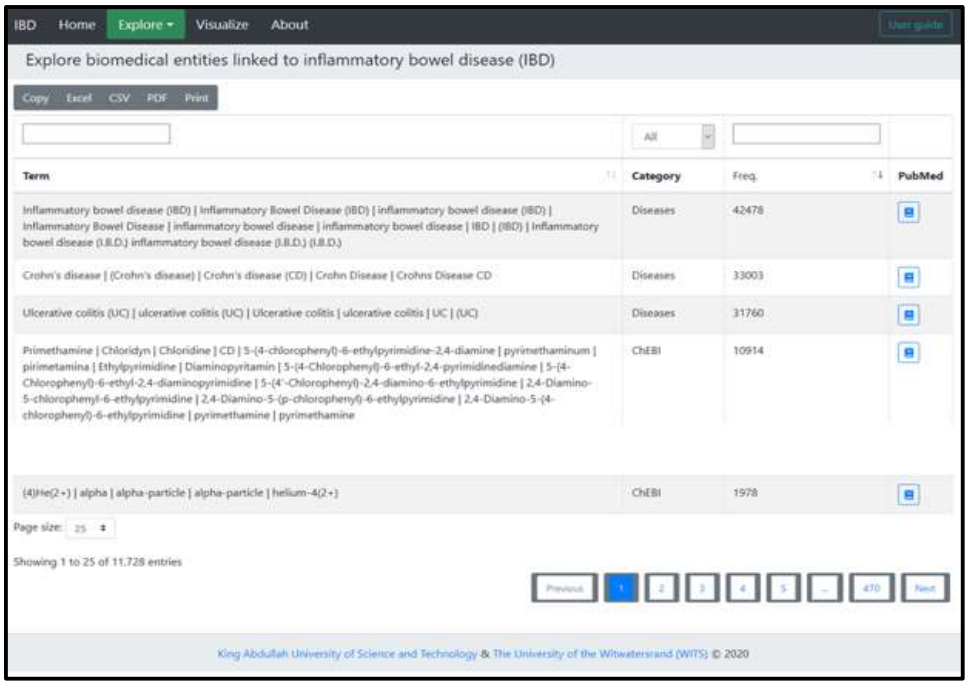


Figure 4. 4 This figure illustrated the Biomedical entities under Explore tab.

4.1.1.4 Text mining application in IBD database

On 106 333 IBD related abstracts obtained from PubMed, text mining was conducted, and the translated text was subsequently deployed into an SQLite database and indexed. From various features of this database, it also focused on text mining process to automate the indexing of the terms, parsing and normalization of the data. This text mining module is based on three-tier (layer) architecture as data tier, logic tier and presentation tier (Fig 4.5).

The data layer is based on SQLite database and when the offline processing is completed then text mining database is used as a data tier. In this tier the data or text relevant to IBD related abstracts are extracted from the PubMed and then these text data are parsed (string of text data converted from one format to another). These parsed data are then deployed to SQLite database and then are indexed using 11 dictionaries with 2 22 726 biomedical entities. This process generated an index of over 11000 terms and close to 3 00 000 co-occurring terms. The data tier contains dictionary table that is embedded with the list of 11 dictionaries names and the glossary table that contains

list of biomedical terms. Indexing engine applied these biomedical terms to the corpus table which contains PubMed records. The indexing module created two new tables by updating the corpus table: 'term table' that contains indices of terms and 'term-pair table' that contain indices of co-occurrence terms. Next tier is the application tier that does the logical thinking.

Logic tier is implemented as several server-side PHP modules interfaced with data layer. It served the ajax (Asynchronous JavaScript and XML) calls initiated from the user interface in the presentation tier. Ajax technique, a website development technique on the client-side that created asynchronous application, it uploads the parts of website pages without reloading the whole page. Here it contained various modules like entities, co-occurrences entities, knowledgebase, entities network and hypotheses explorer. It is hosted on the in-house server. It accesses the tables and combines them to present in the presentation tier.

Presentation tier is built on jQuery/Bootstrap framework and is mobile-friendly. It is compatible on both desktop and mobile as well as on any other devices. It is a website-based layer and is implemented on JavaScript Object Notation (JSON) format. Application programming interface (API) calls are used by the presentation tier to connect with the other tiers.

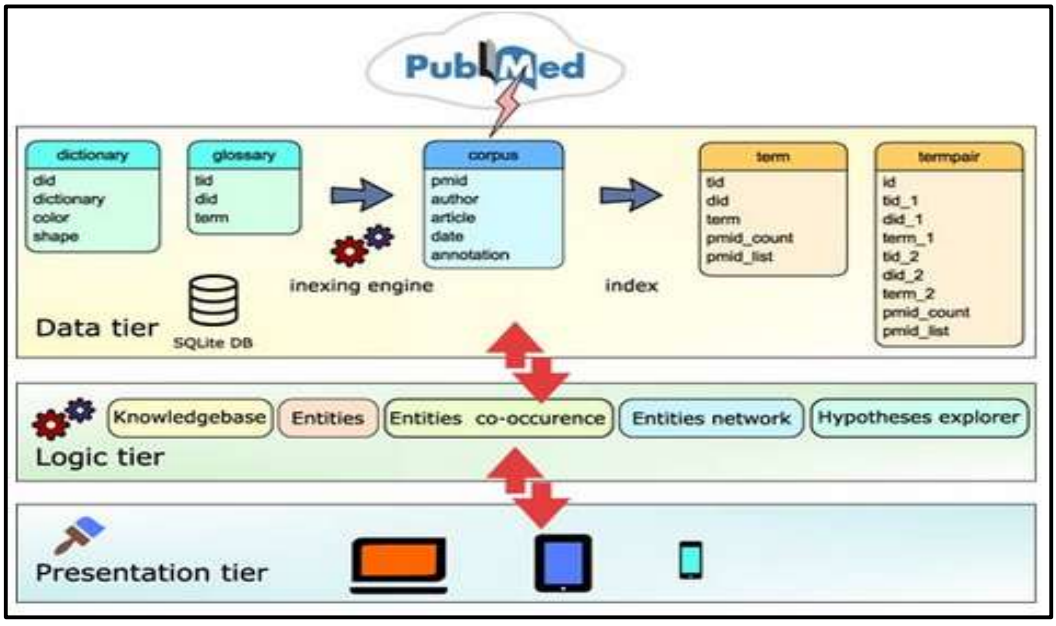


Figure 4. 5 The figure illustrates the text mining approach in IBD database. It showcases the three-tier data tier, logic tier and presentation tier (Khan et al., 2021). The Data layer is based on SQLite database and contains dictionary and glossary table. The Logic tier is implemented as several server-side PHP modules interfaced with data layer. The presentation tier communicates with other tiers by outputting

results to the browser and client tier and it displays information relating to services as browsing merchandise, purchasing, and shopping cart contents.

4.1.1.5 Hypotheses Explorer

One of the unique capabilities of IBDDB database is to generate novel hypotheses. It is based on the theory of linking proposed by Don Swanson where if term 'A' is related to term 'B' and term 'B' is linked to term 'C', then may be term 'A' can relate to term 'C' (Swanson, 1986). This theory has helped researchers in publishing literature accommodating several hypotheses in recent years (Huang et al., 2012). Our theory is based on previously published PubMed research linking lipid profiles to IBD, particularly low plasma levels cholesterol and high triglycerides. We wanted to learn more about the role of lipids (especially cholesterol) in IBD and see if they have an effect on IBD-related disorders like Pyoderma Gangrenosum or PG (Khan et al., 2021) (Fig 4.6). PG is a skin ulceration condition. We searched the term PG under term A using dictionary disease or medical subject headings (MESH) term 'disease' and in term B we selected IBD from dictionary and in term C we selected biological processes using dictionary and term as lipid metabolism. We thus created a novel hypothesis and from this hypothesis we emphasized that further exploration of the possibility of link between cholesterol, PG and IBD should be made. We have thus created a network based on this hypothesis and to our surprise we did not find any published literature available on this linkage.

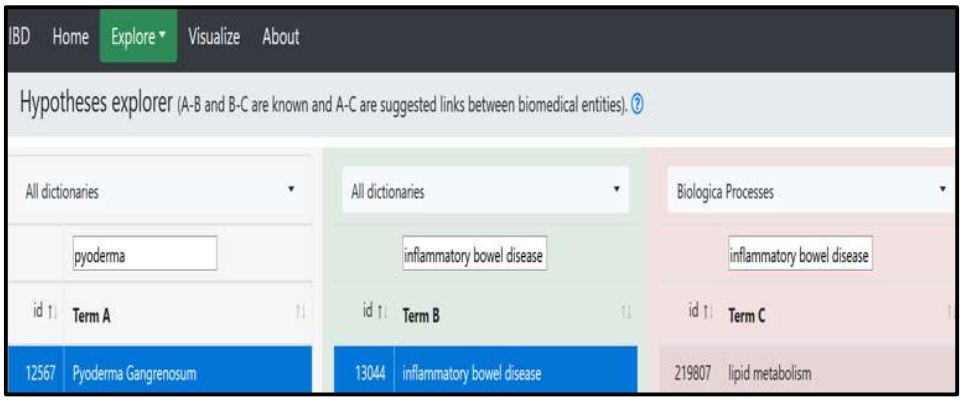


Figure 4. 6 The hypothesis generated by the hypothesis explorer showing that PG may be linked to lipid metabolism and IBD (Khan et al., 2021).

4.1.1.6 Network construction under Visualization tab

One of the other unique features of this database is that it can create interactive network construction. To work with a plenitude of biomedical entities and to use terms from disparate dictionaries can sometimes be a tedious job. This network construction application can be found under 'Visualize tab'. Users

can establish networks connecting various entities, which is based on their frequency in the PubMed literature. Once the user pressed the visualize tab, it takes to the network page which is ingrained with various dictionaries. User can use more than one dictionary while constructing a network. There are instructions provided in the user manual guide (Appendix C). This network construction tool is based on the Cytoscape approach for graphical display and is embedded with a quintessential array of layouts (e.g., concentric, cola, breadthfirst, circle, cose and grid display layout). The network nodes are color-coded with different shapes representing the biomedical entities from the dictionary like star shape node represents diseases, diamond shape represents biological processes, etc, according to the dictionaries selected by the user and the numbering on edges represented the number of PubMed records. Users can trim or delete the edges or links that are not useful by clicking twice. Here I have used the same example of hypotheses generator (section 4.1.1.5) to create interactive network as shown in step wise. We clicked the visualize tab and opened the network page. Bearing in mind that we were curious regarding the role cholesterol play in IBD along with PG. We attempted to build linkages between cholesterol, IBD, and PG by creating a network of biological terminology as illustrated in a step-by-step method to determine the correctness and flexibility of this technique (Fig 4.7).

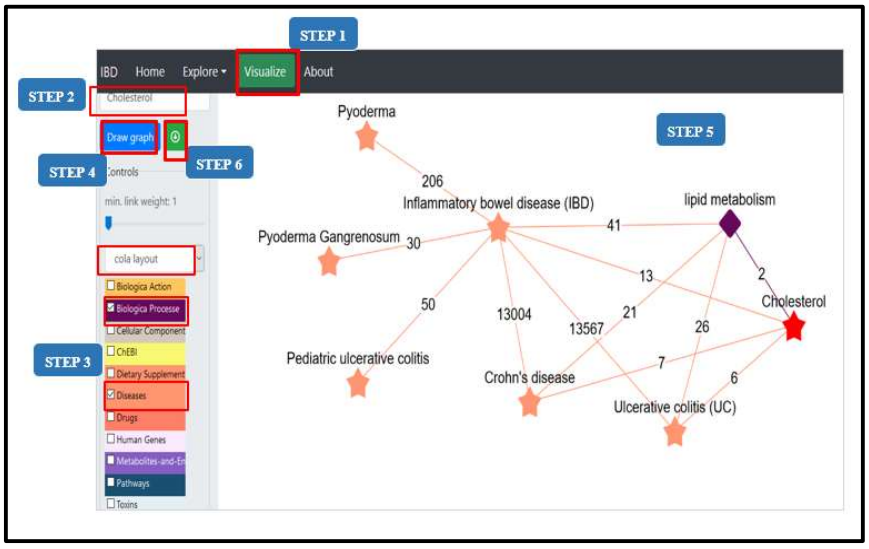


Figure 4. 7 This figure demonstrates the step-by-step network construction to identify the relationship between disease PG and its connection with the pathway lipid metabolism and IBD.

4.1.1.7 IBD database exportable formats

The IBDDB database can create networks and present these in six different interactive layouts (e.g., concentric, cola, breadthfirst, circle, cose and grid display layout) and this network presentation can be downloaded in .PNG (Portable Network Graphics) format. While the IBD genes datasets can easily

be exported in Excel, CSV (comma separated value), Print, PDF, and copy formats (Fig 4.8). Meanwhile IBDDB user manual guide can also be handy as it can be downloaded in PDF format.

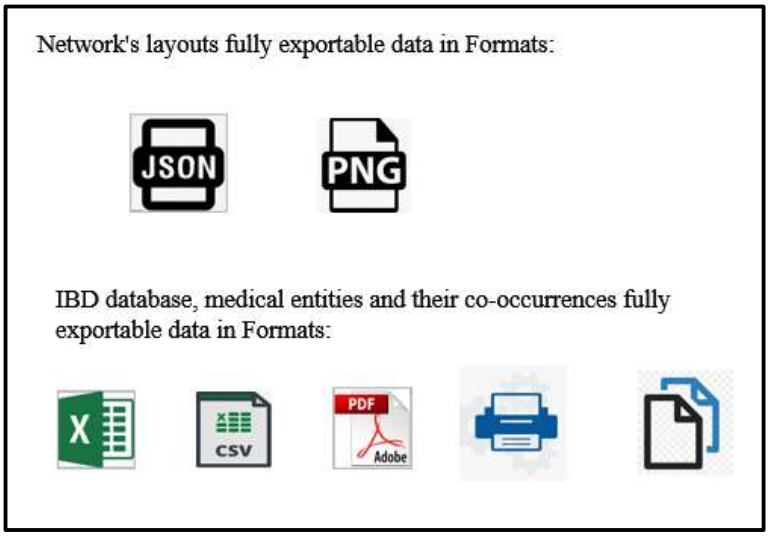


Figure 4. 8 This figure illustrates the exportable data from IBD database in various formats.

4.2 AIM 2

4.2.1 Identification of Biomarkers of IBD Genes

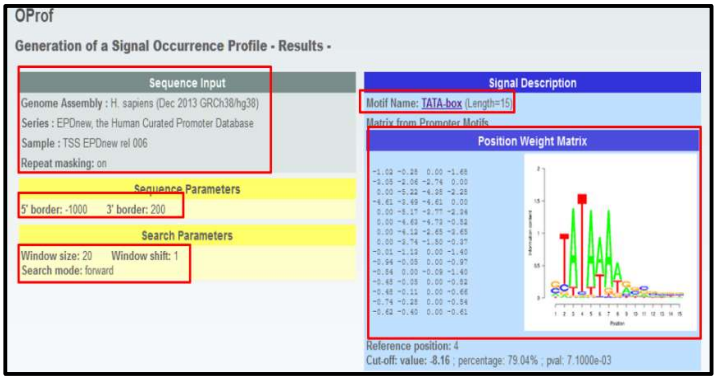
4.2.1.1 Positional distributions of core promoters' elements

For the identification of biomarkers of IBD. The 289 IBD genes from IBDDB database were downloaded in excel format (Table.4.1). We used computational approaches to determine the TFs that putatively control the transcription initiation. We comprehended that the minimal portion of the core promoters, required to properly initiate the transcription is 34 nucleotides upstream from the TSS. We identified the promoter regions range [-1000, +200] relative to the TSS at the 5' ends of genes and were extracted from EPDnew database as shown in (Fig 4.9 A). As a result, we used the OProf program, which elucidated a promoter as a position reference to a Gene bank/European Molecular Biology Laboratory (GenBank/EMBL) (Benson et al., 1993) sequence that determines the TSS location in the associated entry, as shown in the graph (Fig 4.9 B). Subsequently we were interested in knowing the TSS location in the nucleotide sequences for core promoters of IBD genes. Although promoter sequence approach was implemented through pointers around TSS location in the nucleotide entries and TSS of genes were determined based on the non-repetitive collection of eukaryotic Pol II promoters. Meanwhile, we were able to extract promoters of 87 genes out for 289 IBD genes using HGNC gene symbols of IBD genes. We therefore used OProf that scanned the set of DNA and displayed the distribution of TATA box (Conserved DNA sequence help in the initiation) as seen in the graph (Fig 4.9 B). We were able to display the positional distributions of core promoter elements, which were discovered at a set distance from the TSS and had a match to the TATA box weight matrices between positions -30 and -20 relative to the TSS. The parameters used to generate the positional distribution of core promoters' elements are shown in (Fig 4.9 B).

Table 4. 1 An excerpt from the datasets of IBD genes downloaded from IBDDB database in Excel format.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	HGNC gene s	HGNC gene name	chromosome_location	snp_variants	IBD phenotypes	previous_name	omim_id	omim_diseases	uniprotkb	entrez_id	HGNC_id	ensembl_id	refseq_DNA_s	pmid	pmcid	experimental evidences
2	1	NKX2-3	NK2 homeobox 3	chr10q24	rs1466509484, UC		CSX3, NK2 transcriptic	606727		Q8TAU0	159296	7836	ENSG000001	NC_000010.11	20957	295015	qPCR, RT-PCR, Genotyping, We
3	2	NLRP3	NLR family pyrin domai	chr1q44	rs1387455556, UC, CD		AGTAVPRL, AII, AVP, FCAS, FCAS1, FCU, KEFH, MW	Q96P20		114548	16400	ENSG000001	NC_000001.11	19784	274575	SNP genotyping, qPCR, ELISA	
4	3	NOD2	Nucleotide binding olig	chr16q12	rs905115240, r	IBD, CD	ACUG, BLAU, BLAUS, C	605956		Q9HC29	64127	5331	ENSG000001	NC_000016.10	25886	433569	miRNA microarray, RT-qPCR, SI
5	4	NOD1	Nucleotide binding olig	chr7p14	rs5743379, rs1	IBD	CLR7.1, NLRC1, Caspase recruitment domain-conta	Q9Y239		10392	16390	ENSG000001	NC_000007.14	21729	320965	FACS analysis, RT-PCR and Micr	
6	5	ATG16L1	Autophagy related 16 li	chr2q37	rs1432027609,	IBD, CD, UC, PIBD	APG16L beta, ATG16 a	610767	Inflammatory bo	Q676U5	55054	21498	ENSG000000	NC_000002.12	28130	532219	Genotyping, IgA measurement	
7	6	TNF	Tumor necrosis factor	chr6p21	rs757032200, r	IBD, UC, CD	APC1 protein, DIF, TNI	191160	Asthma,Malaria,	P01375	7124	11892	ENSG000002	NC_000006.12	22848	340497	Fish, qPCR, Histopathology, ELI
8	7	RIPK2	Receptor interacting se	chr8q21	rs1484792029,	IBD, CD	CARD-carrying kinase,	603455		O43353	8767	10020	ENSG000001	NC_000008.11	#####	######	Immunoblotting, qPCR and Flo
9	8	TLR2	Toll like receptor 4	chr4q32	rs1462952018,	CD, UC, IBD	CD282, toll/interleuki	603028	Colorectal cance	O60603	7097	11848	ENSG000001	NC_000004.12	28388	538466	Genotyping, SNP analysis, qPC
10	9	TLR4	Toll like receptor 3	chr9q32	rs143756948, r	UC, CD, IBD	ARMD10, CD284, TLR-4	603030	Colorectal cance	O00206	7099	11850	ENSG000001	NC_000009.12	29913	######	qPCR and Western blotting, RF
11	10	IL23R	Interleukin 23 receptor	chr1p31	rs908937526, r	IBD, UC, CD	Interleukin-23 receptc	607562	Psoriasis	Q5VWK5	149233	19100	ENSG000001	NC_000001.11	17447	185311	SNP analysis, FRET, qPCR, ELISA
12	11	IRGM	Immunity related GTPa	chr5q33	rs892534510, r	CD, UC, IBD	LRG-47-like protein, ir	608212		A1A4Y4	345611	29597	ENSG000002	NC_000005.10	#####	######	Microarray expression analysis
13	12	TLR3	Toll like receptor 3	chr4q35	rs1246822233,	IBD, UC	CD283, IIAE2, toll like	603029	Macular degener	O15455	7098	11849	ENSG000001	NC_000004.12	#####	######	1D SDS-PAGE, 2D Gel analysis a
14	13	GAPDH	Glyceraldehyde-3-phos	chr12p13	rs1062429, rs4	IBD, CD	G3PD, HEL-S-162eP, O	138400		P04406	2597	4141	ENSG000001	NC_000012.12	27812	278121	(GWAS) RT-qPCR, Immunoblot
15	14	PTPN2	Protein tyrosine phosph	chr18p11	rs1006143775,	IBD, CD, UC	PTN2, T-cell protein ty	176887	Rheumatoid arth	P17706	5771	9650	ENSG000001	NC_000018.10	#####	######	(GWAS) PCR, DNA extraction, F
16	15	MUC2	Mucin 2, oligomeric mu	chr11p15	rs1192639066,	IBD, UC, CD	MLP, mucin-2, mucin 2	158370		Q02817	4583	7512	ENSG000001	NC_000011.10	#####	######	Immunochip, High throughput
17	16	PIGR	Polymeric immunoglob	chr1q31	rs291102, rs22	IBD, CD, UC	Hepatocellular carcinc	173880		P01833	5284	8968	ENSG000001	NC_000001.11	#####	######	RT-qPCR, Tissue histology and
18	17	BAHD1	Bromo adjacent homolc	chr15q14	rs17856679, rs	UC	Bromo adjacent homc	613880		Q8TBE0	22893	29153	ENSG000001	NC_000015.10	#####	######	Immunohistochemistry, RT-qP
19	18	FFAR2	Free fatty acid receptor	chr19q13	rs1214823681,	IBD	FFA2R, fatty acid recep	603823		O15552	2867	4501	ENSG000001	NC_000019.10	#####		Immunohistochemistry, RT-qP
20	19	FFAR3	Free fatty acid receptor	chr19q13	rs1210689174,	IBD	FFA3R, GPR42, free fat	603821		O14843	2865	4499	ENSG000001	NC_000019.10	#####		Immunohistochemistry, RT-qP
21	20	NCF2	Neutrophil cytosolic fac	chr1q25	rs926521823, r	UC	P67-PHOX, neutrophil	608515		P19878	4688	7661	ENSG000001	NC_000001.11	21900	380648	Western blot, Cell culture and
22	21	RAC2	Rac family small GTPase	chr22q12	rs1359903900,	IBD, UC	Gx, HSPC022, Ras-relat	602049	Immunodeficien	P15153	5880	9802	ENSG000001	NC_000022.11	21900	380648	RT-PCR, Apoptosis analysis, SN
23	22	NCF4	Neutrophil cytosolic fac	chr22q13	rs941794887, r	CD	CGD3, NCF, P40PHOX,	601488		Q15080	4689	7662	ENSG000001	NC_000022.11	21900	######	SNP genotyping and analysis, S
24	23	SEL1L	SEL1L adaptor subunit o	chr14q31	rs1051193, rs1	UC	Hrd3, PRO1063, SEL1-l	602329		Q9UBV2	6400	10717	ENSG000000	NC_000014.9	#####		Western blot, Knockdown, RT-
25	24	GLI1	GLI family zinc finger 1	chr12q13	rs1442166745,	UC	GLI-Kruppel family me	165220		P08151	2735	4317	ENSG000001	NC_000012.12	#####		RT-qPCR, SNP genotyping, Gen
26	25	PTCH1	Patched 1	chr9q22	rs1472021371,	UC	BCNS, protein patched	601309		Q13635	5727	9585	ENSG000001	NC_000009.12	#####		DNA extraction, SNP genotypir
27	26	IHH	Indian hedgehog signali	chr2q33	rs1308941016,	CD	BDA1, HHG2, Indian he	600726		Q14623	3549	5956	ENSG000001	NC_000002.12	19071	######	ELISA, DAVID, RT-qPCR
28	27	AVIL	Advillin	chr12q13	rs866202746, r	IBD	ADVIL, DOC6, NPHS21	613397		O75366	10677	14188	ENSG000001	NC_000012.12	#####		PCR, Mapping of AVIL by Fish, S

A



B

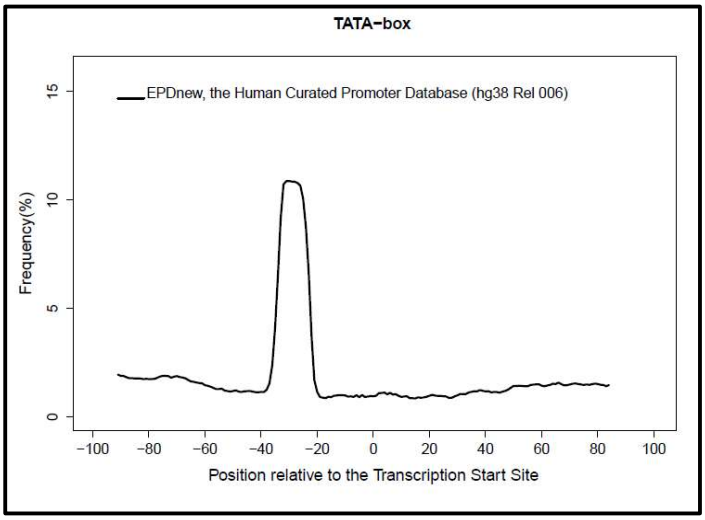


Figure 4. 9 Figure illustrates the TSS location of core promoters. **A.** Parameter used for the OProf program to scan the DNA sequences. **B.** This graph shows positional distributions of core promoters' elements, which are expected to be found at fixed distance from the TSS. The bed file contains genomic TSS coordinates of human promoters that match the TATA-box weight matrix distribution between positions -30 and -20 relative to the TSS.

4.2.1.2 Prediction of Transcription Factor Binding Sites

TFs regulate several of the cell's most core functions, including cell growth, division, differentiation, apoptosis, DNA replication, and DNA repair and since these are the distinct class of proteins that influence the rate of transcription by binding to the DNA (Latchman, 1997; Nebert, 2002). Understanding the molecular mechanisms underpinning gene regulation may be critical for finding essential genes or proteins that can be used as prognostic or diagnostic biomarkers. Therefore, not only we wanted to identify TFs for IBD genes as biomarkers but also, we wanted to know which TFs can be responsible for the progression from IBD to CRC. Therefore, we aimed to explore for CRC progression from IBD during the early stages of disease. We, therefore, downloaded the 2056 CRC genes from CORECG database and 289 IBD genes from IBDDB database. We applied OPOSSUM 3.0 application for the identification of over-representative TFs and its binding sites. The parameters selected were [-1000, +200] upstream and downstream regions in nucleotide. The list of 289 IBD genes were submitted to the OPOSSUM 3 software for TFs query. This list was compared for information content of their base pairs and were annotated for targeted gene hits and targeted TFBS hits and were computed based on Fisher exact test score as seen in (Table 4.2). We selected those TFs from the outcome which controlled maximum number of genes. The criteria for selection of putative TFs were based on one-tailed of fisher exact probability, which should be greater than or equal to 7 (score ≥ 7). The fisher score ranks the over-representation of a TFs putative binding sites from the most strongly over-represented to the least, to aid in selecting potential TFs based on their classes and families.

Table 4. 2 This table shows the TFs regulating IBD genes by number of gene hits and TFBS hits based on Fisher score as the outcome by OPOSSUM 3.

Transcription Factors	Gene hits	TFBS hits	Fisher score
SPIB	154	593	11.710
ELF5	136	401	9.706
SPI1	134	341	13.088
FEV	130	309	14.132
RUNX1	106	196	12.942
NFKB1	69	108	11.856
RELA	57	80	13.501

Further, we explored that OPOSSUM mapped the TFs to mammalian matrices models using information content of TFs matrices profiles from Jaspar 2018 database as shown in outcome of the TFs

query search. The outcome of the analysis was downloaded as the tabulated form with several informative columns about the analysed TFs (Table 4.3). These columns are TFs, JASPAR ID, Class of TF, family of TF, Taxa, Information content (IC), Guanine-cytosine content (GC), Target gene hits, Target gene non-hits, Background gene hits, Background gene non hits, Target TFBS hits, Target TFBS nucleotide rate, Background TFBS hits, Background TFBS nucleotide rate, Z-score and Fisher score which measures the over-represented TFBSs. The IC is the specificity of the TFBS' PWMs profile, and it may influence the rate of occurrences when the IC is either < 9 or > 19 by highlighting in red, alerting the user to the extreme of cases. Likewise in case of GC content of each TFBS binding profile, it highlights when GC is either < 0.33 or > 0.66 alerting the user.

Table 4. 3 This table shows the analysis of IBD genes outcome through OPOSSUM 3.

TF	JASPAR ID	Class	Family	Tax group	IC	GC Content	Target gene hits	Target gene non-hits	Background gene hits	Background gene non-hits	Target TFBS hits	Target TFBS nucleotide rate	Background TFBS hits	Background TFBS nucleotide rate	Z-score	Fisher score
FEV	MA0156.1	Winged Helix-Turn-Helix	Ets	vertebrates	12.121	0.442	130	148	8066	16686	309	0.027	17806	0.0194	16.695	14.132
RELA	MA0107.1	Ig-fold	Rel	vertebrates	14.757	0.567	57	221	2643	22109	80	0.00874	3406	0.00463	18.250	13.501
REL	MA0101.1	Ig-fold	Rel	vertebrates	10.515	0.559	86	192	4725	20027	146	0.0159	7315	0.00995	18.243	13.221
SPI1	MA0080.2	Winged Helix-Turn-Helix	Ets	vertebrates	9.640	0.435	134	144	8544	16208	341	0.0261	21231	0.0202	12.552	13.088
RUNX1	MA0002.2	Ig-fold	Runt	vertebrates	10.401	0.502	106	172	6292	18460	196	0.0235	12041	0.018	12.547	12.942
NF-kappaB	MA0061.1	Ig-fold	Rel	vertebrates	13.345	0.621	69	209	3633	21119	108	0.0118	5106	0.00695	17.635	11.856
SPIB	MA0081.1	Winged Helix-Turn-Helix	Ets	vertebrates	9.060	0.466	154	124	10465	14287	593	0.0453	38079	0.0363	14.655	11.710
IRF1	MA0050.1	Winged Helix-Turn-Helix	IRF	vertebrates	16.008	0.383	41	237	1803	22949	51	0.00668	2315	0.00378	14.289	11.021
ELF5	MA0136.1	Winged Helix-Turn-Helix	Ets	vertebrates	8.693	0.429	136	142	9248	15504	401	0.0394	25420	0.0311	14.419	9.706
IRF2	MA0051.1	Winged Helix-Turn-Helix	IRF	vertebrates	21.134	0.412	9	269	198	24554	9	0.00177	202	0.000495	17.267	7.526

4.2.1.3 TFBS mapping

TFBSs in the promoter sequences were mapped using all the mammalian matrices models like PFMs and PWMs which are embedded in JASPAR 2018 database. TFs sites were scanned by sliding the corresponding PWMs along the sequences and scoring it at each position. Further verifications of the predicted TFs were implemented by mapping them to PFMs models profile using *Homo sapiens* matrix ID and DNA base pair logos as shown in the example for one of the TFs RUNX1 (Fig 4.10). Here it shows the RUNX1 matrix profile MA0002.1 summary on the left side with description of this TF's class, family, JASPAR core vertebrate collection, taxa, species, data type, uniport ID and source. On the right side it depicts two tables, the sequence motif logo table with bases, here the sequences are the binding sites on the x-axis and bits on the y-axis. The other table is frequency matrix PFMs of A, C, G, T bases. These PFMs can be downloaded as sequence alignments under various formats like JASPAR, TRANSFAC, MEME and Raw PFM. The frequency matrix is read as the first 5 or 6 column

lines. Meanwhile the PFMs was downloaded in JASPAR file format as shown in (Fig 4.11). The mapping to PWMs was performed by using PWMs tools that scanned the frequency matrix in JASPAR file format and the outcome is shown in (Fig 4.12).

We selected TFs based on the promoter sets and mapped them with mammalian matrices models like PWMs, PFMs along with the sequences and scoring it at their position pointers around TSS location in each nucleotide entries. The threshold value at minimum relative score used to report the position as a putative binding site. The background set was computed using a threshold of 70%.

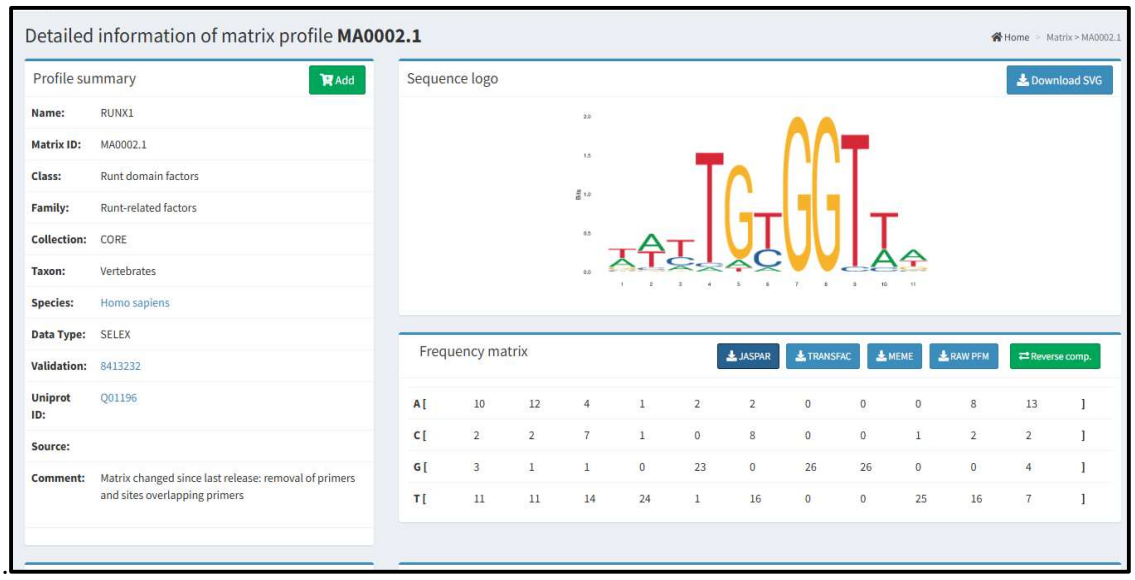


Figure 4. 10 This figure depicts the TFBS of RUNX1 mapped to the JASPAR matrix ID MA0002.1 mammalian matrix model like PFM. On the left- side it shows the summary of TF RUNX1. On the right-side top table, it shows the sequence motif logo representing bits on y-axis and frequency of bases on x-axis. The table below illustrates the frequency matrix of A, C, G, T bases. These TFBS profiles which are stored as sequence alignment data can be downloaded under various formats.

>MA0002.1 RUNX1

A [	10	12	4	1	2	2	0	0	0	8	13]
C [	2	2	7	1	0	8	0	0	1	2	2]
G [	3	1	1	0	23	0	26	26	0	0	4]
T [	11	11	14	24	1	16	0	0	25	16	7]

Figure 4. 11 This is the matrix profile of TF RUNX1 from the core redundant vertebrate collection, displayed as PWMs scores in the RUNX1 TFBS model (weights) retrieved in JASPAR format. PFMs summarize the occurrences of each nucleotide at each position in a set of observed TF-DNA interactions. PFMs are transformed into probability models, that would be used to generate PWMs that can scan any DNA sequence and predict TFBS.

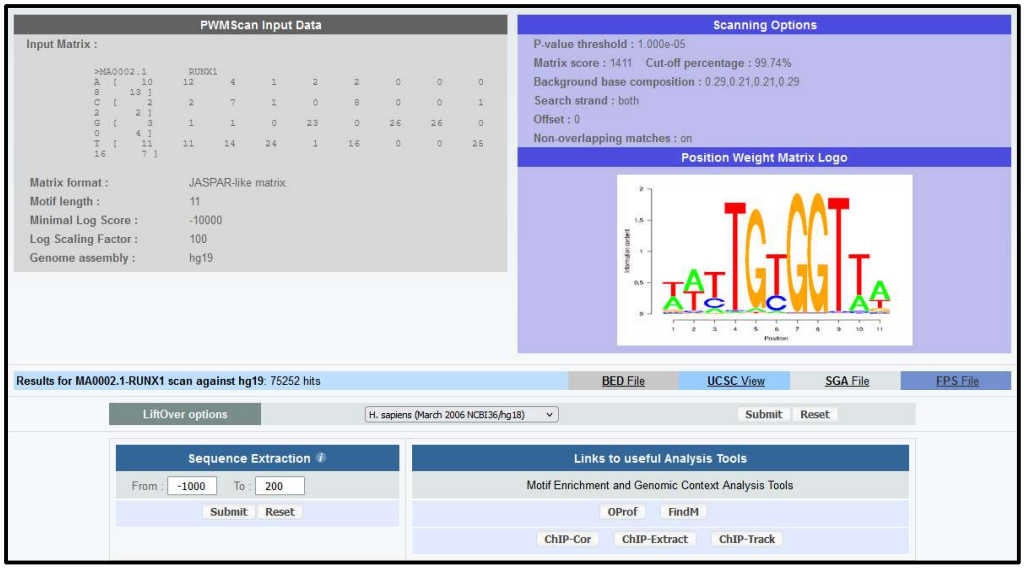


Figure 4. 12 This is the outcome from PWMScan. The input data was JASPAR file format of RUNX1 TF-matrix in the genome assembly hg19 and motif length is 11 as shown on left-side. On the right-side top table displays the scanning option of analysis parameters used are p-value threshold of 1.000e-05, matrix score 1411 and cut-off percentage 99.74% which is searched on for both strands, along with non-overlapping matches. The PWM logo table is shown below, with position frequency scoring of base occurrences on the x-axis and IC content on the y-axis. The matrix profile can be further examined to extract sequence in FASTA format, which can then be downloaded as BED, SGA, or FPS files and viewed in the UCSC genome browser.

4.2.1.4 Enrichment of TFBS patterns and edge identifications

Enrichment of mapped TFBSs found in the targeted gene was determined based on their over-represented index in binding sites by co-expressed genes. We ranked these IBD TFs on: 1) the maximum number of targeted IBD genes controlled by TF and 2) by their maximum fisher score achieved by TFs. The selected TFs of IBD genes were shown in the (Table 4.4). The table shows TFs in the left column followed by family, class matrix Id, gene hits, TFBS hits, rank, fisher score and their base pair logos (PWMs logos). These TFs are also verified in the published literature with their PubMed ID.

Table 4. 4 This representative table shows ranking based on Fisher exact test score of some of the TFs out of the complete set of 289 IBD genes.

TF_Symbol	Family	Class	MATRIX_ID	Literature validation	Number of gene targets hits	TFBS_hits	Ranks	Fisher score	Sequence_Logo
SPIB	E26-transformation-specific	Tryptophan cluster factors	MA0081.1	7624145	154	593	1	11.710	
ELF5	E26-transformation-specific	Tryptophan cluster factors	MA0136.2	35045294	136	401	2	9.706	
SPI1	E26-transformation-specific	Tryptophan cluster factors	MA0080.2	35212366	134	341	3	13.088	
FEV	E26-transformation-specific	Tryptophan cluster factors	MA0156.2	27803750	130	309	4	14.132	
RUNX1	Runt-related factors	Runt domain factors	MA0002.1	27668286	106	196	5	12.942	
NFKB1	NF-kappaB related factors	(RHR) Rel homology region factors	MA0105.2	31572379	69	108	6	11.856	
RELA	NF-kappaB-related factors	(RHR) Rel homology region factors	MA0107.1	34155144	57	80	7	13.501	

4.2.1.5 Transcriptional Regulatory Network of IBD genes

The clusters of genes were elucidated on the basis of the promoters allowed the creation of network of IBD genes that were potentially regulated by individual TFs or their combinations. We demonstrated the 289 IBD genes network and TRN of TF RUNX1 regulating 106 genes by creating interactive network (Fig 4.13 A and B). In (Fig 4.13 A), it demonstrates the network of 289 IBD genes and (Fig 4.13 B) shows TF RUNX1 controlling 106 IBD genes. For example, TF RUNX1 which has been found to have binding 106 genes out of 289 IBD genes. Therefore, the network shows that one TF may control the expression of 106 IBD related genes. Therefore, we can anticipate that this particular TF can control about 35% of 289 IBD genes, thus the identified TF to be at the hub of the gene network linked to IBD.

A B

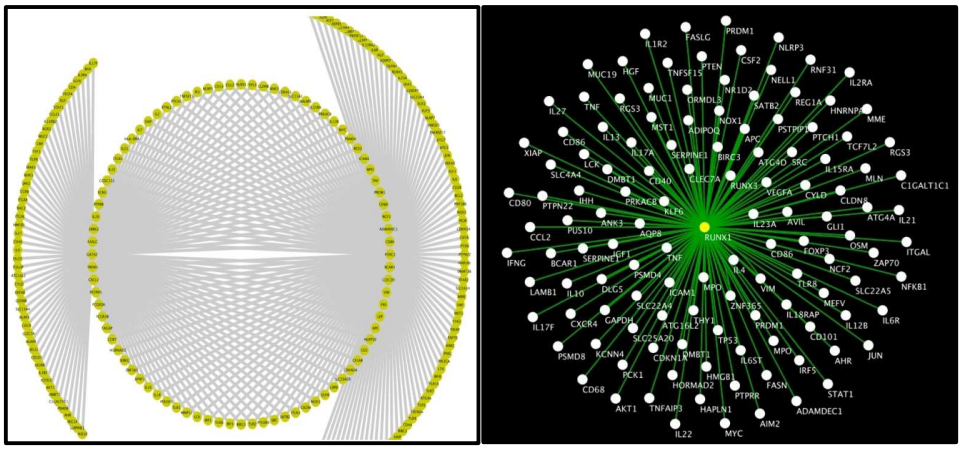


Figure 4. 13 Reconstruction of TRN model of IBD genes. **A.** Illustrates the largest network of 289 IBD genes. **B.** Demonstrates the TF RUNX1 regulating 106 IBD genes. TRN edges represent direct interactions between a transcription factor and its target genes. The plug-in used in Cytoscape was ReNE, to determine the hub of IBD genes controlled by the TFs. ReNE plug-in was installed, and a tabulated data was uploaded as a .txt file format in Cytoscape and the source and target were selected as nodes and edges constructing the data into a regulatory network.

4.2.1.6 Functional and Pathway enrichment analysis of Network of IBD genes

We performed functional enrichment through GO and KEGG pathway analysis (Fig 4.14) using Cluster profiler. Cluster profiler is a Bioconductor package in R script programming (Appendix B section 8.2.1**)**. We used a cut-off p- value < 0.05 to identify significant terms. The enriched pathways with p-value < 0.01 were cytokine-cytokine receptor interaction, Jak-stat Signalling pathway, NOD-like receptor Signalling pathway, Th17 cell differentiation, IBD, Chagas disease and toxoplasmosis. The most significant terms with p- value < 0.05 in biological processes were positive regulation of cytokine production, T-cell activation, regulation of inflammatory response, response to molecule of bacterial origin, leukocyte cell- cell adhesion, lymphocyte differentiation and leucocyte proliferation. In cellular component the most significant terms were external side of plasma membrane, membrane raft, membrane microdomain, membrane region and vesicle lumen. Molecular functions significant terms detected at p- value < 0.05 were cytokine receptor, receptor ligand binding, signalling receptor activator activity, cytokine activity and growth factor receptor binding. From the results of the GO annotations, it was predicted that cytokines play an important key role in the pathogenesis of the gut microbiota in IBD patients. It was found that mucosal macrophage play an important role in the mucosal immune system as played by these genes NLRP3, TNF, TLR2 and TLR4. These activated macrophages are found to be expressed as the major contributors to the production of inflammatory cytokines in the gut and imbalances of these cytokines contributed to the pathogenesis of the IBD disease (Sanchez-Muñoz

et al., 2008). APAF1 gene was found upregulating apoptosis in UC, which could lead or correlate with inflammation tendency in the IBD patients' pouches (Leal et al., 2010). It could be because both inflammatory and apoptosis pathways are related. It was also found that relative mRNA expression of JAK2 was significantly increased in patients with IBD in comparison with healthy controls (Song et al., 2013). In addition, JAK2 mRNA expression was significantly decreased in patients with IBD having thrombosis compared with those without thrombosis. BCL-2 gene family proteins BAX and BCL-2 are major regulators of Apoptosis/cell-death programme in UC patients. They inhibit apoptosis induced by active caspases and promote cell-survival (Patankar and Becker, 2020). It was also observed that innate immune system that played an important part in the IBD was found in various IBD genes. To learn more about these IBD gene expression in patients' samples, the resulted expression level of significant GO terms in relation with IBD genes (Table 8.2 and 8.3) of (Appendix A Table 8.2 and 8.3).

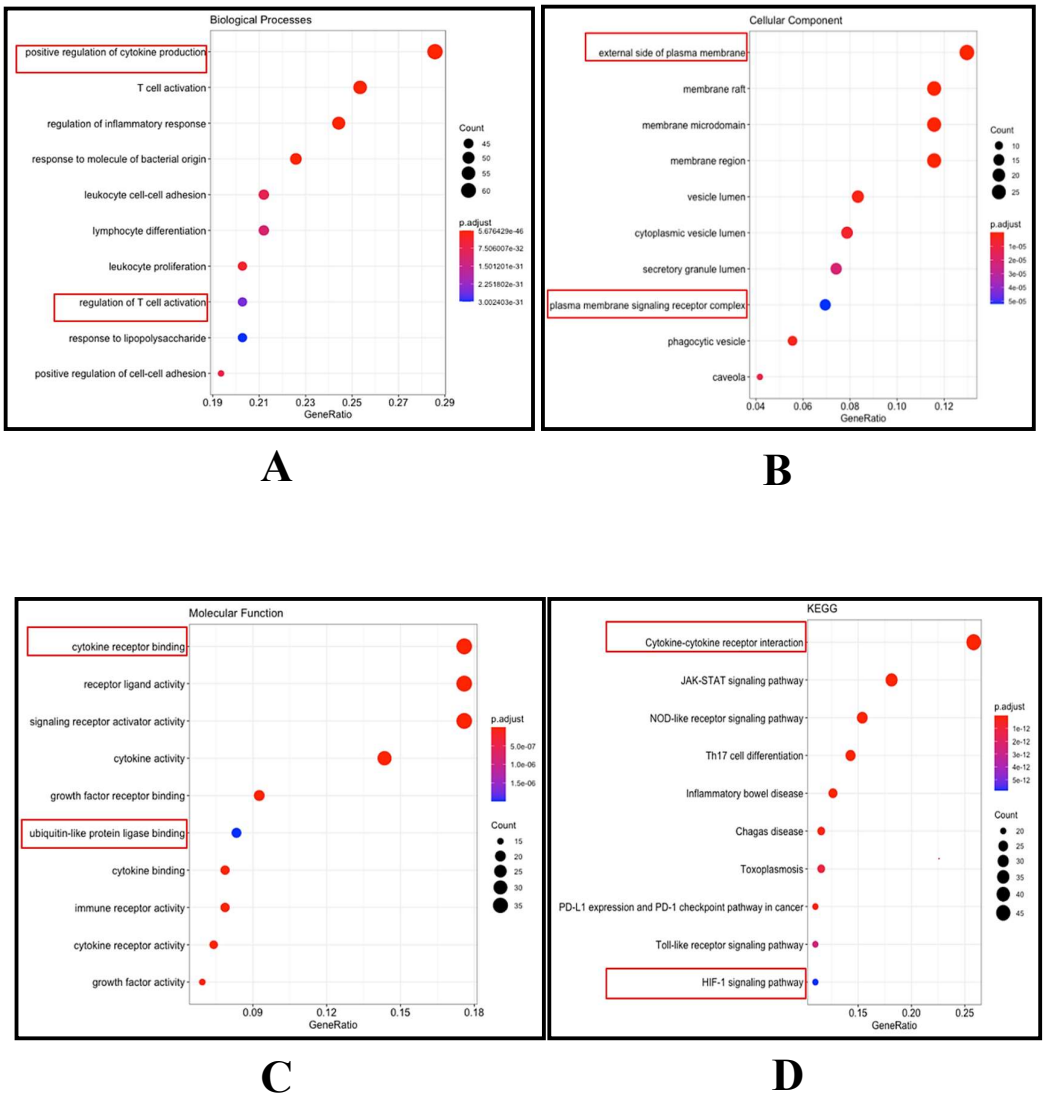


Figure 4. 14 This figure illustrates the GO annotation and KEGG Pathway analysis using R package: ClusterProfiler. **A.** Biological processes. **B.** Cellular components. **C.** Molecular functions. **D.** KEGG Pathway analysis.

4.2.1.7 In silico Validation of Predicted Biomarkers in IBD Patients

The enriched TFs that were ranked based on their maximum number of genes regulated by them were selected as a candidate biomarker for IBD disease diagnosis (section 4.1.2.4 and Table 4.4). An *in silico* approach was used to identify these selected TFs in the development of the disease and its progression through mining of biological database such as ONCOMINE, TCGA, Gene Expression Atlas. The diagnostic values for each TFs' biomarker were retrieved from ONCOMINE database like TP, FP, TN and FN and were thus evaluated. Their sensitivity and specificity were calculated along with their other diagnostic evaluations (section 3.3.10)

for the formulae implemented. More reliable biomarkers for IBD prediction could help in early detection and prevention of CRC. To correctly identify the patients with disease, biomarkers were selected based on their high sensitivity and specificity values (Table 4.5). RUNX1 was selected as the potential biomarker of IBD because of its high sensitivity and specificity as 77 % and 66 % respectively in comparison to other biomarkers. As other biomarkers for example some very high sensitivity but least specificity like RELA, SPIB, NFKB1, ELF5. Meanwhile FEV biomarker had a sensitivity of 53 % and specificity of 50%, but it was very low with what an ideal biomarkers diagnostic value should be. It is often found in clinical settings that sensitivity should be always higher than specificity. Here FEV has a sensitivity of 53% which denote that 53% patients have disease, while 47% do not have a disease but are wrongfully treated as a disease patient. Therefore, ideal characteristics suggest that sensitivity should be high like 80% or so and specificity should be less than sensitivity for a biomarker to be efficient. While SPI1 has a sensitivity of 43 % and specificity 55 %, which is a very low sensitivity but high specificity. If taken example of RELA with 81 % sensitivity and specificity of 8 %, cannot be an ideal biomarker as it can predict that many patients who are disease free were wrong fully diagnosed with disease and can be subjected to further diagnostic investigations. Therefore, we have selected the biomarker which has high sensitivity and specificity that is RUNX1. RUNX1 is selected based on two main factors. Firstly, for its ideal diagnostic values and secondly it can be detected in a non-invasive way in blood. The LRs of RUNX1 were calculated to be LR+ 2.21 and LR- 0.38 which is rather good than other biomarkers as seen in (Table 4.6). RUNX1 TF regulates 106 IBD genes which means it is binding to 106 genes out of 289 IBD genes as seen in (Table 4.4). It can be anticipated that this particular TF basically controls 106 IBD related networks of genes that is about 35 % of 289 genes (Fig 4.13 B). Consequently, with the good diagnostic values in comparison to other biomarkers, RUNX1 can be further tested to be a putative biomarker of IBD.

A key influence of test performance is the intent-to-diagnose population, specifically the prevalence of the disease or condition that the test attempts to diagnose or detect in that community (Landsheer, 2018). Therefore, the need of proper diagnosis necessitates a review of diagnostic biomarker test clinical performance. In (Table 4.5), we presented the columns representing various clinical tests help in identifying the good putative biomarker in identifying patients with disease. The various columns are Biomarkers of IBD, sensitivity, specificity, PPV, NPV, pre-test odds, post-test odds of positive and negative test and LRs for positive and negative

tests along with 95% CI. Here the example taken to define the various diagnostic tests of biomarkers of IBD is TF RUNX1. RUNX1 was selected as the potential biomarker of IBD because of its high sensitivity and specificity as 77 % and 66 % respectively in comparison to other biomarkers. The PPV value of this biomarker is high, 92% among all the other biomarkers. It means the prevalence of the disease to diagnose in a patient is dependent on this test PPV and NPV. Hence 92% of probability of PPV is developed in this group of patients to have the disease. The pre-test probability or odds is the probability prior to the test to have the disease but post-test probability confirms the chances of having the disease after the diagnosis and can be helpful in decision making by the clinicians to diagnose further in the patient. Here the positive post-test probability is 42% of confirming the disease and negative post-test probability is 12% of not having the disease in the patient. The LR is used to determine a diagnostic test's efficacy and to assist in the selection of the optimal diagnostic tests and their sequences. LR has a significant advantage over sensitivity and specificity in that it is less likely to modify with disorder prevalence and may be used to aggregate the findings of many diagnostic tests to assess the likelihood of a target disorder being diagnosed (Hayden and Brown, 1999). The LR for RUNX1 is LR+ 2.21 and LR- is 0.38. Here LR- 0.38 indicates that value 0.38 decreases the probability of disease about 25% but it still presents a disease. As LR- should be between 0 – 0.1 and not 1.0. It also states that a person without IBD is about 3 times (=1/0.38) more likely to have a negative test than someone with IBD (Mathioudakis et al., 2018). Here LR+ should be placed in between the range from 1 – infinity and for LR- it should be between range 0 - 0.1.

Table 4.6 presented with the diagnostic tests based on accuracy, LR and AUC_{ROC}. The accuracy or effectiveness assesses a test's capacity to identify a condition whether it is present or not. It should be noted that the prevalence of disease may have an impact on the test's accuracy. The accuracy of RUNX1 showed 0.74 which is good in comparison to other biomarkers as well as its AUC_{ROC} value which is dependent on sensitivity and 1- specificity cut-off values. The higher the sensitivity, the higher the curve forms on the upper left corner of the graph. The LR is also stable as compared to the other biomarkers of IBD. Hence considering all the aspects of the diagnostic biomarker, TF RUNX1 can be considered as a putative TF- based biomarker of IBD.

Table 4. 5 This table represents the diagnostic values of biomarkers for clinical utility based on Sensitivity, Specificity, PPV, NPV, Pre-test odds, Post-test odds and LR of IBD biomarkers at 95% CI.

Biomarkers	Sensitivity (%)	Specificity (%)	Positive predictive values (PPV) (%)	Negative predictive values (NPV) (%)	Pre-test odds	Post-test odds (+) [95% CI]:	Post-test odds (-) [95% CI]:	Likelihood Ratio (+)	(-)
SPIB	59	3	52	3	0.66	92 [48 – 57%]	21 [81 – 100%]	0.60	16.4
ELF5	69	26	63	31	0.45	30 [58 – 69%]	16 [54 – 81%]	0.93	0.42
SPI1	43	55	65	33	1.27	55 [55 – 74%]	56 [58 – 73%]	0.98	1.02
FEV	53	50	80	22	0.88	48 [68 - 88%]	45 [65 - 87%]	1.06	0.94
NFKB1	50	17	70	7	1	37 [62 – 78%]	75 [77 – 98%]	0.6	2.94
RUNX1	77	66	92	32	0.33	42 [85 - 97%]	12 [55 - 79%]	2.21	0.38
RELA	81	8	78	8	0.23	17 [85 – 97%]	35 [55 – 79%]	0.88	2.25

Table 4. 6 This table represents the diagnostic capabilities of biomarkers based on Accuracy, LR and AUC_{ROC} in predicting risk of IBD at 95% CI.

Biomarkers	**Accuracy**	**Positive Likelihood ratios (LR+) [95% CI]:**	**Negative Likelihood ratios (LR-) [95% CI]:**	**Area under curve AUC_{ROC} [95%CI]:**
SPIB	0.39	0.60 [0.49 - 0.74]	16.4 [2.32 - 12.08]	0.39 [0.6 – 1]
ELF5	0.54	0.94 [0.73 - 1.19]	1.18 [0.62 - 2.22]	0.54 [0.71 – 0.89]
SPI1	0.48	0.98 [0.63 - 1.52]	1.02 [0.72 - 1.44]	0.48 [0.71 – 0.89]
FEV	0.53	1.06 [0.57 - 1.99]	0.93 [0.49 - 1.78]	0.53 [0.71 – 0.88]
NFKB1	0.43	0.6 [0.41 - 0.877]	2.99 [0.82 - 10.97]	0.43 [0.69 – 0.91]
RUNX1	0.74	2.21 [1.00 - 5.09]	0.38 [0.21 - 0.66]	0.74 [0.67- 0.83]
RELA	0.66	0.88 [0.71-1.10]	2.25 [0.31-16.08]	0.66 [0.82-0.98]

4.2.1.8 Expression of IBD Biomarkers in Patients' Samples

Further validation of these TFs was conducted in patient's samples using 'The Gene Expression Atlas' database. The input parameters of TFs were in HGNC format, *Homo sapiens* as species was selected, and disease category was 'inflammatory bowel disease'. The outcome received was expressed in up or down-regulation and was differentiated based on the IBD phenotypes (UC, CD). The mRNA expression level of these TFs was calculated in Log 2-Fold change, in the samples obtained from blood or tissue using low-throughput techniques like low-throughput techniques like, immunohistochemistry (IHC), quantitative polymerase chain reaction (qPCR), western blot (Table 4.7). These TFs were further analysed in CRC disease category to compare the expression of IBD biomarkers in CRC disease. This could help in predicting progression to CRC at its early stage of cancer. We inferred that all the biomarkers of IBD were up regulated in IBD disease but down regulated in CRC. The TF RUNX1 1.3-fold change was found to be up regulated in CD. While it was found to be down regulated by -2.6 in CRC.

Further evaluation of these TFs was carried out using GEPIA2 online tool that sources its data from TCGA and GTEx. Validation of the expression of these TFs of IBD genes in (colon adenocarcinoma) COAD and READ (rectum adenocarcinoma) database was performed using GEPIA box whisker plots. We found that three TFs showed overexpression levels as shown in (Fig 4.15). The parameters applied for the box plot analysis for changes in the fold of TFs (log 2 transformation of gene expression) are TCGA normal and GTEx data, p-value cut off 0.01 and log 2 (TPM+1) for log scale. Expression of RELA, RUNX1 and NFKB1were significantly higher in COAD-READ than in normal tissues in comparison to FEV TF. SPIB, SPI1 and ELF5 expression levels were lowest in COAD-READ than in normal tissues. For RUNX1, the log 2-fold change found in the interquartile region (median) value of the box was 4.0 in both COAD-READ than in normal tissue. In summary, all the TFs were up regulated except for SPI1 TF in which the tumour vs normal expression was down regulated in COAD- READ expression data patients' samples.

Table 4. 7 Expression of IBD biomarkers in patients' samples.

Biomarkers	Expression Log 2-Fold Change (IBD)	IBD	Species	Samples IBD	Expression of Log2-Fold change (CRC)	CRC Stages	Samples CRC	Species	Detection
SPIB	2.7 [Up]	CD	*Homo sapiens*	Bone marrow/ WBC	-2.2 [Down]	I - II	Biopsy	Mice	qPCR/IHC/Western blotting
ELF5	1.5 [Up]	CD	*Homo sapiens*	Intestinal biopsies/WBC	-3.2 [Down]	I - II	Biopsy	*Homo sapiens*	qPCR/IHC/Western blotting
SPI1	1.8 [Up]	UC, CD	Mice	Blood and intestinal biopsies	-2.0 [Down]	I - II	Blood platelets	*Homo sapiens*	qPCR/IHC/Western blotting
FEV	1.2 [Up]	UC	*Homo sapiens*	Tissue /Blood	-1.4 [Down]	I - II	Biopsy	*Homo sapiens*	qPCR/IHC/Western blotting
NFKB1	1.7 [Up]	UC	*Homo sapiens*	WBC/Macrophages	-1.1 [Down]	I - II	Serum/ Serrated polyposis	*Homo sapiens*	qPCR/IHC/Western blotting
RUNX1	1.3 [Up]	CD	*Homo sapiens*	Bone marrow tissue/normal blood cell/causes decrease in blood platelets	-2.6 [Down]	I - II	Blood platelets $<1 \text{ X } 10^5$ µL	*Homo sapiens*	qPCR/IHC/Western blotting
RELA	1.4 [Up]	CD	*Homo sapiens*	Blood and intestinal biopsies	-1.9 [Down]	I – II	Biopsy	*Homo sapiens*	qPCR/IHC/Western blotting

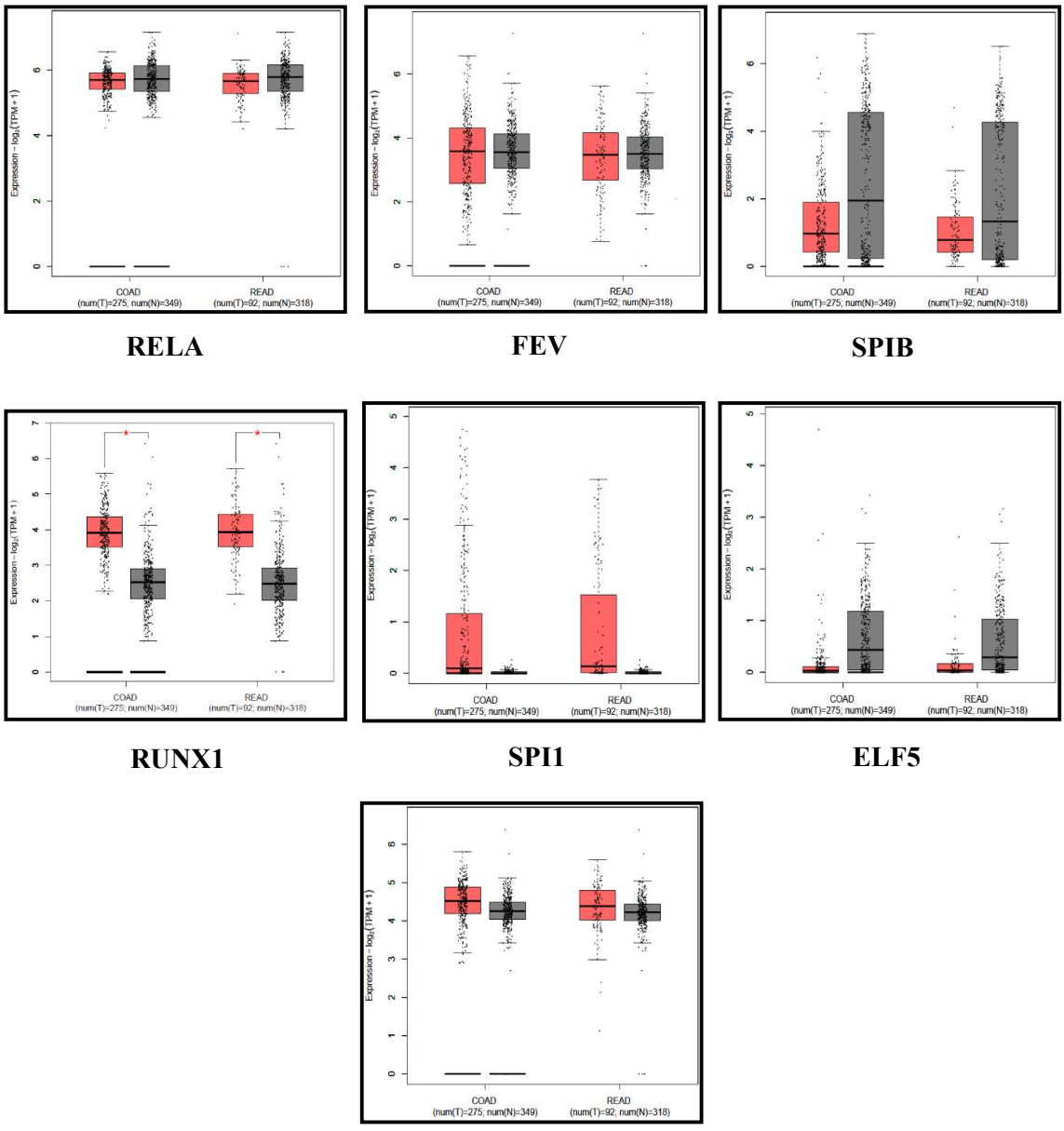


Figure 4. 15 Box-whisker plot shows relative transcription of biomarkers in sub-groups of COAD-READ patients, stratified based on gene expression level in CRC tumor vs Normal status (red for tumor, grey for normal). These box whisker plot span 25th to 75th percentiles, lines in the interquartile region show the median value, whiskers indicate the minimum and maximum value. Significance of difference is estimated by log2-Fold change (TPM+1). The red star denotes statistical significance of p - value < 0.05. Num represents the 275 colon cancer COAD samples vs 349 normal samples and num in READ represents 92 rectal cancer samples vs 318 normal samples.

4.2.1.9 ROC curve analysis of individual IBD Biomarkers

ROC curve analysis was performed on individual biomarkers of IBD using Evidence-based medicine toolbox (EBM calculator) https://ebmtools.knowledgetranslation.net/calculator/diagnostic/. ROC curve was made based on the post-test probability for each of the predicted IBD biomarkers. Post-test probability is not directly provided but can be indirectly obtained by using the ROC curve. Pre-test and post-test probabilities confirm the presence of a condition or disease before and after the diagnostic test is implemented. Post-test probability measurement is very important it helps the clinicians to make decision whether to accept a diagnosis of certain disease or rule out or to implement more tests. This calculator calculated the sensitivity, specificity, LR+, LR-, PPV, NPV along with 95% CI using values provided for TP, TN, FP, FN. Based on its calculations it sketched the ROC curve for each individual IBD biomarkers as shown in the (Fig 4.16). The two axes shown in the ROC curve illustrate (1-specificity) on the x- axis and sensitivity on the y-axis.

Outcome of the post-test probability ROC curve shown for various biomarkers, along with other parameters such as (shown in Table 4.5 and 4.6) sensitivity, specificity, and other diagnostic values can help to identify a most suitable biomarker AUC of 0.5 or below is considered, as no discrimination or poor test, on ability to diagnose patient with or without disease based on a test, while AUC of 0.7 to 0.8 is acceptable and 0.9 AUC is outstanding. RELA shows the probability of negative test to be high than positive test, resulting in patient free from disease, whereas FEV shows positive and negative test post-test probability to be almost equal (Fig 4.16). While SPIB, ELF5, NFKB1 all show negative test as high and positive to be low, falling under the AUC of below < 0.7. Meanwhile, RUNX1 shows positive test to be higher than negative test, considering patients have a disease and the AUC is 0.74.

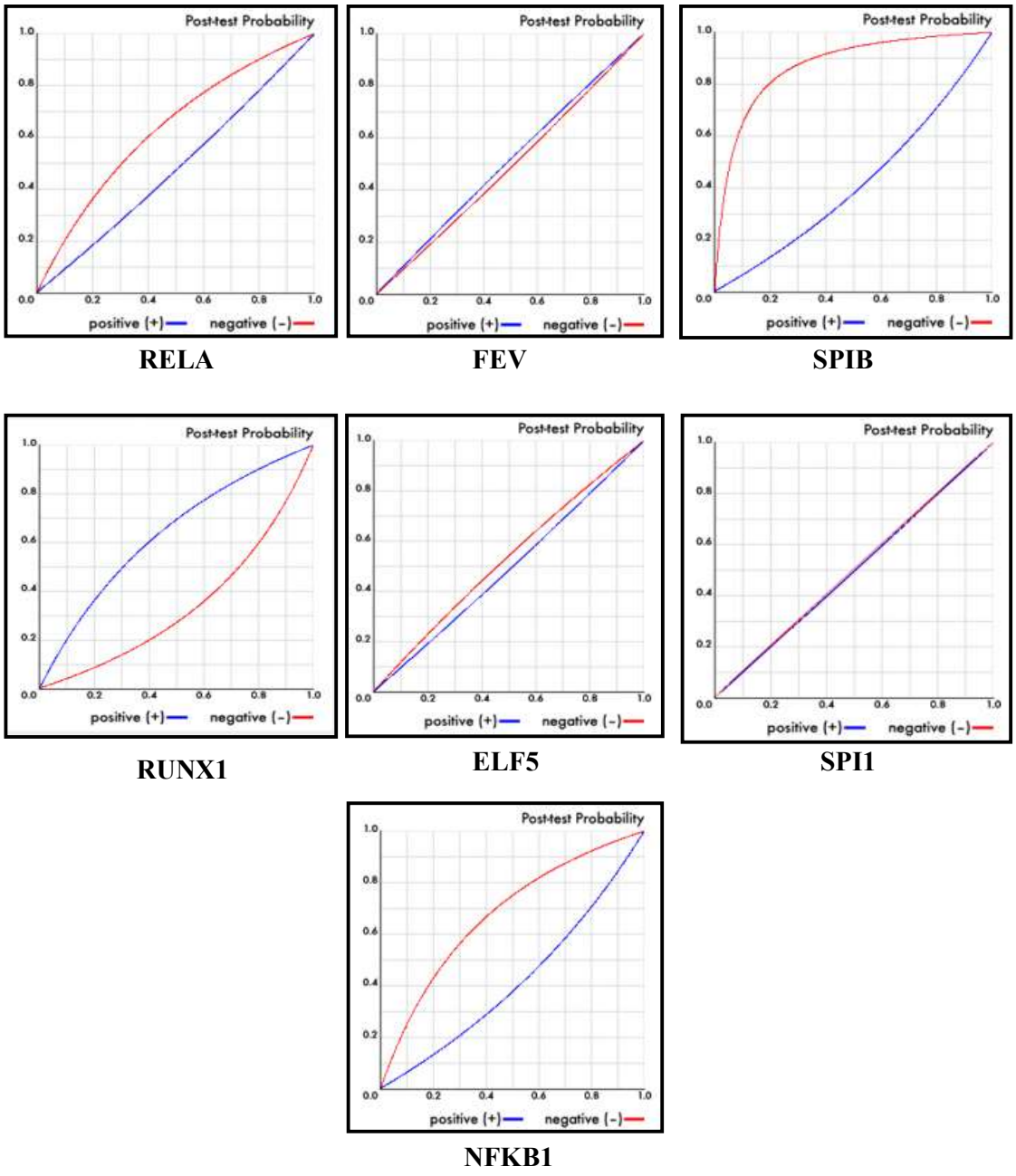


Figure 4. 16 ROC curve analysis of post-test probability for positive and negative test in individual IBD biomarkers. The two axes shown in the ROC is (1-specificity), while on the y-axis is sensitivity. The positive test (blue) indicates the sensitivity, and the negative test (red) indicates the 1- specificity. Greater the AUC and the closer the curve is to the upper-left hand corner, the better the test is at distinguishing between diseased and non-diseased people. The AUC, which can be any value between 0 and 1, is a good indicator of the test's accuracy.

4.2.1.10 Combination and evaluation of IBD Biomarkers

Accuracy in the diagnostic test is very important for the prediction of disease at the early onset of it. However, due to the low sensitivity and low specificity in detecting the disease in asymptomatic patients of IBD can be overwhelming. In this thesis identification and verification of seven TFs as IBD biomarkers is done. Investigating their diagnostic properties, it was found that most of the TFs candidate contain either highest sensitivity and lowest specificity or lowest sensitivity and high specificity, none of them were ideal except for one biomarker that is RUNX1 with 77 % sensitivity and 66 % specificity but still not with very high diagnostic value. Therefore, to improve the accuracy of the diagnostic test to detect IBD at its early stage, it was required to combine the TFs into various sets of combinations. For the application of the formulae applied (section 3.3.10.13).

The (Table 4.8) represents four combined sets that is, combination of seven markers, five markers, four markers and three markers. The outcomes of the combinations are as follows. In combinations set of three markers representing a, b, c and d sets respectively, the accuracy was not improved as in set **a**) (NFKB1 + RELA + RUNX1) it denotes the sensitivity is 97 % and specificity of 74%. In **b**) (SPIB + ELF5 + SPI1) sensitivity is of 93 % and specificity is of 68 %. In **c**) (FEV + NFKB1 + RUNX1) sensitivity is 95 % and specificity is 86 %. In **d**) (RELA + SPIB + FEV) sensitivity is 96 % and specificity is 55 %. In another set of four biomarkers combination (FEV + NFKB1 + RELA + RUNX1), it was found sensitivity of 99 % and specificity of 87 %, which could denote that due to low specificity even after combining 4 markers 87 % were disease free but rest of 13 % of patients remaining can be assumed to be falsely diagnosed with disease. Therefore, this set is again not useful for diagnosis. In a set of five combinations of biomarkers representing a and b set, it was noted that in set **a**) (SPIB + ELF5 + SPI1 + FEV + RUNX1) has a sensitivity of 99 % and specificity of 95 % whereas, in set **b**) (NFKB1 + RUNX1 + RELA + SPIB + ELF5) has a sensitivity of 99 % and specificity 81 %. It is assumed that this set **a)** (SPIB + ELF5 + SPI1 + FEV + RUNX1) has a high sensitivity and specificity within 90 - 99 %. It could be helpful in predicting the accurate diagnosis. Meanwhile, the combination setoff seven biomarkers (SPIB + ELF5 + SPI1 + FEV + NFKB1 + RUNX1 + RELA) has a sensitivity 99.9 % and specificity of 87 %, which again falls in the category of low specificity. Hence the biomarker combination of five (SPIB + ELF5 + SPI1 + FEV + RUNX1) can provide better diagnosis.

Table 4. 8 Evaluation and combinations of biomarkers of IBD.

Combinations	Markers	Sensitivity	Specificity
Single Marker	SPIB	59 %	3 %
	ELF5	69 %	26 %
	SPI1	43 %	55 %
	FEV	53 %	50 %
	NFKB1	50 %	17 %
	RUNX1	77 %	66 %
	RELA	81 %	8 %
Seven Markers	SPIB+ELF5+SPI1+FEV+NFKB1+RUNX1+RELA	99.9 %	87 %
Five markers **(a) Excluded:** NFKB1, RELA **(b) Excluded:** SPI1, FEV	SPIB+ELF5+SPI1+FEV+RUNX1 NFKB1+RUNX1+RELA+SPIB+ELF5	99 % 99 %	95% 81 %
Four Markers	FEV+NFKB1+RELA+RUNX1	99 %	87%
Three Markers (a) Excluded: SPIB, ELF5, SPI1, FEV **(b) Excluded:** FEV, NFKB1, RUNX1, RELA **(c) Excluded:** RELA, SPIB, ELF5, SPI1 **(d) Excluded:** ELF5, SPI1, NFKB1, RUNX1	NFKB1+RELA+RUNX1 SPIB+ELF5+SPI1 FEV+NFKB1+RUNX1 RELA+SPIB+FEV	97 % 93 % 95 % 96 %	74 % 68 % 86 % 55 %

4.2.1.11 Survival analysis of combined biomarkers of IBD

GEPIA2 was used to analyse the relationship between predicted TFs and overall survival (OS) and disease-free survival (DFS) using Kaplan-Meier plotter curve in CRC COAD-READ database. The parameters used to analyse expression of TFs was implemented using (multi gene signature), hazard ratio measured based on Cox PH Model at 95% CI. The five TFs grouped together as a signature (multi gene) expressed in OS showed that patients with high expression of these TFs were primarily associated with shorter OS based on the actual graph (p- value = 0.087) as shown in (Fig 4.17 A). The expression between DFS was also evaluated. It was found that patients with high expression of these TFs were somewhat associated with shorter DFS (p-value = 0.32) (Fig 4.17 B). The p-values do not show significant association however, based on the actual survival graphs especially in the later months, it is expected that the expression of these TFs may play a role in the prognosis of patients with CRC but may not be used as a reliable predictor of survival of patients.

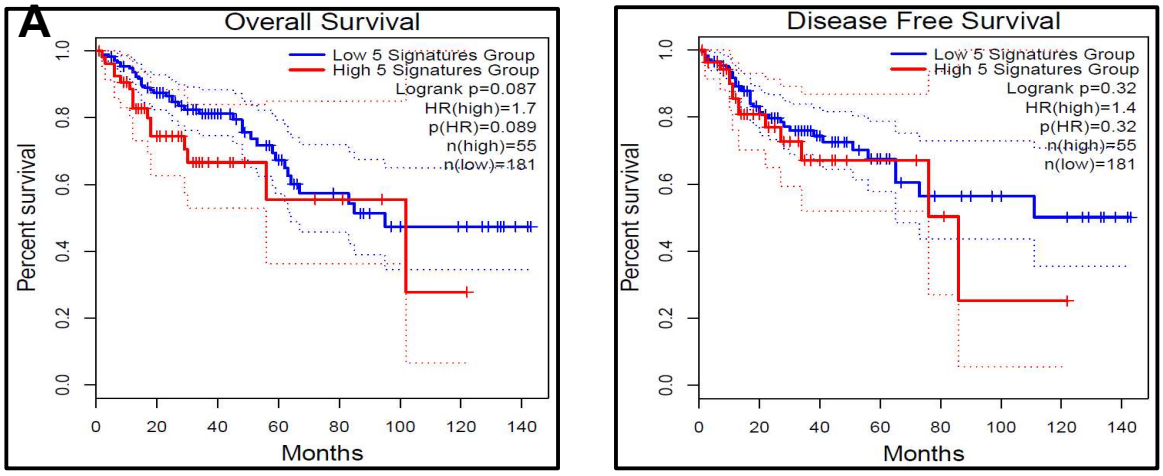


Figure 4. 17 High expression of combination biomarkers was associated with **(A)** worse OS, hazard ratio (HR)=1.7, log rank P= 0.087 and **(B)** worse DFS, HR= 1.4, log rank P=0.32. One-way ANOVA method was used for differential analysis *p-value Cutoff ≤ 0.01.

In summary the aim 2 focuses on the computational identification and validation of biomarkers for early and less-invasive diagnosis and screening of IBD. Here the TFs were identified po-

tentially controlling the IBD related genes. The TFBS were predicted and mapped to mammalian matrices models like PWMs and PFMs, generating information to understand the regulation at molecular level and to identify the underlying aetiology of IBD using GO annotation and pathway analysis. The TRNs of IBD genes were constructed. Further *in silico* validation of these TFs was done and the TFs with most accurate diagnostic value were selected as IBD biomarkers and their expression profiles were validated in IBD, CD, UC patients' samples and were juxtaposed with CRC patients' samples during early stages. Hence, the most suitable combinations of biomarkers were identified (SPIB + ELF5 + SPI1 + FEV + RUNX1). The potential biomarker of IBD identified is RUNX1 that can aid in predicting the risk of early diagnosis IBD patients.

4.3 AIM 3

4.3.1 To identify biomarkers which can serve as potential diagnostic measure to predict early risk of CRC in IBD patients.

An *in silico* approach was used to identify the biomarkers of IBD and CRC in the development of the disease and its progression from IBD to CRC using computational methods. We previously identified IBD biomarkers by downloading genes from the IBDDB database and using bioinformatics approach to identify TFs and their binding sites. Further analysis of these TFs was done by mapping them to matrices models and by reconstruction of the TRNs. We further performed *in silico* analysis on these predicted biomarkers using Gene Expression Atlas in patients' samples. For validation of these biomarkers, we retrieved the diagnostic values like TP, TN, FP, and FN for each biomarker of IBD using ONCOMINE database and calculated different parameters of diagnostic tests by applying various formulae (section 3.1.10). The expression of all the TFs of IBD were analyzed using COAD-READ TCGA database in CRC patients' samples. Finally, we performed survival analysis on the best set of combined biomarkers of IBD (SPIB + ELF5 + SPI1 + FEV + RUNX1). After performing this investigation, we selected TF RUNX1 as a potential biomarker for the early diagnosis of IBD in patients' samples.

The IBD biomarker RUNX1 was selected for early diagnosis of IBD disease based on diagnostic value of sensitivity 77 % and specificity 66 % which was comparatively higher than other IBD biomarkers and based on the non-invasive form, of diagnosis and its expression value detected in samples like blood, Serum, platelets, and tissues in few using low-throughput techniques like immunochemical fecal occult blood test (iFOBT), IHC, reverse transcription polymerase chain reaction (RT-PCR). Furthermore, to evaluate the potential diagnosis of this marker in CRC patient samples to detect early progression to CRC disease, it was required to combine RUNX1 with putative CRC biomarkers. We identified the putative biomarkers of CRC by verifying from the published literature. This RUNX1 biomarker was further analyzed with the known conventional tumor markers that are known to be expressed in the CRC cancer cells and are currently in practical use by clinicians as CRC biomarkers in laboratory settings. The diagnostic values of these CRC biomarkers along with IBD biomarker (Table 4.9 and 4.10).

It was predicted that the known biomarkers of CRC possessed less sensitivity but high specificity values than IBD biomarker RUNX1. CEA has a diagnostic sensitivity of 46 % and specificity of 80 %. TIMP1 has a sensitivity of 52 % and specificity of 60 %, while CA724 and CA199 has a sensitivity of 45 % and 14.39 %, and specificity of 97 % and 89 %, respectively (Table 4.9). Overall, their accuracy (Table 4.10) was comparatively lower than RUNX1. However, after measuring their diagnostic values, it is found that most of these biomarkers acquired high PPV, hence found higher probability in this group of patients to have the disease. The pre-test probability or odds is the likelihood of having the condition before the test, but the post-test probability confirms the chances of having the disease after the diagnosis and can aid clinicians in their decision to diagnose the patient further. The post-test probability was found to be higher in most of these biomarkers.

Meanwhile the positive LR of these individual biomarkers were greater than 1 (LR+ > 1), which was sufficient to interpret the test to be positive. Altogether the good LR, CA724 dominated with LR+ 16.6, which represent an excellent chance of how likely a patient has a disease, yet these biomarkers of CRC currently in use have low diagnostic values and can only predict half (50 %) of the disease in the patients. Therefore, there is a need to combine these biomarkers together to increase their diagnostic accuracy to predict 100 % of the disease in the patients, if possible.

Table 4. 9 This table represents the diagnostic values of biomarkers for clinical utility based on Sensitivity, Specificity, PPV, NPV, Pre-test odds, Post-test odds and LR of IBD to CRC biomarkers at 95% CI.

Biomarker	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Pre-test odds	Post-test odds (+) [95% CI]:	Post-test odds (-) [95% CI]:	Likelihood Ratio (+) (-)
RUNX1	75%; 0.75	66%; 0.66	92	32	0.33	42% [85-97%]	12% [55-79%]	2.2; 0.3
CEA	46%; 0.461	80%; 0.80	89.7	28.4	1.15	73% [84-93%]	43.2% [68-75%]	2.33; 0.67
TIMP1	52%; 0.52	60%; 0.60	42.2	69.5	0.92	54% [34-51%]	42.5% [24-38%]	1.3; 0.8
CA724	45%; 0.45	97%; 0.97	98.4	31.9	1.22	95% [94-100%]	41% [66-70%]	16.6; 0.57
CA199	14.4%; 0.14	89%; 0.89	83.7	21.7	5.948	88.6% [72-91%]	85.1% [77-80%]	1.36; 0.59

Table 4. 10 This table represents the diagnostic capabilities of biomarkers based on Accuracy, LRs and AUC_{ROC} in predicting risk of IBD to CRC progression at 95% CI.

Biomarkers	**Accuracy**	**LR+ [95% CI]:**	**LR- [95% CI]:**	**AUC_{ROC} [95%CI]:**
RUNX1	0.74	2.21 [1.00-5.09]	0.38 [0.21-0.66]	0.74 [0.67-0.83]
CEA	0.53	2.7 [1.44-3.67]	0.7 [0.57-0.79]	0.53 [0.44-0.65]
TIMP1	0.57	1.3 [0.92-1.90]	0.8 [0.57-1.10]	0.57 [0.50-0.71]
CA724	0.56	16.6 [4.20-65.4]	0.57 [0.51-0.63]	0.56 [0.50-0.70]
CA199	0.30	1.36 [0.67-2.77]	0.59 [0.87-1.05]	0.30 [0.17-0.43]

4.3.1.1 Expression of IBD to CRC Biomarkers in IBD and CRC patients' samples

Further validation of these markers was conducted in patient's samples using 'The Gene Expression Atlas database' (Table 4.11). The input parameters of biomarkers were in HGNC format, *Homo sapiens* as species were selected, and disease category was 'IBD' and 'CRC'. The outcome received was expressed in up or down-regulation and was differentiated based on the IBD phenotypes (UC, CD). The expression level of IBD and CRC biomarkers was calculated in Log 2-Fold change, along with the samples obtained from blood or tissue and detected by low throughput techniques as seen in (Table 4.11). This could help in predicting disease progression from IBD to CRC at its early stage of cancer. We identified that IBD biomarker RUNX1 was up regulated (Log 2-FC >1.3) in IBD but down regulated (FC < -2.6) in CRC patients' samples. The CRC biomarkers were found to be up regulated in IBD patients' samples except for CA199 which was found to be down regulated (FC <-1.2). Meanwhile, to predict early-stage diagnosis in CRC patients' samples all the biomarkers were found to be up-regulated except RUNX1 (FC< -2.6). These biomarkers were expressed at their specific level in blood serum and platelet samples in IBD, and CRC patients (Table 4.11).

In the (Fig 4.18). the heat map displayed the expression values of 4 biomarkers CA199, RUNX1, CEA, TIMP1 in IBD, CD, UC vs control in patient samples. The yellow color showed up-regulation of biomarkers in patients' samples and blue colour represents down-regulation of biomarkers in patients' samples. The IBD patients' samples used were RNA-Seq of Ileal biopsies from CD vs non-IBD control, Active UC vs control, CD recurrence vs control, IBD vs control, and newly diagnosed CD vs control. In **A)** expression values (Z-score transformed) indicated CA199 positively correlated only in CD vs control in ileum, whereas TIMP1 was positively correlated in newly diagnosed CD, in active UC and IBD vs control and negatively correlated recurrence in CD patients. CEA is positively up-regulated in CD vs control in terminal ileum and found down-regulated in UC and CD recurrence state in patients. RUNX1 was found to be upregulated in Active UC only. In **B)** heat map for log2 fold changes of 4 serum biomarkers expression in comparison to IBD, CD, UC vs control in patient samples. CA199 showed down-regulation in CD recurrence but up-regulation in CD ileum vs non-IBD control. RUNX1 shows up-regulation in Active UC. Meanwhile CEA indicated the up-regulation in CD, whereas TIMP1 was up-regulated in newly diagnosed CD, IBD and in Active UC and CD vs control.

Further evaluation of these Biomarkers was carried out using GEPIA 2 online tool that source its data from TCGA and GTEx as seen in (Fig 4.19). Validation of the expression of these five biomarkers of IBD-CRC was analysed in COAD and READ database using GEPIA box whisker plots. We found that three biomarkers (CEA, TIMP1 and RUNX1) showed overexpression levels. The parameters applied for the box plot analysis for changes in the fold (log 2 transformation of gene expression) were TCGA normal and GTEx data, p-value cut off 0.01 and log 2 (TPM+1) for log scale. These box whisker plot span 25^{th} to 75^{th} percentiles, lines in the lower and upper interquartile region respectively, showed the median value, whiskers indicate the minimum and maximum value.

Expression of TIMP1, RUNX1 and CEA were significantly higher in COAD-READ than in normal tissues in comparison to CA724 and CA199. In CA724 log 2-fold change found in the interquartile region (median) value of the box was 6.0 in both COAD-READ than in normal tissue and in CA199 log 2-fold change found in the interquartile region (median) value of the box was 4.0 in both COAD-READ than in normal tissue expression levels. CA724 and CA199 expression levels detected were lowest in COAD-READ than in normal tissues. Meanwhile, in CEA log 2-fold change found in the interquartile region (median) value of the box was 11.0 in both COAD-READ than in normal tissue. In TIMP1 log 2-fold change found in the interquartile region (median) value of the box was 9.0 in both COAD-READ than in normal tissue. In RUNX1 log 2-fold change found in the interquartile region (median) value of the box was 4.0 in both COAD-READ than in normal tissue. In summary, all the biomarkers were up regulated except for CA724 and CA199 in which the tumour vs normal expression was down-regulated in COAD-READ expression data patients' samples.

Table 4. 11 Expression of IBD to CRC biomarkers in patients' samples. The expression of 5 biomarkers was compared in IBD and CRC patients' samples at the given concentration unit, predicting progression from IBD to CRC in the early stages of CRC. mRNA Expression was measured from blood serum and platelets using Log 2-Fold change.

Biomarkers	Expression Log 2-Fold change (IBD)	Serum level in IBD	Expression Log 2-Fold change (CRC)	Serum level in CRC	CRC Stages	Samples	Detected by
RUNX1	1.3 [Up]	96-100 IU/mL	-2.6 [Down]	50-100 ng/mL	I – II	Tis-sue/blood platelets	PCR/RT-PCR/IHC/Western Blot
CEA	4.2 [Up]	< 2.5 ng/mL	4.0 [Up]	> 5 ng/mL	I – II	Blood serum	IHC/ immunochemical Faecal occult blood test (iFOBT)
TIMP1	4.3 [Up]	< 98.5 ng/mL	2.6 [Up]	98.5-101.3 ng/mL	I – II	Blood serum	IHC/ immunochemical FOBT (iFOBT)
CA724	2.0 [Up]	< 0.4 IU/mL	3.5 [Up]	0.8 IU/mL	I – II	Blood serum	IHC/ immunochemical FOBT (iFOBT)
CA199	- 1.2 [Down]	< 37 IU/mL	1.3 [Up]	>37 IU/mL	I – II	Blood serum	IHC/ immunochemical FOBT (iFOBT)

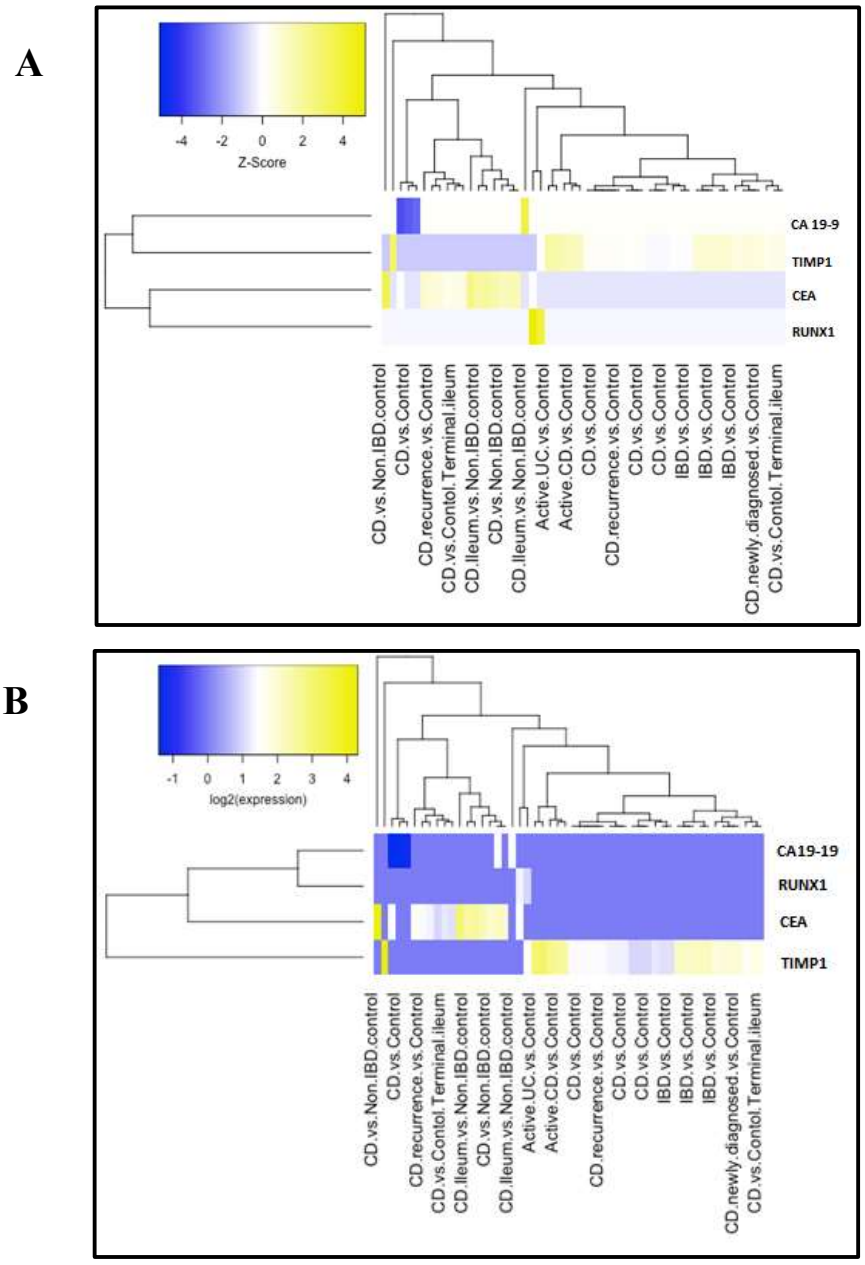
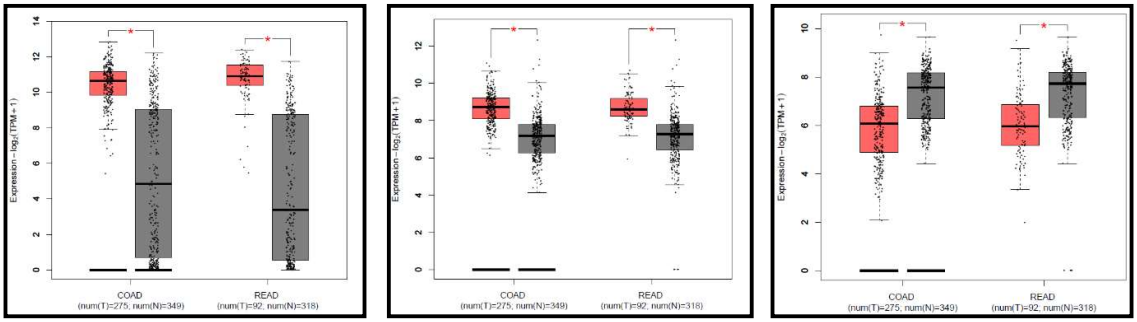
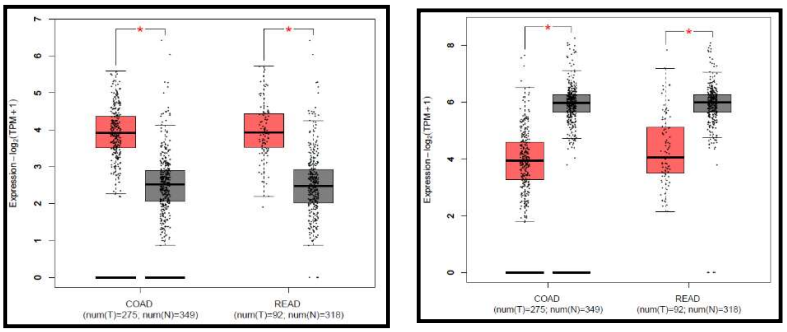


Figure 4. 18 Heat map illustrates the expression values of 4 biomarkers in the patients' samples in IBD. The color (yellow) shows up-regulation of biomarker in patients' samples and color (blue) shows down-regulation in patients' samples by biomarker. The 4 biomarkers are CA199. RUNX1, CEA and TIMP1. The IBD patients' samples used were RNA-Seq of Ileal biopsies of CD vs non-IBD control, Active UC vs control, CD recurrence vs control, IBD vs control, newly diagnosed CD vs control. **A.** This heat map is Z-score transformed based on standard deviations. **B.** This heat map shows expression values in Log2 scale.



CEA **TIMP1** **CA724**



RUNX1 **CA199**

Figure 4. 19 Box-whisker plot shows relative expression of biomarkers in sub-groups of COAD-READ patients, stratified based on gene expression level in CRC tumor vs Normal status (red for tumor, black for normal. These box whisker plot span 25th to 75th percentiles, lines in the interquartile region show the median value, whiskers indicate the minimum and maximum value. Significance of difference is estimated by log2-Fold change (TPM+1). The red star denotes statistical significance p - value < 0.05. The red star denotes statistical significance of p - value < 0.05. Num represents the 275 colon cancer COAD samples vs 349 normal samples and num in READ represents 92 rectal cancer samples vs 318 normal samples.

4.3.1.2 ROC curve analysis of IBD Biomarkers predicting progression to CRC

EBM calculator was used to perform ROC curve analysis. The post-test probability for each of the IBD-CRC biomarkers was used to create a ROC curve. The ROC curve can be used to refer post-test probability, which is not directly presented. As explained before the probabilities indicating the presence of a condition or disease before and after the diagnostic test are known as pre-test and post-test probabilities. Post-test probability estimation is critical because it aids clinicians in deciding whether to accept a diagnosis of a certain condition or rule out or perform more tests. Using the values provided for TP, TN, FP, and FN, this calculator generated sensitivity, specificity, LR+, LR-, PPV, NPV, and 95 % CI. The ROC curve for each individual IBD biomarker was sketched based on its calculations, as illustrated in (Fig 4.20). The two axes shown in the ROC curve illustrated (1-specificity) on the x- axis and sensitivity on the y-axis.

Outcome of the post-test probability ROC curve is illustrated in the respective biomarkers, also refer (Table 4.9 and 4.10) for sensitivity, specificity, and other diagnostic values. As the AUC of 0.5 or below is considered as no discrimination detected or poor test, on ability to diagnose patient with or without disease based on a test, while AUC of 0.7 to 0.8 is acceptable and 0.9 AUC is outstanding (Ray et al., 2010). Biomarker TIMP1 showed the probability of positive test to be very low with no curve or positive and negative test post-test probability to be almost equal, resulting in patient free from disease, AUC is 0.57. Biomarker CA199 also showed the dull curve or positive and negative test to be almost equal, AUC is 0.30. Biomarker CEA showed positive test to be higher in the curve than the negative test, AUC is 0.53. Biomarker CA724 showed the highest positive test than negative test among all the biomarkers with AUC 0.56. Whereas RUNX1 also showed positive test to be higher than negative test, considering patients have a disease and the AUC is 0.74 refer to (Table 4.10). In summary, it is estimated that even if AUC is less, the curve can predict highest sensitivity level.

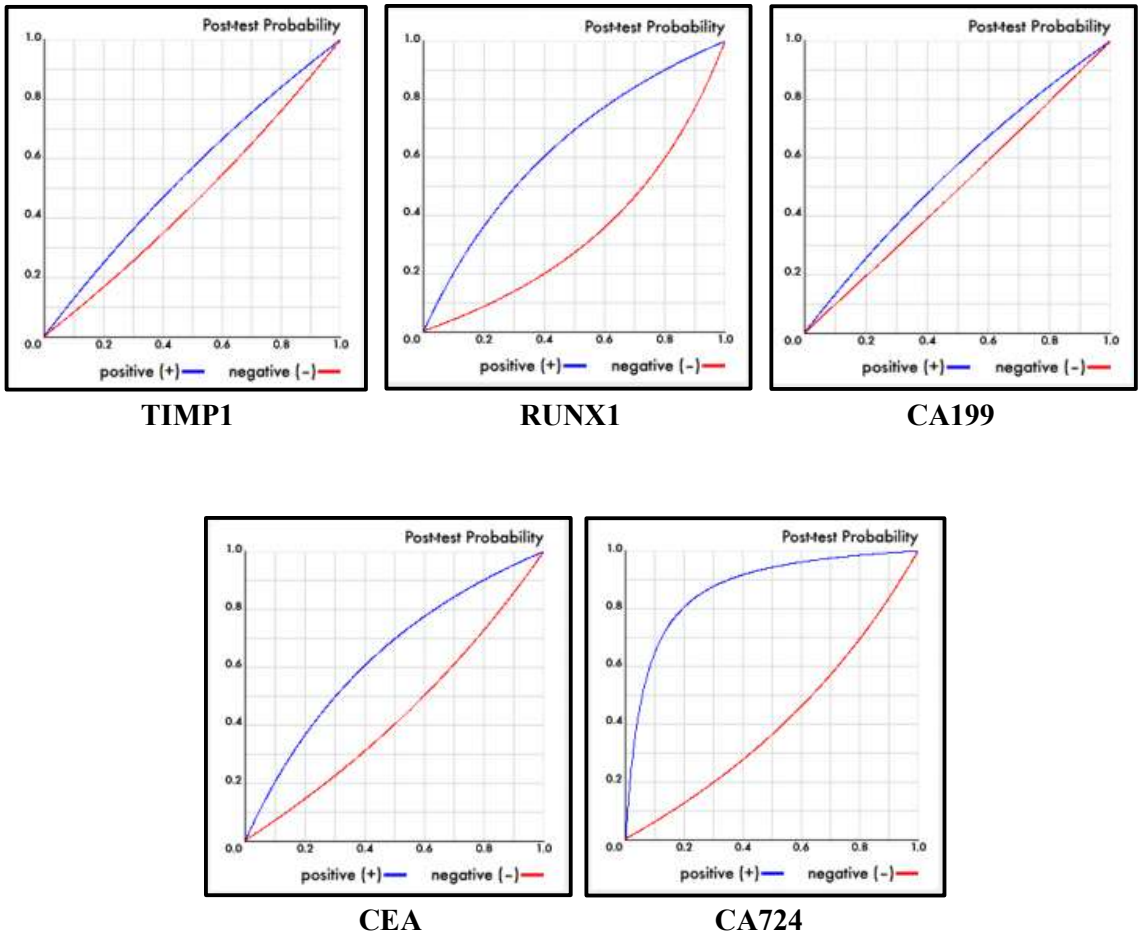


Figure 4. 20 ROC curve analysis of post-test probability for positive and negative test in individual IBD to CRC biomarkers. The two axes shown in the ROC is (1-specificity), while on the y-axis is sensitivity. The positive test (blue) indicates the sensitivity, and the negative test (red) indicates the 1- specificity. Greater the AUC and closer the curve is to the upper-left hand corner, the better the test is at distinguishing between diseased and non-diseased people. The AUC, which can be any value between 0 and 1, is a good indicator of the test's accuracy.

4.3.1.3 Combinations of IBD-CRC biomarkers

After examining the expression and diagnostic values of each of the IBD and CRC markers individually for each condition, we chose the most significant biomarkers. IBD and CRC are both multifaceted diseases. As a result, no single marker will be able to predict disease activity in all cases. Therefore, the prospect of integrating several biomarkers and to increase accuracy is being investigated. We need to be ascertained in finding alternative biomarker combinations to reduce misdiagnosis of the disease at the early stage of progression from IBD to CRC. Therefore, to improve the accuracy of the diagnostic capability to detect IBD to CRC progression at its early stages, it was required to combine these biomarkers into various sets of combinations. For the application of the formulae used (section 3.3.10.13).

The (Table 4.12) shows three combined sets that is, combination of five markers (CEA + TIMP1+ CA724 + RUNX1+ CA199), four markers in two sets **a)** (CEA + TIMP1 + CA724 + RUNX1) and **b)** (CEA + CA724 + RUNX1 + CA199) and three markers set (CEA + CA724 + RUNX1). The outcome of these combinations is as followed. In combination set of five markers the sensitivity increased to 97 % and specificity to 99 %, which was found to be of very high diagnostic value among all the sets of combinations. Whereas, in a combination set of four biomarkers, in **a**) CEA + TIMP1 + CA724 + RUNX1, the sensitivity is 96.5 % and specificity is 99.9 %, which once again proven to be of high accuracy, while in **b**) CEA + CA724 + RUNX1 + CA199, the sensitivity was evaluated to be 94 % and specificity 99 %. In the last set of the three combined biomarkers CEA + CA724 + RUNX1, the sensitivity is 93 % and specificity is 99.7 %.

In summary, all the combinations sets were ranked at the highest diagnostic accuracy in predicting early diagnosis of IBD progressing to CRC in patients. In the clinical settings, the cost of the tests is also an important factor, therefore, from this data it can be concluded that the combination of four markers set (CEA + TIMP1 + CA724 + RUNX1) have the diagnostic capability similar to the five markers set. Therefore, to reduce the cost and time during testing, this four-markers set can be employed without compromising the diagnostic efficiency.

Table 4. 12 Evaluation and combinations of biomarkers from IBD-CRC.

Combinations	**Markers**	**Sensitivity**	**Specificity**
Single Marker	CEA	46.59%	80%
	TIMP1	52%	60%
	CA724	44.80%	97%
	RUNX1	75%	66%
	CA199	14.39%	89%
Five Markers	CEA + TIMP1 + CA724 + RUNX1 + CA199	97%	99%
Four Markers (a) **(Excluded:** CA199**)**	CEA + TIMP1 + CA724 + RUNX1	96.5%	99.9%
(b) (Excluded: TIMP1**)**	CEA + CA724 + RUNX1 + CA199	94%	99%
Three Markers **(Excluded:** TIMP1, CA199)	CEA + CA724 + RUNX1	93%	99.7%

4.3.1.4 Survival analysis of biomarkers predicting progression from IBD to CRC in patients' samples

GEPIA2 was used to analyse the relationship between different biomarkers of IBD – CRC and OS and DFS using Kaplan-Meier plotter curve in CRC COAD-READ database. The parameters used to analyse expression of combinations of biomarkers was implemented using (multi gene signature), hazard ratio measured based on Cox PH Model at 95% CI. The five biomarkers grouped together as a signature (multi gene) expressed in OS showed that the patients with high expression of these biomarkers were primarily associated with poor prognosis of shorter OS based on the actual graph (p-value = 0.2) as shown in (Fig 4.21 A). The expression between DFS was also evaluated. It was found that patients with high expression level of these biomarkers in their blood were somewhat associated with very short DFS (p-value = 0.028) in (Fig 4.21 B). The p-values do not show significant association; however, it is expected that the expression of these biomarkers play a crucial role in the prognosis of patients with CRC and may become a reliable predictor of survival in these patients.

In the clinical settings, when choosing a biomarker for clinical use, the cost of testing is an important factor to consider. From this information it can be concluded that the combination of four markers set (CEA + TIMP1 + CA-724 + RUNX1) has equivalent capabilities to the five markers set. Therefore, to reduce the cost and time during testing, this four-marker set can be employed without compromising the diagnostic capabilities.

Considering the above information, it was found that the 4 biomarkers (CEA + TIMP1 + CA-724 + RUNX1) when grouped together as a multi-gene signature expressed in OS, showed that patients with high expression of these biomarkers were primarily associated with poor prognosis of shorter OS based on the actual graph (p-value = 0.076) as shown in (Fig 4.22 A). The expression of DFS was evaluated and was found that patients with high expression levels of these biomarkers in their blood were significantly associated with short DFS (p-value = 0.014) (Fig 4.22 B). Outcome of this set of four markers suggested that the p-values do not show significant association however, based on the actual survival graphs especially in the later months, it is expected that the expressions of these biomarkers may play a role in the prognosis of disease in patients with CRC better than the five-biomarker set and their combination can become a reliable predictor of survival in these patients.

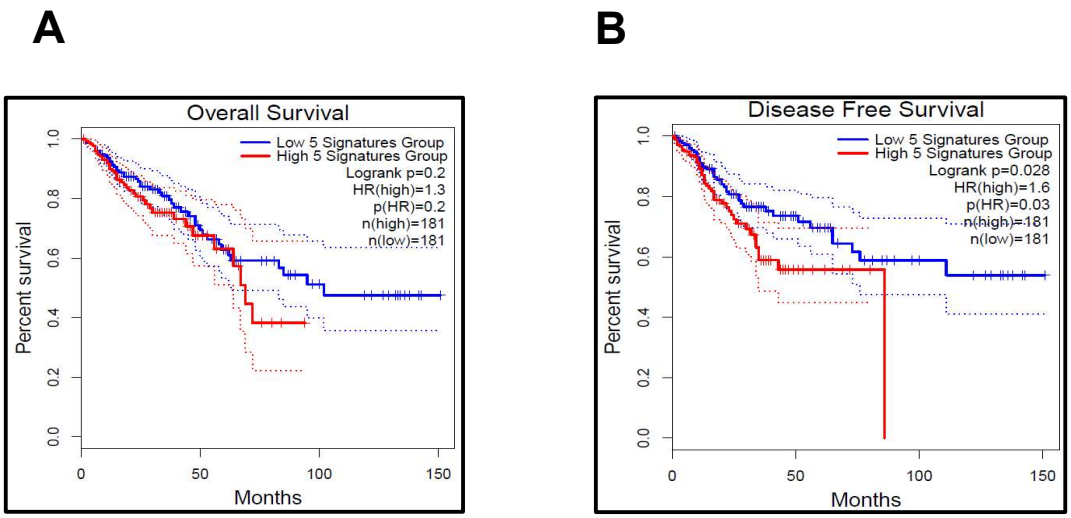


Figure 4. 21 The combination of 5 biomarkers' expression was associated with **(A)** worse OS, hazard ratio (HR) = 1.3, log rank P = 0.2 and **(B)** worse DFS, HR = 1.6, log rank P = 0.028. One-way ANOVA method was used for differential analysis *p-value Cutoff ≤ 0.01.

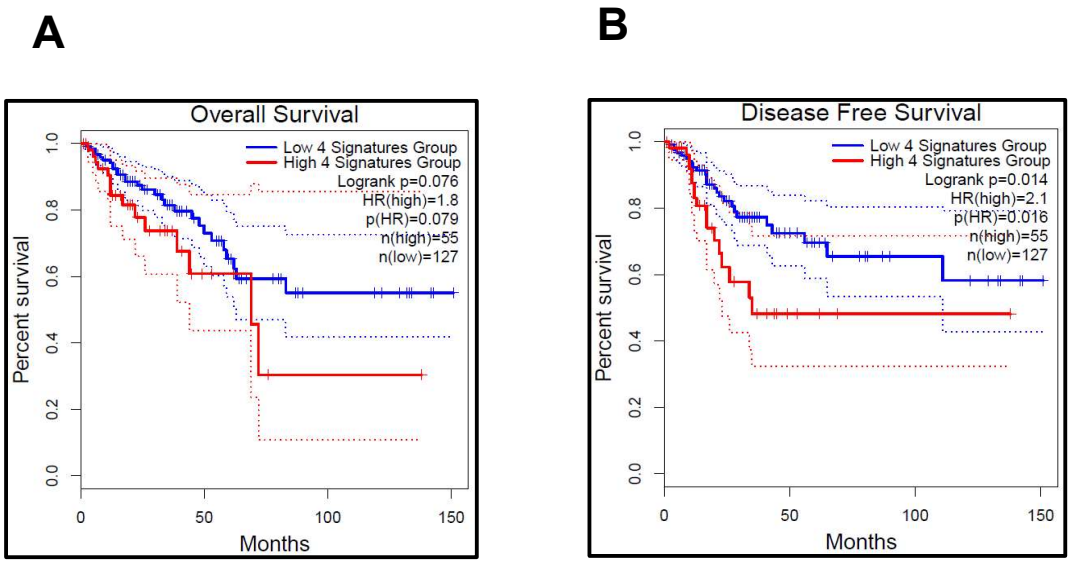


Figure 4. 22 The combination of 4 biomarkers' expression was associated with **(A)** worse OS, hazard ratio (HR) = 1.8, log rank P = 0.076 and **(B)** worse DFS, HR = 2.1, log rank P = 0.014. One-way ANOVA method was used for differential analysis *p-value Cutoff ≤ 0.01.

5 CHAPTER FIVE – DISCUSSION

The goal of the work presented in this thesis is to deploy computational methods that will assist in identifying new diagnostic markers for IBD and to predict progression from IBD to CRC. This thesis focuses on the fulfilment of three aims.

The study's first goal is to create a manually curated IBD related genes database that has been experimentally validated in the lab and is selected from published literature using the PubMed library, as well as to investigate the function of these genes. To the best of my knowledge no other database of curated IBD related genes exists. This database is embedded with some unique features and tools along with being a user-friendly database, for scientists to work with and make meaningful biological conclusions.

The second aim focused on the computational identification and validation of biomarkers for early and less-invasive diagnosis and screening of IBD. Here the TFs were identified. Potentially controlling the IBD related genes. The TFBSs were predicted and were mapped to mammalian matrices models like PWMs and generating information to understand at the molecular level, and to identify the underlying aetiology of IBD using GO annotation and pathways analysis. The TRNs of IBD genes were constructed. Further *in silico* validations of these TFs was done, and the TFs with most accurate diagnostic value were selected as IBD biomarkers and their gene expression profiling were elucidated in IBD, CD, UC patients' samples and were juxtaposed with CRC patients' samples during early stages. Hence the most suitable combinations of biomarkers were identified.

The third aim focused on the development of potential biomarkers that can measure and predict early risk of CRC in IBD patients. In view of this the known biomarkers of CRC currently in use in clinical settings were identified and evaluated for their diagnostic potential using various computational parameters. These individual biomarkers did not possess the optimal diagnostic values. Therefore, a combination analysis was performed. These CRC biomarkers along with the chosen biomarker of IBD were combined to increase the diagnostic values to predict better prognosis of the disease. The findings of the study are discussed in depth in the following sections.

5.1.1 Creation of manually curated IBD genes database

IBD is a complex disorder of GI tract. Recently IBD has become a global phenomenon as it have expanded over the world (Goh and XIAO, 2009). With the emerging data of IBD genes scattered in several publications for the past decades, it has become a very overwhelming issue. Therefore, it is of paramount importance to create a curated database of IBD (IBDDB) (Khan et al., 2021) genes which are experimentally validated under laboratory settings. To create a database, we manually curated 289 IBD genes using published literature from PubMed library.

This database is capable of performing various tasks. It can generate novel hypothesis which is based on Swanson linking theory (Swanson, 1986). The hypothesis that we used was based on previously published literature in PubMed literature. We wanted to connect lipid profiles to IBD and further explore the role of lipids particularly cholesterol in IBD and to find out hypothetically if these affected IBD related conditions such as PG (O'Brien et al., 2020). PG is a skin an ulceration disease (Maronese et al., 2022). We then explored the other capability of the database to create network construction using 11 dictionaries and biomedical entities. The novel hypothesis generated can be visualized under network construction. KB index helps in identification of the enriching terms and its linking with each other in creating novel hypothesis. KB is based on DES system that relies on two key components: dictionary that regulate the vocabularies and text mining using PubMed literature records. DES system is a reliable system found helpful in various published literature (Essack et al., 2009; Sagar et al., 2008).

This database was developed in SQLite, which is an in-process library that created a serverless, transactional SQL database engine that is self-contained and requires no configuration (Kreibich, 2010). SQLite's source code is available in the public domain that is an open-source database and can be used for both personal and business purposes. SQLite reads and writes to regular disk files directly. PHP, a widely used open-source scripting language executed on this server (Chen and Xie, 2008). The suffix 'lite' in SQLite means it is weightless in terms of setup, database management and required resources (Bhosale et al., 2015).

It is distinctive from other databases in several aspects: i)This is the first manually curated database providing extended information about experimentally validated IBD related genes; ii) it provides pre-compiled biomedical text mining information on IBD, which otherwise require complicated computational analyses; iii) it integrates data on 34 concepts such as experimental

techniques used to validate the role of a gene in IBD, sites inflamed in IBD, other diseases linked to the genes, tissue samples used for validation, molecular interactions, pathways, GO, gene expression level in mice or humans, TFs predicted to have binding sites on the promoters of IBD-implicated genes using four different tools, etc.; iv) it contains data on drugs associated with IBD genes; v) it has the hypothesis generation capability by creating networks of genes and other biomedical entities using 11 curated dictionaries; and vi) it can present networks in six different layouts while data can be exported in various formats such as excel, csv and pdf, as well as printing. Despite these benefits, IBDDB has certain limitations when it comes to text mining information (Khan et al., 2021). Text mining information is limited to electronic records alone, but extracted concepts are constrained by the dictionaries' completeness, and co-occurred phrases may or may not imply significant linkage. The accuracy of text mined data could be improved by further advancements in text mining technologies along with uniformity of information presentation in publications. Nonetheless, IBDDB can assist in the generation of new knowledge by assisting in the formation of new potential interactions between terms.

In summary, IBDDB allows users to explore information through various queries, allowing them to look for the most important items or construct linkages between them. It can also come up with intriguing hypotheses, build interactive networks, and export the results. We hope IBDDB, will be a valuable resource for researchers and clinicians.

5.1.2 Identification of Biomarkers of IBD and their validation

Identification of potential biomarkers of IBD genes was carried out using the computational analysis and bioinformatics applications. We identified the potential TFs controlling the IBD related genes.

We identified seven TFs of IBD (SPIB, FEV, ELF5, SPI1, NFKB1, RUNX1, RELA) as a putative biomarker after performing various bioinformatics analysis. Our emphasis from the beginning was to identify and validate biomarkers which should be non-invasive, inexpensive, reliable, with high sensitivity, allowing early diagnosis and thus no interfere in values among both affected patients and controls, which follows the characteristics of an ideal biomarker (Harrison et al., 2020). We investigated using *in silico* validation of these biomarkers in patients' samples. Furthermore, we evaluated the performance of these biomarkers by applying

various parameters of diagnostic tests. The outcome of this analysis was to identify the potential biomarker of IBD disease that can predict early onset of IBD in patients. We found RUNX1 as a potential TF-based biomarker of IBD in comparison with other predicted biomarkers of IBD. Based on the fisher exact test ranking, it was found to have high Fisher score value of 12.942 and it regulates 106 IBD genes. In a study it was reported that Fisher's exact test is a popular statistical significance test for determining the importance of overlap between two sets. It's an exact test in which the significance of the deviance from the null hypothesis is measured rather than predicted using Chi-squared analogue (Fisher, 1922). The outcome of the clinical assays suggested RUNX1 have high sensitivity of 77% and high specificity of 66% than other predicted biomarkers which either have low sensitivity and very high specificity or very high sensitivity and very low specificity. It is reported that the validity and accuracy of diagnostic tests are determined by several parameters when combined. Sensitivity and specificity are the two examples of such variables (Swift et al., 2020). We further evaluated RUNX1 in comparison with other biomarkers of IBD by performing other clinical tests like LR, PPV, NPV, accuracy, pre-test and post-test probability odds and AUC and found RUNX1 scored the highest test values among all the other biomarkers. Finally, calculating AUC values for all, it was observed that RUNX1 occupy 0.74 AUC value and other biomarkers outcome was quite less in comparison to RUNX1. In a statistical report for clinical evaluation of diagnostic tests, it is suggested that AUC range between 0.9 – 1.0 is excellent diagnostic biomarker, 0.8 - 0.9 is good, 0.7 – 0.8 is fair, 0.6 – 0.7 is poor and 0.5 – 0.6 is fail diagnostic value. Therefore, RUNX1 possess between fair to good diagnostic value (Wang et al., 2019; Xia et al., 2013).

For the biomarker to be non-invasive we found that RUNX1 can be considered as a non-invasive biomarker as it is found in blood platelets while other IBD biomarkers were found in tissues. In a study it was reported that RUNX1 variant showed association with thiopurine-induced leukopenia (decrease in white blood cells) in patients with IBD, along with two other markers causing obesity (Kim et al., 2017). Due to the lack of availability of non-invasive biomarker, the only available diagnostic biomarker to predict the disease IBD and CRC in patients is colonoscopy which is considered as a gold standard diagnostic biomarker (Ricciuto et al., 2021). It has its several consequences, as patients have to undergo anaesthesia, especially in children and bowel preparation a day before the procedure was found to be quite complicated (Hart et al., 2018). The after effects in some paediatric patients found was abdominal pain and

nausea after the procedure (Szaflarska-Popławska et al., 2019). Several non-invasive biomarkers of IBD are serum biomarkers like CRP, ESR, pANCA and ASCA biomarkers (Alghoul et al., 2022). CRP indicated patients need to have colectomy with a high degree of certainty, indicating severe continuing and unpredictable inflammation in the GI tract (O'Connell et al., 2020). While another broadly adopted biomarker in the diagnosis of IBD is ESR, which is reported to be higher in active CD than in dormant UC and CD (Longo et al., 2020). pANCA and ASCA help to differentiate between UC and CD in paediatric patients (Horn et al., 2018). But because of low sensitivity of pANCA it is often found to be increased in UC and also in CD patients with UC like symptoms, therefore limiting the reliability (Wang et al., 2019). With the low sensitivity of both ASCA and pANCA biomarkers many studies reported that they are unable to differentiate between UC and CD as oppose to their characteristics and were considered unreliable (Joossens et al., 2002; Lee et al., 2019b). In faecal biomarkers like calprotectin, it is noted that even though faecal calprotectin levels decrease over time, clinicians advised UC patients to take stool samples of their first bowel excretion every two days and note the consistency and time of the faeces. Its role in continuous monitoring while enhancing its deep-rooted and extended consequences has yet to be thoroughly investigated as mentioned in clinician's guidelines for using calprotectin to monitor activity of disease (Bressler et al., 2015). While lactoferrin is found in various bodily fluids and serves as a bactericidal agent in the innate immune system, it degranulates secondary granules throughout the inflammatory process (Baveye et al., 1999; Burton and Roitt, 2016). However, considering its limitation like calprotectin, faecal lactoferrin may increase in amount after using NSAIDs, most likely due to the intestinal obstruction induced, hence clinicians advise against using any non-steroidal medicine before the faecal test for IBD (Gisbert et al., 2009; Patil et al., 2019). Despite the fact that faecal biomarkers are affordable and non-invasive, they do have certain drawbacks, such as the test's clinical relevance, which is limited by the need for gut washout to precisely measure stool proteins (Angriman et al., 2007; Pham and Mohajeri, 2018). Considering all the limitations and unreliability issues of the serum and faecal biomarkers of IBD, we anticipated on validating the TF-based biomarkers which can be predicted in blood and is reliable, affordable, and easy to use. We combined the biomarkers of IBD and found effective diagnostic efficacy by combining the set of five biomarkers (SPIB + ELF5 + SPI1 + FEV + RUNX1).

This biomarker set of five has sensitivity of 99% and specificity of 95%. This can be a potential set of biomarkers for predicting early risk in IBD patients but unfortunately not all the biomarkers are found in blood as many are detected in tissue biopsy. Thus, considering the reliability and non-invasiveness in biomarker, we identified RUNX1 that is found in blood platelets and is also responsible for leukopenia in IBD patients (Kim et al., 2017).

The root cause of IBD still remain ambiguous (Mentella et al., 2020). We further tried to delineate the underlying mechanism of IBD. We found that the functional annotation clustering of these putative TF- based biomarkers of IBD genes emphasized on the enriched terms such as cytokine production, and pathways such as NOD-like receptor signalling pathway, cytokine-cytokine receptor. It is reported that modulation of gut immune system relies completely on cytokines functions, and are produced by monocytes, lymphocytes, macrophages, IECs and fibroblasts (Guan, 2019). Pro-inflammatory cytokines played a crucial role in up-regulation and anti-inflammatory cytokines played in down-regulation of pathogenesis progression (Tatiya-Aphiradee et al., 2019). Speculation in addition to generating intestinal inflammation and accompanying symptoms such as diarrhoea, cytokines regulate extraintestinal manifestation and have systemic effects in IBD were made (Neurath, 2014). Since cytokines and proteins associated in cytokine signalling are encoded by many IBD genes (Graham and Xavier, 2020). It is reported that loss of function mutations in the genes producing IL-10 and the IL-10 receptor, in particular, have been linked to quite early onset of IBD, according to current research (Kotlarz et al., 2012). It can be suggested that cytokines can be the new crucial target for the therapy of inflammatory diseases (Pasqua et al., 2018).

In conclusion, colonoscopy is the only invasive form of diagnosis for IBD patients available today. In this thesis, we have investigated to find the non-invasive form of diagnosis to predict early risk of diagnosis in IBD patients. Taking into considerations the ideal biomarker characteristics that can be reliable, cost-effective, non-invasive and time efficient.

5.1.3 To identify biomarkers which can serve as potential diagnostic measure to predict early risk of CRC in IBD patients.

We identified known biomarkers of CRC currently in use in clinical settings for the early risk of progression of CRC from IBD. It was mandatory to identify the biomarkers of CRC that was already predicted and tested.

Patients with long-term severe colitis were found more likely to acquire CRC (Kameyama et al., 2018). UC patients had a risk of 5.7 times that of the general population, according to population-based research (Mann et al., 2022). Therefore, we identified the known potential biomarkers of CRC currently used as routine clinical marker in laboratory settings. These tumour biomarkers are CEA, TIMP1, CA199 (Rasmussen et al., 2020) and CA724 for CRC (Wu et al., 2020b). From our findings we observed that sensitivity and specificity of CEA is 46% and 80%, TIMP1 is 52% and 60%, CA724 is 45% and 97%, CA199 is 14.4% and 89%, respectively. Apparently due to their low sensitivity and the widespread use of these CRC biomarkers in clinical settings, it became clear that none of them can be used alone to accurately diagnose CRC (TOMA et al., 2018). It was reported in a study that any available blood indicators, such as CEA, CA199, and CA724, are ineffective for early identification of CRC due to their low sensitivity and specificity but preoperative CEA assessment, on the other hand, is beneficial since it can provide autonomous prognostic information, assist with surgical administration, and serve as a baseline for later findings (Xie et al., 2022). It is recommended that CEA levels should be assessed in every few months at least for five years after diagnosis in patients with CRC stage 2 or Dukes B and 3 or Dukes C illness who may be candidates for liver resection (Duffy et al., 2003). It is found that CRC carcinomas can be classed as sporadic, inherited, or familial, depending on the origin of the mutation and that 70% of CRC is sporadic carcinoma (Mármol et al., 2017). Molecular genetics research has discovered certain unfavourable mutations that underpin the aetiology of both sporadic and hereditary CRC so far. As a result, screening tests should be more sensitive and specific than currently available screening tests, and the confirmation of a putative biomarker will offer genetic information on malignancy and metastatic progression (Das et al., 2017). Likewise, CEA that has the potential to detect CRC, TIMP1 can regulate apoptosis, angiogenesis and proliferation in an MMP-independent manner and have a role in CRC carcinogenesis, according to multiple compelling findings (Tuomisto et al., 2019). Many studies show that TIMP1 can be used as a prognostic biomarker for CRC

patients as well as a diagnostic marker for CRC detection (Meng et al., 2018; Sørensen et al., 2008). CA199 and CA724 are found on the surface of many cancer cells, including those with hepatocellular carcinoma, ovarian cancer, gastric cancer, and CRC. CA199, a carbohydrate antigen identified by the monoclonal antibody NS19-9, is extensively used in CRC clinics as a traditional tumour marker (Thurin, 2021). The diagnostic effectiveness of serum CA724 for CRC diagnosis has been investigated in several studies (Yanqing et al., 2018). CA724 and CEA were found to be potential prognostic markers, although Dukes' stage and rectal tumour site were also significant (Sun et al., 2019).

These tumour markers not only serve as auxiliary diagnosis for illness but also act as survival and prognostic tool (Luo et al., 2020). Our main goal was to identify a potential diagnostic tumour marker which is currently in clinical routine use. After finding the markers we aim at improving their diagnostic accuracy to detect early diagnosis of CRC in patients and to predict risk progression from IBD to CRC. To improve their accuracy, we combined these biomarkers in various sets of combinations like biomarker set of five markers (CEA + TIMP1 + CA724 + RUNX1 + CA199) with increased sensitivity of 97% and increased specificity of 99%, set of four markers i) (CEA + TIMP1 + CA724 + RUNX1) with increased sensitivity of 96.5% and increased specificity of 99.9% and ii) (CEA + CA724 + CA199 + RUNX1) with increased sensitivity of 94% and increased specificity of 99% and set of three markers (CEA + CA724 + RUNX1) with increased sensitivity of 93% and increased specificity of 99.7%. We evaluated their combination values and found the best set of combination of five biomarkers (CEA + TIMP1 + CA724 + RUNX1 + CA199) with the highest sensitivity and specificity value of 97% and 99%, respectively, to predict the early diagnosis in disease progression from IBD to CRC. But it can be inferred that the combination of four markers set (CEA + TIMP1 + CA-724 + RUNX1) has the diagnostic capability like the five markers set, which can be used instead, without compromising the diagnostic capabilities. Meanwhile, to predict early-stage of diagnosis in CRC patient’s samples all the biomarkers were found to be up regulated (Wu et al., 2020b) except RUNX1 and CA199 which was found in the result to be down regulated. On the contrary, a study reported RUNX1 to be up-regulated in CRC tissues and its variants to be associated with increased risk for colon and rectal carcinogenesis (Zhou et al., 2017b). it is reported that RUNX1 is considered as a novel tumour suppressor gene for gastrointestinal tumours and suggest that RUNX1 plays an important role in maintaining the balance between the intestinal stem/progenitor cell population and in GI tract epithelial development (Fijneman

et al., 2012). In a study we found that combination set of serum biomarkers was to be of profound importance in detecting early CRC (Liu et al., 2021). Based on the pathophysiology of IBD-associated CRC, CEA can predict the early risk of CRC in UC patients (Zhen et al., 2018). We found that with the combinations of biomarkers, we improved the diagnostic accuracy. Therefore, the selected combination markers (CEA + TIMP1 + CA724 + RUNX1) is set to be able to predict the early risk of progression from IBD to CRC.

5.1.4 Limitations of this study

The involvement of genes in IBD is certainly not limited to 289 genes but due to unavailability of experimentally verified genes, the IBDDB is limited in terms of exploring the genetics of IBD. Another serious limitation is that not many studies have been published that have explored the RNA-Seq of IBD samples (and even CRC samples) especially from Africa and the developed countries where this disease is on rise, therefore limiting the number of genes in database. In terms of TFBSs prediction, one is dependent of the available TF models, hence, we anticipate that the prediction will be more informative in the future when more validated models are added to the databases like JASPAR. An additional limitation to be considered is the inability to test the utility of these biomarkers on actual patients, especially those in African setting. This must be tested in future studies.

5.1.5 Strengths of this study

- Only available manually curated database of verified IBD genes. One stop for all the valuable information on IBD genes and its biological study.
- This database is embedded with some potential features like hypotheses generator, text-mined publications available based on three tier, 11 dictionaries along with 22000 biomedical entities, interactive network creator applications.
- All the genes are HGNC approved.
- These genes were data mined along with their evidence study from literature, phenotypes of disease, drug interactions, TFs, etc.
- Easily exportable formats like excel, csv, print, copy, pdf.
- The *in silico* approach using various bioinformatics tools like python, R scripting and availability of free software packages to pinpoint the predicted biomarkers.

- Web-based applications like GEPIA 2, ONCOMINE, OPOSSUM version 3.0, JASPAR, gene expression atlas, Enrichr, EMBTools calculators.

6 CHAPTER SIX- CONCLUSION AND FUTURE WORK

This is the first study that demonstrated the IBD gene database creation. This database is manually curated with experimentally verified genes under the laboratory settings, extracted from the published literature library PubMed library. It is embodied with various functionality that can enable users to create novel hypotheses by using hypotheses generator and create an interactive novel network construction with different layouts as graphical presentations. This database data mined 289 IBD genes along with its various biological features about those genes. It is a one stop shop for the researchers, biologist, and clinicians to interact with. Till data there is no available IBD genes related database for the users to gather information about it. This study also presents the potentially good biomarkers for diagnosis of IBD like TF RUNX1 and for predicting progression from IBD to CRC are (CEA+TIMP1+RUNX1+CA199+CA724) at the early stages, based on their *in silico* evaluation and validation in the patients' samples. These biomarkers of IBD and CRC when used in combinations can provide an excellent diagnosis accuracy with high sensitivity and specificity. The potential combination set to predict progression from IBD to CRC is (CEA + TIMP1 + CA724 + RUNX1). It is concluded that these biomarkers could serve as an ideal panel of biomarkers for diagnosis, prognosis, and predictive biomarkers for the IBD and IBD to CRC progression in a non-invasive way through blood work analysis.

Future validation studies of these biomarkers must be conducted in IBD and CRC patients' samples and *in vivo*.

The knowledge provided by this study may be useful for the clinicians and patients, as they will not have to undergo the endoscopic procedure and its tedious preparation before the surgery, which is an only available invasive way of diagnosis for IBD and CRC patients including both adults and paediatric patients. Economically, these biomarkers can be an ideal tool in providing inexpensive, non-invasive, and easy to measure diagnostic capabilities, especially in developing countries such as South Africa.

7 REFERENCES

Adam, H., Alqassas, M., Saadah, O.I., and Mosli, M. (2020). Extraintestinal manifestations of inflammatory bowel disease in middle eastern patients. Journal of Epidemiology and Global Health *10*, 298.

Adlerova, L., Bartoskova, A., and Faldyna, M. (2008). Lactoferrin: a review. Veterinarni Medicina *53*, 457-468.

Afrasiabi, A., Parnell, G.P., Swaminathan, S., Stewart, G.J., and Booth, D.R. (2020). The interaction of Multiple Sclerosis risk loci with Epstein-Barr virus phenotypes implicates the virus in pathogenesis. Scientific reports *10*, 1-11.

Agarwal, R., Kumar, B., Jayadev, M., Raghav, D., and Singh, A. (2016). CoReCG: a comprehensive database of genes associated with colon-rectal cancer. Database *2016*.

Agoda-Koussema, L., Anoukoum, T., Djibril, A., Balaka, A., Folligan, K., Adjenou, V., Amouzou, K., N'dakéna, K., and Redah, R. (2012). Ulcerative colitis: a case in Togo. Medecine et sante tropicales *22*, 79-81.

Agoritsas, T., Courvoisier, D.S., Combescure, C., Deom, M., and Perneger, T.V. (2011). Does prevalence matter to physicians in estimating post-test probability of disease? A randomized trial. Journal of general internal medicine *26*, 373-378.

Ahadi, M., Sokolova, A., Brown, I., Chou, A., and Gill, A.J. (2021). The 2019 World Health Organization Classification of appendiceal, colorectal and anal canal tumours: an update and critical assessment. Pathology *53*, 454-461.

Ahlawat, S., Kumar, P., Mohan, H., Goyal, S., and Sharma, K.K. (2021). Inflammatory bowel disease: tri-directional relationship between microbiota, immune system and intestinal epithelium. Critical Reviews in Microbiology *47*, 254-273.

Ahmed, R., Santhirakumar, K., Butt, H., and Yetisen, A.K. (2019). Colonoscopy technologies for diagnostics and drug delivery. Medical Devices & Sensors *2*, e10041.

Alatab, S., Sepanlou, S.G., Ikuta, K., Vahedi, H., Bisignano, C., Safiri, S., Sadeghi, A., Nixon, M.R., Abdoli, A., and Abolhassani, H. (2020). The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet gastroenterology & hepatology *5*, 17-30.

Alende-Castro, V., Alonso-Sampedro, M., Vazquez-Temprano, N., Tuñez, C., Rey, D., García-Iglesias, C., Sopeña, B., Gude, F., and Gonzalez-Quintela, A. (2019). Factors influencing erythrocyte sedimentation rate in adults: New evidence for an old test. Medicine (Baltimore) *98*, e16816.

Alghoul, Z., Yang, C., and Merlin, D. (2022). The Current Status of Molecular Biomarkers for Inflammatory Bowel Disease. Biomedicines *10*, 1492.

Allais, L., Verschuere, S., Maes, T., De Smet, R., Devriese, S., Gonzales, G.B., Peeters, H., Van Crombruggen, K., Bachert, C., and De Vos, M. (2020). Translational research into the effects of cigarette smoke on inflammatory mediators and epithelial TRPV1 in Crohn's disease. PloS one *15*, e0236657.

Almagro, U.A. (1983). Primary signet-ring carcinoma of the colon. Cancer *52*, 1453-1457.

Alzahrani, M.A., Alshahrani, M.A.M., Alshahrani, A.A.A., Shahrani, A.M.A., Alhumaidi, H.M.A., Alsharbi, H.S., Hammad, M.S., Nebrawi, K.Y., Alshehri, A.M., and Shehata, S.F. (2020). Clinical pattern of inflammatory bowel disease in Southern Saudi Arabia: a retrospective cohort study. population *9*, 12.

Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics *34*, 2483-2484.

Andersson, R.E., Olaison, G., Tysk, C., and Ekbom, A. (2001). Appendectomy and protection against ulcerative colitis. New England Journal of Medicine *344*, 808-814.

Andersson, R.E., Olaison, G., Tysk, C., and Ekbom, A. (2003). Appendectomy is followed by increased risk of Crohn's disease. Gastroenterology *124*, 40-46.

Angriman, I., Scarpa, M., D'Incà, R., Basso, D., Ruffolo, C., Polese, L., Sturniolo, G.C., D'Amico, D.F., and Plebani, M. (2007). Enzymes in feces: useful markers of chronic inflammatory bowel disease. Clinica chimica acta *381*, 63-68.

Annese, V. (2020). Genetics and epigenetics of IBD. Pharmacological research *159*, 104892.

Aoki, K., and Taketo, M.M. (2007). Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. Journal of cell science *120*, 3327-3335.

Arjomand Fard, N., Armstrong, H., Perry, T., and Wine, E. (2022). Appendix and Ulcerative Colitis: a Key to Explaining the Pathogenesis and Directing Novel Therapies? Inflammatory Bowel Diseases.

Armaghany, T., Wilson, J.D., Chu, Q., and Mills, G. (2012). Genetic alterations in colorectal cancer. Gastrointestinal cancer research: GCR *5*, 19.

Arnold, C.A., Graham, R.P., Jain, D., Kakar, S., Lam-Himlin, D.M., Naini, B.V., Wu, T.-T., Yeh, M.M., and Torbenson, M.S. (2019). Knowledge gaps in the appendix: a multi-institutional study from seven academic centers. Modern Pathology *32*, 988-996.

Arora, I., and Gupta, A. (2012). Cloud Databases: A Paradigm Shift in Databases.

Aschenbrenner, D., Quaranta, M., Banerjee, S., Ilott, N., Jansen, J., Steere, B., Chen, Y.-H., Ho, S., Cox, K., and Arancibia-Cárcamo, C.V. (2021). Deconvolution of monocyte responses

in inflammatory bowel disease reveals an IL-1 cytokine network that regulates IL-23 in genetic and acquired IL-10 resistance. Gut *70*, 1023-1036.

Azer, S.A. (2013). Overview of molecular pathways in inflammatory bowel disease associated with colorectal cancer development. European Journal of Gastroenterology & Hepatology *25*, 271-281.

Baeyens, J.-P., Serrien, B., Goossens, M., and Clijsen, R. (2019). Questioning the "SPIN and SNOUT" rule in clinical testing. Archives of physiotherapy *9*, 1-6.

Bagny, A., Lawson-Ananissoh, L.M., Kogoe, L.R.-M., Redah, D., and Kaaga, L.Y. (2021). Lower Gastrointestinal Bleeding in the Teaching Hospital Campus Of Lome (Togo). Journal of Gastroenterology and Hepatology Research *10*, 3441-3444.

Bakker, E.R., Hoekstra, E., Franken, P.F., Helvensteijn, W., van Deurzen, C.H., van Veelen, W., Kuipers, E.J., and Smits, R. (2013). β-Catenin signaling dosage dictates tissue-specific tumor predisposition in Apc-driven cancer. Oncogene *32*, 4579-4585.

Bantis, L.E., Tsimikas, J.V., Chambers, G.R., Capello, M., Hanash, S., and Feng, Z. (2021). The length of the receiver operating characteristic curve and the two cutoff Youden index within a robust framework for discovery, evaluation, and cutoff estimation in biomarker studies involving improper receiver operating characteristic curves. Statistics in Medicine *40*, 1767-1789.

Bargen, J.A. (1994). Chronic ulcerative colitis associated with malignant disease. 1928. Diseases of the colon and rectum *37*, 727-730.

Basson, A., Swart, R., Jordaan, E., Mazinu, M., and Watermeyer, G. (2014). The association between race and Crohn's disease phenotype in the Western Cape population of South Africa, defined by the Montreal Classification System. PloS one *9*, e104859.

Baveja, C., and Aggarwal, P. (2017). Statistical analysis of microbiological diagnostic tests. Indian Journal of Medical Microbiology *35*, 184-193.

Baveye, S., Elass, E., Mazurier, J., Spik, G., and Legrand, D. (1999). Lactoferrin: a multifunctional glycoprotein involved in the modulation of the inflammatory process.

Benchimol, E.I., Bernstein, C.N., Bitton, A., Murthy, S.K., Nguyen, G.C., Lee, K., Cooke-Lauder, J., Siddiq, S., Windsor, J.W., and Carroll, M.W. (2019). The impact of inflammatory bowel disease in Canada 2018: A scientific report from the Canadian Gastro-Intestinal Epidemiology Consortium to Crohn's and Colitis Canada. Journal of the Canadian Association of Gastroenterology *2*, S1-S5.

Bennett, M.R., and Devarajan, P. (2011). Characteristics of an ideal biomarker of kidney diseases. In Biomarkers of Kidney Disease, (Elsevier), pp. 1-24.

Benson, A.r., Schrag, D., Somerfield, M.R., Cohen, A.M., Figueredo, A.T., Flynn, P.J., Krzyzanowska, M.K., Maroun, J., McAllister, P., and Van Cutsem, E. (2004). American

Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. Journal of clinical oncology *22*, 3408-3419.

Benson, D., Lipman, D.J., and Ostell, J. (1993). GenBank. Nucleic acids research *21*, 2963-2965.

Berg, E.L., Magnani, J., Warnock, R.A., Robinson, M.K., and Butcher, E.C. (1992). Comparison of L-selectin and E-selectin ligand specificities: the L-selectin can bind the E-selectin ligands sialyl Le(x) and sialyl Le(a). Biochemical and biophysical research communications *184*, 1048-1055.

Berg, K.L., Seymour, T., and Goel, R. (2013). History of databases. International Journal of Management & Information Systems (IJMIS) *17*, 29-36.

Berkowitz, L., Schultz, B.M., Salazar, G.A., Pardo-Roa, C., Sebastián, V.P., Álvarez-Lobos, M.M., and Bueno, S.M. (2018). Impact of cigarette smoking on the gastrointestinal tract inflammation: opposing effects in Crohn's disease and ulcerative colitis. Frontiers in immunology *9*, 74.

Berrington, J. (2007). Databases. Anaesthesia & Intensive Care Medicine *8*, 513-515.

Bhosale, S., Patil, M.T., and Patil, M.P. (2015). Sqlite: Light database system. Int. J. Comput. Sci. Mob. Comput *44*, 882-885.

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Frontiers in genetics *7*, 24.

Borsotti, E., Barberio, B., D'Incà, R., Bonitta, G., Cavallaro, F., Spina, L., Pastorelli, L., Rondonotti, E., Samperi, L., and Dinelli, M. (2021). Low prevalence of colorectal neoplasia in microscopic colitis: A large prospective multi-center study. Digestive and Liver Disease *53*, 846-851.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., and De Vet, H.C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Radiology *226*, 24-28.

Brant, S.R., Okou, D.T., Simpson, C.L., Cutler, D.J., Haritunians, T., Bradfield, J.P., Chopra, P., Prince, J., Begum, F., and Kumar, A. (2017). Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease. Gastroenterology *152*, 206-217. e202.

Brant, S.R., and Shugart, Y.Y. (2004). Inflammatory bowel disease gene hunting by linkage analysis: rationale, methodology, and present status of the field. Inflammatory bowel diseases *10*, 300-311.

Braun, M., Klingelhöfer, D., Oremek, G.M., Quarcoo, D., and Groneberg, D.A. (2020). Influence of second-hand smoke and prenatal tobacco smoke exposure on biomarkers, genetics

and physiological processes in children—An overview in research insights of the last few years. International Journal of Environmental Research and Public Health *17*, 3212.

Bressler, B., Panaccione, R., Fedorak, R.N., and Seidman, E.G. (2015). Clinicians' guide to the use of fecal calprotectin to identify and monitor disease activity in inflammatory bowel disease. Canadian Journal of Gastroenterology and Hepatology *29*, 369-372.

Bruford, E.A., Antonescu, C.R., Carroll, A.J., Chinnaiyan, A., Cree, I.A., Cross, N.C., Dalgleish, R., Gale, R.P., Harrison, C.J., and Hastings, R.J. (2021). HUGO Gene Nomenclature Committee (HGNC) recommendations for the designation of gene fusions. Leukemia *35*, 3040-3043.

Burton, D.R., and Roitt, I.M. (2016). Adversarial strategies during infection. Roitt's Essential Immunology *380*, 321.

Butler, C. (1981). Surgical pathology of acute appendicitis. Human pathology *12*, 870-878.

Cagir, B., Nagy, M.W., Topham, A., Rakinic, J., and Fry, R.D. (1999). Adenosquamous carcinoma of the colon, rectum, and anus. Diseases of the Colon & Rectum *42*, 258-263.

Califf, R.M. (2018). Biomarker definitions and their applications. Experimental Biology and Medicine *243*, 213-221.

Camidge, D.R., Doebele, R.C., and Kerr, K.M. (2019). Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. Nature Reviews Clinical Oncology *16*, 341-355.

Campos-da-Paz, M., Dórea, J.G., Galdino, A.S., Lacava, Z.G., and de Fatima Menezes Almeida Santos, M. (2018). Carcinoembryonic antigen (CEA) and hepatic metastasis in colorectal cancer: update on biomarker for clinical and biotechnological approaches. Recent patents on biotechnology *12*, 269-279.

Can, G., Poşul, E., Yılmaz, B., Can, H., Korkmaz, U., Ermiş, F., Kurt, M., and Dağlı, Ü. (2019). Epidemiologic features of ınflammatory bowel disease in Western Blacksea region of Turkey for the last 10 years: retrospective cohort study. The Korean journal of internal medicine *34*, 519.

Carlomagno, N., Incollingo, P., Tammaro, V., Peluso, G., Rupealta, N., Chiacchio, G., Sandoval Sotelo, M.L., Minieri, G., Pisani, A., and Riccio, E. (2017). Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millennium: a breakthrough in gastric cancer. BioMed research international *2017*.

Carmichael, D.F., Sommer, A., Thompson, R.C., Anderson, D.C., Smith, C.G., Welgus, H.G., and Stricklin, G.P. (1986). Primary structure and cDNA cloning of human fibroblast collagenase inhibitor. Proceedings of the National Academy of Sciences *83*, 2407-2411.

Castagnola, P., and Giaretti, W. (2005). Mutant KRAS, chromosomal instability and prognosis in colorectal cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer *1756*, 115-125.

Chang, K., Taggart, M.W., Reyes-Uribe, L., Borras, E., Riquelme, E., Barnett, R.M., Leoni, G., San Lucas, F.A., Catanese, M.T., and Mori, F. (2018). Immune profiling of premalignant lesions in patients with lynch syndrome. JAMA oncology *4*, 1085-1092.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics *14*, 1-14.

Chen, J.-K., and Lee, W.-Z. (2019). An Introduction of NoSQL Databases based on their categories and application industries. algorithms *12*, 106.

Chen, K., Lu, P., Beeraka, N.M., Sukocheva, O.A., Madhunapantula, S.V., Liu, J., Sinelnikov, M.Y., Nikolenko, V.N., Bulygin, K.V., and Mikhaleva, L.M. (2020). Mitochondrial mutations and mitoepigenetics: Focus on regulation of oxidative stress-induced responses in breast cancers. (Elsevier).

Chen, P., Li, Y., Li, L., Yu, Q., Chao, K., Zhou, G., Qiu, Y., Feng, R., Huang, S., and He, Y. (2019). Circulating microRNA146b-5p is superior to C-reactive protein as a novel biomarker for monitoring inflammatory bowel disease. Alimentary Pharmacology & Therapeutics *49*, 733-743.

Chen, R., and Xie, J. (2008). Open Source Databases and Their Spatial Extensions.

Chi, M., Ma, K., Wang, J., Ding, Z., Li, Y., Zhu, S., Liang, X., Zhang, Q., Song, L., and Liu, C. (2021). The immunomodulatory effect of the gut microbiota in kidney disease. Journal of immunology research *2021*.

Choudhary, D., and Garg, P.K. (2013). 95% confidence interval: a misunderstood statistical tool. Indian Journal of Surgery *75*, 410-410.

Chu, K.M., Watermeyer, G., Shelly, L., Janssen, J., May, T.D., Brink, K., Benefeld, G., and Li, X. (2013). Childhood helminth exposure is protective against inflammatory bowel disease: a case control study in South Africa. Inflammatory bowel diseases *19*, 614-620.

Chung, W.-S., Chung, S., Hsu, C.-Y., and Lin, C.-L. (2021). Risk of Inflammatory Bowel Disease Following Appendectomy in Adulthood. Frontiers in Medicine *8*, 808.

Codd, E.F. (2007). Relational database: a practical foundation for productivity. In ACM Turing award lectures, (Association for Computing Machinery), pp. 1981.

Cohen, A.M., and Hersh, W.R. (2005). A survey of current work in biomedical text mining. Briefings in bioinformatics *6*, 57-71.

Costello, J.F., Frühwald, M.C., Smiraglia, D.J., Rush, L.J., Robertson, G.P., Gao, X., Wright, F.A., Feramisco, J.D., Peltomäki, P., and Lang, J.C. (2000). Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. Nature genetics *24*, 132-138.

Council, N.R. (2011). Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease.

Cui, G., and Yuan, A. (2018). A systematic review of epidemiology and risk factors associated with Chinese inflammatory bowel disease. Frontiers in Medicine *5*, 183.

D'Amico, F., Nancey, S., Danese, S., and Peyrin-Biroulet, L. (2021). A practical guide for faecal calprotectin measurement: Myths and realities. Journal of Crohn's and Colitis *15*, 152-161.

Daiss, W., Scheurlen, M., and Malchow, H. (1989). Epidemiology of inflammatory bowel disease in the county of Tübingen (West Germany). Scandinavian Journal of Gastroenterology *24*, 39-43.

Dalle-Donne, I., Rossi, R., Colombo, R., Giustarini, D., and Milzani, A. (2006). Biomarkers of oxidative damage in human disease. Clinical chemistry *52*, 601-623.

Das, V., Kalita, J., and Pal, M. (2017). Predictive and prognostic biomarkers in colorectal cancer: A systematic review of recent advances and challenges. Biomedicine & Pharmacotherapy *87*, 8-19.

De Maio, A.C., Basile, G., Iacopetta, D., Catalano, A., Ceramella, J., Cafaro, D., Saturnino, C., and Sinicropi, M.S. (2022). The significant role of nutraceutical compounds in ulcerative colitis treatment. Current Medicinal Chemistry.

De Palma, F.D.E., D'argenio, V., Pol, J., Kroemer, G., Maiuri, M.C., and Salvatore, F. (2019). The molecular hallmarks of the serrated pathway in colorectal cancer. Cancers *11*, 1017.

De Zoysa, P., Kingston-Smith, H., Maistry, P., Jayewardene, A.F., and Borody, T.J. (2017). Treatment-Naïve Ulcerative Colitis Patient Treated With Lyophilized Full Spectrum Microbiota: A Case Study: 1983. Official journal of the American College of Gastroenterology| ACG *112*, S1095-S1096.

Debunne, H., and Ceelen, W. (2013). Mucinous differentiation in colorectal cancer: molecular, histological and clinical aspects. Acta Chirurgica Belgica *113*, 385-390.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome biology *4*, 1-11.

Depledge, M.H. (2020). The rational basis for the use of biomarkers as ecotoxicological tools. In Nondestructive biomarkers in vertebrates, (CRC Press), pp. 271-295.

Derynck, R., Zhang, Y., and Feng, X.-H. (1998). Transcriptional activators of TGF-β responses: Smads. cell *95*, 737-740.

Dieckgraefe, B., Stenson, W., Korzenik, J., Swanson, P., and Harrington, C. (2000). Analysis of mucosal gene expression in inflammatory bowel disease by parallel oligonucleotide arrays. Physiological genomics *4*, 1-11.

Diggle, P.J. (2011). Estimating prevalence using an imperfect test. Epidemiology Research International *2011*.

Dodd, L.E., and Pepe, M.S. (2003). Partial AUC estimation and regression. Biometrics *59*, 614-623.

Dreos, R., Ambrosini, G., Périer, R.C., and Bucher, P. (2015). The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic acids research *43*, D92-D96.

Duffy, M., van Dalen, A., Haglund, C., Hansson, L., Klapdor, R., Lamerz, R., Nilsson, O., Sturgeon, C., and Topolcan, O. (2003). Clinical utility of biochemical markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines. European journal of cancer *39*, 718-727.

Dukes, C.E. (1932). The classification of cancer of the rectum. The Journal of Pathology and Bacteriology *35*, 323-332.

Dulai, P.S., Peyrin-Biroulet, L., Danese, S., Sands, B.E., Dignass, A., Turner, D., Mantzaris, G., Schölmerich, J., Mary, J.-Y., and Reinisch, W. (2019). Approaches to integrating biomarkers into clinical trials and care pathways as targets for the treatment of inflammatory bowel diseases. Gastroenterology *157*, 1032-1043. e1031.

Dunn, K.A., Moore-Connors, J., MacIntyre, B., Stadnyk, A.W., Thomas, N.A., Noble, A., Mahdi, G., Rashid, M., Otley, A.R., and Bielawski, J.P. (2016). Early changes in microbial community structure are associated with sustained remission after nutritional treatment of pediatric Crohn's disease. Inflammatory bowel diseases *22*, 2853-2862.

Eaden, J., Abrams, K., and Mayberry, J. (2001). The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut *48*, 526-535.

Ekbom, A., Helmick, C., Zack, M., and Adami, H.-O. (1990). Ulcerative colitis and colorectal cancer: a population-based study. New England Journal of Medicine *323*, 1228-1233.

Essack, M., Radovanovic, A., Schaefer, U., Schmeier, S., Seshadri, S.V., Christoffels, A., Kaur, M., and Bajic, V.B. (2009). DDEC: Dragon database of genes implicated in esophageal cancer. BMC cancer *9*, 1-7.

Ewing, I., Hurley, J.J., Josephides, E., and Millar, A. (2014). The molecular genetics of colorectal cancer. Frontline gastroenterology *5*, 26-30.

Ewongwo, A., Hamidi, M., Alattar, Z., Ayotunde, O.P., Tiwari, H.A., Elquza, E., Scott, A., Hanna, K., and Nfonsam, V. (2020). Contributing factors and short-term surgical outcomes of patients with early-onset rectal cancer. The American Journal of Surgery *219*, 578-582.

Fagan, E., Dyck, R., Maton, P., Hodgson, H., Chadwick, V., Petrie, A., and Pepys, M. (1982). Serum levels of C-reactive protein in Crohn's disease and ulcerative colitis. European journal of clinical investigation *12*, 351-359.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. cell *61*, 759-767.

Feinberg, A.P. (2004). The epigenetics of cancer etiology. In 6. (Elsevier), pp. 427-432.

Feinberg, A.P., and Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature *301*, 89-92.

Fijneman, R.J., Anderson, R.A., Richards, E., Liu, J., Tijssen, M., Meijer, G.A., Anderson, J., Rod, A., O'Sullivan, M.G., and Scott, P.M. (2012). Runx1 is a tumor suppressor gene in the mouse gastrointestinal tract. Cancer science *103*, 593-599.

Filella, X., Molina, R., Grau, J., Pique, J., Garcia-Valdecasas, J., Astudillo, E., Biete, A., Bordas, J., Novell, A., and Campo, E. (1992). Prognostic value of CA 19.9 levels in colorectal cancer. Annals of surgery *216*, 55.

Fiorini, N., Lipman, D.J., and Lu, Z. (2017). Cutting edge: towards PubMed 2.0. Elife *6*, e28801.

Fischer, J.E., Bachmann, L.M., and Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive care medicine *29*, 1043-1051.

Fisher, R.A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society *85*, 87-94.

Flint, H.J., Scott, K.P., Louis, P., and Duncan, S.H. (2012). The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol *9*, 577-589. 10.1038/nrgastro.2012.156.

Flores-Hernández, E., Velázquez, D.M., Castañeda-Patlán, M.C., Fuentes-García, G., Fonseca-Camarillo, G., Yamamoto-Furusho, J.K., Romero-Avila, M.T., García-Sáinz, J.A., and Robles-Flores, M. (2020). Canonical and non-canonical Wnt signaling are simultaneously activated by Wnts in colon cancer cells. Cellular Signalling *72*, 109636.

Fodde, R., Edelmann, W., Yang, K., Van Leeuwen, C., Carlson, C., Renault, B., Breukel, C., Alt, E., Lipkin, M., and Khan, P.M. (1994). A targeted chain-termination mutation in the mouse Apc gene results in multiple intestinal tumors. Proceedings of the National Academy of Sciences *91*, 8969-8973.

Forbes, A., Escher, J., Hébuterne, X., Kłęk, S., Krznaric, Z., Schneider, S., Shamir, R., Stardelova, K., Wierdsma, N., and Wiskin, A.E. (2017). ESPEN guideline: Clinical nutrition in inflammatory bowel disease. Clinical Nutrition *36*, 321-347.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., and Von Mering, C. (2012). STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research *41*, D808-D815.

Gadhave, D.G., and Kokare, C.R. (2019). Nanostructured lipid carriers engineered for intranasal delivery of teriflunomide in multiple sclerosis: optimization and in vivo studies. Drug development and industrial pharmacy *45*, 839-851.

Gaiani, F., Marchesi, F., Negri, F., Greco, L., Malesci, A., de'Angelis, G.L., and Laghi, L. (2021). Heterogeneity of colorectal cancer progression: molecular gas and brakes. International journal of molecular sciences *22*, 5246.

Gao, Y., Wang, J., Zhou, Y., Sheng, S., Qian, S.Y., and Huo, X. (2018). Evaluation of Serum CEA, CA19-9, CA72-4, CA125 and Ferritin as Diagnostic Markers and Factors of Clinical Parameters for Colorectal Cancer. Scientific reports *8*, 2732.

Garber, A., and Regueiro, M. (2019). Extraintestinal manifestations of inflammatory bowel disease: epidemiology, etiopathogenesis, and management. Current gastroenterology reports *21*, 1-13.

García-Gutiérrez, M.S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A., and Manzanares, J. (2020). Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality. Frontiers in psychiatry *11*, 432.

Gerritsen, J. (2015). The genus Romboutsia: genomic and functional characterization of novel bacteria dedicated to life in the intestinal tract. (Wageningen University and Research).

Gisbert, J.P., McNicholl, A.G., and Gomollon, F. (2009). Questions and answers on the role of fecal lactoferrin as a biological marker in inflammatory bowel disease. Inflammatory bowel diseases *15*, 1746-1754.

Glover, M., Mansoor, E., Panhwar, M., Parasa, S., and Cooper, G.S. (2019). Epidemiology of colorectal cancer in average risk adults 20–39 years of age: a population-based national study. Digestive diseases and sciences *64*, 3602-3609.

Goh, K., and XIAO, S.D. (2009). Inflammatory bowel disease: a survey of the epidemiology in Asia. Journal of digestive diseases *10*, 1-6.

Gold, P., and Freedman, S.O. (1965). Demonstration of tumor-specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. The Journal of experimental medicine *121*, 439-462.

Graham, D.B., and Xavier, R.J. (2020). Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature *578*, 527-539.

Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics.

Green, E.D., Watson, J.D., and Collins, F.S. (2015). Human Genome Project: Twenty-five years of big biology. Nature *526*, 29-31.

Greuter, T., Rieder, F., Kucharzik, T., Peyrin-Biroulet, L., Schoepfer, A.M., Rubin, D.T., and Vavricka, S.R. (2021). Emerging treatment options for extraintestinal manifestations in IBD. Gut *70*, 796-802.

Grimes, D.A., and Schulz, K.F. (2005). Refining clinical diagnosis with likelihood ratios. The Lancet *365*, 1500-1505.

Group, F.-N.B.W. (2016). BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet].

Guan, Q. (2019). A comprehensive review and update on the pathogenesis of inflammatory bowel disease. Journal of immunology research *2019*.

Gyde, S., Prior, P., Dew, M., Saunders, V., Waterhouse, J., and Allan, R. (1982). Mortality in ulcerative colitis. Gastroenterology *83*, 36-43.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine *4*, 627.

Halfvarson, J. (2019). Genetic epidemiology of inflammatory bowel disease, early twin and family studies. In Molecular genetics of inflammatory bowel disease, (Springer), pp. 25-45.

Halfvarson, J., Bodin, L., Tysk, C., Lindberg, E., and Järnerot, G. (2003). Inflammatory bowel disease in a Swedish twin cohort: a long-term follow-up of concordance and clinical characteristics. Gastroenterology *124*, 1767-1773.

Hamilton, S.R., and Aaltonen, L.A. (2000). Pathology and genetics of tumours of the digestive system (IARC press Lyon:).

Hammada, T., Lemdaoui, M.-C., Boutra, F., Zoughailech, D., and Asselah, H. (2011). Aspects épidémiologiques des maladies inflammatoires chroniques de l'intestin dans une population algérienne. Journal Africain d'Hépato-Gastroentérologie *5*, 293-302.

Hampe, J., Schreiber, S., Shaw, S.H., Lau, K.F., Bridger, S., Macpherson, A.J., Cardon, L.R., Sakul, H., Harris, T.J., and Buckler, A. (1999). A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. The American Journal of Human Genetics *64*, 808-816.

Hampel, H., Frank, R., Broich, K., Teipel, S.J., Katz, R.G., Hardy, J., Herholz, K., Bokde, A.L., Jessen, F., and Hoessler, Y.C. (2010). Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. Nature reviews Drug discovery *9*, 560-574.

Han, M.K., Anderson, R., Viennois, E., and Merlin, D. (2020). Examination of food consumption in United States adults and the prevalence of inflammatory bowel disease using National Health Interview Survey 2015. PloS one *15*, e0232157.

Haq, A.I., Schneeweiss, J., Kalsi, V., and Arya, M. (2009). The Dukes staging system: a cornerstone in the clinical management of colorectal cancer. The Lancet. Oncology *10*, 1128.

Harada, S., and Morlote, D. (2020). Molecular pathology of colorectal cancer. Advances in Anatomic Pathology *27*, 20-26.

Harrison, S.A., Ratziu, V., Boursier, J., Francque, S., Bedossa, P., Majd, Z., Cordonnier, G., Sudrik, F.B., Darteil, R., and Liebe, R. (2020). A blood-based biomarker panel (NIS4) for non-invasive diagnosis of non-alcoholic steatohepatitis and liver fibrosis: a prospective derivation and global validation study. The lancet Gastroenterology & hepatology *5*, 970-985.

Hart, L., Nael, H., Longmire, N.M., and Zachos, M. (2018). Barriers and facilitators to a good bowel preparation for colonoscopy in children: a qualitative study. Journal of Pediatric Gastroenterology and Nutrition *67*, 188-193.

Hartwig, O., Boushehri, M.A.S., Shalaby, K.S., Loretz, B., Lamprecht, A., and Lehr, C.-M. (2021). Drug delivery to the inflamed intestinal mucosa–targeting technologies and human cell culture models for better therapies of IBD. Advanced Drug Delivery Reviews *175*, 113828.

Hayden, S.R., and Brown, M.D. (1999). Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. Annals of emergency medicine *33*, 575-580.

Hazra, A. (2017). Using the confidence interval confidently. Journal of thoracic disease *9*, 4125.

He, B., and Tan, K. (2016). Understanding transcriptional regulatory networks using computational models. Current opinion in genetics & development *37*, 101-108. 10.1016/j.gde.2016.02.002.

Hearst, M. (2003). What is text mining. SIMS, UC Berkeley *5*.

Hestand, M.S., van Galen, M., Villerius, M.P., van Ommen, G.-J.B., den Dunnen, J.T., and t Hoen, P.A. (2008). CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. BMC bioinformatics *9*, 1-11.

Heuthorst, L., Mookhoek, A., Wildenberg, M.E., D'Haens, G.R., Bemelman, W.A., and Buskens, C.J. (2021). High prevalence of ulcerative appendicitis in patients with ulcerative colitis. UEG Journal *9*, 1148-1156.

Hisamatsu, T., Kanai, T., Mikami, Y., Yoneno, K., Matsuoka, K., and Hibi, T. (2013). Immune aspects of the pathogenesis of inflammatory bowel disease. Pharmacology & therapeutics *137*, 283-297.

Ho Sui, S.J., Mortimer, J.R., Arenillas, D.J., Brumm, J., Walsh, C.J., Kennedy, B.P., and Wasserman, W.W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res *33*, 3154-3164.

Hodges, P., and Kelly, P. (2020). Inflammatory bowel disease in Africa: what is the current state of knowledge? International Health *12*, 222-230.

Hogezand, R.v., Eichhorn, R., Choudry, A., Veenendaal, R., and Lamers, C. (2002). Malignancies in inflammatory bowel disease: fact or fiction? Scandinavian Journal of Gastroenterology *37*, 48-53.

Holliday, R. (1987). The inheritance of epigenetic defects. science *238*, 163-170.

Holmes, M.V., Richardson, T.G., Ference, B.A., Davies, N.M., and Davey Smith, G. (2021). Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. Nature Reviews Cardiology *18*, 435-453.

Hong, S.N., Joung, J.-G., Bae, J.S., Lee, C.S., Koo, J.S., Park, S.J., Im, J.P., Kim, Y.S., Kim, J.W., Park, W.Y., and Kim, Y.-H. (2017). RNA-seq Reveals Transcriptomic Differences in Inflamed and Noninflamed Intestinal Mucosa of Crohn's Disease Patients Compared with Normal Mucosa of Healthy Controls. Inflammatory bowel diseases *23*, 1098-1108.

Horn, M.P., Peter, A.M., Righini Grunder, F., Leichtle, A.B., Spalinger, J., Schibli, S., and Sokollik, C. (2018). PR3-ANCA and panel diagnostics in pediatric inflammatory bowel disease to distinguish ulcerative colitis from Crohn's disease. PLoS One *13*, e0208974.

Hornebeck, W., Lambert, E., Petitfrère, E., and Bernard, P. (2005). Beneficial and detrimental influences of tissue inhibitor of metalloproteinase-1 (TIMP-1) in tumor progression. Biochimie *87*, 377-383.

Hoskins, L.C. (2018). Mucin degradation by enteric bacteria: ecological aspects and implications for bacterial attachment to gut mucosa. In Attachment of organisms to the gut mucosa, (CRC Press), pp. 51-68.

Hossain, M.S., Tonmoy, M.I.Q., Islam, M.N., Islam, M.S., Afif, I.K., Roy, A.S., Fariha, A., Al Reza, H., Bahadur, N.M., and Rahaman, M.M. (2021). MicroRNAs expression analysis shows key affirmation of Synaptopodin-2 as a novel prognostic and therapeutic biomarker for colorectal and cervical cancers. Heliyon *7*, e07347.

Huang, S., He, L., Yang, B., and Zhang, M. (2012). A compound correlation model for disjoint literature-based knowledge discovery. (Emerald Group Publishing Limited).

Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.-P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., and Gassull, M. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599-603.

Hundt, S., Haug, U., and Brenner, H. (2007). Blood markers for early detection of colorectal cancer: a systematic review. Cancer Epidemiology and Prevention Biomarkers *16*, 1935-1953.

Iida, T., Yokoyama, Y., Wagatsuma, K., Hirayama, D., and Nakase, H. (2018). Impact of autophagy of innate immune cells on inflammatory bowel disease. Cells *8*, 7.

Ingelfinger, J.A. (1994). Biostatistics in clinical medicine (McGraw-Hill).

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., and Webster, M. (2009). Genome-wide methylation analysis of human colon cancer reveals similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nature genetics *41*, 178.

Iskandar, H.N., and Ciorba, M.A. (2012). Biomarkers in inflammatory bowel disease: current practices and recent advances. Translational Research *159*, 313-325.

Izdebska, M., Zielińska, W., Grzanka, D., and Gagat, M. (2018). The role of actin dynamics and actin-binding proteins expression in epithelial-to-mesenchymal transition and its association with cancer progression and evaluation of possible therapeutic targets. BioMed research international *2018*.

Jeyakumar, T., Fodil, N., Van Der Kraak, L., Meunier, C., Cayrol, R., McGregor, K., Langlais, D., Greenwood, C.M., Beauchemin, N., and Gros, P. (2019). Inactivation of interferon regulatory factor 1 causes susceptibility to colitis-associated colorectal cancer. Scientific reports *9*, 1-13.

Jiang, C.G., Wang, Z.N., Sun, Z., Liu, F.N., Yu, M., and Xu, H.M. (2011). Clinicopathologic characteristics and prognosis of signet ring cell carcinoma of the stomach: Results from a chinese mono-institutional study. Journal of surgical oncology *103*, 700-703.

Jones, M.W., Lopez, R.A., and Deppen, J.G. (2021). Appendicitis. In StatPearls [Internet], (StatPearls Publishing).

Joossens, S., Reinisch, W., Vermeire, S., Sendid, B., Poulain, D., Peeters, M., Geboes, K., Bossuyt, X., Vandewalle, P., and Oberhuber, G. (2002). The value of serologic markers in indeterminate colitis: a prospective follow-up study. Gastroenterology *122*, 1242-1247.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., and Anderson, C.A. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119-124.

Jung, G., Hernández-Illán, E., Moreira, L., Balaguer, F., and Goel, A. (2020). Epigenetics of colorectal cancer: biomarker and therapeutic potential. Nature reviews Gastroenterology & hepatology *17*, 111-130.

Kaczmarek-Ryś, M., Hryhorowicz, S.T., Lis, E., Banasiewicz, T., Paszkowski, J., Borejsza-Wysocki, M., Walkowiak, J., Cichy, W., Krokowicz, P., and Czkwianianc, E. (2021). Crohn's Disease Susceptibility and Onset Are Strongly Related to Three NOD2 Gene Haplotypes. Journal of clinical medicine *10*, 3777.

Kadkhoda, S., Darbeheshti, F., Rezaei, N., Azizi-Tabesh, G., Zolfaghari, F., Tavakolibazaz, S., Taslimi, R., and Tavakkoly-Bazzaz, J. (2021). Investigation of circRNA-miRNA-mRNA network in colorectal cancer using an integrative bioinformatics approach. Gastroenterology and hepatology from bed to bench *14*, 141.

Kakodkar, S., Mikolaitis, S., Engen, P., and Mutlu, E. (2013). The Bacterial Microbiome of IBD Patients on the Specific Carbohydrate Diet (SCD): 1828. Official journal of the American College of Gastroenterology| ACG *108*, S552.

Kałużna, A., Olczyk, P., and Komosińska-Vassev, K. (2022). The Role of Innate and Adaptive Immune Cells in the Pathogenesis and Development of the Inflammatory Response in Ulcerative Colitis. Journal of Clinical Medicine *11*, 400.

Kameyama, H., Nagahashi, M., Shimada, Y., Tajima, Y., Ichikawa, H., Nakano, M., Sakata, J., Kobayashi, T., Narayanan, S., and Takabe, K. (2018). Genomic characterization of colitis-associated colorectal cancer. World journal of surgical oncology *16*, 1-6.

Kanazawa, T., Watanabe, T., Kazama, S., Tada, T., Koketsu, S., and Nagawa, H. (2002). Poorly differentiated adenocarcinoma and mucinous carcinoma of the colon and rectum show higher rates of loss of heterozygosity and loss of E-cadherin expression due to methylation of promoter region. International journal of cancer *102*, 225-229.

Kaplan, G.G., and Windsor, J.W. (2021). The four epidemiological stages in the global evolution of inflammatory bowel disease. Nature reviews Gastroenterology & hepatology *18*, 56-66.

Katsanos, A.H., Kyriakidi, K., Karassa, F.B., Politis, D., Skamnelos, A., Christodoulou, D.K., and Katsanos, K.H. (2017). Biomarker Development in Chronic Inflammatory Diseases. Biomarkers for Endometriosis: State of the Art, 41-75.

Kaur, A., and Goggolidou, P. (2020). Ulcerative colitis: understanding its cellular pathology could provide insights into novel therapies. Journal of inflammation *17*, 1-8.

Kelly, R.S., Vander Heiden, M.G., Giovannucci, E., and Mucci, L.A. (2016). Metabolomic Biomarkers of Prostate Cancer: Prediction, Diagnosis, Progression, Prognosis, and RecurrenceMetabolomic Biomarkers of Prostate Cancer. Cancer Epidemiology, Biomarkers & Prevention *25*, 887-906.

Kesavelu Sr, D. (2020). The efficacy and safety of combined senna and probiotic-based bowel preparation for colonoscopy in children. Cureus *12*.

Keum, N., and Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol *16*, 713-732.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., Van Der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., and Tan, G. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research *46*, D260-D266.

Khan, A.Q., Kuttikrishnan, S., Siveen, K.S., Prabhu, K.S., Shanmugakonar, M., Al-Naemi, H.A., Haris, M., Dermime, S., and Uddin, S. (2019). RAS-mediated oncogenic signaling pathways in human malignancies. (Elsevier), pp. 1-13.

Khan, F., Radovanovic, A., Gojobori, T., and Kaur, M. (2021). IBDDB: a manually curated and text-mining-enhanced database of genes involved in inflammatory bowel disease. Database *2021*.

Kheirelseid, E.A., Miller, N., and Kerin, M.J. (2013). Clinical applications of molecular profiling in colorectal cancer: review of the literature.

Kim, H.S., Cheon, J.H., Jung, E.S., Park, J., Aum, S., Park, S.J., Eun, S., Lee, J., Rüther, U., Yeo, G.S.H., et al. (2017). A coding variant in FTO confers susceptibility to thiopurine-induced leukopenia in East Asian patients with IBD. Gut *66*, 1926-1935.

Kim, W. (1990). Object-Oriented Databases: Definition and Research Directions. IEEE Trans. on Knowl. and Data Eng. *2*, 327–341.

Kinzler, K.W., and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. cell *87*, 159-170.

Kleinman, A., Eisenberg, L., and Good, B. (1978). Culture, illness, and care: clinical lessons from anthropologic and cross-cultural research. Annals of internal medicine *88*, 251-258.

Kobayashi, Y., Ohfuji, S., Kondo, K., Fukushima, W., Sasaki, S., Kamata, N., Yamagami, H., Fujiwara, Y., Suzuki, Y., and Hirota, Y. (2021). Association of dietary fatty acid intake with the development of ulcerative colitis: a multicenter case-control study in Japan. Inflammatory Bowel Diseases *27*, 617-628.

Kopylov, U., Rosenfeld, G., Bressler, B., and Seidman, E. (2014). Clinical utility of fecal biomarkers for the diagnosis and management of inflammatory bowel disease. Inflammatory bowel diseases *20*, 742-756.

Kotlarz, D., Beier, R., Murugan, D., Diestelhorst, J., Jensen, O., Boztug, K., Pfeifer, D., Kreipe, H., Pfister, E.D., and Baumann, U. (2012). Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy. Gastroenterology *143*, 347-355.

Koutroubakis, I.E., Vlachonikolis, I.G., and Kouroumalis, E.A. (2002). Role of appendicitis and appendectomy in the pathogenesis of ulcerative colitis: a critical review. Inflammatory bowel diseases *8*, 277-286.

Kouzminova, N., Lu, T., and Lin, A.Y. (2010). Molecular basis of colorectal cancer. The New England journal of medicine *362*, 1245-1246; author reply 1246.

Kraus, V.B. (2018). Biomarkers as drug development tools: discovery, validation, qualification and use. Nature Reviews Rheumatology *14*, 354-362.

Kraus, V.B., Burnett, B., Coindreau, J., Cottrell, S., Eyre, D., Gendreau, M., Gardiner, J., Garnero, P., Hardin, J., and Henrotin, Y. (2011). Application of biomarkers in the development of drugs intended for the treatment of osteoarthritis. Osteoarthritis and Cartilage *19*, 515-542.

Kreibich, J. (2010). Using SQLite (" O'Reilly Media, Inc.").

Kroll, W. (2008). Biomarkers–predictors, surrogate parameters–a concept definition. Biomarker. Stuttgart: Schattauer, 1-14.

Krupey, J., Wilson, T., Freedman, S.O., and Gold, P. (1972). The preparation of purified carcinoembryonic antigen of the human digestive system from large quantities of tumor tissue. Immunochemistry *9*, 617-622.

Kucharzik, T., Koletzko, S., Kannengiesser, K., and Dignass, A. (2020). Ulcerative colitis—diagnostic and therapeutic algorithms. Deutsches Ärzteblatt International *117*, 564.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., and Lachmann, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research *44*, W90-W97.

Kumar, G.R., Subazini, T.K., Subha, K., Rajadurai, C.P., and Prabakar, L. (2009). CanGeneBase (CGB)-a database on cancer related genes. Bioinformation *3*, 422.

Kvakova, M., Bertkova, I., Stofilova, J., and Savidge, T.C. (2021). Co-encapsulated synbiotics and immobilized probiotics in human health and gut Microbiota modulation. Foods *10*, 1297.

Lai, L.J., Shen, J., and Ran, Z.H. (2019). Natural killer T cells and ulcerative colitis. Cellular immunology *335*, 1-5.

Lakemeyer, L., Sander, S., Wittau, M., Henne-Bruns, D., Kornmann, M., and Lemke, J. (2021). Diagnostic and Prognostic Value of CEA and CA19-9 in Colorectal Cancer. Diseases *9*, 21.

Lalkhen, A.G., and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. Continuing education in anaesthesia critical care & pain *8*, 221-223.

Lamichhane, S., Sen, P., Dickens, A.M., Orešič, M., and Bertram, H.C. (2018). Gut metabolome meets microbiome: A methodological perspective to understand the relationship between host and microbe. Methods *149*, 3-12.

Landsheer, J.A. (2018). The clinical relevance of methods for handling inconclusive medical test results: quantification of uncertainty in medical decision-making and screening. Diagnostics *8*, 32.

Lane, D.P., and Benchimol, S. p53: oncogene or anti-oncogene.

Langsted, A., and Nordestgaard, B.G. (2019). Smoking is associated with increased risk of major bleeding: A prospective cohort study. Thrombosis and haemostasis *119*, 039-047.

Lao, V.V., and Grady, W.M. (2011). Epigenetics and colorectal cancer. Nature reviews Gastroenterology & hepatology *8*, 686-700.

Latchman, D.S. (1997). Transcription factors: an overview. The international journal of biochemistry & cell biology *29*, 1305-1312.

Leal, R.F., Ayrizono, M.d.L.S., Milanski, M., Fagundes, J.J., Moraes, J.C., Meirelles, L.R., Velloso, L.A., and Coy, C.S. (2010). Detection of epithelial apoptosis in pelvic ileal pouches for ulcerative colitis and familial adenomatous polyposis. Journal of Translational Medicine *8*, 1-6.

Lee, A.Y., Chataway, T., Colella, A.D., Gordon, T.P., and Wang, J.J. (2019a). Quantitative mass spectrometric analysis of autoantibodies as a paradigm shift in autoimmune serology. Frontiers in immunology *10*, 2845.

Lee, S.H., eun Kwon, J., and Cho, M.-L. (2018). Immunological pathogenesis of inflammatory bowel disease. Intestinal research *16*, 26.

Lee, W.-I., Subramaniam, K., Hawkins, C.A., and Randall, K.L. (2019b). The significance of ANCA positivity in patients with inflammatory bowel disease. Pathology *51*, 634-639.

Lees, C.W., Regueiro, M., and Mahadevan, U. (2020). Innovation in inflammatory bowel disease care during the COVID-19 pandemic: results of a global telemedicine survey by the International organization for the study of inflammatory bowel disease. Gastroenterology *159*, 805-808. e801.

Legrand, D., Elass, E., Carpentier, M., and Mazurier, J. (2005). Lactoferrin. Cellular and Molecular Life Sciences *62*, 2549-2559.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1997). Genetic instability in colorectal cancers. Nature *386*, 623-627.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature *396*, 643-649.

León, A.S., Bernstein, C.N., El-Gabalawy, H., and Eck, P. (2018). Variations in the IBD5 locus confer the risk of inflammatory bowel disease in a Manitoban Caucasian Cohort. Clin Nutr *5*, 1-6.

Leventhal, H., Nerenz, D.R., and Steele, D.J. (2020). Illness representations and coping with health threats. In Handbook of Psychology and Health (Volume IV), (Routledge), pp. 219-252.

Liang, Z., Yan, D., Li, G., and Cheng, H. (2018). Clinical analysis of primary colorectal signet-ring cell carcinoma. Clinical Colorectal Cancer *17*, e39-e44.

Liebl, M.C., and Hofmann, T.G. (2021). The role of p53 signaling in colorectal cancer. Cancers *13*, 2125.

Link, A., and Kupcinskas, J. (2018). MicroRNAs as non-invasive diagnostic biomarkers for gastric cancer: Current insights and future perspectives. World Journal of Gastroenterology *24*, 3313.

Liu, S., Ren, J., Wu, X., Ren, H., Yan, D., Wang, G., Gu, G., Li, J., Xia, Q., and Han, G. (2013). Preliminary case-control study to evaluate diagnostic values of C-reactive protein and erythrocyte sedimentation rate in differentiating active Crohn's disease from intestinal lymphoma, intestinal tuberculosis and Behcet's syndrome. The American journal of the medical sciences *346*, 467-472.

Liu, W., Ye, A., Han, F., and Han, J. (2019). Advances and challenges in liposome digestion: Surface interaction, biological fate, and GIT modeling. Advances in Colloid and Interface Science *263*, 52-67.

Liu, X., Jakubowski, M., and Hunt, J.L. (2011). KRAS gene mutation in colorectal cancer is correlated with increased proliferation and spontaneous apoptosis. American journal of clinical pathology *135*, 245-252.

Liu, Y., Liang, Y., and Wishart, D. (2015). PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic acids research *43*, W535-W542.

Liu, Z., Tang, H., Zhang, W., Wang, J., Wan, L., Li, X., Ji, Y., Kong, N., Zhang, Y., and Wang, J. (2021). Coupling of serum CK20 and hyper-methylated CLIP4 as promising biomarker for colorectal cancer diagnosis: from bioinformatics screening to clinical validation. Aging (Albany NY) *13*, 26161.

Lizio, M., Harshbarger, J., Abugessaisa, I., Noguchi, S., Kondo, A., Severin, J., Mungall, C., Arenillas, D., Mathelier, A., and Medvedeva, Y.A. (2016). Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. Nucleic acids research, gkw995.

Longo, S., Chieppa, M., Cossa, L.G., Spinelli, C.C., Greco, M., Maffia, M., and Giudetti, A.M. (2020). New insights into inflammatory bowel diseases from proteomic and lipidomic studies. Proteomes *8*, 18.

Louhimo, J., Carpelan-Holmström, M., Alfthan, H., Stenman, U.H., Järvinen, H.J., and Haglund, C. (2002). Serum HCGβ, CA 72-4 and CEA are independent prognostic factors in colorectal cancer. International journal of cancer *101*, 545-548.

Louis, P., Hold, G.L., and Flint, H.J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. Nature reviews. Microbiology *12*, 661-672.

Loureiro, A.C.C.F., and Barbosa, L.E.R. (2019). Appendectomy and Crohn's disease. Journal of Coloproctology (Rio de Janeiro) *39*, 373-380.

Lu, H., Guo, Y., Gupta, G., and Tian, X. (2019). Mitogen-activated protein kinase (MAPK): new insights in breast cancer. Journal of Environmental Pathology, Toxicology and Oncology *38*.

Lunshof, J.E., Bobe, J., Aach, J., Angrist, M., Thakuria, J.V., Vorhaus, D.B., Hoehe, M.R., and Church, G.M. (2022). Personal genomes in progress: from the human genome project to the personal genome project. Dialogues in clinical neuroscience.

Luo, C.-X., Wen, Z.-H., Zhen, Y., Wang, Z.-J., Mu, J.-X., Zhu, M., Ouyang, Q., and Zhang, H. (2018). Chinese research into severe ulcerative colitis has increased in quantity and complexity. World Journal of Clinical Cases *6*, 35.

Luo, H., Shen, K., Li, B., Li, R., Wang, Z., and Xie, Z. (2020). Clinical significance and diagnostic value of serum NSE, CEA, CA19-9, CA125 and CA242 levels in colorectal cancer. Oncology letters *20*, 742-750.

M'koma, A.E. (2013). Inflammatory bowel disease: an expanding global health problem. Clinical Medicine Insights: Gastroenterology *6*, CGast. S12731.

Ma, Y., Ohmen, J.D., Li, Z., Bentley, G.L., McElree, C., Pressman, S., Targan, S.R., Fischel-Ghodsian, N., Rotter, J.I., and Yang, H. (1999). A genome-wide search identifies potential new susceptibility loci for Crohn's disease. Inflammatory bowel diseases *5*, 271-278.

Mahid, S.S., Minor, K.S., Soto, R.E., Hornung, C.A., and Galandiuk, S. (2006). Smoking and inflammatory bowel disease: a meta-analysis. In 11. (Elsevier), pp. 1462-1471.

Mancini, N.L., Rajeev, S., Jayme, T.S., Wang, A., Keita, Å.V., Workentine, M.L., Hamed, S., Söderholm, J.D., Lopes, F., and Shutt, T.E. (2021). Crohn's disease pathobiont adherent-invasive E coli disrupts epithelial mitochondrial networks with implications for gut permeability. Cellular and molecular gastroenterology and hepatology *11*, 551-571.

Mangifesta, M., Mancabelli, L., Milani, C., Gaiani, F., de'Angelis, N., de'Angelis, G.L., van Sinderen, D., Ventura, M., and Turroni, F. (2018). Mucosal microbiota of intestinal polyps reveals putative biomarkers of colorectal cancer. Scientific reports *8*, 1-9.

Mann, B., Gelos, M., Siedow, A., Hanski, M., Gratchev, A., Ilyas, M., Bodmer, W., Moyer, M., Riecken, E., and Buhr, H. (1999). Target genes of β-catenin–T cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas. Proceedings of the National Academy of Sciences *96*, 1603-1608.

Mann, S., Jess, T., Allin, K., and Elmahdi, R. (2022). Risk of Cervical Cancer in Inflammatory Bowel Disease: A Meta-Analysis of Population-Based Studies. Clinical and translational gastroenterology *13*, e00513.

Maqungo, M., Kaur, M., Kwofie, S.K., Radovanovic, A., Schaefer, U., Schmeier, S., Oppon, E., Christoffels, A., and Bajic, V.B. (2010). DDPC: dragon database of genes associated with prostate cancer. Nucleic acids research *39*, D980-D985.

Marafini, I., Sedda, S., Dinallo, V., and Monteleone, G. (2019). Inflammatory cytokines: from discoveries to therapies in IBD. Expert Opinion on Biological Therapy *19*, 1207-1217.

Mariampillai, A.I., Dela Cruz, J.P., Suh, J., Sivapiragasam, A., Nevins, K., and Hindenburg, A.A. (2017). Evaluation of tumor marker cancer antigen 72-4 (CA 72-4) in the monitoring of metastatic or recurrent tumors of the gastrointestinal tract, lung, breast, and ovaries. American Society of Clinical Oncology.

Markowitz, J., and Swets, J.A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. Perception & Psychophysics *2*, 91-100.

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., and Rodriguez Yoldi, M.J. (2017). Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. International journal of molecular sciences *18*.

Maronese, C.A., Pimentel, M.A., Li, M.M., Genovese, G., Ortega-Loayza, A.G., and Marzano, A.V. (2022). Pyoderma Gangrenosum: An Updated Literature Review on Established and Emerging Pharmacological Treatments. American Journal of Clinical Dermatology, 1-20.

Marotto, D., Atzeni, F., Ardizzone, S., Monteleone, G., Giorgi, V., and Sarzi-Puttini, P. (2020). Extra-intestinal manifestations of inflammatory bowel diseases. Pharmacological Research *161*, 105206.

Marrie, R.A., Graff, L.A., Fisk, J.D., Patten, S.B., and Bernstein, C.N. (2021). The relationship between symptoms of depression and anxiety and disease activity in IBD over time. Inflammatory bowel diseases *27*, 1285-1293.

Mathioudakis, N.N., Everett, E., Routh, S., Pronovost, P.J., Yeh, H.-C., Golden, S.H., and Saria, S. (2018). Development and validation of a prediction model for insulin-associated hypoglycemia in non-critically ill hospitalized adults. BMJ Open Diabetes Research and Care *6*, e000499.

Matsuoka, K., Kobayashi, T., Ueno, F., Matsui, T., Hirai, F., Inoue, N., Kato, J., Kobayashi, K., Kobayashi, K., and Koganei, K. (2018). Evidence-based clinical practice guidelines for inflammatory bowel disease. Journal of gastroenterology *53*, 305-353.

Mazlam, M., and Hodgson, H. (1994). Interrelations between interleukin-6, interleukin-1 beta, plasma C-reactive protein values, and in vitro C-reactive protein generation in patients with inflammatory bowel disease. Gut *35*, 77-83.

McGorry, P., Keshavan, M., Goldstone, S., Amminger, P., Allott, K., Berk, M., Lavoie, S., Pantelis, C., Yung, A., and Wood, S. (2014). Biomarkers and clinical staging in psychiatry. World Psychiatry *13*, 211-223.

Mendoza, J., and Abreu, M. (2009). Biological markers in inflammatory bowel disease: practical consideration for clinicians. Gastroenterologie clinique et biologique *33*, S158-S173.

Meng, C., Yin, X., Liu, J., Tang, K., Tang, H., and Liao, J. (2018). TIMP-1 is a novel serum biomarker for the diagnosis of colorectal cancer: A meta-analysis. PloS one *13*, e0207039.

Mentella, M.C., Scaldaferri, F., Pizzoferrato, M., Gasbarrini, A., and Miggiano, G.A.D. (2020). Nutrition, IBD and gut microbiota: a review. Nutrients *12*, 944.

Meyer, J.S., Robinson, G., Moonah, S., Levin, D., McGahren, E., Herring, K., Poulter, M., Waggoner-Fountain, L., and Shirley, D.-A. (2021). Acute appendicitis in four children with SARS-CoV-2 infection. Journal of Pediatric Surgery Case Reports *64*, 101734.

Milajerdi, A., Ebrahimi-Daryani, N., Dieleman, L.A., Larijani, B., and Esmaillzadeh, A. (2021). Association of dietary fiber, fruit, and vegetable consumption with risk of inflammatory bowel disease: a systematic review and meta-analysis. Advances in Nutrition *12*, 735-743.

Milanesi, E., Dobre, M., Cucos, C.A., Rojo, A.I., Jiménez-Villegas, J., Capetillo-Zarate, E., Matute, C., Piñol-Ripoll, G., Manda, G., and Cuadrado, A. (2021). Whole Blood Expression Pattern of Inflammation and Redox Genes in Mild Alzheimer's Disease. Journal of Inflammation Research *14*, 6085.

Miranda-García, P., Chaparro, M., and Gisbert, J.P. (2016). Correlation between serological biomarkers and endoscopic activity in patients with inflammatory bowel disease. Gastroenterologia y hepatologia *39*, 508-515.

Mirkov, M.U., Verstockt, B., and Cleynen, I. (2017). Genetics of inflammatory bowel disease: beyond NOD2. The lancet Gastroenterology & hepatology *2*, 224-234.

Miyaki, M., Iijima, T., Konishi, M., Sakai, K., Ishii, A., Yasuno, M., Hishima, T., Koike, M., Shitara, N., and Iwama, T. (1999). Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. Oncogene *18*, 3098-3103.

Møller Sørensen, N., Vejgaard Sørensen, I., Ørnbjerg Würtz, S., Schrohl, A.-S., Dowell, B., Davis, G., Jarle Christensen, I., Jørgen Nielsen, H., and Brünner, N. (2008). Biology and potential clinical implications of tissue inhibitor of metalloproteinases-1 in colorectal cancer treatment. Scandinavian Journal of Gastroenterology *43*, 774-786.

Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., and Gori, A.-S. (2018). IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nature communications *9*, 1-18.

Moon, J.M., Kang, E.A., Han, K., Hong, S.W., Soh, H., Park, S., Lee, J., Lee, H.J., Im, J.P., and Kim, J.S. (2020). Trends and risk factors of elderly-onset Crohn's disease: A nationwide cohort study. World journal of gastroenterology *26*, 404.

Moreno, L., Guo, D., Irwin, M.S., Berthold, F., Hogarty, M., Kamijo, T., Morgenstern, D., Pasqualini, C., Ash, S., and Potschger, U. (2021). A nomogram of clinical and biologic factors to predict survival in children newly diagnosed with high-risk neuroblastoma: An International Neuroblastoma Risk Group project. Pediatric Blood & Cancer *68*, e28794.

Mosli, M., Alawadhi, S., Hasan, F., Abou Rached, A., Sanai, F., and Danese, S. (2021). Incidence, prevalence, and clinical epidemiology of inflammatory bowel disease in the Arab World: a systematic review and meta-analysis. Inflammatory Intestinal Diseases, 1-9.

Mundade, R., Imperiale, T.F., Prabhu, L., Loehrer, P.J., and Lu, T. (2014). Genetic pathways, prevention, and treatment of sporadic colorectal cancer. Oncoscience *1*, 400.

Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A., and Board, W.H.O.C.o.T.E. (2020). The 2019 WHO classification of tumours of the digestive system. Histopathology *76*, 182-188.

Nambu, R., Hagiwara, S.-i., Kakuta, F., Hara, T., Shimizu, H., Abukawa, D., Iwama, I., Kagimoto, S., and Arai, K. (2019). Current role of colonoscopy in infants and young children: a multicenter study. BMC gastroenterology *19*, 1-7.

Nebert, D.W. (2002). Transcription factors and cancer: an overview. Toxicology *181*, 131-141.

Neurath, M.F. (2014). Cytokines in inflammatory bowel disease. Nature Reviews Immunology *14*, 329-342.

Newton, K.F., Newman, W., and Hill, J. (2012). Review of biomarkers in colorectal cancer. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland *14*, 3-17.

Ng, S.C., Shi, H.Y., Hamidi, N., Underwood, F.E., Tang, W., Benchimol, E.I., Panaccione, R., Ghosh, S., Wu, J.C., and Chan, F.K. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet *390*, 2769-2778.

Nguyen, H., Kerimoglu, C., Pirouz, M., Pham, L., Kiszka, K.A., Sokpor, G., Sakib, M.S., Rosenbusch, J., Teichmann, U., and Seong, R.H. (2018). Epigenetic regulation by BAF complexes limits neural stem cell proliferation by suppressing Wnt signaling in late embryonic development. Stem cell reports *10*, 1734-1750.

Nielsen, H.J., Brünner, N., Jorgensen, L.N., Olsen, J., Rahr, H.B., Thygesen, K., Hoyer, U., Laurberg, S., Stieber, P., and Blankenstein, M.A. (2011). Plasma TIMP-1 and CEA in detection of primary colorectal cancer: a prospective, population based study of 4509 high-risk individuals. Scandinavian Journal of Gastroenterology *46*, 60-69.

Nigam, Y., Knight, J., and Williams, N. (2019). Gastrointestinal tract 5: the anatomy and functions of the large intestine. Nursing Times *115*, 50-53.

Nishida, A., Inoue, R., Inatomi, O., Bamba, S., Naito, Y., and Andoh, A. (2018). Gut microbiota in the pathogenesis of inflammatory bowel disease. Clinical journal of gastroenterology *11*, 1-10.

Nkosi, B.V.Z., Padayachee, T., Gront, D., Nelson, D.R., and Syed, K. (2022). Contrasting Health Effects of Bacteroidetes and Firmicutes Lies in Their Genomes: Analysis of P450s,

Ferredoxins, and Secondary Metabolite Clusters. International journal of molecular sciences *23*, 5057.

Noble, C.L., Nimmo, E.R., Drummond, H., Ho, G.T., Tenesa, A., Smith, L., Anderson, N., Arnott, I.D., and Satsangi, J. (2005). The contribution of OCTN1/2 variants within the IBD5 locus to disease susceptibility and severity in Crohn's disease. Gastroenterology *129*, 1854-1864.

O'Donnell, P.H., and Dolan, M.E. (2009). Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy. Clinical cancer research : an official journal of the American Association for Cancer Research *15*, 4806-4814.

O'Brien, S., Rai, J.P., Roberts, H.L., Paas, M., Feagins, K., Pierce, E.J., Baumgartner, R.N., and Galandiuk, S. (2020). Pyoderma gangrenosum in inflammatory bowel disease: a systematic review and meta-analysis. Digestive diseases and sciences *65*, 2675-2685.

O'Connell, J., Keohane, S., McGreal-Bellone, A., McDonagh, P., Naimimohasses, S., Kennedy, U., Dunne, C., Hartery, K., Larkin, J., and MacCarthy, F. (2020). Characteristics and outcomes of acute colitis diagnosed on cross-sectional imaging presenting via the emergency department in an Irish academic medical centre. Irish Journal of Medical Science (1971-) *189*, 1115-1121.

Obrocea, F.L., Sajin, M., Marinescu, E.C., and Stoica, D. (2011). Colorectal cancer and the 7th revision of the TNM staging system: review of changes and suggestions for uniform pathologic reporting. Romanian journal of morphology and embryology = Revue roumaine de morphologie et embryologie *52*, 537-544.

Olén, O., Erichsen, R., Sachs, M.C., Pedersen, L., Halfvarson, J., Askling, J., Ekbom, A., Sørensen, H.T., and Ludvigsson, J.F. (2020). Colorectal cancer in ulcerative colitis: a Scandinavian population-based cohort study. The Lancet *395*, 123-131.

Ong, C., Aw, M.M., Liwanag, M.J., Quak, S.H., and Phua, K.B. (2018). Rapid rise in the incidence and clinical characteristics of pediatric inflammatory bowel disease in a South–East Asian cohort in Singapore, 1994–2015. Journal of digestive diseases *19*, 395-403.

Oskooei, A., Manica, M., Mathis, R., and Martínez, M.R. (2019). Network-based biased tree ensembles (NetBiTE) for drug sensitivity prediction and drug sensitivity biomarker identification in cancer. Scientific reports *9*, 1-13.

Overman, M., Adam, L., Raghav, K., Wang, J., Kee, B., Fogelman, D., Eng, C., Vilar, E., Shroff, R., and Dasari, A. (2018). Phase II study of nab-paclitaxel in refractory small bowel adenocarcinoma and CpG island methylator phenotype (CIMP)-high colorectal cancer. Annals of Oncology *29*, 139-144.

Øyri, S.F., Műzes, G., and Sipos, F. (2015). Dysbiotic gut microbiome: A key element of Crohn's disease. Comparative immunology, microbiology and infectious diseases *43*, 36-49.

Palmieri, O., Latiano, A., Valvano, R., D'INCÀ, R., Vecchi, M., Sturniolo, G., Saibeni, S., Peyvandi, F., Bossa, F., and Zagaria, C. (2006). Variants of OCTN1–2 cation transporter genes are associated with both Crohn's disease and ulcerative colitis. Alimentary pharmacology & therapeutics *23*, 497-506.

Papadakis, K.A., and Targan, S.R. (2000). Role of cytokines in the pathogenesis of inflammatory bowel disease. Annual review of medicine *51*, 289-298.

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A.M.-P., George, N., Fexova, S., Fonseca, N.A., Füllgrabe, A., Green, M., Huang, N., et al. (2019). Expression Atlas update: from tissues to single cells. Nucleic acids research *48*, D77-D83.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G.C., and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology *56*, 45.

Parizadeh, S.M., Jafarzadeh-Esfehani, R., Hassanian, S.M., Mottaghi-Moghaddam, A., Ghazaghi, A., Ghandehari, M., Alizade-Noghani, M., Khazaei, M., Ghayour-Mobarhan, M., and Ferns, G.A. (2019). Vitamin D in inflammatory bowel disease: From biology to clinical implications. Complementary Therapies in Medicine *47*, 102189.

Park, S.C., and Jeen, Y.T. (2019). Genetic studies of inflammatory bowel disease-focusing on Asian patients. Cells *8*, 404.

Pasqua, T., Pagliaro, P., Rocca, C., Angelone, T., and Penna, C. (2018). Role of NLRP-3 inflammasome in hypertension: a potential therapeutic target. Current Pharmaceutical Biotechnology *19*, 708-714.

Patankar, J.V., and Becker, C. (2020). Cell death in the gut epithelium and implications for chronic inflammation. Nature reviews Gastroenterology & hepatology *17*, 543-556.

Patil, K.R., Mahajan, U.B., Unger, B.S., Goyal, S.N., Belemkar, S., Surana, S.J., Ojha, S., and Patil, C.R. (2019). Animal models of inflammation for screening of anti-inflammatory drugs: Implications for the discovery and development of phytopharmaceuticals. International journal of molecular sciences *20*, 4367.

Patil, S., Jahagirdar, S., Khot, M., and Sengupta, K. (2022). Studying the role of chromosomal instability (CIN) in GI cancers using patient-derived organoids. Journal of Molecular Biology *434*, 167256.

Paulson, E.K., Kalady, M.F., and Pappas, T.N. (2003). Suspected appendicitis. New England Journal of Medicine *348*, 236-242.

Pellino, G., Gallo, G., Pallante, P., Capasso, R., De Stefano, A., Maretto, I., Malapelle, U., Qiu, S., Nikolaou, S., Barina, A., et al. (2018). Noninvasive Biomarkers of Colorectal Cancer: Role in Diagnosis and Personalised Treatment Perspectives. Gastroenterol Res Pract *2018*, 2397863.

Peng, X., Wang, J., Zhang, C., Liu, K., Zhao, L., Chen, X., Huang, G., and Lai, Y. (2020). A three-miRNA panel in serum as a noninvasive biomarker for colorectal cancer detection. The International Journal of Biological Markers *35*, 74-82.

Pepe, M.S., Janes, H., Li, C.I., Bossuyt, P.M., Feng, Z., and Hilden, J. (2016). Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? Clinical chemistry *62*, 737-742.

Persson, P., Ahlbom, A., and Hellers, G. (1990). Inflammatory bowel disease and tobacco smoke--a case-control study. Gut *31*, 1377-1381.

Pfeifer, G.P., and Rauch, T.A. (2009). DNA methylation patterns in lung carcinomas. In 3. (Elsevier), pp. 181-187.

Pham, V., and Mohajeri, M. (2018). The application of in vitro human intestinal models on the screening and development of pre-and probiotics. Beneficial Microbes *9*, 725-742.

Pierce, J.D., McCabe, S., White, N., and Clancy, R.L. (2012). Biomarkers: an important clinical assessment tool. AJN The American Journal of Nursing *112*, 52-58.

Pino, M.S., and Chung, D.C. (2010). The chromosomal instability pathway in colon cancer. Gastroenterology *138*, 2059-2072.

Politano, G., Benso, A., Savino, A., and Di Carlo, S. (2014). ReNE: a cytoscape plugin for regulatory network enhancement. PloS one *9*, e115585.

Porru, M., Pompili, L., Caruso, C., Biroccio, A., and Leonetti, C. (2018). Targeting KRAS in metastatic colorectal cancer: current strategies and emerging opportunities. Journal of Experimental & Clinical Cancer Research *37*, 1-10.

Porter, R.J., Arends, M.J., Churchhouse, A.M.D., and Din, S. (2021). Inflammatory Bowel Disease-Associated Colorectal Cancer: Translational Risks from Mechanisms to Medicines. Journal of Crohn's and Colitis *15*, 2131-2141.

Powell, S.M., Zilz, N., Beazer-Barclay, Y., Bryan, T.M., Hamilton, S.R., Thibodeau, S.N., Vogelstein, B., and Kinzler, K.W. (1992). APC mutations occur early during colorectal tumorigenesis. Nature *359*, 235-237.

Pyo, J.-S., Sohn, J.H., and Kang, G. (2016). Medullary carcinoma in the colorectum: a systematic review and meta-analysis. Human pathology *53*, 91-96.

Qian, K., Wu, C., Xiao, F., Zheng, Y., Zhang, Y., Yang, Z., and Liu, Y. (2018). Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. (IEEE), pp. 1574-1582.

Ramana, J. (2012). RCDB: renal cancer gene database. BMC research notes *5*, 1-4.

Ramos, G.P., and Papadakis, K.A. (2019). Mechanisms of disease: inflammatory bowel diseases. In 1. (Elsevier), pp. 155-165.

Ranjbar, R., Ghasemian, M., Maniati, M., Khatami, S.H., Jamali, N., and Taheri-Anganeh, M. (2022). Gastrointestinal disorder biomarkers. Clinica Chimica Acta.

Raskov, H., Søby, J.H., Troelsen, J., Bojesen, R.D., and Gögenur, I. (2020). Driver gene mutations and epigenetics in colorectal cancer. Annals of surgery *271*, 75-85.

Rasmussen, L., Nielsen, H.J., and Christensen, I.J. (2020). Early detection and recurrence of colorectal adenomas by combination of eight cancer-associated biomarkers in plasma. Clinical and Experimental Gastroenterology *13*, 273.

Rawla, P., Sunkara, T., and Barsouk, A. (2019). Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Przeglad gastroenterologiczny *14*, 89-103.

Ray, P., Manach, Y.L., Riou, B., Houle, T.T., and Warner, D.S. (2010). Statistical evaluation of a biomarker. The Journal of the American Society of Anesthesiologists *112*, 1023-1040.

Reese, G.E., Constantinides, V.A., Simillis, C., Darzi, A.W., Orchard, T.R., Fazio, V.W., and Tekkis, P.P. (2006). Diagnostic precision of anti-Saccharomyces cerevisiae antibodies and perinuclear antineutrophil cytoplasmic antibodies in inflammatory bowel disease. Official journal of the American College of Gastroenterology| ACG *101*, 2410-2422.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A.M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia (New York, N.Y.) *6*, 1-6.

Ricciuto, A., Carman, N., Benchimol, E.I., Siddiqui, I., Putra, J., Khanna, R., Griffiths, A.M., Walters, T.D., and Church, P.C. (2021). Prospective Evaluation of Endoscopic and Histologic Indices in Pediatric Ulcerative Colitis Using Centralized Review. The American Journal of Gastroenterology *116*, 2052-2059.

Rogler, G., and Biedermann, L. (2015). Clinical utility of biomarkers in IBD. Current gastroenterology reports *17*, 1-9.

Rolli-Derkinderen, M., Leclair-Visonneau, L., Bourreille, A., Coron, E., Neunlist, M., and Derkinderen, P. (2020). Is Parkinson's disease a chronic low-grade inflammatory bowel disease? Journal of neurology *267*, 2207-2213.

Røseth, A., Fagerhol, M., Aadland, E., and Schjønsby, H. (1992). Assessment of the neutrophil dominating protein calprotectin in feces: a methodologic study. Scandinavian Journal of Gastroenterology *27*, 793-798.

Rothfuss, K.S., Stange, E.F., and Herrlinger, K.R. (2006). Extraintestinal manifestations and complications in inflammatory bowel diseases. World journal of gastroenterology: WJG *12*, 4819.

Rothnie, J.B., Bernstein, P.A., Fox, S., Goodman, N., Hammer, M., Landers, T.A., Reeve, C., Shipman, D.W., and Wong, E. (1980). Introduction to a system for distributed databases (SDD-1). ACM Trans. Database Syst. *5*, 1–17.

Rubin, D.T., Traboulsi, C., and Rai, V. (2021). A practical clinical approach to the management of high-risk ulcerative colitis. Gastroenterology & Hepatology *17*, 59.

Rutter, M.D. (2011). Surveillance programmes for neoplasia in colitis. Journal of gastroenterology *46*, 1-5.

Saah, A.J., and Hoover, D.R. (1997). "Sensitivity" and "specificity" reconsidered: the meaning of these terms in analytical and diagnostic settings. American College of Physicians.

Sagar, S., Kaur, M., Dawe, A., Seshadri, S.V., Christoffels, A., Schaefer, U., Radovanovic, A., and Bajic, V.B. (2008). DDESC: Dragon database for exploration of sodium channels in human. BMC genomics *9*, 1-7.

Sakaki, Y. (2019). A Japanese history of the human genome project. Proceedings of the Japan Academy, Series B *95*, 441-458.

Salem, M., Lecka, J., Pelletier, J., Marconato, D.G., Dumas, A., Vallières, L., Brochu, G., Robaye, B., Jobin, C., and Sévigny, J. (2022). NTPDase8 protects mice from intestinal inflammation by limiting P2Y6 receptor activation: identification of a new pathway of inflammation for the potential treatment of IBD. Gut *71*, 43-54.

Salhi, A., Essack, M., Alam, T., Bajic, V.P., Ma, L., Radovanovic, A., Marchand, B., Schmeier, S., Zhang, Z., and Bajic, V.B. (2017). DES-ncRNA: A knowledgebase for exploring information about human micro and long noncoding RNAs based on literature-mining. RNA biology *14*, 963-971.

Salhi, A., Essack, M., Radovanovic, A., Marchand, B., Bougouffa, S., Antunes, A., Simoes, M.F., Lafi, F.F., Motwalli, O.A., and Bokhari, A. (2016). DESM: portal for microbial knowledge exploration systems. Nucleic acids research *44*, D624-D633.

Salibasic, M., Pusina, S., Bicakcic, E., Pasic, A., Gavric, I., Kulovic, E., Rovcanin, A., and Beslija, S. (2019). Colorectal cancer surgical treatment, our experience. Medical Archives *73*, 412.

Salih, A., Widbom, L., Hultdin, J., and Karling, P. (2018). Smoking is associated with risk for developing inflammatory bowel disease including late onset ulcerative colitis: a prospective study. Scandinavian journal of gastroenterology *53*, 173-178.

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., Yan, H., Gazdar, A., Powell, S.M., and Riggins, G.J. (2004). High frequency of mutations of the PIK3CA gene in human cancers. science *304*, 554-554.

Sanchez-Muñoz, F., Dominguez-Lopez, A., and Yamamoto-Furusho, J.K. (2008). Role of cytokines in inflammatory bowel disease. World journal of gastroenterology: WJG *14*, 4280.

Sandier, R.S., Sandler, D.P., McDonnell, C.W., and Wurzelmann, J.I. (1992). Childhood exposure to environmental tobacco smoke and the risk of ulcerative colitis. American journal of epidemiology *135*, 603-608.

Sanz-Garcia, E., Argiles, G., Elez, E., and Tabernero, J. (2017). BRAF mutant colorectal cancer: prognosis, treatment, and new perspectives. Annals of Oncology *28*, 2648-2657.

Sarandakou, A., Protonotariou, E., and Rizos, D. (2007). Tumor markers in biological fluids associated with pregnancy. Critical reviews in clinical laboratory sciences *44*, 151-178.

Sarrabayrouse, G., Landolfi, S., Pozuelo, M., Willamil, J., Varela, E., Clark, A., Campos, D., Herrera, C., Santiago, A., and Machiels, K. (2020). Mucosal microbial load in Crohn's disease: A potential predictor of response to faecal microbiota transplantation. EBioMedicine *51*, 102611.

Sarshekeh, A.M., Advani, S., Overman, M.J., Manyam, G., Kee, B.K., Fogelman, D.R., Dasari, A., Raghav, K., Vilar, E., and Manuel, S. (2017). Association of SMAD4 mutation with patient demographics, tumor characteristics, and clinical outcomes in colorectal cancer. PloS one *12*, e0173345.

Schaefer, A.K., Melnyk, J.E., He, Z., Del Rosario, F., and Grimes, C.L. (2018). Pathogen-and microbial-associated molecular patterns (PAMPs/MAMPs) and the innate immune response in Crohn's disease. In Immunity and Inflammation in Health and Disease, (Elsevier), pp. 175-187.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995). Quantitative Monitoring of Gene Expression Patterns With a Complementary DNA Microarray. Science (New York, N.Y.) *270*, 467-470.

Schramm, A., Köster, J., Marschall, T., Martin, M., Schwermer, M., Fielitz, K., Büchel, G., Barann, M., Esser, D., Rosenstiel, P., et al. (2013). Next-generation RNA sequencing reveals differential expression of MYCN target genes and suggests the mTOR pathway as a promising therapy target in MYCN-amplified neuroblastoma. Int J Cancer *132*, E106-115.

Segal, I., Ou Tim, L., Hamilton, D.G., and Walker, A.R. (1980). The rarity of ulcerative colitis in South African blacks. American Journal of Gastroenterology (Springer Nature) *74*.

Sesay, A.M., Tervo, P., and Tikkanen, E. (2018). Biomarkers in Health Care. Biosensors and Nanotechnology: Applications in Health Care Diagnostics, 17-33.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research *13*, 2498-2504.

Sharara, A.I., Al Awadhi, S., Alharbi, O., Al Dhahab, H., Mounir, M., Salese, L., Singh, E., Sunna, N., Tarcha, N., and Mosli, M. (2018). Epidemiology, disease burden, and treatment challenges of ulcerative colitis in Africa and the Middle East. Expert review of gastroenterology & hepatology *12*, 883-897.

Sher, M.E., Bank, S., Greenberg, R., Sardinha, C.T., Weissman, S., Bailey, B., Gilliland, R., and Wexner, S.D. (1999). The influence of cigarette smoking on cytokine levels in patients with inflammatory bowel disease. Inflammatory bowel diseases *5*, 73-78.

Sheriff, A. (2022). Special Issue "C-Reactive Protein and Cardiovascular Disease: Clinical Aspects". Multidisciplinary Digital Publishing Institute.

Shirmohamadi, M., Eghbali, E., Najjary, S., Mokhtarzadeh, A., Kojabad, A.B., Hajiasgharzadeh, K., Lotfinezhad, P., and Baradaran, B. (2020). Regulatory mechanisms of microRNAs in colorectal cancer and colorectal cancer stem cells. Journal of cellular physiology *235*, 776-789.

Shivji, S., Conner, J.R., Barresi, V., and Kirsch, R. (2020). Poorly differentiated clusters in colorectal cancer: A current review and implications for future practice. Histopathology *77*, 351-368.

Shreffler, J., and Huecker, M.R. (2020). Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios.

Shurtz, S., and Foster, M.J. (2011). Developing and using a rubric for evaluating evidence-based medicine point-of-care tools. Journal of the Medical Library Association : JMLA *99*, 247-254.

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: basic definitions. Ejifcc *19*, 203.

Singh, S., Dulai, P.S., Zarrinpar, A., Ramamoorthy, S., and Sandborn, W.J. (2017). Obesity in IBD: epidemiology, pathogenesis, disease course and treatment outcomes. Nature reviews Gastroenterology & hepatology *14*, 110-121.

Singhvi, N., Gupta, V., Gaur, M., Sharma, V., Puri, A., Singh, Y., Dubey, G.P., and Lal, R. (2020). Interplay of human gut microbiome in health and wellness. Indian journal of microbiology *60*, 26-36.

Smids, C., Horjus Talabur Horje, C.S., Groenen, M.J., van Koolwijk, E.H., Wahab, P.J., and van Lochem, E.G. (2017). The value of serum antibodies in differentiating inflammatory bowel disease, predicting disease activity and disease course in the newly diagnosed patient. Scandinavian Journal of Gastroenterology *52*, 1104-1112.

Smith, L.J., and Samadder, N.J. (2018). Hereditary Colorectal Cancer Syndromes: When to Refer for Genetic Evaluation. GASTROENTEROLOGY *1*.

Sobsey, C.A., Ibrahim, S., Richard, V.R., Gaspar, V., Mitsa, G., Lacasse, V., Zahedi, R.P., Batist, G., and Borchers, C.H. (2020). Targeted and untargeted proteomics approaches in biomarker development. Proteomics *20*, 1900029.

Sokol, H., Leducq, V., Aschard, H., Pham, H.-P., Jegou, S., Landman, C., Cohen, D., Liguori, G., Bourrier, A., and Nion-Larmurier, I. (2017). Fungal microbiota dysbiosis in IBD. Gut *66*, 1039-1048.

Song, L., Zhou, R., Huang, S., Zhou, F., Xu, S., Wang, W., Yi, F., Wang, X., and Xia, B. (2013). High intestinal and systemic levels of interleukin-23/T-helper 17 pathway in Chinese patients with inflammatory bowel disease. Mediators of inflammation *2013*.

Sørensen, N.M., Schrohl, A.S., Jensen, V., Christensen, I.J., Nielsen, H.J., and Brünner, N. (2008). Comparative studies of tissue inhibitor of metalloproteinases-1 in plasma, serum and tumour tissue extracts from patients with primary colorectal cancer. Scandinavian journal of gastroenterology *43*, 186-191.

Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A.K., Greaves, F., Karthikesalingam, A., King, D., Liu, X., and Markar, S.R. (2020). Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nature medicine *26*, 807-808.

Srinivas, P.R., Kramer, B.S., and Srivastava, S. (2001). Trends in biomarker research for cancer detection. The lancet oncology *2*, 698-704.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nature Reviews Genetics *20*, 631-656.

Steinbakk, M., Naess-Andresen, C., Fagerhol, M., Lingaas, E., Dale, I., and Brandtzaeg, P. (1990). Antimicrobial actions of calcium binding leucocyte L1 protein, calprotectin. The Lancet *336*, 763-765.

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., Twik, M., Belinky, F., Fishilevich, S., and Nudel, R. (2016a). VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC genomics *17*, 195-206.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., and Mazor, Y. (2016b). The GeneCards suite: from gene data mining to disease genome sequence analyses. Current protocols in bioinformatics *54*, 1.30. 31-31.30. 33.

Sun, J., Fei, F., Zhang, M., Li, Y., Zhang, X., Zhu, S., and Zhang, S. (2019). The role of mSEPT9 in screening, diagnosis, and recurrence monitoring of colorectal cancer. BMC cancer *19*, 1-10.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians *71*, 209-249.

Sutherland, A.D., Gearry, R.B., and Frizelle, F.A. (2008). Review of Fecal Biomarkers in Inflammatory Bowel Disease. Diseases of the Colon & Rectum *51*, 1283-1291..

Suzuki, K., Suzuki, I., Leodolter, A., Alonso, S., Horiuchi, S., Yamashita, K., and Perucho, M. (2006). Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage. Cancer cell *9*, 199-207.

Swanson, D.R. (1986). Undiscovered public knowledge. The Library Quarterly *56*, 103-118.

Sweeney, L., Moss-Morris, R., Czuber-Dochan, W., Meade, L., Chumbley, G., and Norton, C. (2018). Systematic review: psychosocial factors associated with pain in inflammatory bowel disease. Alimentary pharmacology & therapeutics *47*, 715-729.

Swift, A., Heale, R., and Twycross, A. (2020). What are sensitivity and specificity? Evidence-based nursing *23*, 2-4.

Szaflarska-Popławska, A., Tunowska, D., Sobieska-Poszwa, O., Gorecka, A., and Krogulska, A. (2019). The effectiveness, tolerability, and safety of different 1-day bowel preparation regimens for pediatric colonoscopy. Gastroenterology Research and Practice *2019*.

Takahashi, T. (2020). Roles of nAChR and Wnt signaling in intestinal stem cell function and inflammation. International Immunopharmacology *81*, 106260.

Taman, H., Fenton, C.G., Hensel, I.V., Anderssen, E., Florholmen, J., and Paulssen, R.H. (2017). Transcriptomic Landscape of Treatment—Naïve Ulcerative Colitis. Journal of Crohn's and Colitis *12*, 327-336.

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic acids research *47*, W556-W560.

Tatiya-Aphiradee, N., Chatuphonprasert, W., and Jarukamjorn, K. (2019). Immune response and inflammatory pathway of ulcerative colitis. Journal of basic and clinical physiology and pharmacology *30*, 1-10.

Thomas, D., Fourkala, E., Apostolidou, S., Gunu, R., Ryan, A., Jacobs, I., Menon, U., Alderton, W., Gentry-Maharaj, A., and Timms, J. (2015). Evaluation of serum CEA, CYFRA21-1 and CA125 for the early detection of colorectal cancer using longitudinal preclinical samples. British journal of cancer *113*, 268-274.

Thomson, D., Krupey, J., Freedman, S., and Gold, P. (1969). The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. Proceedings of the National Academy of Sciences *64*, 161-167.

Thurin, M. (2021). Tumor-associated glycans as targets for immunotherapy: the Wistar institute experience/legacy. Monoclonal Antibodies in Immunodiagnosis and Immunotherapy *40*, 89-100.

Tillett, W.S., and Francis Jr, T. (1930). Serological reactions in pneumonia with a non-protein somatic fraction of pneumococcus. The Journal of experimental medicine *52*, 561.

To, N., Ford, A.C., and Gracie, D.J. (2016). Systematic review with meta-analysis: the effect of tobacco smoking on the natural history of ulcerative colitis. Alimentary pharmacology & therapeutics *44*, 117-126.

TOMA, S.C., UNGUREANU, B.S., PATRASCU, S., SURLIN, V., and Georgescu, I. (2018). Colorectal cancer biomarkers-A new trend in early diagnosis. Current Health Sciences Journal *44*, 140.

Torres, J., Petralia, F., Sato, T., Wang, P., Telesco, S.E., Strauss, R., Li, X.-j., Laird, R.M., Gutierrez, R.L., and Porter, C.K. (2020). Serum biomarkers identify patients who will develop inflammatory bowel diseases up to 5 years before diagnosis. Gastroenterology *159*, 96-104.

Toyota, M., and Issa, J.-P.J. (1999). CpG island methylator phenotypes in aging and cancer. In 5. (Elsevier), pp. 349-357.

Treanor, D., and Quirke, P. (2007). Pathology of colorectal cancer. Clinical oncology *19*, 769-776.

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Frontiers in public health *5*, 307.

Tuomisto, A.E., Mäkinen, M.J., and Väyrynen, J.P. (2019). Systemic inflammation in colorectal cancer: Underlying factors, effects, and prognostic significance. World journal of gastroenterology *25*, 4383.

Tweedie, S., Braschi, B., Gray, K., Jones, T.E.M., Seal, R.L., Yates, B., and Bruford, E.A. (2021). Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res *49*, D939-D946.

Tysk, C., Lindberg, E., Järnerot, G., and Floderus-Myrhed, B. (1988). Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. Gut *29*, 990-996.

Ueno, H., Kajiwara, Y., Shimazaki, H., Shinto, E., Hashiguchi, Y., Nakanishi, K., Maekawa, K., Katsurada, Y., Nakamura, T., Mochizuki, H., et al. (2012). New criteria for histologic grading of colorectal cancer. The American journal of surgical pathology *36*, 193-201.

Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nature Reviews Methods Primers *1*, 1-21.

Ungaro, R.C., Chou, B., Mo, J., Ursos, L., Twardowski, R., Candela, N., and Colombel, J.-F. (2022). Impact of COVID-19 on Healthcare Resource Utilization among Patients with Inflammatory Bowel Disease in the USA. Journal of Crohn's and Colitis.

Uthoff, S.M., Eichenberger, M.R., Lewis, R.K., Fox, M.P., Hamilton, C.J., McAuliffe, T.L., Grimes, H.L., and Galandiuk, S. (2001). Identification of candidate genes in ulcerative colitis and Crohn's disease using cDNA array technology. International journal of oncology *19*, 803-810.

Van Assche, G., Dignass, A., Panes, J., Beaugerie, L., Karagiannis, J., Allez, M., Ochsenkühn, T., Orchard, T., Rogler, G., and Louis, E. (2010). The second European evidence-based

consensus on the diagnosis and management of Crohn's disease: definitions and diagnosis. Journal of Crohn's and Colitis *4*, 7-27.

Van De Wetering, M., Sancho, E., Verweij, C., De Lau, W., Oving, I., Hurlstone, A., Van Der Horn, K., Batlle, E., Coudreuse, D., and Haramis, A.-P. (2002). The β-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. cell *111*, 241-250.

Vangoitsenhoven, R., and Cresci, G.A. (2020). Role of microbiome and antibiotics in autoimmune diseases. Nutrition in Clinical Practice *35*, 406-416.

Veloso, F.T., Carvalho, J., and Magro, F. (1996). Immune-related systemic manifestations of inflammatory bowel disease: a prospective study of 792 patients. Journal of clinical gastroenterology *23*, 29-34.

Vermeire, S., Van Assche, G., and Rutgeerts, P. (2004). C-reactive protein as a marker for inflammatory bowel disease. Inflammatory bowel diseases *10*, 661-665.

Verstockt, B., Smith, K.G., and Lee, J.C. (2018). Genome-wide association studies in Crohn's disease: past, present and future. Clinical & translational immunology *7*, e1001.

Vogelstein, B., Fearon, E.R., Hamilton, S.R., Kern, S.E., Preisinger, A.C., Leppert, M., Smits, A.M., and Bos, J.L. (1988). Genetic alterations during colorectal-tumor development. New England Journal of Medicine *319*, 525-532.

Wang, M. (2021). Next-Generation Sequencing (NGS). In Clinical Molecular Diagnostics, (Springer), pp. 305-327.

Wang, Q., Shi, Y.-l., Zhou, K., Wang, L.-l., Yan, Z.-x., Liu, Y.-l., Xu, L.-l., Zhao, S.-w., Chu, H.-l., and Shi, T.-t. (2018). PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer. Cell death & disease *9*, 1-11.

Wang, X., Jiang, X., Zhou, L., Wang, Z., Huang, H., and Wang, M. (2019). LncRNA-NEF is involved the regulation of gastric carcinoma cell proliferation by targeting RUNX1. Molecular Medicine Reports *19*, 2051-2056.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics *10*, 57-63.

Watermeyer, G., Epstein, D., Adegoke, O., Kassianides, C., Ojo, O., and Setshedi, M. (2020). Epidemiology of inflammatory bowel disease in sub-Saharan Africa: a review of the current status. South African Medical Journal *110*, 1006-1009.

Wei, S.-C., Sollano, J., Hui, Y.T., Yu, W., Santos Estrella, P.V., Llamado, L.J.Q., and Koram, N. (2021). Epidemiology, burden of disease, and unmet needs in the treatment of ulcerative colitis in Asia. Expert Review of Gastroenterology & Hepatology *15*, 275-289.

Welgus, H.G., and Stricklin, G. (1983). Human skin fibroblast collagenase inhibitor. Comparative studies in human connective tissues, serum, and amniotic fluid. Journal of Biological Chemistry *258*, 12259-12264.

Westwood, J.H., Bessell, E.M., Bukhari, M.A., Thomas, P., and Walker, J.M. (1974). Studies on the structure of the carcinoembryonic antigen—I. Some deductions on the basis of chemical degradations. Immunochemistry *11*, 811-818.

White, J.H. (2018). Vitamin D deficiency and the pathogenesis of Crohn's disease. The Journal of Steroid Biochemistry and Molecular Biology *175*, 23-28.

Win, A.K., Buchanan, D.D., Rosty, C., MacInnis, R.J., Dowty, J.G., Dite, G.S., Giles, G.G., Southey, M.C., Young, J.P., Clendenning, M., et al. (2015). Role of tumour molecular and pathology features to estimate colorectal cancer risk for first-degree relatives. Gut *64*, 101-110.

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Briefings in bioinformatics *9*, 326-332.

Wittekind, C., Brierley, J.D., Lee, A., and van Eycken, E. (2019). TNM supplement: a commentary on uniform use (John Wiley & Sons).

Wu, M., Kim, Y.S., Ryu, H.-S., Choi, S.C., Kim, K.Y., Park, W.C., Kim, M.S., Myung, J.Y., Choi, H.S., and Kim, E.J. (2020a). MSI status is associated with distinct clinicopathological features in BRAF mutation colorectal cancer: A systematic review and meta-analysis. Pathology-Research and Practice *216*, 152791.

Wu, T., Mo, Y., and Wu, C. (2020b). Prognostic values of CEA, CA19-9, and CA72-4 in patients with stages I-III colorectal cancer. International Journal of Clinical and Experimental Pathology *13*, 1608.

Xi, Y., and Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. Translational oncology *14*, 101174.

Xia, J., Broadhurst, D.I., Wilson, M., and Wishart, D.S. (2013). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics *9*, 280-299.

Xie, W., Huang, X., Wei, C., Mo, X., Ru, H., Zhang, L., Ge, L., Tang, W., and Liu, J. (2022). Preoperative Neutrophil-BMI Ratio As a Promising New Marker for Predicting Tumor Outcomes in Colorectal Cancer. Technology in Cancer Research & Treatment *21*,

Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., and Jagodnik, K.M. (2021). Gene set knowledge discovery with enrichr. Current protocols *1*, e90.

Xu, D., Zhang, J., Xu, H., Zhang, Y., Chen, W., Gao, R., and Dehmer, M. (2020). Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. BMC genomics *21*, 1-17.

Yamamoto-Furusho, J.K. (2007). Genetic factors associated with the development of inflammatory bowel disease. World journal of gastroenterology: WJG *13*, 5594.

Yang, X., Kui, L., Tang, M., Li, D., Wei, K., Chen, W., Miao, J., and Dong, Y. (2020). High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. Frontiers in genetics *11*.

Yanqing, H., Cheng, D., and Ling, X. (2018). Serum CA72-4 as a Biomarker in the Diagnosis of Colorectal Cancer: A Meta-analysis. Open Medicine *13*, 164-171.

Yoo, M., Shin, J., Kim, J., Ryall, K.A., Lee, K., Lee, S., Jeon, M., Kang, J., and Tan, A.C. (2015). DSigDB: drug signatures database for gene set analysis. Bioinformatics *31*, 3069-3071.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology *16*, 284-287.

Zaidi, S.K., Frietze, S.E., Gordon, J.A., Heath, J.L., Messier, T., Hong, D., Boyd, J.R., Kang, M., Imbalzano, A.N., and Lian, J.B. (2017). Bivalent epigenetic control of oncofetal gene expression in cancer. Molecular and cellular biology *37*, e00352-00317.

Zenger, S., Gurbuz, B., Can, U., Bilgic, C., Sobutay, E., Postgil Yilmaz, S., Balik, E., Yalti, T., Kapran, Y., and Bugra, D. (2019). Clinicopathological importance of colorectal medullary carcinoma. European Surgery *51*, 308-314.

Zhang, X., Sun, X.-F., Cao, Y., Ye, B., Peng, Q., Liu, X., Shen, B., and Zhang, H. (2018). CBD: a biomarker database for colorectal cancer. Database *2018*.

Zhang, X., Zhang, W., and Cao, P. (2021). Advances in CpG island methylator phenotype colorectal cancer therapies. Frontiers in Oncology *11*, 629390.

Zhao, H., Ming, T., Tang, S., Ren, S., Yang, H., Liu, M., Tao, Q., and Xu, H. (2022). Wnt signaling in colorectal cancer: pathogenic role and therapeutic target. Molecular Cancer *21*, 1-34.

Zhao, M., Gönczi, L., Lakatos, P.L., and Burisch, J. (2021). The burden of inflammatory bowel disease in Europe in 2020. Journal of Crohn's and Colitis *15*, 1573-1587.

Zhao, Q., Zhan, T., Deng, Z., Li, Q., Liu, Y., Yang, S., Ji, D., and Li, Y. (2018). Glycan analysis of colorectal cancer samples reveals stage-dependent changes in CEA glycosylation patterns. Clinical proteomics *15*, 1-11.

Zhen, Y., Luo, C., and Zhang, H. (2018). Early detection of ulcerative colitis-associated colorectal cancer. Gastroenterology report *6*, 83-92.

Zhou, H., Xiang, N., Xie, J., and Diao, X. (2018). Ecotoxicology: the history and present direction. Encyclop Ecol *2*, 1-9.

Zhou, J., Li, X.-L., Chen, Z.-R., and Chng, W.-J. (2017a). Tumor-derived exosomes in colorectal cancer progression and their clinical applications. Oncotarget *8*, 100781.

Zhou, Q., Shen, Z.-F., Wu, B.-s., Xu, C.-b., He, Z.-q., Chen, T., Shang, H.-t., Xie, C.-f., Huang, S.-y., and Chen, Y.-g. (2019). Risk of colorectal cancer in ulcerative colitis patients: a systematic review and meta-analysis. Gastroenterology research and practice *2019*.

Zhou, Y., Zhang, X., Zhang, J., Fang, J., Ge, Z., and Li, X. (2017b). LRG1 promotes proliferation and inhibits apoptosis in colorectal cancer cells via RUNX1 activation. PloS one *12*, e0175122.

Zou, K.H., Liu, A., Bandos, A.I., Ohno-Machado, L., and Rockette, H.E. (2011). Statistical evaluation of diagnostic performance: topics in ROC analysis (CRC Press).

Zou, K.H., O'Malley, A.J., and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation *115*, 654-657.

Zou, X., Amrit, B., Abu-Izneid, T., Aziz, A., Devnath, P., Rauf, A., Mitra, S., Emran, T.B., Mujawah, A.A., and Lorenzo, J.M. (2021). Current advances of functional phytochemicals in Nicotiana plant and related potential value of tobacco processing waste: A review. Biomedicine & Pharmacotherapy *143*, 112191.

Zuo, T., Kamm, M.A., Colombel, J.-F., and Ng, S.C. (2018). Urbanization and the gut microbiota in health and inflammatory bowel disease. Nature Reviews Gastroenterology & Hepatology *15*, 440-452.

8 APPENDIX

8.1 APPENDIX A

Table 8. 1 STARD is a reporting checklist for diagnostic accuracy research (Bossuyt et al., 2003).

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS		Describe	
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.	
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	How indeterminate results, missing responses and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

Table 8. 2 Up-regulation of IBD genes expression with significant terms in GO annotation and KEGG pathways.

BIOLOGICAL PROCESSES			
GO ID	**TERMS**	**IBD GENES**	**p-value**
GO:0002232; GO:0006915; GO:0006954; GO:0001819; GO:0007165; GO:0001666 GO:0000187; GO:0000060; GO:0002385; GO:0001922; GO:0002237 GO:0000187; GO:0000165; GO:0002523; GO:0002863; GO:0000165; GO:0001774; GO:0001768	Positive regulation of cytokine production, T-cell activation, regulation of inflammatory response, Leukocyte proliferation, response to molecule of bacterial origin, cellular response to cytokine stimulus	FFAR2, FFAR3, IL17A, IL17F, IL21, NOD2, TLR5, TNF, TAGAP, TNFAIP3, XIAP, ADPOQ, BIRC2, BIRC3, EGFR, HGF, IL12B, IL15, IL18, IL12RA, IL2, IL21, IL23A, MYD88, NFKB1, NR1D2, PGLYRP1, PTGER4, PTPN2, SEPINE1, TLR10, TLR2, TLR3, TLR4, TLR7, TLR9, CCL2, CCL5, CX3CL1, CD28, JAK2, MEFV, NLRX1, NLRP1, NLRP12, NLRP3, NLRP6, S100A9, IL15, TAGAP, IL10, CXCL2, CXCL8, CD86, NLRP12, IL13, IL18RAP, MME, CCR7, CXCR4	5.6e-46– 7.5e-32
GO:0000165; GO:0002377; GO:0006468; GO:0002224	Positive regulation of cell-cell adhesion; Lymphocyte differentiation	CCL5, FAS, LCK, TNFRSF17, IL12B, IL2, MYD88, RAC2, CCL2, CCL21, CCR6, CX3CL1	7-5e-32 - 0.0
CELLULAR COMPONENT			
GO:0005737; GO:0005622; GO:0005576; GO:0005622; GO:0005576; GO:0001772; GO:0005769; GO:0005615; GO:0005622; GO:0005886; GO:0005634; GO:0005783; GO:0001772; GO:0005737; GO:0005615	External side of plasma membrane, membrane raft, membrane microdomain, membrane region, vesicle lumen, cytoplasmic vesicle lumen, Phagocytic vesicle, caveola	CCR5, CCR7, CD14, CD19, CD2, CD28, CD4, CD40, CD80, CD86, FAS, FASLG, THY1, ITGAX, ITGB1, ICAM1, IFNG, IL12RB2, IL13, IL17A, IL2RA, IL31RA, IL4, IL6ST, IL6, TLR4, TNF, ABCA1, CCR5, CCR7, FCGR3A, CD4, CFLAR, JAK2, LCK, BIRC2, BIRC3, CASP3, EGFR, ITLN1, TL1, TLR2, TLR6, TNF, ZAP70, ABCA1, ATG5, DMBT1, IRGM, RAC2, HMOX1, HDAC6, LRRK2, MAPK3, PTCH1, LRP6, SRC	1e-05 – 3e - 05
MOLECULAR FUNCTION			
GO ID	**TERMS**	**IBD GENES**	**p-value**
GO:0005125; GO:0004167; GO:0005088;	Cytokine receptor binding, receptor	IL15, IL17A, IL17F, IL21, MIF, IL18, IL2, IL21, IL22, IL23A, IL26, IL27, IL4, IL6, IL7, IL9, LTA, LTB, MIF,	5.0e – 07

GO:0005102; GO:0001968;	ligand binding, cytokine activity, signalling receptor activator activity, growth factor receptor binding,	OSM, TNFSF11, TNFSF15, TNF, VEGFA, FASLG, ADIPOQ, CSF1, CSF2, HMGB1, IFNA1, IFNG, IL10, IL12B, IL13, IL15, IL17A, IL17F, IL18, EGF, HGF, IGF1	
GO:0004896; GO:0005125;	cytokine receptor activity, growth factor activity, immune receptor activity, cytokine binding,	IL10, IL12B, IL2, IL4, IL6, IL7, IL9, LEP, OSM, REG1A, VEGFA, CSF1, CSF2, EGF, HGF, IGF1	1.0e - 06
KEGG PATHWAYS			
KEGG ID	**TERMS**	**IBD GENES**	**p-value**
hsa04060; hsa04010; hsa04014; hsa03050; hsa04010; hsa04062; hsa03050;	Cytokine-cytokine receptor interaction, JAK-STAT signalling pathway, NOD-like receptor signalling pathway, Th17 cell differentiation, IBD	CCL2, CCL21, CCL5, CCR5, CCR6, CCR7, CXCL2, CXCL8, CXCR4, CX3CL1, CD40, FAS, FASLG, TNFRSF17, TNFRSF6B, TNFRSF25, CSF1, CSF2, IFNA1, IFNG, IL1R2, IL10RB, IL10, IL12RB2, IL12B, IL13, IL15RA, IL15, IL17A, IL18RAP, IL18, IL2RA, IL2, IL21, IL22, IL23R, IL23A, IL26, IL4, IL6, IL6ST, IL7, IL9, LEP, LTA, LTB, OSM, TNFSF11, TNFSF15, IL6R, PIK3CA, STAT1, TYK2, MYC, AKT1, JAK2, IL10RB, IL18,	1e – 12 - 2e - 12
hsa04010; hsa04064; hsa00220; hsa00562; hsa04066; hsa04145; hsa04064; hsa04145; hsa04620;	Chagas disease, toxoplasmosis, PD-L1 expression and PD-1 checkpoint pathway in cancer, Toll- like receptor signalling pathway	MAPK3, MYD88, NOS2, NFKB1, PIK3CA, SERPINE1, TLR2, TLR4, TLR6, TLR9, TNF, AKT1, CCL2, CCL5, CXCL8, CFLAR, JUN, FAS, FASLG, IKBKG, IFNG, IL10, IL12B, IL2, IL6	3e - 12

Table 8. 3 Down-regulation of IBD genes expression with significant terms in GO annotation and KEGG pathways.

BIOLOGICAL PROCESSES			
GO ID	**TERMS**	**IBD GENES**	**p-value**
GO:0000165; GO:0001816; GO:0006468; GO:0001525; GO:0002230; GO:0001503; GO:0006915; GO:0006954; GO:0001541; GO:0000060; GO:0000122; GO:0001960; GO:0002237; GO:0001916**;** GO:0007166	Regulation of T-cell activation, response to lipopolysaccharide	CCL2, CCL5, CD4, LCK, THY1, PTPN22, TNFSF11, CCR5, CASP3, CCR7, CXCL2, CXCL8, CD14, CD40, CD80, CD86, FAS, FASLG, JAK2, JUN, NLRP3, SRC, TIRAP, TNFAIP3, TNFRSF25, XBP1, TNFRSF6B, CSF2, ERBIN, ICAM1, IFNG, IRF8, IRAK3	2.25e–31 – 3.0e-31
CELLULAR COMPONENT			
GO ID	**TERMS**	**IBD GENES**	**p-value**
GO:0005576; GO:0004930	Secretory granule lumen, plasma membrane signalling receptor complex	DEFA5, F9, DEFA6, DEFB4A, DUOX2, CA2	3e-05 – 5e-05
MOLECULAR FUNCTION			
GO ID	**TERMS**	**IBD GENES**	**p-value**
GO:0004672; GO:0002020; GO:0008009	Ubiquitin-like protein ligase binding	LEP, GLI1, KLF6, S100A9	1.5e-06
KEGG PATHWAYS			
KEGG ID	**TERMS**	**IBD GENES**	**p-value**
hsa04010; hsa04012; hsa04010; hsa00010; hsa00860; hsa04014; hsa03050; hsa04060	HIF-1 signaling pathway	AKT1, BCL2, CUL2, CDKN1A, EGFR, EGF, GAPDH, HMOX1, IGF1, IFNG, IL6R, IL6	

8.2 APPENDIX B

8.2.1 Scripting Language R

```
farhat <- function(path_file){


 if (!requireNamespace("BiocManager", quietly = TRUE))

  install.packages("BiocManager")


 BiocManager::install(c("clusterProfiler", "enrichplot", "org.Hs.eg.db"))


 library(clusterProfiler)

 library(enrichplot)

 library(org.Hs.eg.db)


 eg_SYM <- read.csv(file=path_file, sep=';')


 kk <- enrichKEGG(gene         = eg_SYM$EntrezID,

                  organism     = 'hsa',

                  pvalueCutoff = 0.05,

                  qvalueCutoff  = 1)


 egoBP <- enrichGO(gene = eg_SYM$EntrezID,

                   OrgDb         = org.Hs.eg.db,

                   ont           = 'BP',

                   pAdjustMethod = "BH",

                   pvalueCutoff  = 0.05,
```

```
            qvalueCutoff  = 1)


 egoCC <- enrichGO(gene = eg_SYM$EntrezID,
            OrgDb         = org.Hs.eg.db,
            ont           = 'CC',
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.05,
            qvalueCutoff  = 1)



 egoMF <- enrichGO(gene = eg_SYM$EntrezID,
            OrgDb         = org.Hs.eg.db,
            ont           = 'MF',
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.05,
            qvalueCutoff  = 1)



 print(dotpot(kk)
 print(dotpot(egoBP))
 print(dotpot(egoCC))
 print(dotpot(egoMF))


}
```

farhat('~/Desktop/GO.csv')

8.2.2 Scripting Language Python


```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 8,
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "    HGNC_Gene_Symbol        Inflammatory Bowel Disease        Pre-
viousname  \\\n",
      "0                NKX2-3                                      IBD1
NKX2C   \n",
      "1                 PSMD8                                      IBD1
NaN   \n",
      "2                 NLRP3                                      IBD1
C1orf7;CIAS1;DFNA34   \n",
      "3                  NOD2                                      IBD1
CARD15;IBD1   \n",
      "4                  NOD1                                      IBD1
CARD4   \n",
      "5               ATG16L1                                      IBD1
APG16L;ATG16L   \n",
      "6                   TNF                                      IBD1
TNFA   \n",
      "7                 RIPK2                                      IBD1
NaN   \n",
      "8                  TLR2                                  IBD1(CD)
NaN   \n",
      "9                  TLR4                                  IBD1(CD)
NaN   \n",
      "10               IL23R                                      IBD1
NaN   \n",
      "11                IRGM                                      IBD1
IRGM1   \n",
      "12                TLR3                                      IBD1
NaN   \n",
      "13               GAPDH                                      IBD1
GAPD   \n",
      "14               PTPN2                                      IBD1
PTPT   \n",
      "15                MUC2                                      IBD1
NaN   \n",
      "16                   PIGR        inflammatory bowel disease 13
NaN   \n",
      "17                  BAHD1        inflammatory bowel disease 14
NaN   \n",
      "18                  FFAR2        inflammatory bowel disease 17
GPR43   \n",
```

```
      "19                  FFAR3        inflammatory bowel disease 17
GPR41   \n",
      "20                    NCF2   inflammatory bowel disease 15(UC)
NaN   \n",
      "21                  RAC2                                      IBD
NaN   \n",
      "28               RALBP1                                     IBD1
NaN   \n",
      "29               TCF7L2                                     IBD1
TCF4   \n",
      "..                  ...                                      ...
...   \n",
      "251               PRDM1                        Crohn's Disease
PRDM1   \n",
      "252               IL1R2                   Ulcerative colitis
IL1RB   \n",
      "253             CDKAL1                              CD and UC
NaN   \n",
      "254               IRF8                                     NaN
ICSBP1   \n",
      "255             DNMT3B                                     NaN
NaN   \n",
      "256                MLN                                      UC
NaN   \n",
      "257               LCN2                                      UC
NaN   \n",
      "258              REG3A                                      CD
PAP   \n",
      "259              CXCL2                                      UC
GRO2   \n",
      "260              REG1A                                      UC
REG   \n",
      "261               RGS3                                      UC
NaN   \n",
      "262              MMP12                                      CD
NaN   \n",
      "263              CLDN8                                  CD; UC
NaN   \n",
      "264            SLC26A2                                   CD;UC
DTD   \n",
      "265               PCK1                                      UC
NaN   \n",
      "266             GUCA2A                                      UC
GUCA2   \n",
      "267               AQP8                                      UC
NaN   \n",
      "268             SLC4A4                                      UC
SLC4A5   \n",
      "269              ABCA8                                      UC
NaN   \n",
      "270              SATB2                                      UC
NaN   \n",
      "271          CBR3-AS1                                     IBD
NaN   \n",
      "272          ADAMDEC1                                      UC
NaN   \n",
```

```
      "273                ANK3                                          UC
NaN   \n",
      "274               NR1D2                                          UC
NaN   \n",
      "277            RUNX1;SPIB;SPI1;FEV;AP1;ELF5;IRF1;RUNX1;SPIB;S...
PRKACB\n",
      "278            RUNX1;SPIB;SPI1;FEV;AP1;ELF5;REL;RUNX1;SPIB;SP...
PTPRR\n",
      "279                                                           NaN
SLC16A1\n",
      "280                                                           NaN
SELENBP1\n",
      "\n",
      "[281 rows x 2 columns]\n"
     ]
    }
   ],
   "source": [
    "import pandas as pd\n",
    "\n",
    "data  =  pd.read_excel(r\"C:\\\\Users\\\\FARHAT  KHAN\\\\Desktop\\\\IBD_DA-
TASETS2.xlsx\")\n",
    "df  =  pd.DataFrame(data,  columns=[\"OPOSSUM_TFs\",\"HGNC_Gene_Sym-
bol\"])\n",
    "print (df)"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": []
  }
 ],
 "metadata": {
  "kernelspec": {
   "display_name": "Python 3",
   "language": "python",
   "name": "python3"
  },
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 3
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython3",
   "version": "3.7.3"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 2
}
```

8.3 APPENDIX C

8.3.1 IBDDB: User manual guide of IBD database

A GUIDE TO USING THE INFLAMMATORY BOWEL DISEASE DATABASE (IBDDB)

OCTOBER 2020

Table of contents

A. Introduction 162
B. Accessing IBDDB 163
C. Exploration of list of reports generated 163
1. IBD database 163
2. Knowledgebase 164
3. Biomedical entities 165
4. Biomedical entities co-occurrence 166
5. Hypotheses explorer 166
D. Network creation and visualization 167
E. Downloading Data 168

A. Introduction

The Inflammatory Bowel Disease Database (IBDDB) is the first manually curated database of experimentally verified IBD genes https://www.cbrc.kaust.edu.sa/ibd/guide.pdf. This database provides pre-compiled biomedical text mining of PubMed (manually cleaned and curated) and PubChem abstracts. The various information embedded in the database is viewed in IBDDB via medical entities identified as a subject-specific set of documents. These entities are terms, assembled into classified dictionaries.

IBDDB exhibits 11 specialized dictionaries:

- Biological Action
- Biological Processes
- Cellular Component
- ChEBI
- Dietary Supplement
- Diseases
- Drugs
- Human Genes
- Metabolites-and-Enzymes
- Pathways
- Toxins

The user can also explore different biomedical entities (about 20,000) and their co-occurrence with other entities (about one million) from 11 dictionaries, which are color-coded, in the indexed PubMed records.

IBDDB database uses various specialized dictionaries that are curated manually. It is important for users to contemplate that these entities are terms, assembled into classified dictionaries. For example, the 'biological process' is a dictionary, and its medical entity is 'amino acid biosynthesis', 'amino acid transport', etcetera. Meanwhile during the use of 'Disease' dictionary, one should know that this dictionary use has also emphasized on broad spectrum. This dictionary not only contains disease names, but also contain symptoms and various other related terms.

B. Accessing IBDDB

IBDDB allows user to explore, dissect and retrieve information using “Explore” tab with ease through an interactive user-friendly interface. By clicking on this tab, it will take to drop down menu displaying several lists of reports discussed below.

The Explore tabs containing list of reports:

- IBD database
- Knowledgebase
- Biomedical entities
- Biomedical entities co-occurrence
- Hypotheses explorer

C. Exploration of list of reports generated

1 IBD database

IBD database is a unique resource as it is a combination of curated and text-mined information which is easily explorable and user-friendly. This is the first manually curated database providing extended information about experimentally validated IBD related genes. It integrates data on 34 subject-specific concepts (listed as columns) as shown in Fig 1.

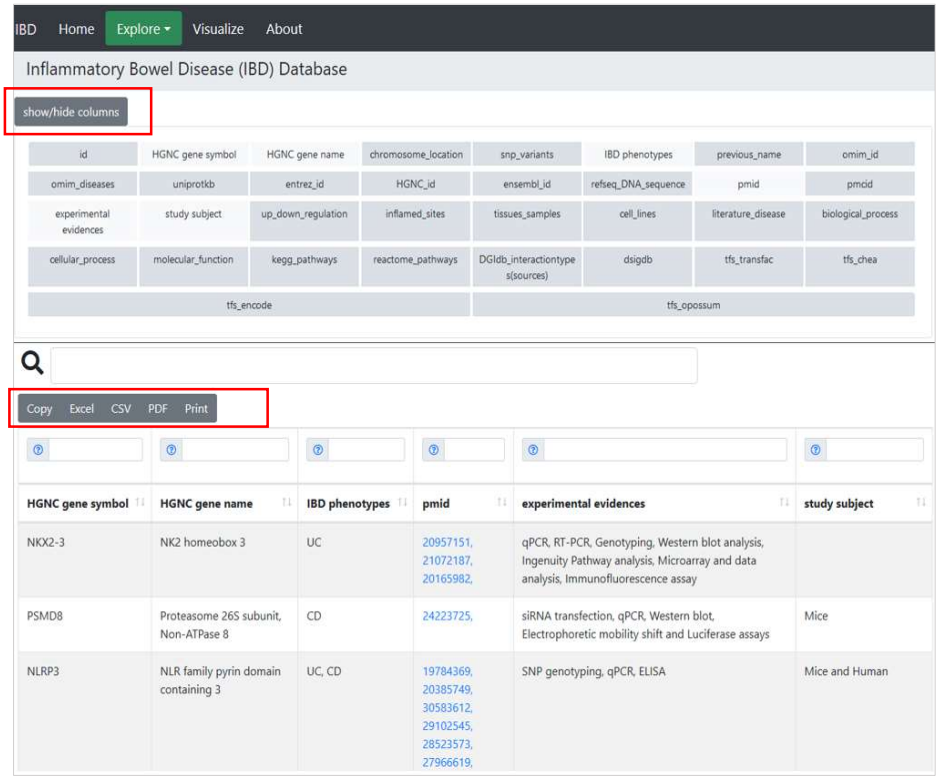


Figure 1 Explore the IBD database.

The information of IBD database page under the Explore tab integrates the data on 34 concepts such as experimental techniques used to validate the role of a gene in IBD, sites inflamed in IBD, other disease linked to the genes, tissue samples used for validation, molecular interactions, gene expression level in mice or humans, and transcription factors of IBD implicated genes using four different tools, and contains data on drugs associated with IBD genes. The result is downloadable under various formats highlighted with red box.

2 Knowledgebase

The in-house text mining Dragon Exploration System (DES), which has been used previously for the creation of several published knowledgebases was used to obtain a list of potential IBD-related genes. The PubMed was searched for relevant literature using queries 'Inflammatory AND bowel AND IBD', 'Crohn's AND CD', and 'Ulcerative AND colitis AND UC'. The titles and abstracts were downloaded and indexed using several dictionaries as shown in Fig 2. We have text-mined the abstract using DES and PolySearch 2.0 tool for more than 700,000 articles and partly done manually along with their PubMed identifiers.

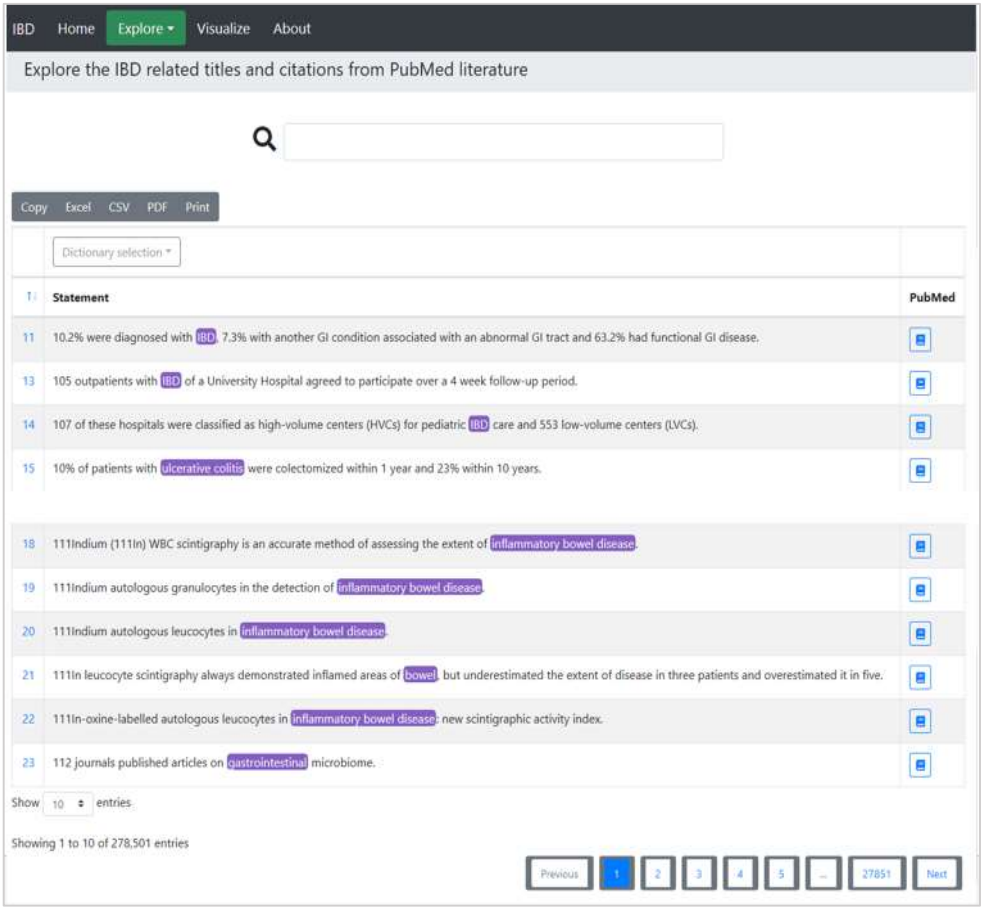


Figure 2 This figure illustrates an example of indexed document for the highlighted term (on the left-side) mentioned in the PubMed literature (on the right-side).

3 Biomedical entities

It is important for users to contemplate that these entities are terms, assembled into classified dictionaries. IBDDB database uses various specialized dictionaries that are curated manually. For example, the 'biological process' is a dictionary, and its medical entity is 'amino acid biosynthesis', 'amino acid transport', etcetera. One medical entity can appear in the various dictionary; therefore, it can be normalized by using different names from the dictionary's list (e.g. cellular component, drug agent, and etcetera). The user can also explore different biomedical entities (about 20,000) from 11 dictionaries in the indexed PubMed records as shown in Fig 3.

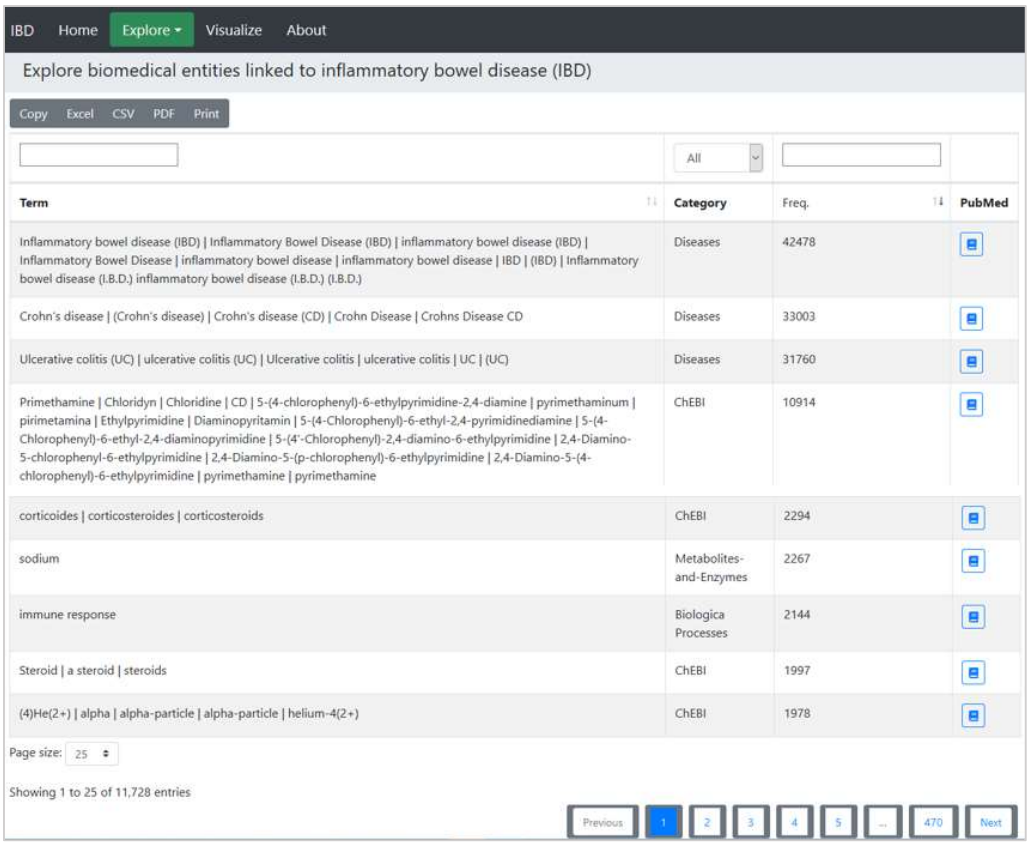


Figure 3 This figure illustrates the biomedical entities under category along with frequency of that occurring in PubMed literatures.

4 Biomedical entities co-occurrence

IBDDB helps user to Explore co-occurrence of biomedical entities linked to IBD. It allows to find co-mention of two entities in the published literature by selecting any two dictionaries (out of 11 listed). The relevant PubMed articles can also be retrieved for further exploration as shown in Fig 4.

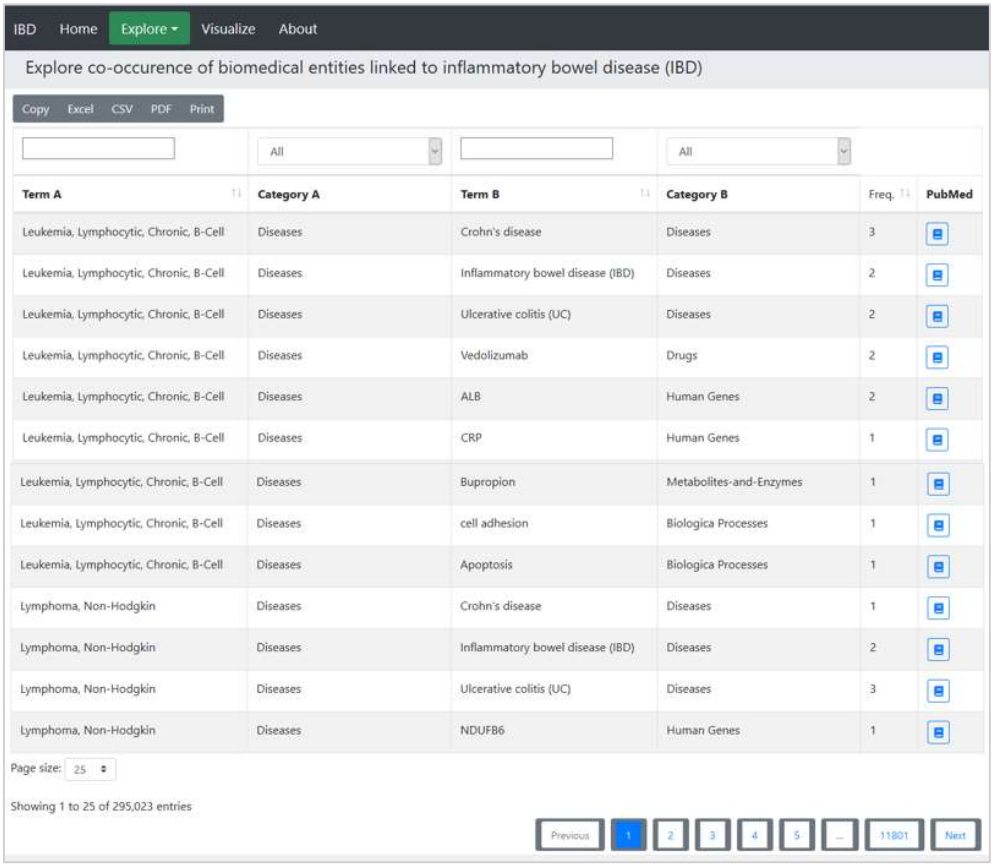


Figure 4 Explore Co-occurrence of biomedical entities linked to IBD. Here terms from Term A and B are co-occurring using biomedical entities from category A and B as denoted by the frequency column. PubMed abstract can be retrieved for these co-occurring terms.

5 Hypotheses explorer

We have text-mined the abstract using OpenFusion tool for more than 700,000 articles and partly done manually along with their PubMed identifiers. The text mining data can create a hypotheses generator, which illustrates the interaction between different identities from various dictionaries and their co-occurrence frequencies. Hypotheses explorer (under Explore tab allows to identify new association among selected terms selected from different dictionaries (A-B and B-C are known and A-C are suggested links between biomedical entities as shown in Fig 6. Users can accept or reject the abstracts based on their active association relationship manually.

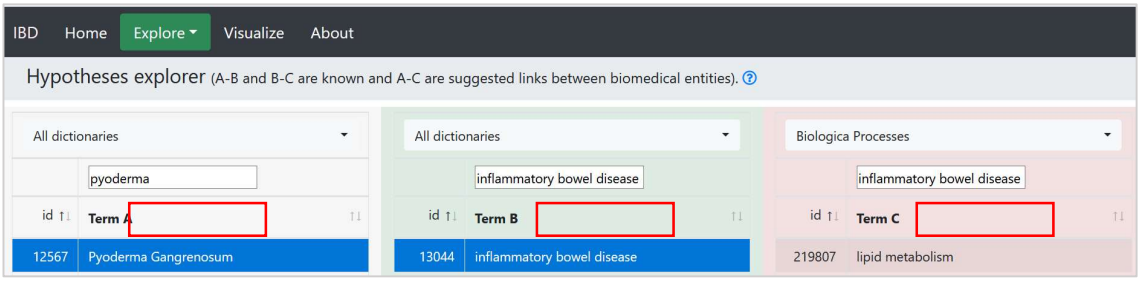


Figure 5 The hypothesis generated by the hypothesis explorer showing that Pyoderma Gangrenosum (PG) may be linked to lipid metabolism.

In the above example (fig 6), we generated a hypothesis. By selecting Pyoderma gangrenosum as 'Term A' and Inflammatory bowel disease' as 'Term B', we search for 'Term C' by selecting biological processes dictionary, 162 terms appeared, and lipid metabolism was a part of these terms. This may mean that lipid metabolism may play a part in pyoderma gangrenosum (Persson et al.).

D. Network creation and visualization

Often working with huge numbers of biomedical entities from various dictionaries can be a complex task. On the contrary, IBDDB provides "Visualize" tab for users to create interactive networks among different entities based on their occurrence in the literature. The user can build the network of their choice by selecting one or more dictionaries at each step and trimming off the links they dislike. The nodes represent the entities from the dictionary and are color-coded while numbering on the edges or links represents the number of PubMed records showing co-occurrence of the entities as highlighted terms. An example of network creation using hypotheses explorer is shown in the Fig 7.

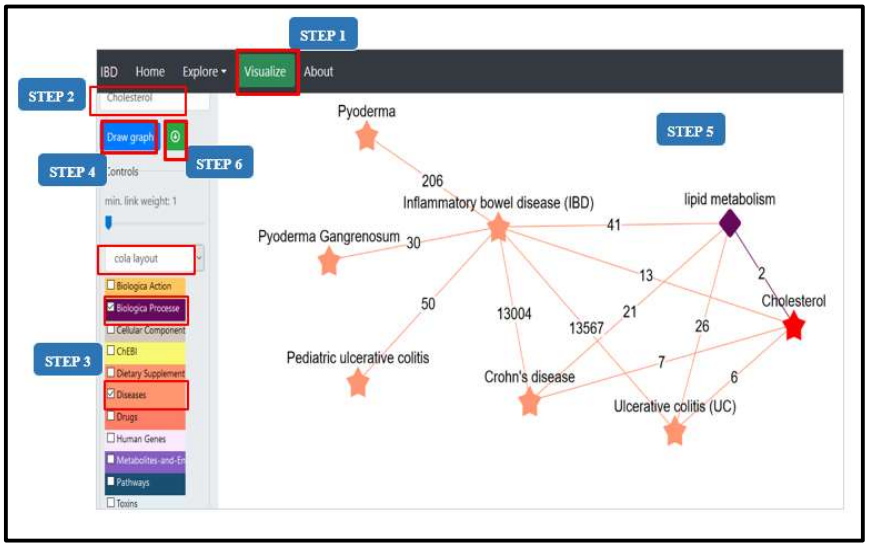


Figure 6 This figure demonstrates the step-by-step network construction to identify the relationship between disease PG and its connection with the pathway lipid metabolism. The red star represents the searched term "cholesterol". The orange stars represent the "Diseases" dictionary. The purple colored

diamond represents the expanded “Biological Processes" dictionary. The colored edges represent the color of their respective dictionary here. The number allocated on each edge showcase the number of publications that link to the associated nodes. The graph can be downloaded by selecting the green box with circled downward arrow to the right of draw graph tab.

E. Downloading Data

This database can make presentations of networks in 6 different layouts and fully exportable data in various formats like. JSON and .PNG as well as printing and direct copying of data to user applications.

The exploration reports like IBD database, biomedical entities, Biomedical entities co-occurrences and hypotheses explorer can be downloaded as Copy, Excel, CSV file, PDF, and Print formats.