IDENTIFICATION AND VALIDATION OF HUB GENES INVOLVED IN ATP-INDUCED CELL DEATH IN THE PATHOGENESIS OF INFLAMMATORY BOWEL DISEASE

ZHANG HAOLONG

UNIVERSITI SAINS MALAYSIA

2025

IDENTIFICATION AND VALIDATION OF HUB GENES INVOLVED IN ATP-INDUCED CELL DEATH IN THE PATHOGENESIS OF INFLAMMATORY BOWEL DISEASE

by

ZHANG HAOLONG

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

February 2025

ACKNOWLEDGEMENT

I would like to take this opportunity to express my heartfelt gratitude to all those who have supported me during this journey. First and foremost, I would like to thank my main supervisor and co-supervisor, Dr. Ahmad Naqib Shuid and Dr. Siti Nurfatimah Binti Mohd Shahpudin for his professional guidance throughout the research process. He provided valuable advice in the academic aspects of critically thinking and problems solving. I am grateful to Assoc. Prof. Dr. Doblin Sandai and Dr. Rosline Sandai for their guidance in my slide's presentation. I would like to express my gratitude to my mentors, Dr. Xingbei Chen, for her invaluable support and guidance throughout this research project. Additionally, I extend my heartfelt thanks to Professor Zhijing Song and Professor Zhongwen Zhang for providing the experimental platform. Furthermore, I want to thank my family and friends. Throughout my entire graduate studies, you have shown me unwavering care and support, encouraging me to overcome challenges and persevere. Once again, I would like to express my sincere gratitude to everyone who has contributed to my research. It is your support and assistance that have helped me overcome this challenging phase of my academic life. I will always remember your help and encouragement and continue to strive for excellence in my future academic and professional endeavours, aiming to give back to society.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ATP	Adenosine Triphosphate
AICD	ATP-induced cell death
AIEC	adherent-invasive Escherichia coli
ANO6	Anoctamin 6
ASK1	Apoptosis signal-regulating kinase 1
AUC	Area Under the Curve
B cells	B lymphocyte
Bax	Bcl-2-associated X protein
Caco-2	Human colon carcinoma cell line
CASP1	Caspase-1
CASP3	Caspase-3
CASP7	Caspase-7
CASP8	Caspase-8
CASP9	Caspase-9
CC	Pancolitis
CCK-8	Cell Counting Kit-8
CCL5	Chemokine ligand 5
CD	Crohn's disease
cDNA	Complementary DNA
CO2	Carbon dioxide
COX-2	cyclooxygenase-2
CRP	C-reactive protein
CXCL2	Chemokine ligand 2
CYTC	Cytochrome C
DCFHDA	2',7'-Dichlorofluorescin diacetate
DEGs	Differentially Expressed Genes
eATP	Extracellular Adenosine Triphosphate

ERK1/2	Extracellular signal-regulated kinase 1/2
FDR	false discovery rate
G908R	Mutation in NOD2 gene
GEO	Gene Expression Omnibus
GO	Gene Ontology
GS	Gene Connectivity
GSDMD	Gasdermin D
GSDME	Gasdermin E
GSEA	Gene Set Enrichment Analysis
HCLS1	Hematopoietic Cell-Specific Lyn Substrate 1
IBD	Inflammatory Bowel Disease
IFN-γ	Interferon-γ
IL	Interleukin
IL-10	Interleukin 10
IL-12	Interleukin 12
IL-17	Interleukin 17
IL-17F	Interleukin 17F
IL-1β	Interleukin-1 beta
IL-21	Interleukin 21
IL-22	Interleukin 22
IL-23	Interleukin 23
IL23R	Interleukin 23 Receptor
IL-6	Interleukin-6
iNOS	inducible nitric oxide synthase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	left-sided colitis
LPS	Lipopolysaccharide
Lyn	Lck/Yes novel tyrosine kinase
MLC	Myosin light chain
MM	Module Membership

mRNA	Messenger RNA
MsigDB	Molecular Signatures Database
Mϕ	Macrophage
NCBI	National Center for Biotechnology Information
NCCD	Nutrients, Cell Growth, and Death
NLRP1	NOD-like receptor protein 1
NLRP3	NOD-like receptor protein 3
NLRP3	NOD-like receptor protein 3
NOD2	Mutation in NOD2 gene
NOD2	Nucleotide-binding Oligomerization Domain 2
NOD2	Nucleotide-binding oligomerization domain-containing protein 2
NOX2	NADPH oxidase 2
NSAIDs	Non-steroidal anti-inflammatory drugs
ORAI1	Calcium Release-Activated Calcium Channel Protein 1
P2Rs	P2 receptors
P2RX3	Purinergic Receptor X3
P2RX4	Purinergic Receptor X4
P2RX5	Purinergic Receptor X5
P2RY1	Purinergic Receptor Y1
P2RY11	Purinergic Receptor Y11
P2RY12	Purinergic Receptor Y12
P2RY13	Purinergic Receptor Y13
P2RY14	Purinergic Receptor Y14
P2RY2	Purinergic Receptor Y2
P2RY4	Purinergic Receptor Y4
P2RY5	Purinergic Receptor Y5
P2RY6	Purinergic Receptor Y6
P2X6	Purinergic Receptor X6
P2X7	Purinergic Receptor P2X7
p38 MAPK	p38 mitogen-activated protein kinase

PBS	Phosphate-buffered saline
PCA	Principal Component Analysis
PI	propidium iodide
PLT	Platelet count
PNAX1	Pneumococcus neuraminidase-activating gene 1
R702W	Mutation in NOD2 gene
ROC	Receiver Operating Characteristic
ROCK I	Rho-associated protein kinase I
RT-qPCR	Reverse Transcription Quantitative Polymerase Chain Reaction
SAPK	Stress-activated protein kinase
STIM1	Stromal interaction molecule 1
T cells	T lymphocyte
TGF-β	Transforming Growth Factor beta
Th1 cells	Type 1 Helper T cell
Th17 cells	Type 17 Helper T cell
TLR4	toll-like receptor 4
TNF-α	Tumor Necrosis Factor alpha
TNF-α	Tumor necrosis factor alpha
TOM	Topological Overlap Matrix
Tregs	Regulatory T cells
UC	Ulcerative Colitis
V-EGFP	Venus-enhanced green fluorescent protein
WGCNA	Weighted Gene Co-expression Network Analysis

PENGENALPASTIAN DAN PENGESAHAN GEN HUB YANG TERLIBAT DALAM KEMATIAN SEL YANG DIDORONG OLEH ATP DALAM PATOGENESIS PENYAKIT KERADANGAN USUS

ABSTRAK

Ekstraselular adenosin trifosfat (eATP) memainkan peranan yang penting bagi penyakit radang usus (IBD) sebagai pengatur utama proses kematian sel. Reseptor P2X7 yang terdapat pada sel usus epitelium didapati seringkali terlibat dengan kematian sel yang melibatkan eATP. Walaubagaimanapun, mekanisme spesifik kematian sel yang disebabkan oleh eATP serta kaitannya dengan IBD masih belum jelas. Oleh yang demikian, tujuan utama kajian ini adalah untuk mengintegrasikan pengumpulan data, analisis bioinformatik, dan teknik eksperimen untuk mengenal pasti dan mengesahkan gen hab yang terlibat dengan kematian sel yang diinduksi oleh eATP dalam penyakit IBD. Pendekatan ini menggabungkan kaedah komputasi dan eksperimen untuk menghasilkan dapatan yang boleh dipercayai dan berkeyakinan tinggi. Pendekatan ini bukan sahaja meningkatkan pemahaman tentang laluan molekul yang terlibat dalam kematian sel yang disebabkan oleh ATP tetapi juga boleh membawa kepada sasaran terapeutik yang berpotensi untuk pengurusan IBD. Eksperimen ini dimulakan dengan pengumpulan data, memberi tumpuan kepada mendapatkan dataset dan set gen yang berkaitan daripada pangkalan data awam dan literatur saintifik. Untuk data transkriptomik, kata kunci seperti "IBD" dan "transcriptome" digunakan untuk mencari pangkalan data GEO (Provide abbreviation). Carian ini berjaya mengekstrak 37 set gen yang relevan dengan kajian IBD. Pada masa yang sama, kata kunci seperti "ATP," "kematian sel," "apoptosis," "autofagi," dan "ATP ekstraselular" digunakan dalam pangkalan data literatur termasuk PubMed, Embase, Web of Science, dan Scopus dan

berjaya mengenal pasti 19 set gen yang berkaitan dengan kematian sel yang disebabkan oleh ATP. Seterusnya data-data yang dikumpulkan di proses menggunakan keadah bioinformatik untuk mengenal pasti sasaran gen yang utama. Analisis ekspresi gen berbeza menggunakan alat visualisasi seperti plot volcano (volcano plot) dan plot kotak (box plot) berjaya mengenal pasti 421 gen yang diekspresi secara berbeza (DEG). Seterusnya, analisis kluster dilakukan untuk memperhalusi penemuan ini dengan menggunakan kaedah seperti ujian Wilcox, analisis subtipe, Analisis Komponen Utama (PCA), dan Analisis Rangkaian Ko-ekspresi Gen Bertimbang (WGCNA). Hasil menggunakan teknik-teknik ini, sejumlah 121 gen teras yang penting dalam mekanisme yang dikaji telah berjaya disaring. Analisis rajah Venn yang mengabungkan gen DEG dan gen teras, seterusnya mengenal pasti HCLS1 sebagai gen utama yang menjadi sasaran molekul kritikal. Gen utama yang dikenalpasti melalui kaedah bioinformatik seterusnya disahkan menggunakan kaedah eksperimental menggunakan model in vitro sel Caco-2. Empat keadaan diuji: 1) sel tidak dirawat (kawalan negatif), 2) sel yang dirawat dengan LPS (kawalan positif), 3) sel yang dirawat dengan ATP, dan 4) sel yang dirawat dengan gabungan LPS dan ATP. Analisa ke atas model-model ini menggunakan RT-qPCR mengesahkan berlakunya ekspresi gen HCLS1. Analisa menggunakan pewarnaan fluoresen dan juga sitometri aliran juga berjaya mengesan dan mengukur apoptosis dan potensi membran mitokondria (MMP). Kesimpulannya, kajian ini berjaya mengenal pasti HCLS1 sebagai gen utama yang terlibat dalam kematian sel yang diinduksi oleh eATP dalam penyakit radang usus (IBD). Dengan menggabungkan pendekatan bioinformatik dan eksperimen, penemuan ini memberikan pemahaman yang lebih mendalam tentang mekanisme molekul yang terlibat dalam proses kematian sel ini. Lebih penting lagi, ia membuka potensi untuk pengembangan sasaran terapeutik

baru bagi pengurusan IBD, yang boleh membawa kepada rawatan yang lebih berkesan di masa hadapan.

IDENTIFICATION AND VALIDATION OF HUB GENES INVOLVED IN ATP-INDUCED CELL DEATH IN THE PATHOGENESIS OF INFLAMMATORY BOWEL DISEASE

ABSTRACT

Extracellular adenosine triphosphate (eATP) plays a critical role in inflammatory bowel disease (IBD) as a key regulator of cell death processes. The P2X7 receptor, found on intestinal epithelial cells, is frequently implicated in cell death mechanisms involving eATP. However, the specific mechanisms of eATP-induced cell death and their connection to IBD remain unclear. Therefore, the primary objective of this study is to integrate data collection, bioinformatics analysis, and experimental validation to identify and confirm hub genes involved in eATP-induced cell death in the context of IBD. This approach combines computational and laboratory methods to produce reliable and high-confidence results. Not only does this approach enhance understanding of the molecular pathways involved in ATP-induced cell death, but it also offers potential therapeutic targets for managing IBD. The experiment begins with data collection, focusing on acquiring relevant datasets and gene sets from public databases and scientific literature. For transcriptomic data, keywords such as "IBD" and "transcriptome" were used to search the GEO database (Gene Expression Omnibus, specify abbreviation). This search successfully extracted 37 gene sets relevant to IBD studies. Simultaneously, keywords such as "ATP," "cell death," "apoptosis," "autophagy," and "extracellular ATP" were utilized to search literature databases, including PubMed, Embase, Web of Science, and Scopus, resulting in the identification of 19 gene sets associated with ATP-induced cell death. These collected data were then processed using bioinformatics methods to identify key target genes.Differential gene

expression analysis using visualization tools such as volcano plots and box plots identified 421 differentially expressed genes (DEGs). Subsequently, cluster analysis was conducted to refine these findings, employing methods such as the Wilcox test, subtype analysis, Principal Component Analysis (PCA), and Weighted Gene Co-Expression Network Analysis (WGCNA). These techniques successfully identified 121 hub genes critical to the studied mechanism. Venn diagram analysis, which integrated the DEG and hub gene data, ultimately identified HCLS1 as the primary gene serving as a critical molecular target. The key gene identified through bioinformatics methods was further validated using experimental approaches with Caco-2 in vitro cell models. Four conditions were tested: untreated cells (negative control), cells treated with LPS (positive control), cells treated with ATP, and cells treated with a combination of LPS and ATP. Analysis of these models using RT-qPCR confirmed the expression of the HCLS1 gene. Fluorescence staining and flow cytometry analyses successfully detected and quantified apoptosis and mitochondrial membrane potential (MMP) changes. Flow cytometry analysis specifically quantified apoptosis in treated cells. In conclusion, this study successfully identified HCLS1 as a key gene involved in cell death induced by eATP in inflammatory bowel disease (IBD). By integrating bioinformatics and experimental approaches, the findings provide a deeper understanding of the molecular mechanisms involved in this cell death process. More importantly, this opens potential for the development of new therapeutic targets for IBD management, which could lead to more effective treatments in the future.

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Overview of Inflammatory Bowel Disease (IBD)

1.1.1(a) Definition and Epidemiology

IBD (Inflammatory Bowel Disease) constitutes a persistent and recurring immune-mediated condition encompassing Crohn's disease, ulcerative colitis, Characterised by sustained inflammation of the intestinal mucosa, aberrant immune system activation, and consequential damage to the intestinal mucosa(Gilliland, Chan, De Wolfe, Yang, & Vallance, 2024; Law et al., 2024; Swaminathan et al., 2024; Zhezhe Tian, Zhao, & Teng, 2024). The incidence and prevalence of IBD have experienced a notable upsurge in recent decades, posing a substantial global health challenge. A comprehensive survey (Ng et al., 2017) revealed that Europe has the highest prevalence of IBD, with Norway having the highest prevalence of ulcerative colitis with 505 cases per 100,000 people, while Germany leading in the incidence of Crohn's disease (322 cases per 100,000 people). Similarly, North America reports elevated levels of IBD, with the United States demonstrating the highest incidence of ulcerative colitis (286 cases per 100,000 people) and Canada reporting the highest incidence of Crohn's disease (319 cases per 100,000 people). IBD affects over 0.3% of the population across various countries in North America, Oceania, and Europe. Notably, the incidence of IBD has been rising in newly industrialised nations in Asia, Africa, and South America since 1990. In the Asia-Pacific region, India records one of the highest instances of IBD, with 9.31 cases per 100,000 person-years (Mak, Zhao, Ng, & Burisch, 2020). From

1990 to 2017, China experienced a decrease in age-standardsed disability-adjusted life years, mortality rates, and years of life lost due to IBD, despite an increase in age-standardised rates of prevalence, incidence, and years lived with disability, suggesting it remains one of the low-endemic regions for IBD (Qiu et al., 2020). In Malaysia, the incidence and prevalence rates of IBD are low but increasing, with significant ethnic disparities, and the incidence of Crohn's disease (CD) is increasing at a faster rate than that of Ulcerative colitis (UC) (Hilmi et al., 2015).

1.1.1(b) Pathogenesis and influencing factors

IBD is a complex chronic intestinal disorder, involving multiple factors, including genetic predisposition, immune dysregulation, environmental influences, and intestinal microbiota (Ouahed, Griffith, Collen, & Snapper, 2024; Subramanian et al., 2024). IBD exhibits notable familial aggregation, with some cases showing a clear genetic tendency within families. Numerous genetic variants associated with disease risk, such as NOD2 and IL23R, which play crucial roles in immune response and intestinal mucosal barrier function, have been identified (Bourgonje, Ungaro, Mehandru, & Colombel, 2024; Masaki, Masuta, Honjo, Kudo, & Watanabe, 2024; Okai et al., 2024). Additionally, the pathogenesis of IBD is closely related to dysregulated immune system reactions. During the course of the disease, an imbalance in the immune response of the intestinal immune system to the gut microbiota leads to excessive inflammation, thereby damaging the intestinal tissue(Garavaglia et al., 2024; Grondin, Jamal, Mowna, Seto, & Khan, 2024). Recent research highlights the pivotal role of intestinal microbiota in the pathogenesis of IBD. The microbial composition in the intestines of patients with the disease typically differs from that of healthy individuals, and certain types of microorganisms may promote disease development by modulating

host immune responses or directly inducing inflammation (Matsuoka & Kanai, 2015; Nishida et al., 2018). Environmental factors such as diet, smoking, medication use, infections, and lifestyle, may also influence disease development. For example, a Western diet and smoking are considered risk factors, while breastfeeding and the use of over-the-counter medications may have protective effects on disease progression (Benninghoff et al., 2020; Majumder & Bano, 2024; Xue et al., 2024). These factors collectively constitute the etiological mechanism of IBD. However, further exploration is required to elucidate the specific pathogenic mechanisms involved.

1.1.1(c) Clinical Manifestations

The clinical manifestations of inflammatory bowel disease (IBD) vary depending on the type of disease, affected areas, and severity. The main symptoms include abdominal pain, diarrhoea, weight loss, and fatigue (Barreiro-de Acosta et al., 2023; Silaghi et al., 2022). Additionally, IBD patients may experience systemic and extraintestinal manifestations (Gordon et al., 2024; Guillo et al., 2022; Kilic et al., 2024).

CD often affects the entire digestive tract, most commonly the terminal ileum, and is charactersed by chronic or recurrent abdominal pain, diarrhoea, weight loss, and malnutrition (McGregor, Tandon, & Simmons, 2023). In severe cases, CD can lead to intestinal obstruction or fistula formation. Extraintestinal manifestations include arthritis, erythema nodosum, uveitis, and hepatobiliary diseases (Dolinger, Torres, & Vermeire, 2024; Lichtenstein et al., 2022).

UC mainly affects the colon and rectum, presenting with diarrhoea (often accompanied by blood and mucus), urgency, tenesmus, and abdominal pain. UC typically alternates between periods of remission and acute flare-ups (Peppercorn & Kane, 2020; Segal, LeBlanc, & Hart, 2021). Like CD, UC patients may also experience

extra intestinal manifestations such as arthritis, skin lesions, and hepatobiliary complications, including primary sclerosing cholangitis (Japparkulova, Aitymbetova, & Tuktibayeva, 2021; Singh & Dulai, 2022).

Both CD and UC patients frequently exhibit systemic symptoms such as fatigue, fever, and weight loss, which are often associated with chronic inflammation and malnutrition (Matsui et al., 2003; Yangyang & Rodriguez, 2017).

1.1.2 ATP-Induced Cell Death

1.1.2(a) Concept of ATP-induced cell death

ATP-induced cell death refers to the process in which extracellular ATP binds to specific receptors, triggering a series of cellular signalling pathways that ultimately lead to programmed cell death, or apoptosis (W. Wang, Zhang, et al., 2023b). In this process, extracellular ATP acts as an important signalling molecule, primarily activating purinergic receptors such as the P2X7 receptor. This activation induces calcium ion influx, membrane pore formation, mitochondrial damage, and inflammation activation. These changes subsequently activate caspase enzymes and other pro-apoptotic pathways, leading to cell death (H.-L. Zhang, S. Doblin, Z.-W. Zhang, Z.-J. Song, B. Dinesh, Y. Tabana, D. S. Saad, M. A. A. Adam, Y. Wang, & W. Wang, 2024). ATP-induced cell death plays a crucial role in maintaining tissue homeostasis and is also associated with various pathological processes, such as inflammation and cell damage in autoimmune diseases(H. Zhang et al., 2024; Z. Zhang et al., 2024). Therefore, the regulation of extracellular ATP concentration and its influence on cell death pathways is key to understanding the mechanisms of numerous diseases.

1.1.2(b) ATP as an Extracellular Danger Signal

ATP is not only the energy currency within cells but also plays a crucial role as a signalling molecule outside the cell. When cells are damaged, infected, or experience other stress responses, ATP is released from the intracellular environment into the extracellular space, acting as a "danger signal" to activate nearby immune cells(K. C.-Y. Huang et al., 2024; Ishikawa et al., 2024; Mrnjavac & Martin, 2025). Extracellular ATP binds to purinergic receptors, such as the P2X7 receptor, triggering inflammatory responses and immune reactions, promoting the release of inflammatory cytokines and the activation of inflammasomes (Kyawsoewin et al., 2024; Santana, de Lima, Silva e Souza, Barbosa, & de Souza, 2024). This signalling pathway is essential for the immune system's rapid response to infections and injuries. However, excessive ATP release and prolonged receptor activation can lead to an overactive inflammatory response, which is closely associated with the development of various chronic inflammatory diseases, such as inflammatory bowel disease (IBD) and autoimmune disorders (Di Virgilio, Dal Ben, Sarti, Giuliani, & Falzoni, 2017). Therefore, ATP, as an extracellular danger signal, not only serves as a critical mechanism for responding to stress but also plays a significant role in certain pathological conditions.

1.1.2(c) Role of ATP-induced cell death in Cellular Function

ATP is not only the energy currency within cells but also plays a crucial role as an extracellular signalling molecule in regulating cell death and maintaining cellular functions(Kepp et al., 2021) (K. C.-Y. Huang et al., 2024; Ishikawa et al., 2024; Mrnjavac & Martin, 2025). ATP-induced cell death occurs through the binding of ATP to purinergic receptors, such as the P2X7 receptor, triggering a series of signalling pathways that lead to apoptosis or other forms of programmed cell death (Ryoden &

Nagata, 2022; W. Wang, Zhang, Sandai, Zhao, Bai, Wang, Wang, Zhang, Zhang, Song, et al., 2023; H.-L. Zhang, S. Doblin, Z.-W. Zhang, Z.-J. Song, B. Dinesh, Y. Tabana, D. S. Saad, M. A. A. Adam, Y. Wang, & W. J. W. J. o. C. O. Wang, 2024). This process is essential for maintaining tissue homeostasis, removing damaged or abnormal cells, and regulating immune responses.

In terms of cellular function, ATP-induced cell death helps to eliminate damaged or dysfunctional cells, preventing them from affecting surrounding healthy tissues. At the same time, by promoting apoptosis, it plays a role in regulating cell renewal and tissue repair(H. Zhang et al., 2023). Moreover, ATP-induced cell death is crucial in immune responses, aiding in the removal of pathogen-infected cells and modulating inflammation. However, excessive ATP-induced cell death can lead to tissue damage and is associated with pathological conditions such as autoimmune diseases and chronic inflammatory diseases (Trautmann, 2009; H. Zhang et al., 2023).

1.1.3 Overview of Apoptosis

1.1.3(a) The concept and process of cell apoptosis

Apoptosis, also known as programmed cell death, is a highly regulated process of self-elimination that plays a key role in maintaining tissue homeostasis, development, and immune defines mechanisms (Gupta, Ambasta, Kumar, & sciences, 2021; Morana, Wood, & Gregory, 2022; Y. Wan, L. Yang, S. Jiang, D. Qian, & J. J. I. b. d. Duan, 2022). Unlike necrosis, which is caused by injury or external stimuli, apoptosis is an orderly, active process characterised by cell shrinkage, chromatin condensation, and membrane blebbing(Morana et al., 2022; Obeng, 2020). The process is typically triggered by intrinsic or extrinsic signals, such as cellular stress, DNA damage, or death

receptor signals, and occurs through two main pathways (Obeng, 2020): the intrinsic pathway, where mitochondria release cytochrome C to activate caspase enzymes, and the extrinsic pathway, where death receptors (such as the Fas receptor) activate the caspase cascade (Green, 2022; Lossi, 2022). The activation of the caspase family leads to the degradation of intracellular proteins, cytoskeleton, and DNA, ultimately resulting in cell shrinkage and apoptotic body formation. These apoptotic bodies are engulfed by macrophages, preventing inflammation. Apoptosis is crucial for growth, development, and disease defined in organisms, and abnormalities in the apoptotic process are closely linked to pathological conditions such as cancer and autoimmune diseases (Y. Wan, L. Yang, S. Jiang, D. Qian, & J. Duan, 2022).

Apoptosis is divided into three main phases: first, the initiation phase, where the cell is stimulated by intrinsic signals (such as DNA damage or oxidative stress) or extrinsic signals (such as the binding of death receptors and death ligands), triggering the apoptosis process. Next is the execution phase, during which the caspase cascade is activated, leading to the degradation of intracellular proteins, the cytoskeleton, and DNA, and causing notable morphological changes in the cell, such as cell shrinkage, chromatin condensation, and membrane blebbing. Finally, in the degradation phase, the apoptotic cell breaks down into apoptotic bodies, which are engulfed and cleared by macrophages, preventing an inflammatory response. Through the precise regulation of these three phases, apoptosis maintains tissue homeostasis and normal cellular functions.

1.1.3(b) Apoptosis detection indicators and assays

The detection of cell apoptosis can be assessed through various indicators. morphological changes such as cell shrinkage, chromatin condensation, and membrane blebbing are typical features of apoptosis, which can be observed under a microscope

(Akçapınar, Garipcan, Goodarzi, Uzun, & Communications, 2021; Chang et al., 2021). Deoxyribonucleic Acid (DNA) fragmentation is another important hallmark of apoptosis, and techniques such as DNA laddering and TdT-mediated dUTP Nick-End Labelling (TUNEL) assay are commonly used to detect this fragmentation (PRABOWO, BINTARA, YUSIATI, SITARESMI, & WIDAYATI, 2023). The increased activity of caspase enzymes (such as caspase-3 and caspase-9) is a critical step in the apoptotic process and can be measured through various methods (Unnisa, Greig, & Kamal, 2023). The externalization of phosphatidylserine, detectable by Annexin V staining, is an early marker of apoptosis. The loss of mitochondrial membrane potential, another early marker, can be detected using JC-1 dye (Gomes et al., 2022). The formation of apoptotic bodies and their clearance by phagocytic cells are characteristic features of late apoptosis, typically detected through flow cytometry or microscopy (Zhao et al., 2021).

1.1.4 Extracellular ATP induced apoptosis in IBD

In inflammatory bowel disease (IBD), extracellular ATP serves as an important signaling molecule that induces apoptosis in intestinal epithelial cells by activating specific receptors, such as the P2X7 receptor, which plays a crucial role in the pathogenesis of IBD. Studies have shown that ATP, through its binding to the P2 receptor family, particularly the P2X7 receptor, not only regulates intestinal immune responses but also promotes cell death, thereby exacerbating intestinal inflammation and damage. Research by Neves et al. (2014) found that in the inflamed intestinal mucosa of CD patients, P2X7 receptor expression is upregulated and colocalizes with dendritic cells and macrophages, promoting the release of inflammatory cytokines and inducing apoptosis. The activation of the P2X7 receptor is closely associated with an increase in pro-inflammatory cytokines such as TNF-α and IL-1β, while inhibition of

the receptor reduces inflammation, suggesting that the P2X7 receptor may be a potential therapeutic target for CD (Neves et al., 2014). Further research by Wan et al. (2016) demonstrated that in experimental colitis models, extracellular ATP regulates the inflammatory response through the P2X7 receptor, inflammasome, and Nuclear Factor kappa-light-chain-enhancer of activated B cells (NF-κB) signaling pathways. Blocking ATP release or degrading ATP effectively alleviates colitis, while inhibiting ATP degradation exacerbates tissue damage, indicating that ATP and its receptor P2X7 play a crucial role in colitis and could be potential therapeutic targets (P. Wan et al., 2016a). Additionally, Liu et al. (2023) showed that the ATP/P2X7R signaling pathway promotes colitis progression by activating macrophages and increasing the release of pro-inflammatory cytokines. In DSS-induced colitis mouse models and Crohn's disease patients, P2X7R activation increases epithelial cell apoptosis and disrupts intestinal barrier function. Inhibition of P2X7R or reduction of ATP levels effectively alleviates inflammation (Y. Liu et al., 2023). In conclusion, extracellular ATP-induced apoptosis via the P2X7 receptor plays an important role in the pathogenesis of IBD. The P2X7 receptor serves as a key regulator of inflammation and apoptosis, making it a potential therapeutic target for future IBD treatments.

1.1.5 Effect of Bioinformatics on IBD

Bioinformatics, the interdisciplinary field that combines biology, computer science, and information technology to analyse and interpret biological data, has significantly advanced the study of Inflammatory Bowel Disease (IBD). By enabling the analysis of vast amounts of genomic, transcriptomic, and proteomic data, bioinformatics has led to the identification of genetic variants associated with IBD, shedding light on the disease's complex pathogenesis. Through bioinformatics,

researchers can explore the interactions between genes, proteins, and environmental factors that contribute to the onset and progression of IBD. Additionally, bioinformatics tools facilitate the discovery of novel biomarkers for early diagnosis, the development of personalsed treatment strategies, and the identification of potential therapeutic targets. As a result, bioinformatics has become an essential component in understanding the underlying mechanisms of IBD and improving patient outcomes. For example, Garza-Hernandez D (2022) identified a total of 197 genes associated with Crohn's disease risk through bioinformatics methods, including 126 from specific experiments and 71 from genome-wide association studies (GWAS) (Garza-Hernandez et al., 2022). Similarly, Hu W (2022) identified 446 differentially expressed genes (DEGs), 9 hub genes, and 7 microRNAs (miRNAs) as potential biomarkers for ulcerative colitis through bioinformatics analysis (W. Hu, Fang, & Chen, 2022). The key research work using bioinformatics in this study is summarized in Table 1.1.

Table 1.1 Summary of Key Research Work

Analysis Method	Key Attributes	Purpose and Application	Data Type/Scope	Advantages and Limitations
Differential Analysis	Compares gene expression differences between groups using statistical methods (e.g., t-test, ANOVA)	Identifies significantly altered genes between different conditions or treatments, helping to uncover potential biomarkers and pathways	Gene expression data (e.g., RNA-Seq or microarray data)	Advantages: Identifies significant gene differences between groups; Limitations: Can be influenced by sample size and noise, may not capture complex regulatory mechanisms
Subtype Analysis	Divides samples into different subtypes based on similarity or differences (e.g., clustering or consistency analysis)	Identifies different biological subtypes, providing a basis for personalized treatment, and revealing biological differences between subtypes	Gene expression data, clinical data	Advantages: Can uncover potential subtypes; Limitations: Difficult to determine the optimal number of subtypes, may be affected by data quality
PCA Analysis	Dimensionality reduction technique that simplifies high-dimensional data by extracting principal components	Visualizes the major variation patterns in data, helping to uncover potential structures, groups, or outliers	High-dimensional data (e.g., gene expression data, phenotype data)	Advantages: Provides a clear visual of the main structure in data; Limitations: Focuses only on the largest variance, potentially missing smaller variations

WGCNA Analysis	Constructs gene co-expression networks and analyzes the associations and modules of genes	Identifies gene modules related to specific phenotypes, revealing potential regulatory relationships and biological functions	Gene expression data	Advantages: Can uncover complex gene relationships and functional modules; Limitations: Requires large sample sizes and high-quality data

1.1.5(a) Difference analysis

Differential analysis is a crucial method in bioinformatics, primarily used to compare gene expression, protein levels, or other biological features under different conditions (e.g., between treatment groups, or between disease and healthy groups) (Y. Chen et al., 2020; Rosati et al., 2024). The goal is to identify genes, molecules, or pathways that exhibit significant differences across conditions from large-scale biological datasets, thereby providing insights into their potential roles in biological processes, disease mechanisms, or drug responses. Differential analysis is widely applied to gene expression data (such as RNA-Seq or microarray data), proteomics data, epigenetic data, and other omics studies (D. Li et al., 2022; Rosati et al., 2024). A study used integrated bioinformatics analysis of two GEO microarray datasets to identify differentially expressed genes (DEGs) associated with gastric cancer (GC), revealing four significantly upregulated genes (OLFM4, IGF2BP3, CLDN1, and MMP1) that may serve as potential biomarkers for early diagnosis and prevention of GC. The findings also highlight key biological functions and pathways, such as the Wnt and tumour signalling pathways, involved in GC progression (C. Yang & Gong, 2021). Common difference analysis methods are Table 1.2 and Advantages and disadvantages of each method are Table 1.3.

Table 1.2 Common difference analysis methods

Method	Principle and Characteristic
T-test	This method is suitable for analyzing gene expression differences between two groups of samples, assuming the data follows a normal distribution.

ANOVA	Used for comparing gene expression across multiple groups of samples, typically applied when comparing three or more groups.
DESeq2 and edgeR	These tools are designed for RNA-Seq data, based on a negative binomial distribution model. They account for both the inherent variability in gene expression and the differences between samples.
limma	Commonly used for both microarray and RNA-Seq data, limma employs a linear model and empirical Bayes methods, allowing it to handle more complex experimental designs effectively.

Table 1.3 Difference analysis of the advantages and disadvantages of various methods

Method	Advantages	Disadvantages
t-test	Simple, easy to implement, intuitive	Only for two-group comparisons, assumes normal distribution
ANOVA	Handles multiple group comparisons, considers interaction effects	Sensitive to assumptions of normality and homogeneity of variances
DESeq2 and edgeR	Suitable for RNA-seq and large-scale data analysis	Not sensitive to low-expression genes, relies on large sample sizes
limma	Broad applicability, controls for multiple testing, handles complex designs	Assumes linearity, not sensitive to small sample sizes

Based on the above, for the transcriptomic data of all genes in IBD, the Limma package in R was employed for differential analysis. Genes exhibiting biological significance in differential expression were identified with a criterion of $|\log2FoldChange| > 1$ and adj. p-value < 0.05, which is well-suited for large-scale gene expression data. For the 19 ATP-induced cell death genes in the IBD transcriptomic dataset, the Wilcoxon test was applied for differential analysis. This test is appropriate for expression data with a non-normal distribution, and it effectively handles small sample sizes, enabling comparison of median differences between two independent samples.

1.1.5(b) Subtype analysis

Subtype analysis is a method used to classify a disease or condition into different subtypes based on specific characteristics such as gene expression patterns, clinical features, or responses to treatment. This analysis helps researchers and clinicians better

understand the heterogeneity of the disease, leading to more accurate diagnoses, personalsed treatments, and improved patient outcomes (Gill et al., 2022).

In diseases like cancer or inflammatory bowel disease, subtype analysis can reveal unique molecular or cellular mechanisms for each subtype, which may respond differently to various treatments. This approach allows for a more tailored therapeutic strategy and can identify distinct prognostic markers, contributing to a more nuanced understanding of disease progression and treatment response. Subgroup analysis typically relies on both unsupervised and supervised learning methods to identify potential subgroups within data (Gong et al., 2021; Han, Kong, Liu, Cheng, & Han, 2021). Xinyue Hu et al. discovered two distinct immune patterns (ClusterA and ClusterB) through the ConsensusClusterPlus method, with patients in ClusterA showing significantly lower levels of resting dendritic cells, M2 macrophages, resting mast cells, activated natural killer cells, and regulatory T cells compared to those in ClusterB (X. Hu, Ni, Zhao, Qian, & Duan, 2022). Common subtype analysis methods and the advantages and disadvantages of each method are shown in Table 1.4 and Table 1.5.

Table 1.4 Common methods of subtype analysis

Method		Principle and Characteristic
Consensus clustering		Through multiple random sampling and clustering operations, a variety of clustering results are generated, and the stability of the clustering structure is determined by calculating the similarities between these results. Consensus clustering effectively addresses the issue of sensitivity to initialization inherent in traditional clustering methods, thereby providing more reliable and stable clustering outcomes.
Clustering Analysis	Hierarchical Clustering	This method displays the hierarchical relationships between samples or genes using a dendrogram, making it particularly suitable for small datasets.
	K-means Clustering	Samples or genes are divided into K clusters, where the elements within each cluster share similar characteristics. This method is suitable for large datasets but requires prior determination of the K value.
	Spectral Clustering	Based on graph theory, this method considers the global structure of the data and is particularly effective in handling nonlinear relationships.

PCA	PCA is a dimensionality reduction technique that maps high-dimensional data (such as gene expression data) into a lower-dimensional space, facilitating visualization and subgroup identification. PCA maximizes the variance of the data by computing its principal components, thereby revealing the primary sources of variation within the data.	
Non-negative Matrix Factorization (NMF)	NMF is a matrix factorization technique commonly used to extract potential subgroups from gene expression data. It assumes that all elements in the data matrix are non-negative, making it well-suited for extracting biologically meaningful feature patterns from gene expression profiles.	
t-SNE and UMAP	t-SNE (t-Distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection) are popular dimensionality reduction techniques for data visualization. They map high-dimensional data into two or three-dimensional space, making it easier to observe the distribution of subgroups. These techniques are particularly effective at revealing complex nonlinear structures within data, thus aiding in the identification and separation of subgroups.	
Supervised Learning Methods	Support Vector Machines (SVM)	Suitable for classification of small-sample, high-dimensional data, SVM can effectively distinguish between different subgroups.
	Random Forest:	By constructing multiple decision trees, this method performs classification or regression analysis and is well-suited for complex, high-dimensional data.
	Logistic Regression	This approach builds a regression model to predict the probability of a sample belonging to different subgroups

Table 1.5 The advantages and disadvantages of each method are subtype analysed

Method	Advantages	Disadvantages
Consensus clustering	The main advantages of consensus clustering lie in its stability and robustness. By integrating multiple clustering results, it significantly reduces the uncertainty associated with single clustering instances, thereby enhancing the reliability of the outcomes. In addition, consensus clustering is less affected by outliers and noise, making it suitable for various clustering algorithms and demonstrating a high degree of flexibility. It reveals underlying structures and patterns in the data, showing particularly strong performance in high-dimensional data analysis. The visualization of the consensus matrix allows researchers to intuitively understand the relationships between samples, thus facilitating the interpretation of clustering results.	However, consensus clustering also has some drawbacks. First, it requires multiple runs of clustering algorithms and the construction of a consensus matrix, leading to high computational complexity and time costs, especially when handling large-scale datasets. Second, the choice of clustering algorithms and parameters for sample resampling may influence the final results, with different settings potentially resulting in varying clustering outcomes. Although consensus clustering provides more stable results, interpreting the biological or practical significance of the clusters in complex data can still be challenging. Finally, when faced with very large or high-dimensional datasets, the storage and computational demands of consensus clustering may become bottlenecks.
Hierarchical Clustering	1. Does not require pre-specifying the number of clusters. 2. Generates a dendrogram, which aids in data interpretation. 3. Suitable for small datasets.	1. High computational complexity, especially with large datasets. 2. Sensitive to noise and outliers. 3. Difficult to scale for large datasets.
K-means Clustering	1. Suitable for large-scale datasets. 2. Simple algorithm with high computational efficiency.	1. Requires pre-specifying the number of clusters (K), and choosing the wrong K can affect the results.

	3. Results are easy to interpret.	2. Sensitive to initial cluster centroids. 3. Only suitable for convex clusters, difficult to handle complex data.
Spectral Clustering	1. Can handle complex, non-linear cluster structures. 2. Does not require strong assumptions about cluster shapes. 3. Suitable for data with graph structures or similarity matrices.	1. High computational complexity, especially with large datasets. 2. Requires computing the graph Laplacian, leading to higher computational and storage costs. 3. Sensitive to the choice of graph and similarity measure.
Principal Component Analysis (PCA)	1. Helps reduce data dimensionality for easier visualization and analysis. 2. Provides an intuitive understanding of data variation. 3. Fast computational speed.	1. Assumes linear structure in the data, which may not capture non-linear relationships. 2. Results can be difficult to interpret, especially when principal components lack physical meaning. 3. Sensitive to outliers.
Non-negative Matrix Factorization (NMF)	1. Provides easily interpretable decomposition, suitable for non-negative data like gene expression or images. 2. Extracts biologically meaningful patterns from complex data. 3. Suitable for extracting patterns from high-dimensional data.	1. Requires pre-specifying the rank (dimensionality) of the decomposition. 2. Computationally expensive, especially with large datasets. 3. Can get trapped in local optima.
t-SNE (t-Distributed Stochastic Neighbor Embedding)	1. Effectively reveals the local structure of the data, especially useful for high-dimensional data visualization. 2. Commonly used to discover cluster structures in data.	1. High computational complexity, particularly with large datasets. 2. Not suitable for precise classification of subtypes. 3. Sensitive to hyperparameter selection (e.g., perplexity).
UMAP (Uniform Manifold Approximation and Projection)	1. High computational efficiency, suitable for large-scale datasets. 2. Retains both local and global data structures. 3. Performs well with complex data structures.	1. Sensitive to hyperparameter selection. 2. Results can be hard to interpret, especially with high-dimensional datasets. 3. Complex to handle certain data types (e.g., sparse data).

Consistency Clustering was chosen for subtype analysis over other clustering methods due to several key advantages. Unlike methods such as K-means, which require the pre-specification of the number of clusters, consistency clustering does not impose such a limitation. A consensus matrix is generated, allowing for the exploration of the natural relationships in the data, and the most meaningful number of subtypes can be selected based on visual inspection. This flexibility is particularly valuable in

this study, where the exact number of subtypes is unknown and may vary depending on the data distribution. Additionally, consistency clustering is well-suited for high-dimensional data, such as transcriptomic data. Complex relationships between genes and samples can be captured, making it more suitable for data with intricate structures compared to methods like K-means, which assume clusters are spherical and homogeneous. Small sample sizes, a common challenge in biological studies, can also be effectively handled by consistency clustering. Furthermore, the results of consistency clustering can be easily integrated with other analyses, such as differential expression analysis and functional enrichment, providing a more comprehensive biological interpretation. Given these considerations, consistency clustering was determined to be the most appropriate method for identifying subtypes in this study.

1.1.5(c) PCA

In bioinformatics research, Principal Component Analysis (PCA) is a commonly used dimensionality reduction method widely applied in subtype analysis (Bielińska-Wąż et al., 2024; Taguchi, 2024). PCA projects high-dimensional data into lower-dimensional space, helping researchers identify major variations among samples and uncover potential subtype structures. In subtype analysis, PCA captures the directions of maximum variance within the dataset, enabling visualization of sample clusters in two- or three-dimensional space, thereby providing an initial view of different subtypes. This approach not only simplifies complex data structures but also lays the foundation for further clustering analysis. Through PCA, researchers can identify natural groupings within samples, offering critical insights into disease heterogeneity and personalsed treatment. In many bioinformatics studies, PCA has become an indispensable tool for subtype analysis, revealing potential subtypes and hidden patterns within the data

(Grant, Skjærven, & Yao, 2021; Privé, Luu, Blum, McGrath, & Vilhjálmsson, 2020). Common dimensionality reduction methods and the advantages and disadvantages of each method in Table 1.6 and Table 1.7.

Table 1.6 Common dimensionality reduction methods

Method	Principle and Characteristic
Principal Component Analysis (PCA)	PCA is a linear dimensionality reduction technique that finds the principal components along the directions of maximum variance in the data and projects the data onto these components, reducing its dimensionality.
t-Distributed Stochastic Neighbor Embedding (t-SNE)	t-SNE is a non-linear dimensionality reduction technique that constructs a low-dimensional representation by preserving the pairwise similarities of neighboring data points in high-dimensional space.
Uniform Manifold Approximation and Projection (UMAP)	UMAP is a non-linear dimensionality reduction method that optimizes both local and global structures to map high-dimensional data to low-dimensional spaces, suitable for data visualization and clustering.
Independent Component Analysis (ICA)	ICA is a technique used to find statistically independent components in data, commonly used in signal processing, such as for blind source separation.
Linear Discriminant Analysis (LDA)	LDA is a supervised dimensionality reduction technique that seeks to find the components that best separate different classes in the data. It maximizes the between-class variance while minimizing the within-class variance.
Multidimensional Scaling (MDS)	MDS reduces the dimensionality of data while preserving the pairwise distances between points as much as possible. It attempts to minimize the error in representing these distances in a lower-dimensional space.

Table 1.7 Dimensionality Reduction Methods and Their Advantages and Disadvantages

Method	Advantages	Disadvantages
Principal Component Analysis (PCA)	Simple and widely used, reduces dimensionality linearly, interpretable	Assumes linearity, sensitive to outliers, may not capture non-linear structures
t-Distributed Stochastic Neighbor Embedding (t-SNE)	Good for visualizing high-dimensional data, non-linear, preserves local structures	Computationally expensive, does not preserve global structures, difficult to interpret
Uniform Manifold Approximation and Projection (UMAP)	Fast, captures both local and global structures, flexible	Sensitive to parameter tuning, results can vary based on parameters
Independent Component Analysis (ICA)	Suitable for finding statistically independent components, useful in signal processing	Assumes data is independent, not always effective for non-linear relationships
Linear Discriminant Analysis (LDA)	Maximizes class separability, works well for supervised tasks	Assumes normally distributed data with equal covariance, limited to supervised settings
Multidimensional Scaling (MDS)	Intuitive, preserves pairwise distances in a low-dimensional space	Can be computationally expensive, may not preserve large-scale structures

PCA was chosen over other dimensionality reduction methods, such as t-SNE or UMAP, due to its superior interpretability and its ability to preserve the linear structure of the data. PCA provided a clearer and more visual foundation for the subtype analysis.

1.1.5(d) WGCNA

In bioinformatics research, Weighted Gene Co-Expression Network Analysis (WGCNA) is a powerful tool used to identify modules and key genes within gene expression data (Langfelder & Horvath, 2008; Pei, Chen, & Zhang, 2017; Zelin Tian et al., 2020). WGCNA constructs gene co-expression networks to group genes into co-expression modules, thereby revealing the relationships between genes and their potential biological functions. The method first calculates the correlation between gene pairs and then converts them into a weighted network. The similarity between genes is assessed using a Topological Overlap Matrix (TOM). Through hierarchical clustering of the network, WGCNA identifies gene modules with similar expression patterns. Each module's "Module Eigengene" represents the overall expression trend of that module, which can be correlated with phenotypic data to help identify key modules and driver genes associated with specific biological processes or diseases. WGCNA is widely applied in the exploration of gene expression data, especially in identifying potential biomarkers and therapeutic targets (Zelin Tian et al., 2020). Liang W (2020) identified susceptibility modules related to cardiovascular disease in diabetes using the WGCNA method and identified HLA-DRB1, LRP1, and MMP2 as key genes (Liang, Sun, Zhao, Shan, & Lou, 2020). Wan Q (2018) identified co-expression modules and key genes associated with uveal melanoma recurrence and clinical traits using the WGCNA

method, suggesting potential prognostic markers (Q. Wan, Tang, Han, & Wang, 2018). The advantages and disadvantages of WGCNA are shown in Table 1.8.

Table 1.8 Advantages and disadvantages of WGCNA

Advantages	Disadvantages
1. Discovery of potential gene modules: Can reveal co-expression relationships between genes and identify functionally related gene modules.	1. High computational complexity: Processing large-scale gene data is computationally intensive, especially with a high number of samples and genes.
2. Unsupervised learning: Does not require pre-set phenotype labels and can automatically discover underlying structures in the data.	2. Sensitive to parameter selection: The construction of the network involves several parameters, and their selection can significantly affect the results, often requiring cross-validation for optimization.
3. Modular analysis: Genes are grouped into functional modules, simplifying complex gene expression data and identifying key modules.	3. Difficult to handle small sample sizes: Small sample sizes may lead to inaccurate gene correlation estimates, affecting network construction and module identification.
4. Easy to visualize and interpret: Results can be visualized using heatmaps, network diagrams, and other methods, making them easier to understand and interpret.	4. Cannot determine causal relationships: WGCNA only reveals correlations between genes and cannot establish causal relationships.
5. Association with clinical phenotypes: Can link gene modules with specific phenotypes, identifying key modules related to particular traits.	5. Association with phenotypes does not imply causality: The correlation between modules and phenotypes does not necessarily indicate a causal relationship.

1.2 Problem Statement

ATP serves as the primary energy currency within the cellular matrix, fuelling essential life processes. In contrast, extracellular ATP acts as a significant regulatory factor, influencing a variety of cellular activities. Studies have shown that extracellular ATP not only plays a crucial role in the progression of Inflammatory Bowel Disease (IBD) but also induces the death of IBD cells. However, the precise mechanisms underlying the effects of extracellular ATP in IBD remain inadequately understood. Current research on ATP and IBD indicates that ATP-induced cell death is closely associated with complex molecular interactions, particularly those involving the ATP channel protein P2X7. Based on this information, we have posed three critical research questions to gain a comprehensive understanding of ATP's role in IBD:

1. Is there a hub gene associated with ATP-induced cell death that plays a pivotal role in the initiation and development of IBD cells?
2. How does extracellular ATP treatment affect the expression of this hub gene in IBD cell lines?
3. Is there a correlation between the expression of these hub genes and the apoptosis rate of IBD cells?

By conducting a thorough analysis of ATP-induced cell death in the context of IBD, we aim to not only gain new insights but also identify potential therapeutic targets. Integrating transcriptomic data from IBD will enable the identification of key genes closely linked to ATP-induced cell death, which may exert significant regulatory influence on the pathogenesis of IBD. Subsequent experimental validation can confirm whether these key genes serve as viable therapeutic targets, thereby facilitating the development of novel therapies targeting the ATP signalling pathway and improving treatment outcomes for IBD.

1.3 Research hypothesis

1. Based on the literature on ATP-induced cell death, it is hypothesized that ATP plays a critical role in the pathogenesis and progression of inflammatory bowel disease (IBD) through the regulation of a specific gene set. Specifically, ATP may modulate the expression of genes associated with cell death, thereby influencing the survival-death equilibrium of intestinal epithelial cells and playing a pivotal role in the pathological processes of IBD. Consequently, it is proposed that genes involved in ATP-induced cell death may encompass several key genes, which contribute significantly to the onset and progression of IBD and may serve as novel therapeutic targets.

2. It is hypothesized that extracellular adenosine triphosphate (eATP) plays a crucial role in the pathogenesis of inflammatory bowel disease (IBD) by regulating the expression of the HCLS1 gene, which inhibits LPS-induced apoptosis in intestinal epithelial cells. Specifically, it is proposed that eATP may reduce cell death in inflammatory intestinal cells through modulation of HCLS1 expression and enhance mitochondrial membrane potential, thereby promoting cell survival. Furthermore, it is hypothesized that HCLS1, as a key gene, is involved in ATP-induced cell death and may also serve as a novel biomarker for IBD diagnosis, revealing potential therapeutic targets for eATP in the treatment of IBD.

1.4 Objectives

a) To identify the gene set associated with ATP-induced cell death via a comprehensive literature search which will be conducted across PubMed, Embase, Web of Science, and Scopus databases using the keywords "ATP," "cell death," "apoptosis," "autophagy," "necrosis," "death," and "extracellular ATP."

b) To determine the hub genes related to ATP-induced cell death in the pathogenesis of IBD using bioinformatics methods.

c) To validate the cytotoxicity effects of ATP towards LPS-induced Caco-2 cells on the identified hub genes via CCK8, cells staining and RT-qPCR.

1.5 Study design

This study consists of several key steps to investigate ATP-induced cell death in IBD (Figure 1.1). Addressing the key genes related to ATP-induced cell death, involved using keywords such as "ATP," "cell death," "apoptosis," "autophagy," "necrosis," "death," and "extracellular ATP" to screen for gene sets associated with ATP-induced cell death in public databases like PubMed, Embase, Web of Science, and Scopus.

Identifying key genes related to ATP-induced cell death in IBD, gene expression matrices from intestinal epithelial cells of healthy individuals and transcriptomic data from intestinal epithelial cells of IBD patients were also downloaded from the GEO database, and the expression levels of genes related to ATP-induced cell death were extracted. Wilcox test analysis was performed to identify ATP-induced cell death-related genes associated with IBD. Based on these differentially expressed genes, we conducted Subtype analysis on IBD samples to classify them into different IBD subtypes. Additionally, WGCNA analysis was used to identify key genes in different subtypes, and these were intersected with IBD-related differentially expressed genes to determine the key ATP-induced cell death genes associated with IBD.

Verifiction of genes identified through bioinformatics analysis are done with RT-qPCR that detect changes in expression of targeted genes during ATP-induced apoptosis. The apoptosis activity was also validated using flow cytometry and fluorescence staining to assess ATP-induced apoptosis in IBD cells.

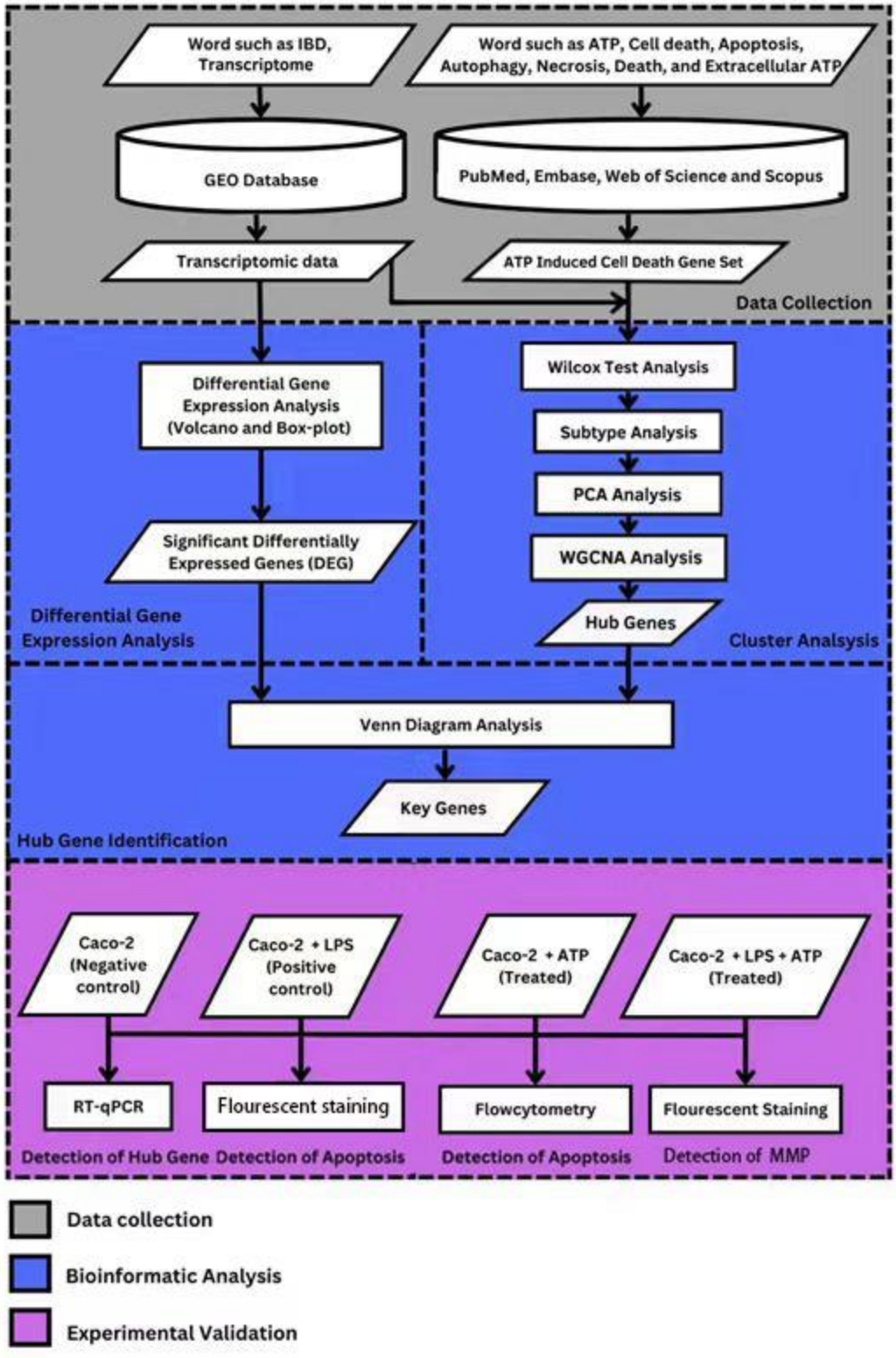

Figure 1.1 Flowchart of the study.

This study investigates ATP-induced cell death by retrieving literature from public databases to identify a relevant gene set. Transcriptomic data from IBD patients and healthy controls were analyzed to identify differentially expressed genes. subtypes

analysis, PCA and WGCNA analysis then classified the IBD subtypes and pinpoint key genes. An inflammation model using LPS-induced Caco-2 cells was created to compare gene expression under ATP intervention, with cell apoptosis assessed through flow cytometry and fluorescence staining.

This study reviews published literature to identify a gene set associated with ATP-induced cell death. Bioinformatics methods are then used to determine key genes related to ATP-induced cell death in Inflammatory Bowel Disease. Combined with cellular experiments, the study explores the potential roles of these key genes in IBD and their involvement in initiating the cell death process. A comprehensive understanding of this mechanism is expected to deepen our knowledge of IBD pathogenesis and provide valuable insights for future therapeutic approaches.

This thesis is structured into the following six chapters:

Chapter 1 Introduces the major concept of this study, including background, problem statement and objective and study design

Chapter 2 Discusses the relationship between ATP-induced cell death and IBD by integrating existing literature and provides a review of the known molecular mechanisms.

Chapter 3 Primarily introduces the bioinformatics analysis methods used in this study, as well as the detailed experimental procedures.

Chapter 4 Summarizes the research findings, with Section 4.1 presenting the results of the bioinformatics analysis, and Section 4.2 summarizing the validation of these results using LPS-induced CaCo-2 cells.

Chapters 5 Discussion of finding related to novelty, other study and gap knowledge.

Chapter 6 Conclusion and future prospects.