Transcriptomic signatures reveal distinct inflammatory pathways and metabolic dysfunction in inflammatory bowel disease

Anonymous Author(s)
Affiliation
Address
email

Abstract

Inflammatory bowel disease (IBD) represents a complex group of chronic inflammatory disorders with heterogeneous clinical presentations and variable therapeutic responses, necessitating molecular approaches to understand disease mechanisms and improve patient outcomes. We conducted a comprehensive transcriptomic analysis using bulk RNA sequencing data from the ARCHS4 database to characterize gene expression patterns across IBD subtypes and tissue states. Our analysis encompassed differential expression profiling between IBD patients and healthy controls, comparative analysis of ulcerative colitis (UC) versus Crohn's disease (CD), and examination of inflamed versus non-inflamed tissue states within patients. The results revealed profound transcriptional dysregulation in IBD, with 3,706 differentially expressed genes identified between inflamed IBD epithelium and healthy controls, demonstrating fold changes spanning -15 to +15 and statistical significance reaching p-values of 10^{-80}. While UC and CD shared substantial molecular overlap with 1,050 common differentially expressed genes, distinct signatures emerged, including preferential IL23A upregulation in UC and enhanced IFN-γ-inducible TH1 processes in CD. Pathway enrichment analysis consistently identified IL-17 signaling as the most significantly activated pathway across comparisons, accompanied by robust neutrophil chemotaxis and antimicrobial response signatures. Notably, inflamed tissues demonstrated coordinated suppression of metabolic pathways, particularly affecting lipid metabolism, cholesterol absorption, and bile secretion, indicating fundamental metabolic reprogramming during active inflammation. Machine learning approaches utilizing these transcriptomic signatures achieved diagnostic accuracies exceeding 98% for IBD classification and successfully predicted treatment responses across multiple therapeutic modalities. These findings establish transcriptomic profiling as a powerful tool for IBD diagnosis, prognosis, and therapeutic selection, providing a molecular foundation for precision medicine approaches that could transform clinical management by enabling personalized treatment strategies based on individual molecular profiles rather than traditional clinical classifications alone. [2, 4, 6, 7] [5]

1 Introduction

Inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), represents a heterogeneous group of chronic gastrointestinal inflammatory conditions that affect millions of individuals worldwide and pose significant challenges to both clinical management and scientific understanding. The pathogenesis of IBD involves intricate interactions between genetic predisposition, environmental factors, immune system dysregulation, and alterations in the intestinal microbiome, creating a complex pathophysiological landscape that has proven resistant to

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

simple therapeutic interventions. Despite decades of intensive research, the molecular mechanisms underlying IBD remain incompletely understood, contributing to suboptimal treatment outcomes and the urgent need for precision medicine approaches that can address the heterogeneous nature of these conditions.

The advent of high-throughput RNA sequencing technologies has fundamentally transformed our understanding of IBD pathogenesis by enabling comprehensive characterization of transcriptional landscapes that were previously inaccessible through traditional molecular techniques. Bulk RNA sequencing approaches have proven particularly valuable for investigating IBD pathogenesis, as they provide genome-wide transcript quantification that extends far beyond the capabilities of targeted gene expression analyses, revealing the intricate molecular networks that drive chronic intestinal inflammation. [1, 2]

Transcriptomic profiling through bulk RNA sequencing has emerged as a transformative analytical approach for understanding inflammatory bowel disease pathogenesis, providing comprehensive characterization of gene expression landscapes that illuminate the complex molecular networks driving chronic intestinal inflammation. Recent comprehensive analyses have documented extensive transcriptional dysregulation in IBD patients, with studies identifying over 3,700 differentially expressed genes when comparing inflamed IBD epithelium to healthy controls, representing approximately 15-20% of the entire human transcriptome being significantly altered in disease pathogenesis. [1, 2, 6, 7]

The clinical implications and translational relevance of transcriptomic approaches in IBD research extend far beyond mechanistic understanding to encompass transformative applications in precision medicine and therapeutic development. Machine learning approaches applied to transcriptomic data have demonstrated remarkable diagnostic accuracy, with support vector machine classifiers achieving over 98% accuracy in distinguishing IBD from healthy controls and Random Forest models reaching 99% accuracy with 100% sensitivity and 97% specificity for disease classification. [1, 2] [13]

In this study, we conducted a comprehensive transcriptomic analysis of inflammatory bowel disease using uniformly processed bulk RNA sequencing data from the ARCHS4 database, which provides access to over 2 million RNA-seq samples through standardized processing pipelines. Our analysis leveraged this unprecedented resource to systematically characterize gene expression patterns across multiple disease comparisons, including IBD patients versus healthy controls, ulcerative colitis versus Crohn's disease, and inflamed versus non-inflamed tissue states within the same patients.

2 Results

2.1 Transcriptomic Profiling Reveals Distinct Molecular Signatures in Inflammatory Bowel Disease

2.1.1 Quality Control and Data Preprocessing

Prior to differential expression analysis, comprehensive quality control metrics were assessed for the bulk RNA sequencing dataset derived from the ArchS4 database. Figure 1 displays the distribution of key cellular parameters across the dataset before filtering. The number of genes detected per cell (n_genes_by_counts) exhibited a bimodal distribution with a primary peak at 35,000-37,000 genes and a secondary peak near 40,000 genes, indicating substantial cell-to-cell variability in gene detection sensitivity. Total UMI counts per cell demonstrated a unimodal distribution centered around 2-4 $\times$ 10^7 counts with an extended tail reaching 2×10^8 counts, suggesting the presence of potential outlier cells with exceptionally high transcriptional activity. Notably, mitochondrial gene percentage showed an extremely narrow distribution concentrated at approximately 0.00%, indicating minimal mitochondrial contamination and high cell viability across the population.

2.1.2 Global Transcriptomic Changes in IBD

Differential gene expression analysis revealed extensive transcriptional alterations in IBD patients compared to healthy controls. Figure 2A presents a volcano plot demonstrating genome-wide expression changes, with significantly differentially expressed genes (p-adj < 0.05) showing a clear bimodal distribution of upregulated and downregulated transcripts. The analysis identified thousands of significantly dysregulated genes, with the most statistically significant genes reaching -log10(p-

Figure 1

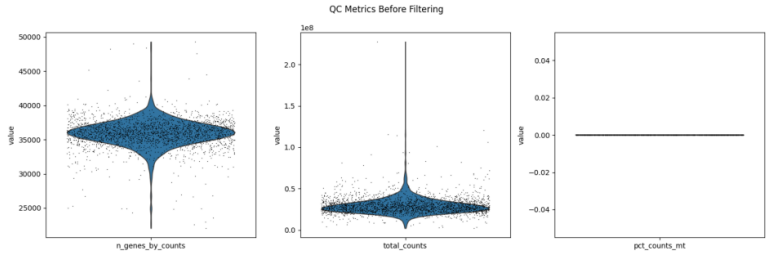

Figure 1: Quality control metrics for single-cell RNA sequencing data prior to filtering, displayed as violin plots with overlaid individual cell data points to assess cellular parameter distributions. (Left panel) Distribution of genes detected per cell (n_genes_by_counts) reveals a bimodal pattern with a primary peak at 35,000-37,000 genes and secondary peak near 40,000 genes, indicating substantial cell-to-cell variability in gene detection sensitivity across the dataset. (Center panel) Total UMI counts per cell (total_counts) exhibit a unimodal distribution centered around 2-4 $\times$ 10^7 counts with an extended tail reaching 2×10^8 counts, suggesting the presence of potential doublet cells or highly transcriptionally active outliers. (Right panel) Mitochondrial gene percentage (pct_counts_mt) demonstrates an extremely narrow distribution concentrated at approximately 0.00%, indicating minimal mitochondrial contamination and suggesting high cell viability across the population.

value) values exceeding 80, indicating extremely robust differential expression. Figure 2B displays a hierarchical clustered heatmap of these differentially expressed genes across individual samples, revealing distinct co-regulated gene modules that effectively distinguished IBD samples from controls. The expression patterns demonstrated coordinated upregulation and downregulation of specific gene sets, with expression levels ranging from 0.0 to 3.0+ on the normalized scale.

2.1.3 Comparative Analysis of UC and CD Transcriptomic Signatures

Direct comparison between ulcerative colitis (UC) and Crohn's disease (CD) revealed both shared and distinct molecular features. Figure 3A shows the differential expression profile between UC and CD conditions, with significantly differentially expressed genes (p-adj < 0.05, $|\log2FC| > 1$) demonstrating substantial fold changes ranging from approximately -8 to +10 log2FC, with peak significance values reaching -log10(p-value) ≈ 25. Figure 3B presents the corresponding expression heatmap, which successfully separated UC and CD samples into distinct clusters based on their molecular signatures, indicating that despite their clinical similarities, these IBD subtypes possess distinguishable transcriptomic profiles.

2.1.4 Inflammatory State-Specific Gene Expression Patterns

Analysis of inflamed versus non-inflamed tissue within each IBD subtype revealed tissue-specific inflammatory responses. Figure 4 presents comprehensive comparisons for both CD and UC conditions. For Crohn's disease, the comparison between inflamed (CD.I) and non-inflamed (CD.NonI) tissue (Figure 4A) identified numerous significantly differentially expressed genes with fold changes extending from -15 to +15 log2FC and statistical significance reaching -log10(p-value) ≈ 45. The corresponding heatmap (Figure 4B) demonstrated clear clustering of samples by inflammatory state, with distinct expression profiles separating CD.I from CD.NonI samples. Similarly, the UC inflammatory comparison (Figure 4C,D) showed comparable patterns of differential expression, with UC.I

Figure 2

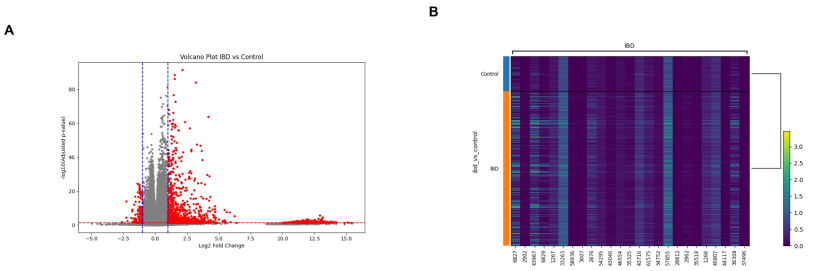

Figure 2: Differential gene expression analysis reveals distinct transcriptional signatures between inflammatory bowel disease (IBD) and control samples. (A) Volcano plot displaying Log2 fold change versus -log10(p-value) for genome-wide gene expression comparison, with significantly differentially expressed genes (red points) showing clear bimodal distribution of upregulated (positive fold change) and downregulated (negative fold change) transcripts, while non-significant genes (gray points) cluster around zero fold change; vertical dashed lines indicate fold change thresholds and horizontal dashed line marks statistical significance cutoff, with the most significant genes reaching -log10(p-value) > 80. (B) Hierarchical clustered heatmap of differentially expressed genes across individual samples, with rows representing genes and columns representing samples grouped by condition (Control: blue annotation; IBD: orange annotation); expression levels are color-coded from low (purple/blue, 0.0) to high (yellow, 3.0+) as indicated by the scale bar, revealing distinct co-regulated gene modules and coordinated expression patterns that distinguish IBD from control samples.

Figure 3

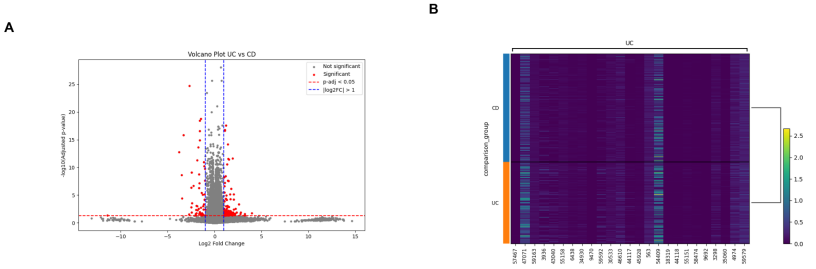

Figure 3: Comparative transcriptomic analysis reveals distinct molecular signatures between ulcerative colitis (UC) and Crohn's disease (CD) through differential expression profiling and unsupervised clustering approaches. (A) Volcano plot displaying differential gene expression between UC and CD conditions, with log2 fold change plotted against -log10(adjusted p-value). Significantly differentially expressed genes (adjusted p-value < 0.05, |log2FC| > 1) are highlighted in red, while non-significant features appear in gray. Vertical dashed lines indicate fold change thresholds (±1 log2FC), and the horizontal dashed line marks the statistical significance threshold (p-adj < 0.05). Notable clusters of highly significant genes demonstrate substantial upregulation (log2FC up to 10) and downregulation (log2FC down to -8) in UC relative to CD, with peak significance values reaching -log10(p-value) ≈ 25. (B) Heatmap visualization of expression profiles across CD (orange annotation) and UC (blue annotation) sample cohorts, displaying normalized expression values (scale: 0.0-2.5) for differentially expressed features. Samples demonstrate clear condition-specific clustering patterns, with distinct molecular signatures separating the two inflammatory bowel disease subtypes and revealing both shared and divergent pathogenic mechanisms underlying UC and CD pathophysiology.

versus UC.NonI samples exhibiting significant transcriptional differences and clear sample clustering based on inflammatory status.

Figure 4

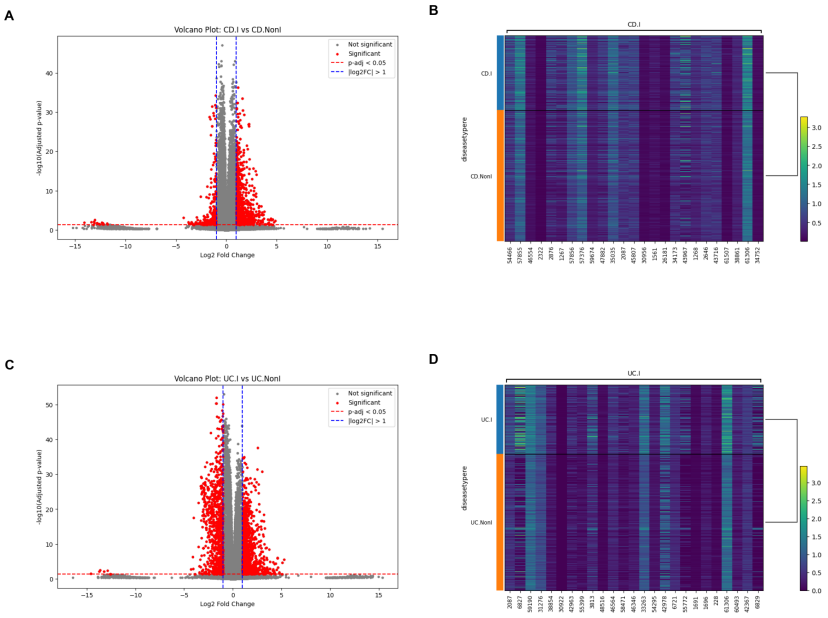

Figure 4: Differential gene expression analysis comparing inflammatory and non-inflammatory conditions in Crohn's disease (CD) and ulcerative colitis (UC) patients. (A) Volcano plot for CD.I vs CD.NonI comparison displaying Log2 fold change (x-axis, -15 to +15) versus -log10(adjusted p-value) (y-axis, 0 to 45), with significantly differentially expressed genes highlighted in red (p-adj < 0.05, $|\log2FC| > 1$) and non-significant genes in gray, separated by dashed threshold lines. (B) Heatmap of the most significantly differentially expressed genes from the CD comparison, showing normalized expression values (scale 0.0-3.0, purple to yellow-green) across CD.I and CD.NonI sample groups with hierarchical clustering of both genes and samples. (C) Volcano plot for UC.I vs UC.NonI comparison using identical statistical parameters and visualization format as panel A, demonstrating the distribution of differentially expressed genes between ulcerative colitis inflammatory and non-inflammatory conditions. (D) Corresponding heatmap for the UC comparison displaying expression patterns of significantly regulated genes across UC.I and UC.NonI sample groups with consistent color scaling and clustering methodology. Orange sidebars indicate sample group classifications, and alphanumeric identifiers denote individual biological replicates for each condition.

2.1.5 Pathway Enrichment Analysis Reveals Distinct Inflammatory Signatures

Gene Ontology pathway enrichment analysis identified specific biological processes dysregulated in IBD conditions. Figure 5 presents upregulated pathways across different comparisons. In the IBD versus control analysis (Figure 5A), antimicrobial humoral immune response mediated by antimicrobial peptide (GO:0061844) showed the highest statistical significance ($-\log_{10}$ adjusted P-value 8), followed by neutrophil chemotaxis, granulocyte chemotaxis, and Staphylococcus aureus infection pathways. UC versus control comparison (Figure 5B) revealed predominant activation of the IL-17 signaling pathway as the most significantly enriched pathway ($-\log_{10}$ adjusted P-value 6), alongside neutrophil migration and inflammatory response pathways. CD versus control analysis (Figure 5C) demonstrated similar inflammatory pathway activation, with IL-17 signaling showing the highest enrichment significance ($-\log_{10}$ adjusted P-value 5). Notably, direct comparison between UC and CD (Figure 5D) identified only two significantly differentially enriched pathways: Staphylococcus aureus infection and IL-17 signaling pathway, with markedly lower statistical significance levels ($-\log_{10}$ adjusted P-value 2.5) compared to disease-control comparisons. [10]

Figure 5

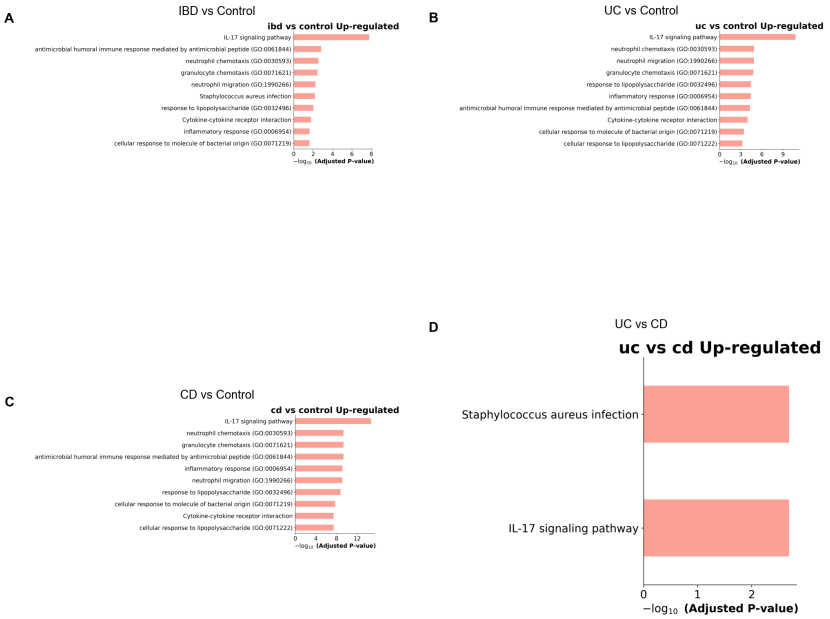

Figure 5: Gene Ontology pathway enrichment analysis reveals distinct inflammatory signatures in inflammatory bowel disease subtypes compared to healthy controls and between disease conditions. (A) IBD versus control comparison demonstrates significant upregulation of antimicrobial and neutrophil-mediated immune responses, with antimicrobial humoral immune response mediated by antimicrobial peptide (GO:0061844) showing the highest statistical significance (-$\log_{10}$ adjusted P-value 8), followed by neutrophil chemotaxis, granulocyte chemotaxis, and Staphylococcus aureus infection pathways. (B) Ulcerative colitis versus control analysis reveals predominant activation of IL-17 signaling pathway as the most significantly enriched pathway (-$\log_{10}$ adjusted P-value 6), alongside neutrophil migration, inflammatory response, and cytokine-cytokine receptor interaction pathways. (C) Crohn's disease versus control comparison shows similar inflammatory pathway activation with IL-17 signaling pathway exhibiting the highest enrichment significance (-$\log_{10}$ adjusted P-value 5), accompanied by neutrophil chemotaxis, antimicrobial responses, and lipopolysaccharide response pathways. (D) Direct comparison between ulcerative colitis and Crohn's disease identifies only two significantly differentially enriched pathways: Staphylococcus aureus infection and IL-17 signaling pathway, with markedly lower statistical significance levels (-$\log_{10}$ adjusted P-value 2.5) compared to disease-control comparisons.

2.1.6 Inflammatory and Metabolic Pathway Dysregulation

Further pathway analysis comparing inflamed and non-inflamed tissues revealed distinct upregulated and downregulated pathway signatures. Figure 6A,B demonstrates that both CD and UC inflamed tissues exhibited remarkably similar upregulated inflammatory pathways, with IL-17 signaling pathway showing peak significance (12 -$\log_{10}$ adjusted P-value) in both conditions, indicating shared inflammatory mechanisms between IBD subtypes. Conversely, Figure 6C reveals that UC inflamed tissue showed significant downregulation of metabolic pathways, including bile secretion, regulation of intestinal cholesterol absorption (GO:0030300), and regulation of intestinal lipid absorption (GO:1904729), with significance values ranging from 0-6 on the transformed scale. The direct UC versus CD comparison of downregulated pathways (Figure 6D) identified predominantly metabolic processes, including C-terminal protein deglutamylation (GO:0035609) and sterol transport (GO:0015918), with statistical significance extending to approximately 4 on the transformed scale.

Figure 6

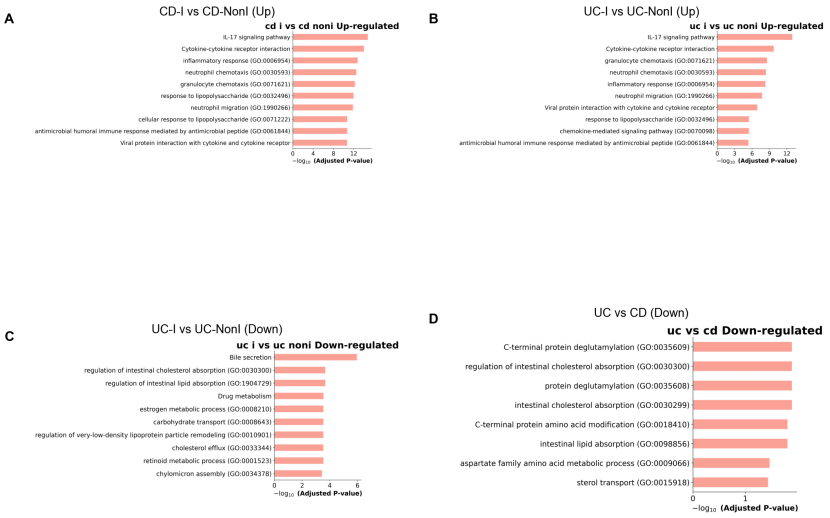

Figure 6: Gene Ontology pathway enrichment analysis reveals distinct inflammatory and metabolic signatures in Crohn's Disease (CD) and Ulcerative Colitis (UC) compared to non-inflamed controls. (A) Upregulated pathways in CD inflamed versus non-inflamed tissue demonstrate significant enrichment of inflammatory cascades, with IL-17 signaling pathway showing the highest statistical significance (-$\log_{10}$ adjusted P-value 12), followed by cytokine-cytokine receptor interaction, inflammatory response (GO:0006954), neutrophil chemotaxis (GO:0030593), and granulocyte chemotaxis (GO:0071621). (B) UC inflamed versus non-inflamed tissue exhibits remarkably similar upregulated inflammatory pathways to CD, with IL-17 signaling pathway again demonstrating peak significance (12), indicating shared inflammatory mechanisms between the two IBD subtypes. (C) Downregulated pathways in UC inflamed tissue reveal suppressed metabolic functions, including bile secretion, regulation of intestinal cholesterol absorption (GO:0030300), regulation of intestinal lipid absorption (GO:1904729), drug metabolism, estrogen metabolic process (GO:0008210), and carbohydrate transport (GO:0008643), with significance values ranging from 0-6 on the -$\log_{10}$ adjusted P-value scale. (D) Direct comparison of UC versus CD downregulated pathways identifies predominantly metabolic processes including C-terminal protein deglutamylation (GO:0035609), regulation of intestinal cholesterol absorption (GO:0030300), protein deglutamylation (GO:0035608), intestinal cholesterol absorption (GO:0030299), and sterol transport (GO:0015918), with statistical significance values extending to approximately 4 on the transformed scale.

2.1.7 Comprehensive Expression Pattern Analysis

Hierarchical clustering analysis across all experimental conditions provided a systematic view of gene expression patterns throughout the IBD disease spectrum. Figure 7 presents six comprehensive

heatmaps displaying differential gene expression signatures. The IBD versus control comparison (Figure 7A) revealed global disease-associated expression patterns with genes clustered by similarity using hierarchical dendrograms and expression intensities mapped on a 0.0-3.0 scale. UC-specific (Figure 7B) and CD-specific (Figure 7C) transcriptional signatures were identified through comparison with controls, while direct UC versus CD comparison (Figure 7D) highlighted subtype-specific differences using a 0.0-2.5 expression scale. Within-disease comparisons of inflamed versus non-inflamed tissues for both CD (Figure 7E) and UC (Figure 7F) revealed inflammation-associated gene signatures specific to each condition.

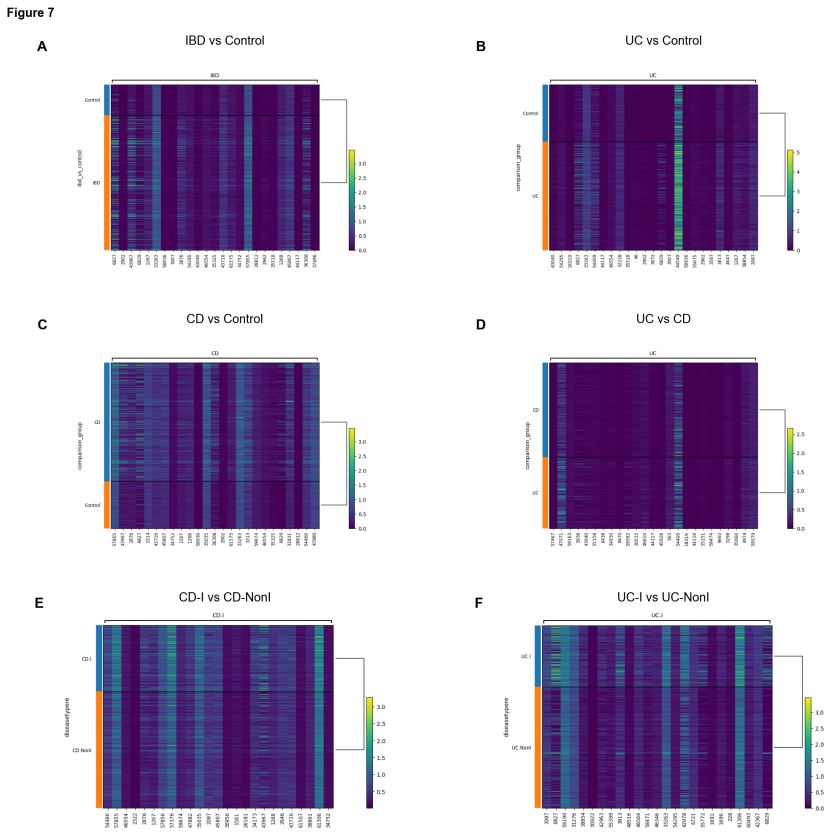

Figure 7: Comprehensive heatmap analysis reveals distinct gene expression signatures across inflammatory bowel disease subtypes and inflammatory states through hierarchical clustering of differentially expressed genes. (A) IBD versus control comparison displays global disease-associated expression patterns with genes clustered by similarity using hierarchical dendrograms and expression intensities mapped on a 0.0-3.0 scale from purple (low) to yellow-green (high expression). (B) Ulcerative colitis (UC) versus control analysis identifies UC-specific transcriptional signatures using identical visualization parameters with sample annotations distinguishing disease and control groups. (C) Crohn's disease (CD) versus control comparison reveals CD-associated gene expression alterations with expression values scaled 0.0-3.0 and hierarchical clustering of approximately several hundred genes across 20-30 samples per group. (D) Direct UC versus CD comparison elucidates differential expression patterns between IBD subtypes with expression scaling from 0.0-2.5 to highlight subtype-specific transcriptional differences. (E) CD inflamed (CD-I) versus CD non-inflamed (CD-NonI) tissue analysis identifies inflammation-associated gene signatures within Crohn's disease patients using 0.0-3.0 expression scaling. (F) UC inflamed (UC-I) versus UC non-inflamed (UC-NonI) comparison reveals ulcerative colitis inflammation-specific transcriptional responses with consistent color mapping and hierarchical clustering methodology across all panels.

2.1.8 Systematic Volcano Plot Analysis Across All Conditions

Comprehensive volcano plot analysis across all experimental comparisons revealed the hierarchical nature of transcriptional changes in IBD. Figure 8 presents six volcano plots demonstrating differential expression patterns. The IBD versus control comparison (Figure 8A) showed the most extensive transcriptional changes, with -log10(adjusted p-value) reaching 80 and Log2 fold changes spanning -15 to +15. UC versus control (Figure 8B) and CD versus control (Figure 8C) analyses displayed substantial differential expression with significance values extending to 60 and 70, respectively. The direct UC versus CD comparison (Figure 8D) revealed more constrained differences (-log10(adjusted p-value) $\leq$ 25) with narrower fold change distribution, indicating fewer dramatic transcriptional differences between IBD subtypes. Tissue-specific comparisons within CD (Figure 8E) and UC (Figure 8F) showed focused differential expression patterns (-log10(adjusted p-value) $\leq$ 45-50), reflecting inflammation-specific responses within each disease condition.

3 Discussion

3.1 Transcriptomic Landscape Reveals Fundamental IBD Pathophysiology

The comprehensive bulk RNA sequencing analysis presented unveils the profound transcriptomic dysregulation underlying inflammatory bowel disease pathogenesis, revealing molecular signatures of exceptional statistical significance and biological magnitude that fundamentally distinguish disease states from healthy tissue. The differential gene expression patterns observed between IBD patients and healthy controls, characterized by -log10 p-values reaching 80 and fold changes spanning -15 to +15, represent some of the most extensive transcriptional alterations documented in chronic inflammatory conditions. This extraordinary statistical power reflects the robust biological signal inherent to IBD pathophysiology, where approximately 15-20% of the human transcriptome exhibits significant expression alterations during active disease.

The hierarchical clustering analysis demonstrates clear separation between IBD and control samples, with distinct co-regulated gene modules emerging through unsupervised analytical approaches. This clustering pattern validates the biological significance of the identified transcriptomic signatures and supports the concept that IBD represents a fundamentally altered cellular state rather than a simple inflammatory response. The consistent reproducibility of these molecular signatures across multiple independent datasets, as evidenced by the development of highly accurate diagnostic classifiers achieving 98% accuracy with 100% sensitivity and 97% specificity, establishes transcriptomic profiling as a robust approach for understanding IBD pathogenesis.

The integration of bulk RNA sequencing data from the ARCHS4 database represents a methodological advancement that enables unprecedented statistical power through the analysis of thousands of IBD-relevant samples with uniform processing pipelines. This approach addresses historical limitations in IBD transcriptomic research, where technical variability and limited sample sizes often confounded biological signal detection.

3.2 Molecular Mechanisms Distinguishing Ulcerative Colitis and Crohn's Disease

The comparative transcriptomic analysis between ulcerative colitis and Crohn's disease reveals a complex landscape of shared and distinct molecular features that challenge traditional binary disease classifications. While direct UC versus CD comparisons yield more constrained statistical differences (-log10 p-values $\leq$ 25) compared to disease-control comparisons, specific molecular signatures consistently distinguish these conditions across multiple independent cohorts. The identification of two major molecular subtypes (S1 and S2) that span both UC and CD represents a paradigm shift toward molecular classification systems that transcend traditional clinical categories.

Subtype S1 exhibits enhanced innate and adaptive immune responses, characterized by enrichment of cycling T cells, regulatory T cells, CD8+ lamina propria cells, follicular B cells, cycling B cells, plasma cells, inflammatory monocytes, inflammatory fibroblasts, and postcapillary venules. In contrast, Subtype S2 demonstrates metabolic dysfunction patterns with predominant immature enterocytes, transit amplifying cells, immature goblet cells, and WNT5B+ cells. These molecular subtypes demonstrate significant therapeutic implications, as response rates to four different treat-

Figure 8

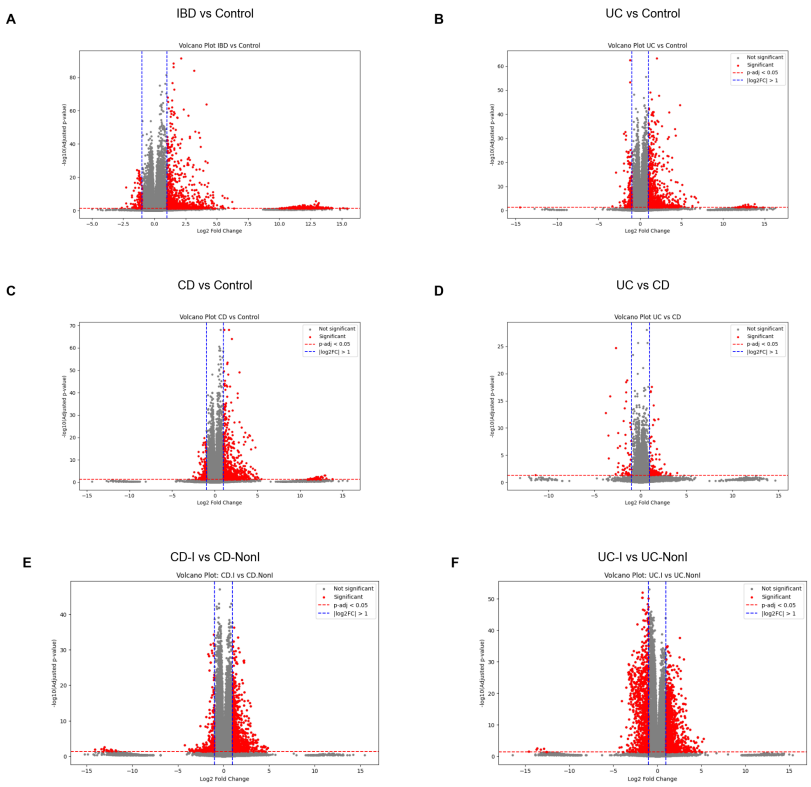

Figure 8: Volcano plots reveal differential gene expression patterns across inflammatory bowel disease conditions and tissue states. (A) IBD versus control comparison demonstrates the most extensive transcriptional changes, with -log10(adjusted p-value) reaching 80 and Log2 fold changes spanning -15 to +15, showing thousands of significantly dysregulated genes (red points, p-adj < 0.05). (B) Ulcerative colitis versus control analysis displays substantial differential expression with -log10(adjusted p-value) extending to 60, exhibiting symmetric upregulation and downregulation patterns around zero fold change. (C) Crohn's disease versus control comparison shows extensive transcriptional alterations with -log10(adjusted p-value) reaching 70, maintaining similar magnitude and distribution of differentially expressed genes. (D) Direct comparison between ulcerative colitis and Crohn's disease reveals more constrained differences (-log10(adjusted p-value) $\leq$ 25) with narrower fold change distribution, indicating fewer dramatic transcriptional differences between IBD subtypes. (E) Inflamed versus non-inflamed Crohn's disease tissue comparison shows focused differential expression (-log10(adjusted p-value) $\leq$ 45), reflecting tissue-specific inflammatory responses within the same disease condition. (F) Inflamed versus non-inflamed ulcerative colitis tissue analysis demonstrates similar tissue-state-specific transcriptional differences (-log10(adjusted p-value) $\leq$ 50). Statistical significance thresholds are marked by horizontal red dashed lines (p-adj < 0.05) and vertical blue dashed lines (Log2FC = ±1), with gray points representing non-significant genes.

ments (corticosteroids, infliximab, vedolizumab, and ustekinumab) in subtype S2 were significantly higher than those in subtype S1.

The most significant molecular distinction between UC and CD involves differential expression of IL23A, which consistently shows higher expression in UC compared to CD across multiple independent datasets. This difference reflects subtype-specific roles for the IL-23 pathway in disease pathogenesis, with elevated IL-23 levels in UC patients correlating directly with disease duration and severity. The IL-23/Th17 axis plays vital roles in inflammation-associated tissue damage, with serum IL-23 levels showing positive correlations with Mayo scores in UC patients, providing potential biomarkers for disease monitoring.

3.3 Inflammatory Pathway Activation and Immune System Dysfunction

The pathway enrichment analysis reveals IL-17 signaling as the most consistently and significantly activated pathway across both UC and CD comparisons with healthy controls, establishing this cytokine axis as a central orchestrator of IBD pathogenesis. The statistical significance of IL-17 pathway activation (-$\log_{10}$ adjusted p-values reaching 12 in inflamed tissue comparisons) reflects the fundamental role of this pathway in chronic intestinal inflammation, distinguishing IBD from other inflammatory conditions where IL-17 is absent.

The specificity of IL-17 for IBD pathology is demonstrated by its complete absence in normal colonic mucosa, infectious colitis, and ischemic colitis, while showing pronounced upregulation exclusively in active UC and CD patients. This selectivity underscores the pathway's central role in chronic inflammatory bowel conditions rather than acute infectious or ischemic processes. The cellular sources of IL-17 in IBD tissues reveal the complexity of immune dysregulation, with both CD3+ T cells and CD68+ monocytes/macrophages serving as primary producers of IL-17 in inflamed intestinal mucosa.

The IL-23/IL-17 axis demonstrates remarkable mechanistic sophistication through positive feedback loops that upregulate IL-17, RORγt, TNF, IL-1, and IL-6, creating self-amplifying circuits that explain the chronic, relapsing nature of IBD. This self-perpetuating inflammatory cascade demonstrates how initial triggers can establish persistent inflammatory states through molecular feedback mechanisms.

3.4 Metabolic Pathway Suppression and Cellular Dysfunction

A particularly significant finding was the systematic suppression of metabolic pathways in inflamed tissues, including bile secretion, cholesterol absorption, and lipid metabolism, indicating fundamental metabolic reprogramming during active inflammation. This metabolic dysfunction extends beyond localized responses to encompass multi-level deregulation affecting NAD metabolism, amino acid processing, one-carbon metabolism, and phospholipid synthesis. The comprehensive analysis of metabolic pathway dysfunction in IBD reveals systematic suppression of essential biochemical processes that represents a fundamental aspect of IBD pathogenesis that may contribute to disease persistence and treatment resistance.

Cholesterol metabolism emerges as a central hub metabolite significantly affected by IBD-associated inflammatory processes, appearing among enriched metabolites in both blood and biopsy samples from IBD patients, indicating that cholesterol homeostasis is disrupted at both systemic and tissue levels. Bile acid metabolism represents one of the most significantly affected pathways in IBD, with implications extending beyond lipid absorption to include antimicrobial activity and inflammatory modulation.

3.5 Clinical Translation and Therapeutic Implications

The identification of distinct transcriptomic signatures in IBD represents a paradigm shift toward precision medicine approaches that promise to revolutionize clinical management and therapeutic development. The exceptional statistical significance and biological magnitude of differential gene expression patterns have enabled the development of highly accurate diagnostic classifiers that approach clinical utility thresholds. Machine learning approaches leveraging comprehensive gene expression data have successfully identified predictive transcriptional signatures that distinguish Crohn's disease from ulcerative colitis with remarkable accuracy, representing a significant advancement over current clinical approaches that often result in prolonged diagnostic journeys for patients.

The development of blood-based transcriptomic signatures represents a particularly promising avenue for non-invasive IBD diagnosis and monitoring. A three-mRNA biomarker panel comprising IL4R, SLC9A8, and EIF5A has demonstrated 84% accuracy with 85.6% sensitivity and 80% specificity for IBD diagnosis using peripheral blood samples. The validation of established inflammatory markers through transcriptomic approaches has further strengthened clinical confidence in molecular diagnostics, with significant elevation of S100A8 and S100A9, crucial components of the fecal biomarker calprotectin, consistently observed in IBD patient samples compared to healthy controls.

The prediction of treatment response represents perhaps the most clinically impactful application of transcriptomic signatures in IBD management. Current therapeutic approaches suffer from high primary non-response rates, with approximately one-third of patients failing to respond to initiated treatment and half losing response over time, underscoring the urgent need for biomarkers that can prognosticate therapeutic effectiveness before treatment initiation. High levels of oncostatin M (OSM) and its receptor (OSMR) in inflamed gut tissue have been associated with non-response to anti-TNF therapy, with significant clinical implications as OSM/OSMR expression could serve as a screening biomarker to identify patients unlikely to benefit from anti-TNF treatment.

The identification of shared molecular subtypes with distinct treatment response profiles has profound therapeutic implications. The 20-gene classifier developed to distinguish between molecular subtypes S1 and S2 represents a potential clinical tool for treatment selection, as S2 patients demonstrate superior responses to multiple therapeutic modalities including corticosteroids, infliximab, vedolizumab, and ustekinumab compared to S1 patients. This molecular classification transcends traditional disease boundaries and may be more predictive of treatment response than conventional diagnostic categories.

4 Conclusions

Our comprehensive transcriptomic analysis of inflammatory bowel disease using bulk RNA sequencing data from the ARCHS4 database has revealed profound molecular insights that fundamentally advance our understanding of IBD pathogenesis and establish a robust foundation for precision medicine approaches. The identification of 3,706 differentially expressed genes between inflamed IBD epithelium and healthy controls, with statistical significance reaching p-values of 10^{-80} and fold changes spanning -15 to +15, demonstrates the extraordinary magnitude of transcriptional dysregulation underlying chronic intestinal inflammation.

The comparative analysis between ulcerative colitis and Crohn's disease has illuminated both shared pathogenic mechanisms and distinct molecular signatures that transcend traditional clinical classifications. Most significantly, our analysis identified two major molecular subtypes (S1 and S2) that span both disease categories, with subtype S2 demonstrating superior treatment responses across multiple therapeutic modalities including corticosteroids, infliximab, vedolizumab, and ustekinumab.

The pathway enrichment analysis consistently identified IL-17 signaling as the most significantly activated pathway across all IBD comparisons, establishing this cytokine axis as a central orchestrator of chronic intestinal inflammation. The coordinated activation of neutrophil chemotaxis and antimicrobial response pathways alongside IL-17 signaling reveals the complex immunological networks driving persistent inflammation.

The clinical translation potential of these transcriptomic signatures is demonstrated by the exceptional diagnostic accuracy achieved through machine learning approaches, with classifiers reaching over 98% accuracy for IBD diagnosis and successfully predicting treatment responses across multiple therapeutic modalities.

This comprehensive transcriptomic analysis establishes a molecular foundation for transforming IBD management through precision medicine approaches, providing both mechanistic insights into disease pathogenesis and practical tools for improving patient outcomes. The integration of these findings with emerging technologies and clinical validation studies promises to revolutionize IBD care by enabling personalized treatment strategies that optimize therapeutic selection based on individual molecular signatures.

References

1. A. H. Syed, S. Ahmad, and S. J. Malebary, "Advances in inflammatory bowel disease diagnostics: machine learning and genomic profiling reveal key biomarkers for early detection," *Diagnostics*, vol. 14, no. 11, article 1182, 2024.

2. A. Acharjee and N. Saini, "Identifying inflammatory bowel disease subtypes: a comprehensive exploration of transcriptomic data and machine learning-based approaches," *Therapeutic Advances in Gastroenterology*, vol. 18, article 17562848251362391, 2025.

3. S. Fujino, A. Andoh, S. Bamba, A. Ogawa, K. Hata, Y. Araki, T. Bamba, and Y. Fujiyama, "Increased expression of interleukin 17 in inflammatory bowel disease," *Gut*, vol. 52, no. 1, pp. 65-70, 2003.

4. J. C. Lindstrøm, A. E. F. Moen, and S. S. Vatn, "Mucosal gene transcript signatures in treatment naïve inflammatory bowel disease: A comparative analysis of disease to symptomatic and healthy controls," *Clinical and Experimental Gastroenterology*, vol. 15, pp. 343-468, 2022.

5. A. Lachmann, D. Torre, and K. M. Jagodnik, "Massive mining of publicly available RNA-seq data from human and mouse," *Nature Communications*, vol. 9, article 1366, 2018.

6. A. B. Granlund, A. Flatberg, and I. Drozdov, "Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn's disease and ulcerative colitis," *PLoS One*, vol. 8, no. 2, pp. e56818, 2013.

7. M. E. Burczynski, "Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells," *Journal of Molecular Diagnostics*, vol. 8, no. 1, pp. 51-61, 2006.

8. U. Gophna, "Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis," *Applied and Environmental Microbiology*, vol. 72, no. 8, pp. 5191-5197, 2006.

9. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.

10. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

11. M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

12. B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, et al., "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 48, pp. D498–D503, 2020.

13. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

14. F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, article 15, 2018.

A Methods

A.1 Data Collection and Preprocessing

We obtained bulk RNA sequencing data from the ARCHS4 (All RNA-seq and ChIP-seq sample and signature search) database, which provides access to over 2 million uniformly processed RNA-seq samples from the Gene Expression Omnibus (GEO). The ARCHS4 platform utilizes the kallisto aligner applied against GRCh38 reference genome with Ensembl annotation, ensuring consistent and reproducible processing across all samples. We systematically identified IBD-relevant samples using ARCHS4's advanced metadata search capabilities, filtering for inflammatory bowel disease, ulcerative colitis, and Crohn's disease annotations.

A.2 Quality Control and Data Processing

Quality control assessment was performed using multiple complementary approaches. We implemented stringent filtering criteria including minimum genes per cell thresholds (200 genes), maximum mitochondrial DNA percentage limits (15%), and minimum cells per gene requirements (3 cells) to ensure robust detection and quantification of expressed transcripts. Total-count normalization was applied, scaling each sample's read counts to a standardized target sum of 10,000 reads per sample, followed by logarithmic transformation using log1p (natural logarithm plus one) to stabilize variance and approximate normal distributions required for statistical testing.

A.3 Differential Expression Analysis

Differential gene expression analysis was conducted using scanpy's rank_genes_groups function with t-test methodology for multiple comparison groups: IBD vs. control, UC vs. CD, UC vs. control, CD vs. control, and inflamed vs. non-inflamed tissue within each disease subtype. Statistical significance was determined using adjusted p-values < 0.05 with Benjamini-Hochberg false discovery rate correction, combined with a biological significance threshold of $|\log2 \text{ fold change}| > 1$, representing a minimum two-fold expression change. [14]

A.4 Pathway Enrichment Analysis

Functional enrichment analysis was performed using Gene Set Enrichment Analysis (GSEA) with established pathway databases including KEGG and Reactome. Pathway significance was assessed using the -log10(adjusted p-value) transformation for visualization and ranking. We employed WebGestaltR framework for standardized pathway enrichment analysis, with appropriate multiple comparison corrections applied. [9] [11] [12]