UNRAVELING CROHN'S DISEASE HETEROGENEITY THOUGH MULTI-OMICS FROM DISEASE INCEPTION TO PROGRESSION

A Dissertation
Presented to
The Academic Faculty

by

Savannah Washburn

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Biology in the
School of Biological Sciences

Georgia Institute of Technology
August 2025

UNRAVELING CROHN'S DISEASE HETEROGENEITY THROUGH MULTI-OMICS FROM DISEASE INCEPTION TO PROGRESSION

Approved by:

Dr. Greg Gibson, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Abigail Lind
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Saurabh Sinha
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: July 21, 2025

To my family and friends who have constantly supported me

ACKNOWLEDGEMENTS

I am incredibly thankful for the experiences and growth throughout the past few years. First, I would like to thank my PhD advisor, Dr. Greg Gibson, for his mentorship and constant support throughout my PhD. I have really enjoyed learning from Greg. He pushed me to be more confident in myself and to grow as a scientist. I am very grateful for the opportunity to work with him and be a member of the Gibson lab.

I would also like to thank my committee members Dr. King Jordan, Dr. Peng Qiu, Dr. Abigail Lind, and Dr. Saurabh Sinha. I appreciate their time and energy, support, and advice. Their expertise helped shape the projects discussed in this thesis. Additionally, I would like to thank my collaborators: Dr. Subra Kugathasan and his lab, Dr. Judy Cho, Dr. Kyle Gettler, and Dr. Mark Lazarev. Each of these projects would not have been possible without their support, and I am very thankful for their contributions and clinical knowledge on Crohn's disease.

My time at Georgia Tech would not have been the same without my friends and lab mates. I feel very fortunate to have been surround by friends who provided mentorship and encouragement. I would like to thank past Gibson lab graduates, Dr. Maggie Brown and Dr. Emily Iannetta, for their friendship and mentorship. I would also like to thank current members of the Gibson lab, Dr. Sini Nagpal, Hira Anis, Varsha Bhat, Siming Zhao, and Desirée Bogen. I appreciated the laughter that we shared and thoughtful scientific discussions over the years.

Finally, I would like to thank my family for their love and support throughout my PhD. I am very grateful for my parents, Vance and Candice Washburn. They always encouraged me to chase my dreams and to set high expectations for myself, knowing I could do anything I set my mind to. From a young age, they fostered my sense of curiosity, which led to my love of science and research. I am also thankful for the support from my sisters, Abbigayle and Maggie. They kept me laughing throughout the ups and downs of my PhD. I would also like to thank my grandma for her love and support. Lastly, I am very grateful for Zachary Mudge. I am thankful for his support and always believing in me.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF SYMBOLS AND ABBREVIATIONS xii

SUMMARY xviii

CHAPTER 1. Introduction 1
1.1 Crohn's disease 2
1.1.1 Phenotypic classification of Crohn's disease subtypes 2
1.1.2 Aberrant immune response in Crohn's disease 4
1.1.3 Treatment approaches for Crohn's disease 5
1.1.4 Managing Crohn's disease in the post-operative setting 7
1.2 Applications of genomic technologies in Crohn's disease research 8
1.2.1 Bulk RNA-sequencing in Crohn's disease research 9
1.2.2 Crohn's disease insights from bulk transcriptomics 10
1.2.3 Alternative splicing in Crohn's disease 13
1.3 Single cell RNA-sequencing and Crohn's disease research 15
1.3.1 Complexities of single cell RNA-sequencing analysis 15
1.3.2 From cell types to signatures: single cell RNA-sequencing in IBD research 18
1.4 Proteomics in Crohn's disease research 22
1.4.1 Multi-omic integration: bridging genetics, transcriptomics, and proteomics 23
1.4.2 Advancing clinical applications of IBD research through proteomic profiling 25
1.5 Personalized medicine for Crohn's disease 25

CHAPTER 2. Persistent inflammation of the rectum in perianal fistulizing Crohn's disease is associated with goblet cell function 28
2.1 Introduction 28
2.2 Materials and Methods 28
2.2.1 Study population for perianal fistulizing Crohn's disease patients 29
2.2.2 Tissue processing for scRNA-seq 29
2.2.3 Gene expression profiling 30
2.2.4 Robustness and batch effect comparison 31
2.2.5 Partitioning epithelial and immune cells 32
2.2.6 Cell type annotation 32
2.2.7 Differential gene expression analysis 32
2.3 Results 33
2.4 Discussion 36

CHAPTER 3. Identification of Crohn's disease subtypes in single cell RNA sequencing signatures of treatment naïve samples across the paediatric gastrointestinal tract **38**
3.1 Introduction **38**
3.2 Materials and Methods **40**
3.2.1 Patient recruitment and ascertainment 40
3.2.2 Phenotypic classifications of B1, B2, and B3 41
3.2.3 Sample preparation, processing and quality control 41
3.2.4 Stability assessment 42
3.2.5 Cell proportion and hierarchical clustering analysis 44
3.2.6 Tucker tensor decomposition 45
3.2.7 Differential gene expression and pathway analysis 46
3.2.8 Cell-cell communication 46
3.2.9 Gene module score 47
3.2.10 Disease behaviour bias in the ileum 47
3.2.11 PCA in colon 48
3.2.12 GWAS and scRNA-seq integration analysis 48
3.3 Results **49**
3.3.1 Patient recruitment and sample preparation 49
3.3.2 Filtering for stably assigned cells improves overall clustering stability and similarity 49
3.3.3 Modest influence of inflammation on cellular proportions 53
3.3.4 scITD stratifies donors into potentially clinically important groups 56
3.3.5 Groups one vs. two in the ileum are biased towards B2 vs. B3 signatures independently identified in the RISK cohort 58
3.3.6 Myeloid cell activation accompanied by inflammation stratifies colon samples 61
3.3.7 Inflammation is associated with interferon gamma in group one donors in the rectum 65
3.3.8 Myeloid cells and T cells are associated with Crohn's disease across tissue 67
3.4 Discussion **69**

CHAPTER 4. Post-operative ileum transcriptomics implicate sex-biased mechanisms in Crohn's disease recurrence **77**
4.1 Introduction **77**
4.2 Materials and Methods **83**
4.2.1 Patient exclusion criteria 83
4.2.2 Biopsy preservation and RNA isolation 83
4.2.3 Library preparation and sequencing 83
4.2.4 RNA-seq data collection and QC 84
4.2.5 Differential transcript usage analysis 84
4.2.6 Tissue specific differential gene expression analysis 85
4.2.7 Tissue associated DEG pathway analysis 85
4.2.8 Stratification of recurrence status 86
4.3 Results **86**
4.3.1 Cohort phenotype/clinical summary and serial sampling 86

4.3.2 Differential splicing and differential gene regulation demonstrate rectalization associated with recurrent ileal disease 88
4.4 Discussion 90

CHAPTER 5. Correlated multi-omic signatures inform potential clinical stratification in post-operative Crohn's disease 94
5.1 Introduction 94
5.2 Materials and Methods 97
5.2.1 Patient recruitment and data collection 98
5.2.2 Sample preparation and RNA sequencing 98
5.2.3 Differential expression analysis 99
5.2.4 Batch effect correction 101
5.2.5 Correlation analysis 101
5.2.6 Permutation test 102
5.2.7 Group comparison 102
5.2.8 Recurrence status stratification 103
5.3 Results 104
5.3.1 Cohort description 104
5.3.2 Serum protein expression implicates recurring disease sex bias 105
5.3.3 Ileal gene expression associated with recurring disease and anti-TNF use 107
5.3.4 Correlation between ileal gene expression and serum protein expression 109
5.3.5 Correlated groups of genes and proteins stratify donors 112
5.3.6 Recurrence status stratification 114
5.4 Discussion 115

CHAPTER 6. Conclusions and Future Directions 121

PUBLICATIONS 127

APPENDIX A. Supplementary Tables 128

APPENDIX B. Supplementary Figures 140

REFERENCES 158

LIST OF TABLES

Table 1 Rutgeerts score to guide classification of recurring disease. 8

Table 2 Summary of post-op cohort characteristics. 87

LIST OF FIGURES

Figure 1	Montreal classification of Crohn's disease.	3
Figure 2	Key differences between scRNA-seq (top) and RNA-seq (bottom).	16
Figure 3	Representation of rectal cell types.	34
Figure 4	Signatures of inflammation in goblet cells.	36
Figure 5	Stability assessment and rigor of ileum clustering results.	50
Figure 6	Clustering results reveal heterogeneity within and across tissue.	54
Figure 7	scITD stratifies donors into clinically important groups.	57
Figure 8	Ileum group 1 vs. 3 demonstrate B2 vs. B3 bias independently identified in the risk cohort.	59
Figure 9	Myeloid cell activation accompanied by inflammation stratifies colon samples.	63
Figure 10	Inflammation associated with interferon gamma is enriched in the rectum of group 1 donors.	66
Figure 11	Cell types associated with CD.	68
Figure 12	Differential transcript usage and expression indicate that recurrence status is associated with altered splicing biased towards the rectum.	89
Figure 13	Sex-bias associated with recurring disease at the proteomic level.	106
Figure 14	Differentially expressed genes associated with recurrence status and sex.	109
Figure 15	Correlation between ileal gene expression and serum protein expression.	110
Figure 16	Coordinated gene and protein expression stratify groups of patients.	113
Figure 17	ROC curves of 3-fold cross-validation of logistic regression models with PCs 1-5 of gene and protein Group signatures.	115

LIST OF SYMBOLS AND ABBREVIATIONS

% Exp	Percent Expressed
AA	African American
AI	Artificial Intelligence
AIC	Akaike Information Criterion
Ambig.	Ambiguous
ANB	Activated Naïve B Cell
ANOVA	Analysis of Variance
Anti-TNF	Anti-Tumor Necrosis Factor
ARI	Adjusted Rand Index
AS	Alternative Splicing
AUC	Area Under the Curve
Avg. Exp.	Average Expression
B/OCol	BEST4/OTOP2 Colonocytes
BH	Benjamini-Hochberg
BIC	Bayesian Information Criterion
CB	Cycling B Cells
CCA	Canonical Correlation Analysis
CD	Crohn's disease
CDAI	Crohn's Disease Activity Index
CHOA	Children's Healthcare of Atlanta
Col	Colonocytes
ColProg	Colonocyte Progenitor

CRP	C-Reactive Protein
DEG	Differentially Expressed Gene
DEP	Differentially Expressed Protein
dIF	Differentially used Isoform Fraction
DNT	Double-Negative T cells
DSS	Dextran Sodium Sulfate
DTU	Differential Transcript Usage
EA	European American
ECM	Extracellular Matrix
EEN	Exclusive Enteral Nutrition
ELISA	Enzyme Linked Immunosorbent Assay
Entero	Enteroendocrine
eQTL	expression Quantitative Trait Loci
ESR	Estrogen Receptor
Exp. Var.	Explained Variance
F	Female
FC	Fold Change
FDR	False Discovery Rate
FMB	FCRL4+ Memory B Cell
$G1_C$	Colon Group 1
$G1_I$	Ileum Group 1
$G1_R$	Rectum Group 1
$G2_C$	Colon Group 2
$G2_I$	Ileum Group 2
$G2_R$	Rectum Group 2

GCB	Germinal Center B Cells
GIMATS	IgG plasma cells, Inflammatory Mononuclear phagocytes, and Activated T and Stromal cells
GO	Gene Ontology
Gob	Goblet Cell
GRC	Genetic Research Center
GSEA	Gene Set Enrichment Analysis
GWAS	Genome Wide Association Study
Har.	Harmony
Hisp.	Hispanic
HVG	Highly Variable Genes
IBD	Inflammatory Bowel Disease
IBDGC	Inflammatory Bowel Disease Genetics Consortium
IF	Inflamed
IF	Isoform Fraction
IFN	Interferon
ILC3	Type 3 Innate Lymphoid Cell
imGob	Immature Goblet Cell
INFLARES	Inflammatory Epithelial Cells
IPA	Ingenuity Pathway Analysis
IQR	Interquartile Range
IRB	Institutional Review Board
JI	Jaccard Index
KO	Knock-Out
Lti-like NCR+ ILCs	Lymphoid Tissue Inducer-like NCR+ Type 3 Innate Lymphoid Cell

M	Male
Macro IF	Macroscopic Inflammation
MAGMA	Multi-marker Analysis of GenoMic Annotation
MAST	Model-based Analysis of Single-cell Transcriptomics
Micro IF	Microscopic Inflammation
ML	Machine Learning
MNP	Mononuclear Phagocytes
mRNA	messenger RNA
MS	Mass Spectrometry
NB	Naïve B Cells
NCD4T	Naïve CD4+ T Cells
NI	Non-Inflamed
NK	Natural Killer
NPX	Normalized Protein eXpression
P. adjust	Bonferroni Adjusted P value
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PEA	Proximity Extension Assay
Perianal-CD	Perianal Fistulizing Crohn's Disease
Perm.	Permuted
PGR	Progesterone Receptor
Post-op	Post-Operative
pQTL	protein Quantitative Trait Loci
PRS	Polygenic Risk Score

PSI	Percentage Spliced In
QC	Quality Control
R0	Non-Recurring
R1	Recurring
RIN	RNA Integrity Number
RMB	Resting Memory B Cell
RNA-seq	RNA-sequencing
ROC	Receiver Operating Characteristic
rPCA	reciprocal Principal Component Analysis
RSEM	RNA-seq by Expectation Maximization
rsq	r-squared
sc-SHC	Single-cell Significance of Hierarchical Clustering
scIBD	Single-cell meta-analysis of Inflammatory Bowel Disease
scITD	Single-cell Interpretable Tensor Decomposition
scRNA-seq	Single-cell RNA-sequencing
SIRE	Self-Identified Race and Ethnicity
SNP	Single Nucleotide Polymorphism
Spearman Cor.	Spearman Correlation
STAR	Spliced Transcipts Alignment to a Reference
TA	Transit Amplifying
Tfh	T Follicular Helper Cells
TH1	Type 1 Helper T Cell
Th17	Type 17 Helper T Cell
UC	Ulcerative Colitis
UMAP	Uniform Manifold Approximation and Projection

UMI Unique Molecular Identifier

SUMMARY

Crohn's disease (CD), one of two inflammatory bowel diseases (IBD), is characterized by chronic, transmural inflammation throughout the intestines. While the cellular and molecular causes of CD remain uncertain, it is thought that an interaction between environmental factors, genetic susceptibility, and gut microbiota may contribute to the development of this disease [1]. Previous genetic studies have identified over 200 loci associated with IBD and have provided insight into disease mechanisms [2, 3]; however, varying disease severity is observed across donors, suggesting alternative mechanisms may promote disease progression. Moreover, therapeutic response may range from sustained efficacy to no response or eventual resistance, highlighting another aspect of patient variability. To investigate patient heterogeneity, "omics" such as transcriptomics and proteomics, offer an opportunity to explore how the role of genetic variation in CD may manifest across diverse environmental conditions and disease courses.

The main questions driving this thesis concern how omic profiles evolve from disease onset to progressive disease in individuals with CD, and how these dynamic profiles may guide personalized medicine approaches. Current treatment strategies for CD aim to prolong remission and induce mucosal healing, yet they remain limited because they do not account for the underlying mechanisms driving disease progression throughout the intestines in individuals. I address this knowledge gap by leveraging transcriptomics and proteomics comprised of intestinal biopsies or serum sampled across different CD cohorts.

Chapter 2 describes a single cell RNA-sequencing cohort containing rectal biopsies of individuals diagnosed with perianal fistulizing CD, a severe complication of CD. This

study identified cell composition differences between inflamed and non-inflamed rectal biopsy samples. Furthermore, goblet cells were found to be associated with inflamed tissue and showed enrichment of inflammatory pathways. This study implicated goblet cells as potentially having a role in sustained inflammation in individuals with perianal fistulizing CD.

Chapter 3 focuses on single cell RNA-sequencing performed on biopsies from the ileum, colon, and rectum of individuals with CD at inception. The first part of this chapter introduces a single cell clustering stability assessment to improve reproducibility and robustness of clustering results. The second part of this chapter is an investigation of heterogeneity across individuals recently diagnosed with CD. Substantial heterogeneity was observed across and within each tissue based on cellular proportion profiles. Additional analyses identified groups of donors within each tissue associated with pathogenic disease mechanisms. Lastly, integrative analysis between genome wide association study (GWAS) summary statistics and single cell data identified T cells and monocytes as potentially associated with CD.

Chapter 4 investigates alternative splicing mechanisms associated with post-operative recurring CD using RNA-sequencing data from ileal biopsies. This study revealed that rectal-like splicing signatures in ileum as well as dysregulation of HP1γ may be drivers of recurring disease. This study also implicated signatures of p53 signaling in recurring disease, historically discussed in the context of colorectal cancer.

Chapter 5 extended the analysis from chapter 4 by jointly assessing serum Olink proteomics and ileal biopsy transcriptomics from the post-operative CD cohort. This study

confirmed sex-bias at the proteomic level. Comparative correlation analysis between serum proteomics and ileal transcriptomics suggested observed correlation patterns were greater than random expectation. Hierarchical clustering analysis identified correlated groups of genes and proteins that were associated with clinical factors, underscoring the utility of multi-omic data for patient stratification.

Collectively, my findings underscore the importance of transitioning from "one-size-fits-all" disease management style to more a personalized medicine approach to manage disease based in part on consideration of the cellular and molecular profile of the patient's gastrointestinal tract.

The research was performed in collaboration with clinical research teams who provided samples, designed the studies, and provided insight into the biology of CD. Specifically, Chapters 2 and 3 were performed with the groups of Dr. Subra Kugathasan (Emory University) and Dr. Peng Qiu (Georgia Tech), Chapter 4 with Dr. Kyle Gettler in the group of Dr. Judy Cho (Mt Sinai, New York), and Chapter 5 with Dr. Mark Lazarev (Johns Hopkins University) and Dr. John Rioux (McGill University). All are leaders of the NIDDK IBD Genetics Consortium.

CHAPTER 1. INTRODUCTION

Scientific advancements in genomic sequencing have propelled research across biomedicine. Continuously evolving technologies and methodologies beg the question of how this "big data" can be utilized to improve human health and investigate disease biology. A revolutionary idea derived from the implementation of genomic data in pharmaceuticals is the concept of personalized or precision medicine, a term coined by Robert Langreth and Michael Waldholz in 1999 [4]. Over the past 25 years, the definition of personalized medicine has gradually transformed, but the general idea is to use one's molecular profile to guide therapeutic decisions and disease management strategies [5]. National initiatives to promote personalized or precision medicine are underway through creation of large biobanks in the United States, underscoring the importance of this concept to transform healthcare [6]. Furthermore, as algorithms like machine learning (ML) and artificial intelligence (AI) become more mainstream in society, understanding how this technology will continue to transform clinical care is paramount. To achieve this, "omic" profiling may be adopted to enable the identification of distinct disease mechanisms, which may hold significant potential for advancing personalized therapeutic strategies.

This thesis investigates an inception cohort comprised of Crohn's disease (CD) patients from whom intestinal transcriptomic data has been obtained. I aim to use it to generate insights concerning the molecular mechanisms acting from disease onset to steady-state disease. Additionally, intestinal transcriptomics and plasma proteomic signatures will be used to characterize progressive CD, illuminating distinct disease mechanisms and potential markers of disease biology. Through the lens of personalized

medicine, the goal of this thesis is to leverage transcriptomic and proteomic data to understand variable disease mechanisms across individuals with CD.

1.1 Crohn's disease

CD, one of two inflammatory bowel diseases (IBD), is characterized by chronic inflammation throughout the gastrointestinal tract, commonly affecting the ileum and colon [7]. Symptoms of CD include abdominal pain, fever, diarrhea, and bowel obstructions with disease onset typically between 20-40 years of age [1, 8]. Historically, industrialized countries and urban areas have higher incidence and prevalence rates of CD; however, the influence of Western-like lifestyle is postulated to have led to a steady increase of incidence and prevalence rates in newly industrialized countries in Asia, Africa, and South America [9]. As CD demographics shift, the global burden of disease increases, posing challenges for healthcare, prevention, and diagnosis [10].

1.1.1 Phenotypic classification of Crohn's disease subtypes

CD is a heterogenous disease exhibiting variable severity and location. The Montreal classification system is used to phenotypically describe disease behavior, location, and age of diagnosis (Figure 1) [11]. Most individuals present initially as non-stricturing/non-penetrating (B1), which is the canonical inflammatory state. Stricturing disease (B2) is a complication of CD characterized by narrowing of the intestinal walls due to relapsing and remitting inflammation [12]. Penetrating disease (B3) is a severe complication of CD attributed to chronic transmural inflammation that can lead to the development of fistulas which are abnormal connections between two different parts of the body [13-15]. A perianal disease modifier is used when there is involvement of the perianal

region due to distinct aggressive disease behavior [11, 16]. Individuals are also screened for extraintestinal manifestations of disease [8, 9]. These tools are implemented to ensure patients receive proper treatment and to improve disease prognosis.

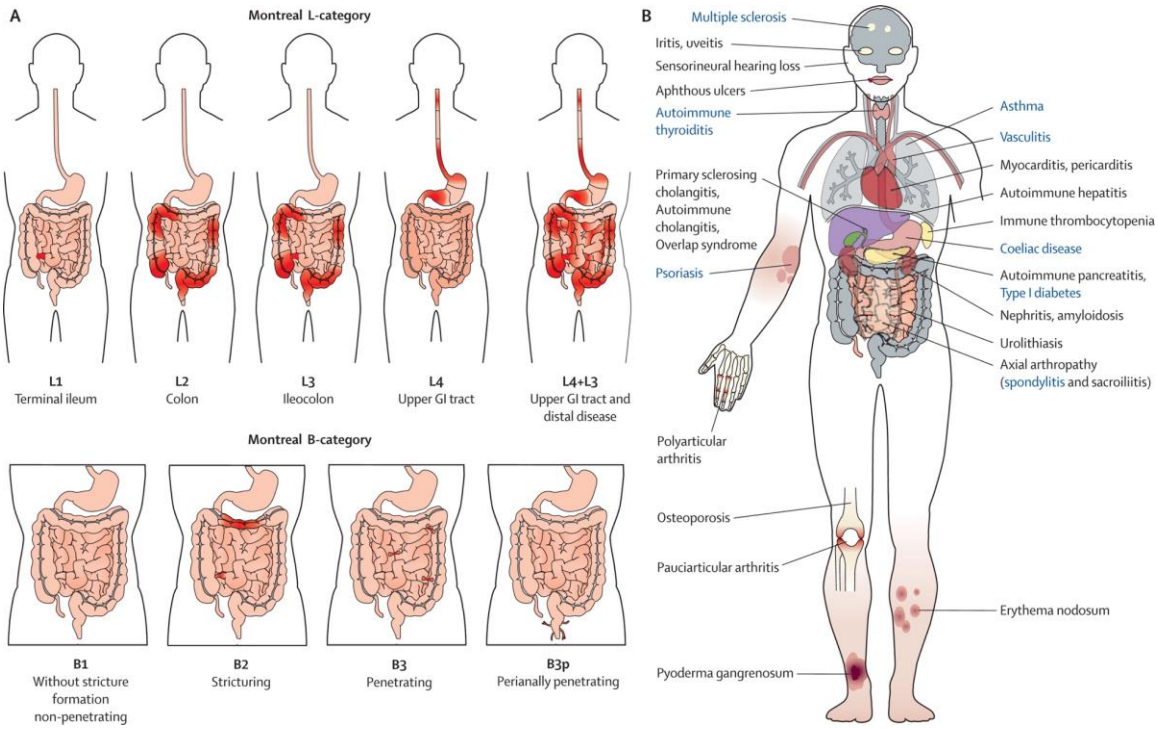


Figure 1 Montreal classification of Crohn's disease. (A) Classification which considers disease location (top) and behavior (bottom). (B) Extraintestinal manifestations of Crohn's disease. From Baumgart and Sandborn (2012) [8].

CD can affect different sections of the intestines including the ileum, part of the small intestine, and the colon, also known as the large intestine. Around 40% of individuals have ileocolonic disease whereas about 30% have ileal- or colonic-dominant disease [17]. Location-specific differences in disease manifestation have been described where those with ileal-dominant disease typically experience more CD-related disease complications like disease progression from B1 to B2 or B3 disease. On the other hand, those with colonic-dominant disease tend to have increased risk for developing extraintestinal

manifestations [17]. Furthermore, therapeutic efficacy is reduced in individuals with ileal-dominant disease compared to colonic-dominant disease. Currently, medical management of CD does not consider tissue-specific differences with respect to therapeutic efficacy [17, 18]. These clinical shortcomings emphasize the need to understand disease mechanisms across tissue to understand therapeutic response and alternative modes of disease.

Earlier age of disease onset has been previously linked to more aggressive and severe course of disease. Many children (< 18 years of age) experience extensive intestinal involvement at initial presentation of disease in addition to potential growth defects from nutrient malabsorption [19]; however, the etiology is thought to be similar among children and adults [20]. A comprehensive analysis of pediatric CD at initial diagnosis across the intestines is essential to elucidate mechanisms that may promote disease behavior and location-specific differences over time.

1.1.2 Aberrant immune response in Crohn's disease

Although CD is considered an idiopathic disease, it is believed that complex interactions among environmental factors, genetic susceptibility, and gut microbiota contribute to its development by triggering abnormal immune response and impairing epithelial barrier integrity leading to inflammation [1]. While disease pathophysiology is heterogeneous, one predominant mechanism of disease is attributed to a breakdown in epithelial barrier function, leading to an increase in microbial invasion to the lamina propria, inducing recruitment of immune cells, and consequently production of pro-inflammatory cytokines, further perpetuating inflammation in the intestines [9].

Interaction between the innate and adaptive arms of the immune system promotes inflammation within the intestines. Neutrophils, part of the innate immune system, may promote inflammation by secreting pro-inflammatory cytokines and other inflammatory molecules like α defensins, which recruit additional neutrophils, monocytes, macrophages, and T cells to the site of inflammation [21]. Furthermore, macrophages and monocytes are also involved in pathology by secreting pro-inflammatory cytokines promoting Type 1 helper T cell (T_H1) and Type 17 helper T cell (T_H17) differentiation and recruitment of monocytes and neutrophils to the site of disease [22, 23]. Previous studies have reported both increased and decreased macrophage infiltration in association with CD development, suggesting alternative modes of disease mechanisms across individuals [23]. T cells, part of the adaptive immune system, are thought to have a central role in disease due to persistent activation. T_H1 and T_H17 cell immune response are hallmark signatures of CD and further promote disease by secreting pro-inflammatory cytokines and recruitment of other immune cells. It is also thought that effector $CD4^+$ T cells in IBD patients are resistant or less responsive to regulatory T cells [9]. The disruption of the careful balance in the immune system can have devastating outcomes for an individual's health and quality of life.

1.1.3 Treatment approaches for Crohn's disease

The primary treatment focus for managing CD is to induce mucosal healing, which is known to lead to improved outcomes. Various therapies including corticosteroids, immune modulators, and biologics are administered to suppress inflammation [24]. The use of biologics such as anti-tumor necrosis factor (anti-TNF) has drastically altered disease course by inducing and maintaining clinical remission in individuals and

decreasing risk for surgery. While the exact anti-TNF mechanism of action remains unknown, it is thought that anti-TNF neutralizes TNF-α produced primarily by macrophages and T cells [25, 26]. While the use of anti-TNF has improved patient outcomes with severe complications, about 30-50% of patients are primary non-responders, and about 50% of initial responders eventually become refractory to treatment (secondary non-responders) [27]. Currently, there is little agreement as to why some individuals are not responsive and/or lose clinical benefits of anti-TNF therapy [28, 29]. To this end, alternative therapies such as anti-integrins, and anti-IL-12/23 have been evaluated to treat and manage CD [24]. Leveraging both clinical and genomic information may improve prediction of therapeutic response, ultimately leading to more cost-effective solutions for patients in addition to improved quality of life.

The aggressive "top-down" approach of administering biologic therapies first has been adopted by many clinicians, however, many individuals still require surgical intervention at some stage throughout their disease course. Risk factors for necessitating surgery include refractory disease, bowel perforation, bowel obstruction, complications from strictures, and intra-abdominal abscess [30, 31]. Different surgical interventions such as bowel resection, stricturoplasty, and abscess drainage are available to treat but not cure the affected portion of the intestine [30]. Timing surgical intervention remains paramount to reduce the risk of surgery-related complications [32]. An interdisciplinary approach, perhaps coupling therapeutic, surgical mediation and lifestyle modification, should be considered based on disease severity and behavior [32]. Understanding the mechanisms differentiating remission from treatment resistance may be key to optimizing intervention strategies.

1.1.4 Managing Crohn's disease in the post-operative setting

Although some individuals achieve remission after therapy and surgery, a subset of individuals continue to experience post-operative (post-op) disease complications. Risk factors for developing post-op complications include age, smoking, penetrating disease, previous CD-related intestinal resections, and duration of disease [30]. Post-op clinical recurrence, defined by the Crohn's disease activity index (CDAI), occurs in approximately 20-40% of patients after 12 months of surgery, and 35-50% of patients after 5 years [30]. The Rutgeerts score is largely based on the distribution and extent of aphthous (ulcerous) lesions and is used to evaluate endoscopic recurrence which precedes clinical and surgical recurrence (Table 1) [33]. Those with Rutgeerts score > i2b are classified as having recurring disease [30, 33]. Endoscopy is the primary method for disease surveillance, which has some limitations due to risk, cost, and patient inconvenience [33]. Fecal calprotectin is an alternative non-invasive method for assessing disease surveillance but lacks diagnostic accuracy [33, 34]. Based on disease activity findings, disease management programs may be re-evaluated by adjusting the intensity of therapy. Anti-TNF has been shown to prevent post-op disease recurrence in CD; however, the effectiveness of other biologics needs to be studied [24, 35]. There remains a significant need for less invasive metrics to reliably monitor post-op disease activity and treatment decisions.

Table 1 Rutgeerts score to guide classification of recurring disease. Rutgeerts score > i2b is considered recurring disease. From Shah and Click (2021) [33].

Score	Endoscopic Findings
i0	No lesions in distal ileum
i1	<5 aphthous lesions
i2	>5 aphthous lesions with normal mucosa between the lesions, skip areas of large lesions
i2a	Lesions confined to the ileocolonic anastomosis
i2b	Lesions in the neoterminal ileum with normal intervening mucosa (with or without anastomotic lesions)
i3	Diffuse aphthous ileitis with diffusely inflamed mucosa
i4	Diffuse inflation with larger ulcers, nodules, and/or narrowing

1.2 Applications of genomic technologies in Crohn's disease research

Since CD was first described in the 1930's, many studies attempting to understand the multifactorial role of disease pathology have emerged. Some of the first twin studies provided compelling evidence for a genetic component to CD, and early linkage studies revealed the role of the *NOD2* locus in CD risk [36-38]. The Human Genome Project completed in the early 2000's paved the way for genetic studies across human populations thereby enabling contemporary methods to understand the genetic architecture of disease. Genome wide association studies (GWAS) are conducted to identify genetic variants associated with disease or complex traits. Over 200 loci associated with IBD, and 140 loci associated with CD have been identified, primarily in European populations, providing insight to mechanisms of disease through the role of pro-inflammatory mechanisms, cellular stress, and shifts in cellular metabolism [1-3, 36]. Ongoing efforts to incorporate

diverse genetic ancestry in studies is underway to map the genetic diversity that contributes to CD risk [2]. Genetic discoveries offer valuable insight into disease pathophysiology and hold promise for future therapeutic interventions.

In the post-GWAS era, successive efforts wielding additional "omics" are used to disentangle the complexity of CD pathogenesis. While GWAS has identified variants associated with CD risk, the estimated heritability is approximately 65-75% from twin studies and the SNP (single nucleotide polymorphism) heritability is approximately 40% suggesting environmental factors play a role in disease risk and progression [1, 39, 40]. Heritability simply estimates the magnitude of the genetic contribution, while other dynamic "omics" including transcriptomics and proteomics, can be leveraged to generate a deeper understanding of the role of genetic variation in CD across different environmental conditions.

1.2.1 Bulk RNA-sequencing in Crohn's disease research

The central dogma of biology states that genetic information flows from DNA, is transcribed to RNA and translated to protein. Messenger RNA (mRNA) is an RNA subtype that has been extensively studied since these RNAs represent an intermediary between DNA and protein [41]. RNA-sequencing (RNA-seq) measures the average gene expression in a sample, enabling transcriptomic profiling across a range of conditions [42]. RNA-seq analysis is performed in three main steps: library preparation, sequencing, and data analysis. The mRNA is extracted from a tissue of interest, sequenced, then aligned to the relevant reference genome. Quality control (QC) metrics like GC content and length biases are evaluated and corrected through normalization practices [43]. After pre-processing the

data, downstream analyses such as differential gene expression and pathway analysis can be performed to gain insight to the biological question of interest.

Pre-mRNA from eukaryotic cells can undergo alternative splicing (AS), giving rise to diverse transcripts, and subsequently, proteins. The spliceosome is the machinery that regulates the combination of exons that eventually produce mature mRNA [44]. AS is tissue specific and has been implicated in numerous diseases including neurological diseases, diabetes mellitus, and cancer [44]. Previous studies have also linked dysregulation of splicing mechanisms to IBD [45]. To facilitate analysis, computational tools are available to estimate transcript isoforms from RNA-seq data, allowing for insight into AS mechanisms. RNA-seq by Expectation Maximization (RSEM) is a tool available to map genes and isoforms to a reference and quantify gene and isoform abundance [46]. From there, downstream analysis like differential transcript usage can be performed to understand the role of AS across conditions of interest. The extent to which AS impacts CD progression and recurrence, and whether alternatively spliced products could be targeted by novel therapeutics remains to be evaluated.

1.2.2 Crohn's disease insights from bulk transcriptomics

Transcriptomic analyses have been pivotal in advancing the understanding of mechanisms of disease initiation and progression. Overall, many studies have identified dysregulation of immune response, oxidative stress, cell adhesion, response to stimuli, and cell migration pathways as enriched in CD [47-51]. For example, RNA-seq analysis of a pediatric CD cohort identified gene modules associated with neutrophil chemotaxis and migration. The authors argued that dysfunctional neutrophils contribute to pathogenesis

rather than merely reflecting aberrant enrichment of pro-inflammatory pathways [49]. Additionally, Ashton et al. identified a gene module enriched for oncostatin-M and NOD signaling pathways, in parallel to enrichment of IL17 signaling pathways after differential gene expression and pathway analysis in treatment naïve pediatric CD patients. These pathways have been previously described as key factors in CD pathogenesis [52]. Other studies have leveraged gene expression signatures to distinguish subtypes of CD [48].

Given clinical variability across CD behaviors, RNA-seq of ileum biopsies was performed to investigate the molecular underpinnings of disease complication in a pediatric inception cohort. This prospectively recruited cohort, referred to as the RISK cohort, followed 1,800 individual's disease course over time [53]. While ileum gene expression was heterogenous, direct comparison of those who developed B2 vs. B3 complications was performed revealing distinct expression patterns. B2 disease complications were associated with up-regulation of extracellular matrix (ECM) accumulation pathways while B3 disease complications were associated with pro-inflammatory pathways and improved outcomes with early anti-TNFα use [53]. Alternatively, pro-fibrotic disease mechanisms were enriched in donors classified as B3 as well as B3 in combination with B2 disease in an adult cohort [54]. Collectively, these findings demonstrate that imbalance between ECM and pro-inflammatory pathways likely drives progression to B2 or B3 disease; however, RNA-seq approaches lack the resolution to delineate the cell type specific contributions to these distinct disease behaviors.

RNA-seq analysis can also be performed to uncover gene signatures that associate immune response with the microbiome. A separate study utilizing the RISK cohort also identified a gene expression signature associated with increased *DUOX2* gene expression

and decreased *APOA1* gene expression [55]. The *DUOX2* signature implicated *LCT* and *MUC4* gene expression with expansion of Proteobacteria taxa whereas *APOA1* signature was associated with increased gene expression of *IFNG* and *CXCL9,* corresponding to Th1 polarization, and associated with depletion of Firmicutes and Bacteroidetes taxa [55]. RNA-seq has also been utilized to identify markers of disease. For example, Hong et al. performed RNA-seq analysis which identified *CXCL1* as up-regulated in CD compared to controls, confirming expression with real-time quantitative reverse transcription polymerase chain reaction and enzyme linked immunosorbent assay (ELISA), potentially serving as a biomarker for disease [48]. These signatures reveal disruption of both epithelial and immune cell homeostasis in CD, implying alternative therapeutic strategies that consider these signatures, alongside microbiome composition, may improve patient outcomes.

Considering the environmental and microbial role in CD pathogenesis, some studies have also used RNA-seq to understand how diet may affect the mucosal immune system. Specifically, Wu et al. investigated lactylation-related genes in CD since increased lactic acid levels have been associated with production of pro-inflammatory cytokines from T cells. Hubs of lactylation genes were able to stratify CD individuals from controls, suggesting these genes may be useful biomarkers of disease [56]. Braun et al. conducted a study comparing urban and rural environmental exposures in newly diagnosed CD patients and healthy controls from China and Israel to elucidate microbiome, diet, and ileal transcriptional CD signatures across environments [57]. Diet-linked metabolites from stool were associated with ileal epithelial cell metabolic signatures whereas microbe-linked metabolites were associated with immune cell function [57]. Collectively, these studies

represent initial attempts to parse exposures connected to the mucosal immune system. The transcriptomic and stool-specific metabolomic signatures may not be fully representative of dynamic changes in transcripts or metabolites in response to environment across different sections of the intestines. Additional efforts to understand the environmental contribution to more progressive forms of CD across the intestines is warranted. Key challenges related to IBD genetics and genomics research are outlined in Gibson et al. [58]. Future multi-omic longitudinal studies incorporating diverse populations with comprehensive clinical phenotyping will be important to dissect the genetic and environmental components that influence disease pathogenesis.

1.2.3 Alternative splicing in Crohn's disease

While the studies previously described identified transcriptomic signatures associated with CD, offering insight into disease heterogeneity, transcriptomic regulation of isoform and AS events were not considered. Initial studies were performed to establish that AS within the intestines is involved in CD pathology through RNA-seq and microarray technologies [59-62]. These studies demonstrated that AS in the intestines may perpetuate disease through up-regulation of pro-inflammatory mechanisms and increased intestinal permeability [61, 62]. Investigation of AS in IBD identified CD-specific splicing factors that could stratify CD vs. ulcerative colitis (UC); however, the extent to which ileum or rectal tissue yielded different splicing patterns was not described [60]. The potential utility of targeting isoforms for therapeutic intervention for IBD has been investigated as well [63]. Despite their contributions, these studies lack a comprehensive view of AS mis-regulation or tissue- and phenotypic-specific variations in disease.

A strong candidate for "global" AS dysregulation in IBD has emerged as demonstrated by Mata-Garrido et al. This study investigated the effects of HP1γ on splicing mechanisms in ulcerative colitis (UC) [64]. HP1γ regulates formation and maintenance of heterochromatin and gene silencing in addition to pre-mRNA processing and AS [65]. In *Cbx3* (encoding HP1γ) knock-out (KO) mice, general deregulation of splicing mechanisms in conjunction with increased inflammatory genes and decreased antimicrobial gene expression was observed in colonic tissue. Alterations in epithelial barrier composition was also observed through an increase in stem cell niche and uncontrolled proliferation. Progerin is an isoform of *LMNA*, involved in Hutchinson Gilford Progeria Syndrome, characterized by accelerated aging [66]. This study also observed detection of progerin in *Cbx3* KO mice and UC patients, potentially serving as a marker of UC. Increase in splicing noise was associated with decreased expression of HP1γ in UC, suggesting this protein is protective of disease. Replication of this signature defined by HP1γ in a CD cohort is needed to better understand the role of this protein in chronic inflammation and disease recurrence.

Tissue-related aberrant splicing has also been implicated in IBD pathology. Berger et al. investigated AS in ileum and rectum biopsies from individuals with IBD [67]. Principal component analysis (PCA) revealed tissue as a driver of variation across donors for both gene expression and percent spliced in (PSI) counts. This analysis also revealed eight ileal samples with intermediate expression profiles, more reflective of rectal profiles, which was partly attributed to an increased proportion of epithelial cells. Further, three of these ileum samples clustered with rectal samples, exemplifying rectal-like splicing patterns. The gene expression patterns of these 3 individuals were intermediate between

ileum and rectum. This signature was coined “spliceopathy” defined by increased splicing. Additional validation of “spliceopathy” in other cohorts and consideration of clinical factors like inflammation, disease behavior, and progression is needed.

Based on previous studies, I hypothesize that the combination of decreased expression of HP1γ and “spliceopathy” signatures may contribute to underlying CD mechanisms that promote more progressive and recurring disease.

1.3 Single cell RNA-sequencing and Crohn’s disease research

After the initial single cell transcriptomic study performed by Tang et al. in 2009, profound advancements in high-throughput technology and microfluidics have enabled sequencing of millions of cells [68]. Single cell RNA-sequencing (scRNA-seq) technology has shed light onto cellular dynamics across a range of biological questions in developmental biology, neurobiology, cancer, immunology, and even microbiology [69]. Importantly, scRNA-seq can be leveraged in biomedical research to understand clinical aspects of disease [70]. The Human Cell Atlas, formed in response to the growing field, aims to map all cell types across the human body, serving as a reference for future studies [71]. Here, we use scRNA-seq to understand molecular mechanisms of CD in initial disease onset and more progressive disease.

1.3.1 Complexities of single cell RNA-sequencing analysis

The process of generating scRNA-seq data is similar to RNA-seq data; however, one of the main differences between single cell sequencing and bulk sequencing is the isolation of single cells from a tissue of interest (Figure 2). Various methods are available

for cell isolation including microfluidics, microwell and droplet-based techniques [69, 70, 72]. After cell isolation, library preparation, sequencing and downstream analysis can be performed to capture mRNAs from each cell in a sample. Through this approach, dynamic expression patterns across cells can be evaluated to investigate the heterogeneity of underlying biological mechanisms.

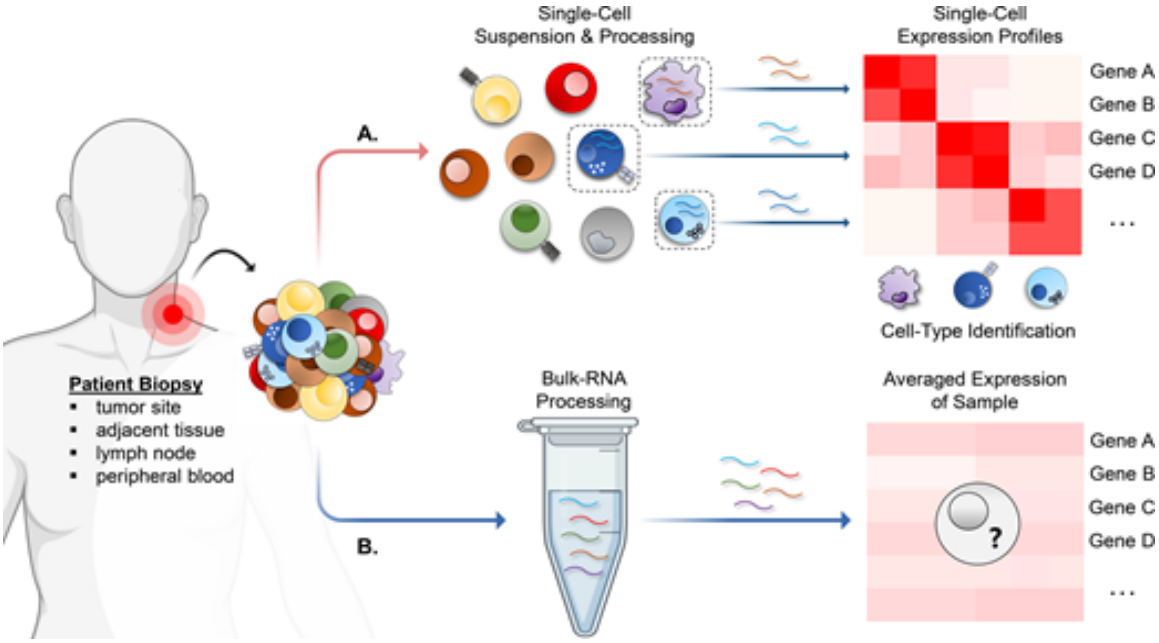


Figure 2 Key differences between scRNA-seq (top) and RNA-seq (bottom). From Guruprasad et al. (2020) [73].

The general workflow for single cell analysis can be summarized into four main steps: initial pre-processing of sequenced data, QC, clustering, and downstream analysis. In initial pre-processing, the transcripts captured per cell in a sample are aligned to the relevant reference transcriptome, generating the count matrices used for analysis [72, 74, 75]. After count matrices are generated, QC is performed to remove damaged, dying, and doublet cells. The next step involves dimensionality reduction and feature selection to reduce dimensionality of the dataset and facilitate clustering analysis. The data is then

clustered based on similar transcriptional profiles, identifying cell populations in tissue [72, 74, 76]. After the dataset is clustered, and the clusters are assigned to cellular identities, downstream analyses like differential gene expression and cell-cell communication analysis are performed. Different tools are available such as Seurat, CellChat, and Monocle for clustering and downstream analysis [77-79]. While general frameworks with best practices have been outlined in various studies, heterogeneity across technical and biological factors makes it challenging to establish compressive guidelines for selecting optimal thresholds and clustering parameters in analysis.

One of the primary goals in scRNA-seq analysis is to identify meaningful cell populations in a tissue. Usually, unsupervised clustering algorithms like k-means, Louvain, or Leiden, are used to identify these cell populations [80, 81]. While these clustering approaches are data-driven and unbiased, many of these algorithms require user-defined tuning parameters which may result in different conclusions with respect to cell type or cell state annotation across and within datasets [80]. Even clustering a dataset with the same parameters may yield slightly different results from uninteresting random variation in the clustering algorithm [82, 83]. Due to this challenge, development of a workflow that evaluates reproducibility and rigor of clustering results in scRNA-seq data is needed since many downstream analyses rely on cluster assignment.

To address the clustering-related challenges, different studies have developed methods or workflows to identify optimal clustering parameters and evaluate clustering stability. One tool proposed to identify the optimal number of clusters is single-cell significance of hierarchical clustering (sc-SHC) [83]. This tool introduces hypothesis testing in a hierarchical clustering framework motivated by identifying clusters not due to

chance. To evaluate clustering stability, Tasic et al. employed a cross-validation clustering approach using PCA and weighted gene co-expression network analysis in conjunction with Ward's method to identify "core" cells, representing cells classified into the same cluster, and "intermediate" cells that most likely represent transitional cells [84]. Tang et al. also used a sub-sampling clustering approach defining stable clusters with the Jaccard Index [85]. While these consider statistical instability in clustering, evaluation of the impact of only retaining the cells that repeatedly go to the same cluster is warranted. I hypothesize that retaining stably assigned cells, which are cells that repeatedly go to the same cluster, will yield more reproducible and robust clustering results.

1.3.2 From cell types to signatures: single cell RNA-sequencing in IBD research

scRNA-seq has emerged as a powerful tool for unraveling cellular and molecular heterogeneity in the intestinal mucosal immune system while offering new insights into intestinal disease pathology. Many landmark publications have established functions of diverse cell types in the intestines over space and time, including functional plasticity during inflammatory events in individuals with CD [86, 87]. For example, $BEST4^{+}$ epithelial cells, a subtype of absorptive cells, are involved in pH sensing and were initially identified in the intestines using scRNA-seq technology [86-89]. The altered abundance of $BEST4^{+}$ epithelial cells observed between inflamed UC tissue and healthy controls suggests potential involvement in disease pathogenesis [89]. Goblet cells, specialized secretory cells producing mucin, were enriched in CD compared to controls, while secretion of WFDC2 by goblet cells emerged as a potential diagnostic marker of UC [88, 90]. Two alternative hypotheses for breakdown of epithelial barrier are: 1) penetration of pathogenic microbes to damaged epithelial barrier stimulate immune response or 2)

microbial interaction with epithelial barrier which induces hyperactivation of inflammatory mediators [91]. Oliver et al. created a gastrointestinal tissue scRNA-seq atlas comprised of healthy and diseased samples. This study identified a metaplastic cell population termed INFLAREs which was associated with inflamed tissue from donors with IBD [92]. The authors hypothesize this metaplastic population is generated to repair the intestinal epithelial barrier in response to injury, providing evidence for the first hypothesis; however, it is not clear if this cell population was associated with severe outcomes of disease (B2 or B3) or the initial inflammatory disease state (B1). Leveraging scRNA-seq at initial disease onset may provide greater understanding of mechanisms mediating epithelial barrier defects and how this contributes to disease progression or mucosal healing.

scRNA-seq approaches have revealed complex functional cellular states and subpopulations in CD that extend beyond traditional classification systems. Garrido-Trigo et al. observed substantial macrophage heterogeneity within healthy, UC, and colonic CD individuals, in addition to interaction of macrophages with inflammatory fibroblasts in UC and colonic CD patients, suggesting this may partly explain patient heterogeneity [93]. Decrease in intraepithelial $CD8^+$ T cells was accompanied by an increase in $CD4^+$ T cells in inflamed tissue whereas the opposite trend was observed in the lamina propria within the ileum [94]. Further, subsets of T_H17 cells exhibited a quiescent or effector phenotype, demonstrating T cell plasticity in CD [94]. Interestingly, concordance across CD and control donor profiles within B and T cells was observed in peripheral blood mononuclear cells suggesting shared activity across immune cells and in response to disease [95]. Previous studies have also exploited scRNA-seq to identify cells or groups of cells that are

associated with different aspects of IBD. The GIMATS (IgG plasma cells, inflammatory mononuclear phagocytes, and activated T and stromal cells) module was identified in a subset of inflamed surgically resected ileum tissue by Martin et al. and was associated with resistance to anti-TNF therapy [96]. Inflammatory fibroblasts were also associated with resistance to anti-TNF therapy in UC [89]. While these studies established intestinal cellular function and alterations in disease, further investigation of these signatures across tissue and disease status is needed.

Regional variation across the small and large intestine in CD has been clinically documented, prompting scRNA-seq studies to investigate tissue-specific pathology in CD. Kong et al. observed marked differences in underlying mechanisms promoting inflammation in the ileum vs. colon. Specifically, immune cell enrichment was seen in the ileum whereas transcriptional changes were enriched in colonic tissue [97]. These findings underscore integration of disease location as a critical variable when investigating CD pathology. To further understand the molecular landscape in IBD over time, a longitudinal cohort of individuals diagnosed with CD or UC was prospectively recruited to understand disparate anti-TNF response. This study described cellular compositional changes over time as well as shared pathways, including an IFN-mediated signaling pathways, that were enriched in epithelial cells of remission donors (individuals who respond to treatment) and monocytes of refractory donors (individuals who do not response to treatment) for both CD and UC [98]. While this cohort provides valuable insights into personalized disease signatures, a key limitation is that CD-related remission and refractory signatures were neither deconvolved across different intestinal compartments nor analyzed in the context

of alternative disease behaviors, potentially obscuring import tissue-specific or behavior-specific factors that influence therapeutic response.

The transition from inflammatory state to more progressive phenotypes of CD are designated by alterations in cell populations and function which can be detected using scRNA-seq. Previous studies have reported an increase in prevalence and severe complications of individuals with African ancestry, motivating studies to investigate the differences across population groups [99, 100]. Our group used bulk RNA-seq to demonstrate that the same pathways that are associated with poor outcomes tend to be upregulated in African Americans at inception [101]. Levantovsky et al. leveraged scRNA-seq and functional genomics to attempt to unravel mechanisms of perianal-fistulizing CD in a cohort of patients with African American ancestry and European American ancestry, identifying cellular differences, both morphological and functional, across populations [102]. A recent study comparing perianal fistulizing CD to idiopathic perianal fistulas observed up-regulation of interferon gamma (IFN-γ) and tumor necrosis factor-alpha (TNF-α) mediated by T_H17 and myeloid cells in perianal fistulizing CD, suggesting therapies targeting IFN-γ may resolve this disease complication [103].

scRNA-seq has been leveraged to understand the cellular contribution of stricturing and fibrotic complications of CD as well. Populations of fibroblasts, activated through interactions with macrophages, have been associated with promoting fibrosis through production of ECM [104-106]. Enrichment of B cells and plasma cells has been associated with stricturing disease emphasizing mucosal reorganization in the fibrotic intestine [105]. CDH11 was nominated as a potential therapeutic target to mediate fibrosis, discovered

from findings in scRNA-seq data [106]. Creeping fat is another CD-related complication associated with stricturing disease and muscle cell hyperplasia, formed in response to microbial invasion and inflammation [107, 108]. scRNA-seq was performed on creeping fat tissue, finding increased expression of PPARγ in a sub-cluster of vascular endothelial cells, suggesting that this may have a crucial role in creeping fat formation [109]. Alternatively, Wu et al. found that L-kynurenine produced by macrophages in response to commensal gut bacteria promoted creeping fat formation [110]. While these studies have identified important cellular mechanisms associated with disease complications, a critical limitation remains in determining whether these signatures represent causative factors driving disease progression or merely adaptive responses to underlying pathological processes.

Atlas level scRNA-seq studies will continue to emerge as this field grows, transforming the understanding of disease by creating cellular taxonomies across patient cohorts, tissues, and disease states. As a result, platforms like scIBD have emerged to facilitate meta-analysis of IBD across different studies [111]. These discoveries may empower development of targeted therapeutic intervention for hard-to-treat CD conditions; nevertheless, careful consideration of study design requires meticulous deliberation of technical artifacts, patient heterogeneity, and inter-study variability in analytical frameworks for integration.

1.4 Proteomics in Crohn's disease research

Although genomics and transcriptomic analyses provide proximal measurements of activity occurring within cells, these approaches may not accurately reflect protein

abundance and function, which ultimately determine phenotypic outcomes across conditions. Proteomics, a rapidly growing field, has evolved from conventional methods like Western blot and ELISA to high-throughput methods like mass spectrometry (MS), to measure protein expression [112, 113]. Proximity extension assay (PEA), the Olink proteomics protocol, quantifies protein expression in serum or plasma samples [114]. Antibodies with oligonucleotides (tags) bind to the target protein and hybridize in proximity which is then followed by extension with DNA polymerase. Amplification with polymerase chain reaction (PCR) is performed on the hybridized tags, called DNA barcodes. Lastly, real-time PCR is performed to quantify protein expression [114]. An advantage of using the PEA technology is the high specificity and sensitivity of protein measurement, even allowing for quantification of low-abundance proteins [115]. As proteomic technologies and analytical strategies advance and become integrated with multi-omic approaches, identification of therapeutic targets and disease markers will improve while bridging the gap between discovery and clinical application.

1.4.1 Multi-omic integration: bridging genetics, transcriptomics, and proteomics

Efforts to map the genetic architecture of plasma and serum proteomics have aimed to understand how inter-individual variation may affect protein structure and function. Previous studies have identified variants associated with plasma protein abundance, termed protein quantitative trait loci (pQTL) [116, 117]. The pQTLs also associated with disease variants from GWAS may be targetable for therapeutic action. Eldjarn et al. compared pQTLs identified from Olink and SomaScan proteomics in the UK biobank and a cohort comprised of Icelandic people [118]. This study found shared pQTLs across platforms and cohorts; however, caution in interpretation of results is warranted because these platforms

may have measured different proteoforms, or pQTL associated with antibody-epitope binding, instead of protein expression [118]. Previous studies have also examined the relationship between genetics and protein expression for those diagnosed with IBD. Cis- and trans-pQTLs were associated with a disease variant in IL10RA, previously reported for monogenic IBD [118]. Additionally, Zhang et al. identified causal proteins, including IL12B, IFNG, SEPTIN8, CXCL9, CCN3, and RSPO3, associated with CD through integration of pQTLs and Mendelian randomization analyses, some of which were also associated with druggable targets [119]. Integrating genetics with proteomics offers insight into how modifiable lifestyle and environmental factors may influence disease risk and development.

Transcriptomic profiling has been adopted widely as a proxy for measuring protein expression to investigate biological changes across a range of tissues and conditions despite well-documented discordances between mRNA and protein abundance [120, 121]. Modest correlation between mRNA and protein levels, typically ranging from 0.3-0.6, has been noted across organisms, cell lines, and tissue [122-126]. This pattern is also consistent across different technologies, including MS and immunofluorescence, to measure protein expression [122, 124]. While biological and technical factors account for discordance in mRNA and protein expression, Koussounadis et al. observed higher correlation between differentially expressed genes (DEGs) and protein expression compared to non-DEGs and protein expression, suggesting that DEGs may have more biological meaning to support inferences of disease pathology [124]. Furthermore, stronger mRNA-protein correlations were enriched among known drug targets [125]. Based on these results, I hypothesize that

higher correlations between mRNA and protein levels may enhance identification of reliable biomarkers or therapeutic targets.

1.4.2 Advancing clinical applications of IBD research through proteomic profiling

The ease at which serum or plasma can be obtained has important clinical considerations for disease monitoring and pinpointing of potential therapeutic targets, as well as delineation of the systemic effects of IBD pharmaceuticals. Previous omic studies have primarily focused on findings from intestinal tissue, however, the extent to which signatures from primary disease location is seen in serum remains to be evaluated. I hypothesize that correlated levels of groups of genes and/or proteins within the intestines and serum may be used as potential markers of CD activity. Furthermore, sex specific differences observed in clinical transcriptomic, and proteomic studies [127-131] warrants investigation of sex-specific differences at the proteomic level in post-op disease. Comparative analysis of tissue-specific transcriptomic signatures with circulating proteomic profiles could enable clinicians to monitor disease activity and predict treatment outcomes through minimally invasive blood-based assays.

1.5 Personalized medicine for Crohn's disease

An emerging concept taking precedence in modern health care across various diseases and conditions is the idea of personalized medicine, which can be defined as tailoring treatment to an individual's molecular and clinical profile [132, 133]. While current strategies aim to manage disease relapse and prolong remission, they remain suboptimal because factoring the underlying processes promoting inflammation tends to be neglected [134, 135]. Given the substantial inter-individual heterogeneity in disease

presentation, progression, and therapeutic response among CD patients, the implementation of personalized medicine approaches represents a critical shift toward optimizing therapeutic intervention through individualized care strategies [135]. The expectation is that molecular signatures of an individual, such as genetic, transcriptomic or proteomic information, can be used to effectively guide clinicians to the most promising treatment protocol. The goal is to transition from the current "one-size fits all" framework to a more individualized treatment strategy, reducing healthcare, physical, and psychological burdens on patients.

Implementation of personalized medicine approaches leveraging transcriptomic information has been described. Rojas-Peña et al. employed longitudinal RNA-seq profiling of individuals with complicated malaria. This study revealed marked patient heterogeneity with engagement of distinct arms of the immune system and inflammation, implying different disease mechanisms [136]. Another study using longitudinal RNA-seq profiling to infer patient specific mechanisms was performed by Banchereau et al., focusing on systemic lupus erythematosus. This study correlated disease activity with several groups of donors based on transcriptional profiles, highlighting separate groups of donors characterized by plasmablast, IFN, and neutrophil signatures [137]. Further validation is needed to confirm patient signatures identified in these studies; however, these examples lay the foundation for incorporating transcriptomic signatures for personalized medicine and patient stratification. Broadly, the goal of this thesis is to highlight the importance of personalized medicine across inception and progressive disease.

Chapter 2 of this thesis examines how persistent inflammation and potential inflammatory mechanisms may perpetuate perianal-fistulizing CD. Chapter 3 addresses

two questions: first, the effect of single-cell clustering stability on robustness and repeatability; and second, inter-individual variation in initial CD onset across the intestines. Chapter 4 investigates how AS may mediate recurring CD by testing associations between tissue specific AS signatures and recurring disease. Finally, Chapter 5 examines the correlation between gene and protein expression in recurring CD, and how coordinated expression patterns may inform patient stratification based on clinical characteristics.

CHAPTER 2. PERSISTENT INFLAMMATION OF THE RECTUM IN PERIANAL FISTULIZING CROHN'S DISEASE IS ASSOCIATED WITH GOBLET CELL FUNCTION

2.1 Introduction

Perianal fistulizing Crohn's disease (perianal-CD) is a debilitating form of rectal CD associated with multiple surgical interventions and lifelong morbidity. Bacterial infection and inflammation in the rectum appear to promote the formation of perianal-CD fistulas by increasing matrix metalloproteinase activity and cytokine levels in the mucosa that instigate an epithelial to mesenchymal transition during fistula tract formation [138-140]. Many patients will resolve the inflammation with anti-tumor necrosis factor (TNF) treatment and undergo healing while others do not respond to therapy, making sustained inflammation a driving factor of disease. The remodeling of the mucosal compartments in response to inflammation during rectal disease with fistulation and its resolution in response to therapy has not been analyzed at the single-cell level. Thus, we compared single-cell transcriptomic profiles of healed rectal mucosa of established perianal-CD patients undergoing anti-TNF therapy to those who remain unhealed and inflamed (IF), hypothesizing that changes in the epithelial compartment might affect the persistence of fistulas in individuals with perianal-CD. Our study confirms previous findings that support the role of pro-inflammatory cytokines in perianal-CD and suggests one mechanism within the epithelial compartment that might affect fistula persistence in these individuals.

2.2 Materials and Methods

2.2.1 Study population for perianal fistulizing Crohn's disease patients

Rectal biopsies were obtained from Crohn's disease (CD) patients with persistent perianal fistulae. Rectal mucosal tissue was obtained from patients undergoing repeat colonoscopy who were classified into two groups: healed perianal-CD with non-inflamed rectal mucosa, and active perianal-CD with inflamed rectal mucosa. One individual (IF4) with a history of severe persistent perianal-CD was transiently inactive at time of sampling but nevertheless had an inflamed profile. Written consent was obtained from patients undergoing colonoscopy at the Division of Pediatric Gastroenterology at Children's Healthcare of Atlanta (Georgia, USA). Patient sample demographics are provided in Supplemental Table 1. All 12 subjects had received anti-TNF biologic therapy before retrieval of the rectal mucosa. During colonoscopy, 6 subjects were classified as inflamed with active fistula and 7 as non-inflamed with inactive/healed fistula.

2.2.2 Tissue processing for scRNA-seq

Three rectal biopsies per patient were immediately processed using a cold protease (Sigma Cat #P5430) protocol [141]. Enzyme mix was prepared using Rho kinase (Y-27632, CAS#872543-07-6) and Caspase inhibitors (CAS#187389-52-2) in DPBS with 0.5M EDTA. Biopsies were minced and digested twice (for 15 minutes each) and passed through a 40 µm filter. Single cell suspensions were washed with buffer containing Benzonase (CAS#9025-65-4) and centrifuged at 400 g for 5 minutes. Cells were resuspended in 200-300 µl of RPMI with 10% FBS and counted. Single cell encapsulation was carried out using 10X reagent mix and loaded into the Chromium microfluidic device according to manufacturer's instructions. This was performed using Chromium Next GEM

Single Cell 3' Reagent Kit v 3.1 Dual Index (PN-100268) with a target of 10,000 cells. Barcoded sequencing libraries were made from the amplified cDNA. Sequencing at a read depth of 50,000 reads per cell was performed on the Illumina NovaSeq S4 sequencing platform at the Emory-Yerkes Genome Core.

2.2.3 Gene expression profiling

The FASTQ files were aligned to GRCh38 human reference genome using 10X Genomics Cell Ranger 6.1.2 "cellranger -count" excluding introns, except for IF1 (Cell Ranger 6.0.2 "cellranger -count" include introns), was used to generate the gene expression matrix for each sample [142]. The resulting filtered feature-barcode matrices (n = 13 samples) were merged for downstream analysis.

Filtered gene expression matrices were loaded into Seurat v. 4.1.1 with min.features = 200 and min.cells = 3 [143]. The following QC parameters were used: number of genes 250-6500, number of transcripts >300, mitochondria percent <30% and log10(Genes/UMI) >0.80. Only genes that were observed to be expressed in 10 or more cells were kept for downstream analysis. After QC, 67,119 cells from rectal tissue were kept for downstream analysis.

The dataset was normalized using log normalization and the highly variable features (n=2000) were selected for clustering using Seurat (v. 4.1.1). Next, the dataset was scaled, and PCA was performed on the highly variable genes. The cells were clustered in this low dimensional setting using "FindNeighbors" and "FindClusters". 15 PCs and a resolution of 0.50 were used for clustering analysis. The clustering results were visualized using UMAP.

2.2.4 Robustness and batch effect comparison

Three batch effect correction methods (Harmony, CCA, and rPCA) were applied to test for the robustness of clustering for this dataset [77, 144]. Epithelial, immune, and stromal cells were included in the batch effect correction comparison. The following parameters were kept the same to facilitate method comparison: log normalization, 15 PCs and 0.50 clustering resolution. Additionally, the samples were integrated across sequencing batch.

The batch effect correction methods were assessed based on clustering stability and similarity. To determine the stability of the clustering results, the Adjusted Rand Index (RI) was computed with the R package Dune v. 1.8, using the "merge" function to determine the maximum agreement between clusters (Supplemental Figure 1) [145]. Similarity of the clustering results was ascertained by generating confusion matrix heatmaps. More "splitting" of clusters was indicative of decreased similarity (Supplemental Figure 1). Lastly, the distribution of each sample contributing to each cluster was assessed to ensure that sample-specific or batch-specific clusters were not present.

To ensure that clustering stability was not strongly affected by including the control sample (NI5), clustering stability and similarity was reassessed after removal of this sample, again contrasting batch effect correction with Harmony, CCA, and rPCA (same parameters as previously mentioned). The RI was computed, and clustering similarity was evaluated. It was determined that clustering stability and similarity was not strongly affected by including the control sample, so the control sample was included in clustering analysis, but excluded for differential gene expression analysis.

2.2.5 Partitioning epithelial and immune cells

To identify additional subtypes of epithelial and immune cells, epithelial and immune cells were partitioned based on the previously annotated cell types from the UMAP. The partitioned epithelial and immune cells were re-clustered using rPCA with the previously described parameters, including integration across sequencing batch, and annotated based on marker genes from literature. Within the immune cell compartment, as observed by others [146, 147], a cluster of cells that exhibited an epithelial and immune cell phenotype was detected. These cells were removed from clustering analysis of the immune cells.

2.2.6 Cell type annotation

Cell types were annotated using marker genes from literature [86, 96]. Additionally, the "FindAllMarkers" function (Wilcoxon rank sum test) from Seurat was employed on the epithelial and immune cells to assist in verification of the annotations. Genes that were detected in at least 25% of the cells were included, and genes that had an adjusted p value < 0.05 were considered significantly differentially expressed for verifying the identified epithelial/immune cell types.

2.2.7 Differential gene expression analysis

MAST was used to compare inflamed and non-inflamed rectal tissue so as to adjust for random effects of individuals [148]. Two samples were excluded from DEG analysis: NI3 (low number of cells) and NI5 (control sample) for a total of 2,863 inflamed and 2,081 non-inflamed cells. The covariate variables include sex, ancestry, and age. DEGs were

filtered for p value <0.05, average log2FC >0.25, and >10% of either inflamed vs. non-inflamed cells expressing the DEG. The list of DEGs were imported to ToppFun for pathway analysis [149]. Pathways were considered significant with a Bonferroni correction p-value < 0.05. IL-6 pathway was identified by the Pathway Interaction Database [150] and Interferon gamma signaling pathway was identified by REACTOME [151].

2.3 Results

During colonoscopy, 6 subjects were classified as IF (5 with active fistula), 6 as non-inflamed (NI) with inactive/healed fistula, and a non-inflammatory bowel disease (IBD) control was included. In total, 67,119 cells from 13 individuals that were dissociated by cold protease digestion and subject to droplet-based single-cell RNA sequencing on the 10X Genomics Chromium platform [142] (methods) were jointly clustered to broadly identify cell types within the rectum.

Batch effect correction was performed on the dataset to facilitate accurate downstream analysis. After testing three methods for reproducibility and rigor based on clustering stability (Adjusted Rand Index), clustering similarity, and proportion of cells from each sample in each cluster (Supplemental Figure 1), reciprocal principal component analysis was chosen. We identified seven broad classes of epithelial and immune cells and a small population of stromal cells within the rectum (Figure 3A). Epithelial cells were the most abundant population within the rectum. Compared to the IF, unhealed mucosa, there were higher proportions of colonocytes in the NI tissue with corresponding increases in the abundance of naïve B cells and plasma cells associated with inflammation (Figure 3B). This observation suggests dysregulation of both the immune component and epithelial

barrier within the rectum of individuals with perianal-CD and highlights heterogeneity of cell type abundance.

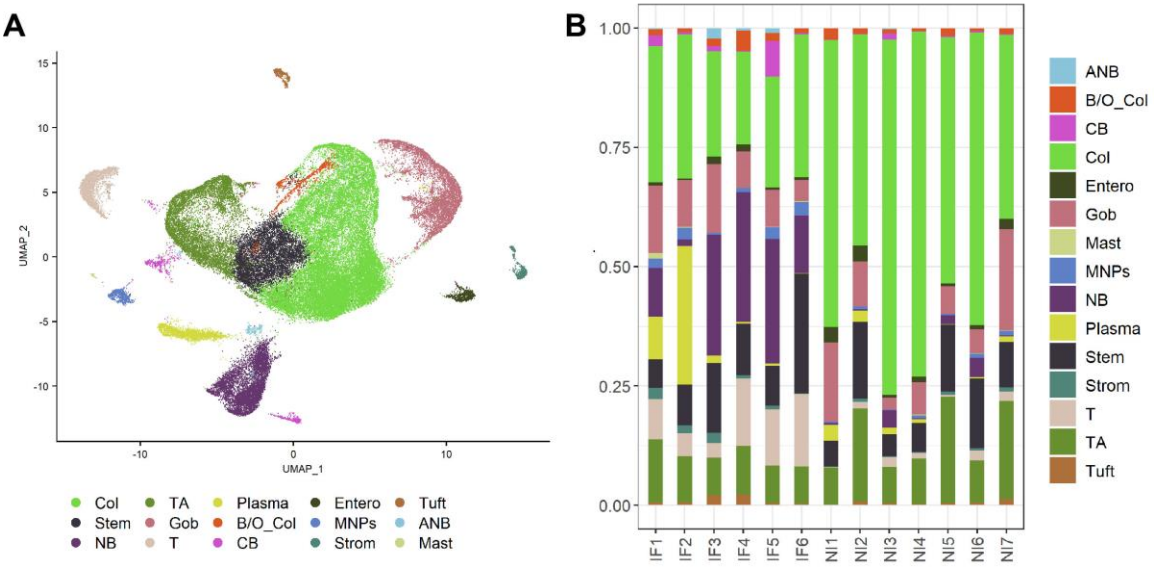


Figure 3 Representation of rectal cell types. (A) UMAP of 15 major cell types detected by cluster analysis of 6 inflamed (IF), 6 non-inflamed CD (NI), and 1 non-IBD control (NI5) samples. Epithelial cells (colonocytes, *BEST4/OTOP2* colonocytes, enteroendocrine, transit amplifying, stem, tuft cells) mostly form the central group with goblet cells to the right, while immune cells (T, cycling, naïve, or activated naïve B, plasma cells, mononuclear phagocytes, and mast cells) tend to be at the periphery of the plot. (B) Stacked bar graph showing the proportion of each cell type in each sample. UMAP, Uniform Manifold Approximation and Projection; ANB, activated naïve B Cell; B/OCol, BEST4/OTOP2 colonocytes; CB, cycling B cells; Entero, enteroendocrine; Gob, goblet; Mast, mast cells; MNPs, mononuclear phagocytes; NB, naïve B cells; Plasma, plasma cell; Stem, stem cell; Strom, stromal cells; T, T cells; TA, transit amplifying cells.

In further analyses of each compartment separately, we sought to identify additional cellular subtypes, leveraging previous single-cell RNA sequencing studies for cell cluster annotation. For consistency, the epithelial cells were re-clustered using reciprocal principal component analysis. Additional sub-populations included immature goblet cells (imGob) (mostly comprised of cells from patient IF1), colonocyte progenitors, and M cells, all of which were mostly present in the IF tissue (Supplemental Figure 2A). The majority of

epithelial cell types, excluding imGob, were identified in both IF and NI rectal tissue (Supplemental Figure 2A,B). We observed an overall trend of increased goblet cell abundance in IF tissue (4/6 > 10% of epithelial cells in IF, 5/7 < 10% in NI), which was offset by higher levels of colonocyte (or colonocyte progenitors in NI2) in NI samples as compared to other epithelial cells (Supplemental Figure 2B).

Similarly, focused analysis on the mucosal immune compartment led to the identification of additional subpopulations within the B and T cell sets, as well as LTi-like NCR^+ ILCs. Although NI tissue had fewer immune cells overall compared to IF tissue [152] (Supplemental Figure 2C), similar distributions of immune cell sub-populations were identified in both classes of samples. However, we did note enrichment of IgG plasma cells in the IF tissue of two donors (IF1 and IF2, Supplemental Figure 2D), contrasting with expansion of IgA plasma cells in two NI samples (NI1 and NI2). Considerable heterogeneity of immune cell proportions was observed, notably for naïve B cells, and surprisingly mononuclear phagocytes were seen in more of the NI samples suggesting they may play a role in healing.

Next, focusing on goblet cells because of their protective function in the epithelial barrier, interaction with gut microbiota [153], and increased abundance in IF tissue of perianal-CD, we performed differential expression analysis to elucidate the altered pathways involved in persistent inflammation. By comparing IF vs NI rectal tissue (n = 6 IF, n = 5 NI: methods), we identified 226 significant up-regulated differentially expressed genes, some of which were enriched for the inflammatory pathways involving interleukin 6 and interferon gamma [140, 153, 154], along with 92 significant down-regulated differentially expressed genes (Figure 4A,B). It is conceivable that increased interleukin 6

and interferon gamma signaling might prevent healing by disrupting epithelial tight junctions [155].

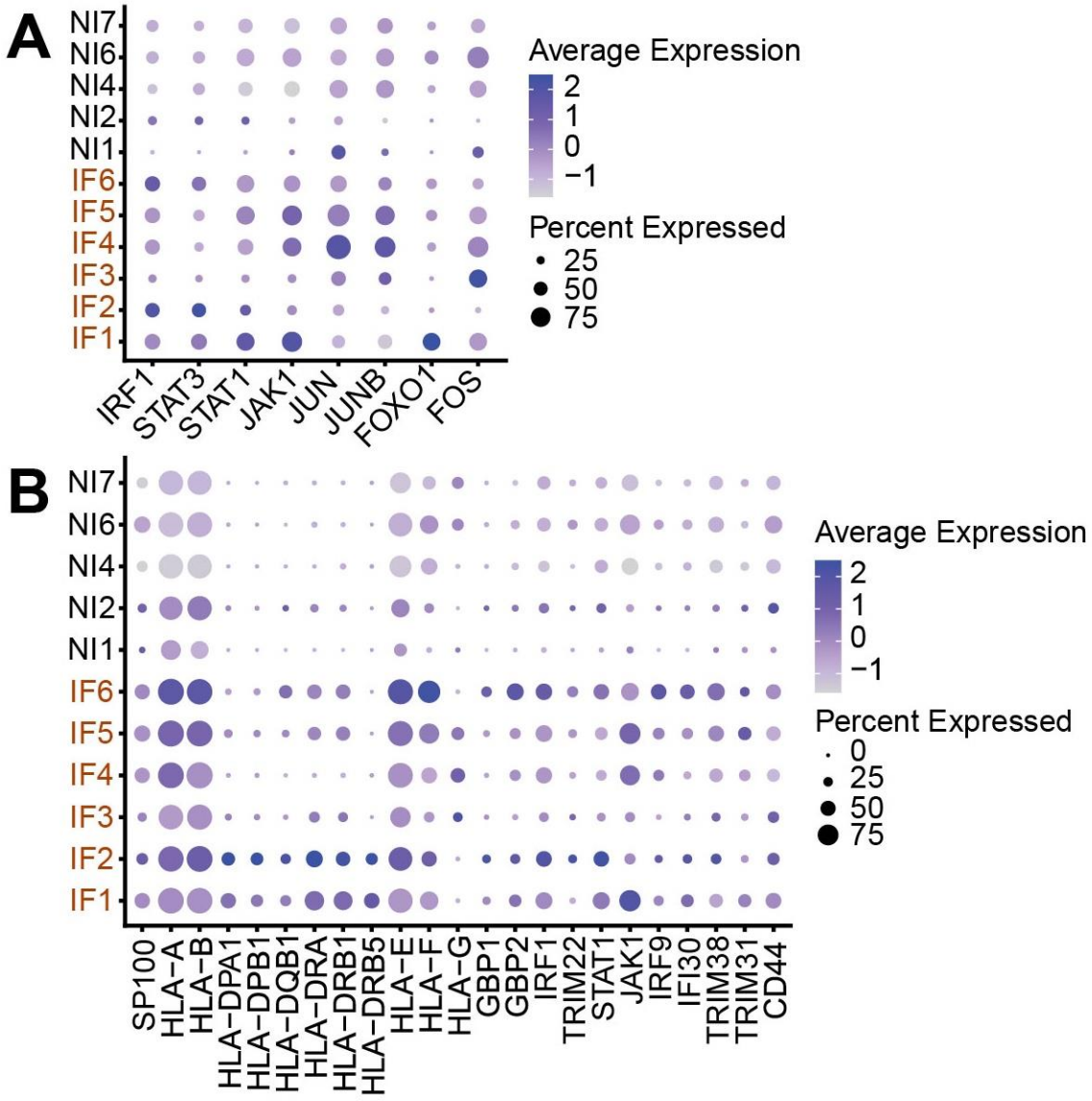


Figure 4 Signatures of inflammation in goblet cells. Dot plots of expression of genes in goblet cells identified in 6 inflamed (IF) and 5 non-inflamed (NI) samples representing (A) IL-6 inflammatory pathways and (B) interferon gamma pathway. The size of each dot corresponds to the percentage of cells expressing each gene, and the color shading represents the average log-normalized expression in cells that express it. IL-6, interleukin 6.

2.4 Discussion

In summary, our findings implicate altered epithelial cell function during persistent inflammation, which could further promote mucosal damage in perianal-CD. Further studies are needed to determine if the higher levels of goblet cells in the IF rectal mucosa are a result of the ongoing inflammatory response or contributing to it. Single-cell transcriptomics of the rectum in cases of idiopathic fistula (non-IBD) should also establish whether altered cytokine levels [156] are due to specific aspects of cellular pathology in CD. Our data suggest that persistent perianal-CD despite anti-TNF treatment may be due to the altered behavior of goblet cells (in one case imGob) and colonocyte precursors in IF tissue that express inflammatory pathways. These findings appear to be related to studies showing imbalances in epithelial subtype composition when pathways downstream of Myd88 and Cox-2 are disrupted during inflammation [157]. Remarkably, *PTGER4*, a known IBD risk locus [158] expressed in the epithelium, encodes a receptor that responds to paracrine stimulation from prostaglandins produced by Cox-2 expressed in the mesenchymal stromal cells [159]. Taken together, our findings suggest that epithelial differentiation is affected by inflammation during perianal-CD and has a negative impact on fistula healing. Longitudinal profiling studies are currently underway to systematically map these cellular changes in mucosal biology over the course of each patient's disease.

CHAPTER 3. IDENTIFICATION OF CROHN'S DISEASE SUBTYPES IN SINGLE CELL RNA SEQUENCING SIGNATURES OF TREATMENT NAÏVE SAMPLES ACROSS THE PAEDIATRIC GASTROINTESTINAL TRACT

3.1 Introduction

Crohn's disease (CD) is a life-long, chronic inflammatory disease of the gastrointestinal tract. This chronic inflammation in CD is patchy in nature, often transmurally involved with remitting and relapsing disease course but progressive in nature [9]. Despite effective therapies such as biologics and small molecules targeting various pathways to counter inflammation, many patients eventually progress to complications such as strictures and internal penetrating disease with abscesses [160, 161]. Further, we observe a therapeutic ceiling that none of these therapies can achieve mucosal healing in more than 50% of CD cases [27, 162, 163]. It is commonly believed that pathogenesis of CD is multifactorial and known to be associated with genetic susceptibility, immune dysregulation and microbial dysbiosis [1, 9]. The onset of CD occurs at any age with peak incidence between 15-25 years; however, clinical and etiological research findings so far point toward the same underlying pathogenesis for both children and adults [20, 164]. Despite extensive research, the precise mechanisms of associated genes, environmental factors and other unknown factors remain poorly understood. This may be due to most research on CD being conducted using biopsies from a single intestinal location, either from patients with established Crohn's disease or from discarded surgical tissue, which

represents the effect of treatment biologics or end stage of the disease rather than early/treatment naïve CD pathogenesis.

Recent advances resulting in many landmark publications with single-cell RNA sequencing (scRNA-seq) technology have provided an unprecedented opportunity for dissection of the complex cellular and molecular landscape of CD in the mucosa at the individual cell-type level [87-89, 92, 96, 165]. These studies have revealed the contribution of cellular diversity in the intestinal microenvironment to the pathogenesis, progression, and treatment outcomes of CD. A key objective of most scRNA-seq analyses is to classify tissue from a heterogenous cell composition into meaningful populations, uncovering cell type specificity in the biological question of interest. While scRNA-seq has enabled discovery of novel cell types and states, correctly identifying the cell types/states remains a challenge due to heuristic clustering parameter selection leading to non-robust clustering results and high dimensional data [81, 83]. Methods that evaluate the robustness of clustering or downstream results from scRNA-seq studies are needed. This problem persists as datasets grow and become more heterogenous. To better understand the underlying mechanisms of initial disease onset across donors and the different regions of the intestines, we approached this question by assessing rigor and repeatability of cell type assignment while evaluating the impact of cell type assignment on downstream results. Different methods to survey clustering stability include bootstrapping, subsampling, and optimal transport alignment [166]. In this study, we employ a bootstrapping method to assess clustering stability of a treatment naïve CD cohort and filter out cells that are considered unstably assigned.

Treatment naïve CD at the time of diagnosis in children with short duration of symptoms before diagnosis may provide the best opportunity to examine the nature of early

inflammatory and pathogenic events in CD. For this study we recruited 34 children at the time of diagnosis, obtaining mucosal biopsies during the diagnostic colonoscopies from three locations including both inflamed and non-inflamed bowel locations for scRNA-seq experiments resulting in 270,268 cells post quality control and the stability assessment. This cohort provides an opportunity to examine how inflammation influences cellular contribution to disease since the subjects are treatment naïve and not confounded by years of inflammation, therapies, diet, or other environmental exposures. We hypothesized that discovery of distinct cellular signatures in immune, epithelial, and stromal cells in treatment naïve CD during early stages of inflammation may provide novel pathways to target new therapies while evaluating signatures of disease management and prognosis. Bulk RNA-seq analysis of the RISK study has, for example, suggested bias towards mesenchymal or immunological engagement in stricturing and penetrating disease, respectively, but could not resolve cellular contributions [53].

Here we show that the higher resolution provided by scRNA-seq allows interrogation of specific cellular contributions to disease behavior while recapitulating signatures observed in bulk RNA-seq. Moreover, we demonstrate that donors can be categorized into clinically meaningful groups based on tensor decomposition with the tool single-cell Interpretable Tensor Decomposition (scITD) [167]. Lastly, we integrate genome wide association studies (GWAS) and scRNA-seq data to identify cell types associated with CD across the intestines.

3.2 Materials and Methods

3.2.1 Patient recruitment and ascertainment

The study adheres to all the relevant guidelines and regulations. Mucosal biopsies were obtained from newly diagnosed Crohn's disease patients (treatment naïve CD) undergoing clinically indicated colonoscopy at Children's Healthcare of Atlanta (CHOA), under a protocol approved by Emory University Institutional Review Boards (IRB). This study includes a diverse pediatric cohort with patients ranging from 3 to 18 years of age. Written consent was obtained from all the patients and/or their legal guardians. Mucosal biopsies were collected from various intestinal locations, including the ileum, ileocecal valve, colon (cecum, ascending/transverse/descending), rectal sigmoid, and rectum. Biopsies collected specifically from the ileocecal valve to the descending colon are categorized as colon and rectal sigmoid/rectum as rectum for downstream analysis. In total, the cohort included 34 donors, with 27 ileum, 36 colon, and 32 rectum samples. For each biopsy, both the macroscopic (endoscopic) and microscopic (pathology from adjacent biopsy) inflammation statuses were determined by reviewing the endoscopic and histological/pathology reports of the donors from whom the biopsies were collected.

3.2.2 Phenotypic classifications of B1, B2, and B3

Non-stricturing, non-penetrating (B1) disease was classified as an uncomplicated disease state. Stricturing disease (B2) refers to persistent luminal narrowing with pre-stenotic dilation. Internal penetrating disease (B3) refers to intra-abdominal fistulizing disease resulting in intra-abdominal or pelvic abscesses or fistulas to an adjacent organ (excluding vagina or perianal region) [53].

3.2.3 Sample preparation, processing and quality control

Biopsies from ileum, colon and rectum regions of each patient were immediately processed using a cold protease (Sigma Cat #P5430) protocol. Enzyme mix was prepared using Rho kinase (Y-27632, CAS#872543-07-6) and Caspase inhibitors (CAS#187389-52-2) in DPBS with 0.5M EDTA. Biopsies were minced and digested twice (for 15 minutes each) and passed through a 40 µm filter. Single cell suspensions were washed with buffer containing Benzonase (CAS#9025-65-4) and centrifuged at 400 g for 5 minutes. Cells were resuspended in 200-300 µl of RPMI with 10% FBS and counted. Single cell encapsulation was carried out using 10X reagent mix and loaded into the Chromium microfluidic device according to manufacturer's instructions. This was performed using Chromium Next GEM Single Cell 3' Reagent Kit v 3.1 Dual Index (PN-100268) with a target of 10,000 cells. Barcoded sequencing libraries were made from the amplified cDNA and sequenced at the Molecular Evolution core at Georgia Tech using the Illumina NOVA S4 kit with 50,000 reads per cell.

Raw sequencing files were aligned to GrCH38-2020-A using CellRanger count (v. 6.1.2) [142], and the resulting gene expression count matrices were assessed with Seurat (v. 4.3.0) [77]. The following quality control (QC) metrics were implemented to retain high quality cells: number of genes per cell 250-3000, mitochondria percent <30%, and keep genes that are expressed in 10 or more cells.

3.2.4 Stability assessment

The clustering stability workflow was performed for each tissue separately (R v.4.2.1). rPCA (reciprocal Principal Component Analysis) batch effect correction was applied to correct for sequencing batch effect in the ileum, colon, and rectum data for the

stability assessment. The typical Seurat rPCA clustering pipeline [77] was implemented with the log-transformed and normalized data (GitHub: https://github.com/GibsonLab-GT/CD_transcriptomics). Each tissue was clustered separately with resolution of 0.20, 15 principal components (PCs), and 2000 highly variable genes (HVG), referred to as the reference dataset. The reference dataset was randomly split into two non-overlapping subsets, called the query. The following query parameters were implemented: resolution of 0.20, 15 PCs, and 2000 HVG to have similar cluster number between reference and query. Cluster annotations were mapped from the reference to the query clusters with Seurat's TransferData function using PCs 1-15. Stably assigned cells were identified by comparing how often cells in each query cluster were assigned to the same reference cluster. We required at least 80% concordance between cells in the query cluster and reference cluster to match clusters. If the query cluster was partitioned into two or more clusters that mapped to the same reference cluster, the clusters containing the largest proportion of cells adding up to 80% were considered to be the same. This process was repeated four additional times for a total of 10 queries that were compared to the reference. Next, a list containing the number of times each cell was assigned to the same cluster, and cells were considered stably assigned if this occurred in four out of the five times the query was made. These clustering results were termed the low-resolution stably assigned cells.

The clustering stability assessment was also performed with a higher clustering resolution for the ileum data. The same workflow was implemented but the reference was clustered at resolution of 1 and the queries with a resolution of 1.5. The following parameters were considered for low-, high-resolution stably assigned cells, and reference clustering: resolution of 0.25, 0.50, 0.75, and 1, PCs 1-10, 1-15, and 1-30, and HVGs of

500, 2,000, and 5,000. The adjusted Rand Index computed using (ARI) Dune (v. 1.10) merge function was used to assess clustering similarity within the low-, high-resolution stably assigned cells, and reference clustering parameters[145]. The top three clustering results based on mean ARI score for low resolution stably assigned cells and median ARI score for high resolution stably assigned cells and reference were further assessed for clustering robustness. The top clustering result of the low-, high-resolution stably assigned cell, and reference were selected based on proportion of cells from each sample and batch in each cluster, as well as segregation of cell type based on marker gene expression. These results are referred to as the clustering set. The Jaccard index (JI) was used to quantify the similarity of cells present within each cell type across all three sets of clusters after cell annotation. Differential gene expression was performed for each set using Seurat's FindAllMarkers function (Wilcoxon Rank Sum Test) with min.pct = 0.25. Spearman correlation of the average log2 fold change (FC) was calculated for the overlapping genes within each cell type for each set.

The top three ARI clustering results based on mean scores for low-resolution stably assigned cell clustering results for ileum and median scores for colon and rectum were assessed for each tissue. The final set of clusters were determined based on criteria previously described. The clustering parameters used for the ileum were resolution 0.5, 1-15 PCs, and 2,000 HVG, parameters for colon were resolution 0.25, 1-30 PCs, and 2,000 HVG, and parameters for rectum were resolution 0.75, 1-15 PCs, and 2,000 HVG. Doublet clusters were removed from each tissue, and the data was then re-clustered with same parameters previously described.

3.2.5 Cell proportion and hierarchical clustering analysis

Speckle (v. 0.99.7) was used to perform the cell type proportion analysis [168]. A t-test or ANOVA test was used to determine if cell type proportion was significantly associated (FDR adjusted p value < 0.05) with macroscopic/microscopic inflammation status, batch, sex, or SIRE group in the ileum and rectum data. Additionally, a t-test was used to determine if cell type proportions were significant across ileum donor group status from scITD. A linear model with individual as a random effect was used to assess relationship between cell type proportion and metadata variables mentioned previously in the colon data.

Hierarchical clustering was performed using the transformed proportion of cells, calculated from Speckle, for each sample with Wards' method using Euclidean distance. The number of groups identified in each tissue was determined by measuring within group sum of squares and visualized with an elbow plot.

3.2.6 Tucker tensor decomposition

The R package scITD (v. 1.4.0) was used to perform Tucker tensor decomposition within each tissue [167]. The tensor was formed using the raw pseudobulked count data corrected for batch with Combat. The following parameters were used to form the tensor: donor_min_cells = 5, norm_method = "regular", scale_factor = 1,000,000, vargenes_method = "norm_var", vargenes_thresh = 500, scale_var = TRUE, and var_scale_power = 0.5. SVD Rank determination was used to establish the number of factors and gene sets used for tensor decomposition. Additionally, stability of the selected ranks were assessed with the run_stability_analysis() function. The number of factors and gene sets used were 2 and 9 in the ileum, 4 and 10 in the colon, and 4 and 10 in the rectum,

respectively. The ‘hybrid’ rotation method and ‘regular’ tucker decomposition method was applied. Positive or negative donor scores used to stratify donors into groups represent the relative degree of a gene expression pattern observed in that donor [167]. Some donors and cell populations were excluded from scITD analysis due to low cell count (Supplementary Table 1).

3.2.7 Differential gene expression and pathway analysis

Differential gene expression analysis comparing groups identified in the tensor decomposition analysis was performed using Dreamlet (v. 1.1.1) in R (v.4.3.1) for each tissue separately [169]. The raw count data was pseudobulked per cell type per sample with default parameters in Dreamlet using the aggregateToPseudobulk() function. Voom style normalization with processAssays() was performed using sequencing batch and as a covariate and individual as a random effect for the colon data and sequencing batch as a covariate for the rectum data with min.cells = 10 for all three tissues. Batch was not used as a covariate for the ileum data. The dreamlet() function was used to perform differential gene expression analysis comparing group status with the same covariates used for normalization. Genes were considered differentially expressed if the FDR adjusted p value < 0.05. Pathway analysis with significant differentially expressed genes was performed using the clusterProfiler (v. 4.10.1) R package with Gene Ontology Biological Pathways [170]. Significant pathways had a Bonferroni adjusted p value < 0.05.

3.2.8 Cell-cell communication

CellChat (v. 2.1.2) was implemented to perform cell-cell communication analysis comparing groups identified from tensor decomposition in each tissue [78]. Default parameters were used to infer cellular communication networks.

3.2.9 Gene module score

Gene module scores were used to quantify the strength of gene expression in a gene set related to a biological process [171]. UCell (v. 2.2.0) was used for module scoring within each tissue [171]. Genes present in the ileum from the Gene Ontology (GO) pathways, metallopeptidase activity (GO:0008237) and chemokine receptor binding (GO:0042379), were used for module scoring. Genes present in colon from GO pathways macrophage activation (GO:0042116) and regulation of inflammatory response (GO:0050727) were used to create module scores. Genes present in rectum from GO pathways type II interferon response (GO:0034341) and response to TNF (GO:0034612) were used to create module scores.

3.2.10 Disease behaviour bias in the ileum

Average gene expression was calculated from the normalized expression values for each group using the AverageExpression() function in Seurat. Log2FC + 1 comparing group 1 vs. group 2 was calculated. Genes with Log2FC difference > 0.25 or < -0.25 were retained. Genes present in the Gene Ontology molecular functions and biological processes pathways identified as enriched in B2 or B3 ileum samples in Table S4, S5, and S6 from Kugathasan et al. were used to assess enrichment for B2 or B3 bias in the ileum group data [53].

3.2.11 PCA in colon

Pseudobulked and normalized myeloid cell gene counts from Dreamlet were used for scaled PCA of the DEGs enriched in macrophage activation pathway (GO:0042116) with prcomp() function (Supplementary Table 2). The normalized mean gene counts were used for PCA of macrophage activation pathway DEGs across all cell types identified in the colon.

3.2.12 GWAS and scRNA-seq integration analysis

The following workflow was adapted from Duncan et al. and MAGMA (v1.10) was used to identify cell type specific gene sets associated with CD [172, 173]. Prior to running MAGMA, the average normalized expression of each gene per cell type was calculated with AverageExpression() in Seurat. The cell type specificity score used to identify cell type specific gene sets, were calculated as the fraction of a gene's expression across all cells per tissue. After pre-processing the data, GWAS summary statistics from Liu et al. were obtained after filtering for high quality variants [174]. MAGMA was implemented to annotate SNPs from GWAS and identify genes associated with CD using the probit transformation of SNP p-values from GWAS summary statistics (snp-wise=mean). Linkage disequilibrium was adjusted for by using the European ancestry panel from phase 3 of 1000 Genomes Project [175]. Then gene property analysis was performed to identify cell types associated with CD using the cell type specificity scores to define a gene-set. A one-sided positive test was used to test for cell type association with CD. Lastly, a conditional analysis with the top cell types after Bonferroni correction was performed to identify signals that were most likely acting independent of other cell types using the same

criteria explained in Duncan et al. [172] Conditional analysis was not performed on the rectal data since only one cell type was retained after Bonferroni correction. This analysis was performed for each tissue separately.

3.3 Results

3.3.1 Patient recruitment and sample preparation

A diverse cohort of treatment naïve pediatric CD patients (ages 3-18) was recruited prospectively within the Children's Healthcare of Atlanta network. Biopsies were obtained from the ileum, different sections of the colon, rectal sigmoid, and rectum from each donor. Samples from the ileocecal valve to descending colon were classified as colon, and rectal sigmoid and rectum were classified as rectum for downstream analysis. When research biopsies were obtained, a corresponding biopsy from same location was acquired for histological analysis. Endoscopic images and histology reports were gathered. A panel re-examined all reports for macroscopic and microscopic inflammation to determine the inflammation status (methods). In total, there were 34 donors, 27 ileum samples, 36 colon samples, and 32 rectum samples included in this cohort. The biopsies were sequenced with 10X Genomics 3' scRNA-seq technology (methods, Supplemental Figure 3A) [142]. Additional patient metadata is provided in Supplemental Table 2.

3.3.2 Filtering for stably assigned cells improves overall clustering stability and similarity

In order to be confident in cell type assignments, we assessed clustering stability by conceptualizing a workflow [82] that tests robustness and repeatability of clustering results while retaining dataset heterogeneity.

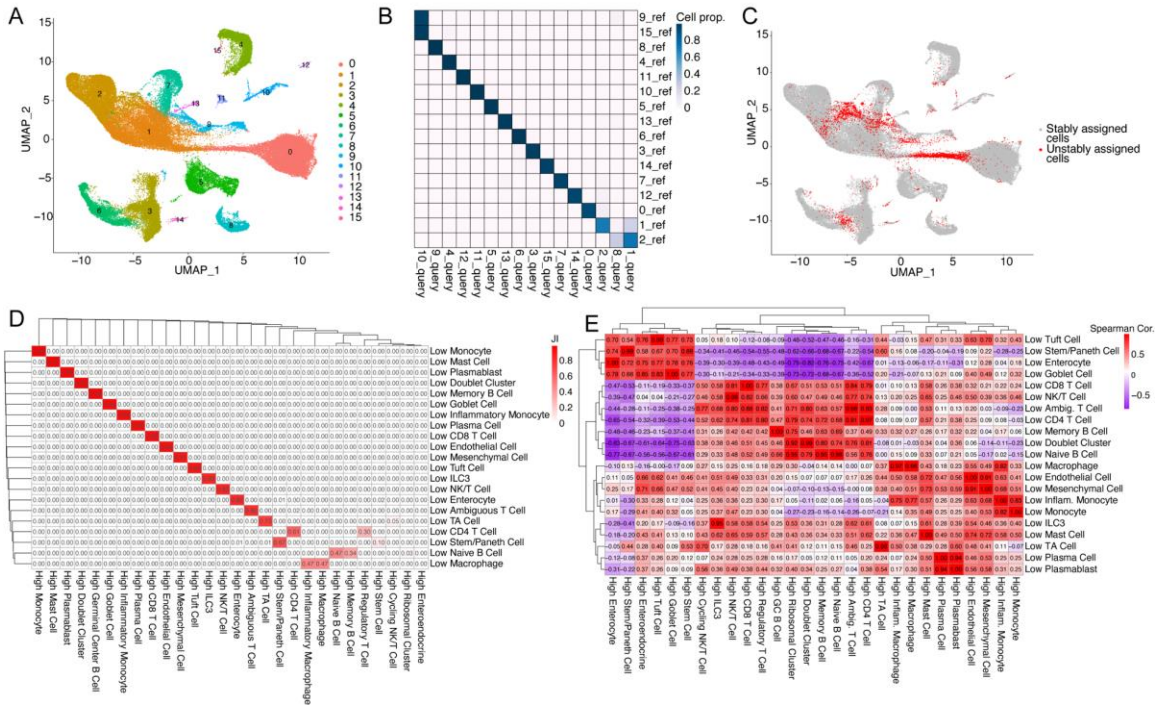


Figure 5 Stability assessment and rigor of ileum clustering results. (A) The "reference" ileum clustering results (87,785 cells). (B) Example of a confusion matrix heatmap showing the proportion of cells shared between query and reference clustering results. Cell prop. = cell proportion, ref = reference. (C) Stably assigned (83,702 cells) and unstably assigned (4,083 cells) cells highlighted by the low-resolution stability assessment. (D) Heatmap comparing the cells present in each cell type between the low and high resolution stably assigned cells scored by the Jaccard Index. JI = Jaccard Index. (E) Spearman correlation of Log2FC values of shared DEGs in the low and high resolution stably assigned cells. FC = fold change, Spearman Cor. = Spearman correlation, ILC3 = type 3 innate lymphoid cell, NK = natural killer, TA = transit amplifying, UMAP = Uniform Manifold Approximation and Projection.

After sequencing, alignment, and obtaining count matrices, quality control (QC) was performed for each tissue separately. The same QC parameters were implemented for each tissue resulting in a total of 87,785 ileum cells, 111,282 colon cells, and 88,355 rectum

cells post QC (methods, Supplemental Figure 3B-D). The QC'ed data, termed reference, was clustered using the Seurat [77] pipeline with rPCA batch effect correction (methods) at a low clustering resolution to identify the broad cell types. An example of the ileum reference clustering results is shown in Figure 5A. Next, the reference was split in half, termed query, and each query was clustered with rPCA batch effect correction. Query cell cluster assignment was compared to the matching reference cell cluster assignment (methods), and clusters that share 80% or more cells were considered to be matched (Figure 5B). There was high concordance of cell type assignment between the reference and query, indicating cell assignment was generally stable. This process was repeated four additional times, for a total of ten queries that were evaluated. A list of cells that repeatedly cluster together, termed stably assigned cells, was generated. We hypothesized that removing the unstably assigned cells would improve clustering stability because they are most likely a mixture of low-quality cells and transitioning cells that do not reliably retain cluster assignment, potentially affecting downstream results and interpretation. For this study, cells that were clustered together at least four times were considered stably assigned and retained for downstream analysis. This general workflow, termed the low-resolution clustering assessment, was executed for each tissue separately. Generally, unstably assigned cells were typically found at the junction between two different clusters (Figure 5C). Most of the cells considered unstably assigned were primarily located in epithelial clusters (Figure 5C). After the stability assessment, 95.34% of the ileum cells, 94.41% of colon cells, and 92.24% of rectum cells were retained.

To further evaluate the robustness of the stably assigned cell clustering results, the same workflow was implemented with a higher clustering resolution for the ileum data,

which was called the high-resolution clustering assessment (methods). The low-resolution stably assigned cells, high-resolution stably assigned cells, and reference were each clustered with a variety of different clustering parameters (methods). The clustering results were compared to evaluate the effect of the stability assessment on downstream results. The adjusted Rand Index (ARI) computed with Dune was used to evaluate clustering similarity across different clustering parameters within a dataset [145]. The low resolution stably assigned cells had a mean and median ARI of 0.87 and 0.86, the high resolution stably assigned cells had a mean and median ARI of 0.81 and 0.83, and the reference had mean and median of 0.82 and 0.84 ARI in the ileum. The high ARI scores indicate that the clustering results are similar within the dataset, regardless of clustering parameters chosen.

The top clustering parameters for the ileum low-resolution, high-resolution stably assigned cells, and reference were annotated using marker genes. Each of these results will be referred to as clustering set. Differential gene expression analysis comparing one cell type vs. the rest was also performed for each clustering set. The Jaccard index (JI) was used to evaluate the similarity of cells present within each annotated cluster. Overall, there was high concordance of cells present within each cell type, except for sub-cell types that were identified in one clustering set and not the others. For example, the ribosomal cluster, likely a technical artifact, was found in the high resolution stably assigned cells and reference (Supplemental Figure 4A,B) but absent in the low resolution stably assigned cells. As a result, this cluster exhibited a low JI (~0) (Figure 5D). The stability assessment removed most cells that comprised this cluster from the low-resolution stably assigned clustering result (Figure 5C). Additionally, pairwise comparisons of the average log2 fold change (log2FC) values from the differential gene expression results were evaluated to assess

similarity in downstream analysis. There was high correlation between differentially expressed gene (DEG) results for each cell type and the broad cell types (Figure 5E). Some distinct cell types had higher concordance in part because this analysis was restricted to comparison of the same DEGs. We used the low-resolution stability assessment workflow for the ileum, colon and rectum data since (i) there was evidence that the different clustering sets yield similar results, (ii) downstream pseudobulk analyses were based on the broad cell type to investigate variation within each cell type, and (iii) the low-resolution retained more cells than the high-resolution stably assigned cell clustering. The mean and median ARI was 0.89 and 0.92 in the colon, and 0.82 and 0.85 in the rectum, respectively.

Overall, these results demonstrate that the stability assessment improves the clustering stability and similarity. Additionally, heterogeneity within each dataset/cell type was preserved while ensuring results are robust to the selected stably assigned cells.

3.3.3 Modest influence of inflammation on cellular proportions

After the stability assessment was performed for each tissue, the low-resolution stably assigned cells were filtered for doublet clusters and re-clustered with the same parameters (methods). The clusters were annotated using marker gene expression and validated by assessing the top DEGs for each cluster (Supplemental Figure 4C). Across each tissue, epithelial, immune, mesenchymal, and a small population of endothelial cells were identified (Figure 6A-C). An ambiguous T cell population was detected in the ileum and rectum data, having some marker gene expression of T cells accompanied by ribosomal gene expression but did not express *CD4* or *CD8A/CD8B.* The same ambiguous T cell population was also observed in the high-resolution stably assigned clustering and the

reference clustering in the ileum data, suggesting that it was not driven by the selected clustering parameter (Supplemental Figure 4A,B). Additionally, an ambiguous B cell population in the rectum was identified, again expressing some B cell marker genes accompanied by ribosomal gene expression. In general, the expected cell types were identified within each tissue, even after filtering for the stably assigned cells.

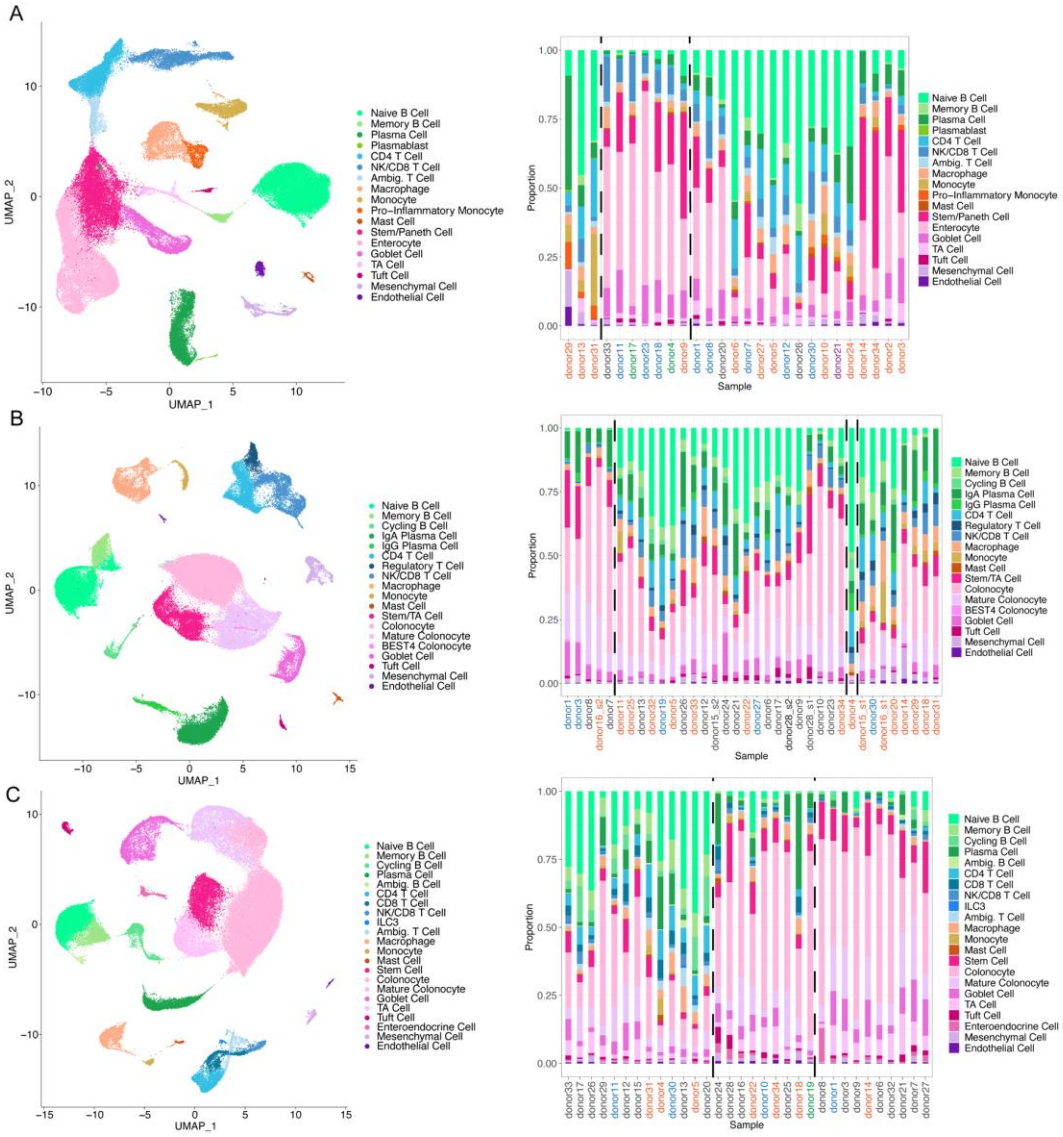


Figure 6 Clustering results reveal heterogeneity within and across tissue. (A-C) (Left) Ileum (top), colon (middle), and rectum (bottom) low-resolution stably assigned cell clustering results. (A-C) (Right) Stacked bar plots of proportion of cells within each

donor grouped by hierarchical clustering results. Dashed lines separate each group identified by hierarchical clustering. (ileum – top, colon – middle, rectum – bottom). Sample labels in black are both macroscopically and microscopically non-inflamed, labels in orange are both macroscopically and microscopically inflamed, labels in blue are macroscopically inflamed and microscopically non-inflamed, labels in green in macroscopically non-inflamed and microscopically inflamed, and label in purple represents macroscopically inflamed but microscopic inflammation is unknown. Ambig. = ambiguous, ILC3 = type 3 innate lymphoid cell, NK = natural killer, TA = transit amplifying.

Post stability assessment, we still observed heterogeneity across and within each tissue. We were first interested in cellular composition differences across the intestines. Directly comparing epithelial cell proportion across each tissue, we observed increased epithelial cell proportion in the rectum compared to the colon, which was not as pronounced in the ileum to rectum or ileum to colon comparison (Supplemental Figure 5A).

Since previous studies observed differences in cellular compartments with respect to inflammation status, we hypothesized that inflammation would affect the immune and epithelial cellular proportions across each tissue [97]. Speckle [168] was used to assess whether cell type composition was associated with macroscopic/microscopic inflammation (macro IF/micro IF) status and other metadata variables (methods). Interestingly, inflammation status was not strongly associated with ileal or rectal cell proportion, most likely due to the high heterogeneity across inflamed and non-inflamed samples. The relatively low variance component due to inflammation may be related to generally high inflammation in an inception cohort but could also be influenced by the relatively small number of donors, and by sampling bias at the biopsy sites. Macro and micro IF status was however associated with monocytes in the colon (FDR adjusted p value = 0.05, Student's

t = -3.2 macro IF status, and FDR adjusted p value < 0.01, Student's t = -5.3 micro IF status, Figure 6B and Supplemental Figure 5E).

Since inflammation was not largely affecting cell composition, we performed hierarchical clustering using the proportion of cells within each sample to identify patterns of cell composition (methods). Across each tissue, we observed that one group tended to have over representation of epithelial cells, and one group tended to have over representation of immune cells (Figure 6A-C, Supplemental Figure 5B-D). In the ileum and colon, some donors were mostly comprised of all immune cells. Alternatively, one group in the rectum was comprised mostly of all epithelial cells [141]. This heterogeneity across tissue and sample could be biologically meaningful or a consequence of sampling bias during tissue collection.

3.3.4 scITD stratifies donors into potentially clinically important groups

To investigate variation of CD across samples within each tissue, we performed Tucker tensor decomposition analysis with scITD [167] (methods). Briefly, scITD identifies patterns of gene expression that covary across cell types and donors to meaningfully stratify patients: rather than independently identifying co-regulated gene sets within each cell type, scITD finds patterns shared by multiple cell types. Most of the variance within the ileum was explained in two factors (Figure 7A). Factor 1 was extreme in three donors in Batch 1 and was weakly correlated with SIRE (self-identified race and ethnicity), and factor 2 was weakly correlated with sex. Within the colon, four factors explained most of the variance (Figure 7B). Factor 1 was significantly associated with batch; factor 3 also captured variation from two donors in batch 1. Factor 2 was weakly

correlated with macro and micro IF status while factor 4 was weakly correlated with macro IF status. Lastly, four factors explained most of the variance within the rectum (Figure 7C). Factors 1 and 2 were weakly correlated with macro IF status. Factor 3 was weakly correlated with SIRE, and factor 4 was weakly correlated with micro IF status and sex. Batch did not drive tensor decomposition within the rectum.

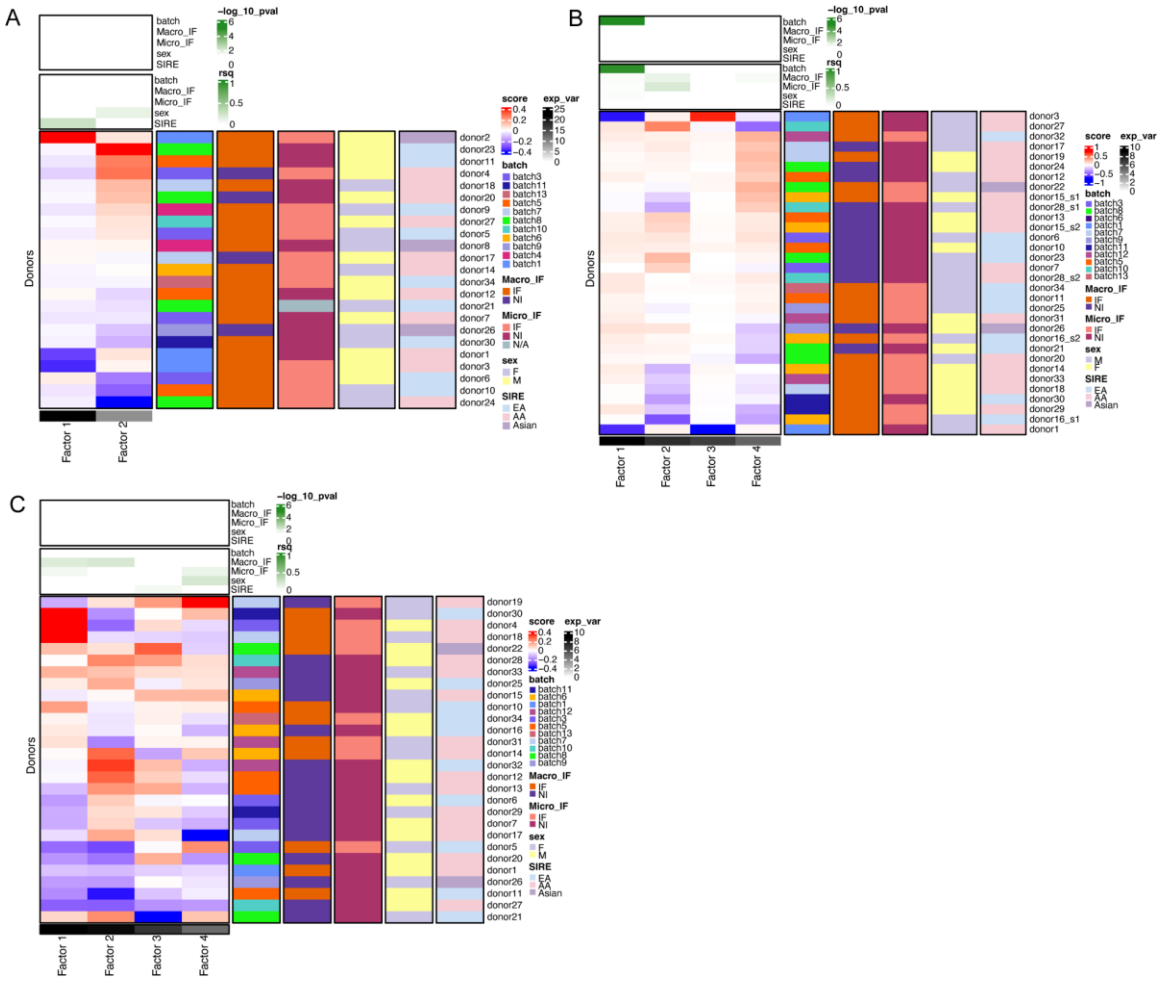


Figure 7 scITD stratifies donors into clinically important groups. (A-C) Ileum (top), colon (middle), and rectum (bottom) scITD donor matrix heatmap showing association (top box) and correlation (bottom box) of metadata variables. Variance explained by each factor is shown under the heatmap for each tissue. Macro_IF = macroscopically inflamed, Micro_IF = microscopically inflamed, IF = inflamed, NI = non-inflamed, F = female, M = male, SIRE = self-identified race and ethnicity, EA =

European American, AA = African American, exp_var = explained variance, rsq = r-squared.

We next were interested in investigating the underlying differences between donors within each tissue. Focusing on factor 2 because batch was driving factor 1 in the ileum, donors with a positive score were called group 1 ($G1_I$, n = 13) and donors with a negative score were called group 2 ($G2_I$, n = 10). Focusing on factor 2 in the colon due also to batch driving factor 1, samples with a positive score were called $G1_C$ (n = 19) and samples with a negative score were called $G2_C$ (n = 13). Colon samples from donors 15 and 16 were stratified into separate groups. For donor 15, one colon sample was inflamed, and the other was non-inflamed which most likely explains the difference in group identity. Although both colon samples from donor 16 were inflamed, sample one had an expansion of immune cells whereas sample two had an expansion of epithelial cells (Figure 6B, Supplemental Figure 5C,E), potentially explaining differences in group identity. Finally, donors with a positive score in factor 1 were called $G1_R$ (n = 13) and negative scores in factor 1 were called $G2_R$ (n = 15) in the rectum. For these group comparisons, the null hypothesis was no differences in group status, whereas the alternative hypothesis was inflammation, or an alternative disease mechanism stratifies donors into groups.

3.3.5 Groups one vs. two in the ileum are biased towards B2 vs. B3 signatures independently identified in the RISK cohort

To study the differences between ileum $G1_I$ and $G2_I$, pseudobulked differential gene expression and pathway analysis was performed (methods). These preliminary findings revealed up-regulation of pro-inflammatory pathways such as response to type II interferon (interferon-gamma, IFN-γ) in enterocytes of $G2_I$ donors (Supplemental Figure

6A,B). These results were recapitulated in the cell-cell communication analysis comparing groups, in addition to displaying enrichment of pro-fibrotic pathways such as THBS, KLK, and CypA in $G1_I$ donors (methods, Supplemental Figure 6C). These results suggested alternative modes of disease mechanism in $G1_I$ vs. $G2_I$ donors.

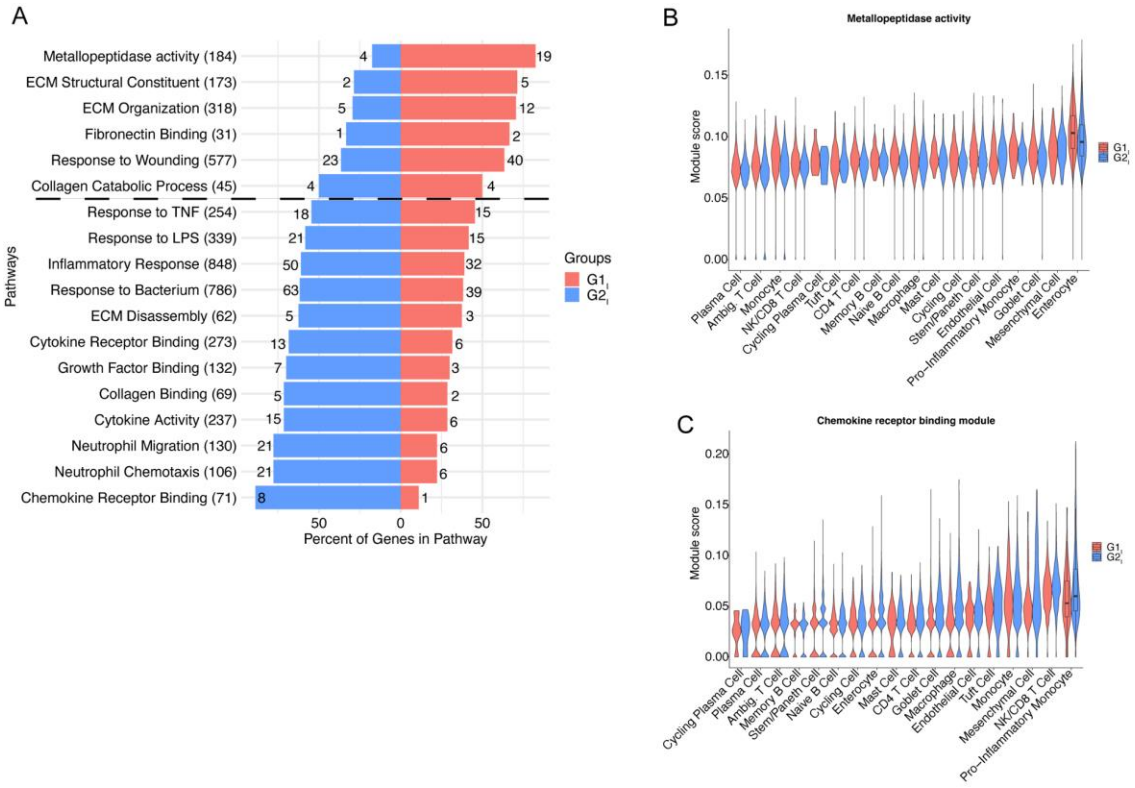


Figure 8 Ileum group 1 vs. 2 demonstrate B2 vs. B3 bias independently identified in RISK study. (A) Proportion of genes enriched in B2 (above dashed line) vs. B3 pathways (below dashed line) identified in RISK study. Number of genes in GO pathway are displayed in parentheses next to the pathway. Number of genes represented in group 1 or 2 are displayed next to each bar. (B) Gene module score of metallopeptidase activity per cell type in group 1 and 2 donors ordered from lowest to highest score. Median module score (line) and interquartile range (box) is shown for enterocytes. (C) Gene module score of chemokine receptor binding per cell type in group 1 and group 2 donors ordered from lowest to highest score. Median module score (line) and interquartile range (box) is shown for enterocytes. ECM = extracellular matrix, $G1_I$ = group 1, $G2_I$ = group 2, Ambig. = ambiguous, NK = natural killer.

The RISK study is an independent study that examined molecular correlates of disease complications in an inception cohort of CD using bulk RNA-seq from ileum biopsies [53]. The authors observed enrichment of extracellular matrix (ECM) remodeling in donors with stricturing (B2) disease and enrichment of pro-inflammatory pathways in donors with internal penetrating (B3) disease [53]. Based on these previous findings, we hypothesized that donors within $G1_I$ and $G2_I$ were exhibiting gene expression bias towards B2 or B3 disease signatures. Since this is an inception cohort, only 4 of the 23 individuals with ileal profiles already showed signs of severe disease (Supplemental Table 2). Whereas 12 of the 19 B1 cases (63%) were $G1_I$, 2 of the 3 B3 were $G2_I$. The sample is too small to draw any inferences about statistical significance, but the trend is consistent with a tendency for $G2_I$ to progress to penetrating disease.

To test our hypothesis, the B2 and B3 enriched pathways were used to determine if donors from $G1_I$ or $G2_I$ were biased towards the RISK cohort B2 or B3 signatures (methods). The mean gene expression for donors in $G1_I$ and $G2_I$ were calculated for genes in pathways identified as enriched in B2 or B3 samples from the RISK study. This showed that donors in $G1_I$ were biased towards the B2 signature (pathways above the dashed line), whereas donors in $G2_I$ were biased towards the B3 signature (pathways below the dashed line, Figure 8A). Supplemental Figure 6D and E show gene expression patterns of donors in each group for the top genes in the metallopeptidase activity pathway and neutrophil chemotaxis pathways. The "collagen binding" and "ECM disassembly" pathways were the only pathways not enriched in the expected group, $G1_I$, but the enrichment in $G2_I$ was slight. The collagen-binding integrin $\alpha_1\beta_1$ has also been associated with inflammation and

promoting monocyte activation in DSS-induced colitis mouse model, suggesting a pro-inflammatory role which is aligned with B3 bias in $G2_I$ donors [176].

Next, we were interested in the cell types driving the B2 and B3 signatures in $G1_I$ and $G2_I$. UCell [171] was used to compute gene module scores from genes in the metallopeptidase activity pathway and chemokine receptor binding pathway (methods). A metallopeptidase activity module score was elevated in enterocytes (Figure 8B). Mesenchymal cell signaling is partly responsible for the careful balance between ECM and epithelial cells, affecting intestinal permeability and fibrotic mechanisms which can promote progression to stricturing disease [106, 177, 178]. These results further support the importance of epithelial cells in initial disease manifestation and interactions that could lead to severe outcomes of CD [89, 96, 106, 179, 180]. Chemokine receptor binding module was elevated in pro-inflammatory monocytes (Figure 8C). The proportion of monocytes and pro-inflammatory monocytes were significantly higher in $G2_I$ donors (methods, FDR adjusted p value < 0.05, pro-inflammatory monocytes Student's t= -3.0, monocytes Student's t= -2.8, Figure 6A). Monocytosis, the increase in circulating monocytes, has been previously associated with penetrating disease complications, providing further evidence for B3 complications in $G2_I$ donors [181]. Together, tensor decomposition can stratify donors exhibiting disease progression signatures before therapeutic intervention.

3.3.6 Myeloid cell activation accompanied by inflammation stratifies colon samples

Similar to the ileum analysis, pseudobulked differential gene expression and pathway analysis was performed to understand underlying differences among colon $G1_C$

and $G2_C$ donors. There were 10 DEGs upregulated in $G1_C$ and 253 DEGs upregulated in $G2_C$ myeloid cells (pseudobulked gene expression of monocytes, macrophages, and mast cells per sample) (Figure 9A). DEGs upregulated in $G1_C$ were mainly involved in cellular homeostasis. Pathways enriched in myeloid cells of $G2_C$ donors include regulation of inflammatory response, myeloid cell differentiation, response to IFN-γ, and response to microbes (Figure 9B). We then hypothesized that myeloid cells are driving inflammation within $G2_C$ donors.

Gene module scores using genes enriched in the macrophage activation pathway corroborated the differential gene expression results in $G2_C$ donors (Supplemental Figure 7A). The module scores were highest in $G2_C$ macrophages/monocytes, and, in general, immune cells, and lower in epithelial cells. Notably, there was variation in module score within each group with donor 16 in $G2_C$ having the highest mean module score (Supplemental Figure 7A).

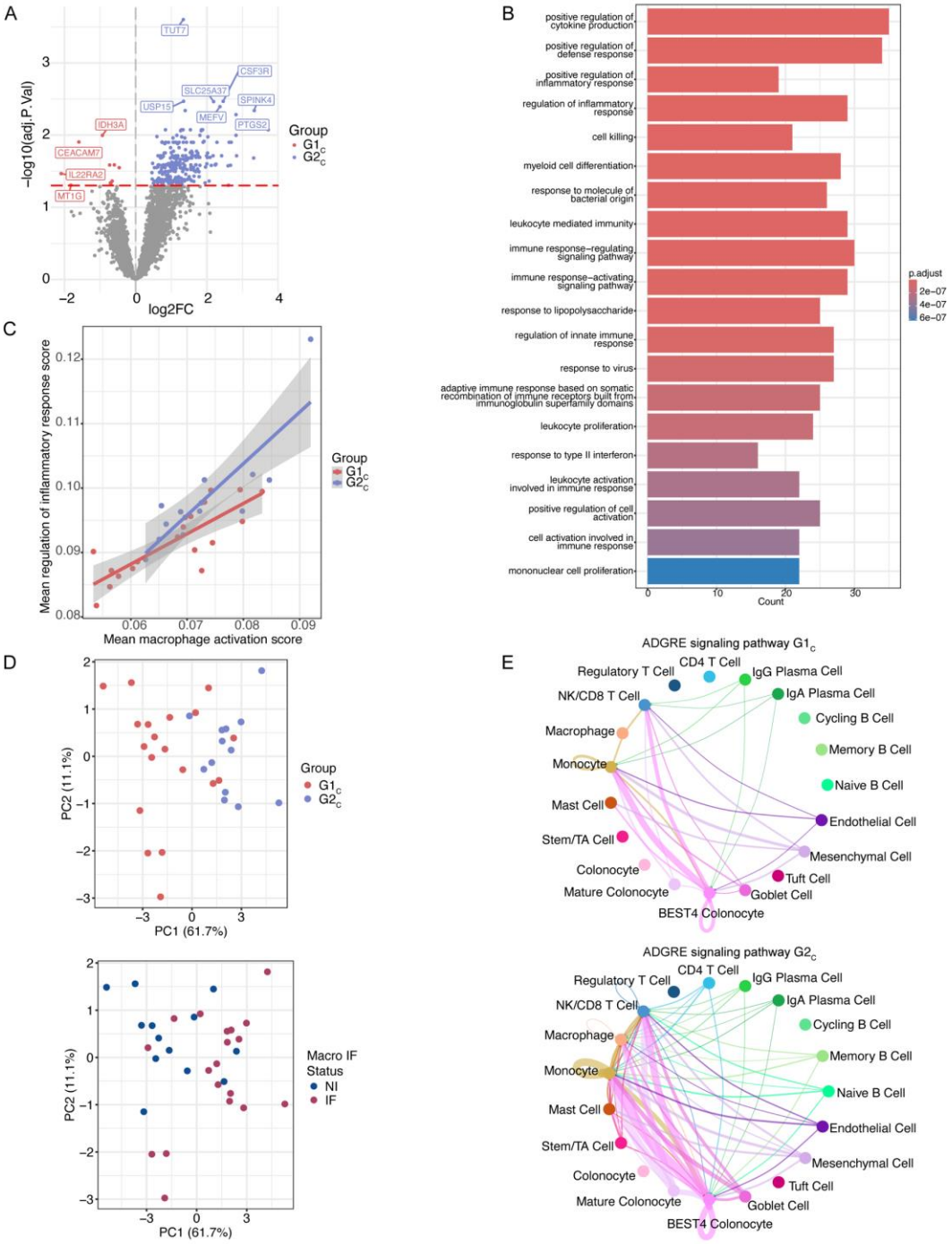


Figure 9 Myeloid cell activation accompanied by inflammation stratifies colon samples. (A) Volcano plot of DEGs identified in the colon group 1 and 2 myeloid cells. (B) Top 20 pathways enriched in myeloid cells of group 2 donors. P.adjust = Bonferroni adjusted p value. (C) Relationship between mean macrophage activation module score and mean response to inflammation module score stratified by group. (D) PCA of DEGs enriched in macrophage activation GO pathway in the myeloid cells colored by group (top) and colored by macroscopic inflammation status

(bottom). Macro IF status = macroscopic inflammation status, IF = inflamed, NI = non-inflamed, PC = principal component. (E) Circle plot of ADGRE signaling pathway in group 1 (top) and 2 (bottom) donors highlights increased engagement of immune of epithelial cells in group 2 donors. Edges are proportional to interaction strength. $G1_C$ = group 1, $G2_C$ = group 2, NK = natural killer, TA = transit amplifying.

We sought to test whether macrophage activation can predict inflammatory response. A module score using genes from the regulation of inflammatory response pathway was calculated (methods). The mean score of macrophage activation across all cells per donor was regressed against the mean score of regulation of inflammatory response across all cells per donor with group interaction as a covariate. While the relationship between macrophage activation and regulation of inflammatory response between groups was marginally significant (p value = 0.051), both groups exhibited the same trend suggesting macrophage activation is contributing to inflammatory response. Patient differences between groups were highlighted, specifically for outlier donor 16 in $G2_C$ (Figure 9C).

To confirm that donors were stratified based on macrophage cell activation, PCA using DEGs involved in the macrophage activation pathway was performed on the pseudobulked normalized gene counts from the myeloid cells. PC1 was significantly associated with group status (Student's t test = -6.2, p value < 0.01) and inflammation status (Student's t test = 3.2, p value < 0.01 Macro IF and Student's t test = 4.4 p value < 0.01 Micro IF), confirming group stratification (Figure 9D). These results were recapitulated using the normalized mean gene counts across all cells identified in the colon (Supplemental Figure 7B).

Cell-cell communication analysis using CellChat [78] validated the pro-inflammatory signature of $G2_C$ samples (methods, Supplemental Figure 7C). Pro-inflammatory effects of myeloid cells could be partly driven by the cellular signaling pathway, ADGRE, which is involved in angiogenesis, tumor progression, and leukocyte recruitment to inflammation [182, 183]. This signaling pathway is enriched in $G2_C$ donors with strong engagement between T cells, monocytes, and epithelial cells (Figure 9E). These results suggest macrophage activation is accompanying inflammation in the colon of $G2_C$ donors.

3.3.7 Inflammation is associated with interferon gamma in group one donors in the rectum

Pseudobulk differential gene expression and pathway analysis was also performed on the rectal compartment (methods). There were 27 DEGs upregulated in $G1_R$ colonocytes and just 3 DEGs upregulated in $G2_R$ colonocytes (Figure 10A). The genes upregulated in $G2_R$ colonocytes are involved in colonocyte cellular homeostasis. The DEGs upregulated in $G1_R$ colonocytes are enriched for pro-inflammatory pathways such as antigen presentation, response to virus, and response to IFN- γ pathways [184] (Figure 10B). Similar results were observed in the rectal stem cells. There were 42 DEGs upregulated in $G1_R$ stem cells and 2 DEGs upregulated in $G2_R$ stem cells (Figure 10C). Pathways enriched in $G1_R$ donors include antigen presentation and response to IFN- γ (Figure 10D). Pro-inflammatory pathways were also enriched in $G1_R$ myeloid cells, including the interferon-mediated signaling pathway (Supplemental Figure 8A). These results suggest donors in $G1_R$ are exhibiting pro-inflammatory response, potentially perpetuated by IFN- γ.

Cell-cell communication analysis was performed to further elucidate the role of inflammation in $G1_R$ donors. Supplemental Figure 8B shows pathways enriched in $G1_R$ and $G2_R$ donors with TNF and TNF related pathways (LIGHT, LT) elevated in $G2_R$. TNF signaling was mediated by ILC3 to target T cells, monocytes, mature colonocytes, mesenchymal cells, and endothelial cells (Supplemental Figure 8C). Based on these initial findings, we questioned whether $G1_R$ donors may be biased towards IFN-γ mediated inflammation and $G2_R$ donors biased towards TNF mediated inflammation.

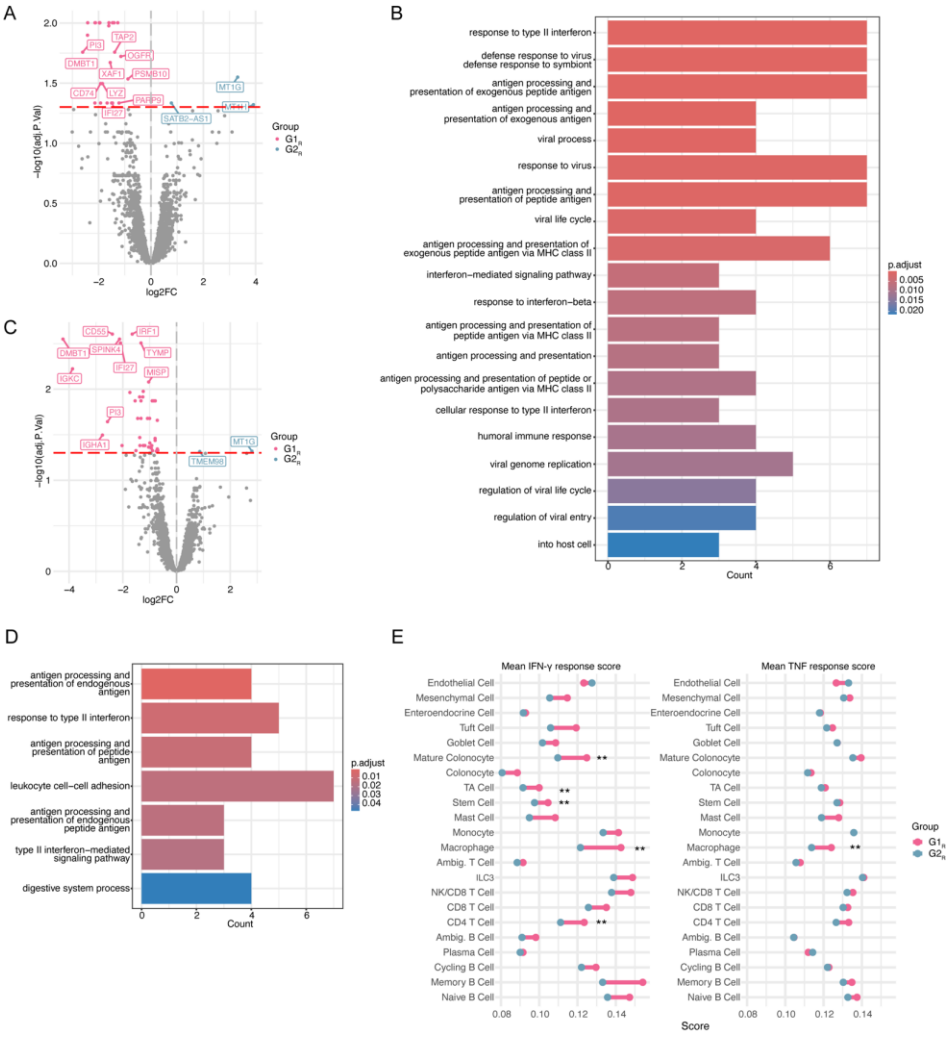


Figure 10 Inflammation driven by interferon gamma is enriched in the rectum of group 1 donors. (A) Volcano plot of DEGs identified in group 1 and 2 colonocytes. (B)

Top 20 pathways enriched in colonocytes of group 1 donors. P.adjust = Bonferroni adjusted p value. (C) Volcano plot of DEGs identified in group 1 and 2 stem cells. (D) Pathways enriched in stem cells of group 1 donors. (E) (left) Mean IFN-γ module score of groups 1 and 2 donors per cell type. (right) Mean TNF response module score of groups 1 and 2 donors per cell type. ** indicates FDR p value < 0.05 from Student's t test. IFN = interferon, TA = transit amplifying, Ambig. = ambiguous, ILC3 = type 3 innate lymphoid cells, NK = natural killer.

To test for IFN-γ vs. TNF inflammation bias, module scores using genes in response to IFN-γ or response to TNF were generated (methods). Memory B cells had the highest mean response to IFN-γ module score in $G1_R$ donors (Figure 10E, Supplemental Figure 8D). There was also donor variation in the mean module score for response to IFN-γ (Supplemental Figure 8D). ILC3 had the highest mean TNF response module score in $G2_R$ donors (Figure 10E, Supplemental Figure 8E). There was less variation in TNF module score per donor, perhaps reflecting low-grade inflammation and disease activity that precedes inflammation in non-inflamed tissue pretreatment [106, 185]. Directly comparing mean module score for each group per cell type, the IFN-γ response score was significantly higher (FDR adjusted p value < 0.05) in mature colonocytes (Student's t = 3.5), TA cells (Student's t = 3.6), Stem cells (Student's t = 3.7), macrophages (Student's t = 4.5), and $CD4^+$ T (Student's t = 3.1) cells in $G1_R$ donors (Figure 10E). The mean TNF response module score was unexpectedly significantly higher in macrophages (Student's t = 3.94, FDR adjusted p value = 0.01) from $G1_R$ donors (Figure 10E). These results suggest that the overall inflammatory response is elevated in $G1_R$ donors and is driven, in part, by IFN-γ engaging immune and epithelial cells in the rectum.

3.3.8 Myeloid cells and T cells are associated with Crohn's disease across tissue

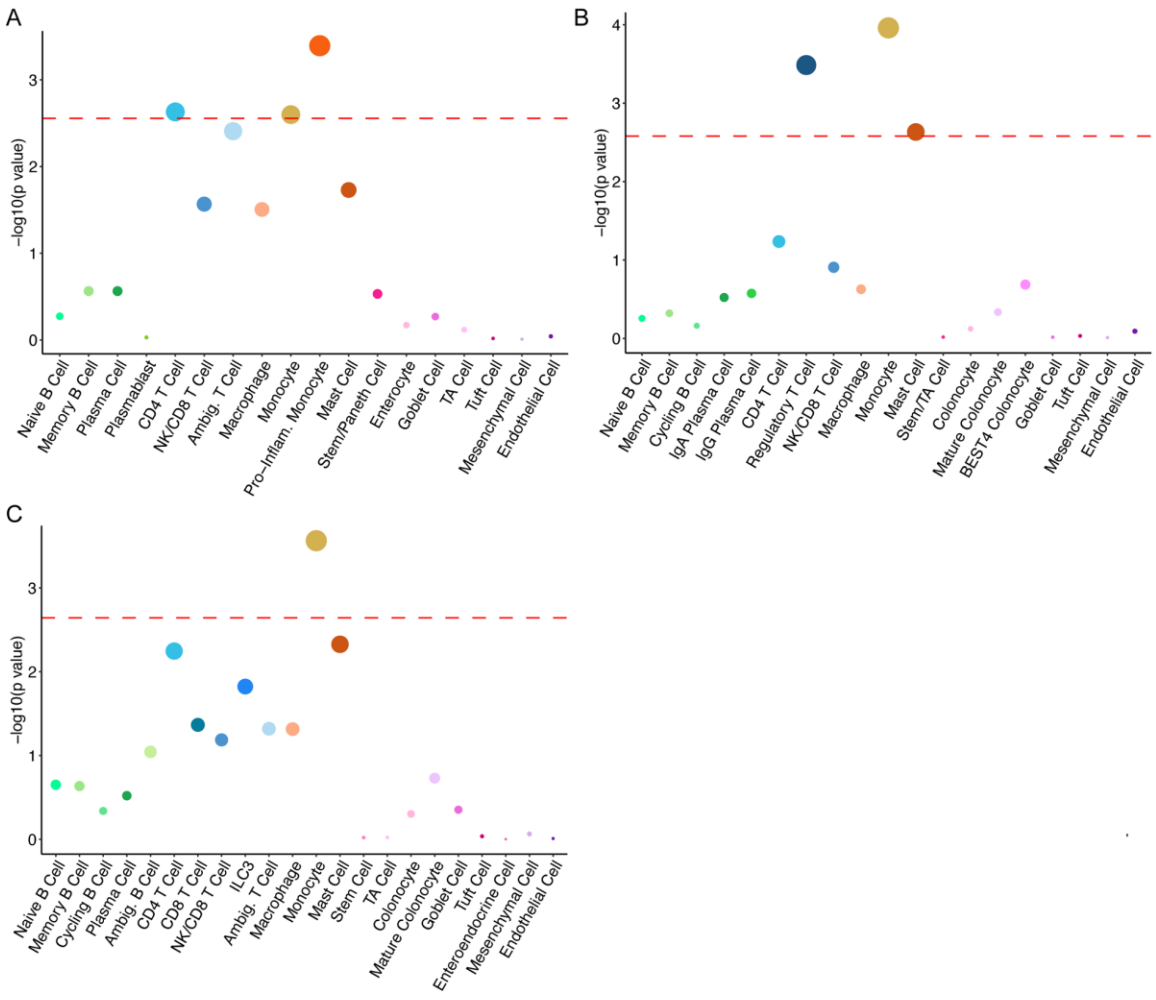


Figure 11 Cell types associated with CD. (A-C) Cell types significantly associated with CD in the ileum (A), colon (B), and rectum (C). Point size represents the Bonferroni-corrected p-value. Dashed red line indicates Bonferroni-corrected p-value. Ambig. = ambiguous, pro-inflam. = pro-inflammatory, NK = natural killer, TA = transit amplifying.

After identifying enrichment of different cell types within groups across the small and large intestine, we next were interested in identifying cell types associated with CD, linking GWAS and scRNA-seq data. Briefly, average gene expression per cell type for each tissue was quantified and the fraction of a gene's expression across all cells was calculated, referred to as the specificity score. MAGMA was used to annotate SNPs from Lui et al. GWAS summary statistics and identify genes associated with CD on the

assumption that genes located in the vicinity of lead variants tend to be causal, though recognizing that this assumption is often violated. [174]. Next, the specificity scores and genes associated with CD were used to perform gene property analysis, identifying cell types associated with CD, following a recently described pipeline for schizophrenia [172]. Conditional analysis was performed to refine cell types most likely acting independent of other cell types in CD pathogenesis.

Within the ileum, pro-inflammatory monocytes and $CD4^+$ T cells were significantly associated with CD after Bonferroni correction and conditional analysis (Figure 11A). Pro-inflammatory monocytes and regulatory T cells were consistently associated with CD in both the high-resolution stably assigned ileum dataset and the reference data. Within the colon, monocytes, regulatory T cells, and mast cells were significantly associated with CD after Bonferroni correction and conditional analysis (Figure 11B). Lastly, in the rectum, monocytes remained significantly associated with CD after Bonferroni correction (Figure 11C). These findings were replicated when all three tissue types were combined, as regulatory T cells and monocytes were associated with CD after Bonferroni correction and conditional analysis. These findings highlight the role of myeloid cells and T cells in CD across tissue, while further validating the cell types driving group status differences in the previous analyses.

3.4 Discussion

Single cell genomics analysis is predicated on consistent assignment of cells to clusters, a process that is both to some degree stochastic and guided by researcher preferences regarding the level of resolution. The inherent sparsity of single cell data often

leads to inconsistent cluster assignments due to statistical uncertainty, which underscores the importance of analytical approaches that evaluate reproducibility. For this reason, we decided to focus our analysis on cells that consistently cluster together at a commonly agreed upon relatively low level of resolution. Unstably assigned cells were defined as cells that do not repeatedly cluster together. Since they were generally identified at the junction of two clusters, they could be transitioning cells, or in some cases, low quality cells, but many are normal members of the cluster that are misassigned simply for statistical reasons. Higher resolution clustering, for example based on as few as a dozen marker genes, may identify novel cell types or states in these regions, but they tend to have lower reproducibility. The stability of clusters should in our view be reported as a standard component of single cell analysis [82], much as bootstrap intervals support phylogenetic analysis.

The clustering stability assessment was conducted to enhance rigor of data analysis, motivated by the of lack of repeatability reported in many scRNA-seq studies. Previous studies have bench-marked different clustering algorithms and batch effect correction methods; however, investigation of the methods after permutation is warranted. Here, we demonstrated that the stability assessment workflow yielded similar and repeatable downstream results based on the JI of cluster assignment and Spearman correlation of differential gene expression results across different sets of the same data (Figure 5). The stably assigned cells are most likely a more representative and reliable transcriptomic measure of the cell type(s) identified in the data. The findings from the stability assessment may be more broadly applicable across different datasets.

After performing the clustering stability assessment for the ileum, colon, and rectum data, we assessed the relationship between cell type proportion and inflammation status. Analysis of each site was performed separately because after joint clustering of the data we observed that epithelial, B cell, and plasma cell compartments in the ileum were very different from these compartments in the large intestine. Previous studies have identified shifts in immune, epithelial, and mesenchymal cell populations with respect to inflammation [97, 98], yet major differences in cell type composition regarding inflamed or non-inflamed tissue were not observed in the ileum or rectum, reflecting high donor variability within this inception cohort [186]. Like previous studies, we also observed a pronounced transcriptional response in the colonic data compared to the ileal data [97]. Monocytes were significantly associated with inflammation status within the colon, underscoring the role of monocytes in inflammation [187]. To further investigate cellular proportion heterogeneity, the donors were hierarchically clustered into groups. One group had over-representation of epithelial cells, and one group had over-representation of immune cells across each tissue (Figure 6, Supplemental Figure 5). We also observed an outlier group in the ileum data and outlier donor in the colon data comprised of mostly immune cells. Longitudinal data collection from the same donors could help decipher whether this pattern is biologically relevant or due to sampling bias.

Since inflamed tissue was not exhibiting obvious disease patterns, we sought to understand interindividual variation by performing tensor decomposition with scITD. This analysis stratified donors into clinically meaningful groups for each tissue, which has important implications for personalized medicine. Within the pro-inflammatory groups, ileum $G2_I$, colon $G2_C$, and rectum $G1_R$, there was one donor that was present in all three

of these groups. There was considerable overlap between group membership in colon $G2_C$ and rectum $G1_R$, and some overlap between ileum $G2_I$ and rectum $G1_R$ (Supplemental Figure 9A). The other groups, ileum $G1_I$, colon $G1_C$, and rectum $G2_R$ had three donors that were present in all three of these groups. Again, there was strong overlap in group membership between colon $G1_C$ and rectum $G2_R$, and some overlap between ileum $G1_I$ and colon $G1_C$, as well as ileum $G1_I$ and rectum $G2_R$ (Supplemental Figure 9B). One possible explanation for decreased overlap in group membership between ileum vs. colon and ileum vs. rectum could be due to tissue specific differences. Additionally, there was some donors that only contributed colon and rectum samples which could also explain group membership differences.

The ileum data was separated into two factors after Tucker tensor decomposition. Focusing on factor 2, preliminary analysis indicated $G1_I$ donors had an enrichment of pro-fibrotic and ECM related pathways whereas $G2_I$ donors had an enrichment of pro-inflammatory pathways (Supplemental Figure 4). These signatures have been previously associated with B2 vs. B3 bias in the RISK study, which we validated at the single-cell level. Tindle et al. also observed similar subtypes of CD, they termed immune deficient infectious CD, and senescence and stress induced fibrotic CD, within their organoid biobank [188]. These subtypes reflect the ileal specific disease signatures identified in this cohort. Further, enterocytes and mesenchymal cells were found to drive the B2 disease bias signal in $G1_I$ donors (Figure 8B). Fibroblasts have been identified as having a main role in stricturing disease, confirming our findings [106]. Epithelial cells may promote fibrosis by secreting cytokines and growth factors that activate fibroblasts and ECM deposition [189]. Pro-inflammatory monocytes were driving the B3 signal in $G2_I$ donors (Figure 8C).

Penetrating disease has been previously associated with increased circulating monocytes and up-regulation of pro-inflammatory pathways.

Our findings suggest single cell profiling might be used as a technique to reveal B2 or B3 bias in the ileum and hence to help guide therapeutic intervention. However, we emphasize that this clinical interpretation must be replicated in a much larger study and that a major limitation of the present study is that only four of the patients with ileal RNA-seq data had progressed to severe disease. We have initiated longitudinal profiling and over the next several years hope to report on whether inception, or subsequent, biopsies are associated with progression to stricturing (B2) or penetrating (B3) complications. The advantage of single cell profiling may lie in increased accuracy and ability to identify relevant cell types relative to bulk RNA-seq, as well as to define categorical profile groups as opposed to a cutoff of a continuous expression signature. While the preliminary data is consistent with B3 individuals more likely being $G2_I$, since the sole B2 donor was also in this group, it is clear that any classifier will not be completely predictive of course of disease. However, the level of accuracy of any genomic score needed for clinical implementation remains to be seen, particularly if the score is considered alongside other markers such as histology, with the intent to guide more likely intervention strategies. Additionally, the RISK study demonstrated that individuals who progressed to B2 disease were not TNF-responsive, suggesting the need for alternative therapeutic strategies. For example, anti-fibrotics (in development but not yet available for CD subjects) may benefit individuals with B2 disease complications, while anti-inflammatories (anti-TNF, anti-IL23, etc.) could help individuals with B3 disease complications, ultimately improving

patient outcomes. Clinical implementation should benefit from increased knowledge of heterogeneity underlying cellular and molecular mechanisms in patient subtypes.

We also investigated underlying disease mechanisms in the colon data. Tensor decomposition was applied to the colon data, separating the data into four factors. Donors were sorted into two groups based on factor 2 scores (Figure 7B). $G2_C$ donors in the colon had enrichment of macrophage activation in addition to upregulation of pro-inflammatory pathways and microbial interactions in the myeloid cells. $G1_C$ donors also exhibited a pro-inflammatory response, albeit at a lower level for both donor and cell type (Figure 9). Garrido-Trigo et al. examined the heterogeneity of macrophage and neutrophils, arguing myeloid cell plasticity was shaped by patient inflammatory microenvironments [187]. Here, we saw heterogeneity within myeloid cell populations across inflamed and non-inflamed colonic tissue, with elevated macrophage activation in a subset of donors. This process may be mediated by ADGRE pathway in dysfunctional myeloid cells leading to tissue repair failure [190]. Macrophage plasticity, influenced by inflammatory environment, could lead to alternative disease progression mechanisms with divergent therapeutic outcomes in these subsets. These results further support the idea that donors with high macrophage and inflammatory activity may benefit from anti-inflammatory therapies compared to donors with lower macrophage activity.

Examining the modes of disease progression in the rectum, the data was separated into four factors based on tensor decomposition. Two groups of donors in the rectum were derived from factor 1 scores (Figure 7C). $G1_R$ donors had an elevated pro-inflammatory signature, partly mediated by IFN-γ which has been shown to affect vasculature through disruption of VE-cadherin in DSS-induced colitis in a mouse model and confirmed in IBD

patients, providing experimental evidence of the IFN-γ contribution [154]. $G2_R$ donors initially displayed a TNF signature enriched in ILC3s, with upregulation of TNF, LIGHT, LT pathways (Supplemental Figure 8B), however, this signature was masked by $G1_R$ donors (Figure 10E). This is most likely due to the synergistic effect of IFN-γ and TNF-α in an inflammatory environment [191, 192]. Differences in pro-inflammatory cytokine profiles can lead to alternative modes of disease progression, such as persistent inflammation/healing leading to variable patient outcomes we see in clinical practice today.

Lastly, we integrated CD GWAS summary statistics with scRNA-seq data to identify cell types associated with CD across the intestines (Figure 11). This study identified an association between monocytes and T cells with CD. Although this interpretation is based on MAGMA assumptions that prioritize causal genes as those most proximal to GWAS peaks, and is not without the limitations of excluding potential environmental influences on cell type gene expression [172], it should be noted that previous studies have also identified increased expression of IBD GWAS genes in regulatory T cells, $CD4^+$ T cells, and monocytes in the ileum and colon [97, 193, 194]. Rare variants associated with CD were enriched in genes related to mesenchymal cell function affected by inflammatory signaling [180]. These findings further reinforce myeloid cells and $CD4^+$ T cells as important targets for therapeutic interventions in pediatric CD.

While this study consists of a treatment-naive CD cohort with samples from different tissue locations of the same individuals, the major limitation is that it is too early to tell to what extent the genomic subtypes correlate with or predict disease complication. The distinct disease signatures of disease progression need to be validated as predictive biomarkers by association with stricturing or penetrative complications as some of the

patients experience adverse outcomes over the next few years. In the interim, follow-up samples from the same donors during clinically indicated biopsy might be analyzed to establish whether profiles persist longitudinal or are restricted to the time of diagnosis. Extended profiling of diverse cohorts with similar strategies will also illuminate the mechanisms of elevated risk of progression in particular populations [101].

In conclusion, treatment naïve CD is highly heterogeneous, exhibiting alternative modes of initial disease onset across tissue, even after filtering for stably assigned cells. This has important implications for guiding therapeutic decisions from clinicians. A more targeted, personalized approach with respect to affected tissue and disease profile should be considered when treating patients even at the time of diagnosis to improve their outcomes. Longitudinal data will inform upon patient profile stability and highlight patient specific features across the intestines.

CHAPTER 4. POST-OPERATIVE ILEUM TRANSCRIPTOMICS IMPLICATE SEX-BIASED MECHANISMS IN CROHN'S DISEASE RECURRENCE

4.1 Introduction

Crohn's disease (CD) is one of two major subtypes of inflammatory bowel disease (IBD) and most commonly affects the terminal ileum; its chronic inflammation can affect any part of the intestine, with frequent skip and segmental involvement, compared to the continuous tracts seen in ulcerative colitis [174, 195, 196]. Treatment of CD aims to reduce inflammation and achieve remission through potent biologics, notably anti-TNF monoclonal antibodies. Resection of inflamed regions due to treatment non-response remains common for refractory cases in the biologic era [197-199], which often results in complications such as ileal strictures and fistulae formation. However, recurrence after resection has been reported to be as high as 70% within 6 months of surgery [200-202].

Consequently, there is an unmet need to identify molecular features associated with recurrent disease following colectomy [53, 203]. This Chapter describes my contribution to a study to this end, submitted to the journal Gastroenterology describing RNA-seq analysis of 339 neoterminal ileal biopsies between 1 and 3 timepoints from 268 post-operative CD patients. The main analyses were carried out by Dr. Kyle Gettler in the laboratory of Professor Judy Cho (Mt Sinai School of Medicine, New York), identifying sex-related differences in gene expression related to recurrence, while eQTL mapping was performed by my colleague in the Gibson lab at Georgia Tech, Dr. Sini Nagpal. Here, I

first describe their results as Introduction before presenting my own findings in relation to the contribution of aberrant splicing to recurrent disease possibly through a process akin to rectalization [204].

In a post-operative setting many genes are differentially expressed when comparing recurrent and non-recurrent samples. Males have more pronounced differences in expression than females when comparing recurring to non-recurring samples (7,037 and 545 genes with adjusted p-values < 0.1 after downsampling, respectively). *OSM* is an example of a gene which was significantly more highly expressed in recurrent samples compared to nonrecurrent samples specifically in males (adjusted p-value = 0.015 in males and adjusted p-value = 0.82 in females). Male and female recurrence-related expression differences show high correlation (r = 0.706) and log2FC values are higher for males. Next, the 20 most significantly differentially expressed genes from the male recurrent vs. male non- recurrent analysis were used to calculate PC1 for the female samples. PC1 values for the females significantly separated recurrent and non- recurrent samples, showing that top predictors of recurrence are relevant to both sexes (p-value = 8.85e-6). Overall, anti-TNF treatment did not substantially modulate gene expression after correction for recurrence state, though there are a small number of genes with very significant changes in expression such as TNFAIP6 and CXCL9 which have increased expression in the absence of anti-TNF treatment; TNFAIP6 is induced in response to TNF [205] and can down-regulate TNF effects [206, 207].

Ingenuity pathway analysis (IPA) of genes differentially expressed when comparing recurring and non-recurring samples was used to identify targets of upstream regulators. Known inflammatory pathways such as LPS, TNF, and IFNG response were activated.

Anti-inflammatory pathways such as TGFB1 are also activated, consistent with its role in mediating peripheral immune tolerance [208]. Conversely, HNF4A pathways are inhibited, consistent with prior reports indicating a protective effect of gut-specific HNF4A against DSS colitis [209]. Interestingly, other top pathways included dexamethasone and sex-hormone regulation by beta-estradiol, progesterone receptor (PGR), estrogen receptor 1 (ESR1), and estrogen receptor 2 (ESR2). The androgen receptor was also predicted to be an upstream regulator activated in recurring samples, though less significantly than estrogen receptors in our IPA results. Dihydrotestosterone was also implicated as a highly significant upstream regulator of recurrence genes (p-value = 1.42E-24), and has recently been linked to induction of TNF production in monocytes in vivo [210]. In ileal single cell data from CD patients, we found that genes regulated by dihydrotestosterone are primarily expressed in fibroblasts, stromal cells, and myeloid cells. Comparing recurrence-related log2(fold change) between males and females within genes downstream of top regulators revealed that some of the largest differences occurred in inflammatory pathways like TNF and in genes regulated by dexamethasone. Similarly, HNF4A and ESR1 response had larger fold changes in males. However, genes downstream of ESR2 and PGR had comparable fold change values between males and females.

The sex-related expression differences observed are robust even when different definitions of recurrence are used. If recurrence is instead defined to be occurring in samples with Rutgeerts scores i1+ and only i0 samples considered to be non-recurring males still show more significant expression changes.

Results were replicated using ileal mucosa samples from an Affymetrix GeneChip microarray study on recurrence risk [211]. Non-recurring female and male sample sets

were compared to inflamed recurring samples (28 inflamed recurring vs. 40 non-recurring for each sex). Logistic regression models were used to assess the relationship between gene expression and recurrence status, with anti-TNF treatment and smoking status included as covariates. We found that males again have more significantly differentially expressed genes compared to females (8,277 genes in males compared to 4,783 in females at a p-value threshold < 0.05 and 5,226 genes in males compared to 1,601 in females at a p-value threshold < 0.01). Beta coefficients for each gene are also highly correlated between females and males in the replication data (Pearson correlation coefficient = 0.622) and have a higher magnitude when comparing recurring and non-recurring males (trendline slope of 0.438).

Testing for differential expression between inflamed CD cases and non-IBD controls using ileal RNA-seq data from the RISK pediatric inception cohort [53] gave similar results to the recurrence analyses, with males showing more pronounced expression differences in response to inflammation. When comparing 12 non-IBD control samples to inflamed CD samples collected from males a total of 6,575 genes were differentially expressed with p-values < 0.1. However, when comparing the same number of female samples there were only 3,604 genes differentially expressed at the same threshold. The gene signatures were very consistent, with almost all genes having comparable but reduced log2FC values in females compared to males. This comparison confirms that the sex bias in gene expression observed in our dataset is not specific to post-operative ileal samples.

Cell-specific markers were used to calculate cell-type representative first principal components (PC) in the post-operative ileal bulk dataset. PC1 calculated using the markers for enterocytes (p-value = 1.58E-8) and goblet cell (p-value = 1.36E-5) clusters were best

able to separate recurrent and non- recurrent samples when included in logistic regression models.

Genome-wide, polygenic risk scores (PRS) conferring 3-5 fold increased risk for 3.2-0.2% of the overall population, respectively have been reported [212, 213]. We calculated a Crohn's specific PRS [174] and identified no evidence (p-value = 0.7) for association to disease recurrence.

To assess whether the cis-regulation of gene expression is conserved between the sexes or modified by recurrence status, cis-eQTL were identified for genes expressed in the ileum. A total of 829 independent eQTL were identified at $p<0.05$, influencing expression of 594 genes. Effect sizes were highly correlated between males and females, but surprisingly 212 of the eQTL were in opposite directions, of which 138 had a significant sex-by-genotype interaction effect at $p<0.001$. This discordance was not observed in the RISK study of ileal biopsies and appears to be exacerbated in the post-operative ileal samples.

To confirm this, the magnitude and sign of the eQTL effects was tested for equivalence between non-recurrent and recurrent samples. Restricting the analysis to the eQTL associated with 81 genes differentially expressed in the recurrent cases indicated further enrichment for opposing sex effects, as 40 genes had the same direction of eQTL in males and females and 41 had opposite signs. The eQTL with consistent direction of effect included several with immune functions (notable genes including *C6*, *TNFSF13*, *CLCN1*, *SETD9*) whereas several opposite direction eQTL were related to the extracellular matrix. The strongest example is the matrix metalloproteinase *MMP3*, expression of which

is elevated for the major allele at chr11:10279772 in recurrent males, but for the minor allele in recurrent females, whereas there is no effect of genotype on expression in the non-recurrent samples. Two-thirds (27/41) of the opposite-sex effect eQTL showed this pattern, sex bias in gene expression is driven by disease recurrence. In contrast, only 11/40 of the same-direction sex effects differed in recurrent samples and just six genes showed elevated directional responses that were consistent in both sexes, one example being *C6*. Across all eQTLs with opposite sex effects, 25 of 41 (61%) transcripts are upregulated and for the largest effect eQTL with opposite sex effects, 19 of 23 (83%) transcripts are upregulated in recurrence group.

While the previous findings characterized recurring disease mechanisms at the genetic and transcriptomic level, additional mechanisms like alternative splicing (AS) may be involved in promoting recurring disease post-surgical intervention. Dysregulation of AS has been previously implicated across different conditions including CD, characterized by microbiome dysbiosis, epithelial barrier breakdown, inflammation, and promotion of fibrosis [45, 60, 214]; however, mechanisms associated with recurring disease are less well understood. Distinct tissue specific AS events have also been described by Berger et al. by identifying a "splicopathy" AS signature in individuals with IBD defined as rectal-like splicing within ileum samples [67]. Evidence from other studies highlighted AS products as actionable therapeutic targets for intervention [62, 215]. Global regulators of splicing have been implicated in IBD pathology. Mata-Garrido et al. found that decreased expression of HP1γ lead to increased splicing noise and was associated with ulcerative colitis [64]. Based on these previous studies, we hypothesize that signatures of "spliceopathy" and HP1γ dysregulation may be mechanisms promoting recurring CD.

4.2 Materials and Methods

4.2.1 Patient exclusion criteria

Participants were excluded from the study if they underwent ileal resection with ileal–ileal anastomosis leaving an intact ileocaecal valve, a sub-total or near sub-total colonic, resection with temporary or permanent diverting ileostomy, or more than two prior surgeries.

4.2.2 Biopsy preservation and RNA isolation

Endoscopic biopsies were immediately added to RNAlater Stabilization Solution (Invitrogen AM7022) and stored at 4°C overnight or up to 48 hours. Subsequently, samples were transferred to -80°C for long-term storage. Samples were shipped in batches to the University of Minnesota Genomics Center (St. Paul, MN), where RNA isolation was performed using the Qiagen RNEasy Plus Kit (74192). RNA quality and quantity was assessed using either Agilent Bioanalyzer or Tapestation analyzers.

4.2.3 Library preparation and sequencing

All samples were normalized to a concentration of either 500ng or 100ng, where 500ng was not available. Samples with RIN scores < 3 were not included in subsequent steps. Libraries were created using the Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero Plus (20040529) using corresponding protocol (1000000124514 v01 – v02). Library quality was assessed using either Agilent Bioanalyzer or Tapestation analyzers and

additional AMpure bead cleanup (Beckman Coulter A63882) was performed if necessary to remove unwanted small fragments. Libraries were then pooled and run on a MiSeq Nano v2 2x150 PE sequencer to ascertain library normalization. Samples were then run on a NovaSeq S4 2x150 PE v1.0 – v1.5 with a minimum read depth of 70M reads per library with mean quality scores above Q30.

4.2.4 RNA-seq data collection and QC

Ileal RNA samples were ribo-depleted and sent for paired-end 150 base pair sequencing using the NovaSeq S4 platform, resulting in 60-70 million reads per sample (339 samples for 267 unique individuals). Reads were aligned to the human genome (hg38) using STAR [216] and quantified with RSEM [46] using the nf-core Nextflow pipeline [217, 218].

4.2.5 Differential transcript usage analysis

To perform Differential Transcript Usage (DTU) analysis, isoform fractions (IF) were calculated as the normalized isoform counts divided by normalized total gene counts. Isoforms were filtered if the gene is not expressed, >95% of individuals express isoform as the only isoform of this gene, and >95% of individuals do not express this transcript. Differentially used isoform fractions (dIF) were calculated as the mean IF in recurring disease – mean IF in non-recurring disease for each isoform. Isoform fractions considered differentially used if dIF > +0.1 or < -0.1 with FDR adjustment p-value < 0.05 [219]. Five samples were excluded from analysis due to either missing recurrence status information (S44 and S150), or missing smoking status information (S67, S157, and S182).

Raw data (n = 10 ileum, n = 10 rectum from N = 10 individuals) was obtained from GSE158952 [67]. The Nextflow-RNA-seq workflow from nf-core was performed [217, 218]. The same workflow to calculate differential transcript usage was applied. Differentially used isoforms were calculated as the mean IF in rectum – mean IF in ileum for each isoform, identifying the tissue-specific DTUs. The same significance cutoffs were applied. Principal Components Analysis (PCA) in the post-operative ileum dataset was performed with the tissue-specific DTUs using prcomp() in R.

4.2.6 Tissue specific differential gene expression analysis

Genes that had a mean normalized expression greater than 5 counts were retained for downstream analysis. Linear mixed models and an ANOVA test were implemented to identify log2 normalized tissue specific DEGs. Genes were considered differentially expressed if the FDR adjusted p-value < 0.05 and log2FC > 1 or < -1. These tissue specific DEGs were used to perform PCA in the post-operative ileum dataset with prcomp() in R. Five samples in the post-operative ileum dataset were excluded from analysis due to either missing recurrence status information (S44 and S150), or missing smoking status information (S67, S157, and S182).

4.2.7 Tissue associated DEG pathway analysis

The top 100 genes in PC1 and PC4 from tissue-specific PCA analysis were used in pathway analysis. PC1 pathways were discovered with enrichGO() from clusterProfiler R package. Pathways with Bonferroni adjusted p value < 0.05 were retained. PC4 pathways were discovered with ToppGene, and pathways with Bonferroni adjusted p value < 0.05 were retained.

Gene set enrichment analysis (GSEA) was used to evaluate whether ileum vs. rectum DEGs had enrichment of the p53 pathway [220, 221]. Genes in the p53 GSEA pathway were also used to perform PCA in the post-op data at both the gene level and the isoform level.

4.2.8 Stratification of recurrence status

A logistic regression model was implemented to stratify patients into recurring and non-recurring disease. PC2 from DTU analysis, PC1 and PC4 from tissue-specific differential expression analysis, sex, smoking status, age at sampling, race, and batch were predictors in the model. The data was divided into 5 folds, where recurrence status was even in each fold, for k-fold cross validation.

4.3 Results

4.3.1 Cohort phenotype/clinical summary and serial sampling

The clinical features of the study cohort are summarized in Table 2 [129]. RNA-seq data from 339 ileal samples (166 female/173 male) were collected from 268 individuals with CD. Phenotypic variables are comparable to similar studies [222, 223]: the median age of the individuals in the study was 33, 41 (25.0%) samples from females and 55 (31.8%) samples from males demonstrated disease recurrence (Rutgeerts scores i2b+), 67 (40.6%) samples from females and 68 (39.3%) samples from males were collected during current anti-TNF treatment, and 18 (10.8%) females and 14 (8.2%) males were reported to smoke at the time of sample collection (Table 2), lower than were reported in the REMIND cohort [222] study, also of post-operative CD recurrence. Recurrence rates between

females and males are comparable in our study because matched samples were selected for RNA sequencing to improve power.

The 339 samples which passed QC include 237 samples collected at the 1st endoscopy (57 of which were recurring), 95 from the 2nd (37 of which were recurring), and 7 from the 3rd or later endoscopy (3 of which were recurring). In total, 70 individuals had multiple samples collected (65 of which were at the first and second endoscopies). Of these, 14 (7 females and 7 males) recurred between post-operative colonoscopies and 7 (2 females and 5 males) changed from recurring to non-recurring. Six individuals started anti-TNF (one of which recovered from recurrence) and 3 stopped anti-TNF treatment between collected endoscopies. Four of the 14 individuals who experienced recurrence after the first timepoint were on anti-TNF and none changed between the two timepoints.

Table 2 Summary of post-op cohort characteristics. RIN = RNA integrity number.

Sample characteristics (n=339)	Female (n=166) Median (IQR) or n (%)	Male (n=173) Median (IQR) or n (%)
Age (years)	33 (26.25 – 47)	33 (25 – 41)
Time between diagnosis and sample collection (years)	9 (4 – 16.75)	11 (4 – 19)
Recurrence (Rutgeerts scores i2b+)	41 (25.0%)	55 (31.8%)
Use of anti-TNF at time of endoscopy (yes)	67 (40.6%)	68 (39.3%)
Smoking status (yes)	18 (10.8%)	14 (8.2%)
RIN values	7 (5.9 – 7.7)	6.7 (5.7 – 7.6)

4.3.2 Differential splicing and differential gene regulation demonstrate rectalization associated with recurrent ileal disease

Given recent evidence that RNA splicing is extensively misregulated in IBD [60, 214], in part in response to down-regulation of the chromatin and RNA-binding protein HP1γ [64], we next asked whether recurrence was associated with differential splicing. A transition from ileal-like to rectal-like splicing patterns was also described in ileal biopsies from 2 of 20 pediatric CD patients [67]. Tissue-specific DTU in a publicly available dataset containing 10 ileum and 10 rectum samples [67] was assessed using the Nextflow RNA-seq workflow for isoform and gene counts. DTU analysis was performed comparing ileum vs. rectum samples, identifying 107 rectal DTUs and 96 ileal DTUs. PCA of these 203 DTUs in the post-operative ileal dataset revealed a strong batch effect explaining PC1, while PC2 was associated with recurrence status (Figure 12A,B). Further analysis showed that ileal disease recurrence is associated with expression of transcript isoforms more commonly observed in the rectum ("rectalization"). No significant difference between males and females was observed with respect to splicing.

Using a similar analysis, apparent rectalization was recapitulated at the gene expression level (Figure 12C,D) where 383 ileum DEGs and 552 rectum DEGs were identified. Recurrence status was associated with tissue specific DEGs in PC1 and PC4. Pathway analysis of the top 100 genes in PC1 highlighted lipid catabolism and transmembrane trafficking as well as organic ion transport (Figure 12E). Similarly, PC2 of the DTUs was enriched for lipid metabolism and membrane trafficking. The same pathway also appeared in aberrantly spliced transcripts associated with gut pathology in CBX3 knock-out mice that do not express HP1γ. PC4, by contrast, implicates pathways associated

with epithelial morphogenesis and extracellular matrix function (Figure 12F), consistent with the involvement of epithelial and stromal cell populations in CD recurrence.

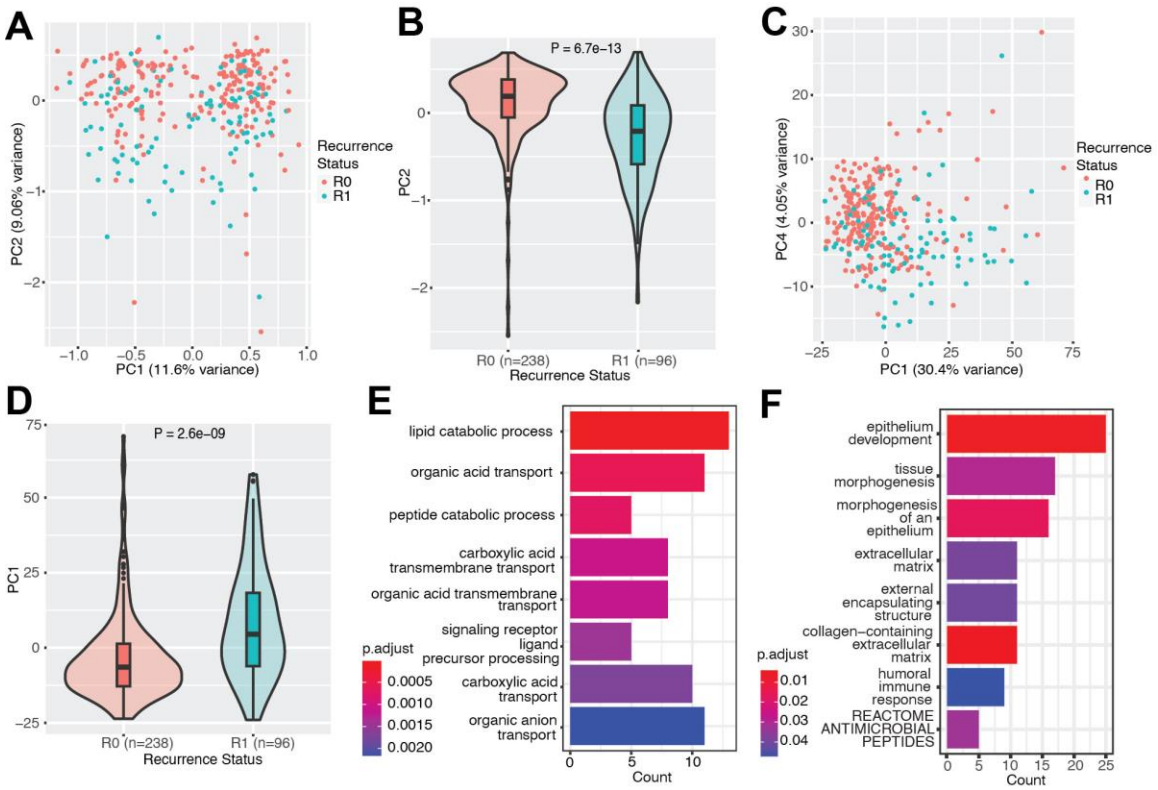


Figure 12 Differential transcript usage and expression indicate that recurrence status is associated with altered splicing biased towards the rectum. (A.) Tissue specific DTUs are associated with recurrence status. (B.) Violin plots of PC2 scores of samples with non-recurring (R0) and recurring (R1) disease show median and interquartile range as boxes, and full kernel density distribution of scores as surrounding shapes. The significant difference is from a Wilcoxon Rank Sum Test. (C.) Tissue specific DEGs are associated with recurrence status. (D.) Violin plot of PC1 scores in non-recurring and recurring disease show recurrence status is associated with tissue biased expression patterns (Wilcoxon Rank Sum Test). (E.) Pathways enriched in top 100 genes in PC1. Shading represents Bonferroni adjusted p-value. (F.) Pathways enriched in top 100 genes in PC4. Shading represents Bonferroni adjusted p-value.

A logistic regression model with PC2 from the DTU analysis, PC1 and PC4 from the differential gene analysis, as well as other controlling variables, had an average AUC

score of 0.75, positive predictive value of 0.7, and negative predictive value of 0.9 after k-fold cross validation where recurrence is the positive class (Supplemental Figure 10).

The epithelial enrichments (Figure 12F) combined with IPA analysis showing the most marked enrichment of the p53 pathway in ileal samples with CD recurrence, motivated PCA analysis of the p53 pathway in the context of the rectalization signature in ileal recurrence. p53 inhibits cell replication and gut-based studies have traditionally focused on its role in colorectal cancer, and p53 pathway genes were enriched in genes differentially expressed in the rectum. PCA analysis of hallmark p53 pathway genes showed that PC2 was able to separate recurrent and non-recurrent ileal samples ($p = 1.1E-7$).

4.4 Discussion

Transcriptomic analysis of post-operative ileal biopsies from 268 CD patients taken six months or more after bowel resection indicates large-scale gene expression remodeling in the approximately one third of individuals who experience recurrent disease. Differential expression is highly correlated between the sexes, but notably more pronounced in males. The consequence of these changes in gene expression appears to be pro-inflammatory, with lipopolysaccharide, tumor necrosis factor, and interferon-γ pathways all activated; induction of regulatory pathways, such as TGFB1 highlights the interplay of pro- and anti-inflammatory pathways in disease progression. Interferon-γ pathway genes, including CXCL9, were also top serum markers of recurrence in the post-operative CD ileal proteome [128]. High OSM expression has previously been linked with failure of infliximab and golimumab treatments [224] and was identified to be specifically

upregulated in male patients but not females after recurrence, which may help to explain why some studies have found that males are more likely to recur and help to improve treatment plans based on patient sex.

The observed sex bias in gene expression might be attributed to sex hormone (dihydrotestosterone, estradiol, estrogen, progesterone, and androgen) activity since genes in these pathways are enriched among recurrence-related genes. In Western populations females have a higher risk of CD than males only after puberty [225], and these pathways have previously been linked to post-pubertal IBD [226, 227]. While our study demonstrated that males are only slightly more likely to experience recurrence (37% versus 30%), a similar European study demonstrated a 2.48-fold increased clinical recurrence rate in males [222]. Interestingly, the 41 genes that show opposite direction effects in the cis-eQTL analysis are as much as two-fold enriched within these pathways. Since these genes are linked to functions such as extracellular matrix deposition, itself implicated in fibrotic disease, whereas cis-regulation of immune activity tends to show the same effects in males and females, epithelial dysregulation is implicated. Furthermore, we see that genes associated with goblet cell and enterocyte cell populations in single cell data are the most strongly associated with recurrence. While sampling of ulcer edges demonstrates the highest transcriptional remodeling [228], development of risk for recurrence scores based on goblet cell and enterocyte transcripts from intact mucosa may be important.

Another aspect of dysregulation is altered splicing involving genes that regulate lipid metabolism and membrane trafficking among other processes. One possibility is that aberrant splicing in stem cells biases epithelial cell differentiation toward fates less commonly observed in the ileum, contributing to increased likelihood of recurrent disease.

Many of these genes are also differentially expressed in recurrent disease samples and have been identified as targets of the chromatin and RNA-binding protein HP1γ which is known to be down-regulated in the presence of pro-inflammatory microbes. This regulator thus emerges as a potential therapeutic target for prevention of recurrence, though we note a limitation of the study is that the biopsies were obtained mostly after onset of recurrent disease.

Given the differences between recurring and non-recurring disease, we sought to evaluate the ability of these gene expression signatures to stratify these patients. A logistic regression model with PC2 from the DTU analysis, and PC1 and PC4 from the differential gene analysis, as well as other controlling variables, had an average AUC score of 0.75, positive predictive value of 0.7, and negative predictive value of 0.9 after k-fold cross validation where recurrence is the positive class (Supplemental Figure 10). This model suggests that patients with recurring vs. non-recurring disease can be stratified based on altered splicing and gene expression factors that also distinguish the ileal and colorectal compartments.

A second potential limitation is that the analysis is of bulk biopsy samples rather than single cell level profiling. However, given the ease of sampling from multiple recruiting sites, combined with the continually lowering costs of sequencing, it may be feasible to develop molecular RNA marker sets more predictive of disease recurrence than current clinical measures. Many extant IBD-based single cell datasets estimate cell-type abundances after cell dissociation. Epithelial cells disproportionately die with dissociation; given the particularly significant contributions of enterocytes and goblet cells in our present

recurrence signatures, we regard bulk RNA-seq analyses as being complementary to single cell approaches.

Finally, polygenic risk scores for CD were not associated with disease recurrence, with well-known clinical factors, like post-operative anti-TNF use being much more impactful. Partitioned polygenic risk scores based on multiomic single cell data may be more relevant [229]. Women are more likely to develop Crohn's disease after puberty [225]; the non-replication of the sex discordant eQTLs in the pediatric onset cohort highlights the complex contributions of age, sex, disease duration and treatment history in linking genetics with bulk RNA-seq data. Some genome-wide association studies of disease progression have been reported, notably in chronic kidney [230] and neurodegenerative diseases, albeit with substantially fewer genetic associations compared to susceptibility trait GWAS. The present study suggests key covariates and pathways associated with CD recurrence, some of which are distinct from disease susceptibility risk factors. A fundamental challenge in Crohn's disease management over time is the frequent loss of response to biologic therapies; response rates for initial biologics often form a ceiling for response to subsequent drugs with different mechanisms of action [231-234]. This may reflect epigenetic and/or stem cell alterations, combined with limitations of pro-inflammatory cytokine blockade via monoclonal antibodies or JAK-level inhibition. This study's novel implication of numerous nuclear hormone pathways in CD recurrence, highlights potential new mechanisms for therapeutic targeting beyond pro-inflammatory cytokine blockade.

CHAPTER 5. CORRELATED MULTI-OMIC SIGNATURES INFORM POTENTIAL CLINICAL STRATIFICATION IN POST-OPERATIVE CROHN'S DISEASE

5.1 Introduction

Crohn's disease (CD) is characterized by chronic transmural inflammation commonly affecting the small and large intestine. While etiology remains unknown, interactions between the immune system, microbiome, and genetic susceptibility most likely play a role in disease manifestation [1, 8]. Different therapeutic strategies are employed to treat and manage disease based on severity, with the primary objective to induce mucosal healing [235, 236]. Disease management strategies include dietary modifications, corticosteroids, antibiotics, immunomodulators, and biologics such as anti-Tumor Necrosis Factor (anti-TNF) therapy [237, 238]. Despite improved outcomes in the post-biologic era, approximately half of all patients require surgical intervention within ten years of diagnosis [239]. Surgery is not curative, and approximately 17%-55% of individuals experience recurring disease, defined as the appearance of new lesions after bowel resection, within five years post-operation (post-op) [129, 240-243]. Risk factors for recurring disease after intestinal resection include smoking, history of CD-related surgery, penetrating disease complications, and younger age [239, 244, 245]. We (Hernandez-Rocha et al. [129]) also identified male sex and non-European ancestry to be associated with increased risk of recurring disease, while anti-TNF use after surgery was associated with decreased risk of recurring disease, in addition to the previous risk factors described. To identify potential biomarkers associated with post-op disease recurrence, a cohort of

individuals with CD was recruited and both serum proteomics and ileal biopsy RNA-sequencing (RNA-seq) analyses were performed.

Various biomarkers and disease scores are used to evaluate disease activity. The Rutgeerts score is commonly utilized to evaluate disease course following ileocolonic resection based on severity of lesions [246, 247]. Usually those with Rutgeerts score > i2 are considered to have recurring disease [246]. A modified Rutgeerts score has been proposed, separating i2 into two groups: i2a characterized by anastomotic lesions and i2b characterized by lesions in the terminal ileum, associated with disease recurrence [248, 249]. Biomarkers such as C-reactive protein (CRP) or fecal calprotectin have been utilized to monitor intestinal inflammation; however, these biomarkers lack specificity to recurring disease and have poor correlation to the Crohn's Disease Activity Index (CDAI) scores [238, 244, 250]. Non-invasive biomarkers to monitor disease recurrence should facilitate optimized disease management. To this end, proteomic and RNA-seq approaches provide an opportunity to improve patient stratification and discover potential biomarkers of disease recurrence.

Inflammatory bowel disease (IBD)-related proteomic studies have been performed to investigate underlying disease pathology, identify potential therapeutic targets and disease biomarkers, and stratify patients. Previous proteomic studies with serum or plasma samples identified IBD subtype biomarkers and found varying classification accuracy of CD vs. Ulcerative colitis (UC) using protein levels which were dependent on primary disease location in individuals with CD [127, 251, 252]. These results further reinforce the importance of identifying tissue-specific therapies. Additional studies have been performed to characterize the association between therapy and specific protein levels in individuals

with IBD. Zwicker et al. (2017) performed a pilot study to investigate proteomic profiles of serum before and after IBD patients were treated with vedolizumab, which binds to $\alpha_4\beta_7$ integrin [253]. These results revealed remodeling of proteomic profiles after therapy and proposed CCL13 as a potential marker for response to vedolizumab [253]. Comparison of inflammatory proteomic profiles before treatment between anti-TNF responders and non-responders showed increased expression of several proteins associated with non-response, in line with increased immune activation playing a role in response to therapy [254]. Investigation of other treatment methods for pediatric IBD, such as exclusive enteral nutrition (EEN), showed that the traditional Type 1 Helper T cell profile associated with CD and Type 17 Helper T cell profile for UC, suggesting alternative disease mechanisms. In this cohort, protein expression data were used to stratify IBD subtypes and predict response to EEN, with a more focused analysis among CD patients suggesting that those with ileal involvement may respond better to EEN than those with colonic CD [255]. Granno et al. were also able to predict development of CD and UC based on protein expression levels [256]. Proteomic studies to investigate disease recurrence in post-op individuals with CD have also been reported. Our group (Walshe et al. [128]) identified strong associations between increased CXCL9 and MMP1 expression and higher Rutgeerts scores in a post-op CD cohort, with elevated CXCL9 and CXCL11 expression having the strongest association with higher Rutgeerts scores within the anti-TNF-treated group. Collectively, these studies indicate that proteomic profiling could serve as a valuable clinical tool for biomarker discovery, disease monitoring, and patient stratification for individuals with IBD.

While important CD-related discoveries have been made with proteomics and transcriptomics alone, incorporating both modalities may improve biomarker recognition and uncover additional pathological mechanisms associated with disease progression [257, 258]. While modest correlation between RNA and protein expression has been observed across tissues [125, 259], Koussounadis et al. also noted increased correlation between differentially expressed genes (DEGs) and protein expression [124]. Preto et al. also performed an integrative analysis of intestinal biopsy RNA-seq data and plasma proteomics from individuals with IBD. In this analysis, they were able to classify CD vs. UC patients and identify subgroups of donors within UC and CD [260]. One of the aims of this work is to perform a comparative analysis between ileal transcript and serum protein expression to gain insight to post-op disease recurrence and how these signatures may be leveraged for patient stratification. The other aim of this study is to investigate potential sex differences associated with recurring disease at the proteomic level since sex-biased differences were observed in recurring disease at the transcriptomic levels [Gettler et al, submitted].

In this work, we identify DEPs and DEGs associated with disease recurrence and anti-TNF use while observing sex-biased protein expression with respect to recurring disease. Furthermore, in a comparative analysis between ileal RNA-seq described by Gettler et al. [261] and described in Chapter 4, and serum protein expression described by Walshe et al [128], hierarchical clustering of correlation patterns revealed two patient subgroups for whom transcript and protein abundance correlated with clinical phenotypes. Lastly, a logistic regression model using the correlated expression patterns was implemented to stratify patients based on recurrence status.

5.2 Materials and Methods

5.2.1 Patient recruitment and data collection

A full description of the patient cohort for this study is described in Hernández-Rocha et al., Walshe et al., and Chapter 4 of this thesis [128, 129]. Briefly, adult patients diagnosed with Crohn's disease were prospectively recruited across Genetic Research Centers (GRCs) in affiliation with the NIDDK Inflammatory Bowel Disease Genetics Consortium (IBDGC). The centers include Icahn School of Medicine at Mount Sinai New York, Cedars-Sinai Medical Center Los Angeles, University of Pittsburgh, Johns Hopkins University, Mount Sinai Hospital Toronto, and University of Montreal. These patients were undergoing ileocaecal or ileocolonic resection, and only patients with confirmed ileal disease were included in the study. Exclusion criteria included ileal-ileal anastomosis with intact ileocaecal valve, colonic resection, temporary or permanent diverting ileostomy, and more than two prior surgeries [128, 129]. This study only included samples from the first endoscopy with known Rutgeerts score documented during endoscopy. Those with Rutgeerts score ≥i2b were considered to have recurring disease while those with Rutgeerts score ≤ i2a were considered to have non-recurring disease [262]. The total number of individuals at first endoscopy used in the study was n = 120 in the proteomics dataset (29 with recurring disease and 91 with non-recurring disease) and n = 235 in the RNA-seq dataset (56 with recurring disease and 179 with non-recurring disease). There were n = 79 individuals (62 with non-recurring disease and 17 with recurring disease) that had both proteomics and transcriptomics data available at first endoscopy.

5.2.2 Sample preparation and RNA sequencing

Serum samples collected on day of colonoscopy were processed with the Olink proteomics inflammation panel (v. 3012) which quantifies abundance of 92 proteins by virtue of DNA barcodes linked to antibodies to each protein. Due to changes in the panel during sample processing, only 91 proteins were considered in downstream analyses. A full description of sample preparation for Olink proteomics is described in Walshe et al. [128].

Ileal biopsies were obtained and preserved at -80°C for long term storage. Library preparation was performed using the Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero Plus (20040529) with protocol 1000000124514 v01 – v02. The cDNA was sequenced on the NovaSeq S4 platform. Reads were aligned to the human genome (hg38) with STAR [216] and quantified with RSEM [46] using the nf-core RNA-seq Nextflow pipeline [217, 218]. A full description of sample and library preparation for RNA-seq data is found in Chapter 4 of this thesis.

For the most direct contrast of gene and protein expression, a subset of 90 matching gene-protein pairs were analyzed only in patients who donated both sample types.

5.2.3 Differential expression analysis

Differential protein expression analysis was performed on a sample of n = 120 subjects with m = 91 proteins considered. A mixed linear regression model was fit using the open source R code *lme4* (v 1.1-35.1) on log2 normalized protein expression (NPX). Fixed effects included recurrence status, anti-TNF use, sex, and sample age (in years at time of assay) with GRC as a random effect. Protein expression was modeled as the outcome variable. P-values obtained from each model were corrected for multiple testing

with Benjamini-Hochberg (BH) false discovery rate (FDR) correction. Proteins significantly associated with a given fixed effect variable (FDR adjusted p-value < 0.05) from the linear regression model were selected for post-hoc testing with the emmeans() function which performs two-sided t-tests on the estimated marginal means of proteins in the data. Comparisons of differential protein expression were performed for recurrence status, anti-TNF use, and sex separately. The same framework was implemented to investigate the interaction between recurrence status and sex. Two models were considered, the first including recurrence status, sample age, sex, the interaction between recurrence status and sex, and GRC as a random effect. The second model included the addition of anti-TNF use with the other fixed and random effects as previously described. The interaction between recurrence status and sex was evaluated using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) applied to the first model. Subsequently, pairwise comparisons of proteins associated with the interaction between recurrence status and sex at FDR adjusted p value < 0.05 was performed with emmeans(). If proteins were significantly associated with one of the comparisons, they were considered differentially expressed.

A similar workflow was used for the differential gene expression analysis for consistency. For the linear regression models per gene, n = 235 subjects and m = 90 genes with log2 normalized transcriptomic data were used. Fixed effects included recurrence status, anti-TNF use, sex and GRC as a random effect. After BH FDR correction, significant genes were considered for post-hoc analysis (two-sided t-tests) with the emmeans() function. Pairwise comparisons of genes associated with a particular variable was performed as previously described. The interaction between recurrence status and sex

was evaluated with recurrence status, sex, and interaction between recurrence status and sex as the fixed effects and GRC as a random effect. Post-hoc analyses were performed with emmeans() and genes with BH FDR corrected p-value < 0.05 were considered differentially expressed.

5.2.4 Batch effect correction

A strong batch effect was identified by performing principal component analysis (PCA), since PC1 was significantly associated with GRC and anti-TNF use for both the proteomic and transcriptomic data (one way ANOVA p value < 0.05). PCA was also performed in the proteomic data (n = 120) for anti-TNF use and no-anti-TNF use separately. PC1 was significantly associated with GRC (one way ANOVA p value < 0.01). This batch effect was corrected for in the proteomics data with ComBat [263] which uses an empirical Bayes framework on normalized data to adjust for known sources of technical variation. Recurrence status was implemented in the model matrix as the outcome of interest. Similarly, the transcriptomic data batch effect was corrected using ComBat-seq [264], which uses the raw count data for batch effect correction. ComBat-seq implements a negative binomial regression model to estimate batch effects [264]. Again, GRC was the batch covariate and recurrence status was used as the outcome of interest. The correlation and group stratification analysis utilized the batch effect corrected data.

5.2.5 Correlation analysis

Spearman correlation was computed to understand coordination between ileal gene expression and serum protein expression. The batch effect corrected, log2-normalized and scaled (across donor) gene and protein expression values were used for this analysis.

Donors with both transcriptomic and proteomic information (n = 79 subjects) and the same gene-protein pairs (m = 90) were included.

Spearman correlation was evaluated between direct gene-protein pairs to identify gene-protein pairs with BH FDR adjusted p value < 0.05. Hierarchical clustering of gene and protein correlations using Euclidean distance and the "complete" clustering method with the pheatmap R package (v. 1.0.12) was implemented to investigate patterns of correlated gene and protein expression. Hierarchical clustering was also performed using correlation distance and "complete" clustering method. Similar clusters of genes and proteins were identified. There were 30 Group A genes, 25 Group B genes, 18 Group 1 proteins, and 20 Group 2 proteins shared between each hierarchical clustering approach. The number of clusters was identified using the within sum of square metric.

This analysis was also performed on the non-batch-effect corrected data with the same methodology as previously described.

5.2.6 Permutation test

To ensure that correlation patterns between gene and protein expression were greater than random expectation, a permutation test was performed on the log2-normalized, batch effect corrected, and scaled data. Protein labels were randomly permuted, and the Spearman correlation was computed for direct gene-protein expression patterns. This process was repeated a total of 10 times. Pseudo p-values were computed as the proportion of permuted correlations as or more extreme than real correlation (Supplementary table 1).

5.2.7 Group comparison

To investigate variables associated with clusters identified from the hierarchical clustering analysis, two separate approaches were implemented. The first approach was to create a "donor score" by computing the average expression of batch corrected, log2 normalized, and scaled (z-score) of genes in group A or B, as well as proteins in group 1 or 2. Next, a Wilcoxon rank sum test performed for each group to test for association between "donor score" and variables of interest (recurrence status, anti-TNF use, sex, and disease behavior). A variable was considered significantly associated with the "donor score" if the BH FDR adj. p value was < 0.05. The second approach was to compare gene or protein expression within each group using a Wilcoxon rank sum test for each of the variables previously tested. P-values were again adjusted using BH FDR correction with p value < 0.05.

To further confirm group stratification based on gene and protein expression correlation patterns, the same groups of genes/proteins were used to compute donor scores for the donors not in the direct transcriptomic-proteomic comparison. Again, log2-normalized, scaled, and batch effect corrected data were used in this analysis. For the proteomics data, n = 41 donors (21 donors with non-recurring disease and 12 donors with recurring disease) were included. For the transcriptomic data, n = 156 donors (117 donors with non-recurring disease and 39 donors with recurring disease) were included. The same variables listed previously were evaluated for association with donor score using the Wilcoxon rank sum test.

5.2.8 Recurrence status stratification

To stratify recurrence status based on groups identified from hierarchical clustering analysis, a 3-fold cross validation logistic regression model was performed. First, PCA was performed on group A and B genes and group 1 and 2 proteins. PCs 1-5 from group A and B genes and group 1 and 2 proteins were used as variables in the logistic regression model. Youden's J index (sensitivity + specificity – 1) was used as the model's accuracy threshold. The positive class was recurring disease.

5.3 Results

5.3.1 Cohort description

A cohort of adult patients (17-75 years old) diagnosed with Crohn's disease (CD) was recruited prospectively across six genetic research centers (GRC) in North America under the auspices of the Inflammatory Bowel Disease Genetics Consortium (IBDGC) network. This cohort includes patients with confirmed ileal disease undergoing ileocaecal or ileocolonic resection (methods). Those with Rutgeerts score ≥i2b were considered to have recurring disease after surgery. Only individuals with data from first colonoscopy were retained for this analysis. Additional patient metadata is described in Chapter 4, Hernández-Rocha et al. and Walshe et al. [128, 129].

Protein abundance in serum samples collected at time of colonoscopy was assessed by barcode sequencing with the Olink proteomics inflammation panel. Ileum pinch biopsies were also collected at time of colonoscopy and subjected to RNA-sequencing (methods) [128, 129]. In total, there were 120 individuals with proteomic information and 235 individuals with transcriptomic information at time-of-first colonoscopy. For the proteomics cohort, there were 29 donors (9 Female and 29 Male) with recurring disease

and 91 donors (41 Female and 50 Male) with non-recurring disease, and 36% of donors were taking anti-TNF at time of colonoscopy (43 donors on anti-TNF). For the transcriptomics cohort, there were 56 donors (21 Female and 35 Male) with recurring disease and 179 donors (88 Females and 91 Males) with non-recurring disease, with 41% of donors on anti-TNF treatment (96 donors on anti-TNF). A subset of these individuals had both proteomic and transcriptomic information, for a total of 79 individuals (non-recurring: 28 Females and 34 Males, recurring: 6 Females and 11 Males) included in a comparative analysis.

5.3.2 Serum protein expression implicates recurring disease sex bias

After observing sex bias with respect to recurring disease at the transcriptomic level (Chapter 4) and some clinical characteristics [129], we were motivated to investigate whether sex bias was also observed at the proteomic level. In order to study the potential interaction between recurring disease and sex, we first established baseline protein expression differences across recurrence status and sex using linear regression models while controlling for anti-TNF use and sample age with GRC as a random effect (methods). Proteins significantly associated with disease recurrence, anti-TNF, or sex, while controlling for other variables after FDR p-value adjustment (adj.) < 0.05 were then considered for post-hoc analysis. A protein was considered differentially expressed if the association between protein expression and the variable of interest had an FDR adj. p-value < 0.05 after post-hoc analysis (methods). Differential protein expression analysis revealed that IL6 and MMP-1 were significantly associated with recurrence status, and unsurprisingly, had elevated expression in the donors with recurring disease (Figure 13A). Both proteins have been previously associated disease activity and increased disease

severity [265-267]. Differentially expressed proteins (DEPs) associated with sex include CXCL5, FDF-21, Ft3l, IL5, OPG, and TRAIL (Figure 13B). Interestingly, TRAIL, part of the TNF family, is the only protein with increased expression in males [268]. TRAIL may have a dual pro-inflammatory and pro-apoptotic role in CD pathophysiology depending on environmental conditions and has been previously associated with upregulation in mononuclear cells during inflammation [269, 270].

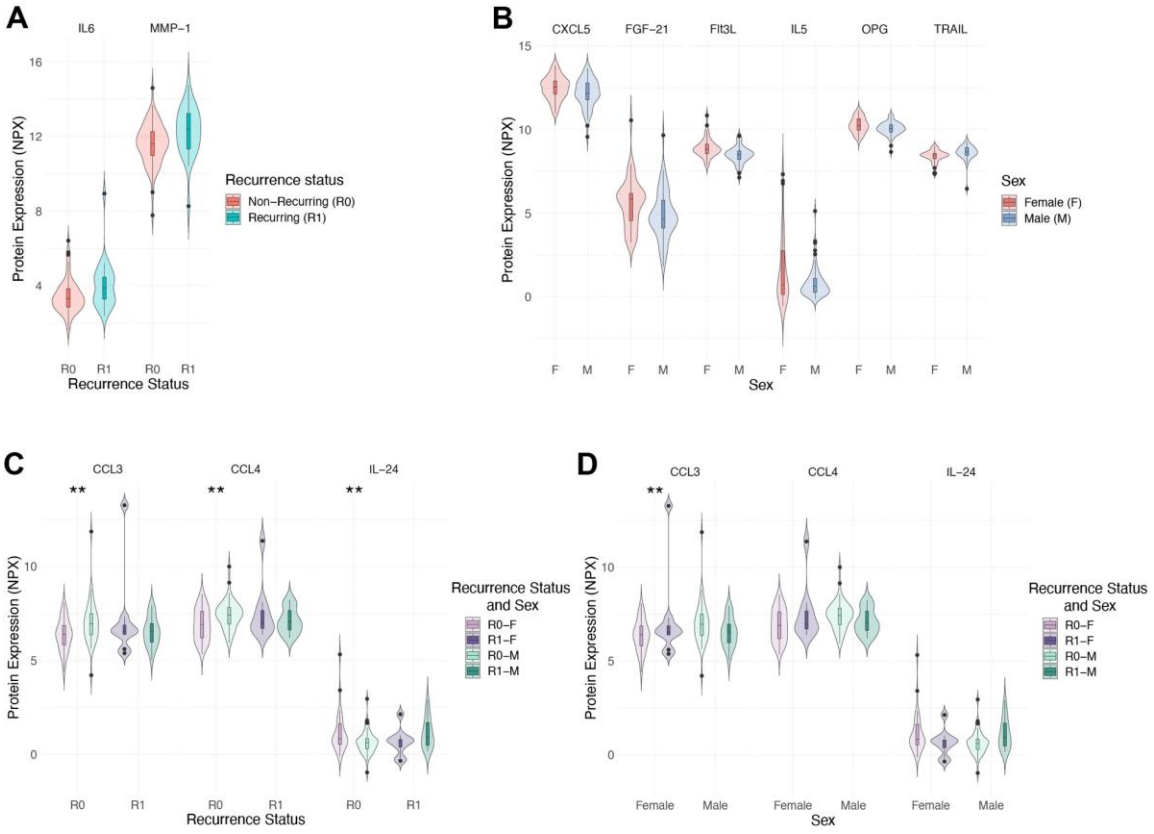


Figure 13 Sex-bias associated with recurring disease at the proteomic level. (A) DEPs associated with recurrence status in n = 120 donors. (B) DEPs associated with sex. (C) DEPs associated with the interaction between recurrence status and sex show expression differences between males and females within non-recurring disease. (D) DEPs associated with the interaction between recurrence status and sex show CCL3 expression difference within females. ** indicate significantly differentially expressed after post-hoc FDR adjusted p value correction. R0 = non-recurring, R1 = recurring, F = female, M = male, NPX = normalized protein expression.

After identifying DEPs associated with recurrence status and sex, we next investigated the interaction between these two variables. Again, linear regression models were implemented to investigate the interaction between recurrence and sex (methods). After post-hoc analysis of the interaction model, DEPs were associated with sex within non-recurring individuals (Figure 13C) and in one case, difference in expression between female recurrence status (Figure 13D). These results further validate sex-bias observed across recurrence status at the proteomic level, though the extent of the bias is much less than characterized in the full ileal RNA-seq data [261] both in terms of number of DEPs and magnitude of the bias.

Since biologic therapies may affect proteomic profiles [253, 254], we next investigated DEPs associated with anti-TNF use. There were seven DEPs associated with anti-TNF use (Supplemental Figure 11A). As expected, each of these proteins had higher expression in individuals who were not taking anti-TNF medication [25, 236]. Notably, CXCL9 was significantly associated with anti-TNF use. This protein has been associated with reduced levels after biologic therapy previously as well as with increased expression in individuals with CD [252, 265]. These findings further support the role of CXCL9 as a clinically important protein in CD pathogenesis.

5.3.3 Ileal gene expression associated with recurring disease and anti-TNF use

To support our proteomic analysis, we also performed a complementary differential gene expression analysis. Similar linear regression models and post-hoc analyses were implemented (methods). There were 22 differentially expressed genes (DEGs) associated with recurrence status after post-hoc analysis (16 DEGs upregulated in recurring disease

and 7 DEGs upregulated in non-recurring disease) (Figure 14A). Interestingly, IL6 and MMP-1 were both DEPs and DEGs associated with recurring disease. *IL10RA* was the only DEG associated with sex (Figure 14B) in this restricted set of genes. It has also been implicated in very early onset IBD [271, 272]. No DEGs were associated with the interaction between recurrence status and sex. Similar results were obtained when identifying DEGs associated with anti-TNF use (Supplemental Figure 11B). Again, *CXCL9* was upregulated in non-anti-TNF users. Additionally, differential expression of *CCL4*, *CXCL10*, and *TNFRSF9* were associated with anti-TNF use in both the proteomic and transcriptomic data (Supplemental Figure 11). The similarity between ileal gene expression and serum protein expression prompted us to investigate the potential correlation between these expression patterns.

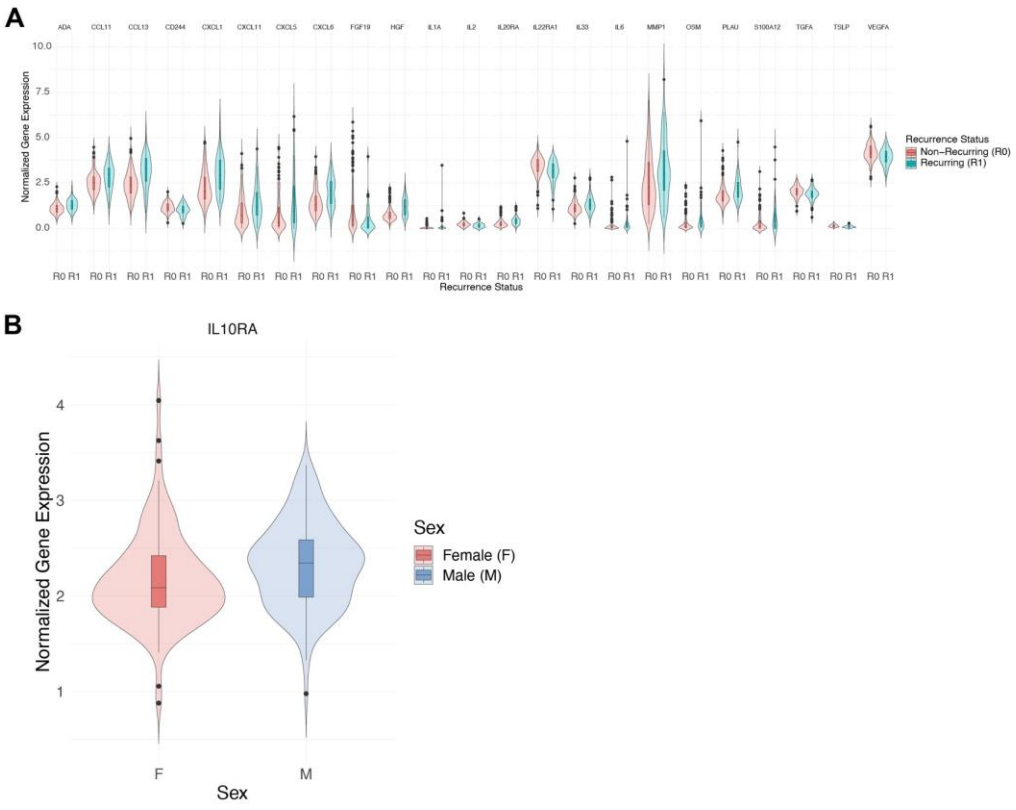


Figure 14 Differentially expressed genes associated with recurrence status and sex. (A) DEGs associated with recurrence status for n = 235 donors. (B) DEG associated with sex. R0 = non-recurring, R1 = recurring, F = female, M = male.

5.3.4 Correlation between ileal gene expression and serum protein expression

To investigate coordinated patterns of expression in the ileum and serum, we analyzed 90 gene-protein pairs from 79 donors with both gene and protein expression data. Prior to correlation analysis, principal component analysis (PCA) was performed which revealed a strong batch effect associated with GRC in both datasets (Supplemental Figure 12A,B). This batch effect in both data modalities was driven by low anti-TNF use in Canadian GRCs (51/69 donors not on anti-TNF in Canada whereas 26/52 donors are not on anti-TNF in the US). PCA performed the proteomic data of anti-TNF users vs. non-anti-TNF users separately continued to show strong batch effect associated with GRC in PC1

(p value < 0.01). ComBat [263] was used to correct for batch effect in the proteomic data whereas ComBat-seq [264] was used to correct for batch effect in the transcriptomic data (methods). Subsequently, the batch effect corrected data were used for the correlation analysis.

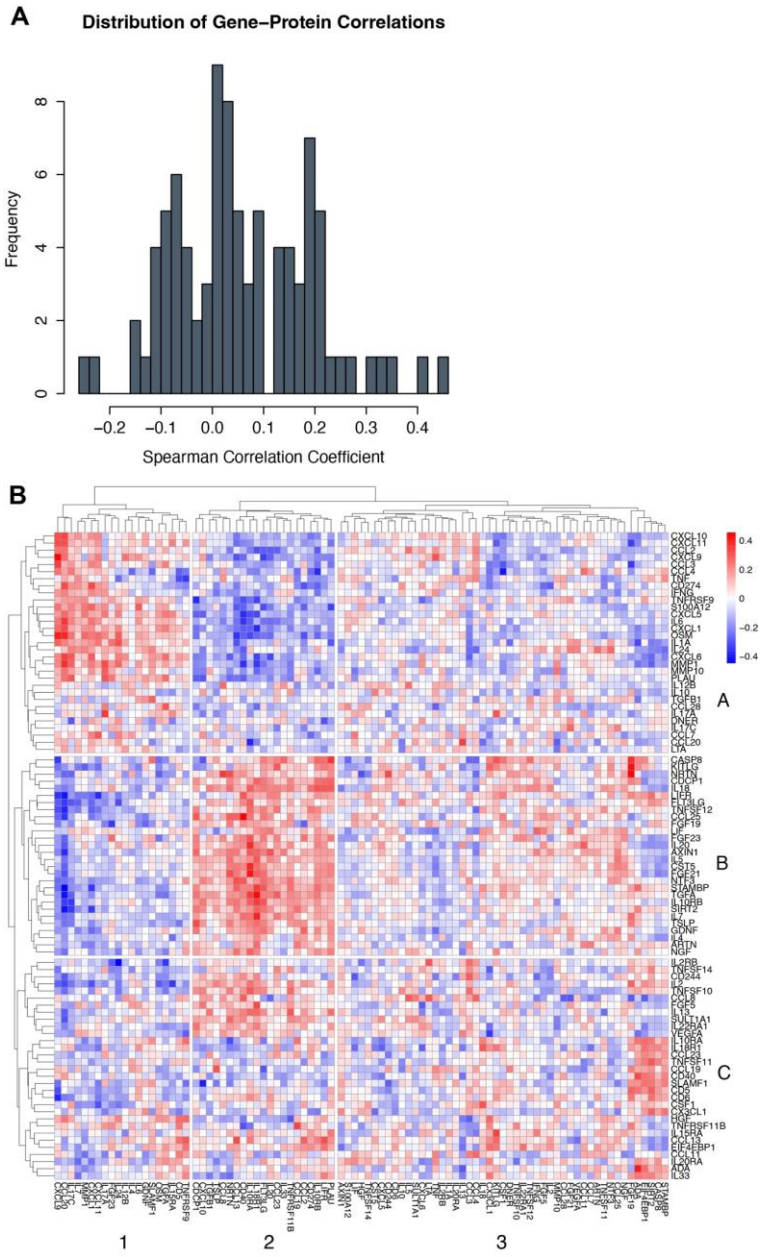


Figure 15 Correlation between ileal gene expression and serum protein expression. (A) Distribution of spearman correlation between each gene-protein pair. (B)

Hierarchical clustering of pairwise comparisons of Spearman correlation values between ileal gene expression and serum protein expression. X-axis: proteins with Group 1-3 labeled, Y axis: genes with Group A-C labeled.

Spearman correlation was first performed for the matching gene-protein pairs (e.g. correlation between *CXCL9* gene and CXCL9 protein, methods). The mean correlation across these pairs was just 0.06, not surprisingly since they are from different tissue sources. Figure 15A shows the distribution of matching gene-protein pair correlations. Many pairs had little correlation, and a few gene-protein pairs had modest correlation around 0.20 [125, 259]. The top correlated gene-protein pairs include ADA, CXCL9, and IL17 (FDR adj. p value < 0.05). To ensure that the observed correlation patterns were greater than random expectation, a permutation analysis was performed (methods). The null hypothesis was that there was not a difference between observed and permuted correlation and the alternate hypothesis was that there was a difference in the correlation distributions Briefly, the protein labels were randomly permuted 10 times, and the gene-protein pair correlation was computed for each permutation. The permuted correlation distribution was qualitatively different from the observed correlation distribution (Supplemental Figure 13A,B) as seen in the rightward shift of the observed distribution relative to all permutation sets, including a shoulder of relatively high positive correlations. 14 gene-protein pairs had a nominally significant pseudo p-value < 0.10, with ADA, CXCL9, and IL17A being three of the 14 gene-protein pairs that remained significantly correlated after the permutation analysis (Supplemental Table 3). Overall, these results nevertheless are consistent with correlated regulation of a minority of the gene-protein pairs across the serum proteome an ileal transcriptome.

To further examine the coordinated gene and protein expression patterns, pairwise correlations were computed for all proteins and transcripts. Figure 15B shows a heatmap of the hierarchical clustering results (methods) of these pairwise comparisons. Hierarchical clustering identified 3 gene clusters (Group A-C) and 3 protein clusters (Group 1-3) which were robust to the choice of clustering algorithm (methods). Group A and Group 1 (Group A1) contained 12 matching gene-protein pairs and Group B and 2 (Group B2) contained 7 matching gene-protein pairs. Similar results were obtained when using the non-batch-effect corrected data to compute correlation and perform hierarchical clustering analysis with pairwise correlations. After identifying these distinct groups of correlated expression patterns, we were next interested in identifying potential clinical factors associated with these patterns.

5.3.5 Correlated groups of genes and proteins stratify donors

Two approaches were taken to investigate the clinical factors associated with GroupA1 or GroupB2 clusters. The first approach was to compute a "donor score," defined as the mean of the batch corrected, normalized, and scaled expression value of genes or proteins identified in each group, calculated per donor. Wilcoxon rank sum tests were then used to assess whether donor scores for each group were associated with different clinical factors (methods). Group A donor score was significantly (FDR adj. p value < 0.05) associated with recurrence status and anti-TNF use (Figure 16A). Group 1 donor score was nominally significant (unadjusted p value < 0.05) with anti-TNF use (Figure 16B). Group A and 1 donor scores had similar trends where donor score was higher in recurring disease and no anti-TNF use. Group B donor score was significantly associated (FDR adj. p value < 0.05) with recurrence status and anti-TNF use (Figure 16C), whereas Group 2 was not

associated with either variable (Figure 16D). Group B had higher donor scores for anti-TNF use and non-recurring disease, showing the opposite trend from GroupA1 donor scores.

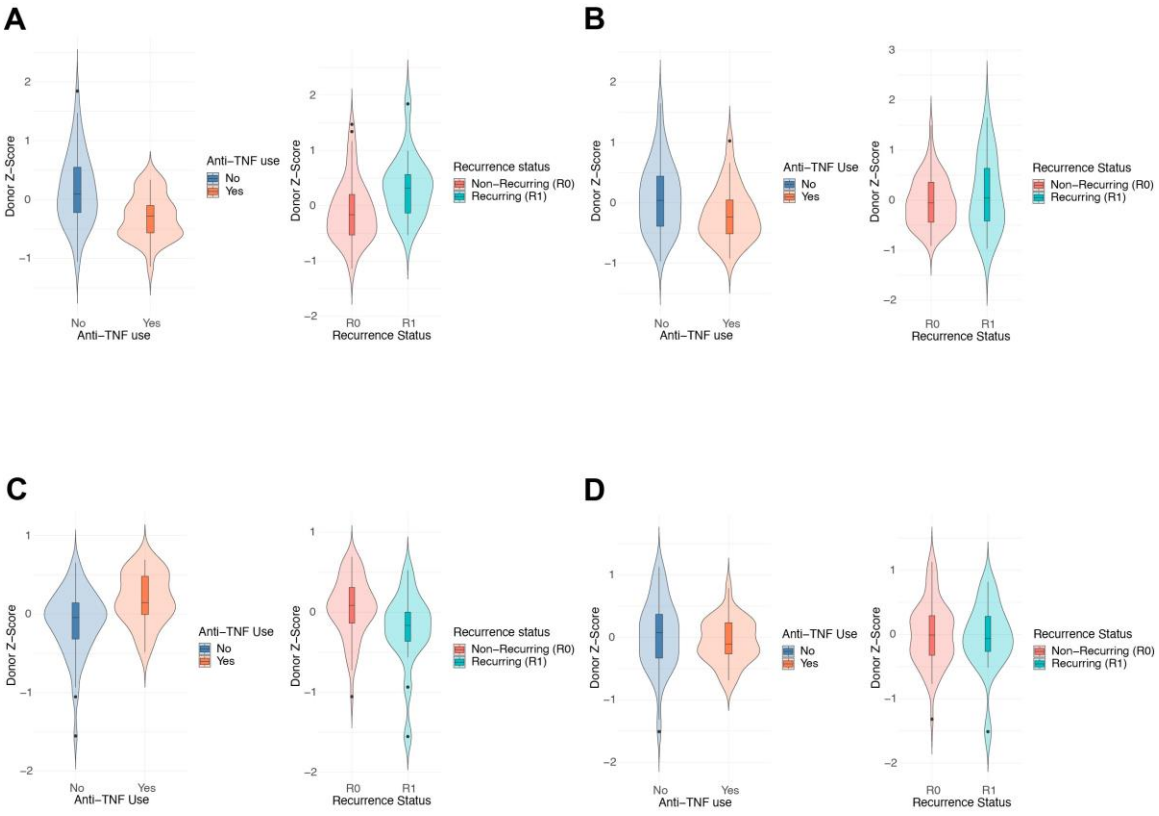


Figure 16 Coordinated gene and protein expression stratify groups of patients. (A) Donor z-score for Group A gene expression. (B) Donor z-score for Group 1 protein expression. (C) Donor z-score for Group B gene expression. (D) Donor z-score for Group 2 protein expression. (A-D) (Left) stratified by anti-TNF use. (Right) stratified by recurrence status. R0 = non-recurring, R1 = recurring.

The donor scores were also computed for the same protein and gene sets on the other individuals not included in the joint gene-protein expression analysis (n = 41 donors with proteomic data and n = 156 donors with transcriptomic data, methods). The general trends were recapitulated for the other donors. Group A was significantly associated (FDR adj. p value < 0.05) with higher donor scores for recurring disease. Group 1 donor score

was not significantly associated with recurrence status or anti-TNF use, but the same trend previously described was observed. Recurrence status and anti-TNF use were not significantly associated with Group B or 2 donor scores, but again, the same trends described earlier were observed. Taking these results together, correlation patterns of gene and protein expression have the potential to stratify donors into clinically relevant groups.

An alternate approach involved direct comparison of gene or protein expression within each Group to associated clinical variables with a Wilcoxon Rank Sum test (methods). Within Group A, 19 of the 31 genes in this group were significantly upregulated in those not on anti-TNF and 10 of the 31 genes were significantly upregulated in those with recurring disease (FDR adj. p value < 0.05, Supplemental Figure 14A,B). All but one gene, *CXCL5*, were associated with both anti-TNF use and recurrence status (Supplemental Figure 14A,B). Two proteins from Group 1, CXCL9 and TNFRSF9, were significantly associated with anti-TNF use, and SLAMF1 was significantly upregulated in males (Supplemental Figure 14C,D). While no Group 2 proteins were significantly associated with recurrence status, 5 genes from Group B were significantly upregulated in those on anti-TNF therapy (Supplemental Figure 14E). These results complement the donor scores that were computed to understand the coordinated ileal gene and serum protein expression patterns.

5.3.6 Recurrence status stratification

We next asked whether the group signatures could stratify donors based on recurrence status. First, PCA was performed using genes in Group A and B and using proteins in Group 1 and 2 (methods). PC1 from Group A and B genes was significantly

associated with anti-TNF use and recurrence status (student's t test < 0.05, Supplemental Figure 15A). PC1 from Group 1 and 2 proteins was also significantly associated with anti-TNF use (student's t test, p value < 0.05, Supplemental Figure 15B). A 3-fold cross validation logistic regression with PCs 1-5 from the gene expression PCA and protein expression PCA was implemented to see whether patients would stratify based on recurring disease (methods, Figure 17). The mean area under the curve (AUC), specificity and sensitivity measures across each fold were 0.63, 0.70, and 0.26, respectively. Overall, the model did not have sufficient predictive power to establish predictive utility with this small sample size, so independent empirical replication is needed.

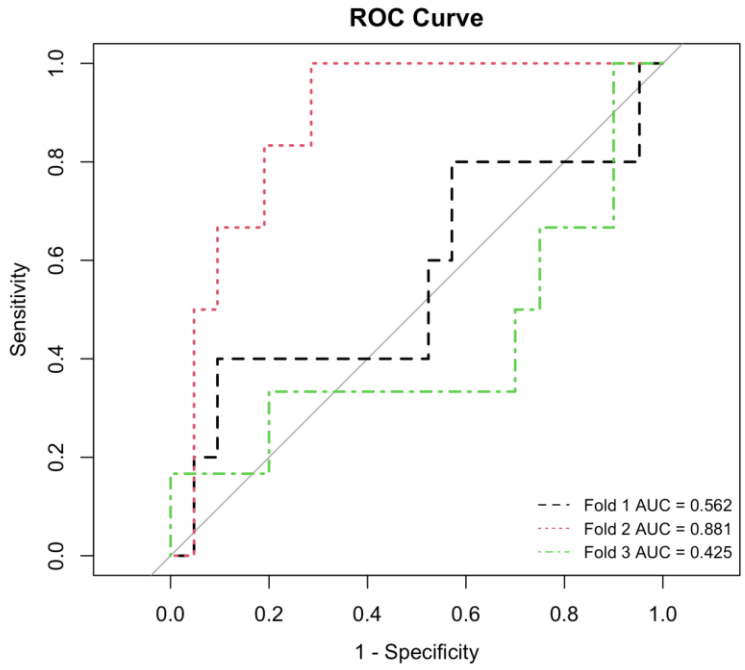


Figure 17 ROC curves of 3-fold cross-validation of logistic regression models with PCs 1-5 of gene and protein group signatures. ROC = receiver operating characteristic, AUC = area under the curve.

5.4 Discussion

While anti-TNF therapy has improved disease prognosis, many individuals with CD still require surgical intervention at some point during disease course [239, 243]. Sequencing technological advancements have enabled biomarker discovery for disease classification and potential therapeutic targets; however, specific, non-invasive biomarkers related to post-op disease recurrence are needed.

Differential expression analysis was leveraged to investigate potential sex-biased protein expression given our recent observations in recurring CD at the transcriptomic level. This analysis revealed that there were a few instances of sex-biased protein expression in recurring CD, though specifically within non-recurring CD (Figure 13C). Notably the set of proteins on the Olink panel are underrepresented in the sex-biased ileal transcript list, so this analysis is not globally representative of sex differences in serum proteins. Nevertheless, previous studies have implicated CCL3 and CCL4, known as macrophage inflammatory proteins, in promoting pro-inflammatory processes through increased neutrophil accumulation in the colon [273-275]. Leveraging CCL4 as a potential therapeutic strategy has been demonstrated by Gong et al. by conjugating CCL4 with nanoparticles which bind to CCR5 effectively alleviating IBD associated inflammation in experimental models of colitis [276]. Interestingly, IL24 was enriched in females with non-recurring disease (Figure 13C). IL24 may have a protective role in CD by suppressing inflammation, potentially engaging pro-fibrotic processes to initiate compensatory mechanisms, attempting to restore homeostasis [277, 278]. CCL3 was also differentially expressed between females with recurring and non-recurring disease, highlighting disease heterogeneity (Figure 13D). Further studies are needed to evaluate the consistency of these

observations and to refine the mechanisms responsible for a sex-biased role of these proteins in therapeutic response.

In addition, the serum proteomics differential expression analysis identified marked enrichment of proteins associated with anti-TNF use relative to recurring disease. Proteins elevated in recurring disease are involved in pro-inflammatory and pro-fibrotic processes which could lead to more progressive disease [279, 280]. Additionally, proteins elevated in patients not using anti-TNF biologics were associated with pro-inflammatory disease mechanisms [273, 281-285]. Our study further supports the role of CXCL9 and CXCL10 in disease recurrence, consistent with studies showing these proteins contribute to disease pathogenesis and potentially influence therapeutic efficacy [128, 282, 286, 287]. Despite previous studies that have reported varying degrees of success utilizing serum biomarkers to monitor CD activity [288-290], our results suggest that serum biomarkers may be useful to monitor post-op therapeutic efficacy in individuals with recurring CD. Future studies are needed to assess whether CXCL9 and CXCL10 are predictive of therapeutic response in the post-op setting.

Differential gene expression analysis complemented the differential protein expression analysis. DEGs associated with higher expression in non-recurring disease generally have a protective effect by initiation of tissue repair in CD [291-297]; however, over expression may lead to fibrosis or angiogenesis in certain cases [298, 299]. Alternatively, DEGs associated with higher expression in recurring disease were generally involved in pro-inflammatory and pro-fibrotic mechanisms [253, 266, 282, 300-306]. Even though there was not a significant interaction between sex and recurrence status in this analysis, the interaction between these variables could be enhanced by other genes not

included in this analysis (Chapter 4). Since similar post-op disease recurrence and anti-TNF signatures were observed for the transcriptomic and proteomic differential expression analysis, we were next interested in performing a comparative analysis between these modalities.

For a comparative analysis between ileal gene expression and serum proteomic expression, we computed spearman correlation for matching gene-protein pairs and for all pairwise comparisons. Even though many genes and proteins were weakly correlated, a few gene-protein pairs had modest but stronger correlation (Figure 15). Koussounadis et al. found that DEGs have higher correlation with protein expression compared to other genes [124]. Similar results were obtained in this study. For example, CXCL9 was one of the top correlated gene-protein pairs, in addition to being both a DEG and DEP associated with anti-TNF use. Overall, these findings suggest that correlated gene and protein expression levels in different tissues may identify biomarkers for disease progression or therapeutic response.

Hierarchical clustering further identified groups of genes and proteins with similar expression patterns (Figure 15B). The groups of genes and proteins were found to be correlated with clinical factors related to disease recurrence. GroupA1 patterns showed elevated expression of genes and proteins in those with recurring disease and no anti-TNF use (Figure 16A,B) whereas the opposite trend was observed in GroupB2 (Figure 16C,D). Multi-omic integration has been used to elucidate alternative underlying disease mechanisms associated with groups of donors, classify IBD subtypes, and predict response to treatment [260, 307, 308]. These results underscore the utility of multi-omic data for patient stratification into clinically informative groups of donors.

Lastly, a logistic regression model was implemented to classify disease recurrence based on the hierarchical clustering derived-Group signatures. Even though this model had limited predictive capabilities, most likely due to the small cohort, this approach applied to a larger cohort may be useful for patient stratification within distinct patient subgroups or clinical factors. A major limitation of the current study is the small sample size despite the combined efforts of multiple GRC. While larger replication studies may confirm our findings, the heterogeneity of molecular response to colectomy means that it is unlikely that combined protein and transcript biomarkers will provide useful prediction of recurrence. Also, since anti-TNF use was different between the Canadian and American sites reflecting in part different therapeutic approaches, the biological correlations are confounded with what may be a batch effect of site. This consideration underscores the practical issues related to robust biomarker development.

Overall, ileal transcriptomic and serum proteomic profiles at first endoscopy post resection showed correlated expression patterns that were associated with clinical factors that stratify CD patients. These results suggest that serum may reflect systemic aspects of post-op disease based on transcriptomic disease signatures in the ileum. We also identified DEPs associated with the interaction between disease recurrence and sex, and anti-TNF use. CXCL9 was one of the top correlated gene-protein pairs and was significantly associated with anti-TNF use at both the transcriptomic and proteomic level, underscoring the potential utility of this biomarker to monitor disease recurrence and/or anti-TNF use. Several proteins were associated with the interaction between disease recurrence and sex, supporting the observation of sex-biased expression in recurring disease. These molecular

signatures may serve as a valuable tool for patient stratification and disease management monitoring in the CD post-op setting.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

Building on early GWAS and linkage studies that revealed the genetic component of CD, omics profiling has been harnessed to further elucidate how genetic variation may lead to alternative mechanisms contributing to CD pathogenesis and progression across a range of environmental conditions [36, 258]. Omics profiling of intestinal tissue from individuals with CD has revealed pathogenic pathways within immune, epithelial, and stromal cell compartments that promote inflammation or response to therapy [92, 96, 309, 310]. Heterogeneity across donors has been previously described, but many studies focused on small or large intestine separately, potentially missing disease heterogeneity even within the same individual. Further, many of these studies were performed on individuals with steady-state disease, warranting investigation of more progressive forms of disease and disease at inception. As omics profiling becomes more popular to investigate disease biology, its potential for advancing personalized medicine in clinical settings remains an active area of research.

Chapter 2 of this thesis focused on investigation of inflammatory mechanisms that may be involved in persistence of perianal-fistulizing CD (perianal-CD), even after therapeutic intervention. This scRNA-seq study revealed compositional remodeling of immune and epithelial compartments in inflamed tissue compared to non-inflamed, healed tissue from those with perianal-CD, while noting heterogeneity even within each group. Generally, there was expansion of epithelial cells in the non-inflamed tissue and expansion of immune cells within the inflamed tissue; however, after analyzing epithelial and immune cell compartments separately, there was an expansion of goblet cells in the inflamed tissue

prompting further investigation of this cell type. Additional analyses revealed pro-inflammatory pathways enriched in goblet cells from inflamed tissue, suggesting goblet cells are involved in disease pathogenesis. Dysregulation of goblet cell function may lead to dysbiosis and perpetuate inflammation [311]. One limitation of this study is that rectal biopsies were collected only from near the fistula, potentially neglecting specific cellular mechanisms occurring in the fistula tract itself, so future single-cell studies addressing this limitation are needed. Furthermore, African Americans are disproportionately affected by perianal-CD [100, 312, 313], so future studies should continue to sample diverse populations to understand factors potentially contributing to progressive disease across different populations.

Progressive disease may exhibit distinct molecular signatures that differ from disease at inception, prompting investigation of this clinical phenotype. There were two main aims of the third chapter of this thesis. The first aim of this chapter was to develop a clustering stability assessment in a treatment naïve CD cohort to promote reproducible and rigorous scRNA-seq analyses. Dependency of current clustering methods applying user-defined parameters may introduce variability in results based on parameter selection [80, 82, 83], necessitating approaches that can identify clusters robust to the selected parameter. The stability assessment workflow yielded similar and repeatable results, suggesting stably assigned cell clustering results were more representative of the cell type(s) identified in the intestines. This workflow could be further extended by investigating the extent to which other clustering parameters impact clustering stability, in addition to further classification of unstably assigned cells into low quality cells or transitional cells, for example. Validation of this approach in other scRNA-seq studies is needed as well. The second aim

of the third chapter was to investigate the molecular signatures that might promote mucosal healing or disease progression across the intestines in the treatment naïve CD cohort. This study revealed substantial cellular composition heterogeneity across and within the regions of the intestine. Since we could not conclude that inflammation status was the main driver of compositional differences, tensor decomposition was used to understand inter-individual variation within each tissue compartment. This analysis stratified donors into distinct groups associated with particular pathological modes of disease. Within the ileum, two groups of donors were identified, one group associated with a stricturing disease signature, and the other group of donors associated with a penetrating disease signature [53, 188]. In the colon, one group of donors had inflammation perpetuated by myeloid cells whereas another mechanism was promoting inflammation in the other group of donors. Tensor decomposition analysis in the rectum revealed two groups of donors with separate inflammatory profiles, one group had an IFN-γ and TNF-α profile whereas the other group had a TNF-α profile. Connecting genetics with transcriptomics, integration of GWAS summary statistics and scRNA-seq data also confirmed association of T cells and monocytes to CD, validating the cell types associated with the groups of donors identified in this cohort. While this unique cohort contained samples across the intestines from treatment naïve CD donors, longitudinal studies with follow-up samples from the same donor are needed to establish whether these signatures are predictive of disease progression, or in certain cases mucosal healing, across the intestines in these individuals.

Post-surgical disease recurrence constitutes another clinically important CD phenotype. The primary focus of chapter 4 was to investigate whether recurring CD is associated with alternative splicing mechanisms in a post-op cohort. Previous studies have

shown that splicing dysregulation in IBD is in part associated with down-regulation of HP1γ and transition from ileal-like to rectal-like splicing patterns in ileal biopsies from CD patients [64, 67]. In this study, tissue-specific differentially used transcripts were able to stratify donors with recurring or non-recurring disease, with rectal transcripts associated with recurring disease in the ileum. These results were also recapitulated at the gene expression level, suggesting that the rectal-like transition may be involved in post-op recurrence. Other signatures of HP1γ dysregulation were observed, including enrichment of lipid metabolism and membrane trafficking [64]. These results suggest that HP1γ could be a potential therapeutic target for preventing disease recurrence. Lastly, this study implicated the p53 pathway, traditionally described in the context of colorectal cancer, in recurring CD. While pathway enrichment analysis suggested that epithelial and stromal cells were involved in recurring CD based on PCA of tissue specific differentially used transcripts, scRNA-seq studies that investigate the distinct cellular contributions to disease recurrence are needed. Moreover, comparison of post-op recurrence across the intestines to earlier stages of disease may enable prediction of disease recurrence.

Extending the analysis from chapter 4, chapter 5 investigated the serum proteomic and ileal transcriptomic signatures in the post-op CD cohort. The first aim of this chapter was to investigate whether sex-bias occurs at the proteomic level with respect to recurring disease. Differential protein expression analysis for samples at first endoscopy post-op revealed protein expression differences between sex in donors with non-recurring disease, and in one case, an expression difference between females with recurring or non-recurring disease, validating sex-bias for some proteins at the proteomic level. The second aim of this study was to evaluate whether incorporation of both transcriptomics and proteomics

data could improve and validate biomarkers for recurring disease. Comparative analysis of correlation between the same serum protein and ileal gene levels validated CXCL9 as a potential biomarker for monitoring post-op recurrence [119, 128]. Hierarchical clustering analysis also identified distinct groups of correlated genes and proteins that stratify donors based on anti-TNF use and disease recurrence. A logistic regression model was implemented to evaluate whether these correlation patterns can stratify donors based on disease recurrence; however the model had low predictive accuracy. A similar analysis in a larger cohort may improve the predictive accuracy of patient stratification based on these signatures. Future studies comparing these potential biomarkers to current disease monitoring strategies (CRP and fecal calprotectin) are needed to determine whether these new markers offer improved monitoring of disease recurrence.

In summary, this thesis explored the applications of omics profiling for characterization of inception and more progressive forms of CD across intestinal tissues, with implications for personalized medicine. Previous studies have primarily focused on differences between CD cases and controls. In this thesis, variation within CD was explored, illuminating alternative mechanisms that might be involved in pathology at both disease inception and progressive disease. As omics cohorts continue to grow and incorporate more complex datasets, novel analytical challenges will inevitably arise. Statistical and technical methods that address related challenges will be needed for reproducible analyses. Despite ongoing barriers to translating personalized medicine approaches into clinical practice, genomics offers valuable insights for guidance on disease management strategies. Overall, this thesis provides a framework for how future omics studies can inform personalized medicine

approaches by considering the pathological pathways underlying disease progression in IBD and other diseases.

PUBLICATIONS

1. **Washburn, S.,** Borowski, K., Briggs, K. Rioux, J., Lazarev, M., & Gibson, G. (2025). Correlated multi-omic signatures inform potential clinical stratification in post-operative Crohn's disease. *In prep.*
2. Gettler, K., Nagpal, S., **Washburn, S.,** Tastad, C., Zhang, J., Sabic, K., Lazarev, M., … & Cho, J.H. (2025). Post-operative ileum transcriptomics implicate sex-biased mechanisms in Crohn's disease recurrence. *Submitted*.
3. **Washburn, S.,** Hwang, Y., Maddipatla, S.C., Murthy, S., Koti, T., Kolachala, V.L. Gibson, G.*, Kugathasan, S.*, & Qiu, P.* Identification of Crohn's disease subtypes in single cell RNA sequencing signatures of treatment naïve samples across the paediatric gastrointestinal tract. (2025). *Under revision*.
4. **Washburn, S.,** Maddipatla, S.C., Murthy, S., Dodd, A., Pelia, R.S., Kolachala, V.L., Geem, D., Matthews, J.D., Gibson, G., & Kugathasan, S. (2024). Persistent inflammation of the rectum in perianal fistulizing Crohn's disease is associated with goblet cell function. *Gastro Hep Advances*, *3*(1), 131-133.

APPENDIX A. SUPPLEMENTARY TABLES

Supplemental Table 1 Patient demographics. [1] Self-reported ancestry: AA, African American; EA, European American; Hisp, Hispanic ethnicity. [2] Age in years. Bins for covariate adjustment in differential gene expression analysis were: <19, pediatric; 19-30 young adult; >30 adult.

Lab ID	Sample ID	Disease	Ancestry[1]	Gender	Age[2]	Infl. Status	Treatment
P5	IF1	perianal-CD	AA	F	41	Inflamed	Infliximab
P7	IF2	perianal-CD	AA	M	31	Inflamed	Humira
P11	IF3	perianal-CD	EA	M	13	Inflamed	Vedolizumab (Entyvio)
P15	IF4	perianal-CD	EA	M	15	Inflamed	Ustekinumab (Stelara)
P16	IF5	perianal-CD	EA	F	21	Inflamed	Ustekinumab (Stelara)
P18	IF6	perianal-CD	EA	M	13	Inflamed	Infliximab (Remicade)
P6	NI1	perianal-CD	EA	M	31	Non-inflamed	Inflectra
P8	NI2	perianal-CD	AA	M	13	Non-inflamed	Inflectra
P13	NI3	perianal-CD	EA	F	19	Non-inflamed	Adalimumab (citrate-free) (Humira (CF) pen)
P17	NI4	perianal-CD	EA	M	51	Non-inflamed	Humira and Methotrexate
P19	NI5	non-IBD	Hisp	F	17	Non-inflamed	Non-IBD
P20	NI6	perianal-CD	EA	F	8	Non-inflamed	Infliximab (Remicade)
P21	NI7	perianal-CD	EA	F	54	Non-inflamed	Vedolizumab (Entyvio)

Supplemental Table 2 Metadata of samples included in this study. Macro IF status = macroscopic inflammation status, Micro IF status = microscopic inflammation status, IF = inflamed, NI = non-inflamed, SIRE = self-identified race and ethnicity, M = male, F = female, AA = African American, EA = European American.

Sample	FastQ Name	Donor	Manuscript ID	Tissue Grouping) (Broad	Tissue	Macro IF status	Micro IF Status	Sex	Age	SIRE	Sequencing Batch	Disease behavior at time of sample	Group (scITD)
1	GCA_31_I	1	donor1	ileum	ileum	IF	NI	M	9	AA	1	B1	1
2	GCA_31_A	1	donor1	colon	ascending colon	IF	NI	M	9	AA	1		1
3	GCA_31_R	1	donor1	rectum	rectum/rectal sigmoid	IF	NI	M	9	AA	1		2
4	GCA_32_I	2	donor2	ileum	ileum	IF	IF	M	13	Asian	1	B3	1
7	GCA_33_I	3	donor3	ileum	ileum	IF	IF	M	11	AA	1	B1	1
8	GCA_33_A	3	donor3	colon	ascending colon	IF	NI	M	11	AA	1		1

9	GCA_33_R	3	donor3	rectum	rectum/rectal sigmoid	NI	NI	M	11	AA	1		NA
24	GCA_40_I	4	donor4	ileum	ileum	NI	IF	M	13	AA	3	B1	1
25	GCA_40_R	4	donor4	rectum	rectum/rectal sigmoid	IF	IF	M	13	AA	3		1
26	GCA_40_A	4	donor4	colon	ascending colon	IF	IF	M	13	AA	3		NA
29	GCA_42_I	5	donor5	ileum	ileum	IF	IF	F	10	EA	3	B1	1
30	GCA_42_R	5	donor5	rectum	rectum/rectal sigmoid	IF	IF	F	10	EA	3		2
31	GCA_42_A	5	donor5	colon	ascending colon	IF	IF	F	10	EA	3		NA
32	GCA_43_I	6	donor6	ileum	ileum	IF	IF	M	16	EA	3	B1	2
33	GCA_43_A	6	donor6	colon	cecum	NI	NI	M	16	EA	3		1
34	GCA_43_R	6	donor6	rectum	rectum/rectal sigmoid	NI	NI	M	16	EA	3		2

35	GCA_44_I	7	donor7	ileum	ileum	IF	NI	M	15	AA	3	B1	2
36	GCA_44_A	7	donor7	colon	cecum	NI	NI	M	15	AA	3		1
37	GCA_44_R	7	donor7	rectum	rectum/rectal sigmoid	NI	NI	M	15	AA	3		2
44	GCA48_I	8	donor8	ileum	ileum	IF	NI	F	18	Asian	4		1
45	GCA48_T	8	donor8	colon	transverse colon	NI	NI	F	18	Asian	4	B1	NA
46	GCA48_R	8	donor8	rectum	rectum/rectal sigmoid	NI	NI	F	18	Asian	4		NA
47	GCA49_I	9	donor9	ileum	ileum	IF	IF	F	12	EA	4	B1	1
48	GCA49_DC	9	donor9	colon	descending colon	NI	NI	F	12	EA	4		NA
49	GCA49_R	9	donor9	rectum	rectum/rectal sigmoid	NI	NI	F	12	EA	4		NA
50	GCA_45_I	10	donor10	ileum	ileum	IF	IF	F	14	EA	5	B3	2

51	GCA_45_R	10	donor10	rectum	rectum/rectal sigmoid	IF	NI	F	14	EA	5		1
52	GCA_45_C	10	donor10	colon	cecum	NI	NI	F	14	EA	5		1
53	GCA_50_I	11	donor11	ileum	ileum	IF	NI	M	16	EA	5	B1	1
54	GCA_50_C	11	donor11	colon	cecum	IF	IF	M	16	EA	5		2
55	GCA_50_R	11	donor11	rectum	rectum/rectal sigmoid	IF	NI	M	16	EA	5		2
56	GCA_51_I	12	donor12	ileum	ileum	IF	NI	M	12	AA	5	B1	2
57	GCA_51_Ce	12	donor12	colon	cecum	NI	NI	M	12	AA	5		1
58	GCA_51_R	12	donor12	rectum	rectum/rectal sigmoid	NI	NI	M	12	AA	5		2
59	GCA_52_I	13	donor13	ileum	ileum	IF	IF	F	7	AA	5	B1	NA
60	GCA_52_Ce	13	donor13	colon	cecum	NI	NI	F	7	AA	5		1

61	GCA52_R	13	donor13	rectum	rectum/rectal sigmoid	NI	NI	F	7	AA	5			2
62	GCA53_I	14	donor14	ileum	ileum	IF	IF	F	3	AA	6	B1		1
63	GCA53_A	14	donor14	colon	ascending colon	IF	IF	F	3	AA	6			2
64	GCA53_R	14	donor14	rectum	rectum/rectal sigmoid	IF	IF	F	3	AA	6			1
65	GCA54_I	15	donor15_s1	colon	ileocecal valve	IF	IF	F	13	AA	6	B2	& B3	2
66	GCA54_Ce	15	donor15_s2	colon	cecum	NI	NI	F	13	AA	6			1
67	GCA54_R	15	donor15	rectum	rectum/rectal sigmoid	NI	NI	F	13	AA	6			2
71	GCA56_Ce	16	donor16_s1	colon	cecum	IF	IF	M	8	EA	6	B1		2
72	GCA56_A	16	donor16_s2	colon	ascending colon	IF	IF	M	8	EA	6			1
73	GCA56_R	16	donor16	rectum	rectum/rectal sigmoid	NI	NI	M	8	EA	6			1

77	GCA58_I	17	donor17	ileum	ileum	NI	IF	M	14	AA	7	B1	2
78	GCA58_T	17	donor17	colon	transverse colon	NI	NI	M	14	AA	7		1
79	GCA58_R	17	donor17	rectum	rectum/rectal sigmoid	NI	NI	M	14	AA	7		2
83	GCA60_I	18	donor18	ileum	ileum	IF	NI	F	15	AA	7	B1	1
84	GCA60_Ce	18	donor18	colon	cecum	IF	IF	F	15	AA	7		2
85	GCA60_R	18	donor18	rectum	rectum/rectal sigmoid	IF	IF	F	15	AA	7		1
86	GCA62_Ce	19	donor19	colon	cecum	IF	NI	F	14	AA	7	B3	1
87	GCA62_R	19	donor19	rectum	rectum/rectal sigmoid	NI	IF	F	14	AA	7		2
88	GCA63_I	20	donor20	ileum	ileum	NI	NI	M	8	AA	8	B1	1
89	GCA63_Ce	20	donor20	colon	cecum	IF	IF	M	8	AA	8		1

90	GCA63_R	20	donor20	rectum	rectum/rectal sigmoid	NI	NI	M	8	AA	8		2
91	GCA64_I	21	donor21	ileum	ileum	IF	N/A	F	15	EA	8	B1	2
92	GCA64_Ce	21	donor21	colon	cecum	NI	NI	F	15	EA	8		1
93	GCA_R	21	donor21	rectum	rectum/rectal sigmoid	NI	NI	F	15	EA	8		1
96	GCA66_SF	22	donor22	colon	splenic flexture	IF	IF	M	15	Asian	8	B1	2
97	GCA66_R	22	donor22	rectum	rectum/rectal sigmoid	IF	IF	M	15	Asian	8		1
98	GCA67_I	23	donor23	ileum	ileum	IF	NI	M	17	EA	8	B1	1
99	GCA67_Ce	23	donor23	colon	cecum	NI	NI	M	17	EA	8		1
101	GCA68_I	24	donor24	ileum	ileum	IF	IF	F	15	AA	8	B1	2
102	GCA68_Ce	24	donor24	colon	cecum	NI	NI	F	15	AA	8		1

103	GCA68_R	24	donor24	rectum	rectum/rectal sigmoid	NI	NI	F	15	AA	8		NA
110	GCA71_Ce	25	donor25	colon	cecum	IF	IF	M	14	EA	9	B1	2
111	GCA71_R	25	donor25	rectum	rectum/rectal sigmoid	NI	NI	M	14	EA	9		1
112	GCA72_I	26	donor26	ileum	ileum	NI	NI	F	7	Asian	9	B1	2
113	GCA72_Ce	26	donor26	colon	cecum	NI	NI	F	7	Asian	9		1
114	GCA72_SC	26	donor26	rectum	sigmoid colon	NI	NI	F	7	Asian	9		2
119	GCA74_I	27	donor27	ileum	ileum	IF	IF	M	13	AA	10	B1	1
120	GCA74_Ce	27	donor27	colon	cecum	IF	NI	M	13	AA	10		1
121	GCA74_R	27	donor27	rectum	rectum/rectal sigmoid	NI	NI	M	13	AA	10		2
128	GCA77_ICV	28	donor28_s1	colon	ileocecal valve	NI	NI	M	17	AA	10	B3	2

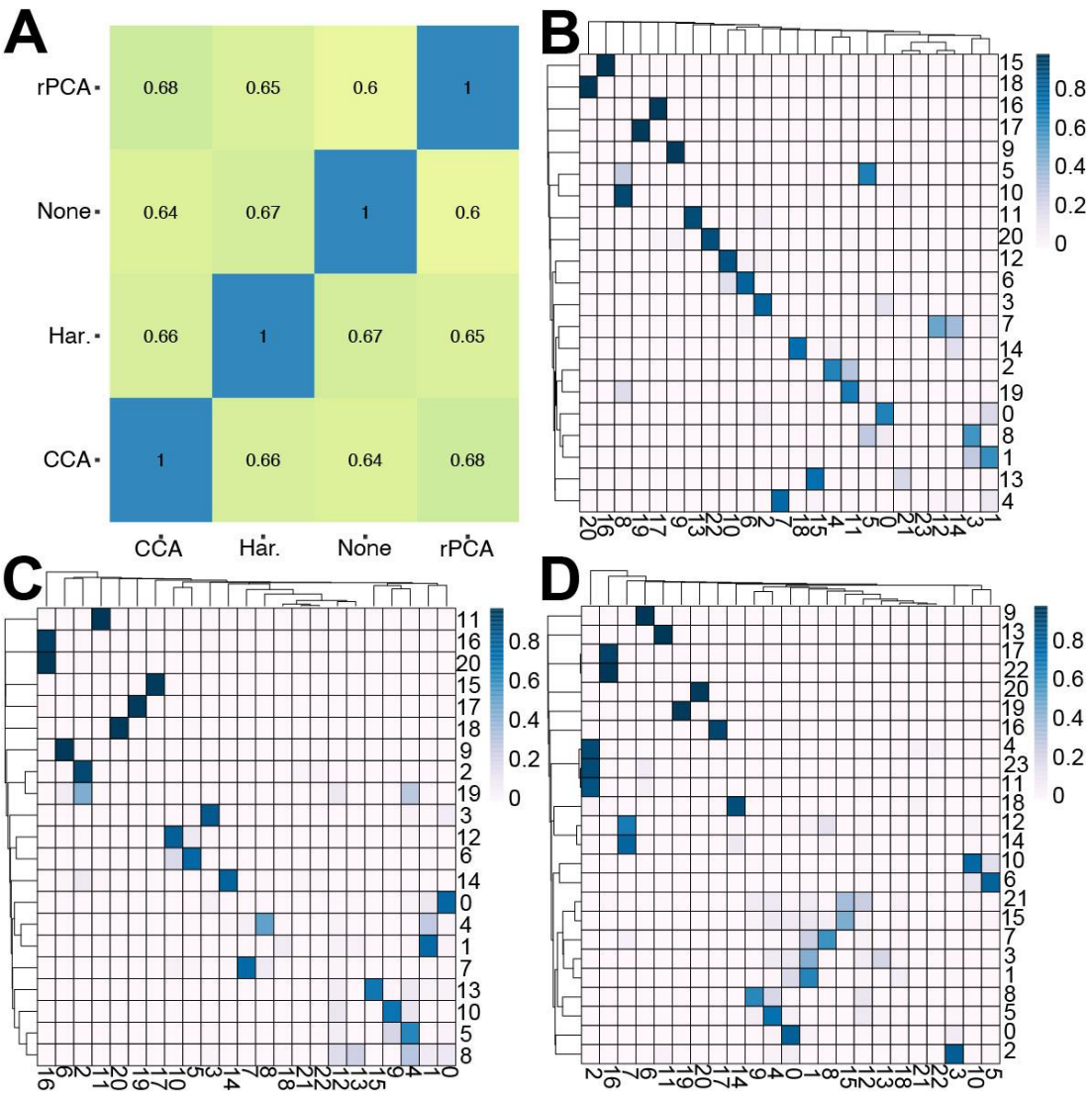
129	GCA77_Ce	28	donor28_s2	colon	cecum	NI	NI	M	17	AA	10		2
130	GCA77_R	28	donor28	rectum	rectum/rectal sigmoid	NI	NI	M	17	AA	10		1
134	GCA79_I	29	donor29	ileum	ileum	IF	IF	F	17	AA	11	B1	NA
135	GCA79_AC	29	donor29	colon	ascending colon	IF	IF	F	17	AA	11		2
136	GCA79_R	29	donor29	rectum	rectum/rectal sigmoid	NI	NI	F	17	AA	11		2
147	GCA84_I	30	donor30	ileum	ileum	IF	NI	F	15	EA	11	B2	2
148	GCA84_Ce	30	donor30	colon	cecum	IF	NI	F	15	EA	11		2
149	GCA_84R	30	donor30	rectum	rectum/rectal sigmoid	IF	NI	F	15	EA	11		1
153	GCA86_I	31	donor31	ileum	ileum	IF	IF	F	9	AA	12	B1	NA
154	GCA86_Ce	31	donor31	colon	cecum	IF	IF	F	9	AA	12		2

155	GCA86_R	31	donor31	rectum	rectum/rectal sigmoid	IF	IF	F	9	AA	12		1
157	GCA87_Ce	32	donor32	colon	cecum	IF	IF	M	12	EA	12	B1	1
158	GCA87_R	32	donor32	rectum	rectum/rectal sigmoid	NI	NI	M	12	EA	12		2
162	GCA89_I	33	donor33	ileum	ileum	NI	NI	F	11	AA	12	B1	NA
163	GCA89_Ce	33	donor33	colon	cecum	IF	IF	F	11	AA	12		2
164	GCA89_R	33	donor33	rectum	rectum/rectal sigmoid	NI	NI	F	11	AA	12		1
171	GCA92_I	34	donor34	ileum	ileum	IF	IF	M	13	EA	13	B3	2
172	GCA92_Ce	34	donor34	colon	cecum	IF	IF	M	13	EA	13		1
173	GCA92_R	34	donor34	rectum	rectum/rectal sigmoid	IF	IF	M	13	EA	13		1

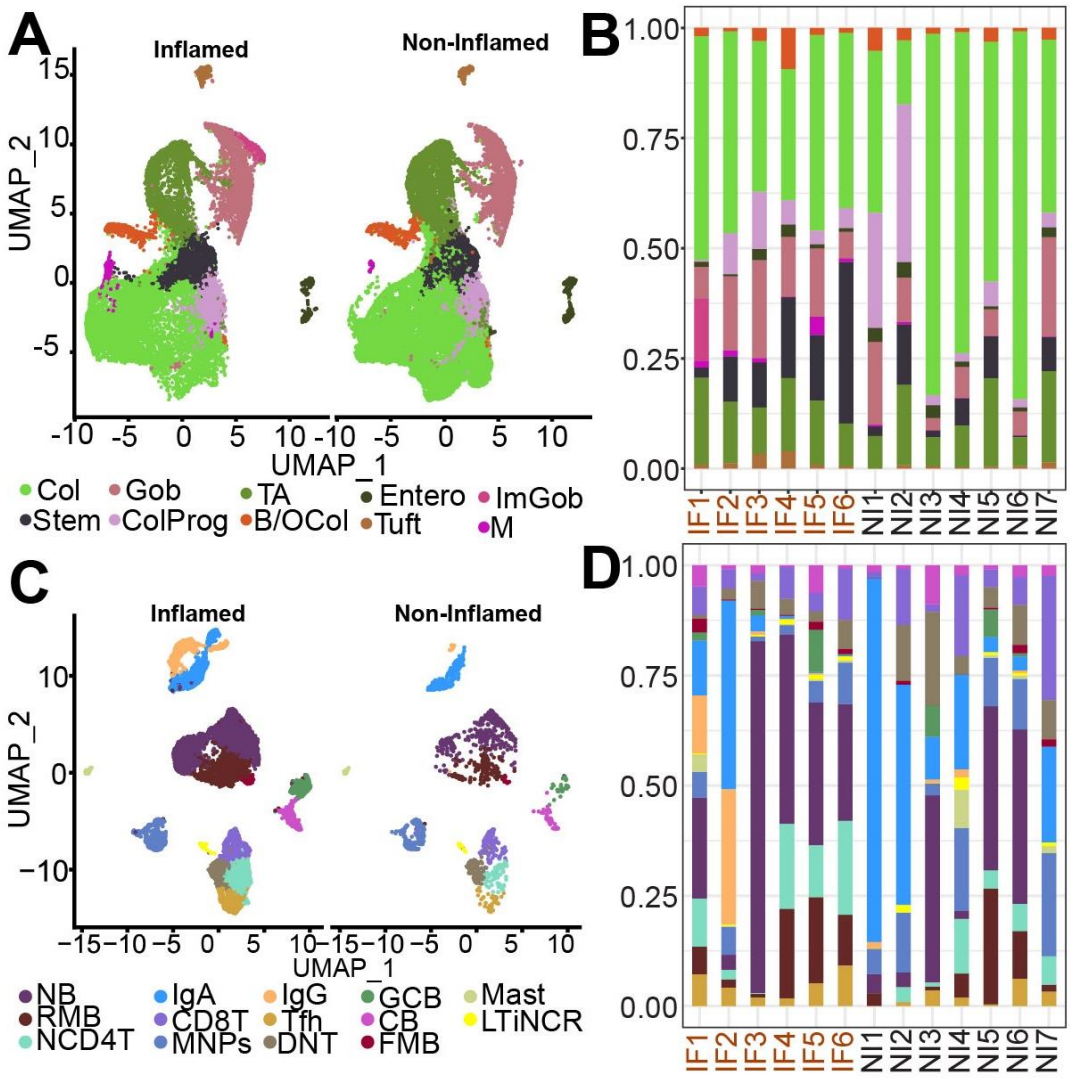
Supplemental Table 3 Gene-protein pairs with pseudo p value < 0.10.

Gene-Protein name	Spearman Correlation
ADA	0.405
CCL13	0.324
CCL23	0.181
CCL8	0.192
CD274	-0.222
CXCL9	0.444
IL10RB	0.202
IL17A	0.358
IL24	0.174
IL6	0.204
MMP10	0.253
NRTN	0.314
TGFA	-0.245
TGFB1	-0.143

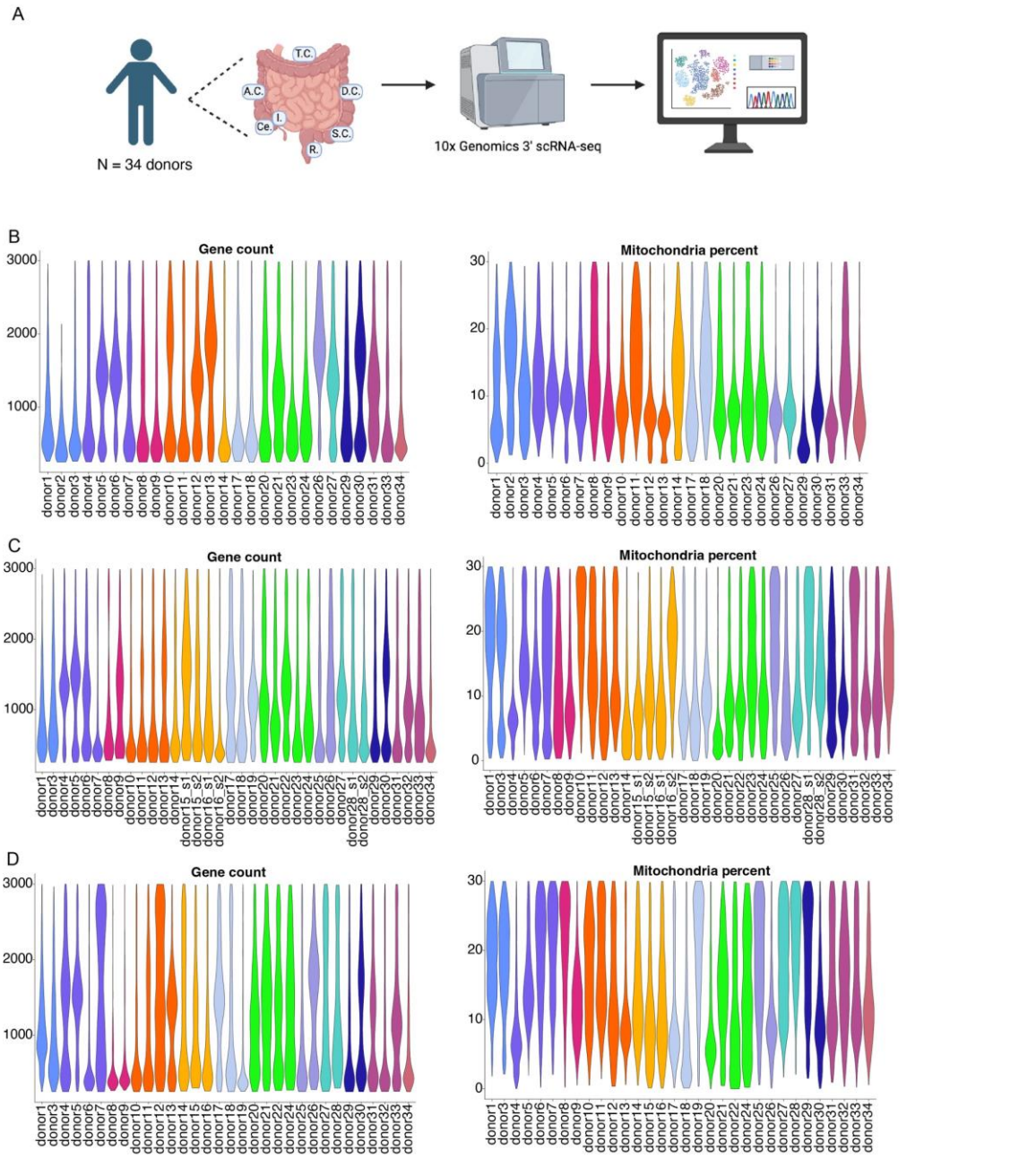
APPENDIX B. SUPPLEMENTARY FIGURES



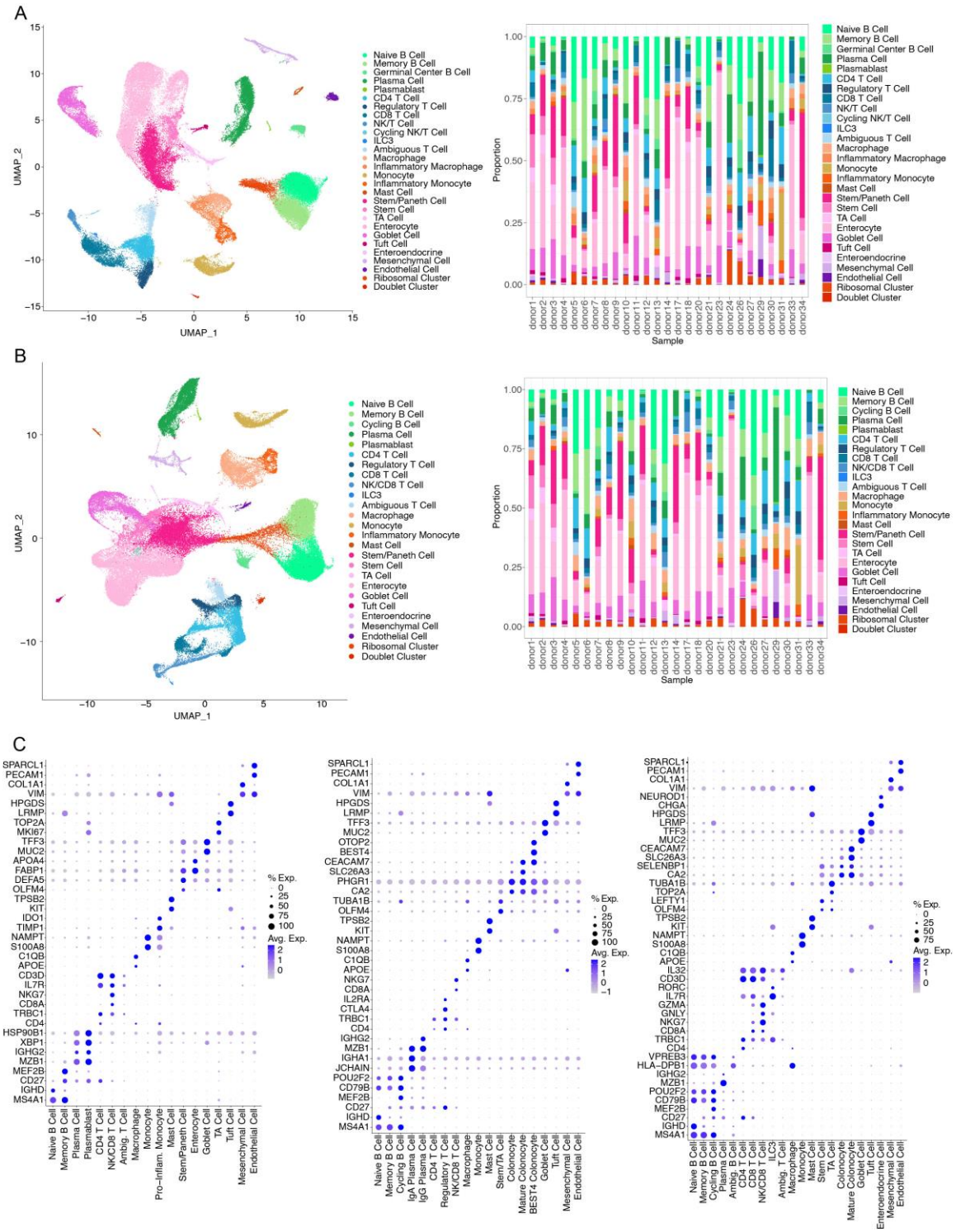
Supplemental Figure 1 Justification of rPCA for Batch Effect Correction. A) Rand Index (RI) comparison of cluster stability contrasting no batch correction (None) with the three batch correction methods, Harmony (Har), Canonical Correlation Analysis (CCA), and reciprocal Principal Component Analysis (rPCA), all implemented in Seurat v4.1.1. An RI of 0 indicates dissimilarity whereas 1 indicates perfect concordance, implying high stability of the clustering. (B-D) Confusion matrix heatmaps showing the proportion of cells assigned to each cluster in the pairwise comparisons of the batch correction methods (darker blue indicates more sharing). (B) rPCA y-axis versus CCA x-axis. (C) rPCA versus Harmony. (D) CCA versus Harmony. rPCA shows the highest concordance with both methods and was chosen for downstream analysis



Supplemental Figure 2 Representation of epithelial and immune cell compartments. (A) UMAP of ten epithelial cell types clustered jointly but split by Inflamed versus Non-Inflamed status. The cell types are Col, colonocytes; ColProg, colonocyte progenitors; B/OCol, *BEST4/OTOP2* positive colonocytes; Gob, goblet cells; ImGob, immature goblet cells; Stem, stem cells; TA, transit amplifying cells; Entero, enteroendocrine; Tuft, tuft cells; M, M-cells. (B) Proportions of each epithelial cell type in each subject, with same color code. (C) UMAP of the immune cell compartment, also jointly clustered but split by inflammation status to show relative overall abundance. The 14 cell types are: NB, naïve B cells; RMB, resting memory B cells; FMB, *FCRL4* positive memory B cells; CB, cycling B cells; GCB, germinal center-like B cells; IgA, IgA-expressing plasma cells; IgG, IgG-expressing plasma cells; NCD4T, naïve CD4-positive T-cells; CD8T, CD8-positive T-cells; LTiNCR, LTi-like NCR-positive type 3 innate lymphoid cells; Tfh, follicular helper T-cells; Mast, mast cells; MNPs, mononuclear phagocytes. (D) Stacked bar graph of immune cell proportions by subject, showing heterogeneity of the compartment.

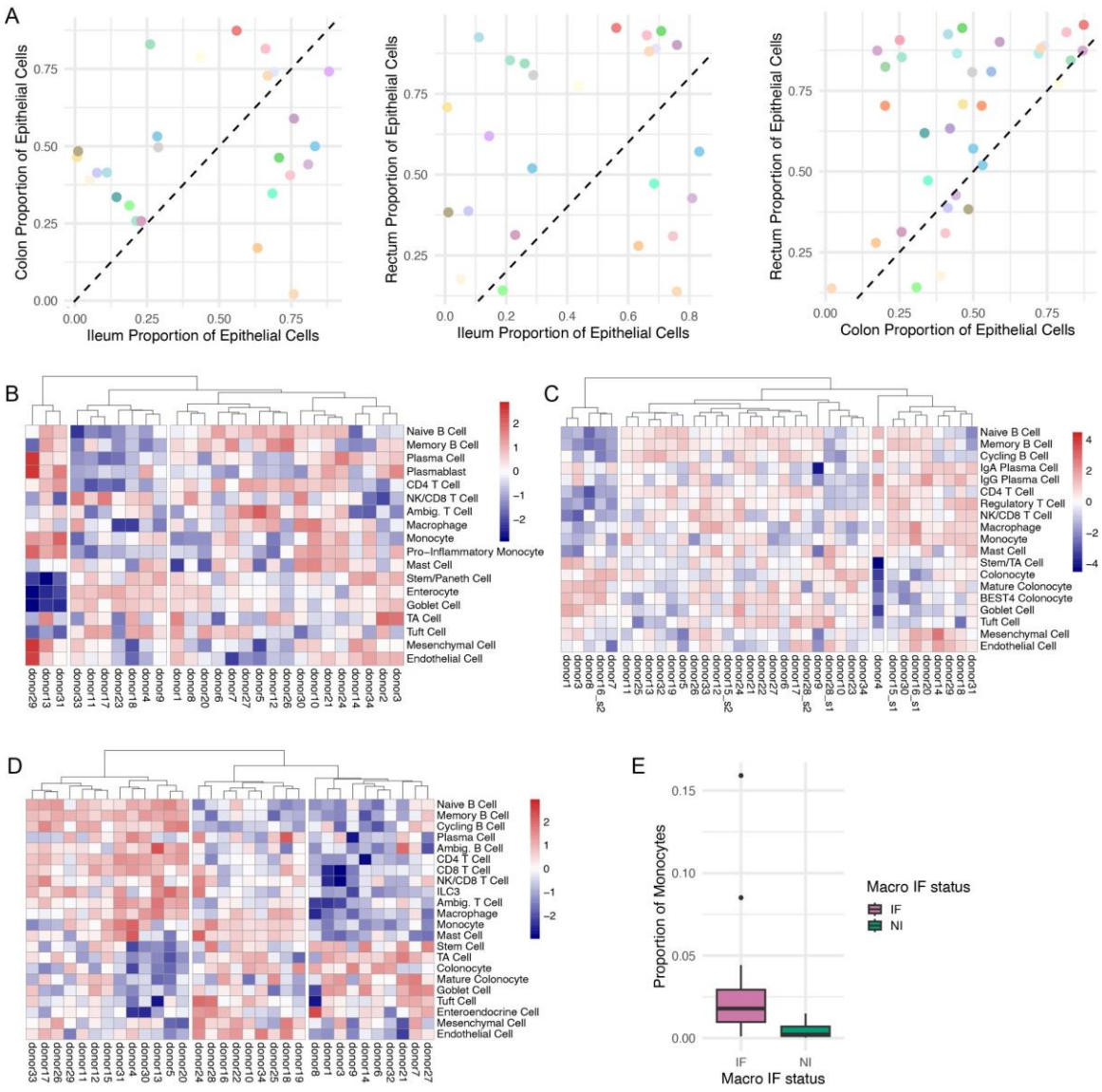


Supplemental Figure 3 Study design and quality control. (A) Overview of study. Total donors = 34 donors. Samples were obtained from the ileum, colon, and rectum, and were processed with 10X Genomics 3' assays. I. = ileum, Ce. = cecum, A.C. = ascending colon, T.C. = transverse colon, D.C. = descending colon, S.C. = sigmoid colon, R. = rectum. Created with BioRender.com. (B) Ileum gene count (left) and mitochondria percent (right) after QC and the stability assessment per sample. Colored by sequencing batch. (C) Colon gene count (left) and mitochondria percent (right) after QC and the stability assessment per sample. Colored by sequencing batch. (D) Rectum gene count (left) and mitochondria percent (right) after QC and the stability assessment per sample. Colored by sequencing batch.



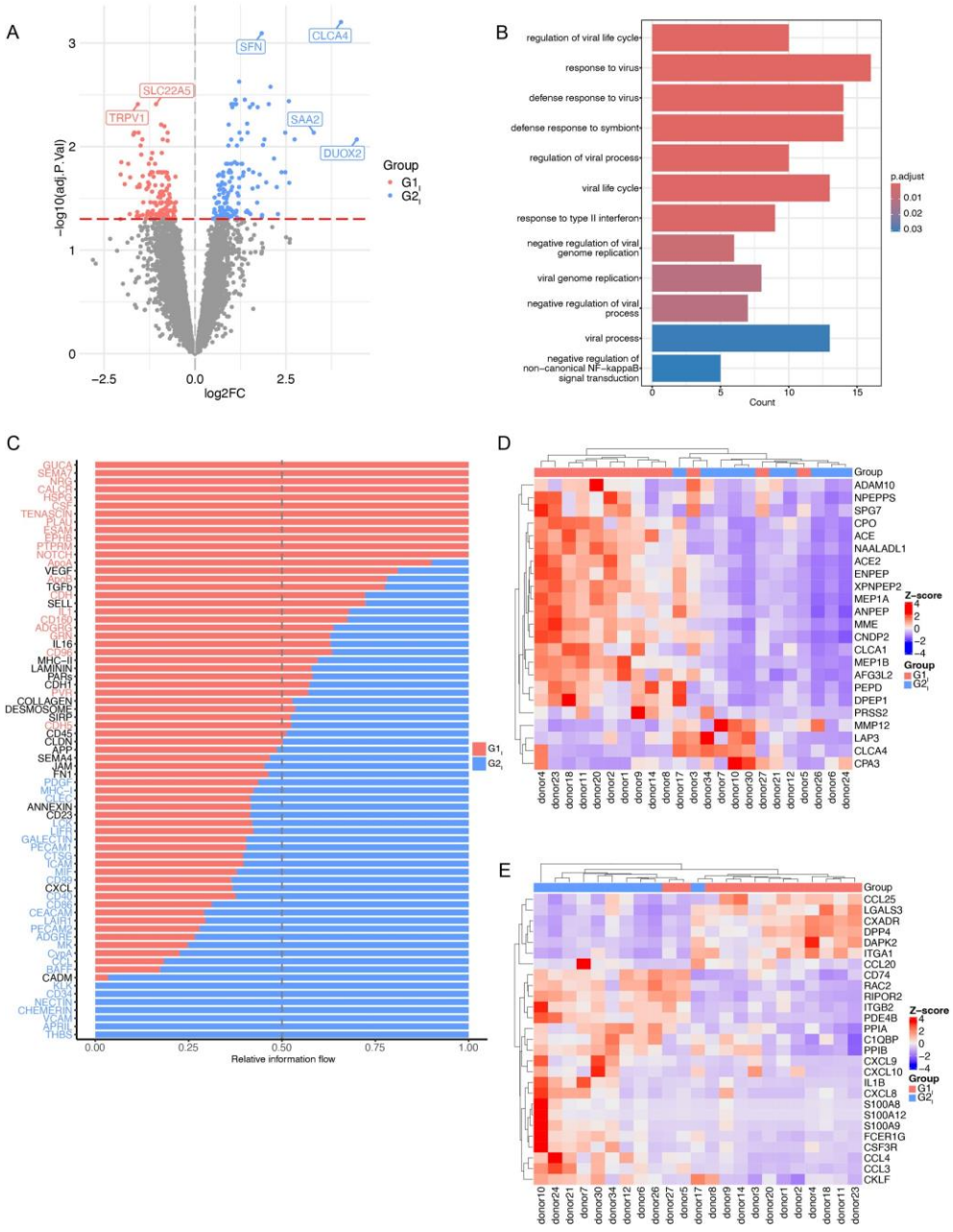
Supplemental Figure 4 Marker genes, ileum high resolution stably assigned cells and ileum reference clustering results. (A) (left) UMAP of the ileum high resolution stably assigned cells clustering results. (right) Stacked bar plots of proportion of cells within each donor. ILC3 = type 3 innate lymphoid cells. (B) (left) UMAP of the ileum

reference clustering results. (right) Stacked bar plots of proportion of cells within each donor. (C) (left) ileum, (middle) colon, and (right) rectum marker genes used to annotate cell types of the stably assigned cell clustering results. Normalized and scaled gene expression counts. % Exp. = Percent expressed, Avg. Exp. = Average Expression, Ambig. = ambiguous, NK = natural killer, Pro-inflam = pro-inflammatory, TA = transit amplifying.

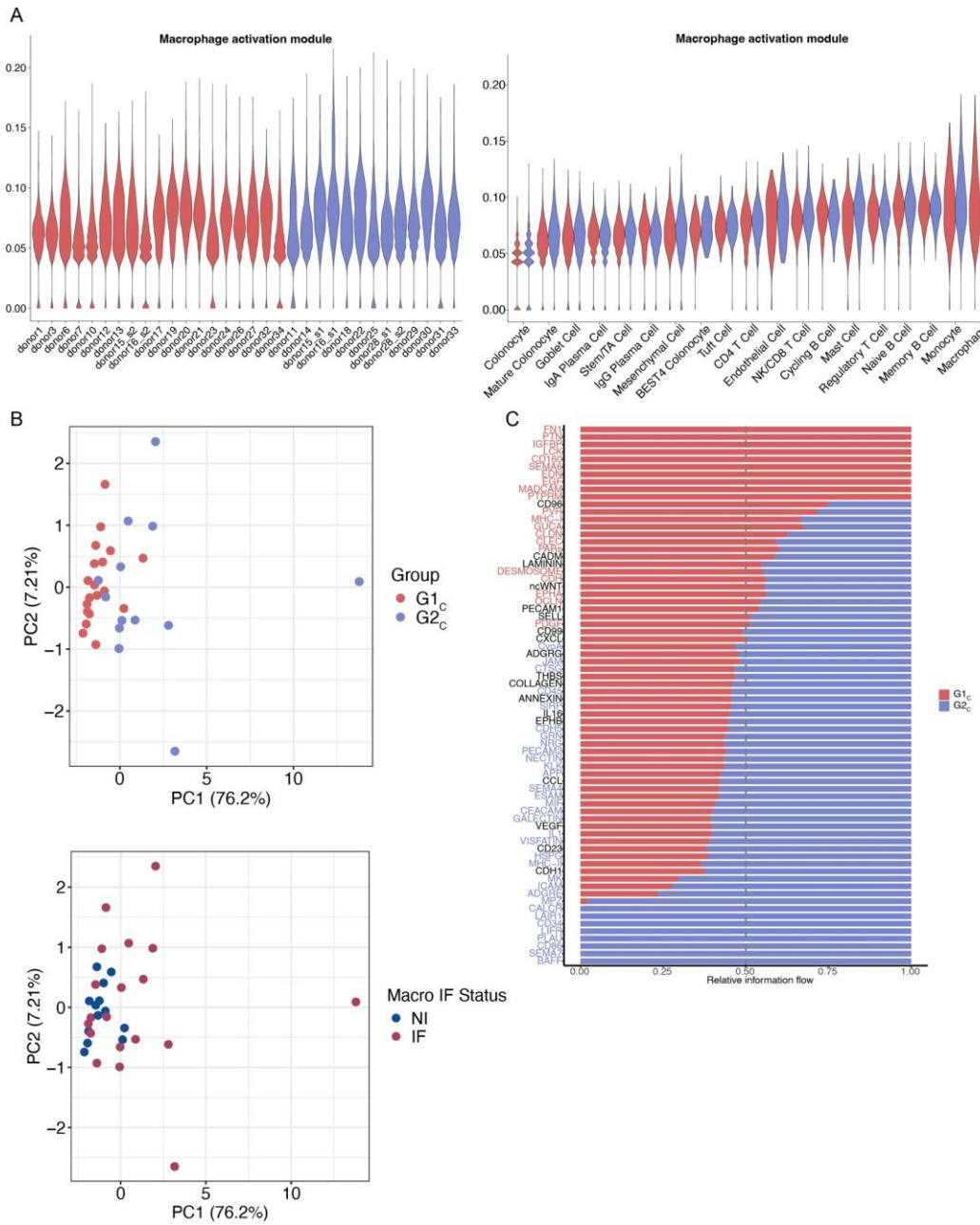


Supplemental Figure 5 Epithelial cell bias across tissue and hierarchical clustering results. (A) (left) Proportion of ileum epithelial cells vs. proportion of colon epithelial cells. (middle) Proportion of ileum epithelial cells vs. proportion of rectum epithelial cells. (right) Proportion of colon epithelial cells vs. proportion of rectum epithelial cells. Each point is colored by sample. Only donors that had a sample in each comparison are plotted. (B) Heatmap of ileum hierarchical clustering results using the proportion of cells per donor. Red indicated high proportion and blue indicated low proportion of cells. Proportions were scaled across donor. (C) Heatmap of colon hierarchical clustering results using the proportion of cells per donor. Red indicated high proportion and blue indicated low proportion of cells. Proportions were scaled across donor. (D) Heatmap of rectum hierarchical clustering results using the proportion of cells per donor. Red indicated high proportion and blue indicated low proportion of cells. Proportions were scaled across donor. (E) Box plot of the

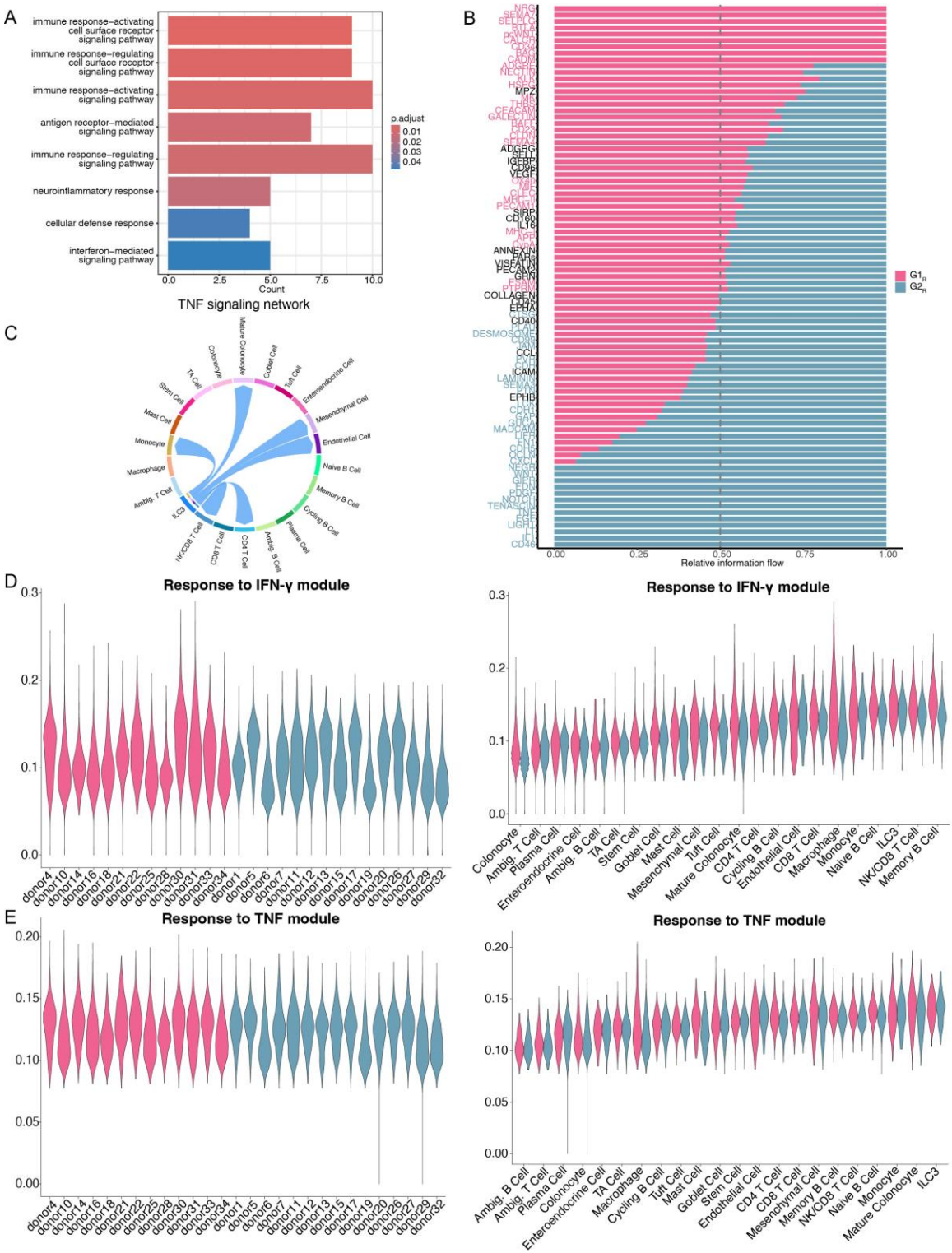
proportion of monocytes in macroscopically inflamed or non-inflamed donors. Shows the median (line), interquartile range (box), and points represent the outliers. Macro IF = macroscopic inflammation status, IF = inflamed, and NI = non-inflamed.



Supplemental Figure 6 Top genes in metallopeptidase activity and neutrophil chemotaxis pathways. (A) Volcano plot of enterocyte DEGs identified in group 1 vs. group 2. (B) Pathways enriched in group 2 enterocytes. P.adjust = Bonferroni adjusted p value. (C) Bar plot of cell signaling pathways from CellChat enriched in group 1 or group 2 donors. Pink pathway names are significantly enriched in group 1 and blue pathway names are significantly enriched in group 2. $G1_I$ = group 1, $G2_I$ = group 2. (D) Normalized and scaled average gene expression counts of top 19 genes in group 1 and top 4 genes in group 2 in metallopeptidase activity module. (E) Normalized and scaled average gene expression counts of top 6 genes in group 1 and top 21 genes from group 2 in neutrophil chemotaxis module.

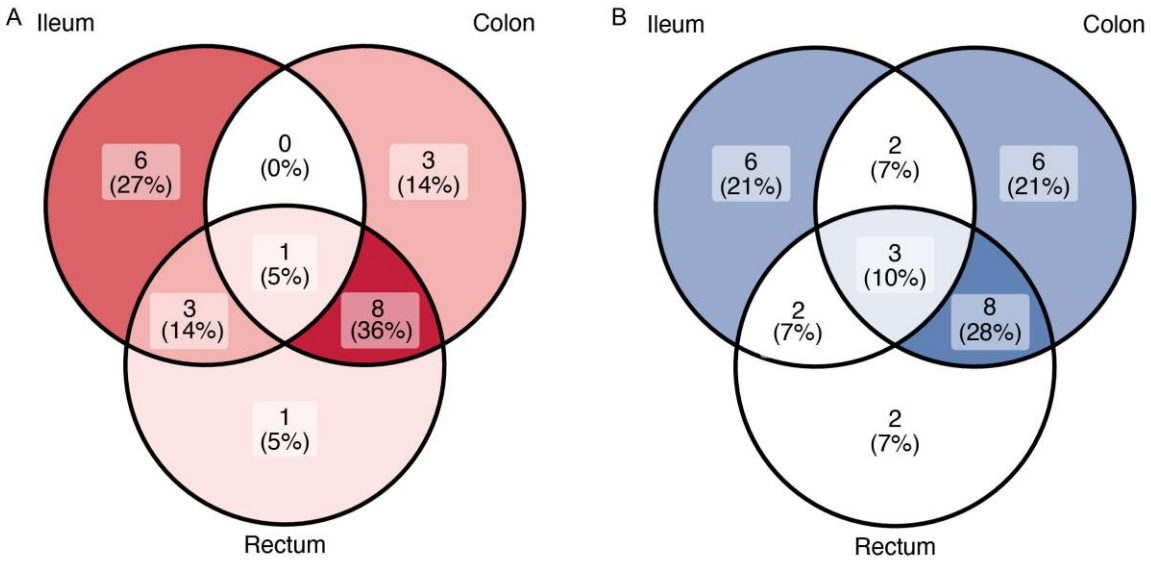


Supplemental Figure 7 Pathway enrichment and macrophage activation in the colon. (A) Violin plot of macrophage activation gene module score per donor (left) and cell type in ascending order (right). (B) PCA of DEGs enriched in macrophage activation pathway using average gene expression of all cells. (top) colored by group status and (bottom) colored by macroscopic IF status. Macro IF = macroscopic inflammation, NI = non-inflamed, IF = inflamed. (C) Bar plot of cell signaling pathways from CellChat enriched in group 1 or group 2 donors. Pink pathway names are significantly enriched in group 1 and blue pathway names are significantly enriched in group 2. $G1_C$ = group 1, $G2_C$ = group 2.

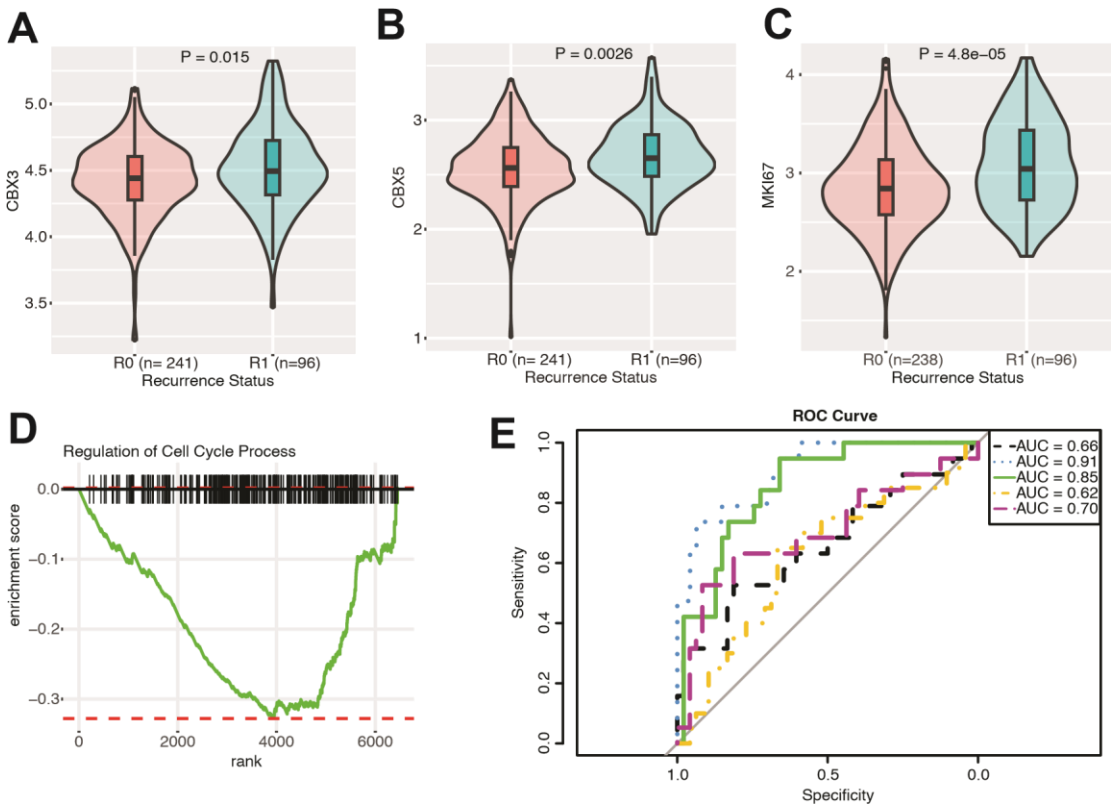


Supplemental Figure 8 Difference in inflammatory pathways in rectum group 1 and 2. (A) Pathways enriched in group 1 myeloid cells. P.adjust = Bonferroni adjusted p value. (B) Bar plot showing cell signaling pathways from CellChat enriched in group

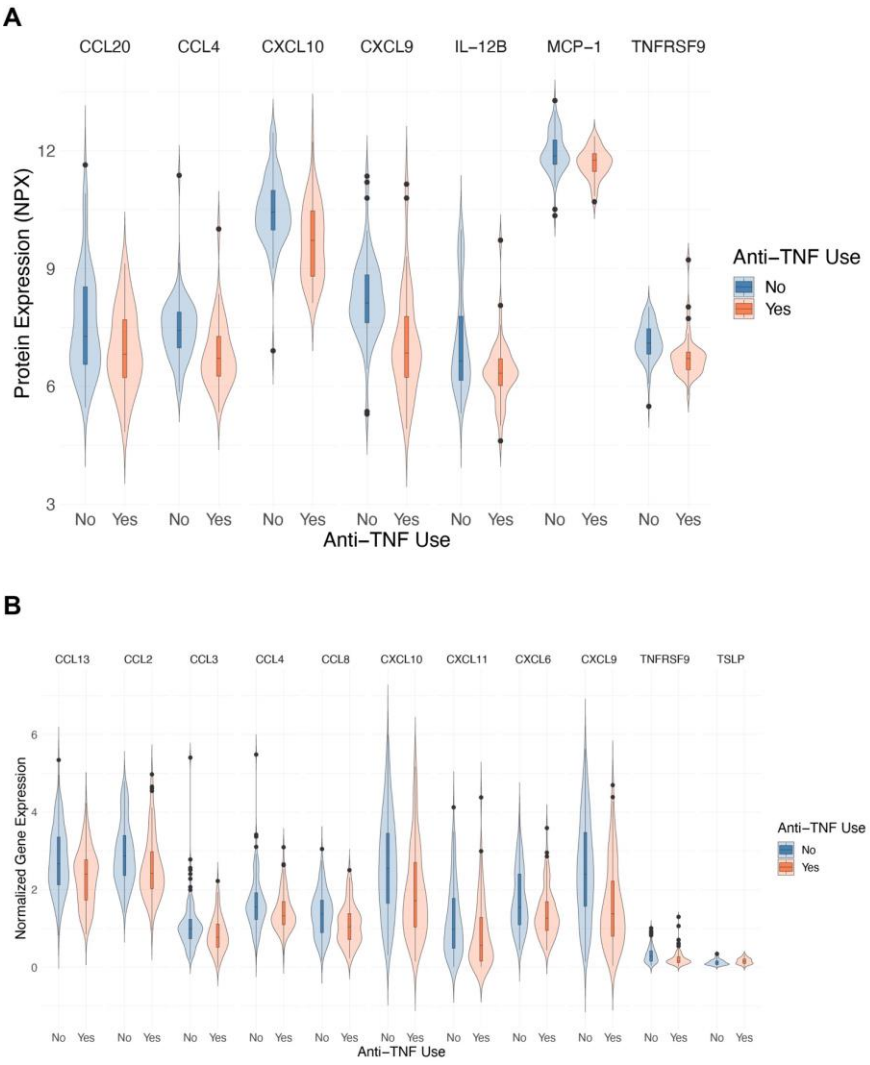
1 or group 2 of the rectum. Pink pathway names are significantly enriched in group 1 and blue pathway names are significantly enriched in group 2. $G1_R$ = group 1 and $G2_R$ = group 2. (C) TNF pathway signaling pathway in group 2. (D) IFN-gamma module score per donor (left) and cell type in ascending order (right). (E) TNF module per donor (left) and cell type in ascending order (right).



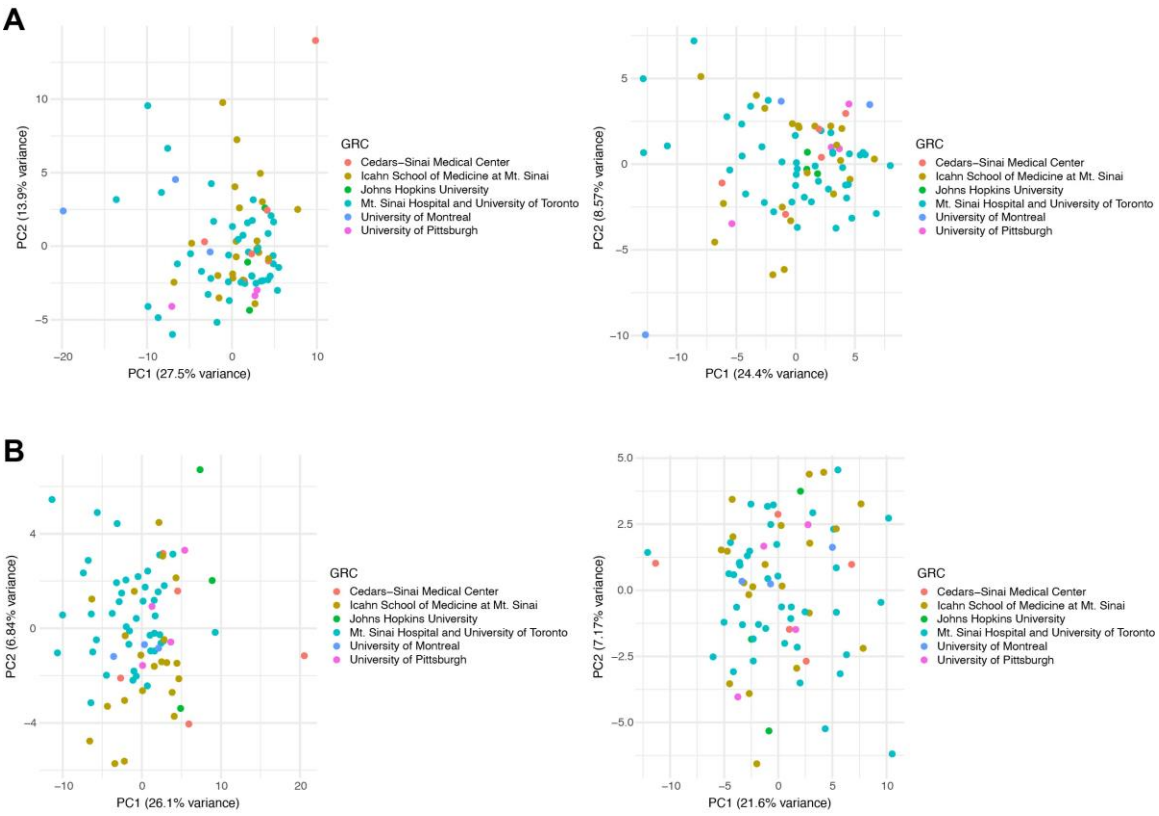
Supplemental Figure 9 Comparison of scITD group membership across tissue. (A) Venn diagram of group membership comparing the "pro-inflammatory" groups. (B) Venn diagram of group membership comparing the "alternative" groups.



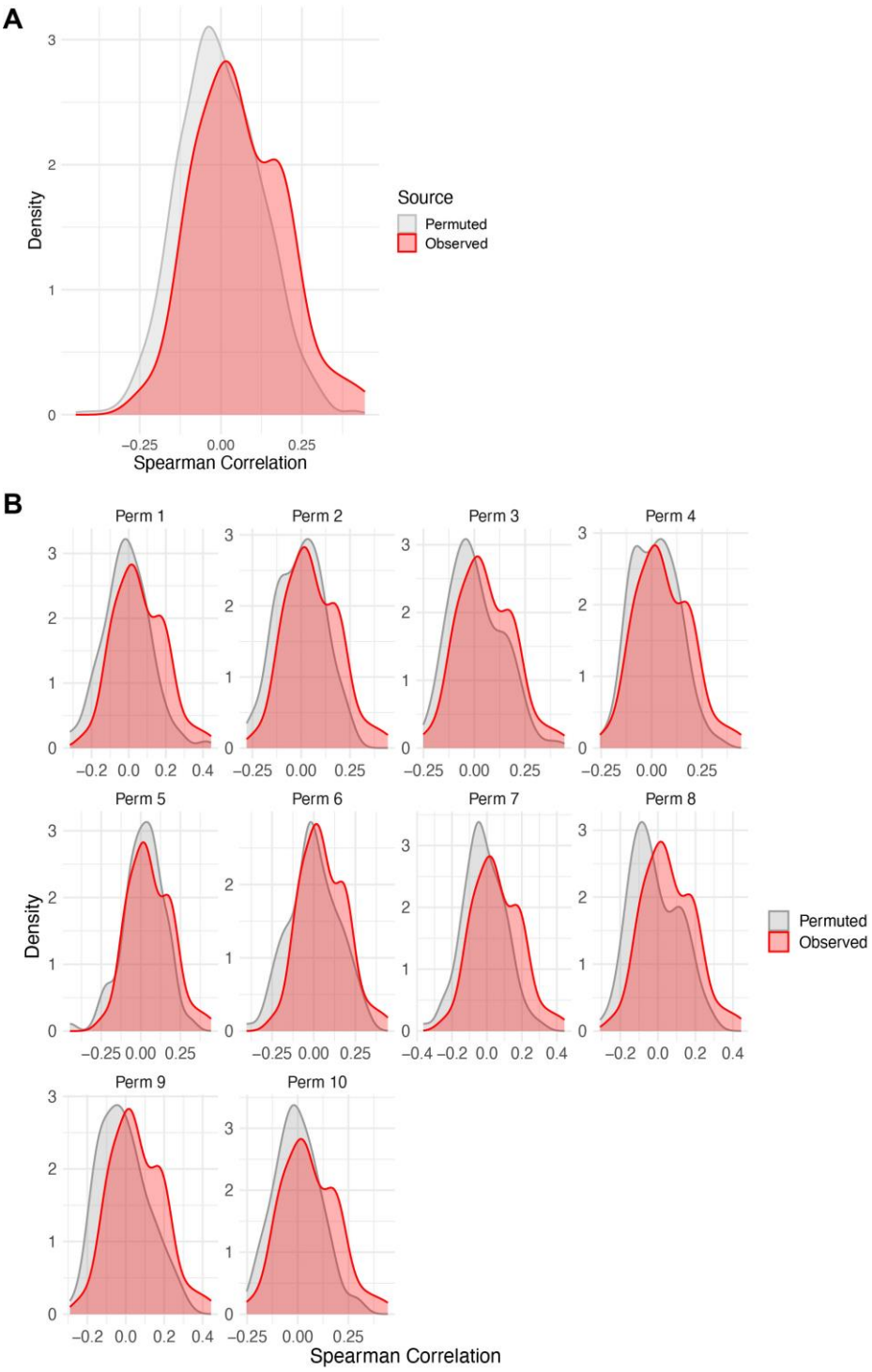
Supplemental Figure 10 Altered splicing signatures of HP1γ in the post-operative ileum dataset. (A) Log2 normalized expression of *CBX3* in R0 vs. R1 samples (Wilcoxon Rank Sum Test). (B) Log2 normalized expression of *CBX5* in R0 vs. R1 samples (Wilcoxon Rank Sum Test). (C) Log2 normalized expression of *MKI67* in R0 vs. R1 samples (Wilcoxon Rank Sum Test). (D) Regulation of cell cycle process is enriched in R1, identified by GSEA and ranked by log2FC. (E) ROC curves for five folds of cross-validation of logistic regression models with PC2, PC1, PC4, sex, batch, smoking status, self-reported population, and age at sample collection.



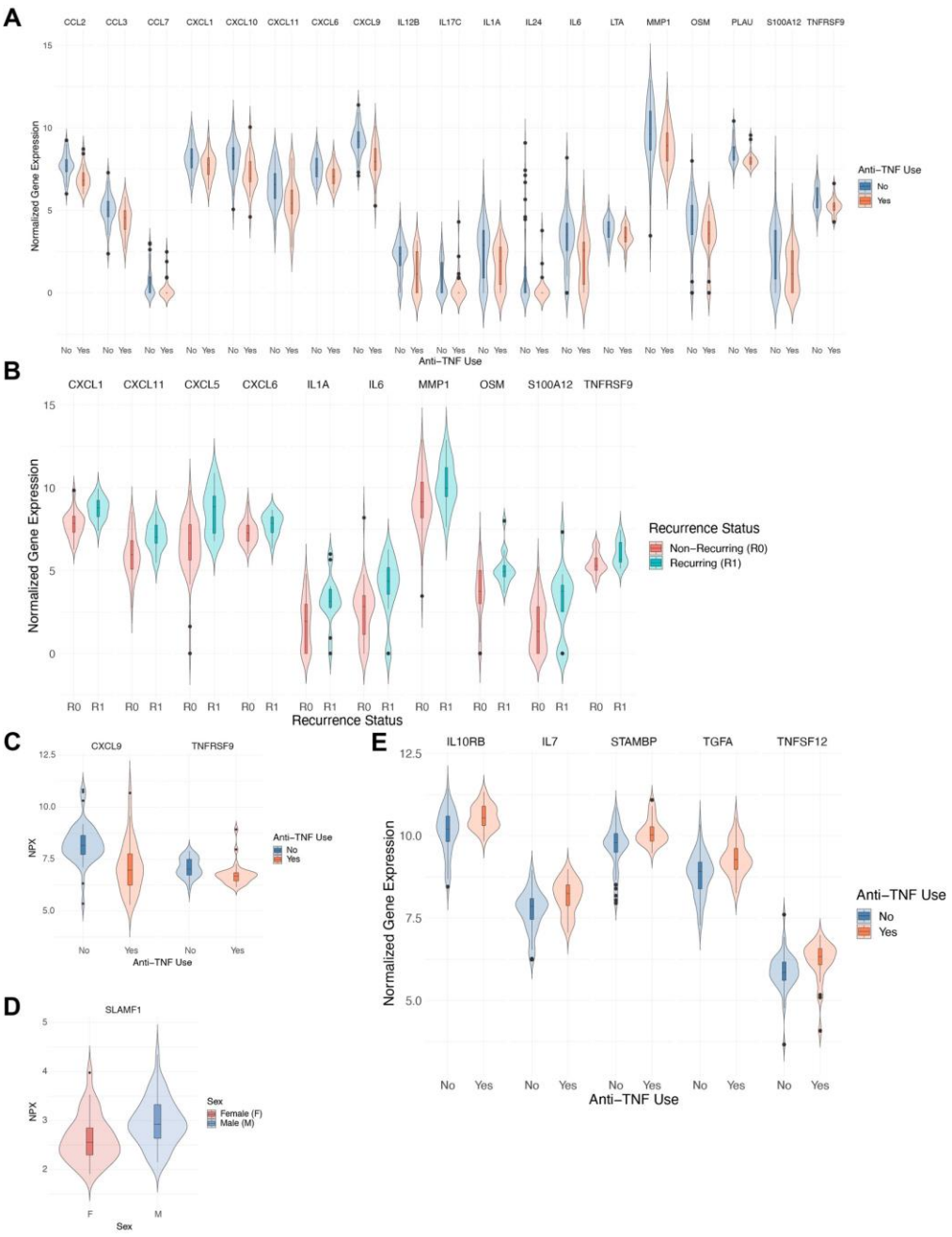
Supplemental Figure 11 Differential expression of proteins and genes associated with anti-TNF use. (A) Differentially expressed proteins associated with anti-TNF use. NPX = normalized protein expression. (B) Differentially expressed genes associated with anti-TNF use.



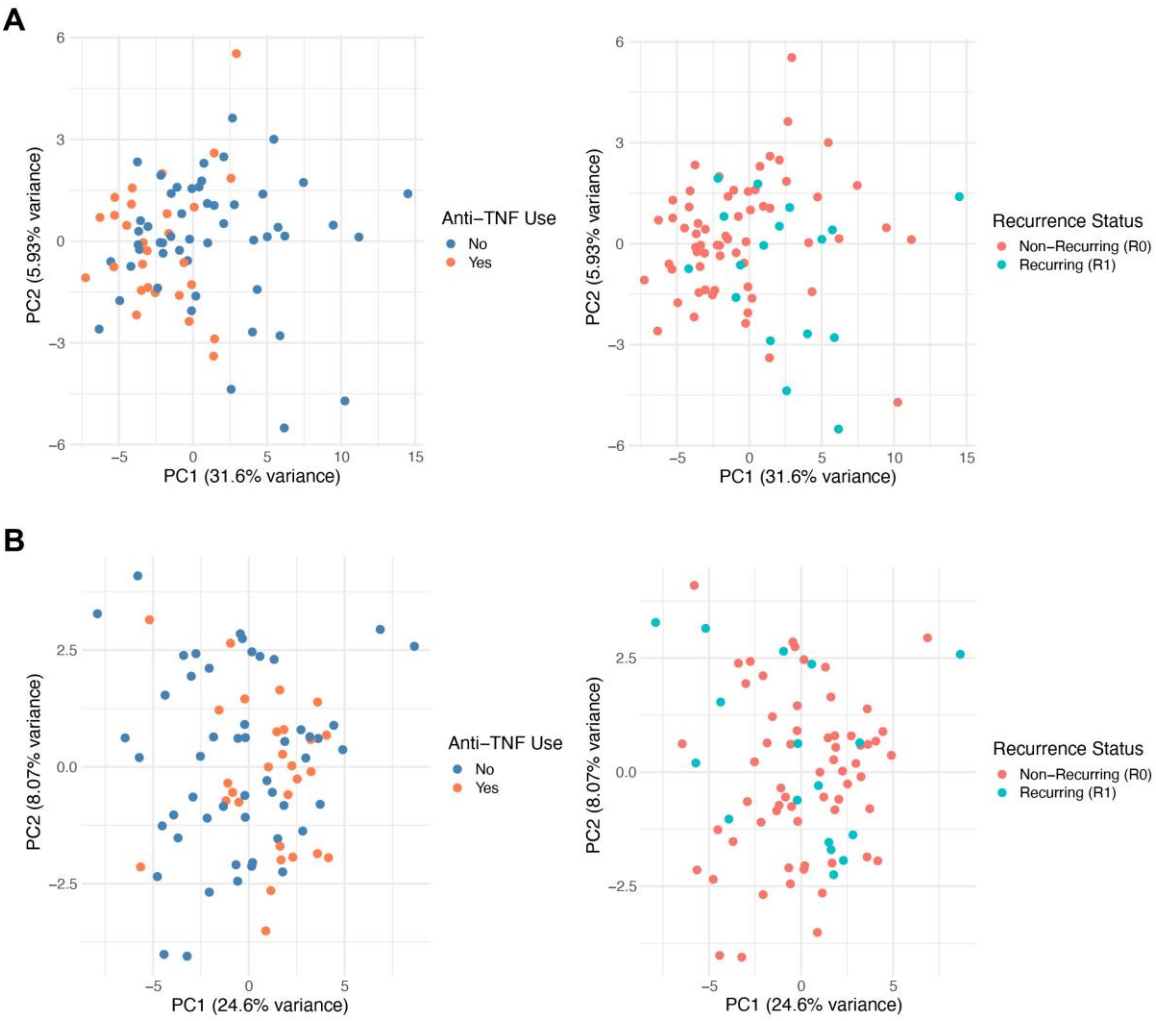
Supplemental Figure 12 Batch effect correction for transcriptomic and proteomic data. (A) PCA of gene expression colored by GRC. (Left) before batch effect correction. (right) after batch effect correction. (B) PCA of protein expression colored by GRC. (Left) before batch effect correction. (Right) after batch effect correction. GRC = genetic research center, PCA = principal component analysis, and PC = principal component.



Supplemental Figure 13 Permutation analysis of gene-protein correlation. (A) Overall distribution of observed spearman correlation between gene-protein pairs (red) and permuted spearman correlation between gene-protein pairs (grey). (B) Distribution of observed spearman correlation between gene-protein pairs and each permutation. Perm = permuted.



Supplemental Figure 14 Group genes or proteins associated with recurrence status, anti-TNF use, and sex. (A) Group A gene expression significantly associated with anti-TNF use. (B) Group A gene expression significantly associated with recurrence status (C) Group 1 protein expression significantly associated with anti-TNF use. (D) Group 1 protein expression significantly associated with sex. (E) Group B gene expression significantly associated with anti-TNF use. R0 = non-recurring, R1 = recurring, F = female, M = male, NPX = normalized protein expression.



Supplemental Figure 15 PCA analysis of Group A and B gene expression and Group 1 and 2 protein expression. (A) PCA of Group A and B genes. (B) PCA of Group 1 and 2 proteins. (A and B) (Left) colored by anti-TNF use. (Right). Colored by recurrence status.

REFERENCES

1. Torres, J., et al., *Crohn's disease.* Lancet, 2017. **389**(10080): p. 1741-1755.

2. Jarmakiewicz-Czaja, S., et al., *Genetic and Epigenetic Etiology of Inflammatory Bowel Disease: An Update.* Genes (Basel), 2022. **13**(12).

3. McGovern, D.P., S. Kugathasan, and J.H. Cho, *Genetics of Inflammatory Bowel Diseases.* Gastroenterology, 2015. **149**(5): p. 1163-1176 e2.

4. Jorgensen, J.T., *Twenty Years with Personalized Medicine: Past, Present, and Future of Individualized Pharmacotherapy.* Oncologist, 2019. **24**(7): p. e432-e440.

5. Redekop, W.K. and D. Mladsi, *The faces of personalized medicine: a framework for understanding its meaning and scope.* Value Health, 2013. **16**(6 Suppl): p. S4-9.

6. Ashley, E.A., *Towards precision medicine.* Nat Rev Genet, 2016. **17**(9): p. 507-22.

7. Ramos, G.P. and K.A. Papadakis, *Mechanisms of Disease: Inflammatory Bowel Diseases.* Mayo Clin Proc, 2019. **94**(1): p. 155-165.

8. Baumgart, D.C. and W.J. Sandborn, *Crohn's disease.* Lancet, 2012. **380**(9853): p. 1590-605.

9. Roda, G., et al., *Crohn's disease.* Nat Rev Dis Primers, 2020. **6**(1): p. 22.

10. Hracs, L., et al., *Global evolution of inflammatory bowel disease across epidemiologic stages.* Nature, 2025.

11. Satsangi, J., et al., *The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications.* Gut, 2006. **55**(6): p. 749-53.

12. Sleiman, J., et al., *Prevention and Treatment of Stricturing Crohn's Disease - Perspectives and Challenges.* Expert Rev Gastroenterol Hepatol, 2021. **15**(4): p. 401-411.

13. Hirten, R.P., et al., *The Management of Intestinal Penetrating Crohn's Disease.* Inflamm Bowel Dis, 2018. **24**(4): p. 752-765.

14. Fan, Y., et al., *Patients With Stricturing or Penetrating Crohn's Disease Phenotypes Report High Disease Burden and Treatment Needs.* Inflamm Bowel Dis, 2023. **29**(6): p. 914-922.

15. Singh, A., et al., *Management of Perianal Fistulizing Crohn's Disease.* Inflamm Bowel Dis, 2024. **30**(9): p. 1579-1603.

16. Ruffolo, C., et al., *Perianal Crohn's disease: is there something new?* World J Gastroenterol, 2011. **17**(15): p. 1939-46.

17. Atreya, R. and B. Siegmund, *Location is important: differentiation between ileal and colonic Crohn's disease.* Nat Rev Gastroenterol Hepatol, 2021. **18**(8): p. 544-558.

18. Atreya, R., et al., *Ileal and colonic Crohn's disease: Does location makes a difference in therapy efficacy?* Curr Res Pharmacol Drug Discov, 2022. **3**: p. 100097.

19. Van Limbergen, J., et al., *Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease.* Gastroenterology, 2008. **135**(4): p. 1114-22.

20. Kelsen, J. and R.N. Baldassano, *Inflammatory bowel disease: the difference between children and adults.* Inflamm Bowel Dis, 2008. **14 Suppl 2**: p. S9-11.

21. Saez, A., et al., *Pathophysiology of Inflammatory Bowel Disease: Innate Immune System.* Int J Mol Sci, 2023. **24**(2).

22. Hegarty, L.M., G.R. Jones, and C.C. Bain, *Macrophages in intestinal homeostasis and inflammatory bowel disease.* Nat Rev Gastroenterol Hepatol, 2023. **20**(8): p. 538-553.

23. Tai, S.L. and A. Mortha, *Macrophage control of Crohn's disease.* Int Rev Cell Mol Biol, 2022. **367**: p. 29-64.

24. Cushing, K. and P.D.R. Higgins, *Management of Crohn Disease: A Review.* JAMA, 2021. **325**(1): p. 69-80.

25. Billmeier, U., et al., *Molecular mechanism of action of anti-tumor necrosis factor antibodies in inflammatory bowel diseases.* World J Gastroenterol, 2016. **22**(42): p. 9300-9313.

26. Evangelatos, G., et al., *The second decade of anti-TNF-a therapy in clinical practice: new lessons and future directions in the COVID-19 era.* Rheumatol Int, 2022. **42**(9): p. 1493-1511.

27. Cui, G., et al., *Evaluation of anti-TNF therapeutic response in patients with inflammatory bowel disease: Current and novel biomarkers.* EBioMedicine, 2021. **66**: p. 103329.

28. Cheah, E. and J.G. Huang, *Precision medicine in inflammatory bowel disease: Individualizing the use of biologics and small molecule therapies.* World J Gastroenterol, 2023. **29**(10): p. 1539-1550.

29. Goll, R., et al., *Pharmacodynamic mechanisms behind a refractory state in inflammatory bowel disease.* BMC Gastroenterol, 2022. **22**(1): p. 464.

30. Gajendran, M., et al., *A comprehensive review and update on Crohn's disease.* Dis Mon, 2018. **64**(2): p. 20-57.

31. Lewis, R.T. and D.J. Maron, *Efficacy and complications of surgery for Crohn's disease.* Gastroenterol Hepatol (N Y), 2010. **6**(9): p. 587-96.

32. Ahmed Ali, U. and R.P. Kiran, *Surgery for Crohn's disease: upfront or last resort?* Gastroenterol Rep (Oxf), 2022. **10**: p. goac063.

33. Shah, R.S. and B.H. Click, *Medical therapies for postoperative Crohn's disease.* Therap Adv Gastroenterol, 2021. **14**: p. 1756284821993581.

34. Seo, J., et al., *Fecal Calprotectin in Patients with Crohn's Disease: A Study Based on the History of Bowel Resection and Location of Disease.* Diagnostics (Basel), 2024. **14**(8).

35. Dasharathy, S.S., B.N. Limketkai, and J.S. Sauk, *What's New in the Postoperative Management of Crohn's Disease?* Dig Dis Sci, 2022. **67**(8): p. 3508-3517.

36. Liu, J.Z. and C.A. Anderson, *Genetic studies of Crohn's disease: past, present and future.* Best Pract Res Clin Gastroenterol, 2014. **28**(3): p. 373-86.

37. Hugot, J.P., et al., *Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease.* Nature, 2001. **411**(6837): p. 599-603.

38. Ogura, Y., et al., *A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease.* Nature, 2001. **411**(6837): p. 603-6.

39. Gordon, H., et al., *Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies.* Inflamm Bowel Dis, 2015. **21**(6): p. 1428-34.

40. Chen, G.B., et al., *Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data.* Hum Mol Genet, 2014. **23**(17): p. 4710-20.

41. Payne, S.H., *The utility of protein and mRNA correlation.* Trends Biochem Sci, 2015. **40**(1): p. 1-3.

42. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

43. Conesa, A., et al., *Erratum to: A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**(1): p. 181.

44. Liu, Q., L. Fang, and C. Wu, *Alternative Splicing and Isoforms: From Mechanisms to Diseases.* Genes (Basel), 2022. **13**(3).

45. Zou, C., et al., *Crosstalk between alternative splicing and inflammatory bowel disease: Basic mechanisms, biotechnological progresses and future perspectives.* Clin Transl Med, 2023. **13**(11): p. e1479.

46. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.

47. Holgersen, K., et al., *High-resolution gene expression profiling using RNA sequencing in patients with inflammatory bowel disease and in mouse models of colitis.* J Crohns Colitis, 2015. **9**(6): p. 492-506.

48. Hong, S.N., et al., *RNA-seq Reveals Transcriptomic Differences in Inflamed and Noninflamed Intestinal Mucosa of Crohn's Disease Patients Compared with Normal Mucosa of Healthy Controls.* Inflamm Bowel Dis, 2017. **23**(7): p. 1098-1108.

49. Xu, L., et al., *Bulk and single-cell RNA sequencing reveal the roles of neutrophils in pediatric Crohn's disease.* Pediatr Res, 2025.

50. Kim, K., et al., *Transcriptomic Profiling and Cellular Composition of Creeping Fat in Crohn's disease.* J Crohns Colitis, 2024. **18**(2): p. 223-232.

51. Chen, K.A., et al., *Post-operative Crohn's Disease Recurrence and Infectious Complications: A Transcriptomic Analysis.* Dig Dis Sci, 2025. **70**(1): p. 203-214.

52. Ashton, J.J., et al., *Ileal Transcriptomic Analysis in Paediatric Crohn's Disease Reveals IL17- and NOD-signalling Expression Signatures in Treatment-naive Patients and Identifies Epithelial Cells Driving Differentially Expressed Genes.* J Crohns Colitis, 2021. **15**(5): p. 774-786.

53. Kugathasan, S., et al., *Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study.* Lancet, 2017. **389**(10080): p. 1710-1718.

54. Tavares de Sousa, H., et al., *Fibrosis-related Transcriptome Unveils a Distinctive Remodelling Matrix Pattern in Penetrating Ileal Crohn's Disease.* J Crohns Colitis, 2024. **18**(11): p. 1741-1752.

55. Haberman, Y., et al., *Corrigendum. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature.* J Clin Invest, 2015. **125**(3): p. 1363.

56. Wu, J., et al., *Immunological profile of lactylation-related genes in Crohn's disease: a comprehensive analysis based on bulk and single-cell RNA sequencing data.* J Transl Med, 2024. **22**(1): p. 300.

57. Braun, T., et al., *Diet-omics in the Study of Urban and Rural Crohn disease Evolution (SOURCE) cohort.* Nat Commun, 2024. **15**(1): p. 3764.

58. Gibson, G., et al., *Eleven Grand Challenges for Inflammatory Bowel Disease Genetics and Genomics.* Inflamm Bowel Dis, 2025. **31**(1): p. 272-284.

59. Li, D., et al., *An alternative splicing signature in human Crohn's disease.* BMC Gastroenterol, 2021. **21**(1): p. 420.

60. Hasler, R., et al., *Alterations of pre-mRNA splicing in human inflammatory bowel disease.* Eur J Cell Biol, 2011. **90**(6-7): p. 603-11.

61. Rankin, C.R., et al., *The IBD-associated long noncoding RNA IFNG-AS1 regulates the balance between inflammatory and anti-inflammatory cytokine production after T-cell stimulation.* Am J Physiol Gastrointest Liver Physiol, 2020. **318**(1): p. G34-G40.

62. Pai, Y.C., et al., *Gut microbial transcytosis induced by tumor necrosis factor-like 1A-dependent activation of a myosin light chain kinase splice variant contributes to IBD.* J Crohns Colitis, 2020. **15**(2): p. 258-72.

63. Bootz, F., A.S. Schmid, and D. Neri, *Alternatively Spliced EDA Domain of Fibronectin Is a Target for Pharmacodelivery Applications in Inflammatory Bowel Disease.* Inflamm Bowel Dis, 2015. **21**(8): p. 1908-17.

64. Mata-Garrido, J., et al., *The Heterochromatin protein 1 is a regulator in RNA splicing precision deficient in ulcerative colitis.* Nat Commun, 2022. **13**(1): p. 6834.

65. Rachez, C., et al., *HP1gamma binding pre-mRNA intronic repeats modulates RNA splicing decisions.* EMBO Rep, 2021. **22**(9): p. e52320.

66. Eriksson, M., et al., *Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome.* Nature, 2003. **423**(6937): p. 293-8.

67. Berger, K., et al., *Altered splicing associated with the pathology of inflammatory bowel disease.* Hum Genomics, 2021. **15**(1): p. 47.

68. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nat Methods, 2009. **6**(5): p. 377-82.

69. Jovic, D., et al., *Single-cell RNA sequencing technologies and applications: A brief overview.* Clin Transl Med, 2022. **12**(3): p. e694.

70. Haque, A., et al., *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.* Genome Med, 2017. **9**(1): p. 75.

71. Regev, A., et al., *The Human Cell Atlas.* Elife, 2017. **6**.

72. Heumos, L., et al., *Best practices for single-cell analysis across modalities.* Nat Rev Genet, 2023. **24**(8): p. 550-572.

73. Guruprasad, P., et al., *The current landscape of single-cell transcriptomics for cancer immunotherapy.* J Exp Med, 2021. **218**(1).

74. Andrews, T.S., et al., *Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data.* Nat Protoc, 2021. **16**(1): p. 1-9.

75. Su, M., et al., *Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications.* Mil Med Res, 2022. **9**(1): p. 68.

76. Luecken, M.D. and F.J. Theis, *Current best practices in single-cell RNA-seq analysis: a tutorial.* Mol Syst Biol, 2019. **15**(6): p. e8746.

77. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587 e29.

78. Jin, S., et al., *Inference and analysis of cell-cell communication using CellChat.* Nat Commun, 2021. **12**(1): p. 1088.

79. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.

80. Kiselev, V.Y., T.S. Andrews, and M. Hemberg, *Challenges in unsupervised clustering of single-cell RNA-seq data.* Nat Rev Genet, 2019. **20**(5): p. 273-282.

81. Peyvandipour, A., et al., *Identification of cell types from single cell data using stable clustering.* Sci Rep, 2020. **10**(1): p. 12349.

82. Gibson, G., *Perspectives on rigor and reproducibility in single cell genomics.* PLoS Genet, 2022. **18**(5): p. e1010210.

83. Grabski, I.N., K. Street, and R.A. Irizarry, *Significance analysis for clustering with single-cell RNA-sequencing data.* Nat Methods, 2023. **20**(8): p. 1196-1202.

84. Tasic, B., et al., *Adult mouse cortical cell taxonomy revealed by single cell transcriptomics.* Nat Neurosci, 2016. **19**(2): p. 335-46.

85. Tang, M., et al., *Evaluating single-cell cluster stability using the Jaccard similarity index.* Bioinformatics, 2021. **37**(15): p. 2212-2214.

86. Elmentaite, R., et al., *Cells of the human intestinal tract mapped across space and time.* Nature, 2021. **597**(7875): p. 250-255.

87. Elmentaite, R., et al., *Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease.* Dev Cell, 2020. **55**(6): p. 771-783 e5.

88. Parikh, K., et al., *Colonic epithelial cell diversity in health and inflammatory bowel disease.* Nature, 2019. **567**(7746): p. 49-55.

89. Smillie, C.S., et al., *Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis.* Cell, 2019. **178**(3): p. 714-730 e22.

90. Kanke, M., et al., *Single-Cell Analysis Reveals Unexpected Cellular Changes and Transposon Expression Signatures in the Colonic Epithelium of Treatment-Naive Adult Crohn's Disease Patients.* Cell Mol Gastroenterol Hepatol, 2022. **13**(6): p. 1717-1740.

91. Cario, E. and D.K. Podolsky, *Differential alteration in intestinal epithelial cell expression of toll-like receptor 3 (TLR3) and TLR4 in inflammatory bowel disease.* Infect Immun, 2000. **68**(12): p. 7010-7.

92. Oliver, A.J., et al., *Single-cell integration reveals metaplasia in inflammatory gut diseases.* Nature, 2024. **635**(8039): p. 699-707.

93. Garrido-Trigo, A., et al., *Author Correction: Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease.* Nat Commun, 2024. **15**(1): p. 857.

94. Jaeger, N., et al., *Single-cell analyses of Crohn's disease tissues reveal intestinal intraepithelial T cells heterogeneity and altered subset distributions.* Nat Commun, 2021. **12**(1): p. 1921.

95. Brown, M., et al., *Concordant B and T Cell Heterogeneity Inferred from the Multiomic Landscape of Peripheral Blood Mononuclear Cells in a Crohn's Disease Cohort.* J Crohns Colitis, 2024. **18**(12): p. 1939-56.

96. Martin, J.C., et al., *Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy.* Cell, 2019. **178**(6): p. 1493-1508 e20.

97. Kong, L., et al., *The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon.* Immunity, 2023. **56**(12): p. 2855.

98. Thomas, T., et al., *A longitudinal single-cell atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease.* Nat Immunol, 2024. **25**(11): p. 2152-2165.

99. Adler, J., et al., *Perianal Crohn Disease in a Large Multicenter Pediatric Collaborative.* J Pediatr Gastroenterol Nutr, 2017. **64**(5): p. e117-e124.

100. Alli-Akintade, L., et al., *Race and fistulizing perianal Crohn's disease.* J Clin Gastroenterol, 2015. **49**(3): p. e21-3.

101. Mo, A., et al., *African Ancestry Proportion Influences Ileal Gene Expression in Inflammatory Bowel Disease.* Cell Mol Gastroenterol Hepatol, 2020. **10**(1): p. 203-205.

102. Levantovsky, R.M., et al., *Multimodal single-cell analyses reveal mechanisms of perianal fistula in diverse patients with Crohn's disease.* Med, 2024. **5**(8): p. 886-908 e11.

103. Cao, S., et al., *Single-Cell and Spatial Multi-omics Reveal Interferon Signaling in the Pathogenesis of Perianal Fistulizing Crohn's Disease.* bioRxiv, 2024.

104. Zhang, Y., et al., *TWIST1+FAP+ fibroblasts in the pathogenesis of intestinal fibrosis in Crohn's disease.* J Clin Invest, 2024. **134**(18).

105. Humphreys, D.T., et al., *Single cell sequencing data identify distinct B cell and fibroblast populations in stricturing Crohn's disease.* J Cell Mol Med, 2024. **28**(9): p. e18344.

106. Mukherjee, P.K., et al., *Stricturing Crohn's Disease Single-Cell RNA Sequencing Reveals Fibroblast Heterogeneity and Intercellular Interactions.* Gastroenterology, 2023. **165**(5): p. 1180-1196.

107. Aggeletopoulou, I., et al., *Creeping Fat in Crohn's Disease-Surgical, Histological, and Radiological Approaches.* J Pers Med, 2023. **13**(7).

108. Borley, N.R., et al., *The relationship between inflammatory and serosal connective tissue changes in ileal Crohn's disease: evidence for a possible causative link.* J Pathol, 2000. **190**(2): p. 196-202.

109. Shu, W., et al., *Single-cell Expression Atlas Reveals Cell Heterogeneity in the Creeping Fat of Crohn's Disease.* Inflamm Bowel Dis, 2023. **29**(6): p. 850-865.

110. Wu, J., et al., *Microbiota-induced alteration of kynurenine metabolism in macrophages drives formation of creeping fat in Crohn's disease.* Cell Host Microbe, 2024. **32**(11): p. 1927-1943 e9.

111. Nie, H., et al., *Single-cell meta-analysis of inflammatory bowel disease with scIBD.* Nat Comput Sci, 2023. **3**(6): p. 522-531.

112. Cui, M., C. Cheng, and L. Zhang, *High-throughput proteomics: a methodological mini-review.* Lab Invest, 2022. **102**(11): p. 1170-1181.

113. Al-Amrani, S., et al., *Proteomics: Concepts and applications in human medicine.* World J Biol Chem, 2021. **12**(5): p. 57-69.

114. Assarsson, E., et al., *Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability.* PLoS One, 2014. **9**(4): p. e95192.

115. Wang, H., et al., *Methods and clinical biomarker discovery for targeted proteomics using Olink technology.* Proteomics Clin Appl, 2024. **18**(5): p. e2300233.

116. Sun, B.B., et al., *Plasma proteomic associations with genetics and health in the UK Biobank.* Nature, 2023. **622**(7982): p. 329-338.

117. Sun, B.B., et al., *Genomic atlas of the human plasma proteome.* Nature, 2018. **558**(7708): p. 73-79.

118. Eldjarn, G.H., et al., *Large-scale plasma proteomics comparisons through genetics and disease associations.* Nature, 2023. **622**(7982): p. 348-358.

119. Zhang, X., et al., *Associations of 2923 plasma proteins with incident inflammatory bowel disease in a prospective cohort study and genetic analysis.* Nat Commun, 2025. **16**(1): p. 2813.

120. Liu, Y., A. Beyer, and R. Aebersold, *On the Dependency of Cellular Protein Levels on mRNA Abundance.* Cell, 2016. **165**(3): p. 535-50.

121. Buccitelli, C. and M. Selbach, *mRNAs, proteins and the emerging principles of gene expression control.* Nat Rev Genet, 2020. **21**(10): p. 630-644.

122. Wang, D., et al., *A deep proteome and transcriptome abundance atlas of 29 healthy human tissues.* Mol Syst Biol, 2019. **15**(2): p. e8503.

123. Qie, J., et al., *Integrated proteomic and transcriptomic landscape of macrophages in mouse tissues.* Nat Commun, 2022. **13**(1): p. 7389.

124. Koussounadis, A., et al., *Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system.* Sci Rep, 2015. **5**: p. 10775.

125. Kosti, I., et al., *Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues.* Sci Rep, 2016. **6**: p. 24799.

126. Wegler, C., et al., *Global variability analysis of mRNA and protein concentrations across and within human tissues.* NAR Genom Bioinform, 2020. **2**(1): p. lqz010.

127. Kalla, R., et al., *Serum proteomic profiling at diagnosis predicts clinical course, and need for intensification of treatment in inflammatory bowel disease.* J Crohns Colitis, 2021. **15**(5): p. 699-708.

128. Walshe, M., et al., *A Role for CXCR3 Ligands as Biomarkers of Post-Operative Crohn's Disease Recurrence.* J Crohns Colitis, 2022. **16**(6): p. 900-910.

129. Hernandez-Rocha, C., et al., *Clinical Predictors of Early and Late Endoscopic Recurrence Following Ileocolonic Resection in Crohn's Disease.* J Crohns Colitis, 2024. **18**(4): p. 615-627.

130. Salem, D.A., et al., *Sex-related differences in profiles and clinical outcomes of Inflammatory bowel disease: a systematic review and meta-analysis.* BMC Gastroenterol, 2024. **24**(1): p. 425.

131. Liu, Z., et al., *Sex-specific comparison of clinical characteristics and prognosis in Crohn's disease: A retrospective cohort study of 611 patients in China.* Front Physiol, 2022. **13**: p. 972038.

132. Chan, I.S. and G.S. Ginsburg, *Personalized medicine: progress and promise.* Annu Rev Genomics Hum Genet, 2011. **12**: p. 217-44.

133. Mathur, S. and J. Sutton, *Personalized medicine could transform healthcare.* Biomed Rep, 2017. **7**(1): p. 3-5.

134. Atreya, R. and M.F. Neurath, *Biomarkers for Personalizing IBD Therapy: The Quest Continues.* Clin Gastroenterol Hepatol, 2024. **22**(7): p. 1353-1364.

135. Ashton, J.J., et al., *Personalising medicine in inflammatory bowel disease-current and future perspectives.* Transl Pediatr, 2019. **8**(1): p. 56-69.

136. Rojas-Pena, M.L., et al., *Individualized Transcriptional Resolution of Complicated Malaria in a Colombian Study.* J Pers Med, 2018. **8**(3).

137. Banchereau, R., et al., *Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients.* Cell, 2016. **165**(6): p. 1548-1550.

138. Scharl, M. and G. Rogler, *Pathophysiology of fistula formation in Crohn's disease.* World J Gastrointest Pathophysiol, 2014. **5**(3): p. 205-12.

139. Kirkegaard, T., et al., *Expression and localisation of matrix metalloproteinases and their natural inhibitors in fistulae of patients with Crohn's disease.* Gut, 2004. **53**(5): p. 701-9.

140. Scharl, M., G. Rogler, and L. Biedermann, *Fistulizing Crohn's Disease.* Clin Transl Gastroenterol, 2017. **8**(7): p. e106.

141. Maddipatla, S.C., et al., *Assessing Cellular and Transcriptional Diversity of Ileal Mucosa Among Treatment-Naive and Treated Crohn's Disease.* Inflamm Bowel Dis, 2023. **29**(2): p. 274-285.

142. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nat Commun, 2017. **8**: p. 14049.

143. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.

144. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nat Methods, 2019. **16**(12): p. 1289-1296.

145. Roux de Bezieux, H., et al., *Improving replicability in single-cell RNA-Seq cell type discovery with Dune.* BMC Bioinformatics, 2024. **25**(1): p. 198.

146. Kumar, V., et al., *Single-Cell Atlas of Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs in Gastric Cancer.* Cancer Discov, 2022. **12**(3): p. 670-691.

147. Tazzari, M., et al., *Complex Immune Contextures Characterise Malignant Peritoneal Mesothelioma: Loss of Adaptive Immunological Signature in the More Aggressive Histological Types.* J Immunol Res, 2018. **2018**: p. 5804230.

148. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biol, 2015. **16**: p. 278.

149. Chen, J., et al., *ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W305-11.

150. Schaefer, C.F., et al., *PID: the Pathway Interaction Database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D674-9.

151. Gillespie, M., et al., *The reactome pathway knowledgebase 2022.* Nucleic Acids Res, 2022. **50**(D1): p. D687-D692.

152. van Unen, V., et al., *Identification of a Disease-Associated Network of Intestinal Immune Cells in Treatment-Naive Inflammatory Bowel Disease.* Front Immunol, 2022. **13**: p. 893803.

153. Birchenough, G.M., et al., *New developments in goblet cell mucus secretion and function.* Mucosal Immunol, 2015. **8**(4): p. 712-9.

154. Langer, V., et al., *IFN-gamma drives inflammatory bowel disease pathogenesis through VE-cadherin-directed vascular barrier disruption.* J Clin Invest, 2019. **129**(11): p. 4691-4707.

155. Sturzl, M., et al., *Angiocrine Regulation of Epithelial Barrier Integrity in Inflammatory Bowel Disease.* Front Med (Lausanne), 2021. **8**: p. 643607.

156. Haddow, J.B., et al., *Comparison of cytokine and phosphoprotein profiles in idiopathic and Crohn's disease-related perianal fistula.* World J Gastrointest Pathophysiol, 2019. **10**(4): p. 42-53.

157. Brown, S.L., et al., *Myd88-dependent positioning of Ptgs2-expressing stromal cells maintains colonic epithelial proliferation during injury.* J Clin Invest, 2007. **117**(1): p. 258-69.

158. Somineni, H.K., et al., *Whole-genome sequencing of African Americans implicates differential genetic architecture in inflammatory bowel disease.* Am J Hum Genet, 2021. **108**(3): p. 431-445.

159. Miyoshi, H., et al., *Prostaglandin E2 promotes intestinal repair through an adaptive cellular response of the epithelium.* EMBO J, 2017. **36**(1): p. 5-24.

160. Dolinger, M., J. Torres, and S. Vermeire, *Crohn's disease.* Lancet, 2024. **403**(10432): p. 1177-1191.

161. Colombel, J.F., et al., *Outcomes and Strategies to Support a Treat-to-target Approach in Inflammatory Bowel Disease: A Systematic Review.* J Crohns Colitis, 2020. **14**(2): p. 254-266.

162. Vasudevan, A., P.R. Gibson, and D.R. van Langenberg, *Time to clinical response and remission for therapeutics in inflammatory bowel diseases: What should the clinician expect, what should patients be told?* World J Gastroenterol, 2017. **23**(35): p. 6385-6402.

163. Schmitt, H., et al., *Expansion of IL-23 receptor bearing TNFR2+ T cells is associated with molecular resistance to anti-TNF therapy in Crohn's disease.* Gut, 2019. **68**(5): p. 814-828.

164. Rosen, M.J., A. Dhawan, and S.A. Saeed, *Inflammatory Bowel Disease in Children and Adolescents.* JAMA Pediatr, 2015. **169**(11): p. 1053-60.

165. Mitsialis, V., et al., *Single-Cell Analyses of Colon and Blood Reveal Distinct Immune Cell Signatures of Ulcerative Colitis and Crohn's Disease.* Gastroenterology, 2020. **159**(2): p. 591-608 e10.

166. Liu, T., H. Yu, and R.H. Blair, *Stability estimation for unsupervised clustering: A review.* Wiley Interdiscip Rev Comput Stat, 2022. **14**(6): p. e1575.

167. Mitchel, J., et al., *Coordinated, multicellular patterns of transcriptional variation that stratify patient cohorts are revealed by tensor decomposition.* Nat Biotechnol, 2024.

168. Phipson, B., et al., *propeller: testing for differences in cell type proportions in single cell data.* Bioinformatics, 2022. **38**(20): p. 4720-4726.

169. Hoffman, G.E., et al., *Efficient differential expression analysis of large-scale single cell transcriptomics data using dreamlet.* bioRxiv, 2024.

170. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.

171. Andreatta, M. and S.J. Carmona, *UCell: Robust and scalable single-cell gene signature scoring.* Comput Struct Biotechnol J, 2021. **19**: p. 3796-3798.

172. Duncan, L.E., et al., *Mapping the cellular etiology of schizophrenia and complex brain phenotypes.* Nat Neurosci, 2025. **28**(2): p. 248-258.

173. de Leeuw, C.A., et al., *MAGMA: generalized gene-set analysis of GWAS data.* PLoS Comput Biol, 2015. **11**(4): p. e1004219.

174. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nat Genet, 2015. **47**(9): p. 979-986.

175. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

176. Krieglstein, C.F., et al., *Collagen-binding integrin alpha1beta1 regulates intestinal inflammation in experimental colitis.* J Clin Invest, 2002. **110**(12): p. 1773-82.

177. Henderson, N.C., F. Rieder, and T.A. Wynn, *Fibrosis: from mechanisms to medicines.* Nature, 2020. **587**(7835): p. 555-566.

178. Vilardi, A., et al., *Current understanding of the interplay between extracellular matrix remodelling and gut permeability in health and disease.* Cell Death Discov, 2024. **10**(1): p. 258.

179. Kinchen, J., et al., *Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease.* Cell, 2018. **175**(2): p. 372-386 e17.

180. Sazonovs, A., et al., *Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility.* Nat Genet, 2022. **54**(9): p. 1275-1283.

181. Anderson, A., et al., *Monocytosis Is a Biomarker of Severity in Inflammatory Bowel Disease: Analysis of a 6-Year Prospective Natural History Registry.* Inflamm Bowel Dis, 2022. **28**(1): p. 70-78.

182. Safaee, M., et al., *CD97 is a multifunctional leukocyte receptor with distinct roles in human cancers (Review).* Int J Oncol, 2013. **43**(5): p. 1343-50.

183. Tseng, W.Y., M. Stacey, and H.H. Lin, *Role of Adhesion G Protein-Coupled Receptors in Immune Dysfunction and Disorder.* Int J Mol Sci, 2023. **24**(6).

184. Murthy, S., et al., *Single-cell transcriptomics of rectal organoids from individuals with perianal fistulizing Crohn's disease reveals patient-specific signatures.* Sci Rep, 2024. **14**(1): p. 26142.

185. Gao, X., et al., *Integrative multi-omics deciphers the spatial characteristics of host-gut microbiota interactions in Crohn's disease.* Cell Rep Med, 2023. **4**(6): p. 101083.

186. Zheng, H.B., et al., *Concerted changes in the pediatric single-cell intestinal ecosystem before and after anti-TNF blockade*. 2023, eLife Sciences Publications, Ltd.

187. Garrido-Trigo, A., et al., *Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease.* Nat Commun, 2023. **14**(1): p. 4506.

188. Tindle, C., et al., *A living organoid biobank of patients with Crohn's disease reveals molecular subtypes for personalized therapeutics.* Cell Rep Med, 2024. **5**(10): p. 101748.

189. Macias-Ceja, D.C., et al., *Role of the epithelial barrier in intestinal fibrosis associated with inflammatory bowel disease: relevance of the epithelial-to mesenchymal transition.* Front Cell Dev Biol, 2023. **11**: p. 1258843.

190. Lu, H., et al., *Monocyte-macrophages modulate intestinal homeostasis in inflammatory bowel disease.* Biomark Res, 2024. **12**(1): p. 76.

191. Fish, S.M., R. Proujansky, and W.W. Reenstra, *Synergistic effects of interferon gamma and tumour necrosis factor alpha on T84 cell function.* Gut, 1999. **45**(2): p. 191-8.

192. Woznicki, J.A., et al., *TNF-alpha synergises with IFN-gamma to induce caspase-8-JAK1/2-STAT1-dependent death of intestinal epithelial cells.* Cell Death Dis, 2021. **12**(10): p. 864.

193. Liu, H., et al., *A broken network of susceptibility genes in the monocytes of Crohn's disease patients.* Life Sci Alliance, 2024. **7**(9).

194. Krzak, M., et al., *Single-Cell RNA Sequencing of Terminal Ileal Biopsies Identifies Signatures of Crohn's Disease Pathogenesis.* medRxiv, 2024: p. 2023.09.06.23295056.

195. Jostins, L., et al., *Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.* Nature, 2012. **491**(7422): p. 119-24.

196. Luo, Y., et al., *Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7.* Nat Genet, 2017. **49**(2): p. 186-192.

197. Murthy, S.K., et al., *Introduction of anti-TNF therapy has not yielded expected declines in hospitalisation and intestinal resection rates in inflammatory bowel diseases: a population-based interrupted time series study.* Gut, 2020. **69**(2): p. 274-282.

198. Lazarev, M., et al., *Small bowel resection rates in Crohn's disease and the indication for surgery over time: experience from a large tertiary care center.* Inflamm Bowel Dis, 2010. **16**(5): p. 830-5.

199. Jeuring, S.F., et al., *Improvements in the Long-Term Outcome of Crohn's Disease Over the Past Two Decades and the Relation to Changes in Medical Management: Results from the Population-Based IBDSL Cohort.* Am J Gastroenterol, 2017. **112**(2): p. 325-336.

200. Olaison, G., K. Smedh, and R. Sjodahl, *Natural course of Crohn's disease after ileocolic resection: endoscopically visualised ileal ulcers preceding symptoms.* Gut, 1992. **33**(3): p. 331-5.

201. Joustra, V., et al., *Natural History and Risk Stratification of Recurrent Crohn's Disease After Ileocolonic Resection: A Multicenter Retrospective Cohort Study.* Inflamm Bowel Dis, 2022. **28**(1): p. 1-8.

202. Rutgeerts, P., et al., *Natural history of recurrent Crohn's disease at the ileocolonic anastomosis after curative surgery.* Gut, 1984. **25**(6): p. 665-72.

203. Haberman, Y., et al., *Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature.* J Clin Invest, 2014. **124**(8): p. 3617-33.

204. Perez, K., et al., *Meta-Analysis of IBD Gut Samples Gene Expression Identifies Specific Markers of Ileal and Colonic Diseases.* Inflamm Bowel Dis, 2022. **28**(5): p. 775-782.

205. Krikos, A., C.D. Laherty, and V.M. Dixit, *Transcriptional activation of the tumor necrosis factor alpha-inducible zinc finger protein, A20, is mediated by kappa B elements.* J Biol Chem, 1992. **267**(25): p. 17971-6.

206. Wertz, I.E., et al., *De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling.* Nature, 2004. **430**(7000): p. 694-9.

207. Ma, A. and B.A. Malynn, *A20: linking a complex regulator of ubiquitylation to immunity and human disease.* Nat Rev Immunol, 2012. **12**(11): p. 774-85.

208. Sanjabi, S., S.A. Oh, and M.O. Li, *Regulation of the Immune Response by TGF-beta: From Conception to Autoimmunity and Infection.* Cold Spring Harb Perspect Biol, 2017. **9**(6).

209. Ahn, S.H., et al., *Hepatocyte nuclear factor 4alpha in the intestinal epithelial cells protects against inflammatory bowel disease.* Inflamm Bowel Dis, 2008. **14**(7): p. 908-20.

210. Lakshmikanth, T., et al., *Immune system adaptation during gender-affirming testosterone treatment.* Nature, 2024. **633**(8028): p. 155-164.

211. Ngollo, M., et al., *Identification of Gene Expression Profiles Associated with an Increased Risk of Post-Operative Recurrence in Crohn's Disease.* J Crohns Colitis, 2022. **16**(8): p. 1269-1280.

212. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nat Genet, 2018. **50**(9): p. 1219-1224.

213. Gettler, K., et al., *Common and Rare Variant Prediction and Penetrance of IBD in a Large, Multi-ethnic, Health System-based Biobank Cohort.* Gastroenterology, 2021. **160**(5): p. 1546-1557.

214. Zhou, J., et al., *The regulatory role of alternative splicing in inflammatory bowel disease.* Front Immunol, 2023. **14**: p. 1095267.

215. Tang, Y., et al., *Alternative Splice Forms of CYLD Mediate Ubiquitination of SMAD7 to Prevent TGFB Signaling and Promote Colitis.* Gastroenterology, 2019. **156**(3): p. 692-707 e7.

216. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

217. Ewels, P.A., et al., *The nf-core framework for community-curated bioinformatics pipelines.* Nat Biotechnol, 2020. **38**(3): p. 276-278.

218. Di Tommaso, P., et al., *Nextflow enables reproducible computational workflows.* Nat Biotechnol, 2017. **35**(4): p. 316-319.

219. Vitting-Seerup, K. and A. Sandelin, *The Landscape of Isoform Switches in Human Cancers.* Mol Cancer Res, 2017. **15**(9): p. 1206-1220.

220. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

221. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. **34**(3): p. 267-73.

222. Auzolle, C., et al., *Male gender, active smoking and previous intestinal resection are risk factors for post-operative endoscopic recurrence in Crohn's disease: results from a prospective cohort study.* Aliment Pharmacol Ther, 2018. **48**(9): p. 924-932.

223. Khoudari, G., et al., *Rates of Intestinal Resection and Colectomy in Inflammatory Bowel Disease Patients After Initiation of Biologics: A Cohort Study.* Clin Gastroenterol Hepatol, 2022. **20**(5): p. e974-e983.

224. West, N.R., et al., *Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel disease.* Nat Med, 2017. **23**(5): p. 579-589.

225. Shah, S.C., et al., *Sex-Based Differences in Incidence of Inflammatory Bowel Diseases-Pooled Analysis of Population-Based Studies From Western Countries.* Gastroenterology, 2018. **155**(4): p. 1079-1089 e3.

226. Jacenik, D., et al., *Sex- and Age-Related Estrogen Signaling Alteration in Inflammatory Bowel Diseases: Modulatory Role of Estrogen Receptors.* Int J Mol Sci, 2019. **20**(13).

227. Pierdominici, M., et al., *Linking estrogen receptor beta expression with inflammatory bowel disease activity.* Oncotarget, 2015. **6**(38): p. 40443-51.

228. Novak, G., et al., *Evaluation of optimal biopsy location for assessment of histological activity, transcriptomic and immunohistochemical analyses in patients with active Crohn's disease.* Aliment Pharmacol Ther, 2019. **49**(11): p. 1401-1409.

229. Udler, M.S., et al., *Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine.* Endocr Rev, 2019. **40**(6): p. 1500-1520.

230. Robinson-Cohen, C., et al., *Genome-Wide Association Study of CKD Progression.* J Am Soc Nephrol, 2023. **34**(9): p. 1547-1559.

231. Singh, S., et al., *Comparative efficacy and safety of biologic therapies for moderate-to-severe Crohn's disease: a systematic review and network meta-analysis.* Lancet Gastroenterol Hepatol, 2021. **6**(12): p. 1002-1014.

232. Peyrin-Biroulet, L., et al., *Loss of Response to Vedolizumab and Ability of Dose Intensification to Restore Response in Patients With Crohn's Disease or Ulcerative Colitis: A Systematic Review and Meta-analysis.* Clin Gastroenterol Hepatol, 2019. **17**(5): p. 838-846 e2.

233. Chanchlani, N., et al., *Mechanisms and management of loss of response to anti-TNF therapy for patients with Crohn's disease: 3-year data from the prospective, multicentre PANTS cohort study.* Lancet Gastroenterol Hepatol, 2024. **9**(6): p. 521-538.

234. Singh, S., et al., *Primary Non-Response to Tumor Necrosis Factor Antagonists is Associated with Inferior Response to Second-line Biologics in Patients with Inflammatory Bowel Diseases: A Systematic Review and Meta-analysis.* J Crohns Colitis, 2018. **12**(6): p. 635-643.

235. Adegbola, S.O., et al., *Anti-TNF Therapy in Crohn's Disease.* Int J Mol Sci, 2018. **19**(8).

236. Levin, A.D., M.E. Wildenberg, and G.R. van den Brink, *Mechanism of Action of Anti-TNF Therapy in Inflammatory Bowel Disease.* J Crohns Colitis, 2016. **10**(8): p. 989-97.

237. Lichtenstein, G.R., et al., *ACG Clinical Guideline: Management of Crohn's Disease in Adults.* Am J Gastroenterol, 2018. **113**(4): p. 481-517.

238. Connelly, T.M. and E. Messaris, *Predictors of recurrence of Crohn's disease after ileocolectomy: a review.* World J Gastroenterol, 2014. **20**(39): p. 14393-406.

239. Lee, K.E., et al., *Post-operative prevention and monitoring of Crohn's disease recurrence.* Gastroenterol Rep (Oxf), 2022. **10**: p. goac070.

240. Nos, P. and E. Domenech, *Postoperative Crohn's disease recurrence: a practical approach.* World J Gastroenterol, 2008. **14**(36): p. 5540-8.

241. Petagna, L., et al., *Pathophysiology of Crohn's disease inflammation and recurrence.* Biol Direct, 2020. **15**(1): p. 23.

242. D'Haens, G. and P. Rutgeerts, *Postoperative recurrence of Crohn's disease: pathophysiology and prevention.* Inflamm Bowel Dis, 1999. **5**(4): p. 295-303.

243. Poulsen, A., et al., *Re-resection Rates and Disease Recurrence in Crohn's Disease: A Population-based Study Using Individual-level Patient Data.* J Crohns Colitis, 2024. **18**(10): p. 1631-1643.

244. Battat, R. and W.J. Sandborn, *Advances in the Comprehensive Management of Postoperative Crohn's Disease.* Clin Gastroenterol Hepatol, 2022. **20**(7): p. 1436-1449.

245. Khoshkish, S., et al., *Risk factors for postoperative recurrence of Crohn's disease.* Middle East J Dig Dis, 2012. **4**(4): p. 199-205.

246. Rutgeerts, P., et al., *Predictability of the postoperative course of Crohn's disease.* Gastroenterology, 1990. **99**(4): p. 956-63.

247. Riviere, P., et al., *Rates of Postoperative Recurrence of Crohn's Disease and Effects of Immunosuppressive and Biologic Therapies.* Clin Gastroenterol Hepatol, 2021. **19**(4): p. 713-720 e1.

248. Bak, M.T.J., et al., *Interobserver agreement of current and new proposed endoscopic scores for postoperative recurrence in Crohn's disease.* Gastrointest Endosc, 2024. **100**(4): p. 703-709 e4.

249. Bak, M.T.J., et al., *Prognostic Value of the Modified Rutgeerts Score for Long-Term Outcomes After Primary Ileocecal Resection in Crohn's Disease.* Am J Gastroenterol, 2024. **119**(2): p. 306-312.

250. Yang, D.H., et al., *Usefulness of C-reactive protein as a disease activity marker in Crohn's disease according to the location of disease.* Gut Liver, 2015. **9**(1): p. 80-6.

251. Andersson, E., et al., *Subphenotypes of inflammatory bowel disease are characterized by specific serum protein profiles.* PLoS One, 2017. **12**(10): p. e0186142.

252. Salomon, B., et al., *Characterization of Inflammatory Bowel Disease Heterogeneity Using Serum Proteomics: A Multicenter Study.* J Crohns Colitis, 2025. **19**(5).

253. Zwicker, S., et al., *Systemic Chemokine Levels with "Gut-Specific" Vedolizumab in Patients with Inflammatory Bowel Disease-A Pilot Study.* Int J Mol Sci, 2017. **18**(8).

254. Winter, D.A., et al., *Biomarkers predicting the effect of anti-TNF treatment in paediatric and adult inflammatory bowel disease.* J Pediatr Gastroenterol Nutr, 2024. **79**(1): p. 62-75.

255. White, B., et al., *Inflammation-related Proteins Support Diagnosis of Inflammatory Bowel Disease and Are Modified by Exclusive Enteral Nutrition in Children With Crohn's Disease, Especially of Ileal Phenotype.* Inflamm Bowel Dis, 2025. **31**(3): p. 733-745.

256. Granno, O., et al., *Preclinical Protein Signatures of Crohn's Disease and Ulcerative Colitis: A Nested Case-Control Study Within Large Population-Based Cohorts.* Gastroenterology, 2025. **168**(4): p. 741-753.

257. Kumar, M., M. Garand, and S. Al Khodor, *Integrating omics for a better understanding of Inflammatory Bowel Disease: a step towards personalized medicine.* J Transl Med, 2019. **17**(1): p. 419.

258. Mu, C., et al., *Multi-omics in Crohn's disease: New insights from inside.* Comput Struct Biotechnol J, 2023. **21**: p. 3054-3072.

259. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.* Nat Rev Genet, 2012. **13**(4): p. 227-32.

260. Preto, A.J., et al., *Multi-omics data integration identifies novel biomarkers and patient subgroups in inflammatory bowel disease.* J Crohns Colitis, 2025. **19**(1).

261. Gettler, K., et al., *Post-operative ileum transcirptomics implicate sex-biased mechanisms in Crohn's disease recurrence*. 2025.

262. Hammoudi, N., et al., *Postoperative Endoscopic Recurrence on the Neoterminal Ileum But Not on the Anastomosis Is Mainly Driving Long-Term Outcomes in Crohn's Disease.* Am J Gastroenterol, 2020. **115**(7): p. 1084-1093.

263. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

264. Zhang, Y., G. Parmigiani, and W.E. Johnson, *ComBat-seq: batch effect adjustment for RNA-seq count data.* NAR Genom Bioinform, 2020. **2**(3): p. lqaa078.

265. Reinisch, W., et al., *Clinical relevance of serum interleukin-6 in Crohn's disease: single point measurements, therapy monitoring, and prediction of clinical relapse.* Am J Gastroenterol, 1999. **94**(8): p. 2156-64.

266. Mudter, J. and M.F. Neurath, *Il-6 signaling in inflammatory bowel disease: pathophysiological role and clinical relevance.* Inflamm Bowel Dis, 2007. **13**(8): p. 1016-23.

267. O'Sullivan, S., J.F. Gilmer, and C. Medina, *Matrix metalloproteinases in inflammatory bowel disease: an update.* Mediators Inflamm, 2015. **2015**: p. 964131.

268. Reenaers, C., et al., *Sensitivity of intestinal fibroblasts to TNF-related apoptosis-inducing ligand-mediated apoptosis in Crohn's disease.* Scand J Gastroenterol, 2008. **43**(11): p. 1334-45.

269. Begue, B., et al., *Implication of TNF-related apoptosis-inducing ligand in inflammatory intestinal epithelial lesions.* Gastroenterology, 2006. **130**(7): p. 1962-74.

270. Brost, S., et al., *Differential expression of the TRAIL/TRAIL-receptor system in patients with inflammatory bowel disease.* Pathol Res Pract, 2010. **206**(1): p. 43-50.

271. Moran, C.J., et al., *IL-10R polymorphisms are associated with very-early-onset ulcerative colitis.* Inflamm Bowel Dis, 2013. **19**(1): p. 115-23.

272. Glocker, E.O., et al., *Inflammatory bowel disease and mutations affecting the interleukin-10 receptor.* N Engl J Med, 2009. **361**(21): p. 2033-45.

273. Trivedi, P.J. and D.H. Adams, *Chemokines and Chemokine Receptors as Therapeutic Targets in Inflammatory Bowel Disease; Pitfalls and Promise.* J Crohns Colitis, 2018. **12**(suppl_2): p. S641-S652.

274. Ajuebor, M.N., S.L. Kunkel, and C.M. Hogaboam, *The role of CCL3/macrophage inflammatory protein-1alpha in experimental colitis.* Eur J Pharmacol, 2004. **497**(3): p. 343-9.

275. Kopiasz, L., K. Dziendzikowska, and J. Gromadzka-Ostrowska, *Colon Expression of Chemokines and Their Receptors Depending on the Stage of Colitis and Oat*

Beta-Glucan Dietary Intervention-Crohn's Disease Model Study. Int J Mol Sci, 2022. **23**(3).

276. Gong, W., et al., *CCL4-mediated targeting of spleen tyrosine kinase (Syk) inhibitor using nanoparticles alleviates inflammatory bowel disease.* Clin Transl Med, 2021. **11**(2): p. e339.

277. Fonseca-Camarillo, G., et al., *Expression of interleukin (IL)-19 and IL-24 in inflammatory bowel disease patients: a cross-sectional study.* Clin Exp Immunol, 2014. **177**(1): p. 64-75.

278. Onody, A., et al., *Interleukin-24 regulates mucosal remodeling in inflammatory bowel diseases.* J Transl Med, 2021. **19**(1): p. 237.

279. Alhendi, A. and S.A. Naser, *The dual role of interleukin-6 in Crohn's disease pathophysiology.* Front Immunol, 2023. **14**: p. 1295230.

280. Biel, C., et al., *Matrix metalloproteinases in intestinal fibrosis.* J Crohns Colitis, 2024. **18**(3): p. 462-478.

281. Kaser, A., et al., *Increased expression of CCL20 in human inflammatory bowel disease.* J Clin Immunol, 2004. **24**(1): p. 74-85.

282. Singh, U.P., et al., *Chemokine and cytokine levels in inflammatory bowel disease patients.* Cytokine, 2016. **77**: p. 44-9.

283. Glas, J., et al., *Analysis of IL12B gene variants in inflammatory bowel disease.* PLoS One, 2012. **7**(3): p. e34349.

284. Reinecker, H.C., et al., *Monocyte-chemoattractant protein 1 gene expression in intestinal epithelial cells and inflammatory bowel disease mucosa.* Gastroenterology, 1995. **108**(1): p. 40-50.

285. Souza, R.F., et al., *Study of tumor necrosis factor receptor in the inflammatory bowel disease.* World J Gastroenterol, 2023. **29**(18): p. 2733-2746.

286. Luther, J., et al., *Loss of Response to Anti-Tumor Necrosis Factor Alpha Therapy in Crohn's Disease Is Not Associated with Emergence of Novel Inflammatory Pathways.* Dig Dis Sci, 2018. **63**(3): p. 738-745.

287. Yao, Y., et al., *Identification of Targets for Subsequent Treatment of Crohn's Disease Patients After Failure of Anti-TNF Therapy.* J Inflamm Res, 2023. **16**: p. 4617-4631.

288. Chen, P., et al., *Serum Biomarkers for Inflammatory Bowel Disease.* Front Med (Lausanne), 2020. **7**: p. 123.

289. Jones, J., et al., *Relationships between disease activity and serum and fecal biomarkers in patients with Crohn's disease.* Clin Gastroenterol Hepatol, 2008. **6**(11): p. 1218-24.

290. Lu, Y., et al., *Serum omentin-1 as a disease activity marker for Crohn's disease.* Dis Markers, 2014. **2014**: p. 162517.

291. Li, L.J., et al., *Role of interleukin-22 in inflammatory bowel disease.* World J Gastroenterol, 2014. **20**(48): p. 18177-88.

292. Cineus, R., et al., *The IL-22-oncostatin M axis promotes intestinal inflammation and tumorigenesis.* Nat Immunol, 2025. **26**(6): p. 837-853.

293. Park, J.H., et al., *Insight into the role of TSLP in inflammatory bowel diseases.* Autoimmun Rev, 2017. **16**(1): p. 55-63.

294. Iliev, I.D., et al., *Human intestinal epithelial cells promote the differentiation of tolerogenic dendritic cells.* Gut, 2009. **58**(11): p. 1481-9.

295. Hormi, K., et al., *Transforming growth factor-alpha and epidermal growth factor receptor in colonic mucosa in active and inactive inflammatory bowel disease.* Growth Factors, 2000. **18**(2): p. 79-91.

296. Gadaleta, R.M., et al., *Fibroblast Growth Factor 19 modulates intestinal microbiota and inflammation in presence of Farnesoid X Receptor.* EBioMedicine, 2020. **54**: p. 102719.

297. Jacobsen, G.E., et al., *Lamina Propria Phagocyte Profiling Reveals Targetable Signaling Pathways in Refractory Inflammatory Bowel Disease.* Gastro Hep Adv, 2022. **1**(3): p. 380-392.

298. Scaldaferri, F., et al., *VEGF-A links angiogenesis and inflammation in inflammatory bowel disease pathogenesis.* Gastroenterology, 2009. **136**(2): p. 585-95 e5.

299. Lin, H. and B. Cao, *Circulating VEGF and inflammatory bowel disease: a bidirectional mendelian randomization.* Front Genet, 2024. **15**: p. 1282471.

300. Bersudsky, M., et al., *Non-redundant properties of IL-1alpha and IL-1beta during acute colon inflammation in mice.* Gut, 2014. **63**(4): p. 598-609.

301. Chiriac, M.T., et al., *IL-20 controls resolution of experimental colitis by regulating epithelial IFN/STAT2 signalling.* Gut, 2024. **73**(2): p. 282-297.

302. Antonioli, L., et al., *Inflammatory Bowel Diseases: It's Time for the Adenosine System.* Front Immunol, 2020. **11**: p. 1310.

303. Nunes, T., C. Bernardazzi, and H.S. de Souza, *Interleukin-33 and inflammatory bowel diseases: lessons from human studies.* Mediators Inflamm, 2014. **2014**: p. 423957.

304. Wang, B., et al., *Identification of hub programmed cell death-related genes and immune infiltration in Crohn's disease using bioinformatics.* Front Genet, 2024. **15**: p. 1425062.

305. Kaiser, T., et al., *Faecal S100A12 as a non-invasive marker distinguishing inflammatory bowel disease from irritable bowel syndrome.* Gut, 2007. **56**(12): p. 1706-13.

306. Li, J., et al., *Identification and multimodal characterization of a specialized epithelial cell type associated with Crohn's disease.* Nat Commun, 2024. **15**(1): p. 7204.

307. Tien, N.T.N., et al., *An exploratory multi-omics study reveals distinct molecular signatures of ulcerative colitis and Crohn's disease and their correlation with disease activity.* J Pharm Biomed Anal, 2025. **255**: p. 116652.

308. Lee, J.W.J., et al., *Multi-omics reveal microbial determinants impacting responses to biologic therapies in inflammatory bowel disease.* Cell Host Microbe, 2021. **29**(8): p. 1294-1304 e4.

309. Gudino, V., R. Bartolome-Casado, and A. Salas, *Single-cell omics in inflammatory bowel disease: recent insights and future clinical applications.* Gut, 2025.

310. Kong, L., et al., *The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon.* Immunity, 2023. **56**(2): p. 444-458 e5.

311. Wang, Z. and J. Shen, *The role of goblet cells in Crohn' s disease.* Cell Biosci, 2024. **14**(1): p. 43.

312. Cross, R.K., et al., *Racial differences in disease phenotypes in patients with Crohn's disease.* Inflamm Bowel Dis, 2006. **12**(3): p. 192-8.

313. Basson, A., et al., *The association between race and Crohn's disease phenotype in the Western Cape population of South Africa, defined by the Montreal Classification System.* PLoS One, 2014. **9**(8): p. e104859.