THE UNIVERSITY OF CHICAGO

ADVANCING INSIGHTS INTO INFLAMMATORY BOWEL DISEASE THROUGH GENETIC AND MULTIOMIC DATA ANALYSIS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

YU ZHAO

CHICAGO, ILLINOIS

MARCH 2025

ABSTRACT

Crohn's disease (CD) and ulcerative colitis (UC) are the two major forms of inflammatory bowel disease (IBD), driven by complex interactions among genetic, environmental, and microbial factors. This dissertation presents a dual investigation into (1) the cell-type-specific regulatory architecture of the intestinal mucosa in CD, and (2) the cellular adaptations that occur in the ileal pouch of UC patients. In Chapter 2, a single-cell atlas of chromatin accessibility from biopsies of active and inactive CD and healthy controls reveals 557,310 candidate cis-regulatory elements (cCREs). Topic modeling using cCREs reveals regulatory programs that govern cell type-specific and anatomic location specific functions, as well as those associated with inflammatory responses. Immune cell types show the strongest enrichment for CD heritability, but fine-mapped variants also highlight epithelial and stromal cell involvement at specific loci. Chapter 3 focuses on ileal pouches from UC patients, using single-cell RNA-seq and ATAC-seq to uncover partial colonic metaplasia in pouch. Together, these studies illuminate the diverse cellular and epigenetic mechanisms underlying gut tissue architecture, IBD risk and the plasticity of the intestinal epithelium.

To my family and friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Supplementary files are available online. Appendix C and D give each file's description.

ACKNOWLEDGEMENTS

First and foremost, I want to express my deepest gratitude to my advisor Dr. Sebastian Pott and Dr. Eugene B Chang for their generous and unwavering support. Without their encouragement and guidance, I would not have been able to successfully finish my PhD. There are three moments that shaped the trajectory of my PhD research. For the first moment is at the beginning of my PhD, and here I want to thank Dr. Jun Huang for allowing me to study T cell immunology and work on the deep learning based imaging analysis projects, at the time when I even didn't know what CD3 is. Although the projects were disrupted mainly because of the pandemic, I really appreciate that he saw my passion of studying immunology and solving problems that are truly meaningful for human beings and gave me the opportunity to learn immunology from the beginning. The second moment is when I talked to Dr. Eugene B Chang and Dr. Candace M Cham about my depression about the failure of my first project. They helped me gain confidence back and find research directions where I can fully utilize the skills I have developed and continue to study disease biology. Without their trust and help, I wouldn't be able to continue doing research. The third moment is when I discussed with Dr. Sebastian Pott about how to practically continue my research and the details about the projects I can work on. I still remember it was an afternoon when he talked with me on phone during his driving, but it was this unofficial call that shaped the whole framework about the rest of my PhD and this dissertation. Seb is very knowledgeable and has very high standards for doing research. He is very good at discussing and collaborating with people from different backgrounds or labs. He is very thoughtful and patient, and always thinks from other people's situation and perspective. Seb's example motivates me to embody the qualities of a virtuous individual. He is the most pivotal advisor in my life.

I would like to thank my committee members Dr. Mark Mimee and Dr. Nicolas Chevrier. I first met Mark in a course on human microbiota taught by him. The main topic in my PhD projects is about IBD so I hoped to learn more about this very important perspective, microbiota, which is a complex factor in IBD pathogenesis. Mark is very knowledgeable and his carefully designed course structure and discussion session helped me gain this knowledge comprehensively. Nicolas is an outstanding scientist specializing in immunology, it's a great opportunity that he can provide this valuable perspective. I am deeply grateful to Mark and Nicolas for agreeing to join my committee and helping me bring my PhD journey to a successful conclusion.

I'm grateful to the faculty members and my colleagues in the University of Chicago for providing valuable discussions and feedback on my work: Dr. Candace M Cham, Dr. Peter Carbonetto, Dr. Cambrian Y Liu, Dr. Xin He, Dr. Andrew W. Dahl, Dr. Xuanyao Liu, Dr. Mengjie Chen, Dr. Zepeng, Mu, Dr. Anindita Basu, Dr. Bingqing Xie, Dr. Kaixuan Luo, Dr. Ayelen Lizarraga and Xiaoyuan Zhong,. I want to thank Dr. Ran Zhou for conducting a significant amount of high-quality experimental work and provide a solid basis for my projects.

Nothing in this dissertation would have been possible without my family and friends. I am deeply thankful to my lifelong friends Guoshuai Cao and Yushan Liang whose influence has been integral in shaping my character and defining who I am. My heartfelt thanks go to Guoshuai, who helped me enormously in my PhD research, and stood by me through life's ups and downs. Yushan, an invaluable friend with me since childhood, has always been reliable and responsible as a trusted companion, and his support has been indispensable during this journey. I want to thank Ming Yang for bringing so much positivity, joy and happiness to my world. In the last few months of my PhD,

Zhiwei Jiang, a kind-hearted friend, passionate weightlifting enthusiast, and a researcher known for his integrity, ambition, and broad knowledge, generously shared his expertise in gym workouts with me and helped me establish a healthy fitness habit. Finally, I would like to express my heartfelt appreciation to my mother Shurui Gao and father Zhenfeng Zhao, who give me everything to support my growth and wellbeing, as well as the complete freedom for me to explore this interesting world and to be myself, without forcing any explicit or implicit expectation that don't resonate with me.

There are countless other individuals to whom I owe my thanks. Without their encouragement and support, this dissertation, and this journey, would not have been possible.

CHAPTER 1: Introduction

Inflammatory bowel disease is a complex disorder with unclear etiology

Inflammatory bowel disease (IBD) compromises a set of chronic inflammatory disorders of the gastrointestinal tract, primarily including two main subtypes: Crohn's disease (CD) and ulcerative colitis (UC)[1]. Globally, the incidence of IBD has risen over the past decades, posing significant clinical challenges and economic burdens[2]. Patients typically experience flares of active disease, characterized by diarrhea, abdominal pain, and systemic symptom, and punctuated by periods of remission[3–5]. This relapsing–remitting pattern negatively impacts the quality of life of patients, increasing the urgency to uncover the pathogenic mechanisms that underlie persistent gut inflammation and to develop more precise treatments.

CD can affect any region of the gastrointestinal (GI) tract, from the mouth to the anus, and is characterized by patchy lesions. Inflammation can be transmural, affecting the entire thickness of the intestinal wall, and often leads to complications like fistulae and strictures[5]. Conversely, UC usually presents with continuous mucosal inflammation limited to the colon and rectum, with histological changes primarily confined to the mucosal layer[3,4].

Therapeutic approaches in IBD aim to reduce inflammation and promote mucosal healing. Currently common treatments include aminosalicylates, steroids, immunomodulators and biologic agents targeting specific cytokines (e.g., TNF-α, IL-12/23) or cell trafficking (e.g., integrins)[6,7]. Nonetheless, a considerable number of patients either fail to respond or lose response over time, underscoring the need for deeper mechanistic insight that can guide the development of alternative therapies.

IBD pathophysiology involves a complex interplay among genetic susceptibility, environmental factors, the gut microbiota, and mucosal immune responses. Genome-wide association studies (GWAS) are large-scale investigations designed to identify genetic variants associated with specific traits or diseases. By scanning the entire genome of individuals—comparing thousands or even millions of single nucleotide polymorphisms (SNPs) between patients and healthy controls—GWAS detect loci that occur more frequently in the disease population[8].

In IBD research, GWAS have uncovered hundreds of risk loci implicated in immunological pathways, epithelial barrier function, and host-microbe interactions[9–11]. For instance, variants in genes like NOD2 have been robustly linked with CD, highlighting the role of innate immune pathways that sense bacterial components. Other discovered loci cluster around cytokine signaling and T cell activation, illuminating how genetic predisposition influences the mucosal inflammatory response[12]. Such insights have broadened our understanding of IBD pathogenesis, revealing that both UC and CD involve overlapping yet distinct sets of genetic risk factors.

However, genotype alone does not fully predict disease onset or severity, highlighting the importance of epigenetic and transcriptomic factors in driving disease heterogeneity. Given this complexity, a comprehensive understanding of IBD requires detailed exploration of the cellular populations and molecular pathways within gut tissue, a challenge that can be partially addressed using single-cell genomic technologies.

Ileal pouch-anal anastomosis surgery is the standard treatment for severe ulcerative colitis

In ulcerative colitis, a subset of patients develop severe, refractory disease or experiences complications such as high-grade dysplasia or carcinoma. In these cases, colectomy (removal of the colon) often becomes necessary to eliminate the diseased tissue and mitigate cancer risk. To avoid a permanent ileostomy, surgeons devised the ileal pouch-anal anastomosis (IPAA), commonly known as J-pouch surgery. This strategy restores continuity of the GI tract by creating a reservoir from the terminal ileum, which is connected to the anus, preserving voluntary defecation and improving quality of life. Despite its efficacy, IPAA surgery can give rise to complications. About 60% of these patients develop an inflammatory condition in the pouch itself, known as pouchitis[13]. It indicates that removing the colon does not completely cure the disease[14–16].

Given the unmet clinical needs of patients with persistent pouchitis and the opportunity to study a well-defined, surgically altered intestinal environment, IPAA pouches offer a valuable window into mucosal immunology. However, a high-resolution single-cell profiling of pouch tissue remains unavailable. Addressing this gap is crucial for elucidating pouch-specific disease pathogenesis and uncovering potential therapeutic targets.

GWAS and functional QTLs decipher genetic risk in IBD

Fine-mapping refines broad association signals identified through GWAS by pinpointing the most likely causal variants within a genetic locus. Traditional GWAS often highlight large genomic regions that encompass multiple correlated SNPs. Fine-mapping leverages statistical algorithms to

calculate posterior probabilities of causality for each SNP in a locus[17]. This process narrows down the variants most likely to exert a true biological effect.

However, pinpointing how individual variants are associated with disease remains challenging because most SNPs identified by GWAS reside in non-coding regions of the genome. Thus, functional genomics approaches such as expression quantitative trait locus (eQTL), chromatin accessibility QTL (caQTL) mapping and single-cell chromatin accessibility profiling, have been developed to interpret the biological impact of these variants[18–20].

eQTLs are genomic loci in which genetic differences correlate with variable transcript levels of specific genes. By linking a candidate SNP to changes in mRNA abundance, eQTL analyses illuminate how the variant might alter downstream cellular functions. caQTLs take a related approach, but focus on whether SNPs affect the accessibility of chromatin at particular sites, potentially influencing transcription factor binding and gene regulation[21]. Crucially, colocalization methods can integrate fine-mapped SNPs with QTL data to assess whether a SNP drives variation in traits such as expression levels and chromatin accessibility[22]. By doing so, colocalization links genetic variation to specific molecular mechanisms, thereby facilitating the exploration of disease pathogenesis and potential therapeutic interventions.

Resolving cellular complexity in IBD with single-cell RNA sequencing

Advances in single-cell genomics have transformed our ability to study complex diseases by enabling high-resolution profiling of individual cells. Traditional bulk sequencing methods, while powerful, obscure the heterogeneity of cellular populations by averaging signals across thousands

or millions of cells. This limitation is particularly problematic in tissues like the intestinal mucosa, where a diverse array of immune, epithelial, and stromal cell types interact to maintain homeostasis and drive disease pathology. By contrast, single-cell approaches such as single-cell RNA sequencing (scRNA-seq) and single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) allow researchers to dissect the intricate cellular ecosystems underlying[23]. These methods provide an unprecedented view into the molecular mechanisms governing cell identity, function, and dysregulation in IBD, offering insights into disease progression and potential therapeutic targets.

scRNA-seq enables the measurement of gene expression at the resolution of individual cells, allowing researchers to map cell types and states with high precision. By applying scRNA-seq to intestinal tissues from IBD patients and healthy individuals, researchers have identified a diverse spectrum of cell populations contributing to disease pathology. Studies in Crohn's disease have highlighted dysregulated T cell subsets with inflammatory transcriptional signatures, as well as perturbed interactions between epithelial and immune cells that may perpetuate chronic inflammation[24–26]. UC-focused research has shown how expression patterns and shifting cell populations may shape disease and drug resistance[27]. These findings underscore that IBD is not driven by a single aberrant cell type but rather by a dynamic network of immune cells, epithelial and stromal cells.

Beyond cellular heterogeneity, scRNA-seq has enabled the discovery of molecular biomarkers linked to clinical outcomes. By correlating transcriptomic signatures with patient responses to biologics such as anti-TNF or anti-integrin therapies, researchers aim to develop predictive models

for treatment efficacy[28]. Furthermore, integrating scRNA-seq data with spatial transcriptomics has provided additional context, revealing how inflammatory niches form within intestinal tissues and how these microenvironments drive localized immune activation[28]. This cell-type-resolved approach offers a foundation for precision medicine in IBD, guiding the development of more targeted therapeutic interventions.

Decoding the regulatory landscape using scATAC-seq

Gene expression is regulated in a cell-type- and context-specific manner through the activity of cis-regulatory elements (CREs)—non-coding DNA sequences such as enhancers and insulators that control transcriptional output[29–32]. These elements function by recruiting transcription factors (TFs) and chromatin-modifying complexes to orchestrate gene activation or repression. The accessibility of CREs is dynamically modulated in response to environmental cues, developmental signals, and disease states, making them key regulatory hubs for cellular identity and function. Dysregulation of epigenetic regulation can contribute to a wide spectrum of diseases, including cancer and chronic inflammatory conditions.

The study of chromatin accessibility provides a direct means to map CRE activity across different cell types, and this is where scATAC-seq plays a crucial role[33]. scATAC-seq profiles open chromatin regions at single-cell resolution, identifying which genomic elements are accessible to transcriptional machinery in specific cell populations. Unlike bulk ATAC-seq, which averages signals across diverse cell types, scATAC-seq allows researchers to pinpoint regulatory landscapes unique to each cellular subset, offering a deeper understanding of cell-state transitions and disease-associated epigenetic changes.

In disease biology, single-cell chromatin accessibility data offers a window into how pathological conditions remodel the epigenome. For instance, in Crohn's disease, scATAC-seq can reveal which transcription factors drive inflammatory immune response. By aligning accessibility peaks with known transcription factor motifs, we can determine the specific transcription factors that underlie pathological cell states. Moreover, mapping CREs helps pinpoint the exact cell populations where disease-associated single nucleotide polymorphisms (SNPs) exert their effects. Such mechanistic insights highlight potential therapeutic targets for selective intervention. Additionally, integrating single-cell scRNA-seq with scATAC-seq further refines our understanding of gene regulation, enabling researchers to connect open chromatin regions to changes in gene expression[29,30,34–37]. This multi-omic perspective unravels the interplay between gene activation and transcription factor binding at enhancers or promoters.

One of the challenges in analyzing single-cell chromatin accessibility data is the complexity of regulatory element activity across diverse cellular states. Traditional clustering approaches, while effective for defining major cell types, may fail to capture continuous regulatory programs that are shared by multiple cell types, associated with functional states, disease processes, or environmental responses. To address this, topic modeling—a computational approach originally developed for uncovering latent themes in textual data—has been adapted for single-cell genomics[38,39].

In the context of scATAC-seq, each cell can be viewed as a "document," while accessible chromatin regions (candidate CREs, or peaks) function as "words." By factorizing a matrix of cells-by-peaks, topic modeling uncovers hidden "topics," where each topic represents a

combination of co-accessible genomic regions[38,39]. Topics can then be interpreted as capturing specific regulatory programs (e.g., cell type specific functions, inflammatory processes) by evaluating their summarized gene scores, enrichment for TF binding sites or biological pathways.

Topic modeling is data-driven in nature: it does not rely on predetermined gene sets or accessibility features but rather allows individual programs calculated directly from the high-dimensional dataset. These regulatory topics can reveal the heterogeneity of the cellular landscape in gut tissue and epigenetic programs associated with diseased states.

Dissertation overview

This dissertation explores the complex etiologies and cellular mechanisms underlying IBD, with a focus on CD and on the surgically constructed ileal pouch following proctocolectomy in UC patients. Both conditions demonstrate intricate interplay among genetic factors, immune pathways, and epigenetic regulatory networks. By harnessing cutting-edge single-cell technologies, the work presented here provides insights into the genetic risks, epigenetic cis-regulatory architecture, cell-type-specific transcriptional programs, and tissue adaptations that could be partially associated with disease pathogenesis and colon-related function.

In the first major study, a comprehensive atlas of cell-type-resolved chromatin accessibility was constructed using human biopsy samples from the terminal ileum and ascending colon in active and inactive CD and in healthy controls. Integrating this atlas with genetic data revealed 557,310 candidate cis-regulatory elements (cCREs) that exhibit cell-type-, location-, and inflammation-specific patterns. Notably, CD heritability was enriched in immune cells, yet fine-mapped non-

coding variants also localized within epithelial and stromal cCREs, underscoring that multiple lineages can drive risk at certain loci. This investigation demonstrates how specific transcription factor binding motifs can be disrupted by genetic variants, thereby influencing local chromatin accessibility. Ultimately, this work establishes a high-resolution chromatin accessibility atlas for understanding gene regulation in CD, paving the way for targeted exploration of risk alleles and therapeutic interventions.

The second study addresses total proctocolectomy with ileal pouch-anal anastomosis, a surgical procedure for patients with severe ulcerative colitis. Using both scRNA-seq and scATAC-seq, an integrated atlas was generated for the ileal pouch and the pre-pouch ileum, compared against healthy ileum and colon. The findings highlight distinctive pouch-specific absorptive and secretory epithelial subpopulations exhibiting a hybrid “colon-like” but partially “ileal-retained” gene expression profile. These results illuminate adaptive tissue remodeling after surgery and provide a high-resolution atlas for understanding pouch physiology and potential inflammatory complications.

CHAPTER 2: Cell-type-resolved chromatin accessibility in the human intestine identifies complex regulatory programs and clarifies genetic associations in Crohn's disease

Note:

The following section (Chapter 2) is reproduced with slight modifications from my first-author publication "Cell-type-resolved chromatin accessibility in the human intestine identifies complex regulatory programs and clarifies genetic associations in Crohn's disease", which is available in medRxiv[40]. Crohn's disease (CD) is a multifactorial disorder influenced by genetic, microbial, and environmental factors, yet the specific contributions of distinct cell types to its pathogenesis remain unclear. This chapter presents a comprehensive analysis of cell-type-resolved chromatin accessibility in the intestinal mucosa, providing insights into the regulatory programs and genetic mechanisms that shape CD risk and inflammatory responses.

Authors:

Yu Zhao[†], Ran Zhou[†], Zepeng Mu, Peter Carbonetto, Xiaoyuan Zhong, Bingqing Xie, Kaixuan Luo, Candace M Cham, Jason Koval, Xin He, Andrew W. Dahl, Xuanyao Liu, Eugene B Chang, Anindita Basu, Sebastian Pott

[†] These authors contributed equally to this work.

Author Contributions:

Y.Z. performed data analysis. R.Z. generated data. Z.M. and P.C. contributed to data analysis. X.Z., B.X., K.L., X.H., A.W.D., X.L. and E.B.C. contributed to data interpretation. C.M.C. and

J.K. contributed to sample collection, processing and experiments. Y. Z. and S.P. wrote the manuscript and created figures. S.P. conceptualized the study. A.B. and S.P. supervised the study.

Abstract

Crohn's disease (CD) is a complex inflammatory bowel disease resulting from an interplay of genetic, microbial, and environmental factors. Cell-type-specific contributions to CD etiology and genetic risk are incompletely understood. Here we built a comprehensive atlas of cell-type-resolved chromatin accessibility comprising 557,310 candidate cis-regulatory elements (cCREs) in terminal ileum and ascending colon from patients with active and inactive CD and healthy controls. Using this atlas, we identified cell-type-, anatomic location-, and context-specific cCREs and characterized the regulatory programs underlying inflammatory responses in the intestinal mucosa of CD patients. Genetic variants that disrupt binding motifs of cell-type-specific transcription factors significantly affected chromatin accessibility in specific mucosal cell types. We found that CD heritability is primarily enriched in immune cell types. However, using fine-mapped non-coding CD variants we identified 29 variants located within cCREs several of which were accessible in epithelial and stromal cells implicating cell types from additional lineages in mediating CD risk in some loci. Our atlas provides a comprehensive resource to study gene regulatory effects in CD and health, and highlights the cellular complexity underlying CD risk.

Introduction

Crohn's disease, one of the main types of inflammatory bowel diseases (IBD), is a chronic inflammatory disease of the gastrointestinal (GI) tract[41]. CD is highly heterogeneous in manifestation and can affect the entire GI tract, diminishing quality of life and decreasing life expectancy[5,42]. Incidence of CD is on the rise globally and is highest in the US and European countries[2]. While current treatments often achieve remission in patients, they do not cure CD, and inflammation recurs in most patients, often requiring surgical intervention[43]. Thus, there is an

urgent need to better understand the cellular and genetic basis of CD and to develop effective strategies for therapeutic intervention and disease prevention.

While the etiology of CD is still poorly understood, it is generally believed to result from a complex interplay of host genetics with environmental factors and microbiota[44]. Untangling this interplay has been challenging due to extensive cellular heterogeneity within the gut and the contribution of multiple cell types to CD pathology. Adaptive and innate immune cells are linked to disease risk and severity[45]. Disruption of the epithelial barrier is characteristic of lesions in CD, directly implicating epithelial cell dysfunction in disease etiology[46,47]. Epithelial cells both react to pro-inflammatory signals and mediate inflammatory responses themselves, and cycles of epithelial healing and injury are characteristic of CD[48]. Fibroblasts are essential mediators in wound healing of the epithelium; however, in chronic inflammatory conditions they contribute to fibrosis and strictures[49,50]. Although CD is known to involve a wide range of cell types, delineating their individual contributions and disentangling causal roles from secondary involvement remains a major challenge in this complex disease.

Genome-wide association studies (GWAS) have linked more than 250 loci to IBD risk, providing a rich resource for untangling IBD biology and identifying therapeutic targets[9–11,51]. IBD risk loci and genes implicate a wide range of biological processes, including microbial sensing, adaptive immunity, and epithelial barrier function[47]. However, most of the putative causal variants are non-coding, thus making it challenging to identify their target genes or characterize their mechanism of action. Moreover, to translate these genetic signals into improved understanding of disease

biology requires the identification of the precise cell types and contexts in which IBD-associated loci exert their effects on disease risk.

Single-cell RNA-seq (scRNA-seq) studies have used cell-type-resolved gene expression profiles to identify changes in cellular composition and gene expression associated with CD[24,26,52], and to interpret genetic associations with IBD, revealing candidate genes and cell types that contribute to disease risk. Still, scRNA-seq alone is insufficient for prioritizing causal non-coding disease variants and nominating putative target genes based on gene expression level is complicated when genes are broadly expressed across cell types.

Open chromatin profiling – in particular, single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) – can complement studies of single-cell gene expression by capturing active regulatory elements that are not directly profiled by scRNA-seq. Indeed, atlases of candidate cis-regulatory elements (cCREs) have led to the discovery of cell-type-specific regulatory features, linked complex disease risk to specific cell types and pathways[29,30,34–37,53–55], and identified quantitative trait loci associated with chromatin accessibility (caQTLs)[56,57]. Importantly, recent cellular atlases of gut mucosa in healthy adults and children with CD, respectively, demonstrated that cell-type-resolved chromatin accessibility in the gut can provide some insights into cell types mediating IBD risk[25,58,59].

Here we generated a cell-type-resolved atlas of chromatin accessibility of the intestinal mucosa using pinch biopsies obtained from the ascending colon (AC) and terminal ileum (TI) of individuals with and without CD (Fig. 2.1a). We used this atlas to identify and characterize gene

regulatory programs in active CD (Fig. 2.1b), which allowed us to identify cell-types that mediate CD risk and prioritize non-coding CD risk variants (Fig. 2.1c).

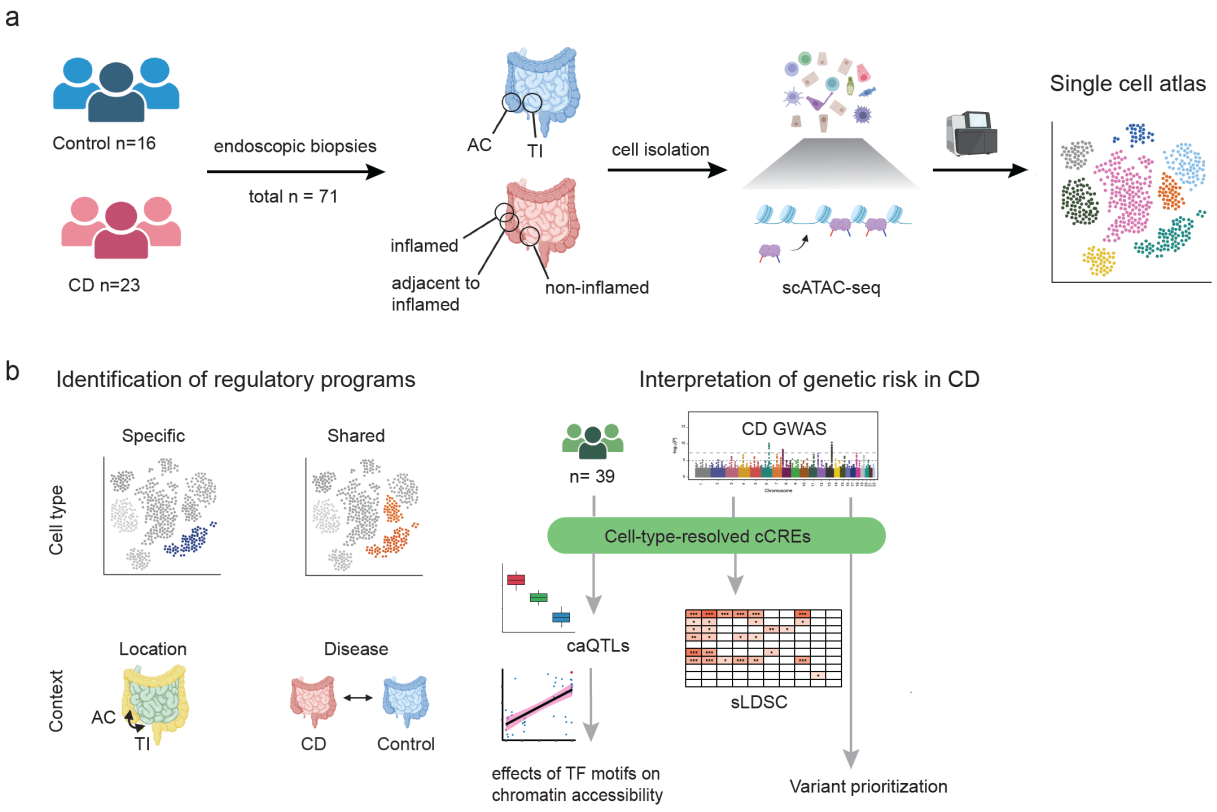


Figure 2.1 Schematic overview of study design.

a, A total of 71 biopsies were collected from the terminal ileum or ascending colon of 16 control individuals and 23 Crohn's disease (CD) patients. These biopsies were stratified into four groups: control tissue, non-inflamed tissue from CD patients, non-inflamed tissue adjacent to inflamed regions, and inflamed tissue. Cell-type-resolved atlas of chromatin accessibility was constructed using scATAC-seq data from each biopsy. **b**, Schematic illustrating biological contrasts explored using this scATAC-seq atlas. **c**, Schematic outlining integrated and complementary approaches to understand effects of common genetic variants on chromatin accessibility, measure cell-type-specific IBD risk, and prioritization of risk variants and cell-types through integration of these data with finemapped CD GWAS variants.

Results

A regulatory atlas of the intestinal mucosa

We obtained terminal ileum (TI) or ascending colon (AC) biopsies from 23 CD patients and 16 healthy controls which we used to isolate single cells and perform scATAC-seq (Fig. 2.1a, Fig.

S2.1). After quality control we retained 178,030 cells and followed an established computational workflow to cluster the data using ArchR[32] (**Methods**). We distinguished 29 clusters which we annotated with cell-type labels from matched scRNA-seq data (Methods, Fig. 2.2a, Fig. S2.2); in-depth analysis of the scRNA-seq data is described elsewhere (Zhou et al*., in preparation*). These cell type clusters show characteristic chromatin accessibility patterns at loci of previously published marker genes[24,60,61] (Methods) (Fig. 2.2c).

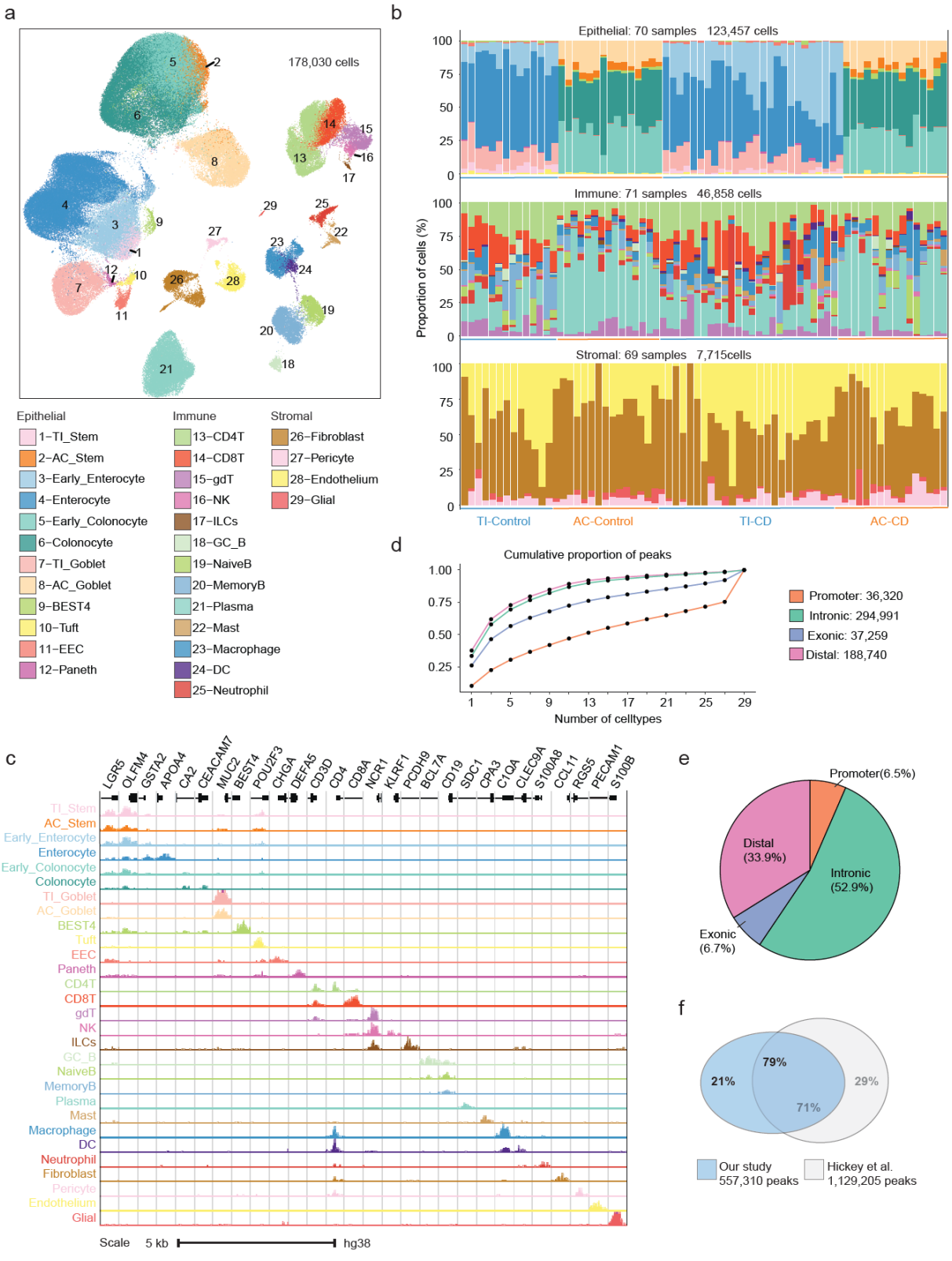


Figure 2.2 Atlas of cell-type-resolved chromatin accessibility in the human gut.

Figure 2.2, continued

a, UMAP visualization of 178,030 cells from 29 distinct cell types across epithelial, immune and stromal lineages in the terminal ileum and ascending colon. **b**, Sample-wise proportions of cell type clusters separated by lineage (epithelial, immune, stromal). **c**, Normalized chromatin accessibility at differentially accessible regions (DARs) near marker genes for each cell type cluster. Gene names and structures are depicted above the accessibility tracks. Individual loci are separated by grey vertical lines. **d**, Comparison of the proportion of peaks shared between cell types in different peak categories (promoter, distal, intronic, exonic). Shown is the cumulative proportion of accessible peaks shared in a given number of cell types in each of the different categories. **e**, Proportions of peaks overlapping distinct genomic categories (promoter, distal, intronic, exonic). **f**, Proportion of overlap between the union peak set generated in this study and the union peak set published in Hickey et al[58].

Our atlas captures cells of the epithelial, immune, and stromal lineages and represents all major cell types previously identified in transcriptome-based cellular atlases[27,58,60,62]. Epithelial cells represent the largest proportion of cells in our dataset (69.3%, 123,457 cells), followed by immune cells (26.3%, 46,858 cells) and stromal cells (4.3%, 7,715 cells) (Fig. 2.2b). Specifically, we identified 12 epithelial cell clusters, most of which further separate into distinct clusters based on their site of origin (AC and TI) (Fig. 2.2a, c, Fig. S2.3a, b). These include large clusters of mature absorptive cells in the TI (Enterocytes) and AC (Colonocytes), as well as progenitor and stem cell populations. We also found prominent clusters of goblet cells from both locations. In addition, we detected clusters representing smaller cell populations, including Paneth cells which are restricted to the TI, enteroendocrine cells (EECs), and Tuft cells. In contrast to epithelial cells, immune and stromal cells do not separate based on site of origin (Fig. S2.3 c, d, e, f).

Major cell types of adaptive immunity are CD4+ T cells and CD8+ T cells, γδT cells, NK and innate lymphoid cells (ILCs), plasma B, memory B, naive B, and germinal center B cells. Among innate immune cells we detected macrophages, mast, neutrophils and dendritic cells (DCs). Plasma B cells are most abundant (30%) among immune cells, followed by CD4+ and CD8+ T cells (Fig.

2.2b). We observed four distinct cell types within the stromal lineage, corresponding to fibroblasts, endothelial cells, pericytes, and glial cells. Due to the smaller number of stromal cells, we did not attempt to subcluster these further.

To identify cCREs for each cell type, we called peaks of accessible chromatin using MACS2[63] on each cluster separately (Methods). We combined all peaks into a union set comprising 557,310 cCREs (Fig. S2.4a). Most of these are located in introns (52.9%) or distal intergenic regions (33.9%), while 6.5% are located within or proximal to promoters (Fig. 2.2e). As observed previously, cCREs in or near promoters are shared across more cell type clusters compared to distal and intronic peaks[64] (Fig. 2.2d). A large proportion of all identified cCREs (79%) overlap regions reported in a multiomic reference atlas of the healthy human intestine[58] (Fig. 2.2f). To test which transcriptional regulators are associated with regulatory features in each cell type cluster, we performed motif enrichment analysis on cell-type-specific peak regions (Fig. S2.4b) (Methods). Many of the top-ranking motifs in each cell type correspond to transcription factors (TFs) known to be master regulators in that cell type (e.g., HNF4a in enterocytes and RUNX3 in T cells), and motif accessibility is correlated with expression level of the respective TF across cell types (Fig. 2.4c). This analysis confirmed that the cell-type-resolved cCREs capture sequence motifs of cell-type-specific transcriptional regulators. Together, these data represent a comprehensive atlas of cell-type-resolved cCREs in the human intestinal mucosa comprising samples from healthy adult and CD patients.

Cell-type-specific, shared, and context-specific regulatory features in the intestinal mucosa

Discrete cell type clusters do not fully capture the complex and dynamic gene regulatory regions and programs in the intestinal mucosa which might be shared across cell types or associated with continuous processes (e.g., differentiation), contexts (e.g., inflammation) or location (e.g., AC vs TI). To complement the cluster-based cell type annotations we therefore adopted a grade of membership approach based on topic modeling[65,39,38]. By allowing for partial memberships, this approach is able to model continuous or indistinct boundaries between cell types, and by allowing for membership in multiple topics, it is able to model regulatory processes that might be shared across multiple cell types. To avoid over-representation of abundant cell types we downsampled each cluster to a maximum of 100 cells per individual (Methods). Examining the accessible regions associated with each topic, we leveraged enriched TF binding motifs to infer active TFs, and identify their putative target genes (Methods, Supplementary File 2-3).

The majority of topics capture specific cell types or multiple closely related cell types within a particular lineage. In addition, we identified topics capturing shared effects, e.g., associated with biopsy location and inflammation status. Finally, several topics are shared broadly across cell types capturing basic features of the gene regulatory architecture.

We considered a topic to be cell type specific if the average topic proportion for cells within a given cluster was at least 60% and 5-fold higher than for the second highest cluster (Methods). Topics defined primarily by a cell type are typically enriched for well-known motifs and genes associated with that cell type. For example, Topic 5 is associated with Goblet cells in the TI (Fig. 2.3a, b, Fig. S2.5a) and topic-peak regions are enriched for motifs of FOXA1/2/3 and KLF3/4/5. Putative target genes of this topic include *MUC2* and *FCGBP* and are enriched for O-linked

glycosylation, maintenance of gastrointestinal epithelium, and other functions (Fig. 2.3b, c, Supplementary File 2-3). Of note, we also detected many topics that are shared between related cell type clusters or associated with specific contexts in our dataset (Fig. 2.3a, b). For example, Topic 18 is strongly enriched in DCs and to a lesser degree, macrophages, likely reflecting the shared origin and overlapping function of these cell types. The associated accessible regions are enriched for motifs of TFs active in DCs, including SPI1 and IRF4, and are located near genes enriched for antigen presentation and chemokine signaling pathways (Fig. 2.3a, c, Fig. S2.3a, Supplementary File 2-3). Topic 29 captured regulatory features associated with both early and mature colonocytes (Fig. 2.3a, b). These regions are enriched for motifs associated with TFs active in colonocytes (e.g., CDX1/2) (Fig. 2.3a, Fig. S2.3a, Supplementary File 2-3) whose targets include genes known to be expressed in colonocytes (e.g., *MS4A12*, *CA2*, and *SATB2*; see Fig. 2.3c, Fig. S2.3a, Supplementary File 2-3).

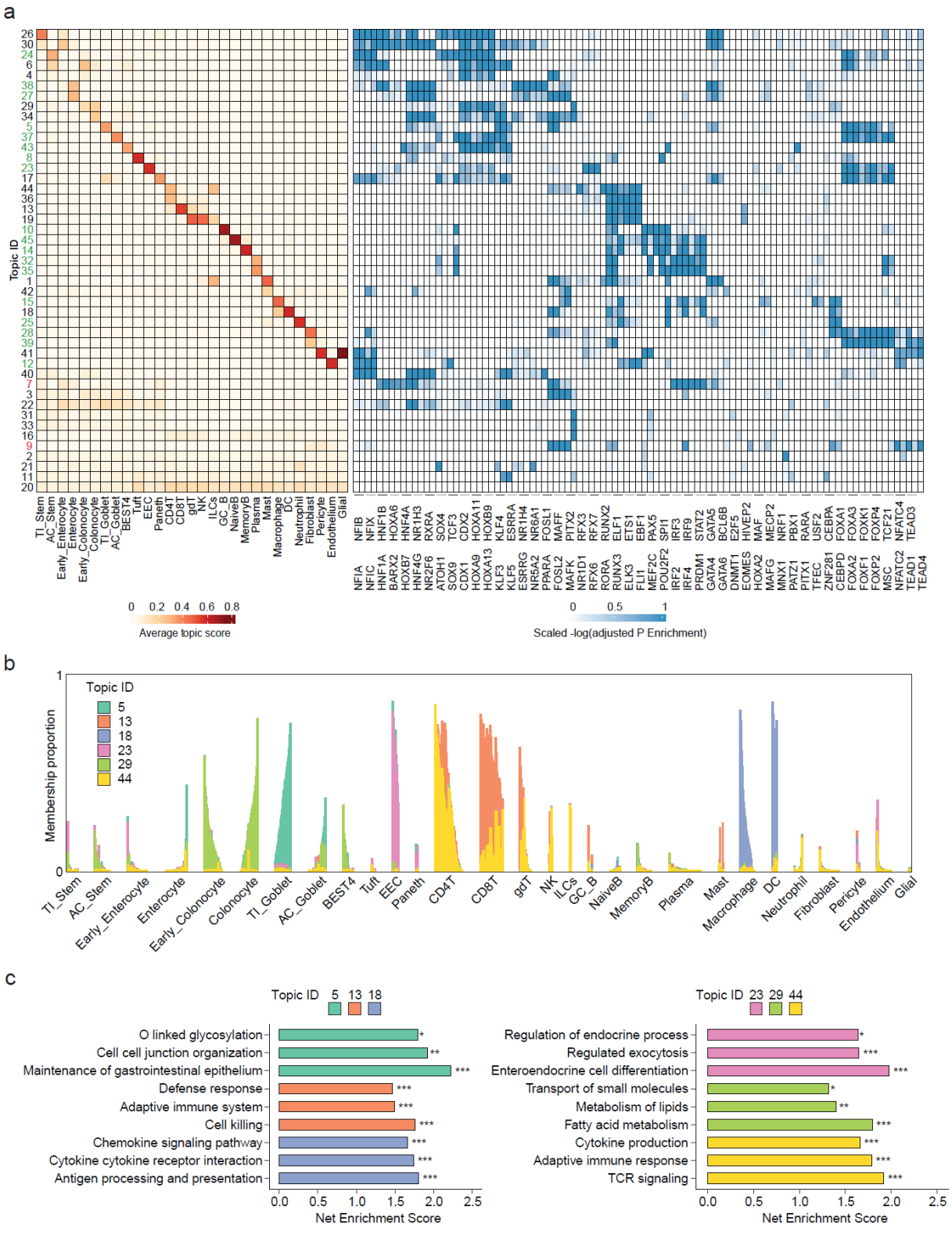


Figure 2.3 Identification of shared, cell-type-, and context-specific regulatory programs through topic modeling.

Figure 2.3, continued

a, Average topic membership for cells grouped by cell type clusters (left). Topics specific to cell type and inflammation are labeled in green and red, respectively. TF motif enrichment for each topic (right). Shown are the significant enriched TF motifs for each topic (adjusted p-value < 0.01). The -$\log_{10}$(adjusted p-value) for motif enrichment was capped at a maximum of 30, with any values exceeding 30 being set to 30. These capped values were then scaled by the maximum value for each TF and plotted. **b**, Structure plot displaying proportion of selected topics in individual cells across all cell types. Each column represents a single cell and cells are grouped by cell types. **c**, Pathway enrichment analysis based on genes associated with each of the selected topics. Pathways significantly enriched for intestine-related biological functions (FDR < 0.05) were selected and visualized. (***: FDR ≤ 0.001, **: FDR ≤ 0.01, *: 0.01 < FDR ≤ 0.05)

Finally, topics 11 and 2 are shared across all cells and represented mostly cCREs with invariant chromatin accessibility across cell types, such as promoters (promoter cCREs make up 38% of topic 11-specific cCREs while the average proportion of promoter peaks is 9.2% (s. d. =11%) across all topics) and distal elements enriched for CTCF motifs (topic 2) (Fig. 2.3a, Fig. 2.5a, Supplementary File 2-3). Topic analysis therefore captures regulatory features that are also identified using discrete cell type clusters. In contrast to cluster-based analysis alone, however, we detected many topics that are shared between related cell type clusters or associated with important biological variables in the intestinal mucosa in healthy individuals and CD patients (Fig. 2.3a, b).

Pervasive changes in regulatory landscape during inflammation in the epithelial lineages

Inflammation is a defining characteristic of CD and underlies its clinical symptoms. To discover cell-type-specific regulatory programs involved in mucosal inflammation, we leveraged topic analysis to identify inflammation-associated cell types, transcriptional regulators, and cCREs.

We found two topics, topics 7 and 9, to be strongly associated with inflammation. Topic 7 proportion is significantly increased in epithelial cells in biopsies taken directly from inflamed

regions but shows only moderate enrichment in patient-matched biopsies from non-inflamed adjacent regions and none in biopsies from healthy controls (Fig. 2.3a, 2.4a, b).

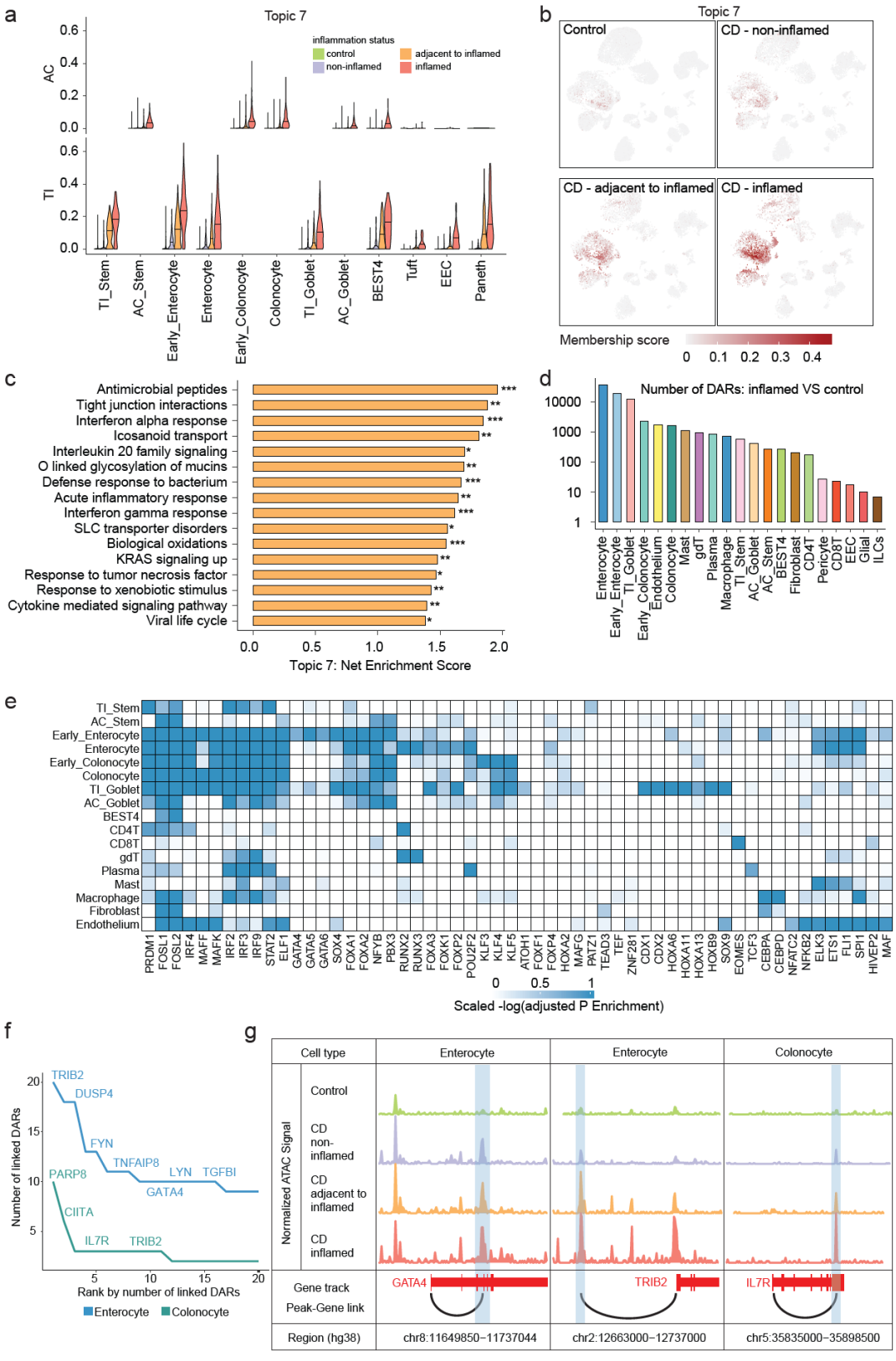


Figure 2.4 Regulatory changes associated with inflammation.

Figure 2.4, continued

a, Proportion of topic 7 membership for each cell. Cells grouped by cell type, tissue location, and inflammation status. **b**, UMAP plots displaying proportion of topic 7 membership for each cell, split by inflammation status. **c**, Pathway enrichment analysis based on genes associated with topic 7 (***: FDR ≤ 0.001, **: FDR ≤ 0.01, *: 0.01 < FDR ≤ 0.05). **d**, Number of inflammation-associated DARs (iDARs) with increased accessibility in inflamed tissue compared with control tissue in each cell type (log fold change ≥ 0.5, FDR ≤ 0.1). **e**, Enriched TF motifs in iDARs. The -$\log_{10}$(adjusted p-value) for motif enrichment was capped at 5, with any values exceeding 5 being set to 5, and then scaled by the maximum value per TF. **f**, Identification of highly regulated inflammatory genes based on the number of linked iDARs. iDARs were linked to genes based on the activity-by-contact (ABC) model[66]. **g**, Examples of genomic loci containing iDARs upregulated in inflamed tissue and putative target genes obtained from ABC model.

Motif enrichment within topic 7 cCREs implicates AP-1 factors (e.g., FOSL1/2) and Interferon response factors (e.g., IRF2/3/9), as the main regulators of inflammation within epithelial cells (Fig. 2.3a). Putative targets are enriched for pathways associated with inflammatory processes, including interferon responses, antimicrobial peptides, and icosanoid transport (Fig. 2.4c). Notably, these target genes included *LCN2*, *NOS2*, *DUOX2, DMBT1, REG1A, and REG1B,* which were recently reported as characteristic of epithelial cells in patients with active CD[26] (Supplementary File 2-3). We thus identified the regulatory features associated with a broad inflammatory response in epithelial cells, suggesting that this response is orchestrated by a defined set of pro-inflammatory TFs. Topic 9 is enriched across stromal cells in inflamed samples (Fig. 2.3a, Fig. S2.5b, c), in particular fibroblasts from AC. Topic 9 cCREs were enriched for motifs for FOSL1/2, STAT2, IRF4, but also included NFKB2.

Our topic analysis revealed inflammatory responses that are shared broadly across many cell types. To test whether inflammation was associated with cell-type-specific changes in cCRE activity we compared 21 cell types from inflamed and control samples to identify differentially accessible regions (DARs) (Methods). We focused our comparison on this contrast since we observed a

strong difference between healthy control and inflamed CD samples in the topic analysis, and because endoscopic assessment of inflammation is often not sufficient to score mucosal inflammation in CD patient. Epithelial cell types yielded the most inflammation-associated DARs (iDARs, Supplementary File 2-4), in particularly in the TI (enterocytes, early enterocytes, goblet cells) (Fig. 2.4d, methods). Epithelial iDARs are enriched in motifs for pro-inflammatory TFs (IRF, STAT2, MAFF, MAFK, FOSL1/2) and NFYB, ELF1 similar to topic 7 cCREs. This analysis also identified motifs enriched in iDARs of one or only a few cell types. For example, iDARs in early enterocytes are enriched for motifs of TFs associated with lineage specification and regeneration/injury (SOX4, FOXA1/2, GATA4/5/6). iDARs in TI Goblet cells were enriched for FOXA1/2, FOXA3 (Hickey et al) SOX4, CDX1/2, SOX9, and ATOH1. Of note, topic 17 is similarly enriched for motifs of ATOH1, SOX4 and associated with Paneth cells in TI in all conditions, this topic appears specifically enriched in TI goblet cells from inflamed samples (Fig. 2.3a). These observations potentially point toward plasticity in the secretory lineage in response to inflammation. In enterocytes and colonocytes we found that iDARs cluster around many genes that were previously associated with inflammatory responses and regeneration, including GATA4 (10 iDARs), TRIB2 (20 iDARs), and IL7R (3 iDARs) (Fig. 2.4f, g).

We identified 1702 iDARs in endothelial cells that are enriched for motifs of pro-inflammatory transcriptional regulators IRF4, STAT2, NFkB2 and AP-1 like factors, suggesting a significant inflammatory transcriptional response in this cell type which plays a key role in the regulation of cellular and molecular access to inflamed regions[67] (Fig. 2.4e). The iDARs we detected in other cell types were enriched for motifs of TFs associated with inflammation or master regulators of specific cell types. For example, among immune cells, CD8+ T cells are enriched for EOMES,

previously implicated in the differentiation of effector CD8+ T cells[68], and Plasma B cells are enriched for POU2F2 (OCT2), a TF critical in Plasma cell maturation[69]. Together these data helped to infer the key regulatory TFs driving cell-type-specific regulatory programs in mucosal cells of patients with active CD.

Testing cell-type-specific regulatory sequence motifs using chromatin accessibility QTLs

Our study generated a catalogue of cell-type-resolved cCREs which we used to infer sequence determinants and TFs regulating chromatin accessibility. To directly test the relationship between DNA sequence (e.g., TF binding motifs) and chromatin accessibility, we leveraged our cohort of 39 individuals to map caQTLs. We obtained genome sequences for each individual directly from scATAC-seq data and imputed their genotypes based on the 1000G panel (Methods). We used RASQUAL to separately map caQTLs in each of the 19 cell types that were represented by at least 20 individuals. We detected between 8533 (early colonocytes) and 21 (neutrophils) caQTLs per cell type (FDR<10%, MAF>5%) (Fig. 2.5a). The number of detected caQTLs depended on the total number of cells constituting the cluster, highlighting substantial differences in power between cell types (Fig. 2.5b). However, applying multivariate adaptive shrinkage (mash[70]) to the identified caQTLs revealed substantial sharing between related cell types (Fig. S2.7h, Method), as did the correlation pattern of caQTL effect sizes between pairs of cell types (Fig. S2.7i). For the subsequent analysis we used the original RASQUAL caQTLs identified for each cell type. Compared to random SNPs from our test set, RASQUAL caQTLs are enriched within promoters, they tend to be located near the center of the caPeaks (Fig. S2.7 b-e), and they are less conserved than non-caQTL Peaks (Fig. S2.7g). caQTL peaks have higher peak scores on average, likely reflecting higher detection power for these regions (Fig. S2.7f).

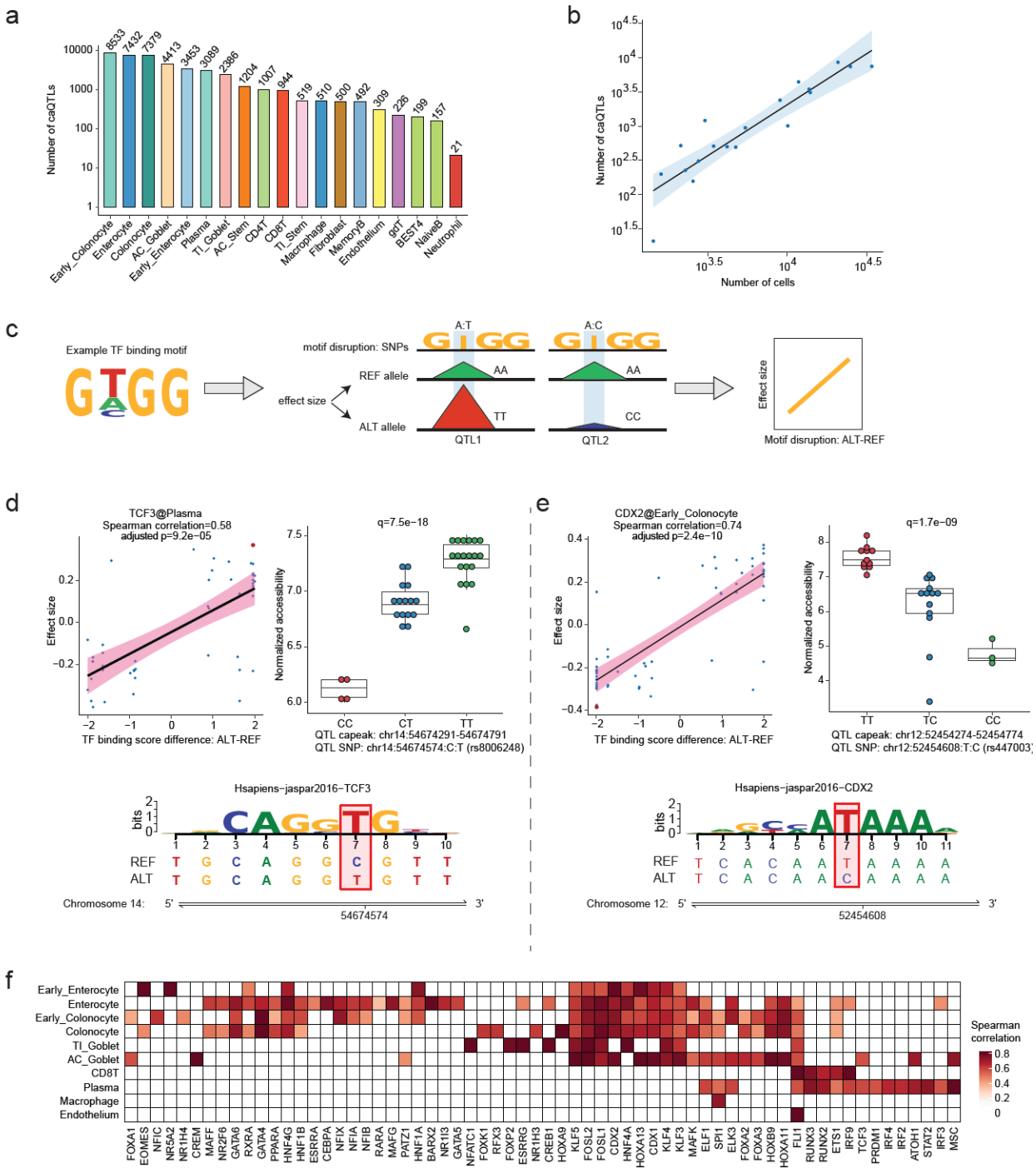


Figure 2.5 Identification of chromatin accessibility quantitative trait loci (caQTLs) in intestinal cell types.

Figure 2.5, continued

a, Number of caQTLs identified in each cell type using RASQUAL. **b**, Relationship between the number of detected caQTLs and the number of cells in each cell type (Spearman correlation rho=0.90, p-value < 2.2e-16). **c**, Illustration of analysis to measure effects of motif disrupting SNPs on chromatin accessibility changes. In the example, differences in motif affinities for the TF at alternative alleles from QTL1 and QTL2 are associated with different accessibilities of the corresponding caPeaks. **d**, Correlation of caQTL effect sizes with motif disruption scores (MotifBreakR) for QTL SNPs predicted to disrupt TF sequence motif for TCF3. Shown are QTL SNPs located within caPeaks in plasma cells. Effect sizes and motif disruption scores were calculated relative to reference allele. Significance of association was calculated using Spearman correlation. Box plots show peak accessibilities in different genotypes illustrating the effect of a specific SNP (highlighted in red in the regression plot). Genome tracks in the bottom show reference and alternative alleles at the QTL SNP site. **e**, Same as in **d**, showing chromatin accessibility at CDX2 motifs in early colonocyte clusters. **f**, Heatmap of Spearman correlation coefficients between QTL motif breaking scores and effect sizes for all TF-cell type pairs. Coefficients are only displayed for significant correlations (FDR <0.05).

In principle, caQTLs allow us to detect TFs that control chromatin accessibility by observing the relationship between TF motif disruption and changes in chromatin accessibility[71,72]. For each cell type, we obtained TF motif instances that overlapped with a caQTL and were predicted to be significantly altered by that SNP (Methods). We then tested the correlation between caQTL effect size and motif disruption score separately for each set of motifs and retained motifs with significant correlation in at least one cell type (Fig. 2.5c). We observed a significant correlation between motif disruption and chromatin accessibility in at least one cell type for 64 TF motifs (adjusted p-value ≤ 0.05, Spearman's correlation coefficient) (Fig. 2.5d-f). For these specific sites, we found strong evidence that disruption of binding motifs of key lineage-specific TFs is associated with reduction in chromatin accessibility at these cCREs; for example, TCF3 in Plasma cells (Fig. 2.5d) and CDX2 in colonocytes (Fig. 2.5e). While the caQTL analysis was performed across all samples regardless of disease status, we detected a significant correlation for IRF9 and IRF3 in enterocytes, providing additional evidence that these proinflammatory TFs regulate chromatin accessibility at cCREs in this cell type (Fig. 2.5f). These results were robust to the selection of SNP sets (Fig

S2.7j). Our analyses thus demonstrate that integration of inter-individual genetic variation into cellular atlases can be used to causally link cell-type-specific sequence determinants to chromatin accessibility and to validate the functional impact of motifs identified through cell-type-specific motif enrichment.

Cell-type-specific enrichment of IBD risk

The CD gut cell atlas provided us with the unique opportunity to ask which cell types and which context-specific regulatory programs might mediate IBD risk. We used stratified LD-score regression (sLDSC[73]) to assess heritability enrichment from GWAS summary statistics for IBD and its subtypes CD and UC in cell-type- and topic-specific chromatin accessibility[10]. As reference traits we also included Celiac disease[74], Rheumatoid arthritis (RA)[75], Alzheimer's disease[76], Parkinson's disease[77,78], Type 1 Diabetes (T1D)[79], Type 2 Diabetes (T2D)[80], and Body Mass Index (BMI).

We found ten of the 29 cell types to be associated with enriched IBD risk based on sLDSC ($\tau^* > 0$, FDR < 10%), all of which represent immune cells (Fig. 2.6a). T cells (CD4, CD8, γδT) and, to a lesser degree, NK cells are enriched for IBD, CD, and UC risk, as well as for Celiacs disease, RA, and T1D. GC, Naïve B, and memory B cells are less strongly enriched for IBD/CD/UC and show higher enrichment for RA and T1D, likely reflecting the humoral component of these autoimmune diseases[81–84]. Of note, we found that risk for CD, but not risk for UC, is strongly enriched in innate immune cells (macrophages, DC, neutrophils). This suggests that known differences in disease etiology and genetic architecture between UC and CD are partly caused by the cell-type-specific contribution to disease risk[85,86].

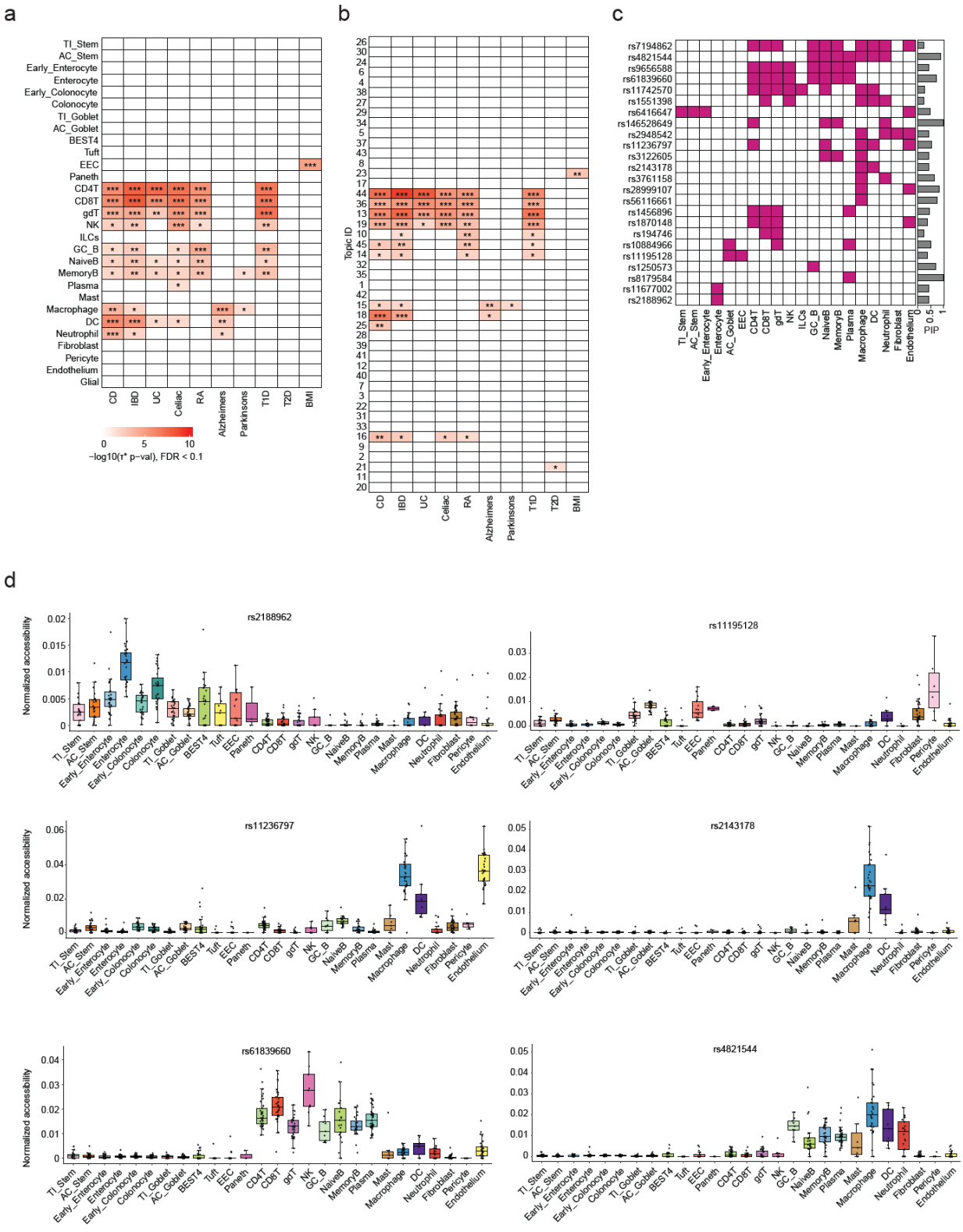


Figure 2.6 Cell-type-specific genetic risk for IBD.

Figure 2.6, continued

a, Enrichment of trait heritability in cell-type-specific peaks using s-LDSC. Shown are cell-type-specific -$\log_{10}$ p-values for positive $\boldsymbol{\tau}^*$ evaluating autoimmune and control traits. Only statistically significant results (FDR<10%) are colored. IBD=Inflammatory Bowel Disease, UC=Ulcerative Disease, RA=Rheumatoid Arthritis, T1D=Type 1 Diabetes, T2D=Type 2 Diabetes, BMI=Body Mass Index. **b**, Topic-specific -$\log_{10}$ p-values for positive $\boldsymbol{\tau}^*$ evaluating autoimmune and control traits. Only statistically significant results (FDR<10%) are colored. **c**, Shown are 24 SNPs were prioritized based on posterior inclusion probabilities (PIPs ≥ 0.2) and overlap with cell-type-specific cCREs. Ruby red color indicates that the fine-mapped SNP is located within cell-type-specific cCREs. PIPs for each SNP are shown as bar graphs on the right of the heatmaps. **d**, Distribution of normalized chromatin accessibility of pseudo-bulk cell groups per cell type per individual in the cCRE overlapping selected finemapped SNP.

In contrast, none of the epithelial or stromal cell populations are enriched for IBD risk. However, we detected strong enrichment for BMI heritability in EEC-specific peak regions, confirming a previous observation[58] and suggesting that part of the genetic contribution to BMI is mediated via EECs. We next used topic-associated peaks to assess association of contexts with IBD risk. Our results confirmed the association of specific immune cell types with IBD risk and the difference in CD and UC risk heritability in innate immune cells (topics 15, 18, 25). Similarly, we found no evidence for enrichment of IBD risk in epithelial (topics 3-7, 17, 22-24, 26, 27, 29-31, 33, 34, 37, 38, 40) or stromal cells (topics 9, 28, 39) using topic cCREs (Fig. 2.6b).

In contrast to cell-type-cluster-based enrichment, topic analysis also captured inflammation-associated cCREs. However, topics 7 and 9, corresponding to epithelial and stromal cells, respectively, were not enriched for IBD, CD or UC heritability. Together these analyses point to immune cell types with large contributions, while suggesting that loci mediating strong inflammatory responses in the epithelium and stromal cells, have a more limited contribution to overall genetic IBD risk. Our analysis provides an example of how context-specific epigenetic data

and genetic data can be combined to disentangle (cell-type-specific) genetic effects from their cellular consequences in complex diseases.

Cell-type-resolved cCREs enable prioritization of candidate regulatory CD variants across mucosal cell types

In contrast to the results of our heritability enrichment analysis, prior data implicated genes with epithelial function in IBD and its subtypes, including *CIITA* and *SLC22A4*[87–89]. sLDSC does not test for cell type specificity of individual loci, but instead obtains a genome-wide estimate of the focal heritability enrichment for a given annotation across all loci (τ^*)[73,90,91]. Thus, sLDSC likely misses cell types for which the contribution to overall genetic risk comes from a small number of loci. To link risk loci to cell-type-specific activity, we fine-mapped the CD GWAS signal and obtained posterior inclusion probabilities (PIPs) for variants in 91 loci (Methods). We noted that variants with high PIP are enriched within cCREs (enrichment=3.5-fold, p-value=0.0096, Fisher's exact test; Methods, Fig S2.8a, b), and in 24% of loci (20/85) at least 50% of the total PIP is associated with cCREs (Fig S2.8c). We identified 29 variants in 24 loci (25 credible sets) with a PIP $\geq$ 0.2 that fall within cCREs (Table 2.1). Where possible, we nominated a putative target gene for these variants using co-accessibility[31,32], published Activity-by-contact (ABC) scores[66,92], or distance as criteria. Out of the 29 variants, 24 overlap with at least one cell-type-specific cCRE (Fig. 2.6c). As expected, most of these regions are accessible in adaptive and innate immune cells. However, we identify 10 variants overlapping a region identified as specific to at least one epithelial or stromal cell cluster, implicating these cell types in mediating IBD risk (Fig. 2.6c). For example, rs2188962 (PIP=0.42) is located within a region broadly accessible in epithelial cells, most strongly in mature enterocytes and colonocytes (Fig. 2.6d), and rs11195128 falls into regions

accessible in secretory epithelial cells (Fig. 2.6d). Additional prioritized SNPs are located in cCREs accessible in T and B cells, Macrophages and DCs, as well as endothelial cells. Several of the prioritized SNPs have been implicated in IBD and validated through functional studies (e.g., rs11236797 and rs61839660[93], Table 2.1). Together, these analyses implicated a range of immune and non-immune cells in the intestinal mucosa in mediating CD risk through regulatory variants.

Of note, we also identified two loci in which two prioritized non-coding variants from the same credible set overlapped separate regulatory elements. In the first locus, rs1456896 (PIP = 0.46) and rs9656588 (PIP = 0.42) which are separated by 2.3kb, overlap adjacent cCREs, both of which are linked to IKZF1 by co-activity (Fig. 2.7a, b, c). Both cCREs display similar accessibility patterns across cell type clusters, with the highest accessibility in T cells. In the second locus, rs7194862 (PIP =0.25) and rs6416647 (PIP =0.49) are separated by 482 bp and located in cCREs linked to *CIITA*[94] (Fig. 2.7d, e, f), a transcriptional co-activator of *MHCI* and *MHCII* genes[94]. In this case, chromatin accessibility of the two cCREs harboring these variants differs significantly across cell types. While rs7194862 is located in a cCRE accessible in macrophages, DC, T cells and B cells and endothelial cells, rs6416647 is accessible in stem cells and absorptive epithelial cells (and Paneth cells), DCs and endothelial cells (Fig. 2.7d, e, f). These two examples highlight the potential contribution of multiple non-coding SNPs to CD risk within a single locus and mediation of this risk by different cell types**.**

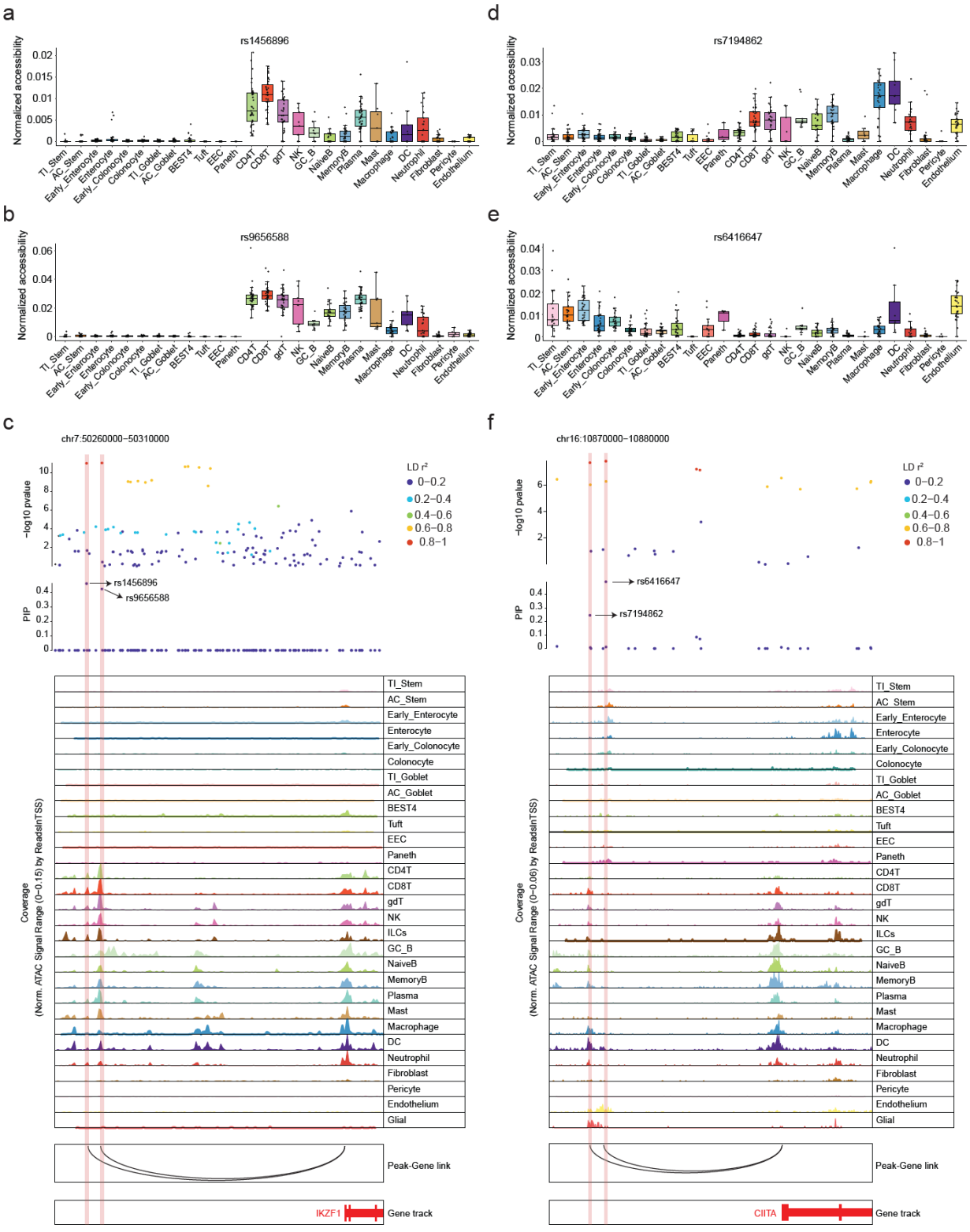


Figure 2.7 Candidate functional risk variant prioritization.

Figure 2.7, continued

a, **b**, **d**, **e**, Distribution of normalized chromatin accessibility of pseudo-bulk cell groups per cell type per individual in the cCRE overlapping the finemapped SNP. **c**, Fine-mapped CD risk SNPs rs1456896 and rs9656588 are from the same fine-mapped credible set. The top track represents the -log10 p-value of SNPs from CD GWAS, with color representing LD with the lead SNP. The middle track shows PIPs from fine-mapping results. In the bottom track, IKZF1 was linked to the cCREs by the co-activity of gene expression and chromatin accessibility. **f**, Fine-mapped IBD risk SNPs rs6416647 and rs7194862 are from the same fine-mapped credible set. In the bottom track, CIITA was linked to the cCREs through the ABC model.

Discussion

We assembled an atlas of cell-type-resolved chromatin accessibility within the intestinal mucosa of adult CD patients and healthy controls and used it to characterize cell-type- and context-specific gene regulatory programs across disease-relevant cell types and states. Our work contextualizes findings from previous scRNA-seq studies, which identified cell-type-specific and inflammation-associated gene expression signatures in CD[24,26]. In particular, we provided direct insights into the cCREs and TFs that define the regulatory landscape of the intestinal mucosa.

Cell-type-specific TF motif enrichment is a common approach for nominating putative regulatory elements[32,95]. If gene expression data is available, the enriched motifs can then be integrated with transcript levels for the cognate TFs. However, such approaches do not assess whether the enriched motifs directly contribute to establishing or maintaining chromatin accessibility in vivo. Our study, which included 71 tissue samples from 39 individuals provided us with the unique opportunity to functionally test the effects of TF motifs in putative cCREs across a range of disease-relevant cell types and contexts. By examining the overlap between caQTLs and putative regulatory SNPs in cCREs, we identified 64 TF motifs that directly affect chromatin accessibility in the intestinal mucosa - even with our moderate sample size. As cohort sizes continue to increase, we expect that

similar approaches will become a standard step in analyzing data from epigenomic atlases and to predict regulatory effects of GWAS SNPs[56].

scATAC-seq has been used across a wide range of tissue contexts to identify cell types mediating genetic risk for a variety of traits[91]. Similar to previous studies in intestinal tissues, we observed strong enrichment for IBD and CD risk in both in adaptive and innate immune cells[25,58]. While UC shares genetic architecture with CD, we found no or little enrichment of UC risk in innate immune cells, reflecting known differences between UC and CD and demonstrating both the sensitivity and specificity of our approach[86]. Unexpectedly, we found that neither epithelial nor stromal cells are enriched for CD risk, nor are inflammation-associated cCREs. This observation is at odds with the evidence for GWAS loci implicating genes with epithelial function[25,47,52,96]. As sLDSC measures enrichment scores across multiple loci (i.e., including but not limited to cCREs), it likely produced conservative estimates of heritability enrichment in cCREs. The combination of context-specific epigenomic data which detects strong inflammation associated changes in epithelial cell types and cell-type-specific risk primarily associated with immune cells prompts us to speculate that we are separating primary genetic risk from secondary effects in CD.

Fine-mapped variants within cCREs have previously been used to identify putative causal non-coding variants with direct effects on gene regulation[53,56]. Among the 29 putative regulatory variants of CD risk, we found 24 falling into cCREs with cell-type-specific accessibility patterns. Of note, several of these variants are located in cCREs with epithelial specific chromatin accessibility. Intriguingly, we also observed cases in which multiple variants in LD with similar PIP scores fall into adjacent regulatory elements, including one example where adjacent elements

implicate distinct cell types. Together, these prioritized variants highlight the gene regulatory complexity underlying CD and provide evidence for a broad involvement of both immune and non-immune cell types mediating CD risk.

rsID	PIP	cCRE	Associated gene	Link method	Lineage specificity	iDAR	Known IBD risk Gene [PMID]
rs8179584	1	chr2:172558327-172558827	PDK1	ABC	Immune		[38103512]
rs146528649	1	chr16:50626815-50627315	SNX20	Co-activity	Immune		[32410873]
rs2284553	0.97	chr21:33404169-33404669	IFNGR2	ABC			[31566580]
rs4821544	0.89	chr22:36862189-36862689	NCF4, CSF2RB	Co-activity	Immune		[18580884, 27377463]
rs28999107	0.83	chr12:6383870-6384370	PLEKHG6, TNFRSF1A, LTBR	Co-activity	Immune, Stromal		[30815272, 33568785]
rs56116661	0.75	chr3:188683082-188683582	LPP-AS1	ABC	Immune		
rs61839660	0.74	chr10:6052545-6053045	IL2RA	ABC	Immune	✓	[34226674]
rs1250573	0.7	chr10:79282308-79282808	ZMIZ1	Nearest	Immune		[30402962]
rs3761158	0.65	chr20:46006257-46006757	MMP9	ABC	Immune	✓	[27542125]
rs11236797	0.53	chr11:76588423-76588923	LRRC32	Co-activity	Immune, Stromal	✓	[32499651]
rs10884966	0.51	chr10:110425817-110426317	DUSP5	ABC	Epithelial, Immune		[26479789]
rs6416647	0.49	chr16:10871644-10872144	CIITA	ABC	Epithelial, Stromal	✓	[35275975]
rs1456896	0.46	chr7:50264832-50265332	IKZF1	Co-activity	Immune		[31057532]
rs3122605	0.44	chr1:206781236-206781736	IL10	Nearest	Immune		[28496525]
rs11677002	0.44	chr2:28391530-28392030	FLJ31356	ABC	Epithelial		
rs2143178	0.44	chr22:39264648-39265148	PDGFB	ABC	Immune		[37995775]
rs2948542	0.44	chr17:27529222-27529722	KSR1	ABC	Immune, Stromal		[34758847]
rs2188962	0.42	chr5:132434879-132435379	P4HA2, SLC22A4, SLC22A5	Co-activity	Epithelial		[38944815, 15503241]
rs9656588	0.42	chr7:50266738-50267238	IKZF1	Co-activity	Immune		[31057532]
rs11195128	0.37	chr10:110426330-110426830	DUSP5	ABC	Epithelial		[26479789]
rs1870148	0.37	chr10:80511347-80511847	LOC102723703	ABC	Immune, Stromal		
rs4807569	0.32	chr19:1122945-1123445	MIDN, EFNA2	Co-activity			[37782162]
rs2031754	0.29	chr9:4965364-4965864	JAK2	Nearest		✓	[32203403]
rs11742570	0.26	chr5:40410171-40410671	PTGER4	ABC	Immune		[36155972]
rs7194862	0.25	chr16:10871018-10871518	CIITA	ABC	Immune, Stromal	✓	[35275975]
rs194746	0.23	chr14:68815969-68816469	ZFP36L1	ABC	Immune		[30402962]
rs7608697	0.23	chr2:60977210-60977710	PUS10	Nearest			[21298027]
rs2024092	0.23	chr19:1123791-1124291	SBNO2	Nearest		✓	[38806456]
rs1551398	0.22	chr8:125527564-125528064	TRIB1	Nearest	Immune	✓	[28187197]

Table 2.1 Prioritized risk SNPs.

Table 2.1, continued

Risk SNPs with a PIP greater than 0.2 that overlapped with a cCRE were listed. The target gene was first linked to the SNP based on co-activity. If no genes showed co-activity with the overlapping peak, the ABC-Max method was used to assign the associated gene. In cases where neither approach was applicable, the nearest gene was assigned. The lineage specificity is summarized from the cell type specificity of the overlapped cCRE. The iDAR column specifies whether the overlapped cCRE is classified as an iDAR.

Methods

Ethics statement

This study protocol (15573A) was approved by the Institutional Review Board of The University of Chicago. Gastroenterologists collected samples and clinical data from patients diagnosed with Crohn's Disease (CD) and non-IBD control. Mucosal samples were collected during routine endoscopic screening in healthy controls, or during diagnosis or follow-up colonoscopy in CD patients. All patients provided signed consent for research use of their mucosal biopsies and storage of their demographic data. Samples, clinical metadata and demographic data were collected by clinical research coordinators and provided anonymously to the research team. This study complied with all relevant ethical regulations.

Patient information

Non-IBD controls were recruited by the time of routine colonoscopy. Non-IBD controls were free of a history of IBD, an auto-immune disease, an ongoing infection in intestines, a history of Familial adenomatous polyposis, a history of primary digestive tract malignancies, or a history of metastatic malignancies in digestive tract. CD patients were recruited at the diagnosis or follow-up colonoscopy. Those patients with prior clinical diagnosis of CD, were reported either in clinical

remission or with active disease. Patient and sample metadata are provided in Supplementary File 2-1.

Tissue biopsy sample collection

We aimed to obtain samples from two anatomic locations, terminal ileum (TI) and ascending colon (AC), and in the case of CD patients with active inflammation from inflamed sites and adjacent regions without macroscopic inflammation. The specific sampling regions within anatomic locations were determined by the gastroenterologists performing the colonoscopy. The number of pinches were limited to 6 per individual site and 12 per patient. CD patients provided up to twelve biopsies, including up to two regions (TI and AC), and three different endoscopic diagnoses (inflamed, adjacent to inflamed, noninflamed)[97]. In most cases with active inflammation, CD patients only provided biopsies from a single anatomic location (i.e. AC or TI).

Non-IBD controls provided up to six biopsies from each region (TI and AC). Mucosal linings were obtained using endoscopic biopsy forceps following standardized care. Mucosal tissues after biopsy were immersed in iced cold Advanced DMEM/F-12 [Thermo Fisher 12634010] in a 1.7 mL eppendorf tube and transferred in a cooler. The maximum wait time between biopsy and sample processing was two hours.

Sample dissociation and nuclei extraction

Single cells from mucosa were obtained following a modified protocol of previous studies[27,98,99]. Mucosas were rinsed and agitated in 1.5 ml ice-cold PBS [Thermo Fisher 10010023] until the supernatant looks clear. Mucosa from the same regions and same endoscopic diagnosis were

minced by iris scissors. Minced tissues were immersed in 2 ml warm (37°C) detachment solution (HBSS Ca/Mg-Free [Thermo Fisher 14175095], 5 mM EGTA [Fisher Scientific 50-255-956], 10 mM HEPES [Thermo Fisher 15630080], and 10% FBS [Gemini Bio 100-106]). The tissues were incubated at 37°C for 15 minutes with end-over-end rotation, followed by 10-minute cool-down on ice. The tissues were agitated (shake 20 times) to separate epithelium from lamina propria. Supernatant containing epithelial sheets, was collected and spun down at 300g for 5 minutes at 4°C. The remnant tissue was rinsed in ice-cold PBS twice. The epithelium was incubated in 2 ml TrypLE express [Thermo Fisher 12605010], at 37°C for 5 minutes with rotation. The remaining tissues were incubated in 2 ml enzyme solution (HBSS Ca/Mg-Free [Thermo Fisher 14175095], 10 mM HEPES [Thermo Fisher 15630080], 200 μg/ml Liberase TM [Millipore Sigma 5401127001] and DNase I [Gold Bio D-301-100]) at 37°C for 20 minutes with rotation. After enzymatic digestion, epithelium and lamina propria were triturated with a P1000 pipette until no tissue chunks is visible. Cells were filtered through a 40 μM cell strainer [Falcon/VWR 21008–949]. The filtrated cells from both digestions were pooled and spun down at 300 g for 5 minutes at 4°C. Cell pellets were resuspended in 200 μl cold wash buffer (HBSS Ca/Mg-Free [Thermo Fisher 14175095] and 10% FBS [Gemini Bio 100-106]). Cell suspension was then diluted by 2 ml RBC lysis buffer [Miltenyi Biotec 130-094-183] and incubated for 3 minutes at room temperature. Cells were then spun at 300g for 5 minutes at 4°C. Cell pellets were resuspended in 200 μl suspension buffer (PBS and 0.04% BSA [Miltenyi Biotec 130-091-376]) and assessed viability with Trypan Blue [Thermo Fisher 15250061] and counted. Only cell suspensions with > 90% viability were used for downstream single cell assays. The detailed processing is fully described here[100]. After cells were partitioned in the scRNA-seq assay, the remaining cells were pelleted again by spinning at 500g for 5 minutes at 4°C. After the removal of supernatant, 100 μl lysis

buffer (CG000169 Rev E, 10x Genomics) was added to each pellet, followed by gently pipetting. The suspension was incubated on ice for 4 minutes and diluted by 1 ml wash buffer (CG000169 Rev E). Nuclei were spun down at 500g for 5 minutes at 4°C. After resuspending in 1 ml wash buffer, nuclei were pelleted again. The nuclei were suspended in 20 µl 1x Nuclei buffer (10x Genomics, PN-2000153/2000207) and assessed for density with Trypan Blue. The final density of nuclei was adjusted to 1.6 million per mL with 1x Nuclei buffer.

scATAC library preparation and sequencing

Single nuclei were isolated from single cell suspensions following the manufacturer's protocol for the Chromium Single cell ATAC (v1.1). Briefly, single nuclei were transposed before loading into chip H. Transposed nuclei were partitioned into GEMS in Chromium controller. GEMs were incubated for barcoding, followed by indexed amplification. Approximately 8,000 nuclei were loaded per channel to target recovery of 5,000 nuclei. Libraries were sequenced on Illumina Nextseq or Novaseq platforms at the Genomics Core Facility at The University of Chicago.

scRNA-seq data preprocessing and analysis

scRNA-seq raw data were obtained from Zhou et al. (GSE266616) and processed using Cell Ranger (v7.0.1) with the 10x human transcriptome GRCh38-2020-A as the reference. SoupX was used with default parameters to estimate and remove cell free mRNA contamination for each sample individually[101]. We used the graph based clustering results from Cell Ranger output to estimate the contamination fraction. Downstream analysis was performed in R (v4.1.0) and Seurat[102–106] (v4.2.0). Individual cells matching the following criteria were retained for analysis: more than 500 UMIs and fewer than 20,000 UMIs; more than 200 unique genes and fewer than

6,000 unique genes; less than 50% of reads mapping to mitochondrial genes; less than 40% of reads mapping to ribosomal genes; and a 'log10GenesPerUMI' score higher than 0.7 (to exclude low-complexity cell types such as red blood cells). Doublet scores were calculated using Scrublet[107] (v0.2.3), and cells with score of 0.2 or above were considered doublets and removed. The cells exhibiting co-expression of different lineage makers (epithelial: EPCAM, T cell: CD3D, B cell: MS4A1, CD79A, IGHA1, myeloid: TYROBP, CD14, CPA3, stromal: IGFBP7, COL3A1, endothelial: PECAM1) were also excluded as doublets. Genes expressed in fewer than 3 cells per sample were filtered out. Transcript counts for each cell were normalized and log-transformed. Cell cycle scores (the difference between the G2M and S phase scores) were calculated with cell cycle genes provided in Seurat (v4.2.0) using the CellCycleScoring Function. Cell cycle scores, the percentage of mitochondrial reads, and the number of unique molecular identifiers (UMIs) were regressed out before data scaling.

We identified the 2,000 most highly variable genes (HVG) by modeling the mean-variance relationship for each gene across all cells and used the count matrix restricted to these HVGs as input for principal component analysis (PCA). We selected the top 30 principal components for subsequent clustering and Uniform Manifold Approximation and Projection (UMAP) embedding[108,109]. Cell clustering was performed using the Louvain algorithm (resolution 0.3–1.5) for modularity optimization with the kNN graph as input[110].

The union dataset was divided into three different lineages (epithelial, immune and stromal) based on clustering analysis and marker gene expression. Cells from each lineage were subset for further analysis and we repeated the process of dimensionality reduction, Louvain clustering and UMAP

visualization for each lineage as described above. We filtered out clusters without specific markers, doublet clusters, and cells likely not belonging to the specific lineage through iterative clustering and annotation procedures. Cell identities for subtypes within each lineage were assigned based on previously reported marker genes[61], with a summary provided in Fig. S2.2.

scATAC data preprocessing

scATAC-seq raw sequencing data were processed using the Cell Ranger ATAC v2.0.0, and the 10x human genome GRCh38-2020-A-2.0.0 as the reference. We used cellranger-atac count to generate scATAC fragment files from fastq files obtained from sequencing core facility. Further filtering of raw data was conducted within the ArchR (v1.0.3) framework[32]. We kept high quality cells for downstream analysis (minimum fragments per cell = 5,000, minimum TSS enrichment score = 6). Doublet scores were calculated using addDoubletScores function (nTrials = 10, dimsToUse = 2:50), and cells predicted to be doublets were removed in downstream analysis (a maximum of 10% of the cells are predicted to be doublets).

scATAC data analysis

We followed the standard ArchR workflow to process the fragment files into a 500 bp tile matrix of fragment counts and generated a gene score matrix using the default gene score model (model 42). Latent semantic indexing (LSI) was performed on the tile matrix, retaining the top 100 LSI vectors by applying the 'addIterativeLSI' function with six iterations and five different clustering resolutions (0.5, 1, 1.5, 2, and 2), and up to 150,000 variable features. The first LSI vector was excluded due to its high correlation with sequencing depth. Cell clusters were identified using the

‘addClusters’ function, and clusters displaying high gene scores for multiple markers representing different lineages (epithelial, immune, stromal) were labeled as doublets and removed, similar with scRNA-seq annotation. We then divided the cell type clusters into the three lineages and performed iterative dimensionality reduction, clustering, and filtering within each lineage. We filtered out the clusters without a specific marker, doublet clusters and cells identified as not belonging to the correct lineage during each iteration.

Within each lineage, the scRNA-seq gene expression matrix was integrated with the scATAC-seq gene score matrix using ArchR’s ‘addGeneIntegrationMatrix’ function. Corresponding cells across datasets were matched using Seurat’s mutual nearest neighbors algorithm, facilitating the transfer of labels from scRNA-seq to scATAC-seq data. The scATAC-seq labels were further refined based on the distribution of marker gene scores.

Accessible chromatin peak calling

To obtain regions of accessible chromatin (‘peaks”) in each cell type cluster, we followed the peak calling workflow introduced in ArchR. Cells within each cell type were merged to generate pseudo-bulk replicates using the addGroupCoverages function. The addReproduciblePeakSet function was then utilized to call peaks for each pseudo-bulk replicate using MACS2. Within each cell type, peak sets from all pseudo-bulk replicates were combined using the iterative overlap peak merging procedure, resulting in a merged peak set for that cell type. Finally, peak sets from all cell types were merged to create a union peak set using the same procedure. The peaks were classified into four categories based on the summit positions: promoters or promoter proximal regions (2000

bp upstream and 100 bp downstream of a TSS), intronic, exonic, distal (>2,000 bp from TSS and intergenic).

Cell-type-specific differential accessible region (DAR) analysis

To identify cell-type specific marker peaks, a single-cell insertion count matrix was created using the function addPeakMatrix in ArchR. Differentially accessible regions (DARs) for each cell type, relative to the others, were identified using Wilcoxon rank-sum tests among cells from all healthy individuals (FDR $\leq$ 0.1, logFC $\geq$ 0.5), excluding the effects of inflammation. To control for technical variation, the single cell TSS enrichment score and number of fragments per cell were considered when selecting a matched null group. The DARs identified in each cell type were defined as cell type specific peak set and provided in Supplementary File 2-4.

Pairwise DAR identification between different inflammation conditions

We conducted pseudo-bulk based differential chromatin accessibility tests following dreamlet (v0.99.6) workflow[111]. Raw counts were first aggregated across cells within each sample and cell type cluster. The pseudobulk counts within each cell cluster were normalized using the voomWithDreamWeights function to account for patient-specific random effects. A regression model was subsequently fitted for each peak and cell cluster to test differential accessibility between pairs of conditions (control, noninflamed, adjacent to inflamed, inflamed), with sex and age included as covariates. iDARs were selected based on a log-fold change greater than 0.5 and an adjusted p-value less than 0.1, when comparing inflamed samples with control samples. The cell type specific iDARs were provided in Supplementary File 2-4.

Topic modeling and pathway enrichment

We used FastTopics[38,39] to fit topic models to the scATAC peak matrix. To reduce computational burden, we subsampled the binarized union scATAC peak matrix by including a maximum of 100 cells per cell type per sample. This also balanced cell counts from different lineages and helped to avoid obtaining a large number of topics associated with epithelial cells at the expense of interpretable topics from other lineages. We fitted Possion non-negative matrix factorizations (NMF) and topic models for the subsampled peak matrix using the fit_poisson_nmf and poisson2multinom functions available in fastTopics[38,39] (v0.6-159). The factors and loadings were initialized uniformly at random and then updated through the expectation-maximization (EM) algorithm for the first 80 iterations. Subsequently, updates were made using the sequential coordinate descent (SCD) algorithm until convergence, achieved after 670 iterations.

To select the number of topics (k), we tested topic resolutions k=10 to k=50 in increments of 5 aiming to balance the desire for high resolution with the ability to interpret the results. We found only a few additional interpretable topics added between 40 to 50 and chose k=45 for subsequent analysis. In many cases additional topics were strongly correlated with previously discovered ones, providing similar sets of peaks, enriched motifs, and putative target genes. To ensure that we did not systematically miss regulatory topics in this combined analysis, we repeated this process for each individual lineage (epithelium, immune, stromal). We increased the number of cells per sample and cell type cluster to 300. Interpretable topics uncovered in individual lineages mapped well to topics uncovered using the union set of all cells. We also assessed whether any of the additional topics associated with new features (e.g. location or inflammation) obtained at higher resolution were enriched for CD heritability but found no new associations. .

To identify regions with chromatin accessibility specifically associated with one topic compared to all other topics, we conducted topic-level differential accessibility analysis using the "grade of membership" (GoM) method implemented in the de_analysis function. The log-fold change relative to the null was calculated without applying adaptive shrinkage. For each topic, we selected the top 30,000 peaks based on two-tailed p-values (p-values were below 0.05 in all cases) and then filtered for positive log-fold changes, leading to peak sets containing slightly fewer than 30,000 regions. The peak sets were defined as topic specific peaks and used in downstream analysis including sLDSC. We validated that all topic-specific peaks met the p-value cutoff of 0.05 for each topic.

For each topic, gene scores were calculated by aggregating topic loadings (z-score obtained from GoM analysis) for each accessible region/peak (weighted by bi-directional exponential decay from the gene TSS). The gene scores were subsequently normalized by the L2 norm of the weights. Gene set categories (H: hallmark gene sets, C2: curated gene sets, C5: ontology gene sets) from the Molecular Signatures Database (MSigDB) were retrieved and analyzed for each topic-specific gene score vector using the msigdbr package (v7.5.1)[112]. Gene set enrichment analysis was performed using the GSEA function in the clusterProfiler package[113].

Transcription factor motif enrichment

We followed the ArchR pipeline to perform enrichment analysis of TF binding motifs within peak sets. We annotate each peak within the union peak set with the position weight matrices (PWMs) from the CisBP[114] database using the addMotifAnnotations function. Background peaks were

computed, controlling for total accessibility and GC content, via the addBgdPeaks function, and per-cell TF ChromVAR activity scores were calculated using the addDeviationsMatrix function. To identify positive TF regulators, we aggregated cells based on cell type, tissue location, and inflammation condition. We then correlated the integrated gene score matrix (scATAC gene score integrated with scRNA data) with the TF activity score matrix across these aggregates. TFs meeting the following criteria were identified as positive regulators: correlation > 0.4, inter-cluster deviation z-score difference > 0.25, and adjusted p-value < 0.01. We further assessed the enrichment of TF motifs in specific peak sets (cell type-specific, inflammation-specific, or topic-specific) using the hypergeometric test, comparing motif overlap with peak sets against random expectations. TFs were prioritized by filtering for positive regulators (Supplementary File 2-3).

Genotyping and quality control

ATAC-seq generates appreciable genome coverage (2.8-24) and we used GLIMPSE2[115] to determine the genotypes of individuals. The b38 reference panel from the 1000 Genomes Project[116] was downloaded from the EBI 1000 Genomes FTP site and filtered to retain only SNPs, removing multiallelic records. To improve processing efficiency, genotype data was excluded from the reference VCF files. The GLIMPSE2_chunk function was then employed to define chunks for imputation and phasing. A binary reference panel was constructed using the filtered reference panel, along with the corresponding genetic map and imputation regions from the previous step. With sorted BAM files from the Cell Ranger pipeline as input, the GLIMPSE2_phase function was applied to impute genotypes within each defined region using the constructed reference panel. Finally, the imputed chunks were ligated using GLIMPSE2_ligate tool, and Eagle (v2.4.1) was subsequently used for further phasing, utilizing the same reference panel[117]. To impute individual-

level genotypes for subjects with multiple samples, the sample-specific BAM files were merged into individual-specific BAM files using the samtools merge function. Both sample-level and individual-level genotypes were generated to validate the robustness of the pipeline. The quality of individual-level BAM files was assessed by evaluating coverage depth (Fig. S2.6a), calculated using samtools depth, and effective coverage (Fig. S2.6b), as described in Li et al.[118], to ensure high imputation accuracy. Non-reference concordance in chromosome 1 was examined in Fig. S2.6c, with all pairwise comparisons of samples from the same individual achieving concordance scores exceeding 0.9, further validating the accuracy of the imputation.

caQTL calling:

We used the RASQUAL[119](v1.1) to compute caQTLs in each cell type. For each individual ID and cell type, we aggregated the scATAC data into pseudo-bulk samples. Only pseudo bulk samples containing more than 20 cells were retained, yielding 19 cell types for analysis. We extracted insertion count matrices for each pseudo bulk sample based on the union peak set across all cell types from ArchR analysis. Allele specific counts were extracted from individual-level bam files and vcf files, with the VCFs filtered to include only variants with a minor allele frequency of at least 5% in our samples, using the createASVCF script. The rasqualTools pipeline (https://github.com/kauralasoo/rasqual/tree/master/rasqualTools) was employed to prepare format-compatible count matrices and GC bias-corrected size factors. Covariates were generated using the size-adjusted peak count matrix via the rasqualMakeCovariates function. We also included age, sex, and top four ancestry principal components (PCs) derived from genotype data from our samples and the 1000 Genomes Project as additional co-variants.

For each peak in the scATAC-seq union peak set, we tested associations between chromatin accessibility and genotype of all variants within a ±10 kb window. Variant-level p-values were transformed into q-values to account for the number of variants tested per peak. To correct for multiple testing genome-wide, we performed three additional association tests using permuted genotypes and calculated an empirical false discovery rate (FDR) for lead variants by comparing the q-values of the real and permuted association results. We obtained QTL mapping results for all tested SNPs and QTLs with an FDR of 10% or below were deemed significant. To visualize chromatin accessibility across different genotypes for significant QTLs, we performed variance-stabilized normalization on the read counts using the varianceStabilizingTransformation function in DESeq2[120] and plotted the resulting data as boxplots.

Variant annotation and TF motif disruption analysis

The variants were annotated with their nearest genes using the matchGenes function implemented in bumphunter (v1.36.0) package[121,122], based on hg38 genome assembly, with a promoter distance cutoff of 2000 bp. Motif disruption analysis was conducted using the information content method implemented in motifbreakR[123] (v2.8.0). SNPs exhibiting a TFMPvalue of less than 1e-4 were classified as significant motif-disrupting variants. The motif disruption score was calculated by subtracting the motif binding score of the reference allele from that of the alternative allele.

mashR

We used mashr[70] (v0.2.79) to estimate the sharing of caQTLs across different cell types. The effect size (effect size = Pi - 0.5) and standard error (SE = abs(beta/sqrt(Chi-square))) for each SNP was

calculated obtained from RASQUAL results, In the mash model fitting, we included all SNPs that were tested across all cell types, with no 'NA' and infinite effect sizes and standard errors. The QTLs that were significantly called in at least one cell type (FDR<10%) were included in the strong test subset, and 2 million peak-SNP pairs from the remaining set were randomly selected as a control subset. During the query dataset construction, the SNP with the lowest q-value for each peak across cell types was selected. For query SNPs, some were not properly tested in certain cell types because of power issues, thus, missing effect size and standard error values (due to absent insertion or allele quantification) were set to 0 and 1e-6, respectively, similar to GTEx study[124]. The fitted mash model was used to reassess effect sizes and significance of the query SNPs. Finally, the SNPs in each cell type were filtered using a local false sign rate (LFSR) cutoff of 0.05 as significant QTLs.

caQTL effect size and motif disruption score correlation

For each transcription factor (TF) within each cell type, we first identified the caQTLs that were significant in the respective cell type and exhibited a significant motif disruption effect for the corresponding TF (p-value < 1e-4). We then computed the Spearman correlation between the motif disruption scores and the QTL effect sizes for each TF-cell-type pair. We only included TF motifs with at least five significant motif-disrupting QTLs in a given cell type. The p-values from these correlations were corrected for multiple testing using the Benjamini-Hochberg procedure. For visualization, the coefficients of non-significant correlations (adjusted p-value cutoff = 0.05) were not displayed. This entire workflow was repeated using QTLs identified as significant by the mash method.

Pairwise caQTL effect size correlation

For each pair of cell types, we selected caQTLs that were significant in either cell type evaluated by mash and retrieved the corresponding effect sizes calculated using the RASQUAL workflow. Missing effect sizes were set to 0. We then computed the Spearman correlation of QTL effect sizes between the two cell types. The correlation p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. For visualization, the coefficients of non-significant correlations were set to zero (adjusted p-value cutoff = 0.05)

Heritability enrichment

We used the 1000G European panel[116] as reference SNPs and HapMap3[125] SNPs as the regression set, excluding SNPs located in the major histocompatibility complex region. Since these references and GWAS data were available in hg19, we lifted over the union set of chromatin accessibility peaks from hg38 to hg19 using the UCSC LiftOver tool. Peak regions were then extended by 1,000 bp both upstream and downstream for SNP annotation, accounting for biases between the reference panel and scATAC-seq-based peak calling. To assess the enrichment of disease or trait heritability within a specified peak set, we used stratified linkage disequilibrium score regression (sLDSC) with the baseline-LD (v2.2) model[73,90]. As a conservative step to ensure cell-type specificity, we additionally conditioned the analysis on all cell-type-specific peaks[91]. We followed the post-processing pipeline described in Samuel et al.[91] to calculate the p-value of τ^* to evaluate the significance of heritability enrichment. τ^* represents the standardized effect size, indicating the proportionate change in per-SNP heritability associated with a one standard deviation increase in

the annotation value. The preprocessed reference files, baseline model and GWAS summary statistics for sLDSC analysis are available from Alkes group: https://console.cloud.google.com/storage/browser/broad-alkesgroup-public-requester-pays/LDSCORE.

Statistical fine-mapping

The GWAS summary statistics (hg19) for Crohn's disease provided by de Lange et al.[10] were utilized for this analysis. Since the sample genotypes and in-sample LD matrices from the meta-analysis GWAS study are not publicly available, we selected two reference panels with matched ancestries to estimate the LD matrix: the 1000G European panel[116] and the UK Biobank panel[83,84], as computed in Zhao et al[126]. Fine-mapping was conducted separately using each reference panel. The GWAS summary statistics were preprocessed and harmonized with the LD reference panel using the pipeline in mapgen (v0.5.9)[55]. The genome was partitioned into LD blocks with LDetect[127], and loci containing at least one SNP with genome-wide significance (p-value < 5e−8) were selected for fine-mapping. We conducted fine-mapping with summary statistics using the sum of single effects (SuSiE) method[128,129], implemented in susieR (v0.12.35), with a uniform prior and permitting up to ten causal variants per locus. SNPs that mismatched the reference panel were filtered. Loci overlapping the region from 25 Mb to 35 Mb on chromosome 6, corresponding to the HLA region, were identified and excluded from prioritization. We provided the complete fine-mapping results, calculated using the 1000G reference panel, in Supplementary File 2-5. Peak-overlapping SNPs with posterior inclusion probabilities (PIPs) greater than 0.2, as determined using both the 1000G and UKB panels, were prioritized and listed in Table 2.1.

Enrichment of causal signals in cCREs

The fold enrichment of causal signals in cCREs was calculated using the following formula: (PIP_cCRE/genome_length_cCRE) / (PIP_non_cCRE/genome_length_non_cCRE). Regions in the black list of hg38[32] were removed in both cCRE regions and non cCRE regions during analysis. To test the significance of the enrichment, we rounded the total PIP in each of the two groups to an integer as causal signals, and compared the number of causal signals among the total number of finemapped variants using Fisher's exact test.

Functional annotation and prioritization of fine-mapped SNPs

We annotated fine-mapped SNPs with the cell-type-resolved chromatin accessibility by directly overlapping SNPs with the union set of peaks.

We linked cCREs containing SNPs to their putative target genes using 3 different methods, co-accessibility, ABC-Max method, and nearest TSS. For co-accessibility, we identified co-activity between gene expression and peak accessibility using the addPeak2GeneLinks function, selecting peak-gene links with a correlation score greater than 0.5, as provided in Supplementary File 2-2. Most Fine-mapped SNPs overlapping with cCREs were annotated with the co-active gene. Additionally, we annotated fine-mapped SNPs using the ABC-Max method. We included the peak-gene links from biosamples that have ABC enhancers enriched for IBD fine-mapped variants, as well as the "small_intestine_fetal-Roadmap" biosample provided in Nasser et al[66,92]. We extended the ABC-Max method to annotate QTL caPeaks by evaluating the ABC scores of all ABC enhancers overlapping the caPeak, assigning the caPeak to the target gene with the highest ABC score. Finally, in cases where neither method provided clear target genes or assignments were ambiguous, we provided the gene with the closest TSS as the likely target. For the small

subset of regulatory candidate variants, we also manually assessed the gene assignment (Supplementary File 2-2).

To identify GWAS variants that are caQTLs in our dataset, we first calculated LD between variants using PLINK (v1.9)[130] based on a merged panel combining our samples and 1000G European samples. Variants exhibiting strong pairwise LD ($r^2 > 0.8$) were classified as overlapping. QTL SNPs that overlapped with fine-mapped SNPs were provided in Supplementary File 2-6.

CHAPTER 3: Multiomic analysis reveals cellular, transcriptomic and epigenetic changes in intestinal pouches of ulcerative colitis patients

Note:

The following section (Chapter 3) is reproduced with slight modifications from my first-author publication "Multiomic analysis reveals cellular, transcriptomic and epigenetic changes in intestinal pouches of ulcerative colitis patients", which is published in Nature Communications[131] with open access. Total proctocolectomy with ileal pouch-anal anastomosis is the standard surgical treatment for patients with severe ulcerative colitis, yet the cellular and molecular adaptations within the ileal pouch remain poorly understood. This chapter presents a cell-type-resolved transcriptional and epigenetic atlas of the ileal pouch, revealing unique epithelial populations and gene regulatory programs that distinguish pouch from both the ileum and colon.

Authors:

Yu Zhao, Ran Zhou, Bingqing Xie, Cambrian Y Liu, Martin Kalski, Candace M Cham, Zhiwei Jiang, Jason Koval, Christopher R Weber, David T Rubin, Mitch Sogin, Sean Crosson, Mengjie Chen, Jun Huang, Aretha Fiebig, Sushila Dalal, Eugene B Chang, Anindita Basu, Sebastian Pott

† These authors contributed equally to this work.

Author Contributions:

E.B.C., A.B. and S.P. conceived the study and supervised this project. Y.Z. performed data analysis. R.Z. performed experiments and generated data. C.M.C., J.K. contributed to sample collection and experiments. B.X., Z.J. and M.C. contributed to data analysis. C.R.W. performed histological assessment of tissue sections. C.Y.L. and M.K. performed IHC and microscopy.

D.T.R., M.S., S.C., J.H., A.F., S.D. contributed to study design and data interpretation. S.P. and Y.Z. wrote the manuscript with input from all authors.

Abstract

Total proctocolectomy with ileal pouch anal anastomosis is the standard of care for patients with severe ulcerative colitis. We generated a cell-type-resolved transcriptional and epigenetic atlas of ileal pouches using scRNA-seq and scATAC-seq data from paired biopsy samples of the ileal pouch and the ileal segment above the pouch (pre-pouch) from patients (male=4, female=2), and paired biopsies of the terminal ileum and ascending colon from healthy individuals (male=3, female=3) serving as reference. Our study finds previously uncharacterized populations of absorptive and secretory epithelial cells within the pouch but not the pre-pouch. These pouch-specific enterocytes express a subset of colon-specific genes, including *CEACAM5* and *CD24*. However, compared to normal colonocytes, expression of these genes is lower, and these enterocytes also express inflammatory and secretory genes while maintaining expression of some ileal-specific genes. This cell-type-resolved transcriptomic and epigenetic atlas of the ileal pouch establishes a reference for investigating pouch physiology and pathology.

Introduction

Patients with medically refractory Ulcerative Colitis (UC) often undergo ileal pouch anal anastomosis (IPAA), which removes the colon in its entirety and thus the organ affected by UC[132]. An ileal pouch (pouch) established from the terminal segment of the ileum restores continence and improves quality of life[132]. Establishment of the pouch in UC patients is frequently associated with histological changes of the mucosa involving both the epithelium and immune cells[133,134], including various degrees of inflammation (neutrophil infiltration) and villus atrophy[134,135]. Furthermore, up to 50% of patients will develop severe inflammation of the pouch (pouchitis) within 1-2 years following surgery[14,15]. In most cases, acute pouchitis can be managed with

antibiotic treatments and biologics; however, about 60% of these patients subsequently develop chronic pouchitis[13]. The etiology for pouchitis remains unclear, hampering efforts to develop specific treatment strategies[16]. Notably, histological changes of the mucosa alone are not predictive for developing pouchitis[135,136].

Transcriptional characterization of newly established ileal pouches for UC patients using bulk RNA-seq revealed a colon-like expression profile in pouch compared to pre-pouch samples[137,138]. However, bulk RNA-seq is unable to provide insights into the cellular composition of the pouch epithelium or detect cell-type-specific gene expression differences between pouch and pre-pouch, and the only single cell RNA-seq study of pouch tissue was limited in scope, focusing only on immune cells (CD45+)[139].

Therefore, it remains unclear whether the ileum-derived mucosa of the pouch undergoes changes in cell type composition and how gene expression changes in specific cell types after pouch functionalization. Addressing these questions is an important first step to understanding the cellular processes that shape this unique tissue and a necessary basis for studies aiming to identify molecular characteristics predisposing patients to pouchitis.

Here we use single cell RNA-seq (scRNA-seq) and single cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) to generate a cell-type-resolved atlas of the UC pouch mucosa and to obtain transcriptional and epigenetic profiles of component cell types. We find and characterize absorptive and secretory cell populations unique to UC pouch, but not pre-pouch. These pouch-specific enterocytes are in part characterized by the expression of a set of colon-

specific genes. However, compared to normal colonocytes, expression of these genes is lower, and these enterocytes also express inflammatory and secretory genes while maintaining expression of some ileal-specific genes. By leveraging published longitudinal bulk RNA-seq data, we find evidence that these cells are present in most UC pouches regardless of the development of pouchitis. Using scATAC-seq we identify regulatory features and transcription factors associated with pouch-specific enterocytes. Together, this cell-type-resolved transcriptomic and epigenetic atlas of the ileal pouch provides a resource for investigating pouch physiology and pathology.

Results

Single-cell RNA-seq profiling reveals cellular composition of ileal pouch tissue

We aimed to develop a cellular atlas for ileal pouches in UC patients post-IPAA surgery, focusing on delineating their cell types and comparing these with pre-pouch samples from the same patients and healthy tissue samples from control subjects. To this end, we performed scRNA-seq (10x Genomics) on paired biopsy samples from the pouch and pre-pouch from 6 UC-IPAA patients with long established pouches (mean=17 yrs post-surgery, range=9-31 yrs). Using paired samples from the same patients allowed us to use the pre-pouch regions as 'internal control' for each patient with the pre-pouch representing the terminal ileum and therefore the tissue from which the pouch was derived. This design aimed to reduce the impact of external and inter-individual variation. While histopathological evaluation indicated the presence of enteritis in the pouch samples, all individuals presented without clinical symptoms of overt pouchitis, this observation agrees with a previous study assessing UC pouch mucosa[135]. Bulk RNA-seq detected increased expression of colon-specific marker genes in pouch compared to pre-pouch samples[137], and we therefore included biopsies from the TI and AC of 6 healthy individuals undergoing routine screening as

reference for healthy tissues and to assess the nature of cellular changes in the pouch (Fig. 3.1a, Fig. S3.1a, Supplementary File 3-1). After quality-control filtering, we obtained 103,772 cells from all scRNA datasets combined (Fig. 3.1b).

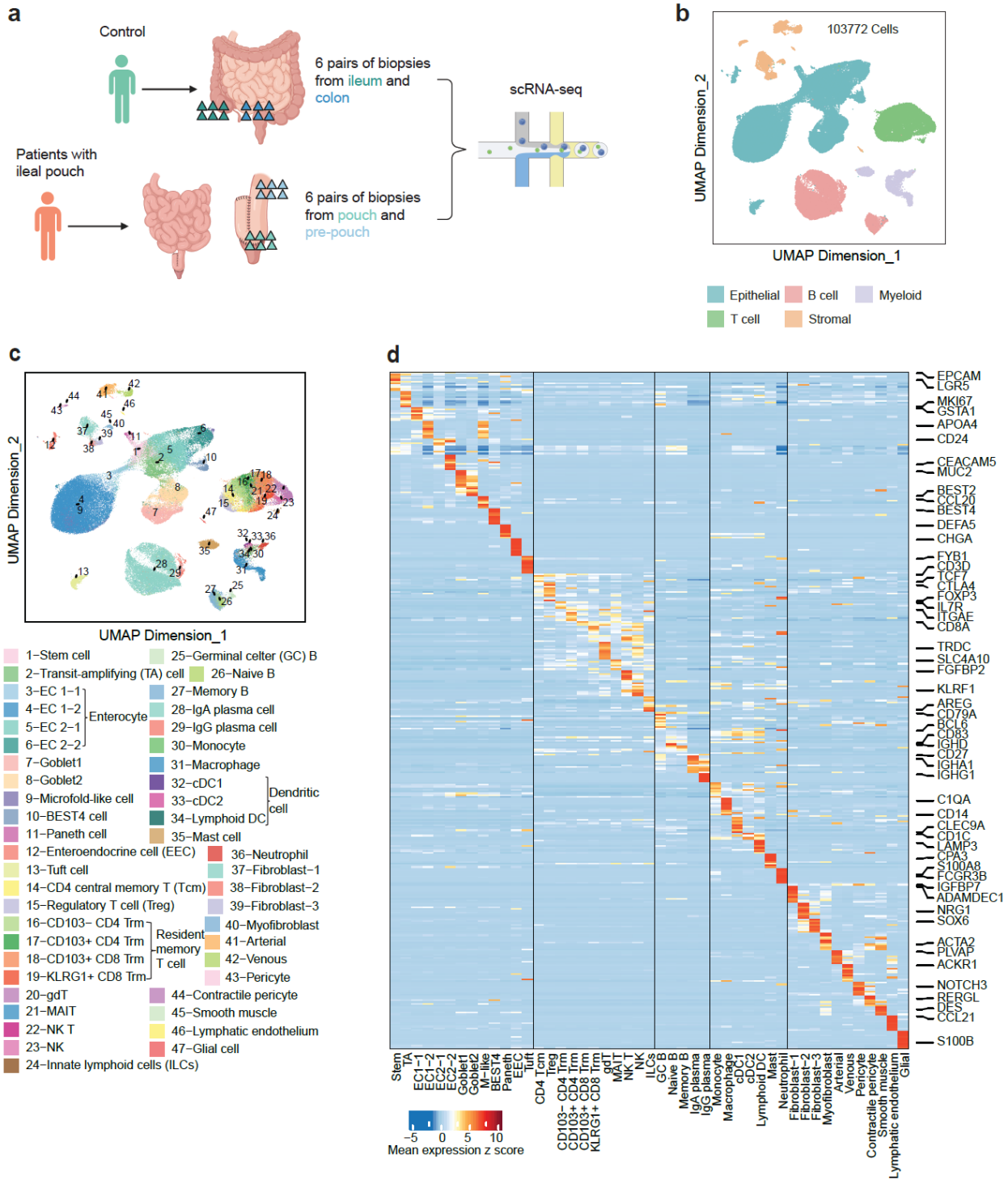


Figure 3.1 Overview for the scRNA-seq profiling of 24 human intestinal biopsies from 4 different anatomical regions.

Figure 3.1, continued

a Schematic of biopsy sampling strategy for scRNA-seq. We profiled 6 pairs of terminal ileum (TI) and ascending colon (AC) biopsies from 6 healthy individuals, and 6 pairs of pre-pouch (PP) and pouch (POU) biopsies from 6 different healthy IPAA patients. Parts of the image were created in BioRender. ZHAO, Y. (2024) https://BioRender.com/b75p140. **b** UMAP dimensionality reduction of 5 major cell lineages in all 24 samples. **c** High resolution UMAP showing 47 cell types in 24 biopsies. **d** Heatmap of top 10 differentially expressed genes and representative cell marker for each cell type within a lineage. Mean expression in each cell type was scaled as z score in each row across all cell types.

Using community-based clustering and marker-gene analysis we identified 5 major clusters comprising cells of the following lineages: epithelial (59,751 cells, marked by *EPCAM*), T-cells (16,389 cells, *CD3D*), B-cells (17,317 cells, *CD79*), myeloid cells (5,375 cells, *TYROBP*), and stromal cells (4,940 cells, *IGFBP7*) (Fig. S3.1c, d). Each individual and sample contributed to all cell lineages (Fig. S3.1b).

We further separated each of the major lineages into multiple cell-types. By combining computational annotation that leveraged previously published data with manual annotation employing well-defined marker genes, we annotated 47 cell types across all lineages (Fig. 3.1c, d, Fig. 3.2-5, see Methods for detailed description of classification).

We identified all major cell types previously reported for the absorptive and secretory epithelial lineages (Fig. 3.1c, d, Fig. S3.6, 7)[60], including stem cells (*LGR5*, *ASCL2*), transit amplifying cells (TA cells) (*MKI67*), absorptive epithelial cells and secretory cells, such as goblet cells (*MUC2*), and Paneth cells (*DEFA5*). Several absorptive and secretory cell populations from the TI and AC clustered separately based on their sampling location. To distinguish these ileal and colonic cell populations we denoted the Enterocyte cell clusters as EC1 (ileal) and EC2 (colon), and similarly identified Goblet1 and Stem1 (ileum) and Goblet2 and Stem2 (colon) clusters. Furthermore, we

distinguished progenitor and mature cells within the EC1 and EC2 cell clusters and denoted them as EC1-1 (*SI*, *GSTA2*) and EC2-1 (*CD24*, *CA2*) and EC1-2 (*SLC15A1*, *APOA4*) and EC2-2 (*CEACAM5*), respectively (Fig. 3.1d).

All samples contained a range of adaptive and innate immune cells representing T-cell, B-cell, and myeloid lineages (Fig. 3.1c, d, Fig. S3.8-10). T cells were further clustered into CD4, CD8, and gamma-delta T cells. CD4 T cells included central memory (*CCR7*), regulatory (*FOXP3*), CD103- and CD103+ resident memory (CD69+, Fig. S3.3) T cells. CD8 T cells include CD103- and CD103+ resident memory T cells. B cells readily separated into germinal center B cells (*BCL6*), naive B cells (*IGHD*), plasma B cells (*SDC1*) and memory/effector B cells (*CD27*) (Fig. 3.1c, d). Innate immune cells included macrophages (*CD163*), monocytes (*FCN1*), dendritic cells (*CLEC9A*, *CD1C*), mast cells (*CPA3*), and neutrophils (*FCGR3B*) (Fig. 3.1c, d). Stromal and other cell types constituted 4.8% (4,940 out of 103,772) of cells. We distinguished several fibroblast populations and identified small clusters of the following cell types: endothelial cells from arteries (*EFNB2*) and veins (*ACKR1*), pericytes (*NOTCH3*), smooth muscle cells (*DES*), lymphatic endothelial cells (*LYVE1*), and glial cells (*S100B*) (Fig. S3.11).

These data constitute the first cellular atlas of UC pouch and pre-pouch tissues, enabling us to compare the cellular and transcriptional changes between pouch and pre-pouch regions and identify similarities and differences to TI and AC reference samples from individuals without IBD history.

Profound differences in the cellular composition of the epithelial lineages between pouch and pre-pouch

To systematically assess the cellular identities and composition of UC pouches, we used data from all 4 regions (pouch, pre-pouch, AC, TI). We noticed stark differences in cellular composition of the epithelial lineage between pouch and pre-pouch samples, with pouch samples featuring two distinct clusters each of absorptive and secretory cells (Fig. 3.2). After integration of all samples using Harmony[140] (Methods), a sizeable proportion of epithelial cells from the pouch clustered with cells from colon samples suggesting some transcriptional similarity. Pouch samples featured two different cell populations for enterocyte, goblet, and stem cells, corresponding to EC1 and EC2, Goblet1 and Goblet2, and Stem1 and Stem2, respectively. (Fig. 3.2b, c, Fig. S3.12). We observed this compositional change in all 6 individuals, suggesting that partial colonic metaplasia might be a common feature of UC pouches. In contrast, the proportion of these colon-like cell populations was low in pre-pouch samples (Fig. 3.2c, Fig. S3.12) and their overall cellular composition was similar to TI.

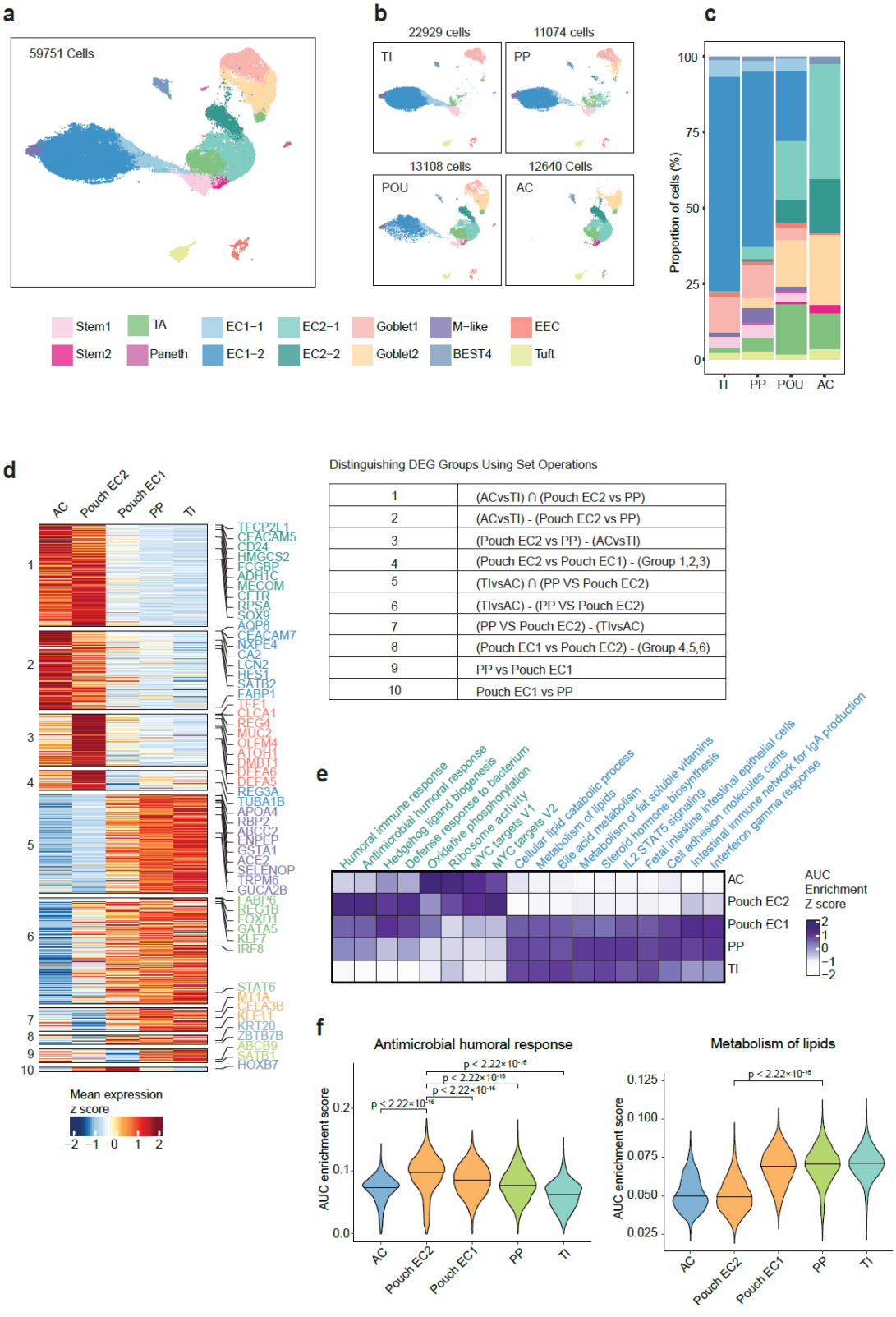


Figure 3.2 Identification of both ileum-like enterocytes and colon-like enterocytes in pouch.

Figure 3.2, continued

a Harmony corrected UMAP of 59,751 cells in epithelial lineage. There are two different populations of enterocytes: ileum-like EC1 and colon-like EC2. EC1-2 and EC2-2 are more mature than EC1-1 and EC2-1 respectively. **b** Harmony corrected UMAP of epithelial cells split based on the four different anatomical regions. **c** Proportions of different cell types in each region. **d** Enterocytes are split into 5 groups: AC EC2, TI EC1, pre-pouch EC1, pouch EC1, and pouch EC2. The DEGs were further grouped into 10 different clusters by the set operations of different contrasts, as summarized in the table. Union of sets (∪) Intersection of sets (∩) Difference of sets (–). **e** Heatmap showing pathway enrichment score in 5 enterocytes groups. The score is scaled across enterocyte groups in each column. **f** AUC enrichment score for antimicrobial humoral response and lipid metabolism related pathway activity. The analysis was conducted between single cells from two cell populations, with each single cell considered a biological replicate. The number of cells in each group is as follows: AC (7,072), pouch EC2 (3,523), pouch EC1 (3,581), PP (7,327), TI (17,633). Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. Notably, all significance tests labeled in the two panels yielded p-values below 2.22×10^{-16}, corresponding to the limit of floating-point precision in R. Source data are provided as a Source Data file.

In addition, we observed smaller proportional shifts in other epithelial cell types. The proportion of TA cells across lineages was higher in pouch compared to pre-pouch samples (14% vs 4.8%, p-value=0.0087), and this difference was also observed in AC compared to TI (12% vs 1.9%, p-value=0.0022). The proportion of TA cells in pre-pouch and TI samples differed slightly (4.8% vs 1.9%, p-value=0.041) (Fig. S3.6). The proportion of M-like cells was significantly higher in pouch and pre-pouch compared to AC and elevated compared to TI (Fig. S3.6). We detected Paneth cells in TI, pre-pouch, and pouch but not in AC (Fig. S3.6).

Pouch-specific changes in non-epithelial cell types

Given the profound cellular changes in the pouch epithelium, we also tested whether similar changes were observed in the immune and stromal compartments. We noted that the changes in the immune compartment were less uniform across patients compared to the epithelial lineage. For example, while we observed an increase in the proportion of IgG plasma B-cells in 3 patients, a change previously observed in UC patients[141], the proportion of these cells in the remaining patients was indistinguishable from pre-pouch and adjacent tissues. Samples from pouch and pre-

pouch contained a slightly higher proportion of CD103+ CD4 Trm cells compared to AC (13%, 11% vs 4.3%, p-value=0.015, 0.0043, respectively), while the proportion of KLRG1+ CD8 Trm was higher in AC, pouch, and pre-pouch compared to TI. Of note, NK T cells were specifically increased in pouch samples compared to pre-pouch samples (1.8%, 0.4%, p-value=0.0087) as were regulatory T cells (13%, 5.1%, p-value=0.0043) (Fig. S3.8). These differences suggest compositional changes of immune cells in pouch samples, possibly reflecting the tissue origin (CD103+ CD4 Trm) and disease status (IgG plasma cells, NK T cells).

Among stromal cell types, we found a significantly higher proportion of the fibroblast-2 (NRG1+) population in the pouch (3.4%) compared to pre-pouch. This population is almost absent in pre-pouch and TI samples (0.2% and 0.3%, respectively) and highest in AC (13%). We complemented these analyses by obtaining spatial transcriptomics data from pouch, pre-pouch, and control samples (Methods). We identified cells expressing NRG1 located underneath the epithelial layer in pouch and AC samples (Supplementary File 3-2). In comparison, fewer NRG1 expressing cells were observed in pre-pouch and TI in this location (Supplementary File 3-2). These observations thus indicated a remodeling of the stromal niche in UC pouch.

In comparison to the observed changes in the pouch epithelium, the differences in the non-epithelial clusters are smaller and more variable across samples. Increasing the number of cells and samples would improve power to detect differences in the immune and stromal lineages. In the remainder of this study, we focused our analysis on the characterization of the partial metaplasia phenotype observed in epithelial cell types in the pouch.

Differential expression in pouch enterocytes suggested partial colonic metaplasia

Our clustering analysis suggested the presence of a colon-like enterocyte population in UC pouches but not in pre-pouch regions. However, global similarities between EC2 cells in the pouch and AC based on clustering and embedding following computational integration with Harmony do not directly imply that cells from these regions are identical, and expression of specific genes might differ substantially. We thus compared and characterized the transcriptional programs of enterocytes in the pouch (EC1 and EC2) and compared these differences to expression differences between similar enterocyte populations from pre-pouch and reference samples. We complemented this analysis with a direct comparison of EC2 population in pouch and AC.

First, we detected a substantial number of DEGs between pouch EC2 and EC1 (1,240 genes up in EC2 and 952 genes up in EC1, false discovery rate (FDR) < 0.05, $\log_2$ fold-change (logFC) > 1) and between AC (EC2) and TI (EC1) from healthy individuals (1,883 genes up in AC, 2,141 genes up in TI). Notably, a substantial proportion of these DEGs were shared across these two comparisons (836 genes upregulated in pouch EC2 vs pouch EC1 and AC vs TI, and 770 genes upregulated between pouch EC1 vs pouch EC2 and TI vs AC). This analysis suggested that EC1 and EC2 in the pouch recapitulate at least a portion of the expression differences observed between enterocytes from colon and ileum. We also noted lower expression of some typical markers of colonic epithelial cells in pouch EC2 cells (e.g., CA2) (Supplementary File 3-3). We further confirmed these common expression differences between EC1 and EC2 cell types through pairwise comparisons across all other conditions and samples (Fig. 3.2d, Fig. S3.12-14). Note that the limited numbers of Goblet1 and 2 and Stem1 and 2 cells in the pouch did not allow us to compare these populations directly within the pouch samples. Likewise, assessment of DEGs in EC2 cells in the pre-pouch was not possible due to the exceedingly small number of cells detected

per individual. Comparison of aggregate expression profiles of pre-pouch enterocytes suggested similarity with EC1 populations in TI and pouch (Fig. S3.13, 14).

Second, we asked whether the sets of DEGs between populations of EC1 and EC2 cells were enriched for specific cellular functions. DEGs associated with EC2 in the pouch were enriched for immune and antimicrobial pathways and shared enrichment for proliferation associated terms and oxidative phosphorylation with enterocytes from AC (Fig. 3.2e, Supplementary File 3-4). We found that DEGs associated with pouch EC1 populations were strongly enriched in lipid and bile acid metabolism terms (Fig. 3.2e, f). This comparison suggested that compared to pouch EC1, pouch EC2 cells appear to lack some of the metabolic functions associated with the ileal epithelium while acquiring some functions also enriched in AC enterocytes.

Third, we directly compared pouch to AC EC2 cells and identified 1,442 differentially expressed genes between EC2 cells from these locations (770 up in EC2 pouch, 672 up in AC EC2, FDR < 0.05, logFC > 1) (Fig. S3.12d, Supplementary File 3-5). Among the most strongly up-regulated genes in the pouch were genes with known ileal expression and function, including *FABP6*, *APOA4* and *APOB*. In contrast, *AQP8*, *HMGCS2*, *SELENBP1*, *SATB2* and *CA2* showed increased expression in AC EC2 compared to pouch EC2. In addition, among genes upregulated in pouch EC2 compared to AC EC2, we noticed several genes previously reported as part of an expression signature in UC/IBD epithelium, including *REG1A*[142], *DMBT1*[143,144], *LYZ*[145], and *REG4*[146]. This finding indicates that while the global transcriptional identity of pouch EC2 cells resembles that of colonic EC2 cells, they maintain substantial differences in their expression profiles which likely

reflect the different tissues of origin, differing microbial and metabolic environments, or disease status.

Among DEGs in EC2 between pouch and AC, we noted higher expression of several IBD-related genes in the pouch. Motivated by this observation, we compared the set of DEGs between AC EC2 and Pouch EC2 to a previous scRNA-seq study of UC samples[27]. We found concordant overlap between a subset of pouch-specific EC2 genes and UC-associated genes from Smillie *et al.*[27] (81 out of 770 pouch EC2 genes) (Methods). Conversely, AC-specific EC2 genes were enriched for genes down-regulated in colonic epithelium of UC patients in that study (29 out of 672 AC EC2 genes) (Fig. 3.3a, Supplementary File 3-5). This comparison suggested a UC-like inflammatory signature in pouch EC2 cells, highlighting a potentially persistent inflammatory environment. While these markers were differentially expressed between AC EC2 and Pouch EC2 cells from all individuals, it is worthwhile noting that expression of these markers was variable between individuals (Fig. 3.3a). These expression differences between Pouch and AC samples were not affected by differences in composition of control and UC populations (Fig. S3.15, 16). We also identified similar inflammation-related changes between pouch and AC Goblet2 cells (Fig. 3.3b, Supplementary File 3-5). Of note, a number of DEGs, including LCN2, REG1, DMBT1 were recently associated with distinct epithelial cell populations in active Crohn's disease[26].

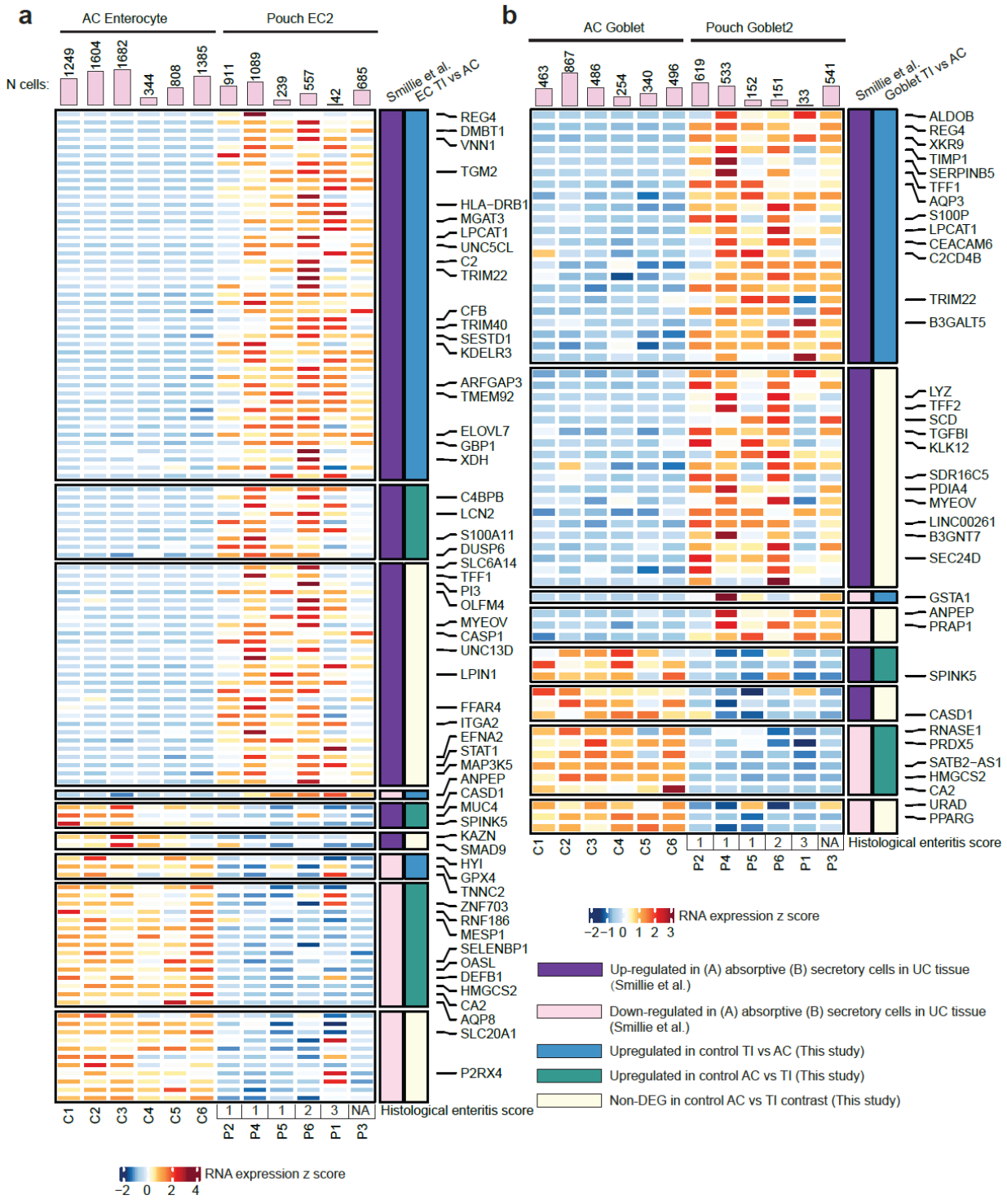


Figure 3.3 Heatmap showing key differentially expressed genes between cell populations.

Figure 3.3, continued

a pouch EC2 vs AC EC2. **b** pouch Goblet2 and AC Goblet. Selected genes were reported in a previous study (Smillie et al.)[27]. Pink bars on top are the number of cells in each sample of the corresponding cell type. Expression is scaled in each row across all samples. The last row shows the histological enteritis score of the sample, as also shown in Fig. 3.4b.

Together, these analyses established the presence of distinct absorptive and secretory enterocyte populations in UC pouches compared to pre-pouches and suggested that the pouch epithelium undergoes partial metaplasia. These comparisons also highlight expression of inflammation-associated genes which were previously identified as part of epithelial IBD signatures suggesting that ongoing inflammation in UC pouches without overt pouchitis affects epithelial programs.

Histopathology confirmed presence of colon-like cells in pouch epithelium

We sought to validate the metaplastic phenotype and characterize the spatial distribution of EC1 and EC2 cells in the pouch epithelium. We obtained tissue sections from pouch and pre-pouch samples from 5 out of 6 individuals profiled with scRNA-seq, for histopathological assessment and immunohistochemistry (IHC) (Fig. 3.4a, Supplementary File 3-6). In addition to mild to severe enteritis (Fig. 3.4b), we observed differences of the epithelial architecture in pouch biopsies, including villus atrophy, in agreement with previous reports[134].

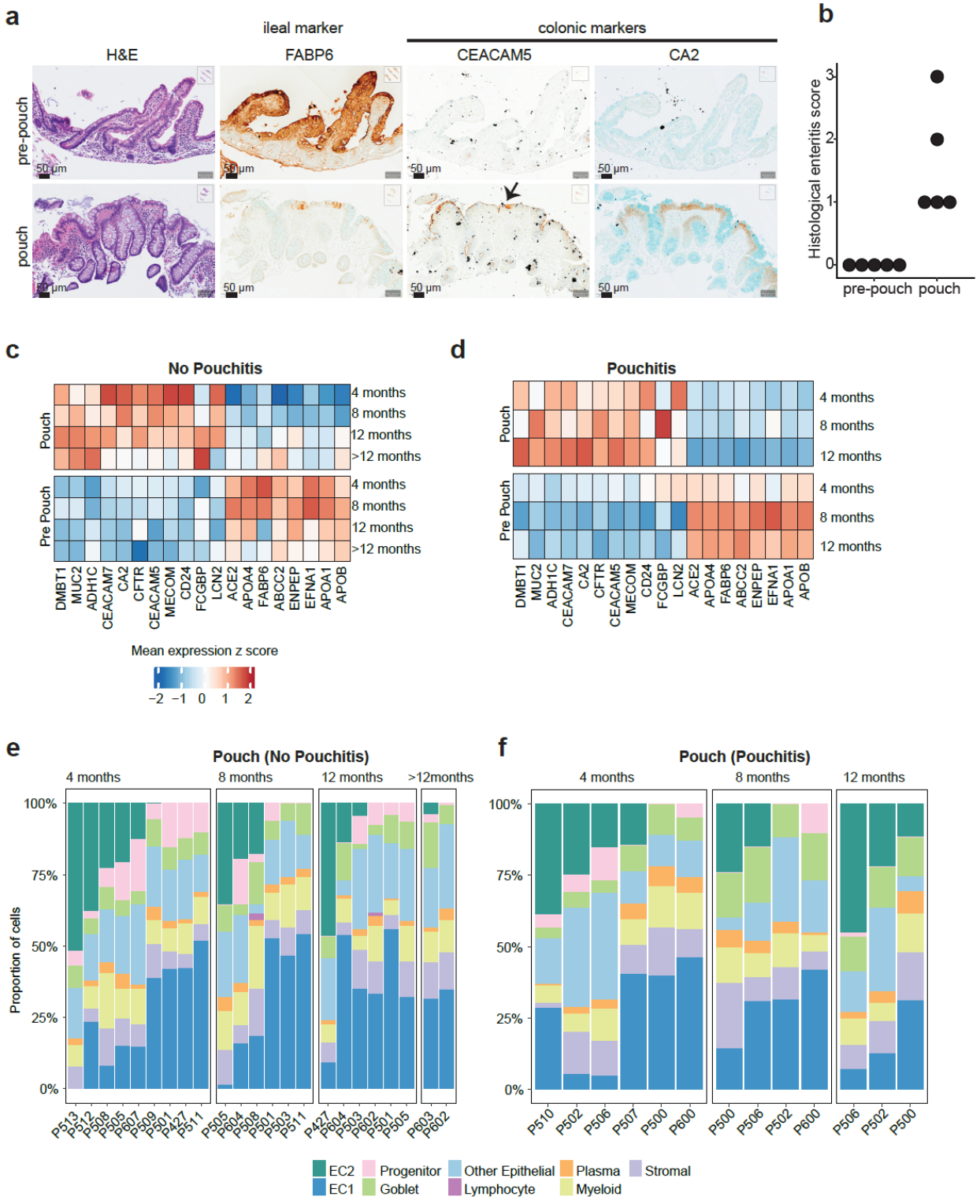


Figure 3.4 Validation of a metaplastic enterocyte population in UC pouches.

Figure 3.4, continued

a Representative immunohistochemistry images from pre-pouch and pouch biopsies for H&E, FABP6, CEACAM5, CA2 staining. We found evidence of FABP6 or APOA4 expression, small-intestinal markers, in 5/5 pouch samples. 4/5 pouch samples showed expression of CEACAM5 (colonic marker), and 1/5 pouch samples showed expression of CA2 (colonic marker). Experiments to titrate antibodies and validate specificity of labeling were performed on ileal or ascending colonic biopsies collected from 2 patients (Supplementary File 3-6). **b** Histological enteritis score of pre-pouch and pouch samples. **c, d** Average gene expression in the bulk RNA-seq data from healthy pouches (**c**) patients and in pouches from patients who developed pouchitis (**d**). Expression for each gene is scaled across pre-pouch and pouch. **e, f** Cell type proportions predicted from bulk RNA-seq data in healthy pouches (**e**) and in pouches from patients who developed pouchitis (**f**). Source data are provided as a Source Data file.

Assessment of sections from pre-pouch biopsies showed staining for the ileal enterocyte marker FABP6 and complete absence of colonic markers CEACAM5 and CA2 (Fig. 3.4a, Supplementary File 3-6). In contrast, expression of FABP6 was greatly diminished in pouch samples and confined to a small subset of cells. In addition, we observed expression of CEACAM5 and CA2 in a subset of epithelial cells from pouch samples (Fig. 3.4a). CEACAM5 was consistently detected in pouch samples on RNA and protein level, CA2 was only detected in some individuals. These observations validated our scRNA-seq findings by confirming the expression of colonic markers in a subset of epithelial cells, while showing diminished expression of ileal marker genes.

In addition, we performed spatial transcriptomics on sections from two pairs of pre-pouch and pouch samples and controls using the Xenium platform from 10x Genomics targeting 422 genes (Methods). To assess differences between pouch and pre-pouch samples, we delineated spatial domains through the integration of spatial transcriptomic data with our single cell reference dataset using IRIS[147] (Methods). This domain-level analysis identified two domains enriched for mature enterocytes and revealed an upregulation of EC2-associated genes in pouch samples compared to pre pouch samples within these epithelial domains (Supplementary File 3-7). These data also detected strong expression of inflammation-associated markers (e.g. *LCN2, DMBT1*), within the

pouch. At the level of individual cells we detected expression of EC1 and EC2 markers in epithelial cells and in agreement with the domain analysis, several EC2 markers, including CEACAM5 and CD24 appeared more highly expressed in pouch samples (Supplementary File 3-7). We detected the expression of ileal marker genes was detected across pouch epithelial cells (e.g. APOA4, FABP6) albeit at lower levels than in pre-pouch, which confirms the scRNA-seq findings.

Pouch metaplasia occurs early after functionalization and independently of subsequent inflammation

We observed colon-like EC2 cells in all our pouch samples; however, our cross-sectional study was limited to six patients with well-established pouches and without clinical symptoms. To generalize our observation, we took advantage of a previously generated bulk RNA-seq dataset[137] which reported a colon-like expression profile in pouch samples. These data allowed us to validate the expression of EC2 markers in pouch samples, to address when EC2 cells appear after pouch functionalization, and to ask whether their presence is associated with to the development of pouchitis. This longitudinal study included 19 patients with newly established pouches, sampled at three time-points within the first year after UC IPAA surgery (4-mo, 8-mo, 12-mo post-functionalization); 6 out of the 19 individuals were subsequently clinically diagnosed with pouchitis[137].

We first assessed the expression of EC1- and EC2-specific markers obtained with our scRNA-seq data in bulk samples from pre-pouch and pouch regions. EC2 markers showed increased expression in bulk samples from pouches, while EC1 markers were increased in pre-pouch samples (Fig. 3.4c, d, Supplementary File 3-6). We then used the cell types identified in UC pouch

individuals (from pouch and pre-pouch samples) with our scRNA-seq data as reference to deconvolve the bulk RNA-seq data from Huang *et al.* using MuSiC[148] (Methods). This approach allowed us to infer the constituent cell types and estimate their respective proportions in each bulk sample. We found that the presence of EC2 cells was predicted for the majority of pouch samples (58%, 21/36) (Fig. 3.4e, f) and in only 3 out of 35 (9%) pre-pouch samples (Supplementary File 3-6). We detected EC2 cells (>1%) in 14 out of 19 individuals in at least one time point. The fraction of EC2 cells in pouch samples (4.0% - 51.9%) was generally inversely correlated to the fraction of EC1 cells (Spearman correlation coefficient = -0.82, p-value = 7.4e-10) in the same sample.

EC2 cells were detected in pouch samples at the earliest time point (4-mo) and there was no significant difference in the proportion of samples or the proportion of cells per sample across timepoints (Fig. 3.4e, f). Similarly, the EC2 cell proportional difference between samples from healthy individuals and those diagnosed with pouchitis is not significant (Mann-Whitney test, p-value = 0.3775) (Fig. 3.4e, f), suggesting that the presence of EC2 cells alone is not causing pouchitis. Together these data suggest the presence of colon-like EC2 enterocytes in most pouches, providing a cellular basis for the previously reported colon-like expression within pouch samples.

Integrative scATAC-seq profiling in ileal pouch and neighboring tissues

To identify key transcription factors, gene regulatory elements, and target genes associated with pouch metaplasia, we measured chromatin accessibility using scATAC-seq[149]. To minimize biological and technical variation, scATAC-seq was performed on cells from the same 24 biopsies used for scRNA-seq (Methods).

After quality-control filtering, we retained 88,595 cells, (sample mean = 3,691, range: 1,704-5,524). Using dimensionality reduction and clustering following previously established procedures, including correcting effects from individuals[32], we identified cells from all major lineages (Epithelium, T cells, B cells, Myeloid, and Stromal/others) (Fig. S3.17-21). We assigned cell-type labels to each cell by computationally transferring cluster labels from scRNA-seq and observed that with few exceptions, specific cell-type labels mapped to discrete locations in the embedding, suggesting good correspondence between the modalities (Fig. 3.5a, Methods). To assess the accuracy of our cell-type annotations, we used gene activity scores as a proxy for gene expression[32] (Methods) and compared the scores for marker genes identified by scRNA-seq across clusters. This comparison further supported our assignment (Fig. S3.22a). Finally, we confirmed that cell-type proportions identified separately by each modality (scRNA-seq and scATAC-seq) were similar for each sample (Fig. S3.22b).

Focusing on cells from the epithelial lineage alone, we retained 59,987 cells across all samples, including 12,474 cells from pouch samples (Fig. 3.5b-d). Similar to scRNA-seq, we detected absorptive cells (EC1-2, EC2-2, BEST4), secretory cells (Goblet1, Goblet2, Paneth, EEC, Tuft) and progenitor cells (Stem, EC1-1, EC2-1). We excluded TA cells and M-cells from annotation since these were more ambiguous, likely due to differences between gene scores and RNA expression for informative markers of these cell types (Methods). Loci harboring marker genes for the annotated cell types showed highly specific chromatin accessibility (Fig. 3.5e).

We confirmed the presence of both ileum-like (EC1) and colon-like enterocyte (EC2) populations in the pouch by comparing clusters from different sampling locations (Fig. 3.5c). As observed with scRNA-seq, the proportion of EC2 cells was significantly increased in pouch compared to pre-pouch (42.47% vs 11.6%) (Fig. 3.5d, Fig. S3.17b). These data confirmed the presence of EC2 cells in the pouch using an independent, chromatin-based assay and thus providing the basis to identify and characterize regulatory regions and transcription factors associated with colonic metaplasia in UC pouches.

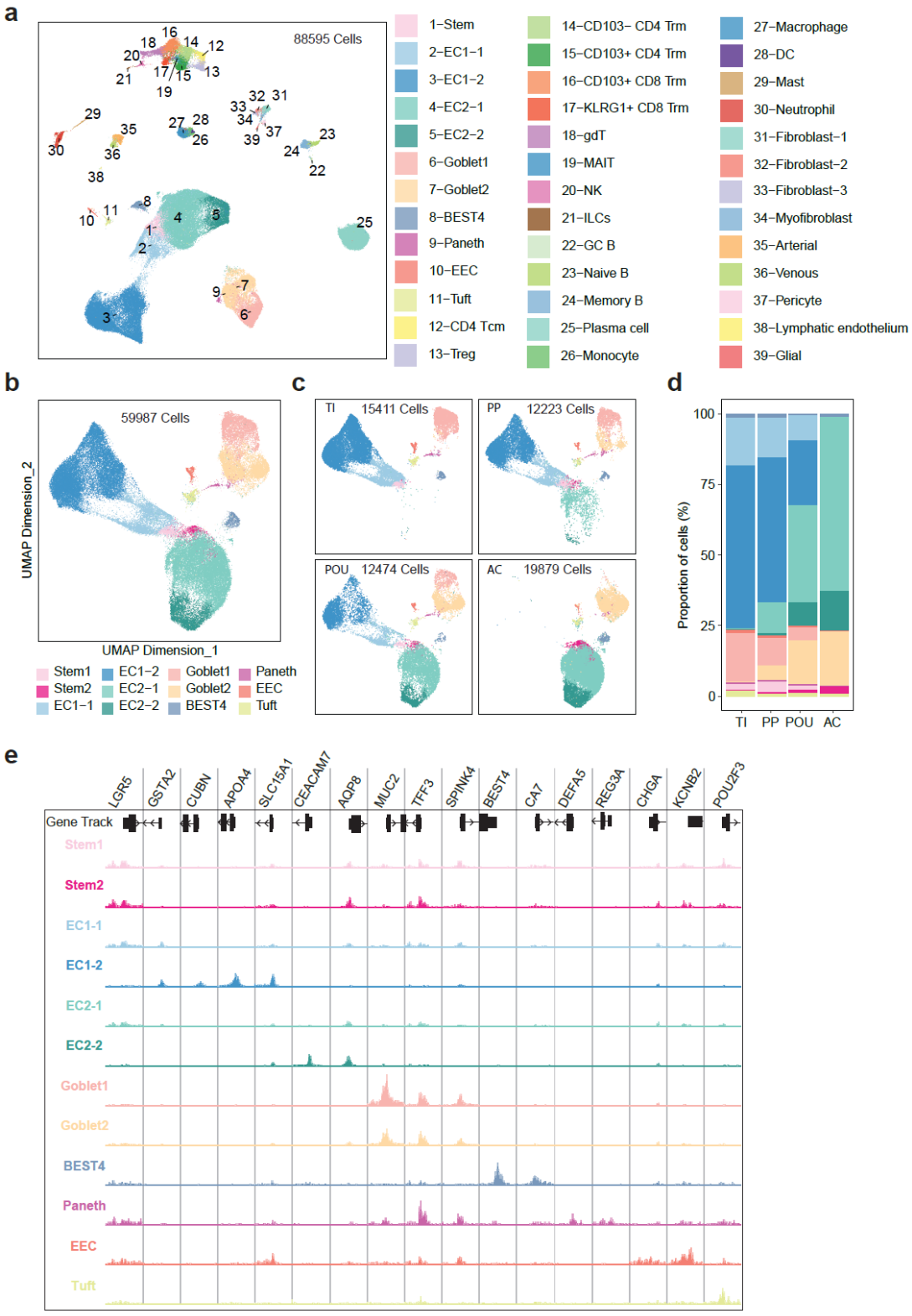


Figure 3.5 scATAC-seq profiling of the same group of 24 biopsies studied in scRNA-seq.

Figure 3.5, continued

a Harmony corrected UMAP and clustering based on single-cell chromatin accessibility identifies 39 distinct cell types from 88,595 cells across all lineages. **b** Harmony corrected UMAP of the epithelial lineages includes all cell types identified in scRNA-seq analysis. **c** Epithelial UMAP split based on four sampling locations. **d** Proportions of cell types in each region. **e** Normalized accessibility in promoter regions of cell- type-specific marker genes. Source data are provided as a Source Data file.

Identification of regulatory regions and sequence motifs associated with colon-like cells in the pouch

To identify cis-regulatory elements (CREs) in each epithelial cell type, we identified significantly accessible regions (peaks) in each cell-type cluster separately (Methods). We identified a union set of 262,177 peaks across all clusters with the number of peaks per cluster ranging from 37,127 to 150,000 (Fig. S3.17d). While most peaks were shared across multiple cell types, we observed significant variation in chromatin accessibility across cell types (Fig. 3.6a). Enrichment analysis within cell-type-specific peaks identified motifs of known master transcription factors (TFs) for these cell types and lineages (Fig. 3.6b). For example, motifs for HNF4A/G were strongly enriched in ileal-like EC1 cells, while motifs for CDX1/2 were most strongly enriched in colon-like EC2 and progenitor cells (Fig. 3.6b). This analysis established the union set of candidate CREs in the epithelial lineage and their use across cell types and helped infer sequence motifs associated with cell-type-specific TFs.

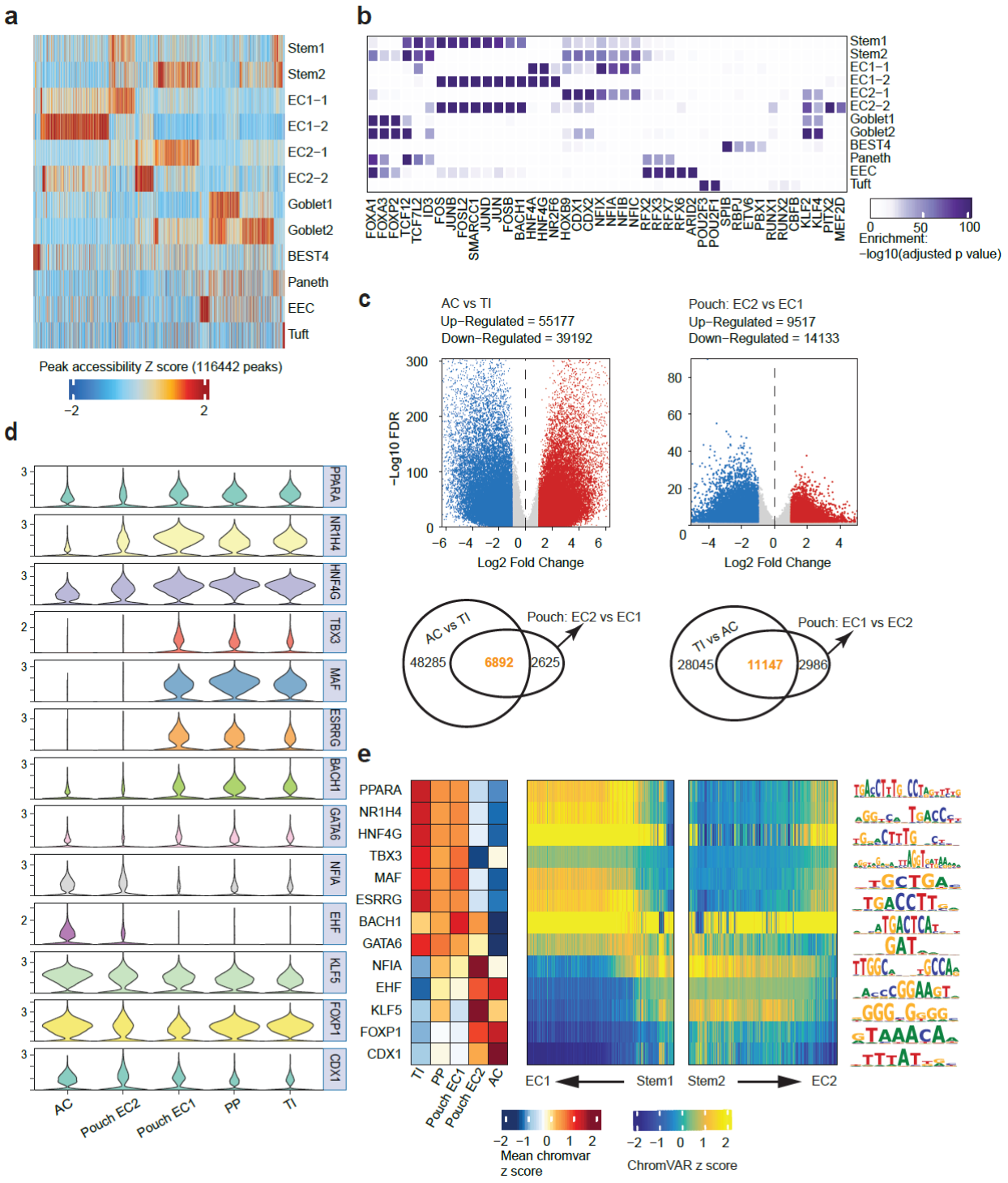


Figure 3.6 Identification of marker peaks and transcription factors (TFs) in each cell type in epithelial lineage.

Figure 3.6, continued

a Marker peaks for each cell type in epithelial lineage. Accessibility is scaled in each column across cell types. **b** Transcription factors enriched in each cell type. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons. **c** Differentially accessible regions in AC EC2 vs TI EC1 and pouch EC2 vs pouch EC1. **d** Stacked violin plot shows the RNA expression of different TFs in each enterocyte group. **e** Heatmap shows the scaled mean ChromVar TF activity score in each enterocyte group. The score is scaled in each row across enterocyte groups (left). TF activity in EC1 lineage and EC2 lineage along inferred differentiation pseudotime (middle, right).

We then used these candidate CREs to identify putative regulators of EC2 cells in the pouch. Analogous to our identification of differentially expressed genes between EC1 and EC2 cells (Fig. 3.2d), we performed pairwise comparisons between EC1 and EC2 cell populations from pouch, pre-pouch, TI, and AC (Methods). We identified 120,569 differentially accessible regions (DARs) in at least one comparison (FDR < 0.01, logFC > 1) (Fig. 3.6c, Fig. S3.23a) and 23,650 DARs between EC1 and EC2 cells in the pouch (14,133 up in EC1, 9,517 up in EC2). In contrast to scRNA-seq, we found substantially more DARs between AC and TI compared to Pouch EC1 and EC2. Nevertheless, this observation provided evidence for epigenetic differences between the two enterocyte cell types in the pouch and suggested plasticity within the epithelial lineages. Of note, 6,892 (72.4%) DARs in pouch EC2 vs EC1 populations were also identified as DARs when comparing EC2 and EC1 cells in AC and TI (Fig. 3.6c, Fig. S3.23a). Similarly, the comparison of EC1 and EC2 cells in the pouch yielded 11,147 (78.9%) DARs also found when comparing TI to AC enterocytes (Fig. 3.6c, Fig. S3.23a). This suggests that differences between EC2 and EC1 cells in the pouch were indeed associated with regulatory regions that broadly differentiated AC from TI. However, we also detected 36,848 DARs in total between EC2 cells in the pouch and AC mirroring the results of our DEG analysis and suggested substantial differences in chromatin accessibility between EC2 cells from these two tissues (Fig. S3.5d, Fig. S3.23a).

These sets of DARs were enriched for a range of motifs for TFs involved in epithelial differentiation as well as regulation of inflammatory processes (Fig. S3.23b). To identify candidate transcription factors associated with these enriched motifs, we computed the correlation between integrated RNA transcript levels and motif accessibility within the epithelial lineage (Methods) (Fig. S3.23c) and retained TF-motif pairs with a correlation > 0.5 (activating TF), and correlation < -0.5 (repressive TF)[34]. Expression (Fig. 3.6d, Fig. S3.23d) and motif accessibility (Fig. 3.6e, Fig. S3.23d, e) reflected differential activity in EC1 and EC2 cell types for most TFs.

Among the identified TFs, CDX1 and 2, NFIA, EHF, and KLF5 showed higher expression and motif accessibility in EC2 cells from both Pouch and AC (Fig. 3.6d, e). HNF4G, PPARA, GATA6, JUND were among the TFs associated with EC1 cells in the TI, pouch, and pre-pouch. Several of these TFs we identified in EC1 and EC2 cells in UC pouches, were previously identified in a multimodal single-cell atlas of the human intestine as having preferential activity in the ileum and colon, respectively[58]. We identified these TFs based on transcript levels and correlated enrichment of their annotated motifs within accessible chromatin. We confirmed that these motifs are occupied in epithelial cells clusters by performing TF footprint analysis (Methods, Fig. S3.24). Finally, our spatial transcriptomics data provided additional evidence for expression of these TFs within epithelial cells in pouch and pre-pouch samples (Supplementary File 3-8).

By comparing accessibility of these TF motifs along pseudotime-based differentiation trajectories from stem cells into EC1 and EC2 cells, respectively, we observed cell-type-specific activation pattern for these TFs during epithelial differentiation (Fig. 3.6e, Fig. S3.25, Supplementary File 3-9).

Notably, TF expression and motif enrichment appeared to be more heterogeneous in pouch EC2 cells compared to AC EC2 cells. Pouch EC2 cells also maintained activity of several TFs associated with EC1 cells, including BACH1 (Fig. 3.6e). Motif enrichment and expression of NFIA was more strongly associated with pouch EC2 highlighting differences in TF activity compared to AC. NFIA was identified in progenitor cells of early colonocytes in our data and by Hickey *et al*[58]. In contrast to AC EC2, NFIA motif activity remained high in pouch EC2 cells, possibly reflecting their less mature state.

Together, these TFs represent a key part of the gene regulatory network associated with the two transcriptionally distinct enterocyte populations observed in the pouch epithelium and provide evidence for the activation of a colon-like regulatory network in pouch EC2 cells.

Activation of a colon-like GRN underlies transcriptional identity of EC2 cells in the pouch

To identify putative target genes of identified cell-type-specific TFs in the pouch epithelium and characterize the regulatory network defining EC2 cells, we linked CREs to genes using co-accessibility of peaks with integrated expression[32] (Methods). We linked 31,808 peaks to 6,312 genes across EC1 and EC2 populations, identifying a large set of putative targets (Fig. 3.7a). We observed a strong enrichment of DARs associated with DEG compared to non-DEGs (Fig. S3.26). For example, distal regulatory elements linked to the colon-specific marker gene CA2 gained chromatin accessibility in pouch EC2 cells (Fig. 3.7b). In contrast, accessibility at CREs linked to ileal marker gene APOA4 was lower in pouch EC2 cells compared to EC1 cells in pouch, pre-pouch and TI (Fig. S3.27).

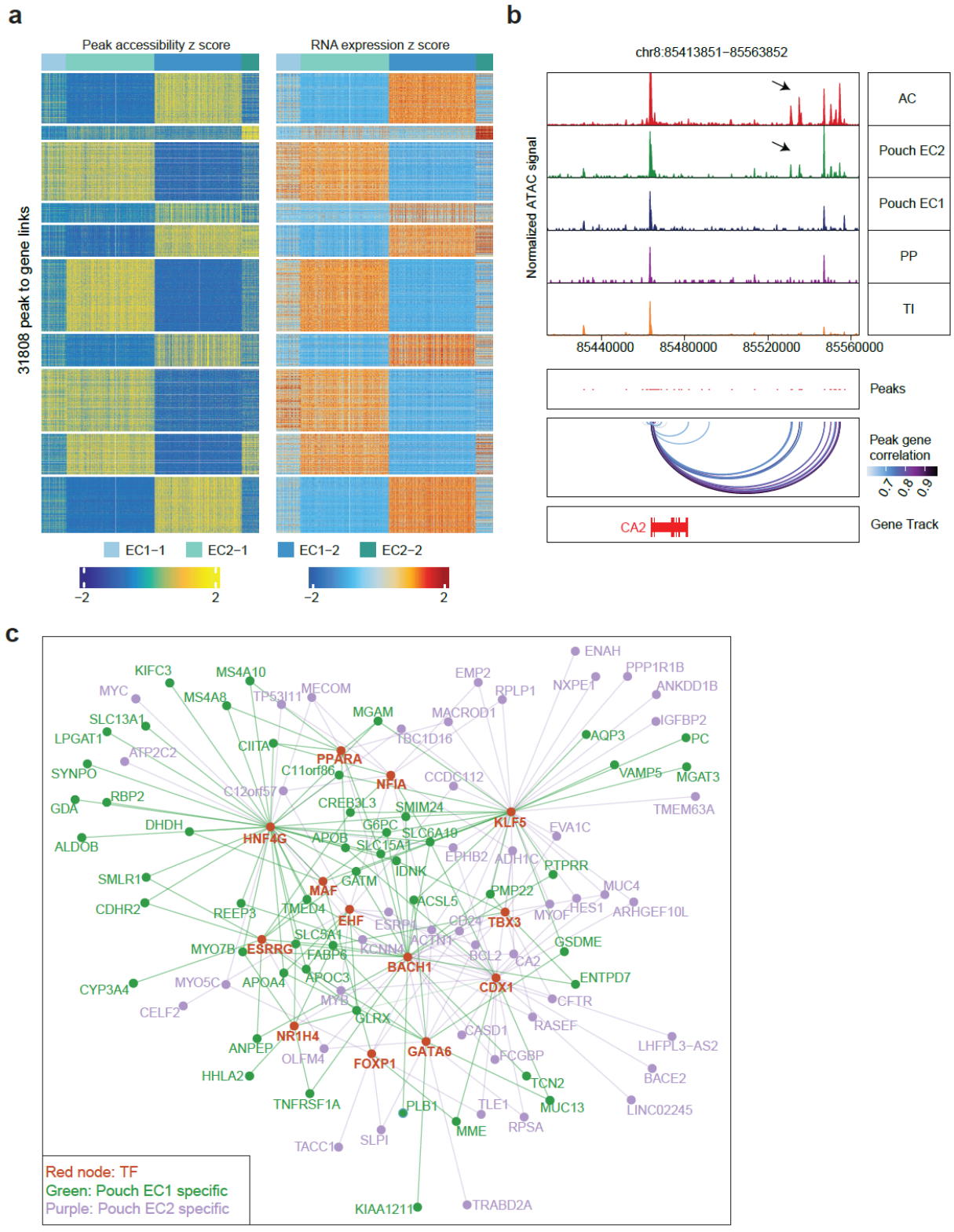


Figure 3.7 Construction of TF-gene regulatory network in pouch enterocytes by peak-gene linkages based on multiomic analysis.

Figure 3.7, continued

a 31,808 peak-gene links are identified based on the correlation of gene expression and peak accessibility with correlation score cutoff>0.6. **b** Peaks highly correlated with CA2 gene expression show higher accessibility in colon and pouch EC2 enterocytes (highlighted by arrows). **c** TF-gene regulatory network shows that the TFs (red nodes) are dominating different groups of genes in pouch EC1 and EC2.

We complemented this identification of target genes using FigR to identify domains of regulatory chromatin (DORCs)[150] which might harbor key target genes and facilitate the identification of regulatory factors. Using this metric, we ranked the genes by the number of associated peaks and identified 765 DORC-associated genes within the epithelial lineage. These putative target genes included both regulatory genes (e.g. ETS2, GATA5, KLF6) and genes associated with cellular function (e.g. OLFM4, MUC2, RBP2) (Supplementary File 3-10).

To summarize the regulatory mechanisms underlying EC1 and EC2 cells, we constructed a partial gene regulatory network by collating information from motif enrichment, TF expression, co-accessibility links, DORC scores, and differential gene expression (Methods). This regulatory network highlighted that the DEGs in the two pouch enterocytes populations are targeted by groups of overlapping TFs with differing activities in EC1 and EC2 cells (Fig. 3.7c). Our data therefore suggested that the colon-like transcriptional program in pouch EC2 cells is the result of the increased activation of TFs predominantly associated with the colon. For example, genes associated with colon function (CA2, CD24, CFTR) are induced in pouch EC2 cells compared to EC1 (Fig. S3.13b). While we showed that this cell type shares regulatory features (CREs and TFs) with enterocytes from the colon and undergoes similar differentiation dynamics, we note that compared to enterocytes from the ascending colon the acquisition of a colonic transcriptional and epigenomic features is incomplete. In addition, we identified several genes associated with ileal

function that are still expressed at significant levels in EC2 cells in the pouch compared to colon (e.g., *FABP6*, *APOA4*, *APOB*) (Fig. S3.12d).

Discussion

To our knowledge, this study is the first single-cell atlas for ileal pouch mucosa in UC patients post-IPAA surgery, leveraging a single-cell multiomic approach. Histological changes associated with UC pouch mucosa are well-documented[134,135], but the molecular basis of the cellular changes remains unclear[134]. Our study contributed to the understanding of these changes in the pouch by measuring the differences in cellular composition and cell-type-specific gene expression in the pouch compared to pre-pouch and control samples from AC and TI.

Through multiomic analysis, we identified and characterized partial colonic metaplasia of epithelial cell types in UC pouches and provided insights into the mechanisms underlying their differentiation. We identified transcriptional regulators active in colon-like pouch cells that were previously associated with colonic epithelial cells and epithelial progenitors, including CDX1, CDX2[151,152], EHF[153] and NFIA[58]. Activity of these TFs was correlated with the expression of genes with colon-specific functions including metabolic (CA2, HMGCS2) and anti-microbial (CEACAM5) activity. While CEACAM5 was expressed in EC2 cells from all samples, CA2 expression was lower compared to AC and only observed in some pouch samples. We also observed an increase in the expression of SATB2 in a subset of pouch EC2 cells. SATB2 is necessary for colonocyte differentiation, and its overexpression is sufficient to convert ileal epithelium into colonic epithelium in mice[152]. However, SATB2 was expressed at lower levels and more heterogeneously in pouch EC2 cells compared to AC EC2. In contrast to AC EC2, pouch

EC2 cells also maintained expression of a subset of genes with generally ileal-specific gene expression pattern (APOA4, APOB, FABP6) (Supplementary File 3-5), and appeared less mature compared to AC EC2. In addition, pouch EC2 and Goblet2 cells showed evidence of a transcriptional signature enriched for immune and secretory functions. Similar expression changes in epithelial cells have been observed in chronic inflammatory conditions in human IBD and animal models[24,26,27,154].

Bulk RNA-seq analyses have revealed distinct gene expression shifts indicative of regionalization in CD[155], and it will be interesting to understand whether this is indicative of metaplasia in these CD cases. Changes in gene expression of the pouch mucosa had previously been observed in bulk RNA-seq data[138]. Our study linked these changes to a particular cell population and comparing our cell-type-resolved data with bulk data provided evidence for the presence of EC2 cells in the majority of pouches. This analysis also suggested that the presence of EC2 cells alone was not linked to development of pouchitis. It is possible that the presence of EC2 cells is unrelated to pouchitis and a consequence of establishing pouches. Alternatively, EC2 cells may be a necessary, but not the sole, contributor to development of pouchitis. Familial adenomatous polyposis patients frequently undergo the same procedure to reduce risk of colon cancer yet reportedly rarely develop pouchitis[136], which suggests that establishment of a pouch alone does not render patients susceptible to inflammation.

Our data underscored the adaptability of the intestinal epithelium in adult tissues, yet the exact progenitor source for the colon-like metaplastic epithelial cells remains to be determined. The differentiation into colon-like and ileum-like enterocyte populations is most pronounced within

the pouch's mature secretory and absorptive cells. However, stem and progenitor populations in pouches diverged into distinct populations as well, and co-clustered closely with their mature counterparts in ileum and colon, suggesting an early stage of differentiation fate. Our analysis integrated cells from different tissues using Harmony which was necessary to account for inter-individual variation. This process facilitated the grouping of broadly similar cells and allowed for joint annotation of cell type clusters within a shared embedding of patient and reference samples. However, this process potentially removed non-technical variation between cell types from patient and controls (e.g. contributed by inflammation) and the resulting embedding might overemphasize similarities between EC2 cells in pouch and AC. Our DEG analysis addressed this concern by directly comparing expression patterns across different cell clusters and identifying both shared and different sets of genes.

While this study provides high-resolution insights into transcriptomic and regulatory signatures using primary human tissues, there are several limitations. The samples were obtained from a small cross-sectional cohort consisting of six individuals. This limitation guided our focus towards epithelial cell types that were both prevalent and exhibited consistent effects across the cohort. The study's focus does not imply that changes in other cell types, especially in stromal lineage, are unimportant but rather reflects our study's limited capacity to explore them comprehensively. For example, identification of colon-enriched NRG1+ fibroblasts[151] in the pouch might suggest the development of a colon-like niche in the pouch. Additionally, single cell genomics data lacks spatial context which is important to study the cellular niches that contribute to the observed cellular changes in the pouch. We generated a small spatial transcriptomics dataset for validation and provided evidence for the expression of key marker genes in pouch and pre-pouch samples.

However, these data do not have sufficient power to be used for discovery and the comprehensive characterization of cellular domains and niches.

Nevertheless, these data are a valuable resource for further cellular and molecular studies of IPAA. Our results suggest that UC pouches provide a unique system to study intrinsic and extrinsic cues contributing to epithelial plasticity and how modulation of epithelial identity relates to disease risks in the pouch and IBD in general.

Methods

Ethics statement

This study was conducted under the translational core research program protocol (15573A), approved by the Institution Review board of University of Chicago. Licensed gastroenterologists collected biopsy samples and clinical data from patients identified as UC-pouch and non-IBD control subjects. Biopsy sample and patients' demographic data were obtained from all participants with signed informed consent. Samples and demographic data were anonymized.

Patient information

Non-IBD controls were consented and recruited by the time of routine colonoscopy. UC-pouch patients were consented and recruited at follow-up pouchoscopy. Non-IBD controls were patients without a history of IBD, an immune related disease, an ongoing infection in intestines, a history of Familial adenomatous polyposis, a history of primary digestive tract malignancies, a history of metastatic malignancies in digestive tract, and symptoms of digestive disorders. The UC-pouch patients had a clinical diagnosis of ulcerative colitis, have a history of ileal pouch-anal anastomosis

and were reported free of active disease via macroscopic assessment from a physician during the pouchoscopy. Demographic data of patients are provided in Supplementary File 3-1.

Tissue biopsy sample collection

All biopsies were collected in locations determined by scoping gastroenterologists. Non-IBD controls had up-to-six biopsies obtained from each location during colonoscopy at two separate anatomic regions (terminal ileum and ascending colon). UC-pouch patients had up-to-six biopsies obtained from each region, at two regions (pre-pouch and pouch), during pouchoscopy. In short, mucosal linings were obtained using biopsy forceps following standardized care. Biopsied mucosal tissues from each region were immersed in Advanced DMEM/F-12 [Thermo Fisher 12634010] in an Eppendorf tube and placed on ice for short-term preservation and transfer. The wait times between biopsy and sample processing were under two hours.

Sample processing

Single cells from mucosal biopsies were acquired using a modified version of previously published protocols[27,156,157]. In brief, intact mucosal tissues were rinsed in 1.5 ml ice-cold PBS [Thermo Fisher 10010023] six times. Mucosal tissues from the same regions were combined and minced by iris scissors. Minced tissues were transferred to 2 ml fresh chelator solution (HBSS Ca/Mg-Free [Thermo Fisher 14175095], 10 mM EGTA [Fisher Scientific 50-255-956], 10 mM HEPES [Thermo Fisher 15630080], and 10% FBS [Gemini Bio 100-106]). The tissues were incubated at 37°C for 15 minutes with end-over-end rotation, then cooled on ice for 10 minutes. The epithelium was separated from lamina propria by mechanical agitation (shaking 20 times). Visualizing epithelial sheets in the solution, the supernatant was collected and spun down at 300xg for 5

minutes at 4°C. The remnant tissue chunks were rinsed in ice-cold PBS and incubated in 2 ml enzyme solution (HBSS Ca/Mg-Free [Thermo Fisher 14175095], 10 mM HEPES [Thermo Fisher 15630080], 200 μg/ml Liberase TM [Millipore Sigma 5401127001] and DNase I [Gold Bio D-301-100]) at 37°C for 20 minutes. The spun-down epithelium was resuspended and incubated in 2 ml TrypLE express [Thermo Fisher 12605010], at 37°C for 5 minutes. Both incubations were applied with end-over-end rotation. Tissues were gently triturated with a P1000 pipette after incubations and no visible tissue chunks remained after trituration. Cells were filtered through a 40 μM cell strainer [Falcon/VWR 21008–949]. After filtration, epithelial cells and cells from lamina propria were combined and spun down at 300xg for 5 minutes at 4°C. Cell pellet from each region were resuspended in 200 μl cold wash buffer (HBSS Ca/Mg-Free [Thermo Fisher 14175095] and 10% FBS [Gemini Bio 100-106]). Cell suspension was then diluted by 2 ml lysis buffer [Miltenyi Biotec 130-094-183] and sat for 3 minutes at room temperature to eliminate red blood cells. Cells in suspensions were spun at 300g for 5 minutes at 4°C. Cell pellets were resuspended in 200 μl suspension buffer (PBS and 0.04% BSA [Miltenyi Biotec 130-091-376]) and assessed for density and viability with Trypan Blue [Thermo Fisher 15250061]. Only cell suspensions with >85% viability were used for downstream single cell assays. The detailed processing is fully described here (https://www.protocols.io/view/preparation-of-single-cell-suspensions-from-human-q26g7bxqklwz/v4).

scRNA library preparation and sequencing

Single cells were loaded onto chip G per the manufacturer's protocol for the Chromium Single Cell 3' Library (v3.1) [10X Genomics 1000121]. In short, single cells were partitioned into Gel Beads in Emulsion (GEMs) in Chromium controller. GEMs were incubated for cell lysis and

barcoded reverse transcription, followed by cDNA amplification. The cDNAs were then enzymatically fragmented and indexed. Approximately 8,300 single cells were loaded to a channel with a target recovery for 5,000 cells. Libraries were sequenced on Illumina NovaSeq platform.

scATAC library preparation and sequencing

Single nuclei were isolated from single cell suspensions following the manufacturer's protocol for the Chromium Single cell ATAC (v1.1). Briefly, single nuclei were transposed before loading to chip H. Transposed nuclei were partitioned into GEMS in Chromium controller. GEMs were incubated for barcoding, followed by fragment indexing. Approximately 8,000 single nuclei were transposed and loaded to a channel with a target recovery for 5,000 nuclei. Libraries were sequenced on Illumina Novaseq platform.

scRNA-seq preprocessing and quality control

scRNA-seq gene expression raw sequencing data were processed using the Cell Ranger software v7.0.1 and the 10X human transcriptome GRCh38-2020-A as the reference. The expected number of cells were auto-estimated in Cell Ranger.

SoupX was used with default parameters for the estimation and removal of cell free mRNA contamination for each sample separately[101]. Clustering information from Cell Ranger output was loaded to estimate the contamination fraction. R (v4.1.0) and Seurat (v4.2.0) were used to pool single cell counts and for downstream analyses[102–104,106,158]. Cells were filtered for more than 500 UMIs, fewer than 20,000 UMIs, more than 200 unique genes, fewer than 6,000 unique genes, less than 50% of reads mapping to mitochondrial genes and less than 40% of reads mapping to

ribosomal genes and a 'log10GenesPerUMI' score higher than 0.7 (to remove contamination with low complexity cell types like red blood cells). Genes expressed in fewer than 120 cells were filtered out (average 5 cells per sample in all 24 samples). Doublet score was calculated using Scrublet (v0.2.3)[107], and a cut-off score of 0.2 was applied to exclude doublets. The cells with unexpected co-expression of different lineage makers such as CD3D and EPCAM were also excluded as doublets. Gene expression for each cell was normalized and log transformed. Cell cycle scores (G2/M phase, S phase) were calculated using the expression of cell cycle genes listed in Seurat (v4.2.0). Cell cycle score (the difference between the G2M and S phase scores), the percentage of mitochondrial reads and unique molecular identifiers (UMIs) were regressed out before scaling the data.

scRNA-seq dimensionality reduction, clustering analysis and visualization

We restricted the expression matrix to the 2000 most highly variable genes identified by fitting the mean-variance relationship and high-quality cells noted above as input for principal component analysis (PCA). Principal components were corrected using Harmony (v0.1.0)[140] to remove patient-specific effects. This strategy facilitated joint clustering and annotation of cells from all samples in a single embedding. Cells were clustered using the Louvain algorithm[159] (resolution 0.3–1.5) for modularity optimization using the kNN graph as input. Cell clusters were visualized in Uniform Manifold Approximation and Projection (UMAP) embedding[160].

The union dataset was divided into different lineages based on clustering analysis and marker gene expression. Cells from epithelial, T cell, B cell, myeloid, stromal lineages were subset for further analysis. For each lineage, we repeated the process of lineage-specific QC, dimensionality

reduction, principal component correction, Louvain clustering and UMAP visualization as described above. Cell identities for cell subtypes in each lineage were assigned based on markers outlined in the following section.

Cell-type annotation based on scRNA-seq data

Cell-type clusters were annotated based on marker genes previously reported as markers of intestinal cell populations[60].

Epithelial lineage cells were defined on the basis of *EPCAM* and *FABP1* expression. Stem cells had specific expression of *LGR5* and *SMOC2*. Transit-amplifying cells were classified based on the highest expression of *MKI67*, *TOP2A* and *TUBA1B*. Enterocyte clusters with ileum functional signatures include EC1-1 (*SI*, *GSTA2*) and EC1-2 (*SI*, *SLC15A1*, *APOA4*, *CUBN*). Enterocytes in EC1-1 cluster is at an earlier differentiation stage than EC1-2, with higher *OLFM4* expression. Enterocyte clusters with colonic signatures include EC2-1 (*CD24*, *CA2*) and EC2-2 (*CD24*, *CA2*, *CEACAM5*). EC2-1 cluster is at an earlier differentiation stage than EC2-2, with higher *OLFM4* expression. Goblet clusters were identified based on high *MUC2* and *TFF3* expression. Goblet2 cluster is a dominating type of goblet cells in the colon with higher *BEST2* and *KLK1* expression compared to Goblet1. Paneth cells (*DEFA5*, *ITLN2*, *REG3A*) and Microfold-like cells (*CCL20*) are observed in ileum, pre-pouch and pouch. BEST4 enterocytes, enteroendocrine cells (*CHGA*, *KCNB2*, *RIMBP2*) and tuft cells (*POU2F3*, *FYB1*) are identified in all four regions.

T cell lineage cells were defined based on positive *CD3D* expression. CD4 T cells had expression of *CD4*, but not *CD8A*, and vice versa for CD8 T cells. Tissue resident memory T cells were characterized by high *CD69* expression. CD4 tissue resident memory T cells were further divided

based on *ITGAE* (encodes CD103) expression. $CD103^{-}$CD4 T cells had higher *IL7R* expression, while $CD103^{+}$CD4 T cells had higher *IL17A* and *IL26* expression. CD8 tissue resident memory T cells include KLRG1 positive and negative clusters, indicating different cytotoxic functions. Central memory T cells had the highest *LEF1*, *TCF7* and *CCR7* expression. Regulatory T cells were characterized by high *CTLA4* and unique *FOXP3*, *IL2RA* expression. Two $TRDC^{high}$ clusters were classified as gamma-delta T cells (ENTPD1, $GZMA^{high}$) and NK cells ($NCR1^{+}KLRF1^{+}CD3D^{-}$) respectively. NK T cells expressed *CD3D*, *FGFBP2*, as well as NK genes *NCR1*, *KLRF1*. Mucosal-associated invariant T (MAIT) cells showed high *SLC4A10* and *NCR3* expression. Innate lymphoid cells (ILCs) (*PRKG1*, *PCDH9*, *AFF3*, *AREG*, *IL1R1*, *IL23R*, *KIT*) are a diverse group of immune cells, but we were not able to further sub-cluster ILCs in this atlas.

B cell lineage cells were defined based on high *CD79A* and *MS4A1* expression, and plasma cell clusters had high *XBP1* and *IGHA1* expression. *BCL6* expression defines germinal center B cells, which is crucial for somatic hypermutation and class switch recombination. Naïve B cells had the highest expression of immunoglobulin heavy chains D (*IGHD*). Memory B cells were characterized by positive *CD27* expression and negative *IGHD*. Class switch IgA and IgG plasma cells showed expression of the *XBP1*, and immunoglobulin heavy chains A and G, respectively.

Myeloid lineage cells were defined based on high expression of *TYROBP* and *CPA3*. Monocytes and macrophages both are *CD14* and *C1QA* positive. Monocytes had high expression of *VCAN*, *FCN1* and *EREG*, and macrophages had highest expression of *CD163* and *MMP12*. Among dendritic cells (DCs) we observed cDC1 (*CLEC9A*) and cDC2 (*CD1C*), and lymphoid DCs

(*LAMP3*). Mast cells were identified by *CPA3*, *KIT* and *TPSB2*. Neutrophils were classified by high expression of *FCGR3B* (encodes Fc gamma receptor 3B).

Stromal lineage cells were defined based on high expression of *IGFBP7* and *COL3A1*. Fibroblasts include three main populations, fibroblast1 (*ADAMDEC1*, *CCL11*, *CCL13*), fibroblast2 (*NRG1*, *NPY*, *PTGS1*), fibroblast3 (SOX6 high). Endothelial cells (*PECAM1*) include arterial endothelium population (*HEY1*, *EFNB2*), venous endothelium population (*ACKR1*) and lymphatic endothelium population (*PROX1*, *LYVE1*, *CCL21*). Pericytes were identified by high *NOTCH3* expression, and a contractile pericyte subset were characterized by high *PLN*, *RERGL* and *KCNAB1* expression. Myofibroblasts and smooth muscle cells both show high actin (*ACTA2*) and transgelin (*TAGLN*) expression, while Myofibroblasts lack smooth muscle marker desmin (*DES*). Glial cells were characterized by *S100B* and *NRXN1*.

Marker gene detection and differential expression analysis

In order to account for patient-specific effects when calculating the differentially expressed genes in enterocytes across regions, we used a linear mixed model (model formula = cluster + (1|Patient_ID)) implemented in variancePartition (v1.21.1)[161,162] to quantify gene expression attributable to individual patients. We verified that other confounding factors are not contributing a significant amount of variance, and we did not include additional factors in the linear model. The cell number threshold to create pseudo-bulk cell clusters was set considering the size of the cell type population (enterocyte groups: 200 cells, goblet cells: 30 cells, stem cell: 30 cells). Differentially expressed genes in different cell type clusters were identified by the 'FindAllMarkers' function (Wilcoxon Rank Sum test with default parameters) in Seurat. The other

pairwise differential tests were performed by the 'FindMarkers' function (Wilcoxon Rank Sum test with default parameters).

Gene set enrichment analysis

Gene set categories (H, C2, C5, C7 and C8) implemented in Molecular Signatures Database (MSigDB) were pulled and examined for each list of differentially expressed genes using msigdbr package (v7.5.1). R package 'org.Hs.eg.db' (v3.14.0) was used to map gene identifiers. Gene set enrichment analyses were performed using the 'clusterProfiler' R package with the 'GSEA' R function (pvalueCutoff = 0.05, pAdjustMethod = "fdr", minGSSize = 10, maxGSSize = 5000)[163].

Enrichment score of signature gene sets

'Area Under Curve' (AUC) was used to calculate the enrichment of a gene set for each cell, as implemented in AUCell (v1.16.0)[164]. For each cell, the genes are ranked from highest to lowest value. The genes with the same expression value are shuffled. Therefore, genes with expression '0' are randomly sorted at the end of the ranking. The AUC value is representing the fraction of genes from the pathway within the top 5% (aucMaxRank=0.05) of all the genes in our dataset.

Immunostaining of human biopsy

Biopsies matched to the patients assessed in single-cell RNA-seq experiments were obtained from the pathology archives at UChicago Medicine under an IRB-approved protocol. Serial sections were prepared from the paraffin-embedded material. Immunohistochemical staining was performed as previously described[165]. Briefly, sections were deparaffinized and rehydrated, heat/pressure-treated for 20 min with citrate buffer (pH 6), bleached with 0.3% hydrogen peroxide

for 30 minutes, and blocked with horse serum. Primary antibodies were prepared in 0.3% Triton-X and 4% horse serum in phosphate-buffered saline (PBS) and incubated with the tissue overnight at 4C. Primary antibodies used were rabbit anti-APOA4 (1:1,000, Sigma HPA001352), rabbit anti-FABP6 (1:200, Sigma HPA012601), mouse anti-CA2 (Santa Cruz Biotechnology sc-48351), and mouse anti-CEACAM5 (R&D Systems MAB41281). After washing with PBS, sections were exposed to horseradish peroxidase-conjugated secondary antibodies (Immpress kit, Vector Labs) for 2 h at room temperature, and signal was developed for 2-10 minutes with a diaminobenzidine (DAB) kit (Vector Labs). Tissue was counterstained with methyl green prior to dehydration, clearing, and coverslipping. A whole-slide scanner (Olympus) mounted with a 20X objective was used to obtain images of the staining. Analysis was performed with QuPath[166].

Histology evaluation and quantitation of enteritis score

Histological assessment of hematoxylin-and-eosin stained biopsy sections was performed by a clinical pathologist (C.W.) blinded to the slide labels. The scoring method used was based on previous assessment[167] of colonic inflammation and mucosal damage. The score ranged from 0 (no abnormalities) to 3 (severe damage), with intermediate grading defined as follows:

Score	Category	Description
0	Quiescent or Inactive	No active inflammation (no intraepithelial neutrophils)
1	Mild Active	Presence of intraepithelial neutrophils
2	Moderate Active	Presence of crypt abscesses
3	Severe Active	Erosions or ulcers

Bulk-RNA data deconvolution

We used cell type specific gene expression from our scRNA-seq data to characterize cell type compositions from bulk RNA-seq data (GSE81266), following the pipeline described in MuSiC (Multi-subject Single-cell Deconvolution, v 1.0.0) package[148], with default parameters.

scATAC analysis workflow preprocessing and quality control

scATAC-seq raw sequencing data were processed using the Cell Ranger ATAC v2.0.0, and the 10X human genome GRCh38-2020-A-2.0.0 as the reference. Further filtering was conducted using ArchR (v1.0.3)[32]. Only high-quality individual cells were kept for downstream analysis (doublet filter ratio = 2.0, minimum fragments per cell = 2,000, minimum TSS enrichment score = 6).

Dimensionality reduction and clustering analysis

We used ArchR to convert the fragments file into a 500bp tile matrix of fragment counts and generated a gene score matrix using the default model (model 42) with default parameters, taking into account chromatin accessibility within a gene body as well as proximally and distally from the TSS. We performed latent semantic indexing (LSI) on the tile matrix and retained the top 100 LSI vectors ('addIterativeLSI' function, 6 iterations including 5 different clustering resolutions: 0.5, 1, 1.5, 2, 2). The first LSI vector was removed because of high correlation with sequencing depth (correlation score > 0.75). The LSI vectors were further corrected Harmony (v0.1.0)[140] using PatientID as variable. The remaining 99 LSI vectors were used to create the UMAP (nNeighbors = 30, minDist = 0.2). Cell clusters were identified using the 'addClusters' function with multiple different resolutions.

Annotation of scATAC cells

The scRNA-seq expression matrix was integrated with the scATAC-seq gene score matrix using the 'addGeneIntegrationMatrix' function from ArchR. It identified the corresponding cells across datasets using Seurat's mutual nearest neighbors algorithm, thus transferring the labels from scRNA data to scATAC data. The labels in scATAC data were further curated based on specific marker gene score distribution. A notable exception were TA cells which did not have a clear corresponding cell type in scATAC-seq. We interpreted this as a consequence of the annotation of TA cells based on the expression of cell cycle genes which might not show strong corresponding changes in chromatin accessibility at their genomic loci. We therefore removed TA designation from scATAC-seq and used progenitor/immature instead. In addition, M-like cells which represent a small proportion of cells in the scRNA-seq data (<1.7%) and which we transcriptionally distinguished from enterocytes solely by expression of CCL20 showed a diffuse distribution, so we removed this label.

Accessible chromatin peak calling

We merged cells within each cell group for the generation of pseudo-bulk replicates and then merged these replicates into a single insertion coverage file. The 'addReproduciblePeakSet' function was used to get insertions from coverage files, call peaks, and merge peaks to get a "Union Reproducible Peak Set" (peaksPerCell = 500, maxPeaks = 150,000, minCells = 50, extsize = 150, cutOff = 0.05, extendSummits = 250).

Differential accessible region (DAR) analysis

To identify cell-type specific marker peaks, a single-cell insertion count matrix was created using the function ‘addPeakMatrix’ in ArchR. To perform DA, we used ‘getMarkerFeatures’ (Wilcoxon rank-sum test, maxCells = 20,000). To control for technical variation, we accounted for "TSSEnrichment" and "log10(nFrags)" when selecting a matched null group for marker feature identification (FDR < 0.01, logFC > 1).

Motif annotation and TF motif enrichment

We used 870 human motif matrices from the CisBP[114] database to annotate each of the peaks in the union peak set for each lineage separately by the ‘addMotifAnnotations’ function. The hypergeometric test was used to assess the enrichment of the number of times a motif overlaps with a given set of peaks, compared to random expectation (adjusted p-value < 0.001). The top 20 transcription factor motifs significantly enriched for each cell type were filtered based on top 3,000 broadly expressed genes (ranked by the proportion of cells expressing this gene). Tn5 bias-subtracted TF footprinting analysis was performed using the getFootprints function in ArchR for pseudo-bulk aggregates of cells in each cell type and tissue location.

ChromVAR deviation score

We computed background peaks controlling for total accessibility and GC-content by the function ‘addBgdPeaks’ and used the function ‘addDeviationsMatrix’ function to calculate the TF *ChromVAR* activity score. The TFs for each enterocyte group were further further filtered based on positive or negative TF activity-expression correlation across enterocyte cell types.

Trajectory analysis

The trajectory analysis was performed using the 'addTrajectory' function in ArchR by specifying the cluster order in two enterocyte lineages (EC1 lineage: Stem1, EC1-1, EC1-2; EC2 lineage: Stem2, EC2-1, EC2-2). We visualized trajectory-dependent changes of TF activity and gene expression along differentiation trajectories by ordering cells by pseudotime.

Domain of regulatory chromatin analysis

The domains of regulatory chromatin (DORCs) were identified using the standard workflow introduced in FigR (v0.1.0)[150]. The integrated gene expression matrix and peak matrix were used to infer cis-regulatory interactions. Here the integrated gene expression matrix was obtained from the multiomic integration analysis introduced above. The minimum number of interactions (cutoff=7) was used to determine a significant DORC gene.

Gene regulatory network construction

Highly correlated links between gene expression and region accessibility were identified by 'addPeak2GeneLinks' (corCutOff = 0.6), as candidate cis-regulatory elements (cCREs). The core transcription factors were first identified by the TF enrichment analysis of different groups of differentially accessible regions between enterocyte groups. These groups of TFs were then filtered for highly expressed genes. Finally, only the most significantly positively and negatively TF regulators in the same TF family were preserved as nodes in the network. We filtered the gene targets for the top 100 most significantly linked genes, the well-known functional genes (within top 300 most significantly linked genes), and top 100 DORC genes (within top 300 most significantly linked genes).

Xenium 10x spatial transcriptomics

Samples were treated following the Xenium 10x Genomics spatial transcriptomics protocol (CG000749). FFPE samples were sectioned onto a Xenium slide by the Human Tissue Resource Center at the University of Chicago. Xenium data were generated at the NuSeqCore at Northwestern University using a set of 422 probes. 322 probes from the pre-designed Xenium Human Colon Gene Expression Panel Kit (1000642) and 100 probes designed to target pouch relevant genes (Supplementary File 3-11). Raw data were processed using Xenium Analyzer. Cell segmentation was performed with the Xenium Cell Segmentation Add-on Kit (1000662).

Spatial domain determination vis IRIS

The xenium 10x spatial transcriptomics data of two pouch and two pre-pouch slices from the same individual were loaded in R version (v4.3.1) using Seurat (v5.1.0)[102–104,106,158] and analyzed with IRIS (v1.0)[147]. In our scRNA-seq data, we merged cell type labels for enterocytes, goblets, T cells, B cells, plasma cells, myeloid cells, stromal cells and endothelial cells. scRNA count matrices were then extracted and used to construct IRIS objects using the function createIRISObject, with minCountGene set to 20 and minCountSpot set to 5. Function IRIS_spatial was called to isolate main spatial domains with numCluster set to 5. Differentially expressed genes were identified using the function FindMarkers from Seurat with default parameters.

CHAPTER 4: Discussion

Understanding how inflammatory bowel diseases manifest at the cellular and molecular levels is central to advancing both basic biology and clinical care. This dissertation addresses two critical aspects of IBD: (1) CD pathology as it unfolds in the terminal ileum and ascending colon, and (2) the adaptive transformations occurring in ileal pouches created for UC patients.

In the second chapter, scATAC-seq was applied to identify 557,310 candidate cCREs across multiple cell types in the intestinal mucosa. This large-scale atlas represents a significant advance in resolving how genetic risk loci are mapped onto specific cell types. Immune cells emerged as predominant mediators of CD heritability when analyzed at a genome-wide level, consistent with the view that T cells, B cells, macrophages, and other immune populations orchestrate chronic inflammation. Nevertheless, the discovery that fine-mapped risk variants frequently reside in regulatory regions specific to epithelial and stromal cells underscores the complex etiology of CD: while immune responses are critical for disease propagation, non-immune cell types also modulate pathology at certain loci. Additionally, caQTL analysis confirmed that SNPs that disrupt cell-type-specific TF binding motifs significantly affected chromatin accessibility in specific cell types.

The prioritized putative causal variants within cell-type-resolved cCREs in CD provide a more nuanced understanding of intestinal inflammation. For example, variants associated with *IL2RA* are implicated in T cell homeostasis and have been associated with multiple autoimmune conditions[168–170]. Similarly, *MMP9* has been highlighted in inflammatory responses[171], particularly in tissue remodeling and extracellular matrix degradation[172]. Other prioritized variants, such as those overlapping cCREs linked to *LRRC32* or *CIITA*, suggest potential roles for T regulatory cell

function and MHC class II antigen presentation, respectively[173]. In epithelial cells, SNPs near *SLC22A4/A5* and *DUSP5* further underscore the importance of barrier function and injury responses in disease onset and progression[174]. These examples illustrate that IBD arise from an interplay of genetic and regulatory alterations rather than a single causal gene. Clinically, pinpointing the precise regulatory mechanisms behind these genetic signals opens up opportunities for biomarker development and targeted therapies.

Chapter 3 focuses on UC patients who undergo total proctocolectomy with IPAA. Although this surgery removes the inflamed colon, it often leaves patients with a newly formed ileal pouch susceptible to inflammation (pouchitis). By using scRNA-seq and scATAC-seq, this dissertation identifies how the pouch epithelium undergoes partial colonic metaplasia: despite originating from ileal tissue, pouch enterocytes adopt a hybrid transcriptional profile that incorporates colon-specific genes (e.g., CEACAM5, CD24), while retaining some ileal-specific markers (e.g., FABP6, APOA4). These cells also exhibit upregulation of inflammatory and secretory signatures reminiscent of colonic epithelium in UC patients, suggesting that pouch formation reshapes the epithelial niche. Notably, this partial metaplastic phenotype also occur in pouches that are without overt pouchitis and early after functionalization.

The corresponding scATAC-seq data illuminate the regulatory mechanisms of this phenomenon. TF binding motifs for colonic regulators such as CDX1, CDX2, and NFIA become accessible in pouch enterocytes, reinforcing that the epigenome is reprogrammed toward a colonic-like identity. However, these cells still differ from classic colonic cells, likely due to persistent ileal transcription factors, subtle ongoing inflammation, or other local cues (e.g., microbiome composition). The

presence of this unique epithelial state could explain the susceptibility of some pouches to chronic inflammation, although direct causality remains an open question.

The CD gut atlas reveals that the majority of prioritized CD risk loci localize to cCREs that are more accessible in T cells, B cells, and myeloid populations. This aligns with established evidence that CD pathology frequently involves immune-driven inflammation throughout the gastrointestinal tract[175]. Conversely, the ileal pouch study and complementary work in UC underscore that UC's risk involves more epithelial-centric programs, particularly in cells mediating barrier function[176]. The pouch data show that UC patients after colectomy and IPAA surgery develop colonic metaplasia in the ileal pouch. These findings imply that while immune mechanisms remain important in UC, the epithelial layer may be central to UC-specific vulnerability, which might explain UC's typical confinement to the colon. Recognizing these lineage-specific contributions helps refine our understanding of disease mechanisms, opening avenues for interventions that focus more specifically on either immune regulation (for CD) or epithelial restoration (for UC).

Both the CD atlas and the UC pouch atlas highlight extensive epigenetic remodeling across tissues, including changes in chromatin accessibility at key regulatory regions. Epigenetic modifications, unlike genetic variants, are potentially reversible, raising the possibility of reversing maladaptive states through targeted therapies[177]. For instance, if aberrant transcription factor binding or chromatin accessibility can be corrected in the intestinal epithelium or specific immune subsets, it could halt or reverse pathologic processes, such as excessive cytokine production or epithelial barrier dysfunction. Pharmacological agents like histone deacetylase inhibitors, bromodomain

inhibitors, or novel small-molecule regulators of chromatin state could, in theory, tip the balance back toward immune tolerance and mucosal healing[178]. While still at a relatively early stage, these epigenome-centered approaches represent a promising direction, underscoring the need for deeper functional characterization of regulatory elements and the TF networks that drive inflammatory or metaplastic states.

In the UC pouch study, the discovery of EC2 epithelial cells with partial colonic transcriptional and epigenetic features, raises intriguing questions about metaplasia and disease localization. These EC2 cells express markers such as CEACAM5 and CD24, making them a partial colonic identity despite originating from the ileum. Determining whether EC2 cells co-localize precisely with areas of significant metaplasia or inflammation could shed light on why UC pathology is typically restricted to the colon. If EC2 cells aggregate in regions where pouch mucosa is most inflamed, then metaplastic transformation might be an early and perhaps necessary step for UC-type lesions to arise. Conversely, minor colocalization could imply that the presence of colon-like cells alone does not drive UC-like inflammation. Future work integrating large-scale spatial transcriptomics and long-term follow-up will clarify the timing and pattern of EC2 cell emergence, as well as whether targeting these cells can reduce the risk of pouchitis or other UC-related inflammation. Ultimately, understanding the cues that spark or sustain EC2 differentiation could yield new opportunities for preventive strategies in both UC and pouch management.

Taken together, both chapters emphasize the power of single-cell approaches to unravel complexity in IBD. Chapter 2 places emphasis on the interplay of genetic risk variants and cell-type-specific regulatory programs in CD, while Chapter 3 highlights the capacity for the intestinal

epithelium to develop different phenotypes under surgical reconfiguration. From a therapeutic perspective, the data provide several major insights.:

1. While immune populations are undoubtedly central, epithelial and stromal cells also play roles in the etiology of disease. Interventions targeted solely at immune-mediated inflammation may underappreciate how non-immune cells contribute to barrier integrity, cytokine production, and overall tissue resilience.
2. The comprehensive atlas of 557,310 cCREs provides a crucial reference for pinpointing how non-coding variants alter regulatory networks in IBD. Future treatments may harness this knowledge, selectively modulating the relevant loci in specific cell types.
3. The partial colonic metaplasia seen in UC pouches reflects the gut's plasticity. This cell-type-resolved transcriptomic and epigenetic atlas of the ileal pouch establishes a reference for investigating pouch physiology and pathology.
4. Despite providing valuable molecular blueprints, both chapters highlight limitations. Larger cohorts would enhance statistical power for caQTL discovery, especially in rarer cell types. Spatial analyses are also crucial for mapping micro-environmental cues. Longitudinal sampling of pouches or CD-affected mucosa could clarify how regulatory landscapes evolve over time or under therapy.

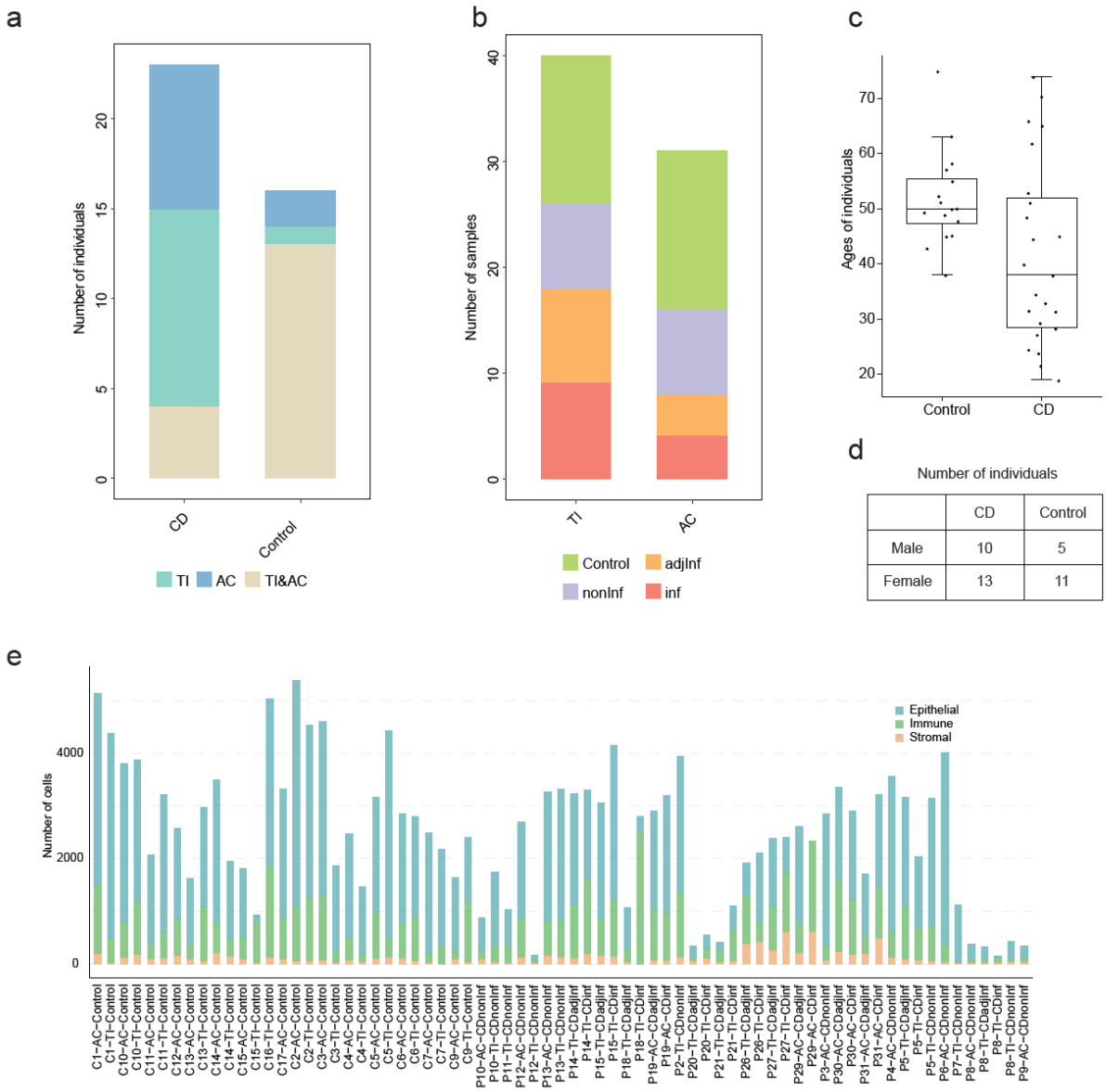
Building on the foundations laid by this dissertation, future studies could focus on the following directions:

1. Longitudinal Multi-Omic Tracking: Repeated sampling from IBD patients at distinct disease stages, especially during transitions from non- to active inflammation, could

disentangle cause-and-effect relationships between transcriptomic or epigenomic remodeling and disease severity.

2. Integrating Microbiome and Metabolomics: Since microbial composition and their metabolites profoundly influence gut epithelial states, pairing single-cell genomics data with high-resolution microbial profiling may uncover the cross-talk that fosters or alleviates inflammation.
3. Functional Validation: CRISPR-based perturbations in *ex vivo* organoids or *in vivo* models can now target specific SNPs or cCREs, allowing direct tests of whether certain variants or TF binding events indeed drive pathogenic programs in the relevant cell types.
4. Translational Applications: Ultimately, these integrated single-cell atlases aim to inform clinical practice by guiding biomarker discovery (e.g., identifying cell-type-specific transcripts or genomic regions that cause disease) and uncovering pathways amenable to therapeutic intervention, moving the field closer to personalized, mechanism-driven treatment strategies for IBD.

Appendix A: Supplementary Figures for Chapter 2



	CD	Control
Male	10	5
Female	13	11

Figure S2.1 Sample overview.

a, The number of individuals that have samples collected from only terminal ileum (TI), only ascending colon, or both regions (TI & AC). **b**, The number of samples in different disease conditions. Control: samples from the control group. nonInf: samples without active inflammation from CD patients. adjInf: non-inflamed tissue adjacent to inflamed site. inf: inflamed tissue. **c**, Distribution of ages of individuals. **d**, Number of individuals stratified by sex and disease condition. **e**, The number of cells across different lineages in each sample.

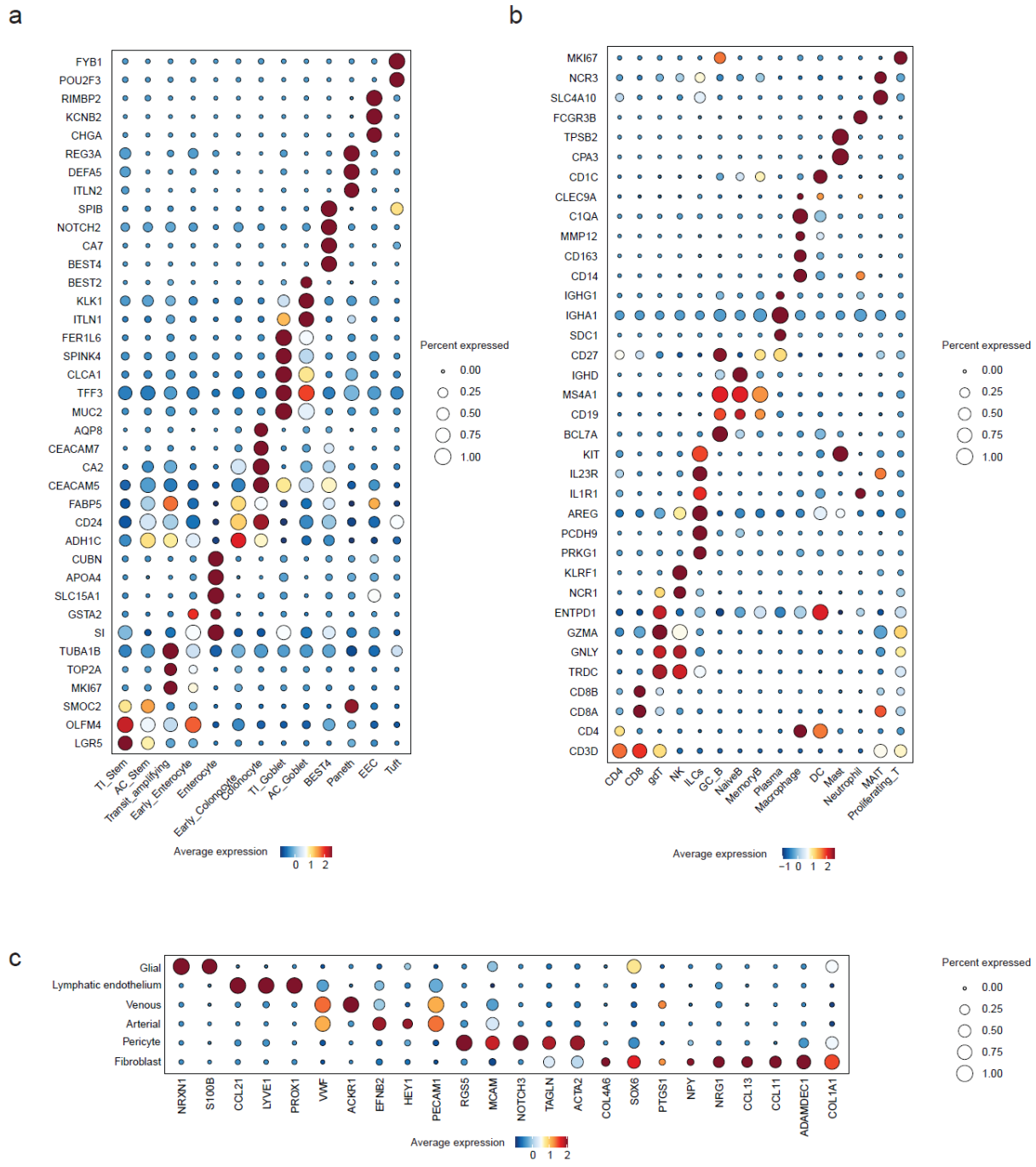


Figure S2.2 Markers for scRNA-seq data annotation.

a, **b**, **c**, Expression of markers for different cell types in each lineage (a, epithelial. b, immune. c, stromal.). The color shows scaled mean expression in each cell type. The dot size represents the proportion of cells expressing that marker.

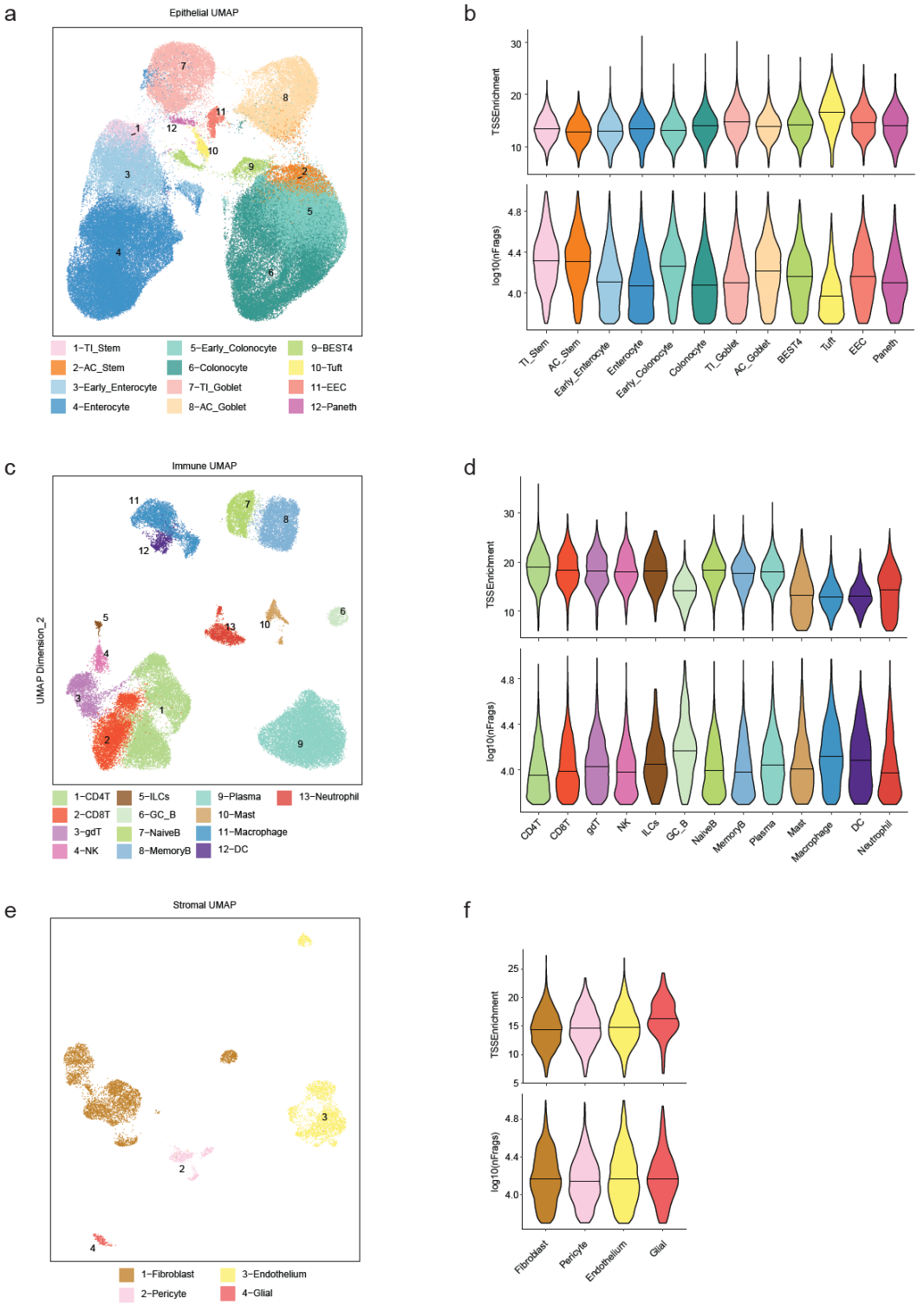


Figure S2.3 Cell type annotation and quality control in scATAC-seq.

Visualization of different cell types in lineage-specific UMAP embedding and the distribution of single cell TSS enrichment scores and fragment counts in each cell type (a, b, epithelial. c, d, immune. e, f, stromal.).

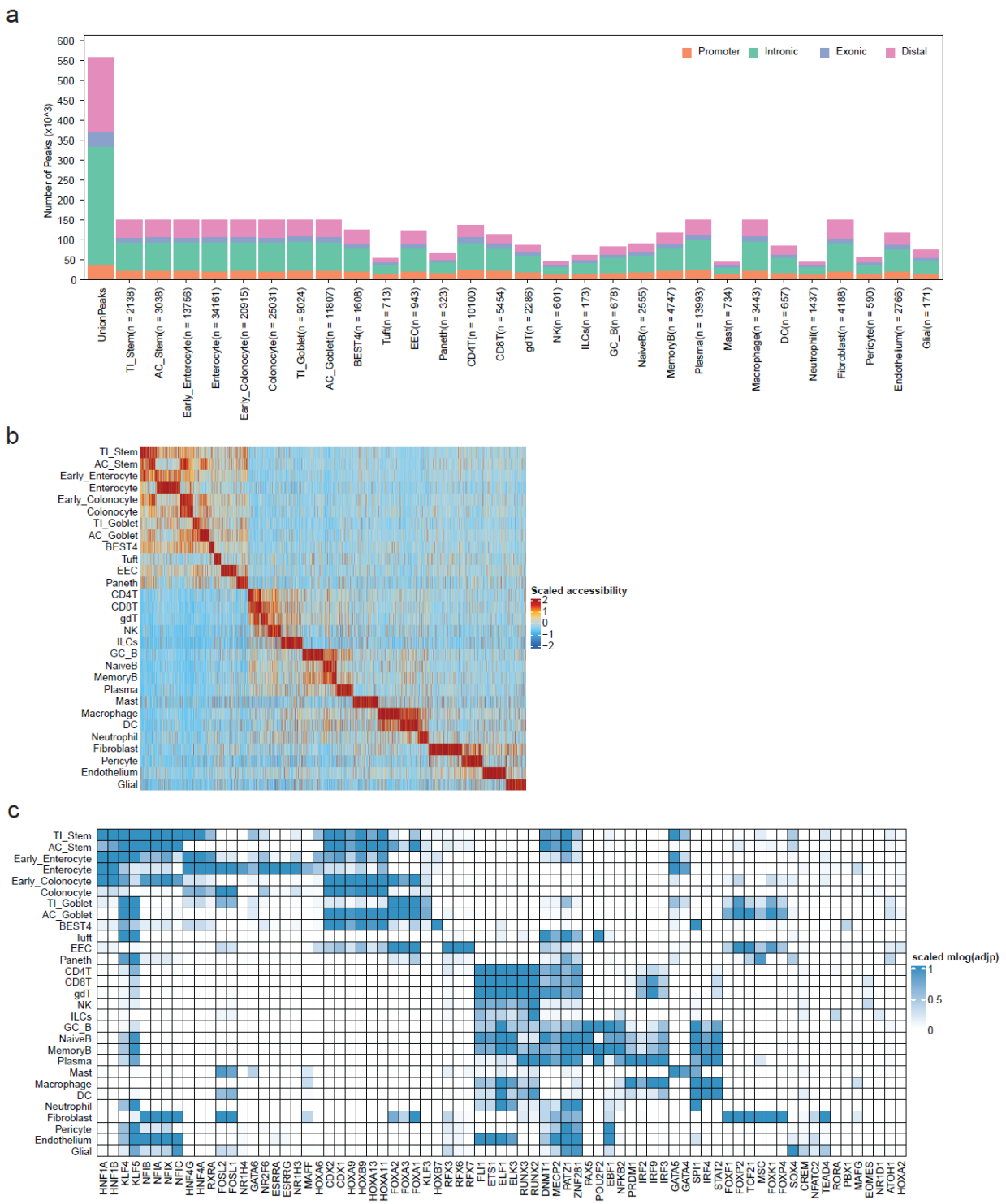


Figure S2.4 Cell type specific DARs and TFs.

a, The number of peaks in the union peak set and the peakset called using MACS2 for each cell type. **b**, Cell-type-specific differentially accessible regions (DARs). Each column is a DAR.

Figure S2.4, continued
c, Cell-type-specific transcription factor motifs enriched from cell-type-specific DARs. The -log10(adjusted p-value) for motif enrichment was capped at 100, with any values exceeding 100 being set to 100, and scaled by the maximum value per TF (right).

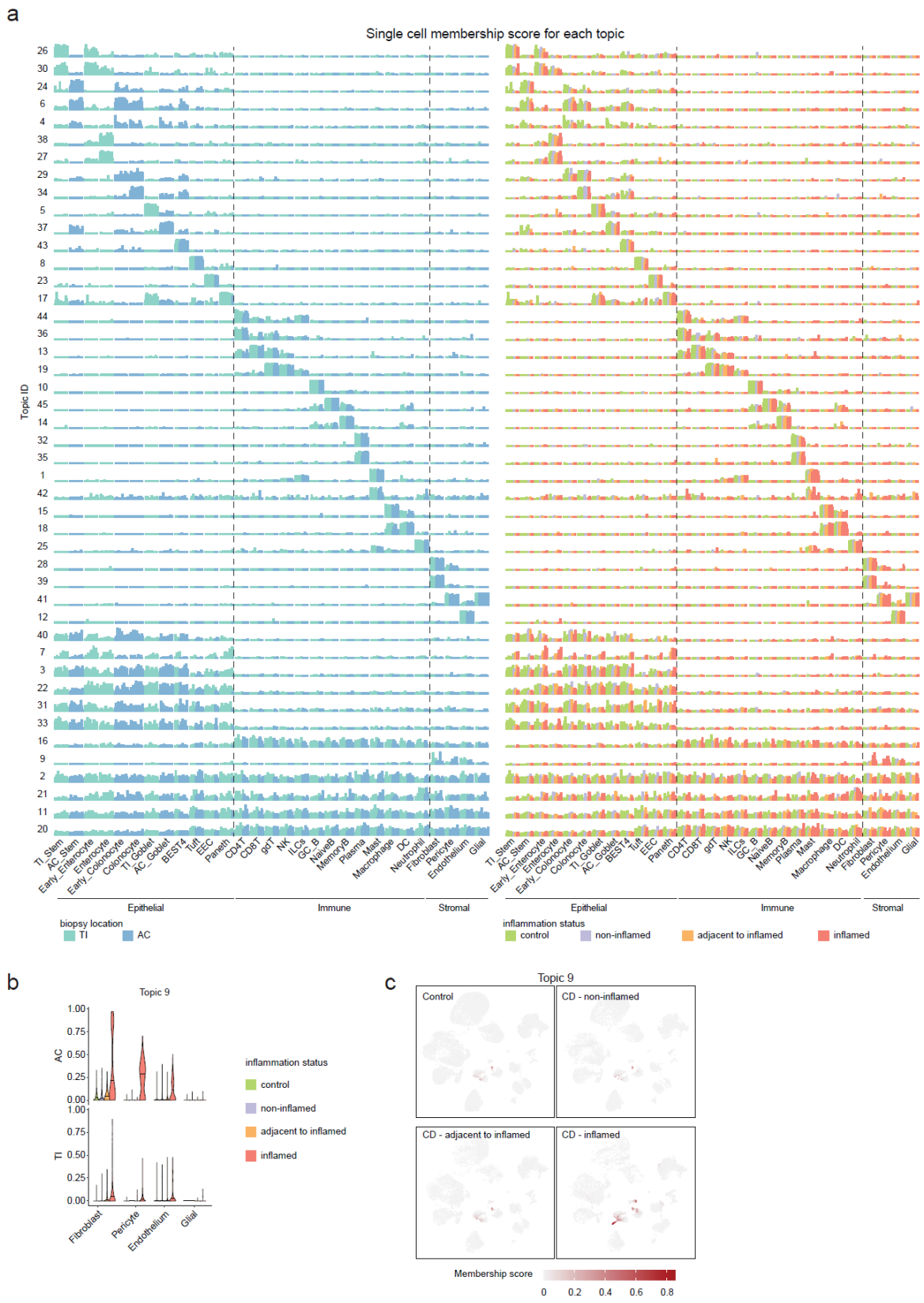


Figure S2.5 Structure plots of all topics.

Figure S2.5, continued

a, Structure plots displaying proportion of topics in individual cells. The cells are colored by biopsy location (left) or biopsy inflammation status (right). Each column represents a single cell and cells are grouped by cell types. Topic scores are scaled by the maximum value per row. **b**, Proportion of topic 9 membership for each cell. Cells grouped by cell type, tissue location, and inflammation status. **c**, UMAP plots displaying proportion of topic 9 membership for each cell, split by inflammation status.

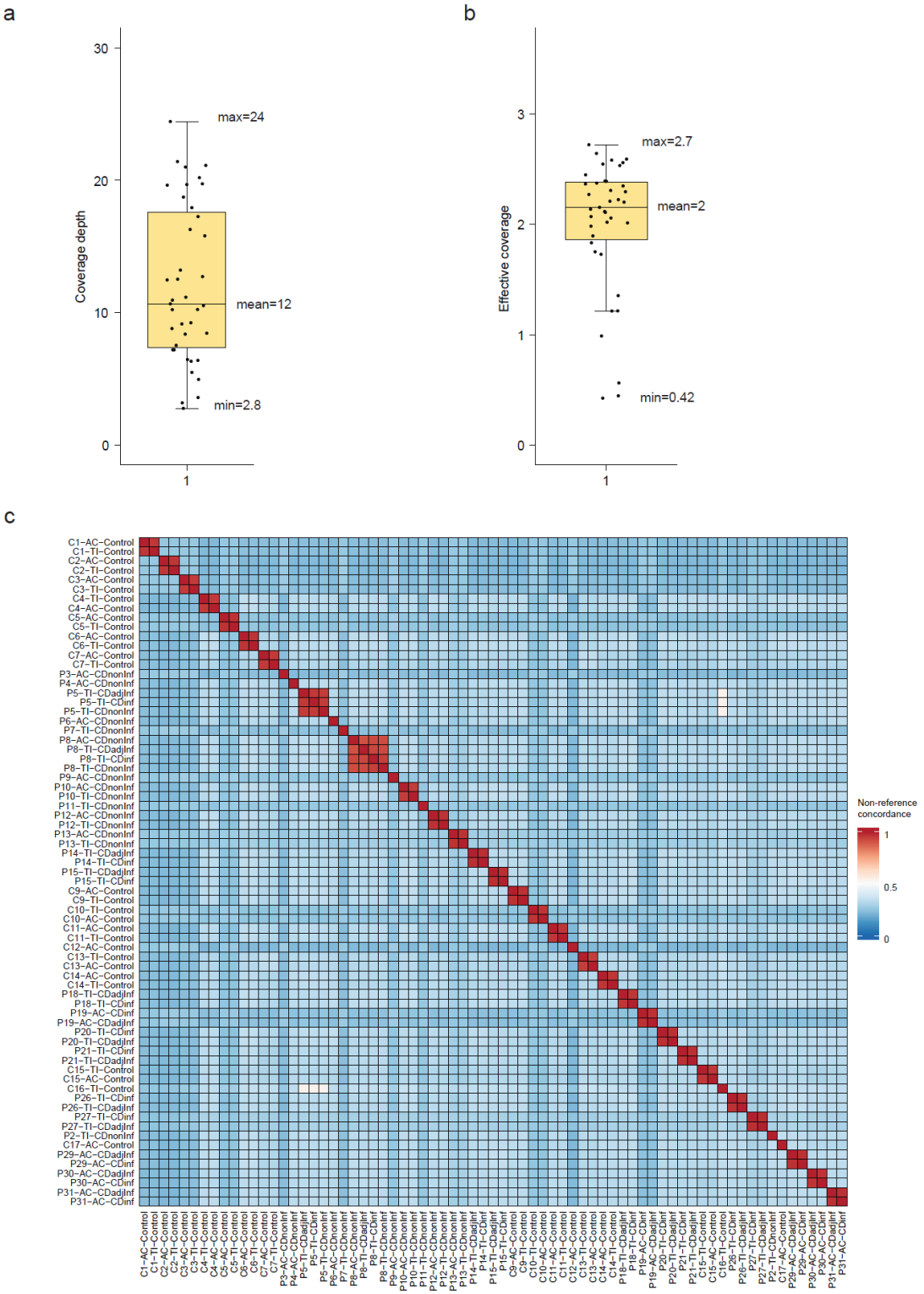


Figure S2.6 Quality control of genotyping.

Figure S2.6, continued

a, Distribution of coverage depths of individual-level BAM files. **b,** Distribution of effective coverages of individual-level BAM files. **c**, Non-reference concordance of imputed variants in chromosome 1 between each pair of samples.

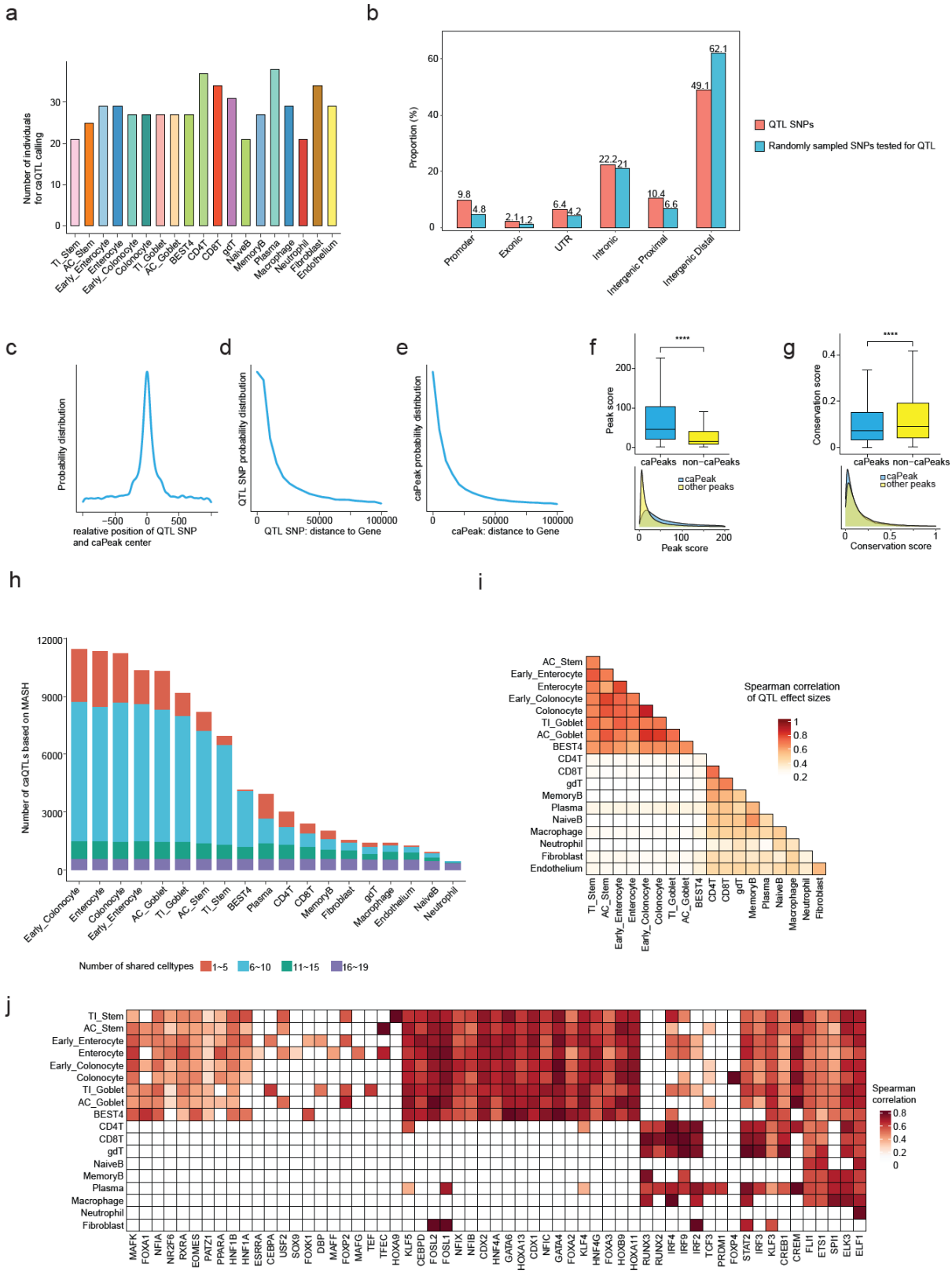


Figure S2.7 Evaluation of caQTLs.

Figure S2.7, continued

a, Number of individuals used for caQTL calling in each cell type. **b**, Proportion of significant QTL SNPs and randomly sampled SNPs across genomics annotations. UTR: Untranslated Region, Promoter: within 2 kb of transcription start site of the nearest gene, Intergenic proximal: within 5 kb of the 5' end of the nearest gene. Intergenic distal: more than 5 kb away from the 5' end of the nearest gene. **c**, Distribution of the distances between each QTL SNPs and associated caPeaks. **d**, Distribution of the distances between each QTL SNPs and the nearest gene. **e**, Distribution of the distance between each caPeak and the nearest gene. **f**, Peak scores obtained from MACS2 peaking calling for caPeaks and non-caPeaks. **g**, Sequence conservation (PhastCons score) for caPeaks and non-caPeaks. **h**, Number of caQTLs significant in each cell type assessed by mashR (LFSR<0.05). The number of shared cell types for caPeaks are indicated by color. **i**, Spearman correlation coefficients of QTL effect sizes for all pairwise cell type comparisons based on mash results. **j**, Heatmap of Spearman correlation coefficients between QTL motif breaking scores and effect sizes for each TF-cell type pair. Coefficients are only displayed for significant correlations (FDR <0.05). The QTLs for this analysis were determined to be significant by mashR.

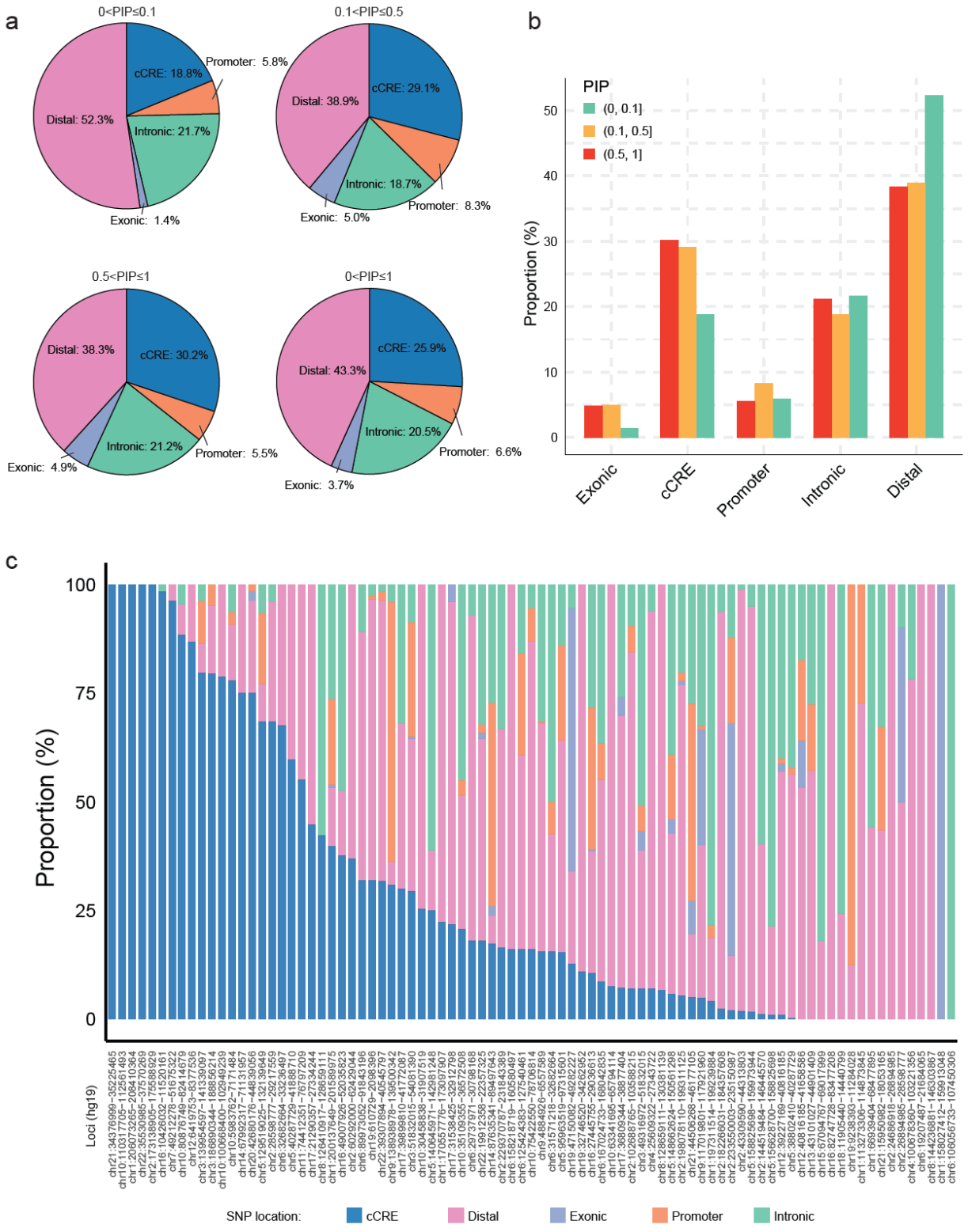


Figure S2.8 Overlap of fine-mapped SNPs with genomic regions.

Figure S2.8, continued

a, b Proportion of summed PIPs in each type of genomic region. SNPs within different ranges of PIP values were calculated separately. **c**, Proportion of summed PIPs in different types of genomic regions at each individual locus.

Appendix B: Supplementary Figures for Chapter 3

a

Individual ID	Biopsies	Years after surgery
P1	Paired pouch and pre-pouch	14
P2	Paired pouch and pre-pouch	9
P3	Paired pouch and pre-pouch	23
P4	Paired pouch and pre-pouch	31
P5	Paired pouch and pre-pouch	10
P6	Paired pouch and pre-pouch	14
C1	Paired ileum and colon	NA
C2	Paired ileum and colon	NA
C3	Paired ileum and colon	NA
C4	Paired ileum and colon	NA
C5	Paired ileum and colon	NA
C6	Paired ileum and colon	NA

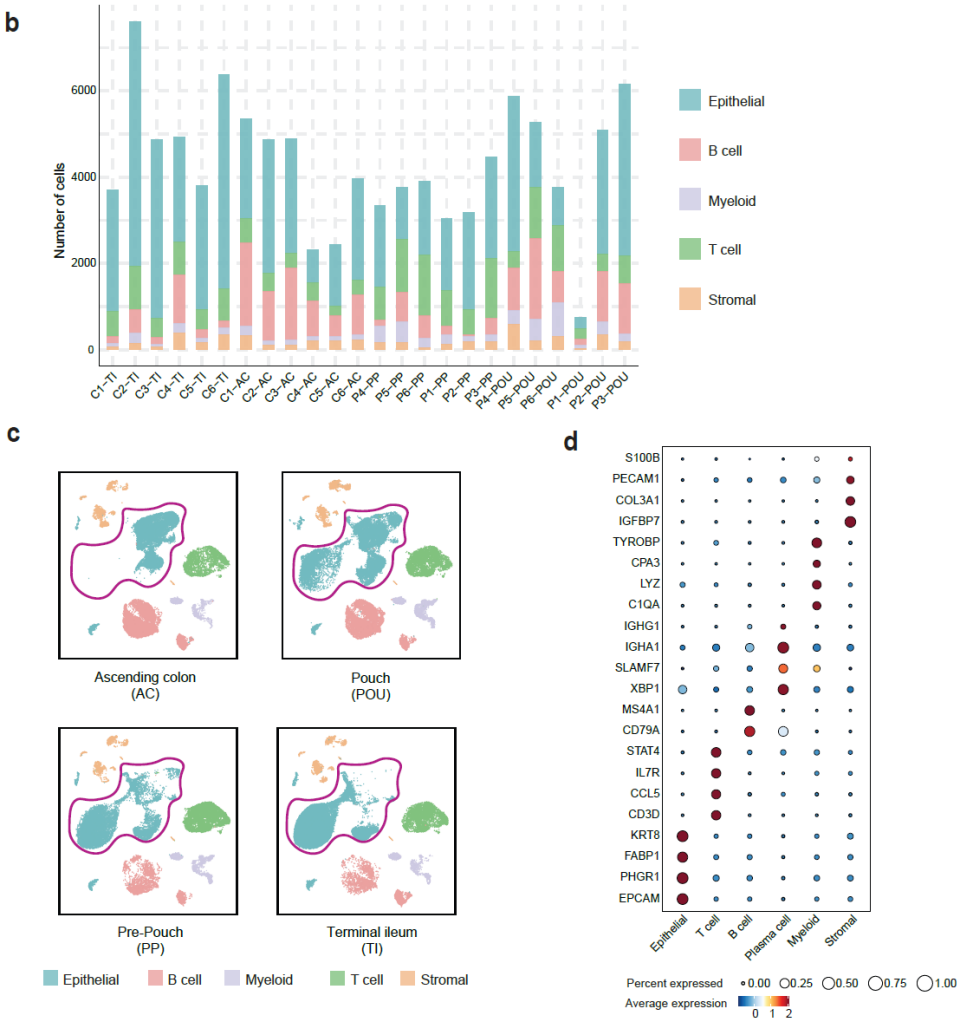


Figure S3.1 Data overview.

Figure S3.1, continued

a Metadata of the patients involved in this study. **b** Number of cells in each biopsy sample that have been analyzed after QC in scRNA-seq analysis. **c** UMAP split by regions highlights that the largest cross-region differences lie in epithelial lineage. **d** Expression of markers for different cell lineages. Color indicates scaled mean expression in each lineage, dot size represents the proportion of cells expressing that marker. Source data are provided as a Source Data file.

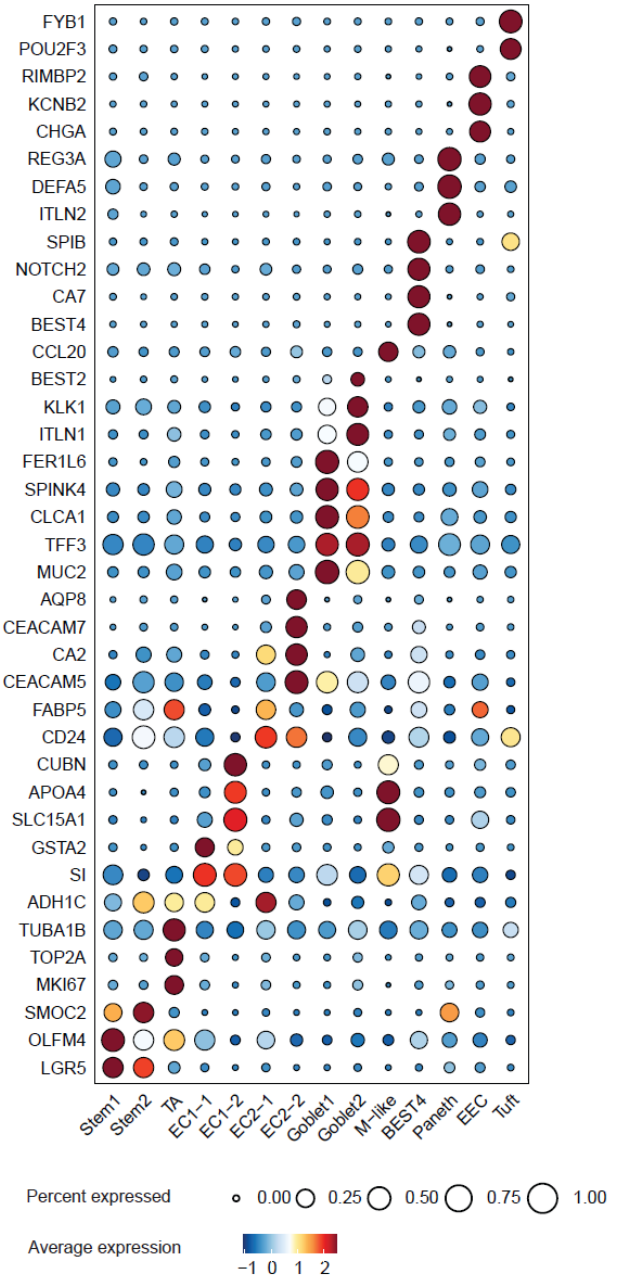


Figure S3.2 Expression of markers for different cell types in epithelial lineage.

The color shows scaled mean expression in each cell type. The dot size represents the proportion of cells expressing that marker.

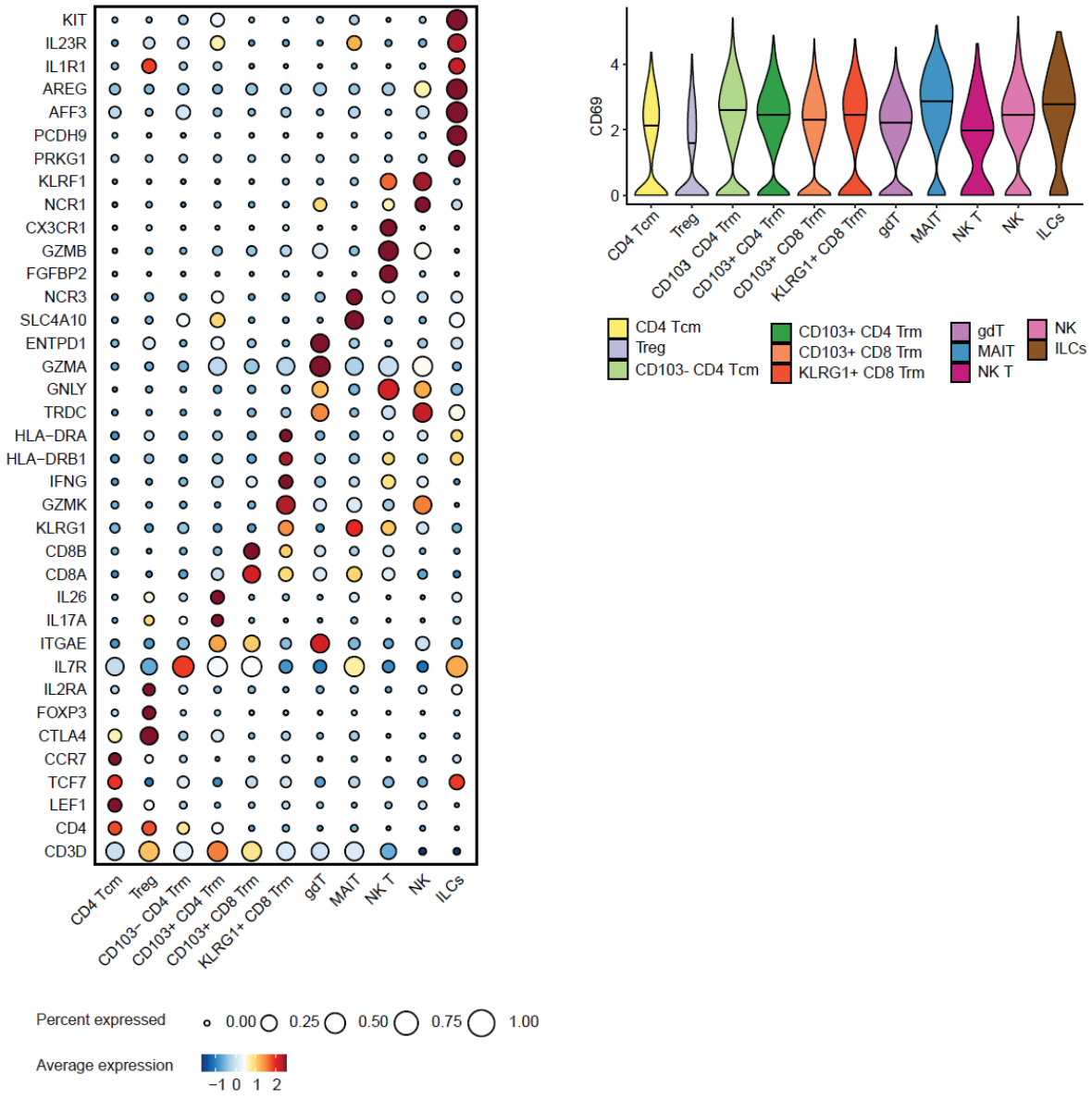


Figure S3.3 Expression of markers for different cell types in T cell lineage.

The color shows scaled mean expression in each cell type. The dot size represents the proportion of cells expressing that marker.

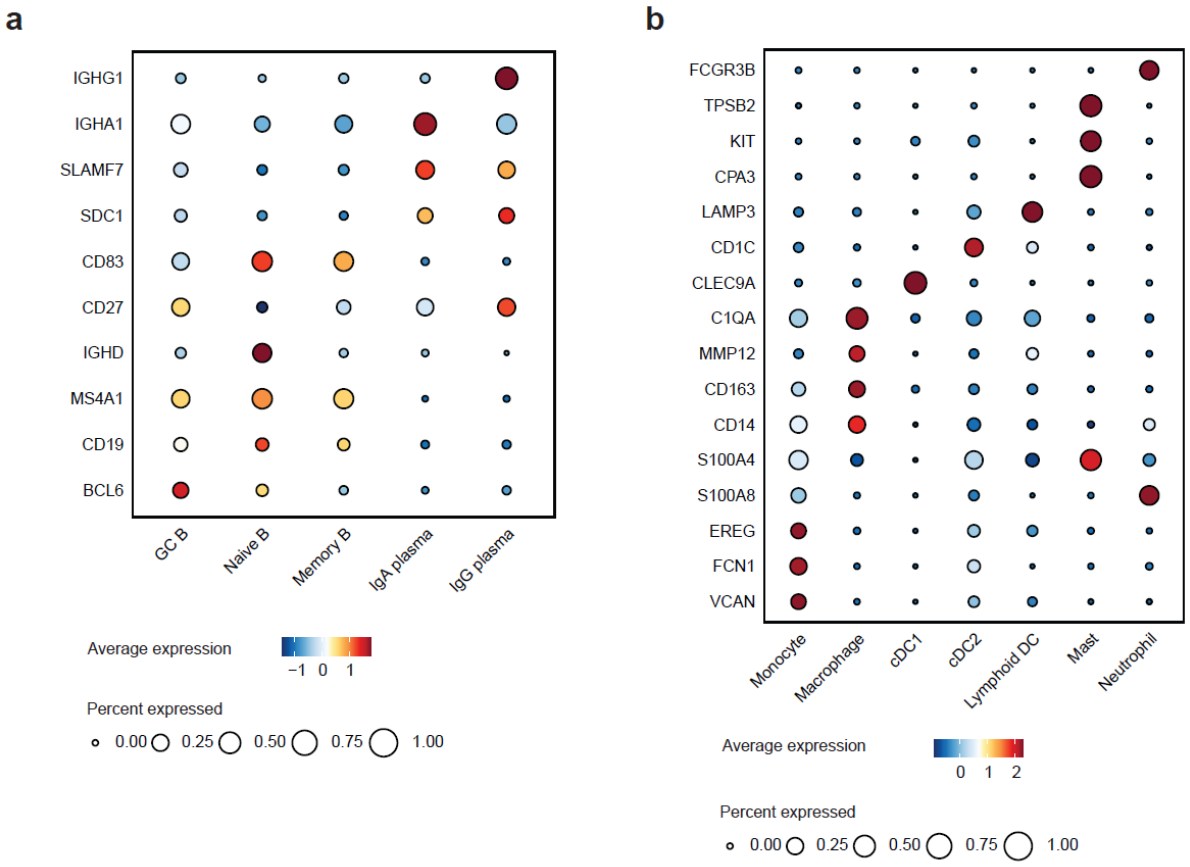


Figure S3.4 Expression of markers for different cell types in (a) B cell lineage (b) myeloid lineage.

The color shows scaled mean expression in each cell type. The dot size represents the proportion of cells expressing that marker.

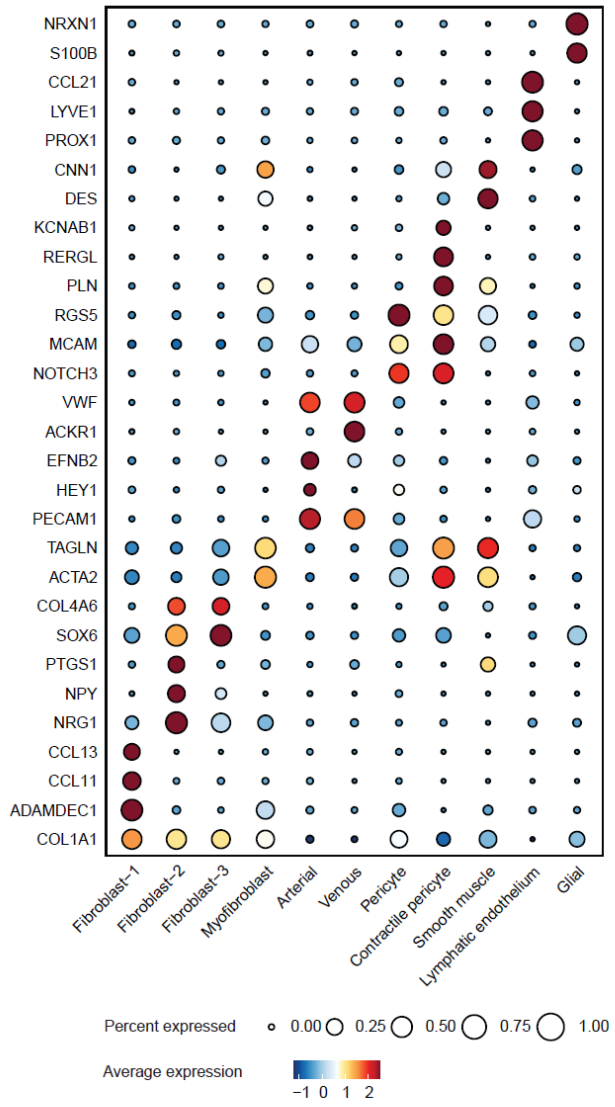


Figure S3.5 Expression of markers for different cell types in stromal cell lineage.

The color shows scaled mean expression in each cell type. The dot size represents the proportion of cells expressing that marker.

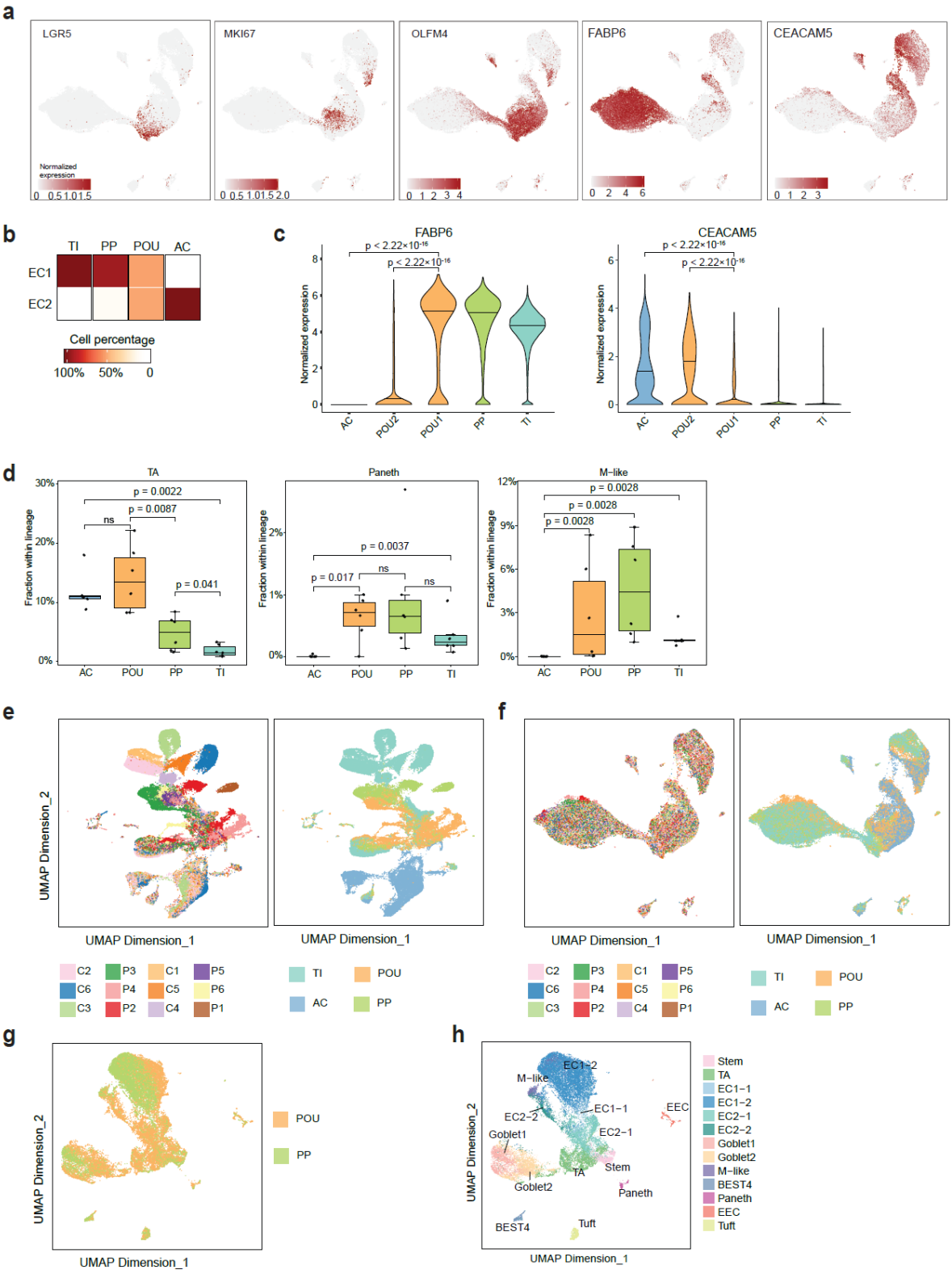


Figure S3.6 scRNA-seq data.

Figure S3.6, continued

a Epithelial UMAPs showing the expression of markers for epithelial stem cells (LGR5), transit-amplifying cells (MKI67), early stage enterocytes (OLFM4), ileum-like enterocytes (FABP6), colon-like enterocytes (CEACAM5). **b** Pouch enterocytes are split into two populations: pouch EC1 and pouch EC2. **c** Violin plots for FABP6 and CEACAM5 expression in 5 enterocyte groups. The analysis was conducted between single cells from two cell populations, with each single cell considered a biological replicate. The number of cells in each group is as follows: AC (7,072), pouch EC2 (3,523), pouch EC1 (3,581), PP (7,327), TI (17,633). Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. **d** Cell proportion in epithelial lineages. A two-sided Wilcoxon rank-sum test was performed to compare samples from different tissue regions, with each sample representing a biological replicate. "ns" indicates $p > 0.05$. **e** UMAP of all epithelial cells without batch effect removal, colored by patient ID and biopsy location, respectively. **f** UMAP of all epithelial cells after batch effect removal, colored by patient ID and biopsy location, respectively. **g** UMAP of epithelial cells from pouch and pre-pouch after batch effect removal, colored by biopsy location. **h** UMAP of epithelial cells from pouch and pre-pouch after batch effect removal, colored by cell type annotation. Source data are provided as a Source Data file.

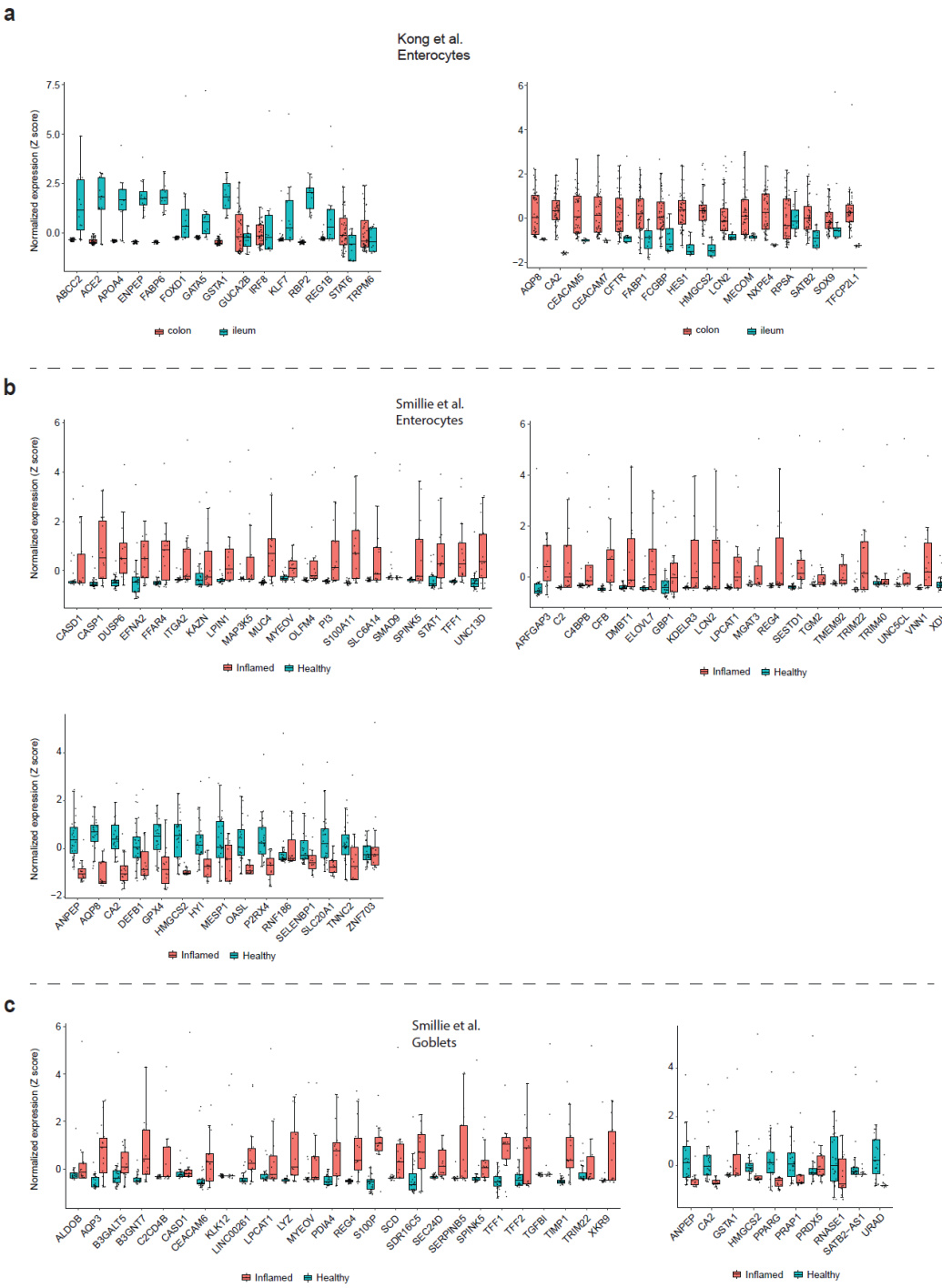


Figure S3.7 Gene expression validation.

The z-score of the average expression of each gene in (**a, b**) Enterocytes or (**c**) Goblets per sample was calculated and plotted across samples for each gene. All highlighted genes in Fig. 3.2d, Fig. 3.3a and Fig. 3.3b were plotted to demonstrate the consistency between the reference dataset **a** Kong et al.[24] or (**b, c**) Smillie et al.[27] and our study. Source data are provided as a Source Data file.

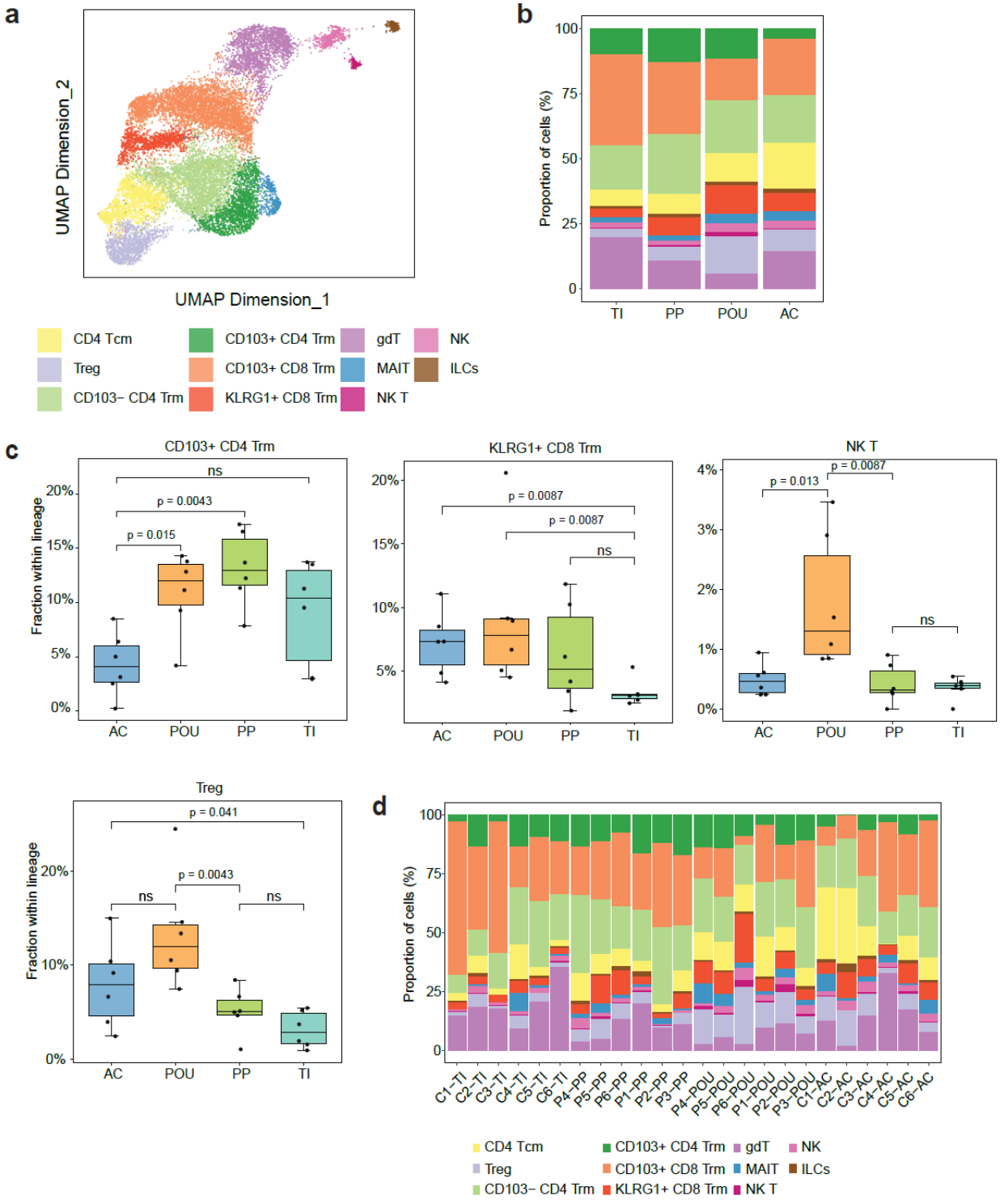


Figure S3.8 scRNA data.

Figure S3.8, continued

a UMAP shows cell types in T cell lineage. **b** The proportion of different cell types in different regions. **c** Cell types that show proportional differences across regions. A two-sided Wilcoxon rank-sum test was performed to compare samples from different tissue regions, with each sample representing a biological replicate. "ns" indicates $p > 0.05$. **d** The cell type proportions in each sample. Source data are provided as a Source Data file.

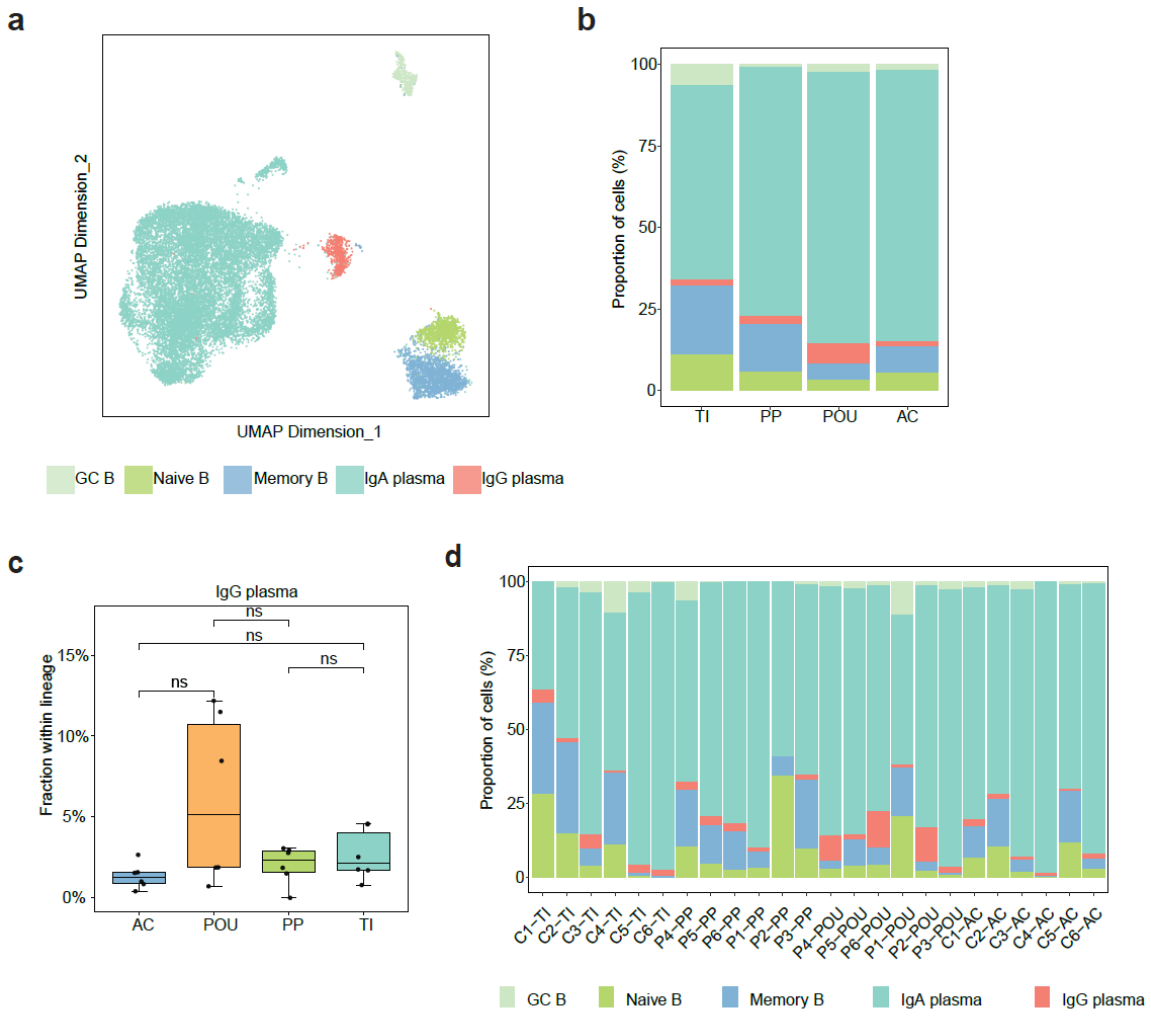


Figure S3.9 scRNA data.

a UMAP shows cell types in B cell lineage. **b** The proportion of different cell types in different regions. **c** The cell type that shows proportional differences across regions. A two-sided Wilcoxon rank-sum test was performed to compare samples from different tissue regions, with each sample representing a biological replicate. "ns" indicates $p > 0.05$. **d** The cell type proportions in each sample. Source data are provided as a Source Data file.

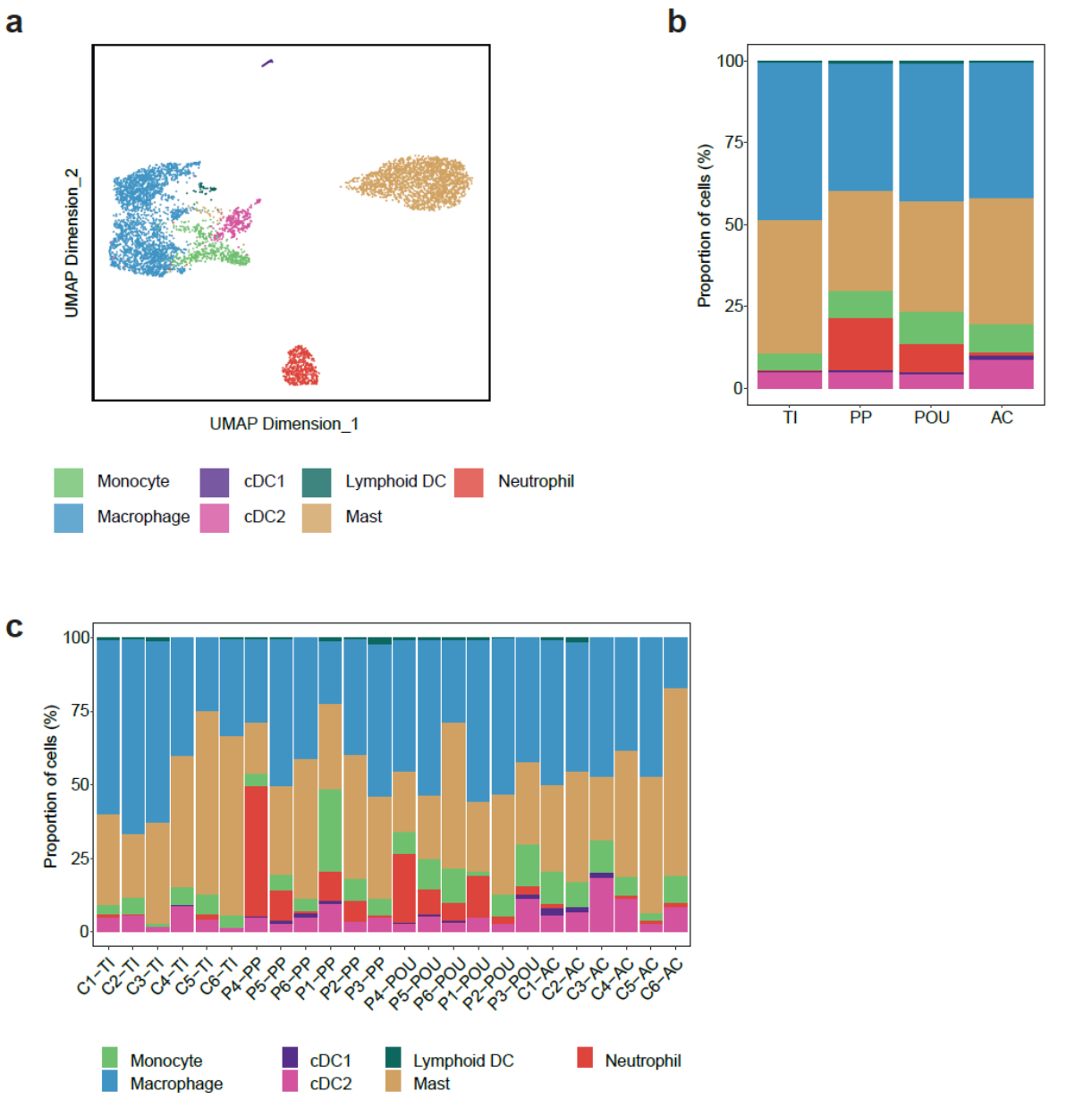


Figure S3.10 scRNA data.

a UMAP shows cell types in myeloid cell lineage. **b** The proportion of different cell types in different regions. **c** The cell type proportions in each sample. Source data are provided as a Source Data file.

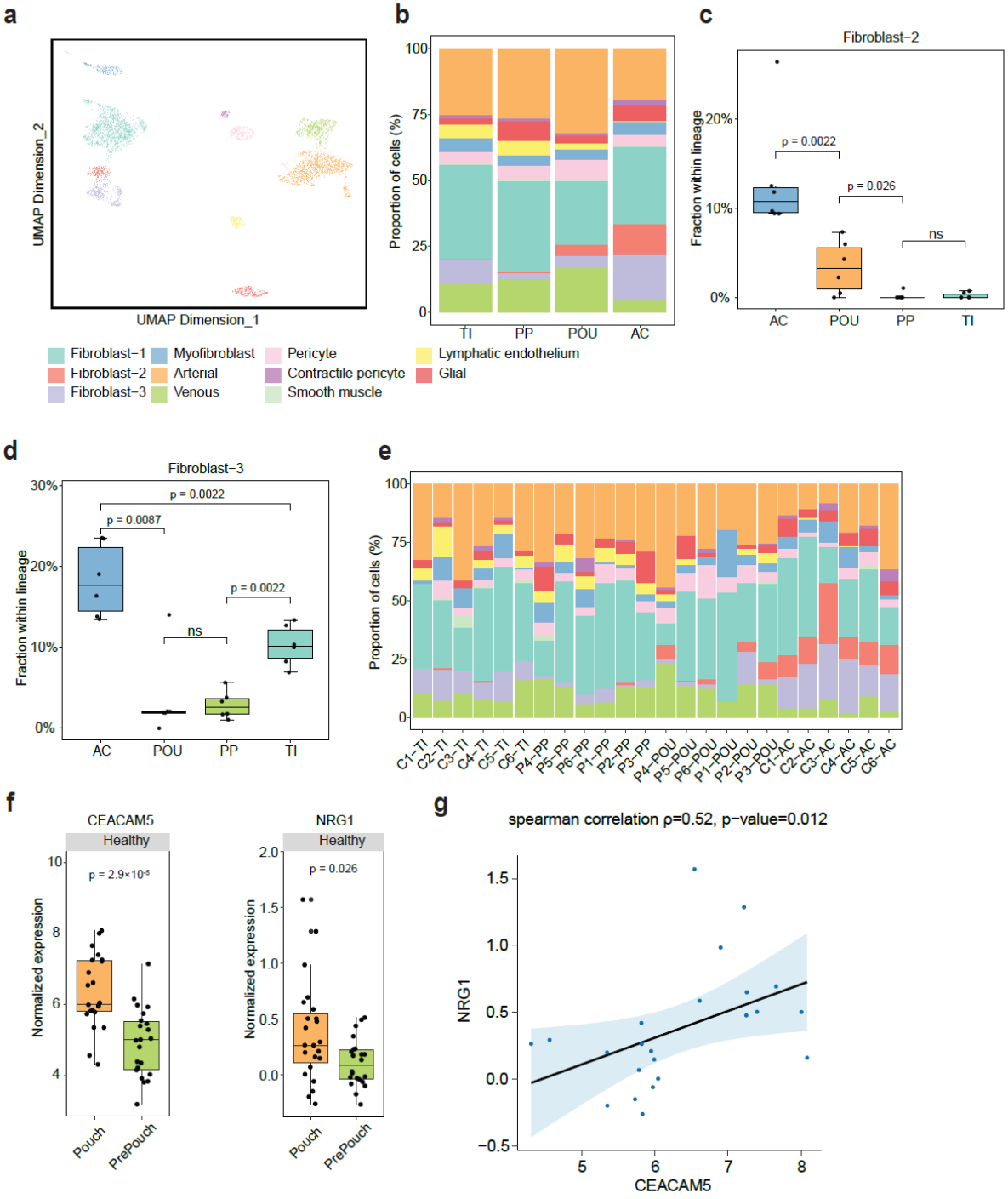


Figure S3.11 scRNA data.

Figure S3.11, continued

a UMAP shows cell types in stromal cell lineage. **b** The proportion of different cell types in different regions. **c, d** Cell types that show proportional differences across regions. A two-sided Wilcoxon rank-sum test was performed to compare samples from different tissue regions, with each sample representing a biological replicate. "ns" indicates $p > 0.05$. **e** The cell type proportions in each sample. **f** Average gene expression from bulk RNA-seq data of pouch and prepouch biopsies without overt pouchitis at the time of pouchoscopy (Huang et al.[137]). A two-sided Wilcoxon rank-sum test was used to compare gene expression between different sample groups, with each sample treated as a biological replicate. **g** A Spearman correlation analysis was performed to evaluate the relationship between *NRG1* expression and *CEACAM5* expression in pouch biopsies without overt pouchitis at the time of pouchoscopy (Huang et al.[137]). Source data are provided as a Source Data file.

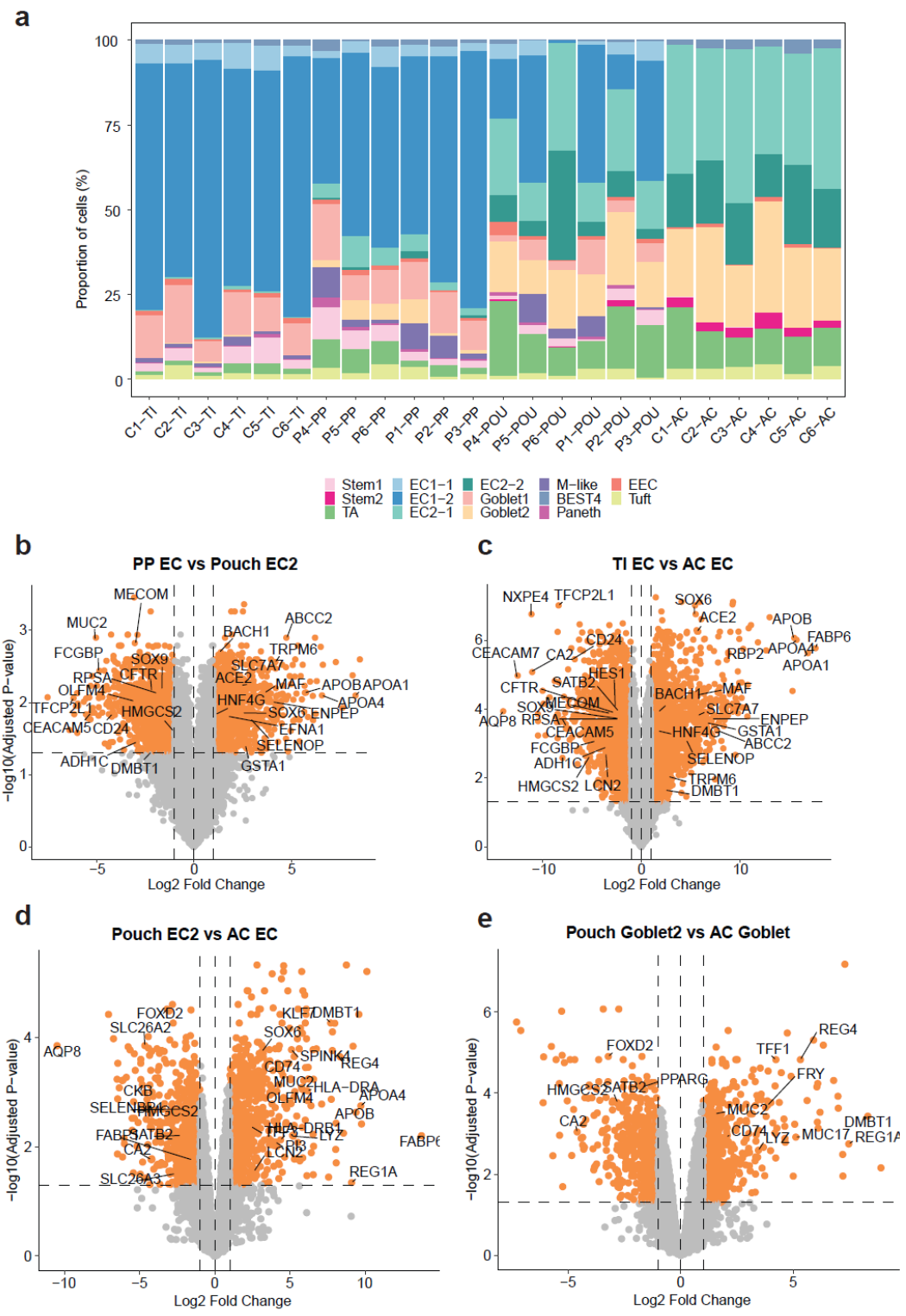


Figure S3.12 scRNA-seq data.

Figure S3.12, continued

a The proportions of epithelial cell types in each sample. Volcano plot shows the top differentially expressed genes comparing different enterocyte groups: (**b**) PP enterocytes vs pouch EC2, (**c**) TI vs AC, (**d**) pouch EC2 vs AC, (**e**) pouch Goblet2 vs AC Goblet. A linear mixed model, accounting for patient-specific effects, was used to evaluate both upregulation and downregulation (two-sided test) of genes (Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method. Source data are provided as a Source Data file.

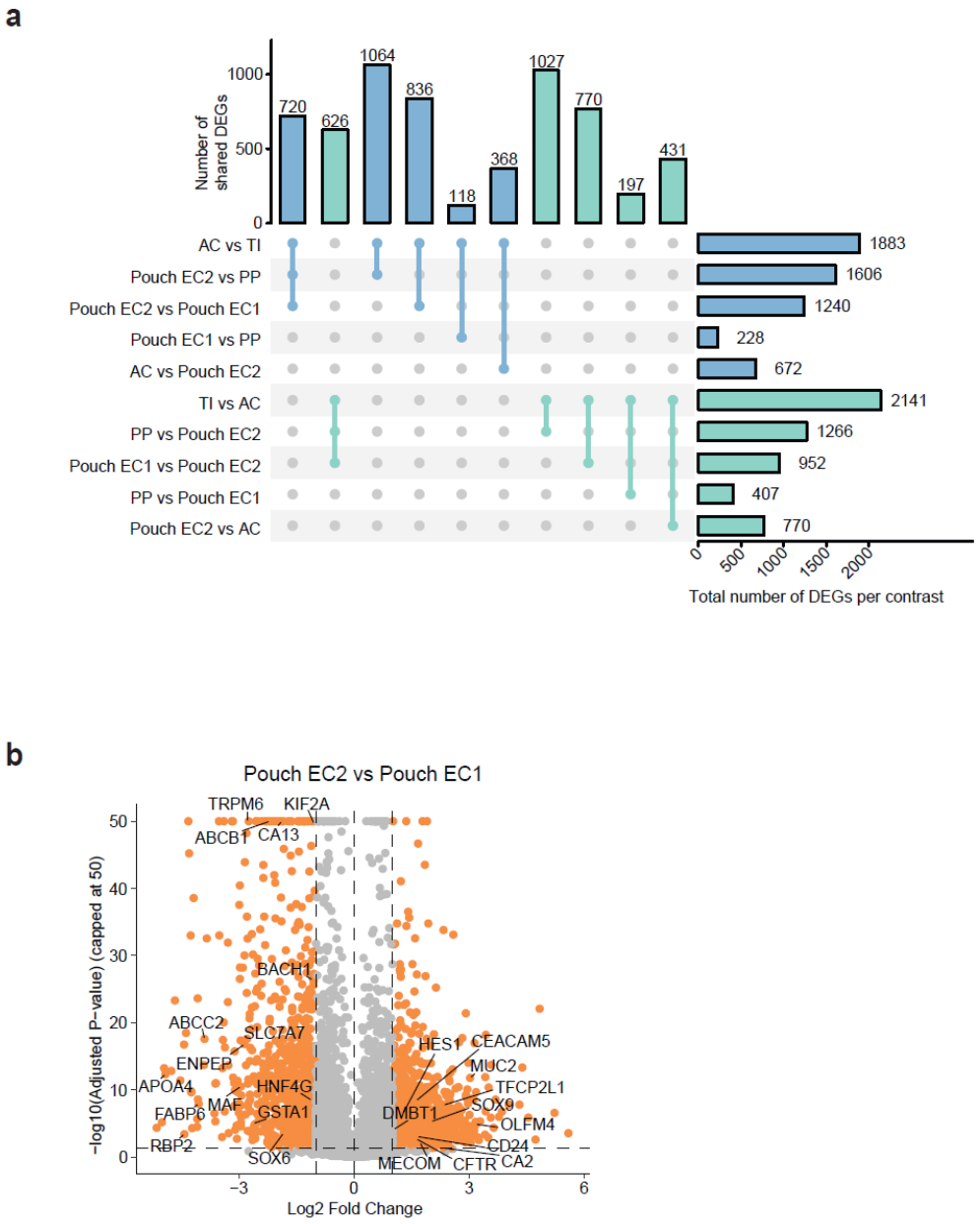


Figure S3.13 Differential expression between select epithelial cell clusters.

a Number of differentially expressed genes that are shared across different contrasts. **b** Volcano plot shows the top differentially expressed genes comparing pouch EC2 vs pouch EC1. In (**a, b**), a linear mixed model, accounting for patient-specific effects, was used to evaluate both upregulation and downregulation (two-sided test) of genes (Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

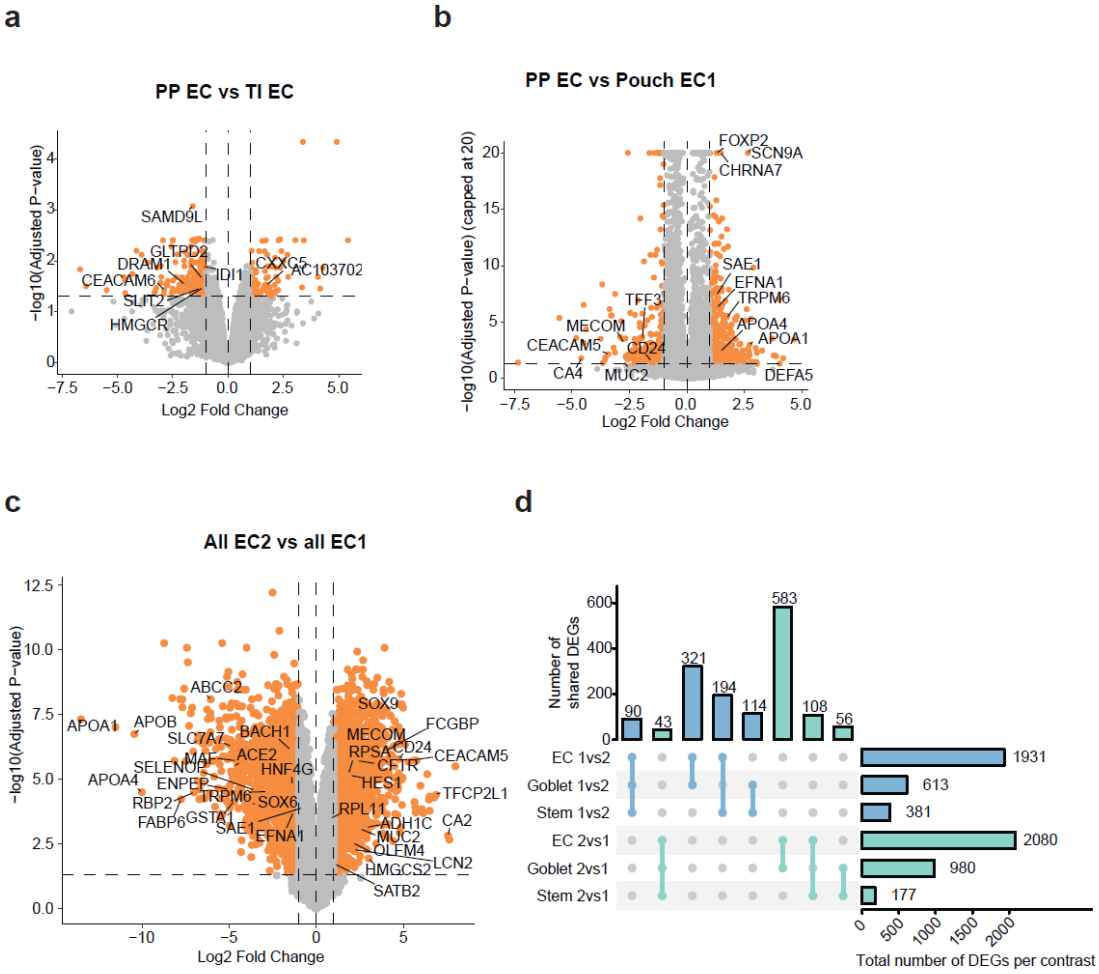


Figure S3.14 scRNA data.

Volcano plot shows the top differentially expressed genes comparing different enterocyte groups (**a**) PP vs TI, (**b**) PP vs pouch EC1. In (**a, b**), a linear mixed model, accounting for patient-specific effects, was used to evaluate both upregulation and downregulation (two sided test) of genes (Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method. (**c**) Differentially expressed genes comparing all EC2 vs EC1. **d** Number of differentially expressed genes that are shared across contrasts in different cell types. In **(c, d)**, differential expression was assessed using a two-sided Wilcoxon rank-sum test, with p-values adjusted for multiple comparisons using the Bonferroni correction.

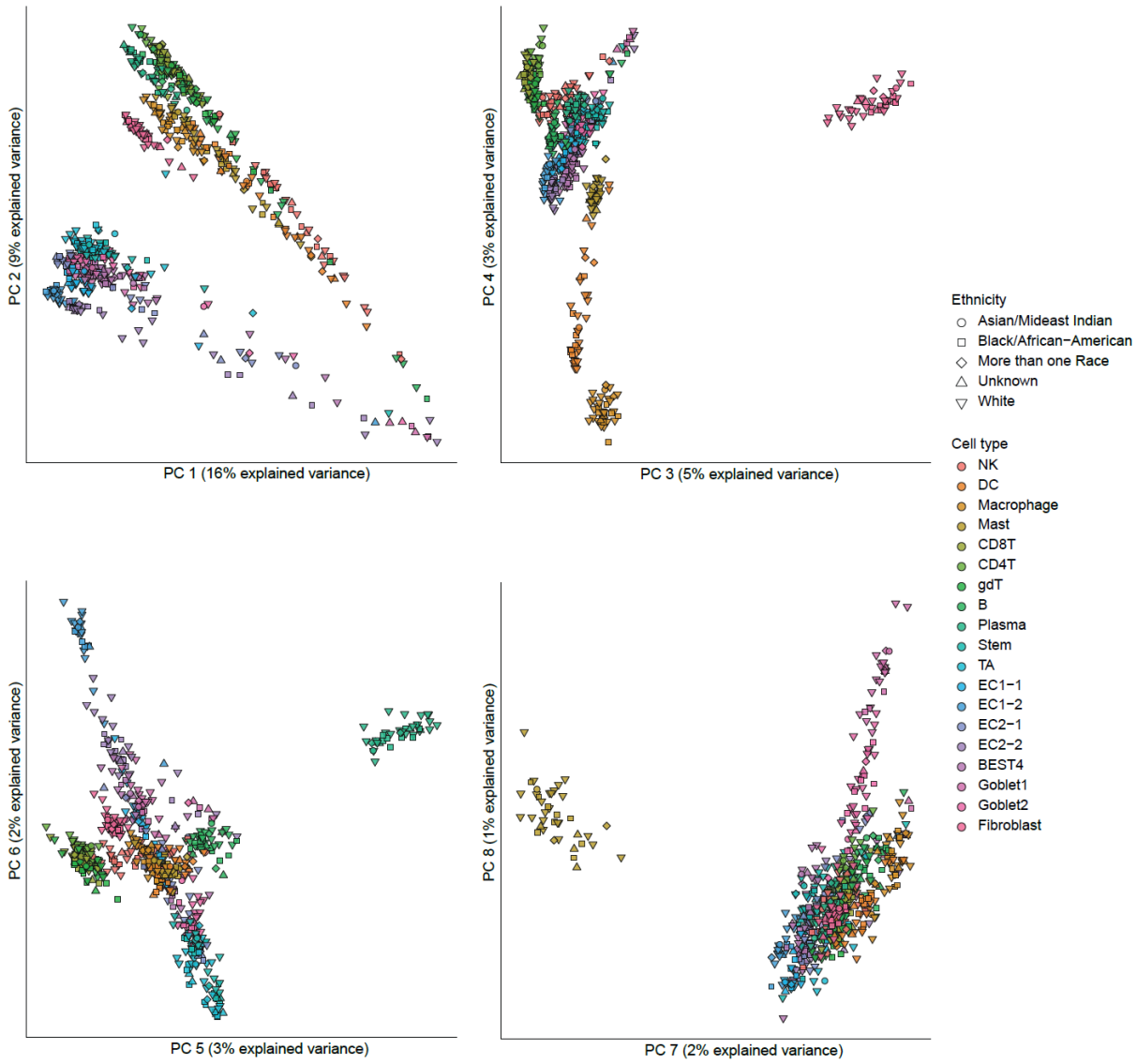


Figure S3.15 Reference validation.

A larger pool of control samples (GSE266616) were used to examine the gene expression variance across different cell type-sample pseudo bulk clusters. Top principal components (explained variance > 1%) calculated from the pseudo-bulk wise gene expression matrix were plotted to highlight the significant role that cell type differences play in genetic variance.

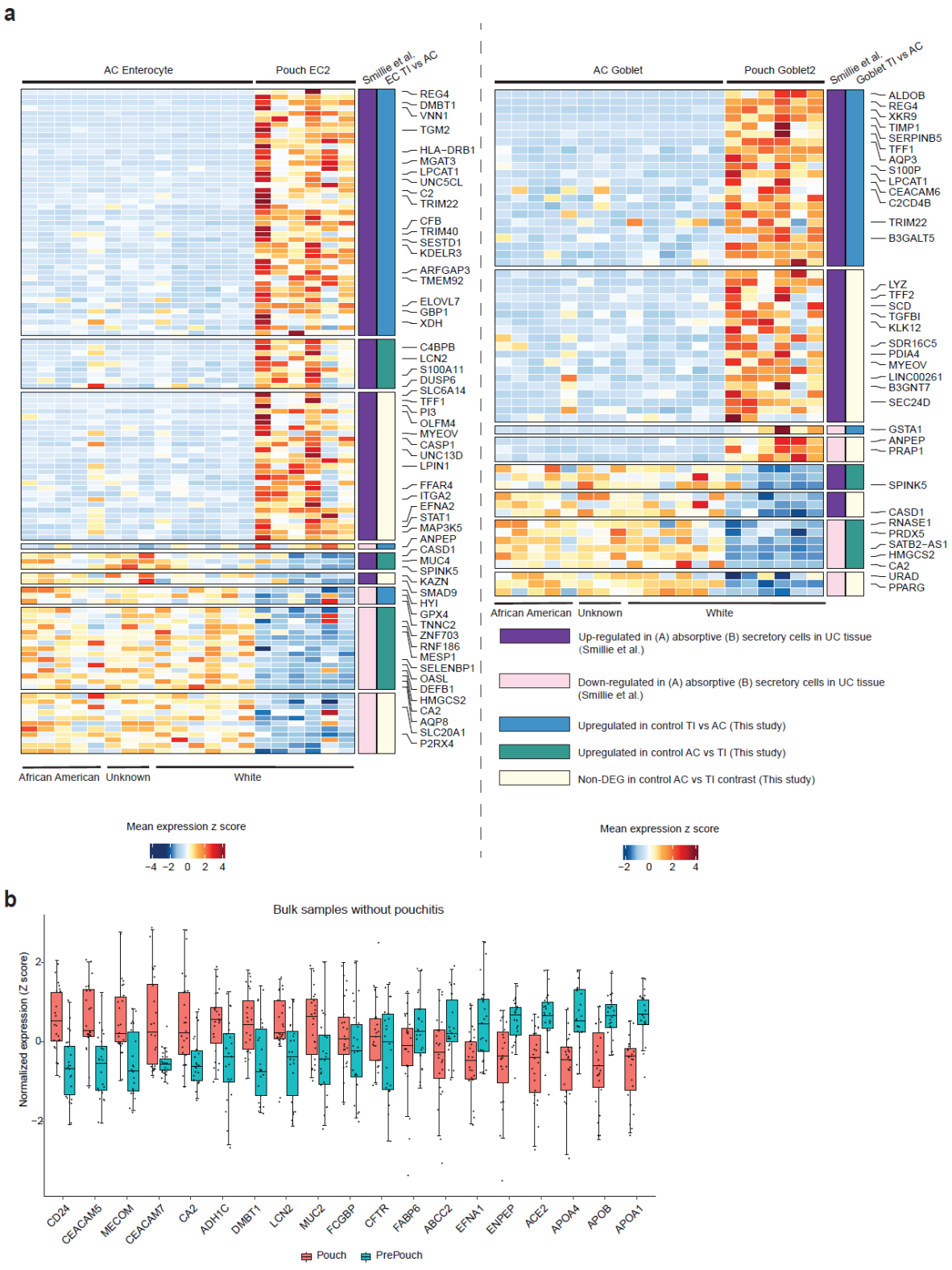


Figure S3.16 Validation.

Figure S3.16, continued

a We extended Fig. 3.3 with a larger pool of control samples to show the consistent gene expression pattern across different ethnic backgrounds. **b** Z scores of average gene expression from bulk RNA-seq data of healthy pouches (Huang et al.[137]) were plotted to confirm the significant gene expression differences between the pouch and prepouch enterocytes identified in our scRNA study. Source data are provided as a Source Data file.

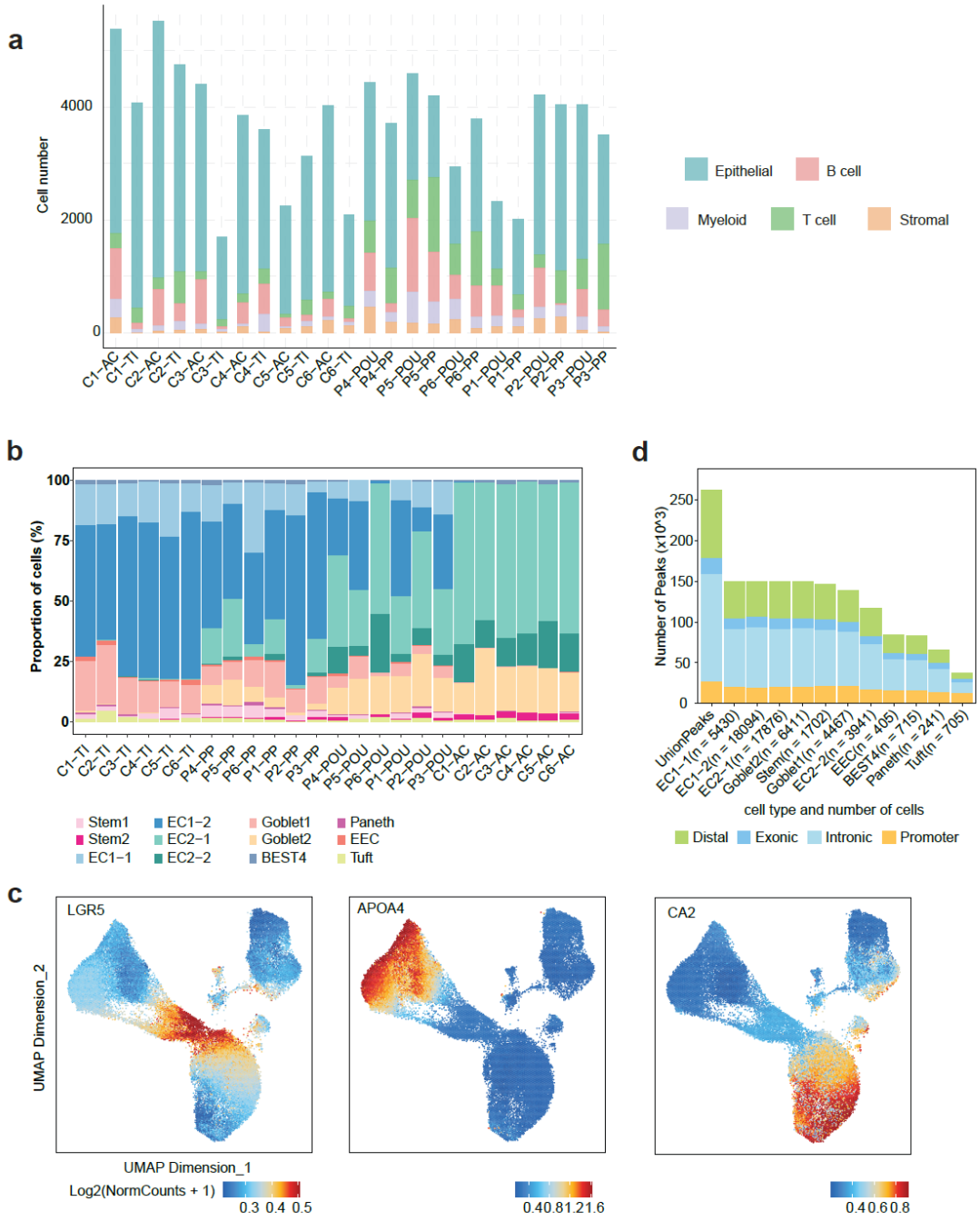


Figure S3.17 scATAC data.

a The number of cells in each biopsy sample that have been analyzed after QC in scATAC-seq analysis. **b** The epithelial cell type proportions in each sample. **c** Gene scores for marker genes of different epithelial subpopulations overlayed onto UMAP of epithelial cells. **d** The number of peaks that were initially called in each cell type. Source data are provided as a Source Data file.

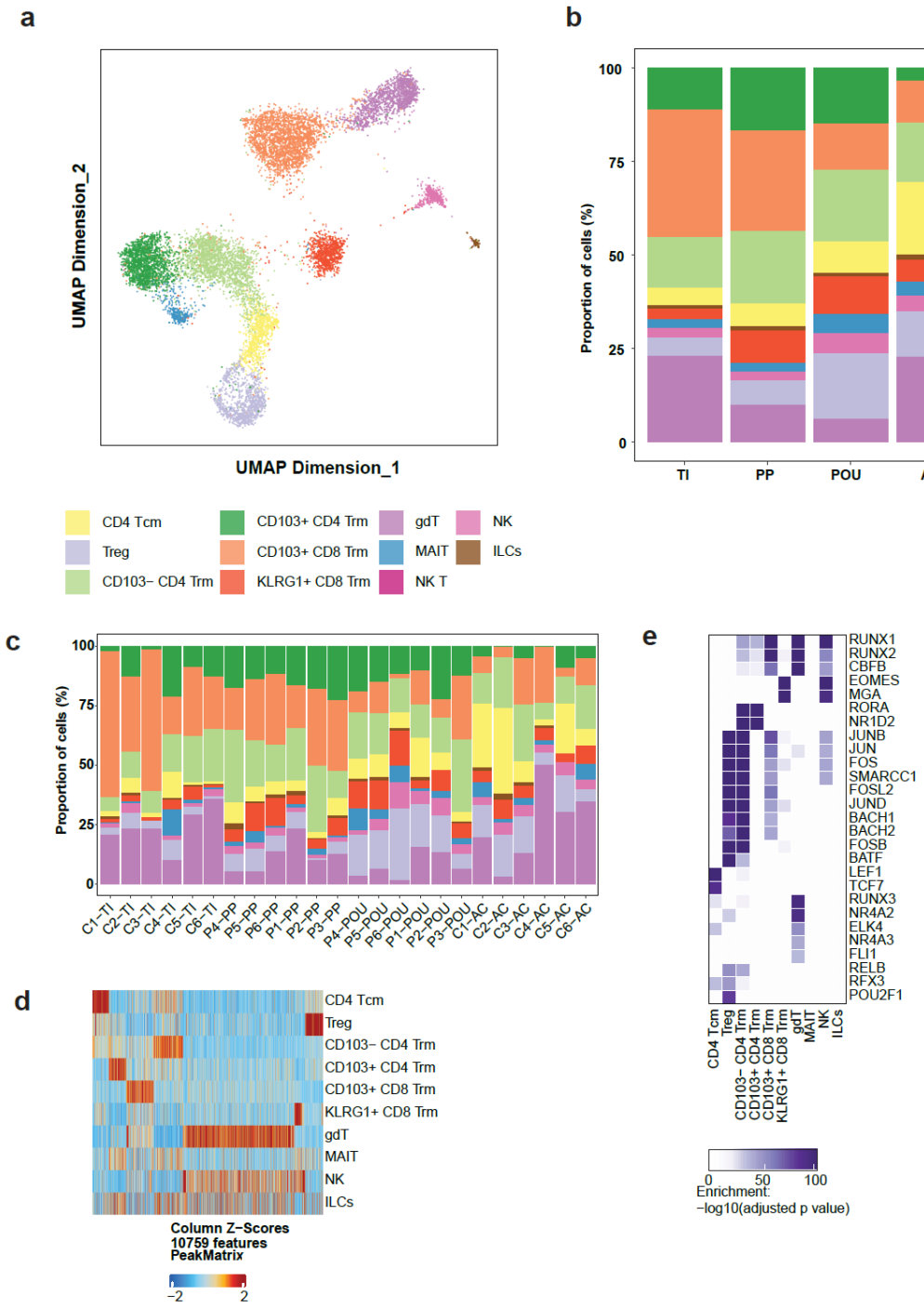


Figure S3.18 scATAC data.

a UMAP shows different cell types in T cell lineage. **b** The proportion of different cell types in different regions. **c** The cell type proportions in each sample. **d** Marker peaks for each cell type. **e** Enriched TF motifs in each cell type. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons. Source data are provided as a Source Data file.

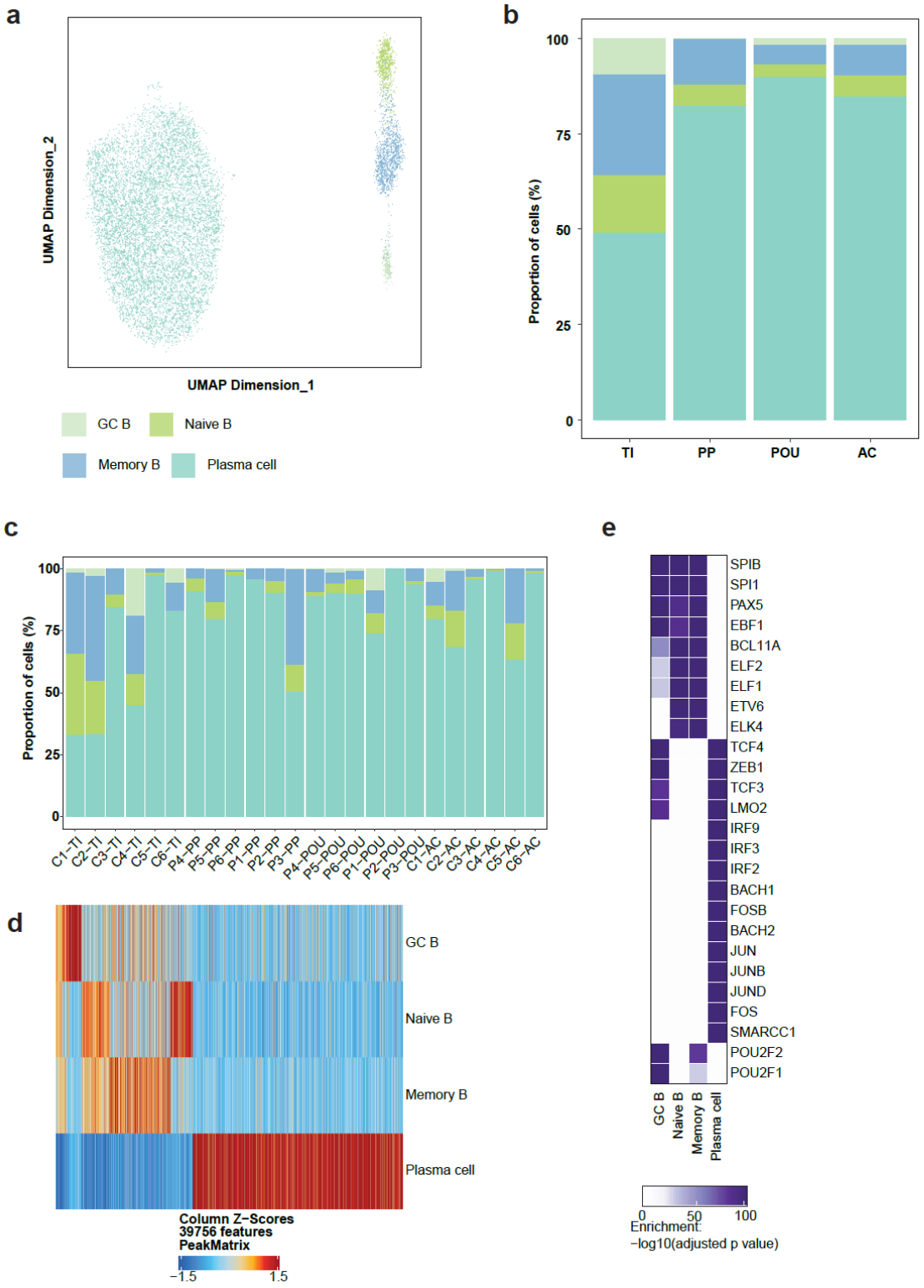


Figure S3.19 scATAC data.

Figure S3.19, continued

a UMAP shows different cell types in B cell lineage. **b** The proportion of different cell types in different regions. **c** The cell type proportions in each sample. **d** Marker peaks for each cell type. **e** Enriched TF motifs in each cell type. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons. Source data are provided as a Source Data file.

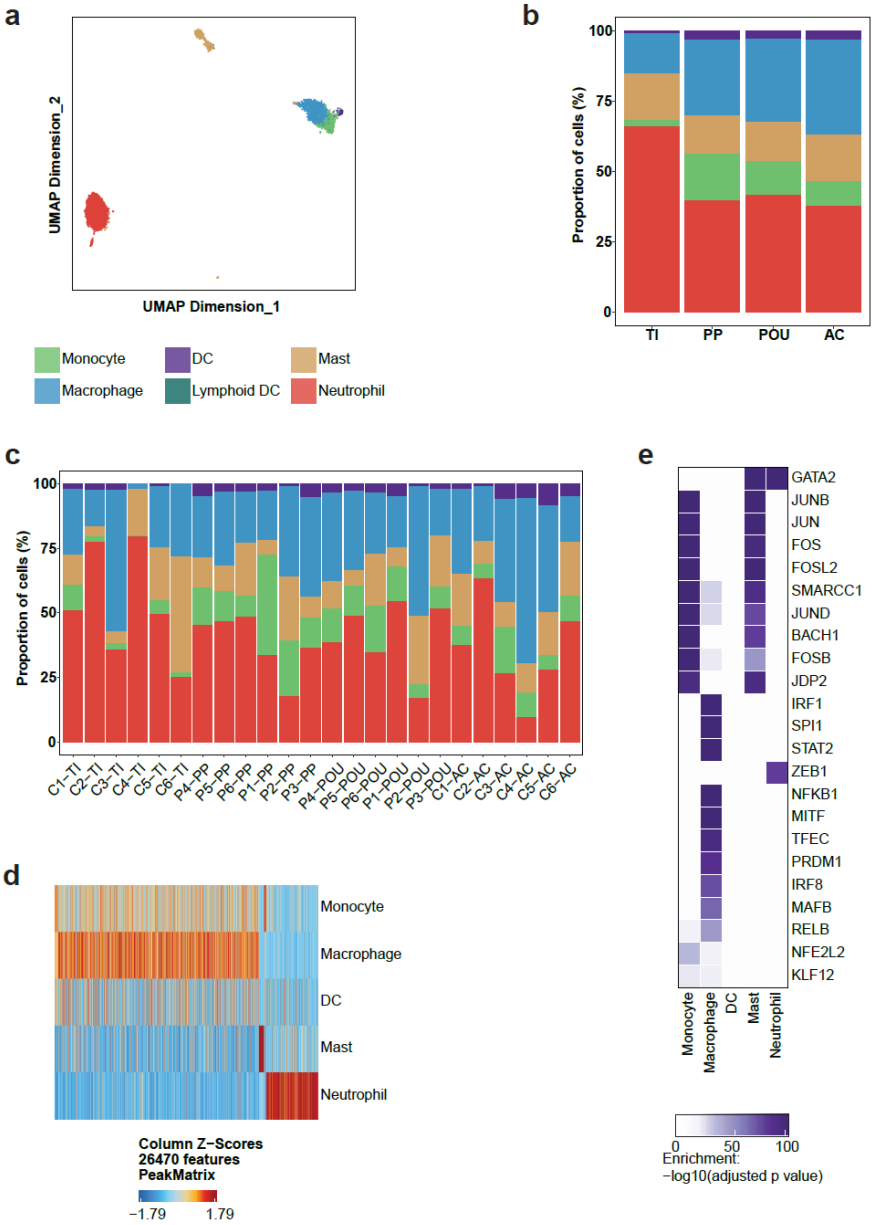


Figure S3.20 scATAC data.

a UMAP shows different cell types in myeloid cell lineage. **b** The proportion of different cell types in different regions. **c** The cell type proportions in each sample. **d** Marker peaks for each cell type. **e** Enriched TF motifs in each cell type. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons. Source data are provided as a Source Data file.

Figure S3.21 scATAC data.

a UMAP shows different cell types in stromal cell lineage. **b** The proportion of different cell types in different regions. **c** The cell type proportions in each sample. **d** Marker peaks for each cell type. **e** Enriched TF motifs in each cell type. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons. Source data are provided as a Source Data file.

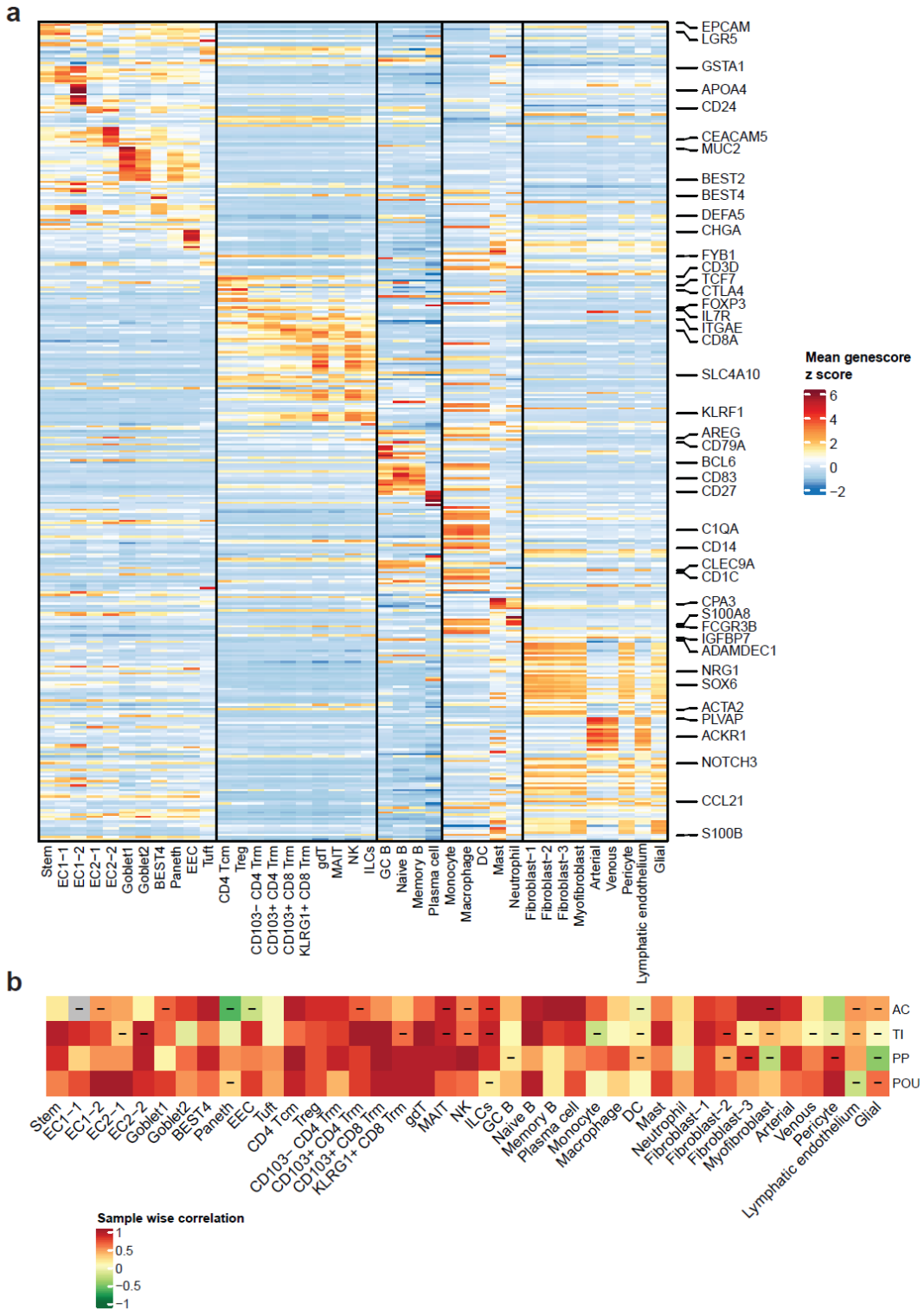


Figure S3.22 Annotation validation.

Figure S3.22, continued

a Gene score of each cell type in scATAC-seq clusters. The markers are the same group of genes as scRNA analysis. **b** Sample wise within lineage cell type proportion in each region. The cell groups with fewer than 50 cells are labeled with '-' to indicate low confidence value.

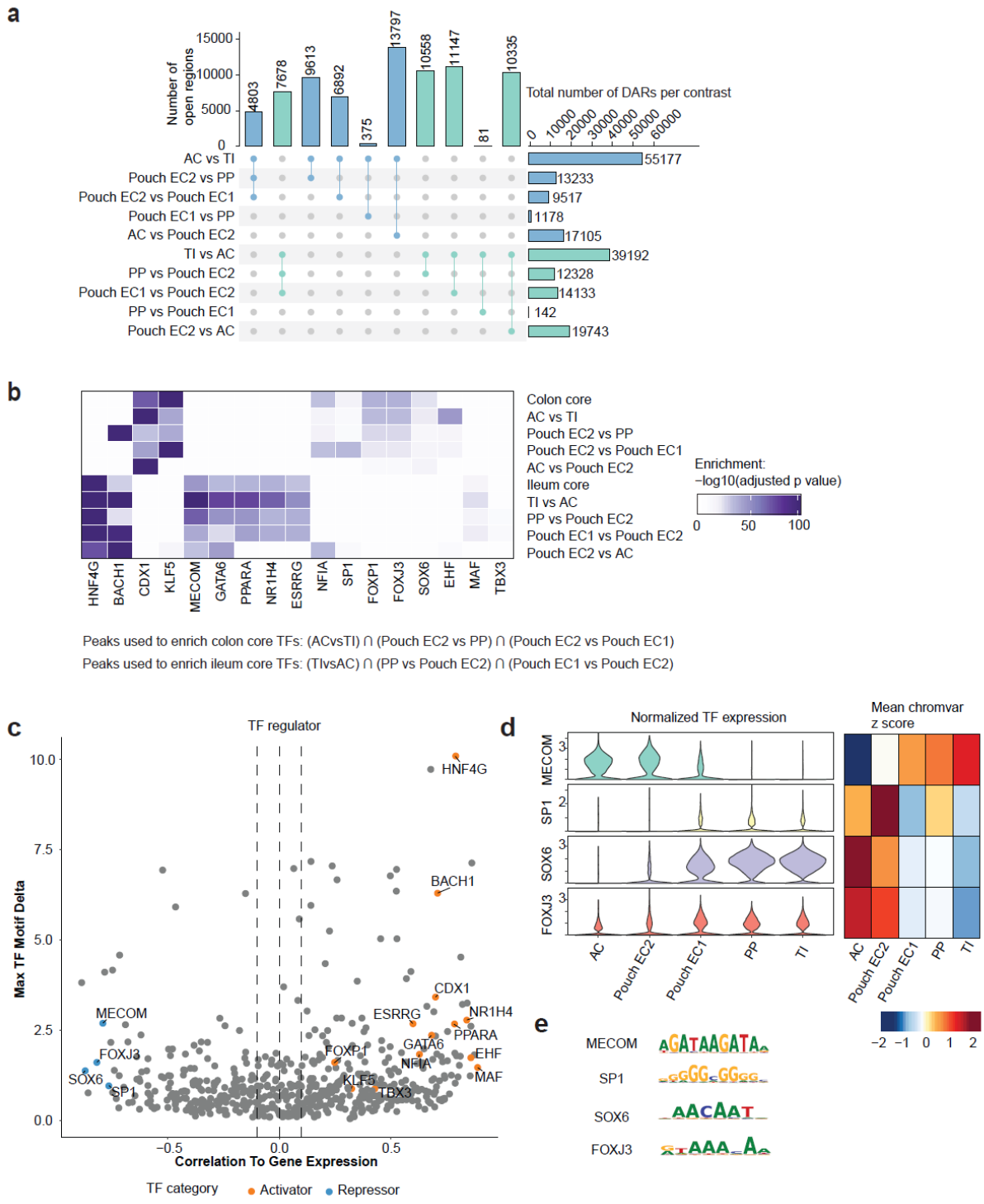


Figure S3.23 scATAC data.

Figure S3.23, continued

a Number of differentially accessible peaks that are shared across different contrasts. **b** Enriched TF motifs in each differential test. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons. **c** TFs are categorized into activators and repressors. **d** The normalized gene expression and scaled average chromVAR score in 5 enterocyte groups. **e** The logos of repressive TFs.

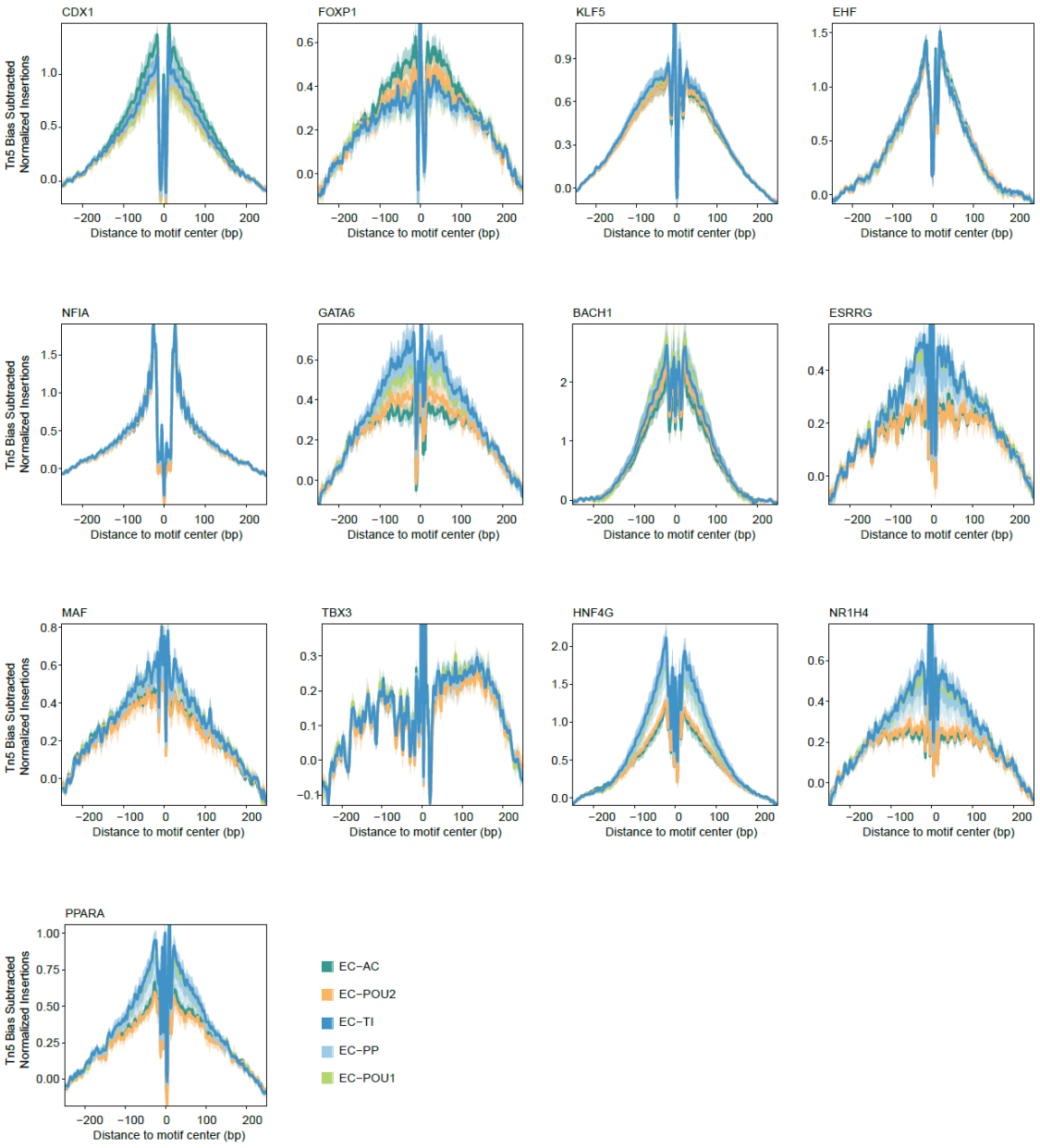


Figure S3.24 TF footprinting.

Tn5 bias-subtracted normalized transcription factor footprinting.

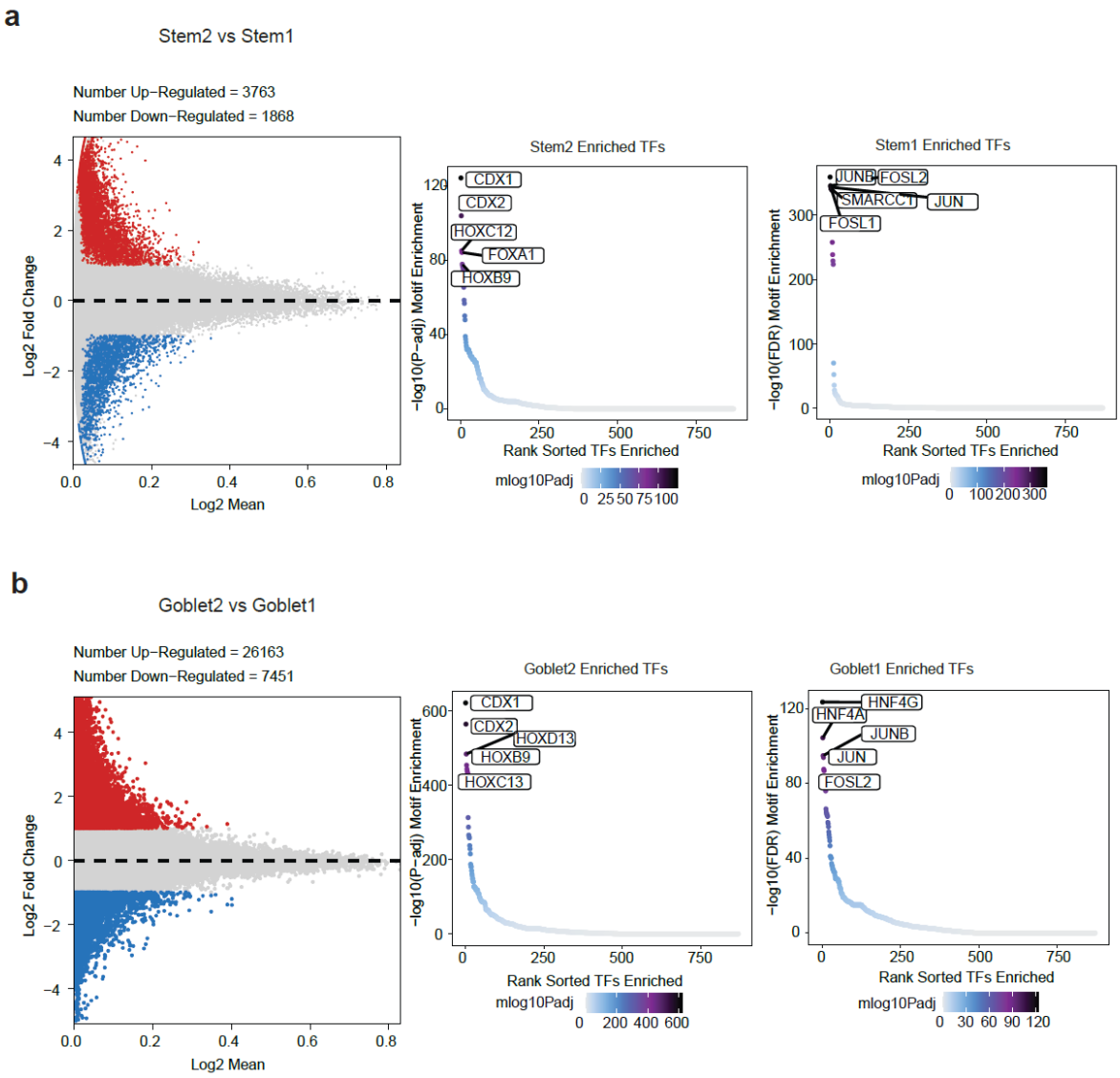


Figure S3.25 scATAC data.

a Differentially accessible peaks comparing Stem2 and Stem1 and enriched TF motifs. **b** Differentially accessible peaks comparing Goblet2 and Goblet1 and enriched TF motifs. In (**a, b**), differential accessibility was assessed using a two-sided Wilcoxon rank-sum test, with p-values adjusted for multiple comparisons using the false discovery rate (FDR) method. Enrichment was assessed using the hypergeometric test (one-sided), with p-values adjusted using the Bonferroni correction to account for multiple comparisons.

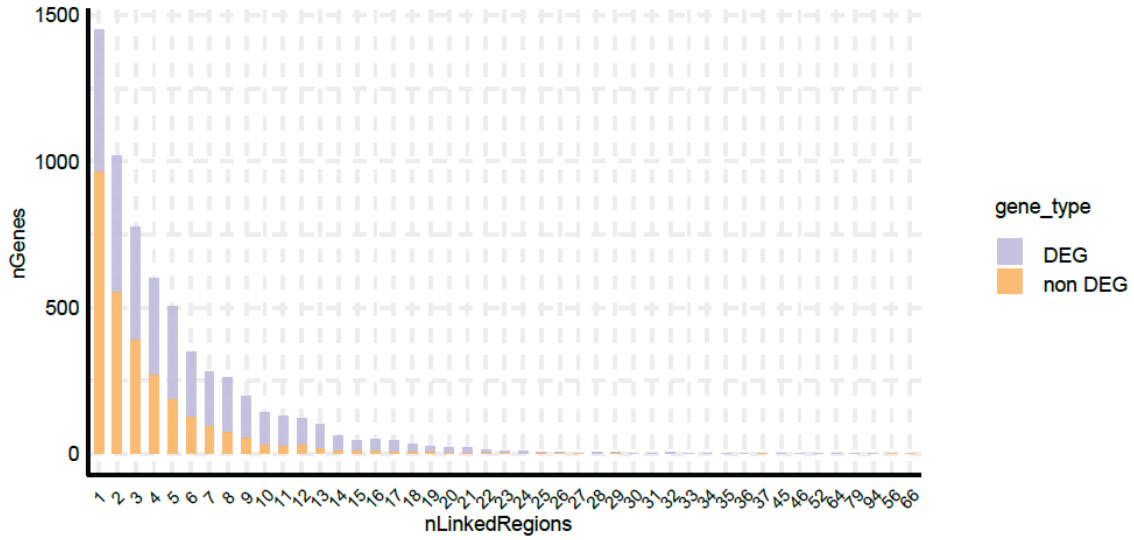


Figure S3.26 Multiomic analysis.

For genes with at least one linked peak (Methods), the number of linked peaks per gene was calculated and plotted as a distribution. Genes with a higher number of linked peaks exhibit a larger proportion of differentially expressed genes (DEGs, purple) compared to those with fewer linked peaks. A linear mixed model, accounting for patient-specific effects, was used to evaluate DEGs (two-sided test; Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method. DEGs were selected in Supplementary File 3-3 from all five comparison groups, with criteria of FDR < 0.05 and |logFC| > 1. Source data are provided as a Source Data file.

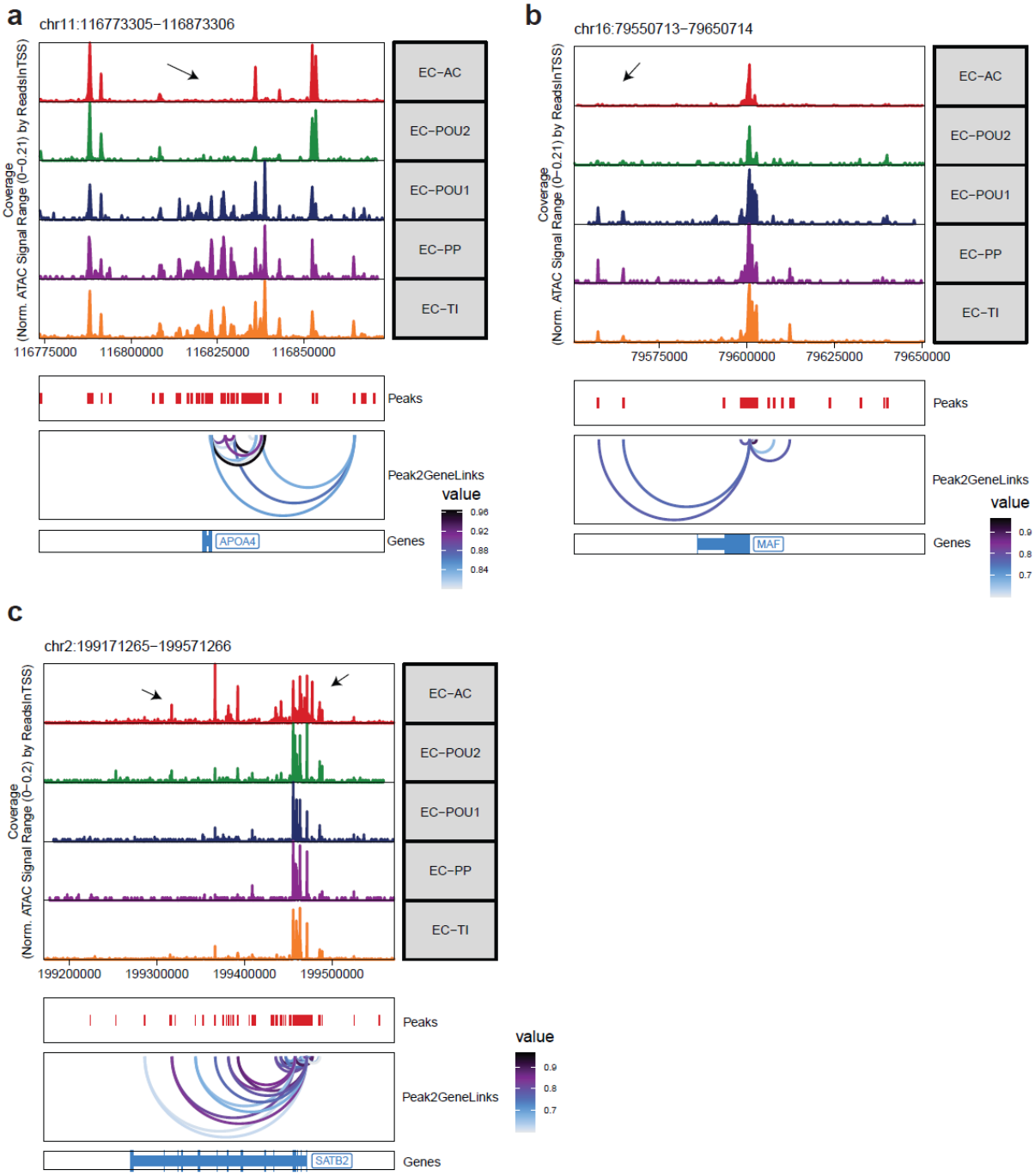


Figure S3.27 Multiomic analysis.

a Peaks highly correlated with *APOA4* expression show higher accessibility in terminal ileum, pre-pouch and pouch EC1 enterocytes. **b** Peaks highly correlated with MAF gene expression show higher accessibility in terminal ileum, pre-pouch and pouch EC1 enterocytes. **c** Peaks highly correlated with *SATB2* expression show higher accessibility in ascending colon and pouch EC2 enterocytes.

Appendix C: Supplementary Files for Chapter 2

Supplementary File 2-1: Sample metadata

Supplementary File 2-2: Peak gene co-activity

Supplementary File 2-3: Top 1000 gene scores for each topic

Supplementary File 2-4: Cell type specific iDARs

Supplementary File 2-5: Cell type specific caQTLs

Supplementary File 2-6: Prioritization of GWAS finemapped SNPs

Appendix D: Supplementary Files for Chapter 3

Supplementary File 3-1: Patient metadata.

Supplementary File 3-2: Spatial distribution of ADAMDEC1 and NRG1 transcripts in ascending colon, terminal ileum, prepouch and pouch biopsies. Color indicates the number of transcripts detected in each segmented cell, with a maximum of three transcripts to maintain a consistent color scale.

Supplementary File 3-3: Comprehensive differential gene expression analysis between different enterocyte populations. A linear mixed model, accounting for patient-specific effects, was used to evaluate both upregulation and downregulation (two sided test) of genes (Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

Supplementary File 3-4: Pathway enrichment analysis. A non-parametric rank-based Gene Set Enrichment Analysis (GSEA) two-sided test was used to evaluate enrichment (Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

Supplementary File 3-5: Differential gene expression analysis performed between pouch EC2 and AC EC, as well as between pouch Goblet2 and AC Goblet cells. A linear mixed model accounting for patient-specific effects was used to assess gene upregulation and downregulation (two-sided test; Methods). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method. Genes identified as differentially expressed (FDR < 0.05, |logFC| > 1) in this study were compared with results published in Smillie et al[27].

Supplementary File 3-6: a, b Experiments to titrate antibodies and validate specificity of labeling were performed on ileal or ascending colonic biopsies collected from 2 patients. c Immunohistochemistry images of pouch biopsies stained for FABP6, CEACAM5, and CA2, obtained from a patient different from the one shown in Fig. 3.4. d The deconvolved cell type proportions in healthy pre-pouches from bulk RNA-seq data. e The deconvolved cell type proportions in inflammatory prepouches from bulk RNA-seq data.

Supplementary File 3-7-1: a Spatial domains in pouch and prepouch replicates from the same donor. b Heatmap to show the marker gene expression of each spatial domain. Each column represents a single cell.

Supplementary File 3-7-2: a, b Differential gene expression analysis between pouch and prepouch in epithelial domains. Positive fold change indicates gene up-regulation in epithelial domains of

pouch samples. Differential expression was assessed using a two-sided Wilcoxon rank-sum test, with p-values adjusted for multiple comparisons using the Bonferroni correction.

Supplementary File 3-7-3: Spatial distribution of CEACAM5, CD24, CA2, FABP6 and APOA4 transcripts in ascending colon, terminal ileum, prepouch and pouch biopsies. Color indicates the number of transcripts detected in each segmented cell, with a maximum of three transcripts to maintain a consistent color scale.

Supplementary File 3-7-4: Spatial distribution of CEACAM5, CD24, CA2, FABP6 and APOA4 transcripts in prepouch and pouch biopsies. Color indicates the number of transcripts detected in each segmented cell, with a maximum of three transcripts to maintain a consistent color scale.

Supplementary File 3-8-1: Spatial distribution of CDX1, FOXP1, KLF5, EHF, NFIA and BACH1 transcripts in prepouch and pouch biopsies.

Supplementary File 3-8-2: Spatial distribution of ESRRG, MAF, TBX3, HNF4G, NR1H4 and PPARA transcripts in prepouch and pouch biopsies.

Supplementary File 3-9-1: Multiomic analysis. a UMAP of enterocytes and stem cells in pouch. b Pseudotime ordering of cells along differentiation from stem cells to EC1 cells. c Pseudotime ordering of cells along differentiation from stem cells to EC2 cells. (d, e) Gene expression changes along pseudotime in each lineage.

Supplementary File 3-9-2: Multiomic analysis. a UMAP of enterocytes and stem cells in AC and TI. b Pseudotime ordering of cells along differentiation from stem cells to EC1 cells. c Pseudotime ordering of cells along differentiation from stem cells to EC2 cells. d In AC and TI, the TF activity changes in EC1 lineage and EC2 lineage along simulated differentiation pseudotime. e Z scores of average gene expression from bulk RNA-seq data of healthy pouches (Huang et al.[137]) were plotted to confirm the transcription factor expression differences between the pouch and prepouch enterocytes identified in our study.

Supplementary File 3-10: Domain of regulatory chromatin (DORC) gene analysis.

Supplementary File 3-11: Additional probes for Xenium experiments.

Supplementary File 3-12: Accession codes for the raw data of each sample.

Bibliography

1. Zhang, Y.-Z. & Li, Y.-Y. Inflammatory bowel disease: Pathogenesis. *World J. Gastroenterol.* **20**, 91–99 (2014).

2. Kaplan, G. G. & Windsor, J. W. The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 56–66 (2021).

3. Ungaro, R., Mehandru, S., Allen, P. B., Peyrin-Biroulet, L. & Colombel, J.-F. Ulcerative colitis. *The Lancet* **389**, 1756–1770 (2017).

4. Kobayashi, T. *et al.* Ulcerative colitis. *Nat. Rev. Dis. Primer* **6**, 1–20 (2020).

5. Torres, J., Mehandru, S., Colombel, J.-F. & Peyrin-Biroulet, L. Crohn's disease. *The Lancet* **389**, 1741–1755 (2017).

6. Liu, J., Di, B. & Xu, L. Recent advances in the treatment of IBD: Targets, mechanisms and related therapies. *Cytokine Growth Factor Rev.* **71–72**, 1–12 (2023).

7. Higashiyama, M. & Hokari, R. New and Emerging Treatments for Inflammatory Bowel Disease. *Digestion* **104**, 74–81 (2022).

8. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

9. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

10. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

11. Liu, Z. *et al.* Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. *Nat. Genet.* **55**, 796–806 (2023).

12. Mirkov, M. U., Verstockt, B. & Cleynen, I. Genetics of inflammatory bowel disease: beyond *NOD2*. *Lancet Gastroenterol. Hepatol.* **2**, 224–234 (2017).

13. Stocchi, L. & Pemberton, J. H. Pouch and pouchitis. *Gastroenterol. Clin. North Am.* **30**, 223–241 (2001).

14. Hoda, K. M., Collins, J. F., Knigge, K. L. & Deveney, K. E. Predictors of pouchitis after ileal pouch-anal anastomosis: a retrospective review. *Dis. Colon Rectum* **51**, 554–560 (2008).

15. Penna, C. *et al.* Pouchitis after ileal pouch-anal anastomosis for ulcerative colitis occurs with increased frequency in patients with associated primary sclerosing cholangitis. *Gut* **38**, 234–239 (1996).

16. Landy, J. *et al.* Etiology of pouchitis. *Inflamm. Bowel Dis.* **18**, 1146–1155 (2012).

17. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

18. Do, C. *et al.* Genetic–epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biol.* **18**, 120 (2017).

19. Aguet, F. *et al.* Molecular quantitative trait loci. *Nat. Rev. Methods Primer* **3**, 1–22 (2023).

20. Akiyama, M. Multi-omics study for interpretation of genome-wide association study. *J. Hum. Genet.* **66**, 3–10 (2021).

21. Gaulton, K. J., Preissl, S. & Ren, B. Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 516–534 (2023).

22. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, (2020).

23. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58**, 610–620 (2015).

24. 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* **56**, 444-458.e5 (2023).

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

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

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

28. Mayer, A. T. *et al.* A tissue atlas of ulcerative colitis revealing evidence of sex-dependent differences in disease-driving inflammatory cell types and resistance to TNF inhibitor therapy. *Sci. Adv.* **9**, eadd1166 (2023).

29. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* 1–23 (2022) doi:10.1038/s41576-022-00509-1.

30. Lalanne, J.-B. *et al.* Multiplex profiling of developmental cis-regulatory elements with quantitative single-cell expression reporters. *Nat. Methods* **21**, 983–993 (2024).

31. Pliner, H. A. *et al.* Cicero Predicts *cis*-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).

32. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

33. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

34. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).

35. Greenwald, W. W. *et al.* Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.* **10**, 2078 (2019).

36. Ober-Reynolds, B. *et al.* Integrated single-cell chromatin and transcriptomic analyses of human scalp identify gene-regulatory programs and critical cell types for hair and skin diseases. *Nat. Genet.* 1–13 (2023) doi:10.1038/s41588-023-01445-4.

37. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e19 (2021).

38. Carbonetto, P. *et al.* GoM DE: interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. *Genome Biol.* **24**, 236 (2023).

39. Carbonetto, P., Sarkar, A., Wang, Z. & Stephens, M. Non-negative matrix factorization algorithms greatly improve topic model fits. Preprint at http://arxiv.org/abs/2105.13440 (2022).

40. Zhao, Y. *et al.* Cell-type-resolved chromatin accessibility in the human intestine identifies complex regulatory programs and clarifies genetic associations in Crohn's disease. 2024.12.10.24318718 Preprint at https://doi.org/10.1101/2024.12.10.24318718 (2024).

41. Baumgart, D. C. & Carding, S. R. Inflammatory bowel disease: cause and immunobiology. *The Lancet* **369**, 1627–1640 (2007).

42. Roda, G. *et al.* Crohn's disease. *Nat. Rev. Dis. Primer* **6**, 1–19 (2020).

43. Feuerstein, J. D. & Cheifetz, A. S. Crohn Disease: Epidemiology, Diagnosis, and Management. *Mayo Clin. Proc.* **92**, 1088–1103 (2017).

44. Imhann, F. *et al.* Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* **67**, 108–119 (2018).

45. Choy, M. C., Visvanathan, K. & De Cruz, P. An Overview of the Innate and Adaptive Immune System in Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **23**, 2–13 (2017).

46. Akdis, C. A. Does the epithelial barrier hypothesis explain the increase in allergy, autoimmunity and other chronic conditions? *Nat. Rev. Immunol.* **21**, 739–751 (2021).

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

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

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

50. Li, J. *et al.* Pathogenesis of fibrostenosing Crohn's disease. *Transl. Res.* **209**, 39–54 (2019).

51. Duerr, R. H. *et al.* A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* **314**, 1461–1463 (2006).

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

53. Gaulton, K. J., Preissl, S. & Ren, B. Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 516–534 (2023).

54. Zilbauer, M. *et al.* A Roadmap for the Human Gut Cell Atlas. *Nat. Rev. Gastroenterol. Hepatol.* (2023) doi:10.1038/s41575-023-00784-1.

55. Selewa, A. *et al.* Single-cell genomics improves the discovery of risk variants and genes of atrial fibrillation. *Nat. Commun.* **14**, 4999 (2023).

56. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat. Genet.* 1–13 (2022) doi:10.1038/s41588-022-01069-0.

57. Benaglio, P. *et al.* Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus ATAC-seq in peripheral blood. *PLOS Genet.* **19**, e1010759 (2023).

58. Hickey, J. W. *et al.* Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 (2023).

59. Wayman, J. A. *et al. Accessible Chromatin Maps of Inflammatory Bowel Disease Intestine Nominate Cell-Type Mediators of Genetic Disease Risk*. http://biorxiv.org/lookup/doi/10.1101/2024.02.09.579678 (2024) doi:10.1101/2024.02.09.579678.

60. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).

61. Zhao, Y. *et al.* Multiomic analysis reveals cellular and epigenetic plasticity in intestinal pouches of ulcerative colitis patients. *medRxiv* 2023.11.11.23298309 (2023) doi:10.1101/2023.11.11.23298309.

62. Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810-826.e23 (2021).

63. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

64. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

65. Dey, K. K., Hsiao, C. J. & Stephens, M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genet.* **13**, e1006599 (2017).

66. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

67. Cromer, W. E., Mathis, J. M., Granger, D. N., Chaitanya, G. V. & Alexander, J. S. Role of the endothelium in inflammatory bowel diseases. *World J. Gastroenterol.* **17**, 578–593 (2011).

68. Pearce, E. L. *et al.* Control of Effector CD8+ T Cell Function by the Transcription Factor Eomesodermin. *Science* **302**, 1041–1043 (2003).

69. Hodson, D. J. *et al.* Regulation of normal B-cell differentiation and malignant B-cell survival by OCT2. *Proc. Natl. Acad. Sci.* **113**, E2039–E2046 (2016).

70. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

71. Currin, K. W. *et al.* Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am. J. Hum. Genet.* **108**, 1169–1189 (2021).

72. Kribelbauer-Swietek, J. F. *et al.* Context transcription factors establish cooperative environments and mediate enhancer communication. *Nat. Genet.* **56**, 2199–2212 (2024).

73. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

74. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).

75. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

76. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

77. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

78. Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).

79. Bradfield, J. P. *et al.* A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. *PLOS Genet.* **7**, e1002293 (2011).

80. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).

81. Rawlings, D. J., Metzler, G., Wray-Dutra, M. & Jackson, S. W. Altered B cell signalling in autoimmunity. *Nat. Rev. Immunol.* **17**, 421–436 (2017).

82. Edwards, J. C. W. & Cambridge, G. B-cell targeting in rheumatoid arthritis and other autoimmune diseases. *Nat. Rev. Immunol.* **6**, 394–403 (2006).

83. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).

84. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

85. Ellinghaus, D., Bethune, J., Petersen, B.-S. & Franke, A. The genetics of Crohn's disease and ulcerative colitis – status quo and beyond. *Scand. J. Gastroenterol.* **50**, 13–23 (2015).

86. Geremia, A., Biancheri, P., Allan, P., Corazza, G. R. & Di Sabatino, A. Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun. Rev.* **13**, 3–10 (2014).

87. Yamazaki, K. *et al.* Association analysis of SLC22A4, SLC22A5 and DLG5 in Japanese patients with Crohn disease. *J. Hum. Genet.* **49**, 664–668 (2004).

88. Del Chierico, F. *et al.* Solute Transporter OCTN1/Slc22a4 Affects Disease Severity and Response to Infliximab in Experimental Colitis: Role of Gut Microbiota and Immune Modulation. *Inflamm. Bowel Dis.* izae135 (2024) doi:10.1093/ibd/izae135.

89. Sæterstad, S. *et al.* Profound gene expression changes in the epithelial monolayer of active ulcerative colitis and Crohn's disease. *PLOS ONE* **17**, e0265189 (2022).

90. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

91. Kim, S. S. *et al.* Leveraging single-cell ATAC-seq and RNA-seq to identify disease-critical fetal and adult brain cell types. *Nat. Commun.* **15**, 563 (2024).

92. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

93. Goldberg, R. *et al.* A Crohn's Disease-associated IL2RA Enhancer Variant Determines the Balance of T Cell Immunity by Regulating Responsiveness to IL-2 Signalling. *J. Crohns Colitis* **15**, 2054–2065 (2021).

94. Singer, D. S. & Devaiah, B. N. CIITA and Its Dual Roles in MHC Gene Transcription. *Front. Immunol.* **4**, (2013).

95. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).

96. Parikh, K. *et al.* Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).

97. Lichtenstein, G. R. *et al.* ACG Clinical Guideline: Management of Crohn's Disease in Adults. *Off. J. Am. Coll. Gastroenterol. ACG* **113**, 481 (2018).

98. Fujii, M. *et al.* Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. *Cell Stem Cell* **23**, 787-793.e6 (2018).

99. Persson, E. K. *et al.* IRF4 Transcription-Factor-Dependent CD103+CD11b+ Dendritic Cells Drive Mucosal T Helper 17 Cell Differentiation. *Immunity* **38**, 958–969 (2013).

100. Zhou, R. R. & Basu, O. Preparation of single cell suspensions from human intestinal biopsies for single cell genomics applications. (2024).

101. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9**, giaa151 (2020).

102. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

103. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

104. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

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

106. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).

107. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

108. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4314.

109. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).

110. Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, 471 (2013).

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

112. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

113. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012).

114. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

115. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).

116. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

117. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

118. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res.* **31**, 529–537 (2021).

119. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).

120. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

121. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).

122. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

123. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).

124. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

125. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

126. Zhao, S. *et al.* Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nat. Genet.* **56**, 336–347 (2024).

127. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

128. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

129. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLOS Genet.* **18**, e1010299 (2022).

130. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

131. Zhao, Y. *et al.* Multiomic analysis reveals cellular, transcriptomic and epigenetic changes in intestinal pouches of ulcerative colitis patients. *Nat. Commun.* **16**, 904 (2025).

132. Fazio, V. W. *et al.* Ileal Pouch Anal Anastomosis: Analysis of Outcome and Quality of Life in 3707 Patients. *Ann. Surg.* **257**, 679 (2013).

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

134. Toritani, K. *et al.* Inflammatory bowel disease-specific findings are common morphological changes in the ileal pouch with ulcerative colitis. *Sci. Rep.* **12**, 20361 (2022).

135. Ettorre, G. M. *et al.* Mucosal changes in ileal pouches after restorative proctocolectomy for ulcerative and Crohn's colitis. *Dis. Colon Rectum* **43**, 1743–1748 (2000).

136. Sriranganathan, D., Kilic, Y., Nabil Quraishi, M. & Segal, J. P. Prevalence of pouchitis in both ulcerative colitis and familial adenomatous polyposis: A systematic review and meta-analysis. *Colorectal Dis.* **24**, 27–39 (2022).

137. Huang, Y. *et al.* Early Transcriptomic Changes in the Ileal Pouch Provide Insight into the Molecular Pathogenesis of Pouchitis and Ulcerative Colitis: *Inflamm. Bowel Dis.* **0**, 1 (2017).

138. Morgan, X. C. *et al.* Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol.* **16**, 67 (2015).

139. Devlin, J. C. *et al.* Single-Cell Transcriptional Survey of Ileal-Anal Pouch Immune Cells From Ulcerative Colitis Patients. *Gastroenterology* **160**, 1679–1693 (2021).

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

141. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci. Immunol.* **5**, eabb4432 (2020).

142. Planell, N. *et al.* Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut* **62**, 967–976 (2013).

143. Fukui, H. *et al.* DMBT1 is a novel gene induced by IL-22 in ulcerative colitis. *Inflamm. Bowel Dis.* **17**, 1177–1188 (2011).

144. Rosenstiel, P. *et al.* Regulation of DMBT1 via NOD2 and TLR4 in Intestinal Epithelial Cells Modulates Bacterial Recognition and Invasion. *J. Immunol.* **178**, 8203–8211 (2007).

145. Dotti, I. *et al.* Alterations in the epithelial stem cell compartment could contribute to permanent changes in the mucosa of patients with ulcerative colitis. *Gut* **66**, 2069–2079 (2017).

146. van Beelen Granlund, A. *et al.* REG gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. *Cell Tissue Res.* **352**, 639–646 (2013).

147. Ma, Y. & Zhou, X. Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics. *Nat. Methods* **21**, 1231–1244 (2024).

148. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).

149. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

150. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genomics* **2**, 100166 (2022).

151. Yu, Q. *et al.* Charting human development using a multi-endodermal organ atlas and organoid models. *Cell* **184**, 3281-3298.e22 (2021).

152. Gu, W. *et al.* SATB2 preserves colon stem cell identity and mediates ileum-colon conversion via enhancer remodeling. *Cell Stem Cell* **29**, 101-115.e10 (2022).

153. Luk, I. Y. *et al.* Epithelial de-differentiation triggered by co-ordinate epigenetic inactivation of the EHF and CDX1 transcription factors drives colorectal cancer progression. *Cell Death Differ.* **29**, 2288–2302 (2022).

154. Parian, A. M. *et al.* Serrated Epithelial Change Is Associated With High Rates of Neoplasia in Ulcerative Colitis Patients: A Case-controlled Study and Systematic Review With Meta-analysis. *Inflamm. Bowel Dis.* **27**, 1475–1481 (2021).

155. Weiser, M. *et al.* Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* **67**, 36–42 (2018).

156. Fujii, M. *et al.* Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. *Cell Stem Cell* **23**, 787-793.e6 (2018).

157. Persson, E. K. *et al.* IRF4 Transcription-Factor-Dependent CD103+CD11b+ Dendritic Cells Drive Mucosal T Helper 17 Cell Differentiation. *Immunity* **38**, 958–969 (2013).

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

159. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

160. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).

161. Hoffman, G. E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinforma. Oxf. Engl.* **37**, 192–201 (2021).

162. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

163. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).

164. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

165. Liu, C. Y. *et al.* Transitional Anal Cells Mediate Colonic Re-epithelialization in Colitis. *Gastroenterology* **162**, 1975–1989 (2022).

166. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).

167. Weber, C. R., Nalle, S. C., Tretiakova, M., Rubin, D. T. & Turner, J. R. Claudin-1 and claudin-2 expression is elevated in inflammatory bowel disease and may contribute to early neoplastic transformation. *Lab. Investig. J. Tech. Methods Pathol.* **88**, 1110–1120 (2008).

168. Smyth, D. J. *et al.* Shared and Distinct Genetic Variants in Type 1 Diabetes and Celiac Disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).

169. Lowe, C. E. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.* **39**, 1074–1082 (2007).

170. Dendrou, C. A. *et al.* Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.* **41**, 1011–1015 (2009).

171. Ram, M., Sherer, Y. & Shoenfeld, Y. Matrix Metalloproteinase-9 and Autoimmune Diseases. *J. Clin. Immunol.* **26**, 299–307 (2006).

172. Stamenkovic, I. Extracellular matrix remodelling: the role of matrix metalloproteinases. *J. Pathol.* **200**, 448–464 (2003).

173. Lehmkuhl, P. *et al.* Dysregulated immunity in PID patients with low GARP expression on Tregs due to mutations in LRRC32. *Cell. Mol. Immunol.* **18**, 1677–1691 (2021).

174. Abreu, M. T. Nod2 in Normal and Abnormal Intestinal Immune Function. *Gastroenterology* **129**, 1302–1304 (2005).

175. Geremia, A., Biancheri, P., Allan, P., Corazza, G. R. & Di Sabatino, A. Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun. Rev.* **13**, 3–10 (2014).

176. Thompson, A. I. & Lees, C. W. Genetics of ulcerative colitis. *Inflamm. Bowel Dis.* **17**, 831–848 (2011).

177. Zhang, Q. & Cao, X. Epigenetic Remodeling in Innate Immunity and Inflammation. *Annu. Rev. Immunol.* **39**, 279–311 (2021).

178. Finley, A. & Copeland, R. A. Small Molecule Control of Chromatin Remodeling. *Chem. Biol.* **21**, 1196–1210 (2014).