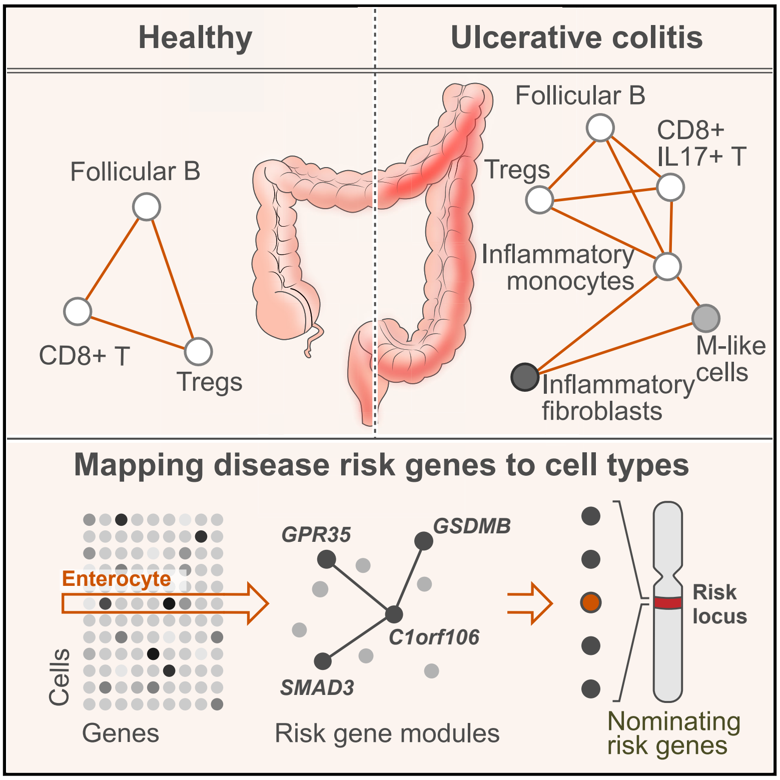
Cell

Resource

Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis

Graphical Abstract



Highlights

- 51 cell subsets in colon mucosa of 18 ulcerative colitis and 12 healthy individuals
- M-like cells, inflammatory monocytes and fibroblasts, and CD8$^+$IL-17$^+$ T cells expand in disease
- Oncostatin M circuit in inflammatory monocytes and fibroblasts may affect drug response
- Co-expression of genes within cells allows inference of causal genes across risk loci

Authors

Christopher S. Smillie, Moshe Biton,
Jose Ordovas-Montanes, ...,
Alex K. Shalek, Ramnik J. Xavier,
Aviv Regev

Correspondence

aananthakrishnan@mgh.harvard.edu
(A.N.A.),
shalek@mit.edu (A.K.S.),
xavier@molbio.mgh.harvard.edu (R.J.X.),
aregev@broadinstitute.org (A.R.)

In Brief

Single-cell analyses of colon biopsy specimens from patients with ulcerative colitis delineate how expression patterns and shifting cell populations may shape disease and drug resistance, and provide a framework for linking GWAS risk loci with specific cell types and functional pathways.

Smillie et al., 2019, Cell *178*, 714–730
July 25, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.cell.2019.06.029

Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis

Christopher S. Smillie,[1,19] Moshe Biton,[1,2,19] Jose Ordovas-Montanes,[1,3,4,5,6,7,19] Keri M. Sullivan,[8] Grace Burgin,[1] Daniel B. Graham,[2,8,9,10,11] Rebecca H. Herbst,[1,12] Noga Rogel,[1] Michal Slyper,[1] Julia Waldman,[1] Malika Sud,[1] Elizabeth Andrews,[8] Gabriella Velonias,[8] Adam L. Haber,[1] Karthik Jagadeesh,[1] Sanja Vickovic,[1] Junmei Yao,[14] Christine Stevens,[9] Danielle Dionne,[1] Lan T. Nguyen,[1] Alexandra-Chloé Villani,[1,13] Matan Hofree,[1] Elizabeth A. Creasey,[14] Hailiang Huang,[15,16] Orit Rozenblatt-Rosen,[1] John J. Garber,[8] Hamed Khalili,[8] A. Nicole Desch,[9,14] Mark J. Daly,[15,16,17] Ashwin N. Ananthakrishnan,[8,*] Alex K. Shalek,[1,3,4,5,6,*] Ramnik J. Xavier,[2,8,9,10,11,14,*] and Aviv Regev[1,18,20,*]

[1]Klarman Cell Observatory, Broad Institute, Cambridge, MA, USA
[2]Department of Molecular Biology, MGH, Boston, MA, USA
[3]Institute for Medical Engineering and Science (IMES), MIT, Cambridge, MA, USA
[4]Department of Chemistry, MIT, Cambridge, MA, USA
[5]Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA
[6]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA
[7]Division of Infectious Diseases and Division of Gastroenterology, Boston Children's Hospital, Boston, MA, USA
[8]Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease, MGH, Boston, MA, USA
[9]Broad Institute, Cambridge, MA, USA
[10]Harvard Medical School, Boston, MA, USA
[11]Center for Microbiome Informatics and Therapeutics, MIT, Cambridge, MA, USA
[12]Department of Systems Biology, Harvard Medical School, Boston, MA, USA
[13]Center for Immunology and Inflammatory Diseases, Department of Medicine, MGH, Boston, MA, USA
[14]Center for Computational and Integrative Biology, MGH, Boston, MA, USA
[15]Medical and Population Genetics, Broad Institute, Cambridge, MA, USA
[16]Analytical and Translational Genetics Unit, MGH, Boston, MA, USA
[17]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
[18]Howard Hughes Medical Institute and Koch Institute for Integrative Cancer Research, Department of Biology, MIT, Cambridge, MA, USA
[19]These authors contributed equally
[20]Lead Contact
*Correspondence: aananthakrishnan@mgh.harvard.edu (A.N.A.), shalek@mit.edu (A.K.S.), xavier@molbio.mgh.harvard.edu (R.J.X.), aregev@broadinstitute.org (A.R.)

https://doi.org/10.1016/j.cell.2019.06.029

SUMMARY

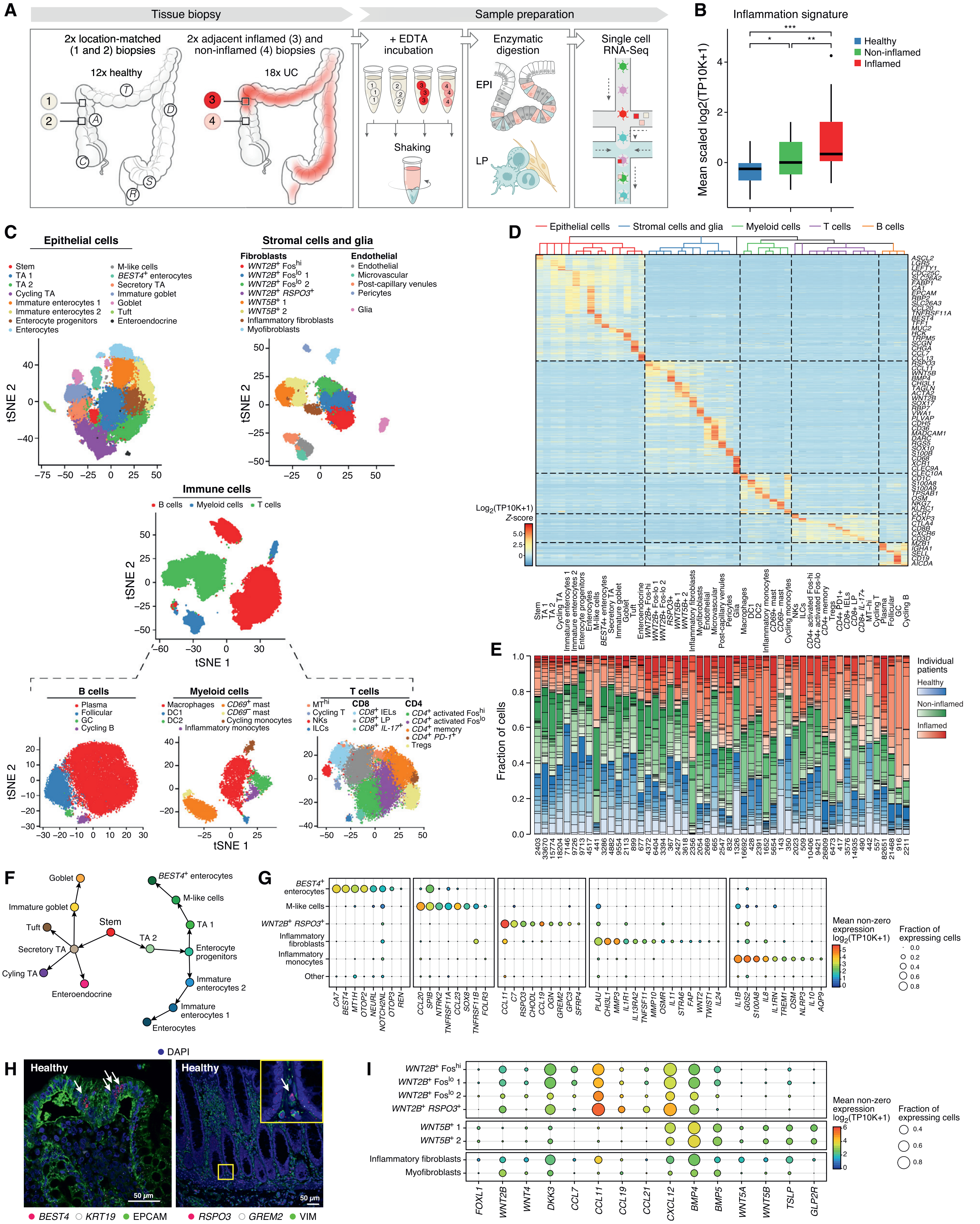
Genome-wide association studies (GWAS) have revealed risk alleles for ulcerative colitis (UC). To understand their cell type specificities and pathways of action, we generate an atlas of 366,650 cells from the colon mucosa of 18 UC patients and 12 healthy individuals, revealing 51 epithelial, stromal, and immune cell subsets, including *BEST4*$^+$ enterocytes, microfold-like cells, and *IL13RA2*$^+$*IL11*$^+$ inflammatory fibroblasts, which we associate with resistance to anti-TNF treatment. Inflammatory fibroblasts, inflammatory monocytes, microfold-like cells, and T cells that co-express *CD8* and *IL-17* expand with disease, forming intercellular interaction hubs. Many UC risk genes are cell type specific and co-regulated within relatively few gene modules, suggesting convergence onto limited sets of cell types and pathways. Using this observation, we nominate and infer functions for specific risk genes across GWAS loci. Our work provides a framework for interrogating complex human diseases and mapping risk variants to cell types and pathways.

INTRODUCTION

Tissues function through the coordinated actions of diverse epithelial, immune, and stromal cell types. Breakdown in any compartment can lead to disease, either because of intrinsic cell dysfunction or the compensatory actions of other cells attempting to restore homeostasis. This interplay can make it difficult to nominate the causal mechanisms that underlie disease. In the specific case of the colonic mucosa, disruptions can lead to ulcerative colitis (UC), a subtype of inflammatory bowel disease (IBD) (Xavier and Podolsky, 2007).

Known disease risk alleles highlight key pathways in the pathogenesis of IBD, including innate and adaptive immunity, gut barrier function, and pathogen sensing and response (Liu et al., 2015; Rivas et al., 2011). However, the underlying genes at risk loci have not been mapped to their cells and pathways of action. Moreover, although histological analysis following endoscopy is the current standard of care (Magro et al., 2017), it fails to capture fine details of disease—e.g., cell proportions, cell type-specific expression, and cell-cell interactions—and does not distinguish between pathways associated with chronic inflammation versus disease restitution.

Single-cell RNA sequencing (scRNA-seq) is helping to advance our understanding of human disease by comprehensively mapping the cell types and states within a tissue,



(legend on next page)

disentangling changes in the expression of gene programs from those in cell frequencies, and connecting them through cell-cell interactions (Tanay and Regev, 2017). Here we apply scRNA-seq to UC, using intestinal biopsies collected from healthy individuals and UC patients to generate and query a single-cell atlas of the healthy and diseased colon.

RESULTS

scRNA-Seq Atlas of Colon Biopsies from Healthy Individuals and UC Patients

We generated 366,650 high-quality single-cell transcriptomes from 68 biopsies (each ~2.4 mm^2) from colonoscopic examinations of 18 UC patients under different treatment regimens and 12 healthy individuals (Figure 1A; STAR Methods; Table S1). We conducted the study in two phases: 115,517 single-cell profiles were collected from 34 colon biopsies of 7 UC patients and 10 healthy individuals as a training set (Figure 1A; STAR Methods; Table S1); another 251,133 were then collected from 34 biopsies of 11 UC patients and 2 healthy individuals as a test set.

To investigate the transitions between healthy and chronically inflamed mucosa while mitigating patient-specific variability, we collected paired samples from each subject in a single procedure. For UC patients, these were endoscopically assessed as adjacent normal tissue (''non-inflamed'') and inflamed or ulcerated tissue (''inflamed'') (Figure 1A; STAR Methods). To estimate intra-subject variation, we obtained two location-matched samples (distance of ~1–2 cm) from each of the 12 healthy subjects as well as from both non-inflamed and inflamed regions of 3 UC patients. We then separated the ''epithelial'' (EPI) and ''*lamina propria*'' (LP) fractions from each sample and performed scRNA-seq (STAR Methods). We confirmed that expression of an inflammation-associated gene set increased from healthy to non-inflamed to inflamed samples (Figure 1B).

A Comprehensive Census of 51 Cell Subsets and Their Molecular Signatures

The single-cell profiles partitioned into 51 subsets by clustering (Figure 1C, after correction for technical and biological variation; STAR Methods), which we annotated by known markers (Figure 1D). The subsets were robust and reproducible because nearly all were represented by all specimens (Figure 1E) and proportionally distributed across patients (Figures S1A and S1B). The discovery and validation cohorts were highly congruent (Figures S1B–S1D), as were replicates collected from the same and even different individuals within the same disease state (Figure S1E; STAR Methods).

The 51 subsets include 15 epithelial subsets, ordered along the differentiation trajectory from intestinal stem cells (ISCs) to mature cell fates (Haber et al., 2017; Figure 1F; STAR Methods). They also include 8 fibroblast, 4 endothelial cell, 1 glial cell, 7 myeloid cell, 4 B cell, 10 T cell (*CD4*$^+$ conventional T helper cell [T$_{\mathrm{conv}}$], regulatory T [T$_{\mathrm{reg}}$], *CD8*$^+$, and $\gamma\delta$), 1 innate lymphoid cell (ILC), and 1 natural killer (NK) cell subsets (Figure 1C). Missing cell types include submucosal enteric neurons, which require isolation by single-nucleus RNA-seq (Habib et al., 2016), plasmacytoid dendritic cells (DCs), and neutrophils (Schelker et al., 2017). Each subset is supported by known and novel markers (Figure 1D; Table S2), including transcription factors (TFs), G protein-coupled receptors (GPCRs), and cytokines (Figures S2A–S2C; Table S3). In most cases, further sub-clusters could not be distinguished by an additional round of clustering (Table S2; STAR Methods). Exceptions included immunoglobulin A (IgA)$^+$ and IgG$^+$ plasma cells, and T$_{\mathrm{reg}}$ cells co-expressing *TNFRSF4/OX40* and *TNFRSF18/GITR*, which may reflect activated versus resting T$_{\mathrm{reg}}$ cells.

Characterization of *BEST4*$^+$ Epithelial Cells and *RSPO3*$^+$ Fibroblasts in the Healthy Colon

Our census revealed that enterocytes expressing *BEST4* are distinct from other epithelial cells (Parikh et al., 2019), and are enriched in genes related to pH sensing and electrolyte balance (validated *in situ*; Figures 1G, 1H, and S1D). These include the otopetrins 2 and 3 (*OTOP2/3*), proton channels that detect pH and underlie sour taste perception (Tu et al., 2018); carbonic anhydrase VII (*CA7*), which catalyzes bicarbonate formation; and bestrophin-4 (*BEST4*), which may export bicarbonate (Qu and Hartzell, 2008). *BEST4*$^+$ enterocytes comprised ~1% of the ileal epithelium from two Crohn's disease (CD) patients (11,473 cells; data not shown).

Multiple fibroblast subsets differ by expression of WNT/bone morphogenetic protein (BMP) signaling genes, likely reflecting distinct positions along the crypt-villus axis (Powell

Figure 1. Single-Cell Atlas of Colon Biopsies from Healthy Individuals and Ulcerative Colitis (UC) Patients

(A) Study design. See also Table S1.

(B) Confirmation of inflammation status. Shown is the mean expression of an inflammation signature (STAR Methods) in cells from healthy (blue), non-inflamed (green), and inflamed (red) biopsies (Wilcoxon test, *p = 0.05; **p = 0.01; ***p = 0.001); boxplots: 25%, 50%, and 75% quantiles; error bars: SD.

(C) Cell census. Shown is t-Stochastic Neighborhood Embedding (t-SNE) of cells colored by cell subset (legend; STAR Methods).

(D) Subset-specific markers. Shown is the expression of marker genes (rows) across cell subsets (columns) ordered by cell lineage relationships (top, color legend; STAR Methods).

(E) Reproducible cell subset distributions across samples (discovery and validation sets). Shown are fractions of cells (y axis) in each cell subset (bars) that are derived from each healthy (blue), non-inflamed (green), or inflamed (red) sample. Bottom: total cell count in the subset (see also Figure S1A).

(F) Epithelial differentiation. Shown is the inferred differentiation trajectory (STAR Methods) for epithelial cell subsets including absorptive (right) and secretory (left) lineages.

(G–I) New colon cell subsets and their markers. Shown are fractions of expressing cells (dot size) and mean expression level in expressing cells (dot color) of select marker genes (columns) across subsets (rows) (G and I) and representative images of combined single-molecule fluorescence *in situ* hybridization (smFISH) and immunofluorescence assay (IFA) of a colon tissue microarray (TMA; STAR Methods) for *BEST4*$^+$ enterocytes (left, white arrow) and *RSPO3*$^+$ fibroblasts (right, white arrow) in healthy colon (H). Inset, ×3 magnification; scale bars, 50 μm.

See also Figures S1 and S2 and Tables S1, S2, and S3.

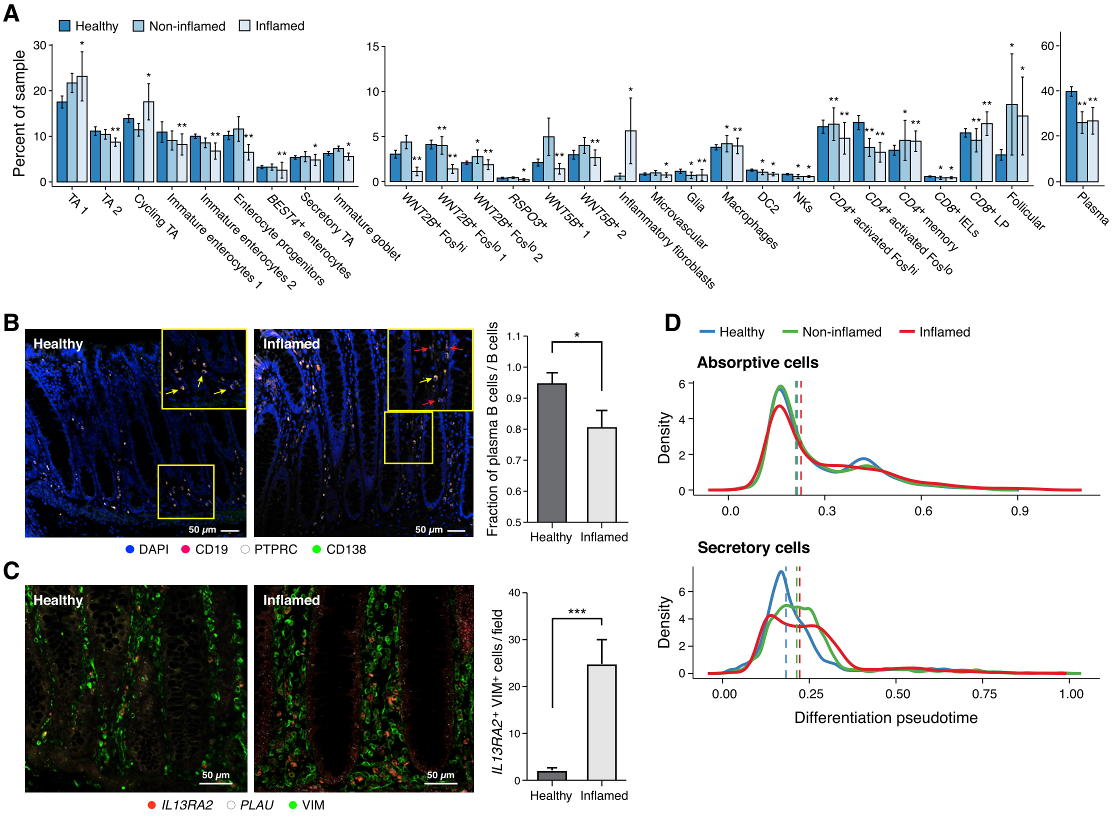


Figure 2. Changes in Cell Composition and Differentiation in UC

(A) Cell proportion changes. Shown are significant changes in cell frequency (y axis) for non-inflamed (light blue) and inflamed (white) samples relative to healthy (dark blue) (Dirichlet-multinomial regression, *adjusted p = 0.05, **adjusted p = 0.01, ***adjusted p = 0.001); error bars: SEM.

(B) Relative reduction in plasma cells among B cells in inflamed colon. Left: representative images of combined smFISH and IFA of plasma cells in TMA from healthy (left) and inflamed (center) human colon; yellow arrows, plasma cell; red arrows, B cell; scale bars, 50 μm; inset, ×2.5 magnification. Right: fraction of plasma cells of total B cells (y axis) in the field of view (n = 9 biopsies per condition; *$p < 0.05$, t test; error bars: SEM).

(C) Expansion of IAFs in inflamed colon. Left: representative images of combined smFISH and IFA of IAFs in TMA from healthy (left) and inflamed (center) human colon; scale bars, 50 μm. Right: number of IAFs (y axis) in the field of view (100 μm^2 per image; n = 9 and n = 7 healthy and inflamed biopsies, respectively; ***$p < 5 \times 10^{-4}$, t test; error bars: SEM).

(D) Reduction in epithelial progenitors with disease. Shown is the distribution of diffusion pseudotimes (STAR Methods) for absorptive (top) and secretory (bottom) epithelial cells, colored by disease state, both significantly shifted to later pseudotimes during disease (likelihood ratio test, $p = 10^{-4}$).

See also Figure S3.

et al., 2011; Shoshkes-Carmel et al., 2018). Some are enriched for *WNT2B*, *WNT4*, and *DKK3*, suggesting that they reside near the crypt, whereas others are enriched for *BMP4*, *BMP5*, and *WNT5A/B* and may reside near the villus (Figure 1I). Many of these genes are reported markers of subepithelial telocytes, a rare population of fibroblasts that support the epithelium (Shoshkes-Carmel et al., 2018); however, in our data, they are broadly distributed across all subsets (e.g., *FOXL1*, *DKK3*, and *WNT5B*; Figure 1I).

One subset of *WNT2B*$^+$ fibroblasts may support the ISC niche by expression of R-spondin-3 (validated *in situ*, Figures 1G–1I), which interacts with the ISC receptor LGR5 (de Lau et al., 2011). *RSPO3*$^+$ fibroblasts express other WNT/BMP signaling genes and several distinct chemokines (Figures 1G and 1I), which may recruit immune cells to the ISC niche (Biton et al., 2018). They are also enriched for genes predictive of poor prognosis in colorectal cancer (CRC) (Calon et al., 2015) and may support tumor growth by promoting a stem-like microenvironment (Figure S5A).

Remodeling of the Colon's Cellular Composition during Disease

The proportions of many cell subsets significantly differed in non-inflamed or inflamed samples versus healthy controls, using both a multivariate test accounting for compositional dependencies (Figure 2A) and univariate tests (Figures S3A and S3B; STAR Methods). These include many known changes in UC patients, such as increases in the proportions of mast cells (King et al., 1992), *CD8*$^+$*IL-17*$^+$ T cells (Tom et al., 2016), and T_{reg} cells (Holmén et al., 2006; Figure 2A, S3A, and S3B).

Microfold (M) cells are typically associated with lymphoid tissue in the human small intestine, where they are important for recognition of the gut microbiota (Mabbott et al., 2013). In the colon, M-like cells were rarely found in healthy individuals but expanded 17-fold during inflammation (validated *in situ*; Figures S3A and S3D). They highly express several chemokines (e.g., *CCL20* and *CCL23*; Figure 1G), suggesting involvement in recruiting immune cells to sites of inflammation.

Mucus layer defects (Xavier and Podolsky, 2007) were not explained by changes in expression (below), suggesting that they may arise post-transcriptionally. Although the frequency of goblet cells did not change, goblet cell progenitors were reduced during inflammation, both as a discrete cell subset (Figure 2A) and along the continuum of differentiation (Figures 1F and 2D; STAR Methods).

Although the overall number of immune cells increased with disease (Danese and Fiocchi, 2011), within the B cell lineage, there was a shift from plasma to follicular (FO) cells (validated *in situ*; Figures 2A and 2B). Among plasma cells, the frequencies of IgA$^+$ relative to IgG$^+$ cells decreased (Figure S3C), suggesting that inflammation is associated with immunoglobulin class switching (Scott et al., 1986).

An Inflammation-Associated Fibroblast Subset Is Unique to the UC Colon

Although most fibroblast subsets were present in both healthy individuals and UC patients, a subset we termed inflammation-associated fibroblasts (IAFs) expanded 189-fold in inflamed tissue of some patients (>1% of LP cells, validated *in situ*; Figures 2A and 2C). IAFs are enriched for expression of many genes associated with colitis, fibrosis, and cancer, including *IL11*, *IL24*, and *IL13RA2* (Figure 1G). Interleukin-11 (IL-11) is a regulator of fibrosis in mice and potentially humans (Schafer et al., 2017). IAFs comprise *WNT2B*$^+$ and *WNT5B*$^+$ subsets (Figure S3E), suggesting that they may reflect a distinct state along the crypt-villus axis.

IAFs express markers of cancer-associated fibroblasts (CAFs) (Figure 1G), including *FAP*, *TWIST1*, and *WNT2* (Erez et al., 2010; Kramer et al., 2017). The expression levels of IAF markers are correlated between IAFs and 414 bulk RNA-seq CRC samples (Cancer Genome Atlas, 2012; Figure S5B; Spearman's $\rho = 0.67$), more than controls ($\rho = 0.33$; $p < 10^{-10}$; Mann-Whitney test), suggesting an expansion of IAFs in tumors (also consistent with increased expression of IAF markers in CRC versus controls; Figure S5B).

Most Expression Changes during Disease Are Shared by Non-Inflamed and Inflamed Tissue

To identify changes in expression associated with disease, we modeled expression as the sum of components reflecting cell subset, disease state (healthy, non-inflamed, or inflamed), and technical covariates while correcting for ambient RNA contamination (Macosko et al., 2015; Figure S1F; STAR Methods). We distinguished between changes shared across cell subsets in epithelial, innate, or adaptive compartments (Figures 3A–3C and S4A–S4C, Table S4) and those unique to each subset (Figures 3D–3F and S4D–S4F; Table S4; STAR Methods).

Despite their endoscopic assessments, non-inflamed and inflamed tissue had similar differentially expressed (DE) gene signatures (Figure 3G; Table S4), suggesting that the transcriptional signature of UC precedes inflammation or persists after resolution. Across epithelial cells, DE genes reflect attempts to restore homeostasis by activating innate immunity, such as antimicrobial and antioxidant defense pathways, mucin biosynthesis, and major histocompatibility complex (MHC) class II machinery (validated *in situ*; Figures 3A, 3D, and 3H; Biton et al., 2018; McDonald and Jewell, 1987). Within the stroma, fibroblasts induced genes for inflammation, fibrosis, and tissue repair (Gieseck et al., 2018), whereas changes in endothelial cells supported tissue vascularization (Figures 3B and 3E). Among immune cells, myeloid and T cells activated co-stimulatory and co-inhibitory genes, and B cells upregulated genes for IgG class switching and affinity maturation (Figures 3C and 3F).

Concerted Metabolic Shifts in Epithelial Cells during Inflammation

Comparison between non-inflamed and inflamed UC tissue (Figures S4A–S4F; Table S4) revealed several metabolic changes in epithelial cells accompanying inflammation. For example, changes in purine metabolism (e.g., *XDH* and *URAD*) may yield uric acid, associated with epithelial damage (Chiaro et al., 2017). Epithelial cells also induced the kynurenine pathway (Figure 4A), associated with disease severity (Sofia et al., 2018). *GPR35*, a kynurenic acid receptor, is a putative risk gene (Huang et al., 2017).

Mapping changes in 239 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways during inflammation (Figures 4B and S5C; STAR Methods) revealed other metabolic alterations in enterocytes, including a shift from oxidative phosphorylation to glycolysis, induction of arginine biosynthesis enzymes (e.g., *NOS2* and *ASS1*), and downregulation of enzymes for the degradation of branched-chain amino acids, particularly *AUH*, a putative risk gene (Liu et al., 2015). Enterocytes also induced HIF1 pathways, a contributor to the glycolytic shift in monocytes (Kelly and O'Neill, 2015). These changes may be driven by impaired production of short-chain fatty acids by gut bacteria (den Besten et al., 2013) because epithelial cells downregulated pathways for β-oxidation and the metabolism of butyrate and propionate but upregulated pathways for dietary fatty acids (e.g., α-linoleic acid).

Induction of a Pro-Inflammatory IL-17 Response and Immune Checkpoints in T Cell Subsets

Among T cells, several *CD4*$^+$ subsets upregulated *IL17A* (Figure 3C), which may reflect both a per-cell increase in *IL17A* expression as well as an expansion of Th17 cells. Surprisingly, a *CD8*$^+$ subset had the strongest overall induction of *IL17A* across both disease states (Figures 4C–4E). *In situ* analysis revealed both *CD4*$^-$ and *CD4*$^+$ cells that co-express *CD8* and *IL17A* (Figures 4D and 4E). Although CD8$^+$IL-17$^+$ T cells have been reported (Cortez et al., 2014; Srenathan et al., 2016), CD4$^+$CD8$^+$IL-17$^+$ T cells are largely uncharacterized. These cells also activated cytotoxic programs and genes related to Th17 pathogenicity in mice (e.g., *RBPJ* and *IL23R*; Figures 3F and 4C; Gaublomme et al., 2015), which may aggravate tissue damage. Most subsets of T cells induced co-stimulatory and co-inhibitory programs (Figure 4C), consistent with attempts to suppress immune activation (Attanasio and Wherry, 2016).

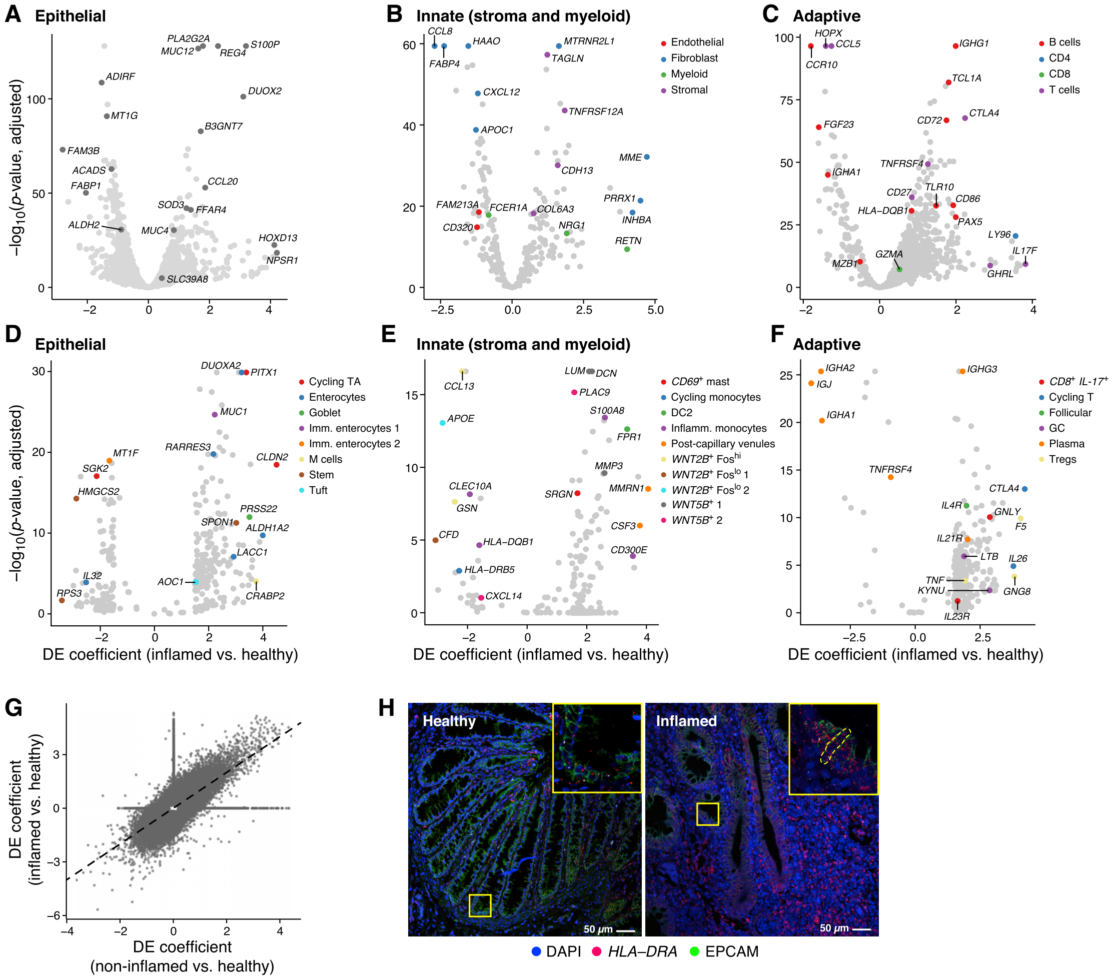


Figure 3. Shared Lineage-Specific and Cell-Specific Expression Changes in Non-inflamed and Inflamed Tissues

(A–G) Lineage- and cell-specific expression changes are shared by non-inflamed and inflamed versus healthy tissue.

(A–F) DE genes shared by the disease states (STAR Methods) with their effect size during inflammation (discrete DE coefficient, x axis) and statistical significance (y axis). Select genes are highlighted; all marker genes are reported in Table S2.

(A–C) Shared changes among multiple cell subsets within (A) epithelial, (B) innate (stromal and myeloid), or (C) adaptive compartments.

(D–F) Unique changes in specific cell subsets within epithelial (D), innate (E), and adaptive (F) compartments.

(G) Discrete DE coefficients estimated for non-inflamed (x axis) and inflamed (y axis) samples versus healthy samples for genes that were significantly DE in at least one disease state (96,445 gene-by-subset coefficients, Spearman's ρ = 0.71, $p < 10^{-16}$).

(H) Upregulation of epithelial MHC class II expression in inflamed colon. Shown are representative images of combined smFISH and IFA of epithelial cells from TMA of healthy (left) and inflamed (right) human colon. Scale bars, 50 μm; inset, ×5 magnification; dashed line, *HLA-DRA*$^+$ epithelial cell.

See also Figure S4 and Table S4.

TNF Expression Shifts to T_{reg}, FO B, and *CD8*$^+$*IL-17*$^+$ T Cells during Inflammation

Monoclonal anti-TNF antibodies are a breakthrough therapy for IBD, but 30% of IBD patients do not respond, and many acquire resistance (Rutgeerts et al., 2004). Tumor necrosis factor (*TNF*) expression shifted during UC, with a prominent role for T_{reg} cells. Baseline expression of *TNF* per cell was highest in *CD8*$^+$ LP and activated *CD4*$^+$*Fos*^{hi} T cells, but in inflamed tissue, *TNF* was induced in T_{reg} and FO B cells (validated *in situ*; Figures 5B and S5D). When estimating the total amount of *TNF* expressed by each cell subset (Figure 5A), T_{reg} cells accounted for 1% of *TNF* expression in healthy tissue but over 14% during inflammation (second only to activated *CD4*$^+$ T cells).

IAFs and Inflammatory Monocytes Are Associated with Resistance to Anti-TNF Therapy

One of the most enriched genes in IAFs is *OSMR*, a putative risk gene (Liu et al., 2015), and the receptor for Oncostatin M (OSM), a cytokine that predicts anti-TNF response (West et al., 2017). *OSM* and *OSMR* are thought to be expressed by unknown myeloid and stromal cells, respectively (West et al., 2017). *OSM* was most enriched in inflammatory monocytes and DC2s,

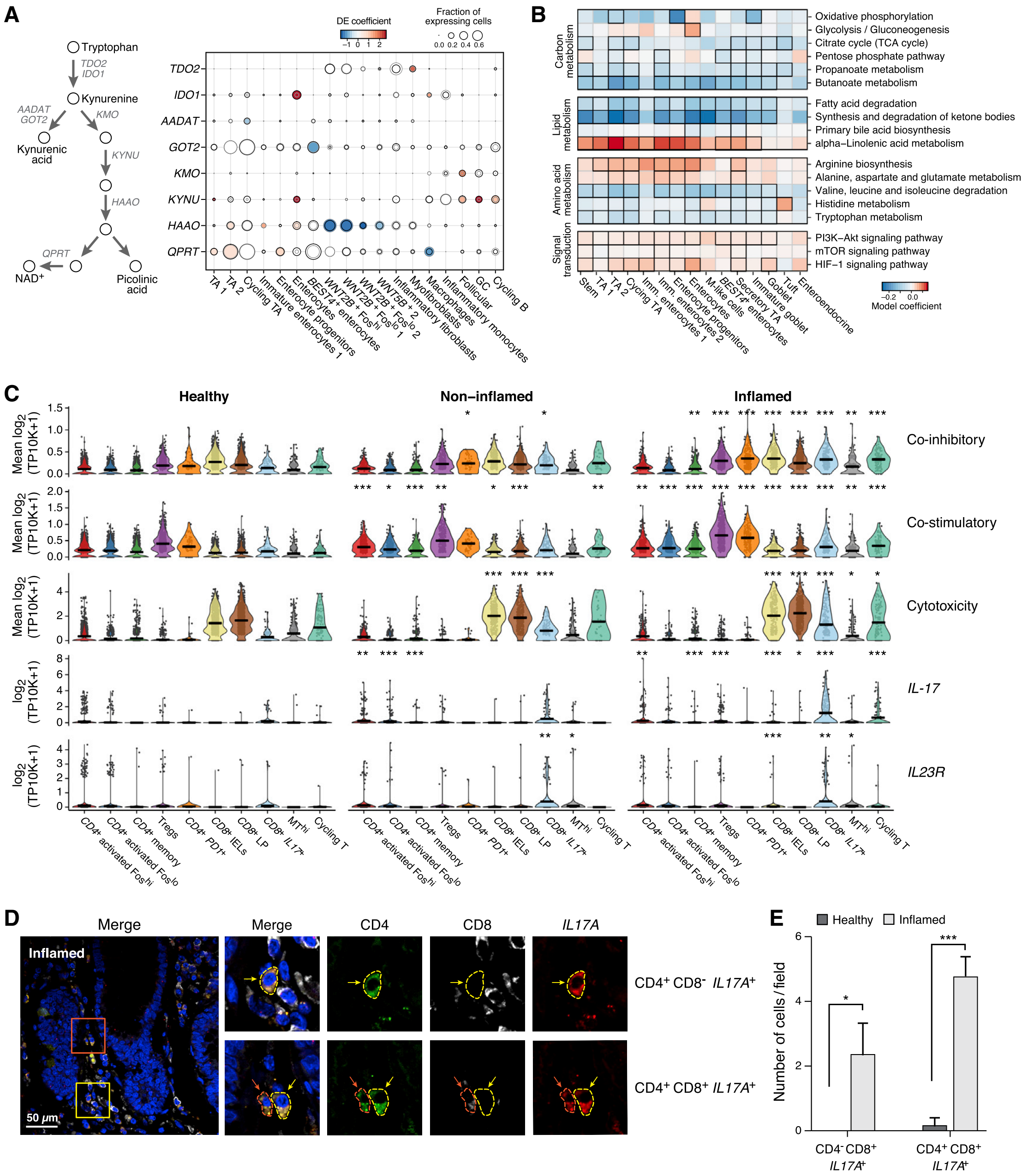


Figure 4. Cell-Specific Expression Changes in UC Highlight Metabolic Reprogramming in Epithelial Cells

(A) Induction of the kynurenine pathway in epithelial cells in UC. Shown are DE genes (rows) from the kynurenine pathway (left) in inflamed versus healthy samples across cell subsets (columns). Dot size, fraction of expressing cells in healthy (gray outline) or inflamed (black outline) samples; dot color, significant DE model coefficients ($q < 0.05$, Model-based Analysis of Single Cell Transcriptomics (MAST) hurdle model, discrete coefficient).

(B) Metabolic reprogramming of enterocytes in UC. Shown are expression changes of KEGG pathways (rows) captured by a mixed linear model (color bar) in inflamed versus healthy samples for epithelial subsets (all subsets in Figure S5C). black outlines, $q < 0.05$.

(C) *CD8*$^+$*IL-17*$^+$ T cells induce *IL17A/F*, *IL23R*, and cytotoxic, co-stimulatory, and co-inhibitory programs in UC. Shown is the distribution of gene and program expression (y axis) in T cells (x axis) from healthy (left), non-inflamed (center), and inflamed (right) samples (Wilcoxon test, *p = 0.05, **p = 0.01, ***p = 0.001); crossbar: mean.

(D) *IL17A* expression by *CD4*$^+$*CD8*$^+$ cells. Shown are representative image of combined smFISH and IFA of *CD4*, *CD8*, and *IL17A* in inflamed human colon TMA (left), showing (inset) *CD4*$^+$*CD8*$^-$*IL17A*$^+$ (yellow outlines; top panels, from the yellow inset) and *CD4*$^+$*CD8*$^+$*IL17A*$^+$ (red outlines, bottom panels, from the red inset) cells. Insets, ×5 magnification.

(E) Number of *CD4*$^-$*CD8*$^+$*IL17A*$^+$ or *CD4*$^+$*CD8*$^+$*IL17A*$^+$ cells in the field of view (250 μm^2). n = 5 samples per condition (*$p < 0.05$, ***$p < 10^{-4}$, t test; error bars: SEM). See also Figure S5.

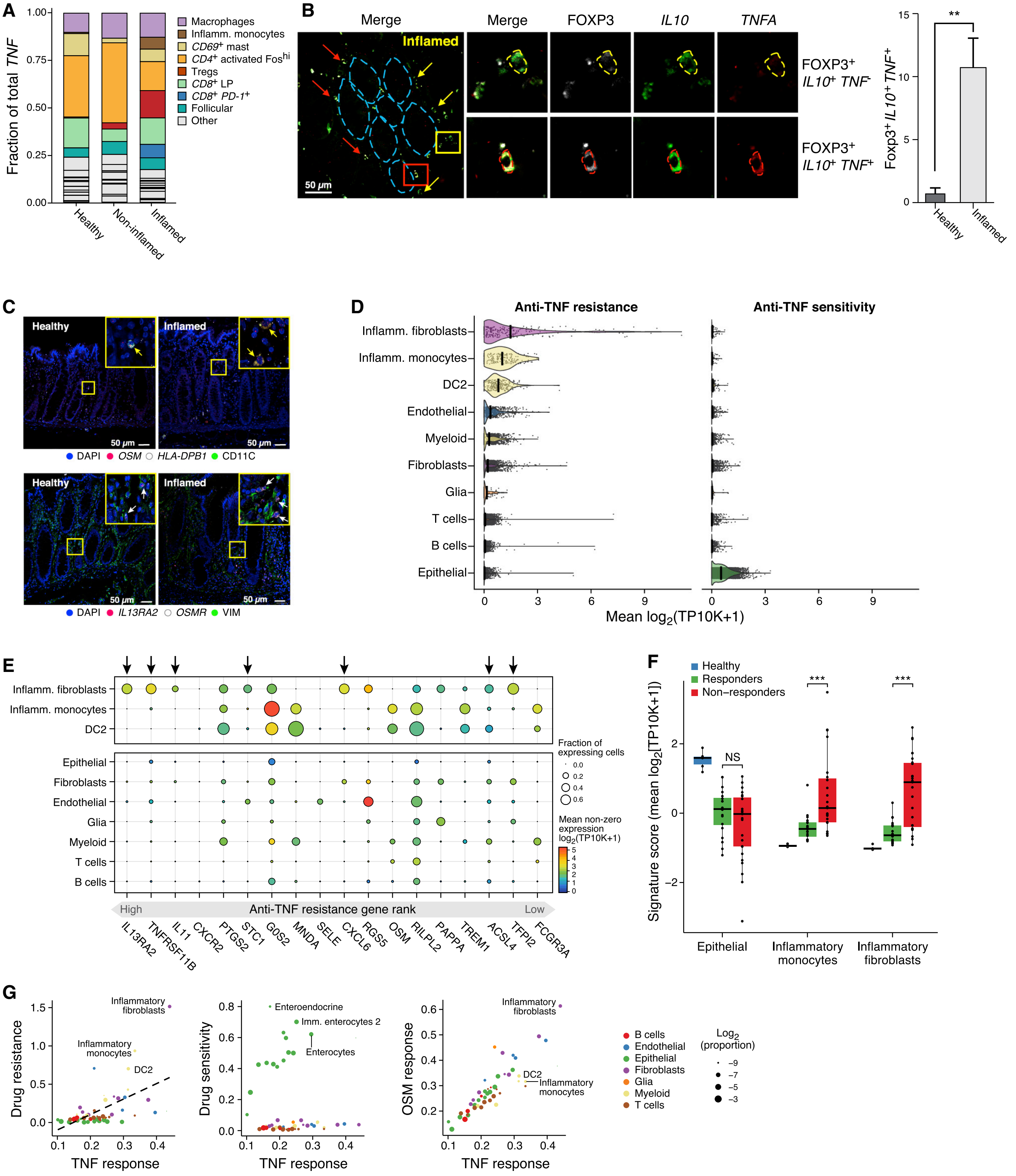


Figure 5. IAFs and Monocytes Are Associated with Anti-TNF Drug Resistance via OSM Signaling

(A and B) T_{reg} cells become major sources of *TNF* expression in UC.

(A) Fraction of total *TNF* transcripts (mean across samples, y axis) expressed by each cell subset in healthy, non-inflamed, and inflamed samples (x axis). Top expressing subsets are highlighted (legend).

(legend continued on next page)

whereas *OSMR* was most enriched in IAFs (validated *in situ*; Figure 5C). Together with the expansion of these subsets during inflammation, this led us to hypothesize that cellular remodeling of the colon may explain, in part, the relationship between OSM and drug resistance.

We therefore scored cell subsets for gene signatures of anti-TNF resistance and sensitivity based on a meta-analysis of bulk expression data from 60 responders and 57 non-responders to therapy (Wang et al., 2016; STAR Methods). The drug resistance signature was strongly enriched in IAFs, inflammatory monocytes, and DC2 cells (Figures 5D and 5E) and the drug sensitivity signature in epithelial cells (Figure 5D). The three genes most associated with drug resistance—*IL13RA2*, TNFRSF11B, and *IL11*—are IAF markers that are rarely expressed in other cells (Figure 5E). An inverse analysis, using the IAF gene signature to infer the pre-treatment levels of IAFs in bulk expression data from 20 drug responders versus 27 non-responders (Figure 5F; STAR Methods), confirmed that IAFs are enriched in patients who are resistant to anti-TNF. Therefore, IAFs may be implicated in the OSM-mediated resistance reported by West et al. (2017).

Potential resistance mechanisms are that OSM synergizes with TNF (West et al., 2017) or phenocopies it. To test these hypotheses, we examined the relationship between TNF and OSM signaling across cell subsets. The signatures were strongly correlated across cell subsets (even after removing shared genes), and both were correlated to the drug resistance signature (Figure 5G). This suggests that OSM phenocopies TNF, activating downstream targets in IAFs. IAFs and inflammatory monocytes may thus compensate during TNF blockade, contributing to resistance.

Rewiring of Cell-Cell Interactions via Inflammation-Associated Cell Subsets during Disease

To more generally chart the rewiring of cell-cell interactions during colitis, we mapped receptor-ligand pairs (Ramilowski et al., 2015) onto cell subsets to construct a putative cell-cell interaction network across disease states (Figures 6A–6C), and identified pairs of cell subsets with significantly more receptor-ligand connections than in a null model (STAR Methods).

Healthy interactions delineated distinct cellular compartments (Figure 6A), whereas DE genes during disease targeted inter-lineage crosstalk and reduced compartmentalization (Figures 6B and 6C), with UC-associated subsets acting as key network hubs (Figure 6D). In healthy mucosa, interactions largely reflected gut homeostasis (e.g., DC1 cells and T cells; Figure 6A; $p < 0.05$). Conversely, non-inflamed interactions were enriched between epithelial cells and fibroblasts and T cells (Figure 6B; all $p < 10^{-4}$), whereas inflamed tissue showed re-wiring of interactions between B cells and T cells and macrophages and $CD8^{+}IL\text{-}17^{+}$ T cells (Figure 6C; all $p < 10^{-4}$). UC-associated subsets (e.g., M-like cells, IAFs, and inflammatory monocytes) were the most central nodes in the network (Figure 6D), indicating that they mediate signals between diverse cell subsets.

Cell-Cell Interactions Predict the Infiltration, Proliferation, and Differentiation of Cell Subsets during Inflammation

We hypothesized that shifts in the proportions of cell subsets could be explained by changes in cell-cell interaction genes expressed by other cells. To test this, we queried all cell subset pairs, examining, for each receptor-ligand pair, whether the ligand's expression level in one cell subset was correlated across samples with the proportions of the cell subset expressing its receptor (including autocrine interactions). This analysis uncovered hundreds of significant interactions (Figure 6E; Table S5; STAR Methods).

For example, *IL18* upregulation by enterocytes during inflammation is correlated with increased proportions of T_{reg} cells, which express its receptor *IL18R1* (Figure 6E; Spearman's $\rho = 0.68$). In mice, IL-18 both inhibits Th17 differentiation and allows for T_{reg} cell-mediated control of gut inflammation (Harrison et al., 2015). However, the role of epithelial cells in recruiting T_{reg} cells to the colon is largely unknown. The frequency of enterocytes, which express *IL22RA1*, was correlated with the expression by $CD4^{+}$ activated Fos^{hi} T cells of *IL22*, which regulates intestinal regeneration (Pelczar et al., 2016; Figure 6E; Spearman's $\rho = 0.55$). We validated this interaction in human colon spheroid culture, where incubation with IL-22 induced an expression program that was significantly enriched in enterocytes versus ISCs (Figure S6A; $p < 10^{-10}$; Wilcoxon test).

Other factors promote recruitment of immune cells (e.g., *CXCL12* for B cells) and expansion of stromal cells (e.g., *PDGFD* for pericytes and *OSM* for IAFs) or are autocrine signals that may

(B) *TNF* expression by T_{reg} cells during inflammation. Left: representative image of combined smFISH and IFA of FOXP3, *IL10*, and *TNFA* in inflamed human colon TMA. $FOXP3^{+}IL10^{+}TNF^{-}$ (yellow outlines; top right, from the yellow inset) and $FOXP3^{+}IL10^{+}TNF^{+}$ (red outlines; bottom right, from the red inset) T_{reg} cells are highlighted. Inset, ×5 magnification; blue dashed lines, crypt position in the tissue. Right: number of $FOXP3^{+}IL10^{+}TNF^{+}$ cells in the field of view (250 μm^2). n = 5 samples per condition (**$p < 0.005$, t test; error bars: SEM).

(C) *OSM* and *OSMR* expression by $MHCII^{+}$ myeloid cells and IAFs, respectively. Shown are representative images of combined smFISH and IFA of TMA from healthy (left) and inflamed (right) human colon. Top: MHC class II^{+} myeloid cells (i.e., inflammatory monocytes or DC2s), yellow arrows. Bottom: IAFs, white arrows. Scale bars, 50 μm. Inset, ×5 magnification.

(D–G) IAF, inflammatory monocyte, and DC2 subsets are associated with anti-TNF resistance.

(D) Distribution of signature scores (x axis) for anti-TNF resistance (left) and sensitivity (right) in select cell subsets (y axis); crossbar: mean.

(E) Mean expression level (color) and fraction of cells (dot size) expressing genes in the anti-TNF resistance signature (columns, ordered by signature rank, bottom bar) in select cell subsets (rows). Arrows, genes whose highest expression is in IAFs.

(F) Distribution of signature scores for cell subsets (x axis) in bulk RNA-seq (Arijs et al., 2009) from human colon biopsies (y axis) of drug responders, non-responders, and healthy controls (Wilcoxon test, ***$p < 0.001$); boxplots: 25%, 50%, and 75% quantiles; error bars: SD.

(G) Signature scores (mean $\log_2$[TP10K+1]) for TNF signaling (KEGG) (x axis) versus drug resistance (left, y axis), drug sensitivity (center, y axis), and OSM signaling (right, y axis) in each cell subset (dots) labeled by lineage (color) and mean proportion across samples (size).

See also Figure S5.

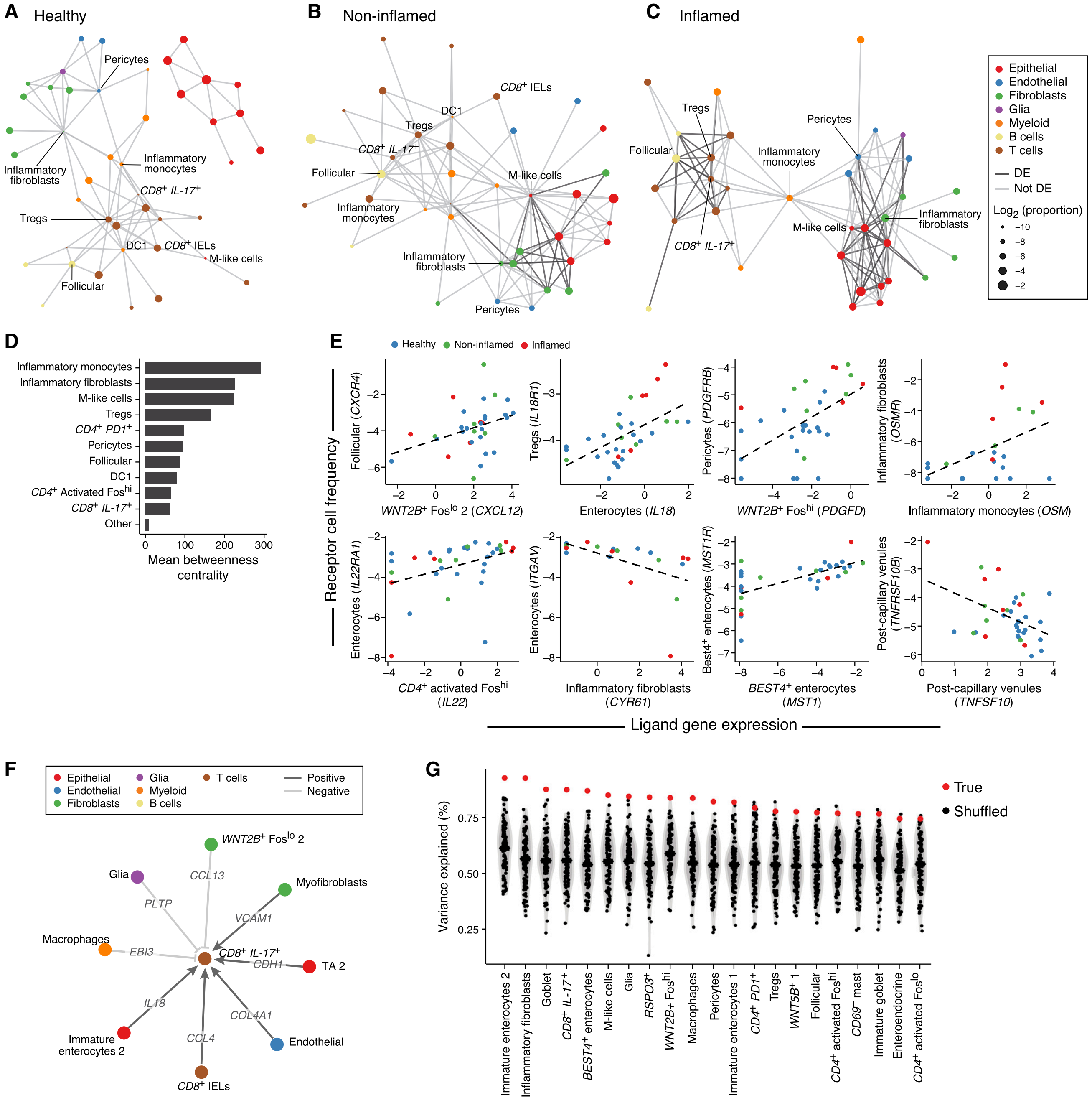


Figure 6. Rewiring of Cell-Cell Interactions Explains Shifts in Cellular Proportions during Disease

(A–C) Increased decompartmentalization with disease. Shown are cell-cell interaction networks estimated in (A) healthy, (B) non-inflamed, and (C) inflamed tissue. Nodes, cell subsets annotated by lineage (color) and mean proportions (size). Edges connect pairs of cell subsets with a significant excess of cognate receptor-ligand pairs expressed (light gray, $p < 0.05$) or DE (dark gray, $p < 0.05$) in a disease state relative to a null model (STAR Methods; Table S5).

(D) Colitis-associated cell subsets are central nodes in the interaction networks. Shown is the mean betweenness centrality (x axis) for each cell subset (y axis) across healthy, non-inflamed, and inflamed networks, showing the 10 highest-ranked cell subsets and the mean across all other subsets (bottom bar).

(E–G) Receptor-ligand interactions explain changes in cell proportions.

(E) Each panel shows, for a pair of cells connected by a receptor-ligand interaction, the mean expression level of the ligand in one cell subset (x axis) and the logit-transformed proportion of the cell subset expressing the receptor (y axis) in each sample, labeled by disease state (color). Dashed line, best linear fit.

(F) Example LASSO model explaining the change in *CD8*$^{+}$*IL-17*$^{+}$ T cell proportions across samples as a function of positive (dark arrows) and negative (light arrows) relationships with ligands (edge label) expressed by other cell subsets colored by lineage.

(legend continued on next page)

regulate cell survival, proliferation, or death (e.g., *MST1* for *BEST4*$^{+}$ enterocytes and *TNFSF10* for post-capillary venules) (Figure 6E). Last, we developed a least absolute shrinkage and selection operator (LASSO) regression model to identify circuits spanning multiple cell types (Figures 6F, 6G, S6B, and S6C; STAR Methods). For example, *CD8*$^{+}$*IL-17*$^{+}$ T cell proportions are explained by a combination of autocrine and paracrine interactions involving epithelial cells, T cells, fibroblasts, and glia (Figure 6F).

Many IBD Risk Genes Are Cell Type or Lineage Specific and Differentially Expressed in Disease

To interrogate IBD genetics using scRNA-seq, we studied 151 risk loci for IBD and UC spanning 346 candidate risk genes (STAR Methods). For most loci, the gene underlying the association signal is unknown; however, in some cases, it is possible to implicate a single gene because it contains a fine-mapped or nonsynonymous coding variant or is resolved to a region of linkage disequilibrium with no other genes. Using this approach, we compiled a set of 57 "GWAS-implicated" risk genes that have a high likelihood of being causally associated with IBD (Table S6; STAR Methods).

Mapping these 57 GWAS-implicated risk genes onto our atlas revealed 29 that were enriched in specific lineages (Figure 7A) and 36 that were significantly DE during disease (Figures S7A and S7B). In addition to known associations (e.g., *NKX2-3* in microvascular cells and *HNF4A* in enterocytes) (Stegmann et al., 2006; Wang et al., 2000), we discovered several new relationships (Table S6). For example, intelectin 1 (*ITLN1*), a lipid raft protein that localizes to the epithelial brush border (Wrackmeyer et al., 2006), is enriched in immature goblet cells. Some cell subsets are enriched for the expression of several GWAS-implicated risk genes (Figure 7B). Notably, M-like cells express many risk genes at higher levels than other cells (e.g., *NR5A2*, *CCL20*, and *JAK2*; Figure 7A; Table S6), suggesting that M-like cell dysfunction may play an important role in the disease.

Co-variation of Gene Expression within a Cell Type Helps Predict Functions for IBD Risk Genes

We hypothesized that variation in gene expression across single cells of the same subset can power us to infer the functions of IBD risk genes. Past approaches often use "guilt by association" across bulk tissue samples but cannot distinguish changes in expression from changes in cell proportions. In contrast, we measured the covariation of genes across single cells within a cell subset, allowing us to isolate co-regulated processes in those cells (Tanay and Regev, 2017; STAR Methods).

In this way, we constructed gene modules for the 57 GWAS-implicated IBD risk genes in all expressing cell subsets and annotated them with putative functions (Table S6; STAR Methods). For example, within healthy epithelial cells, the *C1orf106* module was enriched for tight junction and adherens junction genes ($q < 10^{-6}$ and 10^{-2}, respectively; Fisher's exact test). We recently showed that C1orf106 is involved in cell-cell junctions (Mohanan et al., 2018).

Multiple IBD Risk Genes Co-localize in Shared Gene Modules, Revealing Key Pathways in IBD

In many cases, multiple putative IBD risk genes were members of the same gene module, allowing us to define 10 "meta-modules" spanning over 50% of GWAS-implicated IBD risk genes, which may reflect key disease pathways (empirical $q < 0.05$; Figure 7C; Table S7; STAR Methods). For example, the *PRKCB* meta-module in healthy macrophages contains 5 other risk genes (*GPR65*, *ADCY7*, *PTGER4*, *PTPRC*, and *SH2B3*) and may regulate macrophage activation through cyclic AMP (cAMP) signaling. In addition, the *JAK2* meta-module in UC-associated M-like cells contains 4 other risk genes (*CCL20*, *PTGER4*, *SH2B3*, and *AHR*) and may regulate TNF signaling in M-like cells.

Single-Cell Expression and Co-expression Help Nominate Causal Genes across GWAS Loci

The functional coherence of IBD risk genes suggests that single-cell expression or co-expression can help pinpoint genes that underlie the signal of association at loci with multiple candidate genes. To this end, we identified 20 IBD and UC risk regions (each spanning more than 1 gene) whose candidate gene sets contain a GWAS-implicated risk gene, which we term the "correct" gene for that region (STAR Methods). For each such region, we then tested whether the degrees of (1) expression, (2) differential expression, or (3) co-expression with other candidate genes (iteratively defined across loci; STAR Methods) could recover the "correct" risk gene for the region, relative to a null model in which a gene is randomly selected (STAR Methods). The null model had 33% accuracy, which did not improve when we selected the gene with the largest DE coefficient in either disease state (Figure 7D). However, our predictions significantly improved when we selected the gene with the highest expression in any cell subset (Figure 7D; 55% accuracy, empirical p = 0.03) or belonging to the largest module of other candidate genes (Figure 7D; 63% accuracy, empirical p = 0.001). For CD loci, no method significantly outperformed the null model (Figure 7D), suggesting that the unique risk genes for UC and CD are active in distinct locations or only in diseased tissue.

Finally, we used this co-expression approach to nominate causal genes across all IBD/UC risk loci, including 56 for which the genes driving the association are unknown (Figure 7E; Table S7). We recovered many known IBD and UC risk factors (e.g., *HNF4A*, *IFIH1*, and *GPR35*) but failed to identify others (e.g., RNF186; Figure 7E), highlighting the limitations of our approach (Discussion). In addition, this analysis yielded predictions for 53 genes that were not in the GWAS-implicated set, including *RORC*, *ITGAV*, and *SMURF1* (Figure 7E).

(G) The fraction of variance (y axis) in the proportion of each cell subset (x axis) explained by a LASSO model of cell-cell interactions as in (F) (red dot; STAR Methods) and distribution of this statistic in 100 null models (black dots; STAR Methods). Only subsets with a significant model (empirical p < 0.05) are shown, ordered from left by decreasing fraction of variance explained.
See also Figure S6 and Table S5.

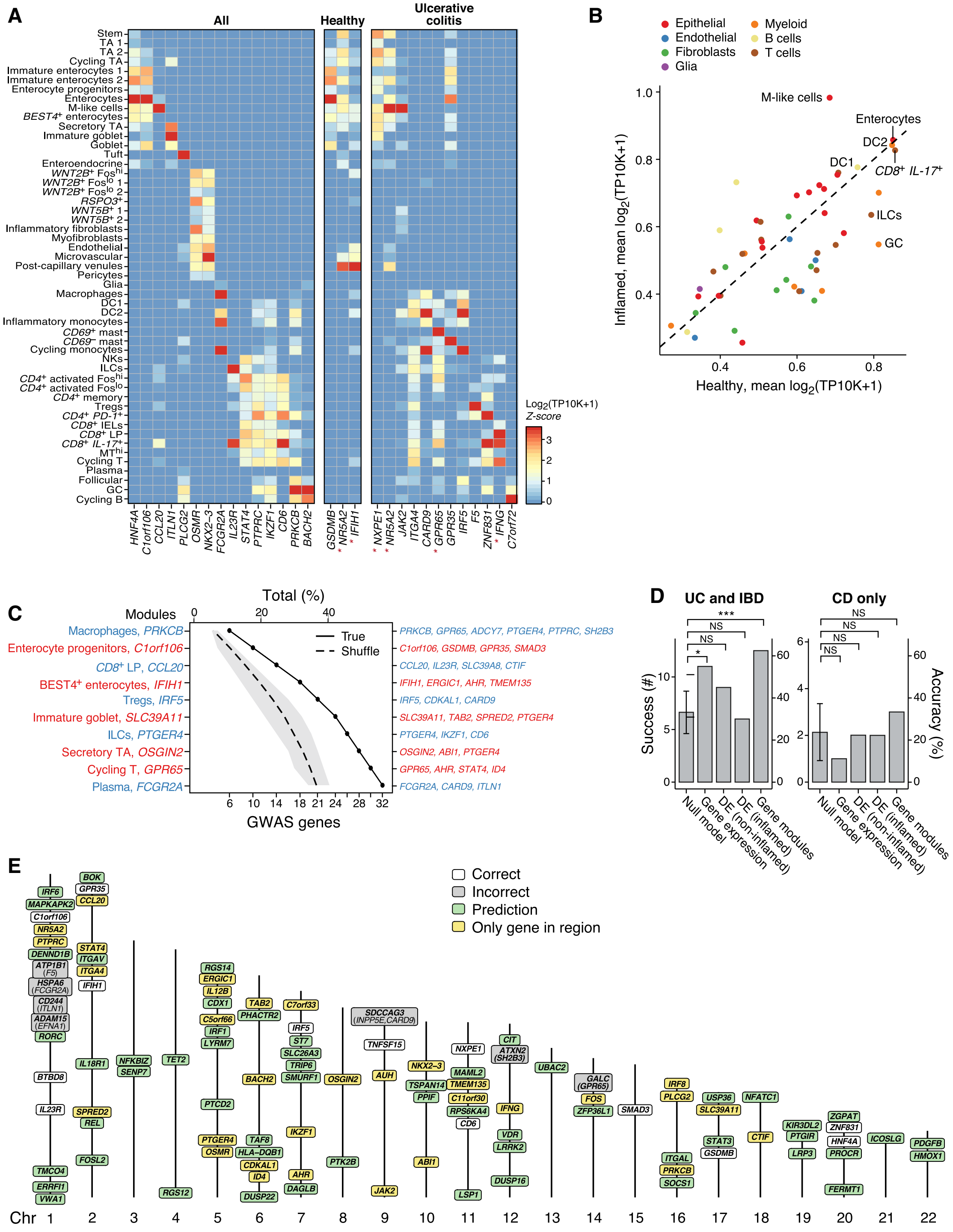


Figure 7. Modules of Co-regulated Risk Genes Help Predict Genes, Pathways, and Cell Types Targeted by IBD

(A) Cell type-specific expression of putative IBD risk genes. Shown is the mean expression of GWAS-implicated IBD risk genes (columns) across cell subsets (rows) that were identified as cell- or lineage-specific in both healthy and UC cells (left), only in healthy cells (center), or only in UC cells (right). Asterisks, genes with significantly changed specificity between health and UC.

(B) Induction of putative IBD risk genes in specific subsets in disease. Shown is the mean expression of GWAS-implicated IBD risk genes across cell subsets (marked by lineage, color) in healthy (x axis) and inflamed (y axis) samples.

(C) Functional annotation of putative IBD risk genes by co-expression meta-modules within a cell subset. Shown are the number (bottom x axis) and percent (top x axis) of GWAS-implicated IBD risk genes captured (solid line) by successive addition of each meta-module seeded by an IBD risk gene (y axis) using

(legend continued on next page)

DISCUSSION

By leveraging scRNA-seq in a clinical context, we assessed cellular composition, gene expression, cell-cell interactions, and IBD risk gene pathways in specific cell subsets from intestinal biopsies. Although distinguishing cause from effect is challenging, relating single-cell data to clinical responses (e.g., IAFs), cell-cell interactions (e.g., enterocytes and T cells), or risk gene expression (e.g., M-like cells) can help inform disease etiology and highlight new opportunities for therapeutic intervention.

M-like cells were rarely detected at baseline but expanded during inflammation and acted as hubs in the cell-cell interaction network (Figure 6D; S3A and S3D). This expansion may reflect tertiary lymphoid tissue or sentinel cells (Mabbott et al., 2013). M-like cells had the highest expression of GWAS-implicated risk genes (Figures 7A and 7B), including *CCL20*, whose expression was correlated to T_{reg} cell frequencies across samples (Table S6). They had the largest module of predicted risk genes (Table S7), enriched in endocytosis and Th17 differentiation genes ($q < 10^{-3}$, Fisher's exact test), which may reflect transcytosis and delivery of antigens.

CD8^{+}*IL-17*$^{+}$ T cells and T_{regs} expand from healthy to non-inflamed to inflamed tissue (Figure 2A) and become major sources of *IL-17* (Figure 4C) and *TNF* (Figure 5A,B) during inflammation, respectively. The former may contribute to T cell pathogenicity and tissue damage (Figures 4D and 4E; 66% co-express *CD4*). Although the latter may have adopted an effector-like state, they are still more enriched for T_{reg} cell markers (e.g., *FOXP3*, *CTLA4*, and *IL10*) than *TNF*$^{-}$ T_{reg} cells (data not shown). Future work will determine whether TNF^{+} T_{reg} cells contribute to disease pathology or anti-TNF resistance (Atreya et al., 2011), as well as the role of CD8^{+} T cell plasticity in gut inflammation.

OSM signaling was implicated in anti-TNF resistance via unknown myeloid and stromal cell types (West et al., 2017). Here we show that inflammatory monocytes and IAFs may mediate resistance via expression of *OSM* and *OSMR*, respectively (Figure 5D). In particular, IAFs were enriched in pre-treatment samples from anti-TNF non-responders (Figure 5F). In addition, we identified that OSM phenocopies TNF, which may explain anti-TNF resistance. Future work will determine whether IAFs are a robust biomarker of drug response or whether combining anti-TNF drugs with inhibition of IAF cytokines and/or receptors can reduce anti-TNF resistance in UC patients.

IAFs uniquely express *IL11*, a potential therapeutic target for fibrosis (Schafer et al., 2017), suggesting involvement in gut fibrosis. Because they express crypt-associated (*WNT2B*$^{+}$) and villus-associated (*WNT5B*$^{+}$) markers (Figures 1I and S3E), IAFs may reflect a distinct fibroblast state. IAFs express several CAF markers, and IAF markers are enriched in CRC tumors (Figure S5B), suggesting a shared origin and/or state. IAF expansion during cancer-associated inflammation may affect the tumor microenvironment. Last, both IAFs and inflammatory monocytes form hubs in the cell-cell interaction network and may affect the proportions of other cells (Figure 6E; Table S5).

By leveraging single-cell co-expression, we mapped more than 50% of risk genes onto 10 meta-modules (Figure 7C) and used these meta-modules to nominate causal risk genes across loci (Figure 7E). However, this approach may fail to identify risk genes that are lowly expressed, active in cells and/or tissues that were not profiled, or not co-expressed with other risk genes. It may also fail when multiple risk genes act at the same locus; however, we find that it improves predictions even when scoring genes irrespective of region (rather than selecting one gene per region) (Figure S7D; STAR Methods). We hope that these analyses will pave the road for combining human genetics and single-cell genomics to better understand polygenic disorders by relating risk gene modules to polygenic risk scores, mapping variants to single-cell phenotypes, and mapping non-coding variants to cells via single-cell allele-specific expression and expression quantitative trait loci (eQTL) analyses.

Our work provides a framework for using scRNA-seq to understand human diseases and their therapeutic responses. We identify changes in cell proportions and gene expression with disease state and integrate these to understand mechanisms of cell-cell signaling and drug susceptibility. Finally, we nominate risk genes across loci, predicting their cells of action and putative functions, and assemble them into the core pathways that underlie disease.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Patients and tissue samples
- METHOD DETAILS
  - ○ Single cell dissociation from fresh biopsies
  - ○ Human spheroid cultures for IL-22-enterocyte interaction validation
  - ○ Droplet-based scRNA-Seq
  - ○ SMART-Seq2 for sequencing of human colon spheroids
  - ○ Immunofluorescence assay (IFA)

healthy (blue) or UC (red) cells relative to a null model (dashed line). Left labels, cell type and seed gene. Right labels, GWAS-implicated IBD risk genes in the meta-module.

(D and E) Meta-modules help nominate causal IBD risk genes from GWAS risk loci.

(D) Mean number of correct predictions (left y axis) and mean percent accuracy (right y axis) across 20 risk regions for IBD and UC (left) and 22 risk regions unique to CD (right) for several methods based on scRNA-seq relative to the null model (x axis). *p = 0.05, **p = 0.01, ***p = 0.001, NS = not significant, Wilcoxon test.

(E) Nominated risk genes. Shown are loci containing GWAS-implicated IBD risk genes with correct (white) or incorrect (gray) predictions, loci associated with a single gene (gold), and all other loci (green). Incorrect predictions are annotated with the predicted (top) and correct (bottom) gene.

See also Figure S7 and Tables S6 and S7.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.06.029.

ACKNOWLEDGMENTS

We thank all consenting participants of this study, the clinical staff at Massachusetts General Hospital, Leslie Gaffney and Anna Hupalowska for help with figures, Tim Tickle for help with the Single Cell Portal, Vijay Yajnik, Frank Colizzo, and Abdifatah Omer. M.B. was supported by a postdoctoral fellowship from the Human Frontiers Science Program. J.O.-M. is an HHMI Damon Runyon Cancer Research Foundation Fellow (DRG-2274-16). Work was supported by the Klarman Cell Observatory (to A.R. and R.J.X.); HHMI (to A.R.); the Manton Foundation (to A.R. and R.X.); Broadnext10 (to A.R. and R.J.X.); DK043351, DK114784, DK117263, the Helmsley Charitable Trust, and the Crohn's and Colitis Foundation (to R.J.X.); a Sloan Fellowship in Chemistry (to A.K.S.); the Searle Scholars Program (to A.K.S.); the Beckman Young Investigator Program (to A.K.S.); and NIH 5U24AI118672 (to A.R. and A.K.S.). This publication is part of the Human Cell Atlas.

AUTHOR CONTRIBUTIONS

M.B. and J.O.-M. designed and performed experiments with A.K.S., R.J.X., A.N.A., A.R., G.B., N.R., M. Slyper, J.W., M. Sud, S.V., J.Y., E.A.C., and A.N.D. C.S.S. designed and performed computational analyses with A.R., R.J.X., D.B.G., R.H.H., A.L.H., K.J., A.-C.V., M.H., H.H., M.B., J.O.-M., and M.J.D. K.M.S., E.A., and G.V. performed clinical work with C.S., J.J.G., H.K., A.N.A., and R.J.X. D.D., L.T.N., and O.R.-R. assisted with scRNA-seq. C.S.S., M.B., J.O.-M., A.K.S., R.J.X., and A.R. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

A.R. is an SAB member of Thermo Fisher Scientific and Syros Pharmaceuticals. A.R. and R.J.X. are cofounders of and equity holders in Celsius Therapeutics. R.J.X. is a consultant to Novartis. A.K.S is compensated for consulting and SAB membership by Honeycomb Biotechnologies, Dot Bio, Cellarity, Cogen Therapeutics, and Dahlia Biosciences. M.B., A.L.H., N.R., R.H.H., J.O.-M., O.R.-R., A.K.S., K.S., C.S.S., A.R., and R.J.X. are co-inventors on PCT/US2018/042554 relating to advances in understanding cellular dynamics, cellular effectors, and risk variants of the human colon in health and UC described in this manuscript.

Received: September 27, 2018
Revised: March 25, 2019
Accepted: June 18, 2019
Published: July 25, 2019

REFERENCES

Arijs, I., Li, K., Toedter, G., Quintens, R., Van Lommel, L., Van Steen, K., Leemans, P., De Hertogh, G., Lemaire, K., Ferrante, M., et al. (2009). Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. Gut *58*, 1612–1619.

Atreya, R., Zimmer, M., Bartsch, B., Waldner, M.J., Atreya, I., Neumann, H., Hildner, K., Hoffman, A., Kiesslich, R., Rink, A.D., et al. (2011). Antibodies against tumor necrosis factor (TNF) induce T-cell apoptosis in patients with inflammatory bowel diseases via TNF receptor 2 and intestinal CD14$^+$ macrophages. Gastroenterology *141*, 2026–2038.

Attanasio, J., and Wherry, E.J. (2016). Costimulatory and Coinhibitory Receptor Pathways in Infectious Disease. Immunity *44*, 1052–1068.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154/1009.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B *57*, 289–300.

Biton, M., Levin, A., Slyper, M., Alkalay, I., Horwitz, E., Mor, H., Kredo-Russo, S., Avnit-Sagi, T., Cojocaru, G., Zreik, F., et al. (2011). Epithelial microRNAs regulate gut mucosal immunity via epitheliaum-T cell crosstalk. Nat. Immun. *12*, 239.

Biton, M., Haber, A.L., Rogel, N., Burgin, G., Beyaz, S., Schnell, A., Ashenberg, O., Su, C.W., Smillie, C., Shekhar, K., et al. (2018). T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. Cell *175*, 1307–1320.e22.

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

Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., and Theis, F.J. (2019). A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods *16*, 43–49.

Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D.V., Byrom, D., et al. (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat. Genet. *47*, 320–329.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Chiaro, T.R., Soto, R., Zac Stephens, W., Kubinak, J.L., Petersen, C., Gogokhia, L., Bell, R., Delgado, J.C., Cox, J., Voth, W., et al. (2017). A member of the gut mycobiota modulates host purine metabolism exacerbating colitis in mice. Sci. Transl. Med. *9*, eaaf9044.

Cortez, V.S., Cervantes-Barragan, L., Song, C., Gilfillan, S., McDonald, K.G., Tussiwand, R., Edelson, B.T., Murakami, Y., Murphy, K.M., Newberry, R.D., et al. (2014). CRTAM controls residency of gut CD4+CD8+ T cells in the steady state and maintenance of gut CD4+ Th17 during parasitic infection. J. Exp. Med. *211*, 623–633.

Danese, S., and Fiocchi, C. (2011). Ulcerative colitis. N. Engl. J. Med. *365*, 1713–1725.

de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. *49*, 256–261.

de Lau, W., Barker, N., Low, T.Y., Koo, B.K., Li, V.S., Teunissen, H., Kujala, P., Haegebarth, A., Peters, P.J., van de Wetering, M., et al. (2011). Lgr5 homologues associate with Wnt receptors and mediate R-spondin signalling. Nature *476*, 293–297.

den Besten, G., van Eunen, K., Groen, A.K., Venema, K., Reijngoud, D.J., and Bakker, B.M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J. Lipid Res. *54*, 2325–2340.

Erez, N., Truitt, M., Olson, P., Arron, S.T., and Hanahan, D. (2010). Cancer-Associated Fibroblasts Are Activated in Incipient Neoplasia to Orchestrate Tumor-Promoting Inflammation in an NF-kappaB-Dependent Manner. Cancer Cell *17*, 135–147.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278.

Gaublomme, J.T., Yosef, N., Lee, Y., Gertner, R.S., Yang, L.V., Wu, C., Pandolfi, P.P., Mak, T., Satija, R., Shalek, A.K., et al. (2015). Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. Cell *163*, 1400–1412.

Gieseck, R.L., 3rd, Wilson, M.S., and Wynn, T.A. (2018). Type 2 immunity in tissue repair and fibrosis. Nat. Rev. Immunol. *18*, 62–76.

Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., et al. (2017). A single-cell survey of the small intestinal epithelium. Nature *551*, 333–339.

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J.J., Hession, C., Zhang, F., and Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science *353*, 925–928.

Harrison, O.J., Srinivasan, N., Pott, J., Schiering, C., Krausgruber, T., Ilott, N.E., and Maloy, K.J. (2015). Epithelial-derived IL-18 regulates Th17 cell differentiation and Foxp3$^+$ Treg cell function in the intestine. Mucosal Immunol. *8*, 1226–1236.

Holmén, N., Lundgren, A., Lundin, S., Bergin, A.M., Rudin, A., Sjövall, H., and Ohman, L. (2006). Functional CD4+CD25high regulatory T cells are enriched in the colonic mucosa of patients with active ulcerative colitis and increase with disease activity. Inflamm. Bowel Dis. *12*, 447–456.

Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C.A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al.; International Inflammatory Bowel Disease Genetics Consortium (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature *547*, 173–178.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

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

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45* (D1), D353–D361.

Kelly, B., and O'Neill, L.A. (2015). Metabolic reprogramming in macrophages and dendritic cells in innate immunity. Cell Res. *25*, 771–784.

King, T., Biddle, W., Bhatia, P., Moore, J., and Miner, P.B., Jr. (1992). Colonic mucosal mast cell distribution at line of demarcation of active ulcerative colitis. Dig. Dis. Sci. *37*, 490–495.

Kramer, N., Schmöllerl, J., Unger, C., Nivarthi, H., Rudisch, A., Unterleuthner, D., Scherzer, M., Riedl, A., Artaker, M., Crncec, I., et al. (2017). Autocrine WNT2 signaling in fibroblasts promotes colorectal cancer progression. Oncogene *36*, 5460–5472.

Lata, S., and Raghava, G.P. (2008). PRRDB: a comprehensive database of pattern-recognition receptors and their ligands. BMC Genomics *9*, 180.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, A.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*, 184–197.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.

Mabbott, N.A., Donaldson, D.S., Ohno, H., Williams, I.R., and Mahajan, A. (2013). Microfold (M) cells: important immunosurveillance posts in the intestinal epithelium. Mucosal Immunol. *6*, 666–677.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Magro, F., Gionchetti, P., Eliakim, R., Ardizzone, S., Armuzzi, A., Barreiro-de Acosta, M., Burisch, J., Gecse, K.B., Hart, A.L., Hindryckx, P., et al.; European Crohn's and Colitis Organisation [ECCO] (2017). Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders. J. Crohn's Colitis *11*, 649–670.

McDonald, G.B., and Jewell, D.P. (1987). Class II antigen (HLA-DR) expression by intestinal epithelial cells in inflammatory diseases of colon. J. Clin. Pathol. *40*, 312–317.

Mohanan, V., Nakata, T., Desch, A.N., Lévesque, C., Boroughs, A., Guzman, G., Cao, Z., Creasey, E., Yao, J., Boucher, G., et al. (2018). *C1orf106* is a colitis risk gene that regulates stability of epithelial adherens junctions. Science *359*, 1161–1166.

Parikh, K., Antanaviciute, A., Fawkner-Corbett, D., Jagielowicz, M., Aulicino, A., Lagerholm, C., Davis, S., Kinchen, J., Chen, H.H., Alham, N.K., et al. (2019). Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature *567*, 49–55.

Pelczar, P., Witkowski, M., Perez, L.G., Kempski, J., Hammel, A.G., Brockmann, L., Kleinschmidt, D., Wende, S., Haueis, C., Bedke, T., et al. (2016). A pathogenic role for T cell-derived IL-22BP in inflammatory bowel disease. Science *354*, 358–362.

Persson, E.K., Uronen-Hansson, H., Semmrich, M., Rivollier, A., Hägerbrand, K., Marsal, J., Gudjonsson, S., Håkansson, U., Reizis, B., Kotarsky, K., and Agace, W.W. (2013). IRF4 transcription-factor-dependent CD103(+)CD11b(+) dendritic cells drive mucosal T helper 17 cell differentiation. Immunity *38*, 958–969.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Powell, D.W., Pinchuk, I.V., Saada, J.I., Chen, X., and Mifflin, R.C. (2011). Mesenchymal cells of the intestinal lamina propria. Annu. Rev. Physiol. *73*, 213–237.

Qu, Z., and Hartzell, H.C. (2008). Bestrophin Cl- channels are highly permeable to HCO3-. Am. J. Physiol. Cell Physiol. *294*, C1371–C1377.

Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., and Forrest, A.R. (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. Nat. Commun. *6*, 7866.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al.; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.

Rosvall, M., and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA *105*, 1118–1123.

Rutgeerts, P., Van Assche, G., and Vermeire, S. (2004). Optimizing anti-TNF treatment in inflammatory bowel disease. Gastroenterology *126*, 1593–1610.

Schaefer, U., Schmeier, S., and Bajic, V.B. (2011). TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. Nucleic Acids Res. *39*, D106–D110.

Schafer, S., Viswanathan, S., Widjaja, A.A., Lim, W.W., Moreno-Moral, A., DeLaughter, D.M., Ng, B., Patone, G., Chow, K., Khin, E., et al. (2017). IL-11 is a crucial determinant of cardiovascular fibrosis. Nature *552*, 110–115.

Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B., and Raue, A. (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat. Commun. *8*, 2032.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Scott, M.G., Nahm, M.H., Macke, K., Nash, G.S., Bertovich, M.J., and MacDermott, R.P. (1986). Spontaneous secretion of IgG subclasses by intestinal mononuclear cells: differences between ulcerative colitis, Crohn's disease, and controls. Clin. Exp. Immunol. *66*, 209–215.

Shoshkes-Carmel, M., Wang, Y.J., Wangensteen, K.J., Tóth, B., Kondo, A., Massasa, E.E., Itzkovitz, S., and Kaestner, K.H. (2018). Subepithelial telocytes are an important source of Wnts that supports intestinal crypts. Nature *557*, 242–246.

Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al. (2018). WikiPathways: a multi-faceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. *46* (D1), D661–D667.

Sofia, M.A., Ciorba, M.A., Meckel, K., Lim, C.K., Guillemin, G.J., Weber, C.R., Bissonnette, M., and Pekow, J.R. (2018). Tryptophan Metabolism through the Kynurenine Pathway is Associated with Endoscopic Inflammation in Ulcerative Colitis. Inflamm. Bowel Dis. *24*, 1471–1480.

Srenathan, U., Steel, K., and Taams, L.S. (2016). IL-17+ CD8+ T cells: Differentiation, phenotype and role in inflammatory disease. Immunol. Lett. *178*, 20–26.

Stegmann, A., Hansen, M., Wang, Y., Larsen, J.B., Lund, L.R., Ritié, L., Nicholson, J.K., Quistorff, B., Simon-Assmann, P., Troelsen, J.T., and Olsen, J. (2006). Metabolome, transcriptome, and bioinformatic cis-element analyses point to HNF-4 as a central regulator of gene expression during enterocyte differentiation. Physiol. Genomics *27*, 141–155.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331–338.

The UniProt Consortium (2018). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *46*, 2699.

Tom, M.R., Li, J., Ueno, A., Fort Gasia, M., Chan, R., Hung, D.Y., Chenoo, S., Iacucci, M., Jijon, H.B., Kaplan, G.G., et al. (2016). Novel CD8+ T-Cell Subsets Demonstrating Plasticity in Patients with Inflammatory Bowel Disease. Inflamm. Bowel Dis. *22*, 1596–1608.

Tu, Y.H., Cooper, A.J., Teng, B., Chang, R.B., Artiga, D.J., Turner, H.N., Mulhall, E.M., Ye, W., Smith, A.D., and Liman, E.R. (2018). An evolutionarily conserved gene family encodes proton-selective ion channels. Science *359*, 1047–1050.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579–2605.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716–729.e27.

Wang, C.C., Biben, C., Robb, L., Nassir, F., Barnett, L., Davidson, N.O., Koentgen, F., Tarlinton, D., and Harvey, R.P. (2000). Homeodomain factor Nkx2-3 controls regional expression of leukocyte homing coreceptor MAdCAM-1 in specialized endothelial cells of the viscera. Dev. Biol. *224*, 152–167.

Wang, Y., Zhu, X., Zhen, N., Pan, Q., and Li, Y. (2016). Gene expression profile predicting the response to anti-TNF antibodies therapy in patients with inflammatory bowel disease: analyses of GEO datasets. Int. J. Clin. Exp. Med. *9*, 23397–23406.

Welch, J., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. (2018). Integrative inference of brain cell similarities and differences from single-cell genomics. bioRxiv. https://doi.org/10.1101/459891.

West, N.R., Hegazy, A.N., Owens, B.M.J., Bullers, S.J., Linggi, B., Buonocore, S., Coccia, M., Görtz, D., This, S., Stockenhuber, K., et al.; Oxford IBD Cohort Investigators (2017). Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel disease. Nat. Med. *23*, 579–589.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018a). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Wolf, F.A., Hamey, F., Plass, M., Solana, J., Dahlin, J.S., Gottgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2018b). Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. bioRxiv. https://doi.org/10.1101/208819.

Wrackmeyer, U., Hansen, G.H., Seya, T., and Danielsen, E.M. (2006). Intelectin: a novel lipid raft-associated protein in the enterocyte brush border. Biochemistry *45*, 9188–9197.

Xavier, R.J., and Podolsky, D.K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. Nature *448*, 427–434.

Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., and Bruford, E.A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. *45* (D1), D619–D625.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
mouse anti-EPCAM	ThermoFisher	Cat#MA1-06502; RRID:AB_558797
mouse anti-Vimentin	Millipore	Cat#MAB3400; RRID:AB_94843
mouse anti-CD19	BioLegend	Cat#302201; RRID:AB_314231
rat anti-CD45 (PTPRC)	ThermoFisher	Cat# MA5-17687; RRID:AB_2539077
goat anti-CD138	R&D Systems	Cat# AF2780; RRID:AB_442186
mouse anti-CD11c	BD Biosciences	Cat#550375; RRID:AB_393646
goat anti-CD4	R&D Systems	Cat#AF-379-NA; RRID:AB_354469
rabbit anti-CD8	Invitrogen	Cat# SP16; RRID:AB_837984
mouse anti-FOXP3	Abcam	Cat#ab20034; RRID:AB_445284
Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Thermo Fisher Scientific	Cat#R37120; RRID:AB_2556548
Goat anti-Rat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 647	Thermo Fisher Scientific	Cat# A-21247; RRID:AB_141778
Goat anti-Rat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Thermo Fisher Scientific	Cat# A-11006; RRID:AB_2534074
Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 647	Thermo Fisher Scientific	Cat#A28181; RRID:AB_2536165
Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 405	Thermo Fisher Scientific	Cat#A-31553; RRID:AB_221604
Donkey anti-Goat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Thermo Fisher Scientific	Cat# A-11055; RRID:AB_2534102
Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 647	Thermo Fisher Scientific	Cat#A-21246; RRID:AB_2535814
Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 594	Thermo Fisher Scientific	Cat#A-11012; RRID:AB_2534079
Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Thermo Fisher Scientific	Cat#A-11008; RRID:AB_143165
Hs-OSMR-tv1-C2	ACDBio	Cat#445691-C2
Hs-IL10	ACDBio	Cat#602051
Hs-TNFA-C2	ACDBio	Cat#310421-C2
Hs-Best4	ACDBio	Custom probe
Hs-KRT19-C2	ACDBio	Custom probe
Hs-RSPON3	ACDBio	Custom probe
Hs-GREM2-C2	ACDBio	Cat#515591-C2
Hs-IL13RA2	ACDBio	Cat#546221
Hs-PLAU-C3	ACDBio	Cat#425001-C3
Hs-HLA-DRA	ACDBio	Cat#475891
Hs-OSM	ACDBio	Cat#456381
Hs-SOX8	ACDBio	Cat#538991
Hs-CCL20-C2	ACDBio	Cat#409611-C2
Hs-IL17A	ACDBio	Cat#310931
SlowFade Diamond Antifade Mountant with DAPI	Thermo Fisher Scientific	Cat#S36964

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Human colon biopsy tissue from healthy and colitis adults	Prospective Registry in Inflammatory Bowel Disease Study at Massachusetts General Hospital (MGH)	PRISM:2004P001067, Table S1
Human colon samples from tissue array	BioMAX	Cat#CO809a; Cat#CO245
Chemicals, Peptides, and Recombinant Proteins		
Recombinant Human IL-22	PrepoTech	Cat#200-22
B-27 Supplement	Thermo Fisher Scientific	Cat#12587-010
N-2 Supplement	Thermo Fisher Scientific	Cat#17502048
N-acetyl-1-cysteine	Sigma-Aldrich	Cat#A9165-5G
Y-276432 dihydrochloride monohydrate	Tocris	Cat#1254
Recombinant Human EGF	PeproTech	Cat#100-47
Matrigel	Corning	Cat#356231
Critical Commercial Assays		
RNAscope Multiplex Fluorescent Detection Kit v2	ACDBio	Cat#323110
Nextera XT Sample Preparation Kit	Illumina	Cat#FC-131-1096
10X Chromium Single Cell 3′ Kit	10X Genomics	Cat#120237
Deposited Data		
Genome Reference Consortium Mouse Build 38	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/
TcoF-DB v2 human transcription factors	Schaefer et al., 2011	http://compbio.massey.ac.nz/apps/tcof/home/
PRRDB human pattern recognition receptors	Lata and Raghava, 2008	http://crdd.osdd.net/raghava/prrdb/
KEGG pathways	Kanehisa et al., 2017	https://www.genome.jp/kegg/
MSigDB pathways	Liberzon et al., 2015	http://software.broadinstitute.org/gsea/msigdb/index.jsp
WikiPathways OSM signaling pathway	Slenter et al., 2018	https://www.wikipathways.org/index.php/WikiPathways
FANTOM5 receptor-ligand database	Ramilowski et al., 2015	http://fantom.gsc.riken.jp/5
Human colon scRNA-Seq FASTQ	This paper	Broad DUOS (https://duos.broadinstitute.org)
Human colon digital gene expression (DGE) matrix	This paper	Single Cell Portal: SCP259 (https://portals.broadinstitute.org/single_cell)
Experimental Models: Cell Lines		
L-WRN	ATCC	CRL-3276
Oligonucleotides		
Reverse Transcription DNA oligonucleotide primer (RNase-free, 100 mM) 5′ -AAGCAGTGGTATCAACGCAGAGTACT(30)VN-3′	IDT	N/A
SMARTER TSO (with LNA) 5′ -AAGCAGTGGTATCAACGCAGAGTACrGrG+G-3′	Exiqon	N/A
PCR oligonucleotide primer 5′ -AAGCAGTGGTATCAACGCAGAGT-3′	IDT	N/A
Software and Algorithms		
CellRanger v2.0	10X Genomics	https://github.com/10XGenomics/cellranger
sva (R package)	Leek et al., 2012	https://www.bioconductor.org/packages/devel/bioc/html/sva.html
Infomap clustering algorithm	Rosvall and Bergstrom, 2008	https://igraph.org/
Barnes-Hut t-SNE algorithm	van der Maaten and Hinton, 2008	https://cran.r-project.org/web/packages/Rtsne/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Scanpy	Wolf et al., 2018a	https://github.com/theislab/scanpy
Gephi	Bastian et al., 2009	https://gephi.org/
MAST	Finak et al., 2015	https://github.com/RGLab/MAST
Rtsne	CRAN	https://cran.r-project.org/web/packages/Rtsne/
PhenoGraph	Levine et al., 2015	https://github.com/jacoblevine/PhenoGraph
MAGIC	van Dijk et al., 2018	https://github.com/KrishnaswamyLab/MAGIC
nlme (R package)	CRAN	https://cran.r-project.org/web/packages/nlme/index.html
rsvd (R package)	CRAN	https://cran.r-project.org/web/packages/rsvd/index.html
Seurat (R toolkit)	Butler et al., 2018	https://satijalab.org/seurat
Analysis code	This paper	https://www.github.com/cssmillie/ulcerative_colitis

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, A.R. (aregev@broadinstitute.org). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patients and tissue samples

Biopsy samples were obtained from Crohn's disease (CD) patients, ulcerative colitis (UC) patients, and healthy individuals after informed consent and approval to the Prospective Registry in Inflammatory Bowel Disease Study at Massachusetts General Hospital (PRISM:2004P001067). Clinical information and metadata for the samples are provided in Table S1. Healthy controls were recruited at the time of routine colonoscopy. Healthy controls were individuals without a history of inflammatory bowel disease (IBD), a 1st degree relative with IBD, histories of autoimmune disease, immune mediated conditions, infectious colitis, and colon cancer, or a family history of colon cancer, and who were overall healthy with no other disease history (Table S1). UC patients were included based on having a clinical diagnosis of ulcerative colitis, and observed to have active disease via macroscopic assessment from a physician during an endoscopy. Two biopsies were obtained during endoscopy, using biopsy forceps that were used in standard of care. Each patient's biopsies were collected in a region determined by the scoping physician. Healthy individuals had two bites of endoscopically normal tissue, while UC patients had either (1) one non-inflamed and one inflamed region biopsied (15 patients; Table S1) or two adjacent non-inflamed and two adjacent inflamed biopsies to account for intra-patient variability (3 patients; Table S1). Biopsy bites were immediately placed into cryovials containing Advanced DMEM F-12 and placed on ice for transport.

For scRNA-seq, all biopsy samples were obtained from UC patients and healthy individuals, including both males and females (Table S1) while spanning a range of ages (20 – 77 years). For human spheroid cultures, biopsies were obtained from IBD patients (2 CD patients (right colon, males and non-smokers) and a UC patient (right colon, female and non-smoker)).

METHOD DETAILS

Single cell dissociation from fresh biopsies

Single-cell suspensions from collected biopsy bites were obtained using a modified version of a previously published protocol (Persson et al., 2013) as detailed below. Typically, two biopsies from the same patient were received directly in hand and processed in parallel with an average time from patient to loading on the 10X GemCode or Chromium platform of 2.5 total hours, and never exceeding 3.5 hours. While intact, biopsy bites were handled using a P1000 pipette applying gentle suction, and all centrifugation steps done in a temperature controlled 4°C centrifuge. Biopsy bites were first rinsed in 30 mL of ice-cold PBS (ThermoFisher 10010-049) and allowed to settle. Each individual bite was then transferred to 10 mL epithelial cell solution (HBSS Ca/Mg-Free [ThermoFisher 14175-103], 10 mM EDTA [ThermoFisher AM9261], 100 U/ml penicillin [ThermoFisher 15140-122], 100 μg/mL streptomycin [ThermoFisher 15140-122], 10 mM HEPES [ThermoFisher 15630-080], and 2% FCS [ThermoFisher 10082-147]) freshly supplemented with 200 μL of 0.5M EDTA. Separation of the epithelial layer from the underlying lamina propria was performed for 15 minutes at 37°C in a rotisserie rack with end-over-end rotation. The tube was then removed and placed on ice immediately for 10 minutes before shaking vigorously 15 times. Visual macroscopic inspection of the tube at this point yielded visible epithelial sheets, and microscopic examination confirmed the presence of single-layer sheets and crypt-like structures. The remnant tissue bite was carefully removed and placed into a large volume of ice-cold PBS to rinse before transferring to 5mL of enzymatic digestion mix (Base: RPMI1640, 100 U/ml penicillin [ThermoFisher 15140-122], 100 μg/mL streptomycin [ThermoFisher 15140-122], 10 mM HEPES [ThermoFisher 15630-080], 2% FCS [ThermoFisher 10082-147], & 50 μg/mL gentamicin [ThermoFisher 15750-060]), freshly supplement immediately before with 100 μg/mL of Liberase TM [Roche 5401127001] and 100 μg/mL of DNase I [Roche 10104159001]), at 37°C with 120 rpm rotation for 30 minutes. During this 30-minute lamina propria (LP) digestion, the epithelial (EPI) fraction was spun down at 400 g for 7 minutes and resuspended in 1 mL of epithelial cell solution before transferring to a 1.5mL Eppendorf tube in order to minimize time spent centrifuging and provide a more concentrated cell pellet. Cells were spun down at 800 g for 2 minutes and resuspended in TrypLE express enzyme [ThermoFisher 12604013] for 5 minutes in a 37°C bath followed by gentle trituration with a P1000 pipette. Cells were spun down at 800 g for 2 minutes and resuspended in ACK lysis buffer [ThermoFisher A1049201] for 3 minutes on ice to remove red blood cells, even if no RBC contamination was visibly observed in order to maintain consistency across samples. Cells were spun down at 800 g for 2 minutes and resuspended in 1 mL of epithelial cell solution and placed on ice for 3 minutes before triturating with a P1000 pipette and filtering into a new Eppendorf tube through a 40 μM cell strainer [Falcon/VWR 21008-949]. Cells were spun down at 800 g for 2 minutes and then resupended in 200 μL of epithelial cell solution and placed on ice while final steps of LP dissociation occurred. After 30 minutes, the LP enzymatic dissociation was quenched by addition of 1ml of 100% FCS [ThermoFisher 10082-147] and 80 μL of 0.5M EDTA and placing on ice for five minutes. Samples were typically fully dissociated at this step and after gentle trituration with a P1000 pipette filtered through a 40 μM cell strainer into a new 50 mL conical tube and rinsed with PBS to 30 mL total volume. This tube was spun down at 400 g for 10 minutes and resuspended in 1 mL of ACK and placed on ice for 3 minutes. Cells were spun down at 800 g for 2 minutes and resuspended in 1 mL of epithelial cell solution and spun down at 800 g for 2 minutes and resuspended in 200 μL of epithelial cell solution and placed on ice.

Human spheroid cultures for IL-22-enterocyte interaction validation

Human biopsies from IBD patients (i.e., 2 CD patients and 1 UC patient, see above) were collected for spheroid culture. Each individual bite was minced and then transferred to 10 mL epithelial cell solution (HBSS Ca/Mg-Free [ThermoFisher 14175-103], 8 mM EDTA [ThermoFisher AM9261], 100 U/ml penicillin [ThermoFisher 15140-122], 100 μg/mL streptomycin [ThermoFisher 15140-122], 10 mM HEPES [ThermoFisher 15630-080]). Separation of the epithelial layer from the underlying lamina propria was performed for 40 minutes at 4°C in a rotisserie rack with end-over-end rotation. The tube was then removed and placed on ice immediately for 10 minutes before shaking vigorously 15 times. Visual macroscopic inspection of the crypt-like structures was performed. Crypt like structures were spun down at 200 g for 3 minutes and washed twice with cold PBS and subsequently were resuspend with Matrigel. Cells were then seeded in 24 well plates and grow with 50% L-WRN media (50% base –Advanced DMEM/F12 [GIBCO 12634-010] + 10% FBS, P/S, GluM, HEPES) +10uM Y27632 [TOCRIS 1254] +10uM SB 431542 [TOCRIS 1614] for 3 days before first splitting. Media was then changed every 2 days.

For IL-22 stimulation, spheroids were grown for 3 days and then split 1:3 with fresh media containing 20ng/ml recombinant human IL-22 [Peprotech 200-22] or mock. After 3 days, spheroids were collected and subjected to bulk RNA-Seq with the SMART-Seq2 protocol (Picelli et al., 2014).

Droplet-based scRNA-Seq

Single cells were processed through the GemCode Single Cell Platform per manufacturer's recommendations using the GemCode Gel Bead, Chip and Library Kits (V1) or single-cell suspensions were loaded onto 3′ library chips as per the manufacturer's protocol for the Chromium Single Cell 3′ Library (V2 and V3) (10X Genomics; PN-120233) (Table S1). Briefly, single cells were partitioned into Gel Beads in Emulsion (GEMs) in the GemCode or Chromium instrument with cell lysis and barcoded reverse transcription of RNA, followed by amplification, shearing (for V1) or enzymatic fragmentation (for V2 and V3) and 5′ adaptor and sample index attachment. Each biopsy bite was sequenced on two channels of the 10X GemCode or Chromium Single Cell Platform, one for the epithelial fraction and the other for the lamina propria fraction in order to recover sufficient numbers of epithelial and lamina propria cells for downstream analyses. An input of 6,000 single cells was added to each channel with a recovery rate of approximately 2,000 cells. Libraries were sequenced on an Illumina Nextseq or Hi-Seq (Table S1).

SMART-Seq2 for sequencing of human colon spheroids

Libraries were prepared using a modified SMART-Seq2 protocol as previously reported (Picelli et al., 2014). RNA lysate cleanup was performed using RNAClean XP beads [Agencourt], followed by reverse transcription with Maxima Reverse Transcriptase [Life Technologies] and whole transcription amplification (WTA) with KAPA HotStart HIFI 2 × ReadyMix [Kapa Biosystems] for 16 cycles. WTA products were purified with Ampure XP beads [Beckman Coulter], quantified with Qubit dsDNA HS Assay Kit [ThermoFisher], and assessed with a high sensitivity DNA chip [Agilent]. RNA-Seq libraries were constructed from purified WTA products using Nextera XT DNA Library Preparation Kit [Illumina, FC-131-1096]. The population and no-cell controls were processed using the same method. The libraries were sequenced on an Illumina MiSeq.

Immunofluorescence assay (IFA)

Staining of human colon samples from tissue array of inflamed and healthy individuals (TMA, US BioMAX, #CO809a and #CO245) was conducted as previously described (Biton et al., 2011). Sections were deparaffinized with standard techniques, incubated with primary antibodies overnight at 4°C, and then incubated with secondary antibodies at room temperature for 30 min. Slides were mounted with Slowfade Mountant+DAPI (Life Technologies, S36964) and sealed.

Single-molecule fluorescence *in situ* hybridization (smFISH)

RNAScope Fluorescent Multiplex and RNAScope Multiplex Fluorescent v2 (Advanced Cell Diagnostics) were used per manufacturer's recommendations with the following alterations. Target retrieval boiling time was adjusted to 12 minutes and incubation with Protease IV at 40°C was adjusted to 15 minutes. Slides were mounted with Slowfade Mountant+DAPI (Life Technologies, S36964) and sealed.

Combined IFA and smFISH

Combined IFA and smFISH was implemented by first performing smFISH and then IFA, as described above, with the following alterations. After horseradish peroxidase (HRP) enzyme blocking, tissue sections were washed in washing buffer, incubated with primary antibodies overnight at 4°C, washed in 1x TBST 3 times and then incubated with secondary antibodies for 30 min at room temperature. Slides were mounted with Slowfade Mountant+DAPI (Life Technologies, S36964) and sealed.

Imaging of tissue sections

Images of tissue sections were taken with a confocal microscope Fluorview FV1200 using Kalman filtering and sequential laser emission to reduce noise and signal overlap. Scale bars were added to each image using the confocal software FV10-ASW 3.1 Viewer. Images were overlaid and visualized using ImageJ software (Schneider et al., 2012).

Antibodies and RNA smFISH probes

Antibodies used for immunofluorescence

Mouse anti-EPCAM (1:500, ThermoFisher MA1-06502), mouse anti-Vimentin (1:500, Millipore MAB3400), mouse anti-CD19 (1:100, BioLegend 302201), rat anti-CD45 (1:200, ThermoFisher MA5-17687), goat anti-CD138 (1:100, R&D Systems AF2780), mouse anti-HLA-DR/DP/DQ (1:200, ThermoFisher MA1-25914), mouse anti-CD11c (1:100, BD Biosciences 550375), goat anti-CD4 (1:100, R&D Systems AF-379-NA), rabbit anti-CD8 (1:100, Invitrogen SP16). Alexa Fluor 488-, 594-, and 647-conjugated secondary antibodies were used (Life Technologies).

Human probes used for single-molecule FISH with RNAscope (Advanced Cell Diagnostics):

BEST4 (C1), KRT19 (C2), RSPON3 (C1), GREM2 (C2), IL13RA2 (C1), PLAU (C3), HLA-DRA (C1), OSM (C1), HLA-DPB1 (C2), OSMR (C2), SOX8 (C1), CCL20 (C2), IL17A (C1).

QUANTIFICATION AND STATISTICAL ANALYSES

Processing FASTQ reads into gene expression matrices

Cell Ranger v2.0 was used to demultiplex the FASTQ reads, align them to the hg19 human transcriptome, and extract their "cell" and "UMI" barcodes. The output of this pipeline is a digital gene expression (DGE) matrix for each sample, which records the number of UMIs for each gene that are associated with each cell barcode. DGE matrices were filtered to remove low quality cells, defined as cells in which fewer than 250 different genes were detected. This cutoff was determined empirically: higher cutoffs led to disproportionate filtering of mast and T cells, whereas lower cutoffs did not affect the cell type distribution, but did reduce overall data quality. To account for differences in sequencing depth across cells, UMI counts were normalized by the total number of UMIs per cell and converted to transcripts-per-10,000 (henceforth "TP10K").

Cell clustering overview

To cluster single cells into distinct cell subsets, we followed the general procedure outlined in Haber et al. (2017) with additional modifications. This workflow includes the following steps: partitioning cells into epithelial, stromal, and immune compartments, followed by clustering the cells within each compartment, which entails the selection of "variable" genes, batch correction, dimensionality reduction (PCA), and graph clustering. Each step of this workflow is detailed below.

Partitioning cells into epithelial, stromal, and immune compartments

Cells were partitioned into epithelial, stromal, and immune compartments based on the expression of known marker genes. First, we clustered the cells *within each sample* by their gene expression profiles (with the clustering procedure below). The clusters were scored for the following gene signatures: epithelial cells (*EPCAM, KRT8, KRT18*), stromal cells (*COL1A1, COL1A2, COL6A1, COL6A2, VWF, PLVAP, CDH5, S100B*), and immune cells (*CD52, CD2, CD3D, CD3G, CD3E, CD79A, CD79B, CD14, CD16, CD68, CD83, CSF1R, FCER1G*). Signature scores were calculated as the mean $\log_2$(TP10K+1) across all genes in the signature. Each cluster was assigned to the compartment of its maximal score and all cluster assignments were manually inspected to ensure the accurate segregation of cells. Finally, the cells within each compartment were assembled into three DGE matrices, comprising all epithelial cells, all stromal cells, and all immune cells.

Variable gene selection

To identify variable genes within a sample, we first calculated the mean (μ) and the coefficient of variation (CV) of expression of each gene. Genes were then grouped into 20 equal-frequency bins (ventiles) according to their mean expression levels. LOESS regression was used to fit the relationship, log(CV) $\sim$log(μ), and the 1,500 genes with the highest residuals were evenly sampled across these expression bins. To extend this approach to multiple samples, we performed variable gene selection separately for each sample to prevent "batch" differences between samples from unduly impacting the variable gene set. A consensus list of 1,500 variable genes was then formed by selecting the genes with the greatest recovery rates across samples, with ties broken by random sampling. This consensus gene set was then pruned through the removal of all ribosomal, mitochondrial, immunoglobulin, and HLA genes, which were found to induce unwanted batch effects in some samples in downstream clustering steps.

Batch correction

We observed substantial variability between cells that had been obtained from different human subjects, which likely reflects a combination of technical and biological differences. In some cases, these "batch effects" led to cells clustering first by patient or disease phenotype, rather than by cell type or cell state. To eliminate these batch differences, we ran ComBat (Johnson et al., 2007) with default parameters on the $\log_2$(TP10K+1) expression matrix, allowing cells to be clustered by cell type or cell state. Importantly, these batch-corrected data were only used for the PCA and all steps relying on PCA (e.g., clustering, diffusion map, t-SNE visualization); all other analyses (e.g., differential expression analysis) were based on the original expression data.

Comparison of batch correction methods

We compared ComBat to two other batch correction methods that were designed specifically for scRNA-Seq data: MultiCCA (Butler et al., 2018) and LIGER (Welch et al., 2018). Both methods were run on the $\log_2$(TP10K+1) expression data for cells from the epithelial, stromal, and immune compartments, using default parameters with n = 20 components to match the original analysis. Following batch correction, cell embeddings were visualized using the Barnes-Hut t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm with default parameters. To visualize congruence between methods, we then projected the cell subsets that were originally defined using the ComBat-transformed data (i.e., Figure 1C) onto the t-SNE coordinates calculated using each of the other methods. Inspection of the batch correction methods revealed that ComBat performed well in comparison to the other methods, in agreement with a recent comparison of batch correction methods (Büttner et al., 2019).

Dimensionality reduction, graph clustering, and t-SNE visualization

Cells were clustered at two stages of the analysis: first, to initially partition the cells within each sample into epithelial, stromal, and immune compartments (single sample clustering), and second, to cluster cells from multiple samples into distinct subsets (multi-sample clustering).

For single-sample clustering, we first ran low-rank PCA on the variable genes of the entire $\log_2$(TP10K+1) expression matrix (as no consensus list needs to be generated). The Infomap graph clustering algorithm (Rosvall and Bergstrom, 2008) was then applied to the *k*-nearest neighbor (*k*-NN) graph defined using PCs 1 to 20 and $k = 50$ nearest neighbors. These parameters were chosen to "over-cluster" the cells, ensuring that cells from distinct compartments were not grouped together.

In contrast, for multi-sample clustering, we ran low-rank PCA on the variable genes of the batch-corrected expression matrix, chosen as described above. We then applied Phenograph (Levine et al., 2015) to the *k*-NN graph defined using PCs 1 to 20 and a varying *k*, which was selected through close inspection of the data (see "Selecting the number of nearest neighbors for graph clustering"): $k = 750$ for epithelial cells, $k = 250$ for stromal cells, and $k = 250$ for immune cells. Although most clusters were stable over a range of *k*, some rare epithelial subsets, such as tuft cells and M cells, were initially merged with larger clusters. We therefore re-clustered the epithelial cells with fewer neighbors ($k = 50$) to achieve higher granularity in the clusters and added clusters corresponding to *BEST4*$^+$ enterocytes, enteroendocrine cells, and M cells to the original set of clusters. Additionally, we partitioned the immune cells into myeloid, B cell, and T cell compartments based on DE genes within each cluster, and repeated the clustering using the *k*-NN graphs defined with PCs 1 to 15 and $k = 50$ for myeloid cells, $k = 100$ for B cells, and $k = 100$ for T cells. After clustering the cells, we merged pairs of clusters that were separated by fewer than 5 differentially expressed (DE) genes with AUC > 0.60, a permissive cutoff that merges only highly similar clusters. Finally, the Barnes-Hut t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm was run on the PCs with perplexity = 20 and for 10,000 iterations to produce two-dimensional embeddings of the data for visualization (Figure 1C).

Selecting the number of nearest neighbors for graph clustering

To select the number of nearest neighbors, *k*, for clustering, we examined a range of choices (typically $k = 25$, 50, 100, 250, 500, and 750). In general, we tried to select a *k* yielding the highest granularity clusters that were still biologically distinct, as determined through inspection of their marker genes. We also looked at the stability of cluster assignments over the full parameter range and tried to select a *k* yielding stable and well-resolved clusters. Therefore, the final choice of *k* reflects both data-driven clustering and expert knowledge. To ensure that all transcriptionally distinct cell subsets were identified, we sub-clustered each cell subset and identified those sub-clusters that were supported by discriminative differentially expressed genes (see "Identifying transcriptionally distinct sub-clusters").

Identifying transcriptionally distinct sub-clusters

To systematically determine whether transcriptionally distinct subpopulations of cells may exist, we sub-clustered each cell subset using its k-NN graph defined using PCs 1 to 10 and $k = 250$ (see "Dimensionality reduction, graph clustering, and t-SNE visualization"). These parameters were selected to yield relatively few sub-clusters, such that that any of the major differences would be found. We then searched for differentially expressed genes that could accurately distinguish between the cells in each sub-cluster versus all other cells with an area under the curve (AUC) exceeding 0.75. Sub-clusters containing highly discriminative marker genes were then flagged for further analysis (Table S2).

Comparison of training and test sets by a classification based approach

To compare the training and test sets, we first trained a Random Forest classifier to predict the subset of each cell, using the cell subsets that were originally defined from the training set (Figure 1C). The Random Forest was trained separately for cells in the epithelial, stromal, and immune compartments, with the input data constructed as follows. First, we used ComBat (Johnson et al., 2007) to generate a batch-corrected $\log_2$(TP10K+1) expression matrix containing the variable genes (X_1). Second, we performed PCA on this matrix to obtain a 20-dimensional embedding of the batch-corrected data (X_2). Third, we used LIGER to compute a separate batch-corrected 20-dimensional embedding of the data (X_3). These three matrices (i.e., X_1, X_2, and X_3) were then combined to form the input feature matrix, X. We note that ComBat and LIGER provide similar but complementary descriptions of the data and both sets of features were deemed important by the Gini Importance measure (data not shown). The Random Forest was trained with 1,000 trees and default parameters, except that in order to account for class imbalances, we weighted each class by the inverse of its class frequency.

Across all cells, the "out of bag" error, which provides an unbiased estimate of the test error, was 10.7%, suggesting this model can be used to accurately infer subsets for the test data. Classification errors were mostly between similar cell subsets (Figure S1C), such as *CD69*$^-$ and *CD69*$^+$ mast cells, with two major exceptions. Our model performed less well for cycling immune cells, which are composed of cells from many different types, and *CD8*$^+$*IL-17*$^+$ T cells.

Next, to confirm that this model can be applied to the test dataset, we co-embedded the single cells from both datasets (see "Dimensionality reduction, graph clustering, and t-SNE visualization") and labeled them according to their predicted cell subsets (Figure S1D). Following removal of doublets from the test dataset (see "Doublet removal"), the two datasets aligned well with no major incongruences. In some cases, the merged dataset contained increased sub-cluster resolution (e.g., immature and mature *BEST4*$^+$ enterocytes versus only one cluster of *BEST4*$^+$ enterocytes), due to an increase in cell number, or discernible patient-specific sub-clusters, but the classifier accurately classified these cells into their larger groups.

Comparison of intra- versus inter-individual variability

To assess levels of biological and technical variation in our scRNA-Seq profiles, we analyzed the epithelial and lamina propria compartments of 18 replicate biopsies collected from the same individual (12 healthy, 3 non-inflamed, 3 inflamed). For each pair of samples, we measured the Pearson correlation between their logit-transformed cell proportions, as well as their mean gene expression levels. We then compared intra-individual and inter-individual correlations across healthy, non-inflamed, and inflamed tissue regions (Figure S1E).

Doublet removal

Following the initial clustering, we removed all clusters consisting of likely cell doublets from epithelial, stromal, myeloid, B cell and T cell compartments, then repeated the steps outlined in "dimensionality reduction, graph clustering, and t-SNE visualization." Doublets were identified through expert annotation of the marker gene lists for each cell cluster and corresponded to clusters with markers from distinct lineages (e.g., clusters with B cell and T cell markers). However, within the immune compartment, cells from distinct lineages sometimes clustered together (e.g., cycling B cells and cycling T cells) and these cells were then separated back into their source lineages. Doublet removal was therefore an iterative process alternating between removing doublets, assigning cells to the correct compartments, and graph clustering and t-SNE visualization.

Cell lineage dendrogram

As an auxiliary tool, cell subsets were manually organized on a dendrogram reflecting known lineage relationships (Figure 1D, top). This tree is organized as follows. Under epithelial cells we split Absorptive and Secretory subtrees. The Absorptive subtree included further subtrees for Transit Amplifying (TA) cells (Absorptive TA 1, Absorptive TA 2), Immature cells (Immature Enterocytes 1, Immature Enterocytes 2, Enterocyte Progenitors), and Mature cells (Enterocytes, *BEST4*$^+$ Enterocytes). The Secretory subtree included subtrees for progenitor cells (Secretory TA, Immature Goblet) and for mature cells (Goblet, Tuft, and Enteroendocrine). Stem cells, Cycling TA cells, and M cells were placed directly under the node corresponding to all epithelial cells. The Stromal subtree had Fibroblast, Endothelial, and Glial subtrees. Fibroblasts were subdivided into *WNT2B*$^+$ (*WNT2B*$^+$*Fos*^{hi}, *WNT2B*$^+$*Fos*^{lo} 1, *WNT2B*$^+$*Fos*^{lo} 2, *RSPO3*$^+$), *WNT5B*$^+$ (*WNT5B*$^+$ 1, *WNT5B*$^+$ 2), inflammatory fibroblast, and myofibroblast subtrees. The Endothelial cell subtree included branches for Endothelial, Microvascular, Post-capillary venules, and Pericytes. The Immune subtree was partitioned into myeloid and lymphoid lineages. Myeloid cells included subtrees for Mast cells (*CD69*$^+$ Mast, *CD69*$^-$ Mast) and Monocytes (Macrophages, Cycling Monocytes, Inflammatory Monocytes, and DCs (DC1s, DC2s)). Lymphoid cells included subtrees for NK cells, ILCs, B cells, T cells (subdivided into *CD4*$^+$ T cells (*CD4*$^+$ Activated *Fos*^{lo}, *CD4*$^+$ Activated *Fos*^{hi}, *CD4*$^+$ Memory, T$_{regs}$, *PD1*$^+$, MT^{hi}) and *CD8*$^+$ T cells (*CD8*$^+$ IELs, *CD8*$^+$ LP, *CD8*$^+$*IL-17*$^+$, and *CD8*$^+$ Cycling)), and B cells (Plasma cells, Follicular (FO) B cells, Germinal Center (GC) B cells, and Cycling B cells).

Scoring samples for inflammation-associated genes

To validate our endoscopic assessments of tissue inflammation, we constructed a gene signature of the following inflammation-associated genes: *IFNG, IFNGR1, IFNGR2, IL10, IL12A, IL12B, IL12RB1, IL12RB2, IL13, IL17A, IL17F, IL18, IL18R1, IL18RAP, IL1A, IL1B, IL2, IL21, IL21R, IL22, IL23A, IL23R, IL2RG, IL4, IL4R, IL5, IL6, JUN, NFKB1, RELA, RORA, RORC, S100A8, S100A9, STAT1, STAT3, STAT4, STAT6, TGFB1, TGFB2, TGFB3,* and *TNF*. We separately scored EPI and LP samples for these signatures, then combined these measurements by calculating their mean z-scores. P values between healthy and non-inflamed or inflamed samples were computed using a one-sided Wilcoxon test. P values between non-inflamed and inflamed samples were computed using a one-sided paired Wilcoxon test within each subject.

Epithelial cell differentiation

The diffusion map and diffusion pseudotime (Figure 2D) for epithelial cells were estimated with Scanpy v0.4.2 (Wolf et al., 2018a) on the $\log_2$(TP10K+1) expression matrix, with the following parameters: *n_pcs* = 20, *n_neighbors* = 30, *n_dcs* = 20, *n_branchings* = 1, *min_group_size* = 0.001. The differentiation map (Figure 1F) was estimated using the partition-based graph abstraction (PAGA) method implemented in Scanpy v1.4 (Wolf et al., 2018b) using the same parameters. In both cases, one of the *LGR5*$^+$ISCs was randomly selected as the root cell. To identify significant changes in epithelial cell differentiation with UC, we estimated diffusion pseudotimes separately for absorptive and secretory cells, and used mixed effects models to assess significance (see "Identifying significant changes in gene signatures and pseudotime during disease").

Estimation of cell proportions

Because EPI and LP samples were separately processed and sequenced, cell proportions estimated from each sample type are not directly comparable. Therefore, rather than combining the cell subset proportions from different sample types (e.g., using a weighted mean across EPI and LP samples), we determined for each cell subset whether it was EPI-associated or LP-associated and calculated its proportions using only samples of that type. As expected, EPI samples mostly consisted of epithelial cells (89% ± 15% epithelial cells on average) with some tissue-resident immune cells, such as *CD69*$^-$ mast cells, *CD8*$^+$ IELs, and *CD8*$^+$*IL-17*$^+$ T cells, whereas LP samples primarily contained immune and stromal cells (84% ± 18% immune and stromal cells on average).

Identifying statistically significant differences in cell proportions

To identify changes in cell proportions between healthy, non-inflamed, and inflamed tissue, we used multiple statistical tests that each capture distinct but complementary types of information: (*1*) a Dirichlet-multinomial regression, (*2*) a Fisher's exact test, and (*3*) a Mann-Whitney test. We describe each of these below. A major concern with the comparison of cell proportions in scRNA-Seq data is that they are not independent of each other. Because all proportions sum to 1, an increase in the proportion of one cell subset will necessarily lead to a decrease in the proportions of other cell subsets. To account for these dependencies, we used a Dirichlet-multinomial regression model, which tests for differences in cell composition between disease states (e.g., inflamed versus healthy), while accounting for the proportions of *all of the other* cell subsets. This regression model and its associated p values were calculated using the "DirichReg" function in the DirichletReg R package. Because this is a multivariate test, its results may at times appear counter-intuitive and may not be congruent with univariate tests, such as a t test, which examine each cell subset independently. We therefore also performed a Fisher's exact test on the numbers of cells from each subset that were isolated from non-inflamed or inflamed specimens versus healthy specimens. This test reflects how enriched each cell subset is in each disease state, but does not account for the sample from which each cell was isolated. Therefore, we also performed a non-parametric Mann-Whitney test on the proportions of each cell subset in non-inflamed or inflamed specimens versus healthy specimens.

Comparison of IgA$^+$ and IgG$^+$ plasma B cells

The mean $\log_2$(TP10K+1) expression levels of the IgA heavy chain genes (*IGHA1, IGHA2*) and IgG heavy chain genes (*IGHG1, IGHG2, IGHG3, IGHG4*) were scored across all plasma cells. After examining the distribution of these scores, we empirically determined that an expression cutoff corresponding to $\log_2$(TP10K+1) = 6 accurately discriminated among IgA$^+$ and IgG$^+$ cells. In total, 94% of all plasma cells were classified as either IgA$^+$ or IgG$^+$, with only 0.2% classified as IgA$^+$IgG$^+$ "double positive" cells (likely corresponding to doublets).

Downsampling single cells for mean expression analysis

To facilitate downstream analyses, a separate dataset was constructed containing 50,375 down-sampled cells. These data were used solely for the estimation of the expression distribution within cell subsets, but all other analyses were based on the full dataset. We verified that the mean expression levels from the full and down-sampled datasets were strongly correlated across cell subsets (mean Pearson's $r = 0.999$). To down-sample cells, we first calculated the number of cells obtained from every cell subset in each sample. We determined a fixed number of cells to retain from each of these subset-sample groups (purposefully not preserving their original proportions) that would yield approximately 50,000 cells in the down-sampled dataset. Finally, the highest quality cells (measured by the number of genes per cell) were retained from each of these groups. Using this method, samples and cell subsets with relatively many cells (e.g., plasma cells) were heavily down-sampled, whereas samples and cell subsets with relatively few cells (e.g., tuft cells) were largely retained intact.

Differential expression analysis

Differential expression (DE) tests were performed using MAST (Finak et al., 2015), which fits a hurdle model to the expression of each gene, consisting of logistic regression for the zero process (i.e., whether the gene is expressed) and linear regression for the continuous process (i.e., the expression level). To reduce the size of the inference problem, separate models were fit for each level of the cell tree (see "Cell lineage dendrogram," above), comparing cells within the given group to all other cells (e.g., ISCs versus non-ISCs). The regression model includes terms to capture the effects of the cell subset and the disease state on gene expression, while controlling for cell complexity (i.e., the number of genes detected per cell).

Specifically, we used the regression formula, $Y_i \sim X + D + N$, where Y_i is the standardized $\log_2$(TP10K+1) expression vector for gene i across all cells, X is a binary variable reflecting cell subset membership (e.g., ISCs versus non-ISCs), D is the disease state associated with each cell, and N is the number of genes detected in each cell. Overall, we fit three types of DE models, which varied by the encoded disease states: (1) to identify cell subset markers and DE genes in UC patients relative to healthy controls, we used three disease states: Healthy, UC non-inflamed, and UC inflamed; (2) to identify DE genes between non-inflamed and inflamed patient samples, we used two disease states: UC non-inflamed and UC inflamed; and (3) to identify genes that are specific to cell subsets in healthy subjects and UC patients, we used two disease states: Healthy and UC. Additionally, a few heuristics were used to increase the speed of the tests: we required all tested genes to have a minimum fold change of 1.2 and to be expressed by at least 1% of the cells *within* the group of interest, and cells were evenly downsampled across groups so that a maximum of 2,500 cells were tested for each cell subset. In all cases, the discrete and continuous coefficients of the model were retrieved and p values were calculated using the likelihood ratio test in MAST. Q-values were separately estimated for each cell subset comparison using the Benjamini-Hochberg correction. Unless otherwise indicated, all reported DE coefficients and q-values correspond to the discrete component of the model (i.e., the logistic regression).

Estimation of the droplet contamination rate and filtering of putative ambient RNA contaminants

Droplets encapsulate single cells with small portions of the extracellular environment, leading to low but persistent levels of contamination by ambient RNA (Macosko et al., 2015). To correct for this, we explicitly modeled droplet contamination. First, we partitioned individual cells into the following groups: epithelial cells, fibroblasts, endothelial cells, myeloid cells, B cells, and T cells. We reasoned that each group should uniquely express a subset of genes that are not found in other cells; for example, B cells uniquely express *IGHA1* and T cells uniquely express *CD3D*. Therefore, the off-target expression of such genes in the incorrect group (*e.g. IGHA1* expression in T cells) should reflect contamination rather than intrinsic gene expression. Moreover, we hypothesized that the levels of such off-target gene expression could serve as an accurate indicator of contamination rates in the entire dataset. To test this hypothesis, we compared the mean expression levels of genes within each group (i.e., in-group expression) to a weighted mean of their expression levels in all other cells (i.e., out-group expression), which is a proxy for the composition of extracellular RNA (e.g., B cells versus non-B cells, Figure S1F, see "Normalization and scaling of expression levels for contamination filtering" below for additional details). As expected, known markers for cell groups were enriched at the edges of the point distribution, where differences between in-group and out-group expression were greatest. For example, known B cell markers were enriched on the left edge of the point distribution (*e.g. IGHA1* and *IGJ*, Figure S1F), while markers for other cell types were enriched on the right edge, likely reflecting contamination (*e.g. CD3D* and *TPSAB1*, Figure S1F). We noticed two other patterns yielding insights into contamination: (1) genes with sufficiently high out-group expression *always* had non-zero in-group expression, and (2) there is a linear relationship between in-group and out-group expression levels, particularly for contaminants on the right edge of the point distribution (Figure S1F). Taken together, these observations suggest that contamination uniformly affects all genes and that the overall levels of contamination for each gene are proportional to its representation in the extracellular RNA pool.

Therefore, to estimate the contamination rate for each cell group, we fit a robust linear model to the genes on the right edge of the point distribution, whose expression is almost entirely driven by contamination. Surprisingly, the fitted models were nearly identical across groups (slope = 1.33 ± 0.07, intercept = −7.22 ± 0.33) and we constructed a consensus model using the mean slope and mean intercept. This model corresponds to a contamination rate between 0.5% and 5% of the total RNA pool in each sample. We used this model to identify potential contaminants in all cell subsets by conservatively flagging genes with residuals < 5 (i.e., 32-fold increase over the estimated contamination rate) and genes in each cell subset whose expression did not exceed 1% of its total expression across all cells. This approach filtered out nearly all identifiable contamination, assessed by manual inspection of the filtered and unfiltered gene lists.

Normalization and scaling of expression levels for contamination filtering

The composition of extracellular RNA is different for each sample; for example, EPI samples have high levels of *MUC2*, while LP samples have high levels of *IGHA1*. Any attempt to identify droplet contamination should therefore account for the distribution of samples that cells were isolated from. For example, the expression of genes in B cells (i.e., in-group expression) should be compared to their pooled expression levels in non-B cells (i.e., out-group expression) using the *same* samples that the B cells were recovered from. Thus, rather than using a simple mean to measure the in-group and out-group expression levels for a gene, we used a weighted mean of its expression in each sample, where the weights were determined as the fraction of in-group cells belonging to that sample. More specifically, the in-group expression of gene i for cell group q is:

$$I_{iq} = \sum_j w_{qj} \cdot \frac{1}{T_j} \sum_{k \in q_j} x_{ik}$$

where x_{ik} is the expression level of gene i in cell k, q_j is the set of all cells that were isolated from sample j that belong to cell group q, T_j is the total number of cells in sample j, and w_{qj} is the weight for cell group q in sample j. Similarly, the out-group expression of gene i for cell group g is:

$$O_{ig} = \sum_j w_{qj} \cdot \frac{1}{T_j} \sum_{k \in Q_j} x_{ik}$$

where Q_j is the set of all cells that were isolated from sample j that do not belong to cell group q. The weight for cell group q in sample j, w_{qj}, is equal to the proportion of cells from cell group q that were isolated from sample j:

$$w_{qj} = \frac{|q_j|}{\sum_k |q_k|}$$

Importantly, the normalization factor, T_j, normalizes the expression to the total number of cells in the sample, ensuring that expression levels are comparable across cell groups.

Gene specificity

For each expressed gene, we tested whether that gene was specific to any cell subset (e.g., T_{reg} cells) or any node of the cell hierarchy dendrogram (*e.g. CD4*$^+$ T cells). We defined a gene as specific to a cell group if it was significantly (i.e., adjusted p value < 0.05) and positively differentially expressed in all pairwise comparisons to non-overlapping cell subsets and its mean expression level within the group was at least 2-fold higher than its mean expression in all non-overlapping cell subsets. In addition, we searched for cases where a gene gained, lost, or changed its cell specificity between health and UC. Note that a change in gene specificity may, however, simply reflect the gain or loss of statistical power, rather than a statistically significant change in gene expression. Therefore, to confirm that a gene was no longer specific to a cell subset in a given cohort (i.e., healthy subjects or UC patients), we required that another cell subset have significantly greater expression of the target gene within that cohort.

Scoring gene signatures and identifying significant changes between health and disease

To prevent highly expressed genes from dominating a gene signature score, we scaled each gene of the $\log_2$(TP10K+1) expression matrix by its root mean squared expression across all cells (using the ‘scale’ function in R with center = FALSE). The signature score for each cell was then computed as the mean scaled expression across all genes in the signature. To identify statistically significant changes in gene signature expression within each cell subset, we compared the change in expression of the gene signature to a null distribution that was estimated from 100 background sets of genes. Each background gene set was selected to have matching expression levels, using 20 equal-frequency expression bins that were defined using the healthy cells within the cell subset. Mixed effects models were used to identify significant changes in background-adjusted expression levels (see “Identifying significant changes in gene signatures and pseudotime with disease”).

Estimation of false discovery rate

Unless otherwise specified, false discovery rates were estimated with the Benjamini and Hochberg correction (Benjamini and Hochberg, 1995), using the “p.adjust” R function with the “fdr” method.

Identifying significant changes in gene signatures and pseudotime with disease

To identify significant changes in diffusion pseudotime (Figure 2D) or in the expression levels of gene signatures (Figure 4B) with disease, we used mixed linear models, which account for the uneven distribution of cells across samples. Mixed linear models were implemented using the "lme" function in the "nlme" R package, using a fixed effect term for disease state (i.e., healthy, non-inflamed, or inflamed) and a random intercept that varies with each sample: $Y_i \sim D + (1 \mid S)$, where Y_i is the vector of covariate i values across cells, D is the disease state associated with each cell, and S is the sample that each cell was isolated from. P values for the fixed terms were estimated with the "anova.lme" function.

Acquisition of gene sets

Human transcription factors were obtained TcoF-DB v2 (Schaefer et al., 2011). Human G-protein coupled receptors were obtained from UniProtKB (search term: family = "g protein coupled receptor," reviewed = "yes," organism = "Homo sapiens (Human) [9606]") (The UniProt Consortium, 2018). Human transporters were obtained from UniProtKB (search term: keyword = "Transport [KW-0813]," reviewed = "yes," organism "Homo sapiens (Human) [9606]"). Human pattern recognition receptors were obtained from PRRDB (Lata and Raghava, 2008) and supplemented with Human Gene Nomenclature Committee (HGNC) "C-type lectin domain containing" gene family members (Yates et al., 2017). Human cytokines were obtained from the KEGG pathway for "Cytokine-cytokine receptor interaction" (Kanehisa et al., 2017).

Acquisition of gene signatures

All pathways related to metabolism, inflammation, and stress were obtained from KEGG (Kanehisa et al., 2017), except in the following cases, for which the pathways were not found in KEGG: IFN-α, IFN-γ, and IL-2/Stat5 pathways were obtained from MSigDB (Liberzon et al., 2015) and the OSM pathway was obtained from WikiPathways (Slenter et al., 2018). The T cell signatures for cytotoxicity (*GNLY, GZMB, GZMK, IFNG, NKG7*), co-inhibition (*CTLA4, PDCD1, TIGIT, HAVCR2, LAG3, BTLA, PDPN, CD160, GP49A, LILRB4, CD274, CD200, CD244, PILRA, SIRPB1, LAIR1, CEACAM1, KLRA7, KLRA3. KLRA9, PTGER4, KLRD1, KLRC1, PROCR)*, and co-stimulation *(CD28, CD226, TNFRSF4, TNFRSF9, ICOS, CD27, TNFSF14, CD80, TNFSF4, CD86, TNFSF11, CD276, CD40LG, TNFRSF18*) were derived from known markers. Gene signatures for resistance and susceptibility to anti-TNF blockade were obtained from a meta-analysis of 60 responders and 57 non-responders (Wang et al., 2016). The gene signature for poor colorectal cancer prognosis was obtained from a meta-analysis of thousands of CRC patients, and the top 25 genes in the signature were used (Calon et al., 2015).

Acquisition of microarray and bulk RNA-Seq datasets

We downloaded the following bulk datasets for comparison to our single-cell data: (*1*) microarray data from colon biopsies of 20 responders and 27 non-responders to TNF blockade, downloaded from the Gene Expression Omnibus (Arijs et al., 2009; GEO: GSE14580); and (*2*) the normalized expression matrix for 414 colon adenocarcinoma samples from The Cancer Genome Atlas (Cancer Genome Atlas, 2012) sequenced on both the Illumina HiSeq and Illumina Genome Analyzer.

Comparison of TNF signaling and response to anti-TNF therapy

We scored each cell subset for gene signatures related to TNF signaling and response to anti-TNF therapy. To ensure that these gene signatures were disjoint for correlation analysis, we removed all shared genes from the gene signatures related to anti-TNF response.

Analysis of bulk RNA-Seq data from human colon spheroids treated with IL-22 versus controls

To test the effects of IL-22 treatment on human colon spheroids, we constructed gene signatures for the top 100 differentially expressed genes in IL-22 treated spheroids versus non-treated controls ("IL-22 signature"). Differential expression was measured as the mean $\log_2$(TP10K+1) fold change between the conditions across all of the bulk RNA-Seq samples. These gene signatures were then scored to enterocytes from healthy individuals (see "Scoring gene signatures and identifying significant changes between health and disease").

Using receptor-ligand pairs to infer cell-cell interactions

To identify cell-cell interactions, we mapped the FANTOM5 database of literature-supported receptor-ligand interactions (Ramilowski et al., 2015) onto our lists of cell subset markers and differentially expressed genes within healthy, UC non-inflamed, and UC inflamed cells. We restricted our analysis to high-confidence interactions by requiring cell subset markers to have a discrete model coefficient greater than 1 and adjusted p value less than 0.05. To identify changes in this network with disease, we also constructed networks where the receptor and/or ligand were significantly differentially expressed, again requiring genes to have a discrete model coefficient with magnitude greater than 1 and adjusted p value less than 0.05. To ensure that these differentially expressed genes had sufficiently high expression, we also required them to be cell subset markers in cells isolated from healthy subjects or the relevant disease state (i.e., UC non-inflamed or UC inflamed).

For all networks, we quantified the interaction strength between two cell subsets as the number of unique receptors and ligands connecting them, resulting in adjacency matrices summarizing all cell-cell interactions within the dataset. Statistical significance was then empirically assessed by permuting the receptors and ligands among all cell subsets in a degree preserving manner (using edge swaps but only for uniquely connecting pairs), thus preserving the number of receptors and ligands encoded within each cell subset, but changing the connectivity between cell subsets. After running 10,000 total permutations, p values were computed as the number of times the edge strength in the permuted network was greater than or equal to the edge strength in the true network.

To plot cell-cell interaction networks, we applied the Fruchterman-Reingold layout algorithm to a network defined using the $-\log_{10}$ transformed p values, using only the edges with p value < 0.05. Although edge weights were used to generate the layout, they were removed from the final visualization for visual clarity.

Using receptor-ligand interactions to predict cell proportions

For each receptor-ligand pair in the cell-cell interaction network, we computed the Spearman correlation coefficient between the mean $\log_2$(TP10K+1) ligand gene expression in the ligand-expressing cell and the logit-transformed proportions of the receptor-expressing cell across samples.

Defining IBD associations and candidate risk genes

We compiled a list of IBD, UC, and CD associations from recent large-scale IBD genome-wide association and fine-mapping studies (de Lange et al., 2017; Huang et al., 2017; Jostins et al., 2012; Liu et al., 2015). Risk variants can act either in *cis* or in *trans* and estimating the precise effect of any given variant is an active area of research that is beyond the scope of this work. We therefore opted to map the genetic associations to all genes in their region of linkage disequilibrium (LD). After removing association signals mapping to more than 50 variants (through fine-mapping if available, or in LD with the best-association SNP with $R^2 > 0.6$), we arrived at 211 associations (comprising 120 associations for both UC and CD, 31 associations unique to UC, and 60 associations unique to CD) that collectively spanned over 531 candidate risk genes (comprising 285 genes for both UC and CD, 63 genes unique to UC, and 199 genes unique to CD).

Defining putative IBD risk genes

Although the gene driving the signal of association is often unknown, in some cases, we can pinpoint a gene that is particularly likely to be associated with disease risk. These putative risk genes were defined as genes containing a fine-mapped or nonsynonymous protein coding variant, or which were the only genes in their region of LD (Table S6). To this set, we added SLC39A8 as an additional IBD risk gene, which contains a fine-mapped variant associated with IBD risk (M.J.D. and R.J.X., unpublished data). In total, we identified 82 putative risk genes (comprising 48 risk genes for both UC and CD, 9 risk genes unique to UC, and 25 risk genes unique to CD).

Construction of gene modules

To construct modules of co-regulated genes, we first used MAGIC v0.1 (van Dijk et al., 2018) to impute gene expression data in the $\log_2$(TP10K+1) matrices for epithelial, stromal, and immune cells. MAGIC was run separately for healthy individuals and ulcerative colitis patients. MAGIC was run with the recommended settings from its GitHub repository (including optimal t selection, $k_a = 4$, and all other parameters set to their default values). To construct gene modules, we calculated the Pearson correlation coefficient between a query gene and all other genes in a cell subset using the imputed expression data. While prior studies of RNA-Seq data have used permutation tests to estimate a null distribution of correlation coefficients to determine cutoffs for gene module membership, this approach did not work well with MAGIC imputation, due to computational constraints. We therefore used a fixed cutoff, retaining the top 100 genes with the largest correlation coefficients for each gene module. We constructed gene modules for all candidate IBD risk genes, using cell subsets where the gene is expressed in at least 1% of all cells. Modules were defined as those containing a significant excess of putative IBD risk genes ($q < 0.05$). To estimate q-values for a given module size, we constructed modules from 100 datasets in which the gene labels were permuted, and modules were calculated with the same seed genes. The false discovery rate was then empirically determined for each of the module sizes ($q = 0.05$, 0.01, and 0.001 for modules with 3, 4, and 5 UC GWAS-implicated genes, respectively). Note that because modules are based on imputed expression data (van Dijk et al., 2018), we verified that their genes were expressed in their respective cell types (Figure S7C).

Optimal set cover of IBD risk gene modules

To identify a minimal number of modules to explain the greatest number of putative IBD risk genes, we used the greedy set cover algorithm. The algorithm is initialized with an empty set of "covered" IBD risk genes. At each step of the algorithm, we add the meta-module with the largest number of "uncovered" IBD risk genes (i.e., genes not in the "covered" set) to this "covered" set.

Nominating IBD risk genes from candidate regions of genetic association

To determine whether scRNA-Seq data can help nominate "causal" genes from candidate gene sets, we first collapsed all risk variants into 165 unique regions (comprising 99 regions for both UC and CD, 24 regions unique to UC, and 42 regions unique to CD), reflecting distinct risk loci. Of these, 99 regions (comprising 57 regions for both UC and CD, 19 regions unique to UC, and 23 regions unique to CD) had candidate gene sets containing more than one gene, including at least one putative risk gene, which we termed the "correct" gene for that region. (In cases where a region contained multiple independent associations each with distinct candidate gene sets, we selected the largest such set). For each candidate gene set, we then identified the gene with either (1) the highest mean expression level across all cell subsets and disease states; (2) the largest DE coefficient in non-inflamed tissue; (3) the largest DE coefficient in inflamed tissue; or (4) the largest module containing other candidate risk genes (iteratively defined; see Nominating IBD risk genes using gene modules). We assessed the probability of selecting the "correct" risk gene using each of these four criteria and compared these results to a null model in which genes were randomly selected from risk regions across 1,000 trials. Missing values were replaced with zeros and ties were broken by random sampling. To estimate statistical significance, we compared the accuracy (defined as predicting the "correct" risk gene) of each method to the null distribution.

Nominating IBD risk genes using gene modules

We developed a method that nominates risk genes based on their degree of co-regulation with other candidate genes from across all IBD risk loci. This method uses no *a priori* knowledge of putative risk genes; instead, it is initialized with the full set of 531 candidate genes defined across all disease risk loci. Our method assumes that IBD risk genes are co-regulated in gene modules within cell subsets, as we observed for GWAS-implicated risk genes (Figure 7C). To measure this co-regulation, we therefore construct gene modules for each candidate gene in each of the cell subsets that express it, yielding 20,630 gene modules for the 531 candidate genes in both healthy and diseased tissue (see Construction of gene modules).

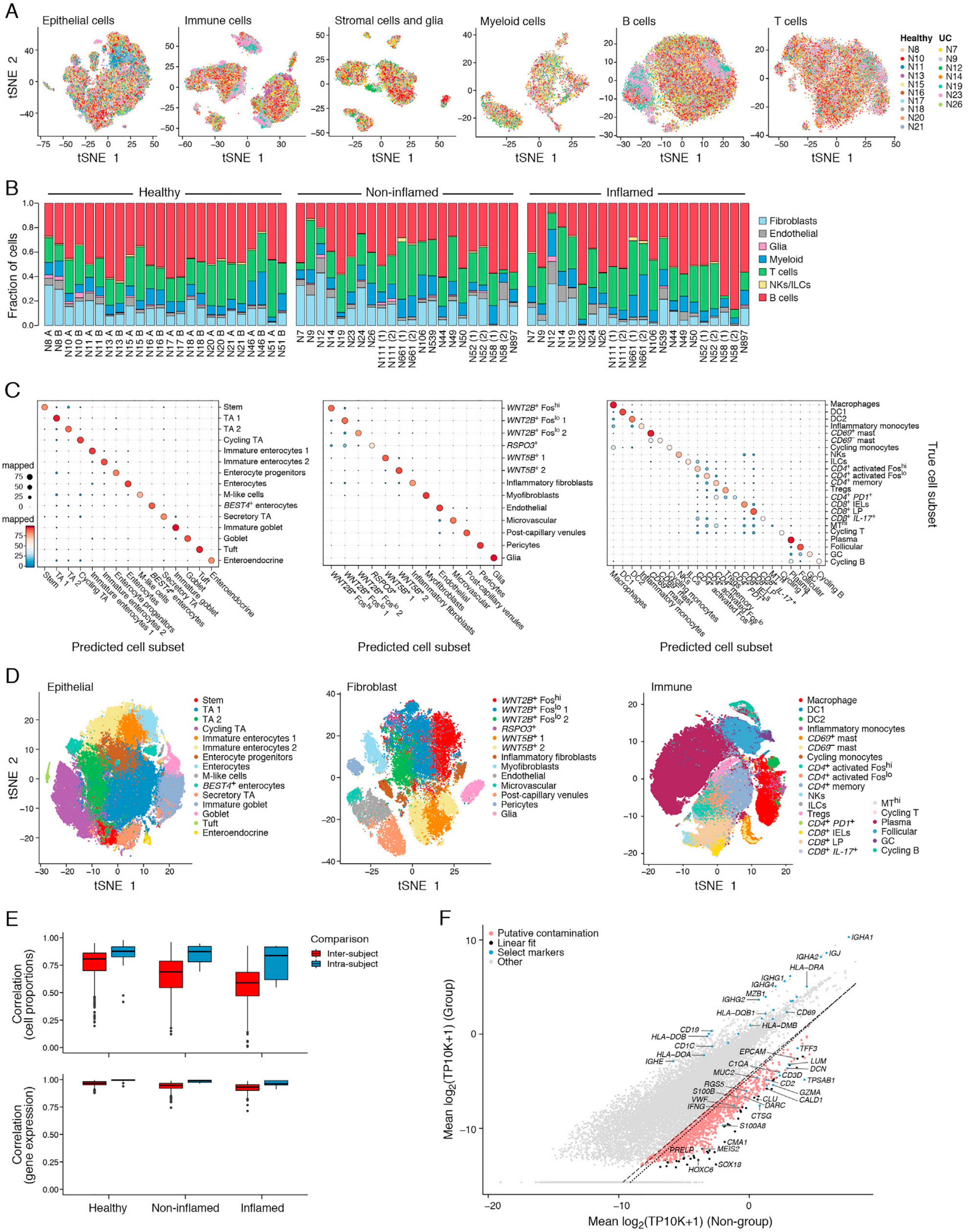
We then iteratively score each gene based on the maximal number of other candidate genes it shares a gene module with, across all such modules: genes belonging to the largest candidate gene modules receive the highest scores. To do so we use an iterative procedure. Because the set of all candidate genes initially contains many false positives, we iteratively weight each gene according to our confidence that it is a risk gene, as follows. First, under the assumption that each risk region contains exactly one risk gene, the weight for gene *i* is initialized according to the probability that it is the risk gene: $w_i = 1/N_i$, where N_i is the size of its candidate gene set (i.e., the number of genes in the risk region). Thus, genes from large candidate gene sets are initially assigned small weights, and those from small candidate gene sets are initially assigned large weights. Next, we score each gene module *J* according to the number of candidate genes that it contains, adjusted by the weight associated with those genes: $x_J = \sum_{j \in J} w_j$, for all genes *j* found in module *J*. Each gene is then mapped to its highest scoring module and the probability for each gene *i* of obtaining its module score, p_i, is estimated from the empirical distribution of module scores. Finally, we update the weights associated with each gene *i* according to the posterior probability that it belongs to the risk module for its risk region: $w_i = (1 - p_i)/\sum_{k \in C_i}(1 - p_k)$ where C_i is the candidate gene set containing gene *i* (i.e., the genes that are in the same LD region as gene *i*). These weights, which reflect our degree of confidence that a given gene is a risk gene, are iteratively updated in this manner until they converge on a final estimate (n = 10 iterations was sufficient).

To relax the assumption that each candidate gene set contains exactly one risk gene, we follow the same procedure outlined above to estimate the weights for each gene. However, rather than using these weights to nominate one risk gene per risk region, we calculate the scores for all genes and use these scores to globally nominate risk genes irrespective of their genetic locus.

DATA AND CODE AVAILABILITY

The accession number for the processed data reported in this paper is Single Cell Portal: SCP259. Raw data will be available for download from the controlled-access data repository, Broad DUOS. Code used in this study will be available at https://www.github.com/cssmillie/ulcerative_colitis.

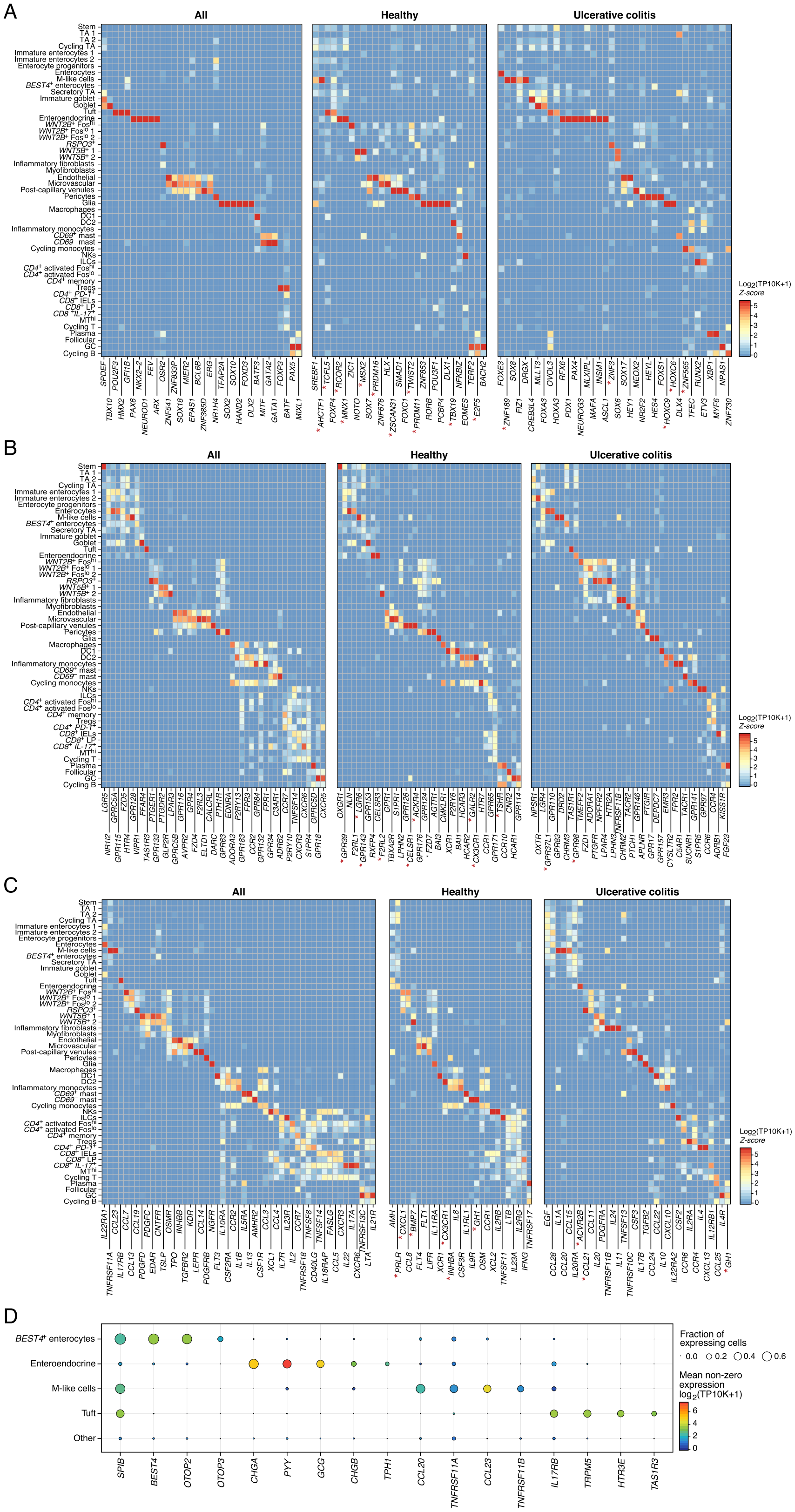
Supplemental Figures



(legend on next page)

Figure S1. Characterization and Validation of the Discovery and Validation Cohorts of the Colon Single-Cell Atlas, Related to Figure 1

A. Cell subsets are evenly distributed across healthy individuals and UC patients in the discovery cohort. t-SNEs of major cell lineages, colored by the individuals. B. Reproducible composition across samples. The fraction of cells (y axis) from each major lineage (color) in each lamina propria sample (x axis) in the discovery and validation cohorts. C. Accurate classification of cell subsets in the discovery set. Confusion matrices for epithelial (left), stromal (middle), and immune (right) cells, showing the percent of cells (dot size and color) from each cell subset in the test cohort (y axis) that are predicted to belong to each cell subset as defined in the discovery cohort (x axis). D. Concordance of discovery and validation cohorts. t-SNEs of epithelial (left), stromal (middle), and immune (right) cells showing co-embedded cells (STAR Methods) from both the discovery and validation cohorts, colored by cell subset assigned in each cohort. E. Reproducible single-cell profiles from samples collected from the same individual and from different individuals. Distribution of correlation coefficients for cell proportions (top) and expression levels (bottom) between replicate samples collected from the same individual (blue) or different individuals (red), for healthy, non-inflamed, and inflamed tissues (x axis). Boxplots: 25%, 50%, and 75% quantiles; error bars: standard deviation (SD). F. Example of approach to correct for ambient RNA contamination. Mean expression level for each gene (dot) in B cells (i.e., “in-group” expression, y axis) and all other cells (i.e., “non-group” expression, x axis), indicating genes used for the robust linear regression (black), genes classified as putative contaminants (red), select marker genes for different cell types (blue), and other genes (gray). Dashed line: robust linear fit used to estimate putative contaminants.



(legend on next page)

Figure S2. Cell Subset-Specific Features of the Colon Single-Cell Atlas, Related to Figure 1

A-C. Cell- and lineage-specific genes in key functional classes. Mean expression across the cell subsets (rows) of genes (columns) encoding cell- and lineage-specific (STAR Methods; Table S3) transcription factors (A), G-protein coupled receptors (B), and cytokines and cytokine receptors (C), expressed those cells in both healthy and UC samples (left), only in healthy samples (center), or only in UC samples (right). Asterisks: genes that significantly changed their cell- or lineage-specificity between health and disease (STAR Methods). D. Specific expression of distinct key signaling genes in sensory epithelial cell subsets. Fraction of expressing cells (dot size) and mean expression level in expressing cells (dot color) of selected signaling genes (columns) across cell subsets (rows).

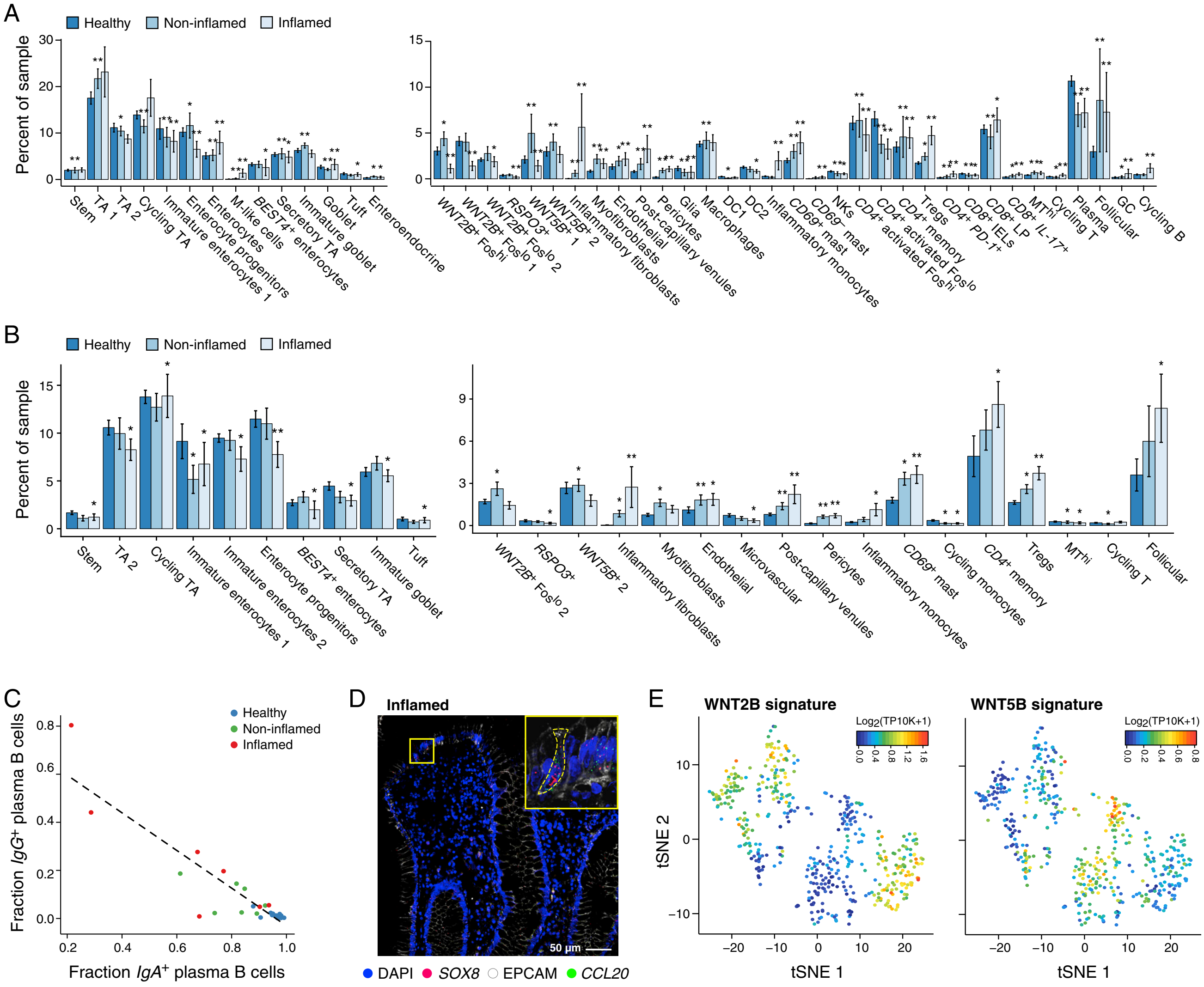


Figure S3. Cell Composition Changes during UC Highlight Changes within Plasma B Cells and in M Cells, Related to Figure 2

A,B. Cell proportion changes. Significant changes in cell frequency (y axis), assessed within a class using a Fisher's exact test (A) or a Mann-Whitney test (B), for non-inflamed (light blue) and inflamed (white) samples relative to healthy samples (dark blue) (adjusted p values, * = 0.05, ** = 0.01, *** = 0.001); error bars: SEM C. Relative increase in IgG$^+$ and decrease in IgA$^+$ B cells among plasma B cells with disease. Fractions of IgA$^+$ (x axis) and IgG$^+$ (y axis) B cells out of all plasma B cells in healthy (blue), non-inflamed (green), and inflamed (red) samples. Dashed line: linear fit. D. Microfold (M)-like cells in inflamed biopsies. Representative images of combined smFISH and IFA of M-like cells in colon TMA showing their presence in inflamed human colon (no M-like cells were observed in healthy tissue from 10 different biopsies). Yellow arrow: M-like cell, scale bar, 50 μm; Inset, x5 magnification. E. *WNT2B*$^+$ and *WNT5B*$^+$ fibroblast markers are expressed in distinct subsets of IAFs. t-SNE of scRNA-Seq profiles from IAFs, colored by the mean expression of markers for *WNT2B*$^+$ fibroblasts (top) or *WNT5B*$^+$ fibroblasts (bottom). IAFs that express *WNT2B*$^+$ markers typically do not express *WNT5B*$^+$ markers (and vice versa).

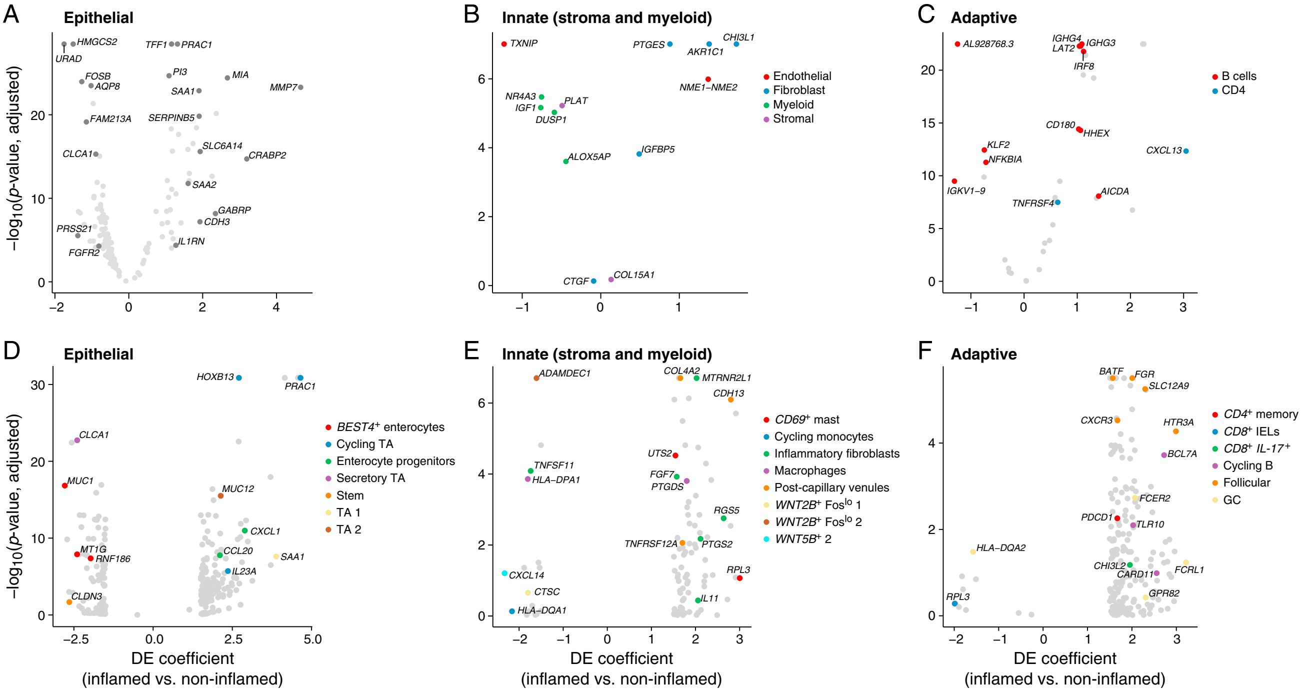


Figure S4. Cell- and Lineage-Specific Expression Changes in Inflamed versus Non-Inflamed Tissues, Related to Figure 3

Volcano plots of genes that are differentially expressed in inflamed cells relative to non-inflamed cells, showing the effect size in inflammation (i.e., discrete DE coefficient, x axis) and statistical significance (y axis). (A-C) General changes that were shared across multiple cell subsets within (A) epithelial, (B) innate (including stromal and myeloid cells), or (C) adaptive immune compartments. (D-F) Unique changes that were specific to cell subsets within these compartments. Selected genes are highlighted, all genes are reported in Table S4.

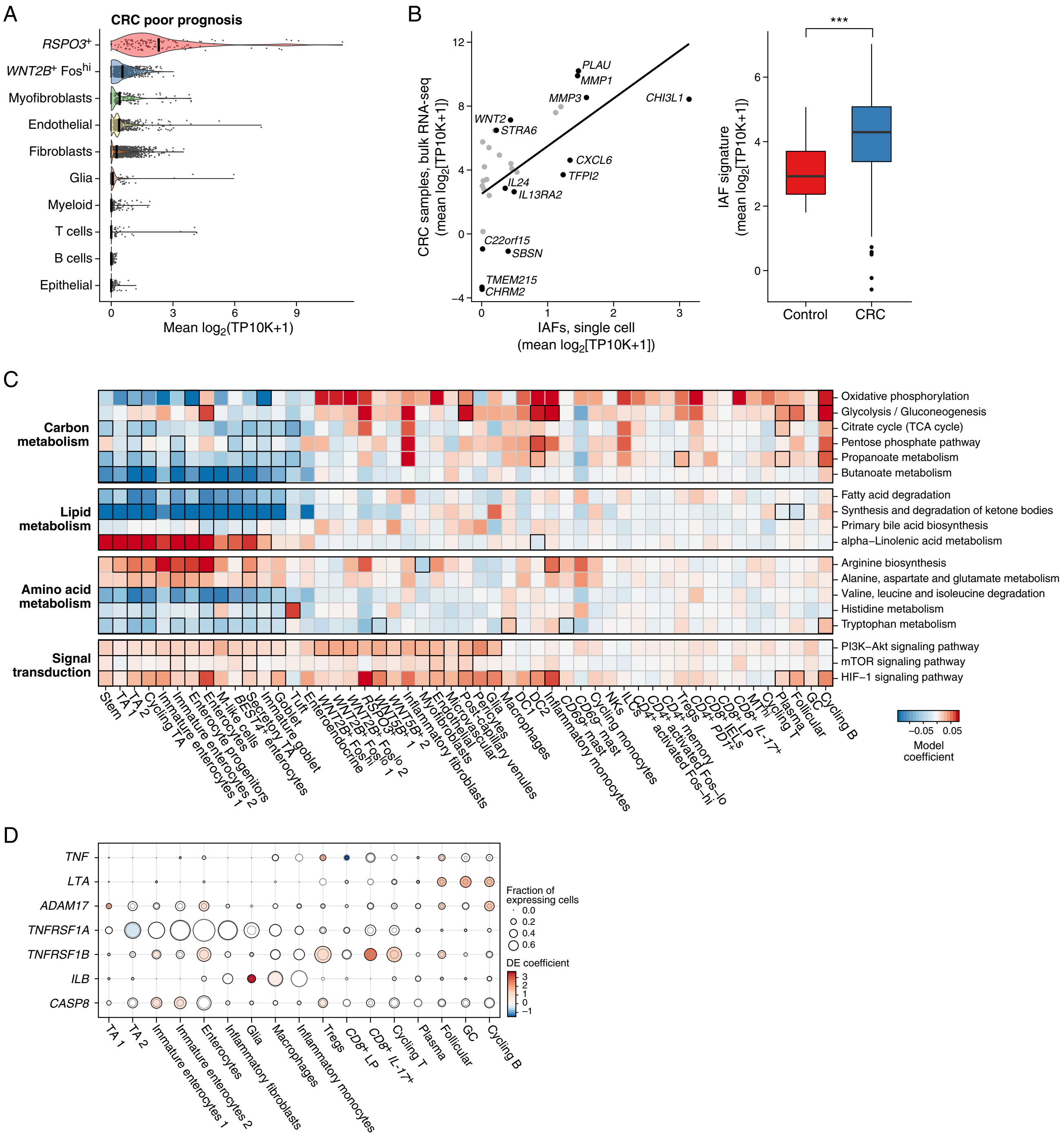


Figure S5. Changes in Transcriptional Programs in UC and Colorectal Cancer, Related to Figure 4

A. *RSPO3*$^+$ cell profiles are enriched in signatures of poor prognosis in colorectal cancer (CRC) (Calon et al., 2015; STAR Methods). Distribution of the mean expression (x axis) of a stromal gene signature of poor prognosis in CRC in the three highest scoring cell subsets and other compartments (y axis); crossbar: mean. B. Inferred expansion of inflammatory fibroblasts with colorectal cancer. Left: mean expression of IAF marker genes in colorectal cancer samples (y axis) and inflammatory fibroblasts (x axis). Black line: linear regression. Select genes annotated. Right: distribution of IAF gene signature scores in bulk RNA-Seq data from colorectal cancer patients (blue) versus healthy controls (red) (***$p < 0.001$, Mann-Whitney test). Boxplots: 25%, 50%, and 75% quantiles; error bars: standard deviation (SD, right). C. Expression changes (model coefficient, color bar) in inflamed cells relative to healthy cells for 23 KEGG pathways (rows) related to carbon, lipid, and amino acid metabolism, and key additional pathways (apoptosis, autophagy, etc., bottom), for each cell subset (columns). Black outlines: significant changes ($q < 0.05$, mixed linear model). D. Differential expression (color bar) of genes related to TNF signaling (rows) in inflamed versus healthy samples across cell subsets (columns). Dot size: fraction of expressing cells in healthy (gray outline) or inflamed (black outline) samples; dot color: significant DE model coefficients ($q < 0.05$, MAST hurdle model).

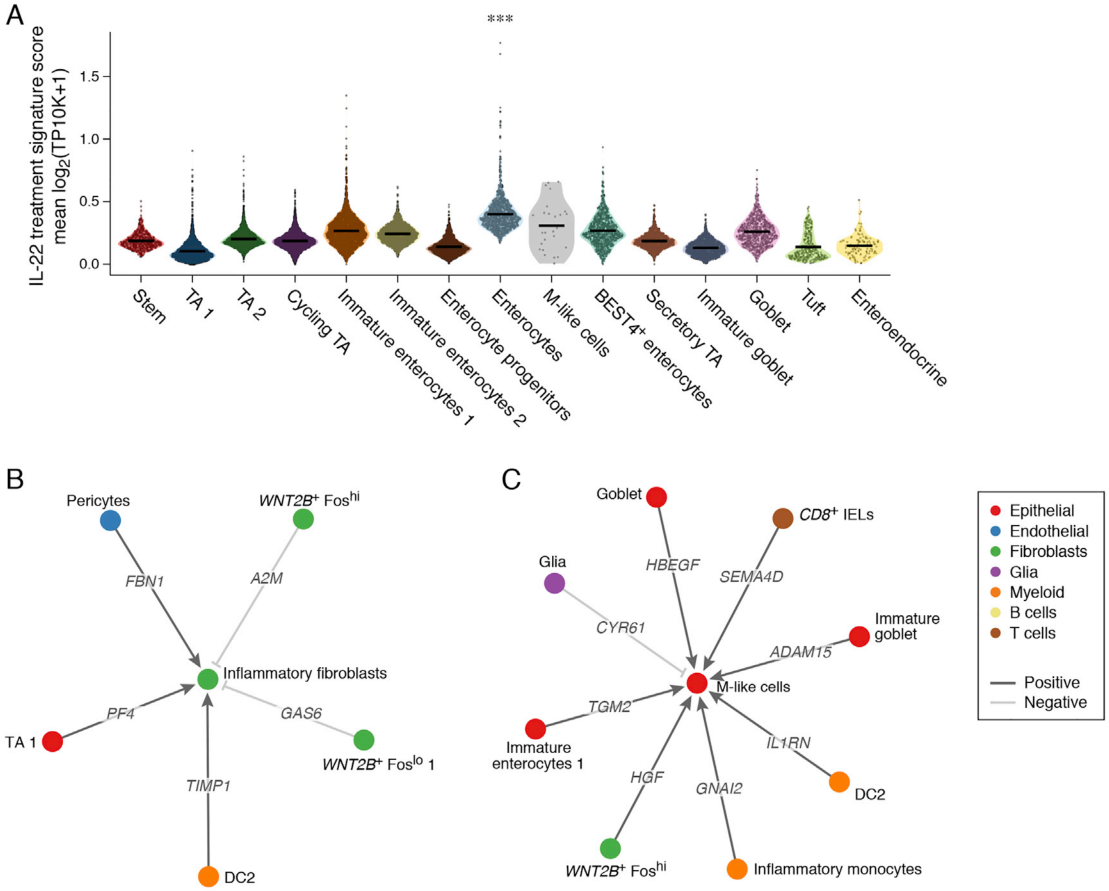


Figure S6. Cell-Cell Interactions May Explain Shifts in Cellular Proportions during UC, Related to Figure 6

A. Treatment of human colon spheroids (n = 3, 2 CD and one UC patients) with IL-22 induces the transcription of genes that are significantly enriched in enterocytes. Distribution of mean expression (y axis) of gene signature enriched in IL-22 treated human colon spheroids across cell subsets (x axis); P value, *** < 10^{-10} for enterocytes versus all other cells; Wilcoxon test. B,C. LASSO based models (STAR Methods) explaining the change in cell proportions across samples in IAFs (B) and M-like cells (C) as a function of both positive (dark gray pointed arrows) and negative (light gray blunt arrows) relations to ligands (edge label) expressed by other cell subsets marked by lineage (color). Shown are all ligands with non-zero coefficients in the regularized LASSO model.

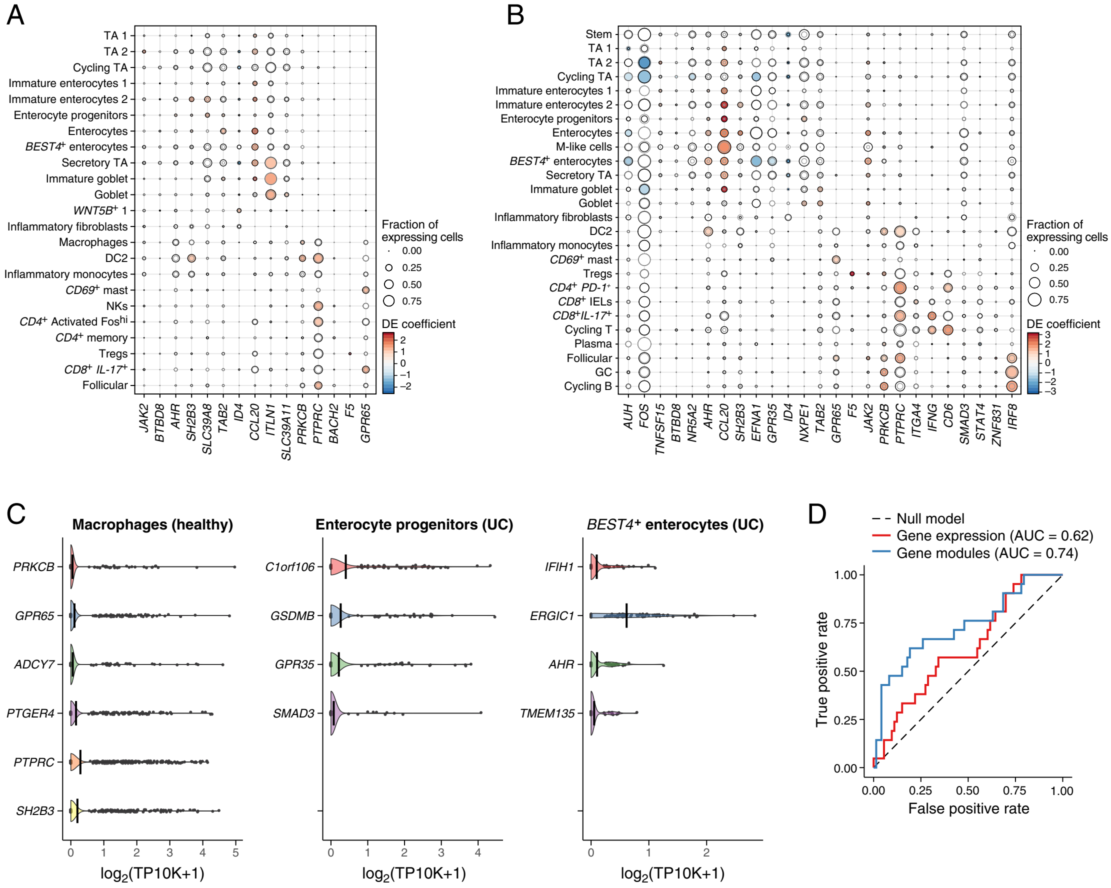


Figure S7. Expression of Risk Genes across Cell Subsets Highlights Key Cell Types and Pathways in UC, Related to Figure 7

A,B. Differential expression of putative IBD risk genes in specific cell subsets. For GWAS-implicated IBD risk genes (columns) that are differentially expressed in non-inflamed (B) or inflamed (C) cells versus healthy cells, shown is the fraction of expressing cells in healthy (gray outline) or diseased samples (black outline) in each cell subset and significant DE model coefficients (color, q < 0.05, MAST likelihood ratio test). C. Co-expression meta-modules are expressed in their respective cell subsets. Distribution of gene expression levels (x axis) in cell subsets (y axis) for each of the putative risk genes in the meta-modules for *PRKCB* in healthy macrophages (left), *C1orf106* in UC enterocyte progenitors (center), and *IFIH1* in UC *BEST4*$^{+}$ enterocytes (right); crossbar: mean. D. Co-expression meta-modules can help nominate multiple risk genes across candidate gene sets. Receiver operating characteristic (ROC) curve showing the true positive rate (y axis) and the false positive rate (x axis) for nomination methods across different cutoffs for gene expression levels (red) and meta-module scores (blue).