## ORIGINAL RESEARCH

# Single-Cell Analysis Reveals Unexpected Cellular Changes and Transposon Expression Signatures in the Colonic Epithelium of Treatment-Naïve Adult Crohn's Disease Patients

Matt Kanke,[1] Meaghan M. Kennedy Ng,[3] Sean Connelly,[2] Manvendra Singh,[4] Matthew Schaner,[2] Michael T. Shanahan,[1] Elizabeth A. Wolber,[2] Caroline Beasley,[2] Grace Lian,[2] Animesh Jain,[2] Millie D. Long,[2] Edward L. Barnes,[2] Hans H. Herfarth,[2] Kim L. Isaacs,[2] Jonathon J. Hansen,[2] Muneera Kapadia,[2] Jose Gaston Guillem,[2] Cedric Feschotte,[4] Terrence S. Furey,[3] Shehzad Z. Sheikh,[2] and Praveen Sethupathy[1]

[1]Biomedical Sciences, Cornell University, Ithaca, New York; [2]Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [3]Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [4]Molecular Biology and Genetics, Cornell University, Ithaca, New York
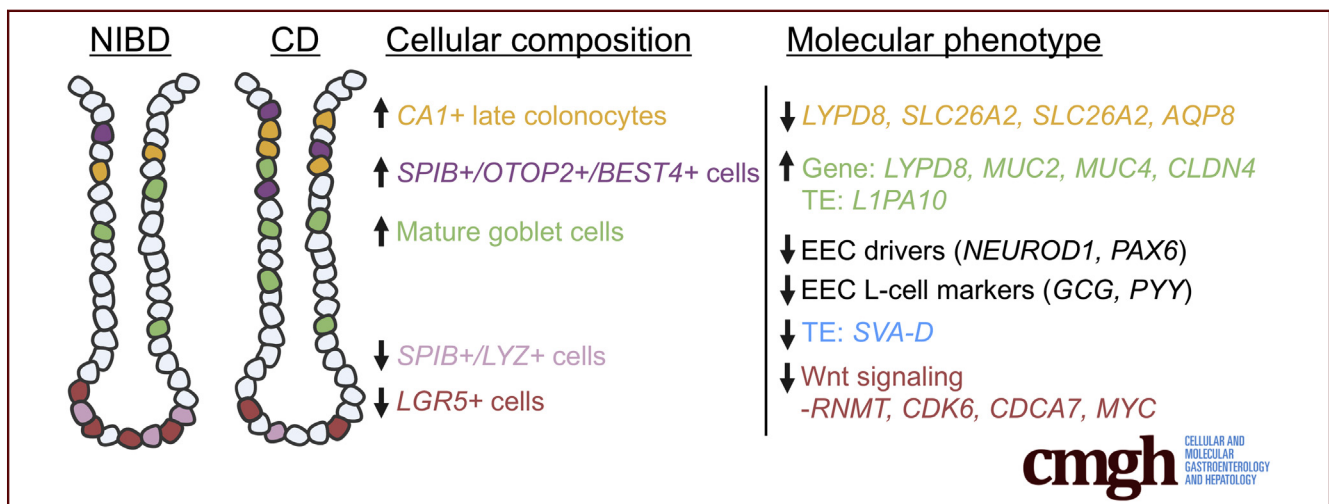
## SUMMARY

The cellular and molecular landscape of Crohn's disease (CD) is still poorly understood. In this study, we performed single-cell analyses of the colonic epithelium from treatment-naïve patients, which showed significant and unexpected shifts in cellular composition and molecular phenotype.

**BACKGROUND & AIMS:** The intestinal barrier comprises a monolayer of specialized intestinal epithelial cells (IECs) that are critical in maintaining mucosal homeostasis. Dysfunction within various IEC fractions can alter intestinal permeability in a genetically susceptible host, resulting in a chronic and debilitating condition known as Crohn's disease (CD). Defining the molecular changes in each IEC type in CD will contribute to an improved understanding of the pathogenic processes and the identification of cell type–specific therapeutic targets. We performed, at single-cell resolution, a direct comparison of the colonic epithelial cellular and molecular landscape between treatment-naïve adult CD and non–inflammatory bowel disease control patients.

**METHODS:** Colonic epithelial-enriched, single-cell sequencing from treatment-naïve adult CD and non–inflammatory bowel disease patients was investigated to identify disease-induced differences in IEC types.

**RESULTS:** Our analysis showed that in CD patients there is a significant skew in the colonic epithelial cellular distribution away from canonical *LGR5+* stem cells, located at the crypt bottom, and toward one specific subtype of mature colonocytes, located at the crypt top. Further analysis showed unique changes to gene expression programs in every major cell type, including a previously undescribed suppression in CD of most enteroendocrine driver genes as well as L-cell markers including *GCG*. We also dissect an incompletely understood *SPIB+* cell cluster, revealing at least 4 subclusters that likely represent different stages of a maturational trajectory. One of these *SPIB+* subclusters expresses crypt-top colonocyte markers and is up-regulated significantly in CD, whereas another subcluster strongly expresses and stains

positive for lysozyme (albeit no other canonical Paneth cell marker), which surprisingly is greatly reduced in expression in CD. In addition, we also discovered transposable element markers of colonic epithelial cell types as well as transposable element families that are altered significantly in CD in a cell type–specific manner. Finally, through integration with data from genome-wide association studies, we show that genes implicated in CD risk show heretofore unknown cell type–specific patterns of aberrant expression in CD, providing unprecedented insight into the potential biological functions of these genes.

**CONCLUSIONS:** Single-cell analysis shows a number of unexpected cellular and molecular features, including transposable element expression signatures, in the colonic epithelium of treatment-naïve adult CD. *(Cell Mol Gastroenterol Hepatol 2022;13:1717–1740; https://doi.org/10.1016/j.jcmgh.2022.02.005)*

The colonic epithelium acts as an essential barrier between the luminal contents of the colon, including a diverse compendium of microbes, and the underlying lamina propria immune system. The epithelium is a heterogenous mix of distinct cell types with a wide range of specialized functions, including absorptive cells that transport nutrients and electrolytes (colonocytes), and secretory cells that emit factors such as mucins (goblet cells) and endocrine hormones (enteroendocrine cells). Stem cells at the base of the colonic crypt are responsible for the continual, rapid renewal of this epithelial layer. Each cell type is crucial for the maintenance of intestinal homeostasis and defects in any could contribute to the onset of inflammatory bowel disease (IBD).

IBD consists of 2 main disease types, ulcerative colitis and Crohn's disease (CD), and is characterized by chronic intestinal inflammation that can lead to severe tissue damage and organ dysfunction. Despite recent advances, the etiology of IBD still remains largely unknown. Unrelenting inflammation is attributed to a complex interaction between genetic, luminal (microbial), and environmental factors that trigger an inappropriate mucosal immune response. Recent studies have begun to unravel the role of different colonic epithelial cell types in IBD using single-cell RNA sequencing (scRNA-seq).[1-3] Although they have advanced our understanding, important challenges remain to be addressed. Notably, the patients included in the studies showed varying disease durations, and some have been treated with therapeutics, which are confounding variables while this scRNA-seq study has investigated treatment-naïve adult IBD. Moreover, the studies have focused mostly on patients with ulcerative colitis while this scRNA-seq study has performed a focused investigation of colonic epithelium in adult CD. Changes to the relative abundance and molecular character of different colonic epithelial cell types during CD pathogenesis is poorly understood and merits deeper examination.

To investigate this, we performed single-cell transcriptional profiling in colonic epithelial cells from a cohort of treatment-naïve adult CD patients and healthy controls. Our analysis revealed that the colonic epithelium of CD patients shows significant and unexpected shifts in cellular composition as well as cell type–specific messenger RNA and transposable element profiles. In addition, through integration of data from genome-wide association studies, we show that genes implicated in CD risk show heretofore unknown cell type–specific patterns of aberrant expression in CD. Taken together, our study provides important clues about the early molecular events that promote the dysfunction of this critical tissue during CD pathogenesis.

## Results

### Single-Cell Analysis Provides a High-Resolution Picture of the Colonic Epithelium From Adult Non-IBD and Crohn's Patients

We harvested mucosal tissue from the ascending colon of treatment-naïve adult individuals with CD (n = 3) and non-IBD healthy controls (NIBD, n = 4) (Table 1). The sample size was restricted by the challenge of obtaining biopsy tissue from treatment-naïve adult CD patients. We then enriched for epithelial cells (the focus of this study), which were dissociated into single-cell suspensions (see the Methods section for more detail). Samples were subjected to RNA sequencing using the 10X Genetics Chromium platform and 13,039 cells remained after filtering (see the Methods section for more detail).

We processed and analyzed the data (CD and NIBD together) using Cell Ranger and Seurat (see the Methods section for more detail) and visualized the cell clusters using Uniform Manifold Approximation and Projection.[4] We identified 21 cell clusters (Figure 1A), 7 of which were determined to be different types of immune cells (Figure 1B), representing 21% of the total. We removed these cells and reclustered the remaining 10,162 epithelial cells, resulting in 14 cell clusters. Based on the expression levels of previously annotated marker genes and highly enriched cluster markers identified here, we assigned these clusters to 14 different colonic epithelial cell types (Figure 2A and B). These cell types include *LGR5+* stem cells; *MUC2+* immature and mature goblet cells; *CHGA+* enteroendocrine cells (EECs); 3 categories of cycling cells including both G2-M-G1 and S-phase transit-amplifying cells, secretory progenitors, and colonocyte progenitors; and at least 4 other cell clusters expressing known markers

Most current article

**Table 1.** Information on the 7 Individuals From Whom the Single Cell Tissue Was Biopsied

| Sample ID | Condition | Age, y | Sex |
|-----------|-----------|--------|-----|
| 206 | NIBD | 55 | M |
| 214 | NIBD | 64 | M |
| 216 | NIBD | 63 | M |
| 217 | NIBD | 72 | F |
| 189 | CD | 25 | M |
| 299 | CD | 56 | F |
| 364 | CD | 68 | M |

F, female; M, male.

of colonocytes including *CA1+* early and late colonocytes, *CEACAM7+* colonocytes, and *SPIB+* cells (the latter of which is similar to the previously reported *BEST4/OTOP2* cells[3] or *BEST4+* enterocytes[2]) (Figure 2*B* and *C*). We found that each of these cell types express a unique set of markers (Methods section, Figure 2*D*). Specifically, the *SPIB+* cells are uniquely marked by *SPIB, NOTCH2,* and *HES4* (*BEST4* and *OTOP2* are present in a subset of the cells in the *SPIB+* cluster); the *CEACAM7+* cells uniquely express several genes including 1 long, noncoding RNA (*LINC01133*) that has been implicated previously in cancer and in the regulation of the Wingless/Integrated (Wnt) signaling pathway[5]; and the *CA1+* late colonocytes share *MALL* expression with *CEACAM7+* cells but uniquely express *CA1* (Figure 2*D*).

Colonic epithelial cells from the crypt bottom to the crypt top represent a gradient of maturation, with stem/progenitor cells at the bottom and mature differentiated colonocytes at the top. Using a previously defined 15-gene signature,[3] we computed a crypt–axis score for every cell in each cluster (Figure 2*E*), which shows good correspondence between our cluster annotations and known cell-type positions within the crypt. RNA velocity analysis[6] confirmed that the data set represents the full maturational spectrum of the colonic epithelium (Figure 2*F*).

### The Colonic Epithelial Cellular Landscape Is Skewed Toward a Crypt-Top Signature in CD

Next, we assessed CD and NIBD data separately to determine whether the distribution of cells along the crypt–axis, based on their crypt–axis score (Figure 2*E*), is altered in CD. We detected a significant shift in the density of cells toward the crypt top in CD relative to NIBD (Figure 3*A*). This likely is driven by specific subtypes of colonocytes because we found that *CA1+* colonocytes (Figure 3*B*), especially *CA1+* late colonocytes (Figure 3*C*), are increased significantly in abundance in CD relative to NIBD. We observed only a slight difference in the relative abundance of mature absorptive cells (combined *CA1+* early, *CA1+* late, and *CEACAM7+* colonocytes) vs mature secretory cells (combined mature goblet cells and EECs) (Figure 3*D*). However, between the 2 major secretory lineages, we discovered a significant skew toward the mature goblet fate in CD (Figure 3*E*). Even within the goblet lineage we observed a

shift toward heightened goblet maturation in CD compared with NIBD (Figure 3*F*).

### Different Colonocyte Clusters Show Unique Changes in Gene Expression in CD

To investigate the genes and pathways most altered in CD in a cell type–specific manner, we first performed differential gene expression analysis in each of the 14 cell types separately. We identified a varying number of significantly (adjusted $P < .05$) differentially expressed genes (DEGs) across cell types (Figure 4*A*). We found that the number of DEGs are roughly proportional to the number of cells in each of the clusters (Pearson correlation coefficient, 0.96), indicating that the variation in the number of DEGs is owing at least in part to variability in statistical power. A notable exception is the *CA1+* late colonocyte cluster, which exhibits as many or more DEGs than several clusters with a greater number of cells, including the *CA1+* early colonocyte, G2–M–G1 transit-amplifying, immature and mature goblet, and immature colonocyte clusters. In addition, even though there are more than 200 fewer cells in the stem cell cluster compared with the mature goblet cell cluster, the stem cells show more DEGs than mature goblet cells (Figure 4*A*), pointing to a robust change in the molecular profile of stem cells in CD. To investigate cell type–specific changes in gene expression in the absorptive and secretory lineages, we first focused on 2 major colonocyte clusters (*CA1+* late colonocyte and *CEACAM7+* colonocyte) and 2 secretory clusters (mature goblet cells and EECs).

We found that 45 genes (30 up regulated, 15 down regulated in CD relative to NIBD) are significantly altered uniquely in *CA1+* late colonocytes (Table 2). Among those significantly up-regulated uniquely in *CA1+* late colonocytes are *PRDM1* and *REL*, and among those down-regulated uniquely are *CA2, SLC26A2,* and *SLC20A1* (Figure 4*B*). *PRDM1* and *REL* have been implicated in anti-inflammatory and microbial-sensing pathways in the colon, and have been reported as harboring variants associated with CD in genome-wide association studies (GWAS)[7,8]; however, the increased expression in CD is observed in only this subtype of colonocytes. *CA2* and *SLC26A2* are not only uniquely down-regulated in *CA1+* late colonocytes, but also are most highly expressed in this cluster (Figure 4*C*), suggesting that normal functions of this cell type such as anion transport are compromised in CD (SLC26A2 and SLC20A1 are sulfate and phosphate transporters, respectively, contributing to solute homeostasis in the colon).

There are 29 genes (7 up regulated, 22 down regulated) significantly altered uniquely in *CEACAM7+* colonocytes (Table 3). Among those down-regulated are *CA4, AQP8, GUCA2A,* and *GUCA2B* (Figure 4*B*), each of which has been implicated in various normal functions of the colonic epithelium, such as the role of Aquaporin 8 (AQP8) in colonic epithelial water transport.[9] Decreased colonic epithelial expression of *AQP8* has been reported in IBD previously,[10] and has been suggested as a candidate therapeutic target for diarrheal diseases.[11] Here, we show that the gene that codes for this important IBD-related protein is
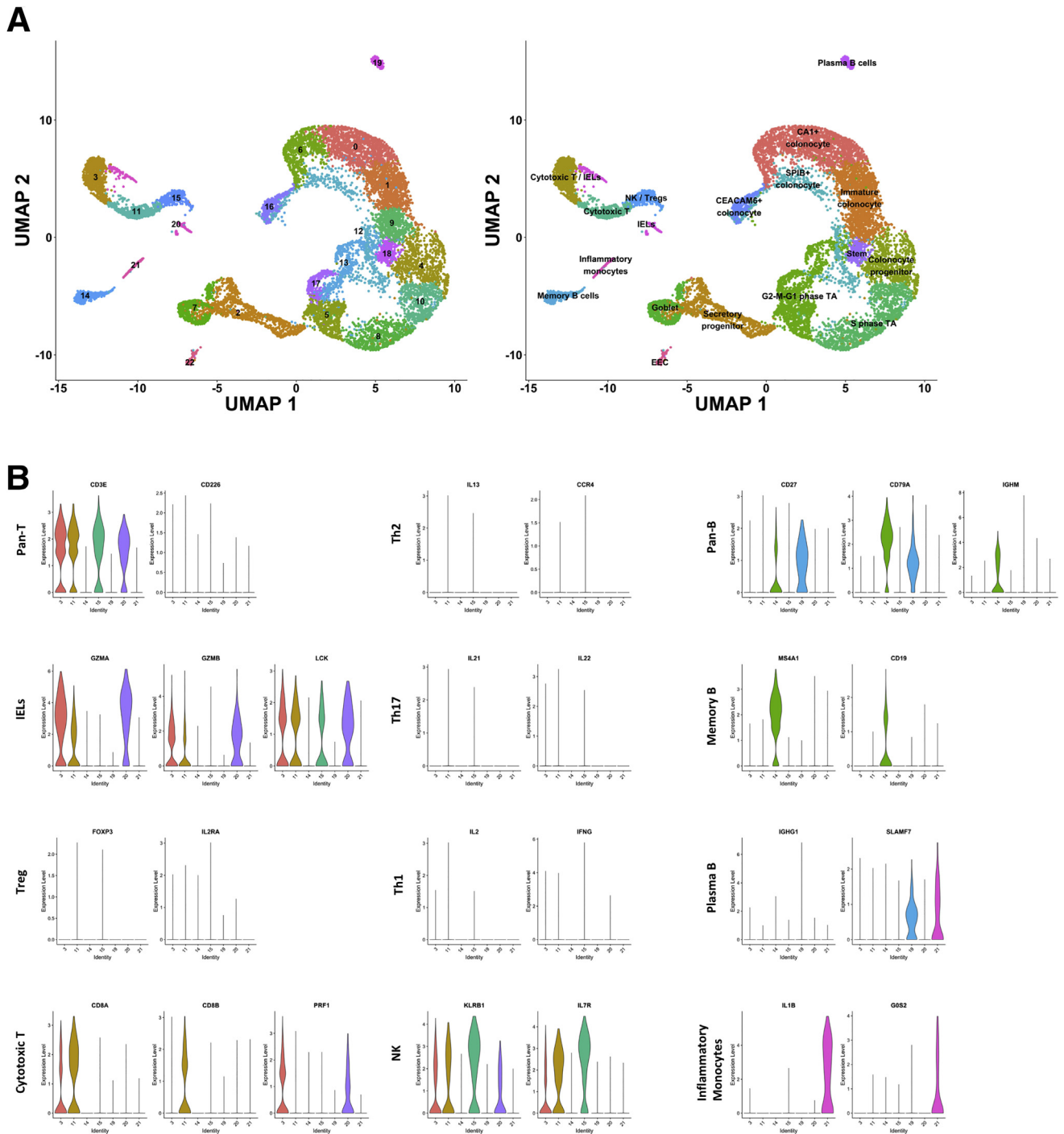
**Figure 1. Clustering and identification of contaminating immune cells.** (*A*) Uniform Manifold Approximation and Projection (UMAP) of initial clusters (*left*) and clusters following cell type annotation using known cell type markers and highly enriched genes (*right*). (*B*) Violin plots showing normalized expression of known markers of different immune cell types on the y-axis and the immune clusters on the x-axis. IEL, intraepithelial lymphocytes; NK, natural killer cells; Pan-B, All B cells; Pan-T, All T cells; Th, T helper cells; Treg, Regulatory T cells.

down-regulated dramatically in CD in *CEACAM7*+ colonocytes only and not in any other cell type of the colonic epithelium (Figure 4*D*).

We found that 4 genes (1 up regulated, 3 down regulated in CD relative to NIBD) are altered significantly in both

*CA1*+ colonocytes and *CEACAM7*+ colonocytes, and in no other cell type. Shared down-regulated genes include *CD177* and *LYPD8* (Figure 4*B*), both of which are in the same family of proteins containing the lymphocyte antigen-6/urokinase plasminogen activator surface receptor (LY6/PLAUR)
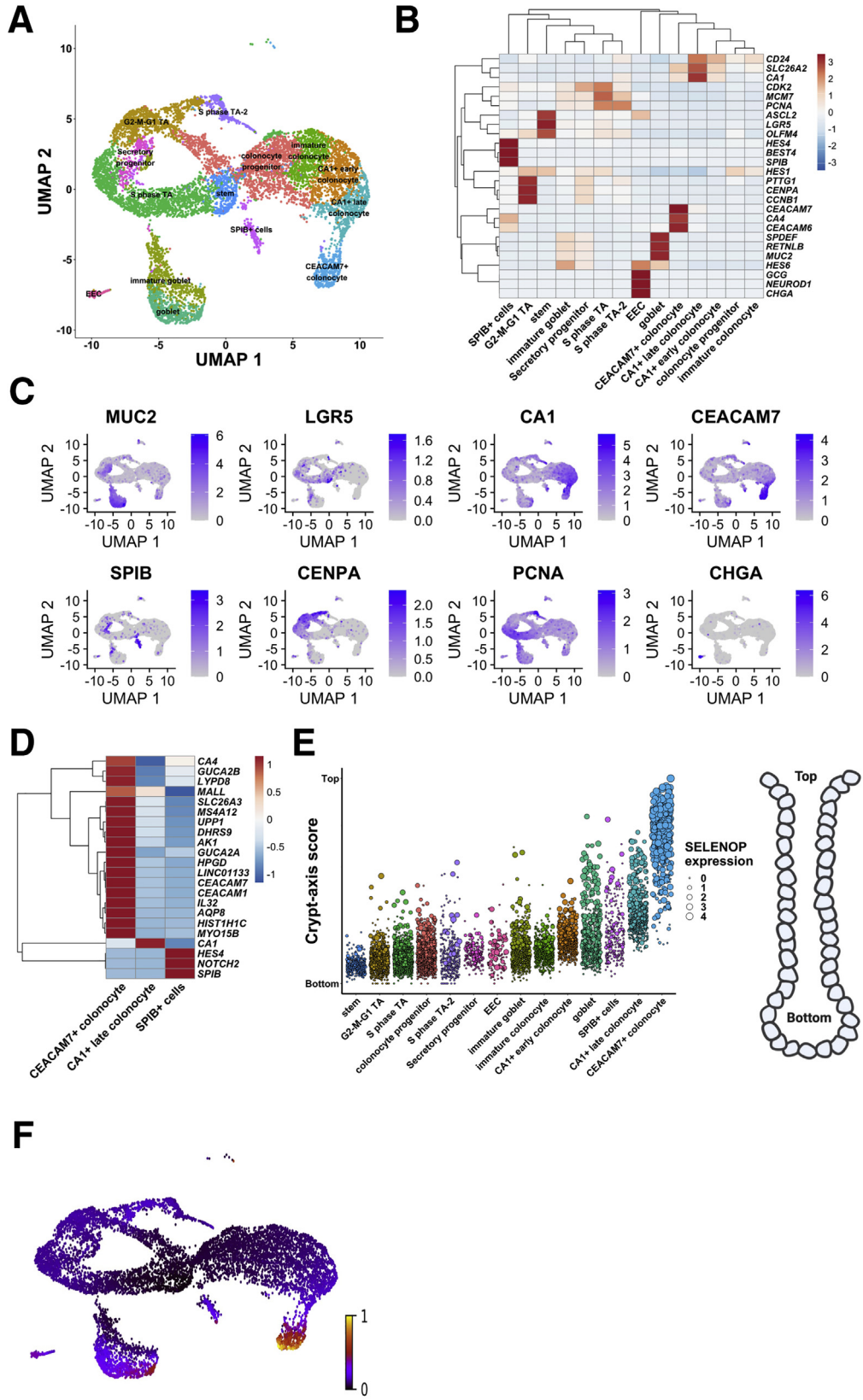
domain. The latter encodes a protein that protects the gut from microbial invasion and is critical for maintaining barrier integrity and preventing intestinal inflammation.[12–14] *LYPD8* is expressed 5-fold greater in *CEACAM7+* colonocytes compared with *CA1+* late colonocytes, but it is downregulated significantly in CD in both cell types (Figure 4*E*). Surprisingly, it also is highly expressed and dramatically reduced in CD in EECs (Figure 4*E*), although this decrease does not achieve significance likely owing to the very small number of EECs, contributing to low statistical power.

### The Mature Goblet Program Is Enhanced Whereas Enteroendocrine Drivers and L-Cell Markers Are Suppressed in CD

Goblet cells secrete mucins to create a protective barrier for the colon from luminal content. In mature goblet cells, classic gene markers of maturity and function, including *MUC2*, *MUC4*, and *TFF1*, are increased significantly in CD, whereas markers of immaturity, including *KLF4*, *DLL1*, and *RETNLB*, are reduced significantly in CD (Figure 4*B*). This is concordant with our finding of a shift toward heightened goblet maturation in CD compared with NIBD (Figure 3*E*). Interestingly, levels of *CLDN4*, previously reported to be highly expressed in EECs[15] and also studied in the context of colonocyte barrier function,[16] are increased significantly in the mature goblet cell cluster and modestly reduced in both *CEACAM7+* colonocytes and EECs (Figure 4*F*). In fact, in CD, the levels of *CLDN4* in mature goblet cells rise to what is observed in *CEACAM7+* colonocytes (Figure 4*F*), the cells in which *CLDN4* is most highly expressed in NIBD.

EECs, which secrete hormones in response to nutrients to maintain metabolic homeostasis, are not well studied in CD. To examine the change in the molecular character of EECs in CD, we first examined the genes encoding key transcription factors (n = 10) that contribute to EEC maturation. We found that 7 of the 10 are altered significantly in CD, all of which are down-regulated (Figure 5*A*), which is consistent with our observation that the abundance of EECs (relative to mature goblet) trends lower in CD compared with NIBD (Figure 3*E*). We next evaluated the genes encoding the major hormones (n = 7) that are produced and secreted from EECs. We found that 2 in particular, *GCG* and *PYY*, both of which are expressed in the colonic L-cell subtype of EECs,[17] are more than 3-fold reduced in CD vs NIBD (Figure 5*B*). *GCG* encodes the key

metabolic hormone glucagon-like peptide 1 (GLP-1), which promotes systemic energy homeostasis,[18] and *PYY* encodes a signal that promotes satiety. Notably, the gene *Nts*, which encodes the proinflammatory peptide neurotensin,[19] is highly up-regulated in CD (Figure 5*B*).

### Transposable Elements Mark Specific Cell Types and Are Expressed Differentially in CD

Transposable elements (TEs) comprise approximately half of the human genome and represent an important source of genetic variation.[20] Retroelements predominate in the human genome and recently or currently active families include Long Interspersed Nuclear Elements 1 (L1), SINE-VNTR-Alu (SVA), and Human Endogenous Retroviruses (HERVs)[21,22] and their long terminal repeats (LTRs). Dysregulation of TE expression has been observed in various disease states.[20–24] Although the regulatory activities of TEs are known to modulate the immune response,[23–25] TE expression in intestinal cell types and in CD have remained uncharacterized. To investigate the expression patterns of TEs, we performed a combined scRNA-seq analysis of genes and TE families using the epithelial cell type classification defined in the previous sections (Methods section, Figure 6*A*). The results showed several TE families that behave as markers of specific cell types. For example, 2 subfamilies MER11A and MER11C, related to the HERVK11 (human mouse mammary tumor virus like-8 [HML8]) family, are markers of goblet cells (Figure 6*B* and *C*) and the *SPIB+* cluster, respectively (Figure 6*D* and *E*). Upon investigating differential expression of TEs in CD relative to NIBD, we found a significant increase in the RNA levels of the *L1PA10* family in goblet cells (Figure 6*F* and *G*) and a significant decrease in *SVA-D* expression in secretory progenitor cells (Figure 6*H* and *I*). Collectively, these results suggest that a small subset of TE families are expressed in a cell type–specific fashion in the colonic epithelium and a few others are dysregulated in CD.

### The Canonical Colonic Stem Cell Signature Is Disrupted in CD

Dysfunction in the intestinal stem cell population (ISC) in CD has been proposed but not rigorously evaluated and documented. To investigate this possibility, we analyzed DEGs in the ISC population in CD relative to NIBD. We found that in ISCs, among the most highly up-regulated genes in

---

**Figure 2.** (*See previous page*). **Single-cell landscape in treatment-naïve Crohn's disease and NIBD patients.** (*A*) Uniform Manifold Approximation and Projection (UMAP) of epithelial cell clusters following assignment of cell type. (*B*) Normalized expression heatmap of genes known to be markers of various epithelial cell populations or found to be highly enriched in a cluster (*rows*) across epithelial clusters (*columns*) confirms cell type assignment. Normalized gene expression is scaled by *row*. (*C*) UMAP overlain with the normalized expression of cell type markers further confirms cluster assignment. *MUC2*, goblet; *LGR5*, stem; *CA1*, *CA1+* colonocytes; *CEACAM7*, *CEACAM7+* colonocytes; *SPIB*, *SPIB+* cells; *CENPA*, G2–M–G1 TA; *PCNA*, S-phase TA; *CHGA*, EEC. (*D*) Normalized expression heatmap of 3 clusters with a colonocytic signature (*CA1+* late colonocytes, *CEACAM7+* colonocytes, and SPIB+ cells) show expression of cluster-specific markers. Normalized expression is scaled by *row*. (*E*) Crypt–axis scores (low near crypt bottom, high near crypt top) of cells shows the expected location of clusters along the crypt axis. Clusters are arranged on the x-axis by mean crypt–axis score (see the Methods section for more detail). The size of the *dot* corresponds to the level of expression of the known marker of the crypt top: *SELENOP*. A diagram of colonic crypt is shown (*right*). (*F*) UMAP overlain with the RNA velocity. RNA velocity is the ratio of spliced transcripts/ unspliced transcripts. Scale is normalized to 1. TA, transit-amplifying.

CD are *PLA2G2A* and *KLF6* (Figure 7A), both of which encode factors that are known to negatively regulate Wnt signaling in the crypts.[26,27] Although we found that *PLA2G2A* is up-regulated in many other clusters also, *KLF6* is increased significantly primarily in ISCs. Accordingly, we observed that numerous genes in the Wnt signaling
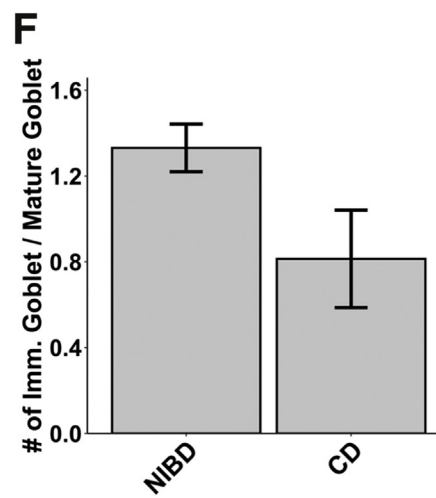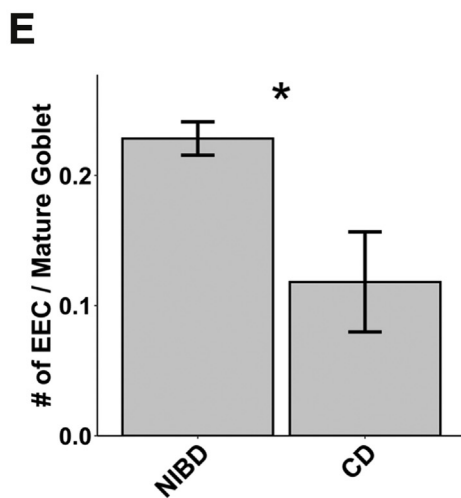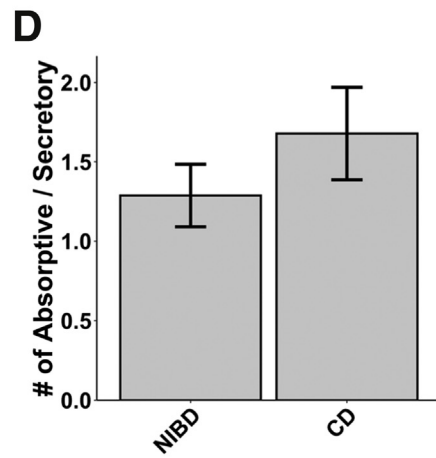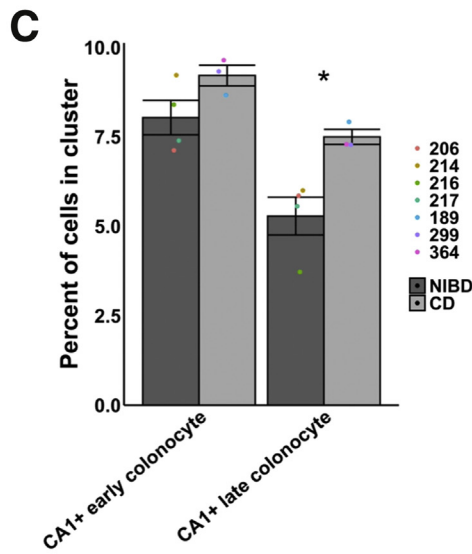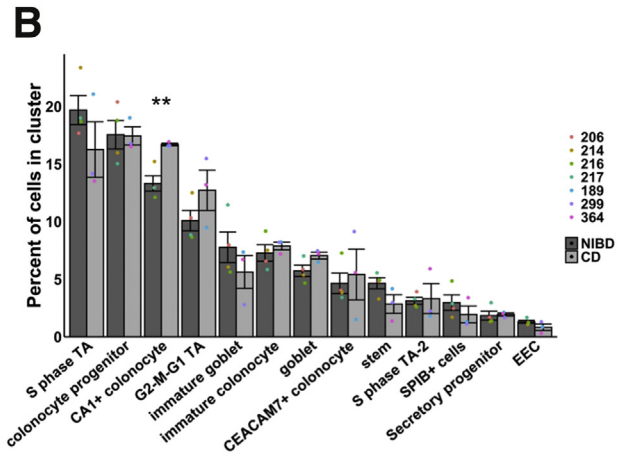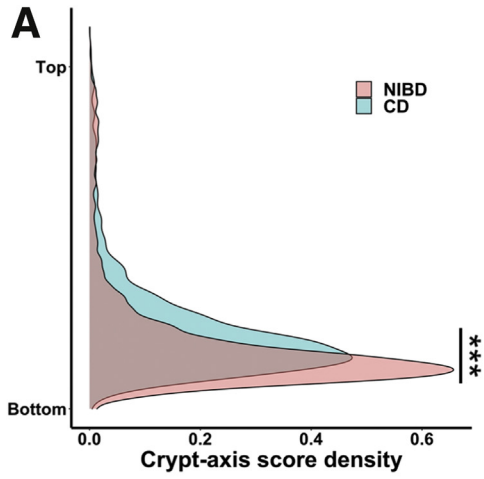
pathway (including *CDCA7*, *CDK6*, *CCDC115*, *MYC*, *RNMT*, *TGIF1*, *YBX1*, and *FOS*) are reduced significantly in expression in the ISC cluster in the CD samples relative to NIBD (Figure 7B). Moreover, in the case of *CCDC115* and *RNMT*, they are altered significantly only in ISCs and not in any other cluster (Figure 7C).

We next assessed whether the Wnt-responsive, canonical marker of colonic stem cells, *LGR5*,[28] is affected in CD. We found that both the percentage of *LGR5+* cells and the expression of *LGR5* in the ISC cluster are reduced significantly in CD relative to NIBD (Figure 7D). As a confirmation of this finding, we also detected a similar trend for *SMOC2* (Figure 7E), which is known to be specifically enriched in the *LGR5+* stem cell compartment.[29]

Intriguingly, we noticed that not all of the cells in the ISC cluster show high expression of *LGR5* or *SMOC2*, even in the NIBD samples (Figure 7F). Previous work has distinguished 3 subtypes of ISCs: ISC-I, ISC-II, and ISC-III. ISC-I comprises canonical *LGR5*-high stem cells, whereas ISC-II and ISC-III are composed of *LGR5*-low stem cells that are more proliferative, more differentiated, and potentially antigen-presenting.[30] We analyzed established markers[30] of the ISC-II and ISC-III subtypes, *CD74* and *LONP1*, respectively, and confirmed that they are indeed present almost exclusively in the cells of the ISC cluster that are *LGR5*-low or *LGR5*-negative (Figure 7G). Moreover, we observed that both the percentage of *CD74+* cells and the expression of *CD74* in the ISC cluster are increased significantly in CD relative to NIBD (Figure 7H). A similar trend was observed for the ISC-II marker gene *LONP1* (Figure 7G and H). We also found that the vast majority of ISC-I marker genes (including but not limited to *LGR5*) are reduced in representation in CD ISCs (Figure 7I). Taken together, these data show that there is a shift away from an ISC-I signature in CD, indicative of a disrupted, nonhomeostatic state in the colonic crypts.

### Detailed Analysis of the SPIB+ Cell Cluster Reveals New, Rare Cell Types Altered in CD

The least understood cell type that we have identified is the *SPIB+* cluster. This cluster is similar to the previously reported *BEST4/OTOP2* cells[3] or *BEST4+* enterocytes.[2] We opted to refer to this cluster as *SPIB+* because the expression of *SPIB* is ubiquitous across the cluster, whereas *BEST4* is prominent only in a subset of the cells in the cluster. To dissect this further, we performed subclustering within this cluster only, and identified 4 subclusters (Figure 8A). We found that each of these subclusters expresses a unique set

of markers (Methods section, Figure 8B and C). Notably, the gene *OTOP2* specifically marks *SPIB+* subcluster 1; *BEST4* (associated specifically with absorptive cells[31]) and *CA7* are present only in *SPIB+* subclusters 0 and 1; *SLC12A2* is enriched in *SPIB+* subclusters 2 and 3; and *LYZ* is increased in *SPIB+* subcluster 3, although it is not highly specific (Figure 8B and C). The crypt–axis score analysis showed a locational gradient of SPIB+ subclusters (Figure 8D). Only 1 of the subclusters (subcluster 1) is closer to the crypt top, whereas the others are closer to crypt bottom (Figure 8D). Notably, we found that the relative abundance of *SPIB+* subcluster 1 (*OTOP2+/BEST4+*) is increased significantly in CD as compared with NIBD (Figure 8E).

We also noted that several markers of proliferation are enriched in *SPIB+* subcluster 3 relative to the other *SPIB+* subclusters. We found that markers of G2–M–G1 proliferating cells (see the Methods section for more detail), such as *TOP2A*, *UBE2C*, *TPX2*, and *CENPF*, are enriched dramatically in *SPIB+* subcluster 3 relative to the other *SPIB+* subclusters (Figure 8F). On the other hand, several classic markers of secretory progenitors, such as *CLCA1*, MUC2, and ITLN1, are not found in *SPIB+* subcluster 3 (Figure 8F). Therefore, we labeled *SPIB+* subcluster 3 as *LYZ+* proliferating cells (Figure 8D). Notably, *LYZ* expression in these cells is reduced significantly in CD relative to NIBD (Figure 8G). We validated the presence of Lysozyme (LYZ) in crypt-bottom cells in NIBD samples and the depletion of this signal in CD samples by immunohistochemistry (Figure 8H). The lineage relationship among the SPIB+ subclusters and other clusters is shown in Figure 8I, supporting a maturational trajectory that mirrors that of the canonical absorptive lineage.

### Genes Implicated in CD Risk Show Cell Type–Specific Patterns of Aberrant Expression Within the Colonic Epithelium

To investigate the colonic epithelial expression of genes implicated in CD, we first defined a list of genes (n = 261) nearest to every genetic variant associated significantly with CD based on GWAS, of which 208 were detected in this data set (see the Methods section for more detail). Proximity of a GWAS signal to a gene does not confirm a link between the gene and the disease; however, it is a standard approach in the field for identifying candidates in the absence of orthogonal data such as chromosome conformation capture. For each of these genes we determined the percentage of cells positive and

---

**Figure 3. (*See previous page*). CD causes shifts in the colonic epithelial landscape.** (*A*) Crypt–axis score density for NIBD and CD cells separately shows a shift toward the colonic crypt top in CD. The significance of the shift was determined using the Kolmogorov–Smirnov test (\*\*\**P* < .001). (*B*) Mean cell abundances across NIBD and CD samples show a significant increase in *CA1+* colonocytes in CD. *Dots* show abundance values of individual samples. (*C*) Mean cell abundances for either *CA1+* early or late colonocyte clusters across NIBD and CD samples shows a significant increase only in the *CA1+* late colonocytes. *Dots* show abundances of individual samples. Abundances of other clusters are not shown. (*D*) Mean ratio of absorptive to secretory cells across NIBD and CD samples. (*E*) Mean ratio of EEC to mature goblet cells across NIBD and CD samples shows a significant shift in the secretory lineage toward mature goblet cells in CD. (*F*) Mean ratio of immature goblet to mature goblet across NIBD and CD samples show a shift toward mature goblet cells in CD. If not indicated, *P* values are calculated by a Student *t* test. \**P* < .05; \*\**P* < .01.
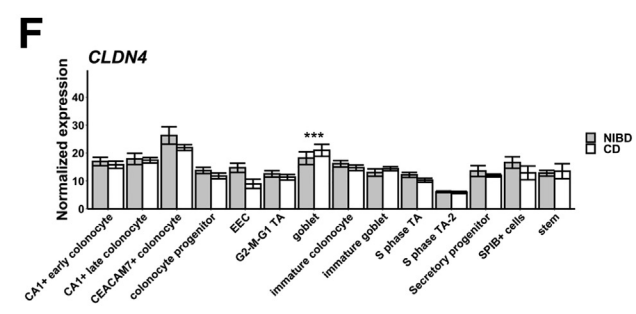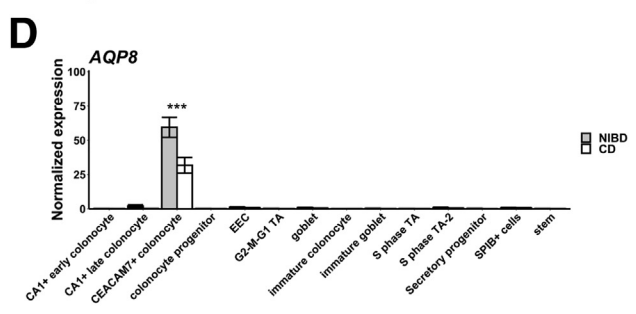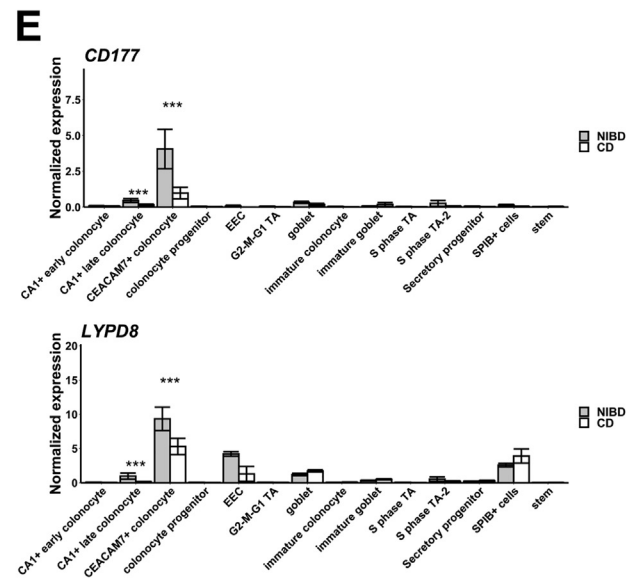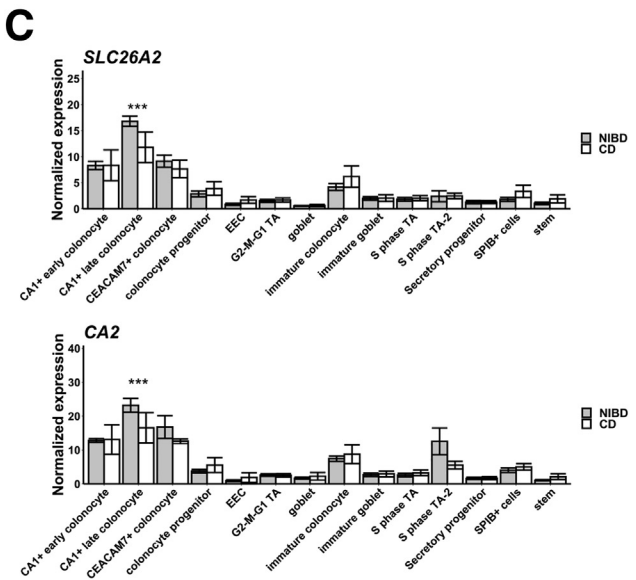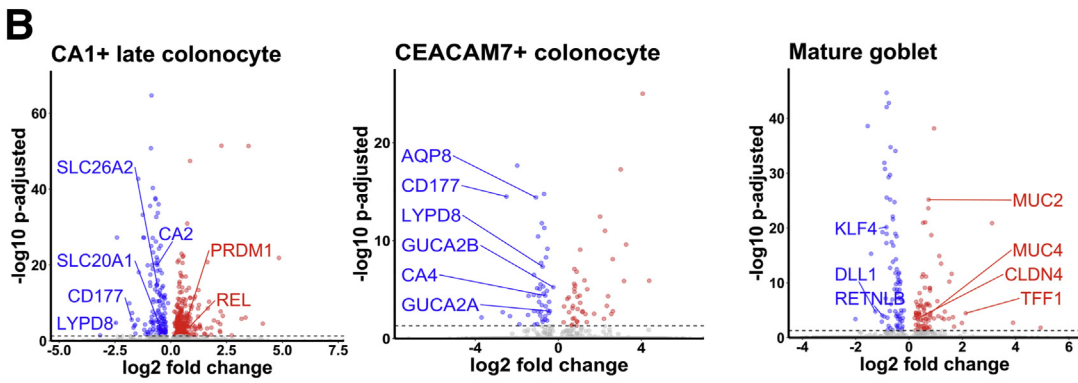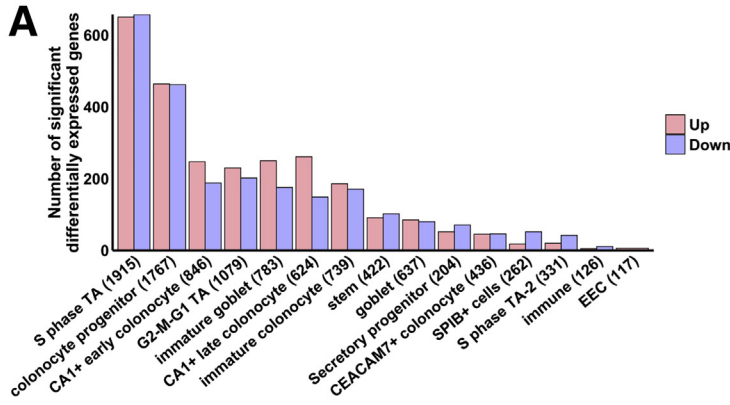
the average expression across cells within each of the 14 colonic epithelial cell clusters (Figure 9A). We confirmed that genes expected to be enriched in the lamina propria, such as *IL10* and *NOD2*, were not detected robustly in any of the epithelial cell clusters (Figure 9A). Indeed, several other genes were in this category as well, including *IL2RA*, *LTA*, *TNFSF8*, *CD244*, and *NELL1*, suggesting that if these genes are involved in CD pathophysiology, their primary roles are most likely outside of the epithelium. We observed that some CD genes are detected at comparable levels across many epithelial cell clusters, such as *ATG5*, *SKAP2*, *TAGLN2*, *PPM1G*, and *PRDX5*, whereas other genes are enriched in only 1 or few cell types, such as *ITLN1* (mature goblet cells) (Figure 9B), *CACNA2D1* (EECs), *COL5A1* (EECs), *RIPOR1* (EECs), *IFNGR2* (*CEACAM7*+ colonocytes), *NOTCH2* (*SPIB*+ cells), and *ATG16L2* (*SPIB*+ cells). Investigating the latter 2 further within the *SPIB*+ cluster, we found that while *NOTCH2* is expressed ubiquitously across all *SPIB*+ subclusters, *ATG16L2* is enriched in *SPIB*+ subcluster 1 (*LYZ*+ proliferating cells) (Figure 9C).

We next sought to determine whether any CD-associated genes are altered significantly in expression in CD relative to NIBD in 1 or more epithelial cell clusters. Some genes show significant changes in gene expression across many clusters (Figure 10A), including HLA family genes such as *HLA-DQB1*, *HLA-DRB1*, and *HLA-DRB5*. These data point to the systematic increase in CD genes encoding for major histocompatibility complex class II factors, generally active in antigen-presenting cells, across most epithelial cells including stem cells as described previously,[30] with some minor exceptions (eg, EECs for *HLA-DRB5* or *CA1*+ early colonocyte for *HLA-DQB1*). Most genes that change in CD show cell type–specific patterns. Notable examples include genes that are nearly specifically altered in EECs, including *CACNA2D1*, *COL5A1*, *RIPOR1*, *RNF123*, and *LEMD2*. The first 3 of these are more highly expressed in EECs in NIBD relative to other cell types (Figure 9A), whereas *RNF123* and *LEMD2* expression levels are comparable across most cell types in NIBD but uniquely suppressed in EECs in CD patients (Figures 9A and 10A).

Other examples include genes preferentially altered in stem cells, including *BACH2*, *GPR65*, and *TNFRSF6B*; in mature goblet cells, most notably *ITLN1*; as well as in *SPIB*+ cells, including *FCHSD2* and *NICN1*. In the case of *FCHSD2*, it is suppressed across all *SPIB*+ subclusters, whereas the change in *NICN1* is driven almost exclusively by *SPIB*+ subcluster 2 (*OLFM4*+/*SLC12A2*+). We found that the CD gene *JAK2* also is increased significantly in *SPIB*+ cells (specifically, subcluster 3 or the *LYZ*+

proliferating cells), although this aberration is shared with EECs as well (Figure 10A). The functions of *FCHSD2* and *NICN1* in CD pathophysiology currently are unknown, whereas *JAK2* is well studied in CD risk and Janus kinase inhibitors are currently Food and Drug Administration–approved for IBD.

## Discussion

While there are other single-cell studies of CD, a unique feature of this study was that the samples were isolated from the colonic epithelium of treatment-naïve individuals, which means that the results are not confounded by the effects of drugs. Importantly, we specifically avoided macroscopically inflamed tissue to focus on cellular reprogramming, which occurs more generally in the epithelial layer within the colon during CD. Inflammation also can lead to destruction of the epithelial layer, making it an unreliable indicator of intestinal epithelial cell (IEC) gene signatures. We identified 14 different types of epithelial cells, including several different subtypes of colonocytes and colonocyte progenitors. Each of the major clusters that express mature colonocyte markers, *CEACAM7*+ cells, *CA1*+ late colonocytes, and *SPIB*+ cells, express a unique set of genes. Notably, we found that *CEACAM7*+ cells are marked by *LINC01133*, which is a long noncoding RNA marker of a colonic epithelial cell type. Although *LINC01133* has been implicated in the control of tumor phenotypes in colon cancer,[32] its role in the normal function of *CEACAM7*+ colonocytes remains unknown and merits further investigation. In addition, among the colonocyte clusters, the *CA1*+ late cluster is the only colonocyte subtype for which relative abundance is altered significantly in CD, and this too warrants study in the future. Intriguingly, tuft cells were not among these epithelial cell types. Upon further investigation, we found that cells expressing *POU2F3*, a tuft cell marker necessary for tuft cell maturation, were removed during the quality filtering stages of the analysis, perhaps suggesting a sensitivity to the cell dissociation method used.

Study of cellular composition showed an increase in late-stage colonocytes in CD compared with healthy controls. However, in-depth analysis of the single-cell data showed aberrancies in the molecular character of these cells. For example, we observed a dramatic reduction of solute and water transporters, suggestive of compromised colonocyte function in CD. Particularly notable is the dramatic decrease in *AQP8* only in *CEACAM7*+ colonocytes. Although *AQP8* has been shown previously to be suppressed in IBD,[10] our study shows that this effect is very likely driven by 1 specific colonocyte subtype. Furthermore, the significant reduction

**Figure 4.** (***See previous page***). **Effect of CD within cell types.** (*A*) Number of significant differentially expressed genes in CD vs NIBD. Clusters are shown on the x-axis followed by the number of cells within the cluster in parenthetical notation and the number of differentially expressed genes is shown on the y-axis. (*B*) Differentially expressed genes for 3 clusters: CA1+ late colonocyte, CEACAM7+ colonocyte, and mature goblet. Log$_2$ fold change is shown on the x-axis and the adjusted *P* value is shown on the y-axis. Mean expression of (*C*) *SLC26A2* and *CA2*, (*D*) *AQP8*, (*E*) *CD177* and *LYPD8*, and (*F*) *CLDN4* across clusters for NIBD and CD samples shows the specificity of expression and differential expression within clusters. *P* values were calculated by the Wilcoxon rank-sum test. ***Adjusted *P* < .001.

**Table 2.** Significant (*P* Adjusted < .05) Differentially Expressed Genes in CD Compared With NIBD That Are Specific to *CA1+* Late Colonocytes

| Gene | p-value | average log fold change | adjusted p-value | average NIBD expression | average CD expression | log$_2$ fold change |
|------|---------|------------------------|------------------|------------------------|----------------------|--------------------|
| SLC26A2 | 4.35E-20 | -0.40 | .00 | 16.95 | 11.06 | -0.62 |
| CA2 | 4.77E-25 | -0.39 | .00 | 23.00 | 15.18 | -0.60 |
| ADIRF | 9.37E-09 | -0.30 | .00 | 3.73 | 2.51 | -0.58 |
| HSPA1B | 5.62E-12 | -0.25 | .00 | 1.23 | 0.74 | -0.73 |
| HIGD1A | 6.10E-07 | -0.25 | .02 | 3.18 | 2.27 | -0.49 |
| SLC20A1 | 6.22E-08 | -0.22 | .00 | 3.69 | 2.77 | -0.41 |
| GPA33 | 1.54E-07 | -0.21 | .00 | 2.03 | 1.45 | -0.48 |
| SCIN | 4.50E-09 | -0.21 | .00 | 0.83 | 0.48 | -0.78 |
| SLC4A4 | 3.63E-08 | -0.19 | .00 | 1.85 | 1.36 | -0.44 |
| C3orf85 | 1.40E-07 | -0.17 | .00 | 0.54 | 0.30 | -0.85 |
| AC015912.3 | 6.87E-07 | -0.17 | .02 | 0.40 | 0.18 | -1.13 |
| NDUFC2 | 8.51E-07 | -0.15 | .02 | 2.73 | 2.22 | -0.30 |
| FASTKD1 | 6.05E-07 | -0.11 | .01 | 0.27 | 0.13 | -1.01 |
| MEF2C | 1.23E-08 | -0.11 | .00 | 0.17 | 0.05 | -1.67 |
| SLC35E2A | 1.32E-06 | -0.05 | .03 | 0.06 | 0.01 | -3.16 |
| GATD3A | 6.54E-07 | 0.07 | .02 | 0.04 | 0.11 | 1.50 |
| NR1I2 | 1.20E-06 | 0.10 | .03 | 0.11 | 0.22 | 1.03 |
| ZNRD2 | 1.17E-06 | 0.10 | .03 | 0.16 | 0.28 | 0.85 |
| TMEM120B | 6.62E-09 | 0.10 | .00 | 0.08 | 0.20 | 1.30 |
| TMC6 | 9.28E-07 | 0.11 | .02 | 0.15 | 0.28 | 0.96 |
| MAP1LC3A | 1.38E-06 | 0.12 | .03 | 0.35 | 0.53 | 0.58 |
| IMP4 | 5.99E-07 | 0.12 | .01 | 0.27 | 0.43 | 0.70 |
| SNRK | 7.57E-07 | 0.12 | .02 | 0.27 | 0.44 | 0.68 |
| SLC25A25 | 5.18E-07 | 0.13 | .01 | 0.25 | 0.42 | 0.74 |
| FUOM | 1.36E-06 | 0.13 | .03 | 0.33 | 0.52 | 0.63 |
| TRAPPC4 | 4.71E-07 | 0.13 | .01 | 0.29 | 0.46 | 0.69 |
| FERMT1 | 2.75E-07 | 0.13 | .01 | 0.56 | 0.78 | 0.48 |
| PSMD1 | 1.85E-07 | 0.14 | .00 | 0.70 | 0.95 | 0.44 |
| RHOF | 2.29E-07 | 0.14 | .01 | 0.34 | 0.54 | 0.67 |
| REL | 8.29E-09 | 0.14 | .00 | 0.25 | 0.44 | 0.83 |
| RNASET2 | 7.75E-07 | 0.15 | .02 | 0.72 | 1.00 | 0.48 |
| GALNT5 | 7.33E-09 | 0.16 | .00 | 0.18 | 0.38 | 1.09 |
| SLC7A1 | 3.48E-10 | 0.16 | .00 | 0.28 | 0.50 | 0.86 |
| KLF13 | 1.54E-07 | 0.16 | .00 | 0.87 | 1.20 | 0.47 |
| PRDM1 | 3.08E-07 | 0.17 | .01 | 0.71 | 1.03 | 0.54 |
| TOR1AIP2 | 4.03E-07 | 0.17 | .01 | 0.69 | 1.01 | 0.54 |
| RAB11FIP1 | 1.18E-07 | 0.18 | .00 | 3.98 | 4.94 | 0.31 |
| INPP1 | 1.83E-07 | 0.18 | .00 | 0.52 | 0.82 | 0.66 |
| QPRT | 2.34E-09 | 0.19 | .00 | 0.17 | 0.41 | 1.29 |
| EMP2 | 7.93E-07 | 0.19 | .02 | 1.14 | 1.59 | 0.47 |
| NEDD4L | 1.30E-07 | 0.19 | .00 | 0.96 | 1.38 | 0.53 |
| ODC1 | 1.06E-06 | 0.21 | .03 | 2.10 | 2.82 | 0.43 |
| FRMD1 | 6.90E-09 | 0.23 | .00 | 0.22 | 0.53 | 1.28 |
| CCND2 | 6.13E-09 | 0.24 | .00 | 0.74 | 1.21 | 0.72 |
| EFNA1 | 2.28E-11 | 0.28 | .00 | 0.74 | 1.29 | 0.81 |

**Table 3.** Significant (*P* Adjusted < .05) Differentially Expressed Genes in CD Compared With NIBD That Are Specific to *CEACAM7*+ Colonocytes

| Gene | p-value | average logFC | adjusted p-value | average NIBD expression | average CD expression | $\log_2$ fold change |
|---|---|---|---|---|---|---|
| LGALS1 | 2.85E-07 | -0.78 | .01 | 1.42 | 0.11 | -3.70 |
| AQP8 | 1.58E-19 | -0.73 | .00 | 60.74 | 28.71 | -1.08 |
| IL32 | 1.60E-08 | -0.51 | .00 | 19.62 | 11.35 | -0.79 |
| AK1 | 7.45E-09 | -0.49 | .00 | 4.51 | 2.37 | -0.93 |
| CFDP1 | 1.19E-10 | -0.46 | .00 | 4.61 | 2.54 | -0.86 |
| PPP1R14A | 2.09E-07 | -0.45 | .01 | 0.83 | 0.17 | -2.33 |
| NQO1 | 1.42E-09 | -0.45 | .00 | 1.85 | 0.81 | -1.18 |
| RHOC | 1.02E-09 | -0.43 | .00 | 8.50 | 5.20 | -0.71 |
| CA4 | 1.68E-09 | -0.41 | .00 | 22.72 | 14.66 | -0.63 |
| DIO3OS | 8.15E-08 | -0.38 | .00 | 0.61 | 0.09 | -2.68 |
| CDKN2B-AS1 | 7.06E-08 | -0.38 | .00 | 2.52 | 1.41 | -0.84 |
| OAZ1 | 2.00E-08 | -0.37 | .00 | 8.83 | 5.79 | -0.61 |
| S100A13 | 1.61E-07 | -0.35 | .00 | 0.86 | 0.31 | -1.46 |
| FTH1 | 5.50E-08 | -0.35 | .00 | 180.01 | 126.94 | -0.50 |
| HSPB1 | 4.66E-07 | -0.35 | .01 | 4.69 | 3.02 | -0.63 |
| SCNN1B | 1.56E-06 | -0.35 | .04 | 1.16 | 0.52 | -1.14 |
| NBL1 | 2.05E-07 | -0.33 | .01 | 3.70 | 2.39 | -0.63 |
| SPINT1-AS1 | 1.30E-06 | -0.30 | .03 | 1.01 | 0.49 | -1.03 |
| JUND | 7.32E-09 | -0.29 | .00 | 14.86 | 10.92 | -0.44 |
| GUCA2A | 6.48E-08 | -0.28 | .00 | 51.86 | 38.99 | -0.41 |
| CELA3B | 1.43E-06 | -0.23 | .04 | 0.40 | 0.11 | -1.88 |
| GUCA2B | 2.26E-10 | -0.16 | .00 | 23.29 | 19.63 | -0.25 |
| REG1A | 3.61E-09 | 0.16 | .00 | 1.00 | 1.34 | 0.43 |
| MT-ND2 | 8.83E-07 | 0.17 | .02 | 64.15 | 76.32 | 0.25 |
| RPS6 | 2.56E-09 | 0.30 | .00 | 5.92 | 8.37 | 0.50 |
| RPL23 | 3.30E-07 | 0.38 | .01 | 1.91 | 3.24 | 0.76 |
| RPS7 | 3.95E-07 | 0.39 | .01 | 3.55 | 5.74 | 0.69 |
| RPL10A | 2.19E-09 | 0.51 | .00 | 2.87 | 5.41 | 0.92 |
| FDPS | 8.94E-09 | 0.52 | .00 | 1.80 | 3.72 | 1.05 |

in CD of *LYPD8*, which encodes an important antimicrobial factor[12–14] in both *CEACAM7*+ colonocytes and *CA1*+ late colonocytes, is a possible indication of the beginning stages of impaired colonocyte contribution to barrier function. We also show that *LYPD8* is reduced in expression in EECs from CD patients, which has not been shown previously.

Of note, we found that *LYPD8* is increased modestly, albeit not significantly, in the mature goblet cell cluster. In addition, mature goblet cells show both an increased abundance in CD as well as increased expression of genes that code for mucins, including *MUC2* and *MUC4*. This result is consistent with another recent study[1] that reported a significant increase in goblet cells in pediatric ileal CD samples. This finding, coupled with alterations in the molecular phenotype of mature colonocytes, raises the possibility of a compensatory response of mature goblet cells to the potentially weakened function of colonocytes in CD. This notion is supported further by significantly increased *CLDN4* expression in mature goblet cells in CD, which encodes for a tight-junction protein that promotes barrier

integrity,[16] to levels that match what has been observed in *CEACAM7*+ colonocytes (where *CLDN4* expression is normally most prominent).

Analysis of the single-cell TE expression data showed a selective set of TE families that behave as markers of specific cell types of the colonic epithelium. Most strikingly, high *MER11A* expression marks goblet cells while *MER11C* marks SPIB+ cells. This is surprising because *MER11A* and *MER11C* are 2 closely related subfamilies of endogenous retroviral LTRs (85% nucleotide similarity in their consensus sequences), yet they mark distinct cell populations (Figure 6B–E). Interestingly, it has been reported that MER11A and MER11C elements are bound by unique combinations of KRAB–zinc finger proteins as measured by chromatin immunoprecipitation-exonuclease (ChIP-exo) assays in human embryonic kidney (HEK) 293 cells.[33] Because Krüppel associated box (KRAB)–zinc finger proteins act as sequence-specific transcriptional repressors of TEs,[34] it would be interesting to investigate whether their differential expression contributes to the cell type–specific
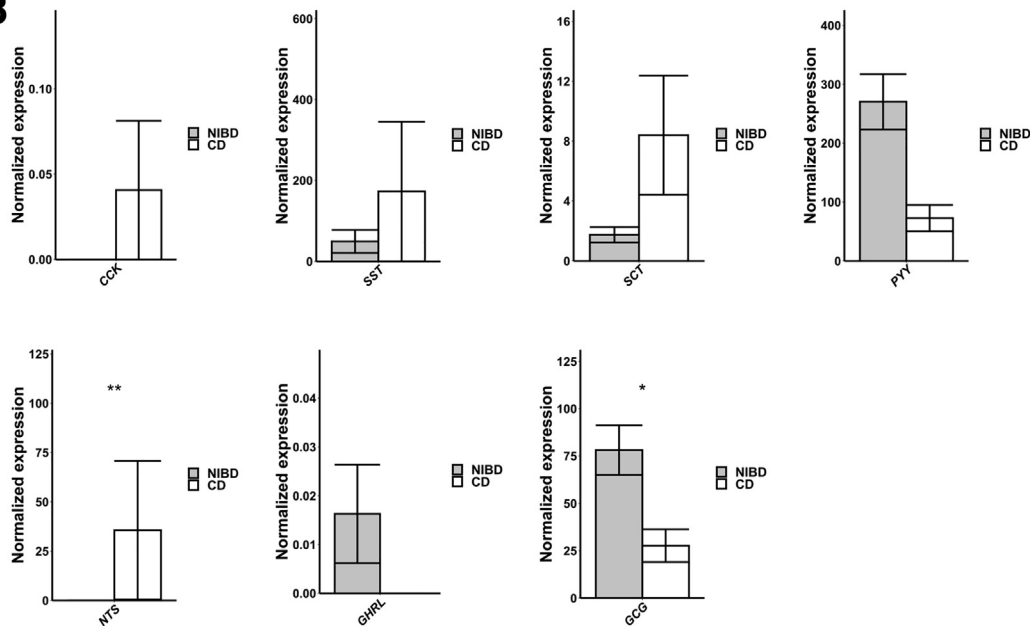
**A**
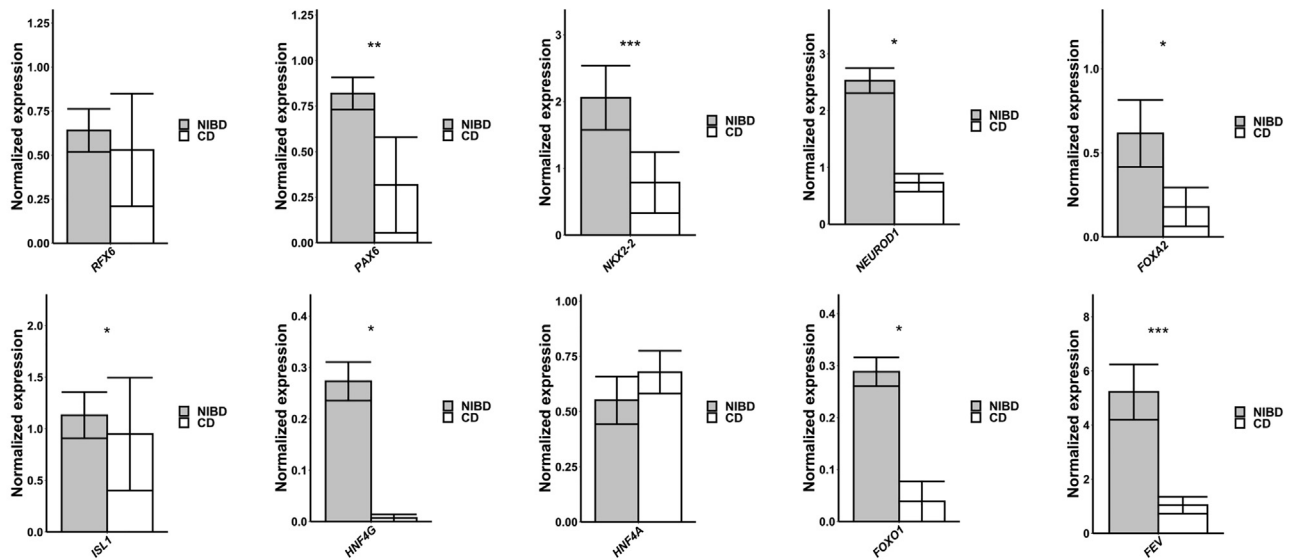


**B**



**Figure 5. CD disrupts EEC homeostasis.** (*A*) Mean normalized expression across 10 transcription factors involved in EEC maturation in NIBD and CD samples. (*B*) Mean normalized expression across 7 EEC hormones in NIBD and CD samples. *P* values were calculated using the Wilcoxon rank-sum test. *$P < .05$, **$P < .01$, and ***$P < .001$.

regulation of MER11A and MER11C in the colonic epithelium. Because endogenous retroviral LTRs often regulate adjacent gene expression,[20,25] it also is possible that these elements are involved in regulating cell type–specific gene expression. For instance, 1 copy of *MER11A* is located immediately adjacent to *ZG16*, which codes for zymogen granule 16, specifically produced and released by goblet cells.[35] The potential regulatory connection between *MER11A* and *ZG16* merits further investigation.

With regard to TEs dysregulated in CD, we found that the *L1PA10* subfamily is up-regulated the most significantly in goblet cells (Figure 6*F* and *G*). Again, the specificity of this response to the *L1PA10* subfamily is intriguing because there are many closely related L1 subfamilies in the human genome, but we do not observe other L1 subfamilies that are dysregulated significantly. It has been shown that L1 expression is up-regulated and correlates with increased inflammation in colon cancer.[36] Our findings suggest that
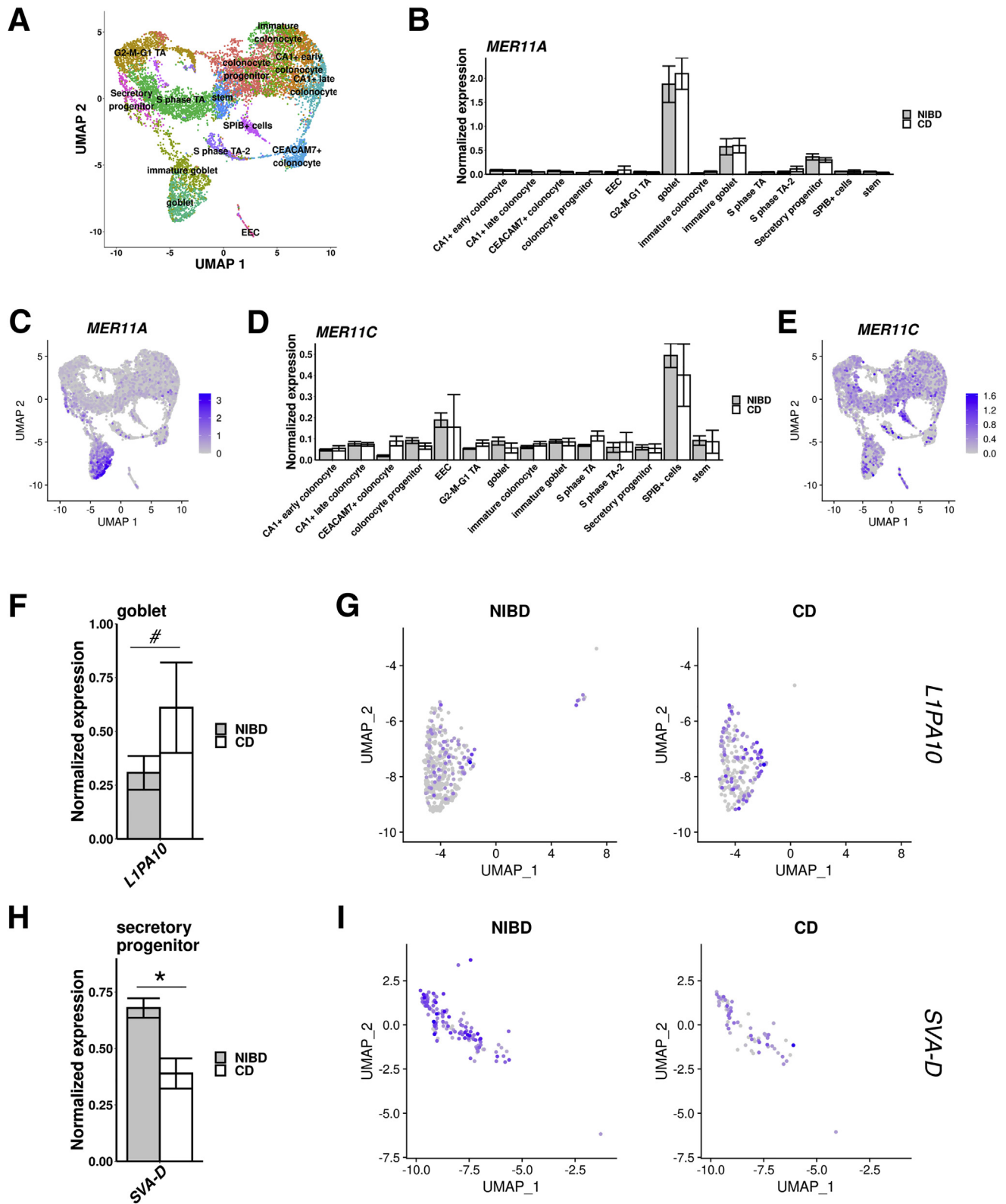
**Figure 6. Transposable elements define cell types and are dysregulated in CD.** (*A*) Uniform Manifold Approximation and Projection (UMAP) of epithelial cell clustering after TE inclusion maintains structure. Mean expression of (*B*) *MER11A* and (*D*) *MER11C* across clusters for NIBD and CD samples shows the specificity of expression within clusters. UMAP overlain with the normalized expression of TE markers (*C*) *MER11A* or (*E*) *MER11C* further confirms cluster specificity. (*F*) Mean expression for *L1PA10* in the goblet cluster, and (*G*) UMAP of the goblet cluster with expression of *L1PA10* overlain, separated by NIBD/CD. (*H*) Mean expression for *SVA-D* in the secretory progenitor cluster, and (*I*) UMAP of the secretory progenitor cluster with expression of *SVA-D* overlain, separated by NIBD/CD. *P* values were calculated by the Wilcoxon rank-sum test. #Adjusted *P* < .1, *adjusted *P* < .05.
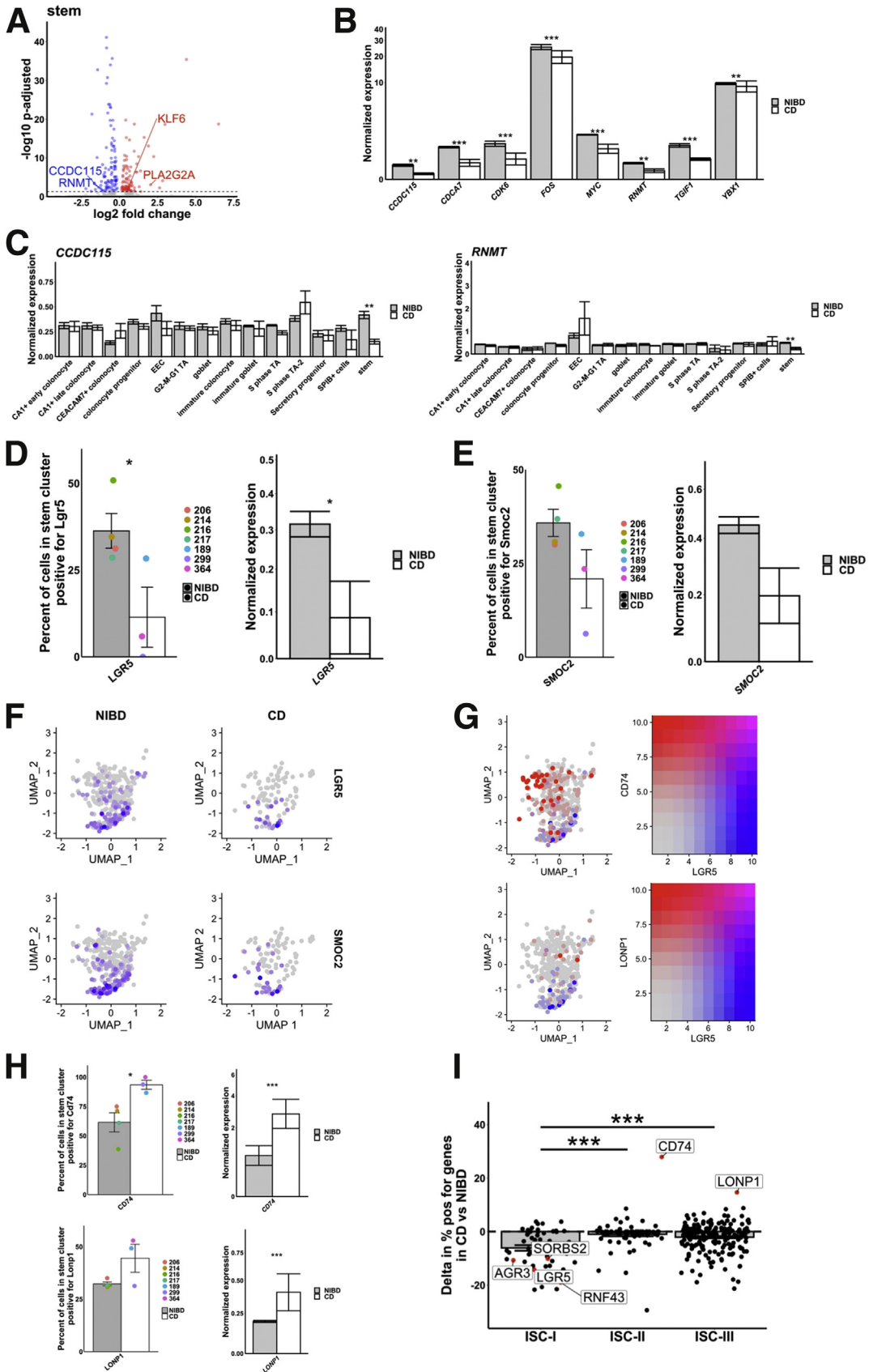
overexpression of L1 elements could contribute to disease pathogenesis and/or exacerbation in CD.

We found a novel shift in ISCs away from the canonical *LGR5*+ signature in CD compared with NIBD, which motivates new avenues of future investigations into the dysfunction in ISCs as well as in lineage determination and colonic epithelial renewal during CD development. Among the numerous pro-Wnt signaling marker genes with reduced expression in ISCs in CD, 1 are uniquely suppressed only in ISCs: *CCDC115* and *RNMT*. RNA guanine-7 methyltransferase (RNMT) is recruited to Wnt signaling gene promoters by MYC,[37] which we show is suppressed in ISCs in CD, and therefore may be involved in mediating the shift away from *LGR5*+ ISCs. We also found increased expression of major histocompatibility complex class II factors such as *CD74*, *HLA-DQB1*, and *HLA-DRB1*, normally associated with antigen presentation, in ISCs of CD patients. A recent study showed that murine small intestinal stem cells act as nonconventional antigen-presenting cells to activate lamina propria T cells, especially under certain conditions such as enteric infection.[30] Whether this also occurs in CD is not known and merits further investigation.

We identified and characterized the *SPIB*+ cells (similar to what was reported previously as *BEST4*+/*OTOP2*+ cells[3] or *BEST4*+ enterocytes[2]) in the context of CD in the colon. We annotated these cells as *SPIB*+ rather than *BEST4*+ because of the near-ubiquitous expression of *SPIB*, whereas *BEST4* is expressed only in a subset of the 4 distinct subclusters. Lineage analysis suggests that these subclusters likely are part of a maturational trajectory that mimics that of the canonical absorptive cells. Among these subclusters, the *OTOP2*+/*BEST4*+ subcluster is increased significantly in relative abundance in CD compared with NIBD. We also identified a subcluster of *SPIB*+ cells comprising *LYZ*+ proliferating cells. Notably, *LYZ* expression is reduced and *LYZ*+ cells are depleted in CD compared with NIBD, both at the RNA and protein level.

Finally, in this study, we link CD GWAS risk genes to specific epithelial cell subtypes in which they are detected and/or aberrantly expressed. These discoveries may offer clues about the potential molecular mechanisms by which these genes contribute to CD etiology. For example, *ATG16L2* is enriched in the *SPIB*+ cluster, particularly within the *SPIB*+/*LYZ*+ subcluster, suggesting that this

gene may contribute to CD etiology through the novel functions of this uncharacterized cluster. We also found that increased *JAK2* expression is prominent not only in the *SPIB*+ cluster but also in EECs, which is completely unexpected and merits more detailed functional investigation given that Janus kinase inhibitors are approved for use in ulcerative colitis and are in clinical trials for CD. Overall, we believe that this study offers a unique picture, at unprecedented resolution, of the cellular and molecular landscape of the colonic epithelium in treatment-naïve adult CD.

Future single-cell studies must expand on this work to include larger cohorts and to assess the impact of age, sex, CD subtype,[38,39] disease region (eg, ileum vs colon), disease duration, and treatment history on cell composition and molecular phenotype. Longitudinal investigations to uncover changes to the cellular landscape and molecular phenotype over the course of disease progression also are merited. Finally, it will be exciting in the future to explore the relationship between specific luminal bacterial changes and alterations at the single-cell level, given the intimate relationship between microbial dysbiosis and CD pathogenesis.

# Methods
## Ethical Statement

The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice. The study protocol was approved by the Institutional Review Board at University of North Carolina at Chapel Hill (approval number: 19-0819 and 17-0236). All participants provided written informed consent before inclusion in the study. All participants are identified by number and not by name or any protected health information.

## scRNA-seq

Colonic mucosa was obtained endoscopically as biopsy specimens from patients with treatment-naive CD and NIBD healthy controls. Cross-sectional clinical data were collected at the time of sampling. All samples were collected from regions of ascending colon without macroscopic inflammation. Isolation of colonic primary IECs was performed as reported previously.[40,41] This method has been shown previously to result in more than 95% purity of IECs.[42] Single-cell libraries were constructed using the Chromium

---

**Figure 7.** (*See previous page*). **CD disrupts stem cell homeostasis.** (*A*) Differentially expressed gene within the stem cluster. Log$_2$ fold change is shown on the x-axis and the adjusted *P* value is shown on the y-axis. (*B*) Mean expression of 8 mediators of the Wnt pathway in NIBD and CD samples show a consistent decrease in response to CD. (*C*) Mean expression of *CCDC115* and *RNMT* across clusters shows a significant decrease only occurs in the stem cluster. Percentage of cells positive for (*D*) *LGR5* (*left*) and (*E*) *SMOC2* (*left*) or the average normalized expression for (*D*) *LGR5* (*right*) and (*E*) *SMOC2* (*right*) in the stem cluster for NIBD or CD samples. *Dots* relate to individual sample value. (*F*) Uniform Manifold Approximation and Projection (UMAP) of the stem cluster with expression of *LGR5* or *SMOC2* overlain, separated by NIBD/CD. (*G*) UMAP of the stem cluster with *LGR5* expression (red) co-overlain with either *CD74* (*top*) or *LONP1* (*bottom*) (both blue) expression. *Right*: Scale for expression is shown. (*H*) Percentage of cells positive for *CD74* (*left top*) and *LONP1* (*left bottom*) or the average normalized expression for *CD74* (*right top*) and *LONP1* (*right bottom*) in the stem cluster for NIBD or CD samples. *Dots* relate to individual sample value. (*I*) Shift in percentage of positive genes in either ISC-I, ISC-II, or ISC-III cell types between NIBD and CD. *P* values for expression bar plots were calculated using the Wilcoxon rank-sum test and the *P* value for the ISC subtype shifts was calculated using a Student *t* test. *P < .05, **P < .01, and ***P < .001.
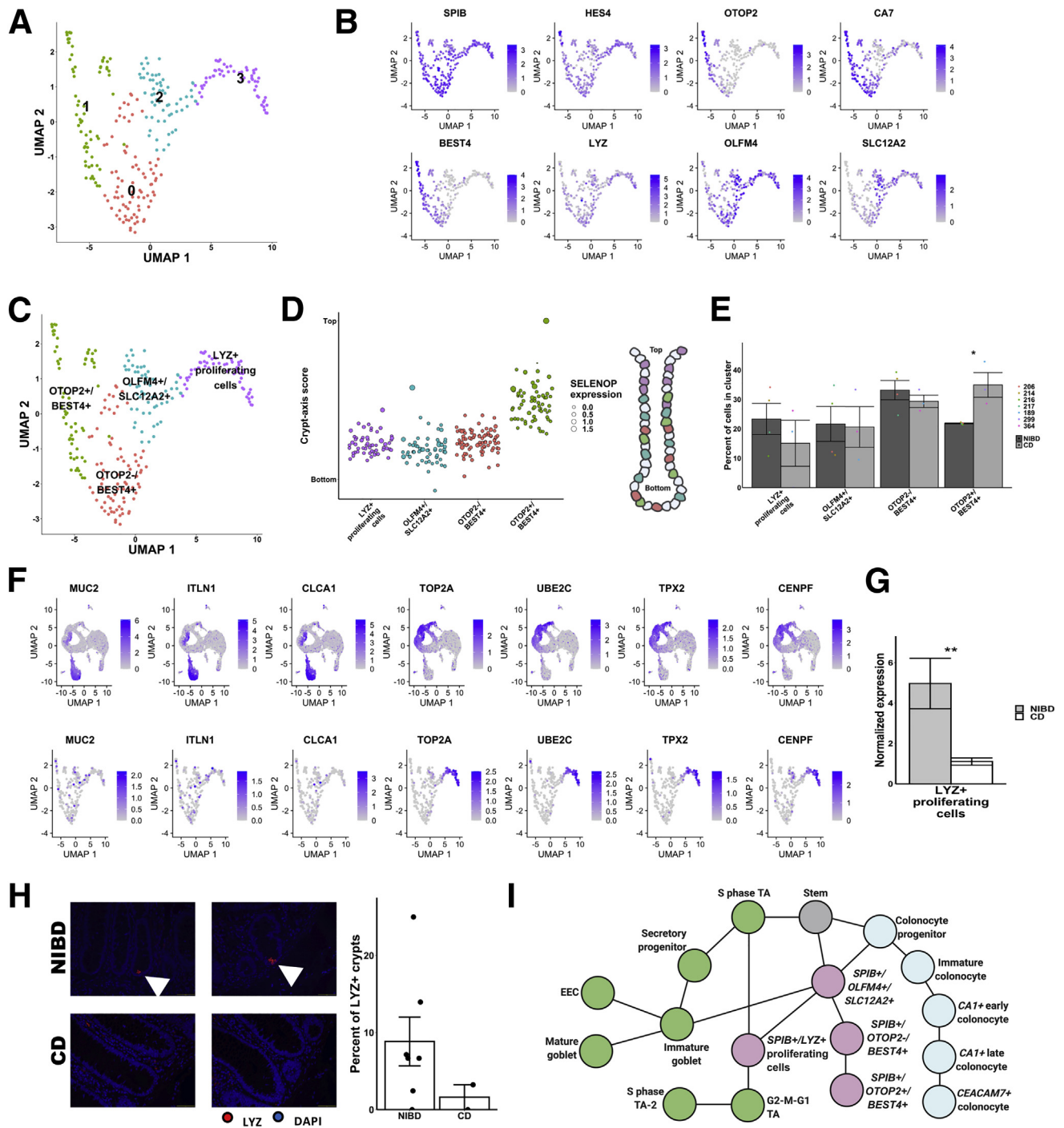
**Figure 8. Subcluster analysis of SPIB+ cells shows distinct lineages.** (*A*) Uniform Manifold Approximation and Projection (UMAP) after subclustering of SPIB+ cells with labeling of the 4 identified subclusters. (*B*) UMAP of SPIB+ cell subclusters overlain with gene expression from genes marking the entire cluster or distinct subclusters. (*C*) SPIB+ cell subcluster UMAP following cell type assignment using highly enriched markers of the subclusters. (*D*) Crypt–axis scores (low near crypt bottom, high near crypt top) of subcluster-assigned cells uncovers a colonocyte signature in the OTOP2+/BEST4+ subcluster. Clusters are arranged on the x-axis by mean crypt–axis score. The size of the *dot* corresponds to the level of expression of *SELENOP,* a known marker of the top of the crypt. (*E*) Mean cluster cell abundances across NIBD and CD samples shows a significant increase in OTOP2+/BEST4+ subcluster in CD. *Dots* show abundances of individual samples. (*F*) UMAP of either all clusters (*top*) or SPIB+ cell subclusters (*bottom*) overlain with the expression of markers of secretory progenitor cells. (*G*) Mean normalized expression of *LYZ* in NIBD and CD samples in the LYZ+ proliferating subcluster. (*H*) LYZ-detected immunofluorescence in colonic crypts from either NIBD (n = 181 crypts, 5 patients) or CD (n = 76 crypts, 2 patients) patient examples (*left*) and quantified by sample (*right*). Mean percentage of crypts positive for LYZ in NIBD or CD. (*I*) Lineage reconstruction as determined by partition-based graph abstraction (PAGA). *P* values are calculated by a Student *t* test. *$P < .05$. DAPI, 4′,6-diamidino-2-phenylindole.

Single Cell 3' Reagent Kits (v3) (10X Genomics, Pleasanton, CA) according to the manufacturer's instructions. Single-cell sequencing data are available under GEO accession GSE164985.

## Single-Cell Transcriptome Data Analysis

Sample single-cell fastqs were aligned to the human genome (GRCh38-3.1.0) using 10x Genomics cellranger count (v4.0.0) (Pleasanton, CA) to obtain gene/cell count matrices. Sample filtering, normalization, integration, clustering, and visualization was accomplished using Seurat (v3.2) (New York, NY). To maintain sample quality, cells with less than 1000 detected genes or greater than 25% of reads aligning to mitochondrial genes were removed. In addition, cells with more than 50,000 reads were removed as potential doublets. To control for read depth, sample counts were log-normalized. Samples were merged using the Seurat integration anchor workflow based on the 2000 most variable genes. Clustering and identification of nearest neighbors relied on 11 principal component analysis (PCA) dimensions. Cells were clustered at a resolution of 0.7. To focus on epithelial cells, immune clusters were identified by expression of known immune cell type markers. Immune clusters were removed, and the integration workflow was repeated using the 2000 most variable genes of the remaining epithelial cells. Clustering and identification of nearest neighbors relied on 10 PCA dimensions. Epithelial cells were clustered using a resolution of 0.8. Highly enriched markers for each cluster were determined using the Seurat function FindAllMarkers, which compares the gene expression within a cluster with all other clusters. Genes that are up-regulated with a log fold change greater than 0.25 and a Wilcoxon rank-sum test $P$ value less than .01 were considered highly enriched markers. Clusters were assigned cell types based on expression of known markers and/or highly enriched markers of the cluster. Despite earlier filtering, one of the resulting clusters was immune cells and therefore was removed from the analysis. Samples were found to have similar count, genes, and mitochondrial read percentage per cell (Table 4). Differential expression of genes between CD and NIBD samples was determined within clusters using the Wilcoxon rank-sum test.

Subclustering of SPIB+ cells was performed to identify discrete cell types within the cluster. The SPIB+ cells were subclustered using the integrated counts. Clustering and identification of nearest neighbors relied on 10 PCA dimensions. SPIB+ cells were subclustered using a resolution of 0.8. Assignment of cell types to subclusters was determined by the highly enriched markers of the subcluster.

## Single-Cell Transcriptome Data Analysis Including Transposable Elements

To estimate the simultaneous expression of genes and TEs, we first extracted the genes (Gencode V19) (Hinxton, Cambridgeshire, UK) and TEs (repeat masked elements) sequences and appended to build a transcriptome, comprising the coding sequences plus untranslated regions of genes and TE sequences in fasta format. We then filtered out those TE sequences that meet the following criteria: (1) overlapping with exons/untranslated regions of genes, (2) <100 bp, and (3) DNA/satellite/short tandem repeats (STRs)/simple repeat transposons and Alu elements. This resulted in approximately 150,000 contigs comprised of distinct genes and TE sequences. To guide the transcriptome assembly, we also curated the gene models (general transfer format [gtf]) as distinct coordinates for each of the contigs. The concatenated gene/TE transcriptome assembly and genome reference was indexed using Salmon (College Park, MD). The demultiplexed reads were aligned to the custum reference genome described above and quantified using Alevin (College Park, MD). Individual TEs often are duplicated throughout the genome, and expression across these loci were collapsed into a single TE family. The resulting gene/TE family and cell count matrices were analyzed using Seurat (v3.2) (New York, NY). Sample filtering, normalization, integration, clustering, and visualization was performed identically to the non-TE analysis. For comparison purposes, only cells present in the non-TE analysis were included and cell cluster identities were maintained. Highly enriched TEs were defined as being up-regulated more than 0.4 in the cluster of interest vs all other clusters and the percentage of cells expressing the TE in the cluster of interest had to be twice the percentage in all other clusters. To be included for the differential expression within clusters, TEs had to have an adjusted $P$ value less than .1 and log fold-change greater than 0.3 or less than -0.3.

## Determination of Highly Enriched Marker Genes

Highly enriched genes were determined by 1 of 2 different thresholds. In both thresholds, the genes had to be up-regulated more than 0.5 log fold change in the cluster of interest vs all other clusters. Using the stringent threshold, a gene had to be expressed in more than 90% of the cells in the cluster of interest and in less than 30% in all other clusters (Figure 6F). Using the slightly relaxed threshold, a gene had to be expressed in more than 80% of cells in the cluster of interest and in less than 40% of cells in all other clusters (Figure 1D).

---

**Figure 9.** (*See previous page*). **Expression of CD-associated risk genes across clusters.** (*A*) Dot plots showing expression and percentage of cells expressing CD-associated risk genes (*columns*) across clusters (*rows*). For each gene, the *dot* color corresponds to the average scaled expression and the size of the *dot* corresponds to the percentage of cells in the cluster expressing the gene. (*B*) Uniform Manifold Approximation and Projection (UMAP) of all epithelial clusters showing *MUC2* (blue) and *ITLN1* (red) expression. (*C*) UMAP of SPIB+ cell subclusters overlain with gene expression from *NOTCH2* or *ATG16L2*. avg.exp.scaled, average expression, scaled; pct.exp, percent of cells expressing gene.
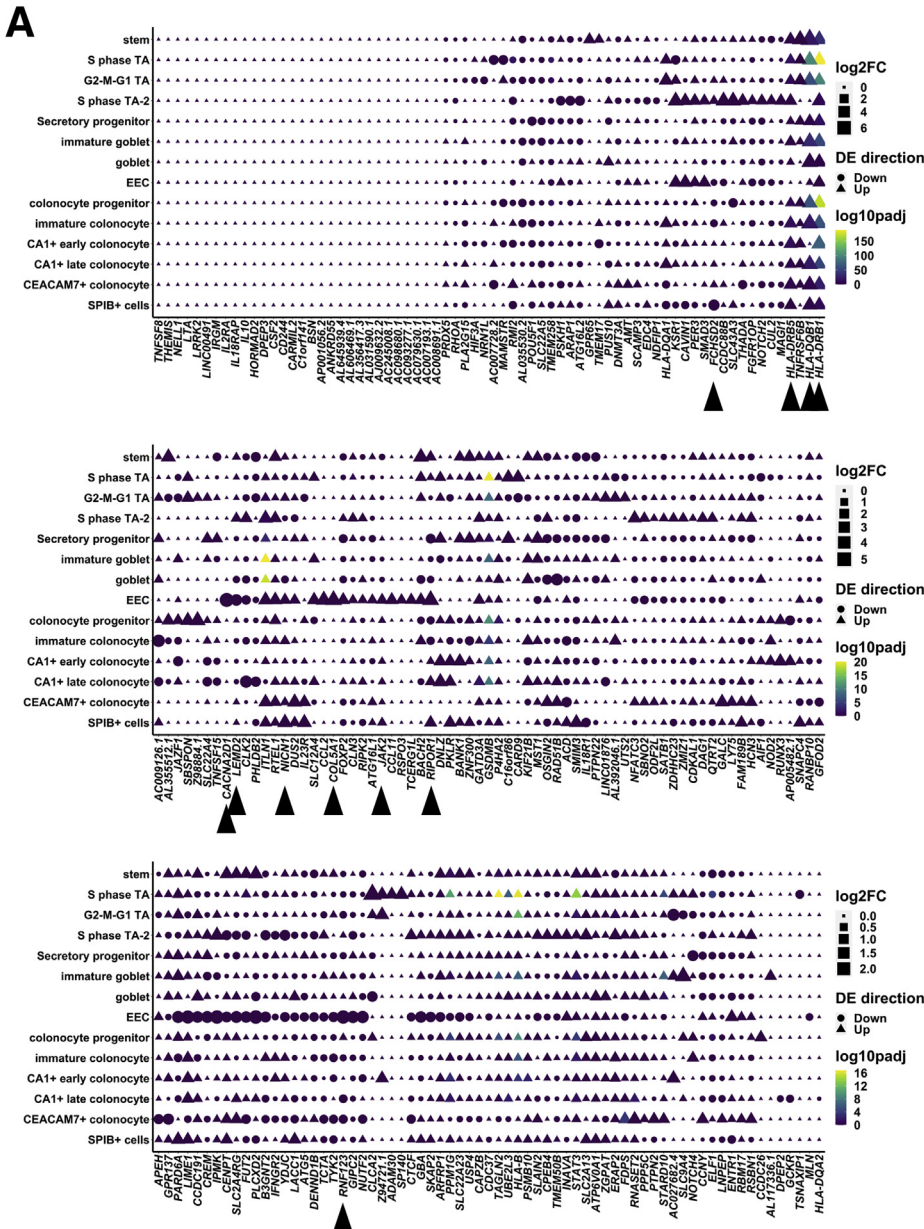
**Figure 10. Differential expression of CD-associated risk genes across clusters.** (*A*) Dot plots showing the log$_2$ fold change of CD compared with NIBD within clusters, the direction of change, and the statistical significance of CD-associated risk genes (*columns*) across clusters (*rows*). For each gene, the *dot* color corresponds to the log$_2$ fold change, the shape of the *dot* corresponds to the direction of change, and the color of the *dot* corresponds to the adjusted *P* value of the comparison. Adjusted *P* value was calculated from the Wilcoxon rank-sum test. DE, differential expression; padj, adjusted p-value.
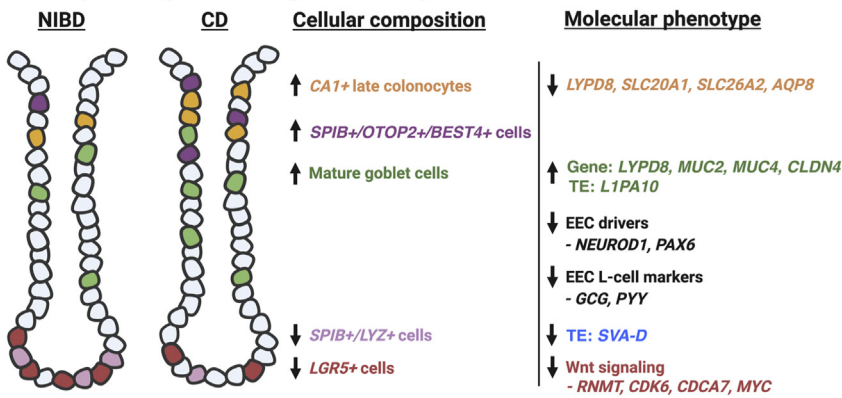
**Table 4.** Average Unique Molecular Identifier Count, Detected Genes, and Percentage of Mitochondrial Genes per Cell Across the 7 Samples Used

| Sample ID | Condition | Average UMI per cell | Average genes per cell | Average mitochondrial reads per cell, % |
|-----------|-----------|----------------------|------------------------|------------------------------------------|
| 206 | NIBD | 21670.1 | 4153.1 | 20.1 |
| 214 | NIBD | 21562 | 4216.2 | 20.1 |
| 216 | NIBD | 23015.5 | 4187.4 | 20.8 |
| 217 | NIBD | 21538.3 | 4104.8 | 18.9 |
| 189 | CD | 19781 | 3895.8 | 20.2 |
| 299 | CD | 17240.9 | 3592.3 | 21.2 |
| 364 | CD | 20114 | 3991.2 | 18.7 |

### Crypt–Axis Score

The crypt–axis score was assigned for each cell and was based on the expression of a previously defined set of genes: *SELENOP*, *CEACAM7*, *PLAC8*, *CEACAM1*, *TSPAN1*, *CEACAM5*, *CEACAM6*, *IFI27*, *DHRS9*, *KRT20*, *RHOC*, *CD177*, *PKIB*, *HPGD*, and *LYPD8*.[3] For each gene, expression within a cell was divided by the maximum expression across all cells to mitigate the weight of highly expressed genes. The maximum expression normalized values places the gene's expression on a 0 to 1 scale. The crypt–axis score is the summation across all genes of the maximum-normalized expression values.

### Intestinal Stem Cell Analysis

Genes constituting the 3 classes of ISCs were obtained from Biton et al.[30] In the stem cluster, the percentage of NIBD or CD cells positive for the genes was calculated and the delta was determined by subtracting the NIBD percentage positive from the CD percentage positive for each gene. The overall shift in an ISC class was determined by averaging the delta across all genes within the ISC class.

### Integrative Analysis With IBD GWAS Results

A list of 5743 CD risk-associated single-nucleotide polymorphisms were identified through GWAS. The closest genes to the disease risk–associated single-nucleotide polymorphisms were identified using bedtools closest (v2.27) (Salt Lake City, UT), resulting in 261 unique genes. Of these genes, 208 were present in the filtered data set. For the figures, the gene order was determined by hierarchical clustering of the Euclidean distance of the $\log_2$ fold change for each gene across clusters.

### RNA Velocity Analysis

The velocity pseudotime figure was generated using scVelo (v0.2.2) (Munich, Germany). The velocities were computed using the dynamic model, and we performed a likelihood-ratio test to test for differential kinetics between clusters and corrected the velocity for differential kinetics. The root index was set as the first indexed stem cell. The velocities were projected onto the Uniform Manifold Approximation and Projection embedding space.

### Generating a Graphic Representation of Cluster Connectivity

The graph representation was generated using partition-based graph abstraction, which was implemented in Scanpy (v1.6) (Munich, Germany). Pruning was performed by setting a minimum edge weight of 0.3. In addition, the tree was rooted at the stem cluster.

### Histologic Analysis

Human proximal colonic tissue was fixed in 4% (vol/vol) neutral-buffered paraformaldehyde, embedded in paraffin, and cut into 10-$\mu$m sections. Immunofluorescent staining of LYZ was performed to visualize Paneth cells. Briefly, sections were incubated with primary antibody (rabbit anti-LYZ, 1:1000 dilution in phosphate-buffered saline [PBS] with 1% [wt/vol] bovine serum albumin) (cat. PA5-16668; Invitrogen, Carlsbad, CA) overnight at 4°C followed by goat anti-rabbit Alexa Fluor 594 secondary antibody (1:1000 in PBS with 1% [wt/vol] bovine serum albumin) (cat. A1102; Invitrogen) incubation for 1 hour at room temperature. Subsequent section incubation with 4′,6-diamidino-2-phenylindole (1:1000 in PBS) (cat. D1306; Invitrogen) for 30 minutes at room temperature was used to visualize nuclei. Images were captured using a BX53 Olympus scope (Olympus, Center Valley, PA). LYZ+ cells were enumerated in longitudinally well-oriented colonic crypts.

### Statistics

The significance of highly enriched markers of clusters and differential expressed genes within clusters was determined by the Wilcoxon rank-sum test and multiple test correction was accomplished using the Benjamini–Hochberg procedure. The significance of the distribution shift along the crypt axis was determined using the Kolmogorov–Smirnov test. The significance of all other comparisons between NIBD and CD was determined by a Student $t$ test.

## References

1. Elmentaite R, Ross ADB, Roberts K, James KR, Ortmann D, Gomes T, Nayak K, Tuck L, Pritchard S,

Bayraktar OA, Heuschkel R, Vallier L, Teichmann SA, Zilbauer M. Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. Dev Cell 2020;55:771–783 e5.

2. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J, Sud M, Andrews E, Velonias G, Haber AL, Jagadeesh K, Vickovic S, Yao J, Stevens C, Dionne D, Nguyen LT, Villani AC, Hofree M, Creasey EA, Huang H, Rozenblatt-Rosen O, Garber JJ, Khalili H, Desch AN, Daly MJ, Ananthakrishnan AN, Shalek AK, Xavier RJ, Regev A. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell 2019;178:714–730 e22.

3. Parikh K, Antanaviciute A, Fawkner-Corbett D, Jagielowicz M, Aulicino A, Lagerholm C, Davis S, Kinchen J, Chen HH, Alham NK, Ashley N, Johnson E, Hublitz P, Bao L, Lukomska J, Andev RS, Björklund E, Kessler BM, Fischer R, Goldin R, Koohy H, Simmons A. Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature 2019;567:49–55.

4. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018;37:38–44.

5. Yang XZ, Cheng TT, He QJ, Lei ZY, Chi J, Tang Z, Liao QX, Zhang H, Zeng LS, Cui SZ. LINC01133 as ceRNA inhibits gastric cancer progression by sponging miR-106a-3p to regulate APC expression and the Wnt/beta-catenin pathway. Mol Cancer 2018; 17:126.

6. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lonnerberg P, Furlan A, Fan J, Borm LE, Liu Z, van Bruggen D, Guo J, He X, Barker R, Sundstrom E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko PV. RNA velocity of single cells. Nature 2018;560:494–498.

7. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature 2011; 474:307–317.

8. Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, Stade B, Bromberg Y, Ellinghaus E, Keller A, Rivas MA, Skieceviciene J, Doncheva NT, Liu X, Liu Q, Jiang F, Forster M, Mayr G, Albrecht M, Hasler R, Boehm BO, Goodall J, Berzuini CR, Lee J, Andersen V, Vogel U, Kupcinskas L, Kayser M, Krawczak M, Nikolaus S, Weersma RK, Ponsioen CY, Sans M, Wijmenga C, Strachan DP, McArdle WL, Vermeire S, Rutgeerts P, Sanderson JD, Mathew CG, Vatn MH, Wang J, Nothen MM, Duerr RH, Buning C, Brand S, Glas J, Winkelmann J, Illig T, Latiano A, Annese V, Halfvarson J, D'Amato M, Daly MJ, Nothnagel M, Karlsen TH, Subramani S, Rosenstiel P, Schreiber S, Parkes M, Franke A. Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. Gastroenterology 2013;145:339–347.

9. Masyuk AI, Marinelli RA, LaRusso NF. Water transport by epithelia of the digestive tract. Gastroenterology 2002; 122:545–562.

10. Ricanek P, Lunde LK, Frye SA, Stoen M, Nygard S, Morth JP, Rydning A, Vatn MH, Amiry-Moghaddam M, Tonjum T. Reduced expression of aquaporins in human intestinal mucosa in early stage inflammatory bowel disease. Clin Exp Gastroenterol 2015;8:49–67.

11. Escudero-Hernandez C, Munch A, Ostvik AE, Granlund AVB, Koch S. The water channel aquaporin 8 is a critical regulator of intestinal fluid homeostasis in collagenous colitis. J Crohns Colitis 2020;14:962–973.

12. Okumura R, Kurakawa T, Nakano T, Kayama H, Kinoshita M, Motooka D, Gotoh K, Kimura T, Kamiyama N, Kusu T, Ueda Y, Wu H, Iijima H, Barman S, Osawa H, Matsuno H, Nishimura J, Ohba Y, Nakamura S, Iida T, Yamamoto M, Umemoto E, Sano K, Takeda K. Lypd8 promotes the segregation of flagellated microbiota and colonic epithelia. Nature 2016; 532:117–121.

13. Hsu CC, Okumura R, Takeda K. Human LYPD8 protein inhibits motility of flagellated bacteria. Inflamm Regen 2017;37:23.

14. Okumura R, Kodama T, Hsu CC, Sahlgren BH, Hamano S, Kurakawa T, Iida T, Takeda K. Lypd8 inhibits attachment of pathogenic bacteria to colonic epithelia. Mucosal Immunol 2020;13:75–85.

15. Nagatake T, Fujita H, Minato N, Hamazaki Y. Enteroendocrine cells are specifically marked by cell surface expression of claudin-4 in mouse small intestine. PLoS One 2014;9:e90638.

16. Watari A, Kodaka M, Matsuhisa K, Sakamoto Y, Hisaie K, Kawashita N, Takagi T, Yamagishi Y, Suzuki H, Tsujino H, Yagi K, Kondoh M. Identification of claudin-4 binder that attenuates tight junction barrier function by TR-FRET-based screening assay. Sci Rep 2017;7:14514.

17. Billing LJ, Smith CA, Larraufie P, Goldspink DA, Galvin S, Kay RG, Howe JD, Walker R, Pruna M, Glass L, Pais R, Gribble FM, Reimann F. Co-storage and release of insulin-like peptide-5, glucagon-like peptide-1 and peptideYY from murine and human colonic enteroendocrine cells. Mol Metab 2018;16:65–75.

18. Barrera JG, Sandoval DA, D'Alessio DA, Seeley RJ. GLP-1 and energy balance: an integrated model of short-term and long-term control. Nat Rev Endocrinol 2011;7:507–616.

19. Castagliuolo I, Wang CC, Valenick L, Pasha A, Nikulasson S, Carraway RE, Pothoulakis C. Neurotensin is a proinflammatory neuropeptide in colonic inflammation. J Clin Invest 1999;103:843–849.

20. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C. Ten things you should know about transposable elements. Genome Biol 2018;19:199.

21. Burns KH. Our conflict with transposable elements and its implications for human disease. Annu Rev Pathol 2020;15:51–70.

22. Kazazian HH Jr, Moran JV. Mobile DNA in health and disease. N Engl J Med 2017;377:361–370.

23. Gorbunova V, Seluanov A, Mita P, McKerrow W, Fenyö D, Boeke JD, Linker SB, Gage FH, Kreiling JA, Petrashen AP, Woodham TA, Taylor JR, Helfand SL, Sedivy JM. The role of retrotransposable elements in ageing and age-associated diseases. Nature 2021; 596:43–53.

24. Jönsson ME, Garza R, Johansson PA, Jakobsson J. Transposable elements: a common feature of neuro-developmental and neurodegenerative disorders. Trends Genet 2020;36:610–623.

25. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 2016;351:1083–1087.

26. Schewe M, Franken PF, Sacchetti A, Schmitt M, Joosten R, Bottcher R, van Royen ME, Jeammet L, Payre C, Scott PM, Webb NR, Gelb M, Cormier RT, Lambeau G, Fodde R. Secreted phospholipases A2 are intestinal stem cell niche factors with distinct roles in homeostasis, inflammation, and cancer. Cell Stem Cell 2016;19:38–51.

27. Cheung P, Xiol J, Dill MT, Yuan WC, Panero R, Roper J, Osorio FG, Maglic D, Li Q, Gurung B, Calogero RA, Yilmaz OH, Mao J, Camargo FD. Regenerative reprogramming of the intestinal stem cell state via Hippo signaling suppresses metastatic colorectal cancer. Cell Stem Cell 2020;27:590–604 e9.

28. Barker N, van Es JH, Kuipers J, Kujala P, van den Born M, Cozijnsen M, Haegebarth A, Korving J, Begthel H, Peters PJ, Clevers H. Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 2007;449:1003–1007.

29. Munoz J, Stange DE, Schepers AG, van de Wetering M, Koo BK, Itzkovitz S, Volckmann R, Kung KS, Koster J, Radulescu S, Myant K, Versteeg R, Sansom OJ, van Es JH, Barker N, van Oudenaarden A, Mohammed S, Heck AJ, Clevers H. The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. EMBO J 2012;31:3079–3091.

30. Biton M, Haber AL, Rogel N, Burgin G, Beyaz S, Schnell A, Ashenberg O, Su CW, Smillie C, Shekhar K, Chen Z, Wu C, Ordovas-Montanes J, Alvarez D, Herbst RH, Zhang M, Tirosh I, Dionne D, Nguyen LT, Xifaras ME, Shalek AK, von Andrian UH, Graham DB, Rozenblatt-Rosen O, Shi HN, Kuchroo V, Yilmaz OH, Regev A, Xavier RJ. T helper cell cytokines modulate intestinal stem cell renewal and differentiation. Cell 2018; 175:1307–1320 e22.

31. Ito G, Okamoto R, Murano T, Shimizu H, Fujii S, Nakata T, Mizutani T, Yui S, Akiyama-Morio J, Nemoto Y, Okada E, Araki A, Ohtsuka K, Tsuchiya K, Nakamura T, Watanabe M. Lineage-specific expression of bestrophin-2 and bestrophin-4 in human intestinal epithelial cells. PLoS One 2013;8:e79693.

32. Kong J, Sun W, Li C, Wan L, Wang S, Wu Y, Xu E, Zhang H, Lai M. Long non-coding RNA LINC01133 inhibits epithelial-mesenchymal transition and metastasis in colorectal cancer by interacting with SRSF6. Cancer Lett 2016;380:476–484.

33. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature 2017;543:550–554.

34. Bruno M, Mahgoub M, Macfarlan TS. The arms race between KRAB-zinc finger proteins and endogenous retroelements and its impact on mammals. Annu Rev Genet 2019;53:393–416.

35. Cronshagen U, Voland P, Kern HF. cDNA cloning and characterization of a novel 16 kDa protein located in zymogen granules of rat pancreas and goblet cells of the gut. Eur J Cell Biol 1994;65:366–377.

36. Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, Haverty PM, Tong AJ, Blanchette C, Albert ML, Mellman I, Bourgon R, Greally J, Jhunjhunwala S, Chen-Harris H. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. Nat Commun 2019; 10:5228.

37. Posternak V, Ung MH, Cheng C, Cole MD. MYC mediates mRNA cap methylation of canonical Wnt/beta-catenin signaling transcripts by recruiting CDK7 and RNA methyltransferase. Mol Cancer Res 2017; 15:213–224.

38. Weiser M, Simon JM, Kochar B, Tovar A, Israel JW, Robinson A, Gipson GR, Schaner MS, Herfarth HH, Sartor RB, McGovern DPB, Rahbar R, Sadiq TS, Koruda MJ, Furey TS, Sheikh SZ. Molecular classification of Crohn's disease reveals two clinically relevant subtypes. Gut 2018;67:36–42.

39. Keith BP, Barrow JB, Toyonaga T, Kazgan N, O'Connor MH, Shah ND, Schaner MS, Wolber EA, Trad OK, Gipson GR, Pitman WA, Kanke M, Saxena SJ, Chaumont N, Sadiq TS, Koruda MJ, Cotney PA, Allbritton N, Trembath DG, Sylvester F, Furey TS, Sethupathy P, Sheikh SZ. Colonic epithelial miR-31 associates with the development of Crohn's phenotypes. JCI Insight 2018;3:e122788.

40. Wang Y, DiSalvo M, Gunasekara DB, Dutton J, Proctor A, Lebhar MS, Williamson IA, Speer J, Howard RL, Smiddy NM, Bultman SJ, Sims CE, Magness ST, Allbritton NL. Self-renewing monolayer of primary colonic or rectal epithelial cells. Cell Mol Gastroenterol Hepatol 2017;4:165–182 e7.

41. Toyonaga T, Steinbach EC, Keith BP, Barrow JB, Schaner MR, Wolber EA, Beasley C, Huling J, Wang Y, Allbritton NL, Chaumont N, Sadiq TS, Koruda MJ, Jain A, Long MD, Barnes EL, Herfarth HH, Isaacs KL, Hansen JJ, Shanahan MT, Rahbar R, Furey TS, Sethupathy P, Sheikh SZ. Decreased colonic activin receptor-like kinase 1 disrupts epithelial barrier integrity in patients with Crohn's disease. Cell Mol Gastroenterol Hepatol 2020; 10:779–796.

42. Camp JG, Frank CL, Lickwar CR, Guturu H, Rube T, Wenger AM, Chen J, Bejerano G, Crawford GE, Rawls JF. Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape. Genome Res 2014; 24:1504–1516.

**Correspondence**
Address correspondence to: Praveen Sethupathy, PhD, Cornell University College of Veterinary Medicine, T7 006D Veterinary Research Tower, Box 17, Ithaca, New York 14853-6401. e-mail: pr46@cornell.edu; Shehzad Z. Sheikh, MD, PhD, Department of Medicine, 7320 Medical Biomolecular Research Building, CB# 7032, Chapel Hill, North Carolina 27599-7032. e-mail: shehzad_sheikh@med.unc.edu; or Terrence S. Furey, PhD, Department of Genetics and Biology, 5022 Genetics Medicine Building, 120 Mason Farm Road, CB#7264, Chapel Hill, North Carolina 27599-7264. e-mail: tsfurey@email.unc.edu; fax: (919) 966-8929.

**Conflicts of interest**
The authors disclose no conflicts.