ARTICLE

https://doi.org/10.1038/s41586-019-0992-y

Colonic epithelial cell diversity in health and inflammatory bowel disease

Kaushal Parikh[1,2,10], Agne Antanaviciute[1,2,3,10], David Fawkner-Corbett[1,2,4,10], Marta Jagielowicz[1,2], Anna Aulicino[1,2], Christoffer Lagerholm[5], Simon Davis[6], James Kinchen[1,2], Hannah H. Chen[1,2], Nasullah Khalid Alham[4], Neil Ashley[7], Errin Johnson[8], Philip Hublitz[7], Leyuan Bao[1,2], Joanna Lukomska[1,2], Rajinder Singh Andev[1,2], Elisabet Björklund[1,2], Benedikt M. Kessler[6], Roman Fischer[6], Robert Goldin[9], Hashem Koohy[3] & Alison Simmons[1,2]*

The colonic epithelium facilitates host–microorganism interactions to control mucosal immunity, coordinate nutrient recycling and form a mucus barrier. Breakdown of the epithelial barrier underpins inflammatory bowel disease (IBD). However, the specific contributions of each epithelial-cell subtype to this process are unknown. Here we profile single colonic epithelial cells from patients with IBD and unaffected controls. We identify previously unknown cellular subtypes, including gradients of progenitor cells, colonocytes and goblet cells within intestinal crypts. At the top of the crypts, we find a previously unknown absorptive cell, expressing the proton channel OTOP2 and the satiety peptide uroguanylin, that senses pH and is dysregulated in inflammation and cancer. In IBD, we observe a positional remodelling of goblet cells that coincides with downregulation of WFDC2—an antiprotease molecule that we find to be expressed by goblet cells and that inhibits bacterial growth. In vivo, WFDC2 preserves the integrity of tight junctions between epithelial cells and prevents invasion by commensal bacteria and mucosal inflammation. We delineate markers and transcriptional states, identify a colonic epithelial cell and uncover fundamental determinants of barrier breakdown in IBD.

Colonic epithelial cells exist in symbiosis with commensal microflora. They coordinate absorptive processes in addition to playing a role in innate and adaptive mucosal immunity[1]. The epithelial barrier comprises specialized cells with diverse functions that emerge from stem cells at the crypt base. Most epithelial cells are absorptive colonocytes, which are interspersed with specialized epithelial lineages, including secretory goblet and enteroendocrine cells (EECs)[1]. Whether other epithelial-cell types exist in the human colon remains unclear.

IBD—which comprises ulcerative colitis and Crohn's disease—results from a breakdown of the symbiotic relationship between the intestinal commensal microflora and the mucosal immune system. Barrier defects characterize both forms of IBD, with goblet cells reportedly being depleted in ulcerative colitis and increased in Crohn's disease[2]. Key examples exist in which disruption of innate epithelial pathways drives colon inflammation (colitis), including autophagy defects[3], endoplasmic-reticulum stress[4], lipid antigen presentation[5] and inflammasome dysfunction[6].

Goblet cells are critical for the maintenance of the colonic barrier, through both the production of mucus and the transportation and presentation of luminal antigens to tolerogenic dendritic cells, particularly the $CD103^+$ type[7]. Luminal secretion of mucins and antimicrobial proteins (AMPs) establishes a physical barrier to microbial contact. This forms inner and outer mucus layers, essential for maintaining homeostasis, with the inner mucus layer being reportedly sterile[8]. In the small intestine, secretion of AMPs by Paneth cells mediates this sterility[9]. However, the colon contains few or no Paneth cells, so the cell types that direct the release of colonic AMPs remain uncharacterized.

It is also unclear whether specific subsets of colonic epithelial cells show intrinsic molecular pathology in IBD. To study this, we used single-cell profiling to create a map of colonic epithelia in health and of clinically inflamed and noninflamed mucosa in ulcerative colitis. These maps identify an absorptive cell with a role in pH balance, as well as goblet-cell drivers of barrier breakdown.

Crypt gradients of absorptive and secretory cells

We isolated colonic biopsies from healthy volunteers or immunomodulator-naive patients with ulcerative colitis, sampled from clinically inflamed and noninflamed mucosa (Supplementary Table 1). Crypts were dissociated to single-cell suspensions and processed using droplet-sequencing technology, capturing 11,175 cells (Extended Data Fig. 1a, b and Methods).

In healthy colons, we identified and visualized ten clusters of cells using *t*-distributed stochastic neighbourhood embedding (*t*-SNE) (Fig. 1a); these clusters included undifferentiated cells, absorptive colonocytes and distinct clusters of goblet cells and EECs (Fig. 1b). EEC populations further divided into L-cells, enterochromaffin cells and precursor-like cells, identifying novel markers of colonic EECs (Extended Data Fig. 1c–e). Isolating undifferentiated cells in silico, we further identified five subclusters: stem cells[10], early transit-amplifying cells, transit-amplifying-like-cells defined by high expression of cell-cycle-related genes, and secretory and absorptive lineage precursor cells (Extended Data Fig. 1f, g). No Paneth cells were detected in the healthy colons (Extended Data Fig. 1b).

Rather than discrete clusters of cells, we observed gene-expression gradients between single epithelial cells, consistent with an axis of

[1]Medical Research Council (MRC) Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine (WIMM), John Radcliffe Hospital, University of Oxford , Oxford, UK. [2]Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK. [3]MRC WIMM Centre For Computational Biology, MRC Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK. [4]Nuffield Department of Surgical Sciences and Oxford National Institute for Health Research (NIHR) Biomedical Research Centre (BRC), John Radcliffe Hospital, University of Oxford, Oxford, UK. [5]Wolfson Imaging Centre Oxford, MRC Weatherall Institute of Molecular Medicine, Oxford, UK. [6]Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [7]MRC Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK. [8]Sir William Dunn School of Pathology, University of Oxford, Oxford, UK. [9]Centre for Pathology, St Mary's Hospital, Imperial College, London, UK. [10]These authors contributed equally: Kaushal Parikh, Agne Antanaviciute, David Fawkner-Corbett.
*e-mail: alison.simmons@imm.ox.ac.uk

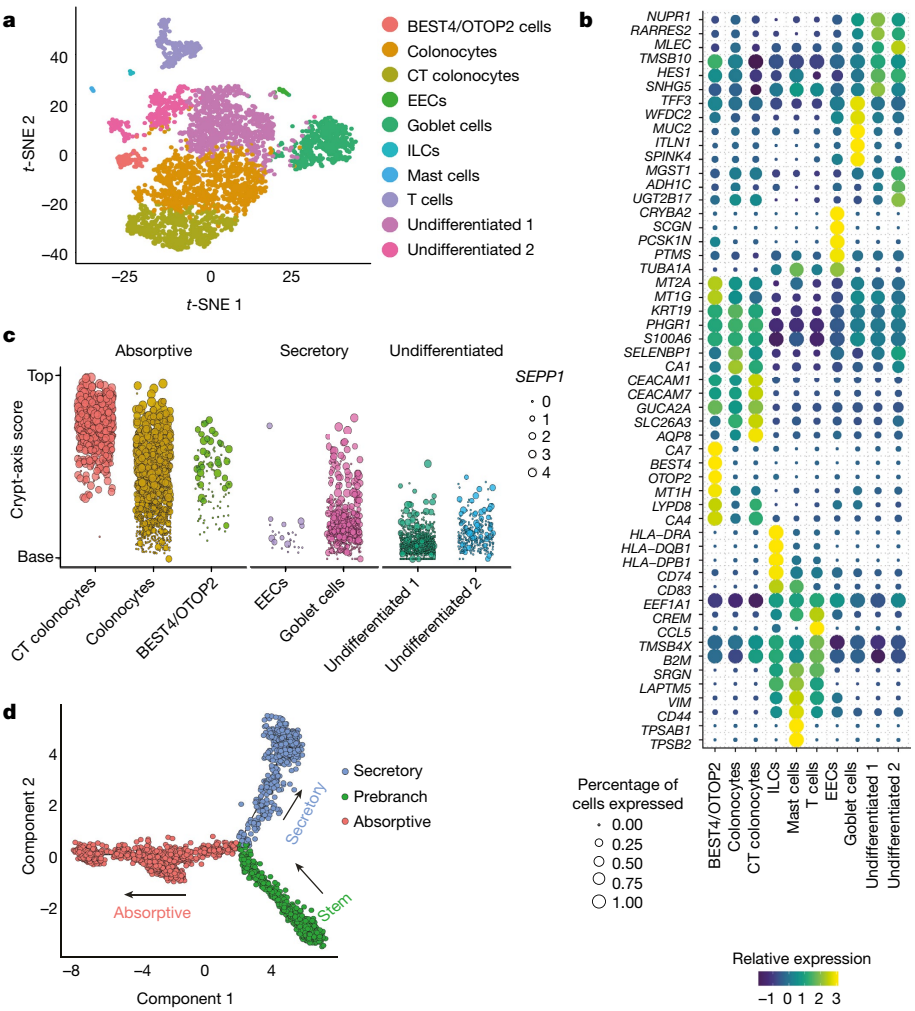


Fig. 1 | Heterogeneity of human colonic epithelial cells in healthy individuals. a, *t*-SNE plot of the healthy human colonic epithelium ($n = 3$; CT, crypt top; ILCs, innate lymphoid cells). **b**, Heat map showing cluster markers, coloured by relative gene expression. The relative size of each dot represents the fraction of cells per cluster that expresses each marker. **c**, Spatial segregation of cell clusters along an axis from the base to the top of the crypt. The *y*-axis represents the axis score, generated from the expression of 15 crypt-axis markers. The sizes of the dots represent the level of expression of *SEPP1* (also known as *SELENOP*), encoding selenoprotein P, a known crypt-axis marker. **d**, Differentiation pseudotime trajectory analysis. Predicted secretory-lineage cells are in blue, absorptive cells in red and uncommitted cells in green ($n = 3$ biological replicates).

differentiation that ascends through the crypt. We in silico-localized cells within the colonic crypt by defining a crypt-axis score (Methods), using 15 genes that are expressed in both absorptive and secretory cells (Fig. 1c). Pseudotime analysis (Methods) confirmed a bifurcating trajectory, arising from stem cells, that separated secretory and absorptive lineages (Fig. 1d), consistent with previously identified clusters and the crypt axis (Extended Data Fig. 1h). Trajectory analysis identified known and putative new factors that could have roles in lineage commitment during differentiation (Extended Data Fig. 1i). Therefore, single-cell RNA sequencing (scRNA-seq) highlights the extent of human colonic heterogeneity and supports the existence of a differentiation hierarchy in the crypt.

Discovery of a novel pH–sensing absorptive colonocyte

Clustering analysis identified a cell cluster (Fig. 1a) predicted to transport salt, ions and metals (Extended Data Fig. 1j). The cells of this cluster expressed markers of mature colonocytes, with distinct expression of the calcium-sensitive chloride channel BEST4 (Fig. 2a, i, iii and Extended Data Fig. 2c, i), the protease cathepsin E (Extended Data Fig. 2c, ii, iii) and the *OTOP2* gene (Fig. 2a, ii, iii), so we designated these cells 'BEST4/OTOP2 cells'. BEST4 marks epithelia that are involved in electrolyte transportation[11], and the OTOP gene family encodes proton-conducting ion channels in various epithelia[12]. The cells of this cluster also expressed the endogenous paracrine hormone and satiety peptide uroguanylin[13] (encoded by *GUCA2B*; Fig. 2a, iv, v, vi), which is required for activation of guanylate cyclase 2C(GC-C) and epithelial cyclic GMP activity. Further, these cells expressed genes belonging to the metallothionein family, which impart defence against free radicals and contribute to metal transport and short-term storage[14] (Fig. 1b).

Trajectory analysis indicated that these cells originate from the absorptive lineage and express the transcription factors *SPIB* and *HES4* (Extended Data Fig. 2c, iv, v), with the latter normally confined to undifferentiated epithelial populations (Extended Data Fig. 2b). We also identified these cells using semisupervised clustering in a human fetal colon dataset[15] (Extended Data Fig. 2d), and found evidence for their loss in inflammation and colorectal cancer[16–21] (Extended Data Figs. 2e, f, and see below).

We next isolated BEST4/OTOP2 cells (Extended Data Fig. 3a, b), and further characterized them using quantitative proteomics (Fig. 2b) and deep scRNA-seq (Smart-Seq2) (Fig. 2c). This enabled the identification of additional messenger RNA and protein markers (Extended Data Fig. 3c and Supplementary Information). Gene Ontology biological-process enrichment analysis of proteomic data highlighted processes that involve ethanol, small molecule and lipid catabolism; icosonoid and fatty acid metabolism; and neutrophil-mediated immunity (Extended Data Fig. 3d).

Functionally, these cells conducted protons into the cell cytosol in response to lowering of the extracellular pH (pH_o); this proton

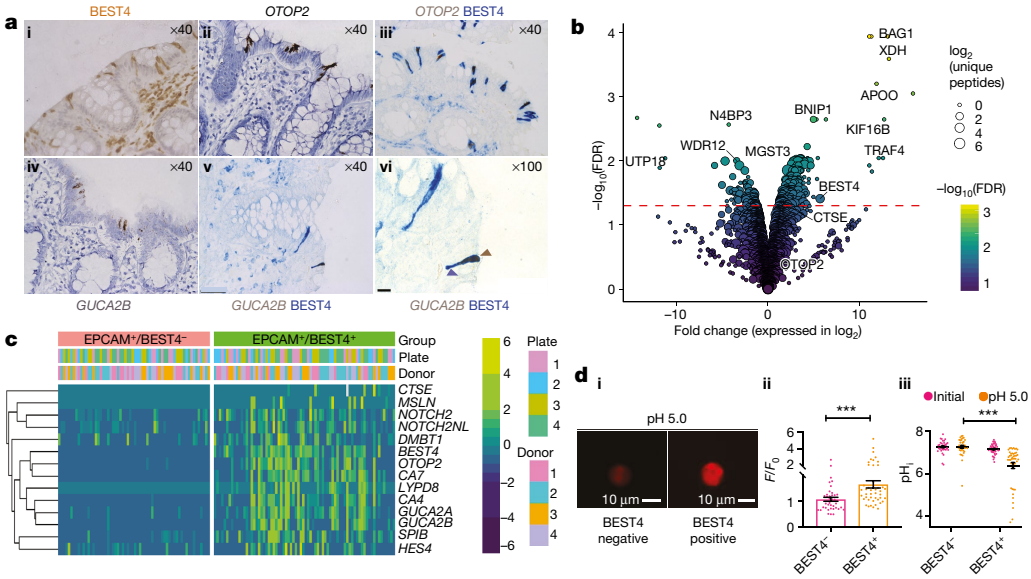


Fig. 2 | scRNA-seq identifies a colonic absorptive cell type. a, **i**–**vi**, Representative images ($n = 3$) of colonic sections stained for BEST4 protein by immunohistochemistry (**i**); for *OTOP2* (**ii**) and *GUCA2B* (**iv**) expression by smISH; and with co-staining and co-localization of BEST4 and *OTOP2* (**iii**) or BEST4 and *GUCA2B* (**v**, **vi**) (the stain colour, brown or blue, relates to the text and magnification shown). **b**, Volcano plot showing proteins that are differentially expressed between BEST4$^+$EPCAM$^+$ cells ($n = 3$; positive fold change (expressed logarithmically)) and BEST4$^-$EPCAM$^+$ cells ($n = 2$; negative fold change (expressed logarithmically)). The red line indicates a 5% FDR (limma linear model empirical Bayes *P* value and Benjamini–Hochberg multiple testing correction). Selected proteins are highlighted. **c**, Heat maps showing selected differentially expressed genes between BEST4$^+$EPCAM$^+$ cells and BEST4$^-$EPCAM$^+$ cells, detected using single-cell Smart-Seq2 ($n = 4$ per group). *NOTCH2NL* is also known as *NOTCH2NLA*. **d**, BEST4$^+$EPCAM$^+$ cells mediate proton influx. **i**, Representative image showing the change in intracellular pH when BEST4$^+$EPCAM$^+$ and BEST4$^-$EPCAM$^+$ cells are exposed to an extracellular buffer of pH 5. **ii**, Normalized fluorescence emission (F/F_0) from pH indicator pHrodo Red in BEST4$^-$EPCAM$^+$ cells ($n = 39$; pink) and BEST4$^+$EPCAM$^+$ ($n = 45$; orange) cells. Responses to pH 5 solution are shown (mean ± s.e.m). **iii**, Intracellular pH_i for the same populations as measured with pHrodo Red AM (see Methods), showing the peak response during each stimulus and the initial starting pH_i (two-sided paired *t*-test, BEST4$^-$EPCAM$^+$ *P* value 0.9873310; BEST$^+$EPCAM$^+$ *P* value 0.0000007768; mean ± s.e.m).

conduction is seen here as an increase in emission of a membrane-permeant pH indicator dye, pHrodo Red (Fig. 2d, i, ii), which corresponded to a substantial change in the intracellular pH (pH_i) (Fig. 2d, iii). Intracellular acidification can be cytotoxic; however, our proteomics data indicate that BEST4/OTOP2 cells express high levels of the anti-apoptotic protein BAG1 (Fig. 2b), which may enable survival following substantial pH changes. Expression of uroguanylin coincident with the ability to sense pH suggests that these cells have a role in setting colonic epithelial cGMP tone in response to luminal pH.

Gene responses in ulcerative colitis

We next sampled clinically inflamed and noninflamed tissue from early-diagnosis, immunotherapy-naive patients with ulcerative colitis (see Methods and Supplementary Table 1). In inflamed tissue, in addition to the clusters that we identified in healthy colons, we detected two clusters that represent inflammation-associated goblet cells and intraepithelial immune cells (Fig. 3a).

Differential gene expression analysis between corresponding cell clusters revealed 1,147 genes (with a false discovery rate (FDR) of less than 1%) that are dysregulated in inflamed ulcerative colitis, with the greatest number of changes being seen in colonocytes (734) and crypt-top colonocytes (676), followed by goblet cells (140), stem cells (65), BEST4/OTOP2 cells (28) and EECs (4) (Fig. 3b and Supplementary Information). We observed universal upregulation of several inflammatory pathways across most cell populations, including interferon-γ signalling, antigen presentation and cytokine production (Extended Data Fig. 4a).

Single-cell profiling enabled us to dissect cell-type-specific responses to colitis. Colonocyte populations downregulated metabolic processes and simultaneously induced genes that are needed to produce reactive oxygen species and for microbial killing (for example, the *SAA1*, *DMBT1* and *PLA2G2A* genes). The BEST4/OTOP2 cell population showed reduced expression of the metallothionein family and other ion-absorption genes (Fig. 3b and Extended Data Fig. 4b). Goblet cells upregulated stress-response genes that actively promoted cell survival in preference to apoptosis (Supplementary Information). *LYZ*, a Paneth-cell gene, was upregulated by lower-crypt goblet cells in inflammation (Fig. 3b and Extended Data Fig. 4e), and may mark the 'deep crypt secretory cells' of the colon[22] that are required to maintain the colonic stem cell niche and to protect stem cells from bacterial damage during colitis. Absorptive and secretory progenitor cells upregulated differentiation and cell-migration pathways, which suggests an active attempt to repair colitis-induced damage. By contrast, stem cells in inflammatory conditions showed downregulated expression of heparin-binding epidermal growth factor (EGF)-like growth factor (HB-EGF; Extended Data Fig. 4d). HB-EGF protects the intestine from injury by preserving Wnt/β-catenin signalling in intestinal stem cells after injury[23]. Failure to upregulate HB-EGF expression in ulcerative colitis may have an effect on Wnt signalling and negatively affect intestinal regeneration. Thus, overall, our data suggest that the outcome of this inflammatory event depends on how these individual cell subtypes balance the dual requirement to restore health and tissue integrity and simultaneously respond to aberrant tissue homeostasis.

IBD susceptibility genes in single colonocytes

As genetic analysis has implicated multiple pathways in IBD pathogenicity, we investigated whether specific genetic-risk genes might operate within distinct epithelial subtypes. Our analysis suggests that IBD susceptibility genes are expressed differently in unique cell populations (Extended Data Fig. 5a).

We used the SNPsea[24] algorithm to test ulcerative-colitis-associated genomic loci[25,26] for enrichment of expression specificity in our single-cell clusters, as well as additional scRNA-seq data from colonic mesenchymal cell populations[27]. We identified intra-epithelial T cells as the most IBD-associated cell type in healthy tissue. This association was driven by high and specific expression of genes such as *IL7R* and *TNFRSF9* (refs. [25,28]; Extended Data Fig. 5b).

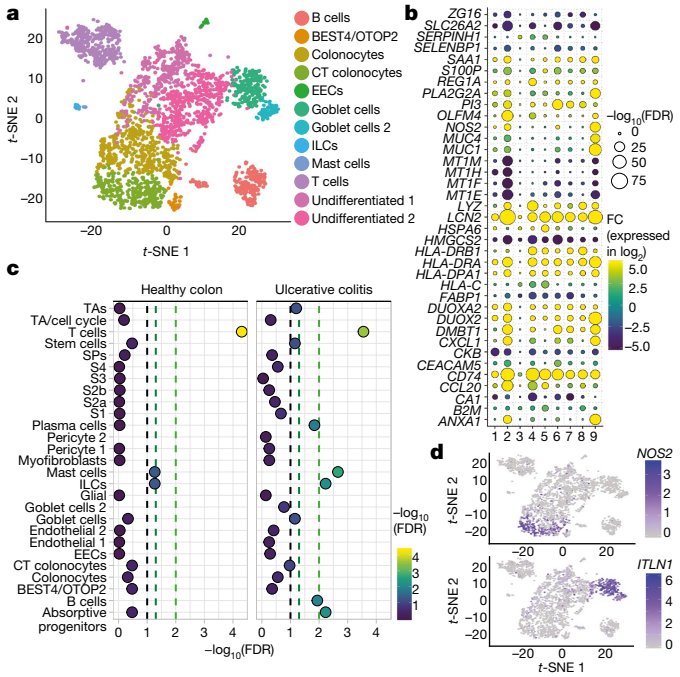


Fig. 3 | Human colonic epithelium in active colitis. a, *t*-SNE plot of single-cell clusters in active ulcerative colitis ($n = 3$). **b**, Heat map showing key differentially expressed genes (FDR less than 1%; two-sided negative binomial likelihood ratio test; Benjamini–Hochberg multiple testing correction) between cell clusters in health and in active ulcerative colitis ($n = 3$). The colour indicates a fold change (expressed in $\log_2$) (dark purple, downregulation; yellow, upregulation). The point size shows the confidence interval for the observation ($-\log_{10}$(FDR)). Point colour indicates fold change (FC (expressed in $\log_2$). 1, BEST4/OTOP2 cells; 2, colonocytes; 3, enteroendocrine cells; 4, goblet cells; 5, stem cells; 6, absorptive progenitors; 7, secretory progenitors (SPs); 8, transit-amplifying cells (TAs); 9, crypt-top colonocytes. **c**, Significance level ($-\log_{10}$(FDR)) of tissue-specific-expression enrichment of ulcerative-colitis-associated GWAS loci in single-cell clusters in colonic epithelium ($n = 3$) and colonic mesenchyme[27] ($n = 2$) in healthy colons and active ulcerative colitis. Dashed lines indicate thresholds for (from left to right) 10%, 5% and 1% FDR cut-offs (SNPsea empirical distribution *P* value; Benjamini–Hochberg multiple testing correction). S1–S4, stromal cells 1–4. **d**, *t*-SNE overlay of selected GWAS-identified, ulcerative-colitis-associated genes expressed specifically in crypt-top colonocytes (*NOS2*) and goblet cells (*ITLN1*) ($n = 3$).

By contrast, using the same approach with inflamed ulcerative colitis samples, we observed a highly significant association (with an FDR of less than 1%) for some immune subsets and absorptive progenitor epithelial cell types (Fig. 3c), with significant associations at alternative cut-offs (FDRs of less than 5% or of 10%) in other immune and epithelial subsets (Fig. 3c). Inflamed crypt-top colonocytes differentially expressed the oxidative-stress-pathway genes *NOS2* (Fig. 3d) and *DDAH2* (ref. [29]), and showed increased expression of *SMAD3* (Extended Data Fig. 5b) and *JAK2*, which are associated with both Crohn's disease and ulcerative colitis[25,30]. IBD-associated *ITLN1* (Fig. 3d) and *IL1R2* (ref. [30]) were expressed by goblet cells, while undifferentiated epithelia expressed IBD-associated *RNF186*, the chemokines *CXCL1*, *CXCL2* and *CXCL3* (ref. [31]), the integrin *ITGB8* (ref. [25]) and the heat-shock-protein *HSPA6* (ref. [28]) (Extended Data Fig. 5b). These results suggest that the effects of diverse, small genetic defects may manifest in different cell types and contribute to the failure to re-establish epithelial barrier function in IBD.

Clinically uninvolved epithelia in ulcerative colitis

Comparing epithelial clusters between healthy, clinically involved and noninvolved ulcerative colitis revealed shifts in relative cell proportions (Extended Data Fig. 5d). Differential expression analysis (Extended Data Figs. 5e, 6a, b) of uninvolved ulcerative colitis versus healthy mucosa identified 207 significantly dysregulated genes (FDR less than 1%). Notably, 59.4% (123 out of 207) of differentially expressed genes in uninvolved ulcerative colitis epithelia were also detected as differentially expressed in involved tissue (Extended Data Fig. 6c); however, gene-expression changes limited to only noninflamed tissue in ulcerative colitis were also observed (Extended Data Figs. 4c, 6a, iii, iv). Furthermore, we fit a generalized linear model to all of the data and found that model coefficients for inflamed and noninflamed samples were correlated, but with smaller effect sizes in noninflamed cells (Extended Data Fig. 6d). Thus epithelia derived from proximal clinically noninflamed colon bear similar—albeit lower-amplitude—transcriptomic hallmarks to those derived from inflamed tissue. This subclinical pathology may arise as result of a dominance of regenerative over damage cues, or as a protective mechanism in anticipation of damage.

Goblet-cell diversity in health and ulcerative colitis

Although dysregulated goblet-cell function contributes to barrier breakdown in colitis, we do not yet know the pathways that underlie this breakdown. Single-cell profiles derived from goblet cells of healthy, inflamed or noninflamed ulcerative colitis tissue in isolation revealed partitioning of this cell group across five clusters (Fig. 4a). We used a score denoting cellular position on the crypt-top to crypt-base axis and unsupervised pseudotime trajectory analysis to infer the localization and maturity of the goblet clusters (Extended Data Fig. 7b, i, ii). For instance, cluster 3 expressed secretory progenitor markers localized to the lower crypt, and cluster 5 localized towards the lumen-facing crypt top (Extended Data Fig. 7a, b, i).

In ulcerative colitis, we observed both spatial and crypt-wide differences at the mRNA and protein levels. This suggested that, although common inflammatory responses exist, goblet cells in spatially distinct regions within the crypt also exhibit highly heterogeneous changes. For instance, the *LCN2* and *REG1A* genes are induced throughout the crypt, whereas *CD74* and *LAMB3* expression was limited to the crypt bottom and top, respectively (Extended Data Fig. 7c). We also observed transcriptional dysregulation, in which genes that are normally limited to the crypt bottom in healthy colons persist in crypt-top cells in inflammation (for example, *SPINK4* and *SPINK1*; Extended Data Fig. 7c, iv, v, d, iii, iv).

Consistent with these observations, we identified the emergence of a disease-associated cluster of goblet cells in cluster 4 (Extended Data Fig. 7e), a counterpart with homology to the crypt-top cluster 5. Ulcerative-colitis-associated goblet cells expressed genes essential for the integrity and homeostasis of the epithelial barrier[32] (Extended Data Fig. 7a, ii, g and Supplementary Information).

We validated these goblet-cell expression patterns by immunofluorescence. Figure 4b shows expression of the BCAS1 (cluster 5), CLCA1 (cluster 1), REGIV (cluster 1) and WFDC2 (cluster 2) proteins together with mucin 2 (MUC2) within goblet cells. CLCA1 and WFDC2 are expressed along a gradient that is higher at the bottom of the crypt (Fig. 4b, ii, iii) and is consistent with our in-silico maturity/crypt gradient predictions (Extended Data Fig. 7b). By comparison, REGIV is observed mainly in the mid-to-upper portions of the crypt (Fig. 4b, iv). Not all goblet cells expressed these proteins, a fact that is also consistent with segregation across subclusters. Double stains for WFDC2 and CLCA1 (Fig. 4b, v) and WFDC2 and REGIV (Fig. 4b, vi) confirmed the heterogeneity of protein expression in goblet cells suggested by the single-cell profiles.

Goblet-cell WFDC2 loss in active ulcerative colitis

We found that the spatial architecture of goblet cells within the inflamed crypt was perturbed and associated with the dysregulation of numerous genes, including *WFDC2*. *WFDC2* is normally highly expressed by crypt-base goblet cells (Extended Data Fig. 7b) but is downregulated in inflammation (Extended Data Fig. 7f and Supplementary Information). We next investigated whether loss of *WFDC2* expression is a hallmark

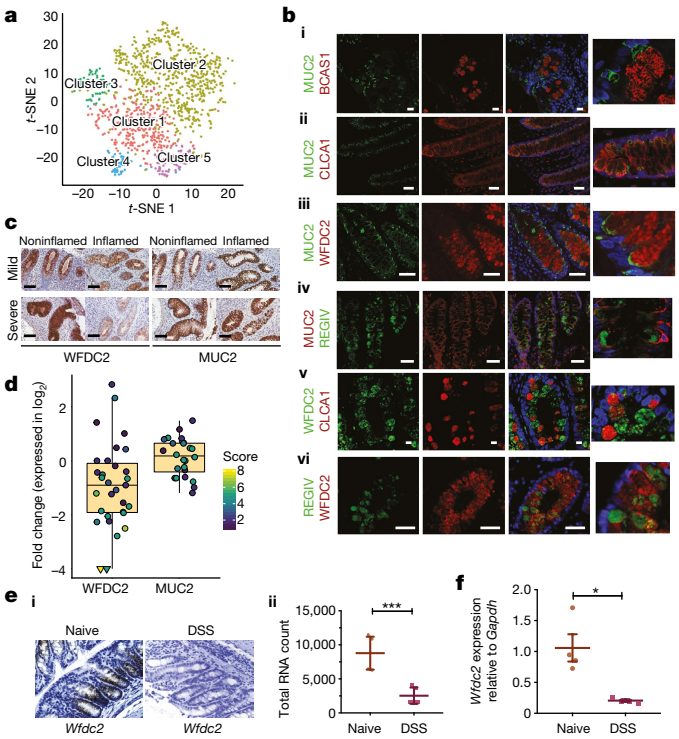


Fig. 4 | Heterogeneity of goblet cells in health and ulcerative colitis. **a**, *t*-SNE plot of subclusters across all captured goblet cells ($n = 3$ per group). **b**, **i**–**iv**, Newly identified and previously known goblet-cell marker proteins validated by immunohistochemistry in healthy human colonic tissue (representative of three patient samples): **i**, BCAS1 (red) with MUC2 (green); **ii**, CLCA1 (red) with MUC2 (green); **iii**, WFDC2 (red) with MUC2 (green); **iv**, REGIV (green) with MUC2 (red). **v**, **vi**, Heterogeneity in expression of these markers is also observed by double staining for WFDC2 (green) with CLCA1 (red; **v**) and for WFDC2 (red) with REGIV (green; **vi**). Scale bars, 20 μm (**i**), 50 μm (**ii**–**iv**, **vi**), 10 μm (**v**). **c**, Representative immunohistochemistry images of colon biopsies from inflamed and noninflamed regions of ulcerative colitis colon stained for WFDC2 ($n = 31$) and MUC2 ($n = 24$). The top panel shows WFDC2 and MUC2 expression in a patient with mild disease, and the bottom panel shows expression in severe inflammation. **d**, Quantification and distribution of change in WFDC2 and MUC2 expression (fold change (expressed in $\log_2$)) in inflamed versus noninflamed tissue from patients with varying disease severity (the 25th, 50th and 75th percentiles are shown). Each point is coloured and sized by severity score (automated quantification; see Methods). Triangles represent outliers (showing a decrease greater than 16-fold). Paired two-sided *t*-test: MUC2 ($n = 24$), $P = 0.2485$; WFDC2 ($n = 31$), $P = 0.0001779$. **e**, **i**, smISH of *Wfdc2* in the colons of naive mice and of animals treated with DSS (a model of acute colitis) ($n = 5$ per group). **ii**, smISH quantification from colons of naive and DSS-treated mice ($n = 5$ per group; $^{***}P = 0.0008$; unpaired two-sided *t*-test; mean ± s.e.m. shown). **f**, Quantification of *Wfdc2* expression relative to *Gapdh* expression by qRT–PCR from the colons of naive and DSS-treated mice ($n = 4$ per group; $^{*}P = 0.0086$; unpaired two-sided *t*-test; mean ± s.e.m. shown).

of colitis in a larger cohort of patients, by analysing clinically non-inflamed and inflamed sections from patients with mild or severe ulcerative colitis by immunohistochemical labelling of WFDC2. As a control for goblet-cell health, we also stained for MUC2. Expression of WFDC2 and MUC2 in patients with varying degrees of mucosal inflammation is shown in Fig. 4c. More-severely inflamed tissue sections showed a clear reduction in goblet-cell WFDC2 expression. Both visual and digital scoring (Fig. 4d and Extended Data Fig. 7h) confirmed substantial protein loss in these cells. Further, we showed a similar loss of *Wfdc2* expression in mouse colonic tissue treated with dextran sodium sulfate (DSS), using RNA in situ hybridization (smISH) (Fig. 4e, i, ii) and quantitative reverse transcription (qRT)–PCR (Fig. 4f).

Loss of WFDC2 expression in colitis cannot be explained by a direct effect of known genetic factors, as it does not segregate with IBD genome-wide association study (GWAS) loci. We hypothesized that the local milieu of inflammatory cytokines may dictate the expression levels of WFDC2 and other proteins, and the degree of residual barrier protection in colitis. Consistent with this hypothesis, we observed the expression of several interferon-induced genes in goblet cells (Extended Data Fig. 7i) with a location-dependent bias, which suggests that at least some of the dysregulation in these cells could be attributed to secreted pro-inflammatory factors. We tested this hypothesis on a human colonic organoid model (Extended Data Fig. 8a, i, ii) that was stimulated with interferon-γ, and observed both distinct organoid morphology and downregulation of *WFDC2* (Extended Data Fig. 8a, iii). Given our data, one possible source of interferon-γ stimulation may be intra-epithelial lymphocytes (Extended Data Fig. 8a, iv).

Barrier integrity requires secreted WFDC2

WFDC2 has been proposed to regulate innate immunity through the inhibition of serine and cysteine proteases[33]. We found WFDC2 to be secreted both basally and apically, and to show increased expression in response to stimulation in HT29-MTX-E12[34] cells (Extended Data Fig. 8b, i, ii). It inhibited the proteolytic activity of the matrix metalloproteinases MMP12 and MMP13, the pathological induction of which in IBD can orchestrate tissue destruction[35] (Extended Data Fig. 8c, i, ii). Furthermore, knockdown of WFDC2 in vitro showed a disturbed cellular morphology with goblet-cell hyperplasia and dysregulated mucus attachment (Extended Data Fig. 8d).

As the inner mucus layer covering the colonic epithelium is sterile[36], we questioned whether WFDC2 secreted into the lumen may perpetuate this sterility via antibacterial activity. We found that recombinant WFDC2 produced a marked dose-dependent reduction in the viability of the Gram-positive bacteria *Staphylococcus aureus* and the Gram-negative bacteria *Escherichia coli* and *Pseudomonas aeruginosa*. However, the viability of other Gram-positive (*Enterococcus faecalis*) and Gram-negative (*Salmonella* Typhimurium LT2) bacteria remained unaffected (Fig. 5a). This selective bactericidal activity of WFDC2 at a concentration comparable to that of other intestinal AMPs[37,38] suggests a potential role in maintaining homeostasis by restricting epithelial–bacterial contact in vivo.

To test this, we explored the function of WFDC2 in vivo using heterozygous ($Wfdc2^{+/-}$) mice, as homozygous deletion of *Wfdc2* was embryonically lethal. smISH (Fig. 5b, i, ii) confirmed reduced levels of *Wfdc2* mRNA in the colons of $Wfdc2^{+/-}$ mice. Transmission electron microscopy (TEM) revealed abnormalities in colonic epithelial intercellular junctions (Fig. 5b, iv), along with irregular distribution of microvilli, in $Wfdc2^{+/-}$ mice (Fig. 5b, vi) compared with wild-type littermates (Fig. 5b, iii, v). $Wfdc2^{+/-}$ mice presented with abnormal histology, with mild-to-modest epithelial hyperplasia (Methods), accompanied by lymphoid infiltration (Fig. 5c and Extended Data Fig. 9a).

We next explored whether the absence of *Wfdc2* facilitates the breakdown of the inner-mucus sterility. Staining for MUC2 in colonic tissues with a preserved mucus layer[39] suggests that the inner mucus layer in heterozygous mice is considerably different (Extended Data Fig. 9b). Gram staining identified colonies from both Gram-positive and Gram-negative bacteria in close proximity to epithelia of $Wfdc2^{+/-}$ mice (Fig. 5d, i, ii). Scanning electron microscopy (SEM) confirmed bacterial attachment and goblet-cell damage in $Wfdc2^{+/-}$ mice (Fig. 5d, iii, iv and Extended Data Fig. 9c). Unlike in wild-type tissues (Fig. 5e, i), TEM analysis of $Wfdc2^{+/-}$ mice showed invading bacteria free in the epithelial cytoplasm within a matrix of vesicles, fibres and membrane fragments (Fig. 5e, ii, iii), along with cellular destruction, epithelial detachment and bacterial aggregates over the epithelial surface (Extended Data Fig. 9d, i–iv). Thus, our data show that WFDC2 is an important, goblet-cell-secreted, antibacterial defence factor that is required to prevent colonization, invasion and epithelial barrier breakdown.

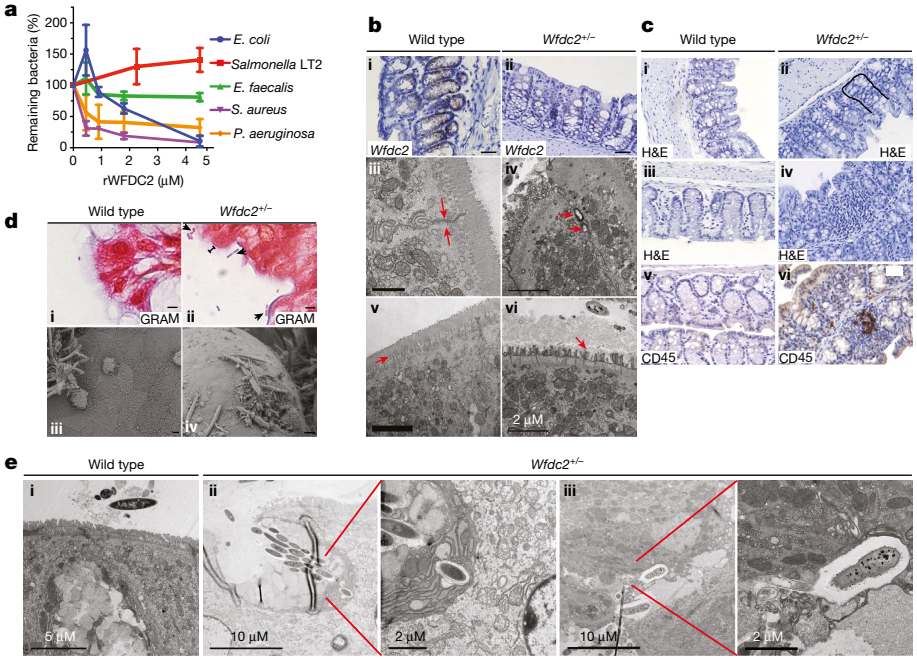


Fig. 5 | WFDC2 shows selective bactericidal activity and influences barrier function. a, Purified recombinant WFDC2 (rWFDC2) was added to mid-logarithmic-phase bacteria for 4 h. Surviving bacteria were quantified by dilution plating. Means ($n = 3$) ± s.d. are plotted. **b**, **i**, **ii**, *Wfdc2* smISH in colons from *Wfdc2*$^{+/-}$ and wild-type littermates. **iii**–**vi**, TEM of colonocytes shows disrupted tight junctions (arrows in **iv**) and scattered microvilli (arrows in **vi**) in *Wfdc2*$^{+/-}$ mice compared with wild-type animals (**iii**, **v**). Scale bars, 100 μm (**i**, **ii**), 1 μm (**iii**), 2 μm (**iv**–**vi**). **c**, **i**–**vi**, Histopathological evaluation of changes in epithelial-cell morphology and mucosal architecture. *Wfdc2*$^{+/-}$ mice (**ii**, **iv**, **vi**) show irregular crypts with variable diameters along the depth of single crypts (**ii**) and focal mucosal infiltration of leukocytes (**iv**, **vi**) compared with wild-type littermates (**i**, **iii**, **v**). Magnification, **i**, ×20; **ii**–**vi**, ×40. H&E, haemotoxylin and eosin. **d**, **i**, **ii**, Gram staining identifies regions that are free from luminal contents above the epithelial layer in wild-type mice (arrow in **i**), but also indicate colonization by both Gram-positive and Gram-negative bacteria (arrows) in *Wfdc2*$^{+/-}$ mice (**ii**). **iii**, **iv**, SEM of the colonic surface shows bacteria invading goblet cells in *Wfdc2*$^{+/-}$ mice (**iv**) compared with wild-type mice (**iii**). Scale bars, 10 μm (**i**, **ii**), 2 μm (**iii**, **iv**). **e**, TEM analysis shows bacteria invading the *Wfdc2*$^{+/-}$ colonic tissue, mostly through goblet cells (**ii**, **iii**). Bacteria were not confined to a membrane-bound compartment but were located free in the cytoplasm. No invasion of epithelial surfaces is observed in the wild-type littermates; the epithelium is intact with preserved colonocytes and goblet cells (**i**). Scale bars, 5 μm (**i**), 10 μm (**ii**, left, **iii**, left), 2 μm (**ii**, right, **iii**, right). In **b**–**e**, $n = 4$ per group.

Discussion

We have presented the first, to our knowledge, large-scale scRNA-seq study of the human colonic epithelium in health and inflammation, which has revealed previously unknown cellular diversity and subtype-specific gene dysregulation in colitis.

We have characterized an absorptive cell type, the BEST4/OTOP2 cells, which are capable of pH sensing and may maintain luminal homeostasis through regulation of the GC-C signalling pathway. These cells selectively express uroguanylin, the endogenous paracrine hormone required for GC-C activation. GC-C receptor signalling occurs in a pH-dependent manner and modulates key physiological functions, including fluid and electrolyte homeostasis, maintenance of epithelial proliferation, barrier function, DNA integrity, epithelial–mesenchymal crosstalk and microbiota composition[40]. Dysregulation of this circuit underlies intestinal transformation. Our data show that these uroguanylin-producing colonic epithelial cells are depleted in IBD and colorectal cancer, which suggests a previously unknown mechanism by which this pathway is dysregulated in these diseases. This provides a further rationale for the use of US Food and Drug Administration (FDA)-approved uroguanylin mimetics, and has wide-ranging implications for future studies.

Furthermore, we have delineated the functional role of a colonic goblet-cell-secreted antibacterial protein, WFDC2, in mucosal barrier homeostasis. We find this protein to be localized towards the bottom half of the crypt in healthy colons and to be dysregulated in ulcerative colitis. Evidence exists for how regional differences in goblet-cell phenotypes may affect key aspects of crypt physiology, such as barrier mucus. Colonic mucus is composed of inner and outer layers, the outer layer being unattached and creating a habitat for microbiota[36]. The antiprotease activity of WFDC2 inhibits the activities of serine and cysteine proteases, preventing the premature conversion of the inner mucus layer to the outer layer in health. Indeed, knockdown of *WFDC2* expression results in abnormalities in mucus-layer formation (Extended Data Figs. 8d, 9b). These mucus defects may allow bacteria to penetrate the mucus layer and to contact epithelial cells (Fig. 5)—hallmarks of ulcerative colitis[41]. Recent mouse studies support the existence of functional subpopulations of goblet cells, with differing rates of mucus production and secretion along the crypt axis[42]. Our work provides a basis for the spatial interrogation of goblet-cell phenotypes, key aspects of crypt physiology and how this specialization breaks down in barrier diseases such as ulcerative colitis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-0992-y.

Received: 7 July 2018; Accepted: 28 January 2019;
Published online 27 February 2019.

1. Peterson, L. W. & Artis, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nat. Rev. Immunol.* **14**, 141–153 (2014)
2. McCauley, H. A. & Guasch, G. Three cheers for the goblet cell: maintaining homeostasis in mucosal epithelia. *Trends Mol. Med.* **21**, 492–503 (2015).
3. Kabat, A. M., Pott, J. & Maloy, K. J. The mucosal immune system and its regulation by autophagy. *Front. Immunol.* **7**, 240 (2016).
4. Hooper, K. M., Barlow, P. G., Henderson, P. & Stevens, C. Interactions between autophagy and the unfolded protein response: implications for inflammatory bowel disease. *Inflamm. Bowel Dis.* https://doi.org/10.1093/ibd/izy380 (2018).
5. Iyer, S. S. et al. Dietary and microbial oxazoles induce intestinal inflammation by modulating aryl hydrocarbon receptor responses. *Cell* **173**, 1123–1134 (2018).
6. Rathinam, V. A. K. & Chan, F. K.-M. Inflammasome, inflammation, and tissue homeostasis. *Trends Mol. Med.* **24**, 304–318 (2018).

7. McDole, J. R. et al. Goblet cells deliver luminal antigen to CD103$^+$ dendritic cells in the small intestine. *Nature* **483**, 345–349 (2012).
8. Johansson, M. E. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nat. Rev. Immunol.* **16**, 639–649 (2016).
9. Ayabe, T. et al. Secretion of microbicidal α-defensins by intestinal Paneth cells in response to bacteria. *Nat. Immunol.* **1**, 113–118 (2000).
10. Barker, N. et al. Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449**, 1003–1007 (2007).
11. Ito, G. et al. Lineage-specific expression of bestrophin-2 and bestrophin-4 in human intestinal epithelial cells. *PLoS ONE* **8**, e79693 (2013).
12. Tu, Y. H. et al. An evolutionarily conserved gene family encodes proton-selective ion channels. *Science* **359**, 1047–1050 (2018).
13. Ikpa, P. T. et al. Guanylin and uroguanylin are produced by mouse intestinal epithelial cells of columnar and secretory lineage. *Histochem. Cell Biol.* **146**, 445–455 (2016).
14. Sato, M. & Bremner, I. Oxygen free radicals and metallothionein. *Free Radic. Biol. Med.* **14**, 325–337 (1993).
15. Gao, S. et al. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018); erratum 20, 1227 (2018).
16. Mojica, W. & Hawthorn, L. Normal colon epithelium: a dataset for the analysis of gene expression and alternative splicing events in colon disease. *BMC Genomics* **11**, 5 (2010).
17. Chu, C. M. et al. Gene expression profiling of colorectal tumors and normal mucosa by microarrays meta-analysis using prediction analysis of microarray, artificial neural network, classification, and regression trees. *Dis. Markers* **2014**, 634123 (2014).
18. Ding, L. et al. Claudin-7 indirectly regulates the integrin/FAK signaling pathway in human colon cancer tissue. *J. Hum. Genet.* **61**, 711–720 (2016).
19. The Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
20. Vanhove, W. et al. Strong upregulation of AIM2 and IFI16 inflammasomes in the mucosa of patients with active inflammatory bowel disease. *Inflamm. Bowel Dis.* **21**, 2673–2682 (2015).
21. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017); correction 50, 1754 (2018).
22. Sasaki, N. et al. Reg4$^+$ deep crypt secretory cells function as epithelial niche for Lgr5$^+$ stem cells in colon. *Proc. Natl Acad. Sci. USA* **113**, E5399–E5407 (2016).
23. Chen, C.-L., Yang, J., James, I. O. A., Zhang, H. Y. & Besner, G. E. Heparin-binding epidermal growth factor-like growth factor restores Wnt/β-catenin signaling in intestinal stem cells exposed to ischemia/reperfusion injury. *Surgery* **155**, 1069–1080 (2014).
24. Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30**, 2496–2497 (2014).
25. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
26. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
27. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386 (2018).
28. Anderson, C. A. et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011); erratum 43, 919 (2011).
29. Leiper, J. M. The DDAH-ADMA-NOS pathway. *Ther. Drug Monit.* **27**, 744–746 (2005).
30. Ellinghaus, D. et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).
31. Jostins, L. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
32. Spenlé, C. et al. Dysregulation of laminins in intestinal inflammation. *Pathol. Biol.* **60**, 41–47 (2012).
33. Chhikara, N. et al. Human epididymis protein-4 (HE-4): a novel cross-class protease inhibitor. *PLoS ONE* **7**, e47672 (2012).
34. Behrens, I., Stenberg, P., Artursson, P. & Kissel, T. Transport of lipophilic drug molecules in a new mucus-secreting cell culture model based on HT29-MTX cells. *Pharm. Res.* **18**, 1138–1145 (2001).
35. O'Sullivan, S., Gilmer, J. F. & Medina, C. Matrix metalloproteinases in inflammatory bowel disease: an update. *Mediators Inflamm.* **2015**, 964131 (2015).
36. Johansson, M. E. V. et al. The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proc. Natl Acad. Sci. USA* **105**, 15064–15069 (2008).
37. Porter, E. M., van Dam, E., Valore, E. V. & Ganz, T. Broad-spectrum antimicrobial activity of human intestinal defensin 5. *Infect. Immun.* **65**, 2396–2401 (1997).
38. Cash, H. L., Whitham, C. V., Behrendt, C. L. & Hooper, L. V. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science* **313**, 1126–1130 (2006).
39. Johansson, M. E. V., Larsson, J. M. H. & Hansson, G. C. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc. Natl Acad. Sci. USA* **108** (Suppl 1), 4659–4665 (2011).
40. Waldman, S. A. & Camilleri, M. Guanylate cyclase-C as a therapeutic target in gastrointestinal disorders. *Gut* **67**, 1543–1552 (2018).
41. Johansson, M. E. et al. Bacteria penetrate the normally impenetrable inner colon mucus layer in both murine colitis models and patients with ulcerative colitis. *Gut* **63**, 281–291 (2014).
42. Johansson, M. E. Fast renewal of the distal colonic mucus layers by the surface goblet cells as measured by *in vivo* labeling of mucin glycoproteins. *PLoS ONE* **7**, e41009 (2012).

Acknowledgements We thank all of the patients who contributed to this study, our endoscopy teams and our clinical research nurses led by S. Fourie, who made this work possible. We acknowledge the support of the Wolfson Imaging Centre, the WIMM flow-cytometry facility, the Discovery Proteomics Facility, R. Dhaliwal for preparing TEM samples, the Oxford NIHR Biomedical Research Centre, the NIHR Clinical Research Network (CRN) Thames Valley, and the Oxford Single Cell Consortium. This work was supported by an NIHR Research Professorship and a Wellcome Investigator Award (to A.S.); the MRC (H.K. and A.S.); Abbvie (K.P.); and Celgene (A. Antanaviciute and H.H.C.). D.F.-C. was supported by a Royal College of Surgeons of England/British Association of Paediatric Surgeons Research Fellowship, an Oxford Wellcome Clinical Training Fellowship and by OHSRC, part of Oxford Hospitals charity. Further acknowledgements are given in the Supplementary Information.

Reviewer information *Nature* thanks Richard Blumberg, Louis Vermeulen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions K.P., D.F-C., A. Antanaviciute and A.S. conceptualized the study. K.P., D.F-C. and M.J. performed and analysed experiments. A. Aulicino, H.H.C., N.A., S.D., J.L., R.S.A and E.B. performed wet laboratory experiments. C.L., N.K.A. and E.J. assisted with all microscopy-related experiments and analysis. R.G. and L.B. assisted with pathology and scoring. P.H. assisted with genetic experiments. A. Antanaviciute, H.K and J.K. performed computational analysis and design. S.D., R.F. and B.M.K. performed proteomic experiments. Writing and editing were carried out by K.P., D.F-C., A. Antanaviciute, H.K. and A.S. H.K. co-supervised and A.S. conceived the study, obtained funding and supervised.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-0992-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-0992-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Isolation of epithelial cells from patient biopsies. Following informed consent, biopsies were collected from volunteers attending endoscopy for routine colonoscopic screening (healthy individuals), or as part of ongoing clinical care (patients with IBD; see Supplementary Table 1 for demographics). For ulcerative colitis, we used tissue derived from immunotherapy-naive patients with a proven histological diagnosis. Tissue was sampled from clinically inflamed distal colons and proximal clinically noninvolved regions. NHS National Research Ethics Service (NRES) research ethics committee (REC) references for the study include GI 16/YH/0247 and IBD 09/H1204/30. Biopsies were incubated in chelation medium (HPGA with 1 mM EDTA) at 37 °C for 80 min with agitation. The supernatant, which was removed and replaced every 20 minutes and contained epithelial crypts, was digested into a single-cell suspension by dissociation with TrypLE Express containing 50 μg ml^{-1} DNase for 1 h at 37 °C. The epithelial single-cell suspension was washed and passed through 70-μm and 40-μm filters. Cell counts and viability were confirmed with a Countess II automated cell counter (Thermo Fisher) with confirmation by manual haemocytometer before further processing.

In all single-cell and exploratory experiments, samples were processed immediately. For some validation experiments (RT–PCR, flow-cytometry panel validation and flow sorting), in which large numbers of samples (more than four) were processed simultaneously, samples were stored by freezing in 1 ml of Cryostor DS10 (Sigma Aldrich). Samples were then thawed and epithelial cells isolated to allow batch processing. Viability and epithelial-cell purity were similar to those of freshly isolated samples (data not shown).

Flow cytometry. Before progressing to scRNA-seq, the purity of epithelial populations was confirmed by fluorescence-activated cell sorting (FACS) analysis using anti-CD90 staining (fluorescein isothiocyanate (FITC); Biolegend), anti-CD326 antibody (epithelial cell adhesion molecule (EPCAM); PeVio, Milteyni), anti-CD45 staining (allophycocyanin (APC); Milteyni) and anti-DNA staining (4′,6-diamidino-2-phenylindole (DAPI); BD Biosciences) as per the manufacturers' instructions. Samples were processed on an Attune NxT flow cytometer (Thermo Fisher) with compensation performed using compensation beads (BD Biosciences) on each run. Once satisfactory viability and EPCAM purity had been demonstrated, samples were directly processed for scRNA-seq.

For validation of the BEST4/OTOP2 cell subpopulation, a similar epithelial staining protocol was used, with addition of a primary anti-BEST4 antibody (Atlas Antibodies), followed by secondary staining for 30 min with Alexa Fluor 488 anti-rabbit secondary antibody (Invitrogen). The staining protocol was validated on the Attune NxT flow cytometer (Thermo Fisher) and then flow sorting was performed on a Sony SH800 cell sorter (Sony) with $BEST4^+$ gates set on fluorescence minus one and secondary control, each of more than 20,000 cells.

Droplet-based single-cell RNA sequencing. Cells were loaded onto the 10X Chromium Single Cell Platform (10X Genomics) at a concentration of 1,000 cells per μl (Single Cell 3′ library and Gel Bead Kit v.2) as described in the manufacturer's protocol (10x User Guide, Revision B). On average, approximately 8,000 cells were loaded across three runs, each with three conditions: healthy, ulcerative colitis inflamed and ulcerative colitis noninflamed. Cells were suspended in phosphate-buffered saline (PBS) with 0.04% bovine serum albumin (BSA). Generation of gel beads in emulsion (GEMs), barcoding, GEM-RT clean-up, complementary DNA amplification and library construction were all performed as per the manufacturer's protocol. Individual sample quality was checked using a Bioanalyzer Tapestation (Agilent). Qubit was used for library quantification before pooling. The final library pool was sequenced on the Illumina NovaSeq6000 instrument using 150-base-pair paired-end reads. Average cell recovery was 1,400 cells per sample, with a total of 11,175 cells captured at a mean depth of 163,822 reads per cell and 1,736 mean genes per cell.

Plate-based scRNAseq, real-time PCR and RNA amplification. Single cells were sorted as previously described and plate-based scRNA-seq was performed as per the Smart-seq2 protocol[43] with minor adaptations. Reverse transcription was carried out with 0.75 units per reaction of SMARTScribe (Clonotech, Takara) and PCR pre-amplification with 5′-biotinylated IS PCR primers (Biomers) for 25 cycles. Post-PCR cleaning was performed with Ampure XP Beads (Beckman Coulter) at a ratio of 0.8/1 (beads/cDNA). cDNA was resuspended in elution buffer (Qiagen) and quality assessed with a high-sensitivity DNA chip (Agilent).

Barcoded Illumina sequencing libraries (Nextera XT, Library preparation kit, Illumina) were generated using the automated platform (Biomek FXp), and libraries were pooled and sequenced using the Illumina NextSeq sequencer.

For bulk RNA amplification in small cell numbers (12,500–25,000), RNA was isolated using the RNAeasy MicroKit (Qiagen) according to the manufacturer's instructions. We then added 1 μl of extracted RNA to a 96-well plate containing lysis buffer and processed it with the same SMART-seq2 protocol with 20 cycles of pre-amplification.

For microfluidic quantitative PCR of small cell numbers, 100 $BEST4^+$ and 100 $BEST4^-$ cells were isolated from three biological replicates. RNA was amplified using a specific targeted amplification strategy targeting the specified gene primers (Taqman, ThermoFisher) in the reverse-transcription mix as per manufacturers' protocols (Biomark, Fluidigm). The primers used are described in Supplementary Table 3. The expression of 12 genes was quantified using an integrated microfluidic chip (Flex 6 IFC) as per manufacturers' instructions (Biomark, Fluidigm). A sample with no reverse transcriptase was included as a control.

For quantitative RT–PCR experiments with larger cell numbers (more than 25,000 cells), total RNA was isolated using the RNeasy microkit (Qiagen) according to manufacturer's instructions. cDNA was then synthesized using the high-capacity RNA-to-cDNA kit (ThermoFisher 4387406), with RT–PCR then performed using applicable Taqman gene expression assays on the QuantStudio 7-Flex system (ThermoFisher). Details of individual gene-expression assays are included in Supplementary Table 3.

Proteomic analysis of BEST4/OTOP2 cell population. To characterize the BEST4/OTOP2 cell population by proteomics, we isolated $BEST4^+EPCAM^+$ and $BEST4^-EPCAM^+$ populations using FACS as previously described. We sorted 6,250 cells in both conditions into 25 μl of lysis buffer, which consists of radioimmunoprecipitation assay (RIPA) buffer (Sigma) with 4% NP-40 (IPEGAL, Sigma). After thawing, 1 μl of benzonase (E1014, Sigma) was added and samples were kept on ice for 30 min. Protein lysates were digested using a modified SP3 protocol[44]. In brief, proteins were reduced with 5 mM dithiothreitol for 30 min and then alkylated with 20 mM iodoacetamide for 30 min at room temperature. We then mixed 2 μl of carboxyl-modified paramagnetic beads (prepared as in ref. [45]) with the samples. Acetonitrile was added to the samples to a final concentration of 70% (v/v). Protein binding to the beads was carried out for 18 min with orbital shaking at 1,000 r.p.m. Beads were then immobilized on a magnet for 2 min and the supernatant discarded. Beads were washed twice with 70% (v/v) ethanol and once with 100% acetonitrile, all on the magnet. Beads were resuspended in 50 mM ammonium bicarbonate containing 25 ng trypsin and digested overnight at 37 °C. After digestion, the beads were resuspended by brief bath sonication. Acetonitrile was added to 95% (v/v), and samples were shaken at 1,000 r.p.m. for 18 min to bind peptides; beads were then immobilized on the magnet for 2 min and the supernatant discarded. Beads were resuspended in 2% dimethylsulfoxide (DMSO) and then immobilized on the magnet for 5 min; the supernatant was transferred to liquid chromatography–mass spectrometry (LC–MS) vials, which were stored at −20 °C until analysis.

Peptides were analysed by nanoscale ultraperformance liquid chromatography coupled to tandem mass spectrometry using a Dionex Ultimate 3000 coupled on-line to an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). A 75 μm × 500-mm C18 EASY-Spray column (Thermo Scientific) with 2-μm particles was used at a flow rate of 250 nl min^{-1}. Peptides were separated using a 60-min linear gradient from 2% buffer B to 35% buffer B (A: 5% DMSO, 0.1% formic acid in water; B: 5% DMSO, 0.1% formic acid in acetonitrile). Precursor scans in the first stage of mass spectrometry (MS1) were performed in the Orbitrap at a resolution of 120,000, a mass-to-charge ratio (m/z) of 200 and a rate of 1 Hz. Precursors were selected for tandem mass spectrometry analysis using an isolation window of 1.6 m/z and were fragmented using higher-energy collision dissociation (HCD) at a normalized collision energy of 28. MS2 fragment spectra were acquired in the ion trap using the Rapid scan rate.

pH imaging. pH imaging was carried out as previously described[12]. In brief, sorted $BEST4^+EPCAM^+$ and $BEST4^-EPCAM^+$ cells were plated onto poly-L-lysine-coated coverslips at 37 °C. After at least 1 hour, cells were loaded with the intracellular pH indicator pHrodo Red AM, using PowerLoad concentrate according to the manufacturer's instructions (Molecular Probes). pH imaging optics and image acquisition were measured using an Olympus DeltaVision II Microscope System. pHrodo red fluorescence intensity for each cell was measured in response to pH 5.0 solution buffered with 2-(*N*-morpholino)ethanesulfonic acid (MES; 150 mM NaCl, 10 mM MES, 2 mM $CaCl_2$). The pHrodo Red fluorescence intensity of each cell was normalized to its baseline fluorescence in pH 7.4 solution (F_0) before the first acid application to determine F/F_0. Cells were then permeabilized with valinomycin and nigericin, and fluorescence measured in high K^+-containing extracellular solutions at pHs 4.5, 5.5, 6.5 and 7.5. A standard curve was generated and the pH_i for each cell was calculated using linear extrapolation.

Immunohistochemistry, immunofluorescence and smISH. For immunohistochemistry, paraffin-embedded tissue sections were deparaffinized through an ethanol gradient, and heat-induced epitope retrieval was performed by boiling at 96 °C for 25 min in either pH 6 citrate or pH 9 Tris/EDTA buffers. Peroxidase was blocked before incubation with an appropriate species-specific serum for primary-antibody incubation of 90 min at room temperature. For full details on the antibodies and concentrations used, see Supplementary Table 2.

Substrate development was performed for each primary antibody using ImmPACT diaminobenzidine (DAB), VectorBlue or NovaRed as appropriate for brown, blue or red development as required (all from Vector Laboratories). In cases of double staining, samples underwent sequential heat-induced epitope retrieval if labelled with same-species antibodies. Haematoxylin and eosin staining was carried out using a kit from Vector Laboratories.

For immunofluorescence, the protocol was followed exactly as described above except that the primary antibodies were incubated overnight at 4 °C. Slides were then washed in PBS and incubated with the appropriate secondary antibodies labelled with Alexa Fluor (Molecular Probes) for 1 h at room temperature in the dark. Slides were washed again and incubated with DAPI for 5 min before washing and mounting using Vectashield (Vector Laboratories).

For smISH, all probes and RNAscope 2.5 HD assay–brown (catalogue number 310035) were purchased from Advanced Cell Diagnostics (ACD) and used according to the manufacturer's instructions. Details of probes are in Supplementary Table 2. Paraffin sections were pretreated with Pretreat 1, 2, and 3 (ACD). Prewarmed (40 °C) probes were added to the slides and incubated in a HybEZ oven (catalogue number 321461; ACD) for 2 h at 40 °C. After six-step signal amplification, tissues were detected using DAB (all part of the RNAscope 2.5 HD assay–brown kit) and counterstained with Mayer's haematoxylin. Slides were mounted with Pertex mounting medium and photographed.

For double staining in smISH and immunohistochemistry (to detect *OTOP2* and BEST4), samples were processed in a manner identical to smISH, with subsequent overnight staining with anti-BEST4 antibodies and development with VectorBlue.

Transwell cultures at 21 days old were fixed using 10% neutral buffered formalin, then membranes were cut out and paraffin-embedded. Haematoxylin and eosin and Alcian blue stains were carried out according to the manufacturers' protocols (Vector Laboratories and Sigma-Aldrich, respectively). For experiments involving staining the mucus layers in mice, the colon was dissected along with faecal content and fixed in chloroform-based Carnoy's fixative[46]. The tissue was fixed overnight, following by washing in methanol and paraffin embedding as usual.

Quantification of patient biopsies using Visiopharm. Slides stained for WFDC2 as described above were scanned using a Leica ScanScope machine (Leica Biosystems) and quantified using Visiopharm (Visiopharm) with a programmed protocol calculated as follows: percentage of positive goblet-cell area = fraction of goblet-cell area × 100, in which the fraction of goblet-cell area was calculated as the goblet-cell area (area inside region of interest (ROI) 1, set on a defined image) as a fraction of the area of interest (total ROI2 in a defined image). The protocol was set to repeat the calculation randomly for 50% of the whole biopsy and give an average for each sample. The result was equated to the percentage of positive/brown stain in goblet cells for each biopsy. For the data presented in Fig. 4d, additional unmatched patient data for WFDC2 were included.

Antibacterial activity. Mid-logarithmic-phase cultures of ATCC12973 *S. aureus*, ATCC13379 *E. faecalis*, ATCC27853 *P. aeruginosa*, *S.* Typhimurium LT2 and *E. coli* were incubated with diluted PBS containing Lucia Bertani (LB), tryptic soy broth (TSB) or brain and heart broth (all Sigma-Aldrich) to a concentration of 2×10^4 colony-forming units per ml. A final volume of 100 μl was used to measure antibacterial activity of recombinant WFDC2 (Abcam) by addition at a concentration of between 0.45–4.50 μM. Following a 4-h incubation at 37 °C, the frequency of bacteria was determined by serial dilution onto LB agar or Columbia blood agar (CBA) plates (Sigma-Aldrich). All plates were incubated overnight and before counting bacterial colonies. Survival was calculated as the percentage of bacteria present at 4 h compared with baseline. All experiments were performed three times.

Antiprotease activity. MMP-inhibitor activity was performed according to the manufacturer's protocol (Enzo Life Sciences: MMP12, catalogue number BML-AK403-0001; MMP13, BML-AK413-0001). In brief, all kit components were diluted according to the recommended concentration. We added 20 μl of each MMP (with varying concentrations; refer to the manufacturer's manual for lot-dependent variations) to control, inhibitor *N*-isobutyl-*N*-(4-methoxyphenylsulfonyl)glycl hydroxamic acid (NNGH) and WFDC2 (25 μg ml^{-1}). The reaction plate was incubated for 30 min at 37 °C to allow inhibitor and enzyme interaction. Next, 10 μl of substrate BML-P277-9090 was added to each reaction and the fluorescence was measured at an excitation/emission of 545 nm/576 nm for 10 min at 37 °C.

Animals. For *Wfdc2* experiments, mice were housed under standard conditions in the MRC Harwell animal facility according to institutional guidelines. Mice heterozygous for the targeted *Wfdc2* allele (*Wfdc2*$^{em1(IMPC)H}$) were generated by injecting targeted embryonic stem cells (obtained from the European Mouse Mutant Archive) into blastocysts (MRC Harwell Transgenic Facility)[47–49].

For DSS colitis experiments, we used 10–12-week-old C57BL/6 male mice (free from *Helicobacter pylori* and murine norovirus; Envigo Laboratories). All procedures were certified according to the UK Home Office Animals (Scientific Procedures) Act 1986 (project license P9B86E6FD). Ten mice were randomized into two treatment groups of five mice each. One group received no treatment and the other received 1.75% DSS (molecular weight 36–50 kDa; MP Biomedicals, lot M9147) in their drinking water from study day 0 until mice were euthanized on the morning of study day 7, and the tissue processed for routine immunohistochemistry as described above. For another independent study, we used a group of eight mice with four mice per group, treated with DSS as described above. RNA isolation from tissue was performed using a Qiagen mini kit (Qiagen), as described earlier.

Wfdc2$^{+/-}$ mice were assigned a subjective colitis severity score based on a modification of the criteria in ref. [50]. Scores for morphology, ulceration and infiltration were ranked on a scale from 0 (normal or absent) to 4 (severe), which were summed to give an overall score.

SEM and TEM. Mice underwent perfusion fixation with 2.5% glutaraldehyde plus 4% paraformaldehyde in 0.1 M sodium cacodylate and the colon was excised. Tissue was cut into pieces of 2–3 mm^3 and then stored at 4 °C in 0.25% glutaraldehyde in 0.1 M sodium cacodylate buffer, until processing for TEM. Samples were washed with 0.1 M sodium cacodylate buffer followed by 50 nM glycine in the same buffer to quench free aldehydes, followed by another wash with 0.1 M sodium cacodylate buffer. Samples were incubated in 1% osmium tetroxide plus 1.5% potassium ferrocyanide in 0.1 M sodium cacodylate buffer for 2 h at 4 °C with vigorous agitation, and then washed with water, before overnight incubation at 4 °C in 0.5% uranyl acetate. Samples were washed with water and dehydrated through a graded series of ice-cold ethanol for 15 min each followed by a final incubation in 100% ice-cold ethanol for 90 min. Samples were then infiltrated with 25% agar low viscosity epoxy resin (Agar Scientific) in ethanol for 3 h and then 50% resin overnight, followed by 75% and 100% resin each for 3 h and then 100% resin overnight. The samples were embedded in flat-dish moulds and polymerized at 60 °C for 48 h. Ultrathin (90-nm) sections were cut with a Diatome diamond knife using a Leica UC7 ultramicrotome and post-stained for 5 min with lead citrate. Sections were viewed on a FEI Tecnai 12 TEM operated at 120 kV equipped with a Gatan OneView digital camera.

For SEM, colons were fixed and processed as above, and dehydrated in a graded ethanol series as above. They were then incubated in absolute ethanol at room temperature, dried by critical point drying (Tousimis Autosamdri-815b), and placed on an SEM stub using conductive silver dag and sputter-coated with gold in a Quorum Q150R ES coating unit. Specimens were imaged using a Zeiss Sigma 300 field emission gun SEM at an acceleration voltage of 2 kV.

Cell culture. A goblet-cell-producing cell line, HT29-MTX-E12 (ref. [34]), was obtained from ATCC and maintained in Dulbecco's modified Eagle's medium (DMEM) glutamax medium (Life Technologies) containing 10% fetal bovine serum (v/v) and 1% (v/v) antibiotics. Cultures were incubated at 37 °C in a humidified 5% (v/v) CO_2 atmosphere and used between passages 10 to 20. For secretion assays, cells were seeded in 24-well culture plates at a concentration of 4.0×10^4 cells per well. The culture medium was changed every two days, and medium without antibiotics and serum was used for the last medium change. Experiments were performed 21 days post-seeding[51]. Cells were stimulated with or without 100 ng ml^{-1} of phorbol 12-myristate 13-acetate (PMA) for 6 h before apical and basal medium collection. WFDC2 was quantified using a human WFDC2 Quantikine ELISA kit according to the manufacturer's instructions (R&D Systems).

Knockdown of *WFDC2* with small hairpin RNA. A small hairpin RNA (shRNA) oligo targeting *WFDC2* was purchased from Sigma (catalogue number SHCLNG-NM_006103) (Supplementary Table 4). HEK-293 T cells were transfected with *WFDC2* shRNA along with packaging vectors, using lipofectamine as per the manufacturer's instructions (Thermo Fisher Scientific). Viral supernatant was collected 72 h after transfection, and concentrated by ultra-centrifugation. Cells were transduced with concentrated lentiviral particles expressing *WFDC2* shRNA in the presence of polybrene as described[52]. Knockdown efficiency was assessed by immunoblotting for WFDC2 and secretion of WFDC2 in culture supernatant.

Organoid cultures. Organoid cultures were established as previously described[53]. In brief, cultures were established from four pairs of colonic biopsies, incubated with 0.4 mg ml^{-1} dispase (Gibco) to establish a single-cell suspension. This was then mixed with 50 μl Matrigel (Corning) and plated on prewarmed 24-well culture dishes. Embedded cells were overlaid with WREN medium (WNT3A conditioned medium, containing lipid-modified WNT3A; ATCC catalogue number CRL:2647TM) and ADF (advanced DMEM-F12 medium; Gibco) 50/50, glutamax (Life Technologies), 10 mM HEPES, N-2 supplement (×1) (Life Technologies), B-27 supplement (×1) (Life Technologies), 10 mM nicotinamide (Sigma Aldrich), 1 mM *N*-acetyl-L-cysteine (Sigma Aldrich), 1 μg ml^{-1} R-spondin 1 (RSPO1; Peprotech), 50 ng ml^{-1} human epidermal growth factor (EGF; Peprotech), 100 ng ml^{-1} human Noggin (Peprotech), 1 μg ml^{-1} gastrin (Sigma Aldrich), 0.05 μM prostaglandin E2 (Sigma Aldrich), 0.1 μM A83-01 (an inhibitor of activin receptor-like kinases), 10 μM p38 inhibitor SB202190, and 10 μM Y27632

(a Rho-associated kinase inhibitor) (all from R&D Systems). Medium was replaced with fresh WREN medium every other day.

For organoid-stimulation experiments, once organoid cultures were established, 100 ng ml^{-1} interferon-γ was added to medium for four days in experimental conditions. For gene-expression quantification, we isolated RNA from organoids and performed RT–PCR as described above.

Computational analysis. *Raw 10× read processing and quality control.* Raw sequence reads were quality-checked using FastQC software. The Cell Ranger version 2.1.1 software suite from 10× Genomics (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to process, align and summarize unique molecular identifier (UMI) counts against the human hg38 assembly reference genome analysis set, obtained from the University of California Santa Cruz (UCSC) ftp site[54].

Corresponding Ensembl gene annotation was obtained using the UCSC Table Browser Tool[55]. Raw, unfiltered count matrices were imported into R for further processing.

Raw UMI count matrices were filtered to remove: barcodes with very low (empty wells) and very high (probably doublets) total UMI counts; matrices for which a high percentage of UMIs originated from mitochondrial features (more than 20%); and matrices for which fewer than 200 genes were expressed. Distributions of UMIs per cell per sample were visualized using the ggplot2 R package (Extended Data Fig. 10a).

The Seurat R package (version 2.3.2) was used to normalize expression values for total UMI counts per cell[56].

Assessment of mRNA content per cell. Cell Ranger software version 2.1.1 was used to downsample data to the level of the lowest coverage sample on the basis of the aligned number of reads per cell. Downsampled UMI count matrices were used to obtain the gene cellular detection rate, in which a gene was considered detected if at least one UMI was assigned to it. The numbers of genes detected per cell for ulcerative colitis inflamed and noninflamed tissue and for healthy samples were visualized as density distributions (Extended Data Fig. 10b). Distributions of gene-detection rates were also visualized on a per-cell-type basis (Extended Data Fig. 10c).

Clustering. Initially an integrated clustering analysis of all samples (Extended Data Fig. 10d) was performed and was then used to guide complementary cluster identification between conditions. Clustering was performed as follows. Cell-cycle-stage annotation was performed using the 'cyclone' function from the R package scran (version 1.6.9)[57]. The resulting G_2M and G_1 phase cell-cycle scores together with total UMI counts per cell, the percentage of mitochondrial features and experimental batches were considered to be a source of unwanted variation and were regressed out using the Seurat package. Variable genes were found either by identifying outliers from fitting the mean-variance relationship in the data, or by fitting the relationship between mean expression and drop-out rate using the R package M3Drop (version 1.6.0). Dimensionality reduction was performed using principal-component analysis (*R* package irlba, version 2.3.2). Scree plots and Jack–Straw permutation tests were used to determine significant principal components (with a *P* value cut-off of less than 0.01) in the data, and cells were clustered in the reduced dimension space using the Seurat package (resolution = 0.7). Cell clusters were visualized using *t*-SNE plots, with all significant principal components (as previously determined by Jack–Straw permutation tests) as input and a perplexity value of 30. We found 20 principal components to be significant and used them to cluster the whole dataset; 8 were used for clustering analysis of undifferentiated cells, 10 for goblet subcluster analysis and 6 for EEC analysis.

Batch-effect assessment. To ensure that clustering was not driven by batch effects, we visualized batch distributions for the dataset using *t*-SNE plots (Extended Data Fig. 10e). We also quantified this effect by computing entropy of batch mixing, as described in ref. [58], for *t*-SNE cell embeddings of sample batches. As a negative control (no batch effect), we assigned each cell a random batch label and computed the expected entropy. Similarly, as a positive control (in which clustering is driven entirely by batch effects), we used cluster identities as batch labels for entropy calculations. Each set of entropies was computed from the neighbourhoods of 100 randomly picked cell locations, bootstrapped 100 times and the distributions visualized as box plots (Extended Data Fig. 10f).

Crypt-axis score. The expression of the following genes was used to define a crypt-axis score: *SEPP1*, *CEACAM7*, *PLAC8*, *CEACAM1*, *TSPAN1*, *CEACAM5*, *CEACAM6*, *IFI27*, *DHRS9*, *KRT20*, *RHOC*, *CD177*, *PKIB*, *HPGD* and *LYPD8*. For each gene, we normalized expression across all cells to a range between 0 and 1, to ensure that the contribution of individual genes to the score was not weighted by baseline expression. The final crypt-axis score for each cell was then defined as the sum of all normalized expression values.

Semisupervised clustering of public scRNA-seq data. To test whether BEST4/OTOP2 cells are present in other single-cell datasets, we downloaded data from the Gene Expression Omnibus (GEO; accessions GSE103239 and GSE81861), processing the data as described above, except that clustering was performed using the a priori identified highly variable genes from our 10× data analysis. Cluster markers were detected as before and compared to the BEST4/OTOP2 cell markers in the 10× data.

Analysis of The Cancer Genome Atlas data. High-throughput-sequencing raw count matrixes were downloaded from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov) for all available patients with colorectal cancer and matched normal samples. Data were normalized using the DESeq2 R package and variance stabilized using the 'rlog' function[59]. Sample clustering and expression of the core BEST4/OTOP2 cell gene signature in this dataset were visualized using the R package pheatmap.

Cluster marker and differentially expressed gene identification. Cluster gene markers were detected with the Seurat package, using the area-under-curve classifier and/or negative binomial likelihood ratio tests. Differentially expressed genes between groups in each cluster were detected using the negative binomial likelihood ratio test. Patient/sample batches, total UMI counts, percentage mitochondrial gene expression and cell-cycle scores were used as model covariates. The Benjamini–Hochberg multiple testing correction was applied and genes with an FDR of less than 1% were considered to be differentially expressed.

The R package MAST[60] was used to estimate generalized linear model coefficients for inflamed and noninflamed samples, using cells from healthy patients as a reference level. We built individual models for all major cell clusters using the 'zmb' function, in which—in addition to ulcerative colitis status (inflamed, noninflamed or healthy)—we modelled gene-detection rate, cell-cycle and donor effects. Correlation between coefficients was visualized as a scatter plot between individual genes.

Differentially expressed gene identification from publicly available microarray data. Data were downloaded from the GEO (accession number GSE59071) for inflamed colon ulcerative colitis samples and healthy controls. The R package limma was used for data normalization and differential expression analysis[61]. The Benjamini–Hochberg multiple testing correction was applied to estimate the FDR.

Ontology enrichment analysis. Biological-process Gene Ontology enrichment of cluster markers and differentially expressed genes/proteins was performed using the R package clusterProfiler version[62] with a Benjamini–Hochberg multiple testing adjustment and an FDR cut-off of 0.05, using all expressed/detected genes as a background control. The results were visualized as dot plots or emap plots using the clusterProfiler and ggplot2 R packages.

Smart-seq2 scRNA-seq data processing and analysis. Raw sequencing data were demultiplexed into one fastq file per plate well using bcl2fastq software, version 2.20.0.422. Reads were quality checked using FastQC software. Illumina Nextera sequencing adapters, Smart-seq2 oligonucleotide sequences, poly-d(T) and poor-quality (fewer than 20-base-pair) sequences were trimmed using Cutadapt software. Reads were aligned to the human hg38 reference genome build (UCSC) using the STAR aligner[63]. Raw read-count matrices were obtained using the featureCounts tool. Data-quality metrics for each well were aggregated using multiQC tool.

The R package scater was used to process raw count data. Poor-quality wells were filtered on the basis of the following criteria: less than 60% of reads uniquely mapped, less than 500 genes detected, more than 20% of reads mapping to mitochondrial features. Library normalization size factors were computed using the R package SCNorm[64]. A small number of contaminating immune cells was identified by expression of CD45/PTPRC and filtered out from the analysis. BEST4/OTOP2 cell marker genes were identified using the R package Seurat as described.

Proteomics data analysis. Label-free quantification of proteins was performed using Progenesis QI for Proteomics (version 4.1, Waters) and proteins were identified using MASCOT (Matrix Science) by searching against the Uniprot reference human proteome (95,128 sequences, database accessed on 18 July 2018). Precursor mass tolerance was set to 10 parts per million; fragment mass tolerance was 0.5 Da; and a maximum of two missed cleavages were allowed. Carbamidomethylation of cysteine was set as a fixed modification; the variable modifications allowed were deamidation of asparagine/glutamine and oxidation of methionine. Peptide FDR was adjusted to 1% and low-scoring peptides (less than 20) were excluded. The R package limma was used for normalization of protein expression and differential expression analysis[61]. The Benjamini–Hochberg multiple testing correction was used to estimate the FDR.

BEST4/OTOP2 cell marker overlap. The intersection of the top 200 markers for BEST4/OTOP2 cells—identified from 10× single-cell data, SmartSeq2 data, quantitative proteomics assays and previous datasets[15,21]—was visualized using the R package circos.

Trajectory and pseudotime analysis. Cell-differentiation trajectories were reconstructed using the R package monocle (version 2.8.0)[65]. Non-epithelial-cell clusters were filtered out and dimensionality reduction was performed with the DDRTree algorithm, using all highly variable genes as inputs and the following residual model formula: ~donor + number of UMIs + percentage of mitochondrial gene expression + G1 score +G2M score. Cell trajectory was then reconstructed using

the orderCells function and the starting state was denoted as the branch encompassing the previously identified stem cells at the most distal end.

To identify putative lineage regulators, we first identified genes that change between secretory and absorptive branches using branched-expression-analysis modelling (negative binomial likelihood ratio test), modelling pseudotime as a covariate. Then, we identified genes that are induced before the trajectory bifurcation point by performing a differential expression test between the cells in the earliest trajectory state, as identified by Monocle, and later prebranch state cells. All significantly upregulated (less than 1% FDR, greater than 0 fold change (expressed in $\log_2$)) genes were then intersected with all previously identified genes that showed significant pseudotime-varying, branch-specific expression. This subset identified genes with branch-specific expression that are also induced before lineage divergence. In all of the above statistical tests, patient/sample batches, cell-cycle scores, cell-size factors and percentage of mitochondrial gene expression were modelled as covariates.

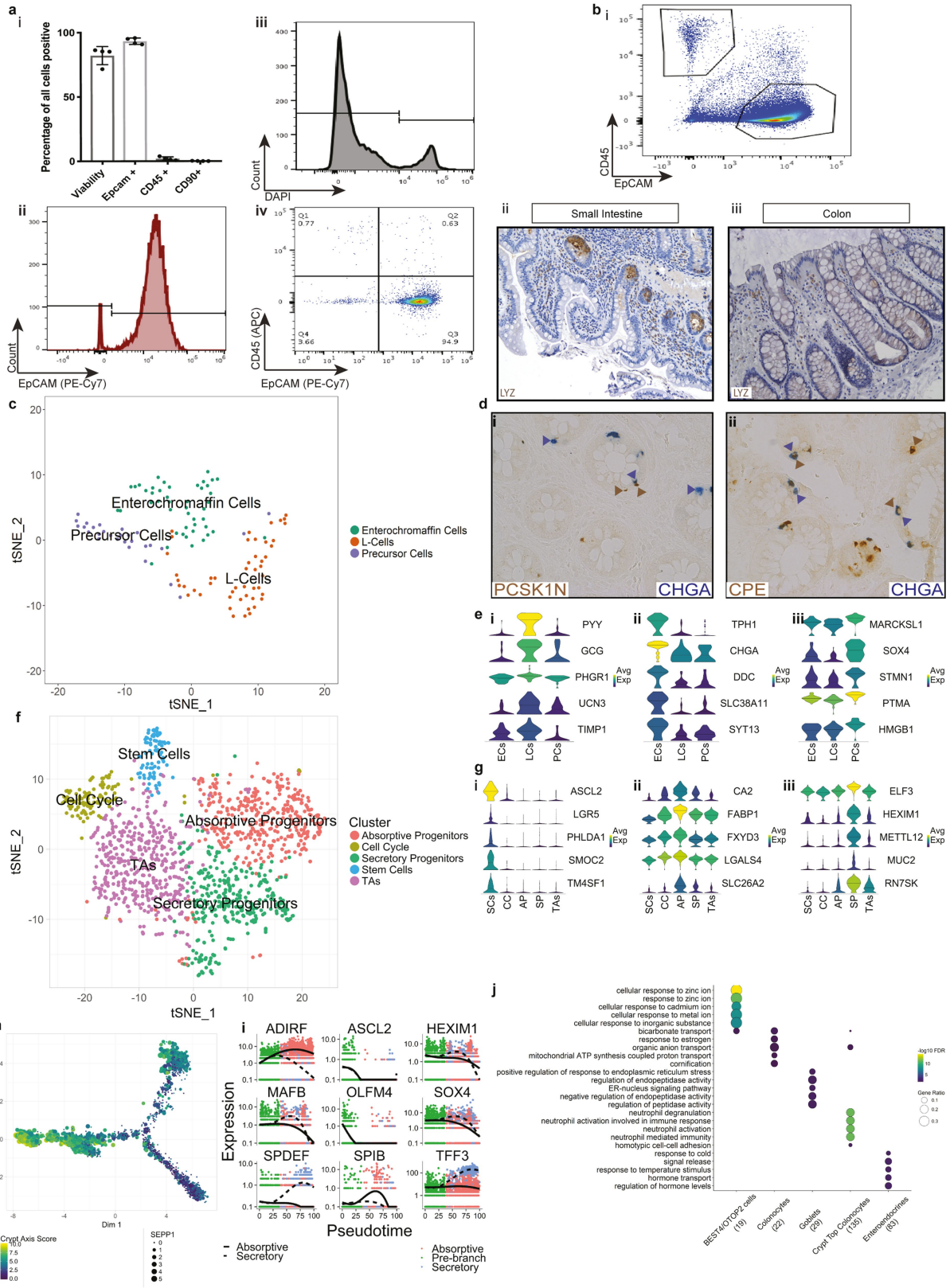
Analysis of tissue-specific expression of GWAS loci. We used the SNPsea algorithm[24] to test for significant enrichment of tissue-specific expression in ulcerative-colitis-associated GWAS loci genes. From the GWAS catalogue[66], we downloaded the ulcerative-colitis-associated loci from refs. [25,26], which report the largest number of ulcerative-colitis-associated loci to date. We used data from the 1000 Genomes Project[67] to sample matched control single-nucleotide polymorphisms (SNPs) and to link SNPs to genes. We first used Gene Atlas gene-expression data (accession number GSE1133) to recapitulate the previous association of T-cell-specific gene-expression enrichment[68] in IBD-associated loci. For single-cell RNA-seq data, we created a 'pseudobulk' dataset for each previously identified cell cluster in health and ulcerative colitis separately by summing all UMI counts for each gene in each cluster. We then normalized the data by computing size factors (R package DESeq2, version 1.20.0) to account for differences in cell-cluster sizes. In all cases, SNPsea was run with the following parameters: –slop 10e3, –threads 8, –null-snpsets 1000, –min-observations 100, –max-iterations 1e7, –score single. Obtained *P* values were further subjected to Benjamini–Hochberg multiple testing correction.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

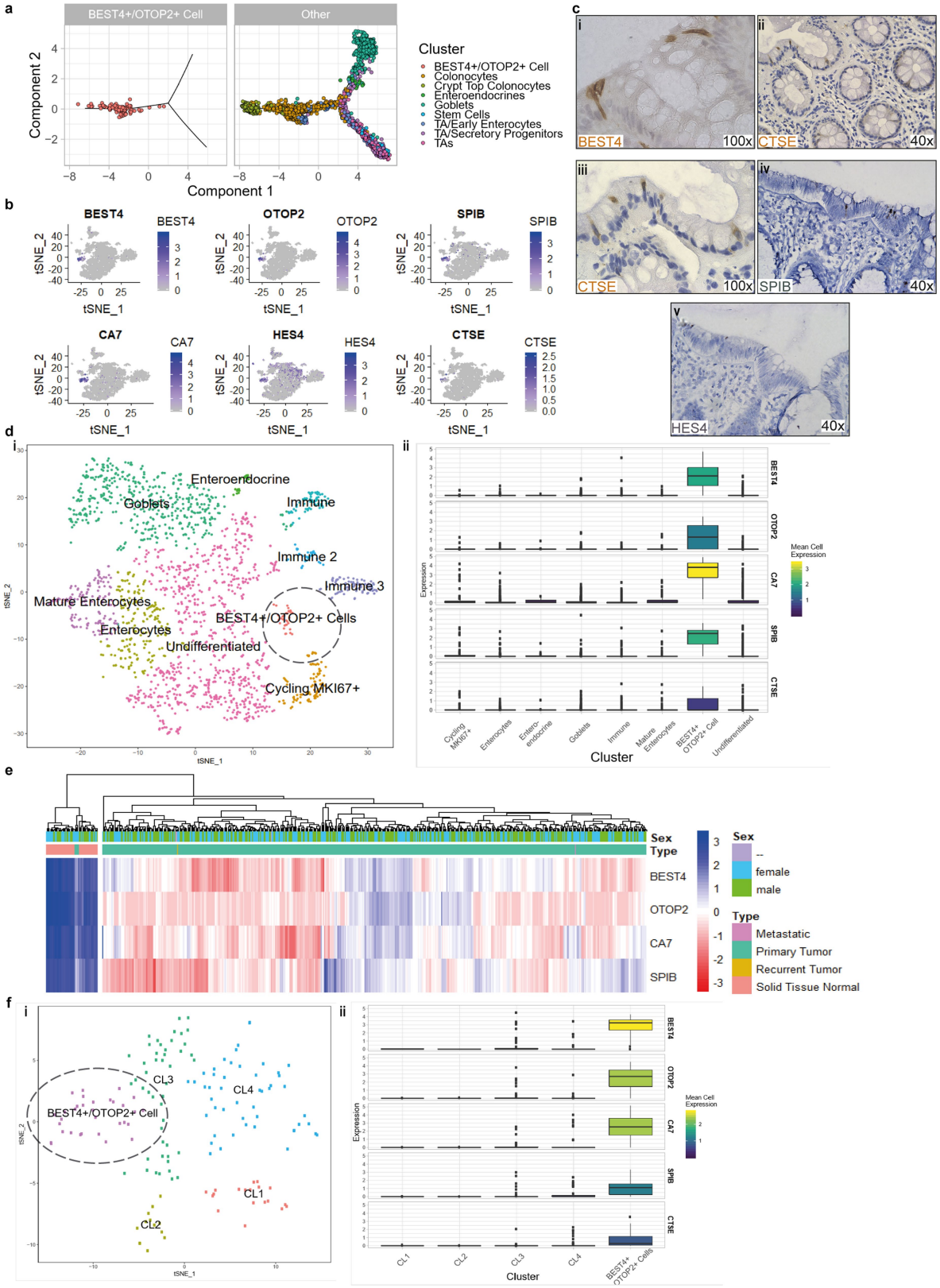
Raw and processed sequencing data files are available under the GEO accession number GSE116222. The source code for analyses has been deposited at https://github.com/agneantanaviciute/colonicepithelium. Proteomics data have been deposited at the ProteomeXchange Consortium (http://www.proteomexchange.org) via the PRIDE[69] partner repository with the dataset identifiers PXD011655 and 10.6019/PXD011655.

43. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171-181 (2014).
44. Sielaff, M. et al. Evaluation of FASP, SP3, and iST protocols for proteomic sample preparation in the low microgram range. *J. Proteome Res.* **16**, 4060–4072 (2017).
45. Hughes, C. S. et al. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **10**, 757 (2014).
46. Johansson, M. E. V. et al. Bacteria penetrate the inner mucus layer before inflammation in the dextran sulfate colitis model. *PLoS ONE* **5**, e12238 (2010).
47. Skarnes, W. C. et al. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342 (2011).
48. Bradley, A. et al. The mammalian gene function resource: the International Knockout Mouse Consortium. *Mamm. Genome* **23**, 580–586 (2012).
49. Pettitt, S. J. et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat. Methods* **6**, 493–495 (2009).
50. Kojouharoff, G. et al. Neutralization of tumour necrosis factor (TNF) but not of IL-1 reduces inflammation in chronic dextran sulphate sodium-induced colitis in mice. *Clin. Exp. Immunol.* **107**, 353–358 (1997).
51. Lesuffleur, T. et al. Differential expression of the human mucin genes MUC1 to MUC5 in relation to growth and differentiation of different mucus-secreting HT-29 cell subpopulations. *J. Cell Sci.* **106**, 771–783 (1993).
52. Berger, G. et al. A simple, versatile and efficient method to genetically modify human monocyte-derived dendritic cells with HIV-1-derived lentiviral vectors. *Nat. Protocols* **6**, 806–816 (2011).
53. Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
54. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
55. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
56. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
57. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
58. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
60. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
61. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
62. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
63. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
64. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).
65. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
66. MacArthur, J. et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.* **45** (D1), D896–D901 (2017).
67. The 1000 Genome Projects Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
68. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
69. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44** (D1), D447–D456 (2016).



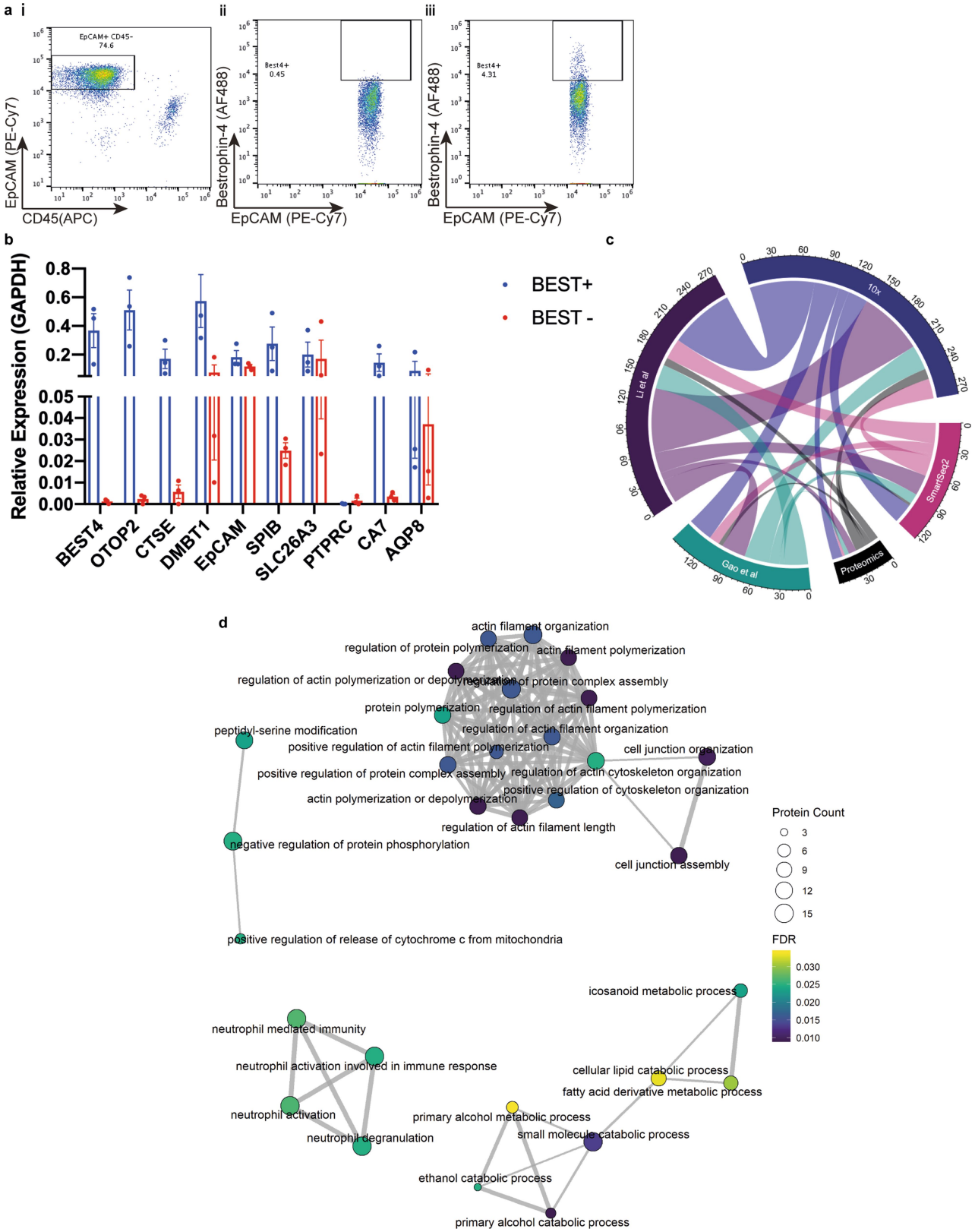
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Identification and validation of epithelial-cell subpopulations. This figure is related to Fig. 1. **a**, **i**–**iv**, Flow-cytometry analysis of cells isolated from biopsies of healthy controls before scRNA-seq (**i**), measuring epithelial viability ($DAPI^-$), purity ($EPCAM^+$), an immune marker ($CD45^+$) and a stromal marker($CD90^+$) ($n = 4$, mean ± s.d.), demonstrating the gating strategy for known epithelial markers (**ii**), viability (**iii**) and immune compartment (**iv**). APC and PE–Cy7 are fluorescent labels. **b**, **i**, FACS purification of $EpCAM^+CD45^-$ isolated epithelial cells ($n = 2$). **ii**, **iii**, Representative images ($n = 3$) of immunohistochemical validation for LYZ expression in human epithelial tissue sections in small intestine (positive control) (**ii**) and colon (**iii**) (images shown at ×20 magnification). **c**, *t*-SNE plot of EEC subclusters. Single cells are coloured by cluster annotation. Descriptive cluster labels are shown ($n = 3$ per group). **d**, **i**, **ii**, Enteroendocrine subsets validated (representative images, $n = 3$) with double-stain immunohistochemistry for CHGA (blue) and two more novel markers identified from scRNA-seq: PCSK1N (**i**, brown) and CPE (**ii**, brown), showing co-localization of both markers in some cells (blue and brown arrowheads) but not in other EECs (blue or brown arrowhead). **e**, **i**–**iii**, Violin plots showing gene expression (*y*-axis) of different top EEC subcluster markers for enterochromaffin cells (ECs), L-cells (LCs) and a precursor-cell population (PCs) ($n = 3$; centre bar indicates median value; colour indicates mean expression). **f**, *t*-SNE plot visualizing undifferentiated colonic epithelial cell subclusters ($n = 3$). **g**, **i**–**iii**, Violin plots of gene expression (*y*-axis) in stem cells (SCs), cell-cycle cluster cells (CC), absorptive progenitor cells (APs), secretory progenitor cells (SPs) and transit-amplifying cells (TAs). The top markers for SCs (**i**), APs (**ii**) and SPs (**iii**) are shown ($n = 3$; centre bar indicates median value; colour indicates mean expression). **h**, Crypt-axis score superimposed over the differentiation trajectory captured by Monocle analysis ($n = 3$). Dimension (Dim) 1 and 2 are on the *x*- and *y*-axis respectively. **i**, Branch-specific expression of selected SC markers, secretory-lineage-specific markers and putative novel lineage-specific transcriptional regulators ($n = 3$). **j**, Selected Gene Ontology terms that show significant enrichment among all marker genes for epithelial clusters. The number of markers identified for each cluster is indicated (*x*-axis). The size of each circle corresponds to the proportion of markers annotated to a given term, while the colour indicates the significance (FDR) ($n = 3$ biological replicates; hypergeometric test and FDR calculated; Benjamini–Hochberg multiple testing correction).

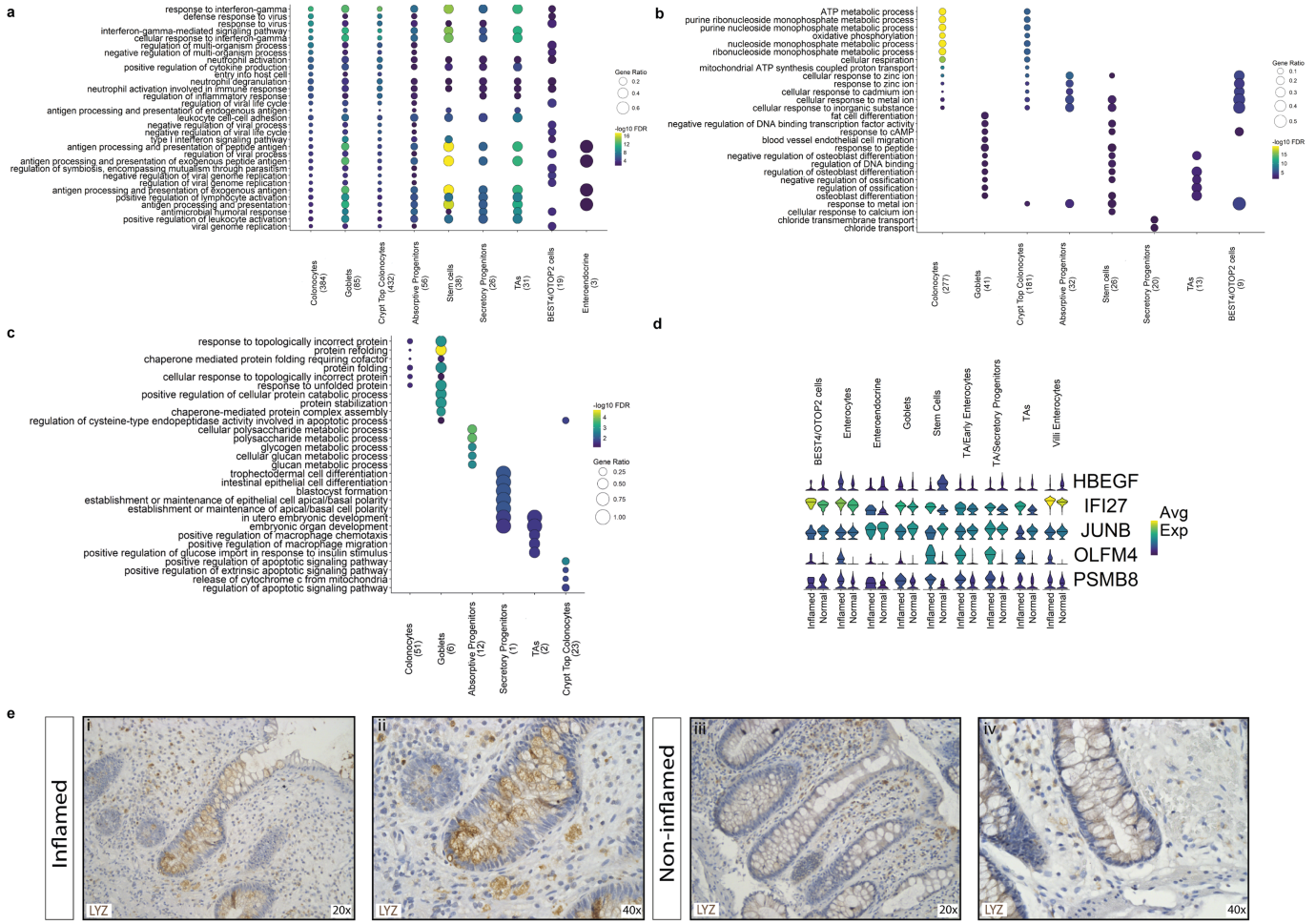


Extended Data Fig. 2 | See next page for caption.

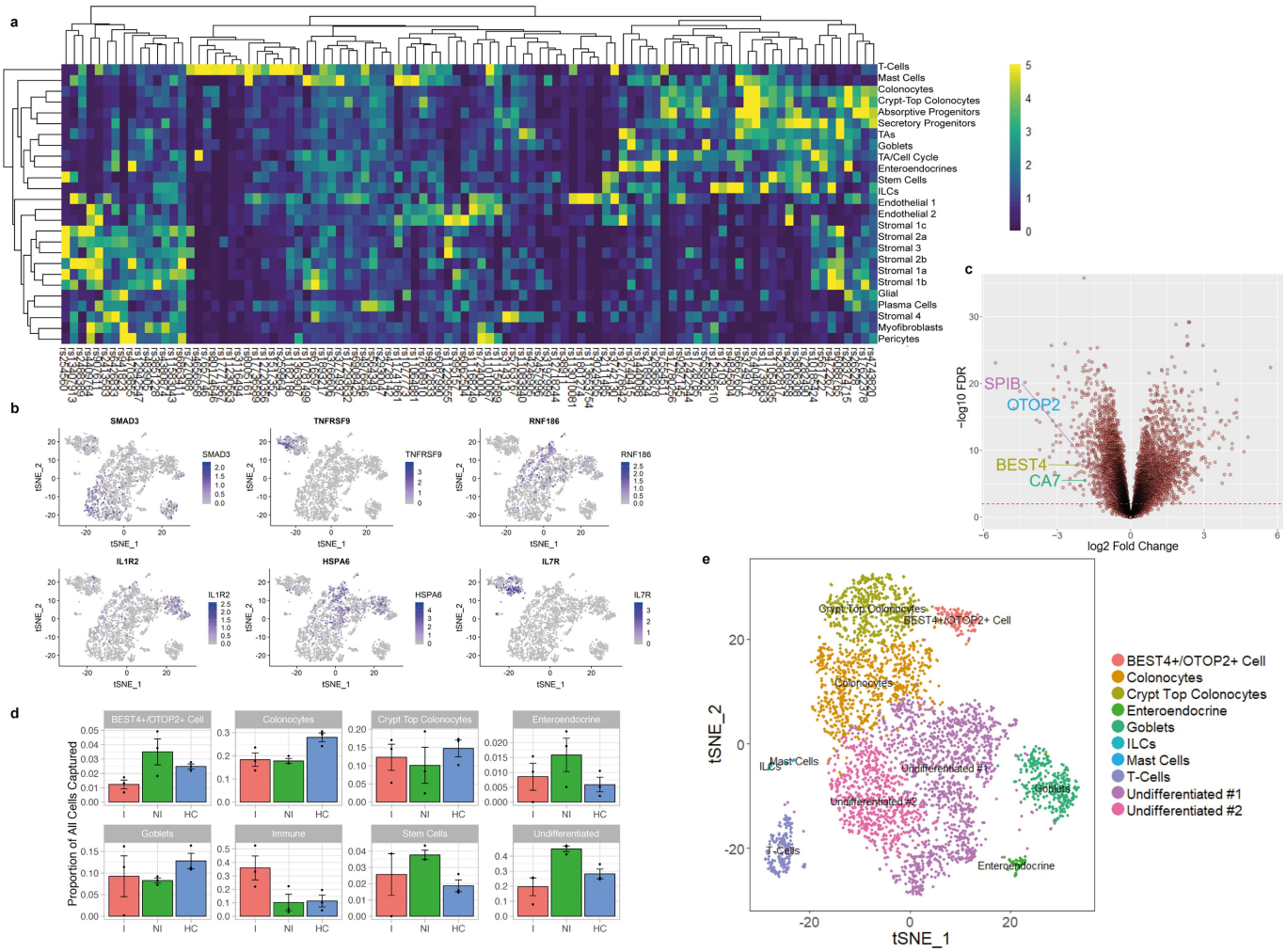
Extended Data Fig. 2 | Validation of the BEST4/OTOP2 cell population. This figure is related to Fig. 2. **a**, Cluster distribution along the differentiation trajectory, captured by Monocle. BEST4/OTOP2 cells are highlighted on the left ($n = 3$). **b**, *t*-SNE gene expression overlay of core BEST4/OTOP2 cell markers ($n = 3$). **c**, **i**–**v**, Representative images ($n = 3$) of colonic sections stained with key BEST4/OTOP2 cell markers by immunohistochemistry to demonstrate BEST4 staining at high magnification (**i**; ×100), cathepsin E (CTSE) staining at low (**ii**; ×40) and high (**iii**) magnification, and additional stains with smISH for SPIB (**iv**) and HES4 (**v**) (each photograph is representative of three samples). **d**, **i**, *t*-SNE visualization of semisupervised clusters of scRNA-seq data identified in a fetal human colon study[15] ($n = 2$). **ii**, Box plot (with 25th, 50th and 75th quantiles) showing co-localized expression of the core BEST4/OTOP2 cell signature. **e**, Heat map showing expression of the core BEST4/OTOP2 cell gene signature in TCGA bulk RNA-seq data in patients with colorectal cancer and matched normal tissue. **f**, **i**, *t*-SNE visualization of semisupervised clustering of scRNA-seq data from matched normal samples from a colorectal cancer study[21] ($n = 10$). Only one BEST4/OTOP2 cell was identified in tumour samples (data not shown). **ii**, Box plot (with 25th, 50th and 75th quantiles) showing localized expression of the core BEST4/OTOP2 cell signature.



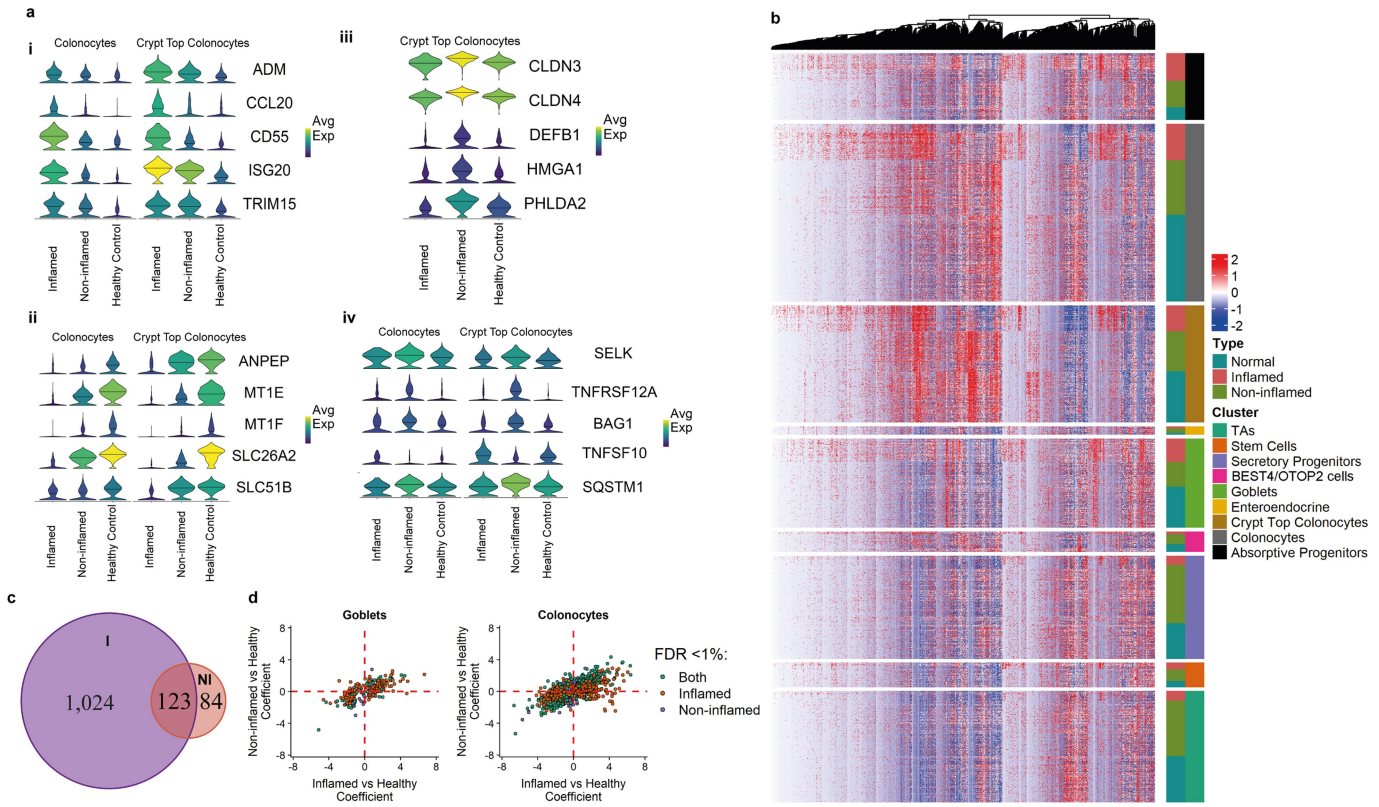
Extended Data Fig. 3 | Isolation and characterization of the BEST4/OTOP2 cell population. This figure is related to Fig. 2. **a**, Flow-cytometry gating strategy for isolating BEST4$^+$ cells. Cells previously gated as live (DAPI$^-$) singlets were selected as EPCAM$^+$CD45$^-$ (**i**) with concurrent staining of a fluorescence minus one (**ii**) to allow placement of a BEST4$^+$ gate on fully stained cells (**iii**). **b**, One hundred EPCAM$^+$BEST4$^+$ and EPCAM$^+$BEST4$^-$ sorted cells ($n = 3$) processed using microfluidic RT–PCR demonstrate increased expression of markers identified from single-cell data relative to GAPDH. Mean ± s.e.m. values are shown. **c**, Circos plot showing overlap between the top 200 BEST4/OTOP2 cell markers detected between 10×, Smart-Seq2, quantitative proteomics and semi-supervised clustering of previously published data[15,21]. **d**, Over-represented Gene Ontology terms in the significantly upregulated protein set in BEST4/OTOP2 cells as identified by quantitative proteomics ($n = 2$ BEST4$^-$ versus $n = 3$ BEST4$^+$; hypergeometric test and FDR-calculated Benjamini–Hochberg multiple testing correction).



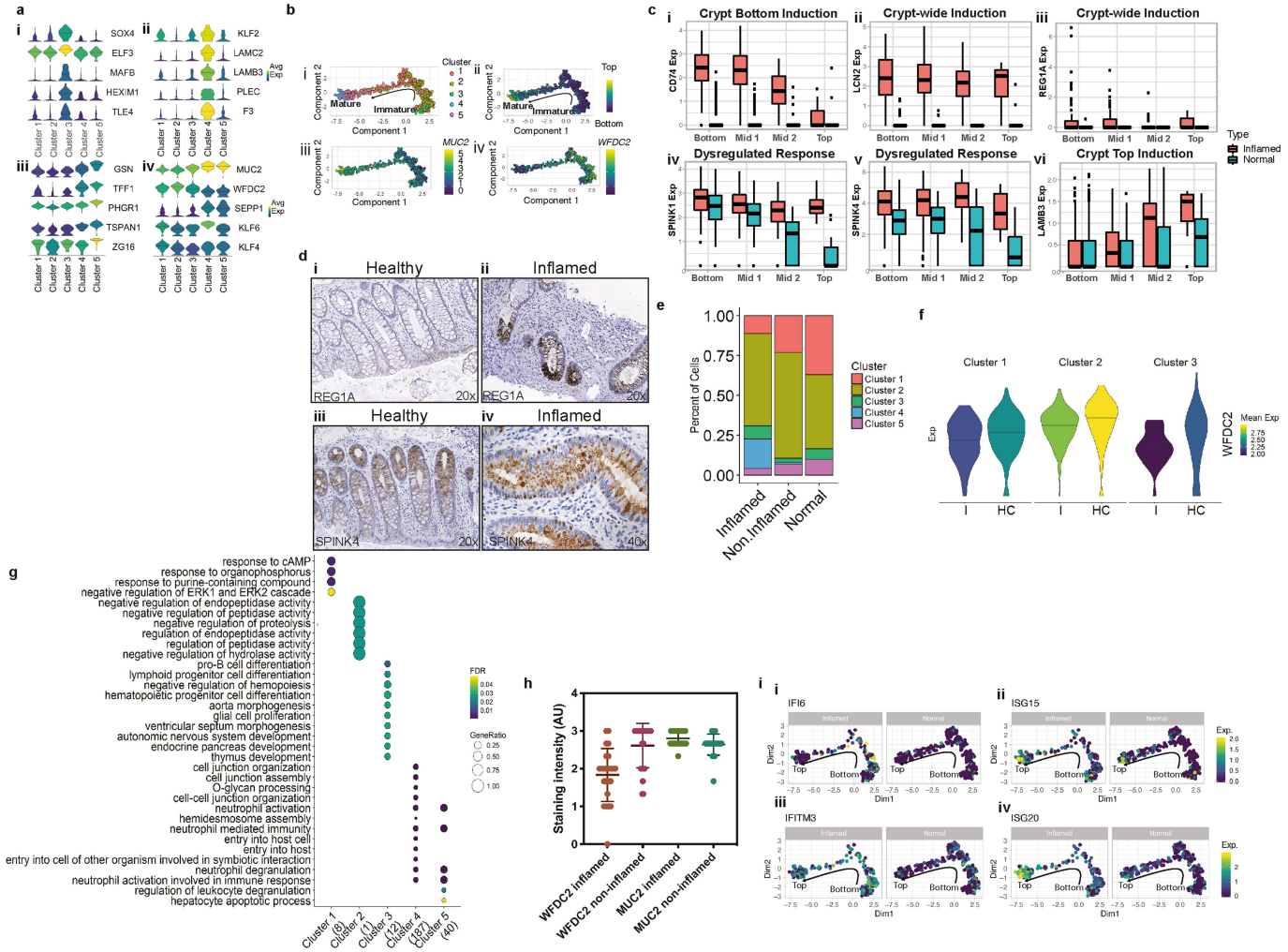
Extended Data Fig. 4 | Gene Ontology enrichment analysis of differentially expressed genes in colonic epithelial-cell clusters. This figure is related to Figs. 3, 4. **a**, Dot plot of Gene Ontology biological-process enrichment in upregulated genes (less than 1% FDR), comparing cell clusters in active colitis and health. **b**, Dot plot of Gene Ontology biological-process enrichment in downregulated genes (less than 1% FDR), comparing cell clusters in active colitis and health. **c**, Dot plot of Gene Ontology biological-process enrichment in differentially expressed genes (less than 1% FDR) in inactive, but not in active, colitis. Points in each dot plot are coloured by enrichment confidence ($-\log_{10}$(FDR)) and sized by the proportion of all genes within the cluster annotated with the Gene Ontology term. For panels **a**–**c**, $n = 3$ per group; hypergeometric test and FDR-calculated Benjamini–Hochberg multiple testing correction. **d**, Violin plots showing expression (*y*-axis) of selected genes that are dysregulated in active colitis ('inflamed') compared with healthy samples ('normal') in stem cells and/or other undifferentiated populations. $n = 3$ per group; centre bar indicates median value; colour indicates mean expression. **e**, **i**–**iv**, Representative immunohistochemistry images ($n = 3$) showing LYZ expression in inflamed and noninflamed colonic tissue sections. **i**, **iii**, ×20 magnification; **ii**, **iv**, ×40 magnification.



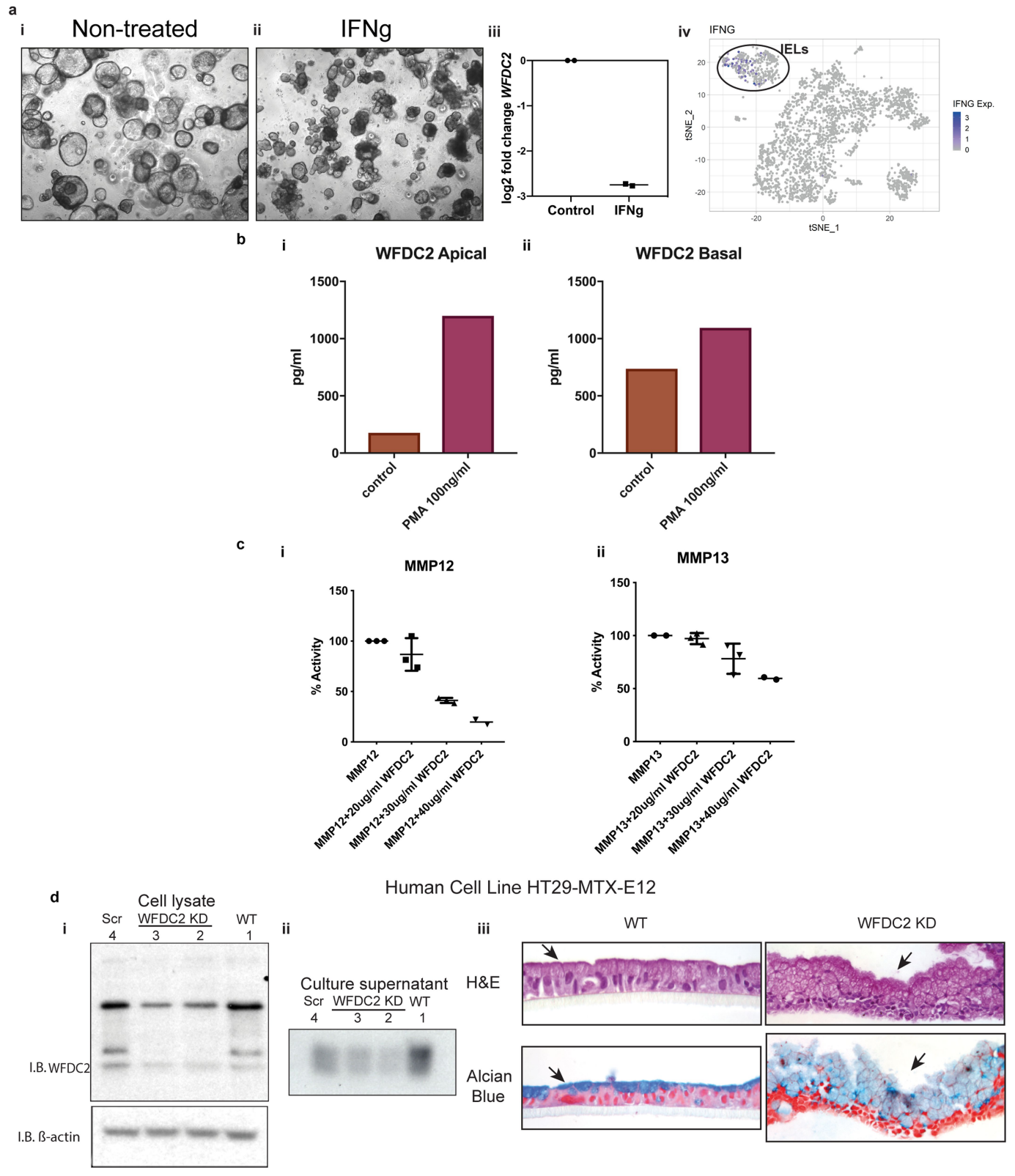
Extended Data Fig. 5 | Human colonic epithelium in clinically noninvolved mucosa and ulcerative-colitis-associated GWAS loci analysis. This figure is related to Fig. 3. **a**, Heat map visualizing the specificity of expression of ulcerative-colitis-associated GWAS loci in immune, epithelial and mesenchymal cell populations. Hierarchical clustering (horizontal) indicates groups of loci with similar expression specificities. **b**, *t*-SNE plots of cells in active colitis ($n = 3$), visualizing selected GWAS ulcerative-colitis-associated gene expression. **c**, Volcano plot showing the differentially expressed genes detected in a microarray study[20], comparing inflamed ulcerative-colitis samples ($n = 74$) with healthy control colon samples ($n = 11$). Significantly downregulated core signature genes from BEST4/OTOP2 cells are highlighted (limma linear model empirical Bayes *P* value and Benjamini–Hochberg multiple testing correction). **d**, Distribution of cluster sizes in healthy colons (HC) and ulcerative-colitis inflamed (I) and noninflamed (NI) samples ($n = 3$ per group), shown as bar charts of proportions of total cells captured. Mean ± s.e.m. values are shown. **e**, *t*-SNE plot of human colonic epithelium single-cell clusters in noninflamed ulcerative colitis ($n = 3$).



Extended Data Fig. 6 | Human colonic epithelium in clinically involved and noninvolved mucosa. This figure is related to Fig. 3. **a**, **i**–**iv**, Violin plots visualizing expression (*y*-axis) of selected differentially expressed genes (less than 1% FDR; two-sided negative binomial likelihood ratio test; Benjamini–Hochberg multiple testing correction) in noninflamed and active ulcerative colitis ($n = 3$). Centre bar indicates median value; colour indicates mean expression. **b**, Heat map visualizing relative expression of all differentially expressed genes (less than 1% FDR; two-sided negative binomial likelihood ratio test; Benjamini–Hochberg multiple testing correction) detected in inflamed (red) and noninflamed (green) colitis compared with healthy tissue (blue) ($n = 3$ per group). **c**, Venn diagram showing the overlap between differentially expressed genes detected in all clusters in clinically inflamed (I) and clinically noninflamed (NI) colitis, compared with healthy tissue. **d**, Comparison between MAST generalized linear model coefficients for significant differentially expressed genes in ulcerative colitis inflamed and noninflamed samples with reference to healthy cells. Correlations for goblet and colonocyte cell clusters are shown ($n = 3$ per group; two-sided Hurdle likelihood ratio test; Benjamini–Hochberg multiple testing correction).

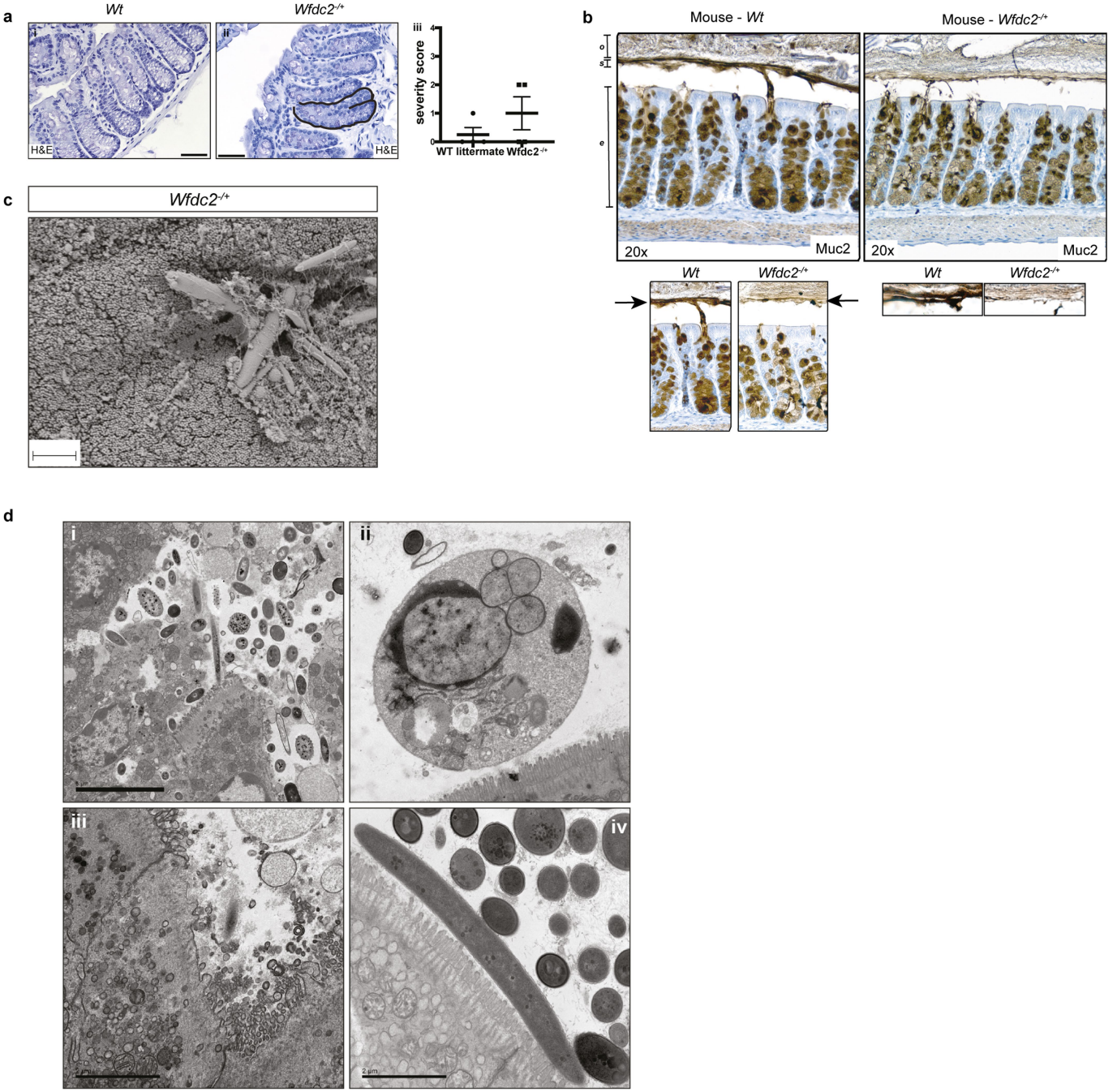


Extended Data Fig. 7 | Goblet-cell remodelling and WFDC2 dysregulation in inflammation. This figure is related to Figs. 4, 5. **a**, **i**–**iv**, Violin plots showing cluster gene expression (*y*-axis) for key marker genes in clusters 1 (**i**), 2 (**ii**) and 3 (**iii**) and common cluster 4 and 5 markers (**iv**) ($n = 3$ per group). Centre bar indicates median value; colour indicates mean expression. **b**, **i**, Pseudotemporal ordering of goblet-cell clusters. **ii**, Crypt-axis score superimposed on trajectory analysis. Cells predicted to reside at the top of the crypt are more mature populations, as inferred by pseudotime ordering, and vice versa. $n = 3$ per group. **iii**, Expression of *MUC2* along the crypt axis. **iv**, Expression of *WFDC2* along the crypt axis. **c**, **i**–**vi**, Gene-expression box plots of selected genes in goblet cells, divided spatially along the crypt axis by binning into four ranges (bottom, mid1, mid2 and top). $n = 3$ per group; 25th, 50th and 75th percentiles shown. Expression of *CD74* (**i**), *LCN2* (**ii**), *REG1A* (**iii**), *SPINK1* (**iv**), *SPINK4* (**v**) and *LAMB3* (**vi**) is shown in health and inflamed ulcerative colitis. **d**, **i**–**iv**, Immunohistochemistry confirms increased expression of REG1A and SPINK4 in inflamed ulcerative-colitis biopsies (**ii**, **iv**) as compared with healthy samples (**i**, **iii**) (representative images of $n = 3$ for each). **e**, Stacked bar chart showing the relative frequency distribution of goblet-cell subclusters (percentage of goblet cells captured) in health and in active (inflamed) and inactive (noninflamed) colitis. **f**, Violin plots showing expression (*y*-axis) of WFDC2 in crypt-bottom goblet-cell clusters in healthy colons (HC) and inflammation (I) ($n = 3$ per group; centre bar indicates median value; colour indicates mean expression). **g**, Comparison of over-represented (hypergeometric test; Benjamini–Hochberg multiple testing correction) Gene Ontology biological-process terms amongst goblet-cell subcluster markers ($n = 3$ per group). **h**, Quantification of WFDC2 and MUC2 expression by immunohistochemistry from patient-matched inflamed and noninflamed sections of 24 patients with ulcerative colitis. Staining intensity was scored from 0 (no staining or weak staining) to 3 (strong staining) by three independent observers. Comparison between WFDC2 inflamed and noninflamed, $P = 0.000148773$; two-sided Wilcoxon matched-pairs signed-rank test, $n = 24$ patients. Comparison between MUC2 inflamed and noninflamed is not significant. Mean ± s.d. shown. **i**, Expression of interferon-induced genes in goblet cells ($n = 3$ per group): *IFI6* (**i**), *ISG15* (**ii**), *IFITM3* (**iii**) and *ISG20* (**iv**).

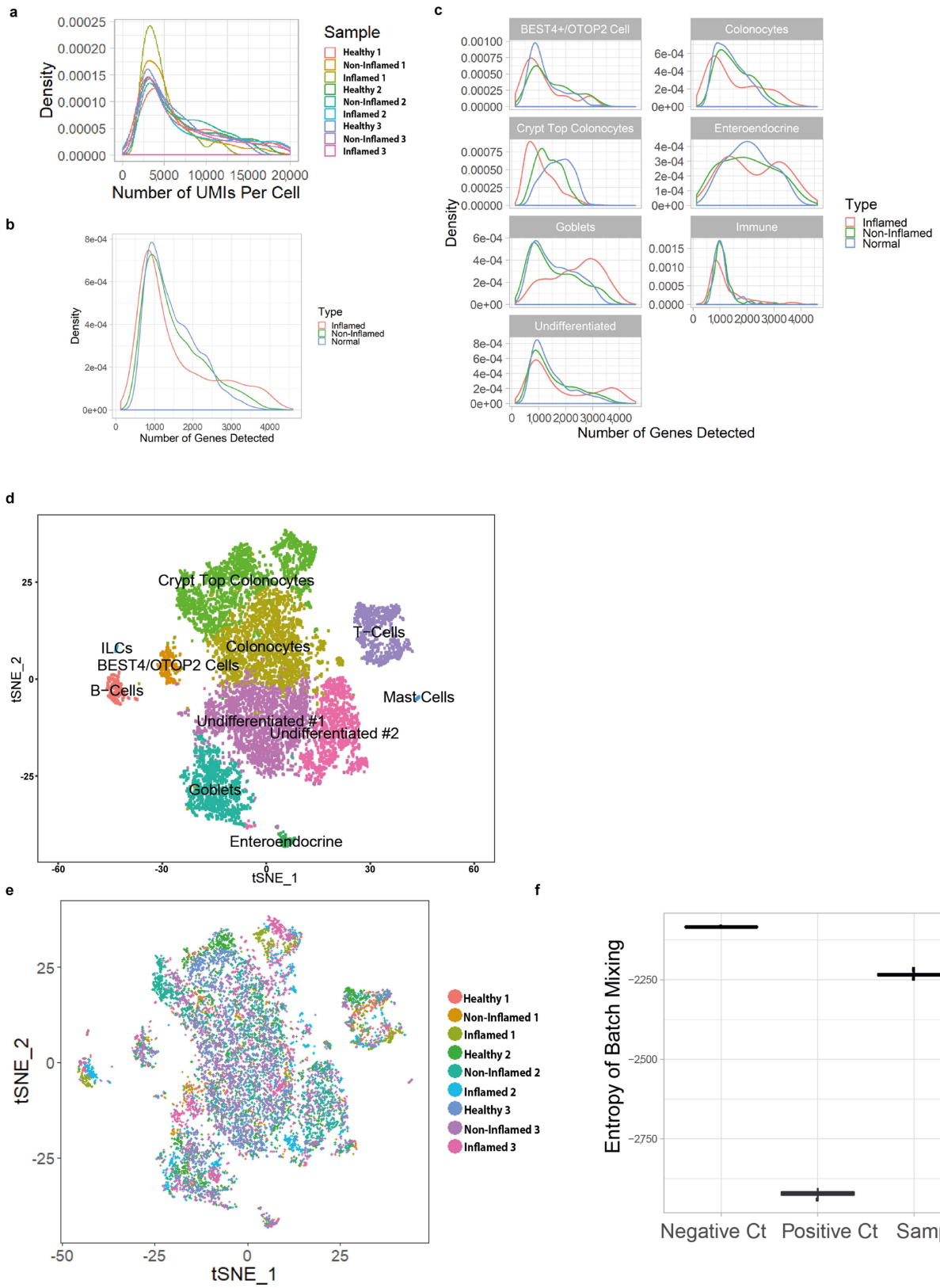


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | In vitro regulation of WFDC2. This figure is related to Figs. 4, 5. **a**, **i**, **ii**, Untreated (**i**) and interferon-γ-treated (**ii**) human colonic organoids in culture. IFNg, interferon-γ. **iii**, qRT–PCR quantification of WFDC2 expression in interferon-γ-treated and untreated organoids ($n = 2$ independent experiments; mean values shown). **iv**, *t*-SNE plot of inflamed epithelium highlighting localized expression of interferon-γ in intra-epithelial lymphocytes (IELs; $n = 3$). **b**, **i**, **ii**, Quantification by enzyme-linked immunosorbent assay of WFDC2 secretion into apical (**i**) or basal (**ii**) medium of HT29-MTX-E12 cells with and without 100 ng ml^{-1} of PMA stimulation for 6 h ($n = 1$). **c**, **i**, **ii**, MMP12 (**i**) and MMP13 (**ii**) activity measured in the absence and presence of various concentrations of WFDC2. Data are presented as percentage of activity remaining. $n = 3$, except for MMP12 + 40 μg ml^{-1} WFDC2 and untreated MMP13, where $n = 2$. Mean ± s.d. shown. **d**, **i**–**iii**, WFDC2 knockdown (KD) in HT29-MTX-E12 cell lines ($n = 2$). **i**, Immunoblot of WFDC2 on cell lysates from nontransfected (lane 1, wild type, WT), *WFDC2* shRNA transfected (lane 2, clone 1; lane 3, clone 2) and scrambled transfected (lane 4; Scr) cells. β-Actin was used as a loading control. I.B., immunoblot. **ii**, Cell-culture supernatants were tested by immunoblotting for secreted WFDC2. **iii**, Cells grown on transwells were stained with haematoxylin and eosin and Alcian blue. Arrows indicate the attached mucus layer and mucin-secreting goblet cells.



Extended Data Fig. 9 | WFDC2 influences barrier function. This figure is related to Fig. 5. **a**, **i**, **ii**, Histopathological evaluation of changes in epithelial-cell morphology and mucosal architecture in wild-type (WT; **i**) and *Wfdc2*$^{+/-}$ (**ii**) mice shows bifurcation at the base of the crypt in the *Wfdc2*$^{+/-}$ mice. **iii**, Mice were assigned a subjective colitis severity score on the basis of a modification of previously published criteria[50]. Scores for morphology, ulceration and infiltration were ranked on a scale from 0 (normal or absent) to 4 (severe), which were summed to give an overall score. **b**, Colonic tissue from *Wfdc2*$^{+/-}$ mice and wild-type littermates was processed to preserve the mucus layers. Immunohistochemistry for MUC2 in the distal mouse colon reveals mucus-filled goblet cells in the epithelium (e) and secreted mucus. The secreted mucus forms two layers: a stratified inner layer (i) and an outer layer (o). Arrows indicate the inner mucus layer. Higher-magnification images are shown in the bottom panels. $n = 4$. **c**, SEM of the colonic surface shows bacteria invading goblet cells in *Wfdc2*$^{+/-}$ mice. Scale bars, 2 μm. **d**, **i**–**iv**, TEM images of colons of *Wfdc2*$^{+/-}$ mice show epithelial-cell damage with destruction of microvilli (**i**), epithelial detachment (**ii**) and destruction (**iii**), as well as bacterial aggregates observed over the surface of *Wfdc2*$^{+/-}$ mice (**iv**). **b**–**d** show representative images; $n = 4$ animals per group.



Extended Data Fig. 10 | Integrated sample analysis and batch distribution. This figure relates to the Methods. **a**, Density distribution of cell UMI counts per sample. **b**, Density distribution of cellular gene-detection rate per condition. **c**, Density distribution of cellular gene-detection rate per condition per cell-type cluster. **d**, *t*-SNE visualization showing integrated clustering analysis of samples across all conditions ($n = 3$ per group). **e**, *t*-SNE visualization of sample batch distribution in the integrated clustering analysis ($n = 3$ per group). **f**, Box plots showing entropy of batch mixing for sample batches ($n = 9$; right); positive controls (Ct), in which clusters were assigned as batches (centre); and negative controls, in which cells were assigned random batch labels in accordance with batch size distribution (left). The entropy of batch mixing for sample batches approaches that of the negative control. Bars show the 25th, 50th and 75th percentiles.

nature research

Corresponding author(s): Prof. Alison Simmons

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	See below (Data Analysis) for full description of collection software as part of data analysis. Software for collection included; Imaging - Visiopharm (Visiopharm, Denmark), Infinity Analyze (lumenera inc, version 6.5); Flow cytometry - Attune NxT software (ThermoFisher); Quantitative RT-PCR Quantstudio software (Thermofisher, version 1.3). Flow sorting - SonySH800 cell sorter software (Sony Biotechnology).
Data analysis	For specified biological experiments data was analysed and presented using Prism (Graphpad software Inc. version 7). For analysis and representation of flow cytometry data was collected using Attune NxT software (Thermofisher) and exported as FCS files for further data analysis using FlowJo (FlowJo Tm, Version 10) with graphical representation of results then presented using Prism (Graphpad). Imaging of immunohistochemistry and histology slides from patient samples was performed using Infinity Analyze software (Lumenera inc, version 6.5) on a Olympus BX60 microscope and Infinity 3s Lumenera Camera. Slides stained for WFDC2 as previously described were scanned using Leica ScanScope machine (Leica Biosystems). To quantify staining we used Visiopharm software (Visiopharm, Denmark) Quantitative RT-PCR was performed on a quantstudio7 (Thermofisher, version 1.3) machine with initial quantification of results using Quantstudio Software (Thermofisher) with raw data then presented using Prism (Graphpad) as previously described. The following software were used to analyse sequencing data: FastQC Version 0.11.7 - Raw sequence data quality control software

CellRanger 2.1.1 - raw 10x single cell data processing, alignment and barcode UMI counting.
SNPsea software, Version 1.0.3 - Over-representation testing of tissue-specific expression of GWAS loci genes.

R Version 3.4.3 - Language for statistical analysis. The following packages were used in addition to base installation:

Seurat Bioconductor R Package, Version 2.3.2 - single cell data QC, normalisation, clustering and differential gene expression analysis.
Monocle Bioconductor R Package, Version 2.8.0 - single cell data trajectory analysis.
Cyclone, scran Bioconductor R Package, Version 1.6.9 - cell cycle score annotation.
M3Drop, R Package, Version 1.6.0 - Highly variable gene selection for single cell dimensionality reduction.
irlba, R Package, Version 2.3.2 - Principal component analysis.
clusterProfiler, Bioconductor R Package, Version 3.6.0 - Gene Ontology enrichment analysis.
ggplot2, R Package, Version 2.2.1 - Plotting package.
DESeq2, Bioconductor R Package, Version 1.20.0 - Differential expression analysis of bulk /pseudobulk RNA sequencing data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data generated in this study are available on Gene Expression Omnibus, accession number GSE116222. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE 101 partner repository with the dataset identifiers PXD011655 and 10.6019/PXD011655

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculation was performed for patient samples which underwent single cell RNA sequencing in the first instance. 3 samples from healthy, non-inflamed UC and inflamed UC was chosen for single cell sequencing to allow for patient variability. For valdiation experiments of gene and protein targets a sample size was not calculated but biological and technical repeats were performed to allow quantification of statistical differences where nescessary with frequency reported for each individual experiment.
Data exclusions	No data exclusion of patient data was performed.
Replication	All single cell experiments, immunohistochemistry stains, in-situ hybridisation stains and flow cytometry quantification experiments were performed on multiple occasions, most frequently >3, in order to ensure reliability and reproducibility of results. Exact quantification for each replicate is described in methods and figure legends.
Randomization	No randomization was performed
Blinding	Interpretation of WFDC2 staining intensity from patient samples was performed by 3 individuals who were blinded as to patient severity score, which was then further quantified using imaging software, as per the description in the methods. In other areas of results blinding was not necessary.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Full details of antibody concentrations and conditions are supplied in a supplementary methods table. These included;
Lyz - Agilent (Dako)- A0099 (lot 20025086) (EC 3.2.1.17);
WFDC2 - Novus Biologicals - Novus NBP2-46360 (clone OTI1E12) (lot W001) ;
MUC2 - Agilent (Dako) - M7313 (Clone CCP58)(lot 20043026);
REG4 - Atlas antibodies - HPA046555 (lot R44544);
CLCA1 - Atlas Antibodies - HPA059301 (lot R82805);
BCAS1 - Atlas Antibodies - HPA051816 (lot R67167);
PCSK1N - Atlas Antibodies - HPA003925 (lot B106794);
CPE - R&D systems - MAB3587 (clone 420709) (lot ZOG021804A)
ChGA - Abcam - b15160 (lot GR244279-2);
CTSE - Atlas Antibodies - HPA012940 (lot R03471),
EpCAM - PeVio770(TM) / PE/Cy7 - Milteyni - 130-099-742 (lot 5171115435);
CD45 APC - Milteyni - 130-098-143 (lot 5171115415);
CD90 FITC - Biolegend - 328107 (clone 5E10) (lot B241989);
DAPI - BD biosciences - BD564907 (lot 7089926)
REG4 (IF) - R&D systems - AF1379-SP (lot IES031802C)
MUC2 (IF) - ThermoFisher Scientific - MA5-12345 (lot TF258930)
WFDC2 (IF) - Abcam - ab24480 (lot GR298303-5)
Bestrophin 4 (IHC and FACS) - Atlas - HPA058564 (Lot R81693)
Alexa Fluor 488 goat anti-rabbit IgG - Invitrogen A11034 (lot 1971948)
SPINK4 (IHC) rabbit - Abcam- ab183347(lotA39120)
REG1A (IHC) rat - R&D systems -Mab49371 (lot cgbr011810A)

Validation

Lyz - IHC - human
- validation reference on supplier insert - Krugliak L, Meyer PR, Taylor CR. The distribution of lysozyme, alpha-1 antitrypsin, and alpha-1-antichymotrypsin in normal hematopoietic cells and in myeloid leukemias. Am J Hematol 1986;21:99-109.
WFDC2 - IHC - human
- there are no published references for this antibody, however the supplier website has validated it for IHC - please check link below - https://www.novusbio.com/products/he4-wfdc2-antibody-oti1e12_nbp2-46360#datasheet. This antibody has also been validated by IHC in the current manuscript and a titration of various concentrations has been carried out to obtain an ideal concentration of 1:250.
MUC2 - IHC - human
- This antibody has been extensively characterised and is routinely used in pathology. The company datasheet offers a wide range of performance characteristics - see https://www.chem.agilent.com/cs/library/packageinsert/public/P02334EFG_02.pdf. Additionally there are also numerous citations mentioned , please see datasheet for complete list.
REG4 - IHC - human
- This antibody has also been extensively validated for IHC by the company. Please see https://atlasantibodies.com/products/REG4-antibody-HPA046555. This antibody has also been used by the Human Protein Atlas (https://www.proteinatlas.org/ENSG00000134193-REG4/tissue).
CLCA1 - IHC -human
- This antibody has been validated by IHC on small intestinal and liver tissue by the manufacturer - Atlas antibodies. Additionally, the validation data is in accordance with their RNA seq data. Please see datasheet - https://atlasantibodies.com/products/CLCA1-antibody-HPA059301. Also, this antibody has been validated by the Human Protein Atlas (https://www.proteinatlas.org/ENSG00000016490-CLCA1/tissue).
BCAS1 - IHC - human
- validated extensively for IHC on human tissue sections by Atlas Antibodies - please see datasheet - https://atlasantibodies.com/products/BCAS1-antibody-HPA051816/references. This antibody has also been validated as part of the efforts of the Human Protein Atlas to map the human proteome using antibodies (https://www.proteinatlas.org/ENSG00000064787-BCAS1/tissue).
PCSK1N - IHC - Human
Extensively validated by supplier (https://www.sigmaaldrich.com/catalog/product/sigma/hpa003925?lang=en&region=GB) and as part of the Human Protein Atlas (https://www.proteinatlas.org/ENSG00000102109-PCSK1N/tissue)
CPE - IHC - Human
Validated by supplier (https://www.rndsystems.com/products/human-carboxypeptidase-e-cpe-antibody-420709_mab3587) and as part of the human protein atlas (https://www.proteinatlas.org/ENSG00000109472-CPE/antibody)
ChGA - IHC - Human
Validated by supplier with 70 references in literature (https://www.abcam.com/chromogranin-a-antibody-ab15160.html)
CTSE - IHC - Human

Validated by supplier with 5 references in literature to date (https://www.sigmaaldrich.com/catalog/product/sigma/hpa012940?lang=en&region=GB) also extensively validated as part of the human protein atlas (https://www.proteinatlas.org/ENSG00000196188-CTSE/tissue)
EpCAM - PeVio770 (TM) / PE/Cy7 - Flow cytometry - Human
Validated on product datasheet with example stains - datasheet available from https://www.miltenyibiotec.com/GB-en/products/macs-flow-cytometry/antibodies/primary-antibodies/cd326-epcam-antibodies-human-hea-125-1-11.html/#pe-vio770:for-100-tests
CD45 APC - Flow Cytometry - Human
Validated on product datasheet with example stains in human samples - datasheet available from https://www.miltenyibiotec.com/GB-en/products/macs-flow-cytometry/antibodies/primary-antibodies/cd45-antibodies-human-5b1-1-50.html#apc:for-100-tests
CD90 FITC - Flow cytometry - Human
Valdiated both internally on product datasheet and in 9 publications - details of both available from https://www.biolegend.com/it-it/products/fitc-anti-human-cd90-thy1-antibody-4113
DAPI - Flow cytometry - human
Validated both by supplier with internal validation on datasheet and with 4 publications - data available from http://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/immunology-reagents/immunology-buffers-and-ancillary-reagents/dapi-solution/p/564907
REG4 (R&D systems) - IF - Human
Validated by supplier with 5 references in literature (https://www.rndsystems.com/products/human-reg4-antibody_af1379). Also extensively validated as part of the human protein atlas (https://www.proteinatlas.org/ENSG00000134193-REG4/antibody)
MUC2 (ThermoFisher Scientific) - IF - Human
Validated by supplier with 4 references in literature (https://www.thermofisher.com/antibody/product/MUC2-Antibody-clone-996-1-Monoclonal/MA5-12345)
WFDC2 (Abcam) - IF - Human
Validated by the supplier (https://www.abcam.com/he4-antibody-ab24480.html) and 7 citations have been reported for it on CiteAb (https://www.citeab.com/antibodies/763161-ab24480-anti-he4-antibody?des=978F621CB0C917CB).
Bestrophin 4 (IHC and FACS) - Atlas - HPA058564 (Lot R81693) validated by supplier (https://atlasantibodies.com/products/BEST4-antibody-HPA058564)
Alexa Fluor 488 goat anti-rabbit IgG - FACS- Invitrogen A11034 (lot 1971948)
519 peer reviewed publications for use in FACS and IF - https://www.thermofisher.com/antibody/product/Goat-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-11034
SPINK4 (IHC) rabbit - Abcam- ab183347(lotA39120)
Validated for use in IHC on manufacturers website (https://www.abcam.com/spink1-antibody-sp166-ab183347.html)
REG1A (IHC) rat - R&D systems -Mab49371 (lot cgbr011810A)
Validated for use in IHC on manufacturers website (https://www.rndsystems.com/products/human-reg1a-antibody-431211_mab49371)

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ECACC General Cell Collection: HT29-MTX-E12. Cat. no. 12040401
Authentication	European Collection of Authenticated Cell Cultures (ECACC)
Mycoplasma contamination	No contamination observed. The cell line has been routinely tested for mycoplasma contamination in the author's laboratory and found to be negative.
Commonly misidentified lines (See ICLAC register)	n/a

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mouse - C57BL/6 - equal mix of males and females for Wfdc2-/+ mouse phenotyping and males for DSS colitis challenge experiments - range between 10-12 weeks of age.
Wild animals	n/a
Field-collected samples	n/a

Human research participants

Policy information about studies involving human research participants

Population characteristics	For single cell analysis and other techniques described, following informed consent biopsies were collected from volunteers attending endoscopy for routine colonoscopic screening (healthy) or as part of ongoing clinical care (IBD patients) (See supplementary Table 1 for demographics). For UC, we utilized tissue derived from immunotherapy naïve patients with a proven histological diagnosis. Tissue was sampled from clinically inflamed distal colon and proximal clinically non-involved regions. Ethical approvals: (Yorkshire and Humber REC, reference:16/YH/0247) and (West Midlands REC, reference: 09/H1204/30).

	Full details of patient demographics including age,gender and disease severity are presented in supplementary table 1. For validation experiments including FACS, proteomics, smart-seq2 and immunohistochemistry / smISH quantification patient samples were collected in a similar manner under the same ethical approvals with healthy controls undergoing routine colonoscopy and samples form patients with ulcerative colitis being from active or inactive inflammation but samples were stored in formalin for further processing as per methods section.
Recruitment	Recruitment was as described above. Measures to avoid bias included balancing of patient genders for single cell analysis and severity blinding in IHC quantification as previously described within the methodology of each experiment.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For flow cytometry quantification samples were processed identically to those methods described in methods section. Following this they were then washed with FACs buffer (50ml PBS with 2% Fetal Calf Serum and 0.25mM EDTA) and were then stained with appropriate antibodies for 30 minutes. Samples were then washed and processed directly as described without fixation. Prior to running samples compensations were calculated with an unstained cellular control, compensation beads (BD) and a stained DAPI control which had been placed on a heat block for 10 minutes prior to antibody addition. This process was performed both for the flow cytometry quantification experiments and sorting experiments.
Instrument	For sample collection without sorting (quantification of populations) FACS was performed on the Attune NxT. For sorting experiment to quantify gene expression in EpCAM+CD45- population (Supplementary figure 1b) sorting was performed on a Facs Aria IIIu. For sorting experiments to validate the BEST4/OTOP2 populations (smartseq2, proteomic, microfluidic RT-PCR) sorting was performed on a SonySH800 (Sony)
Software	As per instrument description samples run on the Attune NxT had data collected with Attune NxT Software v 2.7.0 and for sorting experiments it was collected on the FacsAria IIIu. In both instances FCS files were then exported for further quantitiative analysis using FlowJo (v.10) as described previously.
Cell population abundance	For quantification and sorting experiments epithelial cells were identified by gating on FSC and SSC and then by excluding doublets. In the case of population quantification (supplementary Figure 1a) this was then used as the denominator for calcuating percentage positive of each of the defined populations (Epcam+ / CD45+ /CD90+ and viability = DAPI-) as described in figure legends, with percentages and axis described. Population abundance was calculated using histogram plots of count vs appropriate channel as shown in figure. For sorting experiments the identification of epithelial cells in an identical manner with EpCAM+CD45- cells then sorted as shown in Figure Supplementary 1b. For flow sorting of bestrophin4+ cells an example of gating strategy is given in Extended Figure 2g which demonstrates the placement of a gate for the positive population with a FMO control. As described in methods for each sorting experiment an unstained, fluorescence minus on and secondary control sample was used in order to place the EpCAM+BEST4+/- gates.
Gating strategy	Gating strategies are described above, in figure legends and examples are given in relevant figures as mentioned in "Cell population abundance"

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.