Colonic Epithelial Cell Diversity in Health and IBD

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

[1] *MRC Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK and Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK.*

[2] *MRC WIMM Centre For Computational Biology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK.*

[3] *Wolfson Imaging Centre Oxford, MRC Weatherall Institute of Molecular Medicine, Oxford, UK.*

[4] *Nuffield Department of Surgical Sciences and Oxford NIHR Biomedical Research Centre (BRC), University of Oxford, John Radcliffe Hospital, Oxford, UK.*

[5] *MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK.*

[6] *Sir William Dunn School of Pathology, South Parks Road, Oxford, UK.*

[7] *Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK.*

[8] *Centre for Pathology, St Mary's Hospital, Imperial College, London, UK.*

**These authors contributed equally.*

[§] *Correspondence: alison.simmons@imm.ox.ac.uk*

The colonic epithelium facilitates host-microbe interactions to control mucosal immunity, coordinate recycling and form the mucus barrier. Epithelial barrier breakdown underpins inflammatory bowel disease (IBD). However, we do not know the specific contributions of each epithelial cell subtype to this process. Here, we profiled single colonic epithelial cells in health and IBD. We identified previously unknown subtypes, including crypt gradients of progenitors, colonocytes and goblet cells. We discovered a novel crypt top absorptive cell, expressing OTOP2 and the satiety peptide Uroguanylin, that senses pH and shows dysregulation in inflammation and cancer. In IBD, we observed positional remodeling of goblet cells coinciding with downregulation of WFDC2, a new goblet cell expressed anti-protease that inhibited bacterial growth. *In vivo,* WFDC2 preserved tight junction integrity, prevented commensal invasion and mucosal inflammation. We delineate markers and transcriptional states, identify a new 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 is comprised of specialized cells with diverse functions which emerge from stem cells at the crypt base. The majority of epithelial cells are absorptive colonocytes interspersed with specialized epithelial lineages, including secretory goblet and enteroendocrine cells[1]. Whether other epithelial cell types exist in the human colon remains unclear.

Inflammatory bowel disease (IBD), consisting of Ulcerative Colitis (UC) and Crohn's Disease (CD), results from a breakdown of the symbiotic relationship between the intestinal commensal microflora and mucosal immune system. Barrier defects characterize both forms of IBD, with goblet cells reportedly depleted in UC and increased in CD[2]. Key examples exist where disruption of innate epithelial pathways drives colitis, including defects in autophagy[3], ER stress[4], lipid antigen presentation[5] and inflammasome dysfunction[6].

Goblet cells are critical for the maintenance of the colonic barrier, both through production of mucus, and transportation and presentation of luminal antigens to tolerogenic dendritic cells (DCs), 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, Paneth cell secretion of AMPs mediates this sterility[9]. However, the colon contains few or no Paneth cells, so the cell types directing the release of colonic AMPs remain uncharacterized.

It is also unclear whether specific subsets of colonic epithelia show epithelial cell intrinsic molecular pathology in IBD colitis. To study this, we used single cell profiling to create a map of colonic epithelia in health and clinically inflamed and non-inflamed mucosa in UC. This identifies a new absorptive cell with a role in pH balance and goblet cell drivers of barrier breakdown.

Crypt gradients of absorptive and secretory cells

We isolated colonic biopsies from healthy volunteers or immunomodulator-naïve UC patients sampled from clinically inflamed and non-inflamed mucosa (**Supplementary Table 1**). Crypts were dissociated to single-cell suspensions and processed using drop-seq technology, capturing 11,175 cells (**Methods** and **Extended data Fig. 1a-b**).

In healthy colon, 10 clusters of cells were identified and visualized using t-distributed stochastic neighbourhood embedding (t-SNE) (**Fig. 1a**), including undifferentiated cells, absorptive colonocytes and distinct clusters of goblets (GC) and enteroendocrine cells (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 sub-clusters marking 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 and 1g**). No Paneth cells were detected in the healthy colon (**Extended data Fig. 1b**).

Rather than discrete clusters, we observed gene expression gradients in epithelial cells, consistent with an ascending crypt axis of differentiation. We *in-silico* localised cells within the colonic crypt by defining a crypt-axis score (**Methods**) using 15 genes expressed in both absorptive and secretory cells (**Fig. 1c**)**.** Pseudo-time 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 play roles in lineage commitment during differentiation (**Extended data Fig. 1i**). Therefore, scRNA-seq highlights the extent of human colonic heterogeneity and supports existence of a differentiation hierarchy in the crypt.

Discovery of a novel pH sensing absorptive colonocyte

Clustering analysis identified a novel cell cluster (**Fig. 1a**), predicted to transport salt, ions and metals **(Extended data Fig. 1j**). They expressed mature colonocyte markers with distinct expression of BEST4 (**Fig. 2a(i and iii)**, **Extended data Fig. 2c(i)**), Cathepsin E (**Extended data Fig. 2c(ii-iii)**)*,* and *OTOP2* (**Fig. 2a(ii-iii)**)*,* so we designated these cells "BEST4/OTOP2 cells". BEST4 hallmarks epithelia involved in electrolyte transportation[11], while the *OTOP* family encodes proton conducting ion channels in various epithelia[12]. They also expressed the endogenous paracrine hormone and satiety peptide Uroguanylin[13] (*GUCA2B*) (**Fig. 2a**(**iv-vi**)) that is required for guanylate cyclase 2C(GC-C) activation and epithelial cGMP activity. Further, they expressed genes belonging to the metallothionein family that impart defense 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 expressed 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 semi-supervised 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 5c**).

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 mRNA and protein markers (**Extended data Fig.3c** and **Supplementary Data**). Pathway analysis showed enriched 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 extracellular pH (pH_0), seen here as an increase in emission of a membrane permeant pH indicator dye, pHrodo Red (**Fig. 2d(i-ii)**)**,** which corresponded to a significant change in the intracellular pH (pH_i) (**Fig. 2d(iii)**). Intracellular acidification can be cytotoxic; however, our proteomics data indicated that BEST4/OTOP2 cells express high levels of 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 these cells play a role in setting colonic epithelial cGMP tone in response to luminal pH.

Universal and cell-type specific responses in UC

We next sampled clinically inflamed and non-inflamed tissue from early diagnosis immunotherapy-naïve UC patients (**Methods**, **Supplementary Table 1**). In addition to the previously identified clusters in health, two additional clusters representing inflammation-associated GCs and intraepithelial immune cells were detected (**Fig. 3a**). We also observed shifts in relative cell proportions **(Extended data Fig. 5d**).

Differential gene expression analysis between corresponding cell clusters revealed 1,147 genes (<1% FDR) dysregulated in inflamed UC, with the greatest number of changes in colonocytes (734) and crypt-top colonocytes (676), followed by GCs (140), stem cells (65), BEST4/OTOP2 cells (28) and EECs (4) (**Supplementary Data and Fig. 3b**). We observed universal upregulation of several inflammatory pathways across most cell populations, including interferon gamma signaling, 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 metabolism processes and simultaneously induced genes necessary for reactive oxygen species production and microbial killing (e.g. *SAA1, DMBT1, PLA2G2A*). The BEST4/OTOP2 cell population showed reduced expression of the metallothionein family and other ion absorption genes (**Extended data Fig. 4b** and **Fig.3b**). GCs upregulated stress response genes that actively promoted cell survival over apoptosis (**Supplementary Data**). *LYZ*, a Paneth cell gene, was upregulated by lower crypt GCs in inflammation (**Fig. 3b and Extended data Fig. 4e(i-iv)**) and may mark the “deep crypt secretory cells” of the colon[22] required to maintain the colonic stem cell niche and protect them 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. In contrast, stem cells in inflammation showed downregulated heparin-binding 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 UC may impact 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 gene expression 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 suggested IBD susceptibility genes are expressed differently in unique cell populations **(Extended data Fig. 5a**).

We used the SNPsea[24] algorithm to test UC-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* [25,28](**Extended data Fig. 5b**).

In contrast, we observed a highly significant (FDR < 1%) association for some immune subsets and absorptive progenitor epithelial cell types using the same approach in inflamed UC samples (**Fig. 3c**), with significant associations at alternative cut-offs (FDR < 5%, 10%) in other immune and epithelial subsets (**Fig. 3c**). Inflamed crypt-top colonocytes differentially expressed oxidative stress pathway genes *NOS2* (**Fig. 3d**) and *DDAH2*[29], elevated *SMAD3* (**Extended data Fig. 5b**) and *JAK2*, which are associated with both Crohn's and UC[25,30]. IBD-associated *ITLN1* (**Fig. 3d**) and *IL1R2*[30] were expressed by GCs, while undifferentiated epithelia expressed IBD-associated *RNF186*; chemokines *CXCL1, CXCL2* and *CXCL3*[31]; integrin *ITGB8*[25] and *HSPA6*[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 non-involved UC epithelia

Differential expression analysis (**Extended data Figures 5e and 6(a-b)**) of non-involved UC vs healthy mucosa identified 207 significantly dysregulated genes (<1% FDR). Remarkably, 59.4% (123 out of 207) of differentially expressed genes (DEGs) in non-involved UC epithelia were also detected as differentially expressed in involved tissue (**Extended data Fig. 6c)**; however, gene expression changes limited to only UC non-inflamed tissue were also observed (**Extended data Fig. 6a(iii-iv) and Extended data Fig. 4c**). Furthermore, we fit a generalized linear model to all the data and found that model coefficients for inflamed and non-inflamed samples were correlated, but with smaller effect sizes in non-inflamed cells (**Extended data Fig. 6d).** This suggests that clinically non-inflamed mucosa bears similar transcriptomic hallmarks to inflamed tissue, indicating recovery from inflammation and damage repair; or arising as a protective mechanism in anticipation of damage.

Goblet cell heterogeneity in health and in UC

Although dysregulated GC function contributes to barrier breakdown in colitis, we do not yet know the pathways underlying this breakdown. GC single-cell profiles derived from healthy, inflamed or non-inflamed UC tissue in isolation revealed partitioning of this cell group across 5 clusters (**Fig. 4a**). We used crypt-top/base-axis score and unsupervised pseudo-time trajectory analysis to infer 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**,** while Cluster 5 localized towards the lumen-facing crypt-top (**Extended data Fig 7a-b(i)**).

In UC, we observed both spatial and crypt-wide differences at mRNA and protein levels. This suggested that while common inflammatory responses exist, GCs in spatially distinct regions within the crypt also exhibit highly heterogeneous changes. For instance, *LCN2* and *REG1A* are induced throughout the crypt, while *CD74* and *LAMB3* induction was limited to crypt bottom and top, respectively (**Extended data Figs. 7c**). We also observed transcriptional dysregulation, where genes normally limited to the crypt bottom in health persist in crypt-top cells in inflammation (e.g. *SPINK4* and *SPINK1*) (**Extended data Fig. 7c (iv-v) and d(iii-iv))**.

In line with these observations, we identified the emergence of a disease-associated cluster of GCs in Cluster 4 (**Extended data Fig. 7e)**, a counterpart with homology to crypt-top Cluster 5. UC-associated goblets expressed genes essential for the integrity and homeostasis of the epithelial barrier[32] (**Extended data Fig. 7a(ii), g and Supplementary Data**).

We validated novel GC expressed genes by immunofluorescence (IF). **Fig. 4b(i-iv)** shows expression of BCAS1 (Cluster 5), CLCA1 (Cluster 1), REGIV (Cluster 1) and WFDC2 (Cluster 2) together with MUC2 within GCs. CLCA1 and WFDC2 are expressed along a gradient which is higher at the bottom of the crypt (**Fig. 4b(ii-iii)**) and is in line with our *in-silico* maturity/crypt gradient predictions (**Extended data Fig. 7b**). In comparison, REGIV was mainly observed as expressed in the mid-to-upper portions of the crypt **(Fig. 4b(iv)**). Not all GCs expressed these proteins, which is also consistent with segregation across sub-clusters. Double stains for WFDC2 and CLCA1 (**Fig. 4b(v)**) and WFDC2 and REGIV (**Fig. 4b(vi)**) confirmed the heterogeneity of protein expression in GCs suggested by the single-cell profiles.

Loss of WFDC2 in goblet cells in active UC

Spatial architecture of GCs within the inflamed crypt was perturbed and associated with dysregulation of numerous genes, including *WFDC2*. *WFDC2* was normally highly expressed by crypt-base goblets (**Extended data Fig. 7b**) but downregulated in inflammation (**Extended Data Fig. 7f and Supplementary Data**). We investigated whether *WFDC2* loss is a hallmark of colitis in a larger cohort of patients by IHC for WFDC2 from clinically non-inflamed and inflamed sections from patients with mild or severe UC. As a control for GC

health, we also stained for mucin 2 (MUC2). Expression of WFDC2 and MUC2 in UC patients with varying degrees of mucosal inflammation are shown in **Fig. 4c**. More severely inflamed tissue sections showed a clear reduction in WFDC2 expression. Both visual and digital scoring (**Fig. 4d** and **Extended data Fig. 7h**) confirmed significant protein loss in these cells. Further, we showed a similar loss of *Wfdc2* expression in dextran sodium sulphate (DSS)-treated mouse colonic tissue using RNA *in-situ* hybridisation (smISH) (**Fig. 4e(iii**)) and 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 GWAS loci. We hypothesized that the local inflammatory cytokine milieu may dictate the expression levels of WFDC2 and other proteins, and the degree of residual barrier protection in colitis. In line with this hypothesis, we observed induction of a number of interferon (IFN)-induced genes in GCs (**Extended data Fig. 7i**) with a location-dependent bias, suggesting that at least some of the dysregulation in GCs may 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 IFN-g and observed both distinct organoid morphology and downregulation of *WFDC2* (**Extended data Fig. 8a(iii)**). Given our data, one possible source of IFN-g stimulation may be intra-epithelial lymphocytes (**Extended data Fig. 8a(iv)**).

Barrier integrity requires secreted WFDC2

WFDC2 is proposed to regulate innate immunity via inhibition of serine and cysteine proteases[33]. WFDC2 was secreted both basally and apically and increased in response to stimulation in HT29-MTX-E12[34] cells (**Extended data Fig. 8b(i-ii)**). It inhibited matrix metalloproteinases MMP12 and 13 proteolytic activity whose pathological induction in IBD can orchestrate tissue destruction[35] (**Extended data Fig. 8c(i-ii)**). Furthermore, *in-vitro WFDC2* knockdown showed a disturbed cellular morphology with GC 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 anti-bacterial activity. Recombinant WFDC2 exhibited a marked dose-dependent reduction in the viability of both Gram-positive *Staphylococcus aureus* and Gram-negative *Escherichia coli* and *P. aeruginosa*. However, the viability of other Gram-positive (*Enterococcus faecalis)* and Gram-negative (*Salmonella typhimurium)* bacteria remained unaffected (**Fig. 5a**). This selective bactericidal activity of WFDC2 at a concentration comparable to other intestinal AMPs[37,38] suggested 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 mice (*Wfdc2*$^{-/+}$), as homozygous deletion of *Wfdc2* was embryonically lethal. smISH (**Fig. 5b(i-ii)**) confirmed reduced *Wfdc2* mRNA in the colon of *Wfdc2*$^{-/+}$ mice. Transmission Electron Microscopy (TEM) revealed colonic epithelial intercellular junctional abnormalities (**Fig. 5b(iv)**) along with irregular distribution of microvilli in *Wfdc2-/+* mice (**Fig. 5b(vi)),** compared to wild-

type (*Wt*) littermates (**Fig. 5b(iii and v**)). *Wfdc2-/+* mice presented with abnormal histology with mild to modest epithelial hyperplasia (**Methods**), accompanied with lymphoid infiltration (**Fig. 5c and Extended Data Fig. 9a**)

We explored whether the absence of *Wfdc2* facilitated the breakdown of the inner-mucus sterility. MUC2 staining of colonic tissues with the preserved mucus layer[39] suggested 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 along with GC damage in *Wfdc2-/+* mice (**Fig. 5d(iii-iv) and Extended data Fig. 9c**)**.** Unlike in *Wt* tissues (**Fig. 5e(i)**), TEM analysis of *Wfdc2*$^{-/+}$ mice showed invading bacteria free in the epithelial cytoplasm within a matrix of vesicles, fibers 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 demonstrate that WFDC2 is an important new goblet cell secreted anti-bacterial defense factor required to prevent colonization, invasion and epithelial barrier breakdown.

Discussion

We present the first large scale scRNA-Seq study of the human colonic epithelium in health and inflammation, revealing previously unknown cellular diversity and subtype-specific gene dysregulation in colitis.

We characterise a novel absorptive cell type, BEST4+/OTOP2+ cells, which are involved in pH-sensing and maintaining luminal homeostasis through regulation of the GC-C signaling pathway. They selectively express Uroguanylin, the endogenous paracrine hormone required for GC-C activation. GC-C receptor signaling 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 cross-talk and microbiota composition[40]. Dysregulation of this circuit underlies intestinal transformation. Our data shows that these unique Uroguanylin-producing colonic epithelial cells are depleted in IBD and colorectal cancer suggesting a novel mechanism by which this pathway is dysregulated in these diseases. This provides a new rationale for the use of FDA-approved Uroguanylin mimetics and has wide-ranging implications for future studies.

Furthermore, we delineate the functional role of a novel colonic GC-secreted anti-bacterial protein, WFDC2, in mucosal barrier homeostasis, which we find is localized towards the bottom half of the crypt in health and is dysregulated in UC. Evidence exists for how regional differences in GC phenotypes may impact key aspects of crypt physiology, such as barrier mucus. Colonic mucus is composed of inner and outer layers, the outer layer is not attached and creates a habitat for microbiota[36]. WFDC2 anti-protease activity inhibits the activities of serine and cysteine proteases, preventing the premature conversion of the inner mucus layer to the outer mucus in health. Indeed, knockdown of *WFDC2* expression results in

abnormalities in mucus layer formation (**Extended data Figures 8d and 9b**). These mucus defects may allow bacterial penetration and epithelial contact (**Figure 5**), which is a hallmark of UC[41]. Recent mouse studies support the existence of functional sub-populations of GCs, with differing mucus production and secretion rates along the crypt axis[42]. Our work provides a basis for spatial interrogation of GC phenotypes, key aspects of crypt physiology and how this specialization breaks down in barrier diseases such as UC.

Acknowledgements

We thank all the patients who contributed to this study, the generous support of our endoscopy teams and clinical research nurses led by Ms S. Fourie who made this work possible. We acknowledge support of the Wolfson Imaging Centre, WIMM flow cytometry facility, Discovery Proteomics Facility, Ms. Raman Dhaliwal for preparing TEM samples, Oxford NIHR Biomedical Research Centre, NIHR CRN Thames Valley, the Oxford Single Cell Consortium and life science editors for editorial support. This work was supported by an NIHR Research Professorship, Wellcome Investigator Award (A.S.); the MRC (H.K.; A.S.); Abbvie (K.P.) and Celgene (A.A.; H.C.). D.F-C was supported by a Royal College of Surgeons of England / British Association of Paediatric Surgeons Research Fellowship and the Wellcome Trust. Further acknowledgements in **Supplementary Notes**.

Author Contributions

Conceptualization, K.P., D.F-C., A.A., A.S. Methodology and Experiments, K.P., D.F-C. and M. J. performed and analysed all experiments. A.Au. H.C., N.A., S.D., H.H.C., J.L., R.S.A and E.B. performed wet lab 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.A., H.K and J.K. performed computational analysis and design. S.D. R.F. and B.M.K. performed proteomic experiments. Writing and editing, K.P., D.F-C., A.A., H.K., A.S.

Correspondence and requests for material should be addressed to K.P. (kaushal.parikh@ndm.ox.ac.uk); A.A. (agne.antanaviciute@ndm.ox.ac.uk)**;** D.F-C. (david.fawkner-corbett@balliol.ox.ac.uk) and A.S. (alison.simmons@imm.ox.ac.uk).

Competing interests

The authors declare no competing interests.

References

1 Peterson, L. W. & Artis, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nature Reviews Immunology* **14**, 141, doi:10.1038/nri3608 https://www.nature.com/articles/nri3608#supplementary-information (2014).

2 McCauley, H. A. & Guasch, G. Three cheers for the goblet cell: maintaining homeostasis in mucosal epithelia. *Trends Mol Med* **21**, 492-503, doi:10.1016/j.molmed.2015.06.003 (2015).

3 Kabat, A. M., Pott, J. & Maloy, K. J. The Mucosal Immune System and Its Regulation by Autophagy. *Frontiers in Immunology* **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. *Inflammatory Bowel Diseases*, izy380-izy380, doi: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.e1111, doi:https://doi.org/10.1016/j.cell.2018.04.037 (2018).

6 Rathinam, V. A. K. & Chan, F. K.-M. Inflammasome, Inflammation, and Tissue Homeostasis. *Trends Mol Med* **24**, 304-318, doi:https://doi.org/10.1016/j.molmed.2018.01.004 (2018).

7 McDole, J. R. *et al.* Goblet cells deliver luminal antigen to CD103+ dendritic cells in the small intestine. *Nature* **483**, 345-349, doi:10.1038/nature10863 (2012).

8 Johansson, M. E. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nature Reviews Immunology* **16**, 639-649, doi:10.1038/nri.2016.88 (2016).

9 Ayabe, T. *et al.* Secretion of microbicidal α-defensins by intestinal Paneth cells in response to bacteria. *Nature Immunology* **1**, 113, doi:10.1038/77783 (2000).

10 Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003, doi:10.1038/nature06196 https://www.nature.com/articles/nature06196#supplementary-information (2007).

11 Ito, G. *et al.* Lineage-specific expression of bestrophin-2 and bestrophin-4 in human intestinal epithelial cells. *PLoS One* **8**, e79693, doi:10.1371/journal.pone.0079693 (2013).

12 Tu, Y. H. *et al.* An evolutionarily conserved gene family encodes proton-selective ion channels. *Science* **359**, 1047-1050, doi:10.1126/science.aao3264 (2018).

13 Ikpa, P. T. *et al.* Guanylin and uroguanylin are produced by mouse intestinal epithelial cells of columnar and secretory lineage. *Histochemistry and Cell Biology* **146**, 445-455, doi:10.1007/s00418-016-1453-4 (2016).

14 Sato, M. & Bremner, I. Oxygen free radicals and metallothionein. *Free Radical Biology and Medicine* **14**, 325-337, doi:https://doi.org/10.1016/0891-5849(93)90029-T (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, doi:10.1038/s41556-018-0105-4 (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, doi:10.1186/1471-2164-11-5 (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, doi:10.1155/2014/634123 (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, doi:10.1038/jhg.2016.35 (2016).

19 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (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, doi:10.1097/MIB.0000000000000535 (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, doi:10.1038/ng.3818 (2017).

22 Sasaki, N. *et al.* Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. *Proceedings of the National Academy of Sciences* **113**, E5399 (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, doi:https://doi.org/10.1016/j.surg.2014.01.013 (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, doi:10.1093/bioinformatics/btu326 (2014).
25 de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics* **49**, 256-261, doi:10.1038/ng.3760 (2017).
26 Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics* **47**, 979-986, doi:10.1038/ng.3359 (2015).
27 Kinchen, J. *et al.* Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. *Cell* **175**, 372-386 e317, doi:10.1016/j.cell.2018.08.067 (2018).
28 Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics* **43**, 246-252, doi:10.1038/ng.764 (2011).
29 Leiper, J. M. The DDAH-ADMA-NOS pathway. *Therapeutic drug monitoring* **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. *Nature genetics* **48**, 510-518, doi:10.1038/ng.3528 (2016).
31 Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124, doi:10.1038/nature11582 (2012).
32 Spenlé, C. *et al.* Dysregulation of laminins in intestinal inflammation. *Pathologie Biologie* **60**, 41-47, doi:https://doi.org/10.1016/j.patbio.2011.10.005 (2012).
33 Chhikara, N. *et al.* Human Epididymis Protein-4 (HE-4): A Novel Cross-Class Protease Inhibitor. *PLOS ONE* **7**, e47672, doi:10.1371/journal.pone.0047672 (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 of Inflammation* **2015**, 19, doi:10.1155/2015/964131 (2015).
36 Johansson, M. E. V. *et al.* The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15064-15069, doi:10.1073/pnas.0803124105 (2008).
37 Porter, E. M., van Dam, E., Valore, E. V. & Ganz, T. Broad-spectrum antimicrobial activity of human intestinal defensin 5. *Infection and Immunity* **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 (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. *Proceedings of the National Academy of Sciences* **108**, 4659, doi:10.1073/pnas.1006451107 (2011).
40 Waldman, S. A. & Camilleri, M. Guanylate cyclase-C as a therapeutic target in gastrointestinal disorders. *Gut* **67**, 1543, doi:10.1136/gutjnl-2018-316029 (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, doi:10.1136/gutjnl-2012-303207 (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, doi:10.1371/journal.pone.0041009 (2012).

Figure Legends

Figure 1. **Human Colonic Epithelial Cell Heterogeneity in Health.**

a, t-SNE plot of the healthy human colonic epithelium (n=3, CT-Colonocytes-Crypt Top Colonocytes; EEC-Enteroendocrines; GC-Goblet Cells; Undiff.#1/2-Undifferentiated#1/2). **b,** Heatmap showing cluster markers, coloured by relative gene expression. Relative size of each dot represents the fraction of cells per cluster expressing each marker. **c,** Spatial segregation of cell clusters along crypt base–crypt top axis. Y-axis represents the axis score generated from expression of 15 crypt axis markers. Size of dots represents expression of SEPP1, a known crypt axis marker. **d,** Differentiation trajectory analysis. Predicted secretory lineage cells shown in blue; absorptive in red; and uncommitted in green (n=3).

Figure 2. **scRNA-Seq identifies a novel colonic absorptive cell type.**

a, Representative images (n=3) of colonic sections stained with BEST4 by IHC (i), *OTOP2* (ii) and *GUCA2B* (iv) by sm-ISH with co-staining and co-localisation of BEST4 and *OTOP2* (iii) or *GUCA2B* (v and vi) (stain colour – brown or blue – represented by text and magnification shown). **b,** Volcano plot showing proteins differentially expressed between BEST4+/EPCAM+ cells (n=3) (positive log fold change) and BEST4-/EPCAM+ cells (negative log fold change) (n=2). Red line indicates 5% FDR (limma linear model empirical Bayes p-value and Benjamini-Hochberg multiple-testing correction). Selected proteins are highlighted. **c,** Heatmap showing selected DEGs between BEST4+/EPCAM+ cells and BEST4-/EPCAM+ cells, detected using single-cell Smart-Seq2 (n=4 per group). **d,** BEST4+/EPCAM+ cells mediate proton influx. Representative image of change in intracellular pH when BEST4+/EPCAM+ and BEST4-/EPCAM+ cells are exposed to an extracellular pH5 buffer (i). Normalised fluorescence emission (F/F_0) from pH indicator pHrodo Red in BEST4-/EPCAM+ (n=39; pink) and BEST4+/EPCAM+ (n=45; orange) cells. Responses to pH5 solution shown (ii) (mean +/- SEM). Intracellular pH_i for the same populations as measured with pHrodo Red AM (**Methods**), showing peak response during each stimulus and initial starting pHi (two-sided paired t-test BEST4-/EPCAM+ p-value 0.9873310; BEST+/EPCAM+ p-value 0.0000007768. Mean with SEM shown) (iii).

Figure 3. Human Colonic Epithelium in Active Colitis

a, t-SNE plot of single-cell clusters in active UC (n=3, ILCs-Innate Lymphoid Cells). **b,** Heatmap showing key DEGs (FDR < 1%, two-sided negative binomial likelihood ratio test, Benjamini-Hochberg multiple testing correction) between cell clusters in health and active UC (n=3). Colour indicates log2 fold change (dark purple – downregulation; yellow – upregulation). Point size shows confidence of the observation (-log10 FDR) (Legend: 1-BEST4/OTOP2 cells, 2-Colonocytes, 3-Enteroendocrine, 4-Goblets, 5-Stem Cells, 6-Absorptive Progenitors, 7-Secretory Progenitors, 8-TAs, 9-Crypt Top Colonocytes). **c,** Significance level (-log 10 FDR) of tissue-specific expression enrichment of UC-associated GWAS loci in single-cell clusters in

colonic epithelium (n=3) and colonic mesenchyme[27] (n=2) in health and active UC. Dashed lines indicate thresholds for 10%, 5% and 1% FDR cut-offs (SNPsea empirical distribution p-value, Benjamini-Hochberg multiple testing correction). TAs-Transit-amplifying cells; SPs-Secretory progenitors; S1-S4-Stromal 1-4; APs-Absorptive Progenitors. **d,** t-SNE overlay of selected GWAS UC-associated genes expressed specifically in crypt-top colonocytes and goblets (n=3).

Figure 4. Goblet cell heterogeneity in health and UC.

a, t-SNE plot of subclusters across all captured GCs (n=3 per group). **b**, Novel and previously known GC markers validated by immunohistochemistry (IHC) in healthy human colonic tissue (representative of 3 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). Heterogeneity in expression of these markers is also observed by double staining (v) WFDC2 (green) with CLCA1 (red) and (vi) WFDC2 (red) with REGIV (green) (representative images, n=3 independent experiments) (scales:(i)-20μm, (ii, iii, iv, vi)-50μm, (v)-10μm) **c**, Representative IHC images of colon biopsies from inflamed (I) and non-inflamed (NI) regions of UC colon stained for WFDC2 (n=31) and MUC2 (n=24). Top panel shows WFDC2 and MUC2 expression in a patient with mild disease, while the bottom panel shows expression in severe inflammation. **d**, Quantification and distribution of WFDC2 and MUC2 expression change (log2 fold change) in inflamed vs non-inflamed tissue from patients with varying disease severity (25^{th}, 50^{th} and 75^{th} percentiles shown). Each point coloured and sized by severity score (automated quantification – **Methods**). Triangles/arrows represent outliers (>16-fold decrease). Paired two-sided t-test: MUC2 (n=24): p=0.2485; WFDC2 (n=31): p=0.0001779. **e,** (i) smISH of *Wfdc2* in the colon of naïve and DSS-treated acute colitis mouse model (n=5 per group) (ii) smISH quantification from colons of naïve and DSS-treated mice (p=0.0008, unpaired two-sided t-test, n=5 per group, mean and SEM shown). **f,** *Wfdc2* quantification relative to *Gapdh* by qRT-PCR from colons of naïve and DSS-treated mice (n=4 per group, p=0.0086, unpaired two-sided t-test, mean and SEM shown).

Figure 5. **WFDC2 shows selective bactericidal activity and influences barrier function**

a, Purified recombinant WFDC2 was added to mid-logarithmic phase bacteria for 4h. Surviving bacteria were quantified by dilution plating. Means (n=3) ± SD are plotted. **b**, *Wfdc2* smISH in colons from *Wfdc2*$^{-/+}$ & *Wt* littermates (i-ii). TEM of colonocytes shows disrupted tight junctions (arrows) (iv) and scattered microvilli (vi) in *Wfdc2*$^{-/+}$ mice compared to WT (iii & v) (scales: (i,ii) -100μm, (iii)-1μm, (iv, v, vi)-2μm). **c**, Histopathological evaluation of changes in epithelial cell morphology and mucosal architecture. *Wfdc2*$^{-/+}$ mice show irregular crypts with variable diameters along the depth of single crypts (ii), and focal mucosal infiltration of leukocytes (iv & vi) compared to *Wt* littermates (i, iii & v) (magnification: (i)-20x, (ii-iv)-40x). **d**, GRAM staining identifies regions free of luminal contents above the cell epithelial layer in WT mice (i), but also indicate colonization by both GRAM-positive and GRAM-negative bacteria (arrows) in *Wfdc2*$^{-/+}$

mice (ii). SEM of the colonic surface shows bacteria invading goblet cells in the *Wfdc2*$^{-/+}$ mice (iv) compared to *Wt* mice (iii) (scales: (i,ii)-10μm, (iii,iv)-2μm). **e**, TEM analysis showed that bacteria invaded the *Wfdc2*$^{-/+}$ colonic tissue mostly through GCs (ii-iii). Bacteria were not confined to a membrane-bound compartment but located free in the cytoplasm. No invasion of epithelial surfaces observed in the *Wt* littermates. The epithelium was intact with preserved colonocytes and GCs (i) (Panels b-e, n=4 per group) (scales: (i)-5μm (ii-left)-10μm, (ii-right)-2μm, (iii-left)-10μm, (iii-right)-2μm).

Methods

Isolation of epithelial cells from patient biopsies

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 used 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: (REC reference:16/YH/0247) and (REC reference: 09/H1204/30). Biopsies were incubated in chelation medium (HPGA with 1mM EDTA) at 37^{o} C for 80 min with agitation. The supernatant, which contained epithelial crypts, was digested into a single-cell suspension by dissociation with TrypLE Express containing 50ug/ml DNAse for 1 hour at 37^{o}C. The epithelial single-cell suspension was washed and passed through a 70μm and 40μm filter. 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), where large numbers (>4) samples were processed simultaneously, samples were stored by freezing in 1ml of Cryostor DS10 (Sigma Aldrich). Samples were then thawed and epithelial cells isolated to allow batch processing. Viability and epithelial cell purity was similar to those of freshly isolated samples (data not shown).

Flow Cytometry

Before progressing to scRNA-seq, purity of epithelial populations was confirmed by FACS-analysis using anti-CD90 (FITC, Biolegend), CD326 (EPCAM, PeVio, Milteyni), CD45 (APC, Milteyni) and DAPI (BD) as per manufacturers' instructions. Samples were processed on the Attune NxT Flow Cytometer (Thermo Fisher) with compensation performed using compensation beads (BD) on each run. Once satisfactory viability and EPCAM purity had been demonstrated, samples were then directly processed for scRNA-seq.

For validation of the BEST4/OTOP2 cell sub-population, a similar epithelial staining protocol was used, with addition of a primary anti-BEST4 antibody (Atlas Antibodies), followed by a secondary staining for 30 minutes with Alexa Fluor 488 anti-rabbit secondary antibody (Invitrogen). Staining protocol was validated on the Attune NxT Flow Cytometer (Thermo Fisher) and then flow sorting performed on Sony SH800 Cell sorter (Sony) with BEST4+ gates set on a Fluorescence Minus One (FMO) and Secondary control, each of >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/μl (Single Cell 3' library and Gel Bead Kit v2) as described in the manufacturers protocol (10x User Guide, Revision B). On average, approximately 8,000 cells were loaded across 3 runs, each with 3 conditions - healthy, UC inflamed and UC non-inflamed. Cells

were suspended in PBS with 0.04% BSA at a concentration of 1,000cells/μl. GEM-generation, barcoding, GEM-RT clean-up, cDNA amplification and library construction were all performed as per manufacturer's protocol. Individual sample quality was checked using Bioanalyzer Tapestation (Agilent). Qubit was used for library quantification prior to pooling. The final library pool sequenced on the Illumina NovaSeq6000 instrument using 150bp paired-end reads. Average cell recovery was 1,400 cells per sample, with total 11,175 cells captured at a mean depth of 163,822 reads per cell and 1,736 mean genes per cell.

Plate based single-cell RNA-sequencing, real-time PCR and bulk 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 U/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 re-suspended in elution buffer (Qiagen) and quality was 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 RNAeasy MicroKit (Qiagen) according to manufacturers' instructions. 1μl of extracted RNA was then added to a 96-well plate containing lysis buffer and processed with the same SMART-seq2 protocol with 20 cycles of pre-amplification.

For microfluidic qPCR of small cell numbers, 100 BEST4+ and BEST4- cells were isolated from 3 biological replicates. RNA was amplified using a Specific Targeted Amplification (STA) strategy targeting the specified gene primers (Taqman, ThermoFisher) in the reverse transcription mix as per manufacturers protocols (Biomark, Fluidigm). 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 instruction (Biomark, Fluidigm). A sample with no reverse-transcriptase was included as a control.

For quantitative RT-PCR experiments with larger cell numbers (>25,000 cells), total RNA was isolated using the RNeasy microkit (QIAGEN) according to manufacturers' instructions. cDNA was then synthesized using the high-capacity RNA-to-cDNA kit (ThermoFisher 4387406) with RT-PCR then performed utilizing 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

For characterization of the BEST4/OTOP2 cell population by proteomics, BEST4+/EPCAM+ and BEST4-/EPCAM+ populations were isolated using FACS as previously

described. 6,250 cells were sorted in both conditions into 25µl lysis buffer comprising of 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 minutes. Protein lysates were digested using a modified SP3 protocol[44]. Briefly, proteins were reduced with 5 mM dithiothreitol for 30 minutes and then alkylated with 20 mM iodoacetamide for 30 minutes at room temperature. 2 µL of carboxyl-modified paramagnetic beads (prepared as in[45]) were mixed 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 minutes with orbital shaking at 1,000 rpm. Beads were then immobilised on a magnet for 2 minutes 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 rpm for 18 minutes to bind peptides, then beads were immobilised on the magnet for 2 minutes and the supernatant discarded. Beads were resuspended in 2 % dimethylsulfoxide (DMSO) and then immobilised on the magnet for 5 minutes and the supernatant transferred to LC-MS vials and were stored at -20 °C until analysis.

Peptides were analysed by nano-UPLC-MS/MS using a Dionex Ultimate 3000 coupled on-line to an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). A 75 µm x 500 mm C18 EASY-Spray column (Thermo Scientific) with 2 µm particles was used at a flow rate of 250 nL/min. Peptides were separated using a 60-minute 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). MS1 precursor scans were performed in the Orbitrap at a resolution of 120000 at 200 m/z and a rate of 1 Hz. Precursors were selected for MS/MS analysis using an isolation window with of 1.6 m/z and were fragmented using HCD at a normalised collision energy of 28. MS2 fragment spectra were acquired in the iontrap using the Rapid scan rate.

pH imaging

pH imaging was carried out as described elsewhere[12]. Briefly, sorted EPCAM+/BEST4+ and EPCAM+/BEST4- cells were plated onto poly-L-lysine coated coverslips at 37°C. After at least one hour, cells were loaded with the intracellular pH indicator pHrodo Red AM, using PowerLoad concentrate according to the manufacturers' 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 MES (150mM NaCl, 10mM MES, 2mM $CaCl_2$). The pHrodo Red fluorescence intensity of each cell was normalised 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 pH 4.5, 5.5, 6.5 and 7.5. A standard curve was generated and the pHi for each cell calculated using linear extrapolation.

Immunohistochemistry (IHC), Immunofluorescence (IF) and single molecule in-situ hybridization(smISH)

For IHC, paraffin-embedded tissue sections were deparaffinised through an ethanol gradient and heat-induced epitope retrieval performed by boiling at 96ºC for 25 min in either pH6-citrate or pH9-Tris/EDTA buffers. Peroxidase blocked before incubation with appropriate species-specific serum for primary antibody incubation of 90 min at room-temperature. For full details on antibodies and concentrations used see **Supplementary Table 2**.

Substrate development was performed for each primary antibody using ImmPACT 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 HIER if labelled with same-species antibodies. Haematoxylin and Eosin staining was carried out using a kit from Vector Laboratories.

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

For sm-ISH, all probes and RNAscope 2.5 HD assay - brown (cat. 310035) were purchased from Advanced Cell Diagnostics (ACD, Milan, Italy) and used according to the manufacturer's instructions. Information on probes used are detailed in **Supplementary Table 2**. Paraffin sections were pre-treated with Pretreat 1, 2, and 3 (ACD). Pre-warmed (40ºC) probes were added to the slides and incubated in the HybEZ oven (catalog 321461; ACD) for 2 hours at 40ºC. After 6-step signal amplification, tissues were detected by DAB (all part of the RNAscope 2.5 HD assay - brown kit) and counter-stained with Mayer's hematoxylin. Slides were mounted with PERTEX mounting medium (Gothenburg, Sweden) and photographed.

For double-stains of sm-ISH and IHC (*OTOP2* and BEST4) samples were processed in a manner identical to sm-ISH with subsequent overnight staining with anti-BEST4 and development with VectorBlue.

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

Quantification of patient's biopsy using Visiopharm

Slides stained for WFDC2 as described above were scanned using Leica ScanScope machine (Leica Biosystems) and quantified using Visiopharm (Visiopharm, Denmark) with a programmed protocol calculated as follows: percentage of positive goblet cell area = fraction of goblet cell area x 100, where 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 calculation randomly for 50% of the whole biopsy and give an average for each sample. The result was equated to the percentage positive/ brown stain in goblet cells for each biopsy. For data presented in Figure 4d, additional unmatched patient data for WFDC2 was included.

Anti-bacterial Activity

Mid-logarithmic phase cultures of ATCC12973 *S. aureus*, ATCC13379 *Enterococcus faecalis*, ATCC27853 *P. aeruginosa*, *Salmonella typhimurium* LT2 and *Escherichia 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 x 10^4 CFU/ml. A final volume of 100ul was used to measure antibacterial activity of recombinant WFDC2 (Abcam) by addition at a concentration between 0.45uM-4.50uM. Following 4hr 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 4hr compared to baseline. All experiments were performed three times.

Anti-protease activity

The MMP inhibitor activity was performed according to the manufacturer's protocol (Enzo Life Sciences, Switzerland, MMP12 (BML-AK403-0001) and MMP13 (BML-AK413-0001)). Briefly, all kit components were diluted according to recommended concentration. 20μl of each MMP (concentration varied – refer to manual for lot dependent variations) was added to control, inhibitor NNGH and WFDC2 (25ug/ml). The reaction plate was incubated for 30 minutes at 37°C to allow inhibitor and enzymes interaction. After desired time, 10μl of BML-P277-9090 substrate was added to each reaction and the fluorescence was measured at Ex/Em 545/576nm for 10 minutes at 37°C.

Animals

For *Wfdc2* experiments, animals 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 ES cells (obtained from EMMA (European Mouse Mutant Archive)) into blastocysts (MRC Harwell Transgenic Facility)[47-49].

For DSS colitis experiments, 10–12 weeks old C57BL/6 (Helicobacter pylori-free, murine norovirus-free) male mice (Envigo Laboratories, UK) were used. All procedures were certified according to the UK Home Office Animals (Scientific Procedures) Act 1986 (project license P9B86E6FD). A total of 10 mice were randomised into two treatment groups of five mice each. One group received no treatment and the other received 1.75% DSS (36–50 kDa MW, MP Biomedicals, lot #M9147) in their drinking water from study day 0 until mice were

euthanised on the morning of study day 7, and the tissue processed for routine IHC as described above. For another independent study, we used a group of 8 mice with 4 mice per group, treated with DSS as described above. RNA isolation from tissue was performed using Qiagen mini kit (Qiagen), as described earlier.

Wfdc2$^{-/+}$ mice were assigned a subjective colitis severity score based on a modification of the criteria described by Kojouharoff *et al*, 1997[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.

Scanning and Transmission Electron Microscopy

Mice were perfused fixed with 2.5% glutaraldehyde + 4% PFA in 0.1M sodium cacodylate and the colon was excised. Tissue was cut into 2-3 mm^3 pieces and then stored at 4°C in 0.25% glutaraldehyde in 0.1M sodium cacodylate buffer until processing for TEM. Samples were washed with 0.1M sodium cacodylate buffer followed by 50 nM glycine in the same buffer to quench free aldehydes, followed by another wash with 0.1M sodium cacodylate buffer. Samples were incubated in 1% osmium tetroxide + 1.5% potassium ferrocyanide in 0.1M sodium cacodylate buffer for 2 hours 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 mins 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 hr and then 50% resin overnight, followed by 75% and 100% resin each for 3 hrs and then 100% resin overnight. The samples were embedded in flat dish moulds and polymerised at 60°C for 48hr. Ultrathin (90nm) sections were cut with a Diatome diamond knife using a Leica UC7 ultramicrotome and post-stained for 5 mins with lead citrate. Sections were viewed on a FEI Tecnai 12 TEM operated at 120kV equipped with a Gatan OneView digital camera.
For Scanning Electron Microscopy (SEM), colons were fixed and processed as above dehydrated in graded ethanol series as above. Following this, they were 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 Zeiss Sigma 300 FEG-SEM at an acceleration voltage of 2 kV.

Cell Culture

A goblet cell producing cell line, HT29-MTX-E12[34], was obtained from ATCC and maintained in DMEM Glutamax medium (Life Technologies) containing 10% FBS (v/v) and 1% (v/v) antibiotics. Cultures were incubated at 37 °C in a humidified 5% (v/v) CO2 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 100ng/mL of Phorbol 12-myristate 13-acetate (PMA) for 6 hr before apical and basal

medium collection. WFDC2 was quantified using Human WFDC2 Quantikine ELISA Kit following the manufacturer's instructions (R&D Systems).

WFDC2 shRNA knockdown

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

Organoid Cultures

Organoid cultures were established as originally described by Sato *et al*. [53]. Briefly, cultures were established from four pairs of colonic biopsies, incubated with 0.4mg/mL Dispase (GIBCO) to establish a single cell suspension. This was then mixed with 50uL Matrigel (Corning) and plated on pre-warmed 24-well culture dishes. Embedded cells were overlaid with WREN medium (Wnt3a conditioned medium (L Wnt:3A (ATCC CRL:2647TM)) and ADF (Advanced DMEM-F12 medium - GIBCO) 50:50, Glutamax (Life Technologies), 10mM HEPES, N-2 [1x] (Life Technologies), B-27 [1x] (Life Technologies), 10mM Nicotinamide(Sigma Aldrich), 1mM N-acetyl-L-cysteine (Sigma Aldrich), 1ug/ml R-spondin 1 (RSPO1) (Peprotech), 50ng/mL human epidermal growth factor [EGF] (Peprotech), 100ng/mL human Noggin (Peprotech), 1ug/mL Gastrin (Sigma Aldrich), and 0.05uM PGE2 (Sigma Aldrich), 0.1uM A83-01, 10uM p38 inhibitor SB202190, 10uM Y27632 (all from R&D Systems). Medium was replaced with fresh WREN medium every other day.

For organoid stimulation experiments, once organoid cultures were established, 100ng/ml IFN-gamma was added to medium for 4 days in experimental conditions. For gene expression quantification, we isolated RNA from organoids and performed RT-PCR as described above.

Computational analysis

Raw 10X read processing and QC

Raw sequence reads were quality-checked using FastQC software. The Cell Ranger version 2.1.1 software suite (obtained from 10x Genomics, https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to process, align and summarise UMI counts against human hg38 assembly reference genome analysis set, obtained from University of California Santa Cruz (UCSC) ftp site[54]. Corresponding Ensembl gene annotation was obtained using 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 (likely doublets) total UMI counts; high percentage of UMIs originating from

mitochondrial features (>20%) and fewer than 200 genes expressed. UMI per cell per sample distributions were visualized using 'ggplot2' r package (**Extended data Fig. 10a**).

'Seurat' R package (Version 2.3.2) was used to normalise 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 based on the aligned number of reads per cell. Downsampled UMI count matrices were used to obtain the gene cellular detection rate, where a gene was considered detected if at least one UMI was assigned to it. The number of genes detected per cell for UC inflamed and non-inflamed as well as healthy samples was visualized as density distributions (**Extended data Fig. 10b**). Gene detection rate distributions 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 has been 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 G2M and G1 cell cycle scores together with total UMI counts per cell, percentage of mitochondrial features and experimental batches were considered a source of unwanted variation and were regressed out using 'Seurat' package. Variable genes were identified 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 R package 'M3Drop' (version 1.6.0). Dimensionality reduction was performed using PCA (r package 'irlba', version 2.3.2). Scree plots and Jack-straw permutation tests were used to determine significant principal components (p-value cut-off < 0.01) in the data, and cells were clustered in the reduced dimension space using 'Seurat' package (resolution=0.7). Cell clusters were visualised using tSNE plots, using all significant principal components (as previously determined by Jack-Straw permutation tests) as input and perplexity value of 30. 20 principal components were found to be significant and were used to cluster the whole dataset; 8 for clustering analysis of undifferentiated cells; 10 for goblet sub-cluster analysis and 6 for enteroendocrine cell analysis.

Batch Effect Assessment

To ensure that clustering was not driven by batch effects, batch distributions for the dataset were visualized using tSNE plots (**Extended data Fig. 10e**). We also quantified this effect by computing entropy of batch mixing, as described by Haghverdi *et al*.,[58], for tSNE 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 (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 neighborhoods of 100 randomly picked cell locations, bootstrapped 100 times and the distributions visualized as boxplots (**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*, *LYPD8*. For each gene, we normalised expression across all cells to a range between 0 and 1 to ensure individual gene contribution to the score was not weighted by base line expression. The final crypt-axis score for each cell was then defined as the sum of all normalised expression values.

Semi-supervised clustering of public scRNA-Seq data

To test if BEST4/OTOP2 cells are present in other datasets, we downloaded data from GEO, accessions GSE103239 and GSE81861, and processed it as described above, except clustering was performed using the *a piori* identified highly variable genes from our 10x data analysis. Cluster markers were detected as before and compared to the BEST4/OTOP2 cell markers in the 10x data.

Analysis of TCGA data

Htseq raw count matrixes were downloaded from the TCGA database for all available colorectal cancer patients and matched normal samples. Data were normalized using 'DESeq2' r package and variance stabilized using 'rlog' function[59]. Sample clustering and expression of the core BEST4/OTOP2 cell gene signature in this dataset was visualized using R package 'pheatmap'.

Cluster Marker and Differentially Expressed Gene Identification

Cluster gene markers were detected using 'Seurat' package, using the AUC 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. Benjamini-Hochberg multiple-testing correction was applied and genes with FDR <1% were considered differentially expressed.

R package 'MAST'[60] was used to estimate generalized linear model coefficients for inflamed and non-inflamed samples, using cells from healthy patients as a reference level. We built individual models for all major cell clusters using the 'zmb' function, where in addition to UC status (inflamed, non-inflamed 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 was downloaded from GEO (accession: GSE59071) from inflamed colon UC samples and healthy controls. R package 'limma' was used for data normalization and differential expression analysis[61]. Benjamini-Hochberg multiple testing correction was applied to estimate the false discovery rate.

Ontology Enrichment Analysis

Biological process GO 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 a false-discovery rate cut-off of 0.05, using all expressed/detected genes as a background control. The results were visualised as dot plots or emap plots using '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 oligo sequences, poly-d(T) and poor quality (< 20) sequences were trimmed using Cutadapt software. Reads were aligned to the human hg38 reference genome build using STAR aligner[63]. Raw read counts matrices were obtained using 'featureCounts' tool. Data quality metrics for each well were aggregated using multiQC tool.

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

Proteomics Data Analysis

Label-free quantitation 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 (retrieved 20180718, 95128 sequences). Precursor mass tolerance was set to 10 ppm and fragment mass tolerance was 0.5 Da and a maximum of 2 missed cleavages were allowed. Carbamidomethylation of C was set as a fixed modification and the variable modifications allowed were deamidation of N/Q and oxidation of M. Peptide FDR was adjusted to 1% and low scoring peptides (<20) excluded. R package 'limma' was used for protein expression normalization and differential expression analysis[61]. Benjamini-Hochberg multiple testing correction was used to estimate false discovery rate.

BEST4/OTOP2 cell marker overlap

The intersection of the top 200 markers for BEST4/OTOP2 cells identified from 10x single cell data, SmartSeq2 data, quantitative proteomics assay and datasets from Li *et al,*[21] and Gao *et*

al,[15] was visualized using the r package ‘circos’.

Trajectory and Pseudo-time Analysis

Cell differentiation trajectories were reconstructed using R package ‘monocle’ (version 2.8.0)[65]. Non-epithelial cell clusters were filtered out and dimensionality reduction performed using DDRTree algorithm, using all highly variable genes as input and the following residual model formula: "~Patient + nUMI + percent.mito + G1_score + G2M_score". Cell trajectory was then reconstructed using ‘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 pseudo-time 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 pre-branch state cells. All significantly upregulated (<1% FDR, > 0 log2 Fold change) genes were then intersected with all previously identified genes that showed significant pseudo-time varying, branch-specific expression. This subset identified genes with branch-specific expression that are also induced prior to 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 SNPsea algorithm[24] to test for significant enrichment of tissue-specific expression in UC associated GWAS loci genes. We downloaded UC-associated loci from GWAS catalog[66] database from the following two studies: de Lange *et al*.,[25] and Liu *et al*.,[26], which report the largest number of UC-associated loci to date. 1000 Genomes Project[67] data was used to sample matched control SNPs and link SNPs to genes. We first used Gene Atlas gene expression data (GEO: 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 ‘pseudo-bulk’ dataset for each previously identified cell cluster in health and UC separately by summing all UMI counts for each gene in each cluster. We then normalised the data by computing size factors (‘DESeq2’ r package, 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 subject to Benjamini-Hochberg multiple testing correction.

Methods References

43 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**, 171, doi:10.1038/nprot.2014.006 https://www.nature.com/articles/nprot.2014.006#supplementary-information (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, doi:10.1021/acs.jproteome.7b00433 (2017).

45 Hughes, C. S. *et al.* Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol Syst Biol* **10**, 757, doi:10.15252/msb.20145625 (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, doi:10.1371/journal.pone.0012238 (2010).

47 Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337, doi:10.1038/nature10163 https://www.nature.com/articles/nature10163#supplementary-information (2011).

48 Bradley, A. *et al.* The mammalian gene function resource: the international knockout mouse consortium. *Mammalian Genome* **23**, 580-586, doi:10.1007/s00335-012-9422-2 (2012).

49 Pettitt, S. J. *et al.* Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nature Methods* **6**, 493, doi:10.1038/nmeth.1342 https://www.nature.com/articles/nmeth.1342#supplementary-information (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. *Journal of Cell Science* **106**, 771 (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. *Nature Protocols* **6**, 806, doi:10.1038/nprot.2011.327 (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, doi:10.1053/j.gastro.2011.07.050 (2011).

54 Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* **14**, 144-161, doi:10.1093/bib/bbs038 (2013).

55 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32**, D493-D496, doi:10.1093/nar/gkh103 (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, doi:10.1038/nbt.4096 (2018).

57 Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54-61, doi:https://doi.org/10.1016/j.ymeth.2015.06.021 (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, doi:10.1038/nbt.4091 (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, doi:10.1186/s13059-014-0550-8 (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, doi:10.1186/s13059-015-0844-5 (2015).

61 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).

62 Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : a Journal of Integrative Biology* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

63 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

64 Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14**, 584-586, doi:10.1038/nmeth.4263 (2017).

65 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* **14**, 979-982, doi:10.1038/nmeth.4402 (2017).

66 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**, D896-D901, doi:10.1093/nar/gkw1133 (2017).
67 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
68 Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50**, 621-629, doi:10.1038/s41588-018-0081-4 (2018).
69 Vizcaino, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, D447-456, doi:10.1093/nar/gkv1145 (2016).
70 Travis, S. P. L. *et al*. 2012 Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut 61, 535, doi:10.1136/gutjnl-2011-300486 (2012).

Data availability statement

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

Extended data figure legends

Extended data Fig. 1, related to Fig. 1. Identification and validation of epithelial cell sub-populations.

a, Flow cytometry analysis of cells isolated from biopsies of healthy controls prior to scRNA-seq (i), measuring epithelial viability (DAPI-), purity (EpCAM+), immune (CD45+) and stromal (CD90+) markers (n=4, mean with SD) demonstrating gating strategy for known epithelial markers (ii), viability (iii), and immune compartment (iv). **b,** FACS purification of EpCAM+CD45- isolated epithelial cells (n=2) (i). Representative images (n=3) of IHC validation for LYZ expression in HC epithelial tissue sections in small intestine (positive control) (ii) and colon (iii) (images shown at 20x magnification) **c,** t-SNE plot of enteroendocrine cell sub-clusters. Single cells coloured by cluster annotation. Descriptive cluster labels shown (n=3 per group). **d,** 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-localisation of both markers in some cells (blue and brown arrow) but not in other EEC cells (blue or brown arrow). **e**, Violin plots showing gene expression (y-axis) of top enteroendocrine sub-cluster markers for Enterochromaffin (ECs) (left panel), L-Cells (LCs) (middle panel) and a pre-cursor cell population (PCs) (n=3, center bar indicates median value, colour indicates mean expression). **f,** t-SNE plot visualising undifferentiated colonic epithelium cell sub-clusters (n=3). **g,** Violin plots of gene expression (y-axis) in stem cells (SCs), cell cycle (CC) cluster cells, absorptive progenitor (AP) cells, secretory progenitor (SP) cells and transit-amplifying (TA) cells. Top markers for SC (i), AP (ii) and SP (iii) shown (n=3, center bar indicates median value, colour indicates mean expression). **h,** Crypt-axis score super-imposed over the differentiation trajectory captured by Monocle analysis (n=3). **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 showing significant enrichment among all marker genes for epithelial clusters. The number of markers identified for each cluster indicated (x-axis). Circle size 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, related to Fig. 2. Validation of BEST4/OTOP2 cell population.

a, Cluster distribution along differentiation trajectory captured by Monocle. BEST4/OTOP2 cells are highlighted on the left (n=3). **b,** tSNE gene expression overlay of core BEST4/OTOP2 cell markers (n=3). **c,** Representative images (n=3) of colonic sections stained with key BEST4/OTOP2 cell markers by IHC to demonstrate BEST4 staining at high magnification (i) (100x) and CTSE at low (ii) and high (iii) magnification and additional stains with sm-ISH for *SPIB* (iv) and *HES4* (v) (each representative 3 samples). **d,** (i)

tSNE visualization of semi-supervised clusters of scRNA-seq data from fetal human colon study by Gao *et al*,[15] (n=2). (ii) Boxplot (25th, 50th and 75th quantiles shown) showing co-localized expression of the core BEST4/OTOP2 cell signature. **e,** Heatmap showing expression of the core BEST4/OTOP2 cell gene signature in TCGA bulk RNA-seq data in colorectal cancer patients and matched normal tissue. **f,** (i) tSNE visualization of semi-supervised clustering of scRNA-seq data from colorectal cancer study by Li *et al*,[21] (n=10) (ii) Boxplot (25th, 50th and 75th quantiles shown) showing localized expression of the core BEST4/OTOP2 cell signature.

Extended data Fig. 3, related to Fig. 2: Isolation and characterization of BEST4/OTOP2 cell population.

a, Flow cytometry gating strategy for isolation of 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,** 100 BEST4+/EPCAM+ and BEST4-/EPCAM+ sorted cells (n=3) processed using microfluidic RT-PCR demonstrate increased expression of markers identified from single cell data relative to *GAPDH*. Mean and SEM values shown **c,** Circos plot showing overlap between top 200 BEST4/OTOP2 cell markers detected between 10x, Smart-Seq2, quantitative proteomics and semi-supervised clustering of data from Li *et al*[21] and Gao *et al*[15]. **d,** Over-represented GO terms in significantly upregulated protein set in BEST4/OTOP2 cells as identified by quantitative proteomics (n=2 BEST4- vs n=3 BEST4+, hypergeometric test and FDR calculated Benjamini-Hochberg multiple testing correction).

Extended data Fig. 4, related to Fig. 3 and 4. Gene Ontology Enrichment Analysis of Differentially Expressed Genes in Colonic Epithelial Cell Clusters.

a, Dotplot of GO biological process (BP) enrichment in upregulated genes (<1%FDR) comparing cell clusters in active colitis and health. **b,** Dotplot of GO BP enrichment in downregulated genes (<1% FDR) comparing cell clusters in active colitis and health. **c,** Dotplot of GO BP enrichment in differentially expressed genes (< 1% FDR) in inactive, but not in active colitis. Points in each dotplot coloured by enrichment confidence (-log 10 FDR) and sized by the proportion of all genes within the cluster annotated with the GO 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 showing dysregulation in active colitis (I) when compared to healthy (HC) samples in stem cells and/or other undifferentiated populations. (n=3 per group, center bar

indicates median value, colour indicates mean expression) **e,** Representative IHC images (n=3) of LYZ expression from inflamed (i **–** 20x magnification, ii – 40 x magnification) and non-inflamed (iii – 20x magnification, iv – 40 magnification) colonic tissue sections.

Extended data Fig. 5, related to Fig.3. **Human Colonic Epithelium in clinically non-involved mucosa and UC-associated GWAS loci Analysis**

a, Heatmap visualising UC-associated GWAS loci expression specificity 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) visualising selected GWAS UC-associated gene expression. **c,** Volcano plot showing differentially expressed genes detected in microarray study from Vanhove *et al*[20], comparing inflamed UC samples (n=74) vs healthy control (n=11) colon samples. Significantly downregulated (limma linear model empirical Bayes p-value and Benjamini-Hochberg multiple testing correction) BEST4/OTOP2 cell core signature genes are highlighted. **d,** Distribution of cluster sizes in healthy and UC inflamed and non-inflamed samples (n=3 per group), shown as bar charts of proportions of total cells captured. Mean and SEM values are shown. **e,** t-SNE plot of human colonic epithelium single cell clusters in non-inflamed UC (n=3).

Extended data Fig. 6, related to Fig.3. **Human Colonic Epithelium in clinically involved and non-involved mucosa**

a, Violin plots visualising expression (y-axis) of selected differentially expressed genes (< 1%FDR, two-sided negative binomial likelihood ratio test, Benjamini-Hochberg multiple testing correction) in non-inflamed and active UC (n=3). Center bar indicates median value, colour indicates mean expression. **b,** Heatmap visualising relative expression of all differentially expressed genes (< 1%FDR, two-sided negative binomial likelihood ratio test, Benjamini-Hochberg multiple testing correction) detected in inflamed (red) and non-inflamed (green) colitis compared to healthy tissue (blue) (n=3 per group). **c,** Venn diagram shows the overlap between differentially expressed genes detected in all clusters in active (purple) and inactive (salmon) colitis, compared to healthy tissue. **d**, Comparison between MAST generalized linear model coefficients for significant DEGs in UC inflamed and non-inflamed 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, related to Fig.4 and Fig.5. Goblet cell remodeling and WFDC2 dysregulation in inflammation

a, Violin plots showing cluster gene expression (y-axis) for key marker genes in clusters 1 (i), 2 (ii), 3 (iii) and common cluster 4 and 5 markers (iv) (n=3 per group). Center bar indicates median value, colour indicates mean expression. **b,** (i) Pseudo-temporal ordering of GC clusters. (ii) Crypt-axis score super-imposed on trajectory analysis. Cells predicted to reside at the top of the crypt are more mature populations, as inferred by pseudo-time 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,** Gene expression boxplots of selected genes in goblet cells, divided spatially based on the crypt axis by binning into 4 ranges (Bottom, Mid1, Mid2 and Top) (n=3 per group, 25th, 50th and 75th percentiles shown). (i) Expression of *CD74*, (ii) *LCN2*, (iii) *REG1A*, (iv) *SPINK1*, (v) *SPINK4* and (vi) *LAMB3* are shown in health and inflamed UC. **d,** Increased expression of REG1A and SPINK4 (ii and iv) was confirmed in inflamed UC biopsies as compared to health (i and iii) by IHC (representative images of n=3 for each). **e**, Stacked barchart showing GC sub-cluster relative frequency distribution (% of GC cells captured) in health, active (inflamed) and inactive (non-inflamed) colitis. **f**, Violin plots showing expression (y-axis) of WFDC2 in crypt-bottom GC clusters in healthy samples (HC) and inflammation (I) (n=3 per group, center bar indicates median value, colour indicates mean expression). **g,** Comparison of over-represented (hypergeometric test, Benjamini-Hochberg multiple testing correction) GO BP terms in GC sub-cluster markers (n=3 per group). **h,** Quantification of WFDC2 and MUC2 expression by IHC from patient-matched inflamed and non-inflamed sections of 24 UC patients. Staining intensity scored from 0 – no staining/weak staining to 3 – strong staining by three independent observers. Comparison between WFDC2 inflamed and non-inflamed, p=0.000148773, two-sided Wilcoxson matched pairs signed rank test, n = 24 patients. Comparison between MUC2 inflamed and non-inflamed is not significant. Mean and SD shown. **i,** Expression of interferon-induced genes in goblet cells (n=3 per group), *IFI6* (i), *ISG15* (ii), *IFITM3* (iii), *ISG20* (iv).

Extended data Fig. 8, related to Fig. 4 and Fig. 5. *In vitro* regulation of WFDC2

a, Non-treated (i) and IFN-g treated (ii) human colonic organoids in culture. (iii) qRT-PCR quantification of WFDC2 expression in IFN-g treated and non-treated organoids (n=2 independent experiments, mean values shown). (iv) tSNE plot of inflamed epithelium highlighting localised expression of IFN-g in intra-epithelial lymphocytes (n=3). **b,** Quantification by ELISA of WFDC2 secretion into apical (i) and basal (ii) media of HT29-MTX-E12 cells with and without 100 ng/mL of PMA stimulation for 6 hours (n=1). **c,** MMP12 (i) and MMP13 (ii) activity measured in the absence and presence of various concentrations of WFDC2. Data presented as percent of activity remaining. (n=3, except MMP12+40ug/ml WFDC2 and untreated MMP13, where n=2. Mean and SD shown) **d.** WFDC2 knockdown in HT29-MTX-E12 cell lines (for panels i-iii, n=2). (i) Immunoblot of WFDC2 on cell lysates from

non-transfected (Lane 1), *WFDC2* shRNA transfected (clone 1 – lane 2, clone 2 – lane3) and scrambled transfected (lane 4) cells. Beta-actin was used as a loading control. (ii) Cell culture supernatants were tested by immunoblotting for secreted WFDC2. (iii) Cells grown on transwells were stained by Hematoxylin and Eosin (H&E) and Alcian Blue. Arrows indicate attached mucus layer and mucin secreting goblet cells.

Extended data Fig. 9, related to Fig. 5. WFDC2 influences barrier function

a, Histopathological evaluation of changes in epithelial cell morphology and mucosal architecture in *Wt* (i) and *Wfdc2*$^{-/+}$ (ii) shows bifurcation at the base of the crypt. Mice were assigned a subjective colitis severity score based on a modification of the criteria described by Kojouharoff *et al*, 1997[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 (iii). **b,** Colonic tissue from *Wfdc2*-/+ heterozygous mice and *Wt* 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 (i) and an outer layer (o). Arrows indicate the inner mucus layer. Higher magnification images are shown in the lower panels (n = 4). **c,** SEM of the colonic surface shows bacteria invading goblet cells in the *Wfdc2*$^{-/+}$ mice (scales: 2µm). **d,** TEM images of colons of *Wfdc2*-/+ mice showing epithelial cell damage with destruction of microvilli (i), epithelial detachment (ii) and destruction (iii) and bacterial aggregates were also observed over the surface of *Wfdc2*-/+ mice (iv) (panels b-d show representative images, n=4 animals per group).

Extended data Fig. 10, related to methods. Integrated sample analysis and batch distribution

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,** Boxplots showing entropy of batch mixing for sample batches (n=9) (right); positive controls, where clusters were assigned as batches (center); and negative controls, where cells were assigned random batch labels in accordance to batch size distribution (left). Entropy of batch mixing for sample batches approaches that of the negative control. 25th, 50th and 75th percentiles are shown as bars.