A single-cell survey of the small intestinal epithelium

Adam L. Haber1,*, Moshe Biton1,2,§, Noga Rogel1,*, Rebecca H. Herbst1,3, Karthik Shekhar1, Christopher Smillie1, Grace Burgin1, Toni M. Delorey1,4, Michael R. Howitt1, Yarden Katz3, Itay Tirosh1, Semir Beyaz6,7,8, Danielle Dionne1, Mei Zhang9, Rakhtima Raychowdhury1, Wendy S. Garrett1,5, Orit Rozenblatt–Rosen1, Hai Ning Shi1, Omer Yilmaz1,6,10, Ramnik J. Xavier1,5,11,§ & Aviv Regev1,12,§

Intestinal epithelial cells absorb nutrients, respond to microbes, function as a barrier and help to coordinate immune responses. Here we report profiling of 53,193 individual epithelial cells from the small intestine and organoids of mice, which enabled the identification and characterization of previously unknown subtypes of intestinal epithelial cell and their gene signatures. We found unexpected diversity in hormone-secreting enteroendocrine cells and constructed the taxonomy of newly identified subtypes, and distinguished between two subtypes of tuft cell, one of which expresses the epithelial cytokine Tsp1 and the pan-immune marker CD45, which was not previously associated with non-haematopoietic cells. We also characterized the ways in which cell–intrinsic states and the proportions of different cell types respond to bacterial and helminth infections: Salmonella infection caused an increase in the abundance of Paneth cells and enterocytes, and broad activation of an antimicrobial program; Heligmosomoides polygyrus caused an increase in the abundance of goblet and tuft cells. Our survey highlights previously unidentified markers and programs, associates sensory molecules with cell types, and uncovers principles of gut homeostasis and response to pathogens.

The intestinal mucous membrane interacts dynamically with the external environment. Intestinal epithelial cells sense the contents of and pathogens in the gastrointestinal tract and secrete regulatory products that orchestrate appropriate responses. However, we do not yet know all of the discrete types and subtypes of epithelial cell in the gut, their molecular characteristics, how they change during differentiation or how they respond to pathogenic insults.

A survey of RNA profiles of individual intestinal epithelial cells can help to address these questions. Previous surveys that relied on known markers to purify cell populations1,2 cannot always fully distinguish between cell types, might identify only subsets of types in mixed populations or might fail to detect rare cellular populations or intermediate states. Recent studies3–7 attempted to overcome these limitations by using single-cell RNA sequencing (scRNA-seq), but an extensive characterization of intestinal epithelial cellular diversity has yet to be achieved.

Here we perform a scRNA-seq survey of 53,193 epithelial cells of the mouse small intestine in homeostasis and during infection. We identify gene signatures, key transcription factors and specific G-protein-coupled receptors (GPCRs) for each main type of small intestinal differentiated cell. We distinguish proximal and distal enterocytes and their stem cells, establish a new classification of different enteroendocrine subtypes, and identify previously unrecognized heterogeneity within Paneth and tuft cells. Finally, we demonstrate how these cell types and states adaptively change in response to different infections.

Single-cell survey of small intestinal epithelial cells

We profiled 53,193 individual cells (Supplementary Table 1). First, we used droplet-based massively parallel scRNA-seq (Methods) to profile EpCAM+ epithelial cells from the small intestine of C57BL/6 wild-type and Lgr5–GFP knock-in mice (Fig. 1a). We estimated the required number of cells on the basis of a negative binomial model for random sampling (Methods). If we conservatively assume that 50 sampled cells are required to detect a subset, then profiling 6,873 cells would allow us to detect all known types of intestinal epithelial cell and a hypothetical additional type present at 1% with 95% probability (Methods). We collected 8,882 profiles, and removed 1,402 low-quality cells (Methods) and 264 contaminating immune cells (Methods), retaining 7,216 cells for subsequent analyses (Extended Data Fig. 1a). These profiles demonstrated excellent reproducibility (n = 6 mice, mean r = 0.95; Extended Data Fig. 1c–f).

Unsupervised graph clustering9,10 (Methods) partitioned the cells into 15 groups, which we visualized using t-distributed stochastic neighbour embedding10,11 (t-SNE; Fig. 1b) and labelled by the expression of known marker genes (Extended Data Fig. 1g). Each cluster was associated with a distinct cell type or state, such as enterocyte, goblet, Paneth, enteroendocrine or tuft (Fig. 1b). We identified proliferating cells using a cell-cycle signature12. The enteroendocrine, Paneth, goblet, stem and tuft cells were each represented by a single distinct cluster (Fig. 1b, Extended Data Fig. 1g). Absorptive enterocytes were partitioned across seven clusters representing distinct stages of maturation (Fig. 1b, supplementary information).
Extended Data Fig. 1g). The proportions of most types of differentiated intestinal epithelial cell were consistent with expected abundances given our crypt-enriched isolation (Methods, Extended Data Fig. 1d), although Paneth cells were under-represented (3.6% compared to 5% expected), and the abundances of enteroendocrine and tuft cells were higher than expected (4.3% and 2.3%, respectively, compared to 1% expected). To better capture Paneth cells, we devised a trajectory from stem-like to progenitor to immature enterocytes in pseudo-temporal order (Extended Data Fig. 4a–d). We observed inactivation of the transcription factors that were expressed in different regions of the diffusion map (Methods), we associated them with cell-type signatures. The relative expression level (row-wise score of log2(TPM)) of genes (rows) across cells (columns) is shown, sorted by type. d, e, Mptx2 is a newly identified Paneth cell marker. d, smFISH of Mptx2 (green; left), immunofluorescence assay of the Paneth cell marker Lyz1 (red; middle) and the combined image (right). Dashed line, crypt; arrow, Paneth cell; scale bars, 20 μm. e, In situ hybridization of Mptx2 (red). Scale bar, 50 μm.

Regional cell–type diversity

We next used diffusion maps22 to place the population of enterocytes in pseudo-temporal order (Extended Data Fig. 4a–d). We observed a trajectory from stem-like to progenitor to immature enterocytes (Extended Data Fig. 4a, c) and captured (via diffusion component 2) distinct paths towards enterocytes of the proximal (duodenum and jejunum) and distal (ileum) small intestine (Extended Data Fig. 4b, d). By identifying the transcription factors that were expressed in different regions of the diffusion map (Methods), we associated them with cell-fate commitment to absorptive enterocytes (known22, Sox4; newly identified, Batf2, Mxd3 and Foxm1; Extended Data Fig. 4c, e) or with proximal or distal spatial identity (known23, Gata4 and Nr1h4; newly identified, Creb3l3, Jund, Osr2 and Nr1l3; Extended Data Fig. 4d, e).

To test these predictions, in an independent experiment we profiled 11,665 single cells from epithelial tissue that was extracted separately from the duodenum, jejunum or ileum (n = 2 mice; Fig. 2a). The cells span a continuum that reflects both regional and differentiation ordering (Fig. 2a). Two separable subsets of differentiated enterocytes were populated by cells from either the duodenum or ileum (jejunum cells contributed to both). The signature genes for mature proximal and distal enterocytes that we identified computationally (Methods, Fig. 1c, Supplementary Table 2) were also differentially expressed between cells isolated separately from these regions (FDR < 0.05, Mann–Whitney U test; Fig. 2b), as confirmed by smFISH (Extended Data Fig. 3d). Most of the marker genes for one of the two subsets of Paneth cells (Extended Data Fig. 3b) were enriched (FDR < 0.05) in only the proximal small intestine, and those for the other subset in only the distal small intestine, confirming that the marker genes reflect regional distinctions (Extended Data Fig. 3c); however, the newly identified marker Mptx2 showed no regional specificity (Supplementary Table 10). Finally, the stem cells in each region also express region-specific markers (Extended Data Fig. 3e). When examined in either the non-regional (Extended Data Fig. 4f) or the regional (Fig. 2c) diffusion maps, these markers indicate distinct subsets of stem cells, each probably foreshadowing the eventual distinct type of enteroocyte in the corresponding region (Fig. 2c).

Characterizing subsets of enteroendocrine cells

Enteroendocrine cells are important sensors of nutrients and microbial metabolites4,25,26 that secrete diverse hormones and function as metabolic signal transduction units26. Enteroendocrine cells have been reported to comprise eight distinct subclasses, with cells expressing secretin (Sct), cholecystokinin (Cck), proglucagon (Ggc), glucose-dependent insulinotropic polypeptide (GIP), somatostatin (Sst), neurotensin (Nts), ghrelin (Ghrl) and serotonin (Tph1) traditionally termed S, I, L, K, D, N, A and pentraxins, such as C-reactive protein (CRP) and serum amyloid P component protein (SAP), help to defend against pathogenic bacteria18. In addition, the two subsets of Paneth cells expressed distinct panels of antimicrobial α-defensins (Extended Data Fig. 3b). Next, from the full-length scRNA-seq data we identified enriched transcription factors, GPCRs and leucine-rich repeat proteins (Methods) for each of the main cell types (Extended Data Fig. 2d–f, Supplementary Table 5). The transcription factors identified include Krüppel-like factors such as Klfl4, a known regulator of goblet cell development19, Klfl5 in Paneth cells, and Klfl3 and Klfl6 in tuft cells (Extended Data Fig. 2f). Among the cell-type-enriched GPCRs (Extended Data Fig. 2d, f, Supplementary Table 5), each of the sensory cell types (tuft and enteroendocrine) had more than ten enriched receptors. These included many nutrient-sensing receptors in enteroendocrine cells (such as the bile acid receptor20 Gpbar1 and Gpr119, a sensor for food intake and glucose homeostasis21) and the dopamine receptor Drd3 in tuft cells (Extended Data Fig. 2d). Leucine-rich repeat proteins including several pattern recognition receptors such as Tlr2 were also variably expressed across subsets (Extended Data Fig. 2e).
enterochromaffin cells, respectively. However, substantial crossover between these subtypes has been observed.

To define putative enteroendocrine cell subtypes, we partitioned the 549 enteroendocrine cells (Fig. 1b, 310 cells; Fig. 2a, 239 cells) into 12 clusters (Fig. 3a, b, Extended Data Fig. 5a, Supplementary Table 6, Methods). The cells in four of the clusters expressed markers of enteroendocrine cell precursors (Neurog3, Neurod1 or Sox4); the other eight clusters represent mature enteroendocrine cells. A recent scRNA-seq study of organoid-derived enteroendocrine cells showed regional differences in stem cell differentiation. The diffusion-map embedding of 8,988 cells is coloured by region (first panel), cluster (second panel), or expression (log2(TPM+1)) of newly identified regional markers of induced stem cells (Gkn3, third panel; Bex1, fifth panel) or enterocytes (Fabp1, fourth panel; Fabp6 sixth panel). Diffusion components (DCs) 1, 3 and 4 correspond approximately to differentiation state, proximal or distal identity, and stem-like gene expression, respectively.

Figure 2 | Regional variation in cell type and differentiation. a, Regional surveys. t-SNE of 11,665 cells from the duodenum, jejunum and ileum (n = 2 mice) is shown, coloured by region (left) or assigned cell type (right). b, Regional enteroocyte signatures. The relative expression of genes (rows) across cells (columns) is shown, sorted by region. D, duodenum; J, jejunum; I, ileum. Colour scale as in Fig. 1c. Regional differences in stem cell differentiation. c, Regional variation in cell type and differentiation. t-SNE of 11,665 cells from the duodenum, jejunum and ileum (n = 2 mice) is shown, coloured by region (left) or assigned cell type (right).
tuft and microfold cells. A cluster of 18 cells (Fig. 5c, Methods) was enriched for known microfold cell markers (FDR < 0.05, Mann-Whitney U test; Fig. 5d) and the in vitro microfold cell signature (P < 10^-4, Extended Data Fig. 8g). Using this cluster, we defined an in vivo signature of markers and transcription factors (Fig. 5d, Methods).

The microfold cell signature confirmed that microfold cells from Peyers patches were too rare to detect without specific enrichment of follicle-associated epithelia (only 1 of 7,216 cells in our initial sampling (Fig. 1b) was positive). Discovering any other, as yet unknown, subsets of rare cells would require additional stratification.

**Epithelial response to pathogen infection**

The responses of immune and epithelial cells to pathogens have a key role in maintaining gut homeostasis. We investigated the responses of intestinal epithelial cells to *Salmonella enterica* and to the parasitic helminth *Heligmosomoides polygyrus*. We profiled individual intestinal epithelial cells using droplet-based scRNA-seq two days after *Salmonella* infection (n = 2 mice, 1,770 cells) and three days (n = 2 mice, 2,121 cells) and ten days (n = 2 mice, 2,711 cells) after *H. polygyrus* infection, and matched the results against controls (n = 4 mice, 3,240 cells). We also profiled 389 cells with full-length scRNA-seq. The response to each pathogen incorporated pathogen-specific and pathogen-shared changes in expression and shifts in cell proportions and cell-intrinsic programs.

*Salmonella*-induced genes across all infected intestinal epithelial cells (FDR < 0.25, likelihood-ratio test; Extended Data Fig. 9a, top left, Supplementary Table 9) were enriched for genes encoding elements of pathways involved in the defence response to a bacterium polypeptide: A, ghrelin. d, smFISH of Cck (green; top left), Ghrl (red; top right) and Gcg (white; bottom right), and the combined image (bottom right). Scale bars, 50 μm.

Figure 3 | Enteroendocrine cell taxonomy. a, Unsupervised clustering. t-SNE of 533 enteroendocrine cells (n = 8 mice) is shown, coloured by subcluster. EC, enterochromaffin cell; P, peptide YY-expressing; other cell subtypes are labelled using the notation outlined in section ‘Characterizing subsets of enteroendocrine cells’. b, Signatures of enteroendocrine cell subtypes. The relative expression of subtype-enriched genes (FDR < 0.01; rows) across cells (columns) is shown. Grey bars, elements of pathways involved in the defence response to a bacterium and cell-intrinsic programs.
Figure 4 | CD45+ tuft-2 cells express Tslp. a, Tuft cell subsets. t-SNE of 166 tuft cells (n = 6 mice) is shown, coloured by subcluster. b, Tuft-1 and tuft-2 gene signatures. The relative expression (in droplet-based data) of the top 25 genes (rows) for tuft-1 and tuft-2 cells (columns) is shown (FDR < 0.01 in both datasets). c, Tuft-2 cells express Tslp. The distribution of the expression of Il25 and Tslp in enterocytes (E), but-SSNE of 166 tuft cells (enriched (FDR or newly identified (black) markers (left) or transcription factors (right), associated epithelia cluster (rows) of genes (columns) for known (grey) or newly identified (black) markers (left) or transcription factors (right), enriched (FDR < 0.05; Mann–Whitney U test). d–g, Tuft-2 cells express Ptprc (CD45). (FDR < 0.001, hypergeometric test; Extended Data Fig. 9c), including Reg3b and Reg3g (ref. 39)—protective genes in Salmonella infection (Fig. 6c). Most H. polygyrus-induced genes (62%) were specific to this pathogen and enriched for inflammatory-response genes and tuft cell markers (FDR < 0.25, likelihood-ratio test; Extended Data Fig. 9a, bottom, Supplementary Table 9). Other induced genes (112 of 571; 20%) comprised a non-specific, shared inflammatory response (FDR < 0.25, likelihood-ratio test; Extended Data Figs 9a, 10a middle panels, Supplementary Table 9). Stress gene modules were also upregulated in stem cells after both Salmonella and day-10 helminth infection (FDR < 0.05; data not shown).

Additional responses to Salmonella were cell-type-specific: expression of antimicrobial peptides and Mptx2 was increased in Paneth cells (Extended Data Fig. 9f); 40 genes were induced in enterocytes, mostly (65%) in a Salmonella-specific manner (Extended Data Fig. 9d, Methods), including the pattern-recognition receptor Nlrp6; and the pro-inflammatory apolipoproteins serum amyloid A1 and A2 (encoded by Saa1 and Saa2) were induced in distal enterocytes (Extended Data Fig. 9a, e). Some antimicrobial peptides, such as Reg3a, Reg3b and Reg3d, which are normally enterocyte-specific, were induced in all cell types after Salmonella infection (Fig. 6c, Extended Data Fig. 9b, Supplementary Tables 2, 3, 9).

We distinguished the contribution of changes in cell-intrinsic expression programs from that of shifts in cell composition (determined by unsupervised clustering; Fig. 6a, b). After Salmonella infection, the frequency of mature enterocytes increased substantially (from 13.1% on average in control to 21.7% in infection; Fig. 6b), whereas the proportion of transit-amplifying (52.9% to 18.3%) and stem (20.7% to 6.4%) cells decreased significantly (FDR < 10−10). In agreement with a previous study, the proportions of mature Paneth cells also increased significantly (from 1.1% to 2.3%; FDR < 0.01). (We used another 2,029 cells with sorting optimized to avoid loss of the large Paneth cells; see Methods, Extended Data Fig. 9f, g; n = 4 infected mice.)
subset (FDR < 0.05; Extended Data Fig. 6d), the cells that co-express Ghrl and Gcg genes that encode gut hormones that regulate appetite and satiety. Tuft cells were also enriched for GPCR expression, supporting the results of studies on their specialized chemosensory properties.

Although many studies have demonstrated that goblet cells and, more recently, that tuft cells accumulate and respond to parasites, they are cell-type-specific in homeostatic conditions was broadened across multiple cell types during infection. Overall, our study provides a reference dataset, including cell-type-specific markers, transcription factors and GPCRs, and specific hypotheses for follow-up studies, which could potentially lead to new interventions in inflammatory, metabolic and proliferative gut pathologies.

During infection with H. polygyrus there was a marked increase in the number of goblet cells, which are known to respond to the parasite. A similar increase in goblet cells, which are known to respond to the parasite, was also observed in the tuft cell population by day 10 (17.2% to 43.0%; FDR < 0.05; Wald test). Error bars, s.e.m.

**Discussion**

The intestinal epithelium is the most diverse epithelial tissue in the body. Our high-resolution single-cell survey of the mouse intestinal epithelium reveals further diversity, as well as coherent cell-specific transcriptional programs, which we validated in situ and in prospectively isolated cells. One example of new cellular diversity that we identified in this study is the tuft cell subtypes, one expressing neuron-related and the other Tuft2-recruiting epithelial cytokines, which could provide insight into the mechanisms that underlie food allergies. In addition, CD45 expression by rare epithelial cells highlights the necessity of using a set of markers to identify cell populations unambiguously.

Our survey resolved the cellular populations that are implicated in key sensory pathways at high resolution. For example, we provide a detailed profile of the GPCRs that are expressed by intestinal epithelial cells, including subsets of endocrine cells. Notably, the cannabinoid receptor21 Gpr119 was enriched in the newly identified SILA...
22. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl Acad. Sci. USA 102, 7426–7431 (2005).

23. Basak, O. et al. Mapping early fate determination in Lgr5+ crypt stem cells using a novel K67-RFP allele. EMBO J. 33, 2057–2068 (2014).

24. Beuling, E. et al. GATA factors regulate proliferation, differentiation, and gene expression in small intestine of mature mice. Gastroenterology 140, 1219–1229.e2 (2011).

25. Furness, J. B., Rivera, L. R., Cho, H. J., Bravo, D. M. & Callaghan, B. The gut as a sensory organ. Nat. Rev. Gastroenterol. Hepatol. 10, 729–740 (2013).

26. Worthington, J. J., Reimann, F. & Grubbie, F. M. Enteronecrocin cells-sensory sentinels of the intestinal environment and orchestrators of mucosal immunity. Mucosal Immunol. (2017).

27. Habib, A. M., Richards, P., Rogers, G. J., Reimann, F. & Grubbie, F. M. Co-localisation and secretion of glucagon-like peptide 1 and peptide YY from primary cultured human L cells. Diabetologia 56, 1413–1416 (2013).

28. Gershon, M. D. & Tack, J. The serotonin signaling system: from basic understanding to drug development for functional GI disorders. Gastroenterology 132, 397–414 (2007).

29. Klok, M. D., Jekelobdottir, S. & Drent, M. L. The role of leptin and ghrelin in the regulation of food intake and body weight in humans: a review. Obes. Rev. 8, 21–34 (2007).

30. Karra, E., Chandarana, K. & Batterham, R. L. The role of peptide YY in appetite regulation and obesity. J. Physiol. 587, 19–25 (2009).

31. Gerbe, F. et al. Intestinal tuft cells: epithelial sentinels linking luminal cues to the immune system. Mucosal Immunol. 9, 1353–1359 (2016).

32. Gerbe, F. et al. Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. Nature 529, 226–230 (2016).

33. Bezençon, C. et al. Murine intestinal cells expressing Trpm5 are mostly brush cells and express markers of neuronal and inflammatory cells. J. Comp. Neurol. 509, 514–525 (2008).

34. Biton, M. et al. Epithelial microRNAs regulate gut mucosal immunity via epithelium–T cell crosstalk. Nat. Immunol. 12, 239–246 (2011).

35. de Lau, W. et al. Peyer’s patch M cells derived from Lgr5+ stem cells require SpiB and are induced by RankL in cultured “miniguts”. Mol. Cell. Biol. 32, 3639–3647 (2012).

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

37. Terahara, K. et al. Comprehensive gene expression profiling of Peyer’s patch M cells, villous M-like cells, and intestinal epithelial cells. J. Immunol. 180, 7840–7846 (2008).

38. Peterson, L. W. & Artis, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. Nat. Rev. Immunol. 14, 141–153 (2014).

39. Lopes, L. M. et al. REG3–deficient mice have altered mucus distribution and increased mucosal inflammatory responses to the microbiota and enteric pathogens in the ileum. Mucosal Immunol. 7, 939–947 (2014).

40. Eckhardt, E. R. et al. Intestinal epithelial serum amyloid A modulates bacterial growth in vitro and pro-inflammatory responses in mouse experimental colitis. BMC Gastroenterol. 10, 133 (2010).

41. Martinez Rodriguez, N. R. et al. Expansion of Paneth cell population in response to enteric Salmonella enterica serovar Typhimurium infection. Infect. Immun. 80, 266–275 (2012).

42. Artis, D. et al. RELMbeta/FIZZ2 is a goblet cell-specific immune-effector molecule in the gastrointestinal tract. Proc. Natl Acad. Sci. USA 101, 13596–13600 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Gaffney for help with figure preparation, the Broad Flow Cytometry Facility (P. Rogers, S. Saldi and C. Otis), C. Hafemeister and R. Satija for use of the ‘How Many Cells’ tool, S. Riesenfeld and A. Dixit for statistical advice, and T. Tickle for help with the Single Cell Portal. This study was supported by the Klarmann Cell Observatory at the Broad Institute, NIH RC2DK114784 (A.R. and R.J.X.), HHMI (A.R.), Food Allergy Science Initiative (FASI) at the Broad Institute (A.R. and R.J.X.), and a Broadnxt10 award (A.R. and R.J.X.). M.B. is supported by a postdoctoral fellowship from the Human Frontiers Science Program (HFSP). R.J.X. is supported by NIH DK43351, DK097485 and the Helmsley Charitable Trust.

Author Contributions A.L.H., M.B. and N.R. contributed equally to this study; M.B., R.J.X. and A.R. co-conceived the study; M.B., N.R., A.L.H., R.J.X. and A.R. designed experiments and interpreted the results; N.R. and M.B. carried out all experiments; G.B., T.M.D., M.R.H., S.B., D.D., M.Z. and R.R. assisted with experiments; A.L.H. designed and performed computational analysis with assistance from R.H.H., K.S., C.S., Y.K., I.T. and A.R.; M.R.H. and W.S.G. assisted with tuft and follicle-associated epithelium experiments; M.Z. and H.N.S. assisted with pathogen infections; S.B. and O.Y. assisted with epithelial cell sorting; D.D. and O.R.-R. assisted with scRNA-seq; and A.L.H., M.B., N.R., R.J.X. and A.R. wrote the manuscript, with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.B. (mibton@broadinstitute.org), R.J.X. (xavier@molbio.mgh.harvard.edu) and A.R. (aregev@broadinstitute.org).

Reviewer Information Nature thanks L. Vermeulen and the other anonymous reviewer(s) for their contribution to the peer review of this work.
METHODS

Mice. All mouse work was performed in accordance with the Institutional Animal Care and Use Committees (IACUC) and with relevant guidelines at the Broad Institute and Massachusetts Institute of Technology, with protocols 0055-05-15 and 0612-05-18, respectively. For all experiments, mice were randomly assigned to treatment groups after matching for the sex and age of 7–10-week-old female or male wild-type C57BL/6J or Lgr5–EGFP-IRESCreER<sup>T2</sup> (Lgr5–EGFP) mice, obtained from the Jackson Laboratory (Bar Harbour) or Gfi1b<sup>BFP</sup>– (Gfi1b–GFP) mice, under specific-pathogen-free conditions at the animal facilities at the Broad Institute, Massachusetts Institute of Technology or Harvard T. H. Chan School of Public Health. *Salmonella enterica* and *H. polyergus* infection. C57BL/6J mice (Jackson Laboratory) were infected with 200 third-stage larvae of *Salmonella enterica* or *H. polyergus* and were sorted into an Eppendorf tube containing 50 μl of blood to perform transcriptomic analysis. *Salmonella enterica* maintained under specific-pathogen-free conditions at Massachusetts General Hospital (Charlestown), with protocol 2003N000158. *H. polyergus* was propagated as previously described<sup>44</sup>. Mice were euthanized 3 and 10 days after *H. polyergus* infection. For *Salmonella enterica*, mice were infected with a naturally streptomycin-resistant SL3245 strain of *S. Typhimurium* (10<sup>7</sup> cells) as described previously<sup>44</sup>, and were euthanized 48 h after infection.

Cell dissociation and cryopreservation. Cryopreservation of intestinal epithelial cells was achieved by isolating small sections (0.2–0.5 cm) containing Peyers’ patches from the small intestine of C57Bl/6J or Gfi1b<sup>BFP</sup>–/– mice.

**Cell sorting.** For plate-based single-cell scRNA-seq experiments, a FACs machine (Astro) was used to sort a single cell into each well of a 96-well PCR plate containing 5 μl of PCR buffer with 1% 2-mercaptoethanol. For EpCAM<sup>+</sup> isolation, cells were stained for 7AAD<sup>−</sup> (Life Technologies), CD45<sup>−</sup> (eBioscience), CD31<sup>−</sup> (eBioscience), TER-119<sup>−</sup> (eBioscience), EpCAM<sup>+</sup> (eBioscience); for specific epithelial cells, we also stained for CD24<sup>−1</sup> (eBioscience) and c-KIT<sup>−</sup> (eBioscience). To enrich for specific intestinal epithelial cell populations, cells were isolated from Lgr5–EGFP mice, stained with the antibodies mentioned above and gated on GFP<sup>−</sup> (stem cells), GFP<sup>−</sup> (Tops), GFP<sup>−</sup>CD24<sup>−</sup> (EpCAM<sup>−</sup> secretory lineage) or GFP<sup>−</sup>CD24<sup>−</sup>EpCAM<sup>+</sup> (epithelial cells). For better Paneth cell recovery, we allowed higher side scatter and forward scatter parameters in combination with CD45<sup>−</sup>c-KIT<sup>−</sup> to verify Paneth cell recovery in EpCAM<sup>+</sup> cells. For tuft-2 isolation, epithelial cells from three different mice were stained as above, but using EpCAM<sup>−</sup>CD45<sup>−</sup> to sort 2,000 single cells. We used a lenient sorting gate to ensure that we obtained sufficient numbers of these rare tuft-2 cells, which led to a higher contamination rate of T cells, which we removed in our single-cell analysis using unsupervised clustering.

**Fluoroscopically isolated epithelia isolation.** The 96-well plate was soaked tightly with a Microsolv F and centrifuged at 800g for 1 min. The plate was immediately frozen on dry ice and kept at −80°C until ready for the lysate clean-up. Bulk population cells were sorted into an Eppendorf tube containing 100 μl solution of TCL with 1% 2-mercaptoethanol and stored at −80°C.

For droplet-based scRNA-seq, cells were sorted with the same parameters as for plate-based scRNA-seq, but were sorted into an Eppendorf tube containing 50 μl of 0.4% BSA–PBS and stored on ice until proceeding to the GemCode single-cell platform.

**Plate-based scRNA-seq.** Single cells. Libraries were prepared using a modified SMART-Seq<sup>2</sup> protocol<sup>45</sup>. In brief, RNA lysate clean-up was performed using RNAclean XP beads (Agencourt) followed by reverse transcription with Maxima Reverse Transcriptase (Life Technologies) and whole-transcription amplification (WTA) with Kapa HotStart HIFI 2× ReadyMix (Kapa Biosystems) for 21 cycles. WTA products were purified with Ampure XP beads (Beckman Coulter), quantified with Qubit dsDNA HS Assay Kit (ThermoFisher) and assessed with a high-sensitivity DNA chip (Agilent). RNA-seq libraries were constructed from purified WTA products using Nextera XT DNA Library Preparation Kit (Illumina). On each plate, the population and no-cell controls were processed using the same method as for the single cells. The libraries were sequenced on an Illumina NextSeq 500.

**Droplet-based scRNA-seq.** Single cells were processed through the GemCode Single Cell Platform using the GemCode Gel Bead, Chip and Library Kits (10X Genomics, Pleasanton) as per the manufacturer’s recommendations, and then proceeding with the modified SMART-Seq2 protocol following lysate clean-up. Droplet-based scRNA-seq.

**Combined immunofluorescence and smFISH.** Immunofluorescence. Staining of small intestinal tissues was conducted as described previously<sup>44</sup>. In brief, tissues were fixed for 14 h in formalin, embedded in paraffin and cut into 5-μm-thick sections. Sections were deparaffinized using standard techniques, incubated with primary antibodies overnight at 4°C and then with secondary antibodies at room temperature for 30 min. Slides were mounted with Slowfade Mountant + DAPI (Life Technologies, S36964) and sealed.

**smFISH.** An RNAscope Multiplex Fluorescent Kit (Advanced Cell Diagnostics) was used as per the manufacturer’s recommendations with the following alterations. The target retrieval boiling time was adjusted to 12 min and incubation with Protease IV at 40°C was adjusted to 8 min. Slides were mounted with Slowfade Mountant + DAPI (Life Technologies, S36964) and sealed.

**Image analysis.** Images of tissue sections were taken with a confocal microscope (Fluoview FV1200 using Kalman and sequential laser emission to reduce noise and signal overlap. Scale bars were added to each image using the confocal software FV-10-ASW 3.1 Viewer. Images were overlaid and visualized using Image J software<sup>45</sup>.

**Antibodies and probes.** Antibodies used for immunofluorescence: rabbit anti-DCLK1 (1:200, Abcam ab31704), rat anti-CD45 (1:100, Biolegend 30-911), goat anti-ChgA (1:100, Santa Cruz Sc-1488), mouse anti-E-cadherin (1:100, BD Biosciences 610181), rabbit anti-RELJ<sup>−</sup> (1:200, Peprotech 500-p215), rat anti-lysozyme (1:200, Dako A0099), Lct-2, Fabp-6-c, and Alexa Fluor 488-, 594- and 647-conjugated secondary antibodies, obtained from Life Technologies.

**Probes used for single-molecule RNAscope (Advanced Cell Diagnostics):** Cck (C1), Ghrl (C2), Gg (C3), Tph1 (C1), Retg (C2), Tgf (C1), Ptprc (C1) and Mptz2 (C1).

**Intestinal organoid cultures.** Following cryopreservation, the single-cell suspension was resuspended in Matrigel (BD Bioscience) with 1 μM Jagged-1 peptide (Ana-Spec). Roughly 300 crypts embedded in 25 μl of Matrigel were seeded onto each well of a 24-well plate. Once solidified, the Matrigel was incubated in 500 μl of mTeSR1 medium culture (Advanced DMEM/F12, Invitrogen) with streptomycin/penicillin and glutamax and supplemented with EGF (100 ng ml<sup>−1</sup>, Peprotech), R-spondin-1 (600 ng ml<sup>−1</sup>, Peprotech), Dibutyryl-cAMP (25 μM, Tochris), Y-27643 (10 μM, CalBiochem), noggin (100 ng ml<sup>−1</sup>, R&D), R-spondin-1 (600 ng ml<sup>−1</sup>, R&D), noggin (100 ng ml<sup>−1</sup>, Prepotech), V-SNAP (1 ng ml<sup>−1</sup>, Peprotech), Dibutyryl-cAMP (25 μM, R&D Systems). Fresh media was replaced on day 3, and organoids were passaged by dissociation with TryplE and re-suspended in new Matrigel on day 6 with a 1:3 ratio split. For the following experiments, organoids were additionally treated with RANKL (100 ng ml<sup>−1</sup>, Biolegend). Treated organoids were dissociated and subjected to scRNA-seq using both methods.

**Quantitative PCR.** cDNA of 16 whole-transcriptome-amplified single cells of tuft-1, tuft-2 and random EpCam<sup>+</sup> from the full-length-based scRNA-seq plates were used for the relative qPCR. Gene expression was analysed by quantitative real-time PCR on a LightCycler 480 Instrument II (Roche) using LightCycler 480 SYBR Green master mix (Roche) with the following primer sets: HDTPI<sup>−</sup>-F, GTTAAACAGCTAGACCACCCAAA; HDTPI<sup>−</sup>-R, ACGGCATATCACAACAACTT; UBCF<sup>−</sup> F, CAGCGCTGATCTTCCCAAGCAT; UBCF<sup>−</sup> R, CTGCA GGGTGCGGTTAAT; UBCF<sup>−</sup>-F, TACTCTCAATCCTATCCCTGGCTG; TGFβ-R<sup>−</sup>, CAATTTCTGAGTACCGCTATCCTTC; ALPI<sup>−</sup> R, CTCTACACCTCATTCTCTACT; TCTATGGG, ALPI<sup>−</sup> R, CCAGGCTGCTGTGTTTAG; Dclk1-F, GGTTGAGAACCA
cells were observed in the 10X dataset. These 264 cells were removed by an initial pre-processing of droplet-based scRNA-seq data. De-multiplexing, alignment to the mm10 transcriptome and unique molecular identifier (UMI)-collapsing were performed using the Cellranger toolkit (version 1.0.1) provided by 10X Genomics. For each cell, we quantified the number of genes for which at least one read was mapped, and then excluded all cells with fewer than 800 detected genes. Expression values $E_i$ for gene $i$ in cell $c$ were calculated as UMI counts for gene $i$ by the sum of the UMI counts in cell $c$, to normalize for differences in coverage, and then multiplying by 10,000 to create TPM-like values, and finally computing log(TPM + 1). Batch correction was performed using ComBat as implemented in the R package sva17, using the default parametric adjustment mode. The output was a corrected expression matrix, which was used as an input to further analysis.

Selection of variable genes was performed by fitting a generalized linear model to the relationship between the squared coefficient of variation and the mean expression level in logarithmic space, and selecting genes that deviated significantly ($P < 0.05$) from the fitted curve.41

Pre-processing of SMART-Seq2 scRNA-seq data. BAM files were converted to merged, de-multiplexed FASTQs using the Illumina-provided Bcl2Fastq software package v2.17.1.14. Paired-end reads were mapped to the UCSC mm10 mouse transcriptome using Bowtie with parameters `-q -p 33-quals -n 1 -e 999999999 -I 125 -1 I 1 -X 2000 -a -m 15 -S 6 -p 6`, which allows alignment of sequences with one mismatch. Expression levels of genes were quantified using TPM values calculated by RSEM15,17 v1.2.3 in paired-end mode. For each cell, we quantified the number of genes for which at least one read was mapped, and then excluded all cells with either fewer than 3,000 detected genes or a transcriptome mapping of less than 40%. We then identified highly variable genes as described above.

Dimensionality reduction using PCA and t-distributed stochastic neighbor embedding (t-SNE). We restricted the expression matrix to the subsets of variable genes and high-quality cells noted above, and then centred and scaled values before inputting them into principal component analysis (PCA), which was implemented using the R function `prcomp` from the stats package for the SMART-seq2 dataset. For the droplet-based dataset we used a randomized approximation to PCA, implemented using the `rpack` function from the rsvd R package, with the parameter `k` set to 100. This low-rank approximation was used because it is several orders of magnitude faster to compute for very wide matrices. Given that many principal components explain very little of the variance, the signal-to-noise ratio can be improved substantially by selecting a subset of $n$ 'significant' principal components. After PCA, significant principal components were identified using the permutation test51, implemented using the `permutationPM` function from the jackstraw R package. This test identified 13 and 15 significant principal components in the 10X and SMART-seq2 datasets of Fig. 1b and Extended Data Fig. 2a, respectively. Scores from only these significant principal components were used as the input to further analysis.

For visualization, the dimensionality of the datasets was further reduced using the ' Barnes-hut' approximate version of t-SNE52,53. This was implemented using the `Rtsne` function from the Rtsne package using 20,000 iterations and a perplexity setting that varied from 10 to 30 depending on the size of the dataset. Identifying cell differentiation trajectories using diffusion maps. Before running diffusion-map dimensionality reduction we selected highly variable genes in the gut dataset (7,216 cells; Fig. 1b) showed a cluster of 310 EEC cells and 166 tuft cells. The initial clustering of the whole-gut dataset (7,216 cells, Fig. 1b) showed a cluster of 310 EEC cells and 166 tuft cells. The tuft cells were taken as 'is' for the subanalysis (Fig. 4a, b), whereas the EEC cells were combined with a second cluster of 239 EEC cells that were identified in the regional dataset (Fig. 2a, right) for a total of 549 EEC cells. A group of 16 cells co-expressed EEC markers Chga and Chgb with markers of Paneth cells, including Lys1, Defa2 and Defa22, and were therefore interpreted as doublets and removed from the analysis, leaving 533 EEC cells, which were the basis for the analysis in Fig. 3. To compare expression profiles of enterocytes from the proximal and distal small intestine (Fig. 2b), the 1,041 enterocytes identified from 11,665 cells in the regional dataset (Fig. 2a) were used. Defining cell-type signatures. To identify maximally specific genes for cell types, we performed differential tests between each pair of clusters for all possible pairwise comparisons. Then, for a given cluster, putative signature genes were filtered using the maximum FDR $Q$ value and ranked by the minimum log(fold change). The minimum fold change and maximum $Q$ value represent the weakest effect size across all pairwise comparisons; it is therefore a stringent criterion.

Cell-type signature genes shown in Fig. 1c, Extended Data Fig. 2b, Extended Data Fig. 8e and Supplementary Tables 2–4 and 8 were obtained using a maximum FDR of 0.05 and a minimum log(fold change) of 0.5. In the case of post-mitotic cell-type signatures, all genes passed this threshold in both 3' (Fig. 1c) and full-length (Extended Data Fig. 2b) datasets. In the case of signature genes for subtypes within cell types (Fig. 3b, Fig. 4b, Extended Data Fig. 7b), a combined $P$ value (across the pairwise tests) for enrichment was computed using Fisher's method—a more lenient criterion than simply taking the maximum $P$ value—and a maximum FDR $Q$ value of 0.01 was used, along with a cut-off of minimum log(fold change) of 0.25 for tuft cell subtypes (Fig. 4b, Extended Data Fig. 7b, Supplementary Table 7) and of 0.1 for EEC subtypes.
For the assessment of the significance of spatial distributions of EEC subsets (Fig. 3e), the comparison involved more than two groups. In particular, our null hypothesis was that the proportion of each EEC subset detected in the three intestinal regions (duodenum, jejunum, and ileum) was equal. To test this hypothesis, we used analysis of variance (ANOVA) with a \( \chi^2 \) test on the Poisson model fit described above, implemented using the \texttt{anova} function from the \texttt{stats} package.

**Gene set enrichment and gene ontology analysis.** Gene ontology analysis was performed using the \texttt{goseq} R package\(^\text{59}\), using significantly differentially expressed genes (FDR < 0.05) as target genes, and all genes expressed with \( \log(\text{TPM} + 1) > 3 \) in at least ten cells as background.

**Data availability.** All data are deposited in the Gene Expression Omnibus (GEO; GSE92332) and in the Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/single-intestinal-epithelium).

**Code availability.** R markdown scripts enabling the main steps of the analysis to be performed are available from the corresponding authors on request.

43. Vassen, L., Okayama, T. & Möröy, T. Gfi1b green fluorescent protein knock-in mice reveal a dynamic expression pattern of Gfi1b during hematopoiesis that is largely complementary to Gfi1. Blood 109, 2356–2364 (2007).

44. Su, L. et al. Coinfection with an intestinal helminth impairs host innate immunity against Salmonella enterica serovar Typhimurium and exacerbates intestinal inflammation in mice. Infect. Immun. 82, 3895–3896 (2014).

45. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years and powerful approach to multiple testing. Nat. Methods 9, 671–675 (2012).

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

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

48. Brennecke, R. et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095 (2013).

49. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

50. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

51. Buja, A. & Eyuboglu, N. Remarks on parallel analysis. Multivariate Behav. Res. 27, 509–540 (1992).

52. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. 15, 3221–3245 (2014).

53. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

54. Zelnik, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015).

55. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional and noisy data. Nat. Methods 12, 1129–1131 (2015).

56. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

57. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

58. Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. Science 344, 1492–1496 (2014).

59. Fink, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 16, 278 (2015).

60. Benjamin, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300 (1995).

61. Zhang, H.-M. et al. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res. 40, D144–D149 (2012).

62. Ng, A., Eisenberg, J. M. & Heath, R. Human leucine-rich repeat proteins: a genome-wide bioinformatic catargorization and functional analysis in innate immunity. Proc. Natl Acad. Sci. 108, 4631–4638 (2011).

63. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology enrichment from p-values. Bioinformatics 23, 2700–2705 (2007).

64. Ichimura, A., Hirasawa, A., Hara, T. & Tsujimoto, G. Free fatty acid receptors act as ligand gated ion channels in the mouse ileum. Gut 56, 184–190 (2007).

65. Rubin, D. B. The Bayesian bootstrap. Ann. Stat. 9, 130–134 (1981).

66. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

67. Datta, R. et al. Identification of novel genes selectively expressed in the intestine of mice reveal a dynamic expression pattern of Gfi1b during hematopoiesis that is largely complementary to Gfi1. Blood 109, 2356–2364 (2007).
Extended Data Figure 1 | See next page for caption.
Extended Data Figure 1 | Identifying intestinal epithelial cell types in scRNA-seq data by unsupervised clustering. Related to Fig. 1.

a, b, Quality metrics for scRNA-seq data. Shown are distributions of the number of reads per cell (left), the number of genes detected with non-zero transcript counts per cell (centre) and the fraction of reads mapping to the mm10 mouse transcriptome per cell (right) in the droplet-based 3′ scRNA-seq data (a) and the plate-based full-length scRNA-Seq data (b).

c–f, Agreement across batches. c, Contribution of batches to each cluster. Each pie chart shows the batch composition (colour-coded legend) of each detected cluster (labelling and number of cells are marked above each pie chart) in the droplet-based 3′ scRNA-seq dataset (n = 6 mice). All ten replicates contribute to all clusters, and no major batch effect is observed. d, Cell type proportions across batches. Shown is the proportion of detected cells in each major cell type in the droplet-based 3′ scRNA-seq dataset in each of ten batches (points; n = 6 mice). Grey bar, mean; error bars, s.e.m. e, Agreement in expression profiles across mice. Box and whisker plot shows the Pearson correlation coefficients in average expression profiles (average log2(TPM + 1)) for cells in each cluster, across all pairs of mice. Black bar, median value; box edges, 25th and 75th percentiles; whiskers, a further 1.5 times the interquartile range. Clusters with additional subtypes (such as tuft and EEC cells) show more variation, as expected. f, Scatter plots comparing the average log2(TPM + 1) gene expression values between two scRNA-seq experiments from the droplet-based 3′ scRNA-seq dataset (top), between two scRNA-seq experiments from the plate-based full-length scRNA-Seq dataset (centre), or between the average of a plate-based full-length scRNA-seq and a population control (bottom). Pearson correlation is marked top left.

g, Additional quality control metrics and cluster annotation on the basis of the expression of known cell type markers. t-SNE visualization of 7,216 single cells is shown, where individual points correspond to single cells. Cells are coloured by, from top left to bottom right, their assignment to clusters using a k-nearest-neighbour graph-based algorithm (Methods; legend shows the cluster type) (identical to Fig. 1b), mean expression (log2(TPM + 1)) of several known marker genes for a particular cell type or state (indicated above each plot), the mouse from which they originate (see legend), the number of reads per cell, the number of genes detected per cell, or the number of transcripts as measured by UMIs per cell.
Extended Data Figure 2 | See next page for caption.
Extended Data Figure 2 | Identification and characterization of intestinal epithelial cell types in plate-based full-length scRNA-seq data by unsupervised clustering. Related to Fig. 1. a, Quality control metrics and cluster annotation on the basis of the expression of known cell type markers. t-SNE visualization of 1,522 single cells (n = 8 mice) is shown, where individual points correspond to single cells. Cells are coloured by, from top left to bottom right, their assignment to clusters, mean expression (log(\text{TPM} + 1)) of several known marker genes for a particular cell type or state (indicated above each plot; same as in Extended Data Fig. 1g), the mouse from which they originate (see legend) and its genotype, the FACS gate used to sort them (see legend), the number of reads per cell, or the number of genes detected per cell. b, Cell-type-specific signatures. Heatmap shows the relative expression level (row-wise Z scores) of genes (rows) in cell-type-specific signatures (same genes as in Fig. 1c, with the exception of enterocyte markers), across the individual post-mitotic intestinal epithelial cells (columns) in the full-length scRNA-seq data. Colour code marks the cell type and their associated signatures. c, Mptx2 is a novel Paneth cell marker. t-SNE of the cells from the droplet-based 3’ scRNA-seq (left; as in Fig. 1b) and plate-based full-length scRNA-seq (right; as in a) datasets is shown, coloured by expression (log(\text{TPM} + 1)) of the mucosal pentraxin Mptx2. d, Cell-type-enriched GPCRs. Heatmap shows the relative expression (row-wise Z scores) of genes encoding GPCRs (rows) that are significantly (FDR < 0.001; Mann–Whitney U test; Methods) upregulated or downregulated in the cells (columns) of a given cell type (top; colour coded as in a) compared to all other cells, in the plate-based full-length scRNA-seq data. e, Cell-type-specific leucine-rich repeat proteins (LRRs). Heatmap depicts the mean relative expression (column-wise Z score of mean log(\text{TPM} + 1)) values of genes (columns) encoding leucine-rich repeat proteins that are significantly (FDR < 0.001; Mann–Whitney U test) upregulated or downregulated in a given cell type (rows) compared to all other cells, in the plate-based full-length scRNA-seq data. f, Cell-type transcription factors and GPCRs. Average relative expression (Z score of mean log(\text{TPM} + 1); colour scale) of the top ten transcription factors (left) and GPCRs (right) (columns) enriched in each cell type (rows).
Extended Data Figure 3 | Regional variation in Paneth cell subtypes and stem cell markers. 

a, Paneth cell subsets. t-SNE of 10,396 single cells (points) was obtained using a large cell-enriched protocol (Methods), coloured by cluster annotation (n = 2 mice). b, Paneth cell subset markers. Shown is the expression (row-wise Z score; colour scale) of genes specific (FDR < 0.05; Mann–Whitney U test; log2(fold change) > 0.5) to each of the two Paneth cell subsets (average of 724.5 cells per subtype, down-sampled to 500 for visualization) shown in a. c, Two Paneth subsets reflect regional diversity. Shown is the expression of the same genes (rows) as in b in Paneth cells from each of three small intestinal regions (average of 176.3 cells per region; columns; Fig. 2a). 11 of 11 Paneth-1 markers are enriched in the ileal Paneth cells, whereas 7 of 10 Paneth-2 markers are enriched in duodenal or jejunal Paneth cells (FDR < 0.05; Mann–Whitney U test). d, Validation of regional enterocyte markers. Shown is smFISH of Lct (red) and Fabp6 (white) in the duodenum (proximal; left) and ileum (distal; right). Dotted line, boundary between crypt and villi; green and yellow arrows, proximal and distal enterocytes, respectively; scale bars, 50 μm. e, Regional variation of intestinal stem cells. Expression (row-wise Z score) of genes specific to stem cells from each intestinal region (FDR < 0.05; Mann–Whitney U test; log2(fold change) > 0.5). On average, 1,226.3 cells were obtained from each of the three regions, down-sampled to 500 for visualization (columns).
Extended Data Figure 4 | Differentiation from stem cells to mature enterocytes. a–d. Diffusion-map embedding of 5,282 cells (points) progressing through stages of enterocyte differentiation (Methods). a, b. Cells are coloured by their cluster assignment (Fig. 1b). Diffusion components 1 and 3 (DC-1 and DC-3) are associated with the transition from stem cells to progenitors (a), whereas DC-2 distinguishes between proximal and distal enterocyte fate commitment (b). c, d. Cells are coloured by the expression (log_2(TPM + 1)) of known and newly identified transcription factors associated with stages of differentiation (c), or with proximal or distal enterocyte differentiation (d). e. Transcription factors that are differentially expressed between proximal and distal cell fate.

Heatmap shows the mean expression level of 44 transcription factors differentially expressed between the proximal and distal (rows) enterocyte clusters of Fig. 1b (FDR < 0.05; Mann–Whitney U test). f. Newly identified regional stem cell markers (Extended Data Fig. 3e) identify distinct populations in diffusion-map space. Shown are close-ups of the stem-cell region in diffusion space (b, inset square), coloured by expression level (log_2(TPM + 1)) of pan-stem cell marker Lgr5 (left), proximal stem cell marker Gkn3 (centre) or distal stem cell marker Bex1 (right). Dashed line helps to visualize separation of stem cells on the basis of region-specific markers.
Extended Data Figure 5 | Heterogeneity within EEC cells. Related to Fig. 3. a, EEC subset discovery and regional location. Shown is the t-SNE of the 533 EEC cells identified from the droplet-based datasets for whole small intestine (SI) and regional samples (colour legend; n = 8 mice; Methods).b, Agreement in hormone detection rates between droplet-based 3′ and full-length scRNA-seq. Scatter plot shows the detection rate (fraction of cells with non-zero expression of a given transcript) for a set of known EEC hormones, transcription factors and marker genes (see legend) in EEC cells from the full-length dataset, and from the droplet-based 3′ dataset. Linear fit (dashed line) and 95% confidence interval (shaded) are also shown. c, Expression of key genes across subset clusters. t-SNE plot shows cells coloured by their assignment to the 12 clusters (top left; identical to Fig. 3a) or by the expression (log2(TPM + 1)) of markers of immature EEC cells (Neurog3), genes encoding gut hormones (Sct, Sst, Cck, Gcg, Ghr, GIP, Nts, PYY) or markers of enterochromaffin cells (Tac1, Reg4). d, Co-expression of gastrointestinal hormones by individual cells. Left, heatmap shows the expression of canonical gut hormone genes (rows) in each of 533 individual EEC cells (columns), coloured on the basis of their assignment to the clusters in Fig. 3a (top). Right, heatmap shows for each cluster (columns) the percentage of cells (inset text) in which the transcript for each hormone (rows) is detected.
Extended Data Figure 6 | Classification and specificity of EEC subsets. Related to Fig. 3. **a, b**, Relationships between EEC subsets. 
**a**, Dendrogram shows the relationship between EEC clusters as defined by hierarchical clustering of mean expression profiles of all of the cells in a subset (Methods). Estimates for the significance of each split are derived from 100,000 bootstrap iterations using the R package `pvclust` (*P* < 0.1, **P** < 0.05, ***P** < 0.01, *P* < 0.001; χ² test). 
**b**, Heatmap shows cell–cell similarities (Pearson’s r) between the 11 significant principal component scores (*P* < 0.05; Methods) across the 533 EEC cells (rows, columns). Rows and columns are ordered using cluster labels obtained using unsupervised clustering (Methods). 

**c**, Subset specificity of gut hormones and related genes. Scatter plot shows the specificity of each gene to its marked cell subset (defined as the proportion of cells not in a given subset that do not express a given gene) and its sensitivity in that subset (defined as the fraction of cells of a given type that do express the gene (Methods). Subsets are colour coded as in the legend. Genes are assigned to the subset where they are most highly expressed on average. Genes were chosen on the basis of their known annotation as gut hormones (Cck, Gal, Gcg, Ghrl, GIP, lapp, Nucb2, Nts, Pyy, Scy, Scx), enterochromaffin markers (Tph1, Tac1) and canonical EEC markers (Chga, Chgb).

**d**, GPCRs enriched in different EEC subtypes. Heatmap shows the expression levels (row-wise Z score) averaged across the cells in each of the EEC subtypes (columns) of 11 GPCR-encoding genes (rows) that are differentially expressed (FDR < 0.25; Mann–Whitney U test) in one of the EEC subtypes. The free fatty acid receptors (Ffar) 1 and 4 show specific expression patterns: Ffar1 is highest in SIN cells and is also expressed by the Cck-expressing subsets previously termed I cells (SIL-P, SIL-A and SIK-P), whereas Ffar4 is highest in the GIP-expressing subsets (SIK and SIK-P). These receptors are known to induce the expression of GIP and Gcg to maintain energy homeostasis. Ffar2 was expressed by some progenitors and by enterochromaffin cells, but absent from GIP-expressing cells, whereas the oleoylethanolamide receptor Gpr119, which is important for food intake and glucose homeostasis, is most highly expressed in SILA cells.
Extended Data Figure 7 | See next page for caption.
Characterization of tuft cell heterogeneity and identification of Tslp and the haematopoietic lineage marker Ptprc (CD45) in a subset of tuft cells. Related to Fig. 4. a, Tuft-1 and tuft-2 cells. Shown is t-SNE visualization of 102 tuft cells (points; n = 8 mice) from the plate-based full-length scRNA-seq dataset (Extended Data Fig. 2a), labelled by their subclustering into tuft-1 (orange) and tuft-2 (brown) subtypes. b, Gene signatures for tuft-1 and tuft-2 cells. Heatmap shows the relative expression (row-wise Z scores) of the tuft-1 and tuft-2 marker genes (rows; orange and brown, respectively) across single cells from the plate-based dataset (columns) assigned to tuft-1 and tuft-2 cell clusters (orange and brown, respectively). The top 25 genes are shown for each subtype (all FDR < 0.01 and log_2(fold change) > 0.1 in both plate- and droplet-based datasets). c, Tuft-2 signature genes are enriched in immune functions. Shown are the significantly enriched (Methods; FDR < 0.1; −log_10(Q value)) gene ontology terms in the gene signature for the tuft-2 subset. d, Expression of neuron- and inflammation-related genes in tuft-1 and tuft-2 subsets, respectively. Plot shows for each gene (y axis) its differential expression (x axis) between Tuft-1 and Tuft-2 cells. Bar indicates Bayesian bootstrap 95th estimates of log_2(fold change); hinges and whiskers indicate 25% and 95% confidence intervals, respectively. e, Il33 is not detected in tuft cells. Distribution of expression of Il33 in cell subsets in full-length scRNA-seq. (*FDR < 0.1; Mann–Whitney U test). f, g, Tuft-2 cells are enriched for Tslp. f, Combined smFISH and immunofluorescence of Tslp (green) with DCLK1 (red). Scale bars, 10 μm. g, Relative quantification of mRNA expression by qPCR of Alpi, Tslp and Dclk1 (tuft cell markers) from tuft-1, tuft-2 or randomly selected EpCAM^+ single cells identified from 96-well plate-based full-length scRNA-seq (16 cells per group) (*P < 0.05, **P < 0.005; t test). h, Validation of CD45 expression in tuft-2 cells. Immunofluorescence assay showing co-expression of the tuft cell marker DCLK1 and of CD45 (left) and CD45 (right, with increased brightness); yellow boxes show three representative tuft cells. Scale bars, 200 μm. i, Isolation of tuft-2 cells based on CD45 expression using FACS. Shown is t-SNE of 332 EpCAM^+/CD45^+ FACS-sorted single cells (points; n = 3 pooled mice), coloured by unsupervised clustering (top left), the expression of the Tuft cell marker Dclk1 (top right), or the signature scores for tuft-1 and tuft-2 cells (bottom left and right, respectively).
Extended Data Figure 8 | See next page for caption.
Extended Data Figure 8 | Microfold cells from RANKL-treated intestinal organoids and in vivo. Related to Fig. 5. a–d, Microfold cells in RANKL-treated organoids. a–c, t-SNE of 5,434 single cells (points) from control (left) or RANKL-treated (middle and right) intestinal organoids, or colouring each cell (b, c) by the expression (log2(TPM + 1)) of the canonical microfold cell markers TNF-α-induced protein 2 (Tnaiap2, M-sec; b) and glycoprotein 2 (Gp2; c) (n = 4 pooled wells per treatment condition). d, Expression of microfold cell marker genes55,57,66 in each of the organoid cell clusters. Violin plots show the distribution of expression levels (log2(TPM + 1)) for each of ten previously reported microfold cell marker genes57 (columns), in the cells (points) in each of 13 clusters, including mature microfold cells (red), identified by k-nearest-neighbour clustering of the 5,434 scRNA-seq profiles from organoids. e, f, Microfold cell gene signature in vitro. Heatmaps show for each mature or stem cell cluster of organoid-derived intestinal epithelial cells (columns) the mean expression of genes (rows) for known (grey bars) or newly identified (black bars) microfold cell markers (e) or transcription factors (f), identified as being specific (FDR < 0.05; Mann–Whitney U test) to microfold cells in vitro and in vivo (Methods). g, Congruence of in vitro- and in vivo-derived microfold cell gene signatures. Violin plot shows the distribution of the mean expression of the in vitro-derived signature genes across the in vivo microfold cells (red) and across all other cells derived from the follicle-associated epithelia (grey).
Extended Data Figure 9 | See next page for caption.
Extended Data Figure 9 | Intestinal epithelial cell response to pathogenic stress. Related to Fig. 6. a, Generalized and pathogen-specific response genes. Volcano plots show for each gene (points) the differential expression and its associated significance (−log$_{10}$Q value; likelihood-ratio test) in response to either Salmonella (top) or H. polygyrus (bottom). Genes strongly upregulated in Salmonella (FDR < 10$^{-4}$) or H. polygyrus (FDR < 5 × 10$^{-3}$) are highlighted in purple or red, respectively. All highlighted genes are significantly differentially expressed (FDR < 0.05) in both 3′ and full-length scRNA-seq datasets. Left, all genes differentially expressed in the noted pathogen infection versus uninfected controls; middle, the subset differentially expressed in both pathogens versus control; right, the subset differentially expressed in only the noted pathogen, but not the other (Methods). b, Global induction of enterocyte-specific genes across cells during Salmonella infection. Shown is t-SNE of 5,010 single intestinal epithelial cells from control wild-type mice (left) and mice infected with Salmonella (right). Cells are coloured by the expression of the indicated genes, all specific to enterocytes in control mice (Supplementary Tables 2–4) and strongly upregulated by infection (FDR < 10$^{-10}$ in both 3′ and full-length scRNA-seq datasets). c, Intestinal epithelial cell programs in Salmonella infection. Enriched (−log$_{10}$Q value) gene ontology terms in genes induced in Salmonella-treated intestinal epithelial cells versus control. d, Cell-intrinsic changes after Salmonella infection. Relative expression (row-wise Z scores; colour scale) of 104 genes (top), of which 58 (bottom) are specific to Salmonella infection, significantly upregulated (FDR < 0.05; Mann–Whitney U test; log$_{2}$(fold change) > 0.1) in enterocytes (columns) from Salmonella infection. Ten representative genes are labelled. e, Upregulation of pro-inflammatory apolipoproteins serum amyloid A1 and 2 (Saa1 and Saa2) in distal enterocytes under Salmonella infection. Violin plot shows log$_{2}$(TPM + 1) expression level of Saa1 (top) and Saa2 (bottom) across all post-mitotic cell types from control and Salmonella-treated mice (n = 4 mice; sample identity shown in the legend) (*FDR < 0.01, **FDR < 0.0001; Mann–Whitney U test). f, Upregulation of antimicrobial peptides by Paneth cells after Salmonella infection. Violin plots show log$_{2}$(TPM + 1) expression levels of genes encoding antimicrobial peptides and the mucosal pentraxin Mptx2 in the cells (points) from control and Salmonella-infected mice (n = 4 mice; sample identity shown in the legend) (*FDR < 0.1, **FDR < 0.01, ***FDR < 0.0001; Mann–Whitney U test). g, Paneth cell numbers detected (using graph clustering; Methods) after Salmonella infection. Frequencies of Paneth cells in each mouse (points) under each condition (see legend) (***FDR < 0.01; Wald test). Error bars, s.e.m.
Extended Data Figure 10 | Goblet and tuft cell responses to *H. polygyrus* show a unique defence mechanism. Related to Fig. 6. 

**a**, Genes induced significantly in response to infection in a non-cell-type-specific manner. Shown is t-SNE visualization of 9,842 single intestinal epithelial cells (points) from control wild-type mice (left), mice infected with *H. polygyrus* for 3 or 10 days (middle) and mice infected with *Salmonella* (right). Cells are coloured by the expression (log₂(TPM + 1)) of the indicated genes. Genes were selected as significantly differentially expressed in response to infection in a non-cell-type-specific manner (FDR < 0.001 in both the 3′ scRNA-seq and full-length scRNA-seq datasets). 

**b**, Identification of the tuft-1 and tuft-2 subsets in the dataset of control, *Salmonella*- and *H. polygyrus*-infected cells. **b**, Violin plots of the distribution of the respective signature scores (left and middle) and the expression of *Dclk1* (right; log₂ (TPM + 1)) in cells (points) in each of the tuft subsets. 

**c**, t-SNE mapping of the 409 tuft progenitor, tuft-1 and tuft-2 cells, coloured by the scores for each signature (left and middle) and their assignment to subtype clusters via k-nearest-neighbour graph clustering (right). 

**d**, Induction of antiparasitic genes by goblet cells after helminth infection. Shown is the distribution of expression (log₂(TPM + 1)) of three antiparasitic immunity genes upregulated by goblet cells during *H. polygyrus* infection (FDR < 0.05; Mann–Whitney U test), in control and infected mice. 

**e**, Antiparasitic protein secretion by goblet cells during *H. polygyrus* infection. Immunofluorescence assay of formalin-fixed paraffin-embedded (FFPE) sections of RELMβ (top left; red) and E-cadherin (bottom left; green), and their merged view including DAPI nuclear stain (blue) (right), after 10 days of helminth infection. Arrow, sections of *H. polygyrus*; scale bars, 200 μm.
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

1. Sample size
   Describe how sample size was determined.
   We analyzed over 50,000 single epithelial cells to gain maximum power to detect heterogeneity, and to detect shifts in cell proportions using single cell RNA-seq. We used 'how many cell' tool to estimate the probability of detecting cell types, see 'Single-cell survey of small intestinal cells'.

2. Data exclusions
   Describe any data exclusions.
   No animals were excluded. Low quality, immune cells and doublets were filtered out computationally, see Methods.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   All mouse experiments were repeated at least twice in which we analyzed single cells from 2 mice per group at a single time, most experiments had n=4 mice in total.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   All mouse models (control; C57bl/6J, Lgr5-EGFP-IRES-CreERT2 and Gfi1b-eGFP) that were used in this study were 7-10 weeks old littermates, which assigned randomly or by genotype to groups (Methods).

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   We preformed an unbiased analysis to all datasets from different mouse models on a single cell resolution.
   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).
   - The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - A statement indicating how many times each experiment was replicated
   - The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
   - A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We used available softwares and pipelines to analyze single cell data. All methods are found in the Method section in ‘Computational Analysis’.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used in this study. All materials are found in the Method section.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

All antibodies in this study are found in Methods. in brief:
- rabbit anti-DCLK1 (1:200, Abcam GR245168-1),
- rat anti-CD45 (1:100, Biolegend 30-F11, 103101),
- rabbit anti-RELMBeta (1:200, Peprotech 500-p215),
- mouse anti-E-cadherin (1:100, BD Biosciences 610181),
- rat anti-Lysozyme (Dako, A0099),
- anti-mouse CD31-PE (1:500, e-Bioscience, 12-0311-81),
- anti-mouse TER-119 (1:500, e-Bioscience, 12-5921-81),
- anti-mouse CD326 (EpCAM)-APC-780 (1:300, e-Bioscience, 47-5791-82),
- anti-mouse CD31-PE (1:500, e-Bioscience, 17-0242-82)

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A
c. Report whether the cell lines were tested for mycoplasma contamination.

N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

mouse, C57bl/J6, males or females, the sex was consistent with the experiment, 7-10 weeks of age.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A
Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

- **Data presentation**

  For all flow cytometry data, confirm that:

  1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
  2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
  3. All plots are contour plots with outliers or pseudocolor plots.
  4. A numerical value for number of cells or percentage (with statistics) is provided.

- **Methodological details**

  5. Describe the sample preparation. All cells isolated from mouse small intestine as described in material and methods.

  6. Identify the instrument used for data collection. Data was collected on a Beckman Coulter MoFlo Astrios EQ cell sorter.

  7. Describe the software used to collect and analyze the flow cytometry data. Acquisition was performed with Summit v6.1, and analysis was performed with FlowJo v10.

  8. Describe the abundance of the relevant cell populations within post-sort fractions. Post-sorting we performed single-cell RNA-seq to verify purity of relevant cell populations. See Method section and Figures 1a,b for cell proportions post-sort.

  9. Describe the gating strategy used. The preliminary gating for the starting population used FSC1-Area vs SSC1-Area. Singlets were then gated using SSC1-Area vs SSC1-Width and live/dead was gated using 7AAD-Area vs SSC1-Area. Positive and negative boundaries were determined based on gating of a fluorescence minus one (FMO) control.

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