Title: A high-fidelity in vitro model of the human intestinal epithelium

Author List:
Satoru Fujii[1,2]
Scott T. Espenschied[2]
Vibha Anand[5]
Joao Bettencourt-Silva[6]
Yi Han[2]
Go Ito[1]
Akira Koseki[7]
Akihiro Kosugi[7]
James Kozloski[5]
Ryoma Matsumoto[1]
Shanshan Meng[2]
Natasha Mulligan[6]
Ryan J. Musich[2]
Kevin P. Newhall[2,4]
Eri Ohshina[1]
Shuhei Sekiguchi[1]
Yi Wang[3]
Jianying Hu[5]
Matthew Ciorba[8]
L. David Sibley[9]
Ryuichi Okamoto[1]
Thaddeus S. Stappenbeck[2,10,*]

Affiliations:

[1]Department of Gastroenterology and Hepatology, Tokyo Medical and Dental University, Tokyo 113-8510, Japan
[2]Department of Inflammation & Immunity, The Lerner Research Institute, The Cleveland Clinic, Cleveland, OH 44195, USA
[3]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA
[4]Medical Scientist Training Program, Case Western Reserve School of Medicine, Cleveland, OH 44195, USA
[5]Healthcare and Life Sciences, IBM Research, NY & MA, USA
[6]Healthcare and Life Sciences, IBM Research, Dublin, Ireland
[7]Healthcare and Life Sciences, IBM Research, Tokyo, Japan
[8]Division of Gastroenterology and the Inflammatory Bowel Diseases Center, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
[9]Deparment of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA
[10]Lead Contact
*Correspondence

Preprint: https://doi.org/10.1101/2024.01.01.573838

Classification: Biological Sciences – Cell Biology

Keywords: single cell RNAseq, intestine, epithelial cells, biomedical foundation models

Abstract

Primary intestinal epithelial stem cell culture methods have significantly advanced understanding of mammalian intestinal development and disease. However, progress has been hampered by inconsistent methodological reporting and challenges comparing *in vitro* systems with *in vivo* observations. We previously established a unique method for long-term 2-dimensional (2D) cultivation of mouse IECs using an air-liquid interface (ALI) technique which appears to faithfully recapitulate homeostatic and regenerative features in vitro. Here, we further refined these methods and optimized protocols for long-term, self-organizing 2D cultivation of human IECs. During the culture, we observed that epithelial cells undergo a dynamic morphological transition from squamous to columnar shape. Using single cell transcriptomics, we identified both major lineages and minor populations, including enteroendocrine cells, tuft cells, and BEST4/CA7+ cells. Leveraging the power and scalability of a biomedical foundation model (BMFM) trained on single cell RNA sequencing data, we performed classification tasks to identify cell types across sample sources and to quantitatively benchmark our in vitro differentiated cells against cells collected from patient biopsies. We observed a striking degree of similarity between our in vitro differentiated cells and the corresponding cell types in vivo for multiple differentiated lineages. This novel approach using BMFM holds promise to expand our understanding of the regulatory mechanisms including gene-gene regulation underlying homeostasis and regeneration as well as the functions of rare and poorly understood lineages within the human intestinal epithelia. Moreover, these methods are generalizable to other organs and can be used to assess the correspondence of cells across experimental modalities.

Significance

This manuscript addresses challenges in quantitative comparisons of cell types grown in vitro vs their in vivo counterparts. Here we use the intestinal epithelium as a model system to address this challenge. We devised an in vitro culture platform that supports multipotent intestinal epithelial stem cells and their numerous differentiated progeny. This system gives rise to all known rare and abundant lineages in correct proportions. Novel use of biomedical foundation models pre-trained on publicly available data and then fine-tuned to data from this platform enabled demonstration of high concordance of multiple in vitro differentiated lineages with the corresponding cells in vivo.

INTRODUCTION

Continued advancement in understanding human intestinal development, physiology, and disease requires innovation in model systems. The ability to culture primary stem cells collected from intestinal biopsies was a significant vertical advancement in methods for experimental study of the intestinal epithelium (1). Guided by experimental results, subsequent technological iterations in intestinal epithelial stem cell (ISC) culture methods have identified conditions (including media and substrate/matrix compositions) which are permissive for differentiation of secretory and absorptive lineages. Air-liquid interface (ALI) cultures are a leading-edge technology for *in vitro* studies of mammalian epithelial tissues (2), but have not yet been optimized or applied to the human intestine. However, the diversity of experimental methods has created significant barriers to comparative- and meta-analyses. Coupled with often incomplete methodological reporting, this has effectively created a reproducibility crisis in the field of intestinal biology research.

The timeline of significant developments in 'omics technologies has occurred roughly in parallel with major milestones in ISC culture methods, and have enabled robust genomic, epigenomic, and transcriptomic characterization studies of ISCs *in vitro* (3). Most recently, advances in single cell RNA sequencing (scRNAseq) have uncovered new insights into the diversity and putative functions of intestinal epithelial cells both *in vitro* and *in vivo* (4, 5). However, experimental validation confirming observations from scRNAseq studies has been scarce. Moreover, in spite of the plethora of scRNAseq data published over the last 5 years, computational methods for comparing multiple scRNAseq studies across different experimental modalities are limited, largely due to the lack of sufficient amount of high-quality labeled data, further hampering validation efforts.

Recently a new class of transformer-based machine learning algorithms, known as foundation models (6), has demonstrated remarkable capabilities to learn general representations (also called embeddings) from large amount of unlabeled data, which can then be used to create many task-specific machine learning models with dramatically reduced need for labeled data. While the most well-known demonstration of the power of foundation models have been through Large Language Models (LLMs) such as GPTx (7) and DeepSeek (8) recent studies has demonstrated the ability of such techniques to also ingest, process, and operate on large multimodal biological data classes, including single cell transcriptomics data (6, 9). By pre-training foundation models on scRNAseq data, and then fine tuning on project-specific tasks, there are potential analysis of single cell transcriptomics that were previously not feasible using traditional methods (6).

Here, we couple an advanced *in vitro* ISC culture system with single cell transcriptomics and Biomedical Foundation Models (BMFM) (http://research.ibm.com/projects/biomedical-foundation-models). Using a variety of orthologous and complementary experimental methodologies, we demonstrate that our culture system faithfully recapitulates in vivo intestinal epithelial differentiation programs and lineage specification. We then leverage the power of BMFM to benchmark our *in vitro* intestinal epithelial model system against data from human biopsies by fine-tuning a pretrained BMFM single cell transcriptomics model (pre-trained on scRNAseq data set for the tasks of cell classification) (10). The resulting fine-tuned models were used to perform quantitative evaluation of similarities between *in vitro* differentiated intestinal epithelial lineages and putative concordant cell types present in patient biopsy samples using label transfer. We randomly split the experimental data for re-training or fine-tuning tasks into five sets for label transfer experiments.

RESULTS

To define features of human intestinal epithelial cells (IECs) differentiated *in vitro*, we cultured human rectal epithelial stem cells from non-diseased donors (*SI Appendix* **Table 1**) first as spheroids enriched for epithelial stem and progenitor cells. From these cultures, we then established monolayers using a modified ALI culture technique (**Fig. 1A**) (11). An important difference from an orthologous mouse culture platform (11) was the requirement for including both TGFβ and ROCK small molecule antagonists during the establishment of both 3D and 2D (submerged) stem cell cultures, but both inhibitors are removed to permit differentiation under ALI conditions (*SI Appendix* **Fig. S1A**). Of note, while TGFβ blockade is not required for cultivation of mouse ISCs under maintenance or differentiation conditions, ROCK inhibitor was still required for culture maintenance of human and mouse ISCs (*SI Appendix* **Table S5**). While the addition of recombinant growth factors (EGF, FGF2, and IGF-1) was sufficient to induce proliferation of stem cells, these factors were not required for differentiation in the ALI platform (*SI Appendix* **Fig. S1A,B**). At ALI day 0 (ALId0), the monolayer contained squamous cells with no obvious morphologic features of differentiation. By ALId7, the cells had uniformly adopted a columnar morphology, coincident with the emergence of goblet cell (GC)-like and enterocyte (EC)-like cellular features and resembling the homeostatic architecture of human rectum (**Fig. 1B**). Cell height, a surrogate for differentiation approached 40 µm (**Fig. 1C**), which is comparable to the IEC height observed in human colon tissues (12). (13). EdU incorporation assays revealed that the human IEC monolayer was hypo-proliferative at ALId0, became transiently hyper-proliferative from ALId2-d4, then gradually returned to a less proliferative state thereafter (**Fig. 1D,E**). The transient hyper-proliferation mirrored our previous observations of the ALI culture of mouse IECs as well as regenerating rectal mucosa from ulcerative colitis patients, suggesting that the hyper-proliferative cells around ALId2-d4 may be reminiscent of regenerative processes in the human intestinal mucosa. Collectively, these observations focused our following studies on analysis of the ALI cultures at days 2 and 7.

Cross-platform comparison of scRNAseq from *in vivo* and ALI-cultured intestinal epithelial cells using traditional and novel methods

Given the temporal dynamics of human intestinal epithelial stem cell-derived cultures maintained under ALI conditions, we first used single-cell RNA sequencing to establish the transcriptomic heterogeneity of ALI-cultured human IECs using at ALId2 and ALId7 -. Following QC filtering (*SI Appendix* **Fig S2, Fig S3**), Uniform Manifold Approximation and Projection (UMAP) analysis revealed the presence of similar clusters in cultures established from 3 independent donors (*SI Appendix* **Fig. S3**) (14). Dimensionality reduction analysis of the integrated datasets revealed increasing separation of cell populations as the cultures matured (**Fig. 2A,B**). At ALId2 the cultures appeared transcriptionally homogenous, with generally poor separation of clusters in UMAP space. Manual cell typing using default analysis parameters was unable to confidently identify lineages – in spite of apparent clustering – at ALId2. However, by ALId7 manual cell-typing of the 11 clusters using canonical marker genes revealed the presence of major known cell types including: stem cells (SCs, Cluster #0), transit-amplifying cells (TAs, Cluster #1), enterocytes (ECs, Clusters #2, #3), goblet cells (GCs, Clusters #4-6), enteroendocrine cells (EECs, Cluster #9), and tuft cells (Cluster #10) (**Fig. 2B, Dataset_S01**).

Relative lineage abundances from the ALId7 scRNAseq data were similar when compared to the epithelial cell fraction from scRNAseq of ascending colon (AC) biopsies from 6 non-IBD patients (X^2 = 133.09, df = 136, p = 0.5546) (**Fig. 2C**) (10). These observations suggest the ALI culture platform is permissive for the differentiation and allocation of canonical absorptive and secretory lineages present in the human intestine. Based on these findings, we next utilized single cell data and AI methods to evaluate the degree of similarity of specific differentiated cell types in the colon. We hypothesized that we would find a high degree of statistical concordance between cells evaluated directly from patient biopsies with ALId7 cultured cells, further justifying the utility

of this platform for modeling human cells and tissues.

Cross platform comparison and validation of cell types from different sources and platforms has been challenging since the inception of 'omics approaches. To overcome this challenge, we designed a computational approach leveraging the power of foundation models: COMPARATOR (Computational Omics-based Modeling for Predictive Analysis of Response Alignment and Transfer Optimization in Real-world systems) enables cross-context translation between experimental systems and patient data through advanced label transfer, domain adaptation, and alignment verification technologies. Here 'label transfer' refers to the process of transferring labels from a pre-annotated, reference dataset to a new, unlabeled dataset that is the query (**Fig. 2D**). The goal is to use existing knowledge to efficiently annotate new data sets. Using one of the open-sourced BMFM models (https://github.com/BiomedSciAI/biomed-multi-omic) that was pre-trained for experimental modality-specific cell type annotation. We performed label transfer for all cells in all clusters in both datasets (**Fig. 2D**, **Dataset_S02**) (10). We used radar plots as a visual representation of label transfer results (multivariate data).

Evaluation of BMFM classification results from ALId2 and ALId7 uncovered mixed labeling of many clusters at ALId2 (*SI Appendix* **Table S6 and Table S7**). Consistent with the EdU incorporation assay, the clusters in the ALId2 samples appeared to be mostly comprised of proliferative TA-like cells. Interestingly, the BMFM label transfer experiment uncovered several groups of what appear to be 'proto' lineages (Clusters 3, 4, 5, and 9), which included mixed labels of 'TA' and a mature cell label (e.g. EC, GC), which may partly explain the inability to manually type many of the ALId2 clusters. As the classifier exhibited the strongest performance for identifying major and minor secretory lineages, we elected to focus on these cell types for further analysis.

Goblet cells derived from the ALI platform are functionally and transcriptionally comparable to primary human colonic goblet cells

We found fidelity of GCs at ALId7 in the *in vitro* culture system with GCs in vivo both morphologically and by analysis of single cell transcriptomic data. Transmission electron microscopy (TEM) analysis of ALId7 cells showed ultrastructural features of GCs including a distinct and well-formed apical theca containing mucus granules (**Fig. 3A**). Targeted gene expression analysis showed enriched expression of the mucin-encoding gene *MUC2* at ALId7 at this time point (**Fig. 3B**). Immunofluorescence staining of ALI culture cross sections with an anti-MUC2 antibody indicated the presence of goblet cells with an overlying mucus layer (**Fig. 3C**), which was sufficient to prevent *Shigella* flexneri from directly contacting the IEC monolayer (**Fig. 3D**). These observations support that GCs in the ALI system are functional, and recapitulate protective mechanisms observed in mucosa *in vivo*. We next assessed similarities between *in vitro* differentiated GCs and GCs identified in scRNAseq of biopsies from non-diseased patient donors. When we analyzed GCs from the reciprocal label transfer experiments. Label transfer to the immature ALI culture identified a cluster of proto-goblet cells defined by mixed TA and GC labels (**Fig. 3E**). However, by ALId7, we found similarly high labeling efficiency in both directions (0.93 ± 0.017 from ALI to Biopsy; 0.904 ± 0.0057, from Biopsy to ALI) (**Fig. 3F,G**). Thus, GCs from both data sets, by multiple observational and quantitative metrics, were highly similar to each other.

ALI culture enables differentiation of all known human rectal EEC subsets

We next turned our attention to hormone-producing secretory cells. Plotting gene expression in UMAP embedding space demonstrated enrichment of the EEC marker genes *CHGA* and *NEUROD1* in Cluster 10 (**Fig. 4A,B**). Time course target gene expression analysis revealed relative CHGA expression peaked in the mature ALI culture at d7 (**Fig. 4C**). Amongst human IECs, enteroendocrine cells (EECs) are known to consist of various sub-populations, each of which produces distinct types of hormones (15). By isolating the EEC cluster, and performing

238 sub-clustering, UMAP analysis we identified the presence of 5 discrete sub-populations of EECs
239 in the ALId7 cultures (**Fig. 4D**). Further investigation of the ALI EEC scRNAseq cluster identified
240 L-cells, which express *GCG* (glucagon) and *PYY*, enterochromaffin cells (ECCs), which express
241 *TPH1*, and EEC progenitors, which express *NEUROG3* (**Fig. 4E,F**). Of note, a subset of EECs in
242 the ALId7 monolayer cells forms a small sub-cluster that distinctively expresses *somatostatin*
243 (SST), indicating the presence of D-cells (**Fig. 4E,F**), which have been observed in human colonic
244 tissue sections. Previous scRNAseq studies failed to identify EEC subsets from human colon
245 tissues or organoids (15-17). Immunofluorescence labeling with antibodies that specifically
246 recognize NEUROG3, serotonin (an EC marker), and GLP-1 (an L-cell marker) in ALId7 cultures
247 corroborated the scRNAseq analysis (**Fig. 4F**). We further confirmed the expression of SST at
248 the protein level in mature ALI culture (**Fig. 4F**). Consistently, this finding corroborates the notion
249 that the ALI culture platform supports *in vitro* differentiation of rare IEC secretory lineages found
250 in the human colon.
251 Leveraging the BMFM model for cell type classification, we again performed reciprocal
252 label transfer experiments to assess the similarities between EECs in vitro and in vivo. Of note,
253 we had identified 2 populations of EECs in the biopsy dataset without the need for additional sub-
254 clustering. Using a BMFM fine-tuned on biopsy data, we observed nearly equivalent labeling of
255 ALId2-Cluster10 as EEC_1 and EEC_2; however, at ALDId7 we found the majority of in vitro
256 differentiated EECs were classified as EEC_1 (0.74±0.17), while almost one quarter were
257 classified as EEC_2 (0.23±0.17); when the two labels were combined, the model classified 97%
258 of *in vitro* EECs correctly (**Fig. 4G-I**). Label transfer in the opposite direction was comparably
259 strong when evaluating biopsy EEC_1 cells (0.95±0.02) or biopsy EEC_2 cells (0.94±0.03)
260 classified by a model fine-tuned on our *in vitro* data. Taken together, the results from
261 transcriptional, morphological, and protein characterization of differentiated cells in the ALI system
262 are highly concordant with primary colonic EECs.
263
264 **ALI *in vitro* Differentiated Tuft Cells are Transcriptionally Homologous to Tuft Cells**
265 ***in vivo.***
266 We next compared *in vitro* and *in vivo* differentiated tuft cells (TCs). Given the signals
267 required for experimental expansion of TCs in mice (e.g., helminths, and IL25), we hypothesized
268 that *in vitro* differentiated TCs in the ALI system would be rare and potentially transcriptionally
269 distinct from TCs *in vivo*, as the former lacks requisite extrinsic signals required for the expansion
270 and function (18, 19).
271 We identified a small cluster (Cluster #11) which exhibited distinctive expression of
272 *TRPM5*, a putative marker for TCs in human IECs (20), assigning this to TC cluster (**Fig. 5A,B**).
273 Notably, marker analysis identified 19 transcripts which are highly and uniquely expressed in TCs
274 in the ALI culture (**Fig. 5A**). While *DCLK1* is the most common marker for detecting TCs in mice
275 (21), no specific expression of *DCLK1* was observed in the TC cluster of the homeostatic
276 monolayer cells (**Fig. 5C**). This is consistent with recent immunohistochemical and transcriptomic
277 data suggesting that *DCLK1* is undetectable in human colonic tissues (22, 23), further supporting
278 the notion that *DCLK1* does not identify TCs in humans. *PTGS1* (24), which encodes a critical
279 enzyme in prostaglandin (PG) biosynthesis (25), and *POU2F3*, a transcription factor required for
280 TC differentiation (26), both reliable alternative markers for human TCs, are distinctively
281 expressed in this cluster (**Fig. 5B**, *SI Appendix* **Fig. S5**). *SH2D6*, which has been recently
282 identified as a marker for the rare sub-population of $CD45^+$ TCs in mice, along with *PTPRC* (which
283 encodes CD45) (27), are both enriched in this cluster (**Fig. 5B**, *SI Appendix* **Fig. S5A**), suggesting
284 a potential subdivision of TCs in human intestinal mucosa, as observed in mice.
285 Studies in mice suggest that TCs may be involved in the gut-brain axis and metabolic
286 control, working in cooperation with enteroendocrine cells (EECs) and enteric neurons (21, 28).
287 TCs express chemoreceptors, functioning as sentinels in the intestinal mucosa to regulate
288 homeostasis and regeneration, as well as to respond to infections (18, 29). Intriguingly, pathway

289 analysis on the TC cluster revealed that pathways related to serotonergic synapse and cholinergic
290 synapse signaling are ranked as most prominent (*SI Appendix* **Fig. S5B**). Further analyses
291 identified specific expression of *CHAT*, coding for a critical enzyme for acetyl choline (ACh)
292 synthesis (30), *SLC18A3*, which encodes a vesicular ACh transporter (31), and *HTR3C* and
293 *HTR3E*, genes encoding serotonin receptors (32), in this cluster (*SI Appendix* **S5A**). The TC
294 cluster also distinctively expressed the taste receptor genes *TAS1R1* and *TAS1R3* (*SI Appendix*
295 **Fig. S5A**). These findings corroborate the possibility that TCs in the human intestinal mucosa
296 could function partly as neuronal transmitters/receptors, and/or chemoreceptors, as well as
297 components of PG synthesis machinery. Of note, while studies in mice have clearly suggested
298 that TCs function as a major source of IL-25 in the intestinal mucosa, crucial for initiating type 2
299 immune responses against helminths and protozoan parasites (18), the TC cluster cells did not
300 exhibit specific expression of *IL25* (*SI Appendix* **Fig. S45**). This implies that human TCs might
301 possess little relevance to IL-25 production, or their expression of *IL25* could require activation
302 through certain stimuli.
303 We also leveraged the BMFM label transfer approach to evaluate concordance between
304 *in vitro* differentiated TCs and primary TCs sampled from human biopsy tissue. Of note, the
305 foundation model failed to identify TCs in the ALId2 scRNAseq data. From reciprocal label transfer
306 studies, we observed extremely high confidence in the model's ability to classify TCs (Biopsy to
307 ALI: 0.932±0.01; ALI to Biopsy: 0.919±0.01) (**Fig. 5D,E**). Thus, maintenance of extrinsic signals
308 are not required for human ISCs to allocate TCs *in vitro* and for these TCs to differentiate with
309 similar programs as *in vivo* TCs.
310 The functions of TCs in the human intestine remain largely unknown. However, some
311 studies focusing on IBD have reported a reduction in TCs in the epithelial mucosa of IBD patients
312 (22, 23), suggesting a potential association with IBD pathology. The future application of this novel
313 2D culture technique in researching human intestinal TC functions holds promise for enhancing
314 our understanding of IBD pathophysiology.
315
316 **ALI culture model enables investigation of poorly understood CA7/BEST4 cells**
317 Previously studies reported *BEST4*$^+$ cells as a specialized lineage of mature absorptive
318 enterocytes in human intestinal tissues (33, 34). The *BEST4* gene, which encodes a Ca^{2+}-
319 sensitive chloride channel, has recently gained attention as *BEST4*$^+$ cells may be associated with
320 pathophysiology of UC and colorectal cancer (17, 35). *BEST4*$^+$ cells co-express *OTOP2*, which
321 encodes a proton-conducting ion channel, and both are thought to be engaged in fluid/mineral
322 balance in human intestinal mucosa (17). In addition to *BEST4* and *OTOP2*, cells in Cluster #9
323 distinctly expressed the genes *CA7*, *HES4*, and *NOTCH2* (**Fig. 6A-C**), which were also reported
324 to be expressed in the *BEST4*$^+$ cells *in vivo* (17, 36). We observed more uniform expression of
325 *CA7* compared to *BEST4* in this cluster, prompting us to label this cluster "BEST4/CA7". *In situ*
326 hybridization analysis revealed a small but definitive number of *CA7*-positive cells in both ALI cells
327 and human colonic tissues (**Fig. 6D**), confirming the findings from our scRNAseq analysis and
328 demonstrating the existence of this minor population *in vitro*. We assessed similarities between
329 the CA7$^+$ cluster of cells differentiated *in vitro* under ALI conditions and BEST4$^+$ cells in patient
330 biopsies using the BMFM label transfer approach described above. Transferring labels to the
331 ALId2 cultured cells, we observed heterogeneous labeling of ClusterX wherein ~50% of the cells
332 were classified as BEST4, and a similar fraction identified as the enterocyte sub-type EC4 (**Fig.
333 6E**). At the later ALI time point, we found cells of the CA7 cluster were mostly labeled as BEST4$^+$
334 cells [labeling efficiency of 0.93±0.017 ($\mu\pm\sigma$)] when the model was fine-tuned on scRNAseq of
335 epithelial cells from human biopsy samples. When we fine-tuned the model on our hALId7
336 scRNAseq data and applied labels to the non-IBD dataset, we observed less efficient labeling of
337 the CA7 cluster as BEST+ cells (0.77±0.025); which may be partially explained by the reduced
338 cellular heterogeneity of our *in vitro* culture system (**Fig. 6F,G**).
339 The set of 9 marker genes which distinguish the BEST4/CA7 cluster is too short for

functional pathway enrichment analysis but represents a combinatorially-intractable human literature search problem. We thus devised an *in-silico* approach using natural language processing to query the entire text of primary articles accessible from PubMed Central (see methods in *SI Appendix* for details) for co-occurrence of subsets of marker genes for the BEST4/CA7 cells. This approach identified manuscripts in which pairs or higher-order subsets of genes (≥ 4 genes) from the CA7 marker gene list co-occurred in the same paragraph. Analysis of subsets of the marker genes revealed no co-occurrence of *CPA6* with any of the other genes. However, all higher-order subsets returned scRNAseq studies of the human gastrointestinal tract (**Fig. 6H**, **Dataset_03**) (37, 38). Our search results returned no other organs or tissues, suggesting cells defined by this gene list are unique to the intestine.

352 **DISCUSSION**

353 By integrating single cell RNA sequencing with the power of foundation models and
354 classical molecular and histopathologic approaches, we dissected the features of *in vitro*-
355 differentiated human intestinal epithelial cell lineages. scRNA seq of human rectal epithelial
356 cells cultivated under ALI conditions identified 11 distinct clusters, including mature and
357 functional absorptive and secretory cell lineages. While enterocytes and goblet cells are the
358 most abundant cell types in mature cultures, we also found rare populations including TCs,
359 EECs, and the poorly understood $CA7^{+}$/$BEST4^{+}$cells (**Fig. 1D**). Of note, the frequencies we
360 observed in the ALI culture system are congruent with the frequencies of the same cell types
361 detected by scRNAseq of biopsy samples from healthy donors. We were particularly
362 encouraged by the presence of TCs in our culture system, as these cells are notoriously difficult
363 to differentiate spontaneously *in vitro* (39, 40).

364 We leveraged the power of Biomedical Foundation Models to perform label transfer
365 experiments by first training on scRNAseq from one dataset then performing cell type
366 annotation of cells in a test dataset. Analysis of paired label transfer studies between biopsy and
367 ALId7 datasets (i.e., "reciprocal label transfer experiments") demonstrated the reliability of the
368 foundation model classifier as well as the fidelity of the culture platform, as single cell
369 transcriptional profiles of both abundant and rare epithelial cell lineages were highly concordant
370 with cells recovered from biopsies (**Fig. 2E,F, 3G-I, 4D,E,** and **5D,E**). Similar label transfer
371 analysis to the immature ALId2 cells revealed a variety of notable patterns, including co-labeling
372 of clusters with "TA" and differentiated cell types, suggesting the presence of precursor or proto-
373 cells. That said, not every cluster at ALId2 fit this pattern (e.g. EECs were clearly discernable on
374 their own, while BEST4/CA7 cells imperfectly matched other cell types). Collectively, our
375 findings demonstrate that the optimized ALI platform for culturing human intestinal epithelia
376 faithfully recapitulates lineage allocation and differentiation of mature absorptive and secretory
377 cell types which are highly concordant with cells observed *in vivo*. The methods reported herein
378 describe an enabling technology which will facilitate the next generation of *in vitro* studies of the
379 human intestinal epithelium.

380 Rigor and reproducibility are essential for scientific progress and maintaining public trust
381 in the scientific enterprise. Here, we combine traditional experimental approaches with cutting-
382 edge computational methods to thoroughly establish the fidelity of our new human intestinal
383 epithelial stem cell culture platform. We corroborate expression of scRNAseq marker genes for
384 GCs, EECs, and TCs by immunofluorescence labeling, *in situ* hybridization, and transmission
385 electron microscopy. We further demonstrate GCs form a functional mucus barrier by co-culture
386 with labeled bacteria. However, traditional validation methods are limited and do not scale well:
387 only a few candidates can be tested at a time, and from a modest number of samples. Thus,
388 there is a pressing need for highly dimensional and scalable unbiased data-driven techniques
389 for quantitative comparison of scRNAseq. For example, using cell typing as a BMFM output will
390 greatly expand the capacity for validation of scRNAseq studies. Moreover, these approaches
391 may now be applied to experimental or in silico manipulations where lineage allocation is
392 altered. These *in silico* experiments may subsequently inform the order, scope, or importance
393 for predicting in vitro experimental outcomes.

394 Lastly, we applied a state-of-the art literature search tool
395 (https://research.ibm.com/projects/deep-search) and natural language processing (NLP)-based
396 tools to analyze the marker gene set which defines the CA7+/BEST4+ cells. This was done
397 through iterative and combinatorial searching of the PubMed Central collection (up to April
398 2023). While we identified sets of 5/9 marker genes present in studies reporting scRNAseq
399 analysis of human gastrointestinal samples, no manuscripts were identified with these gene sets
400 in other organs or non-intestinal samples. However, our search space is currently limited to the
401 main text of documents and does not include tabular data often appended to manuscripts as
402 supplemental material. Thus, it is possible that cells with subsets of the CA7/BEST4 marker

gene signature may exist in other organs. Experimental investigation of the ontogeny and function of these cells is an active area of investigation and will require the development of sophisticated cell biological systems in order to fully understand their properties. Ultimately, if applied during the planning stages of a new project, these tools may be further automated and combined with emerging Large Language Models (LLM) to assist researchers in streamlining future study plans by identifying gaps and interesting questions.

The challenge remains with scRNAseq of in vitro models to validate against in vivo results. Here we demonstrate a novel AI-enabled technology that when paired with a robust in vitro platform under differentiation conditions, identifies high degrees of concordance between cultured cells and their *in vivo* counterparts. Culture conditions without differentiation represent an opportunity to perceive an expansion of cell types which may play a role under pathological conditions, such as wound repair. Further use of these tools with other in vitro systems will enable cross-platform validation, thus increasing rigor and reproducibility of data coming from these systems.

METHODS

Human intestinal stem cell lines

Human stem cell cultures were established from rectal and ileal biopsies collected from non-IBD donors during routine endoscopy at the Washington University Digestive Diseases Research Center Core (DDRCC) or at the Tokyo Medical and Dental University Hospital. This study was approved by the Washington University Institutional Review Board, Cleveland Clinic Institutional Review Board), and the Tokyo Medical and Dental University Scientific Ethics Committee. Written informed consent was obtained from all donors. Isolated crypts were cultured as previously described (41). Demographic details of human donors are described in *SI Appendix* **Table S1**.

Human Tissue

Formalin-fixed paraffin blocks of colonic surgical resection from non-IBD donors as healthy controls and UC patients were obtained from Department of Pathology and Immunology at the Washington University in St. Louis or Lerner Research Institute Biorepository Core at Cleveland Clinic. Sections (5 µm) were cut from blocks and stained with H&E, or processed for immunofluorescence or RNAscope *in situ* hybridization.

Single cell RNA sequencing and analysis

Intestinal stem cell cultures were dissociated to single cell suspensions, purified by FACS (≥70% viability, (≥70% singlets), and immediately processed using the 10X Genomics Single Cell 3' Reagents Kit v3.1 Next GEM protocol. Following quality control, libraries were sequenced on an Illumina Novaseq 6000 platform.

Sequence data were mapped to the human genome (GRCh38) using Cell Ranger (v7.0.1) with default parameters(42). Gene-barcode matrices were loaded in Seurat (v5.0.0) in R (v4.3.0)/RStudio (v2023.09.1+494) (43-47). Cells with high mitochondrial ratios (>20%) and/or high total transcript counts (>20,000) were excluded from further analysis. The top 3,000 most variable features were enumerated, and inter-subject variance removed via integration. Additional details are provided in the *SI Appendix*.

Biomedical Foundation Models

IBM's Biomedical Foundation Model for gene expression (biomed-multi-omic/biomed-gene-expression) was developed using the scBERT architecture (9) pre-trained on scRNAseq data from ~1 million cells in the Panglao database (48) using transformer-based deep neural networks. After fine-tuning the model with scRNAseq data of either ALI cultured IECs or primary biopsy samples, the model was asked to classify known (to validate the model for cell type annotation) or unknown cells (to apply label transfer and establish cell type concordance).

DATA AVAILABILITY

Single-cell RNAseq will be publicly available as of the date of publication.

DISCLOSURES

T.S.S. is on the scientific advisory board for Janssen Pharmaceuticals and Amgen. T.S.S. is a founder of Mobius Care, Inc. V.A, A.K, A.K, J.S, N.M, J.K, and J.H are IBM Research employees.

ACKNOWLEDGEMENTS

This work was supported by the Crohn's & Colitis Foundation. We thank the Biorepository Core, Genomics Core, Flow Cytometry Core, and Imaging Core of the Cleveland Clinic Lerner Research Institute, and the Genome Technology Access Center and Digestive Disease Research Core Center at Washington University in Saint Louis.

AUTHOR CONTRIBUTIONS

Conceptualization: TSS, SF, YW; Methodology: SF, AK, AK, JK, VA; Software: AK, AK, VA, JS, NM; Formal Analysis: STE, RJM, AK, JS; Investigation: SF, SM, GI, RM, EO, SS, KPN, YH; Resources: MC, LDS; Data Curation: RJM; Writing – Original Draft: SF, STE, RO, TSS; Writing – Review & Editing: SF, STE, VA, JK, JS, JH, TSS; Visualization: SF, STE; Supervision: JH, TSS; Funding Acquisition: TSS

FIGURE LEGENDS

Fig. 1. *ALI culture platform sustains differentiation of mature absorptive and secretory IECs*
(*A*) Experimental schematic of ALI culture of human intestinal epithelial stem cells. (*B*) Hematoxylin and Eosin (H&E) images of human colon monolayer cells at days 0, 4, and 7 of ALI culture. Micrographs are representative of rectum lines established from 6 independent donors (n = 6); scale bar: 50 µm. (*C*) Epithelial cell height in ALI cultures at days 0, 4, and 7 (n = 4 rectum lines, data presented as mean ± SEM; significance determined ordinary one-way ANOVA with Tukey's multiple comparisons test: **p < 0.005; ***p < 0.0005). (*D*) Representative fluorescence micrographs of EdU incorporation assay from ALI cultures at the indicated time points (nuclei in gray, EdU in green; scale bar = 50 µm). (*E*) Quantification of EdU^+ cells (cultures established from n = 3 independent donors). In C and E significance was determined by ordinary one-wayANOVA with Tukey's multiple comparisons test; **P* < 0.05, ***P* < 0.005, ****P* < 0.0005.

Fig. 2. *scRNAseq analysis of in vitro differentiated ISCs and classification of cell types using a domain-adapted Biomedical Foundation Model.*
(A,B) UMAP representations of integrated scRNAseq data from ALId2 (A) and ALId7(B). *(C)* Relative abundances of the indicated cell types from scRNAseq of ALI day 7 (n = 3) and human ascending colon (n = 6) *(D)* Schematic Biomedical Foundation Model Classifier ("COMPARATOR") and *in silico* label transfer experiment.

Fig. 3. *ALI culture supports differentiation of functional GCs*
(*A*) TEM images showed the morphology of goblet cells at ALId7. Image is representative of 3 independent rectal lines. Scale bar: 5 µm. (*B*) qPCR analysis of *MUC2* expression in human rectum monolayers at ALI day 0, 2, 4 and 7. Results were normalized to the day 0, and shown as mean ± SEM (n = 3 rectum lines per time point). Ordinary one-way ANOVA multiple comparisons: *p < 0.05; **p < 0.005; ***p < 0.0005; ****p < 0.0001. (*C*) Representative fluorescence photomicrographs of ALI culture sections stained with an α-MUC2 antibody. (*D*) Immunofluorescence of *Shigella flexneri* (red) with beta-catenin (upper panel, green), and with MUC2 (lower panel, green), in homeostatic rectum monolayer cells during ALI cultivation, co-cultured with *Shigella flexneri* applied to the apical side of the monolayer. Representative images of three independent experiments. (*E-G*) Radar plot of BMFM label transfer experiment results for GCs (biopsy → hALId2, biopsy → hALId7; and hALId7 → biopsy)

Fig. 4. *Differentiation of all known EEC-subtypes from rectal ISCs cultured under ALI conditions*
(A,B) Feature plot of *CHGA* and *NEUROD1* expression in ALId7 cultures. (*C*) qPCR analysis of *CHGA* expression in human rectum monolayers at ALI days 0, 2, 4 and 7. Results were normalized to day 0. Values shown as mean ± SEM (n = 3 rectum lines per time point). Ordinary one-way ANOVA multiple comparisons: *p < 0.05; **p < 0.005; ***p < 0.0005; ****p < 0.0001. (*D*) Sub-clustering within the EEC cluster showing distinct sub-populations [Progenitor cells (*NEUROG3*+), ECs (*TPH1*+), L-cells (*GCG*+, *PYY*+), and D-cells (*SST*+); UD (undifferentiated, no specific EEC subcluster markers, Stressed_Dying cells have a higher mitochondrial gene ratio than the others]. (*E*) Dot plot analysis of sub-populations within the EEC cluster. (*F*) Immunofluorescence of NEUROG3, serotonin, GLP-1, and SST in the human rectum monolayer at ALI day 7. Representative images of three rectum lines (n =3). Scale bar: 50 µm. (*G-I*) Radar plots of label transfer experiment results for EECs (biopsy → hALId2, biopsy → hALId7; hALId7 → biopsy)

Fig. 5. *ALI culture permits spontaneous differentiation of chemosensory tuft cells*

529 (*A*) Dot plot analysis of TC marker genes in ALId7 cultures. (*B*) UMAP feature plot of *TRPM5.*
530 (*C*) UMAP feature plot of *DCLK1*. (*D-E*) Radar Plots of label transfer experiment results for TCs
531 (left: biopsy → hALId7; right: hALId7 → biopsy)
532
533 **Fig 6.** *ALI culture platform enables investigation of CA7/BEST4+ cells*
534 (A) Dotplot of marker genes for BEST4/CA7 cluster. (*B,C*) UMAP feature plots of *BEST4* and
535 *CA7* expression. (*D*) Representative fluorescence photomicrographs of RNA *in situ* hybridization
536 for *CA7* by RNAscope in the human rectum monolayer cells at ALI day 7 (upper panel) and
537 human colon surgical resection tissue (lower panel) (n =3). Scale bar: 50 µm. (*E-G*) Radar plots
538 of label transfer experiment results for BEST4/CA7 cells (left: biopsy → hALId7; right: hALId7 →
539 biopsy). (H) UpSet plot of NLP literature search results for subsets of the BEST4/CA7 marker
540 gene list.
541

References

1. N. Calzadilla *et al.*, Bile acids as inflammatory mediators and modulators of intestinal permeability. *Front Immunol* **13**, 1021924 (2022).
2. Y. Ji, Y. Sun, Advancements in Organoid Culture Technologies: Current Trends and Innovations. *Stem Cells Dev* **33**, 631-644 (2024).
3. M. Oda, Y. Hatano, T. Sato, Intestinal epithelial organoids: regeneration and maintenance of the intestinal epithelium. *Curr Opin Genet Dev* **76**, 101977 (2022).
4. J. G. Camp, D. Wollny, B. Treutlein, Single-cell genomics to guide human stem cell and tissue engineering. *Nat Methods* **15**, 661-667 (2018).
5. V. Gudino, R. Bartolome-Casado, A. Salas, Single-cell omics in inflammatory bowel disease: recent insights and future clinical applications. *Gut* 10.1136/gutjnl-2024-334165 (2025).
6. R. Bommasani *et al.*, On the Opportunities and Risks of Foundation Models. http://dx.doi.org/10.48550/arXiv.2108.07258.
7. OpenAI *et al.*, GPT-4 Technical Report. http://dx.doi.org/10.48550/arXiv.2303.08774.
8. DeepSeek-AI *et al.*, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. http://dx.doi.org/10.48550/arXiv.2501.12948.
9. F. Yang *et al.*, scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852-866 (2022).
10. Y. Zhao *et al.*, Multiomic analysis reveals cellular, transcriptomic and epigenetic changes in intestinal pouches of ulcerative colitis patients. *Nat Commun* **16**, 904 (2025).
11. Y. Wang *et al.*, Long-Term Culture Captures Injury-Repair Cycles of Colonic Stem Cells. *Cell* **179**, 1144-1159 e1115 (2019).
12. D. Jenkins, A. Goodall, K. Drew, B. B. Scott, What is colitis? Statistical approach to distinguishing clinically important inflammatory change in rectal biopsy specimens. *J Clin Pathol* **41**, 72-79 (1988).
13. V. Greigert *et al.*, Cryptosporidium infection of human small intestinal epithelial cells induces type III interferon and impairs infectivity of Rotavirus. *Gut Microbes* **16**, 2297897 (2024).
14. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 10.1038/nbt.4314 (2018).
15. J. Beumer *et al.*, High-Resolution mRNA and Secretome Atlas of Human Enteroendocrine Cells. *Cell* **182**, 1062-1064 (2020).
16. M. Fujii *et al.*, Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. *Cell Stem Cell* **23**, 787-793 e786 (2018).
17. K. Parikh *et al.*, Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49-55 (2019).
18. J. von Moltke, M. Ji, H. E. Liang, R. M. Locksley, Tuft-cell-derived IL-25 regulates an intestinal ILC2-epithelial response circuit. *Nature* **529**, 221-225 (2016).
19. M. R. Howitt *et al.*, Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science* **351**, 1329-1333 (2016).
20. R. Elmentaite *et al.*, Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250-255 (2021).
21. S. K. Hendel *et al.*, Tuft Cells and Their Role in Intestinal Diseases. *Front Immunol* **13**, 822867 (2022).
22. S. Kjaergaard *et al.*, Decreased number of colonic tuft cells in quiescent ulcerative colitis patients. *Eur J Gastroenterol Hepatol* **33**, 817-824 (2021).
23. A. Banerjee *et al.*, Succinate Produced by Intestinal Microbes Promotes Specification of Tuft Cells to Suppress Ileal Inflammation. *Gastroenterology* **159**, 2101-2115 e2105 (2020).

24. C. H. Petersen *et al.*, Possible predisposition for colorectal carcinogenesis due to altered gene expressions in normal appearing mucosa from patients with colorectal neoplasia. *BMC Cancer* **19**, 643 (2019).
25. R. G. Biringer, The enzymology of the human prostanoid pathway. *Mol Biol Rep* **47**, 4569-4586 (2020).
26. Y. H. Huang *et al.*, POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. *Genes Dev* **32**, 915-928 (2018).
27. Z. Xiong *et al.*, Intestinal Tuft-2 cells exert antimicrobial immunity via sensing bacterial metabolite N-undecanoylglycine. *Immunity* **55**, 686-700 e687 (2022).
28. X. Cheng, U. Voss, E. Ekblad, Tuft cells: Distribution and connections with nerves and endocrine cells in mouse intestine. *Exp Cell Res* **369**, 105-111 (2018).
29. F. Gerbe *et al.*, Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226-230 (2016).
30. B. Schutz *et al.*, Distribution pattern and molecular signature of cholinergic tuft cells in human gastro-intestinal and pancreatic-biliary tract. *Sci Rep* **9**, 17466 (2019).
31. J. Pan, L. Zhang, X. Shao, J. Huang, Acetylcholine From Tuft Cells: The Updated Insights Beyond Its Immune and Chemosensory Functions. *Front Cell Dev Biol* **8**, 606 (2020).
32. B. Niesler, B. Frank, J. Kapeller, G. A. Rappold, Cloning, physical mapping and expression analysis of the human 5-HT3 serotonin receptor-like genes HTR3C, HTR3D and HTR3E. *Gene* **310**, 101-111 (2003).
33. G. Ito *et al.*, Lineage-specific expression of bestrophin-2 and bestrophin-4 in human intestinal epithelial cells. *PLoS One* **8**, e79693 (2013).
34. D. Wang *et al.*, Interferon-responsive intestinal BEST4/CA7(+) cells are targets of bacterial diarrheal toxins. *Cell Stem Cell* 10.1016/j.stem.2025.02.003 (2025).
35. X. S. He *et al.*, Oncogenic potential of BEST4 in colorectal cancer via activation of PI3K/Akt signaling. *Oncogene* **41**, 1166-1177 (2022).
36. J. M. Serigado *et al.*, Ulcerative Colitis: Novel Epithelial Insights Provided by Single Cell RNA Sequencing. *Front Med (Lausanne)* **9**, 868508 (2022).
37. M. Kanke *et al.*, Single-Cell Analysis Reveals Unexpected Cellular Changes and Transposon Expression Signatures in the Colonic Epithelium of Treatment-Naive Adult Crohn's Disease Patients. *Cell Mol Gastroenterol Hepatol* **13**, 1717-1740 (2022).
38. J. Burclaff *et al.*, A Proximal-to-Distal Survey of Healthy Adult Human Small Intestine and Colon Epithelium by Single-Cell Transcriptomics. *Cell Mol Gastroenterol Hepatol* **13**, 1554-1589 (2022).
39. K. E. Boonekamp, T. L. Dayton, H. Clevers, Intestinal organoids as tools for enriching and studying specific and rare cell types: advances and future directions. *J Mol Cell Biol* **12**, 562-568 (2020).
40. L. Huang *et al.*, Tuft cells act as regenerative stem cells in the human intestine. *Nature* **634**, 929-935 (2024).
41. K. L. VanDussen *et al.*, Development of an enhanced human gastrointestinal epithelial culture system to facilitate patient-based assays. *Gut* **64**, 911-920 (2015).
42. G. X. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
43. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).
44. Y. Hao *et al.*, Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293-304 (2024).
45. T. Stuart *et al.*, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
46. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell

644 transcriptomic data across different conditions, technologies, and species. *Nat*
645 *Biotechnol* **36**, 411-420 (2018).
646 47. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of
647 single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
648 48. O. Franzen, L. M. Gan, J. L. M. Bjorkegren, PanglaoDB: a web server for exploration of
649 mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019** (2019).

650

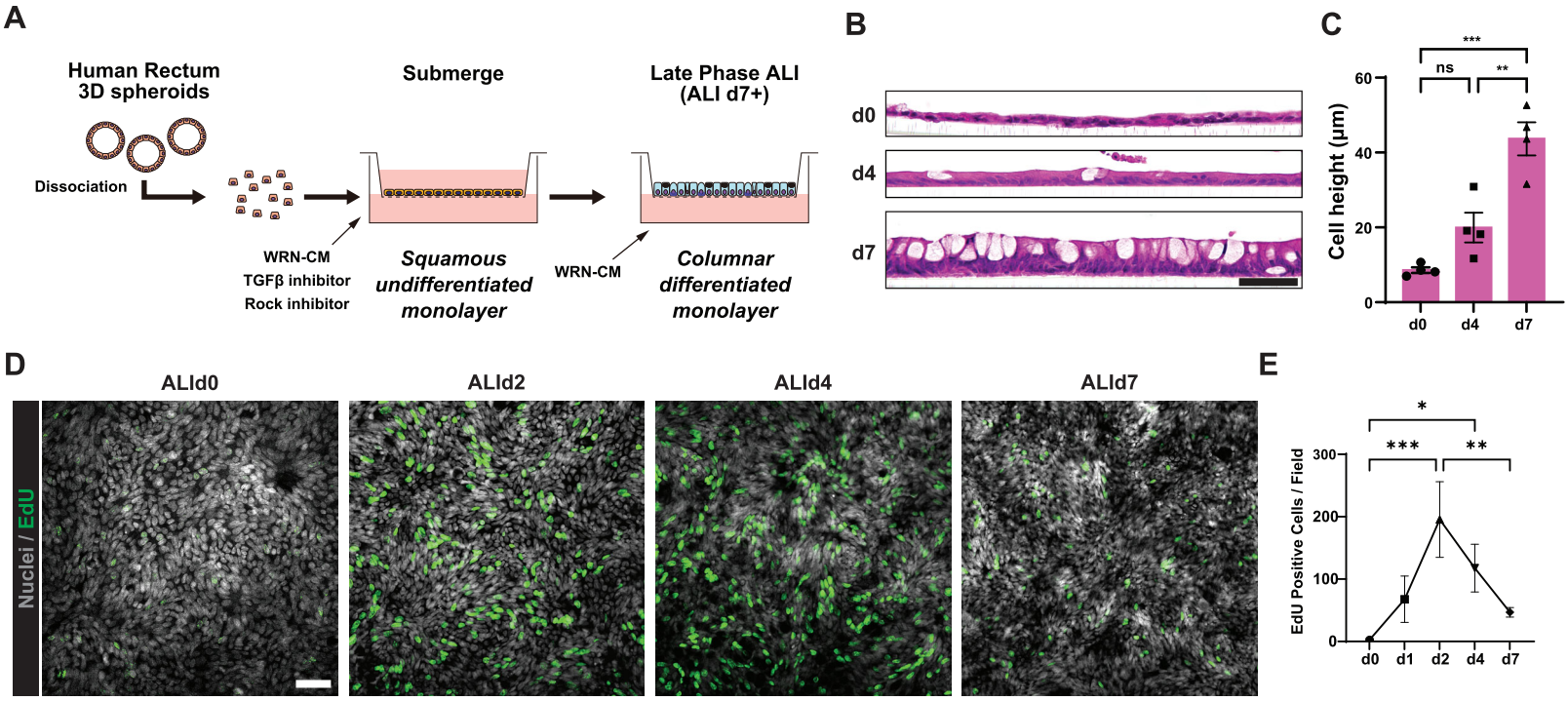
A
Human Rectum
3D spheroids
Dissociation
Submerge
Late Phase ALI
(ALI d7+)
WRN-CM
TGFβ inhibitor
Rock inhibitor
Squamous
undifferentiated
monolayer
WRN-CM
Columnar
differentiated
monolayer
B
d0
d4
d7
C
Cell height (µm)
ns
***
**
d0
d4
d7
D
ALId0
ALId2
ALId4
ALId7
Nuclei / EdU
E
EdU Positive Cells / Field
*
***
**
d0
d1
d2
d4
d7

Figure 1

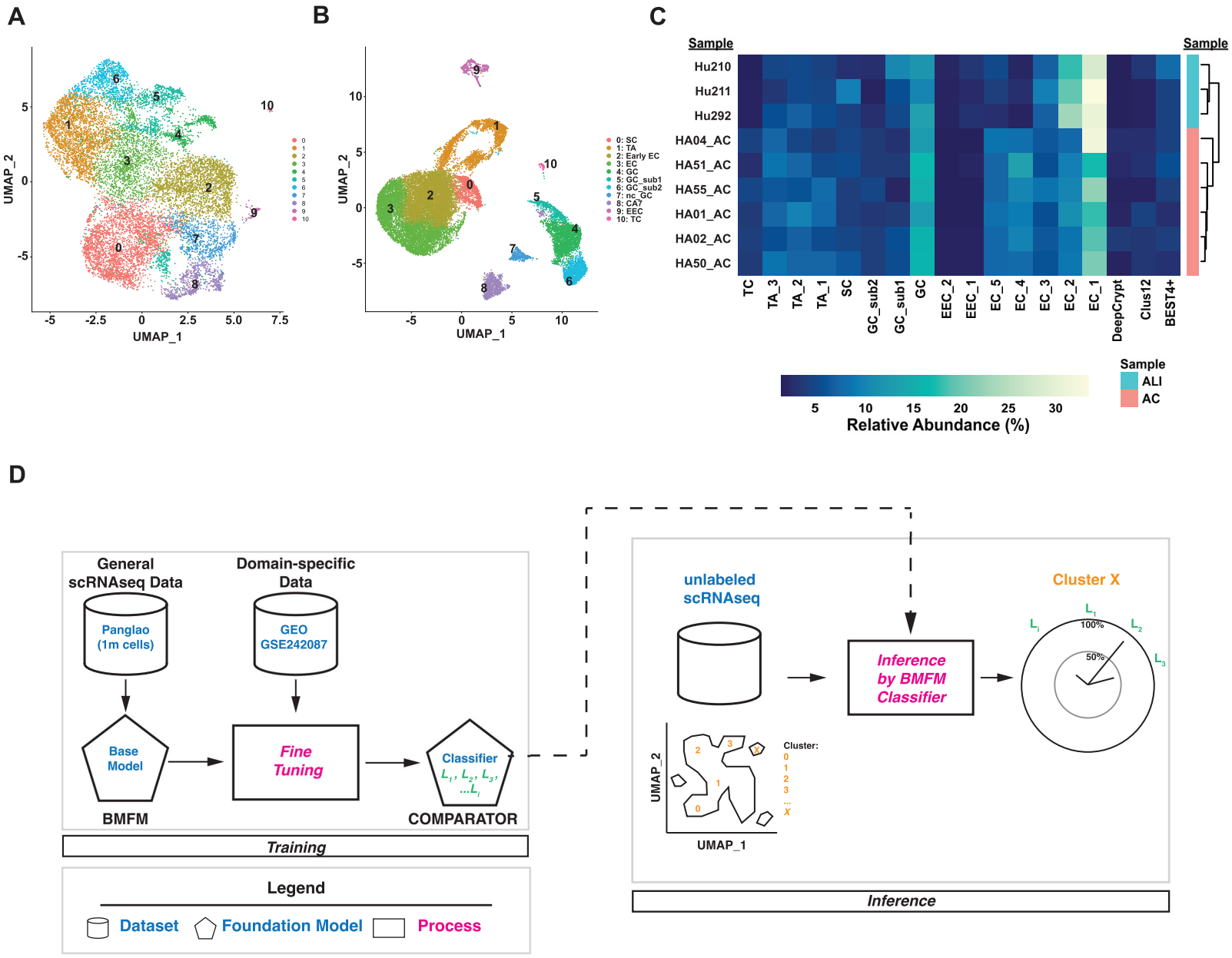
A
B
C
D
UMAP_1
UMAP_2
0: SC
1: TA
2: Early EC
3: EC
4: GC
5: GC_sub1
6: GC_sub2
7: nc_GC
8: CAT
9: EEC
10: TC
Sample
Hu210
Hu211
Hu292
HA04_AC
HA51_AC
HA55_AC
HA01_AC
HA02_AC
HA50_AC
TC
TA_3
TA_2
TA_1
SC
GC_sub2
GC_sub1
GC
EEC_2
EEC_1
EC_5
EC_4
EC_3
EC_2
EC_1
DeepCrypt
Clus12
BEST4+
Relative Abundance (%)
ALI
AC
General scRNAseq Data
Domain-specific Data
Panglao (1m cells)
GEO GSE242087
Base Model
BMFM
Fine Tuning
Classifier
COMPARATOR
Training
Legend
Dataset
Foundation Model
Process
unlabeled scRNAseq
Inference by BMFM Classifier
Cluster X
Cluster:
Inference

Figure 2

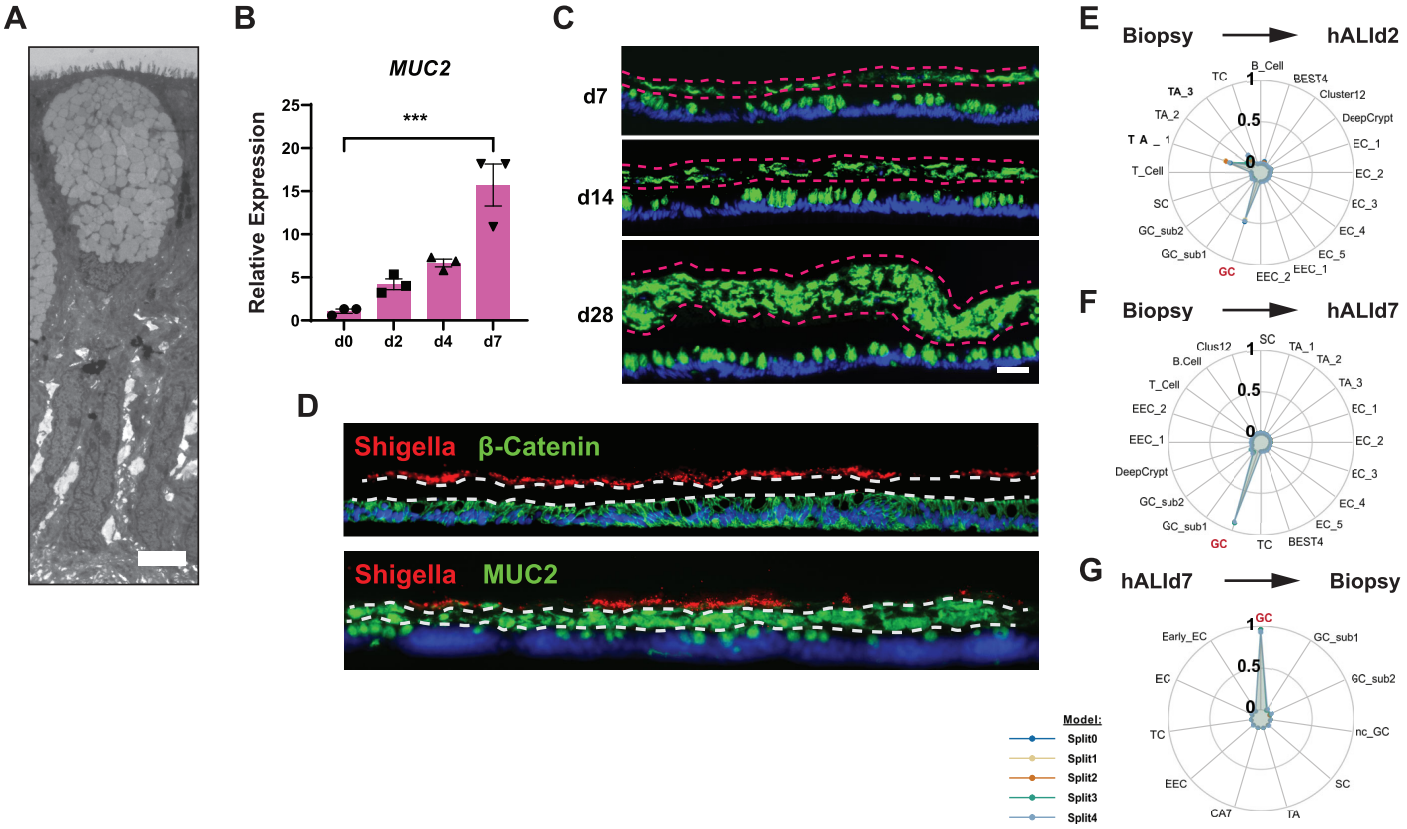
A
B
MUC2
Relative Expression
***
d0
d2
d4
d7
C
d7
d14
d28
D
Shigella β-Catenin
Shigella MUC2
E
Biopsy
hALId2
F
Biopsy
hALId7
G
hALId7
Biopsy
Model:
Split0
Split1
Split2
Split3
Split4

Figure 3

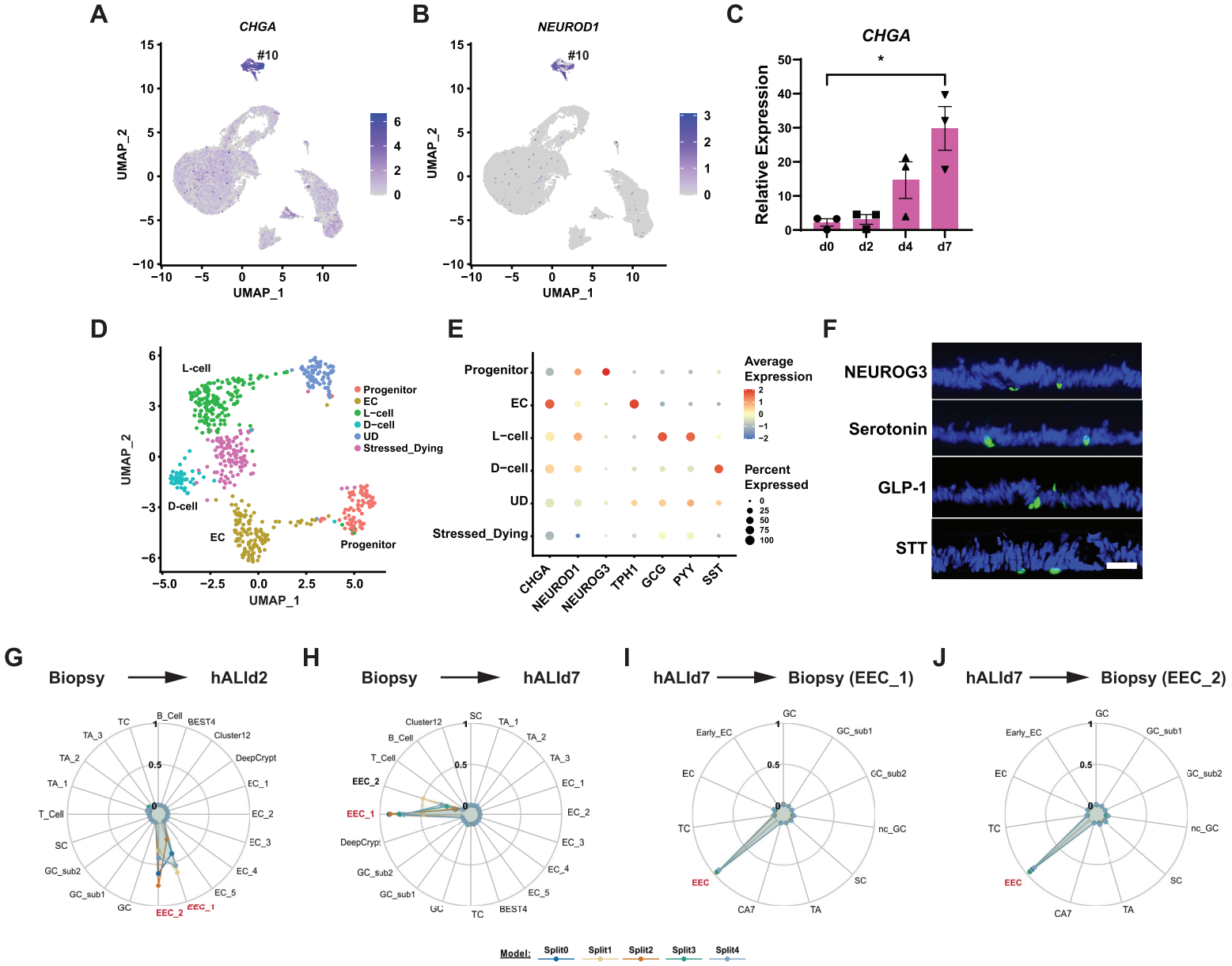
A
CHGA
#10
UMAP_2
UMAP_1
B
NEUROD1
#10
UMAP_2
UMAP_1
C
CHGA
Relative Expression
d0
d2
d4
d7
D
L-cell
D-cell
EC
Progenitor
Progenitor
EC
L-cell
D-cell
UD
Stressed_Dying
UMAP_2
UMAP_1
E
Progenitor
EC
L-cell
D-cell
UD
Stressed_Dying
CHGA
NEUROD1
NEUROG3
TPH1
GCG
PYY
SST
Average Expression
Percent Expressed
F
NEUROG3
Serotonin
GLP-1
STT
G
Biopsy
hALId2
H
Biopsy
hALId7
I
hALId7
Biopsy (EEC_1)
J
hALId7
Biopsy (EEC_2)
Model:
Split0
Split1
Split2
Split3
Split4

Figure 4

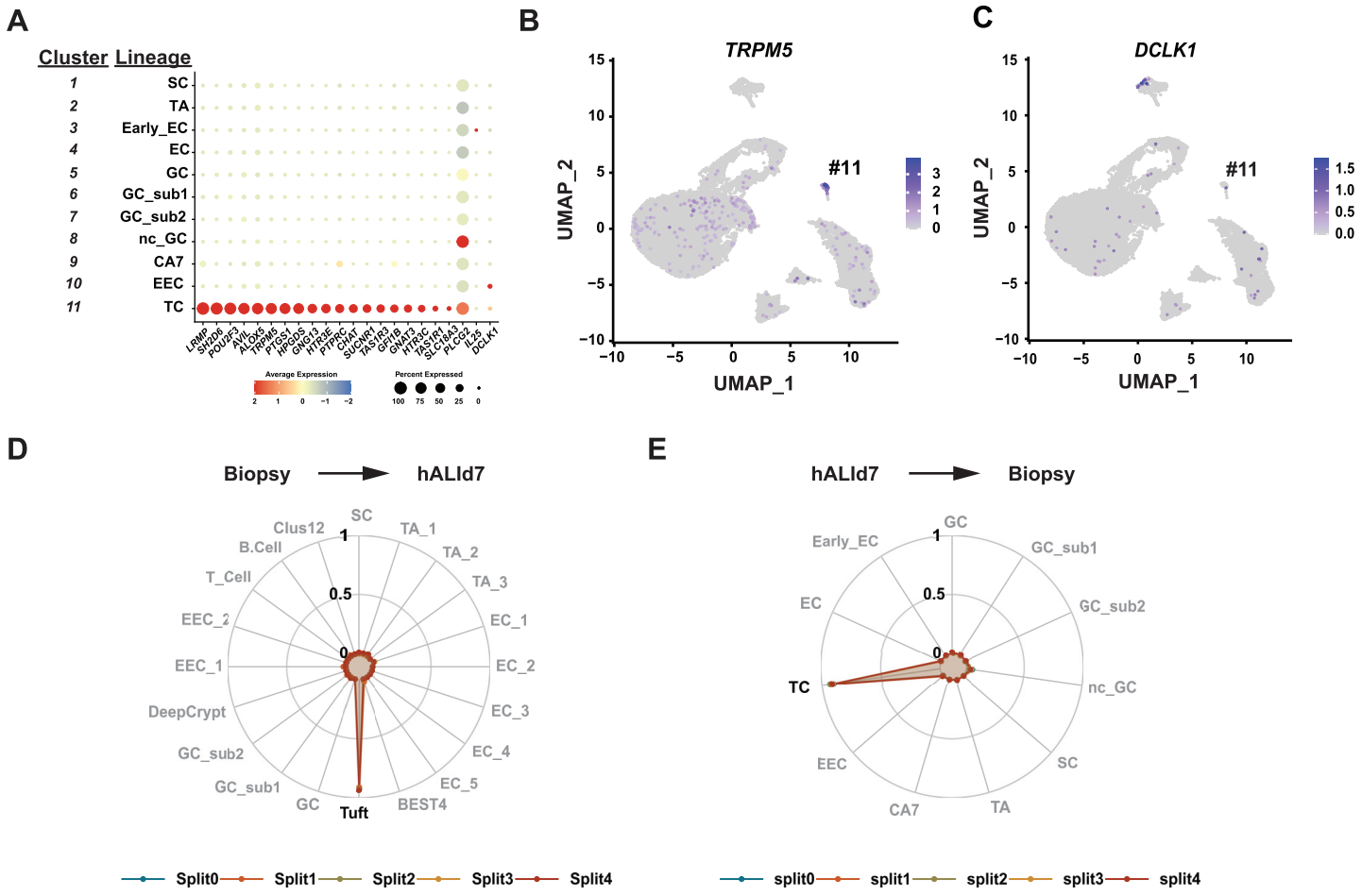
A
Cluster Lineage
1 SC
2 TA
3 Early_EC
4 EC
5 GC
6 GC_sub1
7 GC_sub2
8 nc_GC
9 CA7
10 EEC
11 TC
LRMP SH2D6 POU2F3 AVIL ALOX5 TRPM5 PTGS1 HPGDS GNG13 HTR3E PTPRC CHAT SUCNR1 TAS1R3 GFI1B GNAT3 HTR3C TAS1R1 SLC18A3 PLCG2 IL25 DCLK1
Average Expression
Percent Expressed
B
TRPM5
#11
UMAP_2
UMAP_1
C
DCLK1
#11
UMAP_2
UMAP_1
D
Biopsy
hALId7
SC
TA_1
TA_2
TA_3
EC_1
EC_2
EC_3
EC_4
EC_5
BEST4
Tuft
GC
GC_sub1
GC_sub2
DeepCrypt
EEC_1
EEC_2
T_Cell
B.Cell
Clus12
Split0
Split1
Split2
Split3
Split4
E
hALId7
Biopsy
GC
GC_sub1
GC_sub2
nc_GC
SC
TA
CA7
EEC
TC
EC
Early_EC
split0
split1
split2
split3
split4

Figure 5

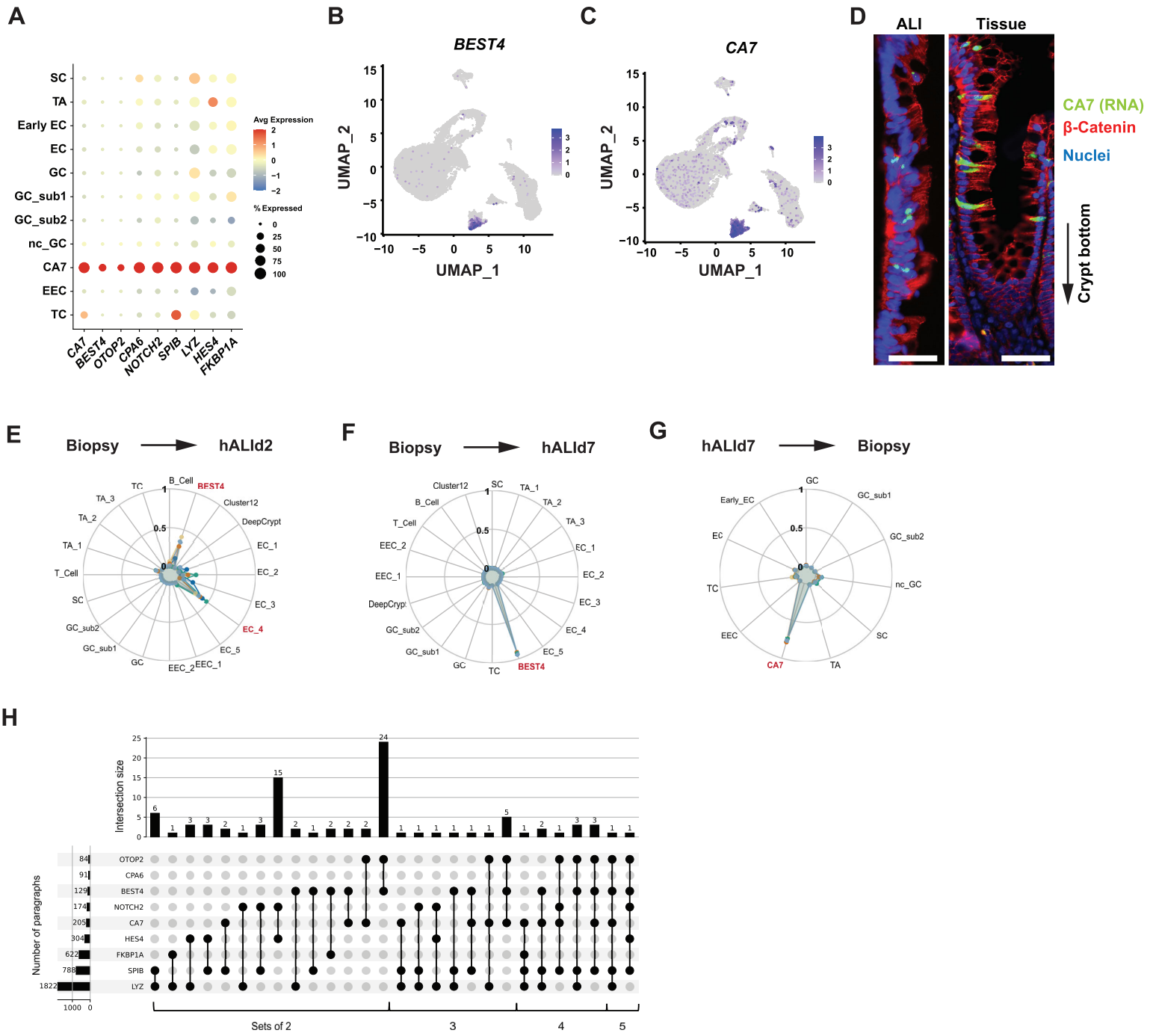
A
SC
TA
Early EC
EC
GC
GC_sub1
GC_sub2
nc_GC
CA7
EEC
TC
CA7
BEST4
OTOP2
CPA6
NOTCH2
SPIB
LYZ
HES4
FKBP1A
Avg Expression
% Expressed
B
BEST4
UMAP_2
UMAP_1
C
CA7
UMAP_2
UMAP_1
D
ALI
Tissue
CA7 (RNA)
β-Catenin
Nuclei
Crypt bottom
E
Biopsy
hALId2
F
Biopsy
hALId7
G
hALId7
Biopsy
H
Intersection size
Number of paragraphs
OTOP2
CPA6
BEST4
NOTCH2
CA7
HES4
FKBP1A
SPIB
LYZ
Sets of 2
3
4
5

Figure 6

Materials and Methods

Crypt isolation and establishment of stem cell cultures

Human intestinal biopsies were minced with fine scissors and digested in 1 mL of 2 mg/mL collagenase type I solution (Gibco, 17100017) with 50 µg/mL gentamicin (Sigma-Aldrich, G1272) in washing media [DMEM/F12 (Sigma-Aldrich, D6421) supplemented with 1X L-glutamine (Gibco, G7513), 1X penicillin/streptomycin (CCF Lerner Research Institute Cell Culture and Media Core) and 10% FBS (Sigma-Aldrich, F2442)] for 20-30 min with pipetting every 5-10 min. Crypts were filtered through a 70-µm strainer, incubated with 9 mL of ice-cold washing media and pelleted by centrifugation at 300 x *g* for 5 min. Pelleted crypts were then suspended in 1 mL washing medium, transferred to a 1.5 mL tube and pelleted by centrifugation at 200 x *g* for 5 min. The supernatant was carefully removed by pipette, and the pellet was re-suspended in Matrigel (Corning, 354234). The Matrigel mixture was plated into 24-well tissue culture plates on ice. Plates were inverted and incubated at 37°C to polymerize the Matrigel and 500 µL of 50% L-WRN conditioned media (1) [a 1:1 mixture of L-WRN CM and fresh primary culture media (Advanced DMEM/F-12 (Gibco, 12634010) supplemented with 20% FBS, 1x GlutaMax (Gibco, 35050061), 1x penicillin/streptomycin)], supplemented with 10 µM ROCK inhibitor (Y-27632) and 10 µM TGF-β inhibitor (SB 431542 or A83-01). Media was refreshed every 2-3 days. Rectal stem cells grown as spheroid cultures were collected in Cell Recovery Solution (Corning, 354253), digested with TrypLE Express (Gibco, 12605010) into small fragments, and passaging every 3-4 days. These human rectal spheroid lines were maintained and passed up to 30 generations for the experiments reported in this paper.

Air Liquid Interface (ALI) culture

Grown human rectal spheroids were dissociated into single cells by TrypLE Express digestion and seeded onto transwell membranes (0.4 µm pore size, Corning, 3470) pre-coated with 10% Matrigel diluted in PBS for 20min. 200,000 cells were seeded per well. Initially, 150 µL and 350 µL of 50% L-WRN media with 10 µM ROCK inhibitor (Y-27632) and TGF-β inhibitor (SB 431542 or A83-01) were added into the upper and lower chambers, respectively, and the cultures were maintained under submerged conditions for 5-7 days until achieving confluency. Medium in the upper chamber was then removed to create an air-liquid interface. Cells were grown in ALI with 50% L-WRN medium (with neither ROCK inhibitor nor TGF-β inhibitor) in the bottom chamber. Cells were cultured for an additional 7-28 days to reach maturity, and media changed every 2-3 days. Details regarding small molecules and recombinant peptides can be found in **Supplementary Table S2**.

Histology

ALI cultured cells were fixed in 4% paraformaldehyde (PFA) for 30 min at room temperature. The fixed cells were then dehydrated in 70% ethanol and were embedded in 2% agar followed by paraffin embedding, sectioning, and staining. Bright field histological images were acquired with an Olympus BX61 microscope equipped with a DP22 camera and cellSens Standard v3.2 software. Cell height was quantified by averaging measurements from two slides per time point for each donor sample using cellSens Imaging software.

Immunostaining

Slides were de-paraffinized and subjected to heat-induced antigen retrieval in Trilogy solution for 20min. Slides were blocked in PBS containing 2% BSA, 5% serum and 0.1% Triton-X for 1hr at room temperature, and then treated with primary antibodies at desired dilutions overnight at 4°C. They were then washed 3 times in PBS and incubated with fluorophore-conjugated secondary antibodies for 1 h at room temperature. Fluorescent images were

acquired with a Zeiss Axiovert 200M or a Zeiss AxioObserver 7 inverted microscope equipped with a Colibri 7 Light Source and Axiocam 702 detector. FIJI (ImageJ) (v1.54f) was used to uniformly adjust brightness and contrast. Information about primary antibodies and dilutions can be found in **Supplementary Table S3**.

Transmission electron microscopy

Cultures grown on transwell inserts were washed once in PBS, then fixed overnight at 4°C in 2.5% Glutaraldehyde (Electron Microscopy Sciences, 16120) and 4% Paraformaldehyde (Electron Microscopy Sciences, 15710) in 0.2M Sodium Cacodylate Buffer (pH 7.4, Electron Microscopy Sciences, 11652). Fixative was removed and samples were stored in 0.2M Sodium Cacodylate Buffer at 4°C until further processing. Transwell membranes were removed from inserts, cut in half, then embedded in Eponate 12 Resin (Pelco, 18012). Sections were cut with a Leica EM UC7 ultramicrotome. Semi-thin sections were transferred to glass microscope slides, stained with toluidine blue and imaged with an Olympus BX43 microscope equipped with a DP22 camera using 20x (NA 0.50, WD 2.10 mm), 40x (NA 0.75, WD 0.51 mm), and 60x (NA 1.25, WD 0.12 mm) objectives. Thin (200 nm) sections were stained with osmium tetraoxide and uranyl acetate, transferred to copper grids, and imaged using an FEI Tecnai G2 Spirit BioTWIN transmission electron microscope equipped with an 11-megapixel Orius 832 CCD camera. Operating voltages were held constant within imaging sessions.

ATP assay

Human rectum 3D spheroids cultured in 50% L-WRN conditioned media with ROCK inhibitor and TGF-beta inhibitor were dissociated into small fragments and mixed with Matrigel at a concentration of 1000 cells/µL. The cell mixture was seeded onto a white opaque 96 well plate at 5 µL per well (5000 cells per well), and 100 µL of test media was overlaid per well. After 72 hours of incubation, an ATP assay was performed using the CellTiter-Glo 3D Cell Viability Assay (Promega, G9682), according to the manufacturer's protocol. Luminescence was measured using a Promega GloMax Discoverer Microplate Reader. Viability was calculated according to the following equations:

$$\text{(Eq. 1) } RLU = Lum_{Sample} - Lum_{Bkg}$$

Where RLU is Relative (corrected) Luminescence, Lum_{sample} is the luminescence value measured in sample wells, and Lum_{Bkg} is background luminescence values from wells with media only (no cells).

$$\text{(Eq. 2) } RLU\ Ratio = \frac{RLU_{Test}}{RLU_{CM}}$$

Where RLU_{CM} is the Relative Luminescence of cells grown in 50% L-WRN conditioned media and RLU_{test} is the Relative Luminescence of cells grown under experimental ("test") conditions.

Quantitative RT-PCR

RNA was extracted from ALI cells using the RNeasy Micro Kit according to the manufacturer's protocol. Briefly, ALI cells were harvested in RLT buffer with 2-mercaptoethanol and stored in -80°C until use. RNA was extracted and eluted in 14 µl of nuclease free water and stored at -80°C until use. RNA was quantified on Cytation 5 by diluting 1:1 with nuclease free water to determine the volume of RNA needed to convert to cDNA. cDNA was prepared using 1 µg of starting RNA and reverse transcribed using iScript reverse transcriptase kit (Bio-Rad, 708841) according to manufacturer's directions. Quantitative PCR reaction was conducted with TB Green Advantage qPCR 2X- premix (Takara, 639676) using a Bio-Rad CFX Connect Real-

Time PCR Detection System. Primers were validated for specificity using a water control and conducting melt curve analysis. Data were analyzed using the Delta-Delta Ct method. Primer sequences for quantitative PCR are listed in **Supplemental Table S4**.

RNAscope based in situ hybridization

ALI culture cells were fixed in 4% PFA for 30 min at room temperature and sequentially cryoprotected in 10% (w/v), 15% (w/v), and 20% (w/v) sucrose in PBS, and frozen in Tissue-Tek OCT compound (Sakura, 4583). Cryosections (5 μm thick) were used for RNAscope-based *in situ* hybridization according to the manufacturer's recommended protocols for RNAscope Multiplex Fluorescent Reagent Kit v2 (ACDBio, 323100). The probes Hs-LGR5 (ACDBio, 311021) and Hs-CA7 (ACDBio, 431961) were used in this study.

Sample preparation for scRNAseq

ALI transwells were washed once with 1X DPBS-EDTA (0.05mM) and incubated with TrypLE Express Enzyme at 37 °C for 25-30 mins, then pipetted 10 times with P1000 pipette and collected into a 15ml conical tube. Cell suspensions were incubated at 37°C until the cells mostly dissociate into singlets. The dissociated cell suspension was filtered through 20 µm strainer to purify cell singlets. A Denovix CellDrop FL automated cell counter was used to evaluate viability and singlet percentage of the prepared cell singlets. If necessary, the prepared cell singlets were sorted with BD FACS Melody Flow Cytometer by using Propidium Iodide and DRAQ5 double staining. The prepared cell singlets from ALI (viability≥70%, singlet percentage≥70%) were immediately processed according to 10X Single Cell 3' Reagent Kits V3.1 Next GEM protocol (document # CG000204). Initially, 10,000 single cells were targeted using the Chromium Controller for cDNA synthesis and barcoding. Subsequent steps involved evaluating cDNA quality and quantity with a Bioanalyzer High Sensitivity DNA assay. The cDNA was then used as input for fragmentation, end-repair, adapter ligation, and sample indexing. Bioanalyzer assays ensured proper library construction. Libraries were pooled, quantified using a Quantabio Q cycler, denatured, and sequenced on an Illumina Novaseq 6000 platform. The sequencing protocol comprised 28 cycles for the forward read and 91 cycles for the reverse read.

scRNAseq data processing

Sequence data were mapped onto the human genome (GRCh38) using CellRanger (v7.0.1) with default parameters. For downstream analysis, filtered gene-barcode matrices generated by CellRanger were read into the Seurat package (v5.0.0) on R (v4.3.0)/RStudio (v2023.09.1+494) (2). Plots were generated using RStudio and ggplot2, and further edited (for size and readability) with Adobe Illustrator CC. In Seurat, dead cells with high mitochondrial ratios (>20%) and potential doublets with high total transcript counts (>20,000) were first filtered out. The resulting data were normalized using SCTransform with default parameters. The top 3,000 most variable features were calculated, and integration was performed to remove patient-to-patient variance for clustering. The first 40 PCs were used for graph-based clustering, which was visualized with UMAP-based dimensional reduction (3). Marker genes for each cluster were calculated using the wilcoxauc function from the presto package (v1.0.0). Visualization of gene expression was performed using various functions within the Seurat package (i.e., FeaturePlot, DoHeatmap, and DotPlot).

Biomedical Foundation Model: Label Transfer Experiments

Biomedical Foundation Models for Targets (BMFM) were developed at IBM research and pre-trained on public datasets using state-of-the-art deep neural network architectures based on BERT models (4, 5). In general, BMFM family of models capture broad knowledge across the biomedical domain and have been open-sourced by IBM (https://github.com/BiomedSciAI) (6,

7). BMFM uses attention mechanisms for learning large-scale, context-specific modeling (8). As transformers, these models are particularly well-suited to understand gene network dynamics as they provide universal representations of knowledge regarding gene expression derived from scRNAseq. Application specific fine-tuning using either the scRNAseq data generated in this study or the data from Zhao et al., enabled both validation of the cell-type annotation task as well as classification of unknown cells (9). Given an input dataset, the model can classify each cell by assigning labels from the fine-tuning dataset and generating a probability score (0-1) for each assignment; subsequently, the label with the greatest probability score is used as the predicted cell-type classification.

For the reciprocal label transfer experiment, the pre-trained BMFM was fine-tuned with either i) the hALId7 scRNAseq dataset (BMFM-hALId7) or ii) the epithelial fraction of both terminal ileum and ascending colon from the non-IBD controls in Zhao et al. (BMFM-ZhaoCtrlEpi) (9). Labels in the datasets used for fine-tuning were generated by manual curation of Seurat outputs informed from historical wet-lab experiments that defined marker genes for different intestinal epithelial cell lineages. This training methodology is schematized in *SI Appendix* **Fig. S5**.

Label transfer experiments were carried out 5 times. For each fine-tuned model, we randomly sampled 80% of the data from the fine-tuning set, 10% of the data for internal test and 10% of the data for internal validation. Label probabilities of each classification from all 5 fine-tuned models were plotted in R using ggplot2 and ggradar packages.

Literature Gene Search and Analysis with Natural Language Processing

Document identification was performed using the Deep Search toolkit (10) and the full PubMed Central collection up to April 2023 was searched. All articles mentioning either one of the genes identified in the CA7/BEST4 cluster were selected and in the case of NOTCH2*, due to the very large volume of documents (N=5,153), only combinations of this gene with any of the others were considered (N=160). Filters were applied to only keep genes that were identified by an automatic biomedical annotator (11) and that were mentioned in paragraphs in the main article text. The general workflow is illustrated in *SI Appendix* **Fig. S6**. Combinatorial results by gene subset were visualized using UpSet plots (12).

ALI co-culture with Shigella flexnerii

Shigella flexneri Castellani and Chalmers (M90T Sm) was cultured in TSB media to an optical density OD_{600} = approximately 1.0. The bacterial culture was pelleted by centrifugation at 3500 x *g* for 3 minutes, after which the bacterial cells were resuspended in 2 mL of advanced DMEM with 20 mM HEPES. Bacterial suspension (100 µL) was applied to ALId19 monolayer cells at a density of 500 bacterial per transwell. The co-culture was centrifuged for 5 min at 500 x *g*, after which the bacterial suspension was aspirated. After 24 hours of co-culture, the cells were fixed and processed for immunostaining and imaging.

Statistical analysis, plotting, and data visualization

Statistical analyses and plotting were performed using Graphpad Prism (v9) and R/Rstudio (build/v). Statistical tests are reported in figure legends. A Mann-Whitney µ-test was used for two-sample comparisons, and ANOVA with Tukey's multiple comparison's test was used for comparing >2 groups. A P value < 0.05 was considered statistically significant. For comparison of lineage frequencies (**Fig. 1E**), a X^2 test was used.

Tables

Supplementary Table S1: Human subject characteristics

Number	Sex	Age	Race	Location	Diagnosis
1	Male	61	Caucasian	Rectum	Healthy Donor
2	Female	55	African	Rectum	Healthy Donor
3	Female	57	Caucasian	Rectum, ileum	Healthy Donor
4	Female	70	Caucasian	Rectum	Healthy Donor
5	Female	56	Caucasian	Rectum, ileum	Healthy Donor
6	Female	67	Caucasian	Ileum	Healthy Donor
7			Asian		Healthy Donor

Supplementary Table S2: Chemicals, Peptides, and other Reagents

Reagent	Vendor	Catalog_Number	Use
Y-27632 dihydrochloride	R&D Systems	1254	ROCK inhibitor
SB 431542	R&D Systems	1614	TGFβ inhibitor
A83-01	Tocris	2939	TGFβ inhibitor
EGF, mouse recoombinant	Peprotech	315-09	growth factor
IGF-1, mouse recombinant	BioLegend	590904	growth factor
FGF2, human recombinant	Peprotech	100-18B	growth factor

Supplementary Table S3: Antibodies

Target	Host	Conjugate	Vendor	Catalog_Number	Clone	Application	Working_Concentration	RRID
MUC2	Mouse	N/A	Santa Cruz Biotechnologies	sc-515032	F-2	immunofluorescence	1:100	AB_2815005
Chromogranin A	Rabbit	N/A	Abcam	ab15160	polyclonal	immunofluorescence	1:2000	AB_301704
5-HT	Goat	N/A	ImmunoStar	20079	polyclonal	immunofluorescence	1:5000	AB_10719366
GLP-1	Rabbit	N/A	Bioss	BS-0038R	polyclonal	immunofluorescence	1:200	AB_10855332
Neurogenin 3	Rabbit	N/A	Bioss	BS-0922R	polyclonal	immunofluorescence	1:200	AB_10856983
SST	Rabbit	N/A	Sigma Aldridch	HPA019472	polyclonal	immunofluorescence	1:1000	AB_1857360
β-Catenin	Mouse	N/A	BD Bioscience	610153	Clone 14/Beta-Catenin	immunofluorescence	1:1000	AB_397554
Shigella	Rabbit	N/A	Abcam	ab65282	polyclonal	immunofluorescence	1:1000	AB_1142846
Rabbit	Donkey	AlexaFluor 488	ThermoFisher	A-21206	polyclonal	immunofluorescence	1:200	AB_2535792
Mouse	Donkey	AlexaFluor 488	ThermoFisher	A-21202	polyclonal	immunofluorescence	1:200	AB_141607
Goat	Donkey	AlexaFluor 488	ThermoFisher	A-11055	polyclonal	immunofluorescence	1:200	AB_2534102
Rabbit	Donkey	AlexaFluor 594	ThermoFisher	A-21207	polyclonal	immunofluorescence	1:200	AB_141637

Supplementary Table S4: Oligonucleotides

Target	ENSEMBL_ID	Application	Amplicon_Length(bp)	Oligo_1 (5'-3')	Oligo_2 (5'-3')
MUC2	ENSG00000198788	qPCR	249	AGGATCTGAAGAAGTGTGTCACTG	TAATGGAACAGATGTTGAAGTGCT
CHGA	ENSG00000100604	qPCR	187	AGAATTTACTGAAGGAGCTCCAAG	TCCTCTCTTTTCTCCATAACATCC

Supplementary Table S5: Summary and Quality Control data for scRNAseq

Datasets

Dataset_S01: Marker gene sets for 11 clusters in differentiated ALI cultures
Dataset_S02: Label Transfer Experiment Results
Dataset_S03: DeepSearch Literature Search and Analysis Query of BEST4/CA7 marker gene signature

Results

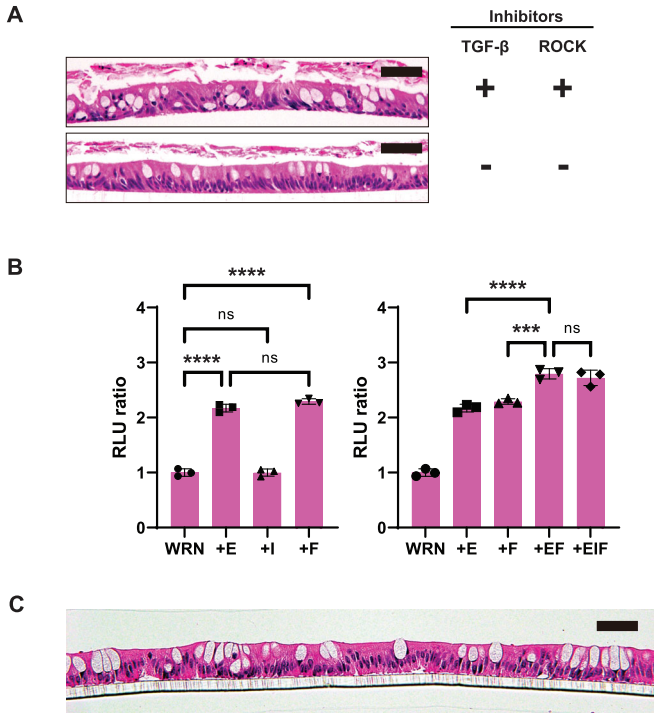

Fig. S1: Optimization of culture conditions for differentiation of human ISCs in ALI platform

(A) Brightfield photomicrographs of H&E-stained sections of human rectal ISCs differentiated under ALI conditions in the presence/absence of TBF-β blockade

(B) Effects of various growth factors on human ISC viability

(C) Brightfield photomicrograph of H&E-stained section of human ileal intestinal stem cells differentiated under ALI conditions

Supplementary Table S5: Small molecule inhibitors required for mouse and human ISC culture under various formats

Format	**3D (Matrigel)**		**2D (submerged)**		**2D (ALI)**	
Inhibitor	**TGF-β**	**ROCK**	**TGF-β**	**ROCK**	**TGF-β**	**ROCK**
Mouse	-	-	-	+	-	-
Human	+	+	+	+	-	-
Shared?	No	No	No	Yes	Yes	Yes

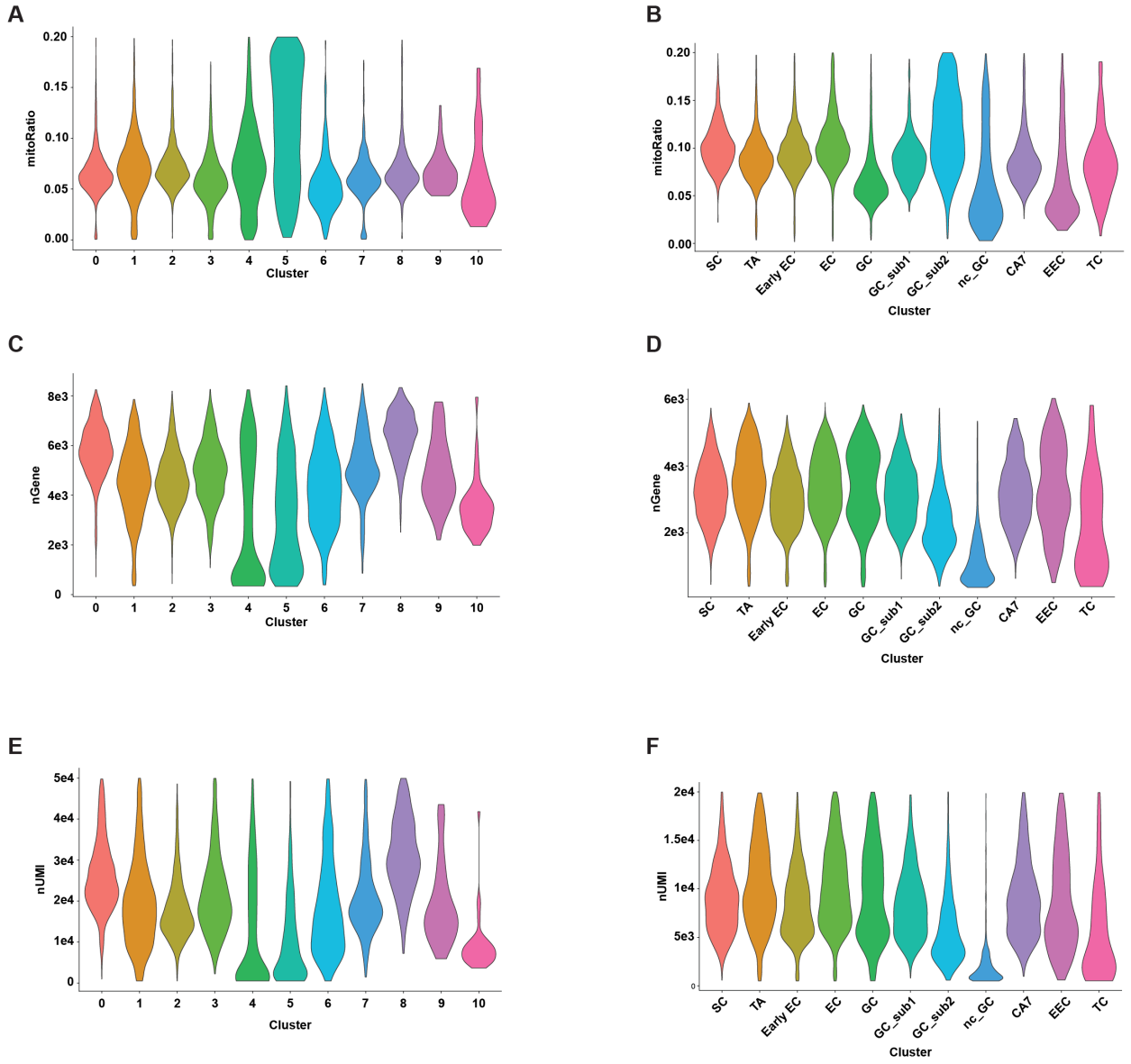

Figure S2. scRNAseq QC analysis by cluster
(A,B) Violin plots of mitochondrial ratio (mitoRatio) split by cluster from ALId2 and ALId7 samples. (C,D) Violin plots of number of genes (nGene) per cell split by cluster from ALId2 and ALId7 samples. (E,F) Violin plots of number of unique molecular identifiers (nUMI) per cell split by cluster from ALId2 and ALId7 samples.

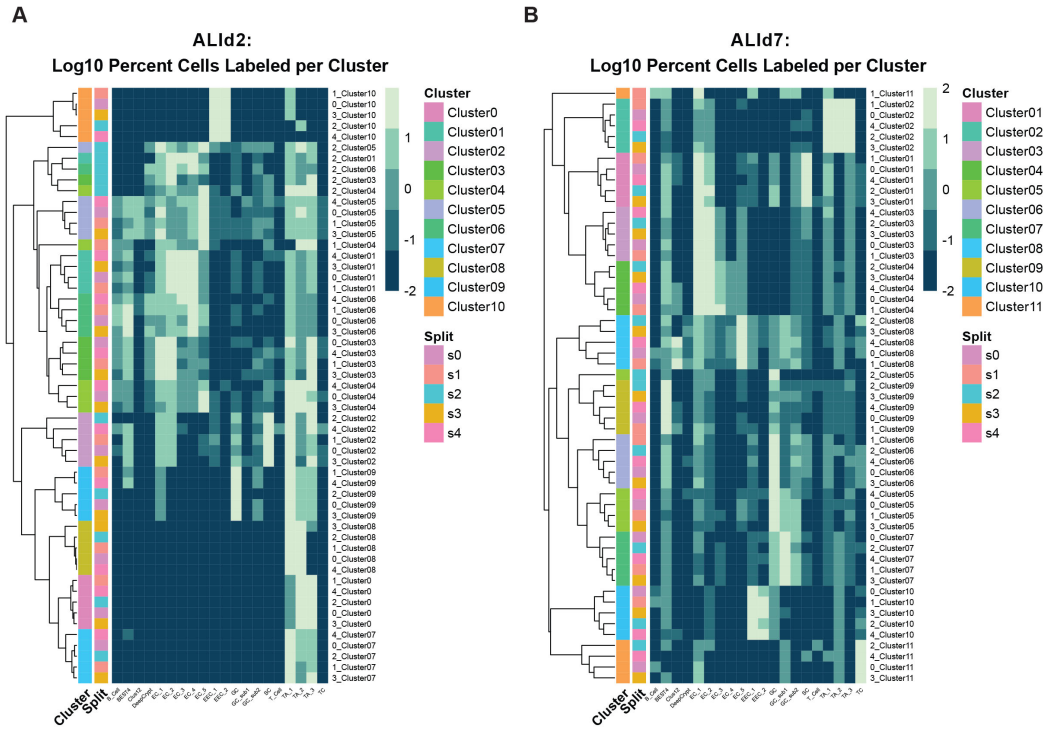

Figure S3: Summary of label transfer results: *in vivo* to *in vitro*

(A) ALI day 2 labelling frequencies by UMAP cluster

(B) ALI day 7 labeling frequencies by UMAP cluster

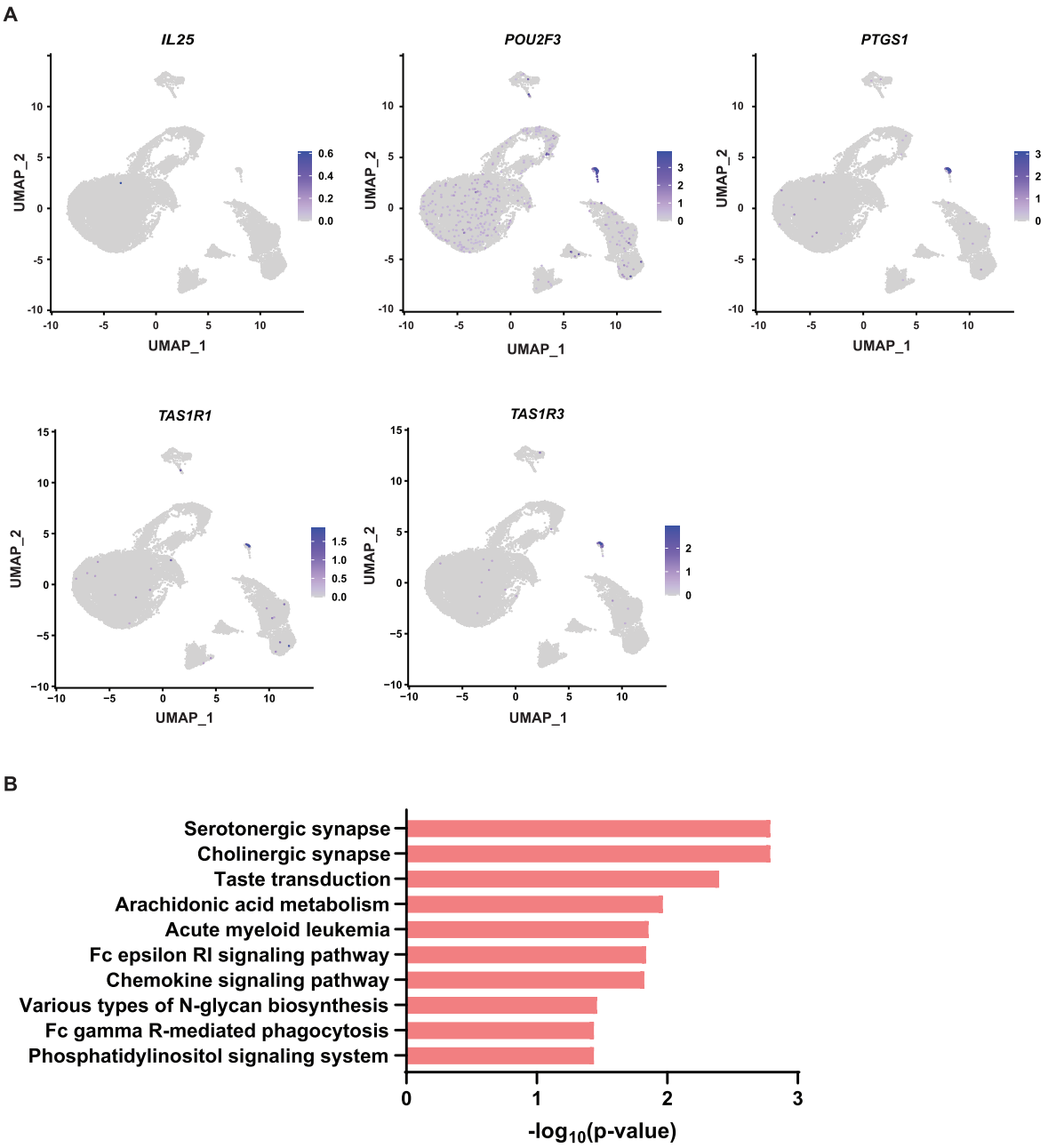

Fig. S4. Expression of canonical and non-canonical TC marker genes in ALI-differentiated human ISC cultures

(A) Feature plots of TC marker gene expression in UMAP space

(B) KEGG Pathway enrichment analysis of top 150 TC genes

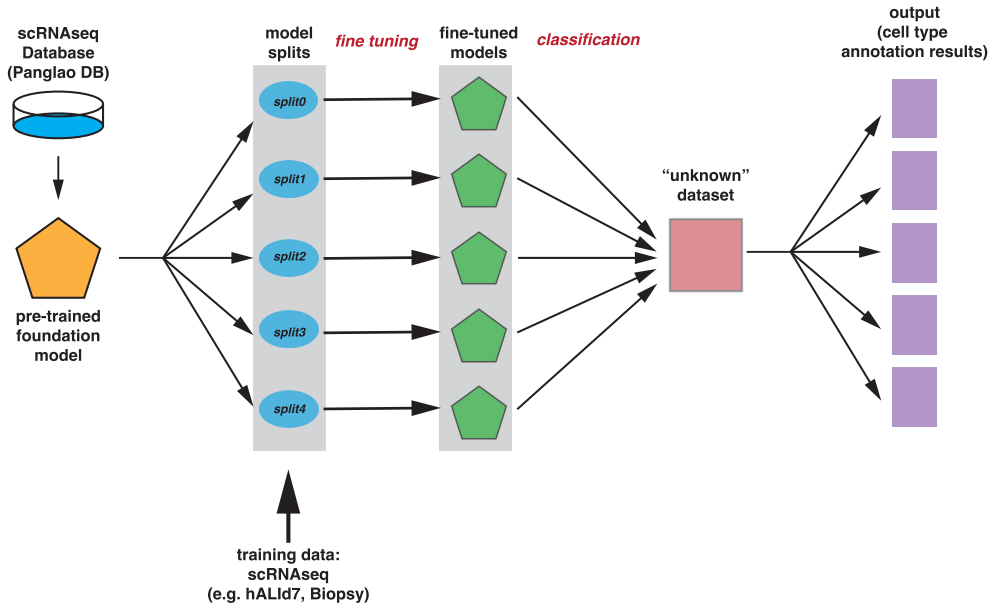

Fig S5. Schematic of BMFM scRNAseq analyses

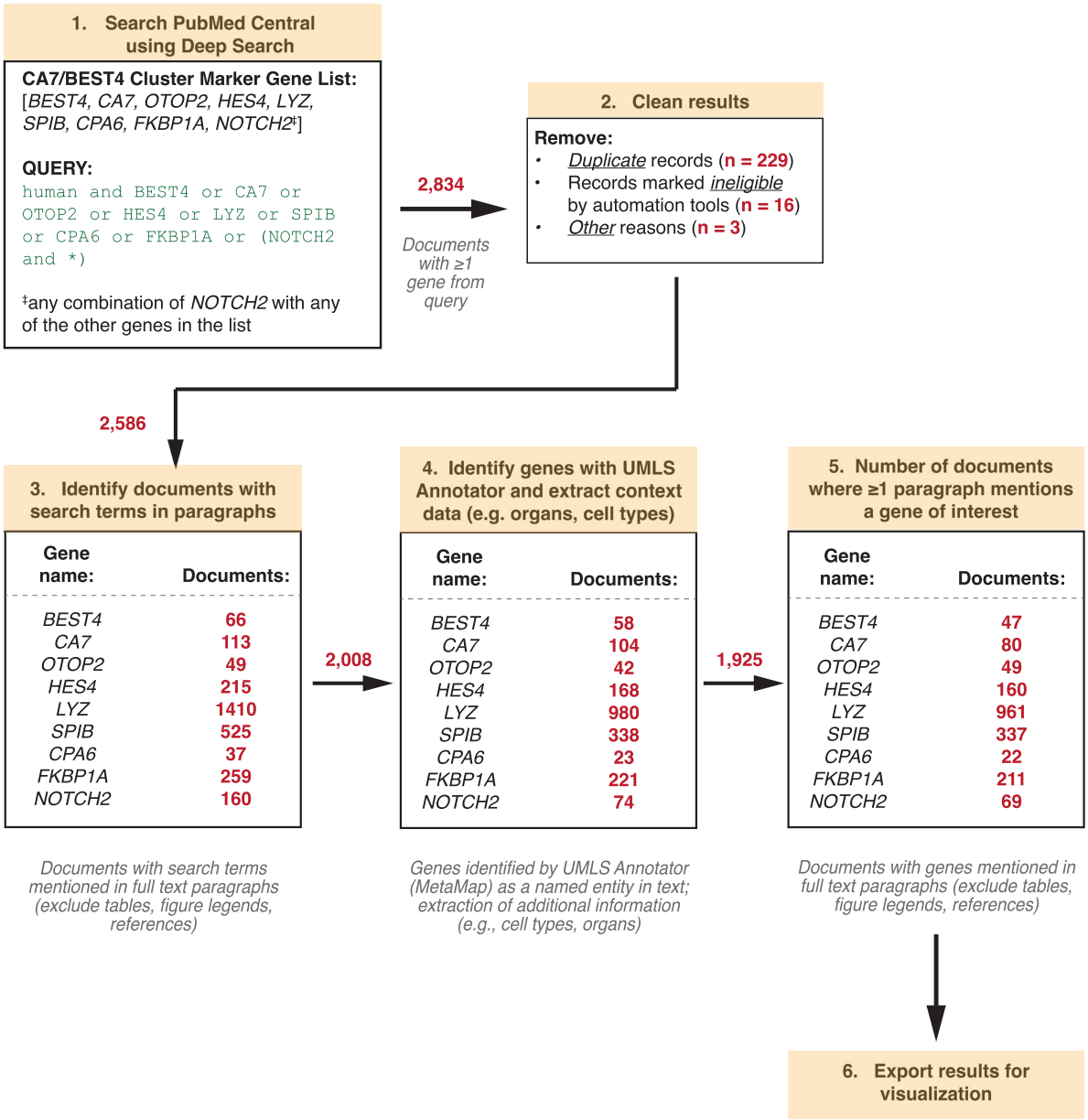

Fig. S6. Deep Search Workflow
Workflow for Literature Search and Analysis of BEST4/CA7 Marker Genes using Deep Search (PubMed Central collection).

References

1. H. Miyoshi, T. S. Stappenbeck, In vitro expansion and genetic modification of gastrointestinal stem cells in spheroid culture. *Nat Protoc* **8**, 2471-2482 (2013).
2. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
3. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 10.1038/nbt.4314 (2018).
4. F. Yang *et al.*, scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852-866 (2022).
5. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://dx.doi.org/10.48550/arXiv.1810.04805.
6. Y. Shoshan *et al.*, MAMMAL -- Molecular Aligned Multi-Modal Architecture and Language. http://dx.doi.org/10.48550/arXiv.2410.22367.
7. P. Suryanarayanan *et al.*, Multi-view biomedical foundation models for molecule-target and property prediction. http://dx.doi.org/10.48550/arXiv.2410.19704.
8. A. Vaswani *et al.*, Attention Is All You Need. http://dx.doi.org/10.48550/arXiv.1706.03762.
9. Y. Zhao *et al.*, Multiomic analysis reveals cellular, transcriptomic and epigenetic changes in intestinal pouches of ulcerative colitis patients. *Nat Commun* **16**, 904 (2025).
10. P. W. J. Staar, M. Dolfi, C. Auer, C. Bekas (2018) Corpus Conversion Service: A Machine Learning Platform to Ingest Documents at Scale. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, London, United Kingdom), pp 774–782.
11. D. Demner-Fushman, W. J. Rogers, A. R. Aronson, MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J Am Med Inform Assoc* **24**, 841-844 (2017).
12. A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983-1992 (2014).