Stem Cell Reports, Volume *4*

Supplemental Information

Transcriptome-wide Analysis Reveals Hallmarks of Human Intestine Development and Maturation In Vitro and In Vivo

Stacy R. Finkbeiner, David R. Hill, Christopher H. Altheim, Priya H. Dedhia, Matthew J. Taylor, Yu-Hwai Tsai, Alana M. Chin, Maxime M. Mahe, Carey L. Watson, Jennifer J. Freeman, Roy Nattiv, Matthew Thomson, Ophir D. Klein, Noah F. Shroyer, Michael A. Helmrath, Daniel H. Teitelbaum, Peter J. Dempsey, and Jason R. Spence

Figure S1, related to Figure 1.

(A) Principal component analysis was conducted HIOs (n=3 independently derived H9-HIOs), adult distal small intestine (n=4 independent biological specimens), adult duodenum (n=2 independent biological specimens), and fetal small intestine (n=6 independent biological specimens) (see also Table S1). The top two principal components account for 76.8% of the variation in the data and show that HIOs group with fetal intestinal tissues. (B) A dendrogram was generated based on hierarchical clustering of the gene sets using Canberra distance (B). Fetal tissues are more similar to each other than to their respective adult tissues and HIOs cluster within the same clade as the fetal tissues. Branch lengths indicate the level of dissimilarity between samples, with longer branch lengths indicating a greater level of dissimilarity. Red labels at branch point correspond to the Approximately Unbiased (AU) p-value. AU > 95 indicates that a given branch assignment is strongly supported by the data. Green labels at each branch point correspond to the bootstrap probability (BP) of a cluster, defined as the frequency of a given relationship among the bootstrap replicates. HuSI=human small intestine, HuSI Dist=human distal small intestine, HuSI Duo=human duodenal small intestine. (C) Spearman ranking was used to cluster samples and generate a heatmap of similarity with dark red indicating the highest level of similarity between samples. The dendrogram indicates that HIOs are most similar to fetal small intestine.

Figure S2, related to Figure 3

HIOs, tHIOs, and adult intestine were stained for markers of differentiated intestinal cells. HIOs express DppIV, a marker of enterocytes, and this expression is maintained after transplantation. Expression resembles staining seen along the brush border of adult intestine.

The goblet cell marker Muc2 is expressed in HIOs and expression is maintained, if not increased, in tHIOs and resembles the staining seen in adult intestine. Lastly, enteroendocrine cells, marked by Chromogranin A (ChgA), are a relatively more rare cell type found in the adult intestine. This is also true in HIOs and tHIOs in which ChgA staining is detected sporadically throughout the tissues.

Figure S3, related to Figure 3

(A) Normalized FPKMs were plotted for HIOs (n=3 independently derived H9-HIOs), fetal small intestine (n=6 independent biological specimens), and adult small intestine (n=6 independent biological specimens). FPKM values show that epithelial specific gene expression does not follow any broad trend. This rules out significant sampling bias for any particular group of samples and justifies the comparison of epithelial gene expression between sample types. Each sample is represented by a single black dot and red bars indicate the average FPKM value for each group. *p-value: 0.01-0.05, **p-value: 0.01-0.001, ***p-value: 0.001-0.0001. (B) Fatty acid binding protein (*FABP*) expression is increased in tHIOs relative to HIOs and is expressed to nearly the same level as in adult intestine (left). Meprin A, beta (*MEP1B*) is an example of a digestion related gene that is not expressed at a higher level in tHIOs compared to HIOs. n=3 independently derived H9-tissues for HIOs and tHIOs. n=3 biologically independent adult intestine tissues. *p-value: 0.01-0.05, **p-value: 0.01-0.001, ***p-value: 0.001-0.0001.

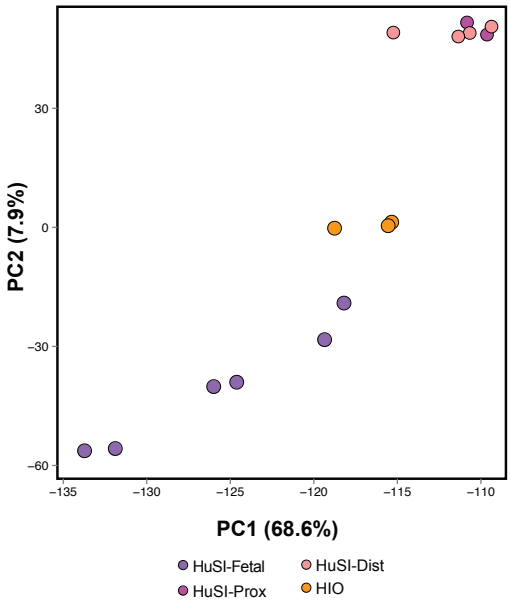
Figure S4, related to Figure 5.

Epithelium was isolated from HIOs through enzymatic and mechanical isolation. Epithelium cultured in the standard media used for growth and maintenance of HIOs does not expand over time (left panel). In 'complete media', which is used to grow human enteroids, the epithelium grows as budding structures that resemble human enteroids (middle and right panel). The epithelial structures derived from HIOs are referred to as HIO-derived epithelium (HDEs).

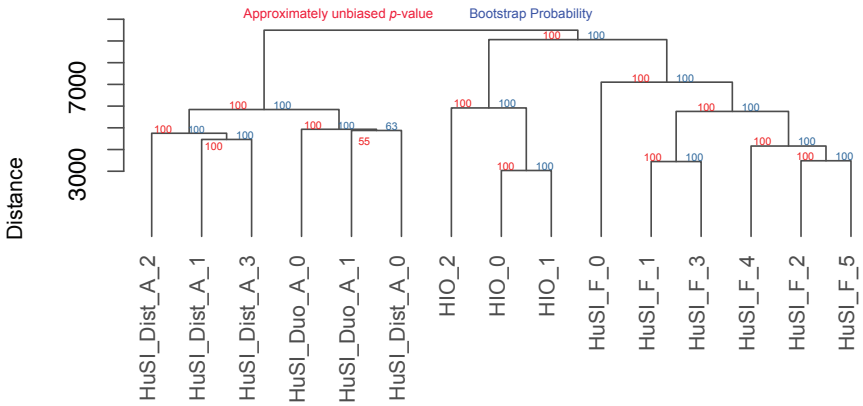
Figure S5, related to Figure 5.
(A) Low magnification image of the LGR5 *in situ* hybridization from the same tissue as shown in Figure 5. LGR5 expression in adult intestine confirms the specificity of the LGR5 probe by showing it is restricted to the crypts. Scale bar = 100um. (B) A slightly higher magnification image that is zoomed in on a villus tip confirms there is no LGR5 signal present on the villus tip. Scale bar = 50um. (C) A slightly higher magnification image that is zoomed in on a crypt highlights the specificity of LGR5 signal to the base of the crypt. Scale bar = 50um.

Fig. S1, related to Figure 1

A.



B.



C.

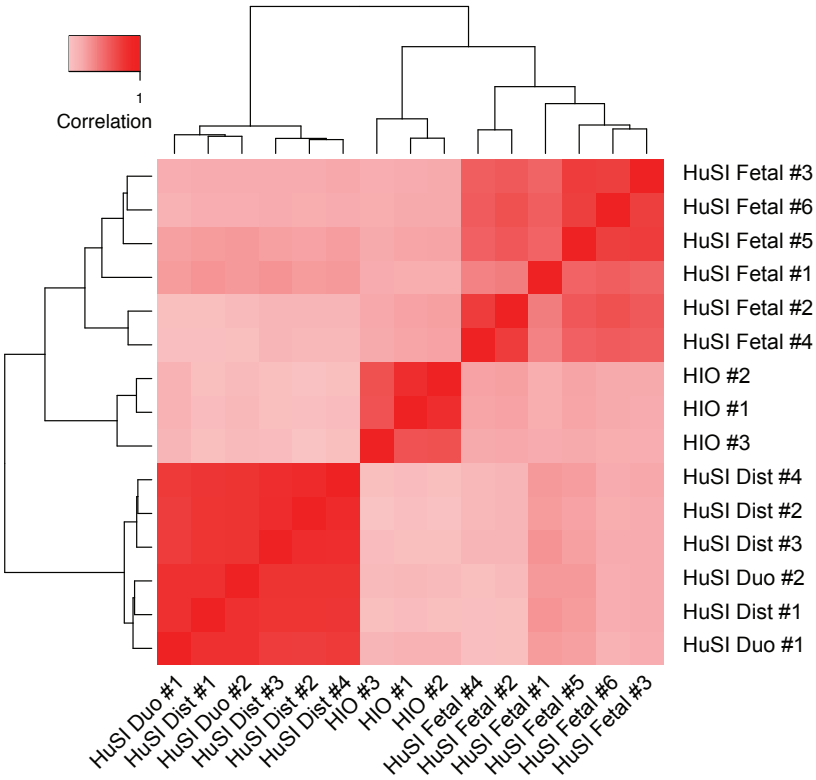


Figure S2, Related to Figure 3

A.

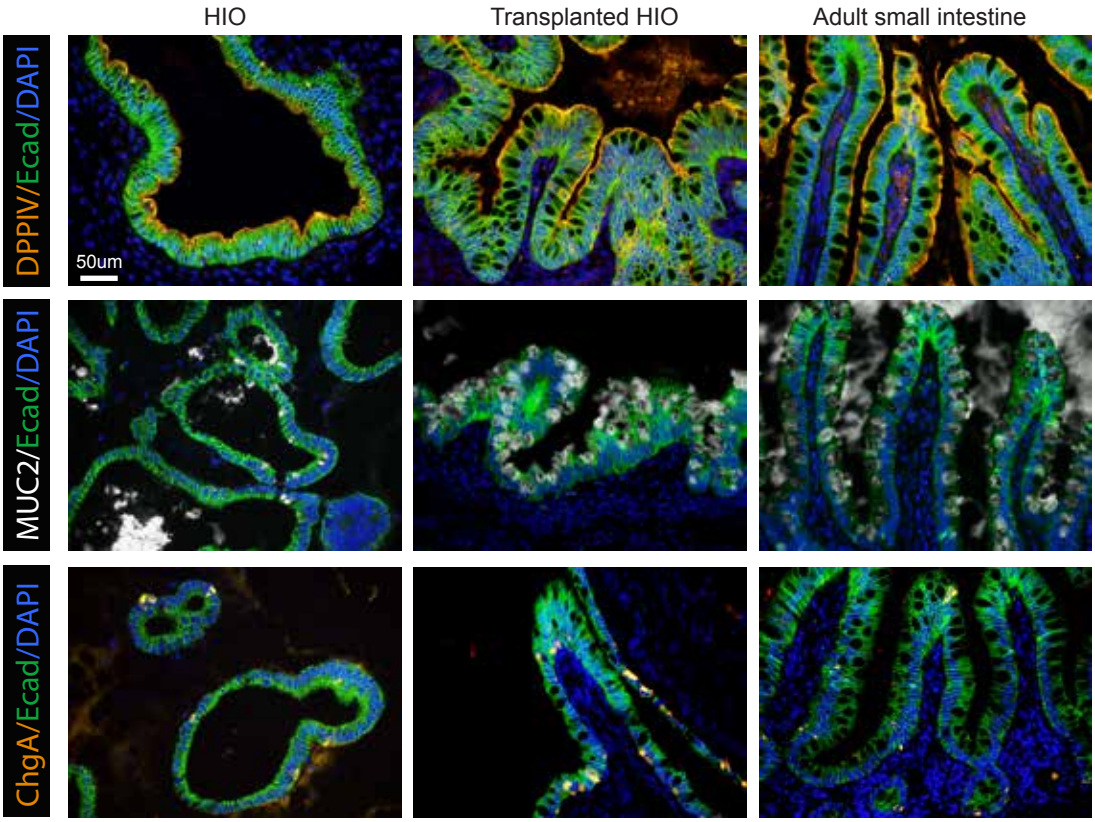
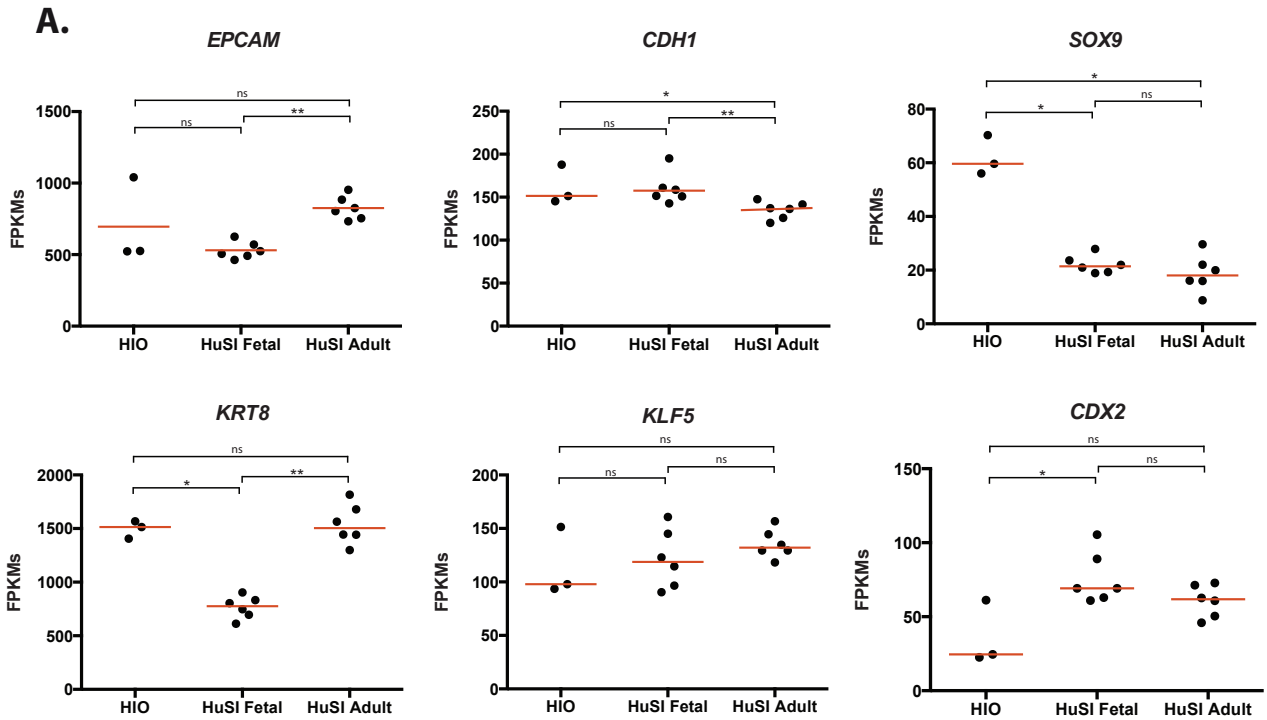


Figure Supplement 3- Related to Figure 3



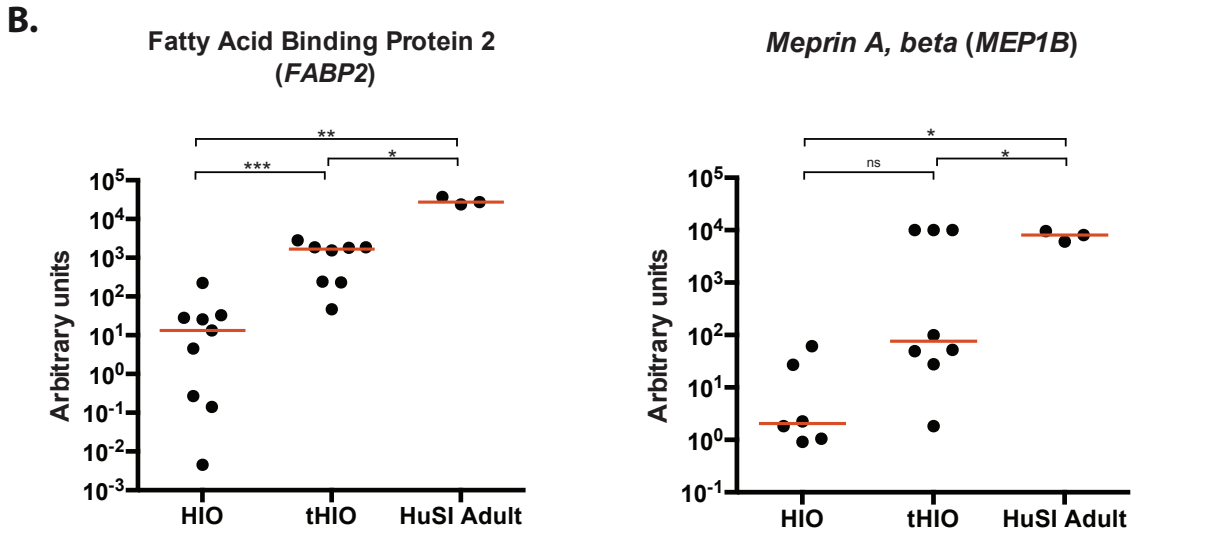


Figure S4 - Related to Figure 5

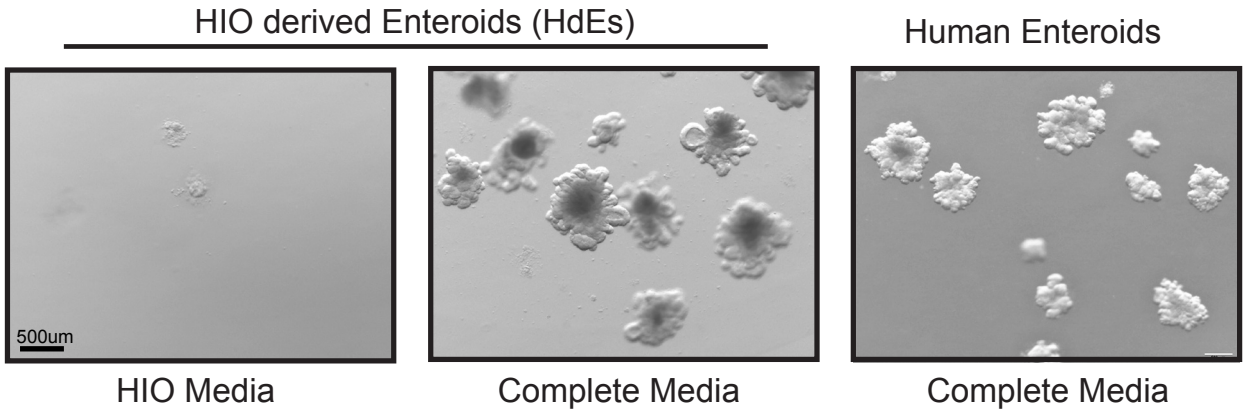


Figure S5, related to Figure 5

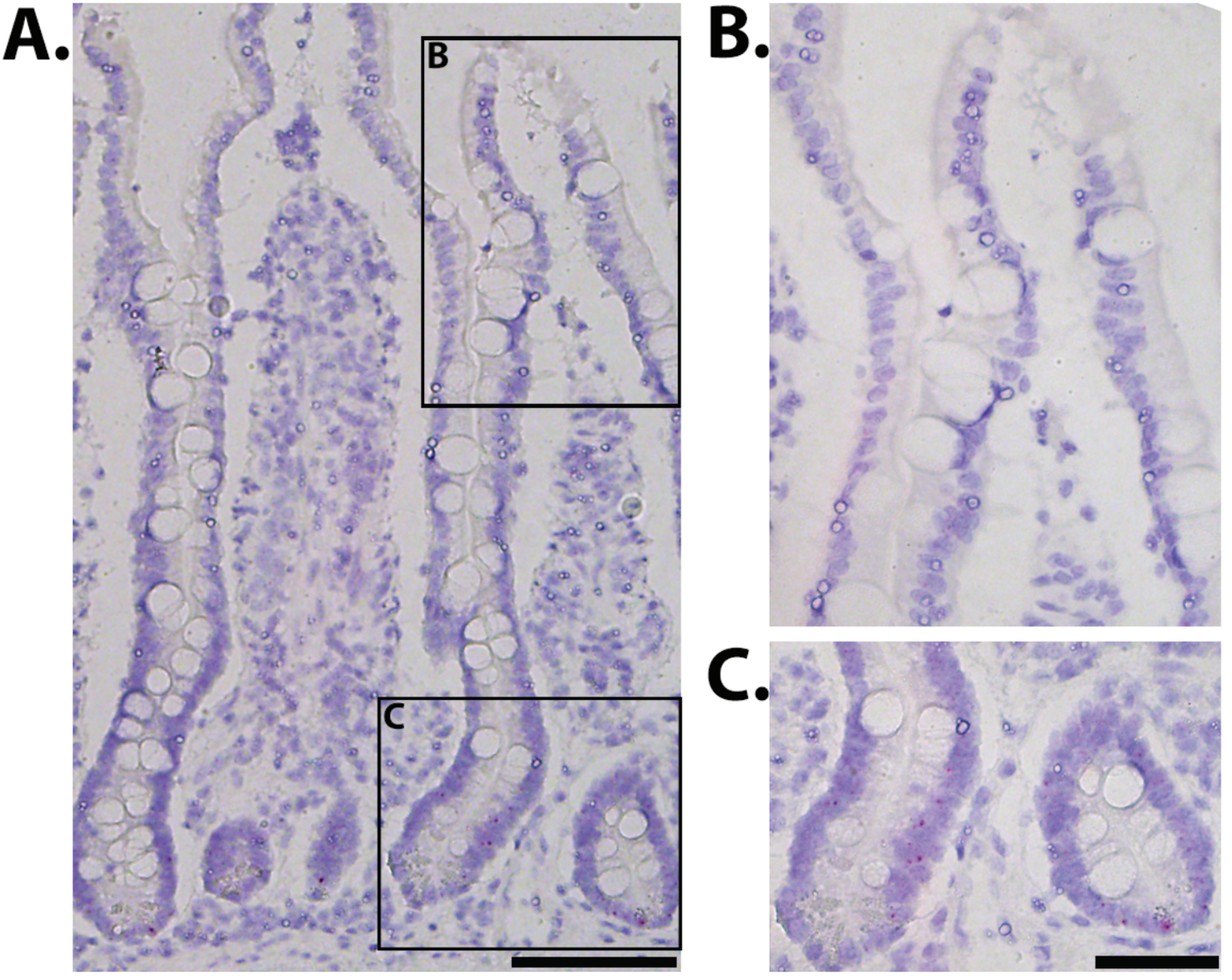


Table 1: RNAseq Datasets Downloaded from Public Databases

Sample Label	Description	Source	Donor ID	Accession #
HuSI.F91	Fetal Day 91, small intestine	GEO Datasets	H-23914	GSM1059508
HuSI.F98	Fetal Day 98, small intestine	GEO Datasets	H-23964	GSM1059521
HuSI.F108	Fetal Day 108, small intestine	GEO Datasets	H-23769	GSM1059486
HuSI.F108	Fetal Day 108, small intestine	GEO Datasets	H-23887	GSM1059507
HuSI.F115	Fetal Day 115, small intestine	GEO Datasets	H-23808	GSM1059517
HuSI.F120	Fetal Day 120, small intestine	GEO Datasets	H-23941	GSM1059519
HuSI.Duo.A1	Adult Duodenum	EMBL-EBI ArrayExpress	V145	E-MTAB-1733 (duodenum_4b)
HuSI.Duo.A2	Adult Duodenum	EMBL-EBI ArrayExpress	V150	E-MTAB-1733 (duodenum_4b)
HuSI.Dist.A1	Distal Small Intestine	EMBL-EBI ArrayExpress	V151	E-MTAB-1733 (small intestine_4a)
HuSI.Dist.A2	Distal Small Intestine	EMBL-EBI ArrayExpress	V152	E-MTAB-1733 (small intestine_4b)
HuSI.Dist.A3	Distal Small Intestine	EMBL-EBI ArrayExpress	V153	E-MTAB-1733 (small intestine_4c)
HuSI.Dist.A4	Distal Small Intestine	EMBL-EBI ArrayExpress	V156	E-MTAB-1733 (small intestine_4d)
HuStomach.F101	Fetal Day 101, stomach	GEO Datasets	H-24125	GSM1220600
HuStomach.F105	Fetal Day 105, stomach	GEO Datasets	H-24005	GSM1220597
HuStomach.F107	Fetal Day 107, stomach	GEO Datasets	H-23758	GSM1220590
HuStomach.F108	Fetal Day 108, stomach	GEO Datasets	H-23887	GSM1220587
HuStomach.F110	Fetal Day 110, stomach	GEO Datasets	H-23604	GSM1220589
HuStomach.F127	Fetal Day 127, stomach	GEO Datasets	H-24078	GSM1220598
HuStomach.A1	Adult stomach	EMBL-EBI ArrayExpress	V18	E-MTAB-1733 (stomach a)
HuStomach.A2	Adult stomach	EMBL-EBI ArrayExpress	V90	E-MTAB-1733 (stomach 3a)
HuStomach.A3	Adult stomach	EMBL-EBI ArrayExpress	V91	E-MTAB-1733 (stomach 3b)
HuColon.A1	Adult colon	EMBL-EBI ArrayExpress	V10	E-MTAB-1733 (colon a)
HuColon.A2	Adult colon	EMBL-EBI ArrayExpress	V11	E-MTAB-1733 (colon_b)
HuColan.A3	Adult colon	EMBL-EBI ArrayExpress	V14	E-MTAB-1733 (colon_c)

SI=small intestine, Dist=distal region of intestine, A=adult, F=fetal (numbers indicate days of gestation)

Table S7: Media Formulations for Enteroid Cultures, Related to Fig. 5

Reagent	Supplier	Catalog #	HIO Media	Complete Media	Basal Media
Advanced DMEM:F12	Life Technologies	12634	YES	YES	YES
B27 Supplement	Life	17504044	1X	1X	1X
N2 Supplement	Life Technologies		---	1X	1X
L-Glutamine	Life Technologies	25030	2mM	2mM	2mM
Pen-Strep	Life Technologies	15140	100/mL; 100ug/mL	100/mL; 100ug/mL	100/mL; 100ug/mL
HEPES	Life Technologies	15630080	15mM	15mM	15mM
EGF	R&D Systems	236-EG	100ng/mL	100ng/mL	100ng/mL
NOGGIN	R&D Systems	6057-NG	100ng/mL	100ng/mL	100ng/mL
R-Spondin 2 conditioned media	---	---	5%	25%	---
Chir99021	Tocris	4423	---	2.5uM	---
WNT3A conditioned media	---	---	---	25%	---
N-acetylcysteine	Sigma	A9165	---	2mM	---
Nicotinamide	Sigma	N0636	---	20mM	---
A-83-01	Tocris	2939	---	0.5 uM	---
SB202190	Sigma	S7067	---	2.5uM	---

Table S8: qRT-PCR Primers

Gene	Primer	Sequence
DEFA5	hDefA5-F	CCTTTGCAGGAAATGGACTC
	hDefA5-R	GGACTCACGGGTAGCACAAC
FABP2	hFABP2-F	GCTGCAAGCTTCCTTTTCAC
	hFABP2-R	CTGAAATCATGGCGTTTGAC
GAPDH	hGAPDH-F	CTCTGCTCCTCCTGTTCGAC
	hGAPDH-R	TTAAAAGCAGCCCTGGTGAC
IAP	hIAP-F	TCAGCTGGGTACTCAGGGTC
	hIAP-R	ATCGCCACTCAGCTCATCTC
LCT	hLCT-F	GGCAGTCTGGGAGTTTTAGG
	hLCT-R	ATGCCAAAATGAGGCAAGTC
LGR5	hLGR5-F	CAGCGTCTTCACCTCCTACC
	hLGF5-R	TGGGAATGTATGTCAGAGCG
LYZ	hLYZ-F	ACAAGCTACAGCATCAGCGA
	hLYZ-R	GTAATGATGGCAAAACCCCA
MEP1B	hMEP1B-F	AGATGGCCTCATACCATTCC
	hMEP1B-R	AGGCGATAACGTTCAAATGC
MGAM	hMGAM-F	GACATTGGCTGGGAGACAAC
	hMGAM-R	CTCCCGTATAGGATATGCCG
OLFM4	hOLFM4-F	ACCTTTCCCGTGGACAGAGT
	hOLFM4-R	TGGACATATTCCCTCACTTTGGA
REG3a	hReg3a-F	CTCCCTGGTGAAGAGCATTG
	hReg3a-R	TGCTACTCCACTCCCAACCT
SI	hSI-F	GGATGGTCACAGATGAAACCT
	hSI-R	TCCTCCCCATCGTCCACTA
TREH	hTREH-F	GGAGCTGTGCCTGCTACTG
	hTREH-R	CTCCCCGTGGCAGTAAATC

All primer sequences came from human qPCR Depot: http://primerdepot.nci.nih.gov/

Table S9: List of Antibodies Used

Antigen	Antibody Species	Company	Catalog #	Dilution
Alpha Defensin 5	mouse	Abcam	Ab90802	1:50
Human nuclear antigen	mouse	Millipore	MAB1281	1:100
Ki67	rabbit	ThermoScientific	RM-9106-S1	1:400
Lysozyme	goat	Santa Cruz	sc-27958	1:100
Olfm4	rabbit	Abcam	ab85046	1:100
Sucrase Isomaltase	goat	Santa Cruz	sc-27603	1:100
Vimentin	mouse	Sigma	V2258	1:200
Smooth muscle actin	mouse	Sigma	C6198	1:400
Pdgfra	rabbit	Santa Cruz	sc-338	1:100
Villin	goat	Santa Cruz	sc-7672	1:50
Mucin 2	rabbit	Santa Cruz	sc-15334	1:1000
Chromagranin A	goat	Santa Cruz	sc-1488	1:100

Supplemental Experimental Procedures

Generation of HIOs

Briefly, H9 embryonic stem cells were treated with Activin A (R&D Systems) for 3 days to generate endoderm. FGF4 (R&D Systems) and Chir99021 (STEMGENT), a Wnt agonist used in place of recombinant Wnt3a, were then used to pattern the endoderm into CDX2+ intestinal cells that spontaneously form floating, 3-dimensional aggregates called spheroids. Spheroids were collected and plated into droplets of Matrigel (BD Biosciences/Corning), a laminin-rich basement membrane complex. Spheroids were cultured in media containing EGF (100ng/mL, R&D Systems), R-Spondin 2 (5% conditioned media (Bell et al., 2008)), and Noggin (100ng/mL, R&D Systems) for 1 week and then in media containing only EGF and R-Spondin 2 thereafter as they grew into HIOs. HIOs used for experiments were grown *in vitro* for 1-2 months prior to use unless otherwise stated. Biological replicates of HIOs consist of pools of 5-8 HIOs per replicate.

RNA Sequencing

RNA sequencing data obtained from of three human intestinal organoid samples was appended to the normalized RNAseq data sets of 6 unique fetal intestinal tissue samples and 6 unique adult small intestinal tissue samples obtained from publicly available sources. The UM Bioinformatics Core concatenated sequence data from the UM sequencing core into a single. fastq file for each sample. Additional sequence data was retrieved from the EBI-AE database (Adult Samples) and NCBI-GEO (SRA) database (Fetal samples) (Table 1). The quality of raw reads for each sample were checked using FastQC (version 0.10.1) to identify features of the data that may indicate quality problems (e.g. low quality scores, over-represented sequences, inappropriate GC content, etc.). Adapter sequences were trimmed from the reads using Cutadapt (version 0.9.5) (Chen et al., 2014). The software package Tuxedo Suite was used for alignment, differential expression analysis, and post-analysis diagnostics (Langmead et al., 2009; Trapnell et al., 2009; 2013). Briefly, reads were aligned to the

reference transcriptome (UCSC hg19) (genome.ucsc.edu) using TopHat (version 2.0.9) and Bowtie (version 2.1.0.0). Default parameter settings were used for alignment, with the exception of allowing software to spend extra time searching for valid alignments and limiting the read mapping to known transcripts. In addition, we used FastQC for a second round of quality control (post-alignment), to ensure that only high quality data would be input to expression quantitation and differential expression analysis. Cufflinks/CuffDiff (version 2.1.1) was used for expression quantitation and differential expression analysis, using UCSC hg19.fa as the reference genome sequence and UCSC hg19.gtf as the reference transcriptome annotation. We used locally developed scripts to format and annotate the differential expression data output from CuffDiff. Briefly, we identified genes and transcripts as being differentially expressed based on three criteria: test status = “OK”, FDR < 0.05 and fold change ≥ 1.5. We annotated genes and isoforms with NCBI Entrez GeneIDs and text descriptions. We further annotated differentially expressed genes with Gene Ontology (GO) (www.geneontology.org) terms using NCBI annotation.

Gene ontology analysis and heatmaps

Biological process gene ontology categories with putative roles in intestinal development were identified using the Gene Ontology Consortium database (Ashburner et al., 2000). Genbank gene symbols classified within each gene ontology (GO) term were retrieved using the BioMart package for Bioconductor in R (Durinck et al., 2009; Gentleman et al., 2004). FPKM data corresponding to gene symbols within a GO term were retrieved from the normalized dataframe if the ANOVA P-value was < 0.05 for the comparison among the three sample groups and mean FPKM was greater than 1 in at least one of the three sample groups. Thus, downstream gene ontology analysis included only genes which differed significantly in expression between the HIO, fetal, and adult samples and were reproducibly expressed in at least one sample group. Normalized Z-scores for each gene within a GO term that met the inclusion criteria for were calculated as follows:

$$z = (x - \mu) / \sigma$$

Where *x* is an FPKM value, μ is the mean FPKM among all samples in all three groups for a given gene, and σ is the standard deviation among all samples in all three groups for a given gene. Heatmaps and associated dendrograms were plotted in R using the function “heatmap.2” in the package gplots (Warnes et al., 2014).

Principal Component Analysis

The complete FPKM matrix, containing frequency counts for all 23,615 known genes for all 33 RNAseq samples, was evaluated using principle component analysis (PCA) to visualize and quantify multi-dimensional variation between samples. Of the 23,615 known human genes, 1,281 (5.4%) were not detected in the RNAseq analysis of any of the 33 samples. Data scaling was applied to give equal weight to each gene as an equal component of the variation between samples. Variance (σ) can be expected to be highly variable among the 22,334 expressed genes contained in the dataset; thus scaling is advisable. Note that scaling is precluded by the presence of zero/missing data. To permit data scaling, it was necessary to remove all genes for which σ is equal to zero across all samples as it would be impossible to scale the variation in the remaining 22,334 genes to σ = 0. The advantage of data scaling is that specific genes with wide variation, perhaps varying in expression by many orders of magnitude, do not outweigh genes that may vary comparatively little but nonetheless may have biological significance. Likewise, in the absence of data scaling, widely variable genes increase the apparent dissimilarity between two samples, which may in fact share similar levels of expression in a large number of other genes. Thus, the range of variation in expression of each gene is normalized across samples such that each gene carries the same weight in the PCA. Data was not centered to avoid distortion of the axis regarding up- vs. down-regulation of transcription. Principle components were calculated using the function ’prcomp’ found in the R (version 3.1.2) statistical programming language (R Core Team, 2014) and plotted using the R package ’ggplot2’ (Wickham,

2009).

Hierarchical Clustering

Hierarchical cluster analysis based on the Canberra distance (Lance and Williams, 1966) between FPKM vectors was used to classify discrete RNAseq samples according to the degree of total transcriptional dissimilarity indicated by the normalized FPKM values. Bootstrap analysis was used to assess the uncertainty in the assigned hierarchical clustering relationships. 10,000 bootstraping iterations were generated by repeatedly randomly sampling the FPKM dataset. The bootstrap probability (BP) of a cluster is defined as the frequency of a given relationship among the bootstrap replicates. Multiscale bootstrap resampling was used to calculate an approximately unbiased (AU) *p*-value for a given relationship, with AU > 95 indicating a high degree of statistical significance. Analyses were conducted using R package 'pvclust' (Suzuki and Shimodaira, 2006).

The Canberra distance is a variation of the Spearman correlation that slightly penalizes the ranked comparisons with the greatest differences (Jurman et al., 2009). This method has been justified in genomics as a correction for intra-experimental variation, which in this experiment could be differences in the cell types present in HIO compared to fetal samples. HIOs contain only epithelium and mesenchyme and lack additional components such as neuronal, vascular, and immune cells that are present in the human fetal stomach and small intestine. The Canberra distance method may therefore exaggerate apparent differences thus resulting in a dendrogram (Fig. 1B) in which fetal small intestine and fetal stomach appear more related than HIOs and fetal small intestine.

Spearman Correlation

Spearman correlation was applied as an additional assessment of the cumulative degree of correlation among RNAseq datasets. We computed Spearman rank correlation coefficients (ρ) in a pairwise manner among all RNAseq samples using the complete normalized FPKM data. The Spearman coefficients were plotted as a heatmap using the function 'heatmap.2' in the R package 'gplots' (Warnes, 2014).

qRT-PCR

RNA was extracted from H9 embryonic stem cells, HIOs, tHIOs, and adult small intestine using the Trizol (Life Technologies) extraction protocol from the manufacturer. cDNA was generated using SuperScript®Vilo™ reverse transcriptase (Life Technologies) and qRT-PCR reactions were carried out using QuantiTect Sybr® Green (Qiagen). All reactions were run with 40 cycles of 95°C for 15 seconds, 55°C for 30 seconds, and 72°C for 45 seconds, followed by a melt curve of 95°C for 15 seconds, 60°C for 1 minute and then increasing temperature up to 95°C at 0.5 degree increments. Primer Sequences used are shown in Table S8.

Microscopy

Frozen or paraffin sections of HIOs, tHIOs, human small intestine, and mouse intestine were stained according to standard immunofluorescence and immunohistochemistry protocols. Generally, samples were blocked and permeabilized with PBS containing 5% donkey serum and 0.5% Triton-X 100. Primary antibodies were applied overnight at 4°C in 5% donkey serum and 0.05% Tween20/PBS. Lysozyme staining required antigen retrieval on both frozen and paraffin sections. For OLFM4 staining, tissues were incubated with primary antibody for 1 hr at room temperature. See Table S9 for a list of antibodies used. Alexa Fluor® conjugated secondary antibodies (Life Technologies) were then applied for 1 hr in 5% donkey serum/0.05% Tween20/PBS. Nuclei were stained with DAPI and mounted with ProLong® Gold antifade mountant (Life Technologies). Ki67 staining was done using a TSA kit (Life Technologies). All images were taken on an Olympus IX71 epifluorescent microscope, Nikon A1 confocal microscope, or a Hitachi H7600 transmission electron microscope. All post acquisition image processing (brightness and contrast) was applied uniformly to all comparable images.

LGR5 *in situ* hybridization

HIO, tHIO, and adult paraffin sections were sent to Advanced Cell Diagnostics for their RNAScope® *in situ* hybridization services. Samples were processed according to their publicly available standard protocol using their RNAScope® 2.0 High Definition Red Kit and returned for subsequent imaging. Images were taken on an Olympus BX-51 light microscope using a 100X objective.

Supplemental References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Bell, S.M., Schreiner, C.M., Wert, S.E., Mucenski, M.L., Scott, W.J., and Whitsett, J.A. (2008). R-spondin 2 is required for normal laryngeal-tracheal, lung and limb morphogenesis. Development *135*, 1049–1058.

Chen, C., Khaleel, S.S., Huang, H., and Wu, C.H. (2014). Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code for Biology and Medicine *9*, 8.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184–1191.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Jurman, G., Riccadonna, S., Visintainer, R., and Furlanello, C. (2009). Canberra distance on ranked lists. Proceedings, Advances in Ranking-NIPS 09 Workshop 22–27.

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

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540–1542.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology *31*, 46–53.

Warnes, G.R., Ben Bolker, Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., et al. (2014). gplots: Various R programming tools for plotting data.

Wickham, H. (2009). ggplot2 (New York, NY: Springer New York).