PNAS
www.pnas.org

Supplementary Information for

A cell atlas of the adult *Drosophila* midgut

Ruei-Jiun Hung[a,1], Yanhui Hu[b,2], Rory Kirchner[c,2], Yifang Liu[a,b], Chiwei Xu[a], Aram Comjean[b], Sudhir Gopal Tattikota[a], Fangge Li[b], Wei Song[a], Shannan Ho Sui[c] and Norbert Perrimon[a,e,1]

[a] Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02115

[b] *Drosophila* RNAi Screening Center, Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA02115

[c] Bioinformatics Core, Harvard T.H. Chan School of Public Health, Boston, MA 02115

[e] Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115

[2] These authors contributed equally

[1] Corresponding authors: Ruei-Jiun Hung, Norbert Perrimon

rjhung@genetics.med.harvard.edu, perrimon@receptor.med.harvard.edu

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1916820117

Supplemental Methods

Single-cell suspension preparation. *Drosophila* guts were dissociated to single cells as previously described (1) with a few modifications. Guts were dissected from 7-day old adult *esg-sfGFP/+, pros-Gal4>RFP/+* females. After pulling out the gut, the crop and midgut/hindgut junction (where the Malpighian tubules branch out of the gut) and Malpighian tubules were removed. Five flies at a time were dissected and immediately transferred into cold PBS containing 1% BSA to avoid exposing the midgut tissue to room temperature for a long period of time. Once 40 guts were dissected, they were transferred to a dissection plate in a droplet and chopped into small pieces using a razor blade. These small fragments were immediately transferred to an eppendorf tube containing 400 µl of 1 mg/ml elastase/PBS solution (Sigma-Aldrich, E0258) and incubated on a shaker at 27°C for 30 min. 1% BSA (final concentration) in PBS was used to stop the digestion reaction and prevent cells from aggregating. The cell suspension was passed through a 100 µm and then 40 µm cell strainer and loaded on the top of Optiprep (Axis-Shield) with a density gradient of 1.12 g/ml. Viable cells were isolated from the top layer of the sample after centrifugation at 800xg for 20 min. Cell viability and number were assessed by 0.4% trypan blue staining and cell counting using a hemocytometer. Cells (>160 cells/µl) were encapsulated at the Single Cell Core at the ICCB-Longwood Screening Facility at Harvard Medical School with inDrop (2). Reverse transcription and library preparation were done at the same facility as previously described (2). For 10x Genomics, the same cell suspension preparation was used and subsequent steps were done following the manufacture's protocol.

High-throughput sequencing. Before sequencing, the fragment size of each library was analyzed on a Bioanalyzer high-sensitivity chip. Libraries were diluted to 1.5 nM and quantified by qPCR using primers that recognize the p5-p7 sequence. InDrop libraries were sequenced on a Nextseq 500 instrument (Illumina) with the following sequencing parameters: 61 bp read 1 – 8 bp index 1 (i7) – 8 bp index 2 (i5) – 14 bp read 2. Sequencing was done at Biopolymers facility at Harvard Medical School.

Datasets processing. Reads were processed using the inDrop v3 pipeline implemented in bcbio-nextgen version 1.0.5a0-9ae5245 (3). Briefly, dual cellular barcodes, sample barcodes and UMIs were detected using umis (4), correcting cellular and sample barcodes of edit distance one from expected barcodes. Cells with less than 500 total reads assigned were discarded. Reads were aligned to the *Drosophila melanogaster* transcriptome from FlyBase, version FB2017_03 (5) using RapMap (6), assigning evidence `e` of 1/N for each read aligning to N different transcripts. Evidence was summed across all transcripts of a gene, and thresholded at a minimum evidence of 1. Quality control was performed using bcbioSingleCell (7), filtering out poor quality cells by keeping cells with the following metrics:
1. >= 100 UMI counts per cell
2. >= 100 genes detected per cell
3. log10(genes detected)/log10(UMI counts per cell) >= 0.7 (complexity)
4. percentage of mitochondria genes <50%

Following these criteria, a total of 7626 cells across two replicates remained.
For 10x Genomics, reads were processed using Cell Ranger. Quality control was filtering out poor quality cells by keeping cells with the following metrics:
1. >= 100 UMI counts per cell
2. >= 100 genes detected per cell
3. log10(genes detected)/log10(UMI counts per cell) >= 0.75 (complexity)
4. percentage of mitochondria genes <25%

Following these criteria, a total of 2979 cells across two replicates remained.
We combined the quality-controlled raw counts from the 10x Genomics and inDrop sequencing runs and performed clustering and cell type identification on the combined counts using Seurat v.3.0.2 (8). We observed a technology-specific effect across a variety of cell types in the clustering, which we corrected for using the FindIntegrationAnchors function in Seurat (8, 9). We then reclustered and reclassified the cells based on their integrated profiles.

The marker genes for each cluster were determined by using the command "FindMarkers (integrated, ident.1=1, test.use="roc")" in Seurat. This approach identifies markers that define clusters via differential expression using ROC analysis (receiver operating curve). There are many algorithms that can be used to try to find markers using Seurat. We tried different algorithms and the ROC analysis consistently gave results that look reasonable. Therefore, we decided to use ROC to identify markers that define clusters. AUC (area under the ROC curve) evaluates if the particular gene alone can be used to classify between two clusters of cells. An AUC value of 1 means that expression values for this particular gene alone can perfectly classify the two clusters. An AUC value of 0 also means that there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Power is calculated as (abs(AUC-0.5)*2), ranked matrix of putative differentially expressed genes.

Trajectory analysis was performed using Slingshot (10) on the integrated data. We chose cells clearly identified as ISC/EB, EC and EE cells (detail cell clusters are ISC/EB, AstA-EE, NPF-EE, AstC-EE, dEC, aEC1-3, pEC1-3 and mEC) and ran PCA on the union of marker genes for the cell types mentioned above. Trajectories were calculated from the first 30 PCA components, specifying ISC/EB cells as the start state and EE cells as the end state. This resulted in identification of three separate trajectories for ISC/EB cells. All the analyses scripts were deposited to the github: https://github.com/hbc/drosophila-midgut-analysis

Gene set enrichment analysis. The marker genes identified using Seurat that passed the ROC test (AUC>0.5) and fold change over other clusters > 2^(0.58) ~ 1.5 fold change were used for gene set enrichment analysis. Gene sets for major functional groups were collected from the GLAD database (11) and the full gene list of each GLAD group is listed in Dataset S3. Gene sets for metabolic pathways were from the KEGG database (12-14) and gene sets for cellular compartments were from gene ontology annotation (15, 16). In addition, the transcriptional target genes of major signaling pathways were assembled manually from the literature. P-value enrichment was calculated based on the hypergeometric distribution. The strength of enrichment was calculated as negative of log10(p-value), which is used to plot the heatmap.

Cell type comparison between *Drosophila* and mammalian intestinal epithelium. Haber et. al used a 3' droplet-based and a full-length plate-based approach to generate scRNA-seq (17). We mainly took the marker genes from the 3' droplet-based approach because inDrop and 10x Genomics are also 3' droplet-based approaches. For cell types such as stem cells that have only one marker, we selected more genes from full marker gene list by lowing the p-value threshold. We used DIOPT (release 7, score>3) (18) to map mouse genes to *Drosophila* orthologs. The Drosophila orthologs of markers identified in various cell types from mammalian datasets were grouped respectively as cell type specific gene sets. Next, we compared the cell type specific gene sets from the mammalian study and the cell type markers from Drosophila. P-value of enrichment was calculated based on the hypergeometric distribution and the similarity of Drosophila markers with mammalian markers is reflected by the negative log10 of the P values.

Transcription factor binding site from the Chip-seq data and co-expression of gut hormones. We looked for common transcription factor binding sites 5 kb upstream of any combination of 4, 3 or 2 gut hormones out of 15 from Chip-seq data (19). The frequency was calculated using the following formula: (number of cells expressing NPF, Tk and Orcokinin / number of cells expressing NPF) x (number of cells expressing NPF, Tk and Orcokinin / number of cells expressing Tk) x (number of cells expressing NPF, Tk and Orcokinin / number of cells expressing Orcokinin), for example.

Fly genetics. The following strains were obtained from the Bloomington *Drosophila* Stock Center: *klu mi05554* (BL44148), *y v; attp2* (landing site only, BL36303), *y v; UAS-LucRNAi, attp2* (BL31603), *UAS-klu RNAi* (BL28731 and BL64967), *zip-Gal4* (BL48187), *lola-Gal4* (BL45325), *lrch-Gal4* (BL63768), *insc-Gal4* (BL25773), *UAS-mCD8.ChRFP* (BL27392). Strains from the

Perrimon lab stock collection are: *esgGal4 UAS-GFP tubGal80^{ts} (EGT)*, *esg-lacZ* (*esg^{k00606}*), *Su(H)-LacZ*, *esg-sfGFP* (generated by David Doupe), and *esg-sfGFP, UAS-mCherryCAAX* (generated by Li He).

Flies were reared on standard cornmeal/agar medium in 12:12 light:dark cycles at 25°C unless noted otherwise. The flies were transferred to fresh food vials every two days. Conditional expression in adult flies using *tubGal80ts* was achieved by maintaining flies at 18°C until four days after eclosion, then shifting young adults to 29°C for 1 week. Each vial typically consisted of 10 females and 5 males. Only females were used in experiments.

Staining and fluorescence imaging. 7-day old (unless otherwise indicated) adult female midguts were dissected in PBS and fixed for 1 hr with PBS containing 4% paraformaldehyde. Samples were then rinsed with PBS three times and incubated with PBST (PBS with 0.2% Triton X-100) containing 5% of normal donkey serum for 45 min. Midguts were stained at 4°C overnight with the primary antibodies in PBST/ Primary antibodies used were: mouse anti-Prospero (1:100; #MR1A, Developmental Studies Hybridoma Bank), chicken anti-GFP (1:200; Aves, GFP-1020), and mouse anti-β-galactosidase (1:1000; Promega). Fly midguts were washed with PBST three times and incubated with secondary antibodies (1:1000) and DAPI (1:1000) in PBST at room temperature for 1 hr in the dark. Secondary antibodies were donkey anti-chicken and anti-mouse conjugated to Alexa-488 or Alexa-594 (Molecular probes). Fly midguts were then washed with PBST three times, mounted in Vectashield (Vector Laboratories). Images were captured with a Zeiss LSM780 confocal microscope equipped with 20x oil lens. All images were adjusted and processed using Fiji.

qRT-PCR. 10 midguts from females were dissected, placed into TRIzol reagent (Thermo Fisher), and homogenized with Bullet Blender (Next Advance, Inc.). RNA was extracted using Direct-zol (Zymo research) and converted to cDNA using SuperScript II reverse transcriptase (Invitrogen). RT-qPCR was performed using SYBR Green Supermix (Bio-Rad) with *rp49* as an internal control. Primers used for RT-qPCR are as follows:

AstA_FW: 5'-GACCTGGCCGACAGAACAAG
AstA_RV: 5'-AAAGTTGAAGGGTTGCGGAC
Tk_FW: 5'-CAATTCCTTTGTGGGGATGCG
Tk_RV: 5'-CTGCTGTTTTCCTCTCAAGTCAT
rp49_FW: 5'-ATCGGTTACGGATCGAACAA
rp49_RV: 5'-GACAATCTCCTTGCGCTTCT

References:

1. Dutta D, Buchon N, Xiang J, & Edgar BA (2015) Regional Cell Specific RNA Expression Profiling of FACS Isolated Drosophila Intestinal Cell Populations. *Curr Protoc Stem Cell Biol* 34:2F 2 1-14.
2. Zilionis R*, et al.* (2017) Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* 12(1):44-73.
3. https://github.com/bcbio/bcbio-nextgen
4. Svensson V*, et al.* (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 14(4):381-387.
5. Gramates LS*, et al.* (2017) FlyBase at 25: looking to the future. *Nucleic Acids Res* 45(D1):D663-D671.
6. Srivastava A, Sarkar H, Gupta N, & Patro R (2016) RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* 32(12):i192-i200.
7. https://github.com/hbc/bcbioSingleCell
8. Butler A, Hoffman P, Smibert P, Papalexi E, & Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36(5):411-420.
9. Stuart T*, et al.* (2019) Comprehensive Integration of Single-Cell Data. *Cell* 177(7):1888-1902 e1821.
10. Street K*, et al.* (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19(1):477.
11. Hu Y, Comjean A, Perkins LA, Perrimon N, & Mohr SE (2015) GLAD: an Online Database of Gene List Annotation for Drosophila. *J Genomics* 3:75-81.
12. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27-30.
13. Kanehisa M, Sato Y, Kawashima M, Furumichi M, & Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44(D1):D457-462.
14. Kanehisa M, Furumichi M, Tanabe M, Sato Y, & Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45(D1):D353-D361.
15. The Gene Ontology C (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 45(D1):D331-D338.
16. Ashburner M*, et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25-29.
17. Haber AL*, et al.* (2017) A single-cell survey of the small intestinal epithelium. *Nature* 551(7680):333-339.
18. Hu Y*, et al.* (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* 12:357.
19. Kudron MM*, et al.* (2018) The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. *Genetics* 208(3):937-949.

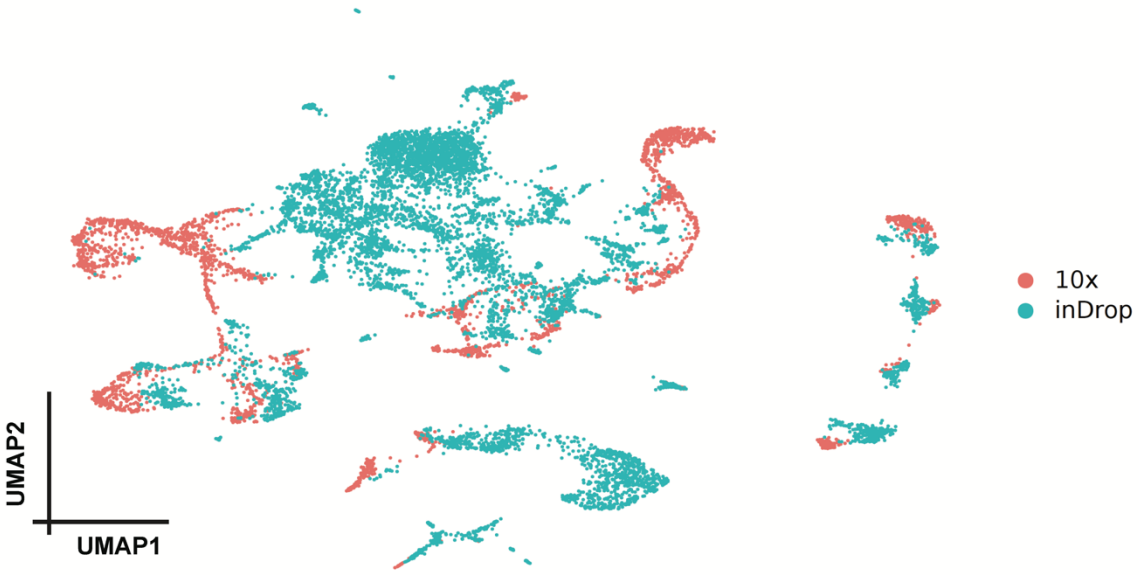

Fig. S1. Batch effects observed between technologies, visualized by UMAP plots. A. UMAP plot showing that cells cluster by cell type and technology.

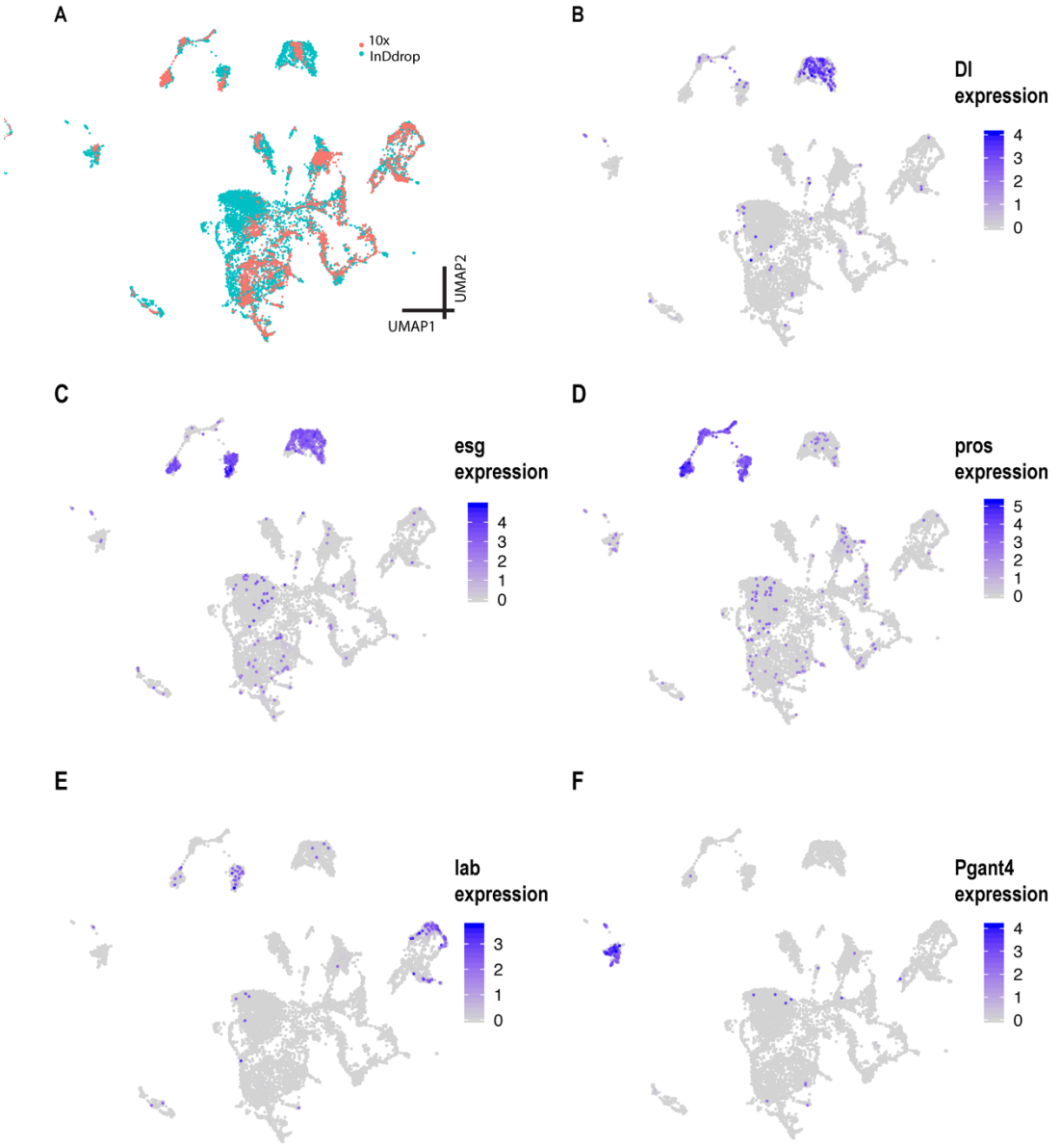

Fig. S2. Expression levels of each marker across different clusters, visualized by UMAP plots. A. The integrated dataset from inDrop and 10x Genomics after alignment. Cells in clusters that expressed *Dl* (B), *esg* (C), *pros* (D), *lab* (E) or *Pgant4* (F) are shown.

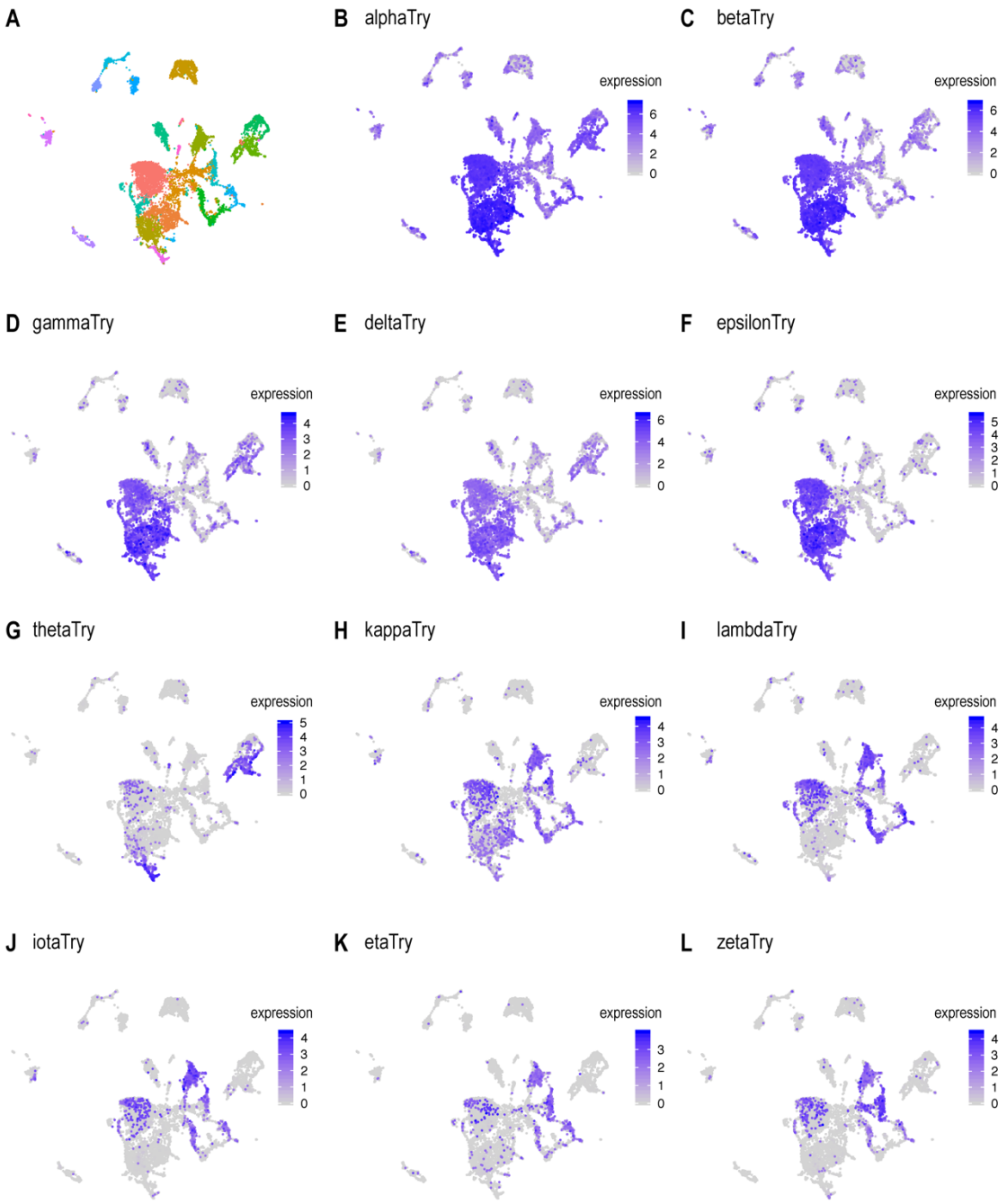

Fig. S3. The expression level of different Trypsins across different clusters, visualized by UMAP plots. (A) UMAP of the integrated dataset. The expression level of *alphaTry* (B), *betaTry* (C), *gammaTry* (D), *deltaTry* (E), *epsionTry* (F), *thetaTry* (G), *kappaTry* (H), *lambdaTry* (I), *IotaTry* (J), *etaTry* (K) or *ZetaTry* (L) in different clusters are shown.

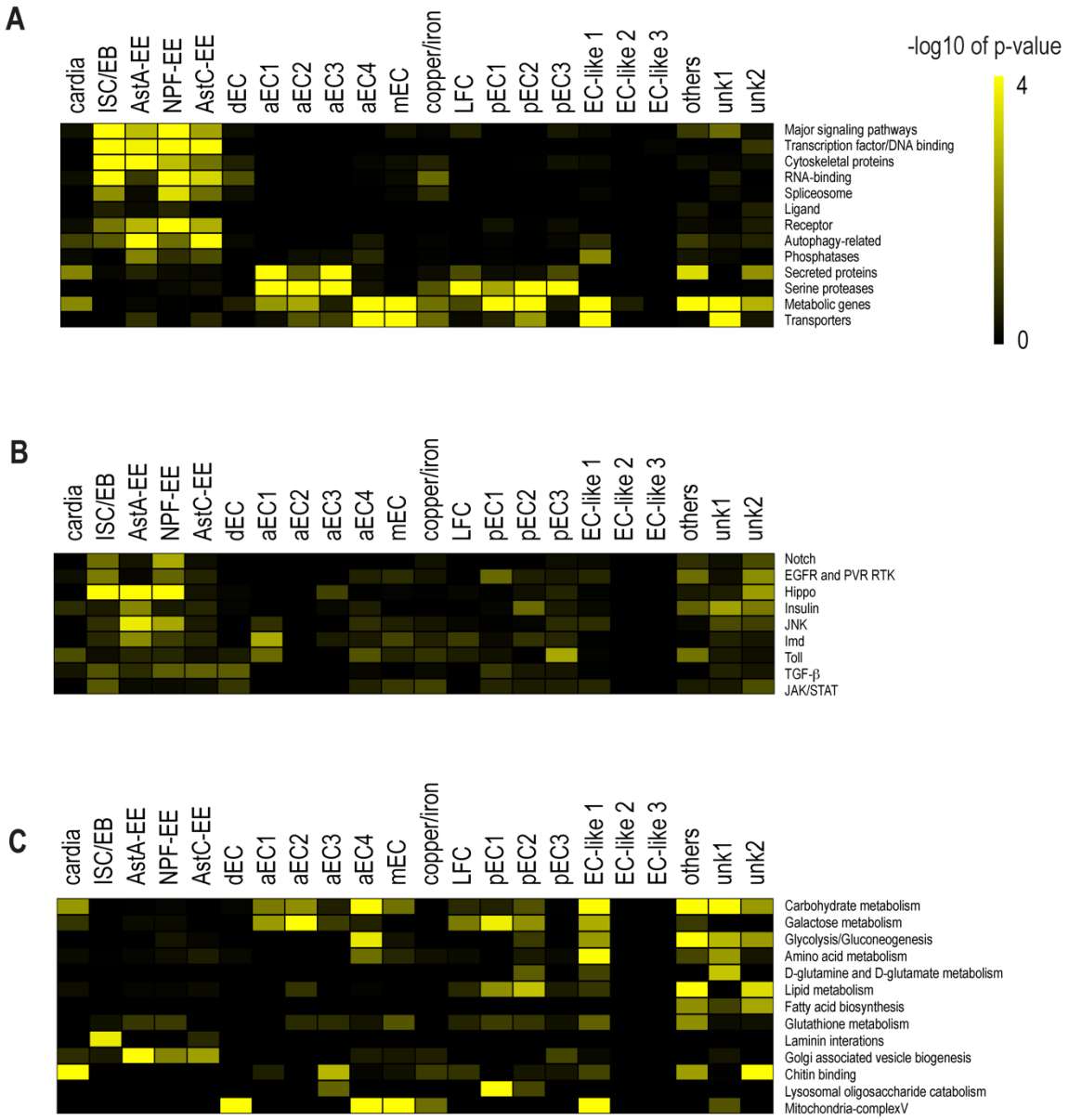

Figure S4. Gene set enrichment analysis. Markers identified by Seurat in different clusters categorized using GLAD gene function groups (A), transcriptional target genes of major signaling pathways (B), and metabolic pathways and others (C).

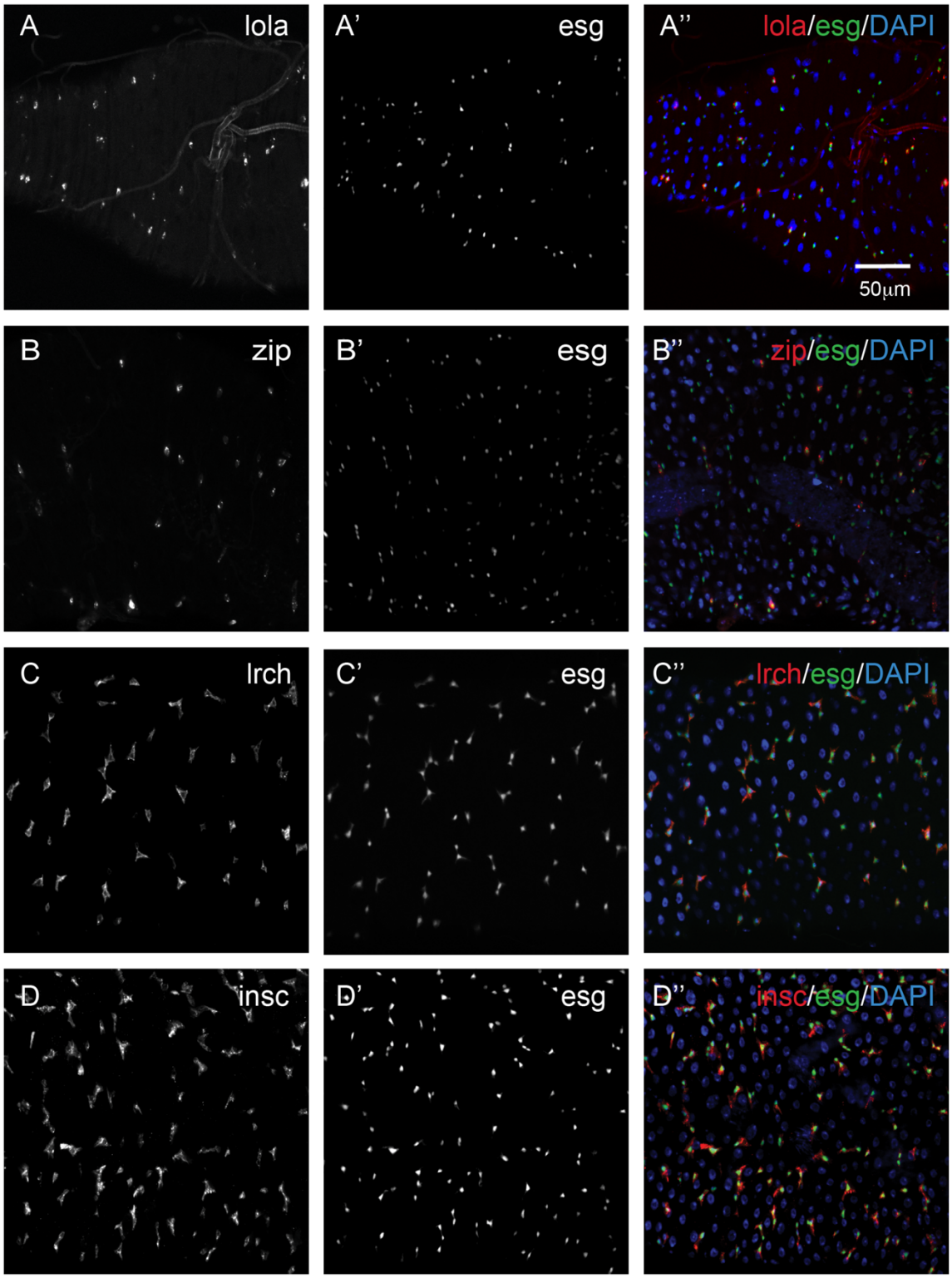

Fig. S5. Validation of new markers expressed in *esg*+ cells. A-A'', Genotype: *esg-sfGFP, UAS-mcherry-CAAX; lola-Gal4*. B-B'', Genotype: *esg-sfGFP, UAS-mcherry-CAAX; zip-Gal4*. C-C'', Genotype: *esg-sfGFP, UAS-mcherry-CAAX, lrch-Gal4*. D-D'', Genotype: *esg-sfGFP, UAS-mcherry-CAAX, insc-Gal4*. Note mcherry –CAAX is membrane bound.

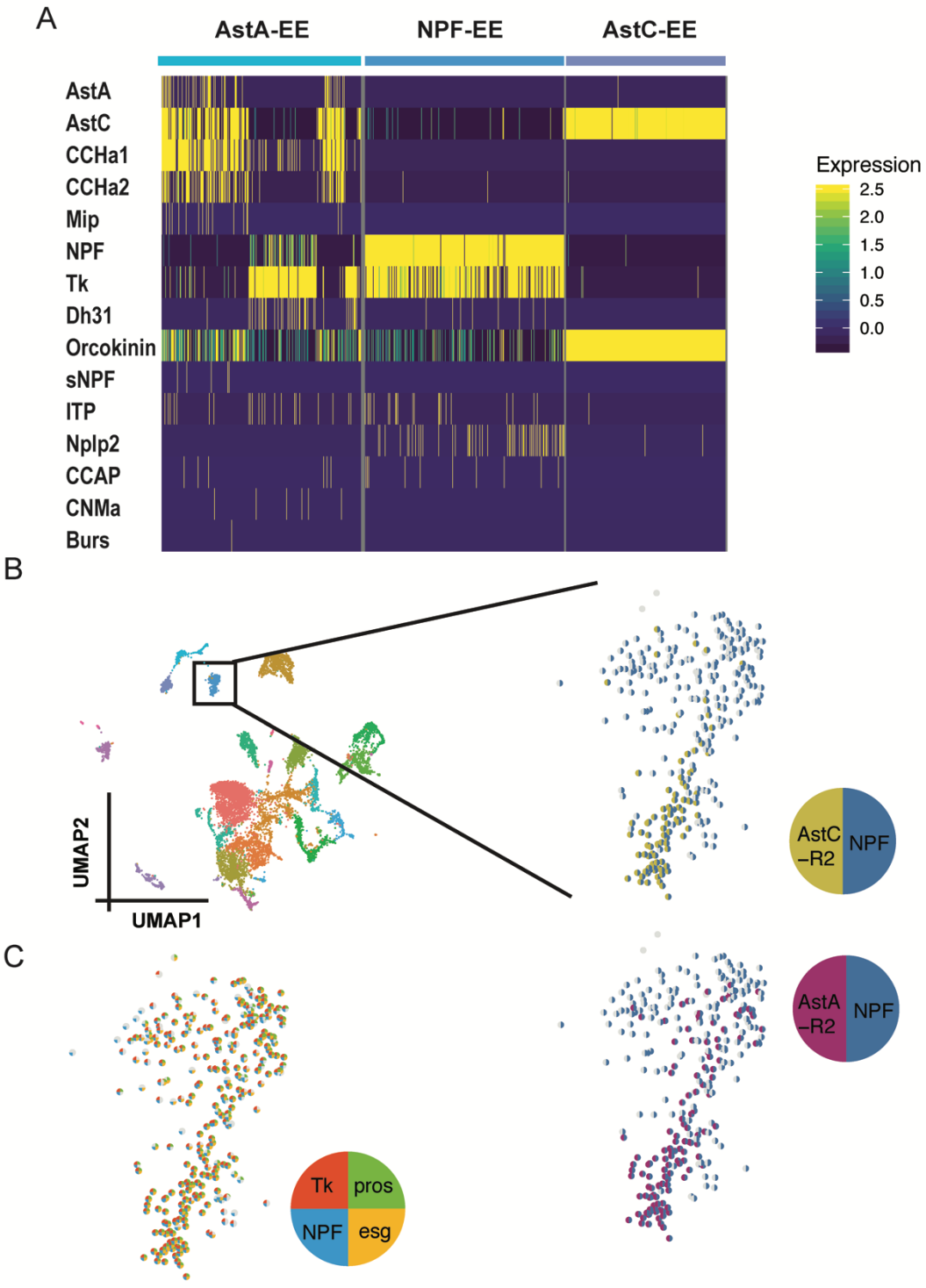

Fig. S6. The diversity of EEs as shown by gut hormone expression. (A) Heatmap showing the expression of 10 known gut hormones and 5 new gut hormones. *AstA, AstC, CCHa1* and *CCHa2* tend to be co-expressed in AstA-EE. NPF, Tk and Orcokinin tend to co-express in NPF-EE. Finally, AstC and Orcokinin tend to be co-expressed in AstC-EE. (B) The UMAP on the right side is a magnified view of the black box on the left. Co-expression of AstC-R2 and NPF (up); AstA-R2 and NPF (bottom) in each individual cell are shown as a pie graph. (C) The *esg+ pros+* cells are also expressed NPF and Tk. Co-expression of four genes are shown as a pie graph.

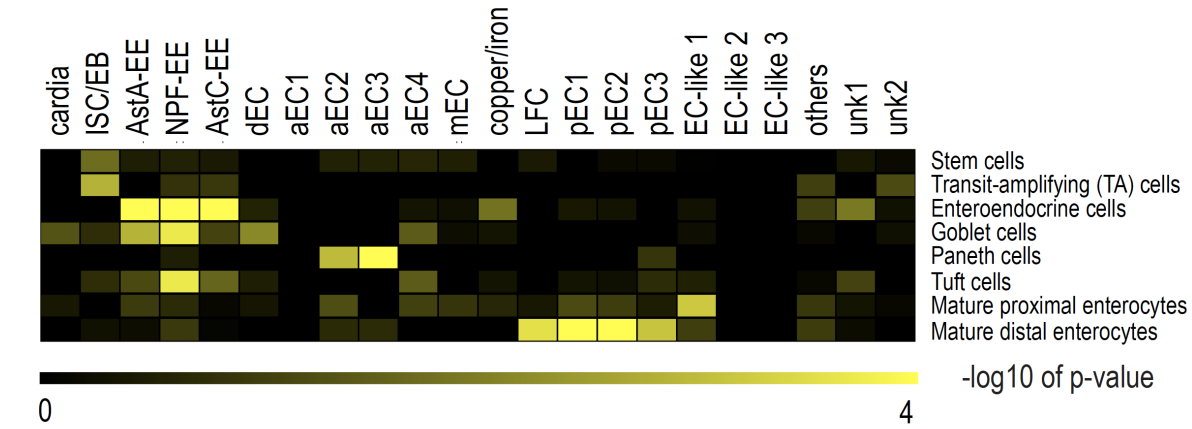

Fig. S7. Comparison of cell type conservation in intestines of flies and mammals. Comparison of fly gut cell types and mammalian intestine cell types using marker genes identified.

Supplemental Datasets

Dataset S1. scRNA-seq experiment statistics. The number of cells for each experiment is listed. The median number of genes per cell, median nUMIs (number of unique molecular identifiers) per cell, total number of genes detected, and the total number of UMIs detected are shown.

Dataset S2. Number of cells detected in each cluster and markers used for cluster identification.

Dataset S3. Information related to gene set enrichment analysis (Figure S4) including gene lists, enrichment p-values and annotation source. Full GLAD gene list is in the second tab.

Dataset S4. Regionalized expression of transcription factors in ECs identified in this study. Expression regions, transcription factors subgroup, mammalian orthologs, mouse gut tube location and flygut-seq expression region are listed.

Dataset S5. Top 32 transcription factors in the ISC/EB cluster identified in this study, including known literature (if available) in intestinal or other stem cells. pct.1 represents the percentage of cells in progenitors expressing this particular gene. pct.2 represents the percentage of cells in other clusters expressing this particular gene. The statistical test applied was ROC analysis (receiver operating curve). AUC (area under the ROC curve) evaluates if the particular gene alone can be used to classify between two clusters of cells. An AUC value of 1 means that expression values for this particular gene alone can perfectly classify the two clusters. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Power is a "predictive power" ranked matrix of putative differentially expressed genes. Genes identified in this study that have not been shown to be expressed in any other stem cells system are highlighted with yellow, and genes identified in this study that have not been shown to be expressed in intestinal stem cells but are expressed in other stem cell systems are highlighted with grey.

Dataset S6. Differentially expressed genes in ISCs and EBs. The p-value, average differential expression, pct.1, pct.2 and adjusted p-value are listed. pct.1 represents the percentage of cells in this cluster expressing this particular gene. pct.2 represents the percentage of cells in other clusters expressing this particular gene. The statistical test applied was Wilcox.

Dataset S7. Differentially expressed genes in each cluster. Only positive marker genes are shown. The statistical test applied was ROC analysis. 22 cluster information is shown on the tabs.

Dataset S8. Predicted transcription factors that could be regulating co-expression of gut hormones. Co-expression of four, three, or two different combinations of gut hormones was analyzed. The frequency to see co-expression of four, three, or two different combinations was calculated. Common transcription factors are shown for each combination of co-expression genes.

Dataset S9. Information related to the comparison of cell types in the *Drosophila* midgut and mammalian intestine (Figure S7). The cell type, source, overlapping genes and enrichment p-values are listed.