Defining cell types and lineage in the *Drosophila* midgut using single cell transcriptomics
Ruei-Jiun Hung¹,³, Joshua Shing Shun Li¹,³, Yifang Liu¹ and Norbert Perrimon¹,²

The *Drosophila* midgut has emerged in recent years as a model system to study stem cell renewal and differentiation and tissue homeostasis. Histological, genetic and gene expression studies have provided a wealth of information on gut cell types, regionalization, genes and pathways involved in cell proliferation and differentiation, stem cell renewal, and responses to changes in environmental factors such as the microbiota and nutrients. Here, we review the contribution of single cell transcriptomic methods to our understanding of gut cell type diversity, lineage and behavior.

**Addresses**

¹ Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, United States
² Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, United States

Corresponding author:
Perrimon, Norbert (perrimon@genetics.med.harvard.edu)
³ Equal contribution.

Two studies monitored the transcriptome of specific gut subregions, one via Affymetrix microarray and the other via bulk-RNAseq analysis [9,10]. Each subregion was defined by the enriched expression of 50–150 genes. Transcripts encoding enzymes involved in the processing of complex macromolecules are enriched in the anterior part of the gut. In contrast, those involved in the processing of simpler nutrients are more abundant in the posterior midgut, possibly reflecting the sequential breakdown and absorption of food along the gut. The region-specific gene profiles imply distinct roles for the gut subregions. Interestingly, many digestive genes are organized in large genomic clusters and are sequentially expressed in different gut regions. For instance, the trypsin genes *alpha-Try*, *beta-Try*, and *gamma-Try* are highly expressed in the anterior midgut, *theta-Try* is highly expressed in the middle midgut, and *kappa-Try*, *lambda-Try*, and *iota-Try* are enriched in the posterior midgut. ISCs also differ between subregions in their division rate, their cell shape, their numbers relative to the number of EEs and ECs, and their abilities to produce tumors [10]. Interestingly, clonal analyses revealed that the subregions of the midguts resemble cellular compartments [10], reminiscent to compartmentalization observed in imaginal tissues [12]. Most ISCs generate progeny that do not cross the subregion borders, suggesting that differences in ISCs may contribute to midgut regionalization.
To further understand the contribution of specific cell types to gut function and regional differences, Dutta et al. generated cell type specific transcriptomes using fluorescence-activated cell sorting (FACS) of each cell type within five gut regions (R1–R5) (Figure 1). Specific cell types were isolated using cell-type specific Gal4-based GFP expression: ISC’s (Dl-gal4), EBs (Su(H)-GBE-gal4), ECs (Myo1A-gal4), EEs (pros-gal4), and visceral muscle (VM) (How-gal4). Interestingly, cell types displayed transcriptional variation along the length of the gut. In particular, R1 and R3 ISC’s are quite distinct from R2, R4, and R5 ISC’s. For example, ISC’s in the acidic R3 region express vacuolar H+ ATPases, indicating an adaptation of ISC’s to their local environment. EBs and ECs also showed clear regional specificities, including altered expression of genes involved in metabolism and digestion along the gut. Among these cell types, EEs showed the most differences between regions due to the diversity of the gut hormones they produce. Interestingly, several morphogens (Wnt1, Wnt4, Wnt6 and WntD) are expressed in VM cells in a gradient dependent manner, which might provide regional specific niches for ISC’s. In addition, an unexpected role of EEs and ECs in producing different combination of AMP secretion upon bacterial infection was observed. Finally, genes that were preferentially expressed in a cell type specific manner were identified and their functions were validated by RNAi. For example, sna was shown to regulate stem cell differentiation in the R4–R5 region.

Altogether, the transcriptomes captured by these previous studies have uncovered fundamental differences between the cell-types within and between regions of the fly midgut. However, both RNAseq of specific regions [9,10] and RNaseq of FACS cell types have limitations. RNAseq of specific regions requires visual discrimination of regional boundaries within the midgut, which can be prone to inconsistencies between individual guts. Additionally, bulk sequencing does not distinguish between cell types within a dissected region, and differences between quiescent and transient states of ISC’s undergoing differentiation are likely missed. With regards to RNAseq of FACS cell types, different expression in GFP levels associated with Gal4-based GFP expression can affect sorting and cell recovery. Further, with FACS, the highly heterogeneous EE subtypes are reduced dimensionally into a single population.

To address some of the limitations of the previous studies, single-cell RNA sequencing (scRNA-seq) was applied on whole guts, as it provides an unbiased approach to survey cell type diversity and function, and to define relationships among cell types [13**]. A total of 22 clusters were identified, including one cluster of ISC/EBs, three clusters of EE’s, 14 clusters of ECs, one cluster of cardia, and three unknown clusters. Among the 14 EC clusters, four clusters (aEC1-4) map to the anterior midgut, three clusters (pEC1-3) to the posterior midgut, one cluster of mEC to the middle midgut, one cluster to copper and iron cells, one cluster to large flat cells (LFC), one cluster to differentiating EC (dEC) and three EC-like clusters (Figure 2). Genes categorized as encoding members of signaling pathways, transcription factors, cytoskeletal proteins, and RNA-binding proteins...
are enriched in ISC/EB progenitor cells. Genes encoding proteins involved in vesicle docking, fusion and secretion are enriched in EEs, reflecting the molecular machinery for hormone secretion. Genes that encode proteins involved in metabolic processes, serine proteases, and transporters are enriched in ECs. Different EC clusters are enriched for different types of metabolic processes. For example, aECs and cardia are enriched in genes involved in carbohydrate metabolism, galactose metabolism, and binding to chitin, a component of the peritrophic matrix secreted by cardia cells. Genes involved in mannosidase activity, galactose metabolism and lipid metabolism are enriched in pECs. Interestingly, ribosomal proteins (Rpl and Rps) are enriched in the ISC/EB and dEC clusters, and transmission electron micrographs of the midgut confirmed that ISC cells exhibit more ribosomes than mature ECs. These observations are consistent with several studies indicating that embryonic stem cells maintain a high level of free inactive ribosomes, which might be essential for rapidly rewiring and priming gene expression toward differentiation [14]. In addition, a study of *Drosophila* germline stem cells demonstrated that the transition from self-renewal to differentiation relies on increased ribosome biogenesis and protein synthesis [15], suggesting that high levels of ribosomes is one characteristic of stem cells.

Analysis of scRNA-seq data also identified previously known and new markers, such as *klu*, *lola* and *Eip75B* [13**]. In particular, *klu* is expressed in cells that are positive for *Sut(HE)GBE-LacZ*, confirming that *klu* is a novel EB marker [13**]. Knocking down *klu* in progenitors resulted in an increase in EEs, which is consistent with an independent study showing that *klu* maintains the EB commitment to EC lineage and suppresses EE fate via indirectly repressing the levels of *scute* (sc) [13**,16**]. Further, *lola* expression overlaps with *esg*, an ISC/EBs marker, and was later found to function in ISC/EBs to suppress ISC proliferation [17]. Finally, *Eip75B* is expressed preferentially in ISC/EBs. Two subsequently studies have shown that Eip75B acts downstream of the edysone receptor (Ecr) to drive ISC daughter cells toward absorptive EC lineage to ensure epithelial growth induced by the steroid hormone edysone [18**,19**].

---

**Figure 2**

Summary of the spatial distribution of intestinal cell clusters defined by Hung et al. (2020). (a) The regional distribution of 22 cell clusters generated by Hung et al. (2020) was mapped using the region-specific gene enrichment (RSGE) algorithm developed by Guo et al. (2019). (b) The bar graph depicts the regional enrichment score for the top 100 region-specific genes in R1 to R5 (derived from Dutta et al. (2015)).
In addition to scRNA-seq of the whole midgut, Guo et al. (2019) conducted scRNA-seq of EEs by isolating EEs via FACS [20**]. They identified 10 major EE subtypes in the midgut and comprehensively characterized the repertoires in these subtypes of the gut hormones (Tk, Dh31, NPF, Nplp2, Gphb5, CCAP, ITP, sNPF, CCHa2, CCHa1, AstA, Orcokinin and AstC). Single-cell analysis revealed that most individual EEs produce approximately 2–5 gut hormones [20**], consistent with [13**]). To retrieve the spatial distribution of all EE subtypes, which is lost during tissue dissociation, Guo generated a region-specific gene enrichment (RSGE) algorithm based on regional bulk EE RNA-seq from previously published data [21] (Figure 3). Interestingly, the expression of gut hormones and their corresponding receptors in EE subtypes revealed both local (paracrine) and long-range communication between gut regions [13**,20**]. For example, AstC-EEs also express NPF receptors, suggesting that NPF from neighboring cells may regulate AstC-EE in a paracrine manner. In addition, the gut hormone CCHa1 is expressed in the posterior midgut while its receptor CCHa1-R is expressed in the anterior and middle midgut, suggesting potential long-range communication. Alternatively and noteworthy, as the gut folds into a stereotypical 3D structure inside the abdomen, gut hormone producing EEs and gut cells expressing their receptors may in some cases be very close to each other.

Transient expression of sc in ISCs specifies pros-expressing EEPs, which divide one more time before terminal differentiation to generate two EEs (pros+) that express AstC (class I EE) or Tk (class II EE) [22,23]. This binary cell fate is specified by Notch activity [22,24,25]. To identify the transcription factors (TFs) that govern EE subtype specification, Guo identified a set of 14 TFs whose binary expression states (ON or OFF) alone are sufficient to classify the subtypes for all EE clusters (with the exception of 2 clusters that share very similar expression signatures). Depleting individual TFs resulted in decreased expression of at least one or more gut hormones. For example, knocking down Ptx1 in EEs decreased AstC expression and loss of AstC expressing EEs. This approach identified two TFs, Ptx1 and Mirror (Mrr), involved in the specification of class I and class II EEs, respectively. In addition, this study identified several regionally expressed TFs that define spatial identity of EEs (e.g. esg in the middle midgut).

Although these findings suggest that EE subtypes can be defined by a combinatorial expression of TFs, expression of some gut hormones did not strictly correlate with EE subtype clusters. One explanation is that extrinsic factors, such as infection or mechanical stress, could influence the expression of gut hormones. Alternatively, EEs could be plastic, as shown for mammalian EEs, which switch their expression as they differentiate [26]. Importantly, the directed approach of isolating cells of interest for scRNA-seq led to a better resolution of EE subtypes than whole gut scRNA-seq had provided. Thus, a more directed approach of isolating cells for scRNA-seq by
using FACS (either based on reporter expression or in situ probes [27]), is likely to provide further characterization of cell types.

An important feature of scRNA-seq is to provide information on lineage relationships between cell types, which have classically relied on genetic recombination and fluorescent reporters to label the starting cells and their progeny. Computational methods, such as Monocle [28], slingshot [29], and STREAM [30], can be applied to scRNA-seq data to construct developmental trajectories and extract transient states from pseudo-temporal predictions. For example, Hung et al. (2020) used slingshot to map cell differentiation trajectories of early states (ISCs) to terminal states (EEs and ECs). Three trajectories of ISCs were identified: Firstly, ISC/EBs → mECs → dECs (differentiating ECs) → aECs; secondly, ISC/EBs → mECs → dECs → pECs; and finally, ISC/EBs → EE. Slingshot also helped to classify dEC as an intermediate state, as dECs did not resemble either ISCs or mature ECs. In the future, the use of DNA sequence barcodes (instead of fluorescent reporters) to encode clonal information should help capture temporal information in the context of lineage tracing. In particular, CRISPR-Cas9 has been used to develop several lineage tracing strategies, including GESTALT [31], LINNAEUS [32], MEMOIR [33], and ScarTrace [34]. In general, these methods use Cas9-induced stochastic mutations to create insertions or deletions (during DNA repair following double-strand breaks) at predefined target sites. Since the alterations occur in the DNA, these methods rely on inferring cell lineage from changes observed over multiple cell divisions. While this works well for constructing trajectories for cell-types that arise from distant progenitors, it falls short in capturing the transient-states over the lifetime of a cell. One possible solution is to use Cas-related technologies to create alterations in predefined RNA reporters. These edited sequences could then be used as barcodes for determining the transient state of single cells and be readily combined with other sequencing-based omics measurements to construct the trajectory of stem cell differentiation.

Surprisingly, while scRNA-seq was able to characterize extensive regional differences in ECs and EEs, it was not able to distinguish regional differences among ISC/EBs clusters. Possibly, the preferential expression of cell type specific genes in ISCs/EBs is more profound than that of regional specific genes. Another possibility is that regional difference genes in ISCs are expressed at low levels, which are not easily detected by scRNA-seq using the commonly used 10× Genomics platform. In this case, use of a more in-depth sequencing approach such as smart-seq [35] to sequence the full-length transcripts rather than only 3’ ends (10× Genomics), and a more targeted approach, such as use of GFP-sorted cells or Probe-seq to enrich for a specific cell type, could be used in future studies.

Future scRNA-seq studies should enable the discovery of the gene regulatory networks that underlie ISC proliferation and differentiation. This would involve mapping TFs that regulate each other and their downstream targets. A few studies have attempted to address this by inferring co-expression networks from scRNA-seq data [36,37]. More recently, SCENIC (Single-Cell eRegulatory Network Inference and Clustering) integrated co-expression analyses with cis-regulatory motif analyses to identify TF-to-target modules (regulons) enriched in single cell clusters [38]. What remains unavailable are computational methods that can map regulon expression onto lineage trajectory predictions. Putative TF-cascade predictions derived from such approaches will be crucial to our understanding of gene networks that underly ISC differentiation.

Finally, although scRNA-seq has the advantage of providing substantial transcriptomics information of individual cells, the spatial information of individual cells in the tissue is missing. Thus, region-specific markers identified by scRNA-seq need to be validated to retrieve spatial information. The wealth of existing data from large-scale in situ RNA hybridization studies and gene expression reporters available for Drosophila will considerably facilitate such studies.

Conflict of interest statement
Nothing declared.

Acknowledgments
We thank Stephanie Mohr for helpful comments on the manuscript. JSSL is supported by a Croucher fellowship for Postdoctoral Research from the Croucher Foundation and NP is an investigator of the Howard Hughes Medical Institute.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- - of outstanding interest


