

Deterministic Genetic Barcoding for Multiplexed Behavioral and Single-Cell Transcriptomic Studies

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted

August 1, 2023 (this version)

Sent for peer review

May 20, 2023

Posted to bioRxiv

April 1, 2023

Jorge Blanco Mendana, Margaret Donovan, Lindsey Gengelbach, Benjamin Auch, John Garbe, Daryl M. Gohl 

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA • Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455, USA

 https://en.wikipedia.org/wiki/Open_access

 <https://creativecommons.org/licenses/by/4.0/>

Abstract

Advances in single-cell sequencing technologies have provided novel insights into the dynamics of gene expression throughout development, been used to characterize somatic variation and heterogeneity within tissues, and are currently enabling the construction of transcriptomic cell atlases. However, despite these remarkable advances, linking anatomical information to transcriptomic data and positively identifying the cell types that correspond to gene expression clusters in single-cell sequencing data sets remains a challenge. We describe a straightforward genetic barcoding approach that takes advantage of the powerful genetic tools available in *Drosophila* to allow *in vivo* tagging of defined cell populations. This method, called Targeted Genetically-Encoded Multiplexing (TaG-EM), involves inserting a DNA barcode just upstream of the polyadenylation site in a Gal4-inducible *UAS-GFP* construct so that the barcode sequence can be read out during single-cell sequencing, labeling a cell population of interest. By creating many such independently barcoded fly strains, TaG-EM will enable a number of potential applications that will improve the quality and information content of single-cell transcriptomic data including positive identification of cell types in cell atlas projects, identification of multiplet droplets, and barcoding of experimental timepoints, conditions, and replicates. Furthermore, we demonstrate that the barcodes from TaG-EM fly lines can be read out using next-generation sequencing to facilitate population-scale behavioral measurements. Thus, TaG-EM has the potential to enable large-scale behavioral screens in addition to improving the ability to reliably annotate cell atlas data, expanding the scope, and improving the robustness of single-cell transcriptomic experiments.

eLife assessment

This **useful** study presents a genetically encoded barcoding system that could not only advance transcriptomic studies but that also has potential further applications, such as in high-throughput population-scale behavioral measurements. The evidence supporting the claims of the authors are currently **inadequate** to demonstrate that the method is indeed greatly superior to existing approaches in behavioural and transcriptomic studies.

Introduction

Spatially and temporally regulated gene expression patterns are a hallmark of multicellular life and function to orchestrate patterning, growth, and differentiation throughout development (Ingham, 1988 [↗](#); Reeves et al., 2006 [↗](#)). In mature organisms, spatial expression patterns both in tissues and within cells define functionally distinct compartments and determine many aspects of cellular and organismal physiology (Martin and Ephrussi, 2009 [↗](#)). In addition, such expression patterns differentiate healthy and diseased tissue and impact disease etiology (Marusyk et al., 2012 [↗](#)). Spatial and temporal expression patterns, which can be used to distinguish between cell types and provide insight into cellular function, also provide a means to understand the organization and physiology of complex tissues such as the brain (Thompson et al., 2014 [↗](#)). Thus, robust and scalable tools for measuring spatial and temporal gene expression patterns at a genome-wide scale and at high resolution would be transformative research tools across many biological disciplines.

Single-cell sequencing technologies have provided insights into the dynamics of gene expression throughout development, been used to characterize somatic variation and heterogeneity within tissues, and are currently enabling the construction of transcriptomic cell atlases (Klein et al., 2015 [↗](#); Macosko et al., 2015 [↗](#); Zheng et al., 2017 [↗](#)). However, linking anatomical information to transcriptomic data and positively identifying the cell types that correspond to gene expression clusters in single-cell sequencing data sets remains a challenge. The cellular identities of gene expression clusters identified in cell atlas data sets are typically inferred from the expression of distinctive gene sets (Hung et al., 2020 [↗](#); Li et al., 2022 [↗](#); Ma et al., 2021 [↗](#)), and the lack of positive identification of gene expression clusters introduces an element of uncertainty in this analysis. Moreover, this process of manual annotation is labor-intensive and often requires additional experiments to determine or confirm the expression patterns of marker genes. Emerging spatial genomics technologies hold promise in linking anatomical and transcriptomic information (Lee et al., 2015 [↗](#); Lein et al., 2017 [↗](#)). Several of the emerging commercial spatial genomics technologies rely on *in situ* sequencing of marker genes allowing droplet-based single-cell transcriptomic data to be mapped onto a tissue. However, these technologies currently suffer from constraints related to cost, content, or applicability to specific model systems.

In addition to descriptive cell atlas projects, studies involving multiple experimental timepoints throughout development and aging, or to assess the effects of experimental exposures or genetic manipulations would benefit from increased ability to multiplex samples. Given the fixed costs of droplet-based single-cell sequencing, generating data for single-cell transcriptomic time courses or experimental manipulations can be costly. These costs are also a barrier to including replicates to assess biological variability; consequently, a lack of replicates is a common shortcoming of single-cell sequencing experiments. Antibody-based cell hashing or feature barcoding approaches have been developed to allow multiplexing of samples in droplet-based single-cell sequencing reactions (Stoeckius et al., 2018 [↗](#), 2017 [↗](#)). In addition, other multiplexing strategies for single-cell sequencing based on alternative methods for tagging cells (Cheng et al., 2021 [↗](#)) or making use of natural genetic variation have been used (Kurmangaliyev et al., 2020 [↗](#)). While such approaches can reduce per-sample costs, typically samples are barcoded at a population level and thus do not enable labeling of cell subpopulations within a sample.

We have developed a straightforward genetic barcoding approach that takes advantage of the powerful genetic tools available in *Drosophila* to allow deterministic *in vivo* tagging of defined cell populations. This method, called Targeted Genetically-Encoded Multiplexing (TaG-EM), involves inserting a DNA barcode just upstream of the poly-adenylation site in a Gal4-inducible *UAS-GFP* construct so that the barcode sequence can be read out during droplet-based single-cell sequencing, labeling a cell population of interest.

Genetic barcoding approaches have been employed in many unicellular systems, cell culture, and viral transfection, to facilitate high throughput screening using sequencing-based readouts (Bhang et al., 2015 [↗](#); Smith et al., 2009 [↗](#); van Opijnen et al., 2009 [↗](#)). In multicellular animals, techniques such as GESTALT have enabled lineage tracing by using CRISPR to create unique barcodes in differentiating tissue (McKenna et al., 2016 [↗](#)), and barcode sequencing has also been employed to map connectivity in the brain (Chen et al., 2019 [↗](#)). In addition to improving single-cell sequencing-based measurements, genetically barcoded fly lines can enable highly multiplexed behavioral assays read out using high throughput sequencing. Flies carrying TaG-EM barcodes can be exposed to different experimental perturbations and then run through assays where flies, larvae, or embryos are fractionated based on behavioral outcomes or other phenotypes. Thus, TaG-EM has the potential to enable large-scale next-generation sequencing (NGS)-based behavioral or other fractionation screens analogous to BAR-Seq or Tn-Seq approaches employed in microbial organisms.

Results

TaG-EM: A novel genetic barcoding strategy for multiplexed behavioral and single-cell transcriptomics

We cloned a fragment containing a PCR handle sequence and a diverse 14 bp barcode sequence into the SV40 3' untranslated region (UTR) sequence just upstream of the polyadenylation sites in the 10xUAS-myr::GFP (pJFRC12, (Pfeiffer et al., 2010 [↗](#))) backbone (**Figure 1A** [↗](#)). A pool containing 29 unique barcode-containing plasmids was injected into *Drosophila* embryos for PhiC31-mediated integration into the attP2 landing site (Groth et al., 2004 [↗](#)) and transgenic lines were isolated and confirmed by Sanger sequencing (**Supplemental Figure 1** [↗](#)). We recovered 20 distinctly barcoded *Drosophila* lines, with some barcodes recovered from multiple crosses (**Supplemental Figure 1** [↗](#)). Such barcoded fly lines have the potential to enable population behavioral measurements, where different exposures, experimental timepoints, and genetic or neural perturbations can be multiplexed and analyzed by measuring barcode abundance in sequencing data (**Figure 1B** [↗](#)). In addition, the barcodes, which reside on a Gal4-inducible UAS-GFP construct, can be expressed tissue-specifically and read out during droplet-based single-cell sequencing, labeling a cell population and/or an experimental condition of interest (**Figure 1C** [↗](#)).

Testing the accuracy and reproducibility of TaG-EM behavioral measurements using structured pools

We conducted initial experiments to optimize amplification of the genetic barcodes using primers targeting the PCR handle inserted just upstream of the 14 bp barcode sequence and PCR primers downstream of the TaG-EM barcode in the SV40 3' UTR sequence (**Supplemental Figure 2** [↗](#)). To test the accuracy and reproducibility of sequencing-based measurements of TaG-EM barcodes, we constructed structured pools containing defined numbers of flies pooled either evenly with each of the 20 barcode constructs comprising 5% of the pool, or in a staggered manner with sets of barcodes differing in abundance in 2-fold increments (**Figure 2A** [↗](#)). To examine the impact of technical steps such as DNA extraction and PCR amplification on TaG-EM barcode measurements, even pools were made and extracted in triplicate and amplicon sequencing libraries were made in triplicate for each independently extracted DNA sample for both the even and staggered pools. The resulting data indicated that TaG-EM measurements are highly accurate and reproducible. Technical replicates (indicated by error bars in **Figure 2B-E** [↗](#)) showed minimal variability. Likewise, the three independently extracted replicates of the even pools produced consistent data with all 20 barcodes detected at levels close to the expected 5% abundance (**Figure 2B-C** [↗](#)). Barcode abundance values for the staggered structured pools was generally consistent with the input values and in most cases, the 2-fold differences between the different groups of barcodes

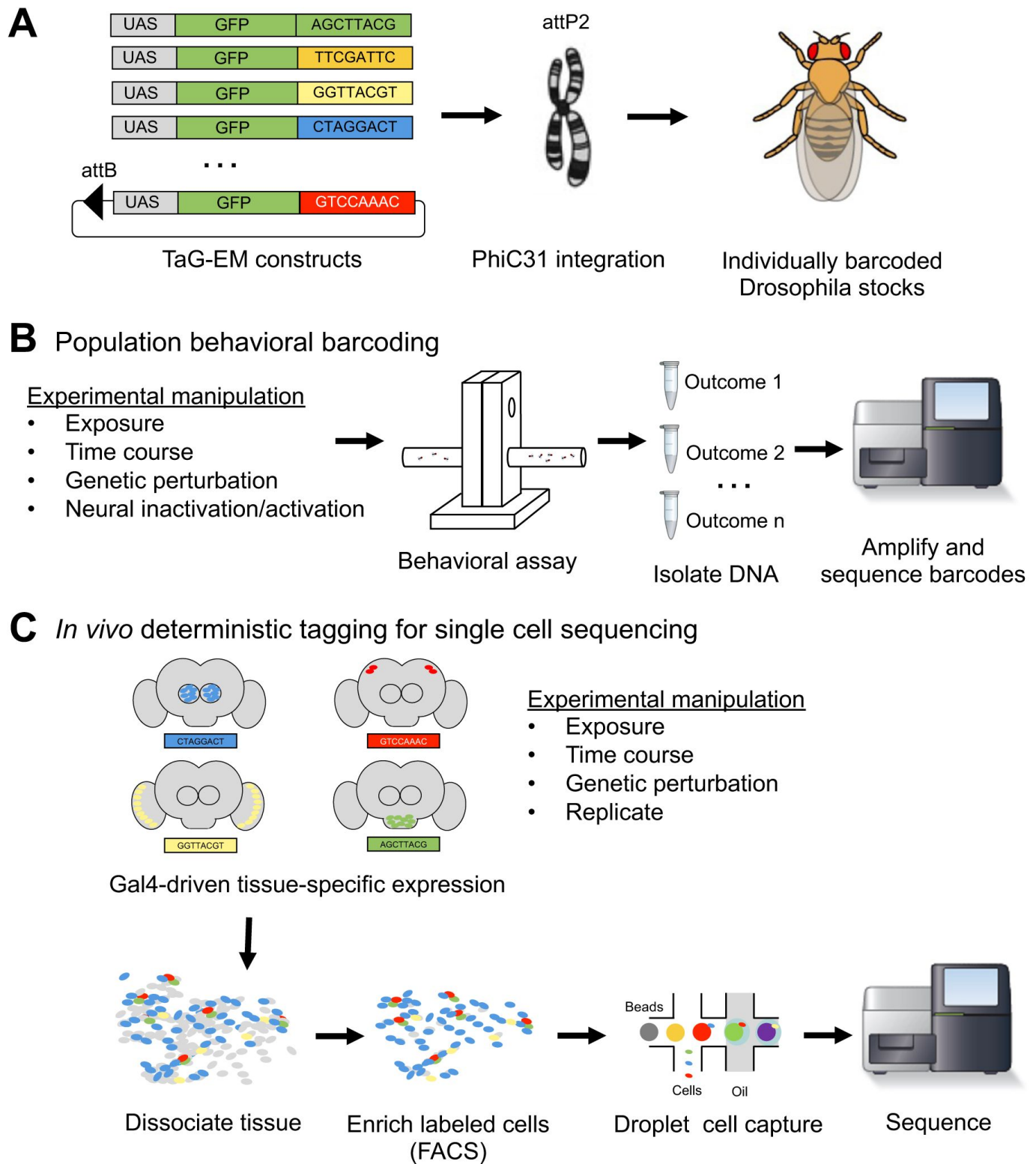


Figure 1.

Overview of TaG-EM system

A) Schematic illustrating the design of the TaG-EM constructs, where a barcode sequence is inserted in the 3' UTR of a UAS-GFP construct and inserted in a specific genomic locus using PhiC31 integrase. B) Use of TaG-EM barcodes for sequencing-based population behavioral assays. C) Use of TaG-EM barcodes expressed with tissue-specific Gal4 drivers to label cell populations *in vivo* upstream of cell isolation and single-cell sequencing.

could be distinguished (**Figure 2D-E** [↗](#)). For the staggered pools, abundances correlated well with the expected values, particularly when multiple barcodes for an input level were averaged, in which case R^2 values were >0.99 (**Figure 2D-E** [↗](#), inset plots). This indicates that a high level of quantitative accuracy can be attained using sequencing-based analysis of TaG-EM barcode abundance, particularly when averaging data for three to four independent barcodes for an experimental condition.

TaG-EM measurement of phototaxis behavior correlate well with video-based measurements

Next, we tested whether TaG-EM could be used to measure a phototaxis behavior. A mixture of barcoded wild type or blind *norpA* mutant flies were run together through a phototaxis assay. At the end of a period of light exposure, test tubes facing toward or away from the light were capped, DNA was isolated, and barcodes were amplified and sequenced for each tube. Raw read counts were scaled in proportion to the number of flies per tube and a preference index was calculated for each barcode (**Figure 3A** [↗](#)). In parallel, individual preference indices were calculated based on manual scoring of videos recorded for each line (**Figure 3B** [↗](#)). Preference indices calculated for the pooled, NGS-based TaG-EM measurements were nearly identical to conventional behavioral measurements for both wild type and *norpA* mutants (**Figure 3A-B** [↗](#)).

TaG-EM measurement of oviposition behavior and age-dependent fecundity

We next tested whether NGS-based pooled measurements of egg laying could be made. Fertilized females from each of the 20 barcode lines were placed together in egg laying cups, embryos were collected, aged for 12 hours to enable cell numbers to stabilize in the developing eggs, and then DNA was extracted from both the pooled adult flies and the embryos. In general, TaG-EM measurements of oviposition correlated with fly numbers, with the exception of barcode 14 which had reduced barcode abundance across multiple trials (**Supplemental Figure 3** [↗](#)). This suggests that despite the fact that the genetic barcode constructs are inserted in a common landing site, differences with respect to specific behaviors may exist among the lines and thus one should test to make sure given lines are appropriate to use in specific behavioral assays.

To determine whether TaG-EM could be used to measure age-dependent fecundity, we collected flies from twelve different TaG-EM barcode lines at four time points separated by one week (three barcode lines per timepoint). We collected eggs from these fly lines individually and scored the number of viable eggs per female. Next, we pooled the barcoded flies from all timepoints and collected eggs from the pooled flies. These eggs were aged, DNA was extracted, and the TaG-EM barcodes were amplified and sequenced. While measurements from individual barcode lines were noisy, both for manual counts and sequencing based measurements, there was a general trend toward declining fecundity over time (**Supplemental Figure 4** [↗](#)), consistent with published reports (David et al., 1975 [↗](#)). Manually scored viable egg numbers and TaG-EM barcode abundances were well correlated across two independent experimental trials (R^2 values of 0.52-0.61 for Trial 1 and 0.74-0.84 for Trial 2). When barcodes from each individual timepoint were averaged, R^2 values for the correlation between manual and sequencing-based measurements were 0.95 for Trial 1 and 0.99 for Trial 2 (**Figure 3C** [↗](#), **Supplemental Figure 4** [↗](#)).

Tissue-specific expression of TaG-EM GFP constructs

To facilitate representation of the TaG-EM barcodes in single-cell sequencing data, genetic barcodes were placed just upstream of the polyadenylation signal sequences and polyA cleavage sites (**Figure 4A** [↗](#)). To verify that the inserted sequences did not interfere with Gal4-driven GFP expression, we crossed each of the barcoded TaG-EM lines to *decapentaplegic-Gal4* (*dpp-Gal4*). We

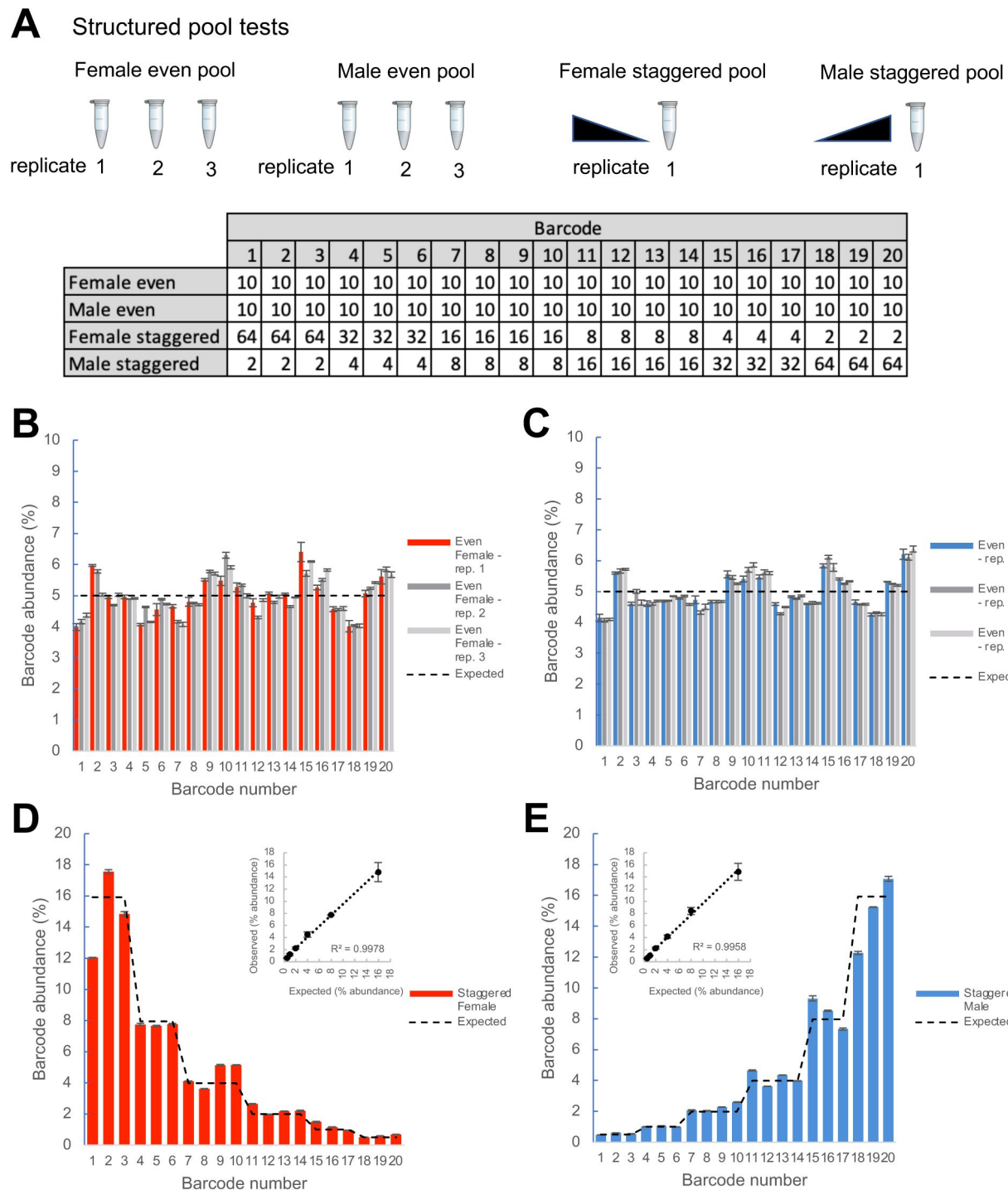


Figure 2.

Structured pool tests

A) Overview of the construction of the structured pools for assessing the quantitative accuracy of TaG-EM barcode measurements. Male and female even pools were constructed and extracted in triplicate. The table shows the number of flies that were pooled for each experimental condition. B) Barcode abundance data for three independent replicates of the female even pool. C) Barcode abundance data for three independent replicates of the male even pool. D) Barcode abundance data for the female staggered pool. Inset plot shows the average observed barcode abundance among lines pooled at each level compared to the expected abundance. E) Barcode abundance data for the male staggered pool. Inset plot shows the average observed barcode abundance among lines pooled at each level compared to the expected abundance. For all plots, bars indicate the mean barcode abundance for three technical replicates of each pool, error bars are $\pm$ S.E.M.

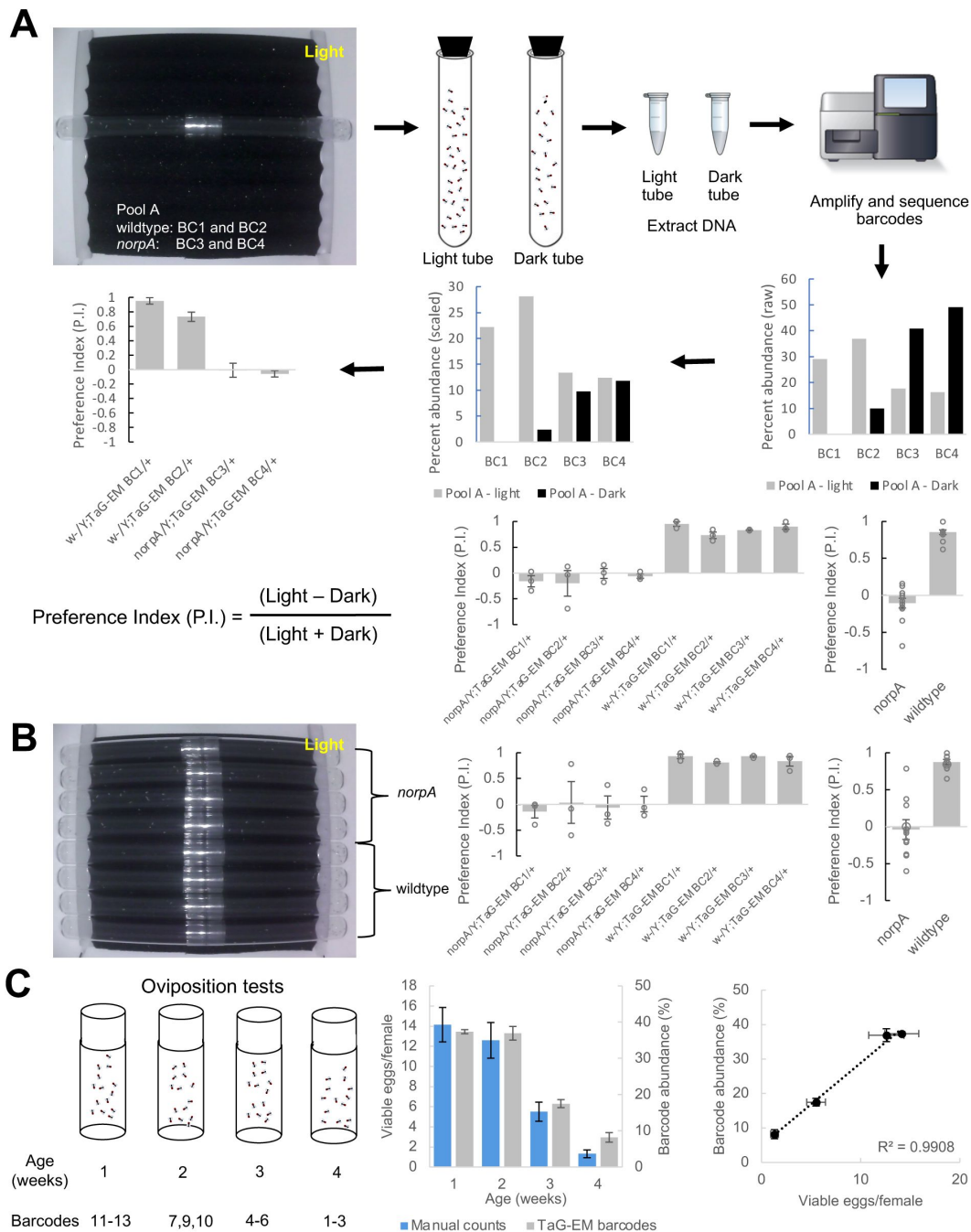


Figure 3.

TaG-EM barcode-based behavioral measurements

A) TaG-EM barcode lines in either a wild-type or *norpA* background were pooled and tested in a phototaxis assay. After 30 seconds of light exposure, flies in tubes facing the light or dark side of the chamber were collected, DNA was extracted, and TaG-EM barcodes were amplified and sequenced. Barcode abundance values were scaled to the number of flies in each tube and used to calculate a preference index (P.I.). Average P.I. values for four different TaG-EM barcode lines in both the wild-type and *norpA* backgrounds are shown ($n=3$ biological replicates, error bars are $\pm$ S.E.M.). B) The same eight lines used for the sequencing-based TaG-EM barcode measurements were independently tested in the phototaxis assay and manually scored videos were used to calculate a P.I. for each genotype. Average P.I. values for each line are shown ($n=3$ biological replicates, error bars are $\pm$ S.E.M.). C) Flies carrying different TaG-EM barcodes were collected and aged for one to four weeks and then eggs were collected and egg number and viability was manually scored for each line. In parallel the barcoded flies from each timepoint were pooled, and eggs were collected, aged, and DNA was extracted, followed by TaG-EM barcode amplification and sequencing. Average number of viable eggs per female (manual counts) and average barcode abundance are shown both as a bar plot and scatter plot ($n=3$ biological replicates for 3 barcodes per condition, error bars are $\pm$ S.E.M.).

observed GFP expression in the expected characteristic central stripe (Teleman and Cohen, 2000) in the wing imaginal disc for 19/20 lines at similar expression levels to the base pJFRC12 *UAS-myr::GFP* construct inserted in the same landing site (Figure 4B, Supplemental Figure 6). No GFP expression was visible for TaG-EM barcode number 8. Gal4-driven expression levels of TaG-EM barcoded GFP constructs were also similar to that of the pJFRC12 base construct for multiple driver lines (Supplemental Figure 7) indicating that the presence of the barcode does not generally impair expression of GFP.

Boosting the GFP signal of TaG-EM constructs to enable robust cell sorting

While with some driver lines, expression of the *myr::GFP* from the TaG-EM construct may be too weak to allow robust enrichment of the tagged cells, adding an additional hexameric GFP construct (Shearin et al., 2014) could boost expression of weak driver lines to levels that are sufficient for robust detection of labeled flies or larvae (Figure 4C) and for labeling of dissociated cells for flow cytometry (Supplemental Figure 8). Stocks with an additional UAS hexameric GFP construct recombined onto the same chromosome as the TaG-EM construct have been established for all 20 TaG-EM barcode lines.

Correlation between expression of TaG-EM barcodes and intestinal cell marker genes in single-cell sequencing data

To test whether we could detect TaG-EM barcodes in single-cell sequencing data, we crossed three TaG-EM barcode lines to two different gut Gal4 driver lines (Ariyapala et al., 2020), one expressing in the enterocytes (EC-Gal4) and the other in intestinal precursor cells (PC-Gal4), including stem cells and enteroblasts (EBs). Due to weak GFP expression with the EC-Gal4 driver, we did not see visible GFP positive cells for this driver line. The PC-Gal4 driver line contained an additional UAS-Stinger (2xGFP) construct and expressed GFP at a level sufficient for flow sorting when crossed to the TaG-EM line (Supplemental Figure 7). Larval guts were dissected, dissociated, stained with propidium iodide (PI) to label dead cells, and flow sorted to recover PI negative and GFP positive cells. Approximately 10,000 cells were loaded into a 10x Genomics droplet generator and a single-cell library was prepared and sequenced. Two clusters were observed in the resulting sequencing data, one of which had high read counts from mitochondrial genes suggesting that this cluster consisted of mitochondria, debris, or dead and dying cells, after filtering the cells with high mitochondrial reads, a single cluster remained (Supplemental Figure 9). This cluster expressed known intestinal precursor cell markers such as *escargot* (*esg*), *klumpfuss* (*klu*), and *Notch* pathway genes like *E(spl)mbeta-HLH* (Supplemental Figure 9). Expression of all three PC-Gal4 driven TaG-EM barcodes was observed in this cluster (Supplemental Figure 9) indicating that TaG-EM barcodes can be detected in single-cell sequencing data. Interestingly, TaG-EM barcode 8, for which no GFP expression was observed was represented in the single-cell sequencing data indicating that the insertion of this specific barcode sequence may affect GFP translation, but not mRNA expression for this line.

A previous study used droplet-based single-cell sequencing to characterize the cell types that make up the adult midgut (Hung et al., 2020). This study took advantage of two fluorescent protein markers, an *escargot* (*esg*)-GFP fusion protein and a *prospero* (*pros*)-Gal4 driven RFP to label the intestinal stem cells (ISCs) and enteroendocrine cells (EECs), respectively (Hung et al., 2020). The authors compared the resulting clusters to a list of known marker genes in the literature, including antibody staining, GFP, LacZ, and Gal4 reporter expression patterns to classify the cells in individual clusters, and also found that the *esg*-GFP expression was present in a broader subset of cells than anticipated. Thus, most of these cell classifications relied upon inference as opposed to direct positive labeling. Recently, a large collection of split-*Gal4* lines were screened for expression in the adult and larval gut (Ariyapala et al., 2020). These include pan-midgut driver

lines, split-*Gal4* lines specific for the EBs, ECs, EECs, and ISC/EBs, as well as driver lines with regionalized gene expression. We crossed four different TaG-EM barcode lines with the pan-midgut driver (PMG-Gal4), and one barcode line to each of the precursor cell (PC-Gal4), enterocyte (EC-Gal4), enteroblast (EB-Gal4) and enteroendocrine (EE-Gal4) drivers (Ariyapala et al., 2020). Larval guts were dissected from these lines and cells were dissociated, flow sorted as described above to select live, GFP positive cells, and approximately 30,000 cells were loaded into a 10x Genomics droplet generator for single-cell sequencing. Using the additional hexameric GFP construct to boost GFP expression resulted in visible fluorescent signal for all eight barcode Gal4 line combinations.

An advantage of cell barcoding both for cell hashing (Stoeckius et al., 2018) and for TaG-EM *in vivo* barcoding is that such labeling facilitates the identification and removal of multiplets, which are an artifact of droplet-based single-cell sequencing approaches. After filtering and removing cells with a high percentage of mitochondrial or ribosomal reads, we searched for cells that co-expressed multiple TaG-EM barcodes. Out of 8,847 cells, we identified and removed 854 cells (9.65%) that expressed two different TaG-EM barcodes (Figure 5A, Supplemental Figure 10). We additionally attempted amplifying the TaG-EM barcodes in parallel to preparation of the gene expression library, following a modified cell hashing enrichment protocol. However, this enrichment was unsuccessful (resulting sequencing reads were predominantly off-target) and thus amplification of TaG-EM barcodes will require further troubleshooting.

After doublet removal, the remaining cells were clustered (Figure 5B) and analyzed using ScanPy (Wolf et al., 2018). Analysis of differentially expressed genes identified clusters expressing marker genes previously reported for adult gut cell types (Hung et al., 2020). These included genes associated with precursor cells (*Notch* pathway genes), enterocytes (trypsin, serine proteases, amylase, mannosidases), and enteroendocrine cells (neuropeptides and neuropeptide receptors) (Supplemental Figure 11, Supplemental Figure 12).

As expected, the barcodes expressed by the pan-midgut driver were broadly distributed across the cell clusters (Supplemental Figure 13). In contrast, expression of the cell type-specific barcodes showed more restricted patterns of expression among the cell clusters and were co-localized with known marker genes for these cell types (Figure 5C-N). For instance, TaG-EM barcode 4 expression (Figure 5F), which was driven by the EC-Gal4 line, was seen in clusters that expressed enterocyte markers such as the serine protease *Jon99Ciii* (Figure 5C) and other enterocyte marker genes such as the amylase, *Amy-d* (not shown). TaG-EM barcode 4 did not label other classes of enterocyte cells such as *beta-Trypsin* or *LManVI* positive enterocytes (Figures 5D-E). TaG-EM barcode 6 (Figure 5J), driven by the EB-Gal4 line, was strongly expressed in cells that also expressed precursor cell markers such as *esg* (Figure 5G), *klu* (Figure 5H), and *Notch* pathway genes such as *e(spl)mbeta-HLH* (Figure 5I) and *e(spl)m3-HLH* (not shown). Finally, expression of TaG-EM barcode 9 (Figure 5N), which was expressed using the EE-Gal4 driver line, was observed in clusters of cells that also expressed enteroendocrine derived neuropeptide genes such as *Dh31* (Figure 5K) and other enteroendocrine markers such as *IA-2*, a tyrosine phosphatase involved in the secretion of insulin-like peptide (Figure 5L). As with the enterocyte labeling, the EE-Gal4 expressed TaG-EM barcode 9 did not label all classes of enteroendocrine cells and other clusters of presumptive enteroendocrine cells expressing other neuropeptides such as *Orcokinin* (Figure 5M), *AstA* (not shown), and *AstC* (not shown), or neuropeptide receptors such as *CCHa2* (not shown) were also observed. The EE-Gal4 driver uses *Dh31* regulatory elements, so it is not surprising that the TaG-EM barcodes specifically labeled *Dh31*-positive enteroendocrine cells and this result further highlights the ability to target specific genetically defined cell types using TaG-EM based on *in vivo* cell labeling. Taken together, these results demonstrate that TaG-EM can be used to label specific cell populations *in vivo* for subsequent identification in single-cell sequencing data.

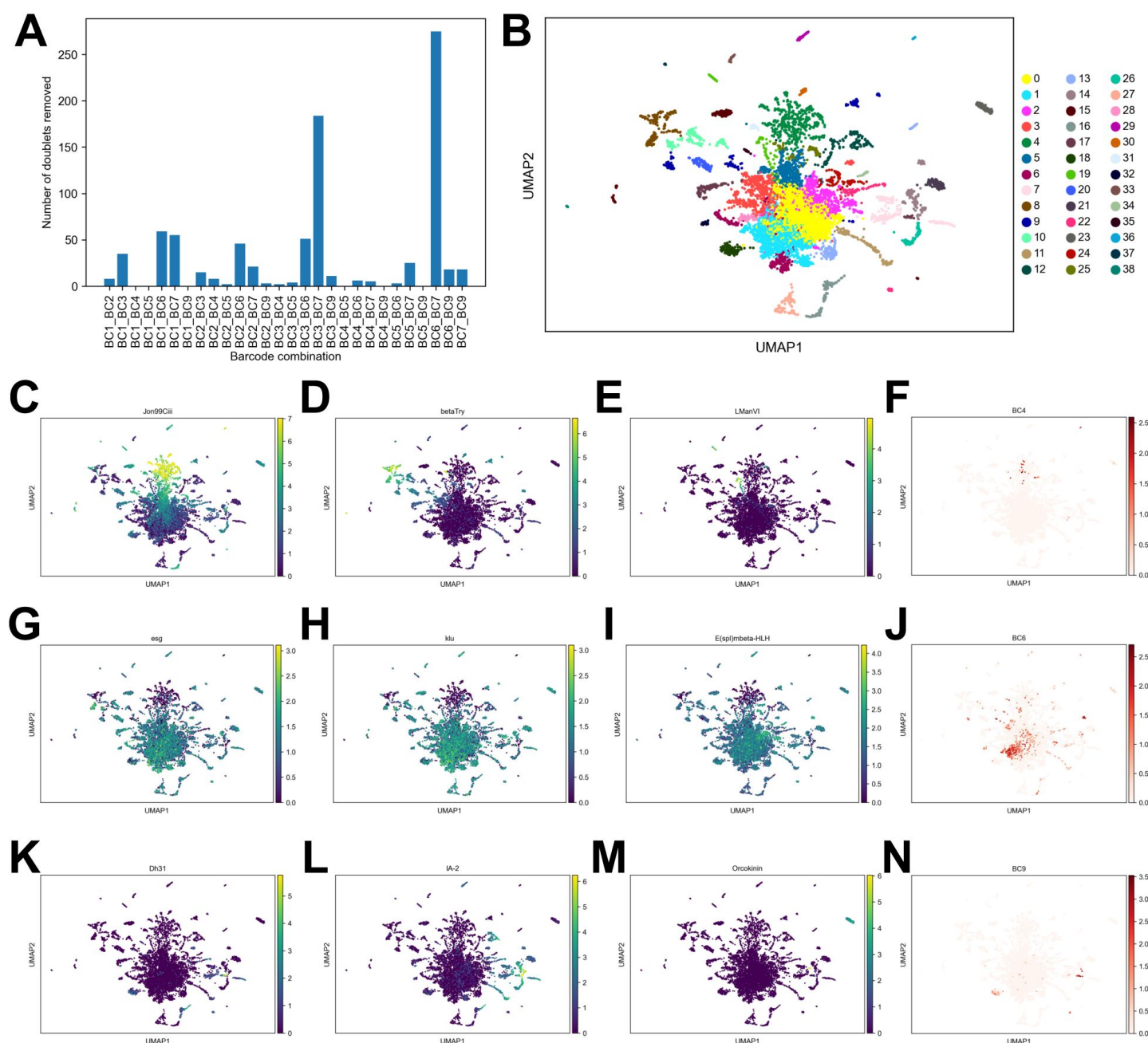


Figure 5.

Expression of TaG-EM genetic barcodes in larval intestinal cell types

A) Number of doublets removed due to co-expression of each of the pairwise combinations of TaG-EM barcodes. B) UMAP plot of *Drosophila* larval gut cell clusters after TaG-EM barcode-based doublet removal. UMAP plots showing gene expression levels of enterocyte marker genes C) *Jon99Ciii*, D) *betaTry*, E) *LManVI*, and F) the TaG-EM barcode (BC4) driven by the EC-Gal4 line. UMAP plots showing gene expression levels of enteroblast/ISC marker genes G) *esg*, H) *klu*, I) *E(spl)mbeta-HLH*, and J) the TaG-EM barcode (BC6) driven by the EB-Gal4 line. UMAP plots showing gene expression levels of enteroendocrine cell marker genes K) *Dh31*, L) *IA-2*, M) *Orcokinin*, and N) the TaG-EM barcode (BC9) driven by the EE-Gal4 line.

Discussion

Advances in next-generation sequencing, as well as single-cell and spatial genomics are enabling new types of detailed analyses to study important biological processes such as development and nervous system function. Here we describe TaG-EM, a genetic barcoding strategy that enables novel capabilities in several different experimental contexts (**Figure 1** [↗](#)).

We demonstrate that the genetic barcodes can be quantified from mixtures of barcoded fly lines using next-generation sequencing. Analysis of structured pools of flies with defined inputs suggests that TaG-EM barcode measurements are highly accurate and reproducible, particularly in cases where multiple barcodes are used to label an experimental condition and averaged (**Figure 2** [↗](#)). Sequencing-based TaG-EM measurements recapitulated more laborious, one-at-a-time measurements in both a phototaxis assay and an age-dependent fecundity assay, demonstrating that TaG-EM can be used to measure behavior or other phenotypes in multiplexed, pooled populations (**Figure 3** [↗](#)). We did note that one line (TaG-EM barcode 14) exhibited poor performance in oviposition assays, suggesting that barcode performance should be verified for a specific assay, though other strategies such as averaging across multiple barcode lines or permutation of barcode assignment across replicates could also mitigate such deficiencies. Currently, up to twenty conditions can be multiplexed in a single pooled experiment with existing TaG-EM lines, but because sequencing indices can be added after amplification in a separate indexing PCR step, hundreds or even thousands of such experiments can be multiplexed in a single sequencing run.

In addition, we show that TaG-EM barcodes can be expressed by tissue-specific Gal4 drivers and used to tag specific cell populations upstream of single-cell sequencing (**Figures 4** [↗](#)–**5** [↗](#)). This capability will allow for positive identification of cell clusters in cell atlas projects and will facilitate multiplexing of single-cell sequencing experiments. Recently, a conceptually similar approach called RABID-Seq (Clark et al., 2021 [↗](#)) has been described, which allows trans-synaptic labeling of neural circuits using barcoded viral transcripts. However, one distinction between the two approaches is that RABID-Seq relies on stochastic viral infection of mammalian cells while TaG-EM allows reproducible targeting of defined cell populations allowing unambiguous cell identification and potentially allowing the same cell populations to be assessed at different timepoints or in the context of different experimental manipulations. One current limitation is that TaG-EM barcodes are not observed in all cells in single-cell gene expression data and low expressing driver lines may result in particularly sparse labelling. In antibody-conjugated oligo cell hashing approaches, this sparsity of barcode representation is overcome by spiking in an additional primer at the cDNA amplification step and amplifying the hashtag oligo by PCR. A similar enrichment approach could be employed to generate dense labeling of the TaG-EM barcodes, improving upon the current mapping of barcode-containing transcripts observed in the gene expression data, and facilitating comparison between replicates and robust detection of multiplets.

In the future, generation of additional TaG-EM lines will enable higher levels of multiplexing. In addition, while the original TaG-EM lines were made using a membrane-localized myr::GFP construct, variants that express GFP in other cell compartments such as the cytoplasm or nucleus could be constructed to enable increased expression levels or purification of nuclei. Nuclear labeling could also be achieved by co-expressing a nuclear GFP construct with existing TaG-EM lines in analogy to the use of hexameric GFP described above.

In summary, combined with the large collections of Gal4 and split-Gal4 lines that have been established in *Drosophila* that enable precise targeting of a high proportion of cell types (Ariyapala et al., 2020 [↗](#); Aso et al., 2014 [↗](#); Davis et al., 2020 [↗](#); Gohl et al., 2011 [↗](#); Namiki et al., 2018 [↗](#); Pfeiffer et al., 2010 [↗](#), 2008 [↗](#); Venken et al., 2011 [↗](#)), TaG-EM provides a means to target and label

cells *in vivo* for subsequent detection in single-cell sequencing. Moreover, these genetic barcodes can be used to multiplex behavioral or other phenotypic measurements. Thus, TaG-EM provides a flexible system for barcoding cells and organisms.

Methods

Drosophila stocks and maintenance

Drosophila stocks were grown at 22°C on cornmeal agar unless otherwise indicated. The following stocks were used in this study:

Stock name	Source
IsoD1	Clandinin Lab, Stanford University
<i>w⁻</i> ; +; + (IsoD1)	Clandinin Lab, Stanford University
<i>attP2</i> line	BDSC #25710
<i>norpA</i>	BDSC# 9048
UAS-myr::GFP (pJFRC12)	BDSC #32197
Hexameric GFP lines	BDSC #91402, #91403
<i>dpp</i> -Gal4 driver	BDSC #1553
Act-Gal4 driver	BDSC #4414
Tub-Gal4 driver	BDSC #5138
Mhc-Gal4 driver	BDSC #55132
PC-Gal4 driver	BDSC #91400 + #91401
EC-Gal4 driver	BDSC #91405
EB-Gal4 driver	BDSC #91398 + #91404
EE-Gal4 driver	BDSC #91402 + #91403
PMG-Gal4 driver	BDSC #91398 + #91399
TaG-EM lines (x20)	This study
TaG-EM lines + 6xGFP (x20)	This study

Design and cloning of TaG-EM constructs

A gBlock with the following sequence containing a part of the SV40 3' UTR with a PCR handle (uppercase, below) and a 14 bp randomer sequence just upstream of the SV40 polyadenylation site (bold and underlined, below) was synthesized (Integrated DNA Technologies, IDT):

caaaggaaaaagctgcactgctatacaagaaaattatggaaaaatattgatgtatagtccttgactagagatcataatcagccata
ccacattttagaggttttactgtcttaaaaaacccacaccccccgaacctgaacataaaatgaatgaattgtgttgaact
tggtttatgcagcttataaCTTCCAACAACCGGAAGTGANNNNNNNNNNNNNNNtggttacaaataaagcaatag
catcacaatttcacaaataaagcattttttcactgcaattctagttgtgtgtgtccaaactcatcaatgtatcttatcatgtctggatcgatct
ggccggccgtttaaacgaattctgaagacgaaaggcctcgatagcctattttataggttaatgtcatgataataatg The gBlock

was resuspended in 20 µl EB, incubated at 50°C for 20 minutes and then cut with *PsiI* and *EcoRI* (New England Biolabs, NEB) using the following reaction conditions: 4 µl gBlock DNA (35 ng), 2 µl 10x CustSmart buffer (NEB), 1 µl *EcoRI* enzyme (NEB), 1 µl *PsiI* enzyme (NEB), and 12 µl nuclease-free water were mixed and incubated at 37°C for 1 hour followed by 65°C for 20 minutes to heat inactivate the restriction enzymes. pJFRC12-10XUAS-IVS-myr::GFP plasmid (Addgene, Plasmid #26222) (Pfeiffer et al., 2010 [DOI](#)) was digested with the following reaction conditions: 5 µl pJFRC12-10XUAS-IVS-myr::GFP plasmid DNA (~3 µg), 5 µl 10x CutSmart buffer (NEB), 1 µl *PsiI* enzyme (NEB), 1 µl *EcoRI* enzyme (NEB), and 38 µl nuclease-free water, were mixed and incubated at 37°C for 1 hour, followed by addition of 1 µl of CIP and incubation for an additional 30 minutes. The digested vector backbone was gel purified using the QiaQuick Gel Purification Kit (Qiagen). The digested gBlock was ligated into the digested pJFRC12-10XUAS-IVS-myr::GFP backbone using the following reactions conditions: 4 µl T4 ligase buffer (10x) (NEB), 20 µl plasmid backbone DNA (0.005 pmol), 5 µl gBlock digest DNA (0.03 pmol), 2 µl of T4 DNA ligase (NEB), and 9 µl nuclease-free water were mixed and incubated at 22°C for 2 hours. 2 µl of the ligation reaction was transformed into 50 µl of TOP10 competent cells (Invitrogen), and the cells were incubated on ice for 30 minutes, then heat shocked at 42°C for 30 seconds, and incubated on ice for 5 minutes. 250 µl SOC was added and the cells were plated on LB+ampicillin plates and incubated overnight at 37°C. DNA was isolated from 36 pJFRC12-gBlock colonies using a QIAprep Spin MiniPrep kit (Qiagen). Expected construct size was verified by diagnostic digest with *EcoRI* and *ApaI*. DNA concentration was determined using a Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific) and the randomer barcode for each of the constructs was determined by Sanger sequencing using the following primers:

SV40_post_R: GCCAGATCGATCCAGACATGA

SV40_5F: CTCCCCTGAACCTGAAACA

Generation of TaG-EM transgenic lines

29 sequence verified constructs were normalized, pooled evenly, and injected as a pool into embryos (Rubin and Spradling, 1982 [DOI](#)) expressing PhiC31 integrase and the carrying the *attP2* landing site (BDSC #25710). Injected flies were outcrossed to w-flies, and up to three *white+* progeny per cross were identified, and the transgenic lines were homozygosed. DNA was extracted (GeneJET genomic DNA purification Kit, Thermo Scientific) and the region containing the DNA barcode was amplified with the following PCR reaction: 2.5 µl 1:10 diluted template DNA, 2 µl 10x Reaction Buffer (Qiagen), 0.2 µl dNTP mix (10 µM), 1 µl 10 µM SV40_5F primer (10 µM), 1 µl SV40_post_R primer (10 µM), 0.8 µl MgCl₂ (3 mM), 0.1 µl Taq polymerase (Qiagen), 12.4 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 95°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 55°C for 30 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes. PCR products were treated with Exo-CIP using the following reaction conditions: 5 µl PCR product, 1 µl Exo-CIP Tube A (NEB), 1 µl Exo-CIP Tube B (NEB) were mixed and incubated at 37°C for 4 minutes, followed by 80°C for 1 minute. The barcode sequence for each of the independent transgenic lines was determined by Sanger sequencing using the SV40_5F and SV40_PostR primers. Transgenic lines containing 20 distinct DNA barcodes were recovered (Supplemental Figure 1 [DOI](#)).

Optimizing amplification of TaG-EM barcodes for next-generation sequencing

The following primers were evaluated to amplify the TaG-EM barcodes upstream of NGS: Forward primer pool; four primers with frameshifting bases to increase library sequence diversity in initial sequencing cycles were normalized to 10 µM and pooled evenly to make a B2_3'F1_Nextera_0-6 primer pool:

B2_3'F1_Nextera:

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCCAACAACCGGAAG*TGA

B2_3'F1_Nextera_2

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCTTCCAACAACCGGAAG*TGA

B2_3'F1_Nextera_4

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGACTTCCAACAACCGGAAG*TGA

B2_3'F1_Nextera_6

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGAGCTTCCAACAACCGGAAG*TGA

The following two reverse primers were tested:

SV40_pre_R_Nextera:

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTTGTGAAATTTGTGATGCTATTGC*T TT

SV40_post_R_Nextera:

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAGATCGATCCAGACA*TGA

SV40_pre_R_nextera is designed to produce a shorter amplicon (200 bp with Illumina adapters and indices added) and SV40-post_R_Nextera is designed to produce a longer amplicon (290 bp with Illumina adapters and indices added).

An initial test was performed with three different polymerases (Qiagen Q5, KAPA HiFi, and Qiagen Taq) at two different annealing temperatures and with both the B2_3'F1_Nextera/ SV40_pre_R_Nextera and B2_3'F1_Nextera/ SV40_post_R_Nextera primer sets to determine whether the primers amplify as expected (**Supplemental Figure 1** [↗](#)). Two different samples were tested:

1. Pool of 8 putative transformant samples (pooled 5 ul each of 1:10 diluted sample)
2. OreR (wild type - diluted 1:10)

Set up the following PCR reactions:

Q5 polymerase

2.5 µl template DNA, 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer (10 µM), 10 µl 2x Q5 Master Mix (Qiagen), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 98°C for 30 seconds, followed by 30 cycles of 98°C for 20 seconds, 55°C or 60°C for 15 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes.

KAPA HiFi polymerase

2.5 µl template DNA, 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 95°C for 5 minutes, followed by 30 cycles of 98°C for 20 seconds, 55°C or 60°C for 15 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes.

Taq polymerase

2.5 µl template DNA, 2 µl 10x Reaction Buffer (Qiagen), 0.2 µl dNTP mix (10 µM), 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer (10 µM), 0.8 µl MgCl₂ (3 mM), 0.1 µl Taq polymerase (Qiagen), 12.4 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 95°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 55°C or 60°C for 30 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes. Samples were run on a 2% agarose gel to verify amplification products (**Supplemental Figure 1** [↗](#)).

Next, the TaG-EM barcode lines were pooled in either an even or staggered manner. To optimize reaction conditions for the barcode measurements, either 5 ng or 50 ng of DNA was amplified in triplicate for each pool for either 20, 25, or 30 cycles with either KAPA HiFi using the B2_Nextera_F 0-6 forward primer pool together with either the SV40_pre_R_Nextera or the SV40_post_R_Nextera reverse primer. Next, PCR reactions were diluted 1:100 in nuclease-free water and amplified in the following indexing reactions: 3 µl PCR 1 (1:100 dilution), 1 µl indexing primer 1 (5 µM), 1 µl indexing primer 2 (5 µM), and 5 µl 2x Q5 master mix. The following indexing primers were used (X indicates the positions of the 8 bp indices):

Forward indexing primer:

AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC

Reverse indexing primer:

CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTCTCGTGGGCTCGG

Reactions were amplified using the following cycling conditions: 98°C for 30 seconds, followed by 10 cycles of 98°C for 20 seconds, 55°C for 15 seconds, 72°C for 1 minute, followed by 72°C for 5 minutes. Amplicons were then purified and normalized using a SequelPrep normalization plate (Thermo Fisher Scientific), followed by elution in 20 µl of elution buffer. An even volume of the normalized libraries was pooled and concentrated using 1.8x AmpureXP beads (Beckman Coulter). Pooled libraries were quantified using a Qubit dsDNA high sensitivity assay (Thermo Fisher Scientific) and libraries were normalized to 2 nM for sequencing on the Illumina MiSeq (see below).

Structured fly pool experiments

Male or female flies from TaG-EM barcode lines were pooled in either an even or staggered manner (**Figure 2A** [↗](#)). For the even pools, three independently pooled samples were constructed in order to assess sample-to-sample variability. DNA was extracted from these structured pools using a protocol adapted from Huang *et al.*, 2009 [↗](#) (Huang *et al.*, 2009 [↗](#)), using homemade SPRI beads (DeAngelis *et al.*, 1995 [↗](#)) in the last purification step and amplified in triplicate using 2.5 µl template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl SV40_post_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 98°C for 5 minutes, followed by 30 cycles of 98°C for 20 seconds, 60°C for 15 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared for sequencing as described above.

Phototaxis experiments

Video-based measurements

A pair of white LED strip lights with Muzata LED Strip Light Diffusers (U1SW WW 1M, LU1) were mounted within a light-tight box and controlled using an Vilros Uno Rev 3 microcontroller. Test tubes containing flies were held in place with Acoustic Foam Panels (1"x12"x12", ALPOWL). Videos and images were acquired using an Arducam 1080P Day & Night Vision USB Camera with an IR filter and using Photo Booth software (Apple). Wild type and *norpA* flies carrying one of 4 different TaG-EM barcodes were tested in three independent experimental replicates. 20 male flies of each genotype were transferred into 25 mm x 150 mm glass test tubes, incubated at 34°C for 10 minutes and then run in the phototaxis assay, where a light at one end of the chamber was turned on for 30 seconds. Videos of all tests were recorded through the end of the 30 second light pulse. Videos were independently scored by two observers to determine the number of flies in the light-facing or dark-facing tubes and the results were averaged. A preference index (P.I.) was calculated using the following formula: $[(\text{number of flies in light tube}) - (\text{number of flies in dark tube})]/(\text{total number of flies})$.

TaG-EM measurements

For TaG-EM barcode-based phototaxis measurements, the following genotypes were consolidated into a single test tube:

Pool A

norpA/Y;TaG-EM BC4/+

norpA/Y;TaG-EM BC3/+

w-/Y;TaG-EM BC2/+

w-/Y;TaG-EM BC1/+

Pool B

norpA/Y;TaG-EM BC2/+

norpA/Y;TaG-EM BC1/+

w-/Y;TaG-EM BC4/+

w-/Y;TaG-EM BC3/+

These pools were individually incubated at 34°C for 10 minutes and then run in the phototaxis assay. Videos of all tests were recorded and at the end of a 30 second light pulse the two test tubes were quickly separated and capped. Flies in each of these tubes were counted, then DNA was extracted from the flies from the light-facing or dark-facing tubes and amplified using 2.5 µl template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl SV40_post_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 95°C for 5 minutes, followed by 30 cycles of 98°C for 20 seconds, 60°C for 15 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared for sequencing as described above.

Oviposition experiments

Newly hatched flies (males and females) from three barcode lines were collected at one-week intervals during 4 consecutive weeks (12 barcode lines in total). Fresh fly food was provided every 3-4 days. Ten days after the last collection, 10 females from each barcode line were taken and pooled together in a collection cage (10 females x 12 barcode lines= 120 females). The remaining females from each barcode line were separated from the males and put in individual collection cages. Two days later, the experiment started and run for 3 consecutive days. Each day a 1-1.5 hour pre-collection was followed by a 6 hour collection, both at 25°C. 100 embryos from each individual collection plate were transferred to new plates and incubated for 2 days at 18°C. The number of hatched larvae were counted and used to calculate the egg survival rate. The pooled collection plate was also incubated at 18°C and the next day the embryos were dechorionated and frozen. The 12 individual collection plates were kept at 4°C and the number of embryos counted in the following days. For the barcode measurements, DNA was extracted from the embryos, and amplified using 2.5 µl template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl SV40_post_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling conditions: 95°C for 5 minutes, followed by 30 cycles of 98°C for 20 seconds, 60°C for 15 seconds, 72°C for 30 seconds, followed by 72°C for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared for sequencing as described above.

Cell dissociation and isolation

Midguts from 3rd instar larvae were dissected in phosphate-buffered saline (PBS) and transferred to microcentrifuge tubes on ice containing PBS + 30% normal goat serum (NGS). After dissection, 150 µL of 2.7 mg/mL elastase was added to each sample tube. The tubes were then incubated at 27°C for one hour. During incubation, samples were mixed by pipetting ~30 times every 15 minutes to improve elastase dissociation of the cells. Samples were then filtered through a 40 µM FlowMi tip filter (Bel-Art) to reduce debris. Afterwards, the samples were quantified on the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems) using 9 µL of sample to 1 µL AO/PI dye to ensure there were enough viable cells for flow sorting.

Once quantified, the samples were brought up to a volume of ~1.1mL with the PBS + 30% NGS solution to facilitate flow sorting. The samples were then fluorescently sorted on a FACSaria II Cell Sorter (BD Biosciences) to isolate GFP+ cells. Following sorting, samples were centrifuged at 300x g for 10 minutes to concentrate the cells. The supernatant was aspirated off until 50 µL cell concentrate remained in each sample. Then, the samples were carefully resuspended using wide bore pipette tips before being combined into one sample tube. This sample was quantified on the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems) as described above. If necessary, cells were centrifuged, concentrated, and re-counted.

Preparation of single-cell sequencing libraries

The resulting pool was prepared for sequencing following the 10X Genomics Single Cell 3' protocol (version CG000315 Rev C). Custom nested primers were used to attempt enrichment of TaG-EM barcodes after cDNA creation using PCR. The remaining cDNA followed the 10X 3' protocol without modification while the enriched TaG-EM portion followed BioLegends' "Total-Seq-A Antibodies and Cell Hashing with 10X Single Cell 3' Reagents Kit v3 or v3.1" Protocol starting at step III. The resulting 3' gene expression library and TaG-EM enrichment library were sequenced together following Scenario 1 of the BioLegends protocol. Later, the TaG-EM library was sequenced alone following Scenario 2 from the same protocol.

Sequencing

Libraries for TaG-EM barcode analysis from structured pools or from phototaxis or oviposition experiments were denatured with NaOH and prepared for sequencing according to the protocols described in the Illumina MiSeq Denature and Dilute Libraries Guides. Single-cell libraries were sequenced on the Illumina NextSeq 2000 or Illumina NovaSeq 6000.

Data analysis

Behavioral experiments

Demultiplexed fastq files were generated using bcl2fastq or bcl-convert. TaG-EM barcode data was analyzed using custom R and Python scripts and BioPython (Cock et al., 2009 [↗](#)). Leading primer sequences were trimmed using cutadapt (Martin, 2011 [↗](#)) and the first 14 bp of the remaining trimmed read were compared to a barcode reference file, with a maximum of 2 mismatches allowed, using a custom script which is available via Github: <https://github.com/darylgohl/TaG-EM> [↗](#).

Single-cell experiments

Data was initially mapped and analyzed using the Cellranger software package (version 7.0.1, 10X Genomics). A custom reference Drosophila genome was created by starting with the BDGP6.28 reference genome assembly and Ensembl gene annotations. Custom gene definitions for each of the TaG-EM barcodes were added to the fasta genome file and gtf gene annotation file and a Cellranger reference package was generated with the cellranger mkref command. Subsequent analysis was performed using Scanpy (version 1.9.3) (Wolf et al., 2018 [↗](#)). Cells expressing less than 200 genes and genes expressed in fewer than three cells were filtered from the expression matrix. Next, cells expressing more than 4,000 genes, cells expressing more than 15% mitochondrial transcripts, and cells expressing more than 30% ribosomal (RpL) genes were filtered out. Expression counts were log normalized and highly variable genes were filtered out (min_mean=0.0125, max_mean=3, min_disp=0.5). Cells expressing multiple TaG-EM barcodes were identified and removed and cells were clustered using the Leiden algorithm (Traag et al., 2019 [↗](#)) and projected with UMAP. Scanpy was used to call and visualize differentially expressed genes.

Declarations

Availability of data, code, and materials

Sequencing data for this project is available through the National Center for Biotechnology Information (NCBI) Sequence Read Archive BioProject PRJNA912199. Fly stocks containing each of the 20 TaG-EM barcodes together with an additional UAS hexameric GFP expression construct will be available from the Bloomington Drosophila Stock Center. The TaG-EM barcode analysis script and barcode reference fasta file are available via Github: <https://github.com/darylgohl/TaG-EM> [↗](#).

Competing interests

None to declare.

Funding

This study was supported by a grant from the Winston and Maxine Wallin Neuroscience Discovery Fund.

Author contributions

J.B.M. conceived and designed experiments, conducted experiments, analyzed data, revised manuscript, M.D. conducted experiments, revised manuscript, L.G. conducted experiments, revised manuscript, B.A. conducted experiments, revised manuscript, J.G. analyzed data, revised manuscript, D.M.G. conceived and designed experiments, conducted experiments, analyzed data, wrote and revised the manuscript.

Acknowledgements

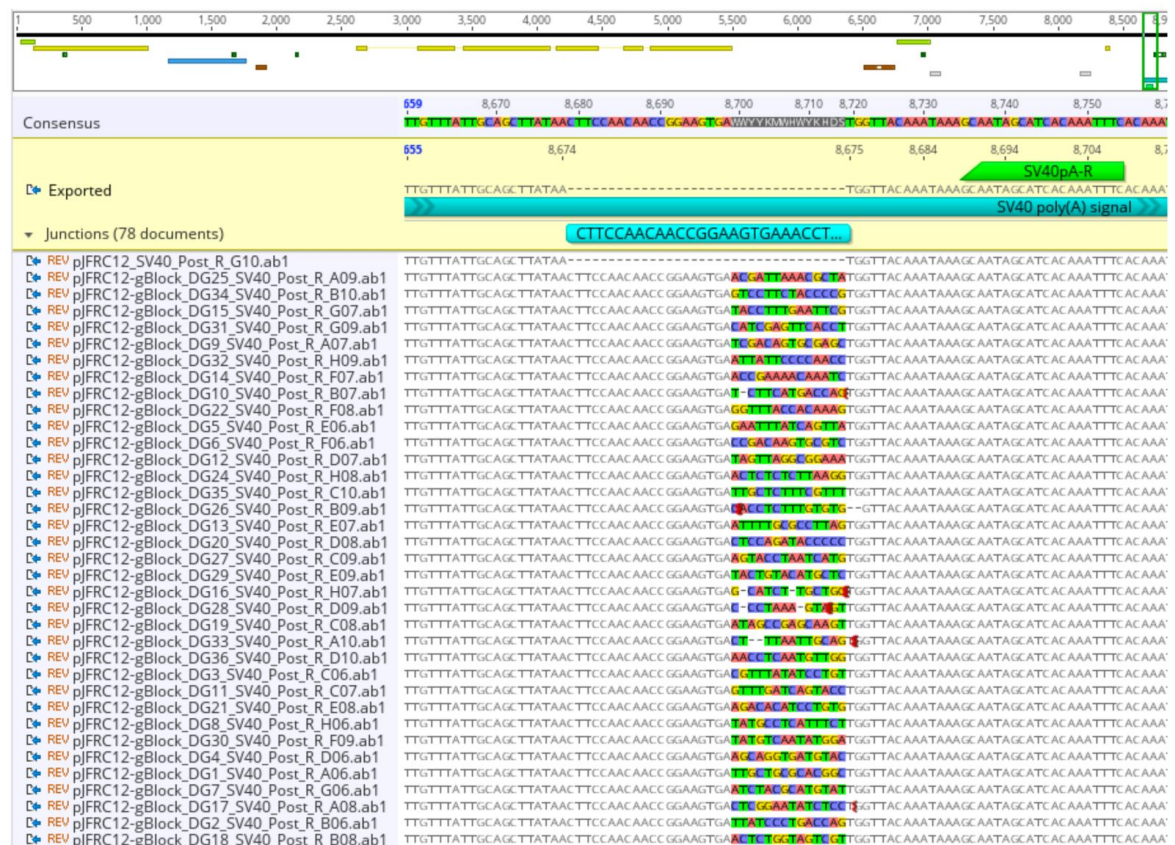
We thank our colleagues in the University of Minnesota Genomics Center (RRID:SCR_012413), in particular Aaron Becker, Dylan Cole, and Logan Silber for help with DNA sequencing, Emma Stanley, Fernanda Rodriguez, and Patrick Grady for assistance and advice on single-cell sequencing, and Kenneth Beckman, Andrew Alegria, Troy Louwagie, and Aaron Barnes for helpful feedback and discussions. This work was supported by the resources and staff at the University of Minnesota University Imaging Centers (RRID:SCR_020997). The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. URL: <http://www.msi.umn.edu> . Stocks obtained from the Bloomington Drosophila Stock Center (NIH P40OD018537) were used in this study.

Supplemental Figures

A

Barcode sequence	Transgenic line
ATTTTGGCGCTTAG	1-2
CCCTAAAGTACCGT	4-3
TAGTTAGGCGGAAA	5-4
TTGCTCTTTCGTTT	6-4
TTGCTGCGCACGGC	7-1
CGTTTATATCCTGT	10-1
AGCAGGTGATGTAC	1-1, 1-3, 10-2
TATGCCTCATTCT	13-2
GTTTGATCAGTACC	13-4
GGTTTACCACAAAG	14-2
AGACACATCCTGTG	15-1
GAATTTATCAGTTA	15-2
TCTTCATGACCAGC	2-1, 2-2, 2-3, 11-1, 11-2, 11-4
TATGTCAATATGGA	3-1, 3-2, 3-3, 9-2
ACTCTCTCTTAAGG	4-1, 4-4
CTTTAATTGCAGTC	5-3, 5-5, 9-1, 9-3
TACTGTACATGCTC	6-2, 10-4, 14-3, 14-4
TACCTTTGAATTCG	6-3, 12-2, 12-4
CATCGAGTTCACCT	8-1, 12-3
GTCCTTCTACCCCG	8-2, 8-4

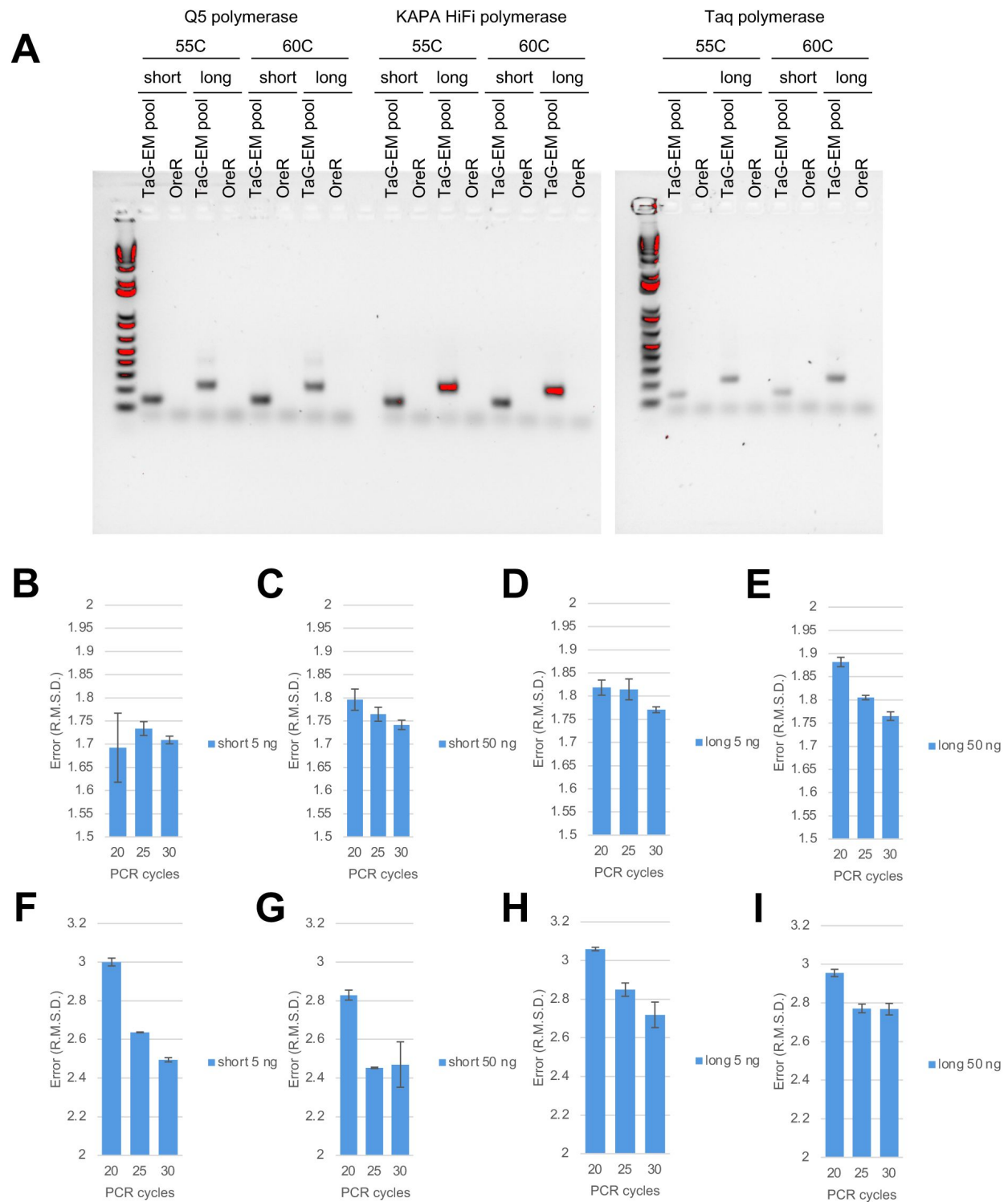
B



Supplemental Figure 1.

Sanger sequencing identification of TaG-EM barcode lines

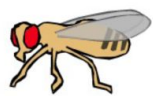
A) Summary of barcode pool injections. Barcode sequence and transgenic vial identifier in which the barcode was identified are shown. B) Sanger sequencing-based confirmation of the barcode sequence and PCR handle in TaG-EM transgenic lines.



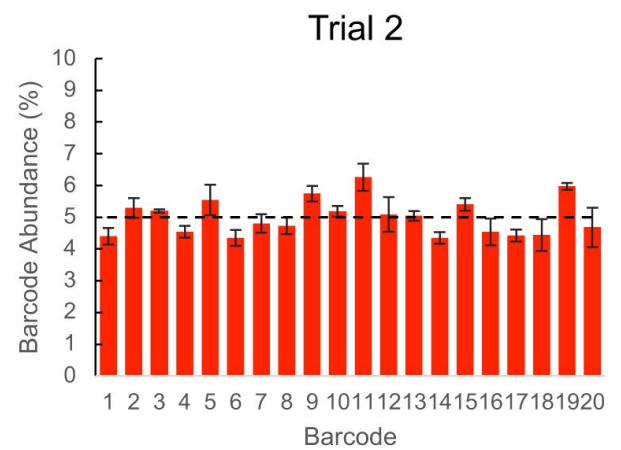
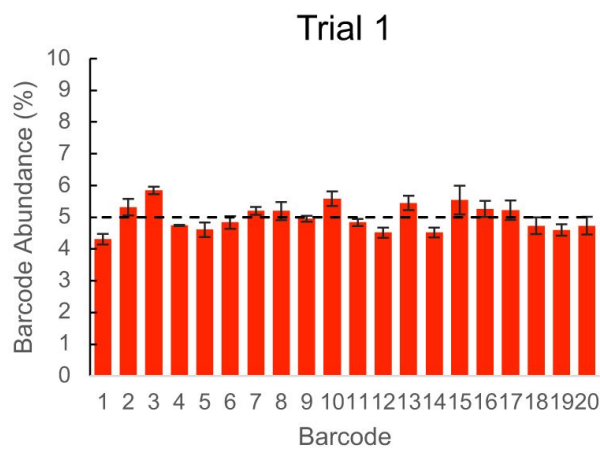
Supplemental Figure 2.

Optimization of TaG-EM barcode amplification

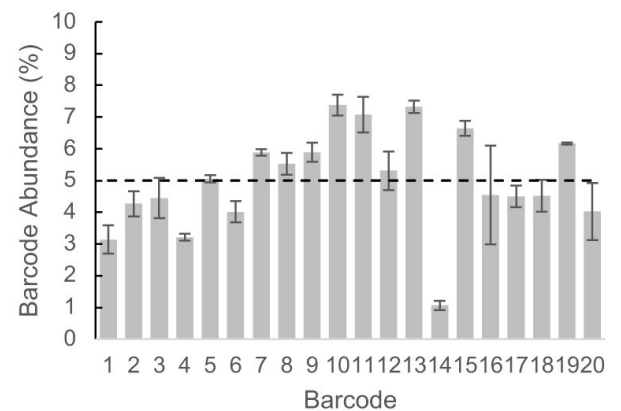
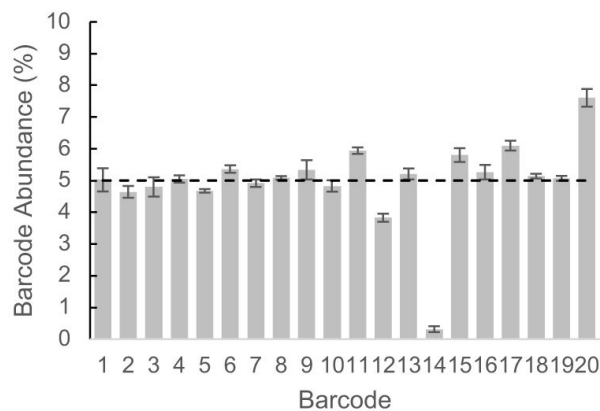
A) Gels showing bands produced when amplifying TaG-EM flies or a wild type control with the indicated polymerase, annealing temperature, and primer pair (short = B2_3'F1_Nextera/ SV40_pre_R_Nextera; long = B2_3'F1_Nextera/ SV40_post_R_Nextera). B-E) Mean error (R.M.S.D. root mean squared deviation from expected value) for even pool amplified with the indicated primer set, input amount, and cycle number using KAPA HiFi polymerase ($n=3$, error bars are $\pm$ S.E.M.). F-G) Mean error (R.M.S.D. root mean squared deviation from expected value) for staggered pool amplified with the indicated primer set, input amount, and cycle number using KAPA HiFi polymerase ($n=3$, error bars are $\pm$ S.E.M.).



Adults



Embryos

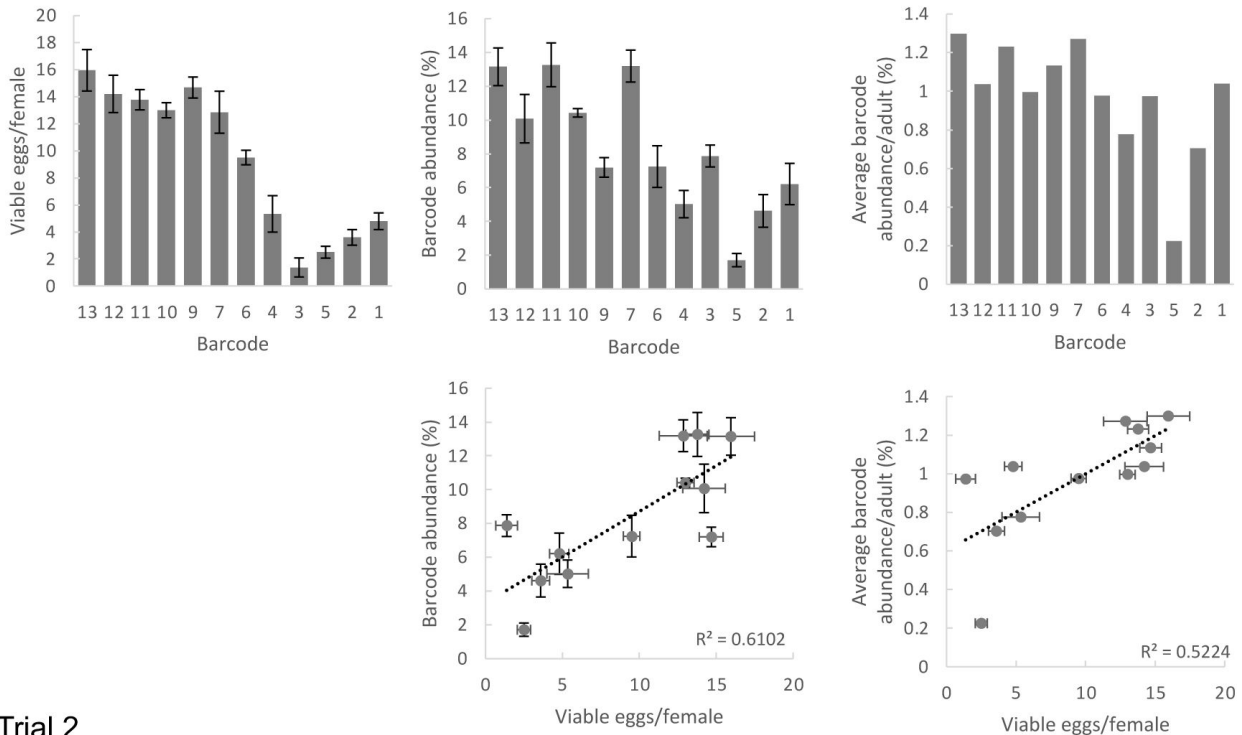


Supplemental Figure 3.

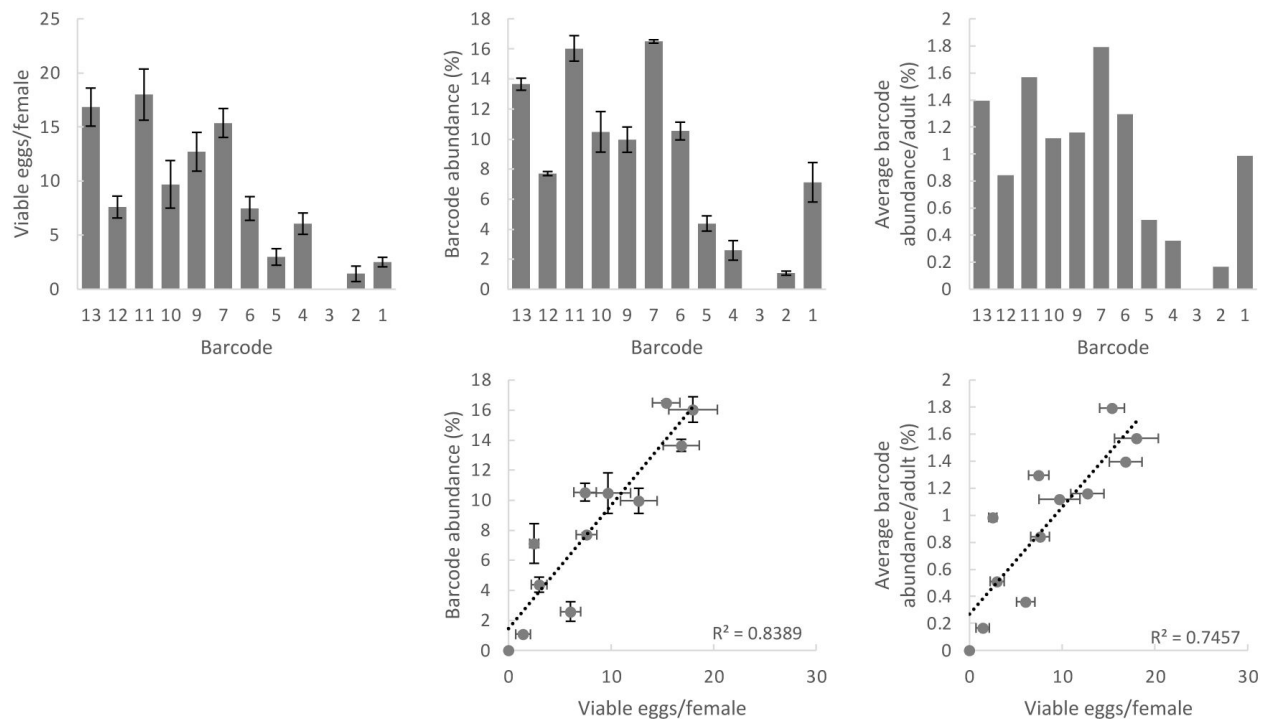
Oviposition tests with TaG-EM barcode lines

Plots showing mean TaG-EM barcode abundance for adult females used in oviposition experiments (top) and eggs collected from these females (bottom). Data from two independent trials is shown ($n=3$ for each trial, error bars are $\pm$ S.E.M.).

Trial 1



Trial 2

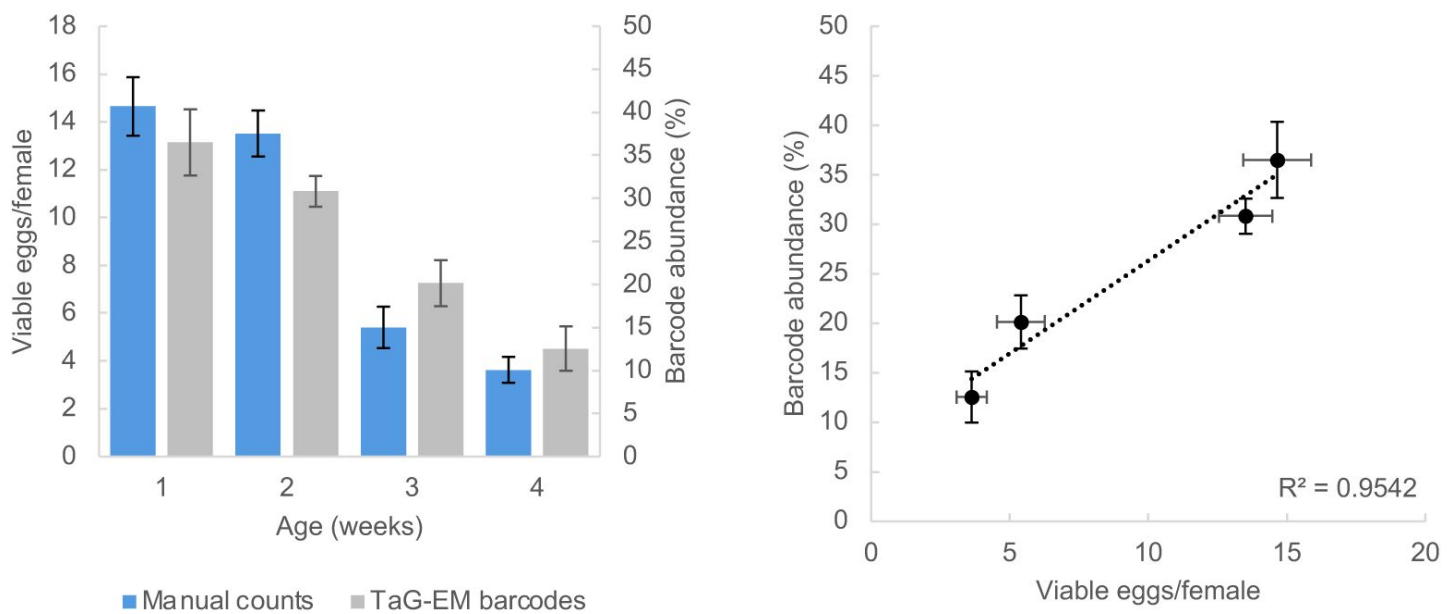


Supplemental Figure 4.

Fecundity data for individual TaG-EM lines

Manually collected data for mean number of viable eggs per female, barcode abundance data, and barcode abundance data normalized to adult fly barcode data for each of the TaG-EM barcode lines used in the age-dependent fecundity experiment. Scatterplots show correlations between manually collected data and barcode sequencing results. Data from two independent trials is shown ($n=3$ for each trial, error bars are $\pm$ S.E.M.).

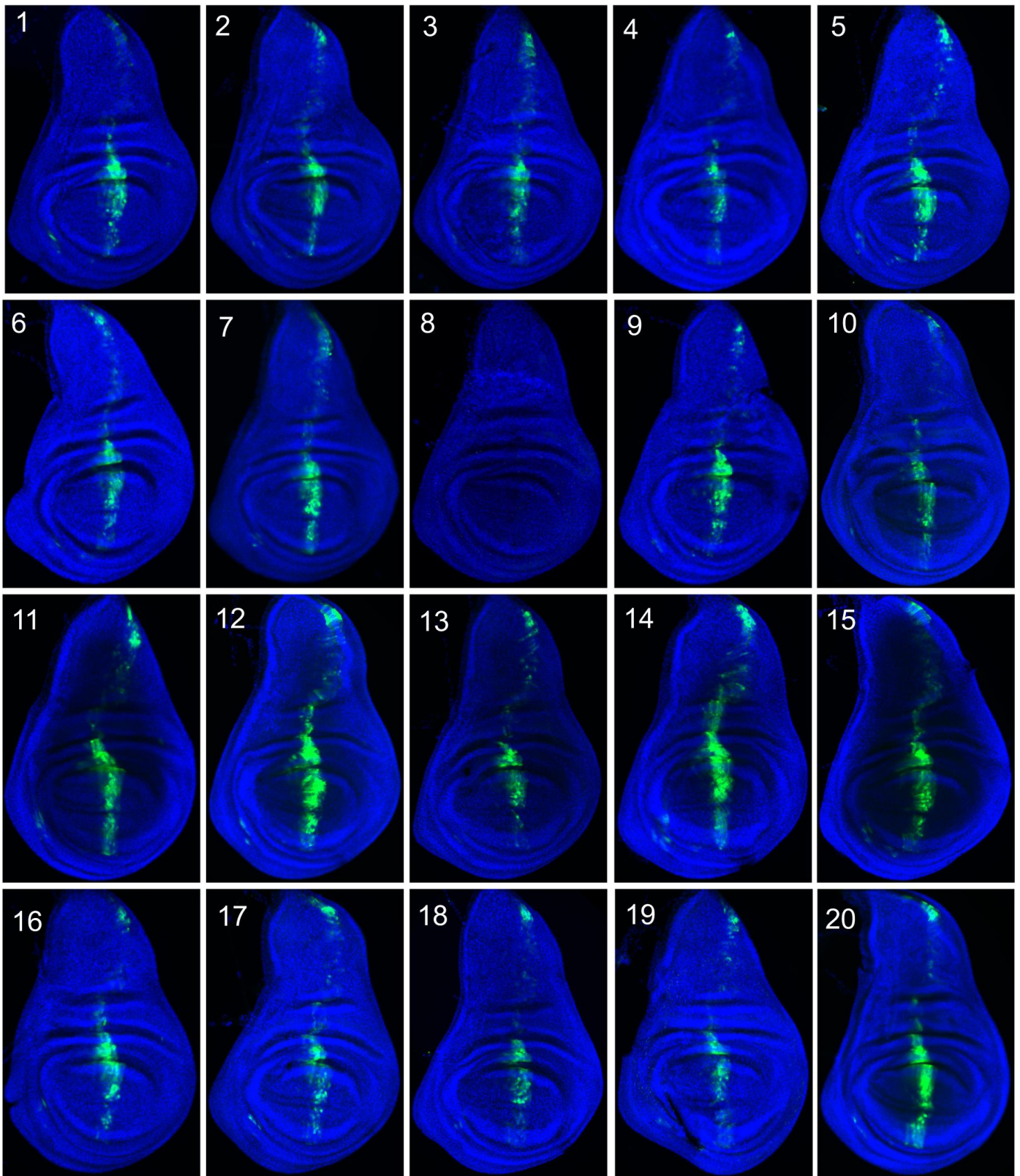
Trial 1



Supplemental Figure 5.

Average age-dependent fecundity data for Trial 1

Average number of viable eggs per female (manual counts) and average barcode abundance are shown both as a bar plot and scatter plot (n=3 biological replicates for 3 barcodes per condition, error bars are $\pm$ S.E.M.). Data from Trial 2 is shown in [Figure 3C](#).

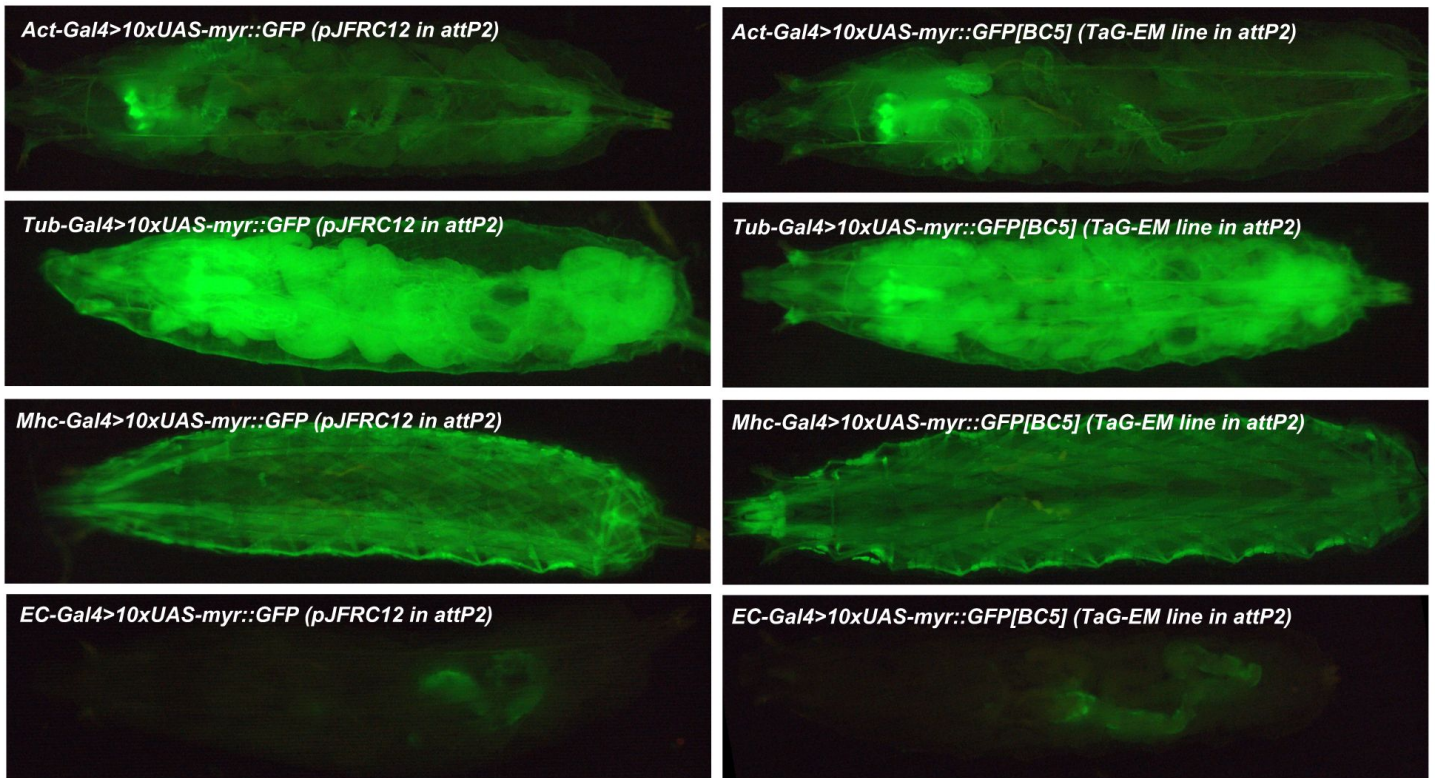


Supplemental Figure 6.

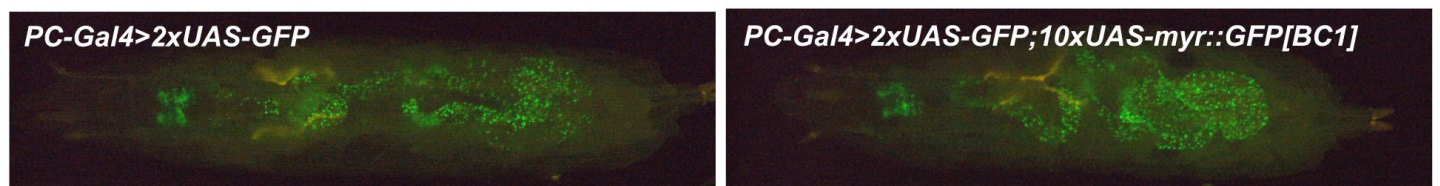
Expression driven by *dpp-Gal4* for 20 TaG-EM lines

GFP antibody staining in the wing imaginal disc for the indicated TaG-EM barcode line driven by *dpp-Gal4*. Wing discs are counterstained with DAPI.

A



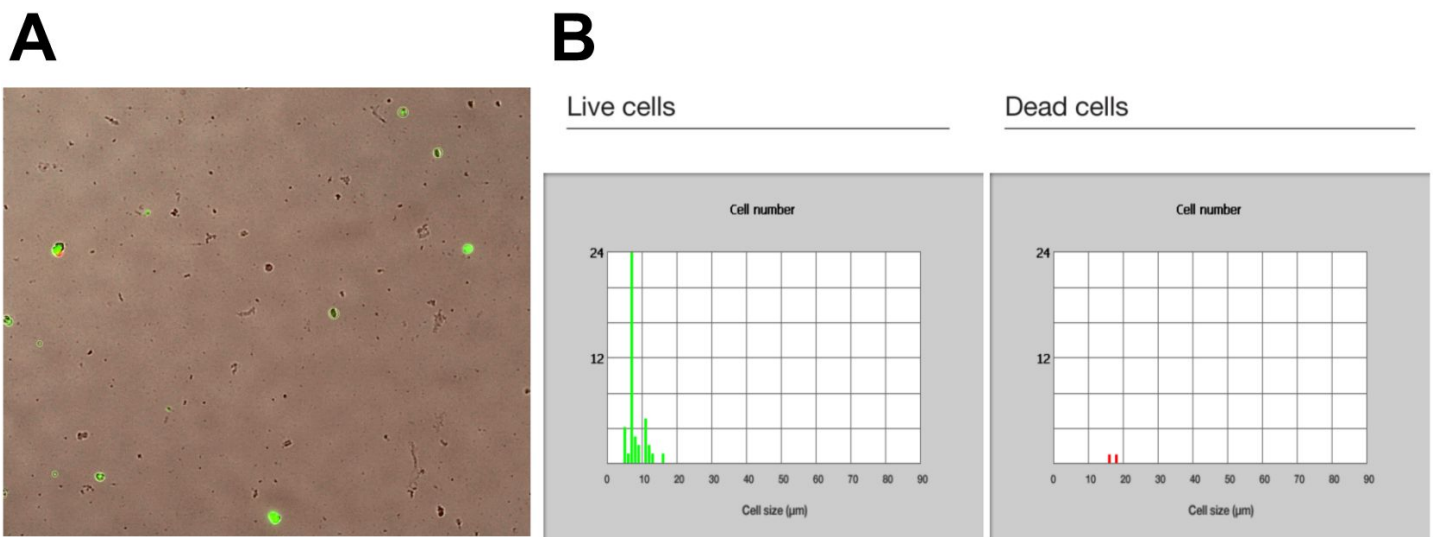
B



Supplemental Figure 7.

TaG-EM line GFP expression driven by different Gal4 drivers

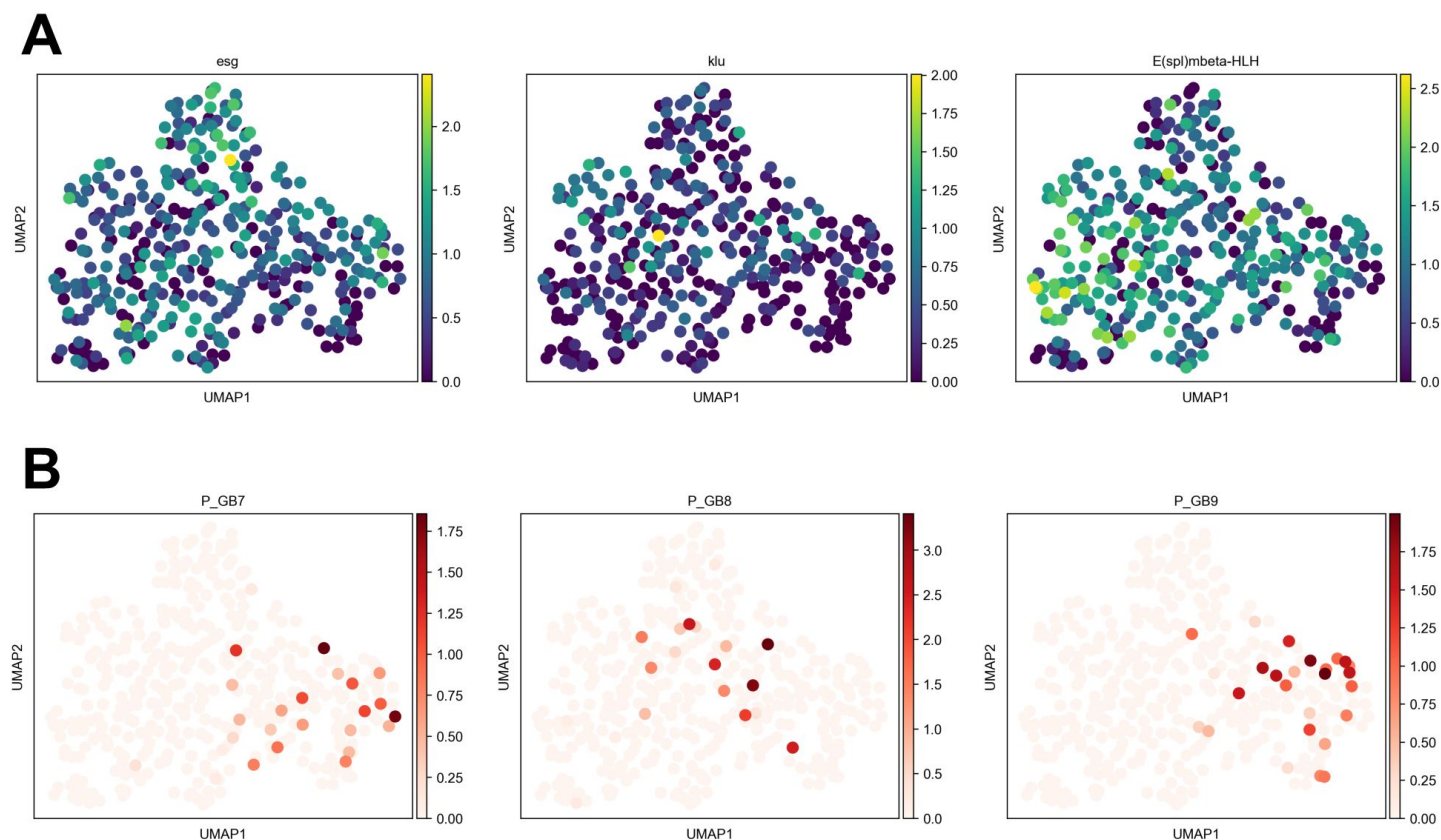
A) Comparison of GFP expression in larvae for the original pJFRC12 construct inserted in the attP2 landing site (left) or for a TaG-EM barcode line (right) expressed under the control of the indicated driver line. B) GFP expression of the PC-Gal (Precursor-Gal4) driver line together with either UAS-2xGFP or a combination of UAS-2xGFP and a TaG-EM barcode line.



Supplemental Figure 8.

Dissociated intestinal cell viability

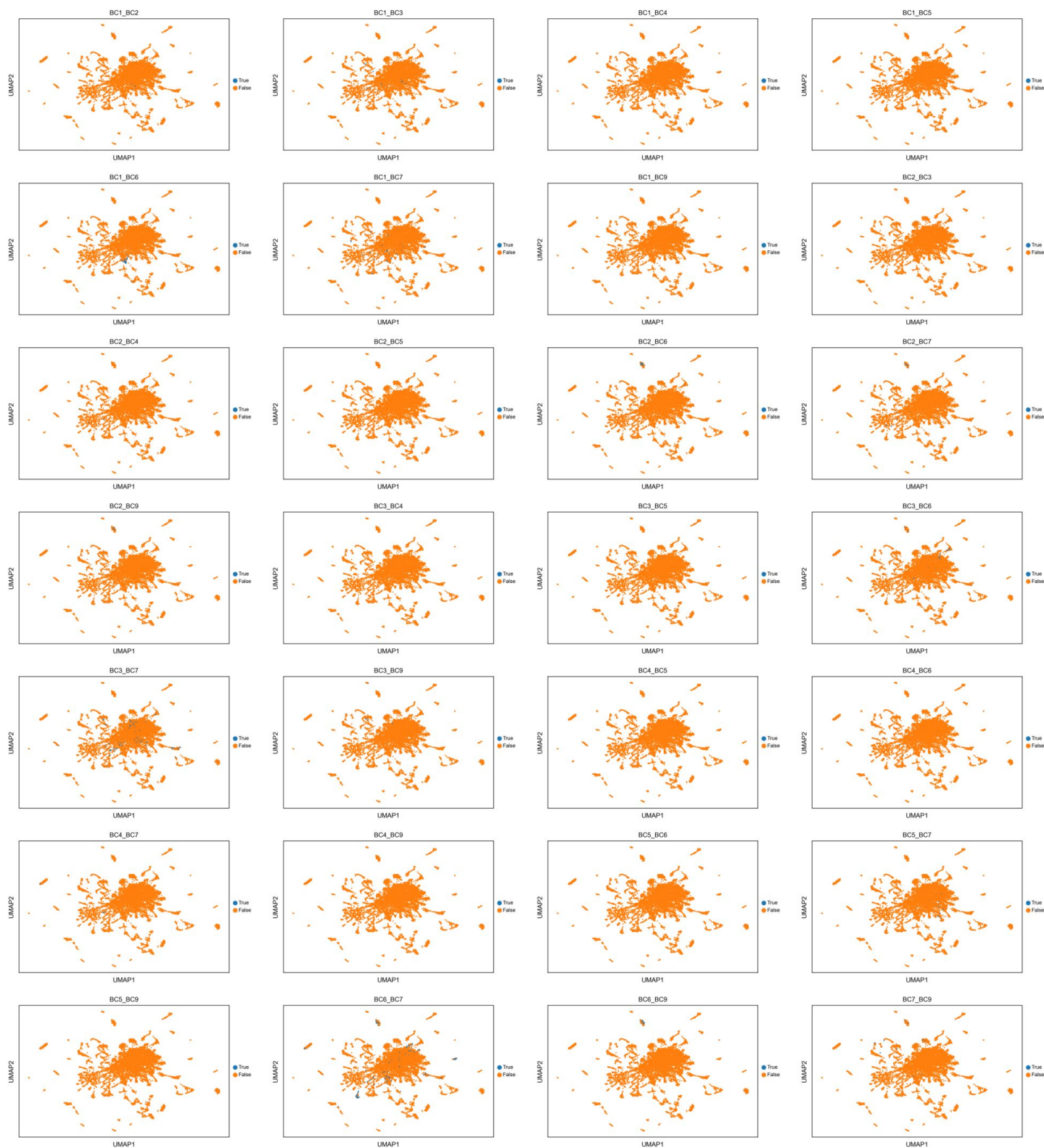
A) GFP expression visualized in dissociated cells from gut driver lines crossed to hexameric GFP and TaG-EM line. B) Proportion of live (left) and dead (right) cells post-isolation and flow sorting as assessed by GFP expression and propidium iodide staining.



Supplemental Figure 9.

Expression of TaG-EM genetic barcodes in larval intestinal precursor cells

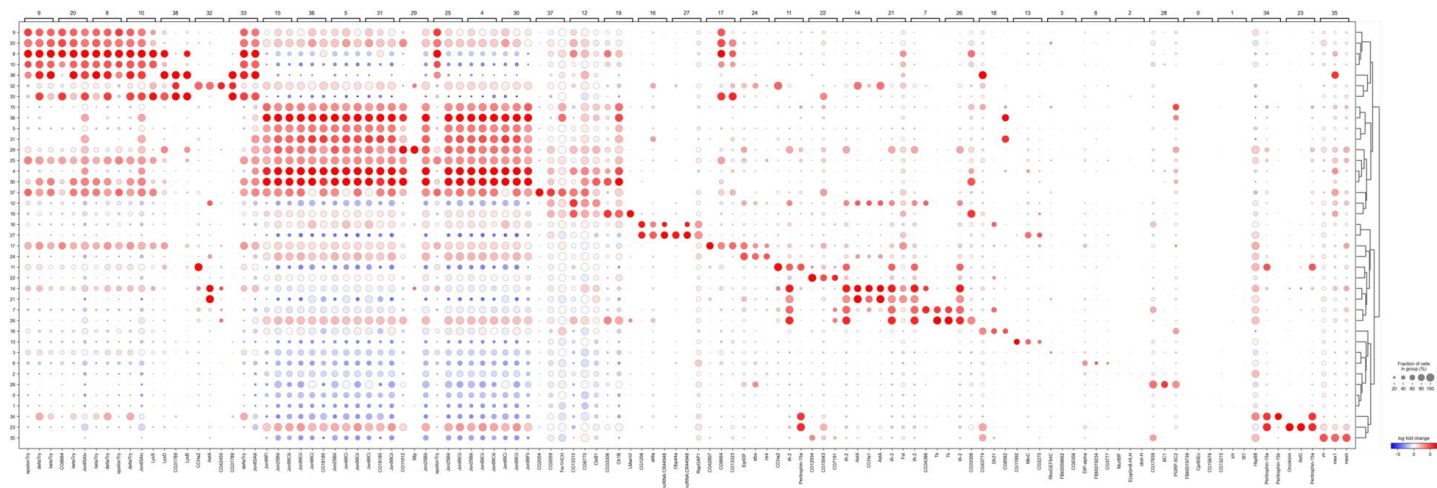
UMAP plots showing gene expression levels of A) enteroblast/ISC marker genes *esg*, *klu*, and *E(spl)mbeta-HLH*; and B) the TaG-EM barcodes 7, 8, and 9 driven by the PC-Gal4 line.



Supplemental Figure 10.

Identification of doublets based on co-expression of TaG-EM barcodes

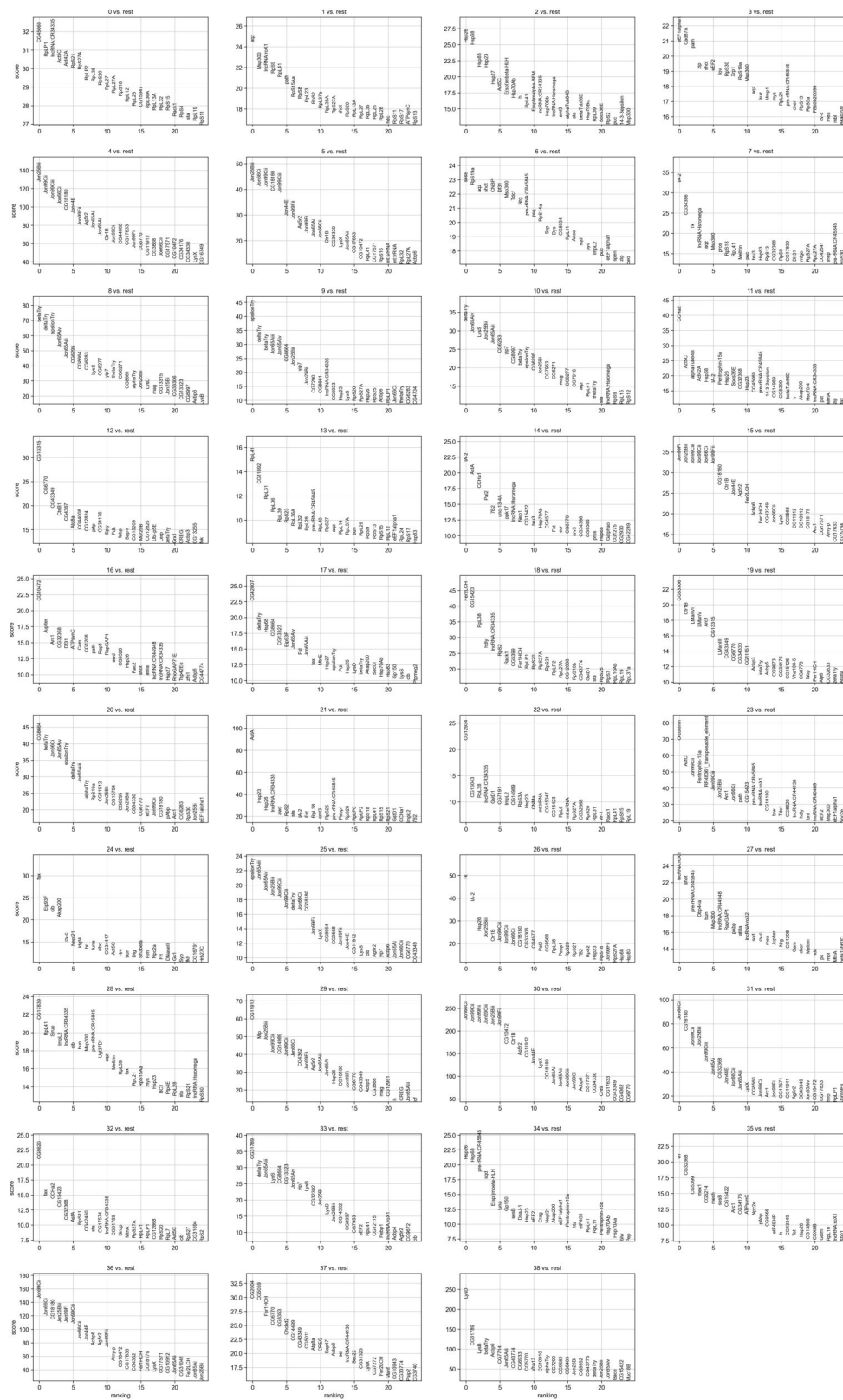
UMAP plots of all 28 pairwise barcode combinations showing cells co-expressing both of the indicated barcodes in blue (True).



Supplemental Figure 11.

Differentially expressed genes

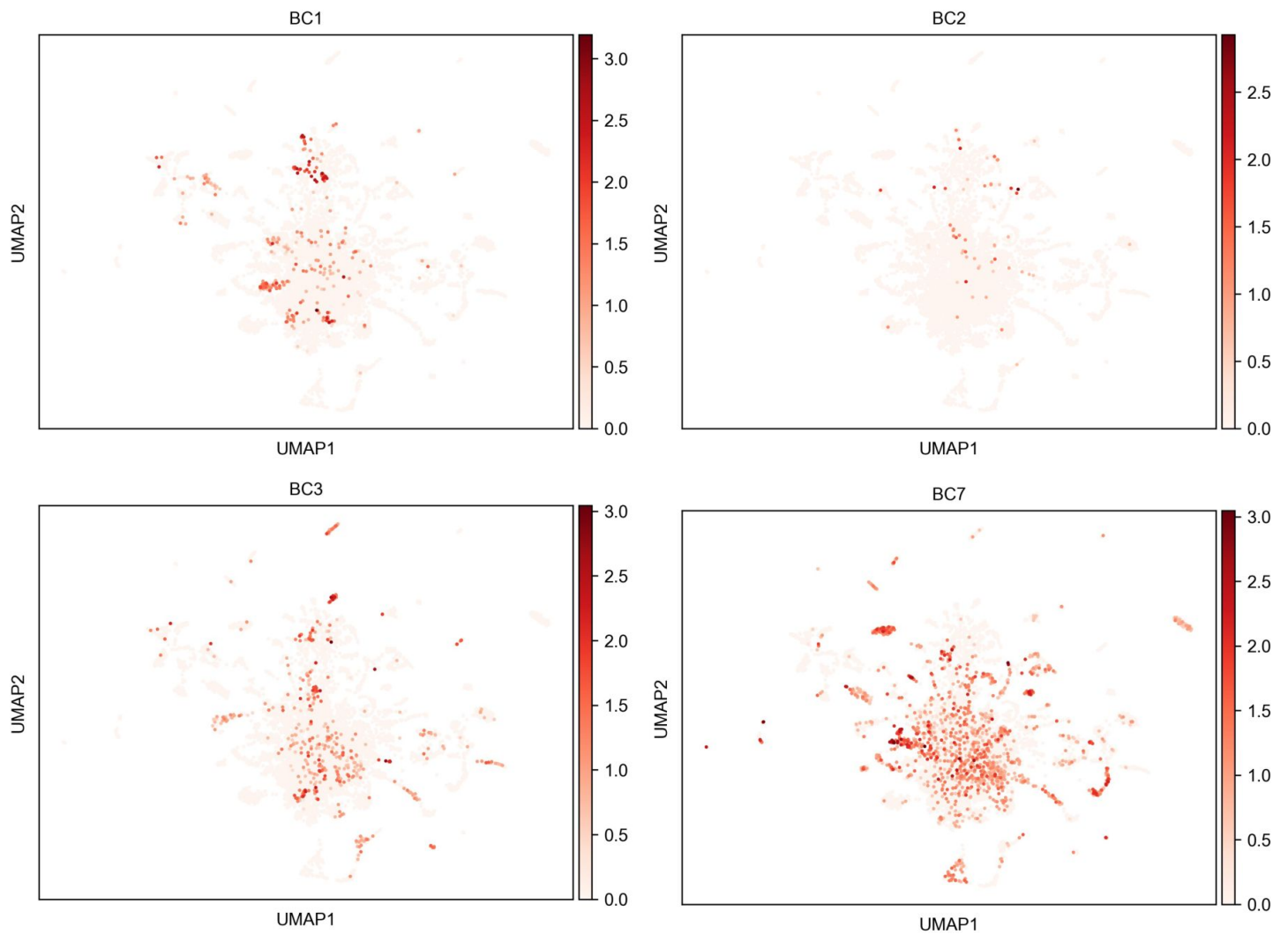
Dotplot showing the top three differentially expressed genes by log-fold change for each cluster.



Supplemental Figure 12.

Top 25 marker genes for each cell cluster

Plots showing t-test-based rankings of the top 25 differentially expressed genes for each cluster.



Supplemental Figure 13.

Expression of the PMG-Gal4 driven TaG-EM barcodes

UMAP plots showing expression of the four PMG-Gal4 driven TaG-EM barcodes (BC1, BC1, BC3, and BC7).

References

- Ariyapala IS, Holsopple JM, Popodi EM, Hartwick DG, Kahsai L, Cook KR, Sokol NS (2020) **Identification of split-GAL4 drivers and enhancers that allow regional cell type manipulations of the drosophila melanogaster intestine** *Genetics* **216**:891–903 <https://doi.org/10.1534/genetics.120.303625>
- Aso Y, Hattori D, Yu Y, Johnston RM, Iyer NA, Ngo TTB, Dionne H, Abbott LF, Axel R, Tanimoto H, Rubin GM (2014) **The neuronal architecture of the mushroom body provides a logic for associative learning** *eLife* **3** <https://doi.org/10.7554/ELIFE.04577>
- Bhang HC, Ruddy DA, Krishnamurthy Radhakrishna V, Caushi JX, Zhao R, Hims MM, Singh AP, Kao I, Rakiec D, Shaw P, Balak M, Raza A, Ackley E, Keen N, Schlabach MR, Palmer M, Leary RJ, Chiang DY, Sellers WR, Michor F, Cooke VG, Korn JM, Stegmeier F (2015) **Studying clonal dynamics in response to cancer therapy using high-complexity barcoding** *Nature Medicine* **21**:440–448 <https://doi.org/10.1038/nm.3841>
- Chen X, Sun Y-C, Zhan H, Kebschull JM, Fischer S, Matho K, Huang ZJ, Gillis J, Zador AM (2019) **High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing** *Cell* **179**:772–786 <https://doi.org/10.1016/j.cell.2019.09.023>
- Cheng J, Liao J, Shao X, Lu X, Fan X (2021) **Multiplexing Methods for Simultaneous Large-Scale Transcriptomic Profiling of Samples at Single-Cell Resolution** *Advanced Science* **8** <https://doi.org/10.1002/advs.202101229>
- Clark IC, Gutiérrez-Vázquez C, Wheeler MA, Li Z, Rothhammer V, Linnerbauer M, Sanmarco LM, Guo L, Blain M, Zandee SEJ, Chao CC, Batterman KV, Schwabenland M, Lotfy P, Tejeda-Velarde A, Hewson P, Polonio CM, Shultis MW, Salem Y, Tjon EC, Fonseca-Castro PH, Borucki DM, de Lima KA, Plasencia A, Abate AR, Rosene DL, Hodgetts KJ, Prinz M, Antel JP, Prat A, Quintana FJ (2021) **Barcoded viral tracing of single-cell interactions in central nervous system inflammation** *Science* **372**
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. (2009) **Biopython: freely available Python tools for computational molecular biology and bioinformatics** *Bioinformatics (Oxford, England)* **25**:1422–3 <https://doi.org/10.1093/bioinformatics/btp163>
- David J, Cohet Y, Fouillet P (1975) **The variability between individuals as a measure of senescence: A study of the number of eggs laid and the percentage of hatched eggs in the case of *Drosophila melanogaster*** *Experimental Gerontology* **10**:17–25 [https://doi.org/10.1016/0531-5565\(75\)90011-X](https://doi.org/10.1016/0531-5565(75)90011-X)
- Davis FP, Nern A, Picard S, Reiser MB, Rubin GM, Eddy SR, Henry GL (2020) **A genetic, genomic, and computational resource for exploring neural circuit function** *eLife* **9**
- DeAngelis MM, Wang DG, Hawkins TL (1995) **Solid-phase reversible immobilization for the isolation of PCR products** *Nucleic Acids Res* **23**:4742–4743
- Gohl DM, Silies MA, Gao XJ, Bhalerao S, Luongo FJ, Lin C-C, Potter CJ, Clandinin TR (2011) **A versatile in vivo system for directed dissection of gene expression patterns** *Nature Methods* **8**:231–237 <https://doi.org/10.1038/nmeth.1561>

- Groth AC, Fish M, Nusse R, Calos MP (2004) **Construction of transgenic *Drosophila* by using the site-specific integrase from phage phiC31** *Genetics* **166**:1775–82 <https://doi.org/10.1534/genetics.166.4.1775>
- Huang AM, Rehm EJ, Rubin GM (2009) **Quick preparation of genomic DNA from *Drosophila*** *Cold Spring Harb Protoc* **2009** <https://doi.org/10.1101/pdb.prot5198>
- Hung RJ, Hu Y, Kirchner R, Liu Y, Xu C, Comjean A, Tattikota SG, Li F, Song W, Sui SH, Perrimon N (2020) **A cell atlas of the adult *Drosophila* midgut** *Proceedings of the National Academy of Sciences of the United States of America* **117**:1514–1523 <https://doi.org/10.1073/pnas.1916820117>
- Ingham PW (1988) **The molecular genetics of embryonic pattern formation in *Drosophila*** *Nature* **335**:25–34 <https://doi.org/10.1038/335025a0>
- Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells** *Cell* **161**:1187–1201 <https://doi.org/10.1016/j.cell.2015.04.044>
- Kurmangaliyev YZ, Yoo J, Valdes-Aleman J, Sanfilippo P, Zipursky SL (2020) **Transcriptional Programs of Circuit Assembly in the *Drosophila* Visual System** *Neuron* **108**:1045–1057 <https://doi.org/10.1016/j.NEURON.2020.10.006>
- Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, Zhang K, Church GM (2015) **Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues** *Nature Protocols* **10**:442–458 <https://doi.org/10.1038/nprot.2014.191>
- Lein E, Borm LE, Linnarsson S (2017) **The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing** *Science* **358**:64–69 <https://doi.org/10.1126/science.aan6827>
- Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA, Brbić M, Spanier K, Leskovec J, McLaughlin CN, Xie Q, Jones RC, Brueckner K, Shim J, Tattikota SG, Schnorrer F, Rust K, Nystul TG, Carvalho-Santos Z, Ribeiro C, Pal S, Mahadevaraju S, Przytycka TM, Allen AM, Goodwin SF, Berry CW, Fuller MT, White-Cooper H, Matunis EL, DiNardo S, Galenza A, O'Brien LE, Dow JAT, Consortium FCA, Jasper H, Oliver B, Perrimon N, Deplancke B, Quake SR, Luo L, Aerts S. (2022) **Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly** *Science* **375** <https://doi.org/10.1126/science.abk2432>
- Ma D, Przybylski D, Abruzzi KC, Schlichting M, Li Q, Long X, Rosbash M (2021) **A transcriptomic taxonomy of *drosophila* circadian neurons around the clock** *eLife* **10**:1–19 <https://doi.org/10.7554/eLife.63056>
- Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets** *Cell* **161**:1202–1214 <https://doi.org/10.1016/j.cell.2015.05.002>
- Martin KC, Ephrussi A (2009) **mRNA localization: gene expression in the spatial dimension** *Cell* **136**:719–30 <https://doi.org/10.1016/j.cell.2009.01.044>
- Martin M (2011) **Cutadapt removes adapter sequences from high-throughput sequencing reads** *EMBnet.journal* **17**:10–12

- Marusyk A, Almendro V, Polyak K (2012) **Intra-tumour heterogeneity: a looking glass for cancer?** *Nature Reviews Cancer* **12**:323–334 <https://doi.org/10.1038/nrc3261>
- McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J (2016) **Whole-organism lineage tracing by combinatorial and cumulative genome editing** *Science* **353** <https://doi.org/10.1126/science.aaf7907>
- Namiki S, Dickinson MH, Wong AM, Korff W, Card GM (2018) **The functional organization of descending sensory-motor pathways in drosophila** *eLife* **7**
- Pfeiffer BD, Jenett A, Hammonds AS, Ngo T-TB, Misra S, Murphy C, Scully A, Carlson JW, Wan KH, Lavery TR, Mungall C, Svirska R, Kadonaga JT, Doe CQ, Eisen MB, Celniker SE, Rubin GM (2008) **Tools for neuroanatomy and neurogenetics in Drosophila** *Proceedings of the National Academy of Sciences of the United States of America* **105**:9715–20 <https://doi.org/10.1073/pnas.0803697105>
- Pfeiffer BD, Ngo T-TB, Hibbard KL, Murphy C, Jenett A, Truman JW, Rubin GM (2010) **Refinement of tools for targeted gene expression in Drosophila** *Genetics* **186**:735–55 <https://doi.org/10.1534/genetics.110.119917>
- Reeves GT, Muratov CB, Schüpbach T, Shvartsman SY (2006) **Quantitative Models of Developmental Pattern Formation** *Developmental Cell* **11**:289–300 <https://doi.org/10.1016/j.DEVCEL.2006.08.006>
- Rubin GM, Spradling AC (1982) **Genetic transformation of Drosophila with transposable element vectors** *Science (New York, NY)* **218**:348–53 <https://doi.org/10.1126/science.6289436>
- Shearin HK, MacDonald IS, Spector LP, Steven Stowers R (2014) **Hexameric GFP and mCherry reporters for the Drosophila GAL4, Q, and LexA transcription systems** *Genetics* **196**:951–960 <https://doi.org/10.1534/GENETICS.113.161141/-/DC1/GENETICS.113.161141-5.PDF>
- Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow C (2009) **Quantitative phenotyping via deep barcode sequencing** *Genome research* **19**:1836–42 <https://doi.org/10.1101/gr.093955.109>
- Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P (2017) **Simultaneous epitope and transcriptome measurement in single cells** *Nature Methods* **14**:865–868 <https://doi.org/10.1038/nmeth.4380>
- Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, Smibert P, Satija R (2018) **Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics** *Genome Biology* **19**:1–12 <https://doi.org/10.1186/s13059-018-1603-1>
- Teleman AA, Cohen SM (2000) **Dpp gradient formation in the drosophila wing imaginal disc** *Cell* **103**:971–980 [https://doi.org/10.1016/S0092-8674\(00\)00199-9](https://doi.org/10.1016/S0092-8674(00)00199-9)
- Thompson CL, Ng L, Menon V, Martinez S, Lee C-K, Glattfelder K, Sunkin SM, Henry A, Lau C, Dang C, Garcia-Lopez R, Martinez-Ferre A, Pombero A, Rubenstein JLR, Wakeman WB, Hohmann J, Dee N, Sodt AJ, Young R, Smith K, Nguyen T-N, Kidney J, Kuan L, Jeromin A, Kaykas A, Miller J, Page D, Orta G, Bernard A, Riley Z, Smith S, Wohnoutka P, Hawrylycz MJ, Puellas L, Jones AR (2014) **A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain** *Neuron* **83**:309–323 <https://doi.org/10.1016/j.NEURON.2014.05.033>

Traag VA, Waltman L, van Eck NJ. (2019) **From Louvain to Leiden: guaranteeing well-connected communities** *Sci Rep* 9 <https://doi.org/10.1038/s41598-019-41695-z>

van Opijnen T, Bodi KL, Camilli A. (2009) **Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms** *Nature methods* 6:767–72 <https://doi.org/10.1038/nmeth.1377>

Venken KJT, Schulze KL, Haelterman NA, Pan H, He Y, Evans-Holm M, Carlson JW, Levis RW, Spradling AC, Hoskins RA, Bellen HJ (2011) **MiMIC: a highly versatile transposon insertion resource for engineering *Drosophila melanogaster* genes** *Nature methods* 8:737–43

Wolf FA, Angerer P, Theis FJ (2018) **SCANPY: large-scale single-cell gene expression data analysis** *Genome Biology* 19 <https://doi.org/10.1186/s13059-017-1382-0>

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH (2017) **Massively parallel digital transcriptional profiling of single cells** *Nature Communications* 8:1–12 <https://doi.org/10.1038/ncomms14049>

Author information

Jorge Blanco Mendana

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA

Margaret Donovan

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA

Lindsey Gengelbach

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA

Benjamin Auch

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA
ORCID iD: [0000-0003-3125-5797](https://orcid.org/0000-0003-3125-5797)

John Garbe

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA

Daryl M. Gohl

University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA, Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455, USA
For correspondence: dmgoehl@umn.edu
ORCID iD: [0000-0002-4434-2788](https://orcid.org/0000-0002-4434-2788)

Editors

Reviewing Editor

Sonia Sen

Tata Institute for Genetics and Society, India

Reviewer #1 (Public Review):

The aim of this paper is to describe a novel method for genetic labelling of animals or cell populations, using a system of DNA/RNA barcodes.

Strengths:

- The author's attempt at providing a straightforward method for multiplexing *Drosophila* samples prior to scRNA-seq is commendable. The perspective of being able to load multiple samples on a 10X Chromium without antibody labelling is appealing.
- The authors are generally honest about potential issues in their method, and areas that would benefit from future improvement.
- The article reads well. Graphs and figures are clear and easy to understand.

Weaknesses:

- The usefulness of TaG-EM for phototaxis, egg laying or fecundity experiments is questionable. The behaviours presented here are all easily quantifiable, either manually or using automated image-based quantification, even when they include a relatively large number of groups and replicates. Despite their claims (e.g., L311-313), the authors do not present any real evidence about the cost- or time-effectiveness of their method in comparison to existing quantification methods.
- Behavioural assays presented in this article have clear outcomes, with large effect sizes, and therefore do not really challenge the efficiency of TaG-EM. By showing a T-maze in Fig 1B, the authors suggest that their method could be used to quantify more complex behaviours. Not exploring this possibility in this manuscript seems like a missed opportunity.
- Experiments in Figs S3 and S6 suggest that some tags have a detrimental effect on certain behaviours or on GFP expression. Whereas the authors rightly acknowledge these issues, they do not investigate their causes. Unfortunately, this questions the overall suitability of TaG-EM, as other barcodes may also affect certain aspects of the animal's physiology or behaviour. Revising barcode design will be crucial to make sure that sequences with potential regulatory function are excluded.
- For their single-cell experiments, the authors have used the 10X Genomics method, which relies on sequencing just a short segment of each transcript (usually 50-250bp - unknown for this study as read length information was not provided) to enable its identification, with the matching paired-end read providing cell barcode and UMI information (Macosko et al., 2015). With average fragment length after tagmentation usually ranging from 300-700bp, a large number of GFP reads will likely not include the 14bp TaG-EM barcode. When a given cell barcode is not associated with any TaG-EM barcode, then demultiplexing is impossible. This is a major problem, which is particularly visible in Figs 5 and S13. In 5F, BC4 is only detected in a couple of dozen cells, even though the *Jon99Ciii* marker of enterocytes is present in a much larger population (Fig 5C). Therefore, in this particular case, TaG-EM fails to detect most of the GFP-expressing cells. Similarly, in S13, most cells should express one of the four barcodes, however many of them (maybe up to half - this should be quantified) do not. Therefore, the claim (L277-278) that "the pan-midgut driver were broadly distributed across the cell clusters" is misleading. Moreover, the hypothesis that "low expressing driver lines may result in particularly sparse labelling" (L331-333) is at least partially wrong, as Fig S13 shows that the same Gal4 driver can lead to very different levels of barcode coverage.
- Comparisons between TaG-EM and other, simpler methods for labelling individual cell populations are missing. For example, how would TaG-EM compare with expression of different fluorescent reporters, or a strategy based on the brainbow/flybow principle?
- FACS data is missing throughout the paper. The authors should include data from their

comparative flow cytometry experiment of TaG-EM cells with or without additional hexameric GFP, as well as FSC/SSC and fluorescence scatter plots for the FACS steps that they performed prior to scRNA-seq, at least in supplementary figures.

- The authors should show the whole data described in L229, including the cluster that they chose to delete. At least, they should provide more information about how many cells were removed. In any case, the fact that their data still contains a large number of debris and dead cells despite sorting out PI negative cells with FACS and filtering low abundance barcodes with Cellranger is concerning.

Overall, although a method for genetic tagging cell populations prior to multiplexing in single-cell experiments would be extremely useful, the method presented here is inadequate. However, despite all the weaknesses listed above, the idea of barcodes expressed specifically in cells of interest deserves more consideration. If the authors manage to improve their design to resolve the major issues and demonstrate the benefits of their method more clearly, then TaG-EM could become an interesting option for certain applications.

Reviewer #2 (Public Review):

In this manuscript, Mendana et al developed a multiplexing method - Targeted Genetically-Encoded Multiplexing or TaG-EM - by inserting a DNA barcode upstream of the polyadenylation site in a Gal4-inducible UAS-GFP construct. This Multiplexing method can be used for population-scale behavioral measurements or can potentially be used in single-cell sequencing experiments to pool flies from different populations. The authors created 20 distinctly barcoded fly lines. First, TaG-EM was used to measure phototaxis and oviposition behaviors. Then, TaG-EM was applied to the fly gut cell types to demonstrate its applications in single-cell RNA-seq for cell type annotation and cell origin retrieving.

This TaG-EM system can be useful for multiplexed behavioral studies from next-generation sequencing (NGS) of pooled samples and for Transcriptomic Studies. I don't have major concerns for the first application, but I think the scRNA-seq part has several major issues and needs to be further optimized.

Major concerns:

1. It seems the barcode detection rate is low according to Fig S9 and Fig 5F, J and N. Could the authors evaluate the detection rate? If the detection rate is too low, it can cause problems when it is used to decode cell types.
2. Unsuccessful amplification of TaG-EM barcodes: The authors attempted to amplify the TaG-EM barcodes in parallel to the gene expression library preparation but encountered difficulties, as the resulting sequencing reads were predominantly off-target. This unsuccessful amplification raises concerns about the reliability and feasibility of this amplification approach, which could affect the detection and analysis of the TaG-EM barcodes in future experiments.
3. For Fig 5, the single-cell clusters are not annotated. It is not clear what cell types are corresponding to which clusters. So, it is difficult to evaluate the accuracy of the assignment of barcodes.
4. The scRNA-seq UMAP in Fig 5 is a bit strange to me. The fly gut epithelium contains only a few major cell types, including ISC, EB, EC, and EE. However, the authors showed 38 clusters in fig 5B. It is true that some cell types, like EE (Guo et al., 2019, Cell Reports), have sub-populations, but I don't expect they will form these many sub-types. There are many peripheral small clusters that are not shown in other gut scRNA-seq studies (Hung et al., 2020; Li et al., 2022 Fly Cell Atlas; Lu et al., 2023 Aging Fly Cell Atlas). I suggest the authors try different data-processing methods to validate their clustering result.
5. Different gut drivers, PMC-, PC-, EB-, EC-, and EE-GAL4, were used. The authors should carefully characterize these GAL4 expression in larval guts and validate sequencing data. For example, does the ratio of each cell type in Fig 5B reflect the in vivo cell type ratio? The

authors used cell-type markers mostly based on the knowledge from adult guts, but there are significant morphological and cell ratio differences between larval and adult guts (e.g., Mathur...Ohlstein, 2010 Science).

6. Doublets are removed based on the co-expression of two barcodes in Fig 5A. However, there are also other possible doublets, for example, from the same barcode cells or when one cell doesn't have detectable barcode. Did the authors try other computational approaches to remove doublets, like DoubleFinder (McGinnis et al., 2019) and Scrublet (Wolock et al., 2019)?

7. Did the authors remove ambient RNA which is a common issue for scRNA-seq experiments?

8. Why does TaG-EM barcode #4, driven by EC-GAL4, not label other classes of enterocyte cells such as betaTry⁺ positive ECs (Figures 5D-E)? similarly, why does TaG-EM barcode #9, driven by EE-GAL4, not label all EEs? Again, it is difficult to evaluate this part without proper data processing and accurate cell type annotation.

9. For Figure 2, when the authors tested different combinations of groups with various numbers of barcodes. They found remarkable consistency for the even groups. Once the numbers start to increase to 64, barcode abundance becomes highly variable (range of 12-18% for both male and female). I think this would be problematic because the differences seen in two groups for example may be due to the barcode selection rather than an actual biologically meaningful difference.

10. Barcode #14 cannot be reliably detected in oviposition experiment. This suggests that the BC 14 fly line might have additional mutations in the *attp2* chromosome arm that affects this behavior. Perhaps other barcode lines also have unknown mutations and would cause issues for other untested behaviors. One possible solution is to back-cross all 20 lines with the same genetic background wild-type flies for >7 generations to make all these lines to have the same (or very similar) genetic background. This strategy is common for aging and behavior assays.

Reviewer #3 (Public Review):

The work addresses challenges in linking anatomical information to transcriptomic data in single-cell sequencing. It proposes a method called Targeted Genetically-Encoded Multiplexing (TaG-EM), which uses genetic barcoding in *Drosophila* to label specific cell populations in vivo. By inserting a DNA barcode near the polyadenylation site in a UAS-GFP construct, cells of interest can be identified during single-cell sequencing. TaG-EM enables various applications, including cell type identification, multiplet droplet detection, and barcoding experimental parameters. The study demonstrates that TaG-EM barcodes can be decoded using next-generation sequencing for large-scale behavioral measurements. Overall, the results are solid in supporting the claims and will be useful for a broader fly community. I have only a few comments below:

Specific comments:

1. The authors mentioned that the results of structure pool tests in Fig. 2 showed a high level of quantitative accuracy in detecting the TaG-EM barcode abundance. Although the data were generally consistent with the input values in most cases, there were some obvious exceptions such as barcode 1 (under-represented) and barcodes 15, 20 (over-represented). It would be great if the authors could comment on these and provide a guideline for choosing the appropriate barcode lines when implementing this TaG-EM method.

2. In Supplemental Figure 6, the authors showed GFP antibody staining data with 20 different TaG-EM barcode lines. The variability in GFP antibody staining results among these different TaG-EM barcode lines concerns the use of these TaG-EM barcode lines for sequencing followed by FACS sorting of native GFP. I expected the native GFP expression would be weaker and much more variable than the GFP antibody staining results shown in Supplemental Figure 6. If this is the case, variation of tissue-specific expression of TaG-EM barcode lines will likely be a confounding factor.

3. As the authors mentioned in the manuscript, multiple barcodes for one experimental condition would be a better experimental design. Could the authors suggest a recommended number of barcodes for each experiential condition? 3? 4? Or more? Also, it would be great if the authors could provide a short discussion on the cost of such TaG-EM method. For example, for the phototaxis assay, if it is much more expensive to perform TaG-EM as compared to manually scoring the preference index by videotaping, what would be the practical considerations or benefits of doing TaG-EM over manual scoring?