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

Jorge Blanco Mendana[1], Margaret Donovan[1], Lindsey G. O'Brien[1], Benjamin Auch[1,†], John Garbe[1], Daryl M. Gohl[1,2,3]

[1] University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA
[2] Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455, USA

[3] Corresponding author: dmgohl@umn.edu
[†] Present address: Phase Genomics, Seattle, WA, 98109, USA

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.

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 studies assessing 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. Outside of descriptive

studies, these costs are also a barrier to including replicates to assess biological variability; consequently, a lack of biological replicates derived from independent samples 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).

Genetically barcoded fly lines can also be used to enable highly multiplexed behavioral assays which can be read out using high throughput sequencing. Flies carrying TaG-EM barcodes can be exposed to different experimental perturbations and then tested in 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 (Figure 1B, 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 1C). 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 1D).

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

147 pools produced consistent data with all 20 barcodes detected at levels close to the expected 5%
148 abundance (Figure 2B-C). Barcode abundance values for the staggered structured pools was
149 generally consistent with the input values and in most cases, the 2-fold differences between the
150 different groups of barcodes could be distinguished (Figure 2D-E). The coefficients of variation
151 were largely consistent for groups of TaG-EM barcodes pooled evenly or at different levels
152 within the staggered pools (Supplemental Figure 3). For the staggered pools, abundances
153 correlated well with the expected values, particularly when multiple barcodes for an input level
154 were averaged, in which case R^2 values were >0.99 (Figure 2D-E, inset plots). This indicates
155 that a high level of quantitative accuracy can be attained using sequencing-based analysis of
156 TaG-EM barcode abundance, particularly when averaging data for three to four independent
157 barcodes for an experimental condition.
158
159 **TaG-EM measurement of phototaxis behavior correlate well with video-based**
160 **measurements**
161 Next, we tested whether TaG-EM could be used to measure a phototaxis behavior. A
162 mixture of barcoded wild type or blind *norpA* mutant flies were run together through a phototaxis
163 assay. At the end of a period of light exposure, test tubes facing toward or away from the light
164 were capped, DNA was isolated, and barcodes were amplified and sequenced for each tube.
165 Raw read counts were scaled in proportion to the number of flies per tube and a preference
166 index was calculated for each barcode (Figure 3A). In parallel, individual preference indices
167 were calculated based on manual scoring of videos recorded for each line (Figure 3B).
168 Preference indices calculated for the pooled, NGS-based TaG-EM measurements were nearly
169 identical to conventional behavioral measurements for both wild type and *norpA* mutants (Figure
170 3A-B).
171
172 **TaG-EM measurement of oviposition behavior and age-dependent fecundity**
173 We next tested whether NGS-based pooled measurements of egg laying could be made.
174 Fertilized females from each of the 20 barcode lines were placed together in egg laying cups,
175 embryos were collected, aged for 12 hours to enable cell numbers to stabilize in the developing
176 eggs, and then DNA was extracted from both the pooled adult flies and the embryos. In general,
177 TaG-EM measurements of oviposition correlated with fly numbers, with the exception of
178 barcode 14 which had reduced barcode abundance across multiple trials (Supplemental Figure
179 4). This suggests that despite the fact that the genetic barcode constructs are inserted in a
180 common landing site, differences with respect to specific behaviors may exist among the lines,

181 and thus one should test to make sure given lines are appropriate to use in specific behavioral
182 assays.
183 To determine whether TaG-EM could be used to measure age-dependent fecundity, we
184 collected flies from twelve different TaG-EM barcode lines at four time points separated by one
185 week (three barcode lines per timepoint). We collected eggs from these fly lines individually and
186 scored the number of viable eggs per female. Next, we pooled the barcoded flies from all
187 timepoints and collected eggs from the pooled flies. These eggs were aged, DNA was
188 extracted, and the TaG-EM barcodes were amplified and sequenced. While measurements from
189 individual barcode lines were noisy, both for manual counts and sequencing based
190 measurements, there was a general trend toward declining fecundity over time (Supplemental
191 Figure 5), consistent with published reports (David et al., 1975). Manually scored viable egg
192 numbers and TaG-EM barcode abundances were well correlated across two independent
193 experimental trials (R^2 values of 0.52-0.61 for Trial 1 and 0.74-0.84 for Trial 2). When barcodes
194 from each individual timepoint were averaged, R^2 values for the correlation between manual and
195 sequencing-based measurements were 0.95 for Trial 1 and 0.99 for Trial 2 (Figure 3C,
196 Supplemental Figure 6).
197
198 **Quantifying food transit time in the larval gut using TaG-EM**
199 Gut motility defects underlie a number of functional gastrointestinal disorders in humans
200 (Keller et al., 2018). To study gut motility in Drosophila, we have developed an assay based on
201 the time it takes a food bolus to transit the larval gut (Figure 4A), similar to approaches that
202 have been employed for studying the role of the microbiome in human gut motility (Asnicar et
203 al., 2021). Third instar larvae were starved for 90 minutes and then fed food containing a blue
204 dye. After 60 minutes, larvae in which a blue bolus of food was visible were transferred to plates
205 containing non-dyed food, and food transit (indicated by loss of the blue food bolus) was scored
206 every 30 minutes for five hours (Supplemental Figure 7).
207 Because this assay is highly labor-intensive and requires hands-on effort for the entire
208 five-hour observation period, there is a limit on how many conditions or replicates can be scored
209 in one session (~8 plates maximum). Thus, we decided to test whether food transit could be
210 quantified in a more streamlined and scalable fashion by using TaG-EM (Figure 4B). Using the
211 manual assay, we observed that while caffeine-containing food is aversive to larvae, the
212 presence of caffeine reduces transit time through the gut (Figure 4C, Supplemental Figure 7).
213 This is consistent with previous observations in adult flies that bitter compounds (including
214 caffeine) activate enteric neurons via serotonin-mediated signaling and promote gut motility

215 (Yao and Scott, 2022). We tested whether TaG-EM could be used to measure the effect of
216 caffeine on food transit time in larvae. As with prior behavioral tests, the TaG-EM data
217 recapitulated the results seen in the manual assay (Figure 4D). Conducting the transit assay via
218 TaG-EM enables several labor-saving steps. First, rather than counting the number of larvae
219 with and without a food bolus at each time point, one simply needs to transfer non-bolus-
220 containing larvae to a collection tube. Second, because the TaG-EM lines are genetically
221 barcoded, all the conditions can be tested at once on a single plate, removing the need to
222 separately count each replicate of each experimental condition. This reduces the hands-on time
223 for the assay to just a few minutes per hour. A summary of the anticipated cost and labor
224 savings for the TaG-EM-based food transit assay is shown in Supplemental Figure 8.
225
226 **Tissue-specific expression of TaG-EM GFP constructs**
227 To facilitate representation of the TaG-EM barcodes in single-cell sequencing data,
228 genetic barcodes were placed just upstream of the polyadenylation signal sequences and polyA
229 cleavage sites (Figure 1A). To verify that the inserted sequences did not interfere with Gal4-
230 driven GFP expression, we crossed each of the barcoded TaG-EM lines to *decapentaplegic-*
231 *Gal4* (*dpp-Gal4*). We observed GFP expression in the expected characteristic central stripe
232 (Teleman and Cohen, 2000) in the wing imaginal disc for 19/20 lines at similar expression levels
233 to the base pJFRC12 *UAS-myr::GFP* construct inserted in the same landing site (Figure 5A,
234 Supplemental Figure 9). No GFP expression was visible for TaG-EM barcode number 8, which
235 upon molecular characterization had an 853 bp deletion within the GFP coding region (data not
236 shown). We generated and tested GFP expression of an additional 156 TaG-EM barcode lines
237 (Alegria et al., 2024), by crossing them to Mhc-Gal4 and observing expression in the adult
238 thorax. All 156 additional TaG-EM lines had robust GFP expression (data not shown). Gal4-
239 driven expression levels of TaG-EM barcoded GFP constructs were also similar to that of the
240 pJFRC12 base construct for multiple driver lines (Supplemental Figure 10) indicating that the
241 presence of the barcode does not generally impair expression of GFP.
242
243 **Boosting the GFP signal of TaG-EM constructs to enable robust cell sorting**
244 While with some driver lines, expression of the myr::GFP from the TaG-EM construct
245 may be too weak to allow robust enrichment of the tagged cells, adding an additional hexameric
246 GFP construct (Shearin et al., 2014) could boost expression of weak driver lines to levels that
247 are sufficient for robust detection of labeled flies or larvae (Figure 5B) and for labeling of
248 dissociated cells for flow cytometry (Supplemental Figure 11). Stocks with an additional UAS

249 hexameric GFP construct recombined onto the same chromosome as the TaG-EM construct
250 have been established for 20 TaG-EM barcode lines.
251
252 **Correlation between expression of TaG-EM barcodes and intestinal cell marker genes in**
253 **single-cell sequencing data**
254 To test whether we could detect TaG-EM barcodes in single-cell sequencing data, we
255 crossed three TaG-EM barcode lines to two different gut Gal4 driver lines (Ariyapala et al.,
256 2020), one expressing in the enterocytes (EC-Gal4: TaG-EM barcodes 1, 2, and 3) and the
257 other in intestinal precursor cells (PC-Gal4: TaG-EM barcodes 7, 8, and 9), which includes stem
258 cells and enteroblasts (EBs). Due to weak GFP expression with the EC-Gal4 driver, we did not
259 see visible GFP positive cells for this driver line. The PC-Gal4 driver line contained an additional
260 UAS-Stinger (2xGFP) construct and expressed GFP at a level sufficient for flow sorting when
261 crossed to the TaG-EM line (Supplemental Figure 12). Larval guts were dissected, dissociated,
262 stained with propidium iodide (PI) to label dead cells, and flow sorted to recover PI negative and
263 GFP positive cells. Approximately 10,000 cells were loaded into a 10x Genomics droplet
264 generator and a single-cell library was prepared and sequenced. Two clusters were observed in
265 the resulting sequencing data, one of which had high read counts from mitochondrial genes
266 suggesting that this cluster consisted of mitochondria, debris, or dead and dying cells. After
267 filtering the cells with high mitochondrial reads, a single cluster remained (Supplemental Figure
268 13). This cluster expressed known intestinal precursor cell markers such as *escargot* (*esg*),
269 *klumpfuss* (*klu*), and *Notch* pathway genes like *E(spl)mbeta-HLH* (Supplemental Figure 13).
270 Expression of all three PC-Gal4 driven TaG-EM barcodes was observed in this cluster
271 (Supplemental Figure 13) indicating that TaG-EM barcodes can be detected in single-cell
272 sequencing data. Interestingly, TaG-EM barcode 8, for which no GFP expression was observed,
273 was represented in the single-cell sequencing data indicating that the lesion in the GFP coding
274 region does not prevent mRNA expression for this line.
275 A previous study used droplet-based single-cell sequencing to characterize the cell
276 types that make up the adult midgut (Hung et al., 2020). This study took advantage of two
277 fluorescent protein markers, an *escargot* (*esg*)-GFP fusion protein and a *prospero* (*pros*)-Gal4
278 driven RFP to label the intestinal stem cells (ISCs) and enteroendocrine cells (EEs),
279 respectively (Hung et al., 2020). The authors compared the resulting clusters to a list of known
280 marker genes in the literature, including antibody staining, GFP, LacZ, and Gal4 reporter
281 expression patterns to classify the cells in individual clusters, and also found that the *esg*-GFP
282 expression was present in a broader subset of cells than anticipated. Thus, most of these cell

283 classifications relied upon inference as opposed to direct positive labeling. Recently, a large
284 collection of split-*Gal4* lines were screened for expression in the adult and larval gut (Ariyapala
285 et al., 2020). These include pan-midgut driver lines, split-*Gal4* lines specific for the EBs, ECs,
286 EEs, and ISC/EBs, as well as driver lines with regionalized gene expression. We crossed four
287 different TaG-EM barcode lines with the pan-midgut driver (PMG-Gal4: TaG-EM barcodes 1, 2,
288 3, and 7), and one barcode line to each of the precursor cell (PC-Gal4: TaG-EM barcode 5),
289 enterocyte (EC-Gal4: TaG-EM barcode 4), enteroblast (EB-Gal4: TaG-EM barcode 6) and
290 enteroendocrine (EE-Gal4: TaG-EM barcode 9) drivers (Ariyapala et al., 2020). Larval guts were
291 dissected from these lines and cells were dissociated, flow sorted as described above to select
292 live, GFP positive cells, and approximately 30,000 cells were loaded into a 10x Genomics
293 droplet generator for single-cell sequencing (Supplemental Figure 14). Using the additional
294 hexameric GFP construct to boost GFP expression resulted in visible fluorescent signal for all
295 eight barcode Gal4 line combinations.
296 An advantage of cell barcoding both for cell hashing (Stoeckius et al., 2018) and for
297 TaG-EM *in vivo* barcoding is that such labeling facilitates the identification and removal of
298 multiplets, which are an artifact of droplet-based single-cell sequencing approaches. After
299 filtering and removing cells with a high percentage of mitochondrial or ribosomal reads, we used
300 DoubletFinder (McGinnis et al., 2019) to computationally identify multiplet droplets. In parallel,
301 we searched for cells that co-expressed multiple TaG-EM barcodes. DoubletFinder identified
302 2,019 multiplet droplets, while TaG-EM barcodes identified 298 such droplets, 198 of which
303 (66.4%) overlapped with those identified by DoubletFinder (Supplemental Figure 15). Thus,
304 TaG-EM help identify an additional 100 doublets that would have otherwise been overlooked
305 using computational doublet identification methods.
306 After doublet removal, the remaining cells were clustered (Figure 6A, Supplemental
307 Figure 16) and analyzed using Seurat (Satija et al., 2015). Analysis of differentially expressed
308 genes identified clusters expressing marker genes previously reported for adult gut cell types
309 (Hung et al., 2020). These included genes associated with precursor cells (*Notch* pathway
310 genes), enterocytes (trypsins, serine proteases, amalyse, mannosidases), and enteroendocrine
311 cells (neuropeptides and neuropeptide receptors) (Supplemental Figure 17, data not shown).
312 TaG-EM barcodes derived from the eight multiplexed genotypes were observed in
313 approximately one-quarter of the cells (Supplemental Figure 17).
314 In antibody-conjugated oligo cell hashing approaches, sparsity of barcode representation
315 is overcome by spiking in an additional primer at the cDNA amplification step and amplifying the
316 hashtag oligo by PCR. We employed a similar approach to attempt to enrich for TaG-EM

317 barcodes in an additional library sequenced separately from the 10x Genomics gene expression
318 library. Our initial attempts at barcode enrichment using spike-in and enrichment primers
319 corresponding to the TaG-EM PCR handle were unsuccessful (Supplemental Figure 18).
320 However, we subsequently optimized the TaG-EM barcode enrichment by 1) using a longer
321 spike-in primer that more closely matches the annealing temperature used during the 10x
322 Genomics cDNA creation step, and 2) using a nested PCR approach to amplify the cell-barcode
323 and unique molecular identifier (UMI)-labeled TaG-EM barcodes (Supplemental Figure 18).
324 Using the enriched library, TaG-EM barcodes were detected in nearly 100% of the cells
325 at high sequencing depths (Supplemental Figure 19). However, although we used a polymerase
326 that has been engineered to have high processivity and that has been shown to reduce the
327 formation of chimeric reads in other contexts (Gohl et al., 2016), it is possible that PCR
328 chimeras could lead to unreliable detection events for some cells. Indeed, many cells had a
329 mixture of barcodes detected with low counts and single or low numbers of associated UMIS.
330 To assess the reliability of detection, we analyzed the correlation between barcodes detected in
331 the gene expression library and the enriched TaG-EM barcode library as a function of the purity
332 of TaG-EM barcode detection for each cell (the percentage of the most abundant detected TaG-
333 EM barcode, Supplemental Figure 19). For TaG-EM barcode detections where the most
334 abundance barcode was a high percentage of the total barcode reads detected (~75%-99.99%),
335 there was a high correlation between the barcode detected in the gene expression library and
336 the enriched TaG-EM barcode library. Below this threshold, the correlation was substantially
337 reduced.
338 In the enriched library, we identified 26.8% of cells with a TaG-EM barcode reliably
339 detected, a very modest improvement over the gene expression library alone (23.96%),
340 indicating that at least for this experiment, the main constraint is sufficient expression of the
341 TaG-EM barcode and not detection. To identify TaG-EM barcodes in the combined data set, we
342 counted a positive detection as any barcode either identified in the gene expression library or
343 any barcode identified in the enriched library with a purity of >75%. In the case of conflicting
344 barcode calls, we assigned the barcode that was detected directly in the gene expression
345 library. This increased the total fraction of cells where a barcode was identified to approximately
346 37% (Figure 6B).
347 As expected, the barcodes expressed by the pan-midgut driver were broadly distributed
348 across the cell clusters (Supplemental Figure 20). However, the number of cells recovered
349 varied significantly among the four pan-midgut driver associated barcodes. Expression of the
350 cell type-specific barcodes showed more restricted patterns of expression among the cell

351 clusters and were co-localized with known marker genes for these cell types (Figure 6C-N). For
352 instance, TaG-EM barcode 6, driven by the EB-Gal4 line, was expressed primarily in cells that
353 were annotated as enteroblasts (Figure 6C-D) and that expressed precursor cell markers such
354 as *esg* (Figure 6E), *klu* (Figure 6F), and Notch pathway genes such as *e(spl)mbeta-HLH* and
355 *e(spl)m3-HLH* (not shown).
356 TaG-EM barcode 4 expression, which was driven by the EC-Gal4 line, was seen
357 primarily in a cluster that was annotated as enterocytes (Figure 6G-H) and that expressed
358 enterocyte markers such as the serine protease *Jon99Ciii* and other enterocyte marker genes
359 such as the amylase, *Amy-d* (not shown), but not the *beta-Trypsin (betaTry)* gene (Figure 6I-J).
360 Detailed characterization of the EC-Gal4 line indicated that although this line labeled a high
361 percentage of enterocytes, expression was restricted to an area at the anterior and middle of
362 the midgut, with gaps between these regions and at the posterior (Supplemental Figure 21).
363 This could explain the absence of subsets of enterocytes, such as those labeled by *betaTry*,
364 which exhibits regional expression in R2 of the adult midgut (Buchon et al., 2013).
365 Finally, expression of TaG-EM barcode 9, which was expressed using the EE-Gal4
366 driver line, was observed in clusters annotated as enteroendocrine cells (Figure 6K-L) and that
367 also expressed enteroendocrine cell derived neuropeptide genes such as *Dh31* (Figure 6M) and
368 other enteroendocrine markers such as *IA-2*, a tyrosine phosphatase involved in the secretion of
369 insulin-like peptide (Figure 6N). Detailed characterization of the EE-Gal4 driver line indicated
370 that ~80-85% of Prospero-positive enteroendocrine cells are labeled in the anterior and middle
371 of the larval midgut, with a lower percentage (~65%) of Prospero-positive cells labeled in the
372 posterior midgut (Supplemental Figure 21). As with the enterocyte labeling, and consistent with
373 the Gal4 driver expression pattern, the EE-Gal4 expressed TaG-EM barcode 9 did not label all
374 classes of enteroendocrine cells and other clusters of presumptive enteroendocrine cells
375 expressing other neuropeptides such as *Orcokinin*, *AstA*, and *AstC*, or neuropeptide receptors
376 such as *CCHa2* (not shown) were also observed. The EE-Gal4 driver uses *Dh31* regulatory
377 elements, so it is not surprising that the TaG-EM barcodes specifically labeled *Dh31*-positive
378 enteroendocrine cells and this result further highlights the ability to target specific genetically
379 defined cell types using TaG-EM based on *in vivo* cell labeling. Taken together, these results
380 demonstrate that TaG-EM can be used to label specific cell populations that correlate with Gal4-
381 driven expression patterns *in vivo* for subsequent identification in single-cell sequencing data.
382

383 **Discussion**

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

388 We demonstrate that the genetic barcodes can be quantified from mixtures of barcoded
389 fly lines using next-generation sequencing. Analysis of structured pools of flies with defined
390 inputs suggests that TaG-EM barcode measurements are highly accurate and reproducible,
391 particularly in cases where multiple barcodes are used to label an experimental condition and
392 averaged (Figure 2). Sequencing-based TaG-EM measurements recapitulated more laborious,
393 one-at-a-time measurements in a phototaxis assay, an age-dependent fecundity assay, and a
394 gut motility assay, demonstrating that TaG-EM can be used to measure behavior or other
395 phenotypes in multiplexed, pooled populations (Figures 3-4). We did note that one line (TaG-EM
396 barcode 14) exhibited poor performance in oviposition assays, suggesting that barcode
397 performance should be verified for a specific assay. We excluded this poor performing barcode
398 line from the fecundity tests, however, backcrossing is often used to bring reagents into a
399 consistent genetic background for behavioral experiments and could also potentially be used to
400 address behavior-specific issues with specific TaG-EM lines. In addition, other strategies such
401 as averaging across multiple barcode lines (3-4 per condition, which yielded R^2 values >0.99 in
402 tests with structured pools) or permutation of barcode assignment across replicates could also
403 mitigate such deficiencies. Currently, up to 176 conditions can be multiplexed in a single pooled
404 experiment with existing TaG-EM lines, but because sequencing indices can be added after
405 amplification in a separate indexing PCR step, many hundreds or even thousands of such
406 experiments can be multiplexed in a single sequencing run. While the utility of TaG-EM
407 barcode-based quantification will vary based on the number of conditions being analyzed and
408 the ease of quantifying the behavior or phenotype by other means, we demonstrate that TaG-
409 EM can be employed to cost-effectively streamline labor-intensive assays and to quantify
410 phenotypes with small effect sizes (Figure 4, Supplemental Figure 8). An additional benefit of
411 multiplexed TaG-EM behavioral measurements is that the experimental conditions are
412 effectively blinded as the multiplexed conditions are intermingled in a single assay.

413 In addition, we show that TaG-EM barcodes can be expressed by tissue-specific Gal4
414 drivers and used to tag specific cell populations upstream of single-cell sequencing (Figures 5-
415 6). This capability will allow for positive identification of cell clusters in cell atlas projects and will
416 facilitate multiplexing of single-cell sequencing experiments. Recently, a conceptually similar

417 approach called RABID-Seq (Clark et al., 2021) has been described, which allows trans-
418 synaptic labeling of neural circuits using barcoded viral transcripts. However, one distinction
419 between the two approaches is that RABID-Seq relies on stochastic viral infection of
420 mammalian cells while TaG-EM allows reproducible targeting of defined cell populations
421 allowing unambiguous cell identification and potentially allowing the same cell populations to be
422 assessed at different timepoints or in the context of different experimental manipulations. One
423 current limitation is that TaG-EM barcodes are not observed in all cells in single-cell gene
424 expression data. It is likely that the strength of the Gal4 driver contributes to the labeling density.
425 However, we also observed variable recovery of TaG-EM barcodes that were all driven by the
426 same pan-midgut Gal4 driver (Supplemental Figure 20). For single cell RNA-Seq experiments,
427 the cost savings of multiplexing is roughly the cost of a run divided by the number of
428 independent lines multiplexed, plus labor savings by also being able to multiplex upstream flow
429 cytometry, minus loss of unbarcoded cells. Our experiments indicated that for the specific
430 drivers we tested TaG-EM barcodes are detected in around one quarter of the cells if relying on
431 endogenous expression in the gene expression library, though this fraction was higher (~37%) if
432 sequencing an enriched TaG-EM barcode library in parallel (Figure 6, Supplemental Figures 18-
433 19).
434 In the future, generation of additional TaG-EM lines will enable higher levels of
435 multiplexing. In addition, while the original TaG-EM lines were made using a membrane-
436 localized myr::GFP construct, variants that express GFP in other cell compartments such as the
437 cytoplasm or nucleus could be constructed to enable increased expression levels or purification
438 of nuclei. Nuclear labeling could also be achieved by co-expressing a nuclear GFP construct
439 with existing TaG-EM lines in analogy to the use of hexameric GFP described above.
440 In summary, combined with the large collections of Gal4 and split-Gal4 lines that have
441 been established in Drosophila that enable precise targeting of a high proportion of cell types
442 (Ariyapala et al., 2020; Aso et al., 2014; Davis et al., 2020; Gohl et al., 2011; Kanca et al., 2022;
443 Namiki et al., 2018; Pfeiffer et al., 2010, 2008; Venken et al., 2011; Zirin et al., 2024), TaG-EM
444 provides a means to target and label cells *in vivo* for subsequent detection in single-cell
445 sequencing. Moreover, these genetic barcodes can be used to multiplex behavioral or other
446 phenotypic measurements. Thus, TaG-EM provides a flexible system for barcoding cells and
447 organisms.
448

449 **Methods**

450

451 **Drosophila stocks and maintenance**

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

454

Stock name	Source
IsoD1	Clandinin Lab, Stanford University
w-; +; + (IsoD1)	Clandinin Lab, Stanford University
attP2 line	BDSC #25710
norpA	BDSC# 9048
UAS-myr::GFP (pJFRC12)	BDSC #32197
Hexameric GFP lines	BDSC #91402, #91403
UAS-6XmCherry-HA	BDSC #52268
UAS-GFP.nls	BDSC #4776
esg-GFP.FPTB	BDSC #83386
dpp-Gal4 driver	BDSC #1553
Act-Gal4 driver	BDSC #4414
Tub-Gal4 driver	BDSC #5138
Mhc-Gal4 driver	BDSC #55132
PC-Gal4 driver	BDSC #73356 + #75528
PC-Gal4 driver (with UAS-Stinger)	BDSC #91400 + #91401
EC-Gal4 driver	BDSC #91406
EB-Gal4 driver	BDSC #91398 + #91404
EE-Gal4 driver	BDSC #91402 + #91403
PMG-Gal4 driver	BDSC #91398 + #91399
TaG-EM lines (x176)	This study, (Alegria et al., 2024)
TaG-EM lines + 6xGFP (x20)	This study

477

478 **Design and cloning of TaG-EM constructs**

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

482 caaaggaaaaagctgcactgctatacaagaaaattatggaaaaatatttgatgtatagtgccttgactagagatcataatcagccata
483 ccacatttgtagaggttttacttgctttaaaaaacctcccacacctccccctgaacctgaaacataaaatgaatgcaattgttgttgttaact
484 tgtttattgcagcttataaCTTCCAACAACCGGAAGTGANNNNNNNNNNNNNNNtggttacaaataaagcaatag
485 catcacaaatttcacaaataaagcatttttttcactg**ca**ttctagttgtggtttgtccaaactcatcaatgtatcttatcatgtctggatcgatct
486 ggccggccgtttaaacgaattcttgaagacgaaagggcctcgtgatacgcctatttttataggttaatgtcatgataataatg
487
488 The gBlock was resuspended in 20 µl EB, incubated at 50ºC for 20 minutes and then cut with
489 *Psi*I and *EcoR*I (New England Biolabs, NEB) using the following reaction conditions: 4 µl gBlock
490 DNA (35 ng), 2 µl 10x CustSmart buffer (NEB), 1 µl *Eco*RI enzyme (NEB), 1 µl *Psi*I enzyme
491 (NEB), and 12 µl nuclease-free water were mixed and incubated at 37ºC for 1 hour followed by
492 65ºC for 20 minutes to heat inactivate the restriction enzymes. pJFRC12-10XUAS-IVS-
493 myr::GFP plasmid (Addgene, Plasmid #26222) (Pfeiffer et al., 2010) was digested with the
494 following reaction conditions: 5 µl pJFRC12-10XUAS-IVS-myr::GFP plasmid DNA (~3 µg), 5 µl
495 10x CutSmart buffer (NEB), 1 µl *Psi*I enzyme (NEB), 1 µl *Eco*RI enzyme (NEB), and 38 µl
496 nuclease-free water, were mixed and incubated at 37ºC for 1 hour, followed by addition of 1 µl
497 of CIP and incubation for an additional 30 minutes. The digested vector backbone was gel
498 purified using the QiaQuick Gel Purification Kit (Qiagen). The digested gBlock was ligated into
499 the digested pJFRC12-10XUAS-IVS-myr::GFP backbone using the following reactions
500 conditions: 4 µl T4 ligase buffer (10x) (NEB), 20 µl plasmid backbone DNA (0.005 pmol), 5 µl
501 gBlock digest DNA (0.03 pmol), 2 µl of T4 DNA ligase (NEB), and 9 µl nuclease-free water were
502 mixed and incubated at 22ºC for 2 hours. 2 µl of the ligation reaction was transformed into 50 µl
503 of TOP10 competent cells (Invitrogen), and the cells were incubated on ice for 30 minutes, then
504 heat shocked at 42ºC for 30 seconds, and incubated on ice for 5 minutes. 250 µl SOC was
505 added and the cells were plated on LB+ampicillin plates and incubated overnight at 37ºC. DNA
506 was isolated from 36 pJFRC12-gBlock colonies using a QIAprep Spin MiniPrep kit (Qiagen).
507 Expected construct size was verified by diagnostic digest with *Eco*RI and *Apa*LI. DNA
508 concentration was determined using a Quant-iT PicoGreen dsDNA assay (Thermo Fisher
509 Scientific) and the randomer barcode for each of the constructs was determined by Sanger
510 sequencing using the following primers:
511 SV40_post_R: GCCAGATCGATCCAGACATGA
512 SV40_5F: CTCCCCCTGAACCTGAAACA
513
514 **Generation of TaG-EM transgenic lines**

515 29 sequence verified constructs were normalized, pooled evenly, and injected as a pool into
516 embryos (Rubin and Spradling, 1982) expressing PhiC31 integrase and the carrying the *attP2*
517 landing site (BDSC #25710). Injected flies were outcrossed to w- flies, and up to three *white*+
518 progeny per cross were identified, and the transgenic lines were homozygosed. DNA was
519 extracted (GeneJET genomic DNA purification Kit, Thermo Scientific) and the region containing
520 the DNA barcode was amplified with the following PCR reaction: 2.5 µl 1:10 diluted template
521 DNA, 2 µl 10x Reaction Buffer (Qiagen), 0.2 µl dNTP mix (10 µM), 1 µl 10 µM SV40_5F primer
522 (10 µM), 1 µl SV40_post_R primer (10 µM), 0.8 µl $MgCl_2$ (3 mM), 0.1 µl Taq polymerase
523 (Qiagen), 12.4 µl nuclease-free water. Reactions were amplified using the following cycling
524 conditions: 95ºC for 5 minutes, followed by 30 cycles of 94ºC for 30 seconds, 55ºC for 30
525 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes. PCR products were treated with
526 Exo-CIP using the following reaction conditions: 5 µl PCR product, 1 µl Exo-CIP Tube A (NEB),
527 1 µl Exo-CIP Tube B (NEB) were mixed and incubated at 37ºC for 4 minutes, followed by 80ºC
528 for 1 minute. The barcode sequence for each of the independent transgenic lines was
529 determined by Sanger sequencing using the SV40_5F and SV40_PostR primers. Transgenic
530 lines containing 20 distinct DNA barcodes were recovered (Supplemental Figure 1). An
531 additional 156 TaG-EM barcode lines were isolated and sequence verified as described in a
532 separate publication (Alegria et al., 2024).
533
534 **Optimizing amplification of TaG-EM barcodes for next-generation sequencing**
535 The following primers were evaluated to amplify the TaG-EM barcodes upstream of NGS:
536 Forward primer pool; four primers with frameshifting bases to increase library sequence
537 diversity in initial sequencing cycles were normalized to 10 µM and pooled evenly to make a
538 B2_3'F1_Nextera_0-6 primer pool:
539 B2_3'F1_Nextera:
540 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCCAACAACCGGAAG*TGA
541 B2_3'F1_Nextera_2
542 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCTTCCAACAACCGGAAG*TGA
543 B2_3'F1_Nextera_4
544 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGACTTCCAACAACCGGAAG*TGA
545 B2_3'F1_Nextera_6
546 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGAGCTTCCAACAACCGGAAG*TGA
547 The following two reverse primers were tested:

548 SV40_pre_R_Nextera:
549 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTTGTGAAATTTGTGATGCTATTGC*T
550 TT
551 SV40_post_R_Nextera:
552 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAGATCGATCCAGACA*TGA
553 SV40_pre_R_nextera is designed to produce a shorter amplicon (200 bp with Illumina adapters
554 and indices added) and SV40-post_R_Nextera is designed to produce a longer amplicon (290
555 bp with Illumina adapters and indices added).
556
557 An initial test was performed with three different polymerases (Qiagen Q5, KAPA HiFi, and
558 Qiagen Taq) at two different annealing temperatures and with both the B2_3'F1_Nextera/
559 SV40_pre_R_Nextera and B2_3'F1_Nextera/ SV40_post_R_Nextera primer sets to determine
560 whether the primers amplify as expected (Supplemental Figure 1). Two different samples were
561 tested:
562 1. Pool of 8 putative transformant samples (pooled 5 ul each of 1:10 diluted sample)
563 2. OreR (wild type - diluted 1:10)
564 Set up the following PCR reactions:
565 Q5 polymerase: 2.5 µl template DNA, 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer
566 (10 µM), 10 µl 2x Q5 Master Mix (Qiagen), 5.5 µl nuclease-free water. Reactions were amplified
567 using the following cycling conditions: 98ºC for 30 seconds, followed by 30 cycles of 98ºC for 20
568 seconds, 55ºC or 60ºC for 15 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes.
569 KAPA HiFi polymerase: 2.5 µl template DNA, 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse
570 primer (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free water. Reactions
571 were amplified using the following cycling conditions: 95ºC for 5 minutes, followed by 30 cycles
572 of 98ºC for 20 seconds, 55ºC or 60ºC for 15 seconds, 72ºC for 30 seconds, followed by 72ºC for
573 5 minutes.
574 Taq polymerase: 2.5 µl template DNA, 2 µl 10x Reaction Buffer (Qiagen), 0.2 µl dNTP mix (10
575 µM), 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer (10 µM), 0.8 µl $MgCl_2$ (3 mM), 0.1
576 µl Taq polymerase (Qiagen), 12.4 µl nuclease-free water. Reactions were amplified using the
577 following cycling conditions: 95ºC for 5 minutes, followed by 30 cycles of 94ºC for 30 seconds,
578 55ºC or 60ºC for 30 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes.
579 Samples were run on a 2% agarose gel to verify amplification products (Supplemental Figure 1).
580

581 Next, the TaG-EM barcode lines were pooled in either an even or staggered manner. To
582 optimize reaction conditions for the barcode measurements, either 5 ng or 50 ng of DNA was
583 amplified in triplicate for each pool for either 20, 25, or 30 cycles with either KAPA HiFi using the
584 B2_Nextera_F 0-6 forward primer pool together with either the SV40_pre_R_Nextera or the
585 SV40_post_R_Nextera reverse primer. Next, PCR reactions were diluted 1:100 in nuclease-free
586 water and amplified in the following indexing reactions: 3 µl PCR 1 (1:100 dilution), 1 µl indexing
587 primer 1 (5 µM), 1 µl indexing primer 2 (5 µM), and 5 µl 2x Q5 master mix. The following
588 indexing primers were used (X indicates the positions of the 8 bp indices):
589 Forward indexing primer:
590 AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC
591 Reverse indexing primer:
592 CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTCTCGTGGGCTCGG
593 Reactions were amplified using the following cycling conditions: 98°C for 30 seconds, followed
594 by 10 cycles of 98°C for 20 seconds, 55°C for 15 seconds, 72°C for 1 minute, followed by 72°C
595 for 5 minutes. Amplicons were then purified and normalized using a SequalPrep normalization
596 plate (Thermo Fisher Scientific), followed by elution in 20 µl of elution buffer. An even volume of
597 the normalized libraries was pooled and concentrated using 1.8x AmpureXP beads (Beckman
598 Coulter). Pooled libraries were quantified using a Qubit dsDNA high sensitivity assay (Thermo
599 Fisher Scientific) and libraries were normalized to 2 nM for sequencing on the Illumina MiSeq
600 (see below).
601
602 **Structured fly pool experiments**
603 Male or female flies from TaG-EM barcode lines were pooled in either an even or staggered
604 manner (Figure 2A). For the even pools, three independently pooled samples were constructed
605 in order to assess sample-to-sample variability. DNA was extracted from these structured pools
606 using a protocol adapted from Huang *et al*., 2009 (Huang et al., 2009), using homemade SPRI
607 beads (DeAngelis et al., 1995) in the last purification step and amplified in triplicate using 2.5 µl
608 template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl
609 SV40_pre_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free
610 water. Reactions were amplified using the following cycling conditions: 98ºC for 5 minutes,
611 followed by 30 cycles of 98ºC for 20 seconds, 60ºC for 15 seconds, 72ºC for 30 seconds,
612 followed by 72ºC for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared
613 for sequencing as described above.
614

615 **Phototaxis experiments**

616 Video-based measurements

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

630

631 TaG-EM measurements

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

634 Pool A

635 norpA/Y;TaG-EM BC4/+
636 norpA/Y;TaG-EM BC3/+
637 w-/Y;TaG-EM BC2/+
638 w-/Y;TaG-EM BC1/+

639

640 Pool B

641 norpA/Y;TaG-EM BC2/+
642 norpA/Y;TaG-EM BC1/+
643 w-/Y;TaG-EM BC4/+
644 w-/Y;TaG-EM BC3/+

645

646 These pools were individually incubated at 34ºC for 10 minutes and then run in the phototaxis
647 assay. Videos of all tests were recorded and at the end of a 30 second light pulse the two test
648 tubes were quickly separated and capped. Flies in each of these tubes were counted, then DNA

649 was extracted from the flies from the light-facing or dark-facing tubes and amplified using 2.5 µl
650 template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl
651 SV40_pre_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free
652 water. Reactions were amplified using the following cycling conditions: 95ºC for 5 minutes,
653 followed by 30 cycles of 98ºC for 20 seconds, 60ºC for 15 seconds, 72ºC for 30 seconds,
654 followed by 72ºC for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared
655 for sequencing as described above.
656
657 **Oviposition experiments**
658 Newly hatched flies (males and females) from three barcode lines were collected at one-week
659 intervals during 4 consecutive weeks (12 barcode lines in total). Fresh fly food was provided
660 every 3-4 days. Ten days after the last collection, 10 females from each barcode line were taken
661 and pooled together in a collection cage (10 females x 12 barcode lines= 120 females). The
662 remaining females from each barcode line were separated from the males and put in individual
663 collection cages. Two days later, the experiment started and run for 3 consecutive days. Each
664 day a 1-1.5 hour pre-collection was followed by a 6 hour collection, both at 25ºC. 100 embryos
665 from each individual collection plate were transferred to new plates and incubated for 2 days at
666 18ºC. The number of hatched larvae were counted and used to calculate the egg survival rate.
667 The pooled collection plate was also incubated at 18ºC and the next day the embryos were
668 dechorionated and frozen. The 12 individual collection plates were kept at 4ºC and the number
669 of embryos counted in the following days. For the barcode measurements, DNA was extracted
670 from the embryos, and amplified using 2.5 µl template DNA (50 ng), 1 µl 10 µM B2_Nextera_F
671 0-6 primers (10 µM), 1 µl SV40_pre_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix
672 (Roche), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling
673 conditions: 95ºC for 5 minutes, followed by 30 cycles of 98ºC for 20 seconds, 60ºC for 15
674 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes. Amplicons were indexed,
675 normalized, quantified, and prepared for sequencing as described above.
676
677 **Larval gut motility experiments**
678 <u>Preparing Yeast Food Plates</u>
679 Yeast agar plates were prepared by making a solution containing 20% Red Star Active Dry
680 Yeast 32oz (Red Star Yeast) and 2.4% Agar Powder/Flakes (Fisher) and a separate solution
681 containing 20% Glucose (Sigma-Aldrich). Both mixtures were autoclaved with a 45-minute liquid
682 cycle and then transferred to a water bath at 55ºC. After cooling to 55ºC, the solutions were

683 combined and mixed, and approximately 5 mL of the combined solution was transferred into
684 100 x 15 mm petri dishes (VWR) in a PCR hood or contamination-free area. For blue-dyed
685 yeast food plates, 0.4% Blue Food Color (McCormick) was added to the yeast solution. For the
686 caffeine assays, 300 µL of a solution of 100 mM 99% pure caffeine (Sigma-Aldrich) was
687 pipetted onto the blue-dyed yeast plate and allowed to absorb into the food during the 90-minute
688 starvation period.
689
690 Manual Gut Motility Assay
691 Third instar Drosophila larvae were transferred to empty conical tubes that had been misted with
692 water to prevent the larvae from drying out. After a 90-minute starvation period the larvae were
693 moved from the conical to a blue-dyed yeast plate with or without caffeine and allowed to feed
694 for 60 minutes. Following the feeding period, the larvae were transferred to an undyed yeast
695 plate. Larvae were scored for the presence or absence of a food bolus every 30 minutes over a
696 5-hour period. Up to 8 experimental replicates/conditions were scored simultaneously.
697
698 TaG-EM Gut Motility Assay
699 Third instar larvae were starved and fed blue dye-containing food with or without caffeine as
700 described above. An equal number of larvae from each experimental condition/replicate were
701 transferred to an undyed yeast plate. During the 5-hour observation period, larvae were
702 examined every 30 minutes and larvae lacking a food bolus were transferred to a
703 microcentrifuge tube labeled for the timepoint. Any larvae that died during the experiment were
704 placed in a separate microcentrifuge tube and any larvae that failed to pass the food bolus were
705 transferred to a microcentrifuge tube at the end of the experiment. DNA was extracted from the
706 larvae in each tube and TaG-EM barcode libraries were prepared and sequenced as described
707 above.
708
709 **Dissection and immunostaining**
710 Midguts from third instar larvae of driver lines crossed to UAS-GFP.nls or UAS-mCherry were
711 dissected in 1xPBS and fixed with 4% paraformaldehyde (PFA) overnight at 4ºC. Fixed samples
712 were washed with 0.1% PBTx (1xPBS + 0.1% Triton X-100) three times for 10 minutes each
713 and blocked in PBTxGS (0.1% PBTx + 3% Normal Goat Serum) for 2–4 hours at RT. After
714 blocking, midguts were incubated in primary antibody solution overnight at 4ºC. The next day
715 samples were washed with 0.1% PBTx three times for 20 minutes each and were incubated in
716 secondary antibody solution for 2–3 hours at RT (protected from light) followed by three washes

717 with 0.1% PBTx for 20 minutes each. One µg/ml DAPI solution prepared in 0.1% PBTx was
718 added to the sample and incubated for 10 minutes followed by washing with 0.1% PBTx three
719 times for 10 minutes each. Finally, samples were mounted on a slide glass with 70% glycerol
720 and imaged using a Nikon AX R confocal microscope. Confocal images were processed using
721 Fiji software.
722
723 The primary antibodies used were rabbit anti-GFP (A6455,1:1000 Invitrogen), mouse anti-
724 mCherry (3A11, 1:20 DSHB), mouse anti-Prospero (MR1A, 1:50 DSHB) and mouse anti-Pdm1
725 (Nub 2D4, 1:30 DSHB). The secondary antibodies used were goat anti-mouse and goat anti-
726 rabbit IgG conjugated to Alexa 647 and Alexa 488 (1:200) (Invitrogen), respectively. Five larval
727 gut specimens per Gal4 line were dissected and examined.
728
729 **Cell dissociation and isolation**
730 Midguts from 3rd instar larvae were dissected in phosphate-buffered saline (PBS) and
731 transferred to microcentrifuge tubes on ice containing PBS + 30% normal goat serum (NGS).
732 After dissection, 150 µL of 2.7 mg/mL elastase was added to each sample tube. The tubes were
733 then incubated at 27°C for one hour. During incubation, samples were mixed by pipetting ~30
734 times every 15 minutes to improve elastase dissociation of the cells. Samples were then filtered
735 through a 40 µM FlowMi tip filter (Bel-Art) to reduce debris. Afterwards, the samples were
736 quantified on the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems) using 9 µL of
737 sample to 1µL AO/PI dye to ensure there were enough viable cells for flow sorting.
738
739 Once quantified, the samples were brought up to a volume of ~1.1mL with the PBS + 30% NGS
740 solution to facilitate flow sorting. The samples were then fluorescently sorted on a FACSAria II
741 Cell Sorter (BD Biosciences) to isolate GFP+ cells. Following sorting, samples were centrifuged
742 at 300x g for 10 minutes to concentrate the cells. The supernatant was aspirated off until 50 µL
743 cell concentrate remained in each sample. Then, the samples were carefully resuspended using
744 wide bore pipette tips before being combined into one sample tube. This sample was quantified
745 on the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems) as described above. If
746 necessary, cells were centrifuged, concentrated, and re-counted.
747
748 **Preparation of single-cell sequencing libraries**
749 The resulting pool was prepared for sequencing following the 10x Genomics Single Cell 3'
750 protocol (version CG000315 Rev C), At step 2.2 of the protocol, cDNA amplification, 1 µl of

751 TaG-EM spike-in primer (10 μM) was added to the reaction to amplify cDNA with the TaG-EM
752 barcode. Gene expression cDNA and TaG-EM cDNA were separated using a double-sided
753 SPRIselect (Beckman Coulter) bead clean up following 10x Genomics Single Cell 3' Feature
754 Barcode protocol, step 2.3 (version CG000317 Rev E). The gene expression cDNA was created
755 into a library following the CG000315 Rev C protocol starting at section 3. Custom nested
756 primers were used for enrichment of TaG-EM barcodes after cDNA creation using PCR.
757 The following primers were tested (see Supplemental Figure 18):
758 UMGC_IL_TaGEM_SpikeIn_v1:
759 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTCCAACAACCGGAAGT*G*A
760 UMGC_IL_TaGEM_SpikeIn_v2:
761 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCAGCTTATAACTTCCAACAACCGGAA
762 GT*G*A
763 UMGC_IL_TaGEM_SpikeIn_v3:
764 TGTGCTCTTCCGATCTGCAGCTTATAACTTCCAACAACCGGAAGT*G*A
765 D701_TaGEM:
766 CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTC
767 CGATCTGCAGC*T*T
768 SI PCR Primer: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC*T*C
769 UMGC_IL_DoubleNest:
770 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCAGCTTATAACTTCCAACAACCGG*A*
771 A
772 P5: AATGATACGGCGACCACCGA
773 D701:
774 GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACTCGATCTCGTATGCCGTCTTCTG
775 CTTG
776 D702:
777 GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCGGAGAATCTCGTATGCCGTCTTCT
778 GCTTG
779 After multiple optimization trials, the following steps yielded ~96% on-target reads for the TaG-
780 EM library (Supplemental Figure 18, note that for the enriched barcode data shown in Figure 6
781 and Supplemental Figure 19, a similar amplification protocol was used TaG-EM barcodes were
782 amplified from the gene expression library cDNA and not the SPRI-selected barcode pool).
783 TaG-EM cDNA was amplified with the following PCR reaction: 5 μl purified TaG-EM cDNA, 50
784 μl 2x KAPA HiFi ReadyMix (Roche), 2.5 μl UMGC_IL_DoubleNest primer (10 μM), 2.5 μl

785 SI_PCR primer (10 µM), and 40 µl nuclease-free water. The reaction was amplified using the
786 following cycling conditions: 98ºC for 2 minutes, followed by 15 cycles of 98ºC for 20 seconds,
787 63ºC for 30 seconds, 72ºC for 20 seconds, followed by 72ºC for 5 minutes. After the first PCR,
788 the amplified cDNA was purified with a 1.2x SPRIselect (Beckman Coulter) bead cleanup with
789 80% ethanol washes and eluted into 40 µL of nuclease-water. A second round of PCR was run
790 with following reaction: 5 µl purified TaG-EM cDNA, 50 µl 2x KAPA HiFi ReadyMix (Roche), 2.5
791 µl D702 primer (10 µM), 2.5 µl p5 Primer (10 µM), and 40 µl nuclease-free water. The reaction
792 was amplified using the following cycling conditions: 98ºC for 2 minutes, followed by 10 cycles
793 of 98ºC for 20 seconds, 63ºC for 30 seconds, 72ºC for 20 seconds, followed by 72ºC for 5
794 minutes. After the second PCR, the amplified cDNA was purified with a 1.2x SPRIselect
795 (Beckman Coulter) bead cleanup with 80% ethanol washes and eluted into 40uL of nuclease-
796 water. The resulting 3' gene expression library and TaG-EM enrichment library were sequenced
797 together following Scenario 1 of the BioLegend "Total-Seq-A Antibodies and Cell Hashing with
798 10x Single Cell 3' Reagents Kit v3 or v3.1" protocol. Additional sequencing of the enriched TaG-
799 EM library also done following Scenario 2 from the same protocol.
800
801 **Sequencing**
802 Libraries for TaG-EM barcode analysis from structured pools or from phototaxis or oviposition
803 experiments were denatured with NaOH and prepared for sequencing according to the
804 protocols described in the Illumina MiSeq Denature and Dilute Libraries Guides. Single-cell
805 libraries were sequenced on the Illumina NextSeq 2000 or Illumina NovaSeq 6000. One of the
806 single cell enriched TaG-EM barcode libraries was sequenced on an Element Aviti sequencer
807 following the manufacturers loading instructions.
808
809 **Data analysis**
810 Behavioral experiments: Demultiplexed fastq files were generated using bcl2fastq or bcl-
811 convert. TaG-EM barcode data was analyzed using custom R and Python scripts and BioPython
812 (Cock et al., 2009). Leading primer sequences were trimmed using cutadapt (Martin, 2011) and
813 the first 14 bp of the remaining trimmed read were compared to a barcode reference file, with a
814 maximum of 2 mismatches allowed, using a custom script (TaG-EM_barcode_analysis.py)
815 which is available via Github: https://github.com/darylgohl/TaG-EM.
816
817 Single-cell experiments:

818 Datasets were first mapped and analyzed using the Cell Ranger analysis pipeline (10x
819 Genomics). A custom Drosophila genome reference was made by combining the BDGP.28
820 reference genome assembly and Ensembl gene annotations. Custom gene definitions for each
821 of the TaG-EM barcodes were added to the fasta genome file and .gtf gene annotation file. A
822 Cell Ranger reference package was generated with the Cell Ranger *mkref*
823 command. Subsequent single-cell data analysis was performed using the R package Seurat
824 (Satija et al., 2015). Cells expressing less than 200 genes and genes expressed in fewer than
825 three cells were filtered from the expression matrix. Next, percent mitochondrial reads, percent
826 ribosomal reads cells counts, and cell features were graphed to determine optimal filtering
827 parameters. DecontX (Yang et al., 2020) was used to identify empty droplets, to evaluate
828 ambient RNA contamination, and to remove empty cells and cells with high ambient RNA
829 expression. DoubletFinder (McGinnis et al., 2019) to identify droplet multiplets and remove cells
830 classified as multiplets. Clustree (Zappia and Oshlack, 2018) was used to visualize different
831 clustering resolutions and to determine the optimal clustering resolution for downstream
832 analysis. Finally, SingleR (Aran et al., 2019) was used for automated cell annotation with a gut
833 single-cell reference from the Fly Cell Atlas (Li et al., 2022). The dataset was manually
834 annotated using the expression patterns of marker genes known to be associated with cell types
835 of interest. To correlate TaG-EM barcodes with cell IDs in the enriched TaG-EM barcode library,
836 a custom Python script was used (TaG-EM_barcode_Cell_barcode_correlation.py), which is
837 available via Github: https://github.com/darylgohl/TaG-EM.
838

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 20 of the TaG-EM barcodes together with an additional UAS hexameric GFP expression construct will be available from the Bloomington Drosophila Stock Center. Additional TaG-EM barcode stocks are available upon request. Single cell analysis code and the TaG-EM barcode analysis script and barcode reference fasta files 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, analyzed data, 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 Thomas Clandinin, Nam Chul Kim, 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. Antibodies were obtained from the Developmental Studies Hybridoma Bank (DSHB). Prospero (MR1A) was deposited to the DSHB by C.Q. Doe and Nub 2D4 was deposited to the DSHB by Michalis Averof.

References

Alegria AD, Joshi AS, Mendana JB, Khosla K, Smith KT, Auch B, Donovan M, Bischof J, Gohl DM, Kodandaramaiah SB. 2024. High-throughput genetic manipulation of multicellular organisms using a machine-vision guided embryonic microinjection robot. *Genetics* **226**:iyae025. doi:10.1093/genetics/iyae025

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**:163–172. doi:10.1038/s41590-018-0276-y

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. doi:10.1534/genetics.120.303625

Asnicar F, Leeming ER, Dimidi E, Mazidi M, Franks PW, Al Khatib H, Valdes AM, Davies R, Bakker E, Francis L, Chan A, Gibson R, Hadjigeorgiou G, Wolf J, Spector TD, Segata N, Berry SE. 2021. Blue poo: impact of gut transit time on the gut microbiome using a novel marker. *Gut* **70**:1665–1674. doi:10.1136/gutjnl-2020-323877

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**:e04577. doi: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. doi:10.1038/nm.3841

Buchon N, Osman D, David FPA, Yu Fang H, Boquete J-P, Deplancke B, Lemaitre B. 2013. Morphological and Molecular Characterization of Adult Midgut Compartmentalization in *Drosophila*. *Cell Reports* **3**:1725–1738. doi:10.1016/j.celrep.2013.04.001

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.e19. doi: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**:2101229. doi: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. doi: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. doi: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**.

929 DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the
930 isolation of PCR products. *Nucleic Acids Res* **23**:4742–4743.
931 Gohl DM, Silies MA, Gao XJ, Bhalerao S, Luongo FJ, Lin C-C, Potter CJ, Clandinin TR. 2011. A
932 versatile in vivo system for directed dissection of gene expression patterns. *Nature*
933 *Methods* **8**:231–237. doi:10.1038/nmeth.1561
934 Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson
935 TJ, Hunter R, Knights D, Beckman KB. 2016. Systematic improvement of amplicon
936 marker gene methods for increased accuracy in microbiome studies. *Nature*
937 *Biotechnology* **34**:942–949. doi:10.1038/nbt.3601
938 Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of transgenic Drosophila by using
939 the site-specific integrase from phage phiC31. *Genetics* **166**:1775–82.
940 doi:10.1534/genetics.166.4.1775
941 Hristova K, Wimley WC. 2023. Determining the statistical significance of the difference between
942 arbitrary curves: A spreadsheet method. *PLoS One* **18**:e0289619.
943 doi:10.1371/journal.pone.0289619
944 Huang AM, Rehm EJ, Rubin GM. 2009. Quick preparation of genomic DNA from Drosophila.
945 *Cold Spring Harb Protoc* **2009**:pdb.prot5198. doi:10.1101/pdb.prot5198
946 Hung RJ, Hu Y, Kirchner R, Liu Y, Xu C, Comjean A, Tattikota SG, Li F, Song W, Sui SH,
947 Perrimon N. 2020. A cell atlas of the adult Drosophila midgut. *Proceedings of the*
948 *National Academy of Sciences of the United States of America* **117**:1514–1523.
949 doi:10.1073/pnas.1916820117
950 Ingham PW. 1988. The molecular genetics of embryonic pattern formation in Drosophila. *Nature*
951 **335**:25–34. doi:10.1038/335025a0
952 Kanca O, Zirin J, Hu Y, Tepe B, Dutta D, Lin W-W, Ma L, Ge M, Zuo Z, Liu L-P, Levis RW,
953 Perrimon N, Bellen HJ. 2022. An expanded toolkit for Drosophila gene tagging using
954 synthesized homology donor constructs for CRISPR-mediated homologous
955 recombination. *eLife* **11**:e76077. doi:10.7554/eLife.76077
956 Keller J, Bassotti G, Clarke J, Dinning P, Fox M, Grover M, Hellström PM, Ke M, Layer P,
957 Malagelada C, Parkman HP, Scott SM, Tack J, Simren M, Törnblom H, Camilleri M,
958 International Working Group for Disorders of Gastrointestinal Motility and Function.
959 2018. Expert consensus document: Advances in the diagnosis and classification of
960 gastric and intestinal motility disorders. *Nat Rev Gastroenterol Hepatol* **15**:291–308.
961 doi:10.1038/nrgastro.2018.7
962 Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner
963 MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem
964 cells. *Cell* **161**:1187–1201. doi:10.1016/j.cell.2015.04.044
965 Kurmangaliyev YZ, Yoo J, Valdes-Aleman J, Sanfilippo P, Zipursky SL. 2020. Transcriptional
966 Programs of Circuit Assembly in the Drosophila Visual System. *Neuron* **108**:1045-
967 1057.e6. doi:10.1016/J.NEURON.2020.10.006
968 Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL,
969 Lee HS, Aach J, Zhang K, Church GM. 2015. Fluorescent in situ sequencing (FISSEQ)
970 of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*
971 **10**:442–458. doi:10.1038/nprot.2014.191
972 Lein E, Borm LE, Linnarsson S. 2017. The promise of spatial transcriptomics for neuroscience
973 in the era of molecular cell typing. *Science* **358**:64–69. doi:10.1126/science.aan6827
974 Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA,
975 Brbić M, Spanier K, Leskovec J, McLaughlin CN, Xie Q, Jones RC, Brueckner K, Shim J,
976 Tattikota SG, Schnorrer F, Rust K, Nystul TG, Carvalho-Santos Z, Ribeiro C, Pal S,
977 Mahadevaraju S, Przytycka TM, Allen AM, Goodwin SF, Berry CW, Fuller MT, White-
978 Cooper H, Matunis EL, DiNardo S, Galenza A, O'Brien LE, Dow JAT, FCA Consortium,
979 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**:eabk2432. doi: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. doi: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. doi:10.1016/j.cell.2015.05.002

Martin KC, Ephrussi A. 2009. mRNA localization: gene expression in the spatial dimension. *Cell* **136**:719–30. doi: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. doi:10.1038/nrc3261

McGinnis CS, Murrow LM, Gartner ZJ. 2019. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems* **8**:329-337.e4. doi:10.1016/j.cels.2019.03.003

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**:aaf7907. doi: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, Laverty TR, Mungall C, Svirskas 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. doi: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. doi: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. doi: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. doi:10.1126/science.6289436

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**:495–502. doi:10.1038/nbt.3192

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. doi: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. doi: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. doi: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. doi:10.1186/s13059-018-1603-1

Teleman AA, Cohen SM. 2000. Dpp gradient formation in the drosophila wing imaginal disc. *Cell* **103**:971–980. doi: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, Puelles L, Jones AR. 2014. A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. *Neuron* **83**:309–323. doi:10.1016/J.NEURON.2014.05.033

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. doi: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.

Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, Campbell JD. 2020. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology* **21**:57. doi:10.1186/s13059-020-1950-6

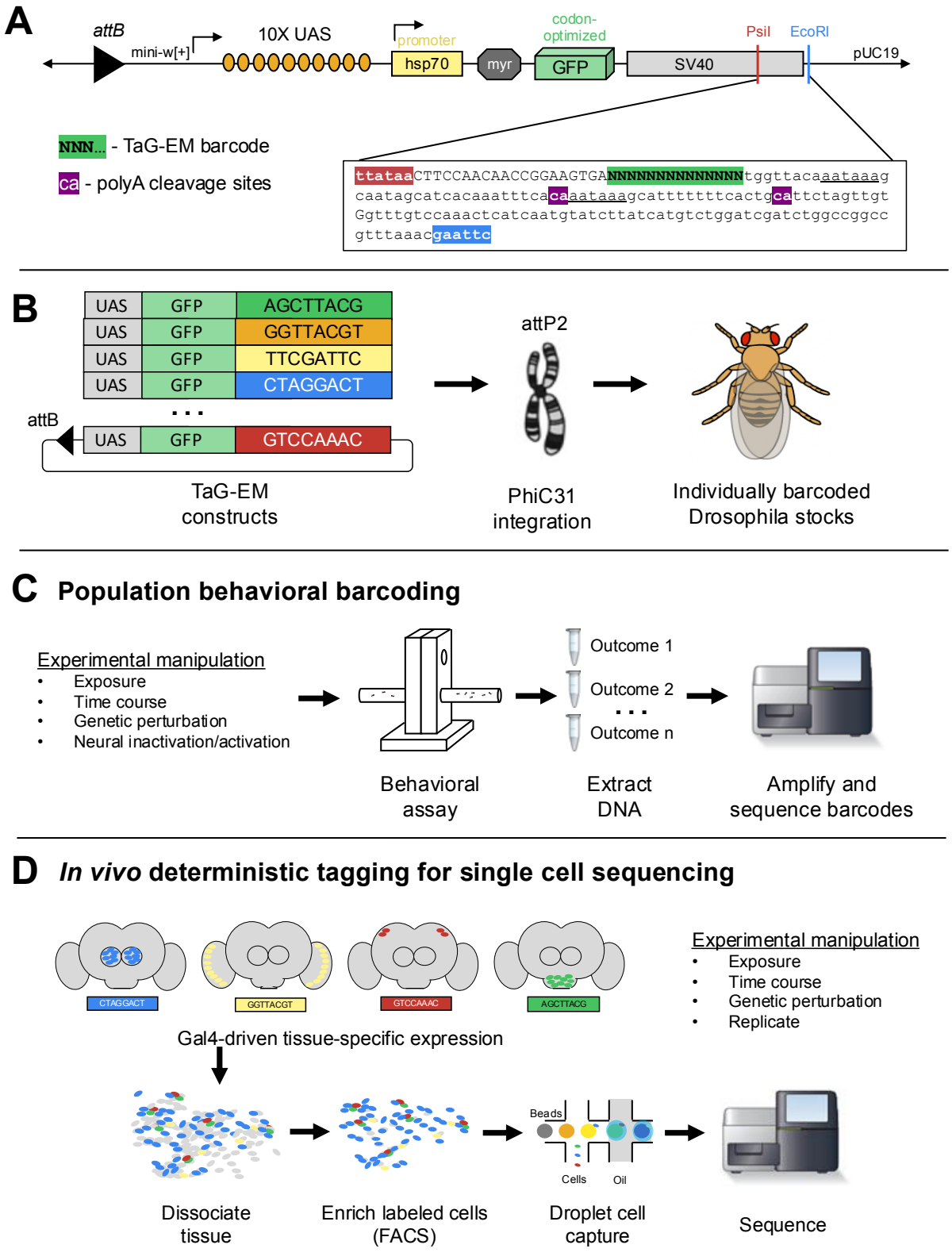
Yao Z, Scott K. 2022. Serotonergic neurons translate taste detection into internal nutrient regulation. *Neuron* **110**:1036-1050.e7. doi:10.1016/j.neuron.2021.12.028

Zappia L, Oshlack A. 2018. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**:giy083. doi:10.1093/gigascience/giy083

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. doi:10.1038/ncomms14049

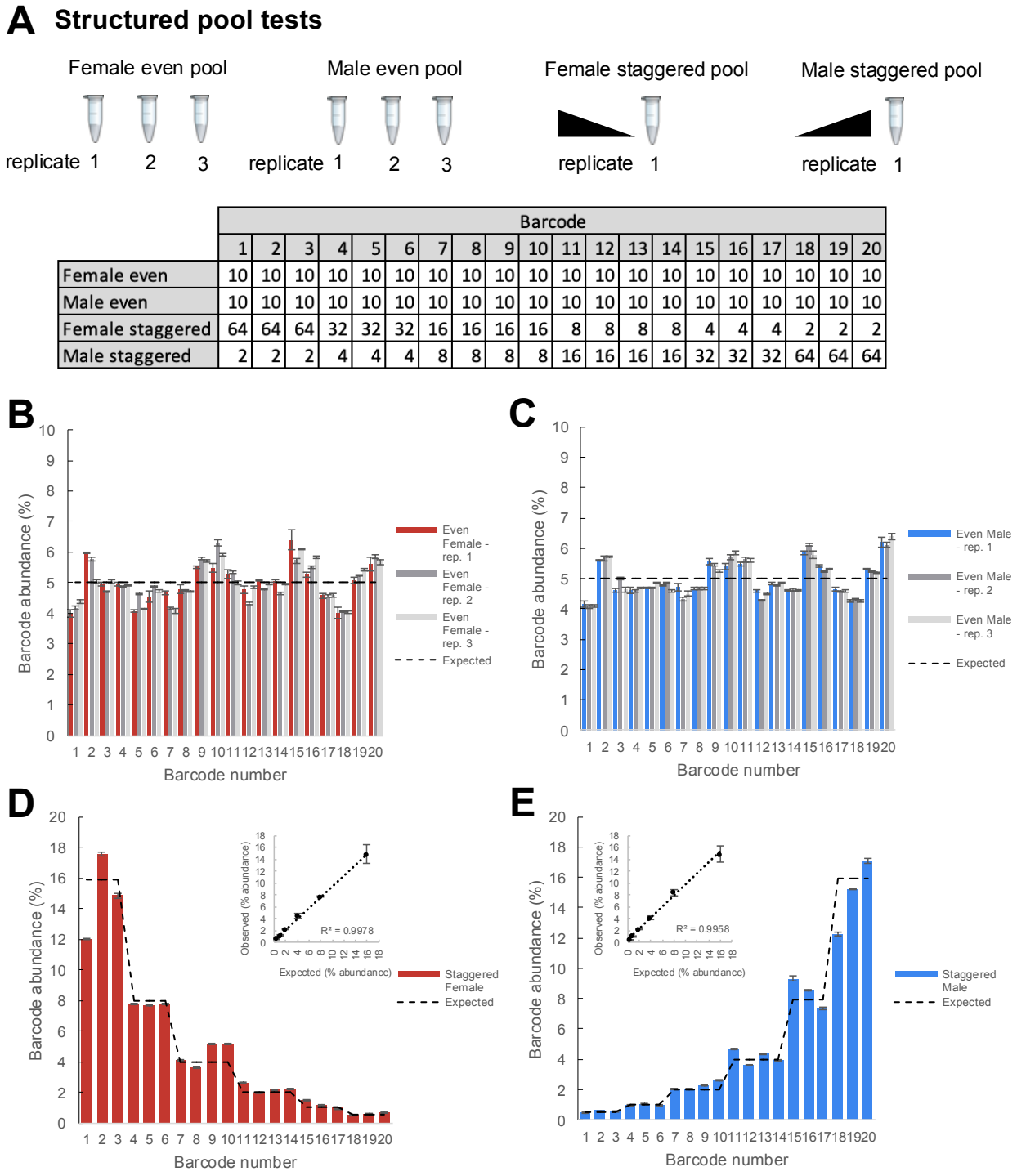
Zirin J, Jusiak B, Lopes R, Ewen-Campen B, Bosch JA, Risbeck A, Forman C, Villalta C, Hu Y, Perrimon N. 2024. Expanding the Drosophila toolkit for dual control of gene expression. *eLife* **12**:RP94073. doi:10.7554/eLife.94073

1066 **Figures**



1067

1068 **Figure 1. Overview of TaG-EM system.** A) Detailed view of the 3' UTR of the TaG-EM
1069 constructs showing the position of the 14 bp barcode sequence (green highlight) relative to the
1070 polyadenylation signal sequences (underlined) and polyA cleavage sites (purple highlights).
1071 pJFRC12 backbone schematic is modified from Pfeiffer *et al.*, 2010. B) Schematic illustrating
1072 the design of the TaG-EM constructs, where a barcode sequence is inserted in the 3' UTR of a
1073 UAS-GFP construct and inserted in a specific genomic locus using PhiC31 integrase. C) Use of
1074 TaG-EM barcodes for sequencing-based population behavioral assays. D) Use of TaG-EM
1075 barcodes expressed with tissue-specific Gal4 drivers to label cell populations *in vivo* upstream
1076 of cell isolation and single-cell sequencing.
1077



	Barcode																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Female even	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Male even	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Female staggered	64	64	64	32	32	32	16	16	16	16	8	8	8	8	4	4	4	2	2	2
Male staggered	2	2	2	4	4	4	8	8	8	8	16	16	16	16	32	32	32	64	64	64

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

1084 replicates of the female even pool. C) Barcode abundance data for three independent replicates
1085 of the male even pool. D) Barcode abundance data for the female staggered pool. Inset plot
1086 shows the average observed barcode abundance among lines pooled at each level compared
1087 to the expected abundance. E) Barcode abundance data for the male staggered pool. Inset plot
1088 shows the average observed barcode abundance among lines pooled at each level compared
1089 to the expected abundance. For all plots, bars indicate the mean barcode abundance for three
1090 technical replicates of each pool, error bars are +/- S.E.M.

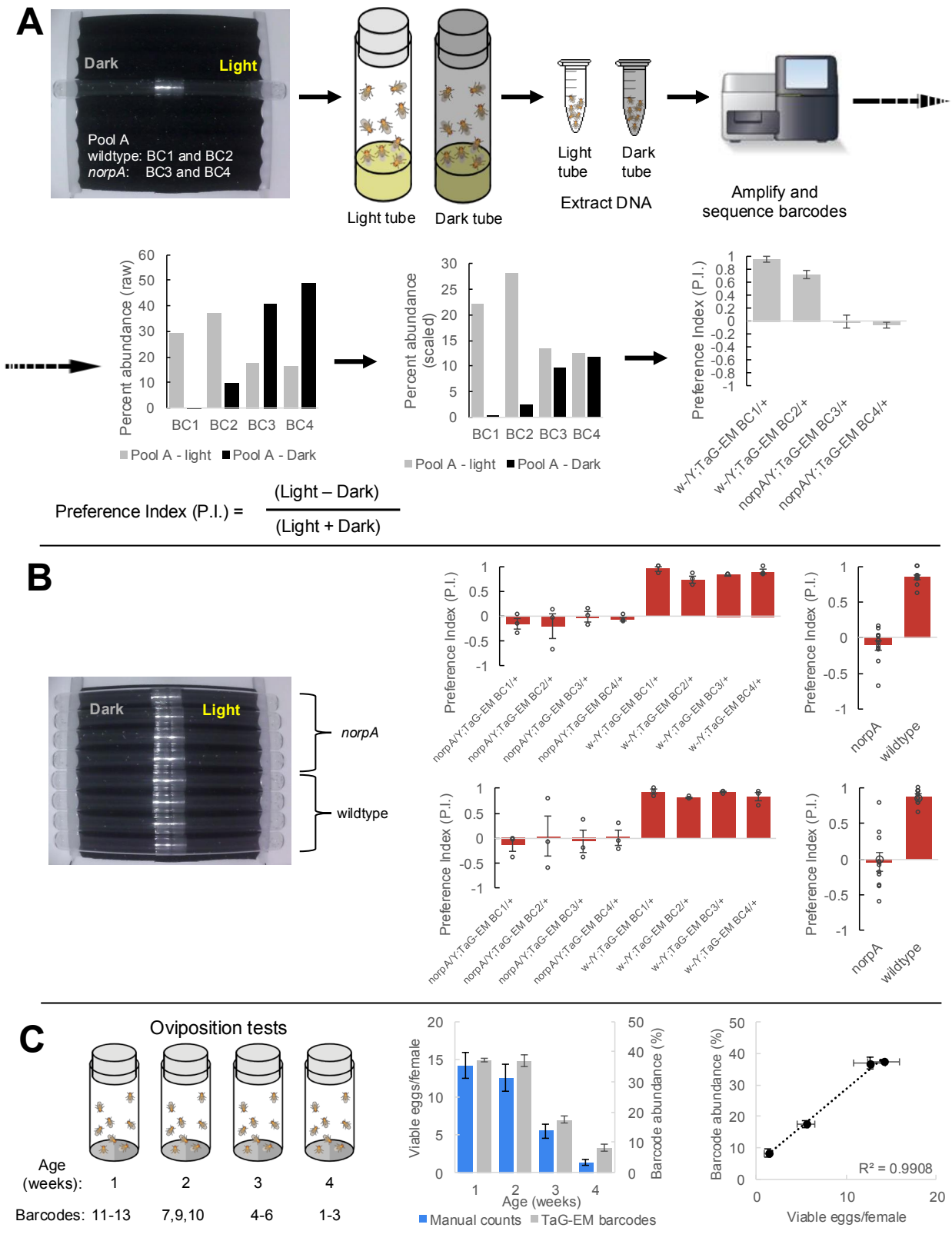


Figure 3. TaG-EM barcode-based behavioral measurements. A) TaG-EM barcode lines in either a wild-type or *norpA* background were pooled and tested in a phototaxis assay. After 30

1094 seconds of light exposure, flies in tubes facing the light or dark side of the chamber were
1095 collected, DNA was extracted, and TaG-EM barcodes were amplified and sequenced. Barcode
1096 abundance values were scaled to the number of flies in each tube and used to calculate a
1097 preference index (P.I.). Average P.I. values for four different TaG-EM barcode lines in both the
1098 wild-type and *norpA* backgrounds are shown (n=3 biological replicates, error bars are +/-
1099 S.E.M.). B) The same eight lines used for the sequencing-based TaG-EM barcode
1100 measurements were independently tested in the phototaxis assay and manually scored videos
1101 were used to calculate a P.I. for each genotype. Average P.I. values for each line are shown
1102 (n=3 biological replicates, error bars are +/- S.E.M.) for TaG-EM-based quantification (top) and
1103 manual video-based quantification (bottom). C) Flies carrying different TaG-EM barcodes were
1104 collected and aged for one to four weeks and then eggs were collected and egg number and
1105 viability was manually scored for each line. In parallel the barcoded flies from each timepoint
1106 were pooled, and eggs were collected, aged, and DNA was extracted, followed by TaG-EM
1107 barcode amplification and sequencing. Average number of viable eggs per female (manual
1108 counts) and average barcode abundance are shown both as a bar plot and scatter plot (n=3
1109 biological replicates for 3 barcodes per condition, error bars are +/- S.E.M.).
1110

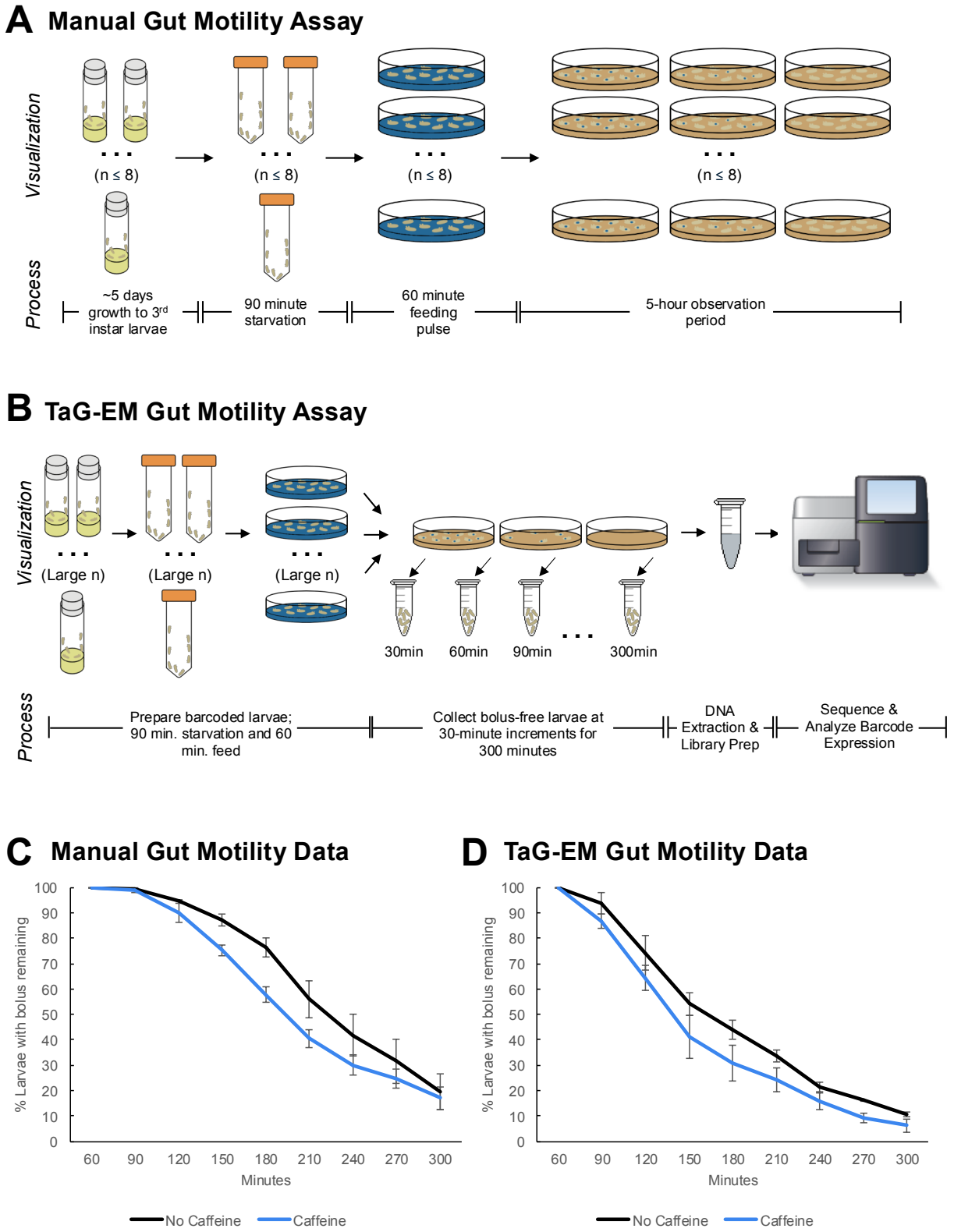


Figure 4. TaG-EM barcode-based quantification of larval gut motility. Schematics depicting A) manual and B) TaG-EM-based assays for quantifying food transit time in Drosophila larvae.

1114 C) Transit time of a food bolus in the presence and absence of caffeine measured using the
1115 manual assay ($p = 0.0340$). D) Transit time of a food bolus in the presence and absence of
1116 caffeine measured using the TaG-EM assay ($p = 0.0488$). A modified Chi-squared method was
1117 used for statistical testing (Hristova and Wimley, 2023).

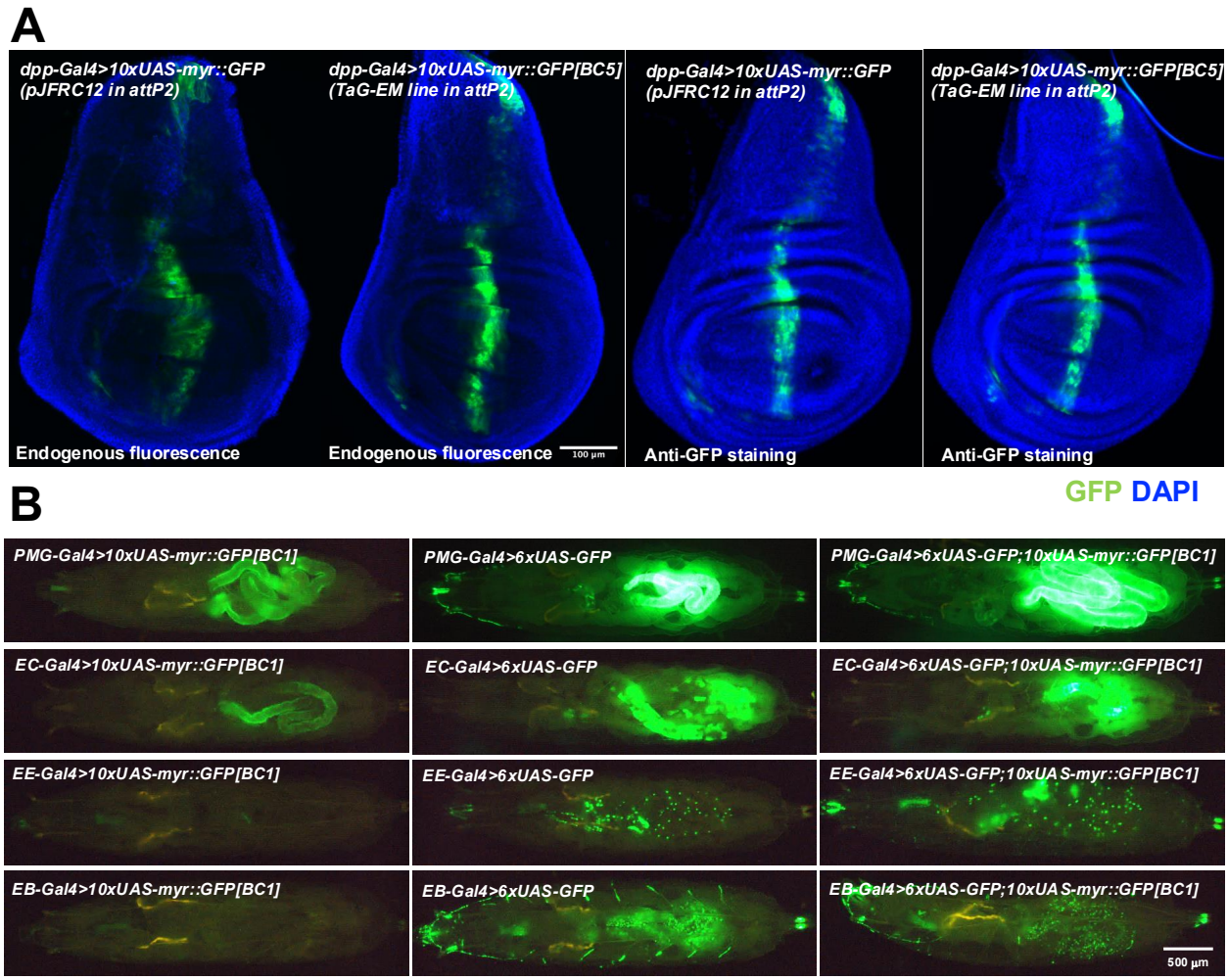


Figure 5. Gal4 driven expression of GFP from TaG-EM lines. A) Comparison of endogenous GFP expression and GFP antibody staining in the wing imaginal disc for the original pJFRC12 construct inserted in the attP2 landing site or for a TaG-EM barcode line driven by *dpp-Gal4*. Wing discs are counterstained with DAPI. B) Endogenous expression of GFP from either a TaG-EM barcode construct (left column), a hexameric GFP construct (middle column), or a line carrying both a TaG-EM barcode construct and a hexameric GFP construct (right column) driven by the indicated gut driver line (PMG-Gal4: Pan-midgut driver; EC-Gal4: Enterocyte driver; EE-Gal4: Enteroendocrine driver; EB-Gal4: Enteroblast driver).

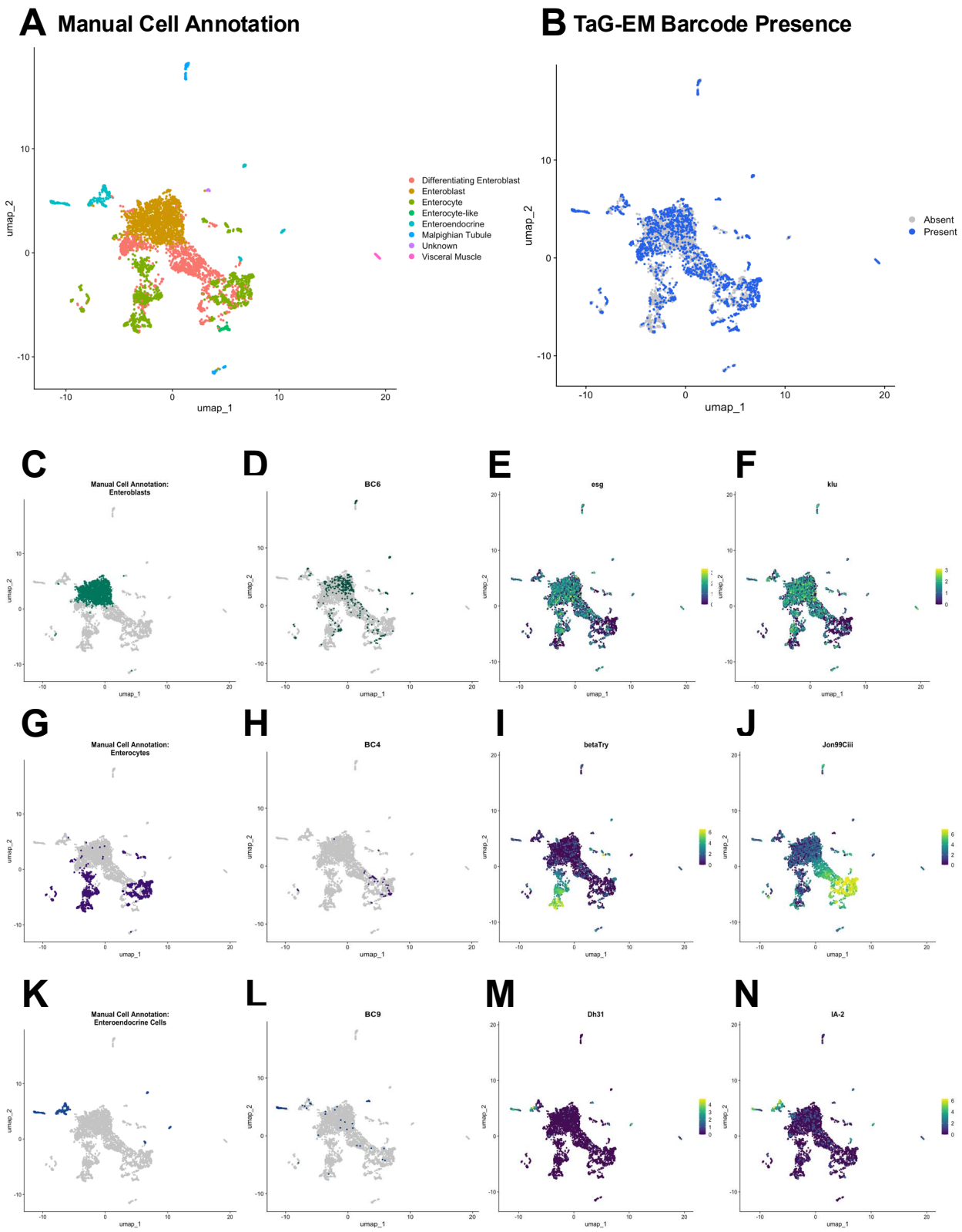


Figure 6. Expression of TaG-EM genetic barcodes in larval intestinal cell types. A) UMAP plot of Drosophila larval gut cell types. B) Annotation of cells associated with a TaG-EM barcode

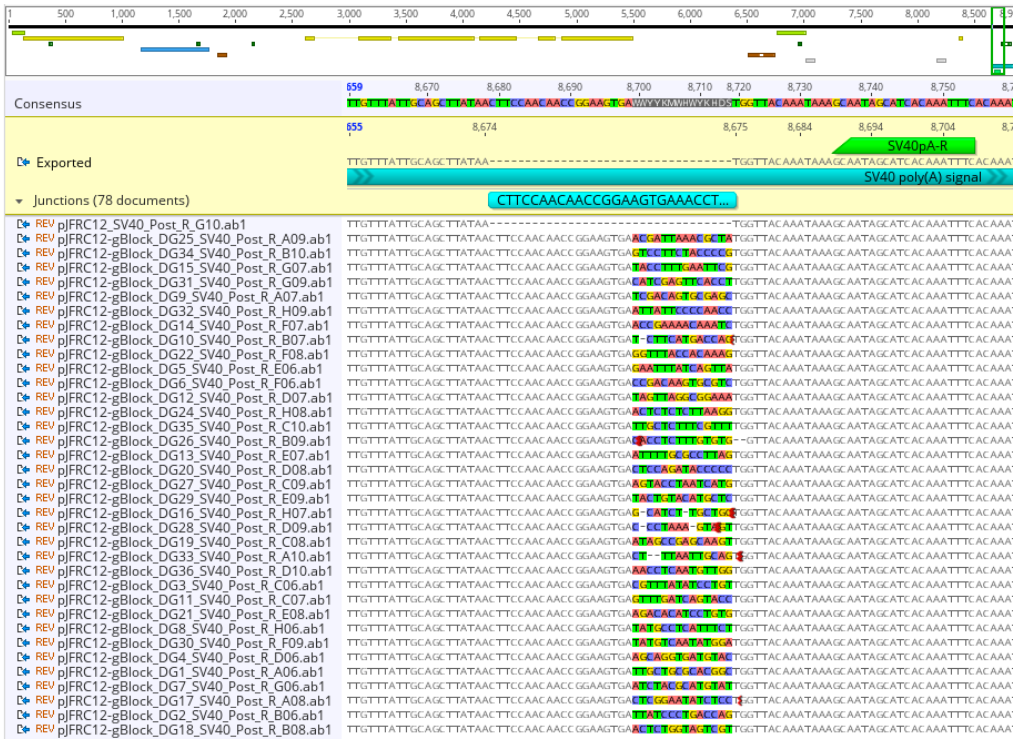
1132 across all 8 multiplexed experimental conditions using data from the gene expression library
1133 and an enriched TaG-EM barcode library. C) Annotated enteroblast cells. D) Presence of TaG-
1134 EM barcode (BC6) driven by the EB-Gal4 line using data from the gene expression library and
1135 an enriched TaG-EM barcode library. Gene expression levels of enteroblast marker genes E)
1136 *esg*, F) *klu*. G) Annotated enterocyte cells. H) Presence of TaG-EM barcode (BC4) driven by the
1137 EC-Gal4 line using data from the gene expression library and an enriched TaG-EM barcode
1138 library. Gene expression levels of enterocyte marker genes I) *betaTry*, J) *Jon99Ciii*. K)
1139 Annotated enteroendocrine cells. L) Presence of TaG-EM barcode (BC9) driven by the EE-Gal4
1140 line using data from the gene expression library and an enriched TaG-EM barcode library. Gene
1141 expression levels of enteroendocrine cell marker genes M) *Dh31*, N) *IA-2*.

Supplemental Figures

A

Barcode sequence	Transgenic line
ATTTTGCGCCTTAG	1-2
CCCTAAAGTACCGT	4-3
TAGTTAGGCGGAAA	5-4
TTGCTCTTTCGTTT	6-4
TTGCTGCGCACGGC	7-1
CGTTTATATCCTGT	10-1
AGCAGGTGATGTAC	1-1, 1-3, 10-2
TATGCCTCATTTCT	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

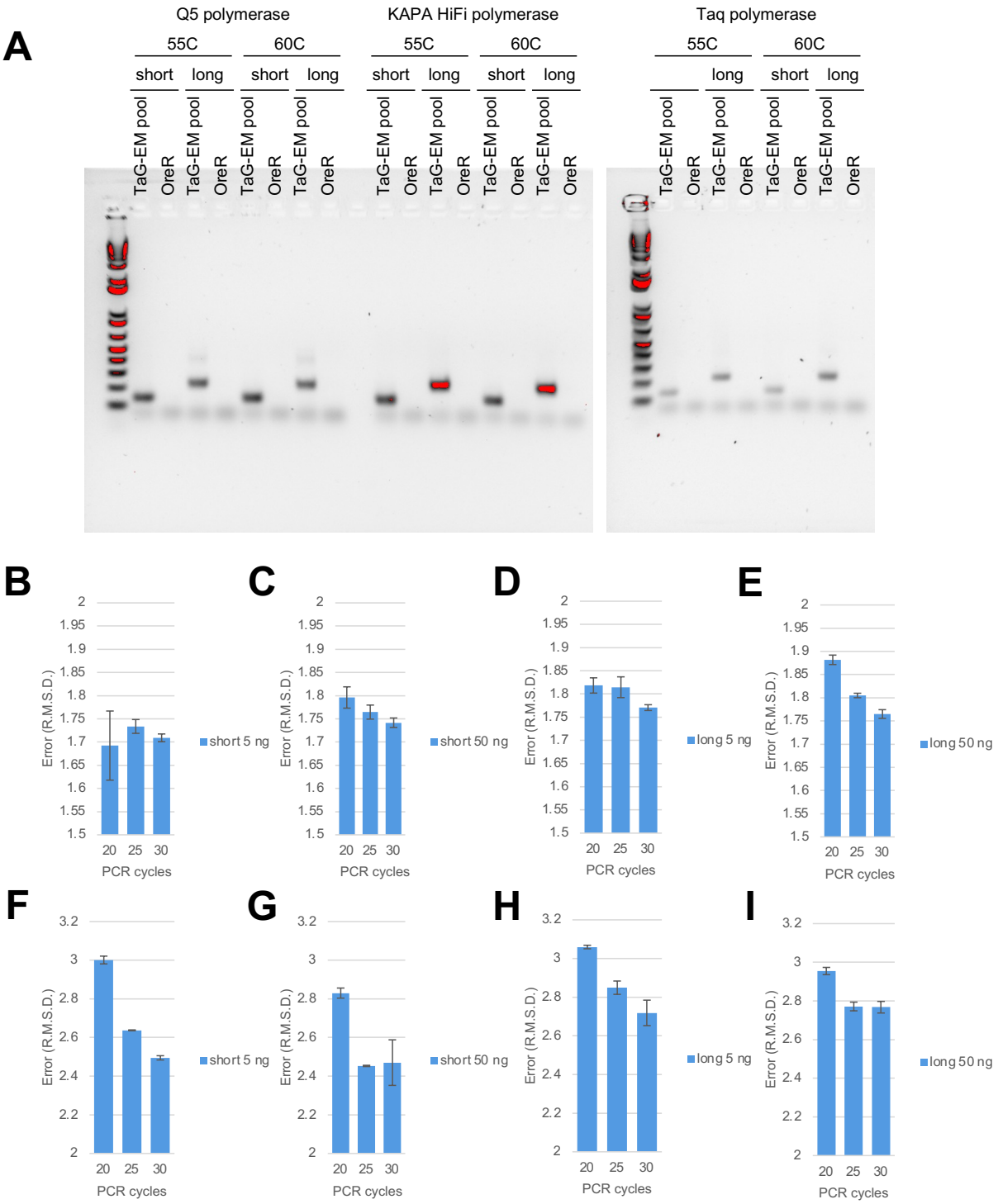
B



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. Because the TaG-EM barcode constructs were injected as a pool of 29 purified plasmids, some of the transgenic lines had

1149 inserts of the same construct. In total 20 unique lines were recovered from this round of
1150 injection.
1151

1152



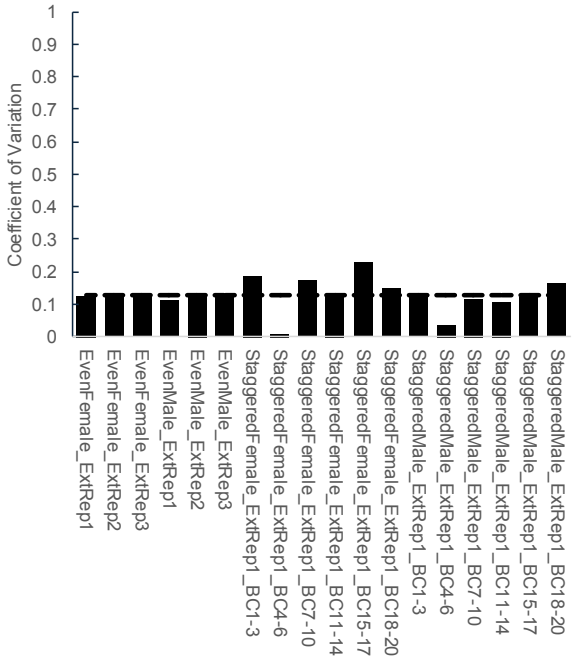
1153

1154 **Supplemental Figure 2. Optimization of TaG-EM barcode amplification.** A) Gels showing

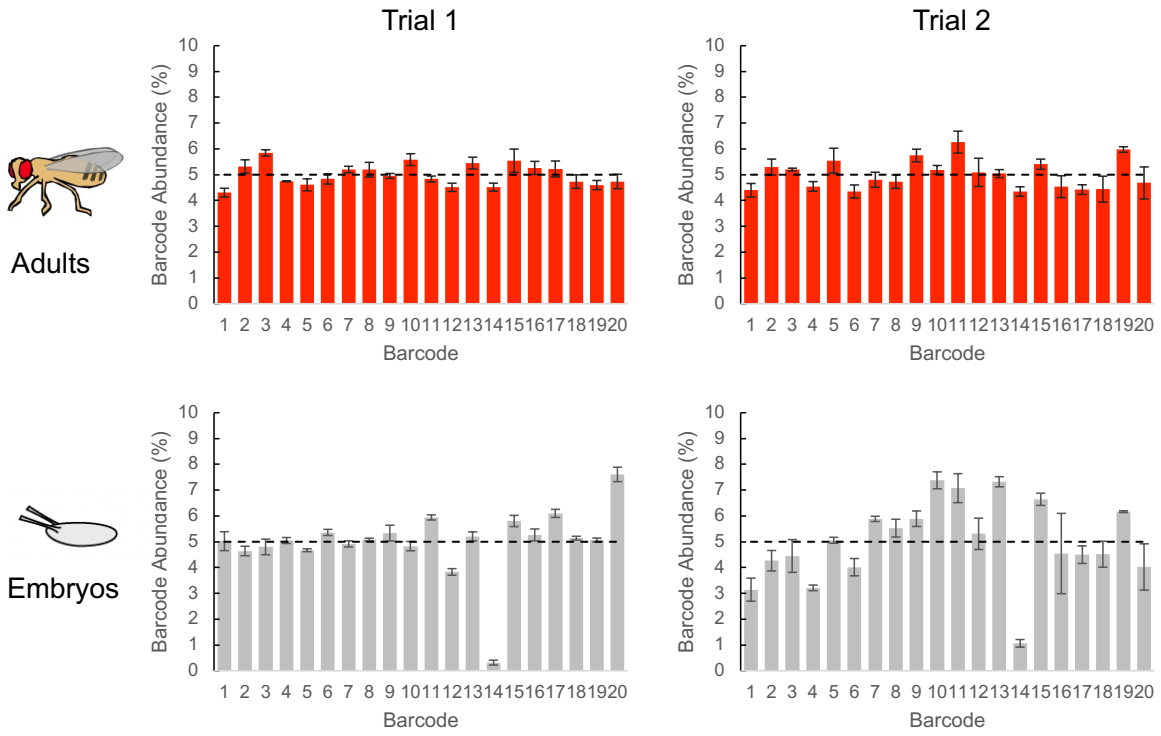
1155 bands produced when amplifying TaG-EM flies or a wild type control with the indicated

1156 polymerase, annealing temperature, and primer pair (short = B2_3'F1_Nextera/

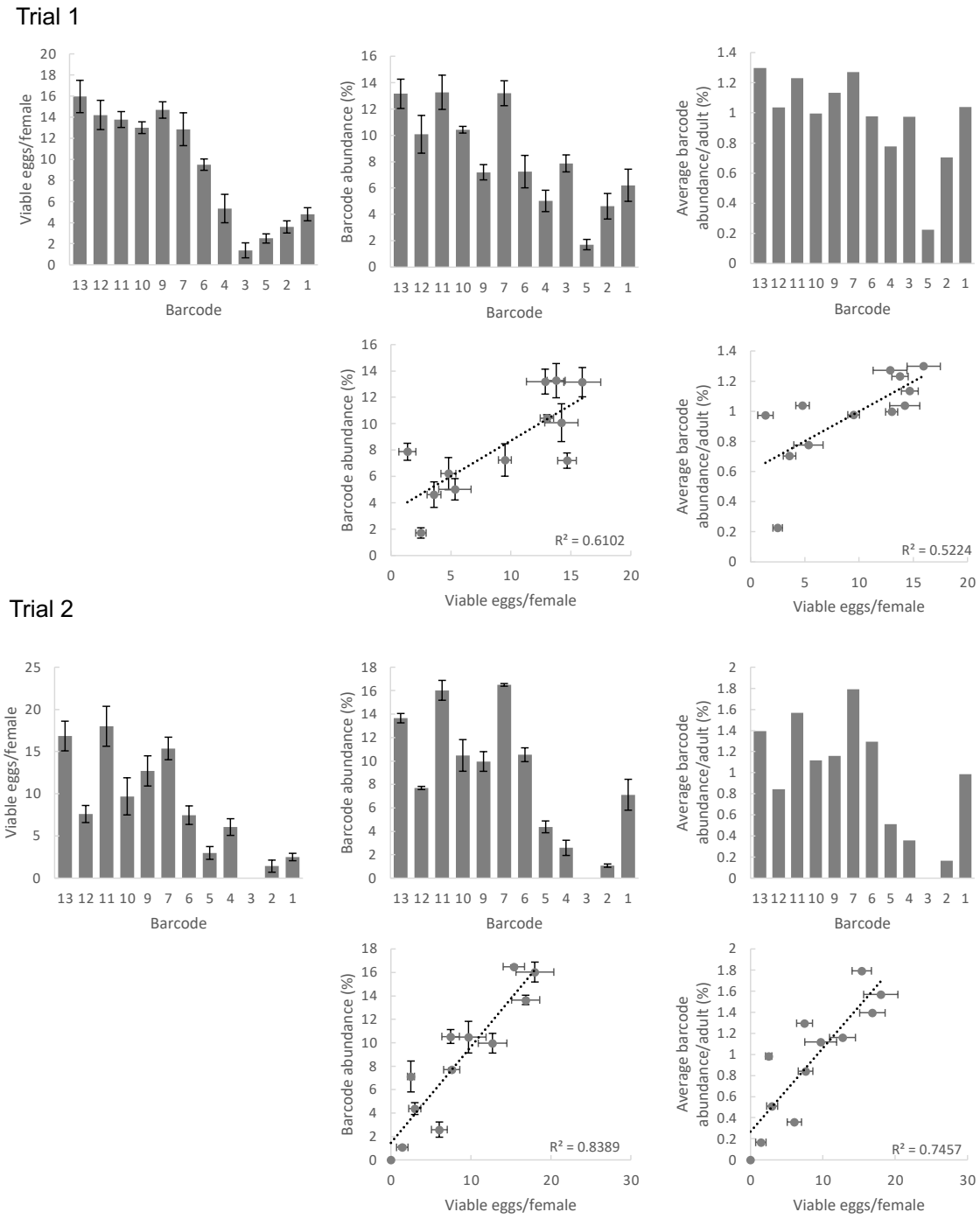
1157 SV40_pre_R_Nextera; long = B2_3'F1_Nextera/ SV40_post_R_Nextera). B-E) Mean error
1158 (R.M.S.D. root mean squared deviation from expected value) for even pool amplified with the
1159 indicated primer set, input amount, and cycle number using KAPA HiFi polymerase (n=3, error
1160 bars are +/- S.E.M.). F-G) Mean error (R.M.S.D. root mean squared deviation from expected
1161 value) for staggered pool amplified with the indicated primer set, input amount, and cycle
1162 number using KAPA HiFi polymerase (n=3, error bars are +/- S.E.M.).
1163



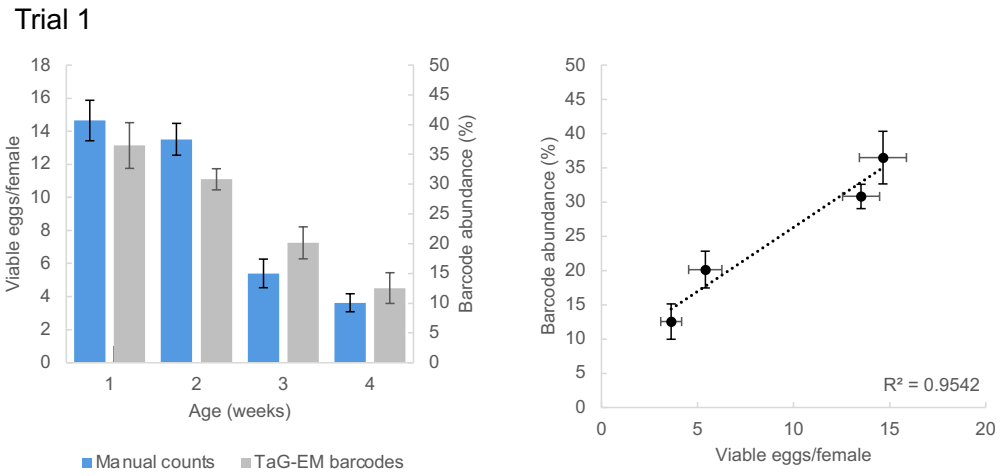
Supplemental Figure 3. Coeffient of variation for TaG-EM structured pools. Plot showing coefficient of variation for different groups of TaG-EM barcodes in the structured pools. Dashed line indicates the mean coefficient of variation across all conditions.



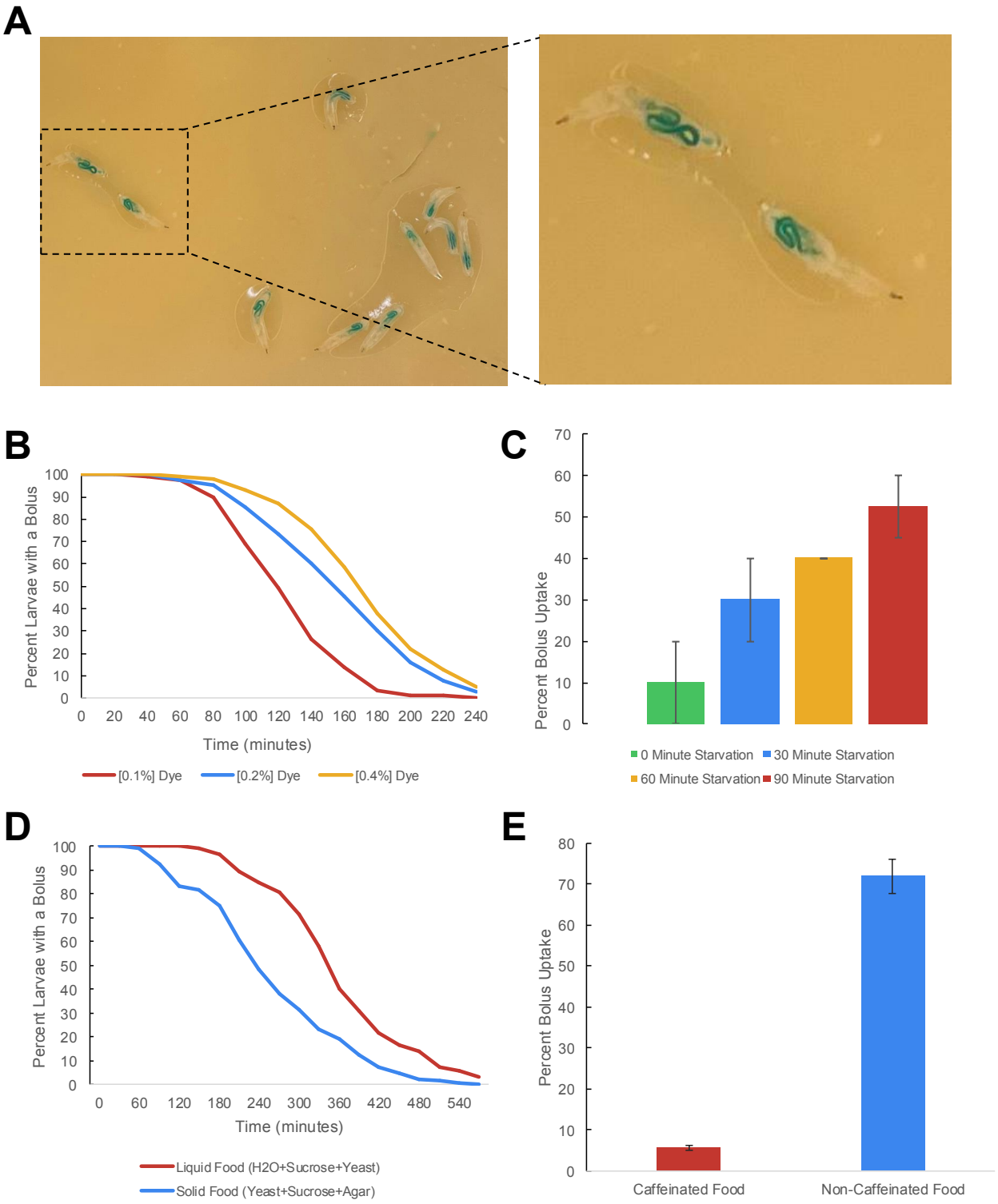
Supplemental Figure 4. Oviposition tests with TaG-EM barcode lines. Plots showing mean TaG-EM barcode abundance for adult females used in oviposition experiments (top) and eggs collected from these females (bottom). Data from two independent trials is shown (n=3 for each trial, error bars are +/- S.E.M.). Dashed lines indicate the expected abundance values.



Supplemental Figure 5. 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 +/- S.E.M.).



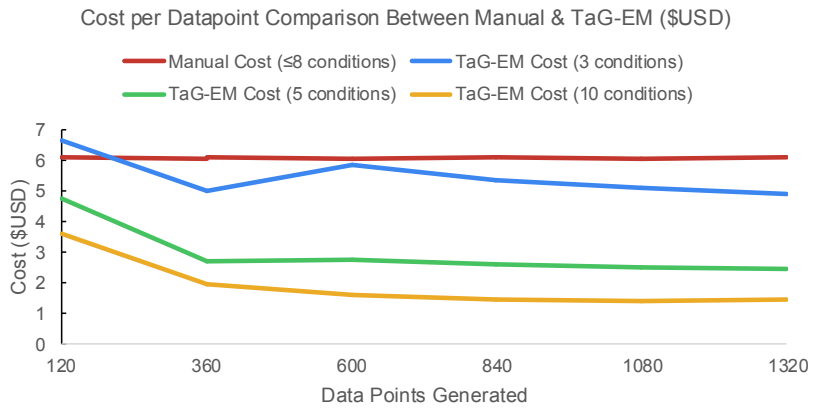
Supplemental Figure 6. Average age-dependent fecundity data for Trial 1. Average number of viable eggs per female (manual counts) and average barcode abundance are shown both as a bar plot and scatter plot (n=3 biological replicates for 3 barcodes per condition, error bars are +/- S.E.M.). Data from Trial 2 is shown in Figure 3C.



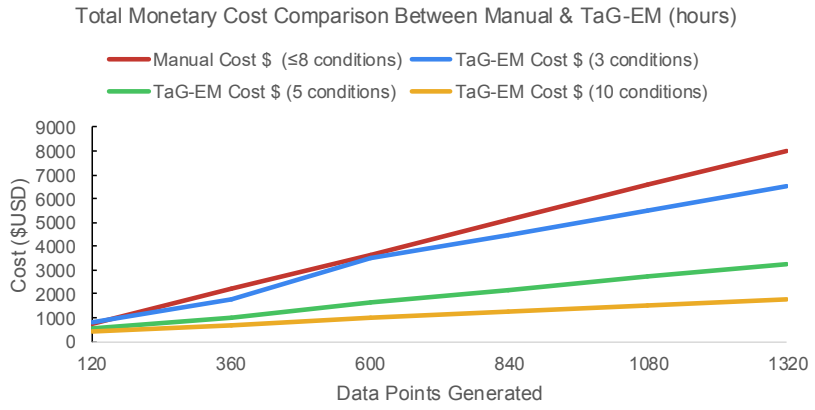
Supplemental Figure 7. Larval gut motility assay parameters. A) Images of larvae fed with blue-dyed yeast agar. B) Effect of dye concentration on food transit time. C) Effect of starvation

1192 time on feeding and uptake of the dyed food bolus. D) Effect of liquid versus solid diet on food
1193 transit time. E) Aversive effect of caffeine on food bolus uptake.

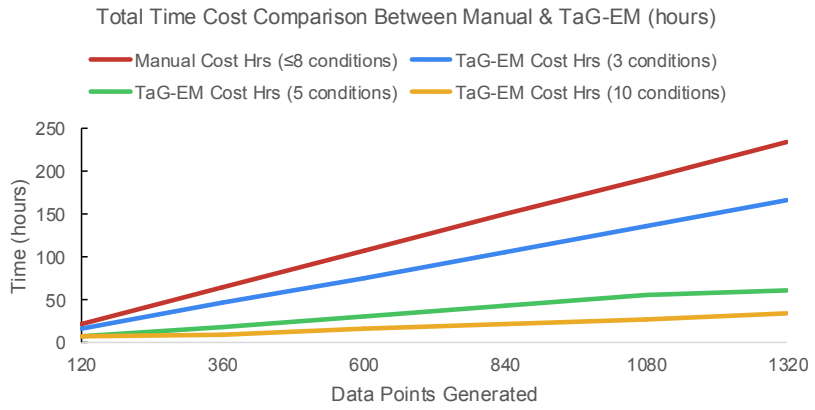
A Cost per Datapoint Analysis



B Monetary Cost Analysis ($USD)

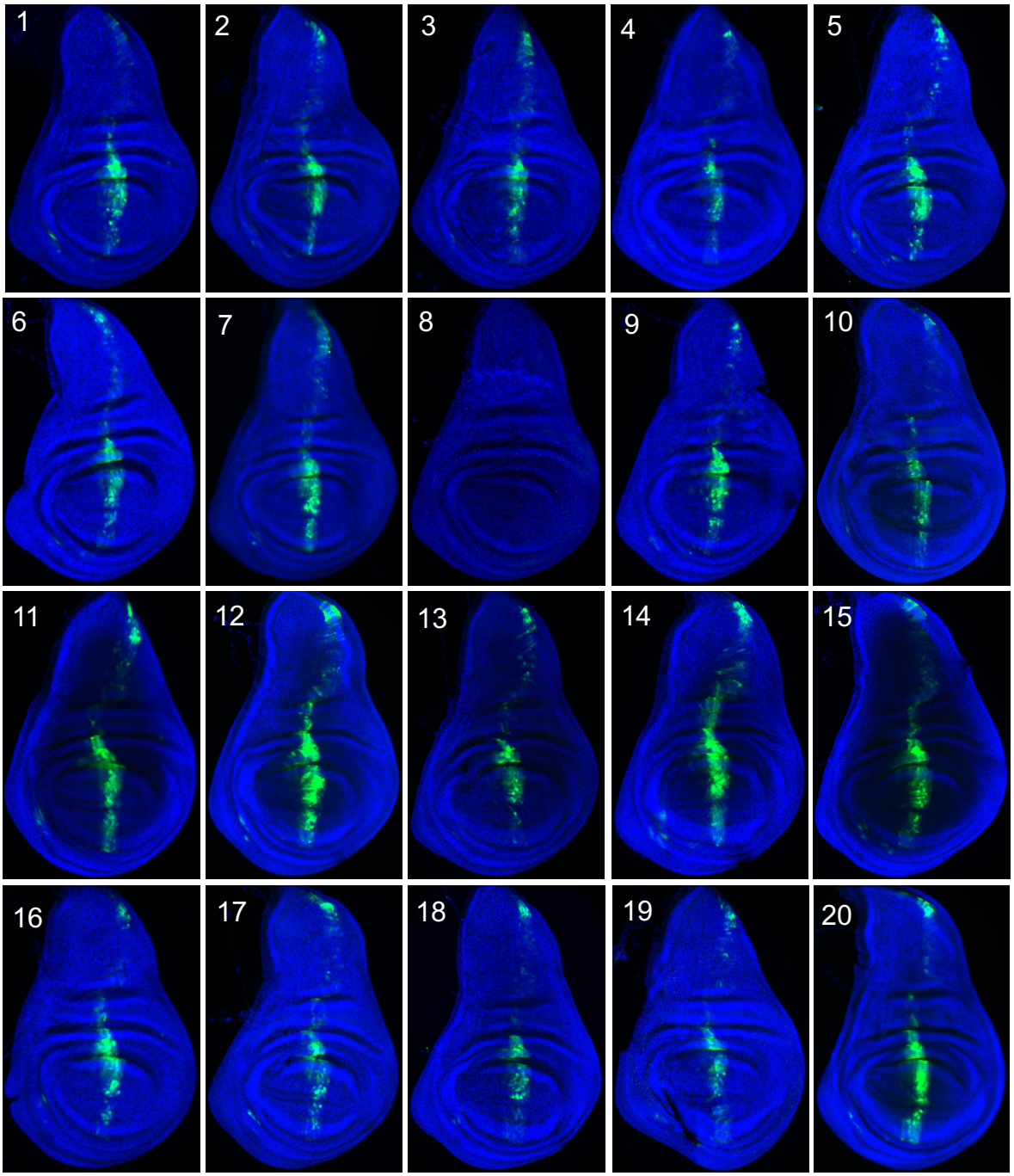


C Labor Time Cost Analysis (Hours)

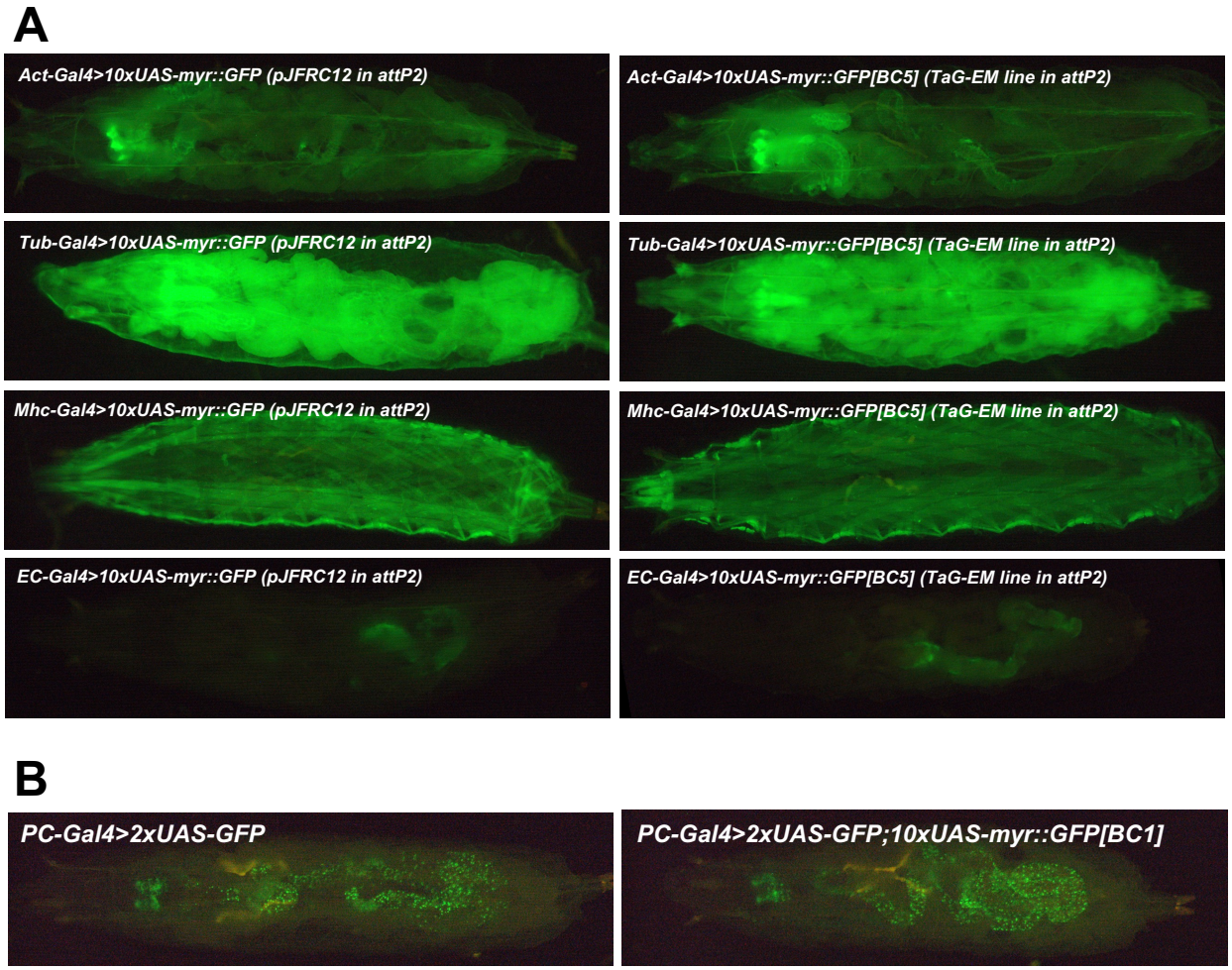


Supplemental Figure 8. Cost comparisons for manual and TaG-EM gut motility assays. A) Cost per data point as a function of the number of data points generated and the number of

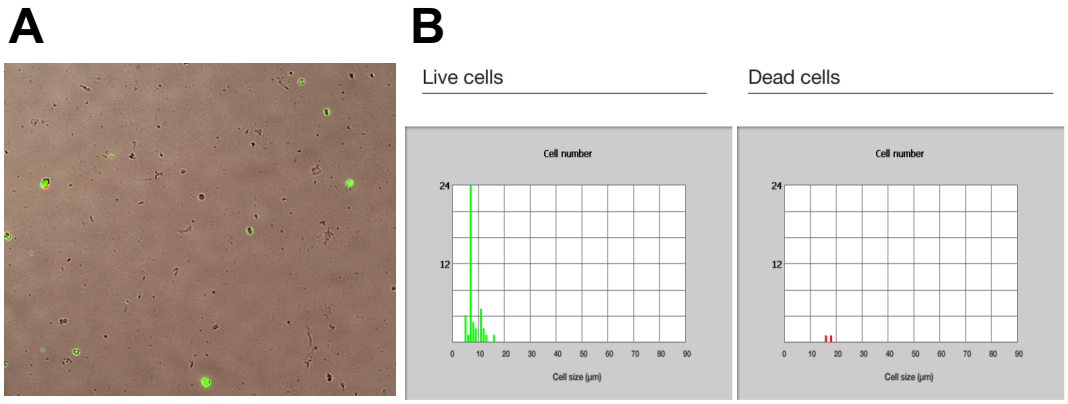
experimental conditions screened. B) Overall experiment cost and C) labor effort as a function of the number of data points generated and the number of experimental conditions screened.



Supplemental Figure 9. 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.



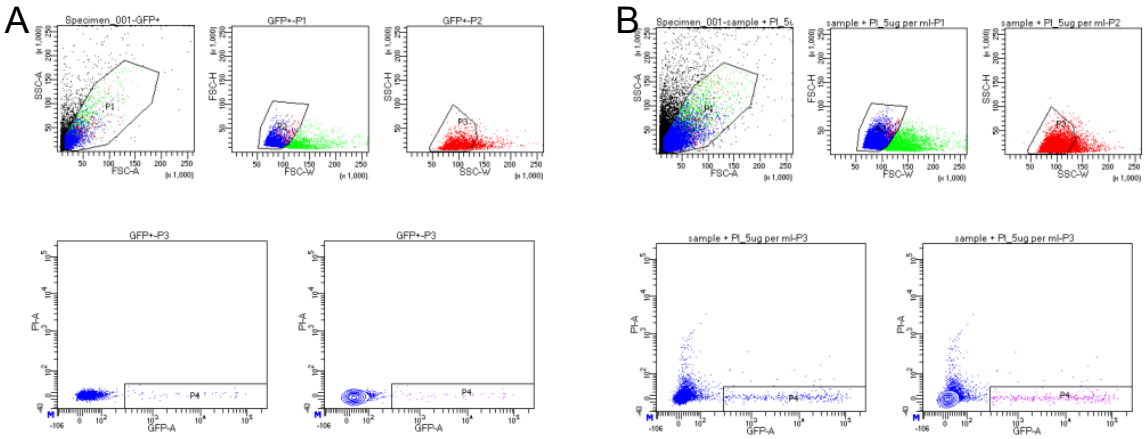
Supplemental Figure 10. TaG-EM line GFP expression driven by different Gal4 drivers. A) Comparison of endogenous GFP expression in larvae for the original pJFRC12 construct inserted in the attP2 landing site (left) or for a TaG-EM barcode line (right) expressed under the control of the indicated driver 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.



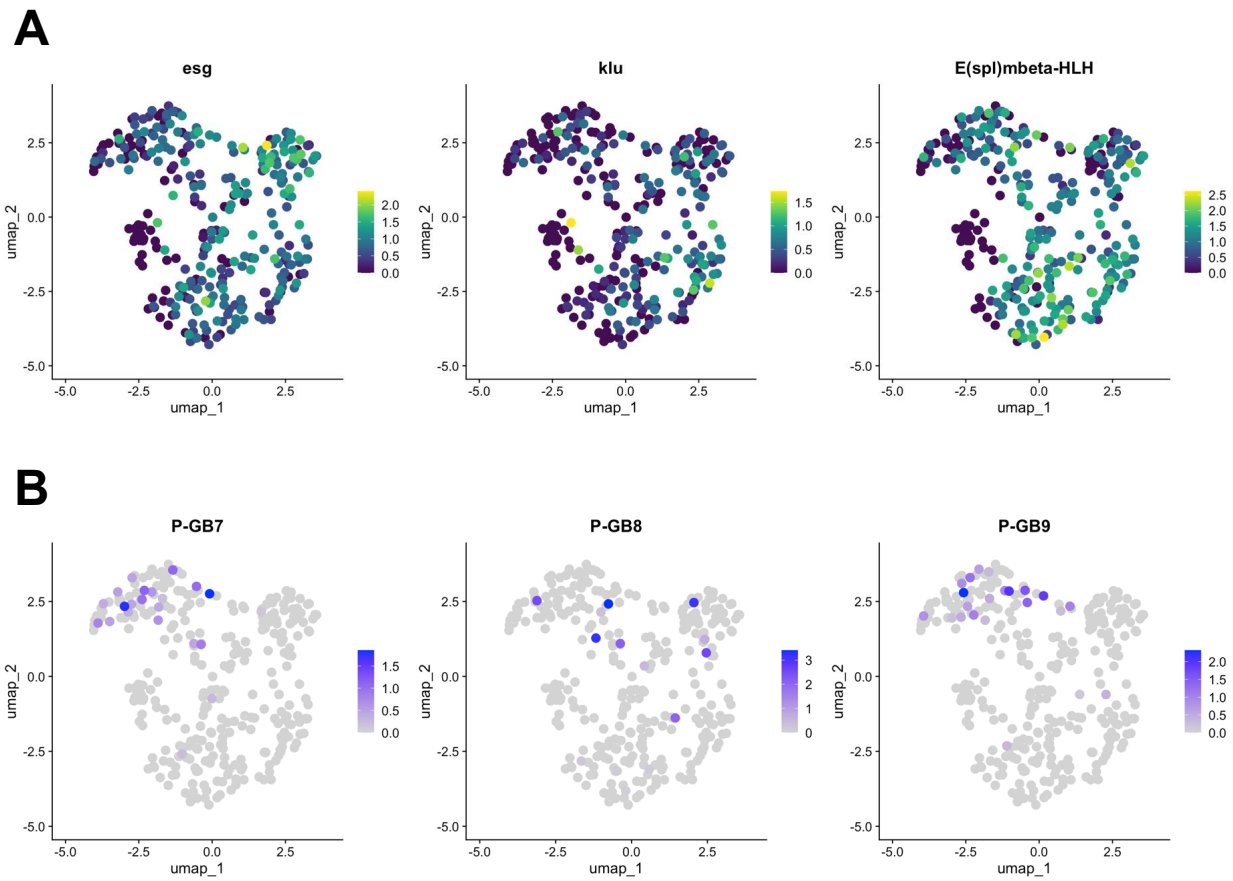
1210

1211 **Supplemental Figure 11. Dissociated intestinal cell viability.** A) GFP expression visualized
1212 in dissociated cells from gut driver lines crossed to hexameric GFP and TaG-EM line. B)
1213 Proportion of live (left) and dead (right) cells post-isolation and flow sorting as assessed by GFP
1214 expression and propidium iodide staining.

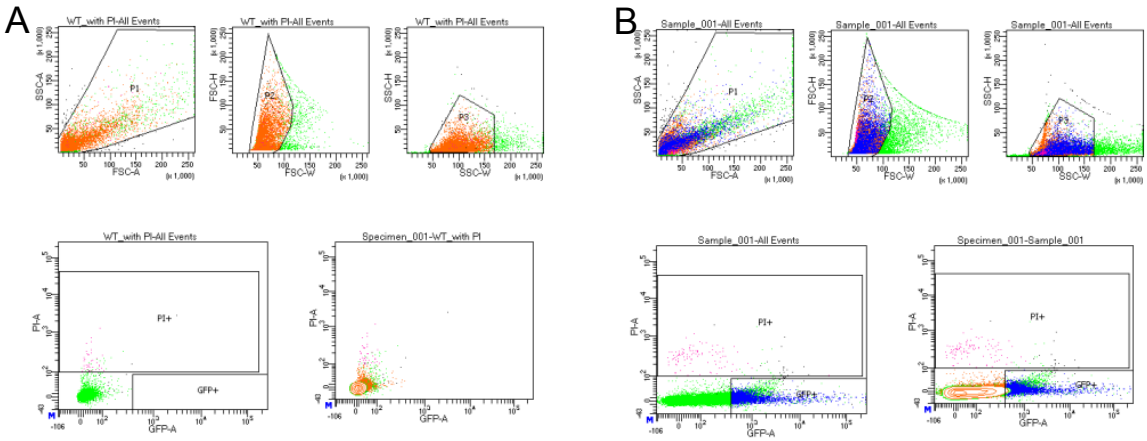
1215



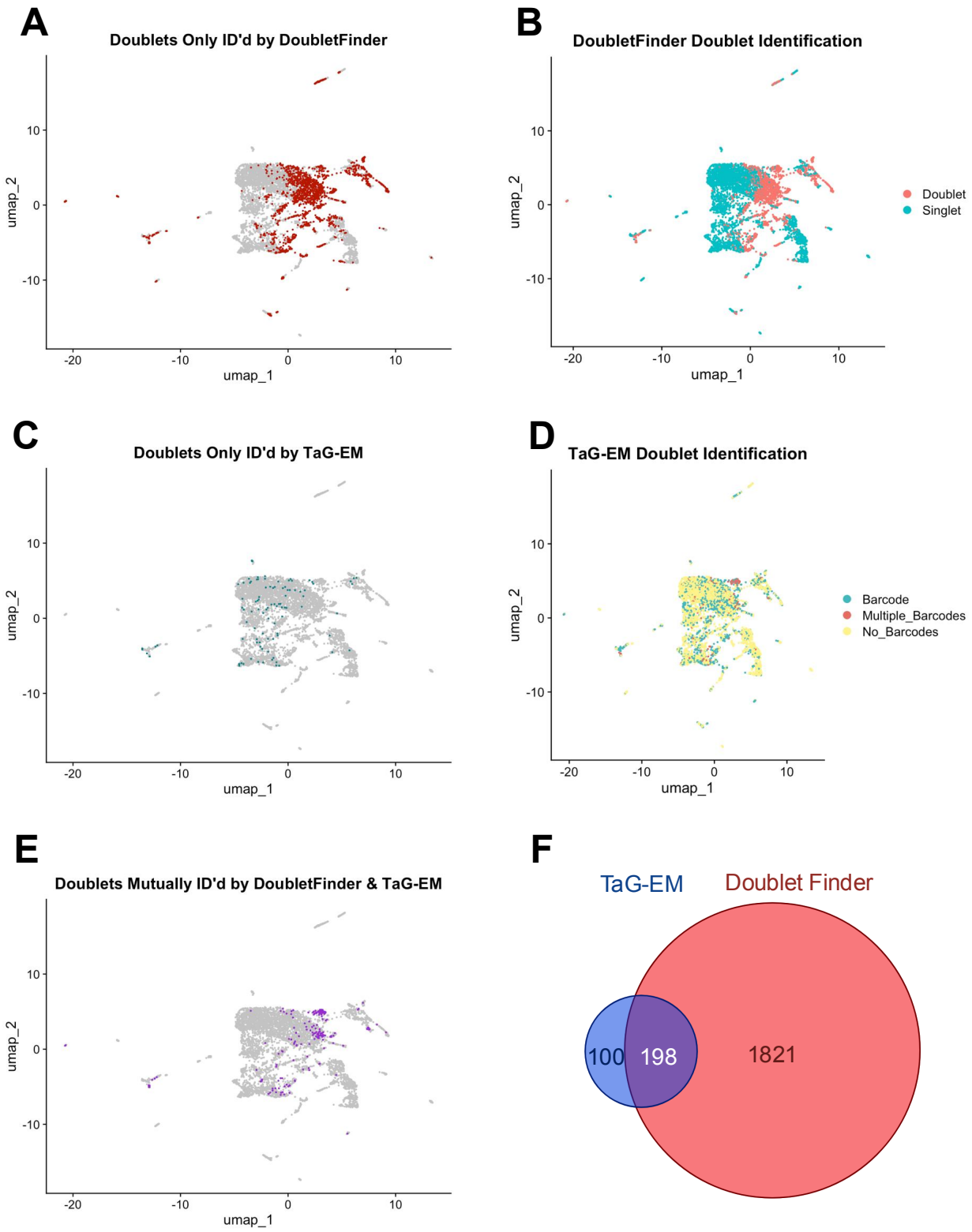
Supplemental Figure 12. BD FACSDiva 8.0.1 gating for sorted cells. A) GFP gating created by analyzing a pool of GFP positive and negative cells. B) Flow gating for Drosophila gut cells with TaG-EM GFP expression driven in intestinal precursor cells (PC-Gal4) and enterocytes (EC-Gal4).



Supplemental Figure 13. 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.

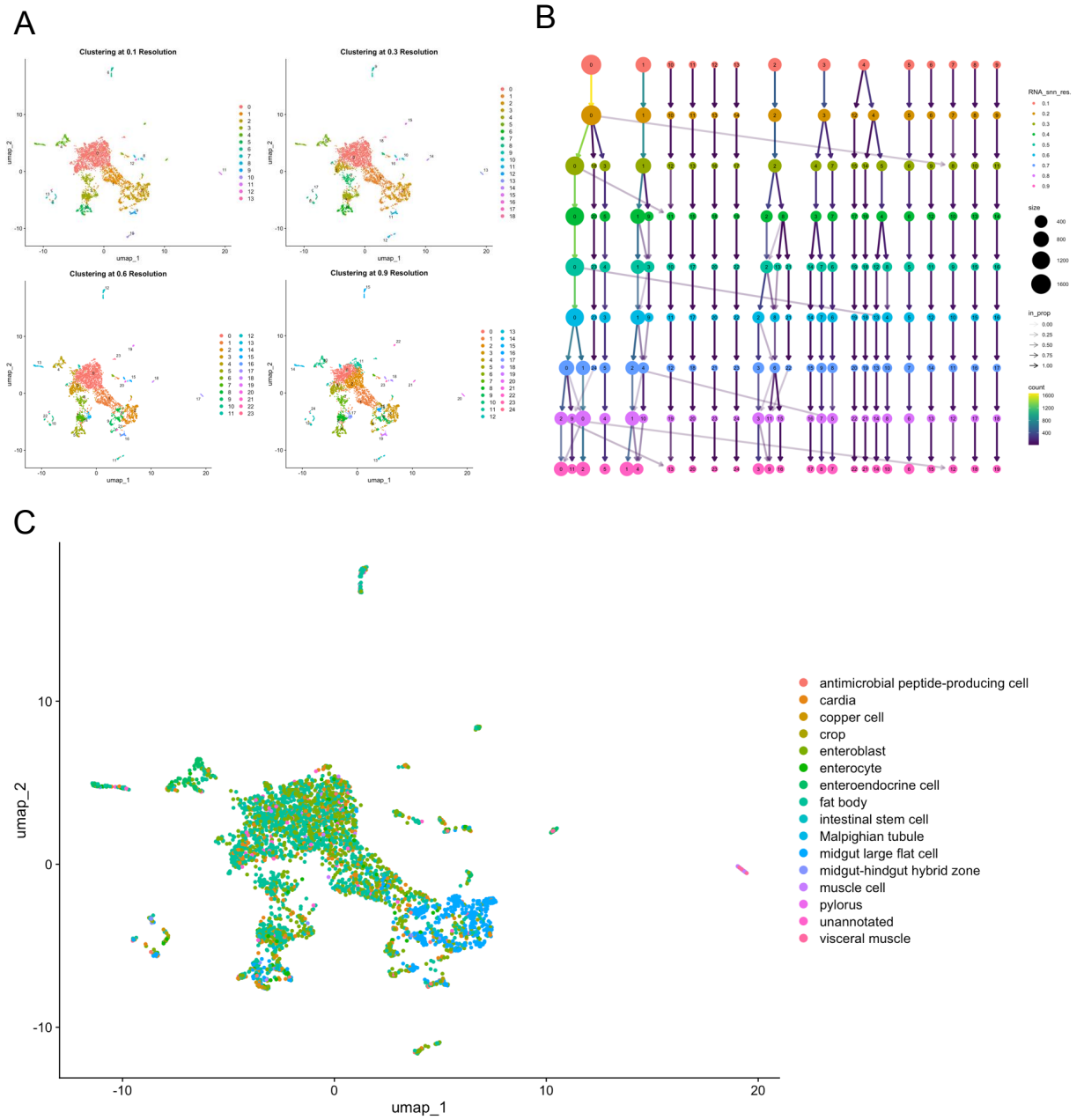


Supplemental Figure 14. BD FACSDiva 8.0.1 gating for sorted cells. A) Dead cell gating created by staining sample with propidium iodine (PI). B) Flow gating for *Drosophila* gut cells with TaG-EM and hexameric GFP expression driven by the pan-midgut, enteroblast, enterocyte, enteroendocrine, and precursor cell drivers.

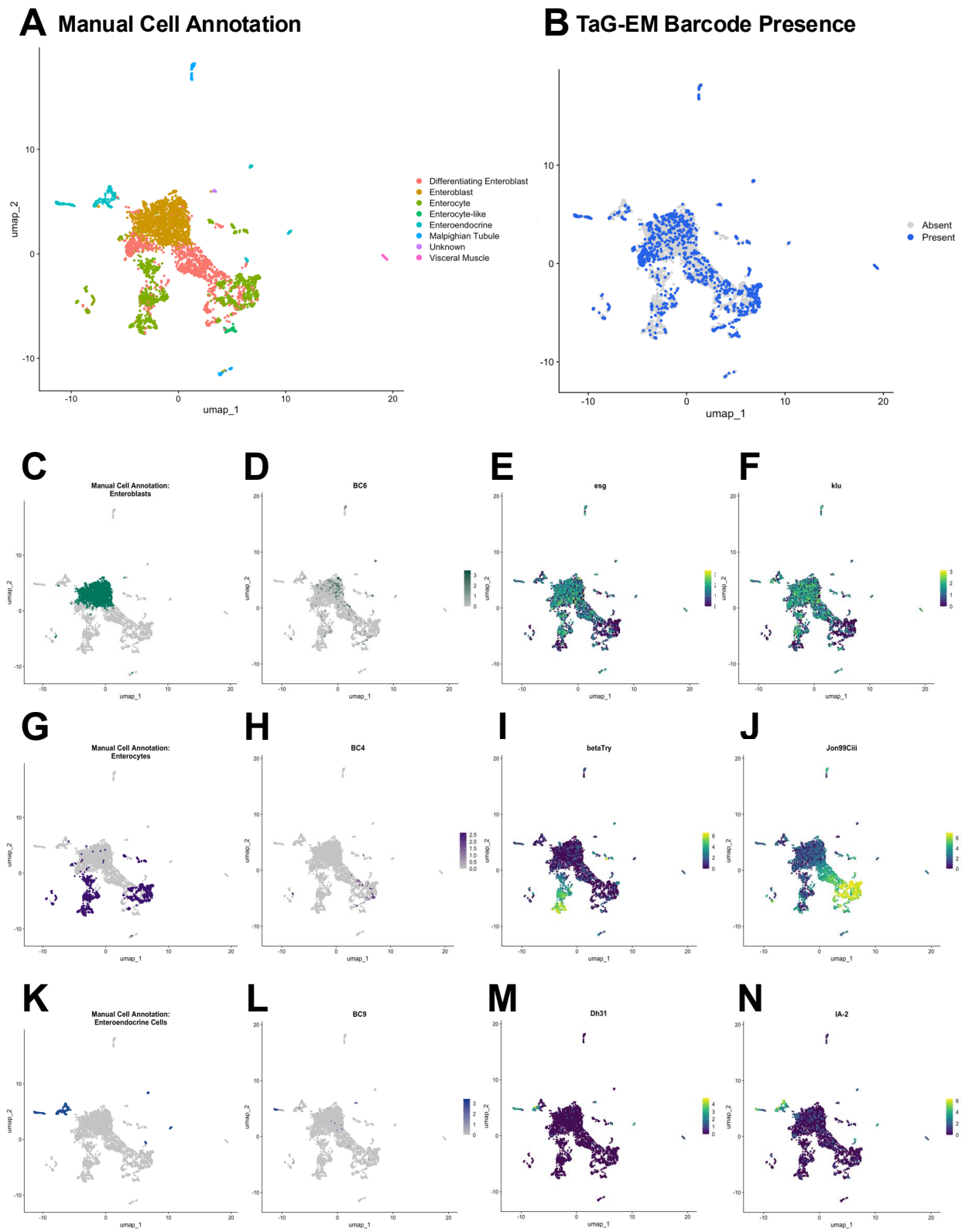


Supplemental Figure 15. TaG-EM-based doublet identification. UMAP plots pre-doublet removal showing A) doublets uniquely identified by DoubletFinder, B) all doublets identified by

1240 DoubletFinder, C) doublets uniquely identified by TaG-EM barcodes, D) all doublets identified
1241 by TaG-EM barcodes, E) doublets mutually found by TaG-EM and DoubletFinder, F) Venn
1242 diagram of overlap between doublets identified by TaG-EM and DoubletFinder.
1243
1244

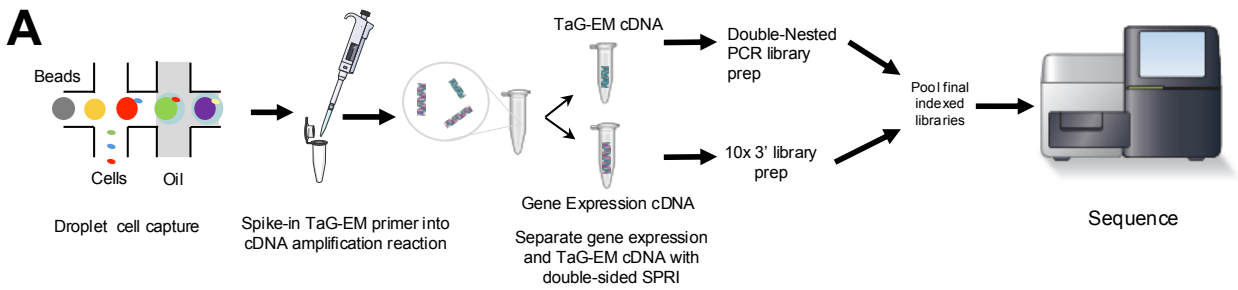


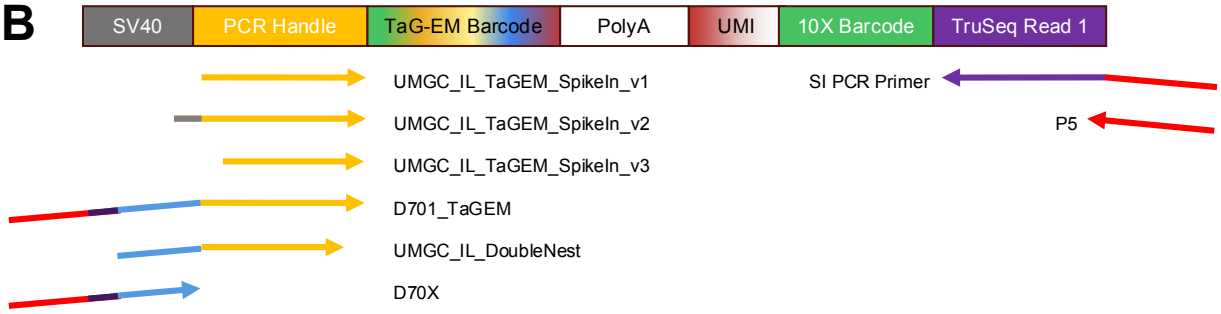
Supplemental Figure 16. Clustering and automated annotation. A) UMAP plots clustered at different resolutions. B) Clustree analysis of the effect of clustering resolution. C) Automated cell type annotation using data from the Fly Cell Atlas.



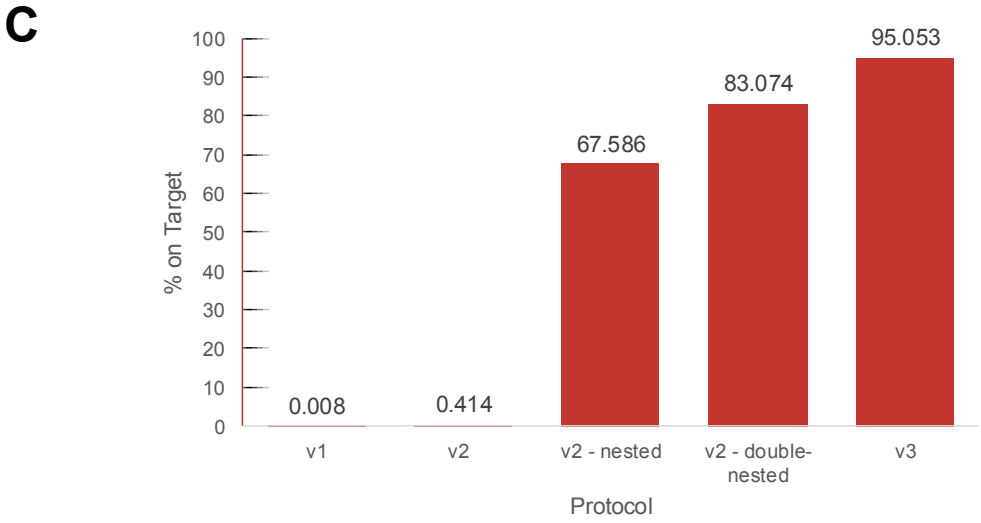
Supplemental Figure 17. Expression of TaG-EM genetic barcodes in larval intestinal cell types. A) UMAP plot of Drosophila larval gut cell types. B) Annotation of cells associated with a

1253 TaG-EM barcode across all 8 multiplexed experimental conditions using data from the gene
1254 expression library only. C) Annotated enteroblast cells. D) Expression level of TaG-EM barcode
1255 (BC6) driven by the EB-Gal4 line using data from the gene expression library only. Gene
1256 expression levels of enteroblast marker genes E) *esg*, F) *klu*. G) Annotated enterocyte cells. H)
1257 Expression level of TaG-EM barcode (BC4) driven by the EC-Gal4 line using data from the gene
1258 expression library only. Gene expression levels of enterocyte marker genes I) *betaTry*, J)
1259 *Jon99Ciii*. K) Annotated enteroendocrine cells. L) Expression level of TaG-EM barcode (BC9)
1260 driven by the EE-Gal4 line using data from the gene expression library only. Gene expression
1261 levels of enteroendocrine cell marker genes M) *Dh31*, N) *IA-2*.
1262



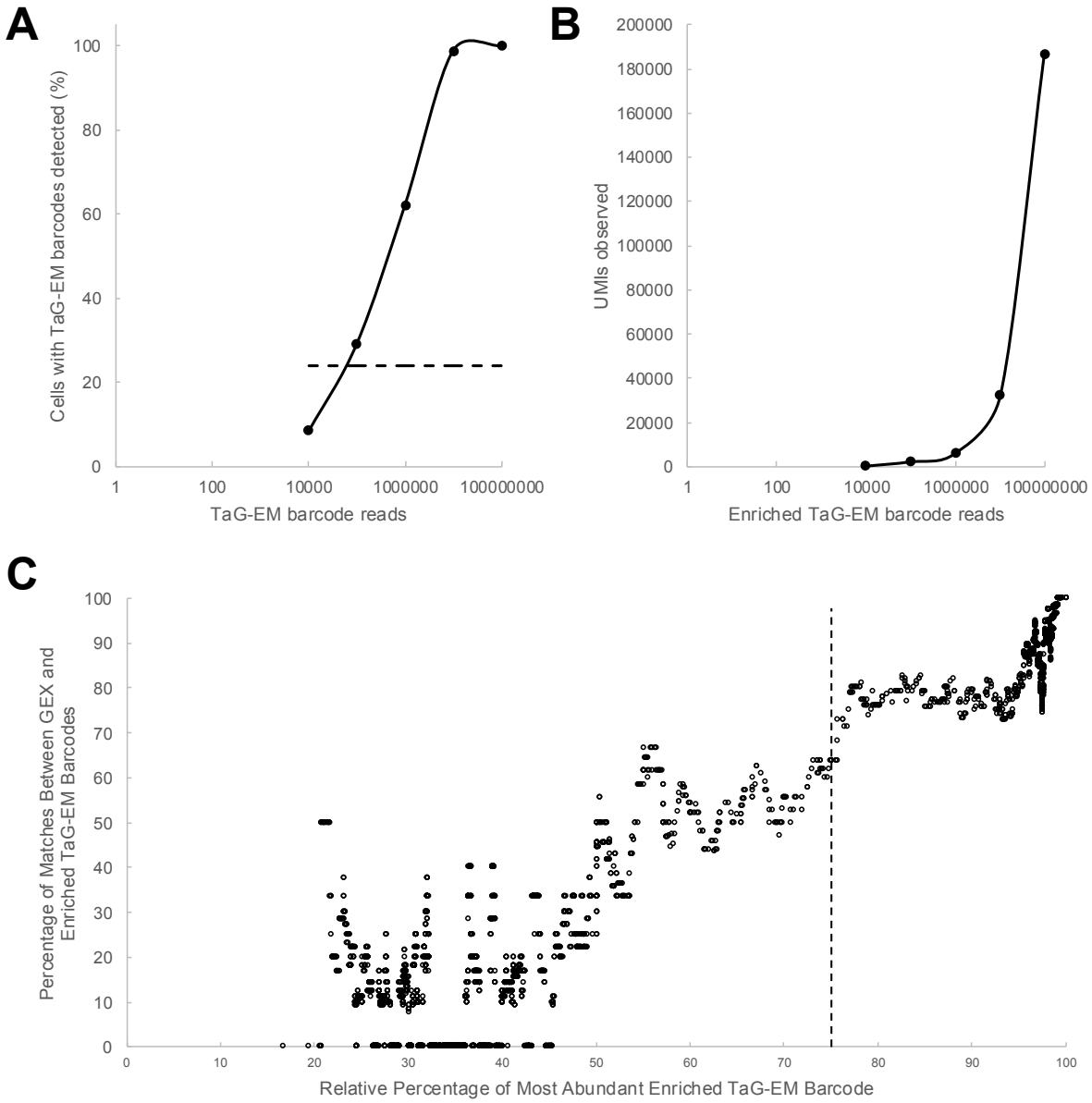


Protocol	v1 PCR	v2 PCR	v2 Nested	v2 Double Nested	v3 Double Nested
Spike In Primer	UMGC_IL_TaGEM _SpikeIn_v1	UMGC_IL_TaGEM _SpikeIn_v2	UMGC_IL_TaGEM _SpikeIn_v2	UMGC_IL_TaGEM _SpikeIn_v2	UMGC_IL_TaGEM _SpikeIn_v3
PCR 1	NA	NA	D701_TaGEM & P5	UMGC_IL_Double Nest & SI PCR	UMGC_IL_Double Nest & SI PCR
PCR 2	NA	NA	NA	D701 & P5	D702 & P5



Supplemental Figure 18. Optimizing amplification of the TaG-EM barcode library. A) Workflow for single-cell capture; cDNA amplification with added spike-in primer for TaG-EM library followed by a SPRI size-selection clean-up, then PCR(s) to create library for sequencing. B) Spike-in primers and amplification primers used to enrich TaG-EM barcodes. Table

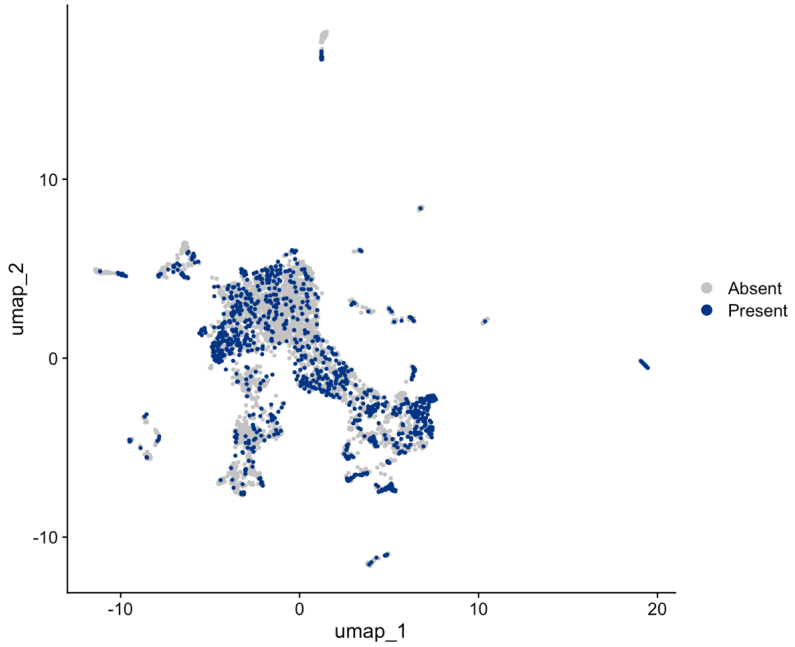
1268 summarizes different protocols tested to amplify the TaG-EM barcodes and create an enriched
1269 sequencing library. C) Percent of on-target reads for each enriched TaG-EM barcode library.
1270
1271



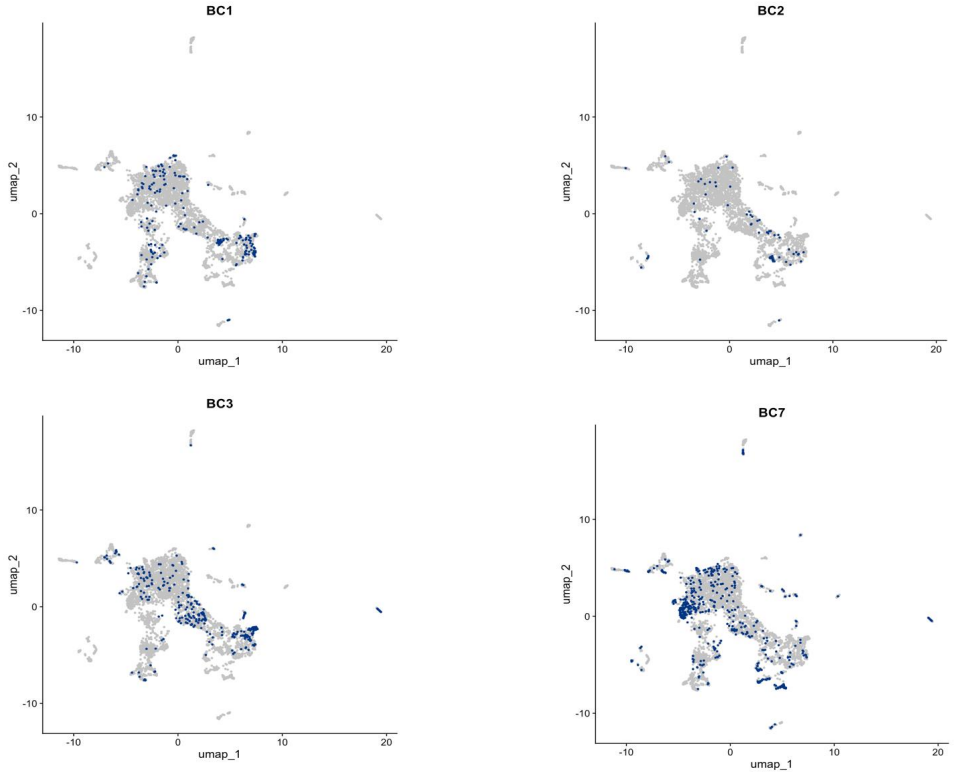
Supplemental Figure 19. Performance of the enriched TaG-EM barcode library. A) Proportion of cells with at least one barcode read assigned as a function of read depth for the enriched TaG-EM barcode library. Dashed line indicated percentage of cells with TaG-EM barcodes detected in the gene expression library B) Number of unique UMIs observed as a function of read depth. C) Correlation between barcodes detected in the gene expression (GEX) library and the enriched TaG-EM barcode library as a function of the purity of TaG-EM barcode assignment to the corresponding cell barcode. Dashed line indicates the threshold used for TaG-EM barcode calling in the enriched TaG-EM barcode library.

A

Pan-midgut-associated TaG-EM barcode expression (BC1, BC2, BC3, BC7)

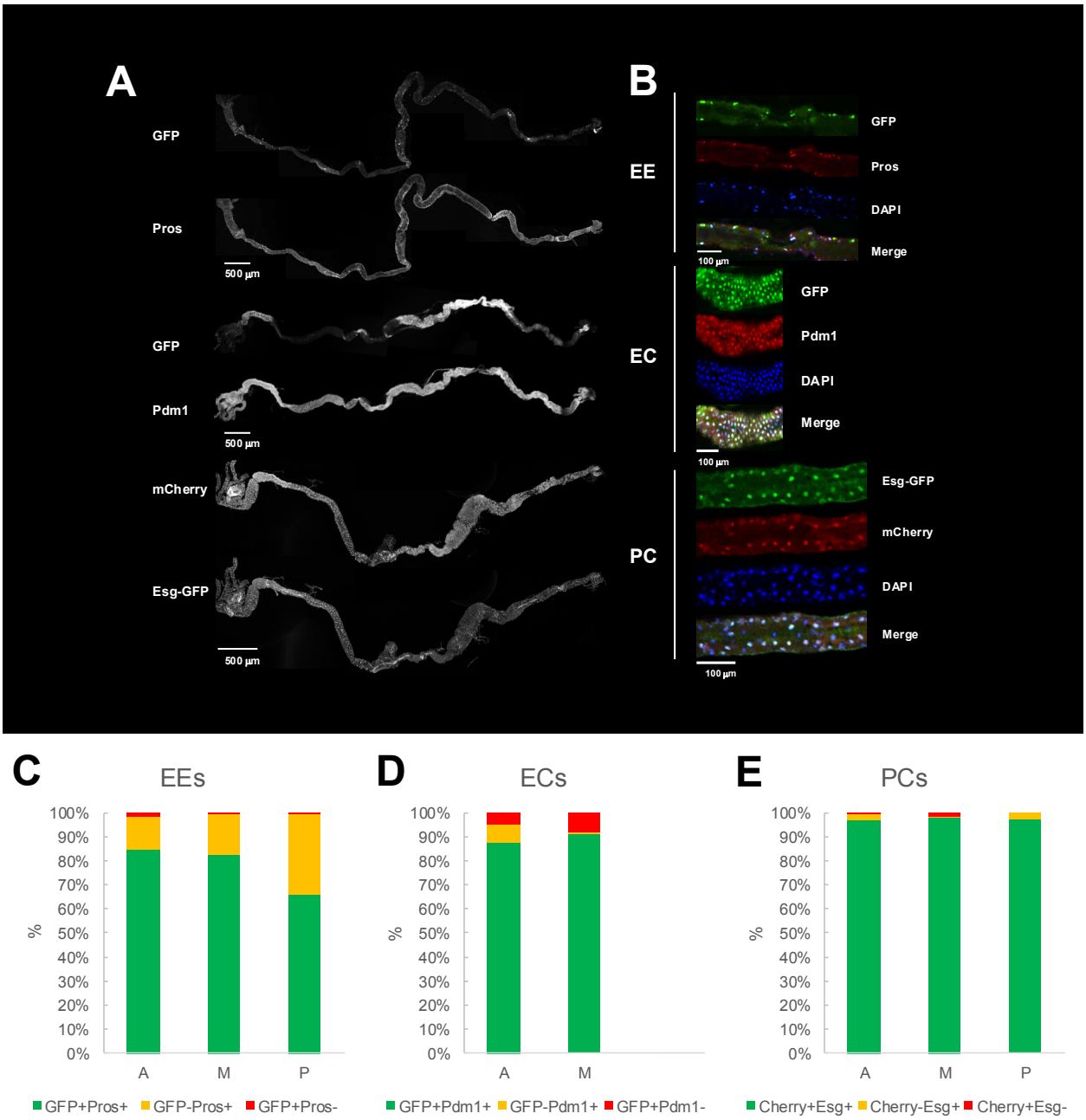


B



1281

1282 **Supplemental Figure 20. Expression of the PMG-Gal4 driven TaG-EM barcodes.** UMAP
1283 plots showing expression of the four PMG-Gal4 driven TaG-EM barcodes (BC1, BC2, BC3, and
1284 BC7) either A) in aggregate or B) individually.
1285



Supplemental Figure 21. Characterization of Gal4 line expression in the larval gut. A) Confocal images of third instar midguts showing Gal4-driven fluorophore expression (GFP or mCherry) and comparison with immunostainings of the gut cell markers Prospero (enteroendocrine), Pdm1 (enterocyte) and Esg-GFP (progenitor cell). For each image, Z projections of the stacks recorded along the length of the midgut were manually stitched together. B) Representative single frames confocal images of a small region of the midgut showing immunostainings of the different gut cell markers and the Gal4-driven fluorophores. Quantification of overlapping and non-overlapping expression between the Gal4-driver fluorophore expression

1295 and the cell type marker in the anterior (A), middle (M), and posterior (P) regions for C)
1296 enteroendocrine cells (EC-Gal4), D) enterocytes (EC-Gal4), E) precursor cells (PC-Gal4). Five
1297 specimens for each Gal4 line were examined. In the case of the enterocyte-specific driver, only
1298 anterior and middle regions were analyzed since the driver is largely inactive in the posterior part
1299 of the midgut.
1300