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

Jorge Blanco Mendana[1], Margaret Donovan[1], Lindsey Gengelbach[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 to assess the effects of experimental exposures or genetic manipulations would benefit from increased ability to multiplex samples. Given the fixed costs of droplet-based single-cell sequencing, generating data for single-cell transcriptomic time courses or experimental manipulations can be costly. These costs are also a barrier to including

replicates to assess biological variability; consequently, a lack of replicates is a common shortcoming of single-cell sequencing experiments. Antibody-based cell hashing or feature barcoding approaches have been developed to allow multiplexing of samples in droplet-based single-cell sequencing reactions (Stoeckius et al., 2018, 2017). In addition, other multiplexing strategies for single-cell sequencing based on alternative methods for tagging cells (Cheng et al., 2021) or making use of natural genetic variation have been used (Kurmangaliyev et al., 2020). While such approaches can reduce per-sample costs, typically samples are barcoded at a population level and thus do not enable labeling of cell subpopulations within a sample.

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

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

Results

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

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

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

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

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

180 To determine whether TaG-EM could be used to measure age-dependent fecundity, we
181 collected flies from twelve different TaG-EM barcode lines at four time points separated by one
182 week (three barcode lines per timepoint). We collected eggs from these fly lines individually and
183 scored the number of viable eggs per female. Next, we pooled the barcoded flies from all
184 timepoints and collected eggs from the pooled flies. These eggs were aged, DNA was
185 extracted, and the TaG-EM barcodes were amplified and sequenced. While measurements from
186 individual barcode lines were noisy, both for manual counts and sequencing based
187 measurements, there was a general trend toward declining fecundity over time (Supplemental
188 Figure 4), consistent with published reports (David et al., 1975). Manually scored viable egg
189 numbers and TaG-EM barcode abundances were well correlated across two independent
190 experimental trials (R^2 values of 0.52-0.61 for Trial 1 and 0.74-0.84 for Trial 2). When barcodes
191 from each individual timepoint were averaged, R^2 values for the correlation between manual and
192 sequencing-based measurements were 0.95 for Trial 1 and 0.99 for Trial 2 (Figure 3C,
193 Supplemental Figure 4).
194
195 **Tissue-specific expression of TaG-EM GFP constructs**
196 To facilitate representation of the TaG-EM barcodes in single-cell sequencing data,
197 genetic barcodes were placed just upstream of the polyadenylation signal sequences and polyA
198 cleavage sites (Figure 4A). To verify that the inserted sequences did not interfere with Gal4-
199 driven GFP expression, we crossed each of the barcoded TaG-EM lines to *decapentaplegic-*
200 *Gal4* (*dpp-Gal4*). We observed GFP expression in the expected characteristic central stripe
201 (Teleman and Cohen, 2000) in the wing imaginal disc for 19/20 lines at similar expression levels
202 to the base pJFRC12 *UAS-myr::GFP* construct inserted in the same landing site (Figure 4B,
203 Supplemental Figure 6). No GFP expression was visible for TaG-EM barcode number 8. Gal4-
204 driven expression levels of TaG-EM barcoded GFP constructs were also similar to that of the
205 pJFRC12 base construct for multiple driver lines (Supplemental Figure 7) indicating that the
206 presence of the barcode does not generally impair expression of GFP.
207
208 **Boosting the GFP signal of TaG-EM constructs to enable robust cell sorting**
209 While with some driver lines, expression of the myr::GFP from the TaG-EM construct
210 may be too weak to allow robust enrichment of the tagged cells, adding an additional hexameric
211 GFP construct (Shearin et al., 2014) could boost expression of weak driver lines to levels that
212 are sufficient for robust detection of labeled flies or larvae (Figure 4C) and for labeling of
213 dissociated cells for flow cytometry (Supplemental Figure 8). Stocks with an additional UAS

214 hexameric GFP construct recombined onto the same chromosome as the TaG-EM construct
215 have been established for all 20 TaG-EM barcode lines.
216
217 **Correlation between expression of TaG-EM barcodes and intestinal cell marker genes in**
218 **single-cell sequencing data**
219 To test whether we could detect TaG-EM barcodes in single-cell sequencing data, we
220 crossed three TaG-EM barcode lines to two different gut Gal4 driver lines (Ariyapala et al.,
221 2020), one expressing in the enterocytes (EC-Gal4) and the other in intestinal precursor cells
222 (PC-Gal4), including stem cells and enteroblasts (EBs). Due to weak GFP expression with the
223 EC-Gal4 driver, we did not see visible GFP positive cells for this driver line. The PC-Gal4 driver
224 line contained an additional UAS-Stinger (2xGFP) construct and expressed GFP at a level
225 sufficient for flow sorting when crossed to the TaG-EM line (Supplemental Figure 7). Larval guts
226 were dissected, dissociated, stained with propidium iodide (PI) to label dead cells, and flow
227 sorted to recover PI negative and GFP positive cells. Approximately 10,000 cells were loaded
228 into a 10x Genomics droplet generator and a single-cell library was prepared and sequenced.
229 Two clusters were observed in the resulting sequencing data, one of which had high read
230 counts from mitochondrial genes suggesting that this cluster consisted of mitochondria, debris,
231 or dead and dying cells, after filtering the cells with high mitochondrial reads, a single cluster
232 remained (Supplemental Figure 9). This cluster expressed known intestinal precursor cell
233 markers such as *escargot* (*esg*), *klumpfuss* (*klu*), and *Notch* pathway genes like *E(spl)mbeta-*
234 *HLH* (Supplemental Figure 9). Expression of all three PC-Gal4 driven TaG-EM barcodes was
235 observed in this cluster (Supplemental Figure 9) indicating that TaG-EM barcodes can be
236 detected in single-cell sequencing data. Interestingly, TaG-EM barcode 8, for which no GFP
237 expression was observed was represented in the single-cell sequencing data indicating that the
238 insertion of this specific barcode sequence may affect GFP translation, but not mRNA
239 expression for this line.
240 A previous study used droplet-based single-cell sequencing to characterize the cell
241 types that make up the adult midgut (Hung et al., 2020). This study took advantage of two
242 fluorescent protein markers, an *escargot* (*esg*)-GFP fusion protein and a *prospero* (*pros*)-Gal4
243 driven RFP to label the intestinal stem cells (ISCs) and enteroendocrine cells (EECs),
244 respectively (Hung et al., 2020). The authors compared the resulting clusters to a list of known
245 marker genes in the literature, including antibody staining, GFP, LacZ, and Gal4 reporter
246 expression patterns to classify the cells in individual clusters, and also found that the *esg*-GFP
247 expression was present in a broader subset of cells than anticipated. Thus, most of these cell

248 classifications relied upon inference as opposed to direct positive labeling. Recently, a large
249 collection of split-*Gal4* lines were screened for expression in the adult and larval gut (Ariyapala
250 et al., 2020). These include pan-midgut driver lines, split-*Gal4* lines specific for the EBs, ECs,
251 EECs, and ISC/EBs, as well as driver lines with regionalized gene expression. We crossed four
252 different TaG-EM barcode lines with the pan-midgut driver (PMG-Gal4), and one barcode line to
253 each of the precursor cell (PC-Gal4), enterocyte (EC-Gal4), enteroblast (EB-Gal4) and
254 enteroendocrine (EE-Gal4) drivers (Ariyapala et al., 2020). Larval guts were dissected from
255 these lines and cells were dissociated, flow sorted as described above to select live, GFP
256 positive cells, and approximately 30,000 cells were loaded into a 10x Genomics droplet
257 generator for single-cell sequencing. Using the additional hexameric GFP construct to boost
258 GFP expression resulted in visible fluorescent signal for all eight barcode Gal4 line
259 combinations.

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

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

277 As expected, the barcodes expressed by the pan-midgut driver were broadly distributed
278 across the cell clusters (Supplemental Figure 13). In contrast, expression of the cell type-
279 specific barcodes showed more restricted patterns of expression among the cell clusters and
280 were co-localized with known marker genes for these cell types (Figure 5C-N). For instance,
281 TaG-EM barcode 4 expression (Figure 5F), which was driven by the EC-Gal4 line, was seen in

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

302 **Discussion**

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

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

322 In addition, we show that TaG-EM barcodes can be expressed by tissue-specific Gal4
323 drivers and used to tag specific cell populations upstream of single-cell sequencing (Figures 4-
324 5). This capability will allow for positive identification of cell clusters in cell atlas projects and will
325 facilitate multiplexing of single-cell sequencing experiments. Recently, a conceptually similar
326 approach called RABID-Seq (Clark et al., 2021) has been described, which allows trans-
327 synaptic labeling of neural circuits using barcoded viral transcripts. However, one distinction
328 between the two approaches is that RABID-Seq relies on stochastic viral infection of
329 mammalian cells while TaG-EM allows reproducible targeting of defined cell populations
330 allowing unambiguous cell identification and potentially allowing the same cell populations to be
331 assessed at different timepoints or in the context of different experimental manipulations. One
332 current limitation is that TaG-EM barcodes are not observed in all cells in single-cell gene
333 expression data and low expressing driver lines may result in particularly sparse labelling. In
334 antibody-conjugated oligo cell hashing approaches, this sparsity of barcode representation is
335 overcome by spiking in an additional primer at the cDNA amplification step and amplifying the

336 hashtag oligo by PCR. A similar enrichment approach could be employed to generate dense
337 labeling of the TaG-EM barcodes, improving upon the current mapping of barcode-containing
338 transcripts observed in the gene expression data, and facilitating comparison between
339 replicates and robust detection of multiplets.
340 In the future, generation of additional TaG-EM lines will enable higher levels of
341 multiplexing. In addition, while the original TaG-EM lines were made using a membrane-
342 localized myr::GFP construct, variants that express GFP in other cell compartments such as the
343 cytoplasm or nucleus could be constructed to enable increased expression levels or purification
344 of nuclei. Nuclear labeling could also be achieved by co-expressing a nuclear GFP construct
345 with existing TaG-EM lines in analogy to the use of hexameric GFP described above.
346 In summary, combined with the large collections of Gal4 and split-Gal4 lines that have
347 been established in Drosophila that enable precise targeting of a high proportion of cell types
348 (Ariyapala et al., 2020; Aso et al., 2014; Davis et al., 2020; Gohl et al., 2011; Namiki et al., 2018;
349 Pfeiffer et al., 2010, 2008; Venken et al., 2011), TaG-EM provides a means to target and label
350 cells *in vivo* for subsequent detection in single-cell sequencing. Moreover, these genetic
351 barcodes can be used to multiplex behavioral or other phenotypic measurements. Thus, TaG-
352 EM provides a flexible system for barcoding cells and organisms.
353

354 **Methods**

355

356 **Drosophila stocks and maintenance**

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

359

360 **Stock name**	**Source**
361 IsoD1	Clandinin Lab, Stanford University
362 *w*-; +; + (IsoD1)	Clandinin Lab, Stanford University
363 *attP2* line	BDSC #25710
364 *norpA*	BDSC# 9048
365 UAS-myr::GFP (pJFRC12)	BDSC #32197
366 Hexameric GFP lines	BDSC #91402, #91403
367 dpp-Gal4 driver	BDSC #1553
368 Act-Gal4 driver	BDSC #4414
369 Tub-Gal4 driver	BDSC #5138
370 Mhc-Gal4 driver	BDSC #55132
371 PC-Gal4 driver	BDSC #91400 + #91401
372 EC-Gal4 driver	BDSC #91405
373 EB-Gal4 driver	BDSC #91398 + #91404
374 EE-Gal4 driver	BDSC #91402 + #91403
375 PMG-Gal4 driver	BDSC #91398 + #91399
376 TaG-EM lines (x20)	This study
377 TaG-EM lines + 6xGFP (x20)	This study

378

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

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

383 caaaggaaaaagctgcactgctatacaagaaaattatggaaaaatatttgatgtatagtgccttgactagagatcataatcagccata
384 ccacatttgtagaggttttacttgctttaaaaaacctcccacacctccccctgaacctgaaacataaaatgaatgcaattgttgttgttaact
385 tgtttattgcagcttataaCTTCCAACAACCGGAAGTGANNNNNNNNNNNNNNNtggttaca<u>aataaa</u>gcaatag
386 catcacaaatttcaca<u>aataaa</u>gcatttttttcactg**<u>ca</u>**ttctagttgtggtttgtccaaactcatcaatgtatcttatcatgtctggatcgatct
387 ggccggccgtttaaacgaattcttgaagacgaaagggcctcgtgatacgcctatttttataggttaatgtcatgataataatg

388

389 The gBlock was resuspended in 20 µl EB, incubated at 50ºC for 20 minutes and then cut with
390 *Psi*I and *EcoR*I (New England Biolabs, NEB) using the following reaction conditions: 4 µl gBlock
391 DNA (35 ng), 2 µl 10x CustSmart buffer (NEB), 1 µl *Eco*RI enzyme (NEB), 1 µl *Psi*I enzyme
392 (NEB), and 12 µl nuclease-free water were mixed and incubated at 37ºC for 1 hour followed by
393 65ºC for 20 minutes to heat inactivate the restriction enzymes. pJFRC12-10XUAS-IVS-
394 myr::GFP plasmid (Addgene, Plasmid #26222) (Pfeiffer et al., 2010) was digested with the
395 following reaction conditions: 5 µl pJFRC12-10XUAS-IVS-myr::GFP plasmid DNA (~3 µg), 5 µl
396 10x CutSmart buffer (NEB), 1 µl *Psi*I enzyme (NEB), 1 µl *Eco*RI enzyme (NEB), and 38 µl
397 nuclease-free water, were mixed and incubated at 37ºC for 1 hour, followed by addition of 1 µl
398 of CIP and incubation for an additional 30 minutes. The digested vector backbone was gel
399 purified using the QiaQuick Gel Purification Kit (Qiagen). The digested gBlock was ligated into
400 the digested pJFRC12-10XUAS-IVS-myr::GFP backbone using the following reactions
401 conditions: 4 µl T4 ligase buffer (10x) (NEB), 20 µl plasmid backbone DNA (0.005 pmol), 5 µl
402 gBlock digest DNA (0.03 pmol), 2 µl of T4 DNA ligase (NEB), and 9 µl nuclease-free water were
403 mixed and incubated at 22ºC for 2 hours. 2 µl of the ligation reaction was transformed into 50 µl
404 of TOP10 competent cells (Invitrogen), and the cells were incubated on ice for 30 minutes, then
405 heat shocked at 42ºC for 30 seconds, and incubated on ice for 5 minutes. 250 µl SOC was
406 added and the cells were plated on LB+ampicillin plates and incubated overnight at 37ºC. DNA
407 was isolated from 36 pJFRC12-gBlock colonies using a QIAprep Spin MiniPrep kit (Qiagen).
408 Expected construct size was verified by diagnostic digest with *Eco*RI and *Apa*LI. DNA
409 concentration was determined using a Quant-iT PicoGreen dsDNA assay (Thermo Fisher
410 Scientific) and the randomer barcode for each of the constructs was determined by Sanger
411 sequencing using the following primers:
412 SV40_post_R: GCCAGATCGATCCAGACATGA
413 SV40_5F: CTCCCCCTGAACCTGAAACA
414

415 **Generation of TaG-EM transgenic lines**

416 29 sequence verified constructs were normalized, pooled evenly, and injected as a pool into
417 embryos (Rubin and Spradling, 1982) expressing PhiC31 integrase and the carrying the *attP2*
418 landing site (BDSC #25710). Injected flies were outcrossed to w- flies, and up to three *white*+
419 progeny per cross were identified, and the transgenic lines were homozygosed. DNA was
420 extracted (GeneJET genomic DNA purification Kit, Thermo Scientific) and the region containing
421 the DNA barcode was amplified with the following PCR reaction: 2.5 µl 1:10 diluted template

422 DNA, 2 µl 10x Reaction Buffer (Qiagen), 0.2 µl dNTP mix (10 µM), 1 µl 10 µM SV40_5F primer
423 (10 µM), 1 µl SV40_post_R primer (10 µM), 0.8 µl $MgCl_2$ (3 mM), 0.1 µl Taq polymerase
424 (Qiagen), 12.4 µl nuclease-free water. Reactions were amplified using the following cycling
425 conditions: 95ºC for 5 minutes, followed by 30 cycles of 94ºC for 30 seconds, 55ºC for 30
426 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes. PCR products were treated with
427 Exo-CIP using the following reaction conditions: 5 µl PCR product, 1 µl Exo-CIP Tube A (NEB),
428 1 µl Exo-CIP Tube B (NEB) were mixed and incubated at 37ºC for 4 minutes, followed by 80ºC
429 for 1 minute. The barcode sequence for each of the independent transgenic lines was
430 determined by Sanger sequencing using the SV40_5F and SV40_PostR primers. Transgenic
431 lines containing 20 distinct DNA barcodes were recovered (Supplemental Figure 1).
432
433 **Optimizing amplification of TaG-EM barcodes for next-generation sequencing**
434 The following primers were evaluated to amplify the TaG-EM barcodes upstream of NGS:
435 Forward primer pool; four primers with frameshifting bases to increase library sequence
436 diversity in initial sequencing cycles were normalized to 10 µM and pooled evenly to make a
437 B2_3'F1_Nextera_0-6 primer pool:
438 B2_3'F1_Nextera:
439 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCCAACAACCGGAAG*TGA
440 B2_3'F1_Nextera_2
441 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCTTCCAACAACCGGAAG*TGA
442 B2_3'F1_Nextera_4
443 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGACTTCCAACAACCGGAAG*TGA
444 B2_3'F1_Nextera_6
445 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGAGCTTCCAACAACCGGAAG*TGA
446 The following two reverse primers were tested:
447 SV40_pre_R_Nextera:
448 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTTGTGAAATTTGTGATGCTATTGC*T
449 TT
450 SV40_post_R_Nextera:
451 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAGATCGATCCAGACA*TGA
452 SV40_pre_R_nextera is designed to produce a shorter amplicon (200 bp with Illumina adapters
453 and indices added) and SV40-post_R_Nextera is designed to produce a longer amplicon (290
454 bp with Illumina adapters and indices added).
455

456 An initial test was performed with three different polymerases (Qiagen Q5, KAPA HiFi, and
457 Qiagen Taq) at two different annealing temperatures and with both the B2_3'F1_Nextera/
458 SV40_pre_R_Nextera and B2_3'F1_Nextera/ SV40_post_R_Nextera primer sets to determine
459 whether the primers amplify as expected (Supplemental Figure 1). Two different samples were
460 tested:
461 1. Pool of 8 putative transformant samples (pooled 5 ul each of 1:10 diluted sample)
462 2. OreR (wild type - diluted 1:10)
463 Set up the following PCR reactions:
464 Q5 polymerase: 2.5 µl template DNA, 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer
465 (10 µM), 10 µl 2x Q5 Master Mix (Qiagen), 5.5 µl nuclease-free water. Reactions were amplified
466 using the following cycling conditions: 98ºC for 30 seconds, followed by 30 cycles of 98ºC for 20
467 seconds, 55ºC or 60ºC for 15 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes.
468 KAPA HiFi polymerase: 2.5 µl template DNA, 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse
469 primer (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free water. Reactions
470 were amplified using the following cycling conditions: 95ºC for 5 minutes, followed by 30 cycles
471 of 98ºC for 20 seconds, 55ºC or 60ºC for 15 seconds, 72ºC for 30 seconds, followed by 72ºC for
472 5 minutes.
473 Taq polymerase: 2.5 µl template DNA, 2 µl 10x Reaction Buffer (Qiagen), 0.2 µl dNTP mix (10
474 µM), 1 µl 10 µM Forward primer (10 µM), 1 µl Reverse primer (10 µM), 0.8 µl $MgCl_2$ (3 mM), 0.1
475 µl Taq polymerase (Qiagen), 12.4 µl nuclease-free water. Reactions were amplified using the
476 following cycling conditions: 95ºC for 5 minutes, followed by 30 cycles of 94ºC for 30 seconds,
477 55ºC or 60ºC for 30 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes.
478 Samples were run on a 2% agarose gel to verify amplification products (Supplemental Figure 1).
479
480 Next, the TaG-EM barcode lines were pooled in either an even or staggered manner. To
481 optimize reaction conditions for the barcode measurements, either 5 ng or 50 ng of DNA was
482 amplified in triplicate for each pool for either 20, 25, or 30 cycles with either KAPA HiFi using the
483 B2_Nextera_F 0-6 forward primer pool together with either the SV40_pre_R_Nextera or the
484 SV40_post_R_Nextera reverse primer. Next, PCR reactions were diluted 1:100 in nuclease-free
485 water and amplified in the following indexing reactions: 3 µl PCR 1 (1:100 dilution), 1 µl indexing
486 primer 1 (5 µM), 1 µl indexing primer 2 (5 µM), and 5 µl 2x Q5 master mix. The following
487 indexing primers were used (X indicates the positions of the 8 bp indices):
488 Forward indexing primer:
489 AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC

490 Reverse indexing primer:
491 CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTCTCGTGGGCTCGG
492 Reactions were amplified using the following cycling conditions: 98°C for 30 seconds, followed
493 by 10 cycles of 98°C for 20 seconds, 55°C for 15 seconds, 72°C for 1 minute, followed by 72°C
494 for 5 minutes. Amplicons were then purified and normalized using a SequalPrep normalization
495 plate (Thermo Fisher Scientific), followed by elution in 20 µl of elution buffer. An even volume of
496 the normalized libraries was pooled and concentrated using 1.8x AmpureXP beads (Beckman
497 Coulter). Pooled libraries were quantified using a Qubit dsDNA high sensitivity assay (Thermo
498 Fisher Scientific) and libraries were normalized to 2 nM for sequencing on the Illumina MiSeq
499 (see below).
500
501 **Structured fly pool experiments**
502 Male or female flies from TaG-EM barcode lines were pooled in either an even or staggered
503 manner (Figure 2A). For the even pools, three independently pooled samples were constructed
504 in order to assess sample-to-sample variability. DNA was extracted from these structured pools
505 using a protocol adapted from Huang *et al*., 2009 (Huang et al., 2009), using homemade SPRI
506 beads (DeAngelis et al., 1995) in the last purification step and amplified in triplicate using 2.5 µl
507 template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl
508 SV40_post_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free
509 water. Reactions were amplified using the following cycling conditions: 98ºC for 5 minutes,
510 followed by 30 cycles of 98ºC for 20 seconds, 60ºC for 15 seconds, 72ºC for 30 seconds,
511 followed by 72ºC for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared
512 for sequencing as described above.
513
514 **Phototaxis experiments**
515 <u>Video-based measurements</u>
516 A pair of white LED strip lights with Muzata LED Strip Light Diffusers (U1SW WW 1M, LU1)
517 were mounted withing a light-tight box and controlled using an Vilros Uno Rev 3 microcontroller.
518 Test tubes containing flies were held in place with Acoustic Foam Panels (1”x12”x12”,
519 ALPOWL). Videos and images were acquired using an Arducam 1080P Day & Night Vision
520 USB Camera with an IR filter and using Photo Booth software (Apple). Wild type and *norpA* flies
521 carrying one of 4 different TaG-EM barcodes were tested in three independent experimental
522 replicates. 20 male flies of each genotype were transferred into 25 mm x 150 mm glass test
523 tubes, incubated at 34ºC for 10 minutes and then run in the phototaxis assay, where a light at

524 one end of the chamber was turned on for 30 seconds. Videos of all tests were recorded
525 through the end of the 30 second light pulse. Videos were independently scored by two
526 observers to determine the number of flies in the light-facing or dark-facing tubes and the results
527 were averaged. A preference index (P.I.) was calculated using the following formula: [(number
528 of flies in light tube) - (number of flies in dark tube)]/(total number of flies).
529
530 TaG-EM measurements
531 For TaG-EM barcode-based phototaxis measurements, the following genotypes were
532 consolidated into a single test tube:
533 Pool A
534 norpA/Y;TaG-EM BC4/+
535 norpA/Y;TaG-EM BC3/+
536 w-/Y;TaG-EM BC2/+
537 w-/Y;TaG-EM BC1/+
538
539 Pool B
540 norpA/Y;TaG-EM BC2/+
541 norpA/Y;TaG-EM BC1/+
542 w-/Y;TaG-EM BC4/+
543 w-/Y;TaG-EM BC3/+
544
545 These pools were individually incubated at 34ºC for 10 minutes and then run in the phototaxis
546 assay. Videos of all tests were recorded and at the end of a 30 second light pulse the two test
547 tubes were quickly separated and capped. Flies in each of these tubes were counted, then DNA
548 was extracted from the flies from the light-facing or dark-facing tubes and amplified using 2.5 µl
549 template DNA (50 ng), 1 µl 10 µM B2_Nextera_F 0-6 primers (10 µM), 1 µl
550 SV40_post_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix (Roche), 5.5 µl nuclease-free
551 water. Reactions were amplified using the following cycling conditions: 95ºC for 5 minutes,
552 followed by 30 cycles of 98ºC for 20 seconds, 60ºC for 15 seconds, 72ºC for 30 seconds,
553 followed by 72ºC for 5 minutes. Amplicons were indexed, normalized, quantified, and prepared
554 for sequencing as described above.
555
556 **Oviposition experiments**

557 Newly hatched flies (males and females) from three barcode lines were collected at one-week
558 intervals during 4 consecutive weeks (12 barcode lines in total). Fresh fly food was provided
559 every 3-4 days. Ten days after the last collection, 10 females from each barcode line were taken
560 and pooled together in a collection cage (10 females x 12 barcode lines= 120 females). The
561 remaining females from each barcode line were separated from the males and put in individual
562 collection cages. Two days later, the experiment started and run for 3 consecutive days. Each
563 day a 1-1.5 hour pre-collection was followed by a 6 hour collection, both at 25ºC. 100 embryos
564 from each individual collection plate were transferred to new plates and incubated for 2 days at
565 18ºC. The number of hatched larvae were counted and used to calculate the egg survival rate.
566 The pooled collection plate was also incubated at 18ºC and the next day the embryos were
567 dechorionated and frozen. The 12 individual collection plates were kept at 4ºC and the number
568 of embryos counted in the following days. For the barcode measurements, DNA was extracted
569 from the embryos, and amplified using 2.5 µl template DNA (50 ng), 1 µl 10 µM B2_Nextera_F
570 0-6 primers (10 µM), 1 µl SV40_post_R_Nextera (10 µM), 10 µl 2x KAPA HiFi ReadyMix
571 (Roche), 5.5 µl nuclease-free water. Reactions were amplified using the following cycling
572 conditions: 95ºC for 5 minutes, followed by 30 cycles of 98ºC for 20 seconds, 60ºC for 15
573 seconds, 72ºC for 30 seconds, followed by 72ºC for 5 minutes. Amplicons were indexed,
574 normalized, quantified, and prepared for sequencing as described above.
575
576 **Cell dissociation and isolation**
577 Midguts from 3rd instar larvae were dissected in phosphate-buffered saline (PBS) and
578 transferred to microcentrifuge tubes on ice containing PBS + 30% normal goat serum (NGS).
579 After dissection, 150 µL of 2.7 mg/mL elastase was added to each sample tube. The tubes were
580 then incubated at 27°C for one hour. During incubation, samples were mixed by pipetting ~30
581 times every 15 minutes to improve elastase dissociation of the cells. Samples were then filtered
582 through a 40 µM FlowMi tip filter (Bel-Art) to reduce debris. Afterwards, the samples were
583 quantified on the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems) using 9 µL of
584 sample to 1µL AO/PI dye to ensure there were enough viable cells for flow sorting.
585
586 Once quantified, the samples were brought up to a volume of ~1.1mL with the PBS + 30% NGS
587 solution to facilitate flow sorting. The samples were then fluorescently sorted on a FACSAria II
588 Cell Sorter (BD Biosciences) to isolate GFP+ cells. Following sorting, samples were centrifuged
589 at 300x g for 10 minutes to concentrate the cells. The supernatant was aspirated off until 50 µL
590 cell concentrate remained in each sample. Then, the samples were carefully resuspended using

591 wide bore pipette tips before being combined into one sample tube. This sample was quantified
592 on the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems) as described above. If
593 necessary, cells were centrifuged, concentrated, and re-counted.
594
595 **Preparation of single-cell sequencing libraries**
596 The resulting pool was prepared for sequencing following the 10X Genomics Single Cell 3'
597 protocol (version CG000315 Rev C). Custom nested primers were used to attempt enrichment
598 of TaG-EM barcodes after cDNA creation using PCR. The remaining cDNA followed the 10X 3'
599 protocol without modification while the enriched TaG-EM portion followed BioLegends' "Total-
600 Seq-A Antibodies and Cell Hashing with 10X Single Cell 3' Reagents Kit v3 or v3.1" Protocol
601 starting at step III. The resulting 3' gene expression library and TaG-EM enrichment library were
602 sequenced together following Scenario 1 of the BioLegends protocol. Later, the TaG-EM library
603 was sequenced alone following Scenario 2 from the same protocol.
604
605 **Sequencing**
606 Libraries for TaG-EM barcode analysis from structured pools or from phototaxis or oviposition
607 experiments were denatured with NaOH and prepared for sequencing according to the
608 protocols described in the Illumina MiSeq Denature and Dilute Libraries Guides. Single-cell
609 libraries were sequenced on the Illumina NextSeq 2000 or Illumina NovaSeq 6000.
610
611 **Data analysis**
612 Behavioral experiments: Demultiplexed fastq files were generated using bcl2fastq or bcl-
613 convert. TaG-EM barcode data was analyzed using custom R and Python scripts and BioPython
614 (Cock et al., 2009). Leading primer sequences were trimmed using cutadapt (Martin, 2011) and
615 the first 14 bp of the remaining trimmed read were compared to a barcode reference file, with a
616 maximum of 2 mismatches allowed, using a custom script which is available via Github:
617 https://github.com/darylgohl/TaG-EM.
618
619 Single-cell experiments:
620 Data was initially mapped and analyzed using the Cellranger software package (version 7.0.1,
621 10X Genomics). A custom reference Drosophila genome was created by starting with the
622 BDGP6.28 reference genome assembly and Ensembl gene annotations. Custom gene
623 definitions for each of the TaG-EM barcodes were added to the fasta genome file and gtf gene
624 annotation file and a Cellranger reference package was generated with the cellranger mkref

625 command. Subsequent analysis was performed using Scanpy (version 1.9.3) (Wolf et al., 2018).
626 Cells expressing less than 200 genes and genes expressed in fewer than three cells were
627 filtered from the expression matrix. Next, cells expressing more than 4,000 genes, cells
628 expressing more than 15% mitochondrial transcripts, and cells expressing more than 30%
629 ribosomal (RpL) genes were filtered out. Expression counts were log normalized and highly
630 variable genes were filtered out (min_mean=0.0125, max_mean=3, min_disp=0.5). Cells
631 expressing multiple TaG-EM barcodes were identified and removed and cells were clustered
632 using the Leiden algorithm (Traag et al., 2019) and projected with UMAP. Scanpy was used to
633 call and visualize differentially expressed genes.

Declarations

Availability of data, code, and materials

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

Competing interests

None to declare.

Funding

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

Author contributions

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

Acknowledgements

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

667 http://www.msi.umn.edu. Stocks obtained from the Bloomington Drosophila Stock Center (NIH
668 P40OD018537) were used in this study.
669

670 **References**

671 Ariyapala IS, Holsopple JM, Popodi EM, Hartwick DG, Kahsai L, Cook KR, Sokol NS. 2020.
672 Identification of split-GAL4 drivers and enhancers that allow regional cell type
673 manipulations of the drosophila melanogaster intestine. *Genetics* **216**:891–903.
674 doi:10.1534/genetics.120.303625
675 Aso Y, Hattori D, Yu Y, Johnston RM, Iyer NA, Ngo TTB, Dionne H, Abbott LF, Axel R,
676 Tanimoto H, Rubin GM. 2014. The neuronal architecture of the mushroom body provides
677 a logic for associative learning. *eLife* **3**:e04577. doi:10.7554/ELIFE.04577
678 Bhang HC, Ruddy DA, Krishnamurthy Radhakrishna V, Caushi JX, Zhao R, Hims MM, Singh
679 AP, Kao I, Rakiec D, Shaw P, Balak M, Raza A, Ackley E, Keen N, Schlabach MR,
680 Palmer M, Leary RJ, Chiang DY, Sellers WR, Michor F, Cooke VG, Korn JM, Stegmeier
681 F. 2015. Studying clonal dynamics in response to cancer therapy using high-complexity
682 barcoding. *Nature Medicine* **21**:440–448. doi:10.1038/nm.3841
683 Chen X, Sun Y-C, Zhan H, Kebschull JM, Fischer S, Matho K, Huang ZJ, Gillis J, Zador AM.
684 2019. High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ
685 Sequencing. *Cell* **179**:772-786.e19. doi:10.1016/j.cell.2019.09.023
686 Cheng J, Liao J, Shao X, Lu X, Fan X. 2021. Multiplexing Methods for Simultaneous Large-
687 Scale Transcriptomic Profiling of Samples at Single-Cell Resolution. *Advanced Science*
688 **8**:2101229. doi:10.1002/advs.202101229
689 Clark IC, Gutiérrez-Vázquez C, Wheeler MA, Li Z, Rothhammer V, Linnerbauer M, Sanmarco
690 LM, Guo L, Blain M, Zandee SEJ, Chao CC, Batterman KV, Schwabenland M, Lotfy P,
691 Tejeda-Velarde A, Hewson P, Polonio CM, Shultis MW, Salem Y, Tjon EC, Fonseca-
692 Castro PH, Borucki DM, de Lima KA, Plasencia A, Abate AR, Rosene DL, Hodgetts KJ,
693 Prinz M, Antel JP, Prat A, Quintana FJ. 2021. Barcoded viral tracing of single-cell
694 interactions in central nervous system inflammation. *Science* **372**.
695 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff
696 F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for
697 computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*
698 **25**:1422–3. doi:10.1093/bioinformatics/btp163
699 David J, Cohet Y, Fouillet P. 1975. The variability between individuals as a measure of
700 senescence: A study of the number of eggs laid and the percentage of hatched eggs in
701 the case of Drosophila melanogaster. *Experimental Gerontology* **10**:17–25.
702 doi:10.1016/0531-5565(75)90011-X
703 Davis FP, Nern A, Picard S, Reiser MB, Rubin GM, Eddy SR, Henry GL. 2020. A genetic,
704 genomic, and computational resource for exploring neural circuit function. *eLife* **9**.
705 DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the
706 isolation of PCR products. *Nucleic Acids Res* **23**:4742–4743.
707 Gohl DM, Silies MA, Gao XJ, Bhalerao S, Luongo FJ, Lin C-C, Potter CJ, Clandinin TR. 2011. A
708 versatile in vivo system for directed dissection of gene expression patterns. *Nature*
709 *Methods* **8**:231–237. doi:10.1038/nmeth.1561
710 Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of transgenic Drosophila by using
711 the site-specific integrase from phage phiC31. *Genetics* **166**:1775–82.
712 doi:10.1534/genetics.166.4.1775
713 Huang AM, Rehm EJ, Rubin GM. 2009. Quick preparation of genomic DNA from Drosophila.
714 *Cold Spring Harb Protoc* **2009**:pdb.prot5198. doi:10.1101/pdb.prot5198
715 Hung RJ, Hu Y, Kirchner R, Liu Y, Xu C, Comjean A, Tattikota SG, Li F, Song W, Sui SH,
716 Perrimon N. 2020. A cell atlas of the adult Drosophila midgut. *Proceedings of the*
717 *National Academy of Sciences of the United States of America* **117**:1514–1523.
718 doi:10.1073/pnas.1916820117

Ingham PW. 1988. The molecular genetics of embryonic pattern formation in Drosophila. *Nature* **335**:25–34. doi:10.1038/335025a0

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**:1187–1201. doi:10.1016/j.cell.2015.04.044

Kurmangaliyev YZ, Yoo J, Valdes-Aleman J, Sanfilippo P, Zipursky SL. 2020. Transcriptional Programs of Circuit Assembly in the Drosophila Visual System. *Neuron* **108**:1045-1057.e6. doi:10.1016/J.NEURON.2020.10.006

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, Zhang K, Church GM. 2015. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols* **10**:442–458. doi:10.1038/nprot.2014.191

Lein E, Borm LE, Linnarsson S. 2017. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**:64–69. doi:10.1126/science.aan6827

Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA, Brbić M, Spanier K, Leskovec J, McLaughlin CN, Xie Q, Jones RC, Brueckner K, Shim J, Tattikota SG, Schnorrer F, Rust K, Nystul TG, Carvalho-Santos Z, Ribeiro C, Pal S, Mahadevaraju S, Przytycka TM, Allen AM, Goodwin SF, Berry CW, Fuller MT, White-Cooper H, Matunis EL, DiNardo S, Galenza A, O'Brien LE, Dow JAT, FCA Consortium, 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

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

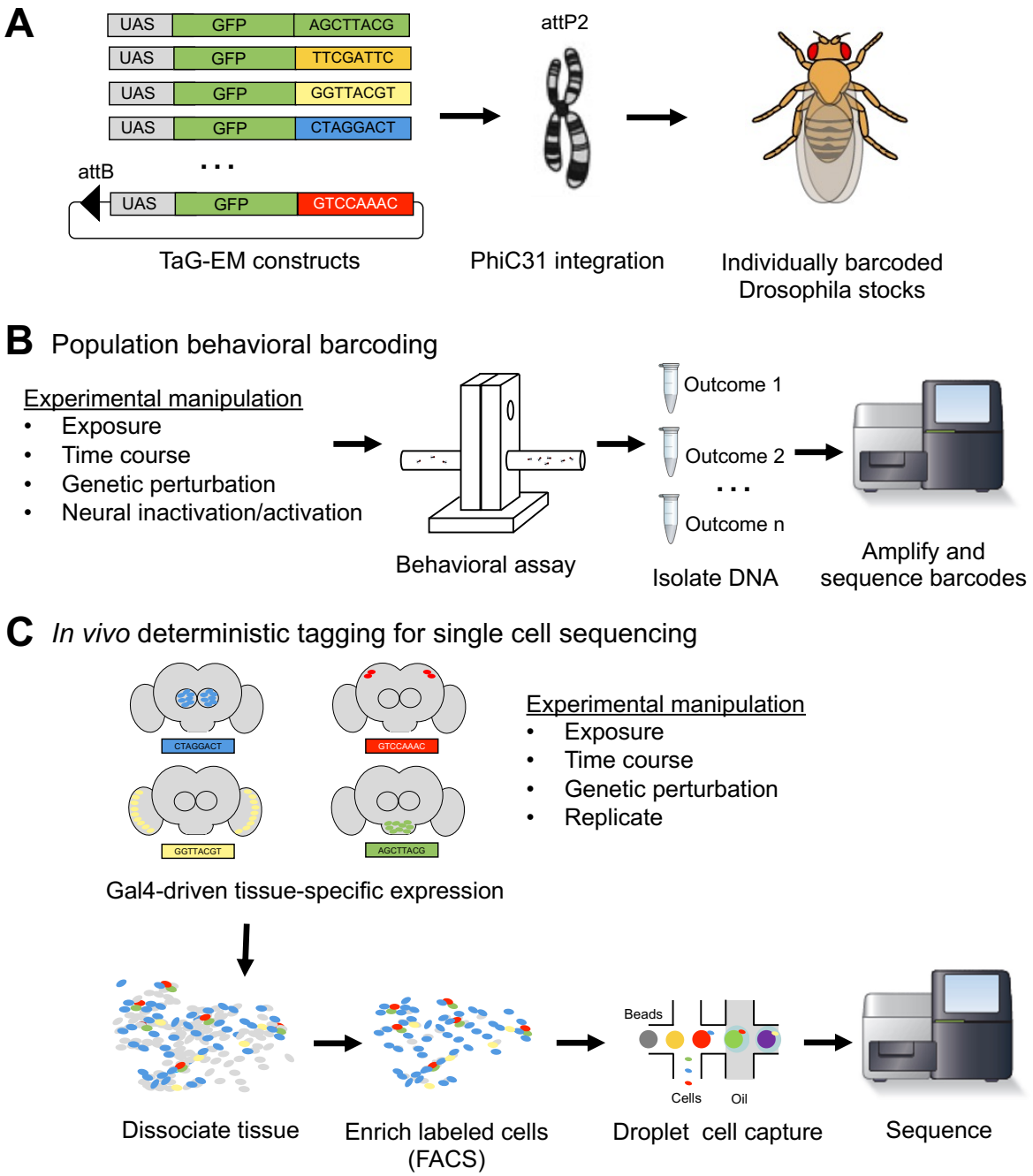
Pfeiffer BD, Ngo T-TB, Hibbard KL, Murphy C, Jenett A, Truman JW, Rubin GM. 2010. Refinement of tools for targeted gene expression in Drosophila. *Genetics* **186**:735–55. doi:10.1534/genetics.110.119917

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

770 Rubin GM, Spradling AC. 1982. Genetic transformation of Drosophila with transposable element
771 vectors. *Science (New York, NY)* **218**:348–53. doi:10.1126/science.6289436
772 Shearin HK, MacDonald IS, Spector LP, Steven Stowers R. 2014. Hexameric GFP and mCherry
773 reporters for the Drosophila GAL4, Q, and LexA transcription systems. *Genetics*
774 **196**:951–960. doi:10.1534/GENETICS.113.161141/-/DC1/GENETICS.113.161141-
775 5.PDF
776 Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow
777 C. 2009. Quantitative phenotyping via deep barcode sequencing. *Genome research*
778 **19**:1836–42. doi:10.1101/gr.093955.109
779 Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H,
780 Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in
781 single cells. *Nature Methods* **14**:865–868. doi:10.1038/nmeth.4380
782 Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, Smibert P, Satija R.
783 2018. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection
784 for single cell genomics. *Genome Biology* **19**:1–12. doi:10.1186/s13059-018-1603-1
785 Teleman AA, Cohen SM. 2000. Dpp gradient formation in the drosophila wing imaginal disc.
786 *Cell* **103**:971–980. doi:10.1016/S0092-8674(00)00199-9
787 Thompson CL, Ng L, Menon V, Martinez S, Lee C-K, Glattfelder K, Sunkin SM, Henry A, Lau C,
788 Dang C, Garcia-Lopez R, Martinez-Ferre A, Pombero A, Rubenstein JLR, Wakeman
789 WB, Hohmann J, Dee N, Sodt AJ, Young R, Smith K, Nguyen T-N, Kidney J, Kuan L,
790 Jeromin A, Kaykas A, Miller J, Page D, Orta G, Bernard A, Riley Z, Smith S, Wohnoutka
791 P, Hawrylycz MJ, Puelles L, Jones AR. 2014. A High-Resolution Spatiotemporal Atlas of
792 Gene Expression of the Developing Mouse Brain. *Neuron* **83**:309–323.
793 doi:10.1016/J.NEURON.2014.05.033
794 Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected
795 communities. *Sci Rep* **9**:5233. doi:10.1038/s41598-019-41695-z
796 van Opijnen T, Bodi KL, Camilli A. 2009. Tn-seq: high-throughput parallel sequencing for fitness
797 and genetic interaction studies in microorganisms. *Nature methods* **6**:767–72.
798 doi:10.1038/nmeth.1377
799 Venken KJT, Schulze KL, Haelterman NA, Pan H, He Y, Evans-Holm M, Carlson JW, Levis RW,
800 Spradling AC, Hoskins RA, Bellen HJ. 2011. MiMIC: a highly versatile transposon
801 insertion resource for engineering Drosophila melanogaster genes. *Nature methods*
802 **8**:737–43.
803 Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data
804 analysis. *Genome Biology* **19**:15. doi:10.1186/s13059-017-1382-0
805 Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD,
806 McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG,
807 Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R,
808 Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson
809 NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively
810 parallel digital transcriptional profiling of single cells. *Nature Communications* **8**:1–12.
811 doi:10.1038/ncomms14049
812

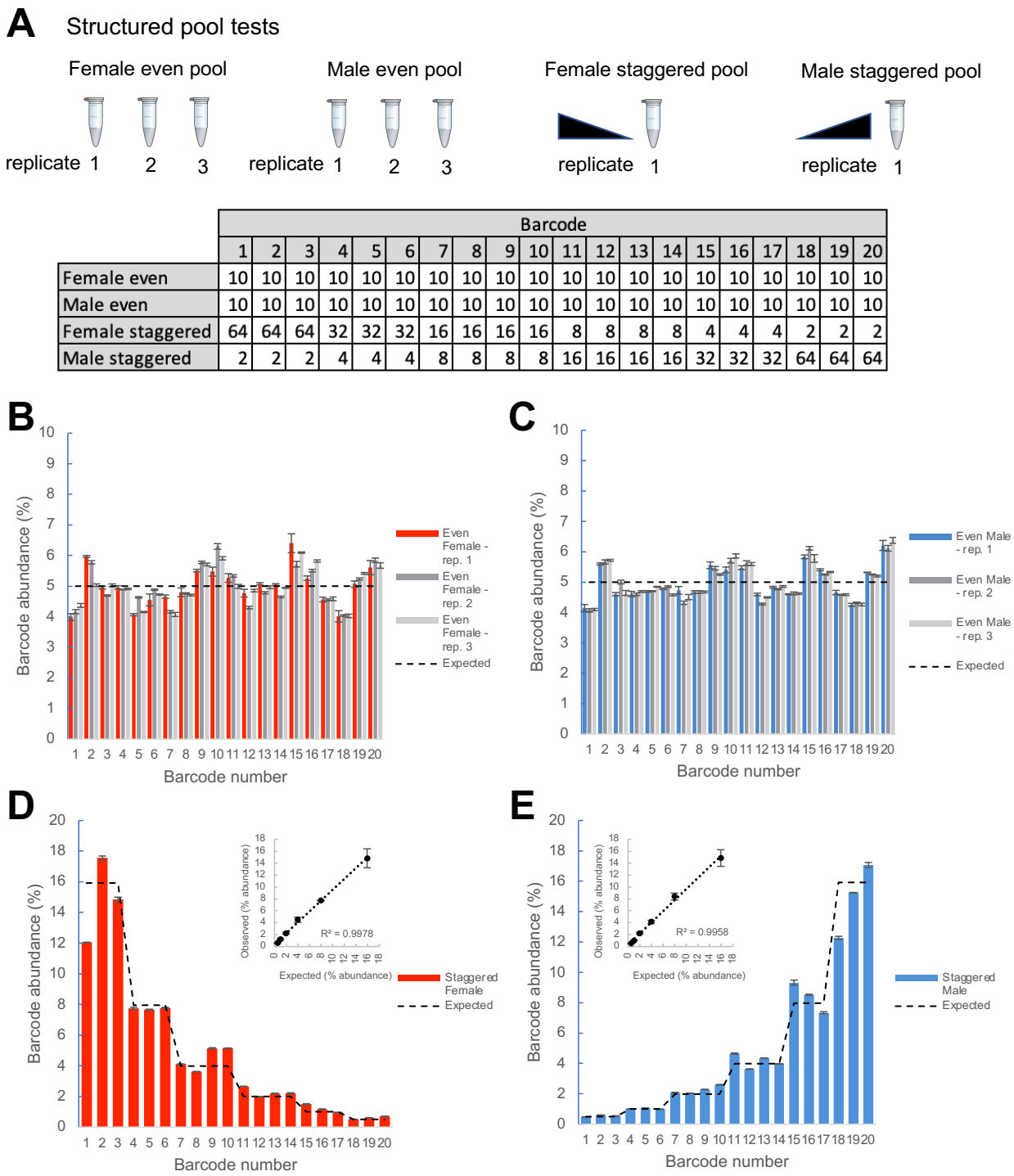
813

814 **Figures**



815

816 **Figure 1. Overview of TaG-EM system.** A) Schematic illustrating the design of the TaG-EM
817 constructs, where a barcode sequence is inserted in the 3' UTR of a UAS-GFP construct and
818 inserted in a specific genomic locus using PhiC31 integrase. B) Use of TaG-EM barcodes for
819 sequencing-based population behavioral assays. C) Use of TaG-EM barcodes expressed with
820 tissue-specific Gal4 drivers to label cell populations *in vivo* upstream of cell isolation and single-
821 cell sequencing.



	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 replicates of the female even pool. C) Barcode abundance data for three independent replicates

828 of the male even pool. D) Barcode abundance data for the female staggered pool. Inset plot
829 shows the average observed barcode abundance among lines pooled at each level compared
830 to the expected abundance. E) Barcode abundance data for the male staggered pool. Inset plot
831 shows the average observed barcode abundance among lines pooled at each level compared
832 to the expected abundance. For all plots, bars indicate the mean barcode abundance for three
833 technical replicates of each pool, error bars are +/- S.E.M.

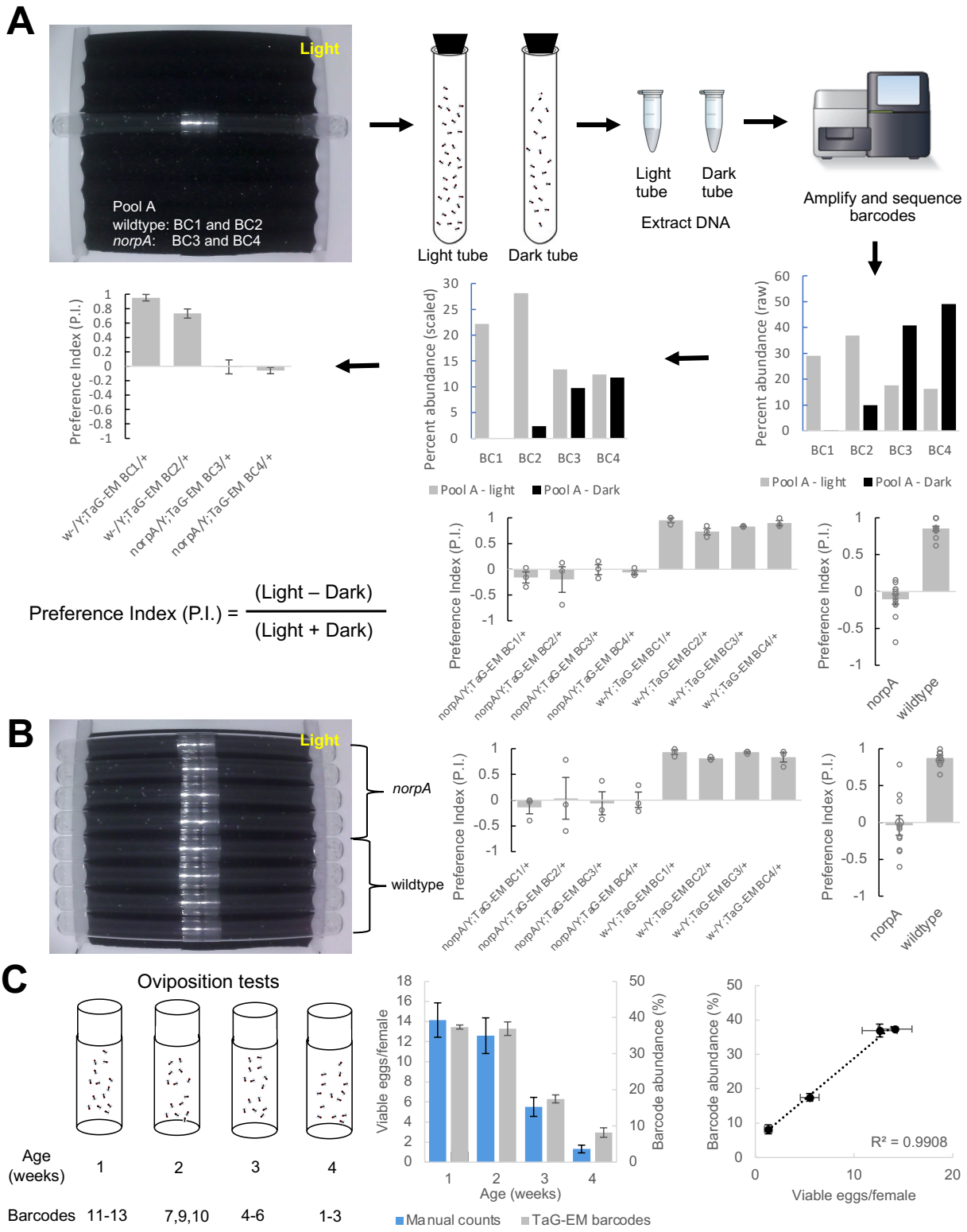


Figure 3. TaG-EM barcode-based behavioral measurements. A) TaG-EM barcode lines in either a wild-type or *norpA* background were pooled and tested in a phototaxis assay. After 30 seconds of light exposure, flies in tubes facing the light or dark side of the chamber were

838 collected, DNA was extracted, and TaG-EM barcodes were amplified and sequenced. Barcode
839 abundance values were scaled to the number of flies in each tube and used to calculate a
840 preference index (P.I.). Average P.I. values for four different TaG-EM barcode lines in both the
841 wild-type and *norpA* backgrounds are shown (n=3 biological replicates, error bars are +/-
842 S.E.M.). B) The same eight lines used for the sequencing-based TaG-EM barcode
843 measurements were independently tested in the phototaxis assay and manually scored videos
844 were used to calculate a P.I. for each genotype. Average P.I. values for each line are shown
845 (n=3 biological replicates, error bars are +/- S.E.M.). C) Flies carrying different TaG-EM
846 barcodes were collected and aged for one to four weeks and then eggs were collected and egg
847 number and viability was manually scored for each line. In parallel the barcoded flies from each
848 timepoint were pooled, and eggs were collected, aged, and DNA was extracted, followed by
849 TaG-EM barcode amplification and sequencing. Average number of viable eggs per female
850 (manual counts) and average barcode abundance are shown both as a bar plot and scatter plot
851 (n=3 biological replicates for 3 barcodes per condition, error bars are +/- S.E.M.).

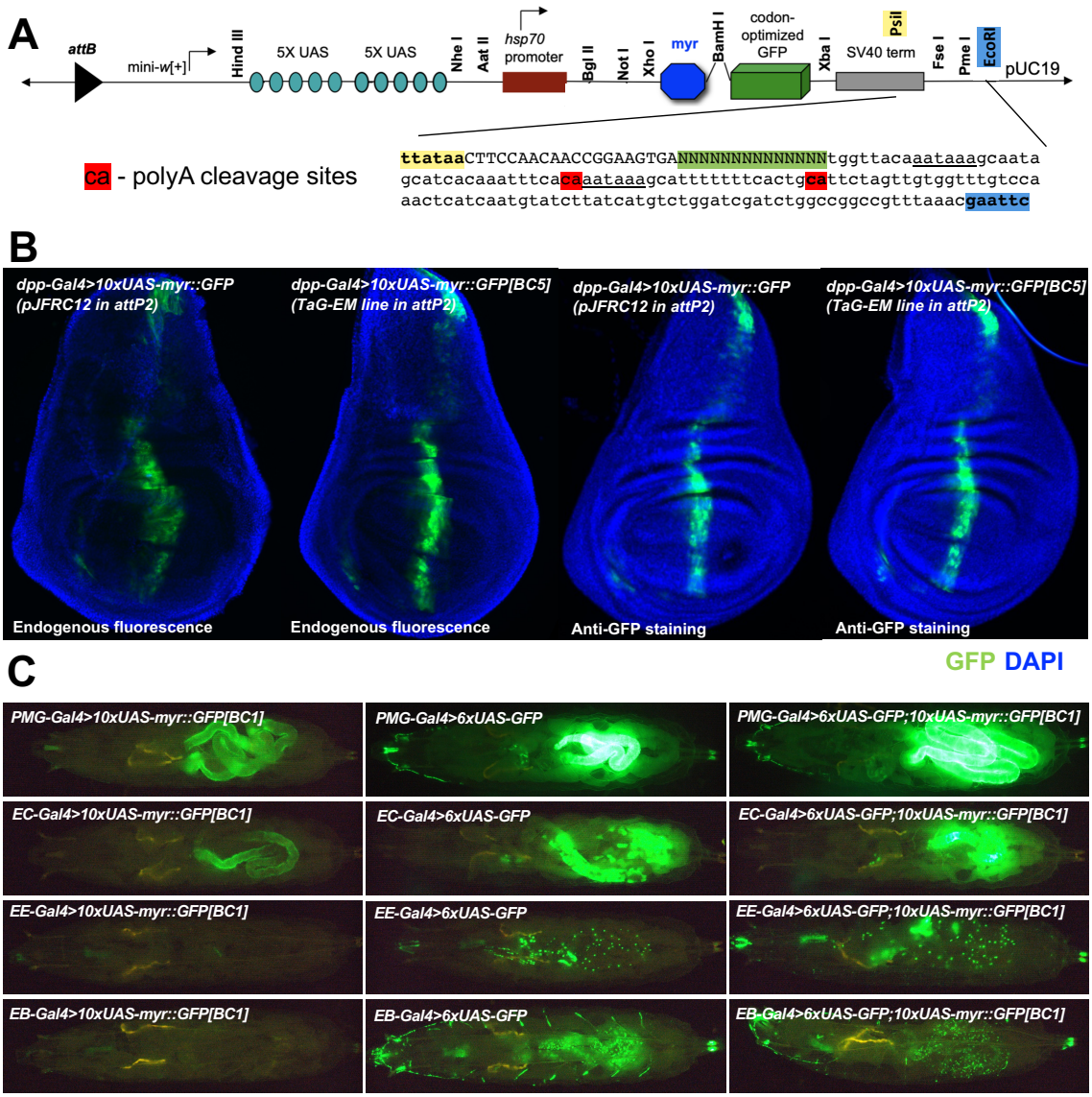


Figure 4. Gal4 driven expression of GFP from TaG-EM lines. A) Detailed view of the 3' UTR of the TaG-EM constructs showing the position of the 14 bp barcode sequence (green highlight) relative to the polyadenylation signal sequences (underlined) and polyA cleavage sites (red highlights). pJFRC12 backbone schematic is from (Pfeiffer et al., 2010). B) 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. C) Expression of GFP from either a TaG-EM barcode construct (left column), a hexameric GFP construct (middle column), or both 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).

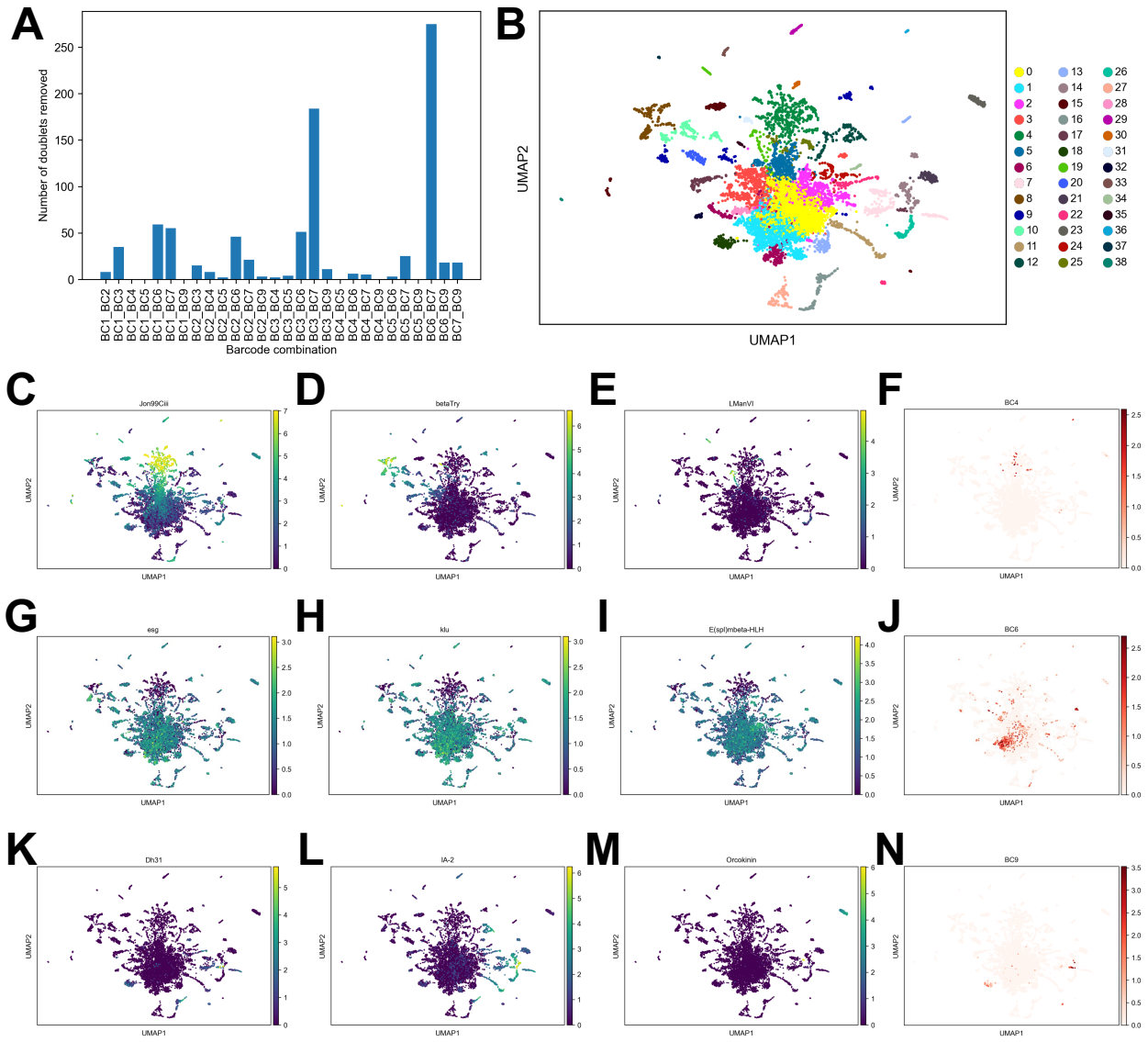


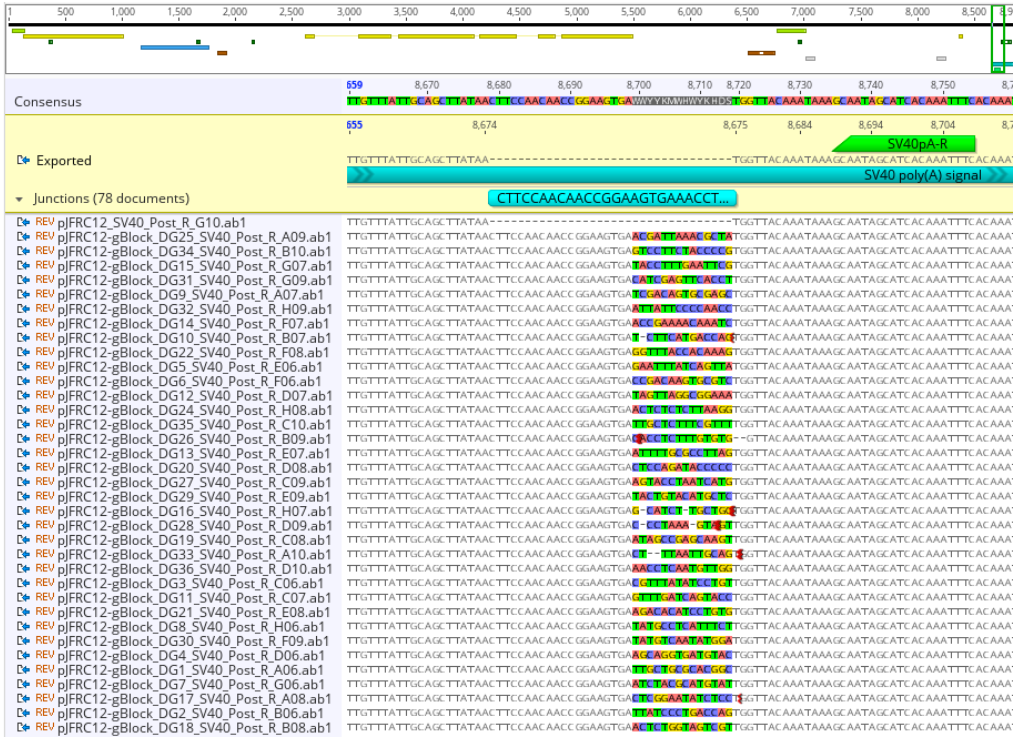
Figure 5. Expression of TaG-EM genetic barcodes in larval intestinal cell types. A) Number of doublets removed due to co-expression of each of the pairwise combinations of TaG-EM barcodes. B) UMAP plot of Drosophila larval gut cell clusters after TaG-EM barcode-based doublet removal. UMAP plots showing gene expression levels of enterocyte marker genes C) *Jon99Ciii*, D) *betaTry*, E) *LManVI*, and F) the TaG-EM barcode (BC4) driven by the EC-Gal4 line. UMAP plots showing gene expression levels of enteroblast/ISC marker genes G) *esg*, H) *klu*, I) *E(spl)mbeta-HLH*, and J) the TaG-EM barcode (BC6) driven by the EB-Gal4 line. UMAP plots showing gene expression levels of enteroendocrine cell marker genes K) *Dh31*, L) *IA-2*, M) *Orcokinin*, and N) the TaG-EM barcode (BC9) driven by the EE-Gal4 line.

Supplemental Figures

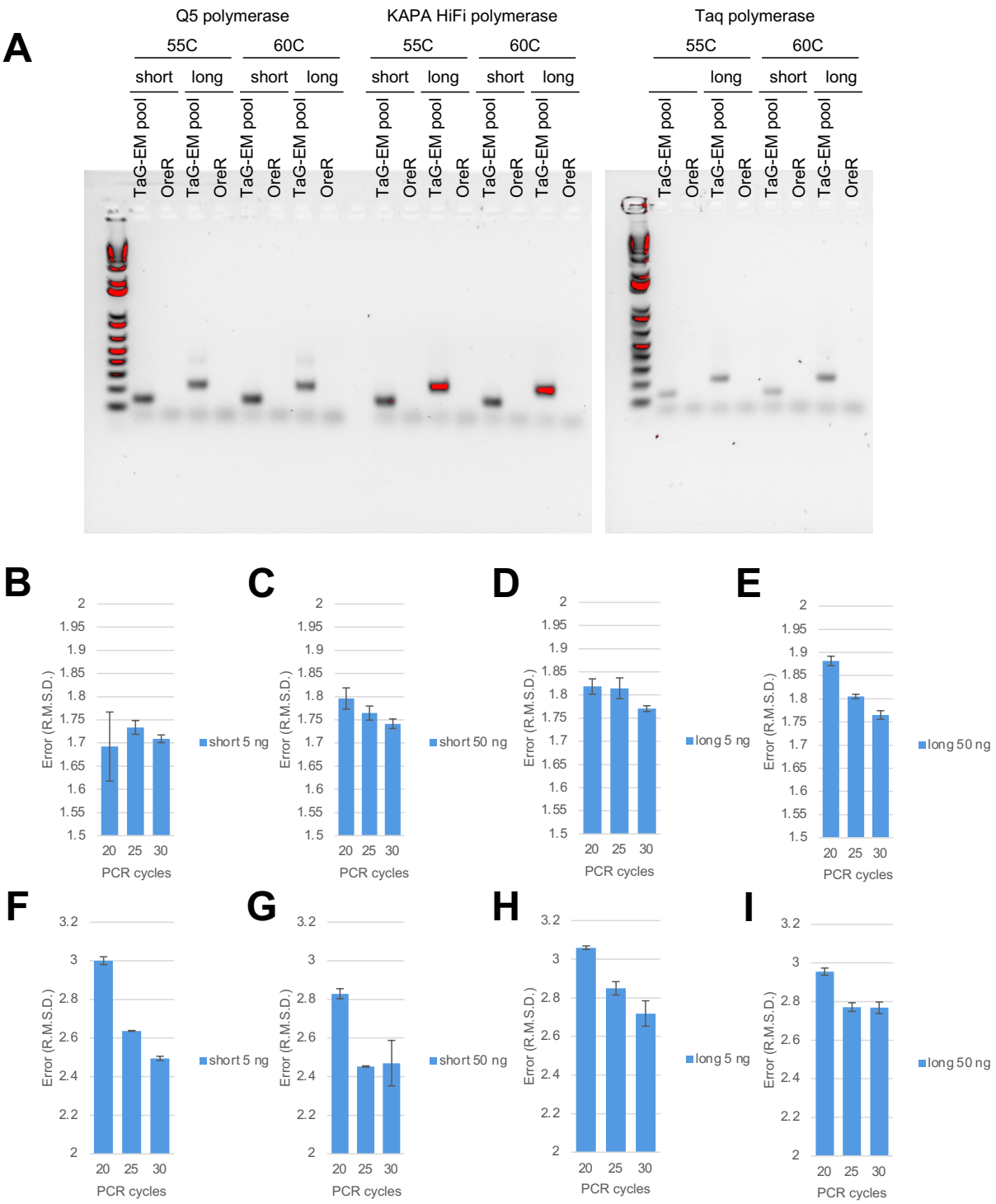
A

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



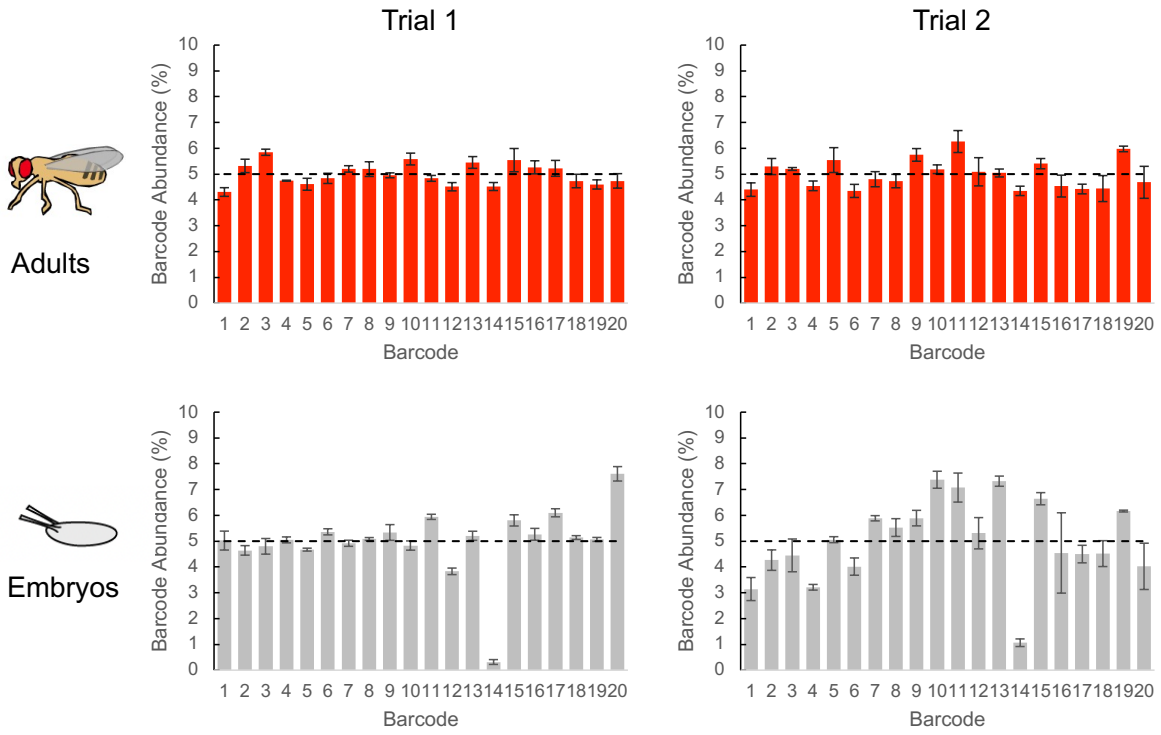
Supplemental Figure 1. Sanger sequencing identification of TaG-EM barcode lines. A) Summary of barcode pool injections. Barcode sequence and transgenic vial identifier in which the barcode was identified are shown. B) Sanger sequencing-based confirmation of the barcode sequence and PCR handle in TaG-EM transgenic lines.



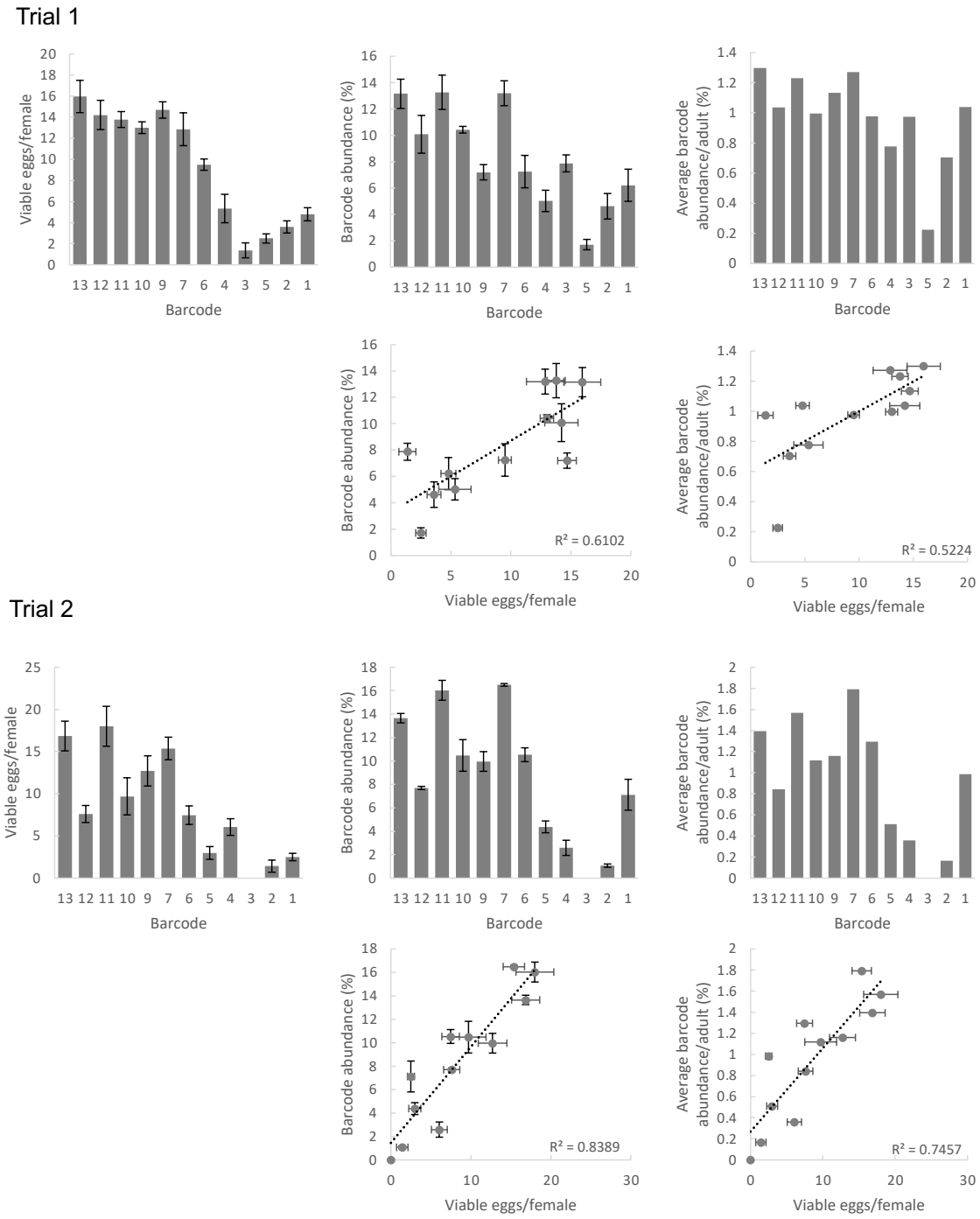
881

882 **Supplemental Figure 2. Optimization of TaG-EM barcode amplification.** A) Gels showing
883 bands produced when amplifying TaG-EM flies or a wild type control with the indicated
884 polymerase, annealing temperature, and primer pair (short = B2_3'F1_Nextera/
885 SV40_pre_R_Nextera; long = B2_3'F1_Nextera/ SV40_post_R_Nextera). B-E) Mean error

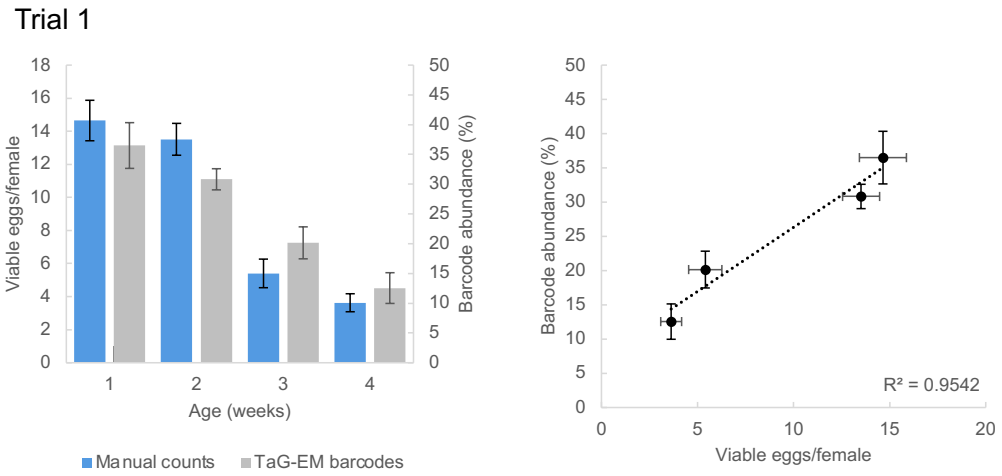
886 (R.M.S.D. root mean squared deviation from expected value) for even pool amplified with the
887 indicated primer set, input amount, and cycle number using KAPA HiFi polymerase (n=3, error
888 bars are +/- S.E.M.). F-G) Mean error (R.M.S.D. root mean squared deviation from expected
889 value) for staggered pool amplified with the indicated primer set, input amount, and cycle
890 number using KAPA HiFi polymerase (n=3, error bars are +/- S.E.M.).
891



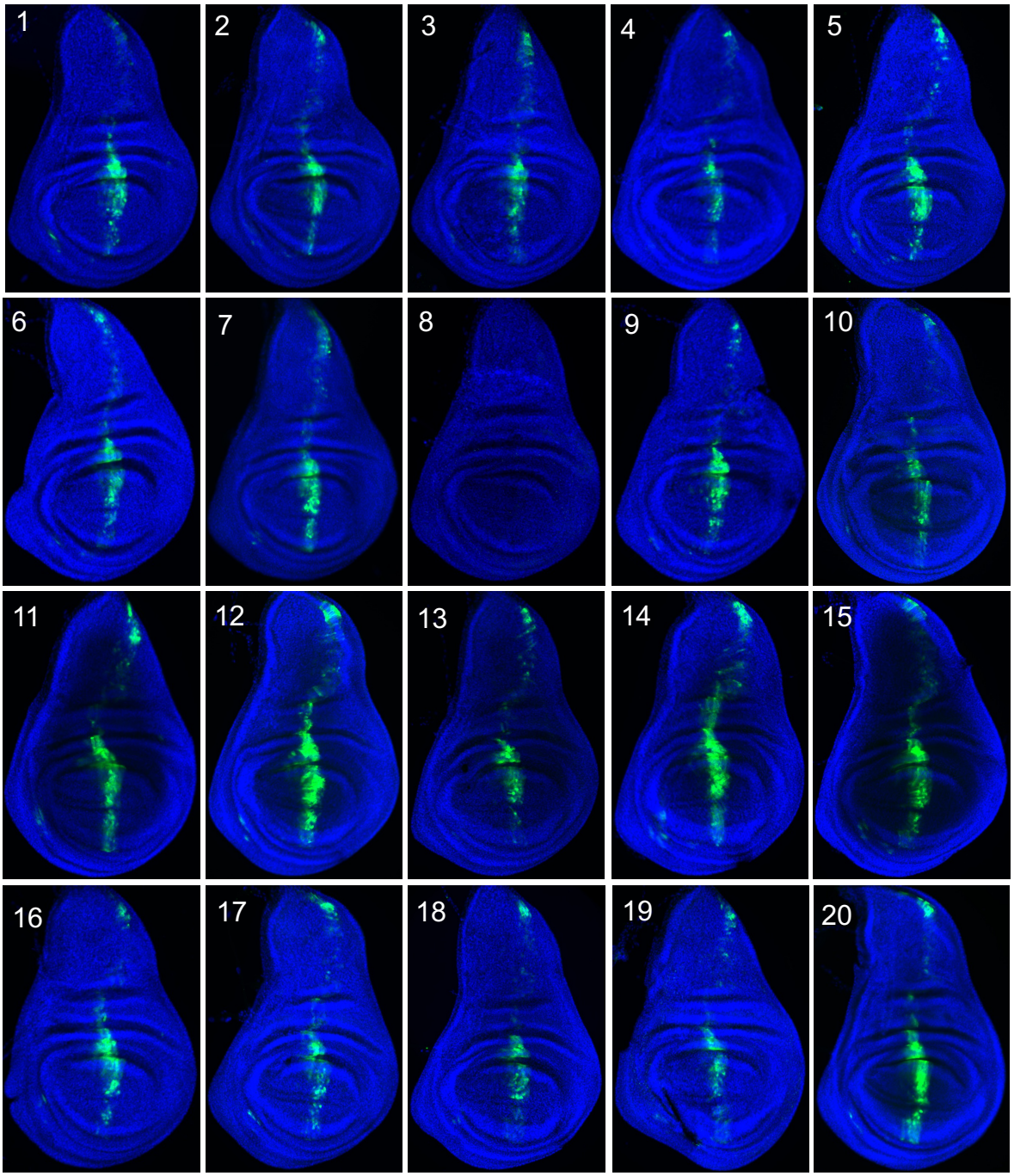
Supplemental Figure 3. Oviposition tests with TaG-EM barcode lines. Plots showing mean TaG-EM barcode abundance for adult females used in oviposition experiments (top) and eggs collected from these females (bottom). Data from two independent trials is shown (n=3 for each trial, error bars are +/- S.E.M.).



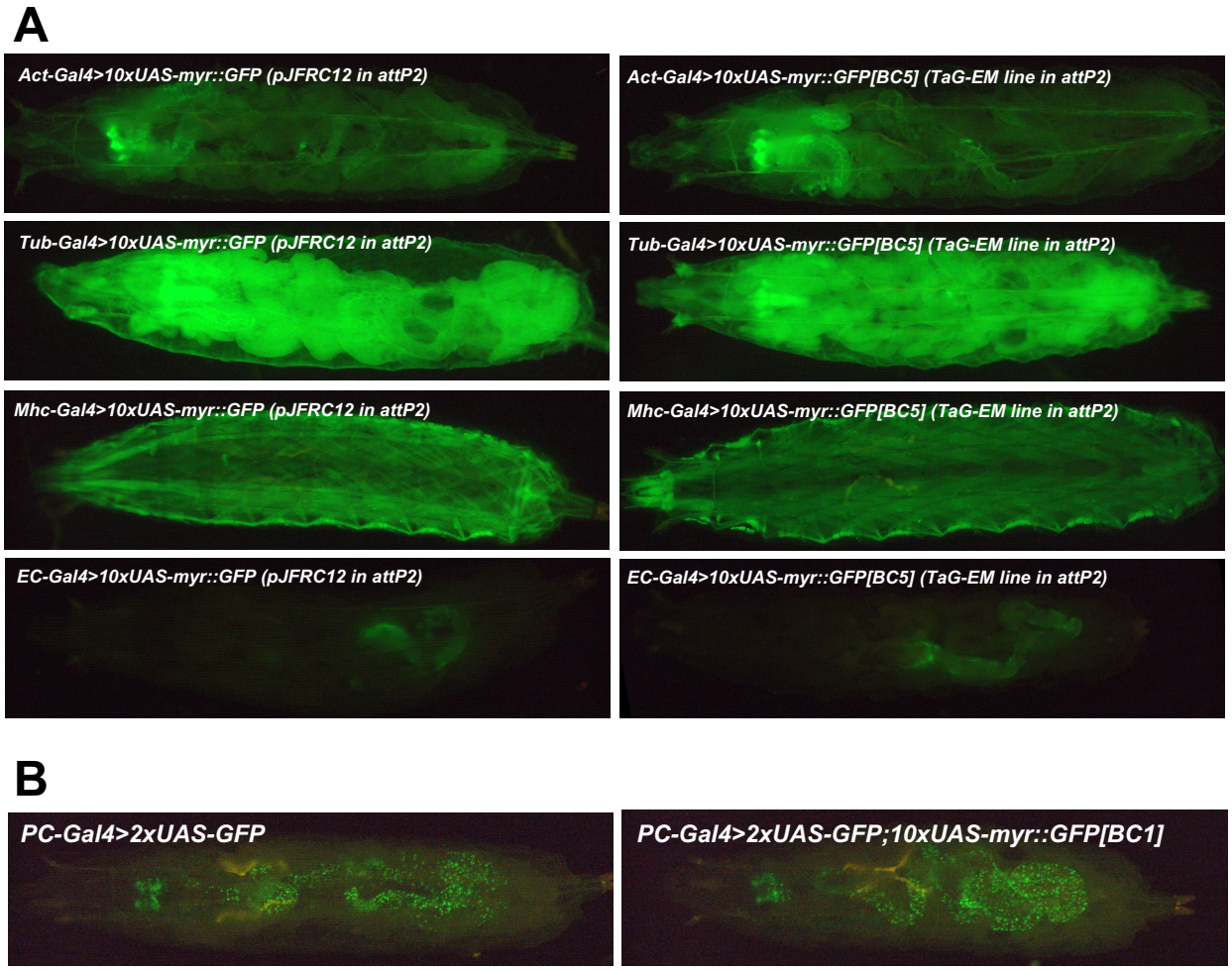
Supplemental Figure 4. Fecundity data for individual TaG-EM lines. Manually collected data for mean number of viable eggs per female, barcode abundance data, and barcode abundance data normalized to adult fly barcode data for each of the TaG-EM barcode lines used in the age-dependent fecundity experiment. Scatterplots show correlations between manually collected data and barcode sequencing results. Data from two independent trials is shown (n=3 for each trial, error bars are +/- S.E.M.).



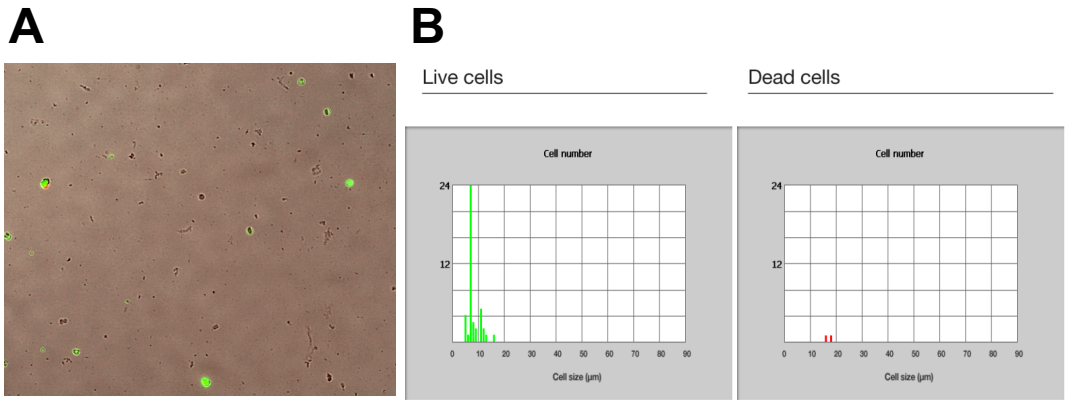
Supplemental Figure 5. Average age-dependent fecundity data for Trial 1. Average number of viable eggs per female (manual counts) and average barcode abundance are shown both as a bar plot and scatter plot (n=3 biological replicates for 3 barcodes per condition, error bars are +/- S.E.M.). Data from Trial 2 is shown in Figure 3C.



Supplemental Figure 6. Expression driven by *dpp-Gal4* for 20 TaG-EM lines. GFP antibody staining in the wing imaginal disc for the indicated TaG-EM barcode line driven by *dpp-Gal4*. Wing discs are counterstained with DAPI.



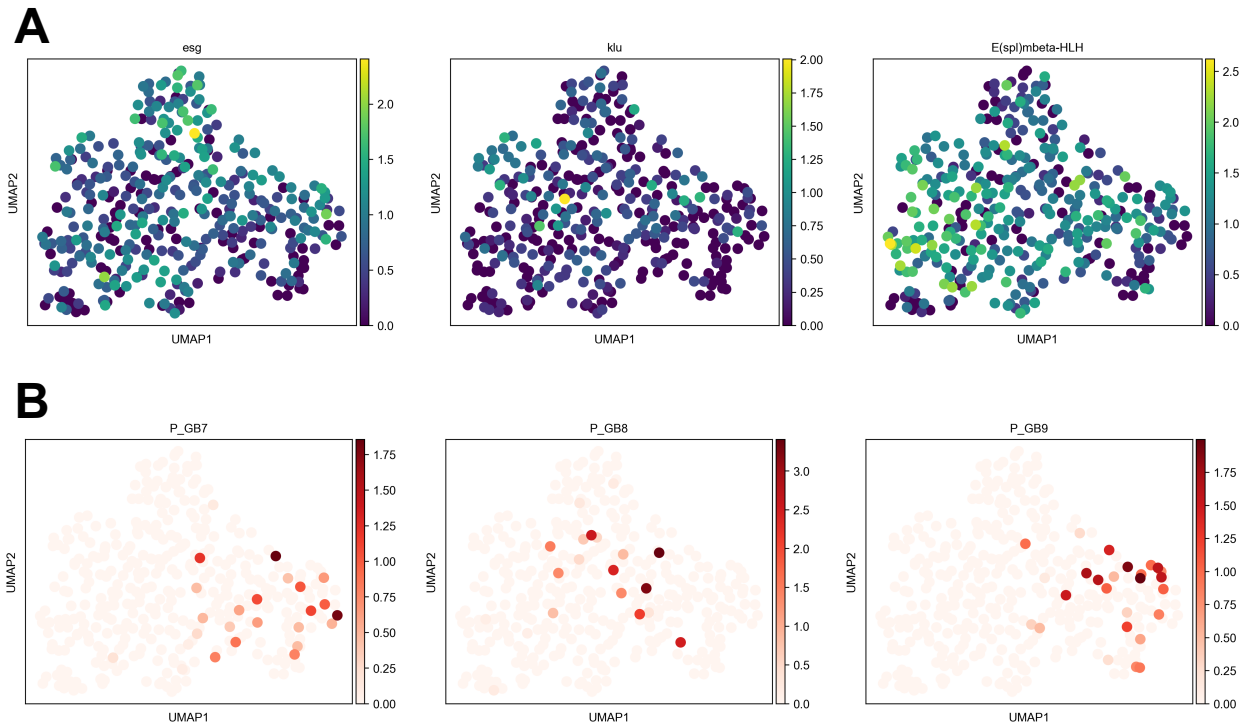
Supplemental Figure 7. TaG-EM line GFP expression driven by different Gal4 drivers. A) Comparison of GFP expression in larvae for the original pJFRC12 construct inserted in the attP2 landing site (left) or for a TaG-EM barcode line (right) expressed under the control of the indicated driver line. B) GFP expression of the PC-Gal (Precursor-Gal4) driver line together with either UAS-2xGFP or a combination of UAS-2xGFP and a TaG-EM barcode line.



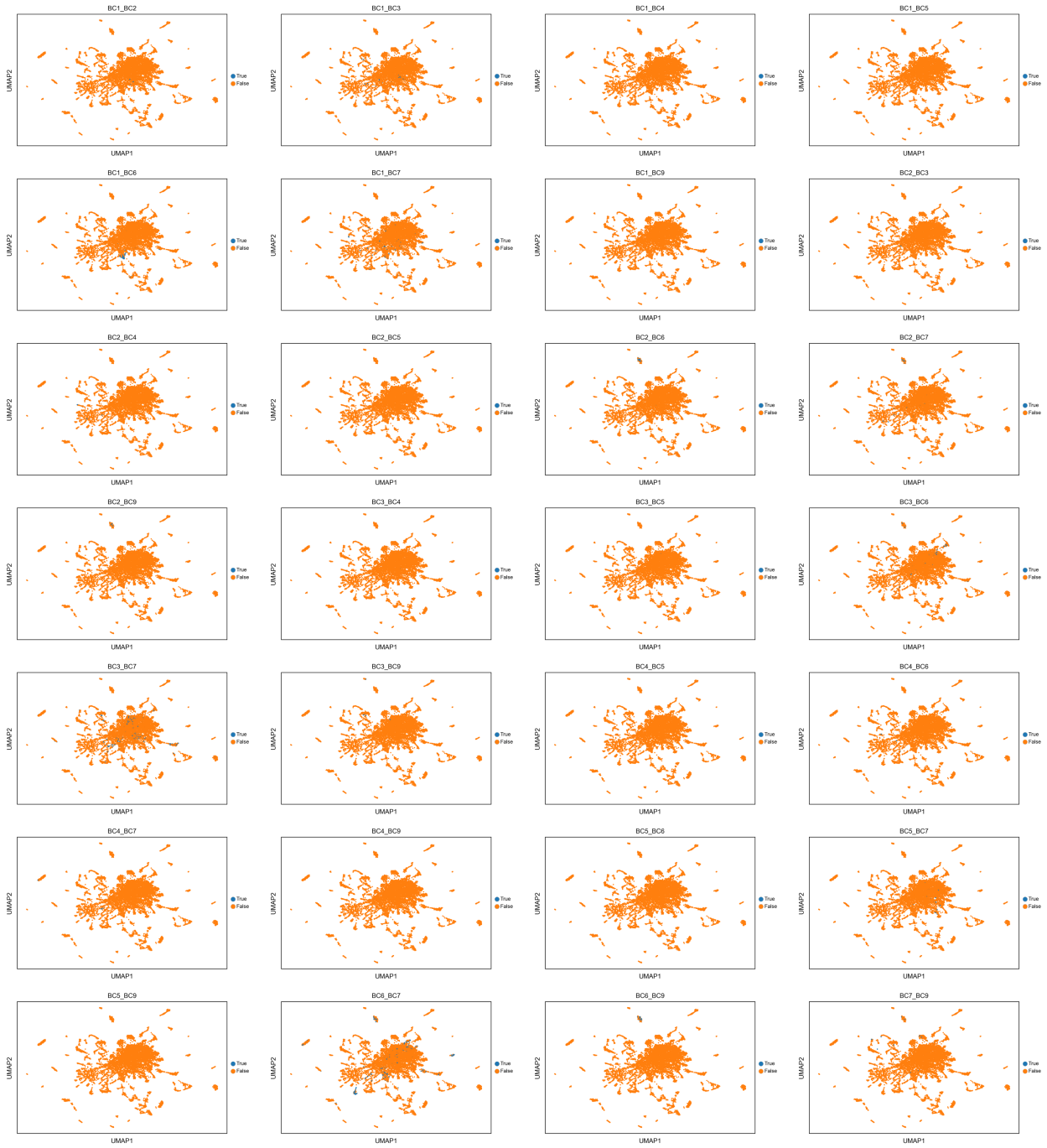
921

922 **Supplemental Figure 8. Dissociated intestinal cell viability.** A) GFP expression visualized in
923 dissociated cells from gut driver lines crossed to hexameric GFP and TaG-EM line. B)
924 Proportion of live (left) and dead (right) cells post-isolation and flow sorting as assessed by GFP
925 expression and propidium iodide staining.

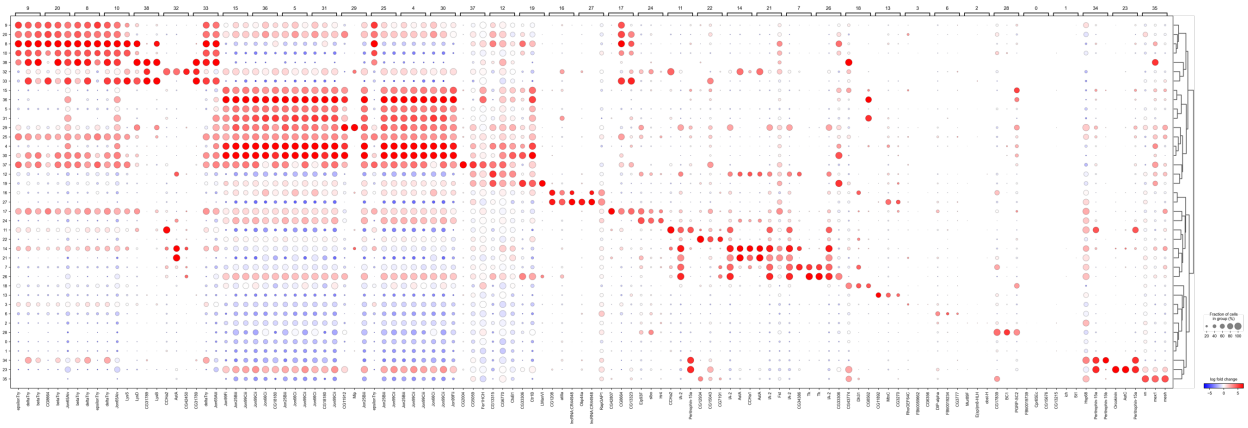
926



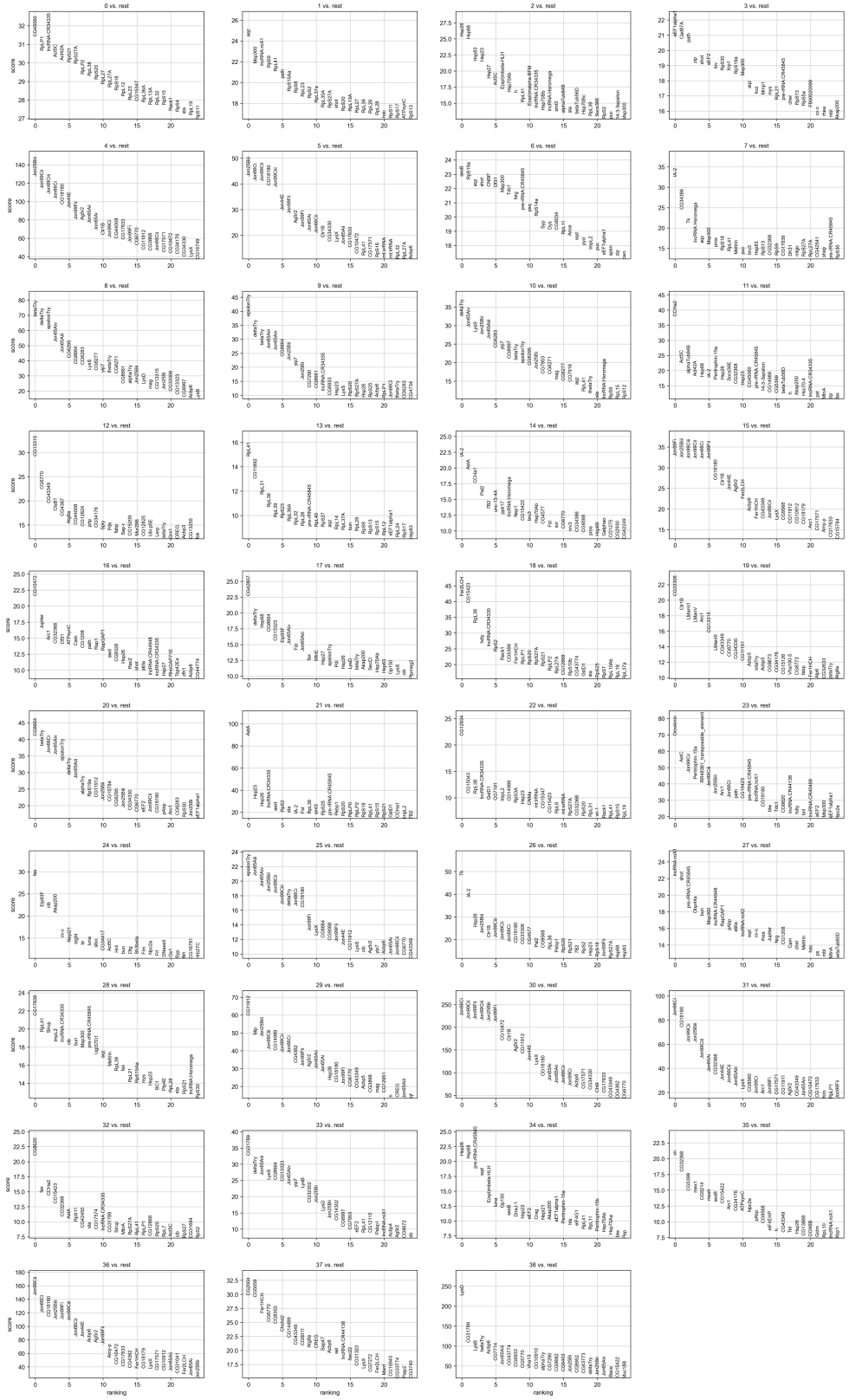
Supplemental Figure 9. Expression of TaG-EM genetic barcodes in larval intestinal precursor cells. UMAP plots showing gene expression levels of A) enteroblast/ISC marker genes *esg*, *klu*, and *E(spl)mbeta-HLH*; and B) the TaG-EM barcodes 7, 8, and 9 driven by the PC-Gal4 line.



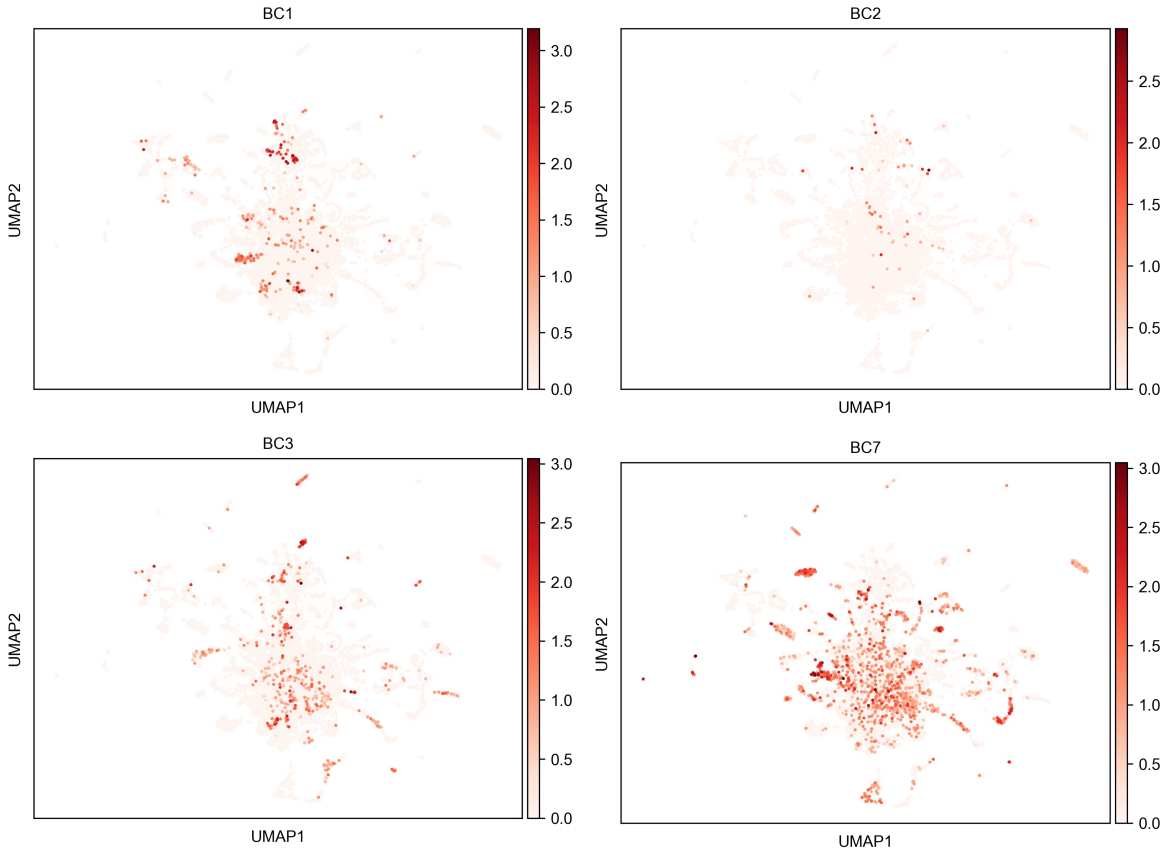
Supplemental Figure 10. Identification of doublets based on co-expression of TaG-EM barcodes. UMAP plots of all 28 pairwise barcode combinations showing cells co-expressing both of the indicated barcodes in blue (True).



Supplemental Figure 11. Differentially expressed genes Dotplot showing the top three differentially expressed genes by log-fold change for each cluster.



Supplemental Figure 12. Top 25 marker genes for each cell cluster. Plots showing t-test-based rankings of the top 25 differentially expressed genes for each cluster.



Supplemental Figure 13. Expression of the PMG-Gal4 driven TaG-EM barcodes. UMAP plots showing expression of the four PMG-Gal4 driven TaG-EM barcodes (BC1, BC1, BC3, and BC7).