PNAS RESEARCH ARTICLE | NEUROSCIENCE

Using single-cell RNA sequencing to generate predictive cell-type-specific split-GAL4 reagents throughout development

Yu-Chieh David Chen[a,1], Yen-Chung Chen[a], Raghuvanshi Rajesh[a,b], Nathalie Shoji[a], Maisha Jacy[a], Haluk Lacin[c], Ted Erclik[d,e], and Claude Desplan[a,b,1]

Contributed by Claude Desplan; received May 3, 2023; accepted July 3, 2023; reviewed by Hugo Bellen, Tom R. Clandinin, and Lawrence L. Zipursky

Cell-type-specific tools facilitate the identification and functional characterization of the distinct cell types that form the complexity of neuronal circuits. A large collection of existing genetic tools in *Drosophila* relies on enhancer activity to label different subsets of cells and has been extremely useful in analyzing functional circuits in adults. However, these enhancer-based GAL4 lines often do not reflect the expression of nearby gene(s) as they only represent a small portion of the full gene regulatory elements. While genetic intersectional techniques such as the split-GAL4 system further improve cell-type-specificity, it requires significant time and resources to screen through combinations of enhancer expression patterns. Here, we use existing developmental single-cell RNA sequencing (scRNAseq) datasets to select gene pairs for split-GAL4 and provide a highly efficient and predictive pipeline (scMarco) to generate cell-type-specific split-GAL4 lines at any time during development, based on the native gene regulatory elements. These gene-specific split-GAL4 lines can be generated from a large collection of coding intronic MiMIC/CRIMIC lines or by CRISPR knock-in. We use the developing *Drosophila* visual system as a model to demonstrate the high predictive power of scRNAseq-guided gene-specific split-GAL4 lines in targeting known cell types, annotating clusters in scRNAseq datasets as well as in identifying novel cell types. Lastly, the gene-specific split-GAL4 lines are broadly applicable to any other *Drosophila* tissue. Our work opens new avenues for generating cell-type-specific tools for the targeted manipulation of distinct cell types throughout development and represents a valuable resource for the *Drosophila* community.

Single-cell RNA sequencing | Split-GAL4 | *Drosophila* visual system | MiMIC/CRIMIC

Significance

Understanding the function of individual cell types in the nervous systems has remained a major challenge for neuroscience researchers, partly due to the incomplete identification and characterization of neuronal cell types. To study the development of individual cell types and their function in health and disease, specific experimental access to each cell type is an essential prerequisite. Here, we establish a pipeline to generate gene-specific split-GAL4 lines guided by single-cell RNA sequencing datasets. These intersectional lines show high accuracy for labeling targeted cell types and can be applied to any tissue in *Drosophila*. The gene-specific split-GAL4 and scMarco, a Graphical User Interface designed to enable identification of marker gene pairs, will represent valuable resources to the fly research community.

Proper function of the nervous system relies on the interactions between a large number of different neurons to form circuits. Understanding the tremendous neuronal diversity in the central nervous system and the role of each neuronal type requires cell-type-specific genetic manipulations. In *Drosophila*, binary expression systems such as GAL4/UAS (1), LexA/LexAop (2), and QF/QUAS (3) systems have been widely used for creating cell-type-specific genetic reagents. For example, large collections of enhancer-based GAL4 driver lines (e.g., GMR-GAL4 and VT-GAL4) were created. Each of these lines contains 2 to 3 kb of genomic DNA sequences near genes expressed in the nervous system that represent distinct enhancer fragments that allow genetic access to different subsets of cells (4, 5). These lines are expressed in restricted patterns that often do not reflect the expression of the nearby gene, and a large screening effort was made to determine their expression. Searchable databases for expression patterns in the central brain as well as the ventral nerve cord are available for most of the enhancer-based GAL4 driver lines, facilitating the identification of genetic tools labeling cell types of interest. Although the enhancer-based GAL4 lines have restricted expression patterns, genetic intersectional techniques such as split-GAL4/LexA systems are often needed to further restrict the expression pattern to achieve cell-type specificity (6–8).

The split-GAL4/LexA system has significantly improved the labeling specificity by splitting the GAL4 or LexA transcription factors into GAL4 or LexA DNA binding domain (DBD) and a transcriptional activation domain (AD), each of which is expressed under the control of different enhancers (7, 9–11). Genetic intersection is achieved when cells express both enhancer lines, thereby reconstituting functional GAL4 or LexA expression. While there are thousands of split-GAL4 lines created for having the same expression pattern as the original GAL4 lines, the expression pattern of these split-GAL4 is often inconsistent with the original GAL4 lines (6, 8), which limits the use of the GAL4 expression database for searching specific elements to drive split-GAL4 components. Therefore, there is often a need to test several combinations of split-GAL4 lines to obtain the desired expression pattern. Although the color maximum intensity projection images (color MIPs) algorithm or NeuronBridge softwares have been recently developed to expedite the search for split-GAL4 combinations (12, 13), it remains challenging to efficiently create cell-type-specific

Author contributions: Y.-C.D.C., Y.-C.C., and C.D. designed research; Y.-C.D.C., Y.-C.C., R.R., N.S., and M.J. performed research; Y.-C.D.C., Y.-C.C., H.L., and T.E. contributed new reagents/analytic tools; Y.-C.D.C. and Y.-C.C. analyzed data; and Y.-C.D.C., Y.-C.C., and C.D. wrote the paper.

Reviewers: H.B., Baylor College of Medicine; T.R.C., Stanford University; and L.L.Z., University of California, Los Angeles.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: ycc4@nyu.edu or cd38@nyu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2307451120/-/DCSupplemental.

Published July 31, 2023.

split-GAL4 lines with high prediction power. Furthermore, the existing collection of enhancer-based GAL4 or split-GAL4 driver lines often have expression patterns that do not reflect the expression of the nearby gene and therefore cannot be predicted from the expression of these genes.

The time and resource consumption for screening and making cell-type-specific split-GAL4 lines is even more challenging for developmental studies. The many existing cell-type-specific genetic tools developed for functional studies in the adult central nervous system are often not expressed during development. This limits the ability to visualize and conduct genetic perturbations in developing neurons.

In this study, we used the developing *Drosophila* visual system as a model to establish a pipeline (scMarco) for generating highly cell-type-specific split-GAL4 lines to label specific neurons throughout development. Recent studies have produced single-cell transcriptomic atlases for all neurons at different stages of development of the *Drosophila* visual system and identified nearly 200 distinct cell types (14, 15). We used these datasets to develop a systematic pipeline for analyzing gene expression in each cell type across various developmental stages. Instead of relying on screening for expression patterns, we developed a marker finding algorithm (scMarco) to identify pairs of genes that are expressed together in only one or very few clusters. We adapted the "Trojan-GAL4" approach (16) with T2A-split-GAL4 cassettes in-frame to the coding sequence of the targeted genes, which allows the translation of split-GAL4 proteins from the same transcript. We found that these split-GAL4 drivers precisely reproduce the expression of the genes identified in the developmental scRNAseq datasets at any given developmental stage. These gene-specific T2A-split-GAL4 lines can be generated through the Recombinase-mediated cassette exchange (RMCE) of T2A-split-GAL4 elements from the large existing collection of MiMIC/CRIMIC lines (16–19) or by N- or C-terminal knock-in of T2A-split-GAL4 elements through CRISPR genome editing. We find high accuracy in labeling predicted cell types using these split-GAL4 lines whose expression is consistent with the scRNAseq expression at the different developmental stages. This allowed us to generate a collection of highly cell-type-specific developmental split-GAL4 lines that represent powerful tools to annotate unidentified clusters in scRNAseq datasets and to identify novel cell types. Additionally, to expand its applicability, we generated donor flies for making gene-specific split-GAL4 lines by genetic crosses. Taken together, the scRNAseq-guided split-GAL4 strategy provides highly cell-type-specific developmental genetic tools to study key genes controlling various neuronal developmental processes. The gene-specific split-GAL4 toolkit developed in our study is adaptable by design to any other tissue in *Drosophila* where the genes used for the optic lobes might be expressed and will be a great resource for the fly community.

Results

Pipeline for Generating Gene-Specific T2A-Split-GAL4 Lines. Recent scRNAseq datasets of developing *Drosophila* optic lobe (14, 15) identified about 200 clusters that presumably represent at least 200 different cell types (we will treat these clusters as cell types in the study and assume each cluster is homogenous unless stated otherwise) (Fig. 1*A*), some of which are categorized into several broad classes such as distal medulla (Dm), lamina (L), lamina wide-field (Lawf), lobula columnar (LC), lobula plate intrinsic (LPi), medullar intrinsic (Mi), transmedullary (Tm), and transmedullary Y (TmY) neurons. Each of these broad cell types has distinct morphological features (Fig. 1*B*) (20). As there are very few individual genes that only label one cluster in the developmental scRNAseq datasets, we looked for pairs of genes whose expression overlaps in one or very few clusters and used the split-GAL4 genetic intersectional strategy in which GAL4DBD and GAL4AD (either GAL4AD, p65, or VP16) are under the control of the regulatory elements of the gene pair (Fig. 1*C*) (6, 8, 9). We chose the gene-specific T2A-split-GAL4 approach to better recapitulate the endogenous target gene expression (16, 17). There are two main ways to create gene-specific T2A-split-GAL4 lines: *i*) T2A-split-GAL4 cassette can be knocked into either the N- or C-terminal of the endogenous gene by CRISPR (Fig. 1 *D* and *E*); *ii*) T2A-split-GAL4 transgenes can be integrated into coding introns of targeted genes by RMCE (Fig. 1*F*) (16–19). For the latter, a T2A-split-GAL4 donor cassette flanked by attB sequences can be exchanged with a large collection of coding intronic MiMIC/CRIMIC insertions flanked by attP sequences in the presence of ΦC31 integrase. The T2A-split-GAL4 cassette is subsequently spliced into the endogenous RNA transcript via splice acceptor and donor sequences in the cassette. After translation of the integrated transgenic cassette, the split-GAL4 protein is released from the endogenous truncated protein through the self-cleaving T2A sequence (21). The existing coding intronic MiMIC/CRIMIC lines provide an off-the-shelf and economical method to generate gene-specific split-GAL4 lines. However, they also generate a mutant allele for the targeted gene. For genes without readily convertible MiMIC/CRIMIC lines, one can generate CRISPR knock-in of T2A-split-GAL4 at either N-terminal or C-terminal of the gene. One would choose N-terminal over C-terminal knock-in if the target gene has multiple isoforms with different C-termini. However, C-terminal knock-in would likely maintain endogenous expression of the target gene as compared to N-terminal knock-in that mutates it. In all of these methods, the expression of these gene-specific T2A-split-GAL4 lines should have the same expression pattern as the endogenous target gene since they both use the same native gene regulatory elements.

We next aimed to identify pairs of genes that label each specific cluster of interest. Due to the sparse nature of scRNAseq, many droplets would still have zero transcript detected even when a gene is in fact expressed. Ambient RNAs could also introduce low read counts into some droplets even when a gene is not expressed. It is thus challenging to determine whether a given gene is expressed in a given cluster by a fixed threshold for the expression level at single-cell level (lognorm value) (Fig. 1*G*). To address this issue, we first binarized the gene expression in the scRNAseq dataset via mixture modeling (15, 22) so that each gene was assigned an expression probability score [probability (ON)] ranging from 0 to 1 in each cluster, 0 indicating no expression while 1 indicates strong expression. We set the probability score of 0.5 as the expression threshold and classified every gene in every cluster as either ON or OFF (Fig. 1*H*). For example, after mixture modeling binarization, two transcription factors Traffic jam (Tj) and Knot (Kn), were predicted to be expressed in 13 and 8 adult clusters, respectively, but overlapped only in one cluster (TmY14) (Fig. 1 *G* and *H*). The binarization of gene expression in scRNAseq data allows faithful selection of gene pairs for split-GAL4 genetic intersection targeting any given cluster.

Validation of Cell-Type Specificity for Gene-Specific T2A-Split-GAL4 Lines in the Adult Optic Lobe. We next generated a collection of gene-specific T2A-split-GAL4 lines predicted to label various known cell types in the optic lobes. Each cell type tested comprises unique, unambiguous morphological features projecting to specific layer(s) in the neuropil(s). To better ascertain the identity of neurons, we performed sparse labeling using the MultiColor FlpOut (MCFO) technique (23). We validated

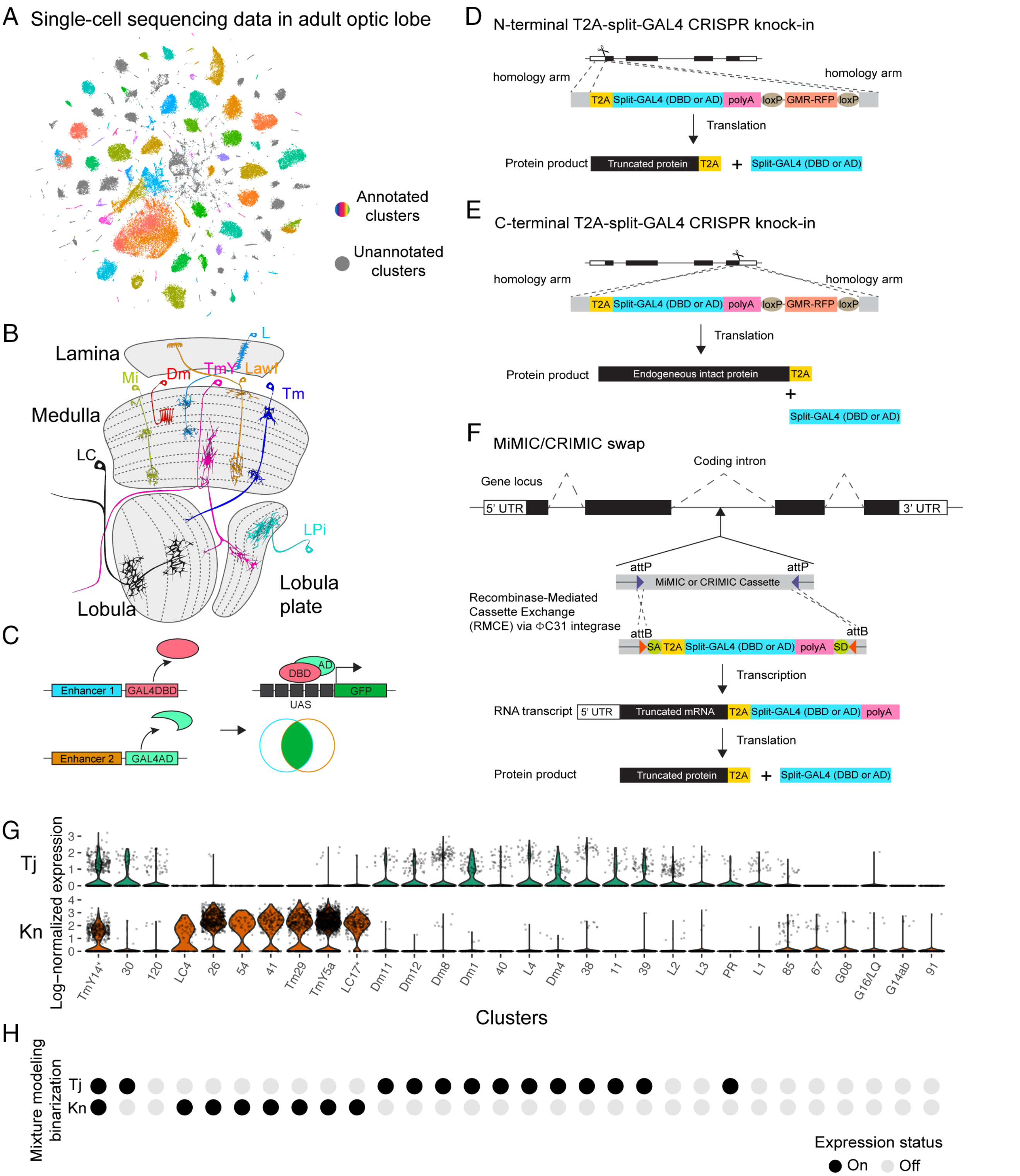


Fig. 1. (*A*) t-SNE plot of the scRNAseq dataset of the adult *Drosophila* optic lobe. Clusters in gray represent unannotated clusters. Clusters with other colors were annotated in a previous study (15). (*B*) Schematic diagram of the *Drosophila* visual system with representative cell types of several major classes. There are four main neuropils in the optic lobe: lamina, medulla, lobula, and lobula plate. Distal medulla (Dm), lamina (L), lamina wide-field (Lawf), lobula columnar (LC), lobula plate intrinsic (LPi), medullar intrinsic (Mi), transmedullary (Tm), transmedullary Y (TmY). (*C*) Schematic diagram of the split-GAL4 system. GAL4 can be split into GAL4 DNA-binding domain (GAL4DBD) and activation domain (AD). Labeling specificity is further restricted when GAL4DBD and AD are under the control of two different enhancers and only the cells with both enhancers active have a functional reconstituted GAL4 for driving reporter gene expression. (*D*–*F*) Schematic diagram of the gene-specific split-GAL4 generation. Gene-specific split-GAL4 lines can be generated either through N- (*D*) or C-terminal T2A-split-GAL4 knock-in (*E*), or through RMCE of T2A-split-GAL4 elements from the large existing collection of MiMIC/CRIMIC lines (*F*). The expression of a split-GAL4 reporter depends on the native gene regulatory network and is expected to recapitulate the endogenous transcript expression detected in the scRNAseq dataset. (*G*) Log-normalized expression of Tj (*Top*) and Kn (*Bottom*) for different clusters. Top 30 clusters expressing Tj or Kn are shown. (*H*) Binarization of gene expression by mixture modeling (15, 22). A probability (ON) score ranging from 0 to 1 is assigned to each cluster: 0 indicates no expression while 1 indicates strong expression. We set the probability score of 0.5 as an expression threshold and classified every gene in every cluster as either ON or OFF. Note that it is practically challenging to define whether a given gene is expressed in each cluster. For example, Tj showed similar expression level between cluster 39 and L2, yet, cluster 39 but not L2 is predicted to express Tj based on mixture modeling binarization.

our in silico prediction by crossing pairs of split-GAL4 lines with a UAS-GFP reporter for full expression, or with MCFO lines for sparse labeling. For all the split-GAL4 combinations used throughout this work, we used gene 1 (DBD hemi driver) ∩ gene 2 (AD hemi driver) for simplicity. We tested six split-GAL4 combinations that were predicted to be coexpressed in only one cluster and thus should specifically label only one cell type: TmY14 was predicted to be labeled by Tj ∩ Kn (Fig. 2*A*), Dm1 by CG5160 ∩ DIPα (Fig. 2*B*), Dm12 by Tj ∩ Dop1R2 (Fig. 2*C*), Dm4 by Ab ∩ Tj (Fig. 2*D*), Mi15 by CngA ∩ Ple (Fig. 2*E*), and LPLC2 by Acj6 ∩ DIPη (Fig. 2*F*). We did notice that additional cell types other than the predicted cell type were sometimes observed. For example, sparse labeling of the Dm4 split-GAL4 driver (Ab ∩ Tj) showed expression in both L2 and L4 neurons in adults (*SI Appendix*, Fig. S1*A*). In addition, as there are more than one split-GAL4 combinations labeling the same cluster, we tested whether we could target the same cell type by using different combinations of genes. Dm1 was predicted to also be labeled by CG5160 ∩ Tj (in addition to CG5160 ∩ DIPα). This split-GAL4 line indeed labeled Dm1, although sparse labeling showed the presence of additional neurons, such as Dm11 and L2 (*SI Appendix*, Fig. S1*B*).

To test further examples, we used other combinations that were predicted to each label two known neurons. Dm11 and LC12 were labeled by Reck ∩ CG14322 (Fig. 2*G*) as well as Lawf1 and Lawf2 by Fer2 ∩ Eya (Fig. 2*H*). Our results showed that all target cell types were labeled by the predicted split-GAL4 combinations, although additional cell types were observed in some cases.

Gene-Specific T2A-Split-GAL4 Lines Facilitate Cluster Annotations in the scRNAseq Datasets. Our ability to precisely label target cell types prompted us to examine if this approach could reveal

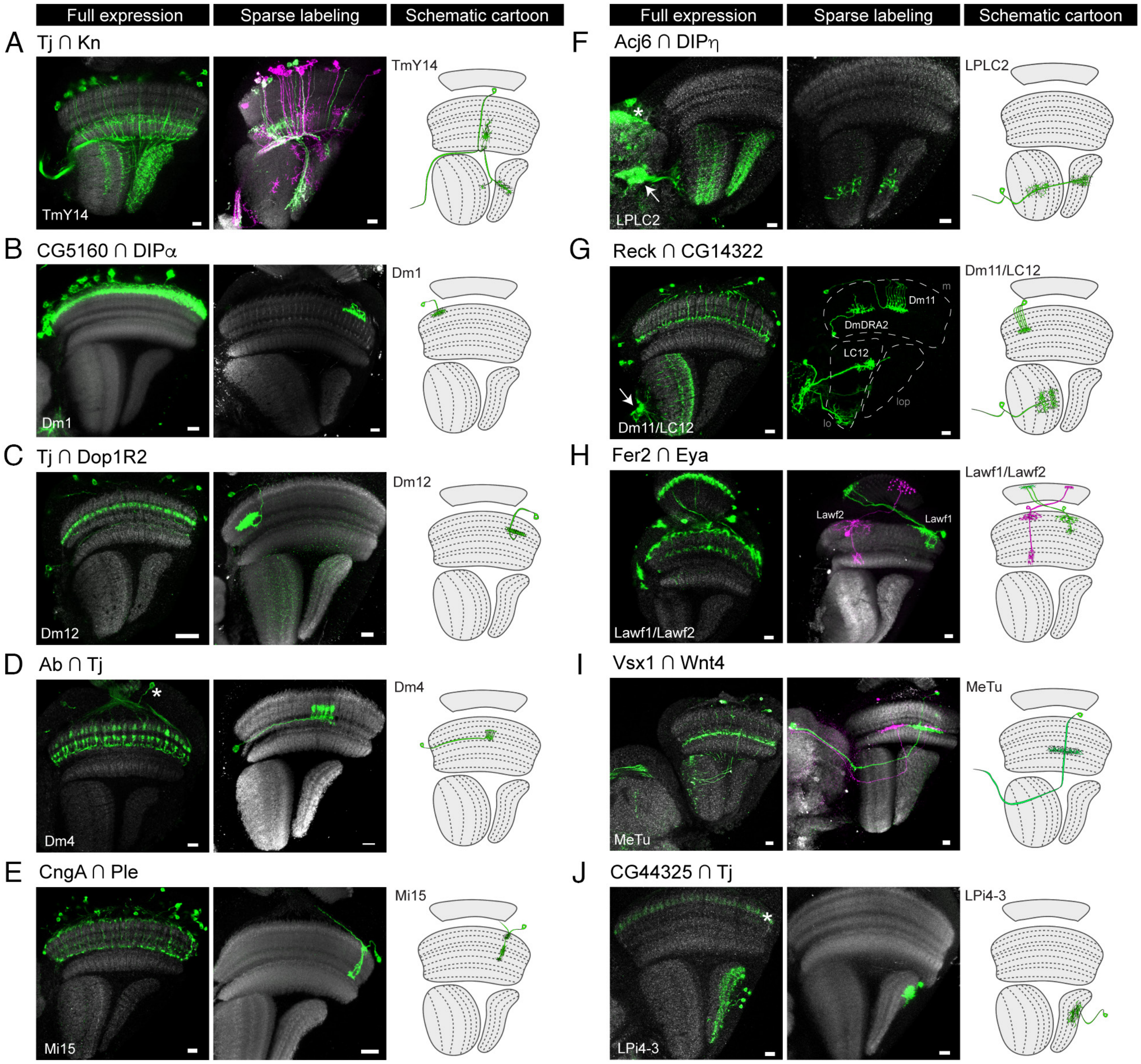


Fig. 2. Characterization of selected gene-specific split-GAL4 lines targeting different cell types/clusters. Targeted cell types predicted by scRNAseq expression are shown in the lower left corner for each split-GAL4 line. The expression pattern of each split-GAL4 line is shown either with UAS-myr-GFP reporter for full expression (*Left*) or with UAS-MCFO lines for sparse labeling (*Middle*). A schematic diagram of each cell type is shown on the right. (*A*–*F*) Examples of split-GAL4 lines targeting single cell types. (*G* and *H*) Example of split-GAL4 lines targeting two cell types. (*I* and *J*) Examples of split-GAL4 lines targeting unannotated cell types in the scRNAseq dataset. Anti-NCad staining (gray) is used for visualizing neuropils. Images are substack projections of full expression labeling or segmented single cells from sparse labeling to show distinct morphological features of distinct cell types. Asterisks indicate expression in the cell types not predicted by mixture modeling. (Scale bar: 10 μm.)

the identity of neurons in unannotated clusters in our scRNAseq dataset. These unannotated clusters could either belong to known cell types or to new cell types. We first targeted cluster 37 that expresses both Vsx1 and Wnt4. Only one cell type innervating the proximal medulla while projecting to the anterior optic tubercle (AOTU) was labeled by Vsx1 ∩ Wnt4 (Fig. 2*I*). The neuronal morphology matched medullo-tubercular neurons (MeTu) (24). Interestingly, the somata of MeTu neurons labeled by this split-GAL4 line were restricted to the dorsal half of the medulla cortex (*SI Appendix*, Fig. S1*C*), resembling both the MeTu_{il} and MeTu_{im} subtypes described in a previous study (24).

Unannotated cluster 40 was predicted to specifically express CG44325 ∩ Tj (Fig. 2*J*). This split-GAL4 combination labeled lobula plate intrinsic neurons (LPi) innervating lobula plate layers 3 and 4. Given that the cell bodies were located in the lobula plate cortex, this cell type matches LPi4-3 (25). This line also labeled L2 neurons in the lamina. We examined another combination of split-GAL4 for cluster 40 (Tj ∩ Mip), which again labeled LPi4-3 (*SI Appendix*, Fig. S1*D*) but did not label L2 neurons.

Gene-Specific T2A-Split-GAL4 Line Uncovers Unique Cell Types. The high accuracy in predicted clusters using the gene-specific T2A-split-GAL4 lines facilitates not only assigning known cell types to unannotated clusters in our scRNAseq data (Fig. 2 *I*–*J*), but also identifying unique cell types that have not been described. In our scRNAseq dataset, TkR86C is expressed in 12 clusters and CG14322 in 8 clusters (Fig. 3*A*) that overlap only in the unannotated cluster 30 (Fig. 3*A*). TkR86C ∩ CG14322 labeled a unique type of medulla projection neuron whose cell body is located in the medulla cortex and its neurites innervate medulla layer M7 (the serpentine layer) (Fig. 3 *B* and *C*). Sparse labeling revealed that cluster 30 neurons bifurcate at the M7 layer with one neurite branch innervating multiple visual columns (Fig. 3 *D* and *E*) and another branch running along the border of medulla and lobula and projecting to the superior posterior slope (Sps) in the central brain (Fig. 3 *F* and *G*). The bifurcation of neurites in the medulla is bidirectional (Fig. 3*E*) and all medulla columns are innervated by a total of 40 to 50 cells (Fig. 3 *B* and *C*). We named this unique medulla projection neuron Medulla-Superior posterior slope, MeSps.

Since MeSps neurons have never been described before, we performed immunofluorescence staining for additional transcription factor markers that together define cluster 30 in the scRNAseq data (Fig. 3*H*). This confirmed that MeSps neurons are positive for Toy, Tj, Pros, and Fd59a (Fig. 3 *I*–*J*), but are negative for Kn (*SI Appendix*, Fig. S2). In conclusion, we identified a unique MeSps medulla projection neuron (Fig. 3*K*), and assigned cluster 30 to this neuron.

In addition to MeSps, we used another split-GAL4 line (CG11317 ∩ Tey) to target unknown cluster 44 (*SI Appendix*, Fig. S3*A*) We found neurons that are similar to Y3 neurons but lack the fork-like structures in the superficial lobula layer (20), which we named Y3-like neurons (*SI Appendix*, Fig. S3 *B*–*D*). Altogether, our results demonstrate a highly efficient strategy for generating gene-specific split-GAL4 lines guided by scRNAseq datasets and annotating unknown clusters.

Gene-Specific T2A-Split-GAL4 Lines Label Neurons from Early Developmental Stages to Adulthood. Although the existing split-GAL4 lines generated by the enhancer-based approach specifically label various cell types in the optic lobe (22, 26, 27), most of them do not label these cell types during development or are expressed in different neuronal types during development. Taking advantage of the availability of single-cell transcriptomes of all optic lobe neurons at five pupal stages (P15, P30, P40, P50, and P70) (15), we selected split-GAL4 lines that were coexpressed at most, if not all, pupal stages. We tested seven such different split-GAL4 combinations targeting the TmY14 (Tj ∩ Kn, Fig. 4*A*), Dm1 (CG5160 ∩ Tj, Fig. 4*B*), Mi15 (CngA ∩ Ple, Fig. 4*C*), Dm11/LC12 (Reck ∩ CG14322, Fig. 4*D*), Lawf1/Lawf2 (Fer2 ∩ Eya, Fig. 4*E*), LPi4-3 (CG44325 ∩ Tj, Fig. 4*F*), and MeSps (TkR86C ∩ CG14322, Fig. 4*G*) and examined their expression at two earlier developmental stages (P15 and P50). Most split-GAL4 drivers consistently labeled the predicted cell types at these time points (Fig. 4 *A*–*G*). We found that CngA ∩ Ple did not show any expression in the optic lobe at P15, consistent with the mixture modeling prediction that this gene pair was not coexpressed in Mi15 at this stage (Fig. 4*C*). We also noted that additional cell types were observed in some of the split-GAL4 lines. For example, lamina neurons were labeled along with Dm1 (Fig. 4*B*) and LPi4-3 (Fig. 4*F*) split-GAL4 drivers. A lobula columnar neuron was labeled by the MeSps split-GAL4 driver (Fig. 4*G*). Another split-GAL4 combination (Beat-IIIc ∩ DIPα) predicted to label Lpi4-3 from P15 to P70, showed Lpi4-3 labeling exclusively at P15 but also showed two additional cell types in adults (one in lobula and another in medulla) (*SI Appendix*, Fig. S4). These results support the consistency of predicted cell types labeled across developmental stages by the scRNAseq-guided gene-specific split-GAL4 approach.

Extensive Predicted Optic Lobe Cluster Coverage by Marker Gene Pairs. There are around 200 clusters in recent single-cell developmental transcriptomic datasets (14, 15). After merging developmental subclusters and excluding immature clusters, ganglion mother cells (GMCs), or transient extrinsic (TE) neurons that are only present during development and not in the adult, there are 198 clusters in the adult. We aimed to determine how many clusters could be predicted to be labeled by combinations of gene pairs throughout development (Fig. 5). We first defined consistently expressed gene pairs as those that label only a single cluster for all six developmental stages, while none of the other clusters show any expression by this gene pair at any developmental stages (Fig. 5*A*). We found 91 clusters that can be labeled by such consistently expressed gene pairs for which cell-type-specific split-GAL4 driver lines targeting these clusters could theoretically be generated (Fig. 5*D*). When including gene pairs that are transiently expressed in 1 or 2 additional clusters, 33 additional clusters could be labeled (Fig. 5*D*). This suggests that 62.6% of the clusters (124/198) can be labeled by split-GAL4 gene pairs that can be used as developmental drivers predicted to be expressed from P15 to adult.

A broader search of potential gene pairs can be performed by relaxing the definition of consistently expressed gene pairs as a given gene pair expressed in the adult as well as at three or more of the five developmental stages. While the lack of expression at given times could reflect biological variations or expression below the threshold in the mixture modeling binarization, it is likely that the split-GAL4 would perdure during stages when the genes are not detected. We always required the consistently expressed gene pairs to label the cluster at the adult stage to ensure unambiguous identification of neuronal morphology. When we only looked for expression at adult plus any four out of five (Fig. 5*B*) or any three out of five developmental stages (Fig. 5*C*), then most clusters (185/198, 93.4%) were covered by a pair of genes (Fig. 5*D*). Each cluster can often be labeled by more than one combination of gene pairs, and all possible combinations of gene pairs that are expressed at all six developmental stages are listed in *SI Appendix*, Table. S3.

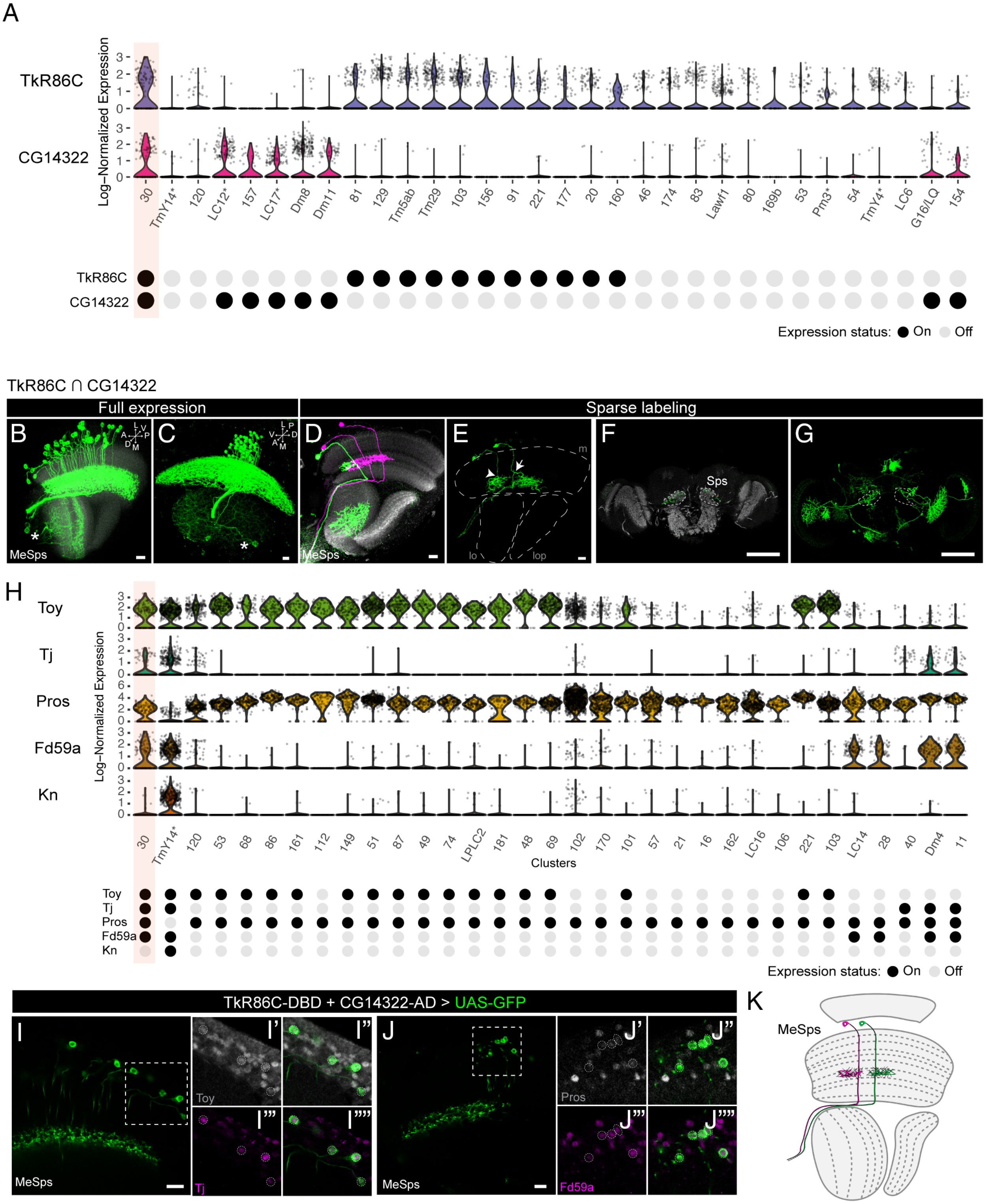


Fig. 3. (*A*) Log-normalized expression of TkR86C and CG14322 for different clusters (*Top*). Mixture modeling binarization of expression status for both genes is shown at the *Bottom*. Note that cluster 30 is predicted to be the only cluster intersected by TkR86C and CG14322. Top 30 clusters expressing TkR86C or CG14322 are shown. (*B* and *C*) The full expression pattern of TkR86C ∩ CG14322 line is shown with UAS-myr-GFP reporter. Note that there is an additional cell type in the lobula with only 2 to 3 cells labeled (marked by asterisks). The cell bodies of MeSps neurons are restricted to the medulla cortex. (*D* and *E*) Sparse labeling of MeSps neurons using MCFO. The bifurcation of neurites at M7 layer can be bidirectional (arrowhead to the anterior projection; arrow to the posterior projection). m: medulla; lo: lobula; lop: lobula plate. (Scale bar: 10 μm.) (*F* and *G*) Sparse labeling of MeSps neurons using MCFO show their projection to the superior posterior slope (Sps). Single optical section (*F*) of substack maximum projections (*G*) are shown. (Scale bar: 100 μm.) Anti-NCad staining (gray) is used for visualizing neuropils. (*H*) Log-normalized expression of selected transcription factors (Toy, Tj, Pros, Fd59a, and Kn) (*Top*). Mixture modeling binarization of expression status for selected genes are shown at the bottom. Note that cluster 30 is predicted to be the only cluster positive for Toy, Tj, Pros, and Fd59a. Kn is used to serve as a negative marker for cluster 30. Top 30 clusters expressing Toy, Tj, Pros, Fd59a, and Kn are shown. (*I–J*) Costaining of MeSps neurons labeled by TkR86C ∩ CG14322 with anti-Toy (Gray) and anti-Tj (Magenta) in *I* (region highlighted in the dotted white square was shown in *I′–I″″*) and anti-Pros (Gray) and anti-Fd59a (Magenta) in *J* (region highlighted in the dotted white square was shown in *J′–J″″*). MeSps neurons expressing GFP reporter are outlined in dotted circles. (Scale bar: 10 μm.) (*K*) Schematic diagram of MeSps neurons, a unique type of medulla projection neurons.

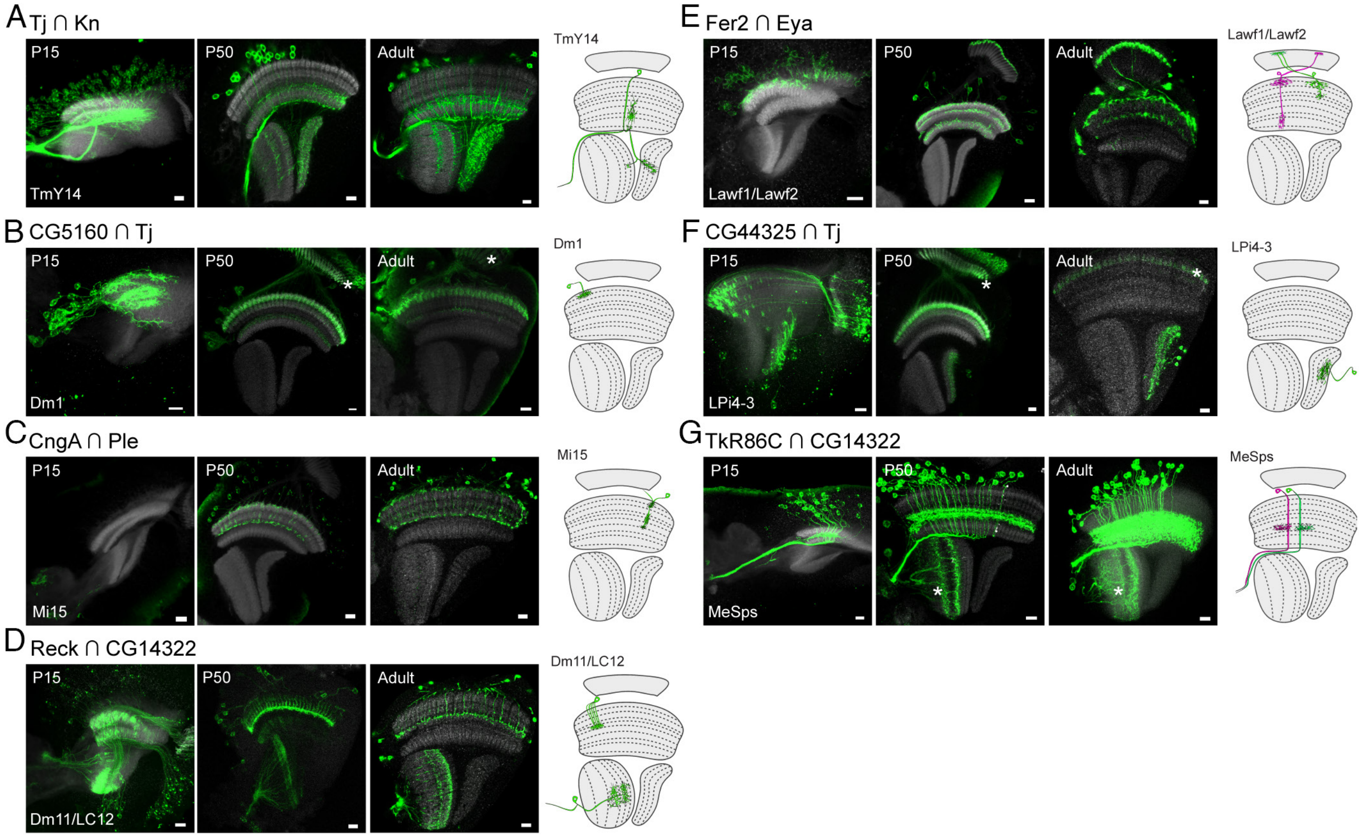


Fig. 4. (*A–F*) Developmental characterization of selected gene-specific split-GAL4 lines targeting different cell types/clusters. Targeted cell types predicted by scRNAseq expression are shown in the lower left corner for each split-GAL4 line. The full expression pattern of each split-GAL4 line is shown with UAS-myr-GFP reporter at P15, P50, and adult stages. A schematic diagram of targeted cell types in adults is shown on the right in each panel. Asterisks indicate expression in the cell types not predicted by mixture modeling. Although other cell types are observed, the targeted cell types are always observed at multiple developmental stages. Anti-NCad staining (gray) is used for visualizing neuropils. Images are substack projections of full expression labeling to show distinct morphological features of distinct cell types. (Scale bar: 10 µm.)

Given that there are many genes with coding intronic MiMIC or CRIMIC lines that are readily split-GAL4 convertible, we also examined the number of clusters with consistently expressed gene pairs currently available for cost-efficient split-GAL4 conversion (Fig. 5*E*). We found that 94 out of 198 clusters have consistently expressed gene pairs with available coding intronic MiMIC or CRIMIC for adult and five, 136 for adult and four, and 176 for adult and three developmental stages (Fig. 5*E*).

There are still 112 out of 198 clusters in the optic lobe that remain unannotated in the optic lobe developmental atlases (14, 15). We therefore performed similar analyses and found over 90% cluster coverage (107/112 for all genes and 102/112 for available MiMIC/CRIMIC lines) among the 112 unannotated clusters (*SI Appendix*, Fig. S5). These results suggest that combinations of genes can be found to generate split-GAL4 lines for most if not all the clusters in the optic lobe and most clusters have gene pairs with readily convertible split-GAL4 reagents.

In Vivo Swapping for Generating Gene-Specific T2A-Split-GAL4 Lines from the Collection of Coding Intronic MiMIC/CRIMIC. There is a large collection of coding intronic MiMIC/CRIMIC lines that can be readily swapped into either GAL4DBD or AD split-GAL4 drivers via embryo injection of an appropriate donor DNA template (16). We adapted the in vivo swapping method for making T2A-GAL4 lines (16) by replacing the GAL4 in the donor cassettes with either GAL4DBD or AD (Fig. 6*A*). Briefly, the T2A-split-GAL4 cassettes with reading frames 0, 1, and 2 were flanked by three lox sequence variants where a circular donor can be excised in the presence of Cre recombinase. By providing both Cre recombinase and ΦC31 integrase, the T2A-split-GAL4 cassette with the right orientation (forward or reverse) and correct splicing phase (phase 0, 1, or 2) can be selected by genotyping PCR. We generated triple donor flies that carry donor cassettes for all three reading frames and tested the in vivo swapping by using the bru1[MI00135] coding intronic MiMIC line (*SI Appendix*, Fig. S6*A*). bru1 is encoded in the "+" strand in phase 1 (Fig. 6*B*); therefore, the successful swap that generates bru1-split-GAL4 should have a forward integration with phase 1 donor. We performed GAL4DBD and AD triple donor crosses with bru1[MI00135] coding intronic MiMIC line and selected 273 y^- males for GAL4DBD and 248 y^- males for AD lines for further molecular mapping. We observed 38.8% (forward orientation) and 61.2% (reverse orientation) for GAL4DBD swapping and 41.6% (forward orientation) and 58.4% (reverse orientation) for AD swapping (Fig. 6*C*). We performed PCR genotyping for the individuals with forward integration and identified no individuals with phase 1 for GAL4DBD donors and 4 individuals with phase 1for AD donors (Fig. 6*C*). While the triple donor simplifies the crossing scheme when performing large-scale in vivo swapping, the simplicity came at the cost of reducing the success of obtaining a useful split-GAL4 line to one-third. To increase the success rate for in vivo swapping, we therefore generated single donor split-GAL4 lines for each phase so that users can select the donor lines with the correct phase (*SI Appendix*, Fig. S6*B*). In sum, we generated both triple-donor and single-donor T2A-split-GAL4 flies for in vivo swapping and provided a proof-of-principle example with

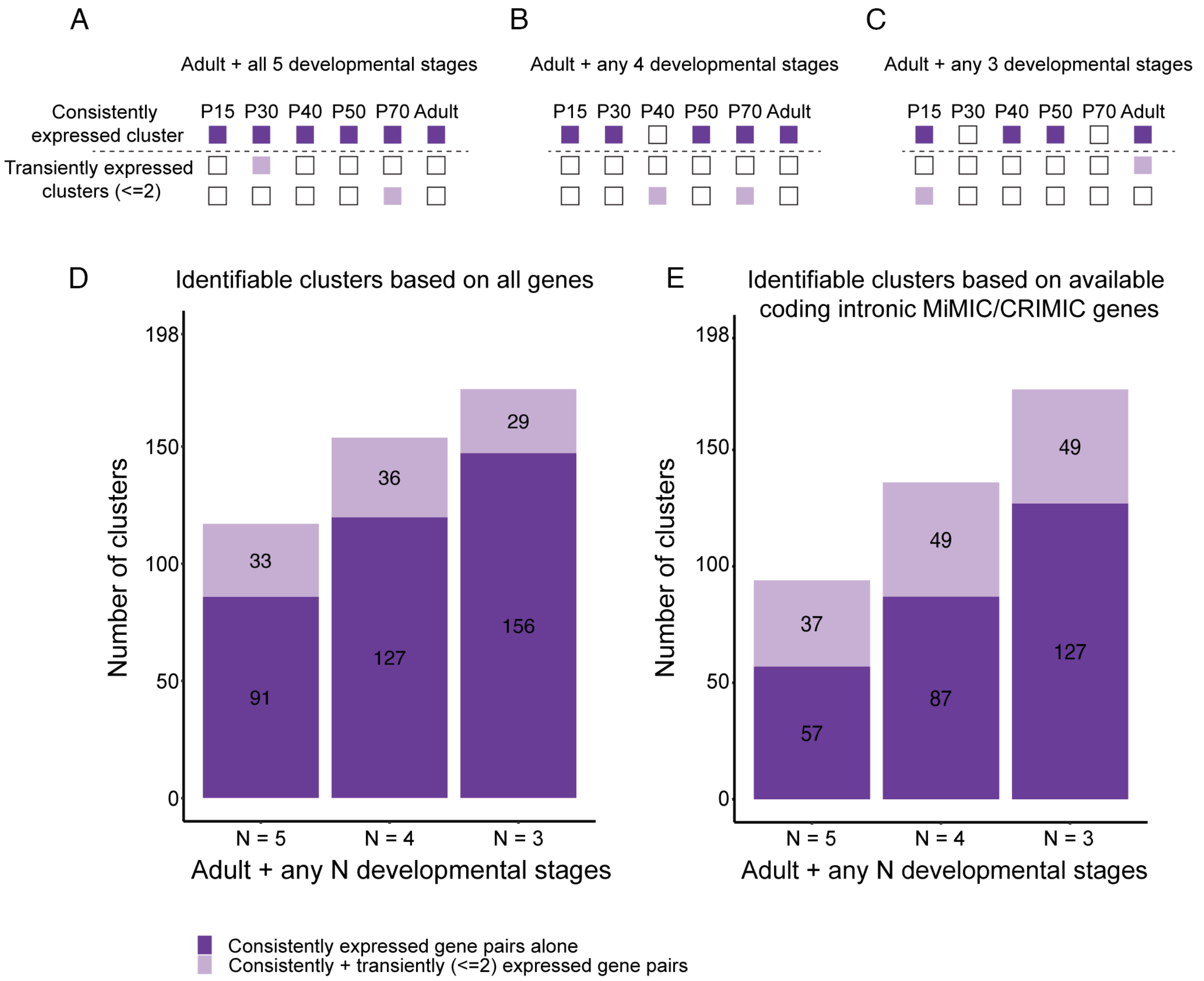


Fig. 5. (*A*–*C*) Illustration of strategies to identify gene pairs that mark a cluster of interest throughout development with different stringency. Dark violet boxes indicate the stages when a gene pair is predicted to be on in the cluster of interest while light violet boxes indicate when the gene pair is predicted to be transiently active in other clusters. (*D*) Number of clusters that are predicted to be identified with gene pairs suggested by each strategy when all genes detected in the atlas are considered. (*E*) Number of clusters that are predicted to be identified with gene pairs suggested by each strategy when only considering genes with coding intronic MiMIC or CRIMIC lines available for RMCE (3,637 genes).

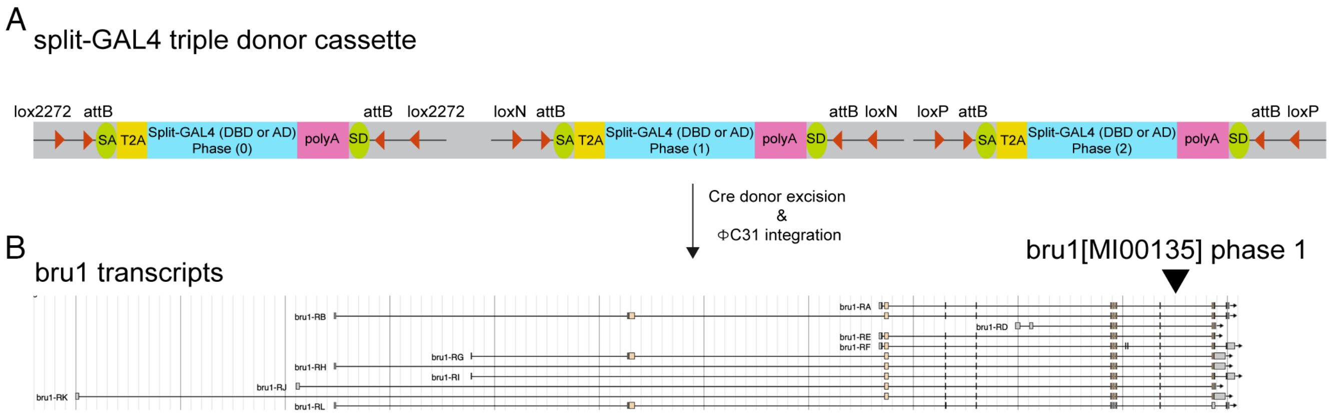
a successful swapping of bru1-AD. These tools will be valuable for the fly community to generate gene-specific T2A-split-GAL4 lines at a low cost.

scMarco, an R/Shiny-Based Application that Allows Marker Combination Selection from Any scRNAseq Dataset. Our results showed that using marker gene pairs for cell-type-specific labeling is a promising strategy. Most individual marker genes presented in this study were expressed in more than 10 clusters (e.g., Tj and CG14322), but the combination of two genes resulted in highly cell-type-specific expression. To provide a graphical user interface (GUI) for selecting marker combinations, we developed scMarco (https://apps.ycdavidchen.com/scMarco), an R/Shiny-based application with a Bayesian mixture model to binarize gene expression in any scRNAseq dataset. scMarco facilitates the identification of marker genes for intersection in other research models even when split-GAL4 generation is not available. We provide detailed step-by-step instructions of how to use the application (https://docs.ycdavidchen.com/scMarco). scMarco can be run from within RStudio or as a stand-alone web application.

Discussion

The availability of cell-type-specific genetic tools is often the prerequisite for many genetic manipulations used to study development and function of the nervous system. Using the *Drosophila* visual system as a model for which developmental transcriptomic atlases exist (14, 15), we showed that scRNAseq-assisted gene-specific split-GAL4 lines provide highly predictable and specific labeling for most cell types. The gene-specific split-GAL4 system provides several advantages over the enhancer-based split-GAL4:

1) Gene-specific split-GAL4 lines have high predictive accuracy. For all 15 gene-specific split-GAL4 lines tested in this study, we always observed the expression of these split-GAL4 lines in the predicted cell types. This facilitates the annotation of unknown clusters in scRNAseq datasets. We also took advantage of the high predictive accuracy to select gene pairs that are also expressed throughout most, if not all, developmental stages in the developing optic lobe. Our expression analyses showed that these split-GAL4 drivers indeed label the targeted cell types during development and are therefore very useful developmental


A
split-GAL4 triple donor cassette
lox2272 attB
SA T2A Split-GAL4 (DBD or AD) Phase (0) polyA SD
attB lox2272
loxN attB
SA T2A Split-GAL4 (DBD or AD) Phase (1) polyA SD
attB loxN
loxP attB
SA T2A Split-GAL4 (DBD or AD) Phase (2) polyA SD
attB loxP
Cre donor excision & ΦC31 integration
B
bru1 transcripts
bru1[MI00135] phase 1
bru1-RA
bru1-RB
bru1-RD
bru1-RE
bru1-RF
bru1-RG
bru1-RH
bru1-RI
bru1-RJ
bru1-RK
bru1-RL

C

	bru1-GAL4DBD	bru1-VP16
Total male screened	273	248
y- males (including w+ and w-)	19	30
Sterile y- males	1	6
PCR genotyping	18	24
Forward orientation	7	10
Reverse orientation	11	14
Phase 0	3	4
Phase 1	0	4
Phase 2	4	2

Fig. 6. (*A*) Schematic diagram showing T2A-split-GAL4 triple donor cassette. The cassette is modified from T2A-GAL4 triple donor (16) by replacing T2A-GAL4 with T2A-GAL4DBD or T2A-AD. Split-GAL4 donors with three different splicing phases (phase 0, 1, and 2) are flanked by attB sequences and lox sequence variants. Supplement of Cre recombinase and ΦC31 integrase will allow the integration of T2A-split-GAL4 into targeted MiMIC insertion. (*B*) Gene structure annotation of bru1 (encoded in a "+" orientation) from JBrowse using *D. melanogaster* (r6.49) ref. 28. Phase 1 coding intronic MiMIC insertion (MI00135) is highlighted by an arrowhead. (*C*) Summary table of in vivo swapping by triple T2A-split-GAL4 donor. The genetic crossing scheme is shown in *SI Appendix*, Fig. S6*A*.

drivers. Although additional cell types are sometimes observed, it might be due to false negatives in our mixture modeling-based binary determination of gene expression in a given cluster. Alternatively, these cells might represent rare cell types that are hidden in heterogenous clusters of our scRNAseq dataset. It is important to note that all the gene-specific split-GAL4 lines used here also label neurons outside the optic lobes, which would hinder the interpretation of functional perturbation of neuronal activities for behavioral studies.

2) The generation of a single fly stock containing both split-GAL4 hemi drivers that is ready-to-use is made easy since each hemi driver is at precise known locations throughout the genome. In contrast, the large collection of enhancer-based split-GAL4 lines are inserted at either of the two locations: the attP2 (the third chromosome) and/or attP40 (the second chromosome) sites. This reduces the possible combinations when both GAL4DBD and AD are inserted at the same location. For example, all the VT-GAL4DBD lines and one-third of VT-AD lines are inserted at attP2, making single ready-to-use split-GAL4 stock in the same animal impossible (8). In contrast, the gene-specific split-GAL4 lines can be recombined when present on the same chromosome and their location predicts the recombination frequency (e.g., TkR86C ∩ CG14322 for MeSps, Reck ∩ CG14322 for Dm11/LC12, and CG11317 ∩ Tey for Y3-like in this study). The options of selecting gene-specific split-GAL4 lines with the ability to recombine on a single chromosome provides flexibility to generate further crosses without using all chromosomes.

3) Similar to the enhancer-based split-GAL4 lines, there are multiple gene-specific split-GAL4 combinations that are predicted to target the same cell type, providing independent drivers for genetic manipulations. It should be noted that coding intronic MiMIC/CRIMIC-derived split-GAL4 are mutant alleles of the targeted gene. Performing experiments in a heterozygous background is necessary to alleviate this problem in most cases. Some genes, especially transcription factors, might have dose-dependent effects or synergistic interactions between two targeted genes when in a heterozygous background. Therefore, proper controls are necessary to identify potential phenotypes, including performing experiments with another combination of split-GAL4 targeting the same cell type.

4) The gene-specific split-GAL4 lines are adaptable to other tissues in *Drosophila*. With the development of scRNAseq technology, single-cell transcriptomes for various tissues are rapidly accumulating and are expanding the list of tissues that are applicable to generate split-GAL4 lines for genetic manipulation. Our pipeline in binarizing scRNAseq expression matrix using mixture modeling can be readily applied to other systems with existing scRNAseq data. These tools would further simplify the search for desired split-GAL4 combination, with many of the individual split-GAL4 lines already existing for other tissues.

5) Thousands of gene-specific split-GAL4 combinations can be readily generated. There are currently available coding intronic MiMIC/CRIMIC lines corresponding to 3,637 genes, and additional new CRIMIC lines are still being generated. The triple- and

single-donor split-GAL4 lines generated in this study are perfectly suited for a large-scale in vivo swapping of these coding intronic MiMIC/CRIMIC lines into split-GAL4 lines simply through genetic crosses. For genes without coding intronic MiMIC/CRIMIC lines, CRISPR genome editing methods can perform T2A-split-GAL4 knock-in into essentially any target.

In sum, we developed an scRNAseq-guided approach for generating highly predictable cell-type-specific T2A-split-GAL4 lines that can be adaptable to any tissue in *Drosophila*. We noted that a contemporaneous study from the Perrimon lab describes the use of split-intein GAL4 to create GAL80 repressible genetic intersectional lines (29). These complementary genetic reagents enable genetic access for labeling specific neuronal or other heterogeneous populations in the fly throughout development. The adaptability of these gene-specific split-GAL4 lines described here will be a valuable community resource for understanding the regulation of gene networks that confer cellular morphology, physiology, function, and identity. For researchers working outside of *Drosophila*, scMarco, the R/Shiny-based application for selection of marker combinations developed in this study, will provide an excellent method for finding cell-type-specific marker genes where in situ hybridization of multiple genes can be done to characterize a given cell type.

Materials and Methods

Molecular Biology and *Drosophila* Strains. Flies were reared on molasses-cornmeal-agar food at 25 °C. See *SI Appendix f*or the details of the molecular constructs and transgenic flies generated in this study. Full genotypes of flies used in this study are described in *SI Appendix*, Tables S1 and S2.

Immunohistochemistry. Optic lobes of flies were dissected in Schneider's *Drosophila* Medium and fixed with 4% paraformaldehyde in 1X Dulbecco's phosphate-buffered saline. The primary and secondary antibodies used in the paper are described in *SI Appendix*, Table S1. Samples were mounted in VECTASHIELD antifade mounting medium. Fluorescent images were acquired using a Leica SP8 confocal microscope with 400 Hz scan speed in 1,024 × 1,024 pixel formats. Image stacks were acquired at 0.5 to 1 μm optical sections. Unless otherwise noted, all images were presented as maximum projections of the *z* stack generated using Leica LAS AF software. See *SI Appendix* for full details.

Identification of Marker Gene Pairs with Mixture Modeling–Inferred Binarized Expression. To identify marker gene pairs (two genes) that are specific to a cluster, we implemented a greedy search algorithm to minimize the number of clusters that express a given gene pair. To determine whether a gene is expressed in a cluster, we assigned probability of whether a gene is expressed (P(ON)) to each cluster at each stage as previously described (15, 22). See *SI Appendix* for full details.

scMarco. scMarco (source code: https://github.com/chenyenchung/scMarco; an example site with optic lobe developmental atlas (15): https://apps.ycdavidchen.com/scMarco; documentations: https://docs.ycdavidchen.com/scMarco) is a Shiny application that provides interactive selection of marker combinations with modeled expression probability as an SQLite database. See *SI Appendix* for full details.

Data, Materials, and Software Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank Fengqiu Diao and Ben White for providing technical assistance on designing split-GAL4 triple donor constructs. We thank all the Desplan lab member for critical feedback on this work, especially Isabel Holguera, Bogdan Sieriebriennikov, Ryan Loker and Sromana Mukherjee. Stocks were obtained from the Bloomington *Drosophila* Stock Center (NIH P40OD018537). This work was funded by grants from the NIH (R01EY013010 and R01EY13012 to C.D.). Y.-C.D.C. is supported by NIH NRSA (5F32EY032750). Y.-C.C. was supported by the MacCracken Program at New York University, by a NYSTEM institutional training grant (Contract #C322560GG), and by Scholarship to Study Abroad from the Ministry of Education, Taiwan. R.R. was supported by the MacCracken Program at New York University. All cartoon schematics of different cell types were created with BioRender.com.

Author affiliations: [a]Department of Biology, New York University, New York, NY 10003; [b]Center for Genomics and Systems Biology, New York University, Abu Dhabi 51133, United Arab Emirates; [c]Division of Biological and Biomedical Systems, University of Missouri - Kansas City, Kansas City, MO 64110; [d]Department of Biology and Cell, University of Toronto - Mississauga, Mississauga, ON L5L 1C6, Canada; and [e]Department of Systems Biology, University of Toronto - Mississauga, Mississauga, ON L5L 1C6, Canada

1. A. H. Brand, N. Perrimon, Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Dev. Camb. Engl.* **118**, 401–415 (1993).
2. S.-L. Lai, T. Lee, Genetic mosaic with dual binary transcriptional systems in Drosophila. *Nat. Neurosci.* **9**, 703–709 (2006).
3. C. J. Potter, B. Tasic, E. V. Russler, L. Liang, L. Luo, The Q system: A repressible binary system for transgene expression, lineage tracing, and mosaic analysis. *Cell* **141**, 536–548 (2010).
4. A. Jenett *et al.*, A GAL4-driver line resource for Drosophila neurobiology. *Cell Rep.* **2**, 991–1001 (2012).
5. E. Z. Kvon *et al.*, Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).
6. H. Dionne, K. L. Hibbard, A. Cavallaro, J.-C. Kao, G. M. Rubin, Genetic reagents for making split-GAL4 lines in Drosophila. *Genetics* **209**, 31–35 (2018).
7. C.-Y. Ting *et al.*, Focusing transgene expression in Drosophila by coupling Gal4 with a novel split-LexA expression system. *Genetics* **188**, 229–233 (2011).
8. L. Tirian, B. J. Dickson, The VT GAL4, LexA, and split-GAL4 driver line collections for targeted expression in the Drosophila nervous system. BioRxiv [Preprint] (2017). https://doi.org/10.1101/198648 (Accessed 22 November 2022).
9. H. Luan, F. Diao, R. L. Scott, B. H. White, The Drosophila split Gal4 system for neural circuit mapping. *Front. Neural Circuits* **14**, 603397 (2020).
10. H. Luan, N. C. Peabody, C. Vinson, C. R. Vinson, B. H. White, Refined spatial manipulation of neuronal function by combinatorial restriction of transgene expression. *Neuron* **52**, 425–436 (2006).
11. B. D. Pfeiffer *et al.*, Refinement of tools for targeted gene expression in Drosophila. *Genetics* **186**, 735–755 (2010).
12. G. W. Meissner *et al.*, A searchable image resource of *Drosophila* GAL4-driver expression patterns with single neuron resolution. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2020.05.29.080473 (Accessed 29 December 2022).
13. H. Otsuna, M. Ito, T. Kawase, Color depth MIP mask search: A new tool to expedite Split-GAL4 creation. bioRxiv [Preprint] (2018). https://doi.org/10.1101/318006 (Accessed 29 December 2022).
14. Y. Z. Kurmangaliyev, J. Yoo, J. Valdes-Aleman, P. Sanfilippo, S. L. Zipursky, Transcriptional programs of circuit assembly in the drosophila visual system. *Neuron* **108**, 1045–1057.e6 (2020).
15. M. N. Özel *et al.*, Neuronal diversity and convergence in a visual system developmental atlas. *Nature* **589**, 88–95 (2021).
16. F. Diao *et al.*, Plug-and-play genetic access to drosophila cell types using exchangeable exon cassettes. *Cell Rep.* **10**, 1410–1421 (2015).
17. P.-T. Lee *et al.*, A gene-specific T2A-GAL4 library for Drosophila. *ELife* **7**, e35574 (2018).
18. J. R. Bateman, A. M. Lee, C.-t. Wu, Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange. *Genetics* **173**, 769–777 (2006).
19. K. J. T. Venken *et al.*, MiMIC: A highly versatile transposon insertion resource for engineering Drosophila melanogaster genes. *Nat. Methods* **8**, 737–743 (2011).
20. K.-F. Fischbach, A. P. M. Dittrich, The optic lobe of Drosophila melanogaster. I. A Golgi analysis of wild-type structure. *Cell Tissue Res.* **258**, 441–475 (1989).
21. F. Diao, B. H. White, A novel approach for directing transgene expression in Drosophila: T2A-Gal4 in-frame fusion. *Genetics* **190**, 1139–1144 (2012).
22. F. P. Davis *et al.*, A genetic, genomic, and computational resource for exploring neural circuit function. *ELife* **9**, e50901 (2020).
23. A. Nern, B. D. Pfeiffer, G. M. Rubin, Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2967–E2976 (2015).
24. J. J. Omoto *et al.*, Visual input to the Drosophila central complex by developmentally and functionally distinct neuronal populations. *Curr. Biol.* **27**, 1098–1110 (2017).
25. A. S. Mauss *et al.*, Neural circuit to integrate opposing motions in the visual field. *Cell* **162**, 351–362 (2015).
26. J. C. Tuthill, A. Nern, S. L. Holtz, G. M. Rubin, M. B. Reiser, Contributions of the 12 neuron classes in the fly lamina to motion vision. *Neuron* **79**, 128–140 (2013).
27. M. Wu *et al.*, Visual projection neurons in the Drosophila lobula link feature detection to distinct behavioral programs. *ELife* **5**, e21022 (2016).
28. L. S. Gramates *et al.*, Fly Base: A guided tour of highlighted features. *Genetics* **220**, iyac035 (2022).
29. B. Ewen-Campen *et al.*, split-intein Gal4 provides intersectional genetic labeling that is repressible by Gal80. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2304730120 (2023).