bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

A genetic, genomic, and computational resource for exploring neural circuit function

Fred P. Davis[1,2,*,X], Aljoscha Nern[1,*], Serge Picard[1,3], Michael B. Reiser[1], Gerald M. Rubin[1], Sean R. Eddy[1,4], and Gilbert L. Henry[1,5,X]

[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA. [2]Molecular Immunology and Inflammation Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892, USA. [3]Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544, USA. [4]Howard Hughes Medical Institute and Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [*]Contributed equally. [X]Correspondence to fred.davis@nih.gov (FPD), henry@cshl.edu (GLH)

August 5, 2018

Abstract. The anatomy of many neural circuits is being characterized with increasing resolution, but their molecular properties remain mostly unknown. Here, we characterize gene expression patterns in distinct neural cell types of the *Drosophila* visual system using genetic lines to access individual cell types, the TAPIN-seq method to measure their transcriptomes, and a probabilistic method to interpret these measurements. We used these tools to build a resource of high-resolution transcriptomes for 100 driver lines covering 67 cell types. Combining these transcriptomes with recently reported connectomes helps characterize how information is transmitted and processed across a range of scales, from individual synapses to circuit pathways. We describe examples that include identifying neurotransmitters, including cases of co-release, generating functional hypotheses based on receptor expression, as well as identifying strong commonalities between different cell types.

Keywords: neural circuit, gene expression, visual system

Highlights

1. Transcriptomes reveal transmitters and receptors expressed in *Drosophila* visual neurons
2. Tandem affinity purification of intact nuclei (TAPIN) enables neuronal genomics
3. TAPIN-seq and genetic drivers establish transcriptomes of 67 *Drosophila* cell types
4. Probabilistic modeling simplifies interpretation of large transcriptome catalogs

Introduction

The anatomy of neural circuits is being characterized with increasing resolution and throughput, in part following a dramatic increase in the size of circuits amenable to detailed electron microscopy reconstruction (Swanson and Lichtman, 2016) and the development of genetic tools to access individual cell types (Luo et al., 2018). These efforts reveal anatomy at unprecedented detail, but not the molecular properties of cells. In principle, the genes expressed in each cell of a neural circuit should serve as a molecular proxy for cell physiology. However, most genomic efforts have focused on surveying neuronal diversity rather than characterizing circuit function (Ecker et al., 2017). To develop a resource exploring molecular correlates of circuit function, here we use an approach that genetically targets cell types within a well-characterized brain region to measure high-quality transcriptomes that can be integrated with connectomes.

Drosophila affords an ideal system to study neural circuits in detail, as both excellent genetic tools and high resolution connectomes are available. Here we focus on the repeating, columnar circuits of the visual system, found in the optic lobes, a widely used model for studying circuit development and function with an extensive genetic toolbox and well-described anatomy (Nériec and Desplan, 2016; Silies et al., 2014; Apitz and Salecker, 2014). This network begins with photoreceptor neurons and contains several layers of connected neurons which process incoming luminance signals into multiple parallel streams of visual information. Many of its cellular components have been described by light microscopy, including classical Golgi studies (Fischbach and Dittrich, 1989) and recent analyses using genetic methods (Morante and Desplan, 2008; Otsuna and Ito, 2006; Nern et al., 2015; Wu et al., 2016). Electron microscopy reconstruction work

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

has characterized the synaptic connections of many optic lobe neurons (Meinertzhagen and O'Neil, 1991; Meinertzhagen and Sorra, 2001; Rivera-Alba et al., 2011; Takemura et al., 2013; Takemura et al., 2015; Takemura et al., 2017). Comparative studies have also explored the evolution of this ancient brain structure (Strausfeld, 2009). However, many of its fundamental properties remain unknown, including the neurotransmitters used at many of its synapses. As we show later, knowing these fundamental properties is critical for understanding the mechanisms behind more complex circuit functions, such as motion detection in the visual system.

Measuring the genes expressed in specific cells of the brain is challenging due to its compact and complex organization. RNA sequencing (RNA-seq) addresses this challenge by profiling either single cells or genetically labeled populations of cells (Ecker et al., 2017). The latter approach requires genetic tools to access individual cell types but provides more direct access to cells of interest than sampling of unmarked single cells, especially for sparse cell types. Profiling identified cell types provides a direct link to previous work on the anatomy and physiology of those cell types. Cell type-specific drivers also facilitate follow-up experiments, for example evaluating the role of individual genes in individual cells. In *Drosophila*, large collections of GAL4 driver lines (Jenett et al., 2012; Tirian and Dickson, 2017) and the possibility to further refine these patterns with intersectional methods such as split-GAL4 (Luan et al., 2006; Dionne et al., 2018) enable genetic access to many neuronal populations (see, for example, Tuthill et al., 2013; Aso et al., 2014; Wu et al., 2016). We therefore chose the genetic, rather than single cell, approach to build a genomics resource to explore circuit function. This approach also complements single cell efforts by providing reference transcriptomes upon which unidentified single cell data can be mapped. These reference landmarks are critical for interpreting single cell data, which is made challenging by measurement noise and sparsity (Kolodziejczyk et al., 2015).

We previously developed an Isolation of Nuclei Tagged in a specific Cell Type (INTACT) method (Deal and Henikoff, 2010) to measure transcriptomes and epigenomes of genetically-marked neuronal populations in *Drosophila* (Henry et al., 2012) and mouse (Mo et al., 2015). Here, we develop a tandem affinity purification of INTACT nuclei (TAPIN) method with increased specificity, sensitivity, and throughput. By combining this method with an extensive set of new driver lines with predominant expression in specific cell types and a new probabilistic method to interpret transcript abundance, we build a resource of high-quality transcriptomes for one hundred driver lines. These data provide expression information for 67 *Drosophila* cell types as well as several broader cell populations. Through validation experiments and comparisons to the literature we demonstrate that this resource is useful both for identifying individual genes expressed in specific cell types and for revealing broader patterns such as the expression of all members of a gene family across many cell types. As an example, we provide details of the expression of neurotransmitters and their receptors. We show how this information, when combined with connectomes, leads to specific hypotheses about circuit mechanisms in the *Drosophila* visual system.

Results

Genetic tools for labeling the visual system

To enable transcriptome analyses of defined cell populations, we first assembled a collection of genetic drivers to access them. We focused on cell types in the optic lobes, the fly's visual system, but also included neuronal populations in two central brain regions, the mushroom body and central complex, primarily to serve as informative outgroups (Figure 1A). The optic lobe contains anatomically diverse neurons and glia arranged in a series of neuropil layers: the lamina, medulla, and lobula complex (consisting of lobula and lobula plate) (Figure 1B,C). Each neuropile region can be further divided into sublayers (which largely represent regions of synaptic contacts). The optic lobes have a repetitive structure of ~750 retinotopically arranged visual columns of similar cellular composition. Some optic lobe cell types are present at one cell per column, others are less numerous with cells that each contribute to several columns. For example, the main synaptic region of the first optic lobe layer, the lamina, contains processes of some 13,000 cells but these belong to only 17 main cell types: 14 neuronal and 3 glial (Figure 1C, top row). A small number of additional neurons (lamina tangential cells, Lat) project to a region just distal to the main lamina neuropile. A few additional glial types, not specifically targeted in this study, are located in the space between the lamina and compound eye. Outside the lamina, we sought to include representatives of neurons of different major types (such as local interneurons and projection neurons that connect optic lobe regions or the optic lobe with the central brain). Our driver lines were selected to include the major cell types of the circuits that compute the direction of visual motion (Mauss et al., 2017).

To characterize new driver lines, we imaged expression patterns across the entire fly brain to determine overall driver specificity (Figure 1D, S1) and examined anatomical features such as layer patterns in higher resolution images to identify specific cell types (Table S1, Figure S2). For most lines, we further confirmed the identity of labeled cells by examining the morphology of individual cells using stochastic labeling (Figure S2). We noted that a few patterns also include some additional contaminating cells (Table S1).

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 1

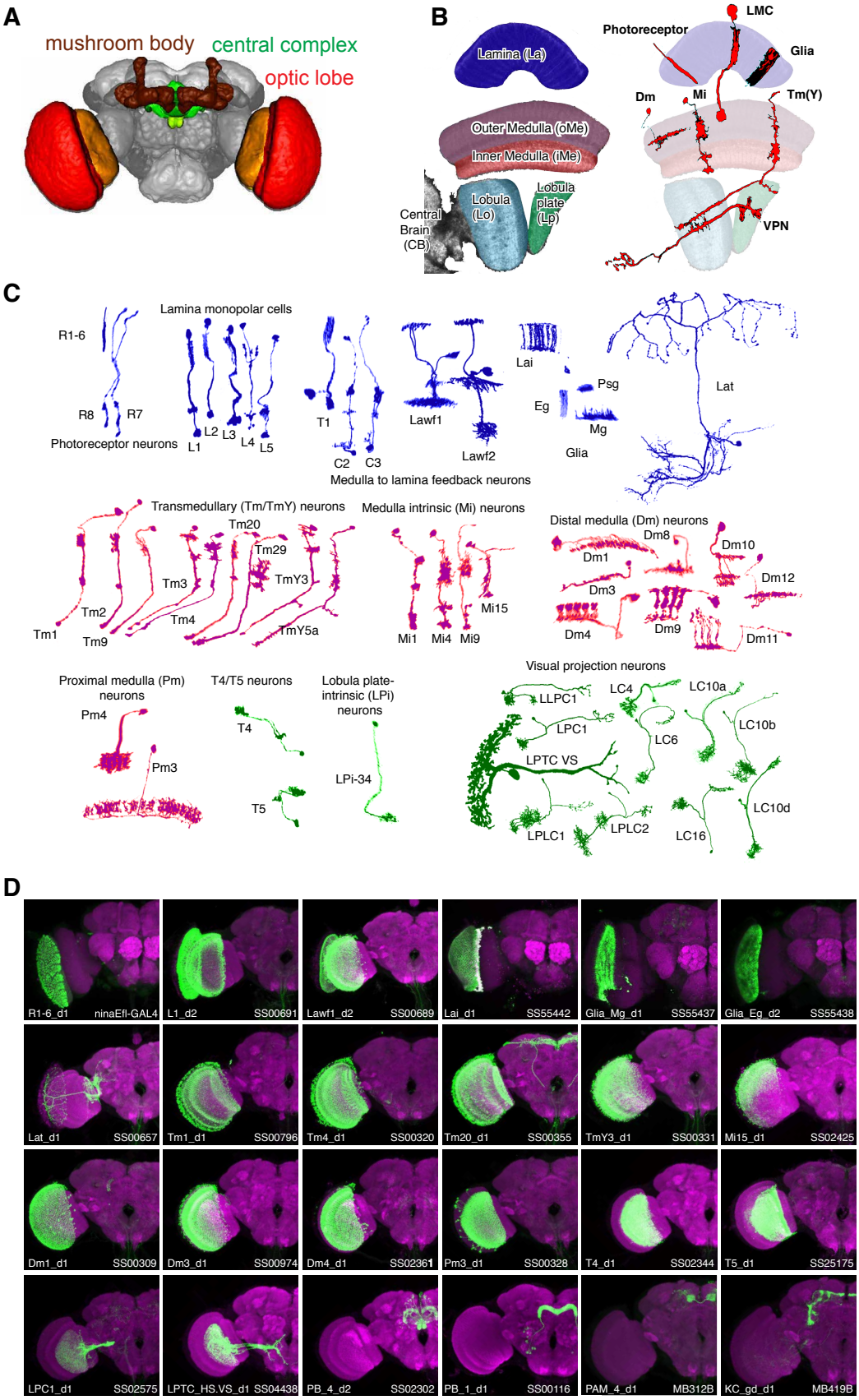


Figure 1: Genetic tools to access cell types in the visual system. A. Major brain regions profiled in this study (brain image from Jenett et al., 2012). B. Left, subregions of the early visual system. Right, examples of layers and neuropil patterns of various classes of visual system neurons. C. We profiled cell types arborizing in the lamina (blue), medulla (purple) and lobula complex (green) of the visual system. Many cells contribute to multiple neuropiles so other groupings are possible. D. Representative expression patterns of driver lines that target specific cell types. Each image is a maximum intensity projection of a whole brain confocal stack (only one optic lobe is shown). In each image the brain is counterstained (magenta) with a neuropil marker and both the targeted cell type and the driver are indicated in the lower left and right corner, respectively. Additional images (focusing on drivers first described in this study) are shown in Figures S1 and S2. Imaging parameters and brightness and contrast were adjusted individually for each image. For genotypes and image details see Table S5.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Purifying nuclei with INTACT and TAPIN

Next, we employed an improved INTACT method to measure nuclear transcriptomes in genetically defined cell populations (Henry et al., 2012), and we also developed a new variant of the method that permits higher throughput with increased purity and sensitivity. In both approaches, nuclei are purified using a nuclear tag whose expression is driven in a cell population of interest by either a standard or split GAL4 driver (Figure 2A). The INTACT protocol adapts a method we previously described in the mouse (Mo et al., 2015) in which we purify nuclei by differential centrifugation, and then bead capture tagged nuclei (Figure S3A). The new variant protocol, tandem affinity purification of INTACT nuclei (TAPIN), uses a bacterial protease (IdeZ) to specifically cleave antibodies in the hinge region separating their Fc and antigen binding F(ab')$_2$ fragments (Figure 2B, S3B). Treating protein A magnetic bead-bound nuclei with this protease generates both nucleus-F(ab')$_2$ and bead-Fc complexes. Soluble nucleus-F(ab')$_2$ is then recaptured on protein G magnetic beads, removing non-specifically bound material from the first capture. INTACT successfully profiled many of the abundant cell types in the optic lobe (> 1000 cells per brain), but failed for sparser cell types and those whose nuclei were difficult to purify by differential centrifugation (photoreceptors, glia, T4, T5). We solved these problems with TAPIN, which does not purify nuclei prior to bead capture.

The greatest advantage of TAPIN is its ability to purify nuclei from sparse cell types ($<$ 50 cells/brain) (Table S1). INTACT is not suitable for these lines because of loss during differential centrifugation. This difficulty cannot be overcome by processing more brains per experiment because differential centrifugation is difficult to scale. TAPIN solves this problem by running a first capture on crude extracts generated from hundreds to thousands of fly heads. The substantial background in this first capture is reduced 5- to 6- fold in a second capture with only a modest decline in both the yield of nuclei and amplified cDNA (Figure 2C).

Measuring transcriptomes with INTACT- and TAPIN-seq

We applied INTACT and TAPIN to the cell populations defined by the genetic drivers we described above (Table S2). Most drivers express in a single anatomically defined cell type or a small group of related cell types. Others target more heterogeneous cell populations sharing a common property (e.g., driver lines aimed at recapitulating the expression of a neurotransmitter marker). Altogether, we built 250 RNA-seq libraries from 242 samples of purified nuclei (46 using INTACT and 196 using TAPIN) and 8 manually dissected samples (Table S2). We estimated relative transcript abundance in each library using kallisto (Bray et al., 2016). Libraries built from more nuclei yielded more cDNA (Figure 2D), had greater numbers of detected genes (Figure 2E), were estimated to have greater physical numbers of transcripts (Figure S3C), had more reproducible transcript abundance (Figure 2F), and exhibited less bias in coverage across gene bodies (Figure S3D,E). We focused on 203 libraries that had at least 8,500 genes detected, 3μg cDNA yield, and 0.85 Pearson's correlation of transcript abundances in two biological replicates. These 203 libraries consist of at least two biological replicates built from 100 drivers that covered 67 cell types (53 visual system, 7 mushroom body, 5 central complex, 2 muscle), 6 broader cell populations (ChAT, Gad1, VGlut, Kdm2, Crz, and NPF), and 2 manually dissected tissues (the lamina and remainder of the optic lobe) (Methods). We provide the read and abundance data for the remaining sub-optimal libraries (47 libraries covering 24 cell types) in the event they may be informative, but we do not consider these to be of sufficient quality and do not consider them further here.

We did not sort the sex of flies when preparing our RNA-seq libraries, as we did not expect large differences between sexes. As a test of this assumption, we prepared libraries for the T4.T5 combination driver using exclusively female or male flies. These two transcriptomes were largely similar, except for differential expression of a small number of genes with known sex-specific regulation including the noncoding genes RNA on X 1 (*roX1*) and *roX2* in males (Amrein and Axel, 1997) and yolk proteins 1 and 3 (*Yp1*, *Yp3*) in females (Belote et al., 1985) (Figure 2G).

We were encouraged by the clear enrichment of previously identified markers in cell types where they were expected. For example, we recovered transcription factors (TFs) previously found in the developing monopolar interneurons and inner photoreceptors (Tan et al., 2015; Figure 2H). While some of these genes showed great separation between the most highly expressed and next highest expressing cell (e.g. *svp*: 1244 TPM in L1 versus 8 TPM in R7), others showed a more continuous spectrum of abundance (e.g. *bab2*: 186 TPM in L2 versus 47 TPM in R7). We further confirmed our measurements by comparing TAPIN-seq results for twelve cell types that were also recently profiled by FACS-seq (Konstantinides et al., 2018; Figure S3G) and found concordant expression of cell type-enriched genes. This concordance also argues against major differences between nuclear and cytoplasmic transcriptomes. In combination with the technical quality of our libraries, this confirmation by independent gene expression measurements validated our approach, and also motivated us to explore how to best interpret a large dataset of relative abundances.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 2

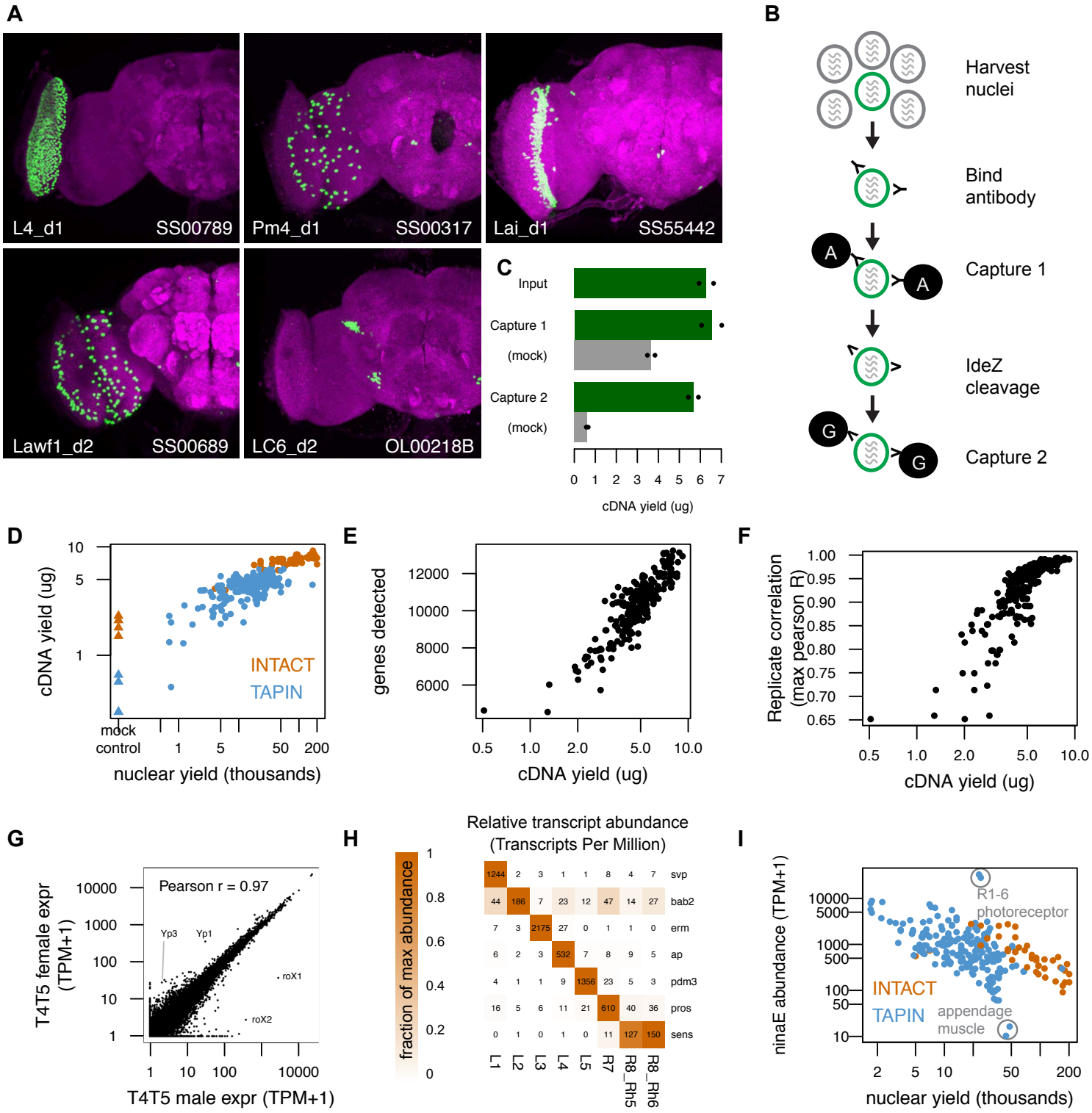


Figure 2: Tandem-affinity purification of INTACT nuclei (TAPIN) enables neuronal genomics. A. Cell type-specific drivers enable expression of the UNC84-2XGFP nuclear tag (green) in specific populations of cells. Both the targeted cell type and driver are indicated in the lower left and right corner, respectively. B. Following nuclei harvest, two rounds of magnetic bead capture serially purify target nuclei. After the first round of protein A bead capture, bacterial protease IdeZ cleaves the anti-GFP antibody in the flexible hinge region, allowing a second round of bead capture with protein G, which recognizes the F(ab')$_2$ region. Protein G, unlike Protein A, can bind both the Fc and F(ab')$_2$ regions of an immunoglobulin. C. Two capture rounds reduce the level of non-specific background (grey bars, mock IgG control) while maintaining the cDNA yield from the captured target nuclei (green bars). Bars represent the mean of two replicates (shown as points). D. RNA-seq libraries created with more nuclei yield more cDNA (circles). TAPIN libraries had lower non-specific background than INTACT (blue vs orange triangles). E. Libraries with more cDNA detect more genes. F. Libraries with more cDNA have more reproducible transcript abundances. G. T4.T5 transcriptomes of female (y-axis) and male (x-axis) flies are well correlated, but also recover known sex-specific genes including RNA on X 1 (*roX1*) and *roX2* and yolk protein 1 (*Yp1*) and *Yp3*. H. Previously identified markers of lamina monopolar and inner photoreceptor neurons (Tan et al., 2015) are enriched in the expected cells. I. Libraries with fewer nuclei had greater carry-over of *ninaE* transcript, which encodes the abundant rhodopsin in the fly eye. The upper outliers are libraries made from R1-6 photoreceptors, the only cells that express *ninaE*. The lower outliers are appendage muscle libraries created after heads are removed from the fly bodies, effectively eliminating *ninaE* carry-over from photoreceptors. See also Figure S3.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Interpreting transcript abundance with mixture modeling

Deriving biological insights from a matrix of transcript abundances is not straightforward. Two main complications arise: (1) cross-contamination between cells during cell separation and library construction; and (2) determining when a low level of expression is biologically meaningful. To address these issues, we developed a statistical model to account for transcript carry-over and to discretize the expression calls. First, we observed expression of photoreceptor marker genes, such as *ninaE*, in unexpected samples, and in inverse correlation with the number of nuclei used to build each library (Figure 2I), suggesting that these transcripts resulted from contamination. *ninaE* is also unexpectedly detected in non-photoreceptor cells in a recent single cell RNA-seq study of the brain (Davie et al., 2018). Other reports have also described unexpected photoreceptor transcripts in both bulk and single cell profiling of the mouse retina (Siegert et al., 2012; Macosko et al., 2015), and attributed them to photoreceptors lysing during tissue homogenization. We optimized our biochemical protocols to minimize such carry-over and then turned to a computational method to address it further. Second, while a cell's expression of a gene can be used to infer a specific functional property of that cell, the level of expression that is needed to establish confidence in such an inference is much less clear. For example, expressing the vesicular acetylcholine transporter (*VAChT*) implies that a neuron is cholinergic. However, *VAChT* transcript abundance exhibits a wide distribution and it is not clear, *a priori*, what level is necessary to conclude that a cell is cholinergic (Figure 3A). Addressing these two issues requires a principled way of deciding which genes are expressed in each sample.

We used mixture modeling to address this challenge by describing the expression levels of each gene as arising from a mixture of two log-normal distributions representing binary 'on' and 'off' states (Figure 3A; Methods). Genes can of course express in more than two states, but we show through extensive validation that this simplifying assumption is a useful one. For example, we modeled *VAChT* expression in the high-quality TAPIN/INTACT-seq libraries to estimate the probability that the gene was expressed in each driver (Figure 3B). The model unambiguously inferred *VAChT* states for all drivers. The most ambiguous call was for the broad and heterogeneous Kdm2 driver, which we estimated to express *VAChT* with a probability (confidence) of 0.95. The model also correctly inferred that only R1-6 photoreceptors expressed *ninaE*, although *ninaE* abundance in other cells reached as high as 2,702 TPM (PAM_1) (Figure 3C,D). Gene-specific models are critical because of differences in dynamic range: 1,000 TPM reflects the off state for genes like *ninaE*, but the on state for more modestly expressed genes like *VAChT* (Figure 3B,D). We used this method to transform our catalog of transcript abundances to probabilities of expression (Figure 3E). To further simplify these probabilities, we discretized them into on ($p \geq 0.8$) and off ($p \leq 0.2$) states, and otherwise considered them to be ambiguous ($0.2 < p < 0.8$). The expression states inferred for replicates had a median 95% concordance (Figure S4A). We combined information from replicates to infer expression at the driver and cell type levels (Methods).

We found many genes that express in all cell types, and many that express in only one, with a range in between (Figure 3F,G). As expected given their roles in specifying identity, homeobox transcription factors (TF) expressed more specifically than transcription factors in general (Figure 3G). Neuropeptides also expressed specifically, while genes with the more general function of synaptic vesicle endocytosis were broadly expressed. We explore these functional properties in more detail later (Figure 4C). Across all genes, we observed a wide spectrum of transcriptional output with a median on-state abundance of 10 TPM and dynamic range of 5-fold between on and off states (Figure 3H,I).

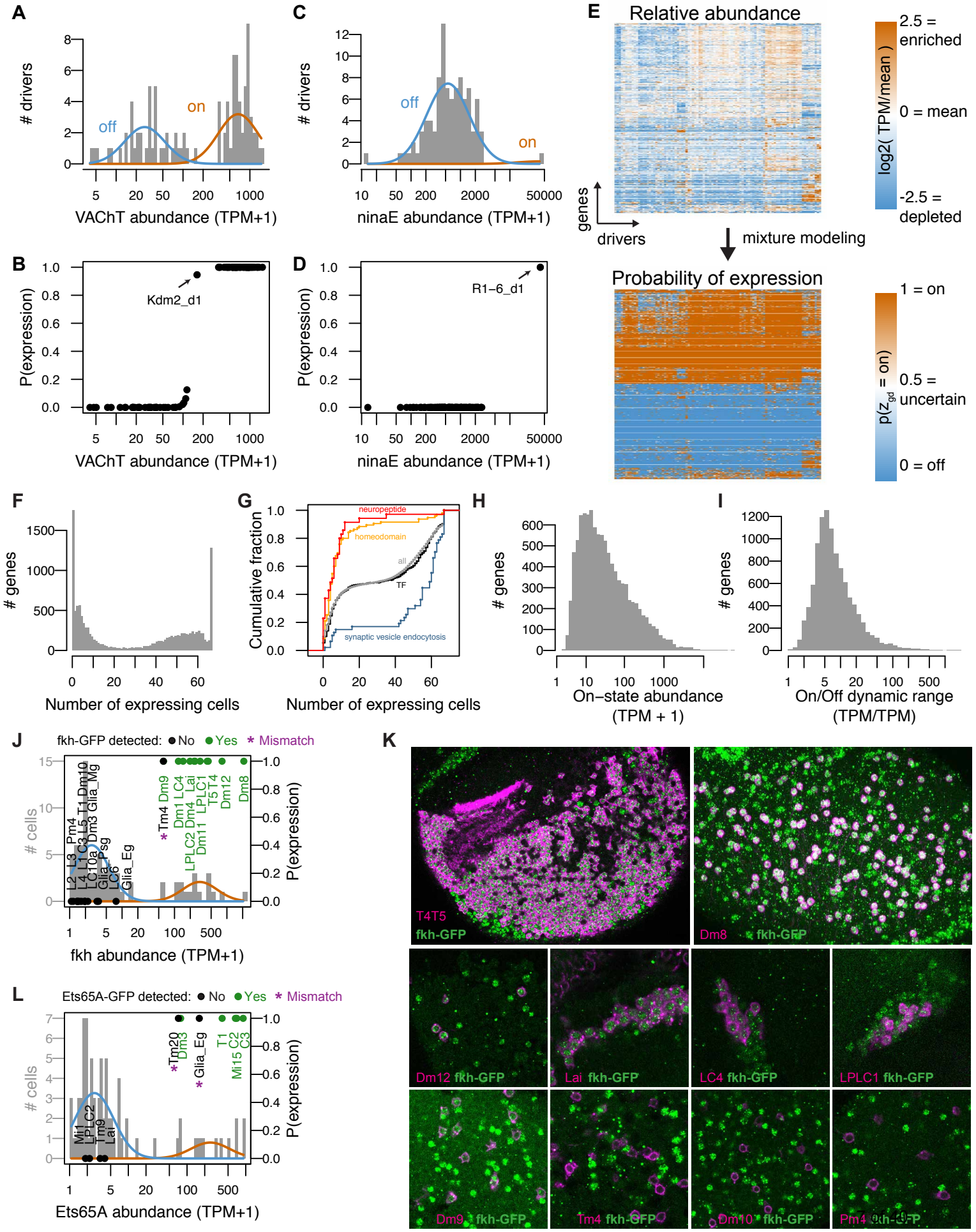
Evaluating accuracy of TAPIN-seq measurements

To validate our TAPIN-seq measurements, we first compared our inferred expression states to FlyBase curated reports of protein expression (n=197 data points of gene/cell pairs; 4 negative points, 193 positive points; n=22 cells; n=69 genes, Table S3) and found 93% concordance (183 matches; 14 mismatches from six genes; 0 mismatches for negative benchmark points; Figure S4B). The benchmark mismatches fell into three categories: expression levels near the transition between inferred on and off components (Figure S4C-E), genes with a wide dynamic range of expression (*Syx*, *Rab11*; Figure S4F-G), and genes with undetected transcript but previously detected protein (*Myo61F*; Figure S4H). The first two categories likely arise from imprecision in the model's fitted components and its inability to describe continuous, rather than, bimodal expression. The third category (conflicting transcript and protein levels) could reflect either technical issues (low sensitivity in our measurements, or false positives in the prior work due to antibody cross-reaction) or biological complexities (e.g., long-lived transcripts, subcellular localization).

To further evaluate our results for genes expressed across a wide range of levels, we compared the model output to protein expression patterns for two transcription factors: Forkhead (*fkh*) and *Ets65A*. We visualized each protein using a C-terminal GFP tag; the tagged proteins were expressed from BAC transgenes with large flanking sequences to ensure a near native genomic context (Kudron et al., 2018). From the transcript data, we inferred *fkh* gene expression in 14 cell types across a 35-fold range of abundance (60 to 2,103 TPM). Of 28 cell

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 3



bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure 3: Mixture modeling accurately interprets TAPIN-seq measurements. A. The distribution of Vesicular acetylcholine transporter (*VAChT*) abundance fit with a mixture of two log-normal components. B. Interpreting these components as "off" and "on" states unambiguously infers expression state in essentially all drivers. C,D. Modeling the distribution of ninaE correctly distinguishes true expression by R1-6 from transcript carry-over in the remaining samples. E. Mixture modeling transforms our catalog of relative transcript abundances (top) to discretized expression states (bottom). F. Histogram of expression breadth per gene. G. Cumulative distributions of expression breadth for all genes (gray), transcription factors (black), homeobox TFs (orange; InterPro domain IPR001356), neuropeptides (red), and genes involved in synaptic vesicle endocytosis (blue). H. Distribution of mean on-state transcript abundance across all modeled genes. I. Distribution of dynamic range across all modeled genes. J,K. Evaluating *fkh* modeling results at the protein level with a BAC transgenic (See Figure S4I). J. Histogram bars represent raw abundance of all cells in our catalog. Blue and orange curves represent the inferred off and on components, respectively. Points represent the cells tested for transgene expression showing either detectable GFP (Green) or no signal (Black). The points' vertical position reflect the estimated probability of gene expression. K. Forkhead-GFP expression in selected cell types. Fkh-GFP (mainly nuclear, in green) and cell type-specific expression of a membrane marker (in magenta) are shown. Because of the wide range of fkh expression levels, imaging parameters and brightness and contrast adjustments are not identical for different panels. Cells with detectable nuclear GFP signal above the background in the same image were scored as expressing fkh. L. As in J, to evaluate *Ets65A* modeling results (See Figure S4J). See also Figure S4.

types that we visualized at the protein level, fkh was detected in all but one that we expected from TAPIN-seq (Figure 3J,K, S4I). The sole exception, Tm4, has a *fkh* abundance (60 TPM) near the border between the inferred off and on states (Figure 3J). However, we did detect protein in Dm9, which had a near identical raw transcript abundance (61 TPM). Similarly evaluating *Ets65A* expression identified two mismatches out of 11 tested cells (Figure 3L, S4J). Ets65a protein was not detected in Tm20 (70 TPM) and epithelial glia (161 TPM), while it was weakly detected in Dm3 (77 TPM). These results further support the accuracy of TAPIN-seq and our statistical model even for genes with a wide dynamic range. The agreement between our transcript on/off calls and protein expression encouraged us to use the discretized on/off calls for all further analyses; the unprocessed relative abundances in TPM are reserved for deeper analysis when needed.

Identifying genes that mark cell types and groups

Examining the relation between cell types using transcriptomes

To study the relation between cell types, we built a dendrogram based on inferred expression states and estimated the support for each branch point with bootstrap resampling (Figure 4A). The broad groupings were well supported and mostly intuitive: muscle were outgroups, followed by a mushroom body cell type (PAM_4), the glia, the photoreceptors, and the remaining neurons. Several fine groupings of anatomically closely related neurons were also well supported (e.g., Kenyon cells; C2,C3; Lawf1, Lawf2; T4, T5; LPLC1, LPLC2). However, mid-level branchings were not well supported, indicating the lack of a simple hierarchical relationship. Neurons were generally grouped by region: central complex, mushroom body, and optic lobe. One surprise was the grouping of Tm20 and Dm1, away from all other optic lobe cell types. Upon closer examination, the identity of genes expressed exclusively in these two lines (*lz*, *Pdh*, *bw*) suggest that this grouping is driven by shared pigment cell contamination in the GAL4-tagged patterns of these driver lines. Similarly, the unusual position of PAM_4 is likely due to some unidentified non-neuronal cells in the driver. These are examples of imperfections in the GAL4 driver lines. While they can lead to some false positives for the main target cell types, they can also provide additional information. For example, analyzing the overlap between Tm1 and Dm20 allowed us to infer marker genes expressed in the pigment cell population.

Transcriptomes identify genes enriched in individual cell types and groups

We next identified genes that marked cell groups in the tree, using three criteria: genes that expressed in all the cells within a group, at most two cells outside this group, and with transcript abundance higher than all cells outside the group (For simplicity, we will hereon refer to cell type as just cell). We used these criteria to identify markers for photoreceptors (n=108), glia (n=60), and muscle (n = 76) (Figure 4B, Table S4). These genes included many known as well as new markers. For example, genes enriched in photoreceptors include signaling components (*Arr2*, *Galphaq*) and transporters (*trpl*, *Eaat2*) with known physiological roles as well as uncharacterized orphan transporters (e.g., *CG8468*). We also identified 18 markers for pigment cells using the Tm20 and Dm1 profiles. In addition to the three types of lamina glia we profiled, several other glia types are present in both the lamina and the medulla. Genes expressed exclusively in the dissected samples (lamina, remainder of optic lobe) and not in the TAPIN libraries identified marker genes for optic lobe cells that we did not directly profile, such as glia. Indeed, the genes identified in this way included several known markers for astrocytes (*alrm*, *wun2*, *Obp44a*) (Huang et al., 2015).

We examined the breadth of expression of different functional groups of genes, as defined by FlyBase gene group curation. HOX-like homeobox TFs were among the most specifically expressed group, while groups of core cellular machinery (e.g., beta importins, mitochondrial complexes) were among the most broadly expressed groups (Figure 4C). Some groups included both broadly and very specifically expressed genes. For example, among cell adhesion molecules, we noted an

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 4

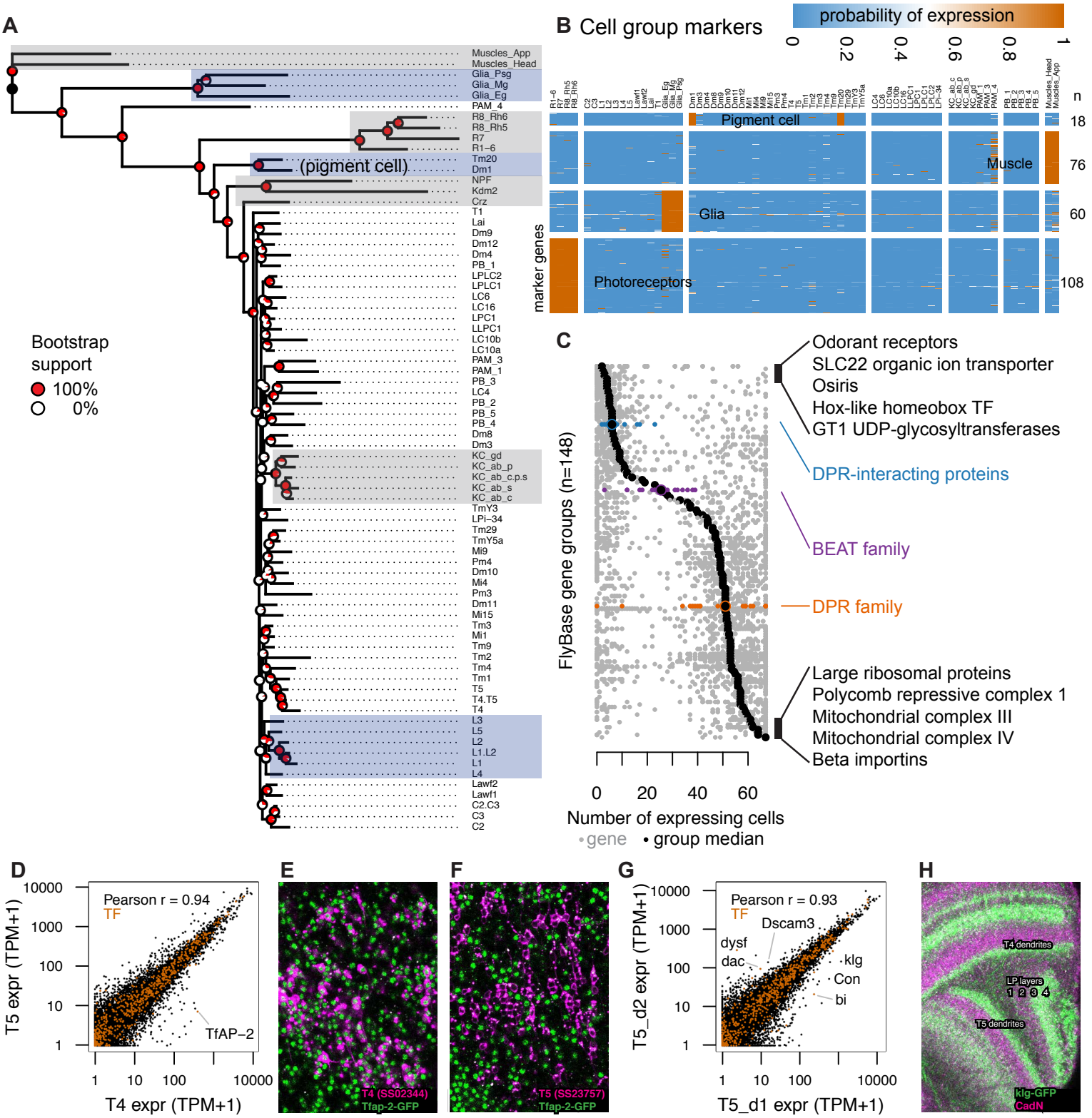


Figure 4: TAPIN-seq profiles identify genes enriched in cell types and groups. A. Cells grouped by a minimum evolution tree of their inferred expression states. B. Heatmap of marker genes enriched in photoreceptors, glia, muscle, and pigment cells. C. Distribution of expression breadth for genes in terminal FlyBase gene groups with more than 10 members in our expression probability matrix. The least- and most- broadly expressed gene groups are labeled, along with the DPR-interacting, beat and DPR family of extracellular proteins. D. *TfAP-2* transcription factor distinguishes closely related cell types T4 and T5. E,F. TfAP-2 protein is specifically expressed in T4 and not in T5, confirming TAPIN-seq. GFP-tagged Tfap-2 (mainly nuclear, in green; see Table S5 and Methods) is shown together with a membrane marker (magenta) expressed in T4 (F) or T5 (G) cells. G. Identification of genes with differential expression in very closely related cell types probed by driver lines for T5 that differentially label layers of the lobula plate (corresponding to different subtypes of T5 cells). H. Confirming our TAPIN-seq data, klg protein (detected using a GFP tag (green); see Table S5 and Methods) is expressed in T4/T5 cells with the expected layer specificity (layers 3 and 4) in the lobula plate (LP). A neuropil marker is shown in magenta. See also Figure S5.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

interesting distribution for three gene groups proposed to be involved in protein-protein interactions that underlie synaptic connectivity (Özkan et al., 2013; Tan et al., 2015). While the 11 DPR-interacting proteins (DIP) were among the most specifically expressed genes (expressed in a median of 6 cells), beat (median, 25.5 cells) and DPR (median, 51 cells) genes were more broadly expressed (Figure S5A-D). As physical interactions among these and other extracellular proteins have been systematically characterized (Özkan et al., 2013), we combined their expression and interaction patterns to estimate the number of potential interaction between cells in the lamina (Figure S5E), many of which are in actual contact (Figure S5F). We found that every pair of lamina cells expressed tens of interacting protein pairs, highlighting the broad potential for cell-cell interactions not only in the developing (Tan et al., 2015) but also adult optic lobe. However, except for a clear paucity of interacting protein pairs expressed by glia, these global expression-based patterns did not correlate well with connectivity in the lamina.

Transcriptomes can distinguish closely related cell types and subtypes

We asked if we could identify genes distinguishing closely related cell types. For example, T4 and T5 had similar transcriptomes and were neighbors in the phylogenetic tree, but we found one transcription factor, *TfAP-2*, that was expressed nearly two orders of magnitude higher in T4 (390 TPM) than T5 (6 TPM) (Figure 4D). We confirmed this pattern at the protein level (Figure 4E,F).

T4 and T5 cells can each be further divided into four subtypes that preferentially respond to motion in one of four cardinal directions and differ in anatomical details such as the lobula plate layer to which they project axons. While our split-GAL4 lines do not isolate single T4/T5 subtypes, the T5_d1 and T5_d2 drivers show differences in subtype expression (Figure S1B,B',C,C'). Comparing the transcriptomes of these two drivers confirmed previously described markers (*Con*, *bi*, *dac*; Apitz and Salecker, 2018) that distinguish T4/T5 cells of lobula plate layers 1/2 and 3/4, and indicated additional genes, including a transcription factor (*dysf*) and cell adhesion molecules (*klg*, *Dscam3*) with selective expression in these subtypes (Figure 4G). As a further confirmation of this finding, we verified that klg shows a layer-specific protein pattern (Figure 4H).

Reference bulk transcriptomes are necessary to interpret single cell transcriptomes

While preparing our paper, single cell RNA-seq (scRNA-seq) maps of the brain (Davie et al., 2018) and optic lobe (Konstantinides et al., 2018) were reported. scRNA-seq is commonly used to survey cellular diversity, however (as also noted by Konstantinides et al.) the 52 single cell clusters (7 of which are glia) found in the optic lobe far under-estimates its over one hundred anatomically distinct neuronal cell types. This result suggests that either scRNA-seq misses some cell types or that multiple cell types can be clustered together. To discern these possibilities, we compared the single cell map to our transcriptome catalog (Figure S7C). Using the reported marker genes, we found that only a few single cell clusters had markers clearly enriched in one or two cell types (e.g., C3, Lawf1, Lai, T1, T4/T5), and that most clusters had markers either enriched in multiple cell types, or without clear enrichment in our data. Although this result could also arise from major errors in our TAPIN-seq profiles, this possibility is unlikely given our earlier validation results and the concordance between our TAPIN-seq profiles and cell type-enriched genes identified from independent FACS-seq measurements (Figure S3G). A more likely explanation is that noisy scRNA-seq measurements make it challenging to accurately identify clusters and marker genes, and subsequently assign cell types. Highlighting the challenge of assigning cell types, we found that the number of cells in each single cell cluster often under-represented or over-represented the true abundance of the reported cell type labels (ranging from 3.4 times fewer T4/T5 cells to 7.9 times more Pm3 cells than expected), indicating differential representation in the scRNA-seq map or inaccurate cell type assignments (Figure S7D). Altogether, these results suggest that cell type-identified data is critical for interpreting single cell maps, as these maps may not proportionally represent every cell type in a tissue, and the inferred cell clusters can each correspond to multiple cell types.

Profiles reveal neurotransmitter output for most neuron types

The proteins that synthesize and transport neurotransmitters are well known, enabling us to use their expression to predict neurotransmitter phenotype. We used histamine decarboxylase (*Hdc*), glutamate decarboxylase (*Gad1*), the vesicular acetylcholine transporter (*VAChT*), and the vesicular glutamate transporter (*VGlut*) to identify potential histaminergic, GABAergic, cholinergic, and glutamatergic cell types, respectively (Figure 5A-D). Our model unambiguously inferred expression states for these genes and indicated a single transmitter (from this group) for nearly all neurons we profiled. A second cholinergic marker, choline acetyltransferase (*ChaT*), matched *VAChT* expression almost perfectly (the two genes also share an exon). The sole exception, apparent expression of *ChAT* but not *VAChT* in R7 photoreceptors, likely results from a subset of dorsal rim R8 cells labeled by the R7 driver line (further discussed below, also see Table S1). In contrast to *Gad1*, we found that the vesicular GABA transporter (*VGAT*; Fei et al., 2010) expressed in nearly all cells (except R-cells, glia and muscles), making it an unreliable marker of GABAer-

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

gic neurons; it may have additional functions, or it may be post-transcriptionally regulated, which would be consistent with observed restricted VGAT immunostaining (Enell et al., 2007; Fei et al., 2010).

Besides these four neurotransmitters that we identified by one or two marker genes, we also identified candidate dopaminergic neurons based on the combined expression of tyrosine hydroxylase (*ple*), dopa decarboxylase (*ddc*), vesicular monoamine transporter (*Vmat*) and dopamine transporter (*DAT*). While *DAT*, *ple*, and *ddc* were also expressed individually in several cell types that did not express *Vmat*, only known dopaminergic cell types and one medulla neuron (Mi15) expressed this combination (Figure 5A).

Transmitters for nearly half of our cell types have been previously proposed and generally agree with our results. For example, *VAChT*/*ChaT* expression in Kenyon cells supports recent reports showing they are cholinergic (Barnstedt et al., 2016; Crocker et al., 2016). Fluorescence in situ hybridization and immunolabeling guided by our measurements confirmed the expression of *ChaT*, *Gad1*, and *VGlut* in Mi1, Mi4, and Mi9, respectively (Long et al., 2017; Takemura et al., 2017). However, we see considerable differences between our assignments and some previous work that used reporter transgenes (Varija Raghu et al., 2011; Raghu and Borst, 2011; Raghu et al., 2013), which we generally attribute to unfaithful transgene expression patterns. We believe our assignments to be more reliable, however they are not without problems. For example, one assignment we made that seems unlikely and is not supported by other available data is the presence of *Gad1* in Mi9, which was not detected in the FISH or antibody experiments mentioned above. Given the presence of some contaminating Mi4 cells in at least one Mi9 driver and the lower *Gad1* abundance (mean 276 TPM in Mi9; 2165 TPM in Mi4; 1870 mean TPM in predicted GABAergic cells), we attribute the Mi9 *Gad1* signal to contaminating contributions from other GABAergic cells such as Mi4.

Transcriptional regulation of neurotransmitter output

We next tried to identify transcriptional regulators of neurotransmitter output, by searching for TF genes expressed in strong correlation with transmitter phenotype. However, we only found such TFs for histaminergic output (Figure S6A). This observation agrees with work on neuronal identity showing that single TFs rarely encode transmitter identity, but rather different TF and TF combinations are used to specify the same neurotransmitter output (Hobert, 2016). We thus expanded our search to TFs whose expression was informative about transmitter phenotype (i.e., cells expressing TF A are likely to produce neurotransmitter B; even if not all cells producing neurotransmitter B express TF A; Figure S6A). This search identified candidate TFs for nearly all neurotransmitter types. For example, the 19 neuronal types (including the broad chat-GAL4 line) expressing apterous (*ap*) are cholinergic. Its worm ortholog, *ttx-3*, regulates the cholinergic phenotype of the AIY neuron (Wenick and Hobert, 2004). Several other TFs we identified also have worm or mouse orthologs implicated in neuronal identity (Figure S6B). Several TFs appeared to identify a transmitter phenotype within a group of cell types but not across the entire dataset. For example, *Lim3* distinguishes the GABAergic Dm10 from the other Dm cell types in our dataset and is also expressed in several other GABAergic cells (Mi4, Pm3, Pm4) but was also detected in the cholinergic LC4 and the glutamatergic TmY5a and Tm29. We confirmed the differential Lim3 protein expression in Dm10 and Dm12 cells (Figure S6C). Several of the transcription factors that we found to be informative of neurotransmitter output were also implicated by single cell RNA-seq data, including *ap* (cholinergic), *tj* (glutamatergic), and *Lim3* (GABAergic) (Konstantinides et al., 2018). Our data also indicate exceptions to these patterns (i.e., neurons expressing *tj* and *Lim3* but with a different neurotransmitter phenotype; Figure S6A). These observations indicate that neuronal features are likely regulated in a context-dependent and combinatorial manner, and that transcriptomes can identify putative regulators.

Examples of non-canonical transmission

Although the transcriptomes implicated a single canonical neurotransmitter for most neuron types, there were a handful of interesting exceptions that suggest either no canonical neurotransmitters or co-transmission. We also see examples of expression of neurotransmitter-associated genes by cells that do not themselves release transmitter, such as glia, which likely provide evidence for transmitter recycling mechanisms (Figure 5D).

One neuronal cell type, T1, expressed none of the neurotransmitter markers *VGlut*, *VAChT*, *Vmat*, and *Gad1* (Figure 5A). Although T1 does express most pan-neuronal genes, it does not express bruchpilot (*brp*), a key component of presynaptic active zones. Consistent with this result, EM reconstruction has identified very few T1 presynaptic specializations (Takemura et al., 2008).

Co-release of multiple neurotransmitters can enhance the signaling capacity of neurons and neural circuits. For example, the same cell type might release different transmitters under distinct conditions or use them to elicit distinct responses in different target cells. In addition to Mi9 (discussed above as being likely due to contamination), we observed two cases of potential co-transmission involving the canonical small molecule neurotransmitters. Both Mi15 drivers express dopaminergic and cholinergic markers, and both R8 drivers expressed cholinergic and histaminergic markers. We confirmed expression of Vmat protein in Mi15 (Figure 5E), the first identified dopaminergic cell type within the optic lobe, and further

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 5

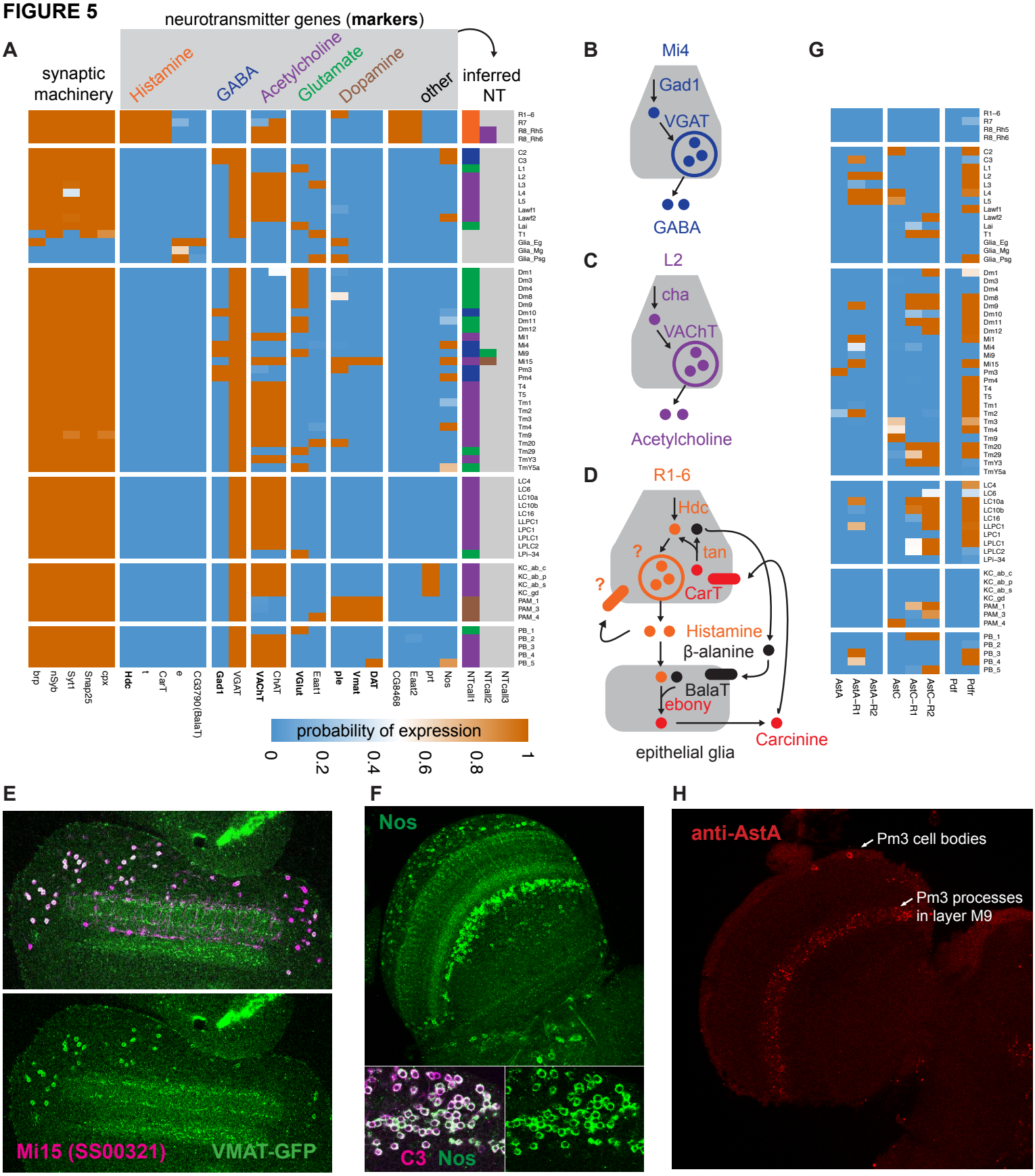


Figure 5: Expression of synthesis and transport genes establish neurotransmitter phenotypes. A. Expression of neurotransmitter marker genes indicate the neurotransmitters produced in nearly all profiled cells. B, C. Example of marker genes for canonical small molecule transmitter GABA and Acetylcholine. D. Proposed histamine recycling mechanism supported by expression of beta-alanine transporter BalaT in epithelial glia. E,F. We confirm TAPIN-seq results at the protein level (green) for (E) Vesicular monoamine transporter (Vmat) expressed in Mi15 (magenta) and (F) Nitric oxide synthase (NOS) in C3 (magenta). Top panel in (G) shows a section through the optic lobe, lower panels C3 cell bodies. G. Several neuropeptides and receptors also express specifically (examples). H. Allatostatin A (AstA) protein expression in the optic lobe matches the distribution and layer pattern of Pm3 cells. See also Figure S6.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

below we confirm the unexpected *VAChT* expression in R8 (Figure 7A).

Evidence for co-transmission involving additional molecules, such as neuropeptides or nitric oxide, appears frequently in our data set. Nitric oxide is a widely conserved signaling molecule that can act on many kinds of cells, including neurons (Lowenstein and Snyder, 1992). We observed very specific expression of its synthesizing enzyme, nitric oxide synthase (*Nos*), in the lamina (C2, C3, and Lawf2) and medulla (Mi4, Pm4, Tm4 and Mi15). To further validate these results, we confirmed Nos expression at the protein level in C3 neurons (Figure 5F). Nitric oxide can be released extra-synaptically, potentially enabling signaling between neurons that are not synaptic partners.

Several neuropeptides and their receptors were also expressed quite specifically, suggesting widespread peptidergic signaling in the visual system (Figure 5G). In some cases, expression of neuropeptides and their receptors aligned with specific known synaptic connections (Takemura et al., 2013), for example the *AstC* neuropeptide in L4 and the *AstC-R1* receptor in TmY3. Other cases suggest volume transmission or peptide release from cell types not in our dataset. *AstA* is only clearly expressed in the Pm3 cells of the medulla, while the *AstA-R1* receptor expressed in Mi1, Tm2, Mi15, and Dm9. Consistent with transcript levels, published AstA expression patterns (Hergarden et al., 2012) include Pm3-like cells in the medulla; we confirmed this AstA protein expression in Pm3 cells (530 TPM, p(on) = 1) but did not detect expression in Tm2 cells (26 TPM, p(on) = 0.22) (Figure 5H). As expected, pigment-dispersing factor (*Pdf*) was not detected in any of the high quality libraries (but is present in the Pdf neuron and lLNv samples). By contrast, we observed broad (though not ubiquitous) expression of the pigment-dispersing factor receptor (*Pdfr*) in the optic lobe, consistent with the extensive arborizations of Pdf-expressing neurons at the surface of the medulla. Previous work has reported Pdfr expression in several clock neuron types but not in optic lobe neurons (Im and Taghert, 2010).

While we focused on genes with well known functions, our expression patterns can also suggest new functions for poorly characterized genes (Figure 5A,D). For example, photoreceptors specifically expressed *CG8468*, an orphan transporter in the solute carrier 16 (SLC16) family. This gene might represent a candidate vesicular or plasma membrane transporter of histamine, which remains unidentified in any species.

Broad and patterned expression of neurotransmitter receptors

Since the functional consequences of the release of a neurotransmitter depend on which receptors for this transmitter are expressed in the receiving cell, measuring the expression of both neurotransmitter input and output genes is necessary to assign potential synaptic signs to connectomes. For example, glutamatergic transmission in *Drosophila* may be either inhibitory or excitatory, depending on the receptors.

In general, neurotransmitter receptors are broadly expressed, qualifying each cell type to detect multiple neurotransmitters (Figure 6A). Patterns for individual receptors (or receptor subunits) varied widely. Some receptors, such as the *GluClalpha* glutamate-gated chloride channel, were expressed in most but not all cell types (Figure 6A,B). Expression of others was much more restricted, such as the EKAR glutamate receptor subunit only detected in photoreceptor neurons. Nearly all cells expressed receptors for acetylcholine, GABA, and glutamate, as expected from the combination of predicted transmitter phenotypes and connectomics data. Receptors for neuromodulators such as serotonin, dopamine, octopamine, and neuropeptides in general were also widespread. For example, octopamine receptors were expressed in broad, yet gene- and cell-type specific patterns, consistent with widespread octopaminergic modulation of visual processing (for example, Arenz et al., 2017; Strother et al., 2018; Tuthill et al., 2014). We confirmed Oamb expression at the protein level in specific lamina neurons and glia, including Lawf2 cells previously shown to be octopamine sensitive (Tuthill et al., 2014) (Figure 6C).

Combining transcriptomes and connectomes

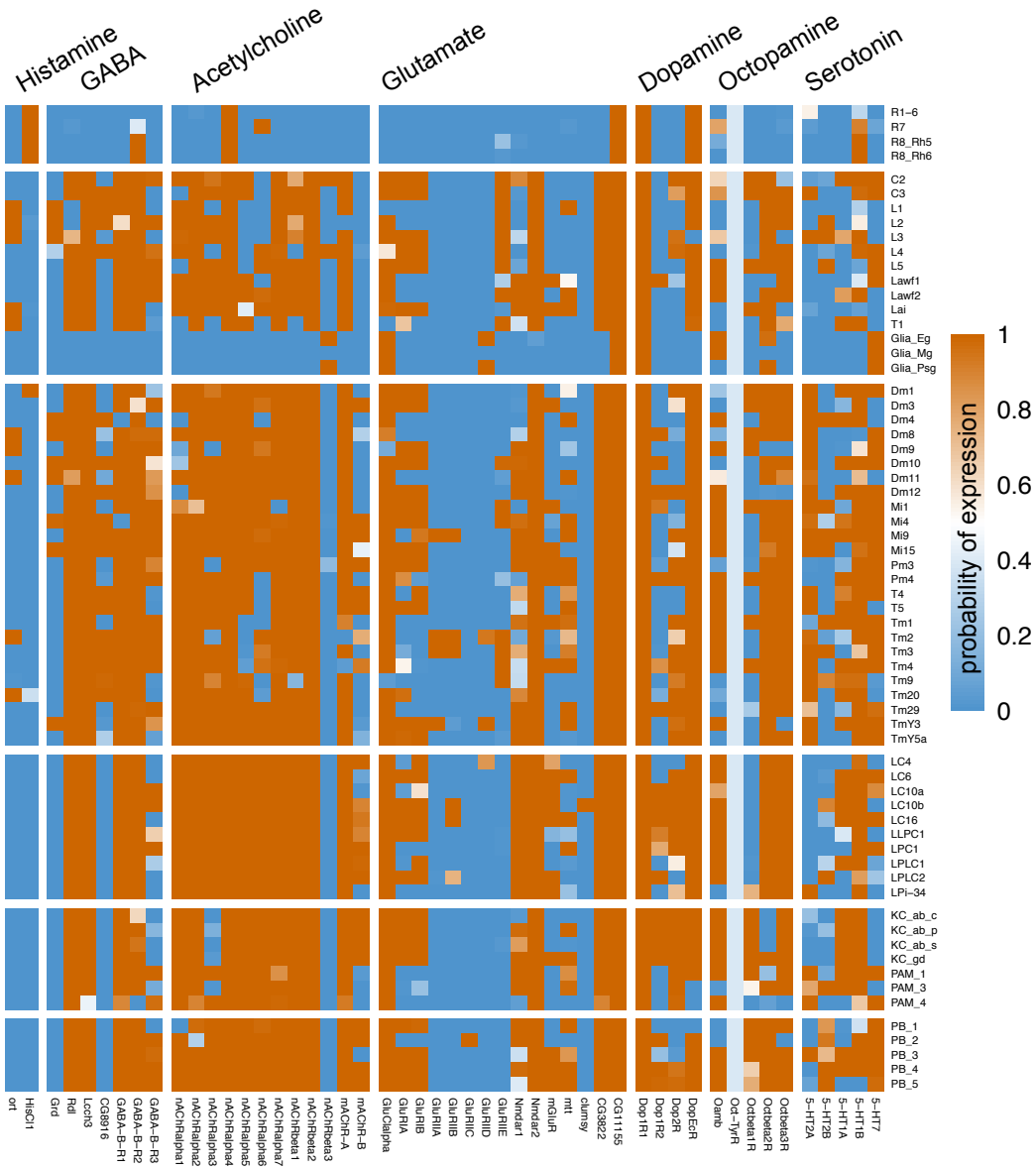
A principal goal of our work is to provide a foundation for combining neurotransmitter and receptor expression patterns with anatomical or functional connectivity data. One application of expression information is to constrain mechanistic models of neural circuits such as the extensively studied motion detection circuit in the fly eye (reviewed in Mauss et al., 2017). For example, for the ON and OFF motion detection pathways that supply inputs to directionally sensitive T4 and T5 neurons, respectively (Takemura et al., 2017), our results show that all of the inputs to T5 (Tm1, Tm2, Tm4, and Tm9) are cholinergic, whereas the inputs to T4 are a mixture of GABAergic (C3, Mi4), cholinergic (Mi1, Tm3), and glutamatergic (Mi9), suggesting different input signs (Figure S7A). Discovering the functional signs of inputs to the directionally selective neurons is an essential step in understanding the mechanism of this long-studied neuronal computation (Strother et al., 2017). In addition, our data reveals aspects of the motion pathway that have not yet been functionally examined, such as the identification of other signaling components (e.g. *Nos*; Figure S7B).

The combined availability of expression and connectomics data for many cell types in a brain region also makes it possible to systematically identify and further explore unusual patterns of receptor or transmitter expression; for example, cell types in which an otherwise

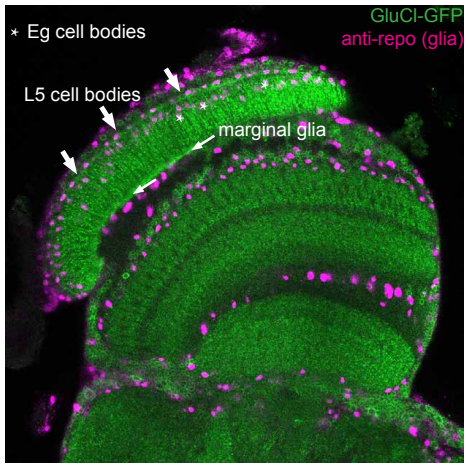
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 6

A Neurotransmitter input



B



C

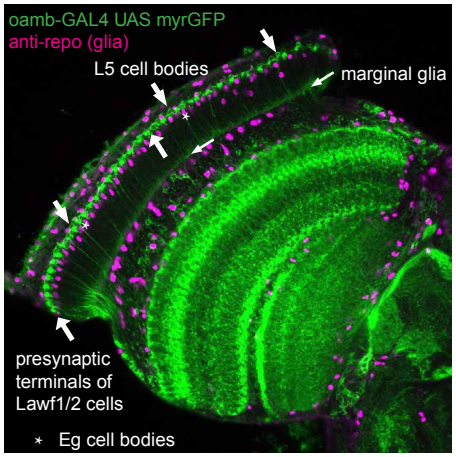


Figure 6: Patterns of neurotransmitter receptor expression. A. Neurotransmitter receptors are widely expressed in specific patterns. With the exception of histamine, most cells express receptors or receptor subunits for nearly all neurotransmitters. B. Expression of the glutamate-gated chloride channel (GluClalpha), detected using a GFP-tag (green), in the optic lobe. The lamina pattern includes L5 neurons and proximal satellite, epithelial and marginal glia. A glia-specific nuclear marker (anti-repo) is shown in magenta. C. Octopamine receptor (Oamb) expressing cells in the optic lobe detected with a protein-trap GAL4 driving expression of a membrane targeted GFP (green). Anti-repo (magenta). In the lamina (to the top and left of the image), Lawf1/2 and L5 neurons and marginal glia are recognizable.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

widely expressed receptor is absent or cells with unusual combinations of receptor subunits. Below we discuss three examples, focused on potential signs of synaptic transmission, of how such patterns can lead to specific, unexpected hypothesis about circuit function. The first, focused on photoreceptor output, originated from a global search for mismatches between neurotransmitter expression and the presence of appropriate receptors in postsynaptic partners identified by EM. The second uses expression patterns of GABA-A receptor subunits to suggest sites and molecular indicators of non-canonical depolarizing GABA-ergic transmission. The third uses differential expression of glutamate receptor subunits to draw inferences about the similarity of two neuron types.

i. R8 photoreceptors are cholinergic as well as histaminergic

Fly photoreceptors have long been known to release histamine (Hardie, 1987; Sarthy, 1991). Our data indicate that inner (color vision) R8 photoreceptors also express the cholinergic markers ChAT and VAChT, suggesting an unexpected additional cholinergic phenotype (Figure 5A). We independently confirmed these results by using a genetic approach (Pankova and Borst, 2017) that allowed us to visualize a tagged VAChT protein (VAChT-HA), expressed from the endogenous locus, selectively in photoreceptor cells. These experiments showed VAChT-HA labeling in medulla terminals of R8 cells (Figure 7A), including the specialized polarized light-responsive R8-cells in the dorsal rim of the medulla. The latter express the rhodopsin Rh3 (which is otherwise expressed in R7s; Fortini and Rubin, 1990), consistent with the presence of Cha and VAChT transcripts in the R7 driver line (for which the model inferred expression for VAChT but not ChaT). By contrast, we did not detect VAChT-HA in R1-6 and R7 photoreceptors outside the dorsal rim using this method.

We asked whether the apparent co-transmitter phenotype of R8 neurons was reflected in the expression of neurotransmitter receptors in their different postsynaptic partners. Histaminergic transmission by photoreceptors occurs via the histamine-gated chloride channels ort and HisCl1 (Pantazis et al., 2008). Postsynaptic partners of R8 cells identified by electron microscopy reconstructions (at least 5 synapses in Takemura et al., 2013) include seven cell types in our dataset: Dm9, Mi1, Mi4, Mi15, R7, L1 and Tm20 (Figure 7B) (Takemura et al., 2013; Takemura et al., 2015). All of these express one or more nAChR subunits (Figure 6A). *HisCl1* and *ort* expression was more selective (Figure 7B,C): L1, Tm20 and Dm9 express *ort*, consistent with previous reports (Gao et al., 2008), while *HisCl1* transcripts were detected in the R7 as well as R8 driver lines, in agreement with another recent report (Schnaitmann et al., 2018; Tan et al., 2015). However, we did not find evidence of expression of *ort* or *HisCl1* in Mi4, Mi1 and Mi15, consistent with R8 signaling via a transmitter other than histamine.

We were interested in whether release of ACh and histamine might occur at spatially distinct locations or whether the two transmitters could potentially be co-released. Insects synapses often consist of multiple postsynaptic sites apposed to the same presynapse (Figure 7D). We used EM reconstruction data (Takemura et al., 2013) to map the predicted expression of histamine receptors in postsynaptic cells at the single synapse level for all presynaptic sites of one reconstructed R8 cell (Figure 7E). The resulting pattern indicates that processes of cell types with and without histamine receptor expression are often grouped at the same R8 presynapse (Figure 7E), whereas this is not the case for a reconstructed R7 cell (Figure 7F). This is consistent with the VAChT-HA labeling observed throughout the medulla terminals of R8s (but not in the axons of these cells in the lamina) (Figure 7A).

A combined cholinergic and histaminergic phenotype has been reported for a small group of extraretinal photoreceptors (the Hofbauer-Buchner eyelet) located near the lamina (Yasuyama and Meinertzhagen, 1999) but was unexpected for R-cells of the compound eye. Establishing the functional significance of potential acetylcholine release by R8 cells will require further experiments. However, we note that double mutants lacking both histamine receptors are not completely blind (Gao et al., 2008), consistent with histamine-independent transmission by photoreceptor neurons. In view of the widespread acetylcholine receptor expression and the grouping of postsynaptic processes described above (Figure 7D,E), acetylcholine co-release could also influence the response of R8 targets that express *ort* or *HisCl1*.

ii. Potentially excitatory GABA-A receptors in lamina monopolar cells

Fast GABAergic transmission via GABA-A receptors is a major source of inhibition in the nervous system. However, some GABA-A subunit combinations could mediate depolarizing GABA-signaling: in vitro assays indicate that homomeric Rdl or heteromeric Rdl/Lcch3 receptors are typical GABA-gated chloride channels (Zhang et al., 1995), while Lcch3/Grd form GABA-gated cation channels (Gisselmann et al., 2004). However, the *in vivo* significance of this difference is unknown. *Rdl* and *Lcch3* were expressed in nearly all neurons in our dataset (Figure 6A, Figure 7G,H), consistent with the general inhibitory nature of GABA signaling. By contrast, *Grd* and another predicted GABA-A receptor subunit, *CG8916*, were expressed in a minority of cell types (Figure 6A, Figure 7G,H). Photoreceptor neurons, for which no major GABAergic inputs have been identified by connectomics, expressed none of the four transcripts (Figure 7H). Lamina monopolar L1 and L2 were the only neurons other than photoreceptors that did not express significant lev-

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE 7

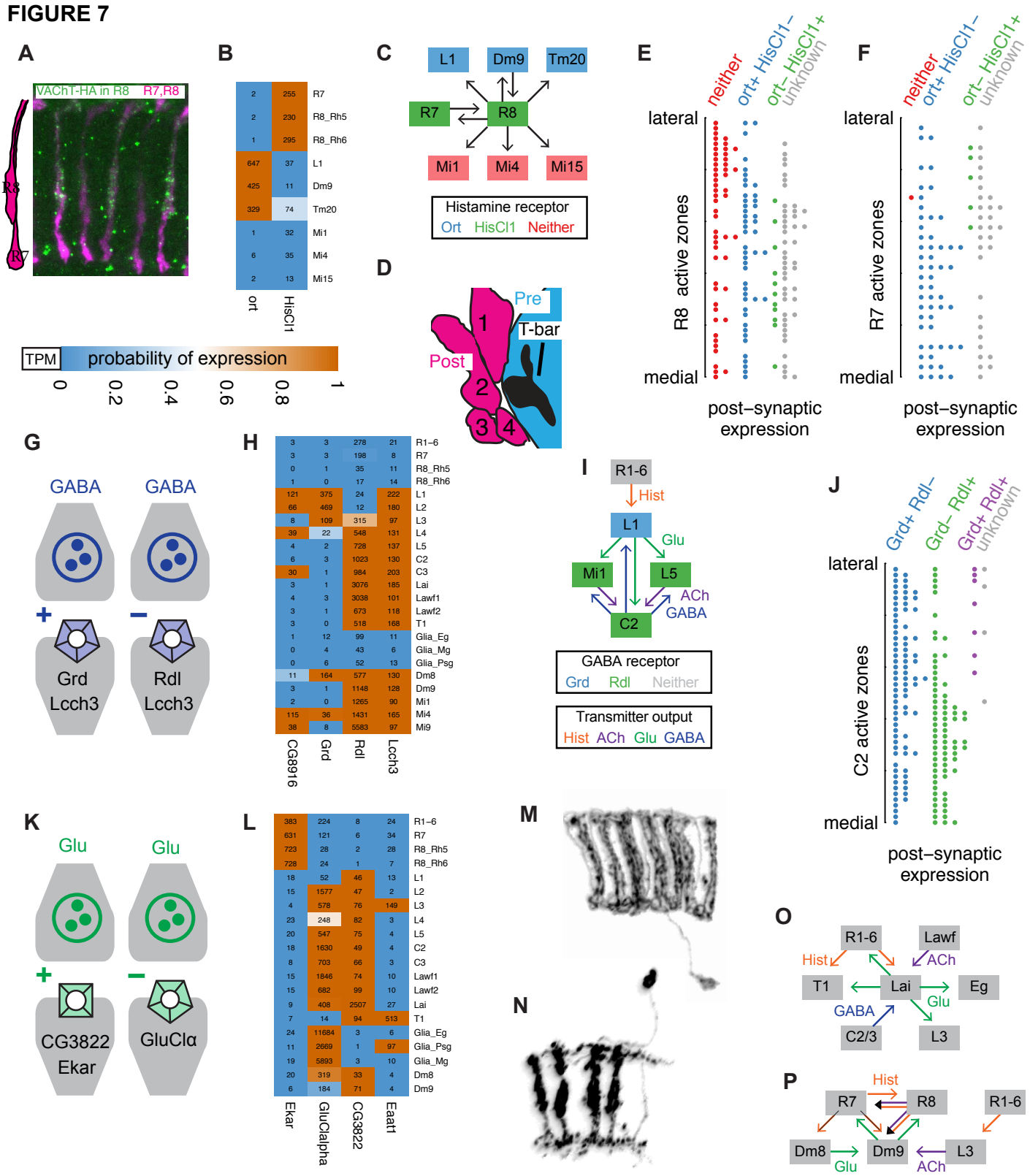


Figure 7: Using gene expression to functionally interpret circuit structure. A. Expression of VAChT in R8 cells. Expression of a HA-tagged VAChT was induced in R8 cells by recombinase-mediated excision of an interruption cassette from a modified genomic copy of the VAChT gene (Pankova and Borst, 2017). Single confocal section shows R7 and R8 cells in magenta and anti-HA immunolabeling in green. B. Heatmap of receptor expression probabilities (color) and relative abundance (numbers; transcripts per million) in R8 targets identified by EM (Takemura et al., 2013). C. Connectivity network for R8 cells, overlaid with receptor expression. Only cells with five or more presynaptic inputs from R8 that are included in the RNASeq dataset are shown. D. Individual R8 active zones can interact with multiple postsynaptic partners. E. Classification of postsynaptic cells at individual R8 active zones based on histamine receptor expression. F. Same analysis as in E but for an R7 cell. G. Different properties of GABA-A receptors in *Drosophila* observed in *in vitro* studies. GABA-A receptor subunits can form either cation or anion channels depending on subunit composition. H. Expression of GABA-A subunits in selected cell types, as in B. I. L1 and two of its target cells form strong reciprocal connections with C2 neurons. J. Distribution of *Rdl* and *Grd* expressing cells at individual C2 synapses. K. Glutamate receptors can also be excitatory or inhibitory. L. Examples of expression patterns for selected glutamate receptors and transporters, as in B. M,N. Morphology of Lai (M) and Dm9 (N) cells. Illustrations based on MCFO images of single cells. O, P. Input and output pathways of Lai (O) and Dm9 (P) neurons. See also Figure S7.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

els of *Rdl*. However, both express *Grd*, *Lcch3* and also *CG8916*. Together with the in vitro findings mentioned above, this result suggests that some or all GABA-A receptors in L1 and L2 may be cation rather than chloride channels. Remarkably, lamina monopolar cells in the housefly *Musca*, which are thought to have very similar functional properties to those in *Drosophila*, depolarize in response to GABA (Hardie, 1987) but hyperpolarize in response to histamine (via ort-containing chloride channels). Thus our data identify a potential link between in vivo electrophysiology, in vitro receptor properties and cell type differences in GABA-A subunit (Rdl or Grd) expression.

Based on synapse counts and our transmitter data, the main GABAergic inputs to L1 and L2 are C2 and C3 neurons (Meinertzhagen and O'Neil, 1991; Rivera-Alba et al., 2011; Takemura et al., 2013; Takemura et al., 2015). Conversely, L1 is the main input to both C2 and C3 cells, followed by the cholinergic L1 targets L5 and Mi1. These strong connections (illustrated for C2 in Figure 7I) indicate that the effective sign of GABA input to L1 and L2 is almost certainly of functional significance. In the illustrated circuit (Figure 7I), L1 cells hyperpolarize in response to luminance increases (as histamine from photoreceptors opens ort chloride channels). The resulting reduced secretion of glutamate is thought to depolarize L1 targets such as Mi1 (via closing of GluClalpha channels). One plausible, though speculative, scenario, is that, similar to Mi1, C2 cells also depolarize in response to light. In this case, GABA-gated cation channels in L1 (formed by Grd and Lcch3) would enable negative feedback (counter-acting) from C2 to L1, which for example could return the membrane potential closer to resting levels – speeding up the response to subsequent luminance changes. By contrast, opening of conventional GABA-A receptors (GABA-gated chloride channels) in L1 would resemble a light response (opening of histamine-gated chloride channels), and thus provide positive (reinforcing) feedback in this case. The latter possibility appears less consistent with the transient nature of the L1 (and L2) response to light (Järvilehto and Zettler, 1971; Laughlin and Hardie, 1978). Distinguishing these and other possibilities will of course require future experimental work.

Similar to the findings for histamine receptors described above (Figure 7E), we observed that cells with different GABA-A profiles can be postsynaptic at the same synapse (Figure 7J). In addition to L1 and L2, *Grd* expression indicated several other candidates for cells with unusual GABA responses (Figure 6A, 7H). In these neurons (e.g., Dm8 or Mi4), *Rdl* and *Grd* were detected together, raising questions such as whether their subcellular distribution is synapse-specific or whether these subunits might co-assemble into channels with yet unexplored properties.

iii. Similarities between glutamatergic interneurons Lai in lamina and Dm9 in medulla

The glutamate gated chloride channel GluClalpha, thought to be the main mediator of inhibitory glutamatergic transmission in flies, was broadly expressed but predicted to be absent from some neurons, including photoreceptor cells (Figure 6A, 7K). Another glutamate receptor subunit, *EKAR* (*CG9935*), was only detected in photoreceptors, consistent with previous work (Hu et al., 2015). These unusual receptor expression patterns prompted us to explore cellular sources of and potential functions for glutamatergic signaling to photoreceptors.

Photoreceptor neurons function over an extremely wide range of light levels, from moonlight to bright sunlight. One potential mechanism enabling this behavior has been proposed whereby a depolarizing feedback signal from photoreceptor targets increases photoreceptor output under low light conditions, but is reduced at higher light intensities (Zheng et al., 2009). As Lai cells express ort, and thus, like other ort-expressing photoreceptor targets, are thought to hyperpolarize in response to light, increased glutamate release from Lai could provide such light-dependent feedback via EKAR in R-cells, consistent with reduced photoreceptor responses at low light intensities after reduction of Lai output or EKAR function (Hu et al., 2015). Lai is the only source of vesicular glutamate release in the lamina, although T1 and L3 might also influence glutamate levels in the lamina via the Eaat1 plasma membrane glutamate transporter. (The strong expression of this transporter in T1 rather than glia is another unusual feature of this cell type that is probably a clue to its enigmatic function; Tuthill et al., 2013.) Other Lai targets in the lamina differ from photoreceptors in their glutamate receptor profiles: e.g., epithelial glia highly express *GluClalpha*, which is absent from photoreceptor neurons, but not *EKAR* (Figure 7L). Lai itself also expresses several glutamate receptors, in particular the glutamate receptor subunit *CG3822*. Since Lai is not postsynaptic to any glutamatergic cells, these receptors must be pre- or extrasynaptic. Indeed, CG3822 was recently reported to function presynaptically in homeostatic control of signaling at the neuromuscular junction (Kiragasi et al., 2017). These examples highlight the diversity of glutamatergic signaling in the lamina and add to the list of examples in which transmitter released by a neuron is predicted to have very different effects on target cells depending on the receptors they express.

Connectomic data identify Dm9 as a potential counterpart of Lai, serving a similar role in the medulla. Dm9 expresses *ort* and is both a major pre- and postsynaptic partner of R7 and R8; it is the only identified R7/R8 target with these properties (other known R7 or R8 targets appear to form few if any feedback synapses on these cells). The overall anatomy of Dm9 cells is also similar to Lai (Figure 7M,N): Both Lai and Dm9 cells span multiple visual columns but the precise number and distribu-

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

tion of columns innervated by each individual cell is variable. Based on connectivity and gene expression (Figure 7L,P), Dm9 cells are predicted to receive hyperpolarizing R7 and R8 input via ort and excitatory input from the photoreceptor targets L3 and Dm8. Thus, similar to Lai (Figure 7O), Dm9 appears qualified to increase photoreceptor output in the medulla under low light conditions, similar to the proposed function of Lai in the lamina.

However, there are also notable differences between Lai and Dm9 associated circuits. For example, there are no obvious counterparts of the interactions of Lai with T1 and glia in the medulla, though this could be partly due to less complete anatomical and expression data (i.e., we did not profile medulla glia, and they are also not included in current connectomes). In contrast to Lai, Dm9 cells receive input from photoreceptor neurons with different spectral tuning. This input involves direct (R7, R8) and indirect pathways (R7 via Dm8, R1-6 via L3) (Figure 7P). This integration of multiple spectral inputs could support a role of Dm9 in color processing. Indeed, Dm9 matches the anatomical and predicted functional properties of an as yet unidentified ort expressing cell type proposed to contribute to color opponent signaling between R7 and R8 cells (Schnaitmann et al., 2018).

Discussion

We present an approach to characterize the function of neural circuits by combining genetic tools to access their component cells, TAPIN-seq to measure their transcriptomes, and a probabilistic model to interpret these measurements. We used this approach to establish an extensive resource of the genes expressed in 67 *Drosophila* cell types, including 53 in the visual system, covering photoreceptors, lamina, and components of the motion detection circuit. Our approach enables an extensive analysis of neurotransmission in the *Drosophila* visual system, including the neurotransmitters sent and received across the network as well as transcription factors that potentially regulate neurotransmitter identity. We also provide specific examples of integrating transcriptomes and connectomes to illuminate circuit function.

Many recent studies have explored gene expression in neurons. However, only a few of these were aimed at neurons in genetically tractable organisms and brain regions for which detailed anatomical data, especially at the level of synaptic connections, are available. Previous work in the mouse retina has used both genetic (Siegert et al., 2012) and single cell approaches (Macosko et al., 2015) to characterize transcriptional regulators as well as classify cell types. More recent work in *Drosophila* used single cell RNA-sequencing to characterize heterogeneity in olfactory projection neurons (Li et al., 2017), the midbrain (Croset et al., 2018), the optic lobe (Konstantinides et al., 2018), and the whole brain (Davie et al., 2018). The expression patterns of many genes have also been mapped in worm neurons, whose connectivity has long been known, although these studies typically focus on individual genes rather than genome-wide catalogs (Hobert, 2016). The unique combination of an extensive genetic toolbox to access individual cell types in the *Drosophila* visual system, and systematic efforts to map its connectivity, make it well suited for exploring whether a comprehensive catalog of gene expression is useful for understanding circuit function. Towards this end, we profiled a diverse array of cell types including all of the neuronal cell types that populate the lamina and a subset of cell types in the medulla and lobula complex including those known to play a central role in the detection of motion. We also analyzed a number of cell types residing in deeper brain structures such as the mushroom body and central complex.

Our approach requires genetic driver lines to obtain transcriptomes of specific cell populations. For this study, we combined drivers from existing collections for cell types in the lamina (Tuthill et al., 2013), the mushroom body (Aso et al., 2014), and the lobula (Wu et al., 2016) with new driver lines for many additional optic lobe cell types and also some neurons of the central complex (Wolff et al., 2015; Wolff and Rubin, 2018; T. Wolff, personal communication). Nearly all of these drivers were generated using an intersectional method, split-GAL4, to refine expression patterns of GAL4 driver lines. The recent availability of large collections of reagents for split-GAL4 intersections (Dionne et al., 2018; Tirian and Dickson, 2017) make it possible to obtain such lines for virtually any cell type of interest. This expanding genetic toolbox works well with our TAPIN-seq method to profile transcriptomes.

In some cases, available driver lines, including some used in this study, may label some additional cell types. While drivers with even higher specificity could be obtained through testing of additional split-GAL4 intersections or perhaps triple intersections (Dolan et al., 2017), we did not find the contributions of small numbers of “off-target” cells to be a major limitation for many applications of expression data. Our results indicate that employing multiple drivers for a cell type, a common strategy used in behavioral studies, may also be a viable approach for refining expression data. In general, the transcriptomes support the high specificity of the intersectional lines we used to access visual system cells (Figure 1). For example, we found specific expression of known marker genes (Figure 2H, 4B) and also that most neurons only express genes for a single neurotransmitter type (Figure 5A). The availability of these genetic tools also makes it possible to validate our transcriptome measurements in a way that is otherwise difficult for single cell RNA-seq studies. Driver lines also permit repeated access to the same cell type in multiple animals at defined time points, enabling the study of behavioral or circadian conditions in individual cell types without having to sequence the whole brain or dissected brain regions.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Modifying the one-step affinity capture in the original INTACT method to a two-step capture in TAPIN-seq increased its specificity, sensitivity, and throughput without the need for time-consuming and labor-intensive centrifugation steps (Figure 1). We initially tried improving the original INTACT method by using density gradient centrifugation to purify nuclei prior to the bead capture step, but this was cumbersome, low throughput, and ineffective for sparse cell types. In addition, for reasons that remain unclear, both photoreceptors and T4 cells consistently yielded few nuclei with this approach. Even with TAPIN, the libraries obtained with some sparser driver lines did not meet the quality control standards we applied. We suspect that the quality of these sub-optimal libraries can be improved by starting with more flies, which is simplified by TAPIN-seq's ability to use frozen material, enabling the collection of many flies on multiple days at defined time points. In contrast, manual or FACS sorting of dissociated cells is more challenging to scale up, because these more labor-intensive tissue procurement schemes cannot be simplified in the same way. It is also worth noting that our tandem affinity purification approach can improve the specificity of any immunopurification method that uses a capture antibody that is cleavable by IdeZ (all IgG subclasses), without requiring expression of a traditional TAP tag (Rigaut et al., 1999).

TAPIN-seq complements single-cell RNA-seq studies of neurons in several ways (Ecker et al., 2017; Konstantinides et al., 2018). First, our high-resolution transcriptomes will serve as a reference for interpreting single-cell measurements. In particular, comparing our expression catalog to a recent single cell map of the optic lobe highlights the challenges in interpreting single cell measurements. Several cell types that we profiled don't appear as clusters in the single cell map, while others are grouped into the same cluster. Having both deep bulk transcriptomes and single cell maps of the same tissue provides an opportunity for developing new analytical tools that can harness available cell type-identified information while clustering single cell data. Second, combining our approach with single-cell profiling could more efficiently profile heterogeneity within a brain region or genetically defined cell population. Finally, the complementarity between bulk and single-cell measurements extends to other genomic features that can be measured in TAPIN-seq purified nuclei, including accessible chromatin and modified histones. We expect this combination of genomic tools to help decipher the transcriptional and epigenetic regulation of neuronal expression programs.

Transcriptome measurements can be of limited utility because it is challenging to interpret relative transcript abundance. In this study we developed a probabilistic mixture modeling approach to classify relative abundances into binary on and off states. Although the expression of some genes are not accurately described with a simplified two-state model (e.g., *Rab11*; Figure S4G), this model was a useful guide for interpreting our expression measurements. Even for specific genes where expression is more continuous than bimodal (e.g., DPR family members; Figure S5D), the results still offer a useful family-wide summary of expression patterns (e.g., DPR genes are more broadly expressed than DIP genes; Figure 4C). Despite our model's utility, it is important to remember the many potential sources of error (minor cell types in driver line patterns, transcript carry over during TAPIN, biases in RNA-seq library construction and sequencing, etc) that can affect measurements of relative transcript abundance and the resulting model inferences. Having observed most discrepancies between our modeling results and protein-level expression near the boundary between on and off states, it is prudent to treat these cases more carefully. Our bimodal model could also help interpret other genomic measurements, such as chromatin accessibility or histone modification, that capture inherently binary genomic processes.

The resource provides additional foundation for systematic functional and molecular studies of the *Drosophila* visual system. We illustrated how the resource can characterize neurotransmission in the network, particularly when combined with connectome information detailing connectivity between cell types as well as the grouping of post-synaptic partner cell types. We determined neurotransmitters used by every cell we profiled and found two likely cases of co-transmission (Figure 5A). The expression patterns of the major fast-acting transmitters histamine, acetylcholine, glutamate and GABA were comparatively simple: Nearly all neuronal cell types in our catalogue appear to express exactly one of these three transmitters. However, the transcriptomes suggest that many cells also have the potential to release specific neuropeptides, other chemical messengers such as nitric oxide, or form gap junctions with other cells.

While selected transmitter markers (e.g. Gad1 or VGlut) could also be assigned to cell types using methods such as immunolabeling or in FISH, these approaches are not practical for comprehensive sampling of markers across these different modes of cell-cell communication. This is particularly clear when the expression patterns of neurotransmitter receptors are also considered (Figure 6A). Our results suggest that, for canonical small molecule transmitters, neurotransmitter output space is tightly tuned while input space is not: neurons typically speak just one main language but can understand many (Figures 5 and 6). The expression patterns of neurotransmitter receptors provide further context for determining circuit mechanisms (Figure 7). Our results also implicate transcription factors involved in regulating neurotransmitter phenotype, including several that appear to have conserved roles in specifying neuronal identity in other species (Figure S6).

The availability of connectivity data for many neurons in the visual system allowed us to interpret neurotransmitter use and receptor distribution in the context of cir-

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

cuit architecture (Takemura et al., 2013; Takemura et al., 2015; Rivera-Alba et al., 2011). For example, the co-transmission suggested by expression of both histaminergic and cholinergic markers in R8 photoreceptors was corroborated by receptor expression in its synaptic targets identified by electron microscopy (Figure 7E). In contrast, R7 only expresses the histaminergic marker *Hdc*, and all of its targets express a histamine receptor (Figure 7F).

Finally, our approach especially complements ongoing efforts to map circuit connectivity, which is complete for *C. elegans*, and is becoming accessible on a whole brain level for *Drosophila* (Zheng et al., 2018), and for portions of the mouse brain such as the retina. Methods to obtain and interpret serial electron micrographs, array tomography and other methods for mapping connectivity are rapidly progressing (Swanson and Lichtman, 2016; Micheva and Smith, 2007; Kebschull et al., 2016). All told, we are entering a period in neuroscience where connectomics will become pivotal. We expect that genomic approaches, such as the methods for data collection and analysis that we describe here, will enhance these efforts by using transcriptomes to provide, at high-throughput, a molecular proxy for physiological features that are otherwise inaccessible to connectomic methods.

Methods

Contact for reagent and resource sharing

Further information and requests for resources and reagents should be directed to the Lead Contact, Gilbert L. Henry (henry@cshl.edu). A detailed description of split-GAL4 hemidrivers (`https://bdsc.indiana.edu/stocks/gal4/split_intro.html`) and cell-type specific split-GAl4 lines is also available (`https://www.janelia.org/split-GAL4`).

Experimental models and subject details

Flies were reared on standard cornmeal/molasses food at 25°C. For profiling experiments adults, 4-7 days of age, were entrained to a 12:12 light:dark cycle and anesthetized by CO_2 at ZT8 - ZT12. Samples can be stored indefinitely at -80°C after flash freezing in liquid N_2. We used female flies for all anatomical characterizations.

Method details

Anatomical analyses

Details of individual genotypes and labeling methods used in the characterization of the driver lines and other anatomical experiments are summarized in Table S5. Details of the driver lines are provided in Table S1. For the naming of RNA-seq samples, we identified all drivers with a main cell type or cell types (e.g. Mi9_d1). Most of these cell types have been described in detail and were identified based on prior descriptions (see references in Table S1). The driver names do not attempt to include additional cells present in some drivers. A few of our cell types are strictly groups of related cell types (for example, the muscle cells or, at a different level of a cell type hierarchy, the T4 and T5 cells, with four subtypes each, or R7 photoreceptor neurons, which include R7s of pale and yellow ommatidia).

Generation and characterization of new driver lines

Split-GAL4 and GAL4 driver lines (Table S1) were used to express UNC84-2XGFP in defined cell populations. New split-GAl4 lines were generated as in previous work (Tuthill et al., 2013; Wu et al., 2016). Briefly, we first identified GAL4 lines with expression in the cell type of interest by screening images of the expression patterns of large collections of such lines (Jenett et al., 2012; Tirian and Dickson, 2017). Typically, several candidate combinations of AD- and DBD-hemidrivers were tested to identify lines with sufficient specificity.

To characterize new driver lines, we examined both overall expression pattern in the brain and optic lobe and, for most lines, confirmed the identity of the main cell type or types using MultiColor FlpOut (MCFO)-labeled single cells (Nern et al., 2015). Since details of the expression patterns of GAL4 or split-GAL4 driver lines can depend on the particular UAS reporter used, we re-imaged 20 drivers with the TAPIN nuclear marker used for the profiling experiments (Figure 2A). In general, the distribution of labeled nuclei in these images appeared to match the expression patterns and specificity expected from the driver line's original characterization using a membrane marker. As expected, a small number of off-target cells were detectable (often more weakly labeled) in many driver lines.

Validation experiments

For validation experiments, we examined expression patterns of tagged proteins expressed in a near native genomic context using either large BAC-transgenes or modifications of the endogenous loci (Nagarkar-Jaiswal et al., 2015; Diao et al., 2015; Kudron et al., 2018).

We classified fkh-GFP and Ets65A-GFP as expressed or not expressed by comparing nuclear GFP signal in cells of interest (identified using a split-GAL4 driver) to background labeling in surrounding cells. Because of considerable differences in the GFP signal for different cell types, confocal settings and post-imaging adjustments were done individually for different cell types for these experiments.

The following transgenes were used (also see Table S5): PBac{y[+mDint2] w[+mC]=fkh-GFP.FPTB}VK00037 (RRID:BDSC_43951), PBac{y[+mDint2]

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

w[+mC]=Ets65A-GFP.FLAG}VK00037 (RRID:BDSC_38640), Mi{PT-GFSTF.0}Nos[MI09718-GFSTF.0] (RRID:BDSC_60278), Mi{Trojan-GAL4.1}Oamb[MI12417-TG4.1] (RRID:BDSC_67506), Mi{Trojan-GAL4.1}Lim3[MI03817-TG4.1] (RRID:BDSC_67450), Mi{PT-GFSTF.1}klg[MI02135-GFSTF.1] (RRID:BDSC_59787), Mi{PT-GFSTF.2}GluClalpha[MI02890-GFSTF.2] (RRID:BDSC_60533), Mi{PT-GFSTF.0}TfAP-2[MI04611-GFSTF.0] (RRID:BDSC_61776), pJFRC12-10XUAS-IVS-myr::GFP in attP2 (RRID:BDSC_32197), pJFRC19-13XLexAop2-IVS-myr::GFP in su(Hw)attP8 (RRID:BDSC_32211), and pJFRC21-10XUAS-IVS-mCD8::RFP in attP18. The VAChT-FRT-STOP-FRT-HA transgene (TI{TI}VAChT[FRT-STOP-FRT.HA] (RRID:BDSC_76021) described in (Pankova and Borst, 2017) was used to examine VAChT expression in photoreceptor neurons. Flp-recombinase, either sens-FLP (expressed in R8 cells; Chen et al., 2014) (fly stock w[*] P{y[+t7.7] w[+mC]=sens-FLPG5.C}attP18; wg[Sp-1]/CyO; sens[Ly-1]/TM6B, Tb[1] (RRID:BDSC_55768) or ey3.5FLP (expressed in all R-cells; Bazigou et al., 2007) (fly stock P{w[+mC]=ey3.5-FLP.B}1, y[1] w[*]; CyO/In(2LR)Gla, wg[Gla-1] PPO1[Bc] (RRID:BDSC_35542) was used to induce VAChT stop-cassette excision.

Histology

Visualization of split-GAL4 driver line expression patterns with pJFRC51-3XUAS-IVS-Syt::smHA in su(Hw)attP1 and pJFRC225-5XUAS-IVS-myr::smFLAG in VK00005 (Nern et al., 2015) or, in a few cases, 20XUAS-CsChrimson-mVenus in attP18 (Klapoetke et al., 2014) as reporters was performed as described (Aso et al., 2014; Wu et al., 2016). Detailed protocols are also available online (`https://www.janelia.org/project-team/flylight/protocols` under “IHC - Anti-GFP”, “IHC - Polarity Sequential” and “DPX mounting”). Multicolor Flp-out (MCFO) markers were detected by immunolabeling with antibodies against HA, FLAG and V5 epitopes as described (Nern et al., 2015). Detailed protocols are also available online (`https://www.janelia.org/project-team/flylight/protocols` under “IHC - MCFO”).

For other experiments, brains of female flies were dissected in insect cell culture medium (Schneider’s Insect Medium, Sigma Aldrich, #S0146) and fixed with 2% PFA (w/v) (prepared from a 20% stock solution, Electron Microscopy Sciences: 15713) also in cell culture medium for 1 h at room temperature. Brains were washed with 0.5 % (v/v) TX-100 (Sigma Aldrich: X100) in PBS and incubated in PBT-NGS (5% Goat Serum [ThermoFisher: 16210-064] in PBT) for at least 30 min. Incubations with primary antibodies and subsequently, after additional PBT washes, secondary antibodies, were in PBT-NGS at 4°C overnight. After additional washes with PBT and then PBS, brains were mounted in SlowFadeGold (ThermoFisher: S36937) and imaged on a Zeiss LSM 710 confocal microscope using 20x 0.8 NA, 40x NA 1.3 or 63x 1.4 NA objectives. A few specimens were mounted in DPX following the protocol described in Nern et al., 2015. For experiments using only native fluorescence, brains were fixed as above and mounted and imaged after the initial post-fixation washes.

Primary antibodies used in each experiment are indicated in Table S5. Primary antibodies were anti-GFP rabbit polyclonal (ThermoFisher: A-11122, RRID:AB_221569; used at 1:1000 dilution), anti-GFP mouse monoclonal 3E6 (ThermoFisher: A-11120, RRID:AB_221568; dilution 1:100), anti-dsRed rabbit polyclonal (Clontech Laboratories, Inc.: 632496, RRID:AB_10013483; dilution 1:1000), anti-HA rabbit monoclonal C29F4 (Cell Signaling Technologies: 3724S, RRID:AB_1549585; dilution 1:300), anti-FLAG rat monoclonal (DYKDDDDK Epitope Tag Antibody [L5], Novus Biologicals: NBP1-06712, RRID:AB_1625981; 1:200), DyLight 549 or DyLight 550 conjugated anti-V5 mouse monoclonals (AbD Serotec: MCA1360D549GA or MCA1360D550GA, RRID:AB_10850329 or RRID:AB_2687576; 1:500 dilution), anti-cockroach allatostatin (Ast7) mouse monoclonal 5F10 (Stay et al., 1992) (also detects *Drosophila* AstA Hergarden et al., 2012; Developmental Studies Hybridoma Bank (DSHB): RRID:AB_528076; dilution 1:5), anti-CadN rat monoclonal DN-Ex #8 (DSHB: RRID:AB_528121; dilution 1:20) (Iwai et al., 1997) and anti-Brp mouse monoclonal nc82 (Wagh et al., 2006) (DSHB: RRID:AB_2314866; dilution1:30).

Secondary antibodies (all from Jackson ImmunoResearch Laboratories, Inc) were DyLight 488-AffiniPure Donkey Anti-Mouse IgG (H+L): 715-485-151, 1:500 dilution; DyLight 594 AffiniPure Donkey anti Rabbit IgG (H+L): 711-515-152, 1:300 dilution; Alexa Fluor 647 AffiniPure Donkey Anti-Rat IgG (H+L): 712-605-153, 1:300 dilution; Alexa Fluor 594 AffiniPure Donkey Anti-Mouse IgG (H+L): 715-585-151,1:300 dilution; Alexa Fluor 647 AffiniPure Donkey Anti-Mouse IgG (H+L): 715-605-151, 1:300 dilution and Alexa Fluor 488 AffiniPure Donkey Anti-Rabbit IgG (H+L): 711-545-152, 1:1000 dilution.

Image processing

Image analyses and processing were mainly done using Fiji (`http://fiji.sc`) and Vaa3D (Peng et al., 2010). Brightness and contrast were adjusted separately for individual images and channels. Figure panels were as-

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

sembled using Adobe Indesign. This included selection of fields of view and adjustments of image size. Some images were rotated or mirrored. In some panels with rotated images, empty space outside the original image was filled in with zero pixels. Most of the images in Figure S1C,C' and S2 show resampled views that were generated from three dimensional image stacks using the Neuronannotator mode of Vaa3D and exported as TIFF format screenshots.

INTACT purification of nuclei

Frozen adult flies were decapitated by vigorous vortexing. Heads or wings/appendages were then collected on cooled metal sieves (H&C Sieving Systems: 1296, 1297, 1298, 1301). Both flies and purified frozen material can be stored indefinitely at -80°C. In a typical experiment 100-500 frozen heads were added to 5ml of 20mM β-glycerophosphate pH7, 200mM NaCl, 2mM EDTA, 0.5% NP40, 0.5mM spermidine, 0.15mM spermine, 1mM DTT, 1X complete protease inhibitor (Sigma: 5056489001), 3mg/ml BSA (ThermoFisher: AM2618), 1mg/ml torula yeast RNA (ThermoFisher: AM7118), 0.6mg/ml carboxyl coated Dynabeads (ThermoFisher: 14306D) and 2μg anti-GFP antibody (ThermoFisher: G10362, RRID:AB_2536526). Homogenization was carried out on ice by 50 tractions in a Dounce homogenizer using the tight pestle followed by filtration over a 10μm cup filter (Partec: 0400422314). Released chromatin and broken nuclei were adsorbed to carboxyl coated magnetic beads for 30 minutes at 4°C with constant rotation. Beads were removed on a magnetic stand and the supernatant was diluted to 50ml with 20mM β-glycerophosphate pH7, 200mM NaCl, 2mM EDTA, 0.5% NP40, 0.5mM spermidine, 0.15mM spermine, 1mM DTT and 1X complete protease inhibitor (Sigma: 5056489001), filtered over a 1μm cup filter (Pluriselect: 435000103) and split into two equal volumes. A 40% Optiprep (Sigma: D1556), 20mM β-glycerophosphate pH7, 2mM EDTA and 0.5% NP40 solution was then gently placed under each aliquot, followed by a lower layer of 50% Optiprep, 20mM β-glycerophosphate pH7, 2mM EDTA and 0.5% NP40. Nuclei were then pelleted on to the 50% layer for 30 minutes at 2300Xg. Purified nuclei were passed over a 10μm cup filter, diluted to 10ml with 20mM β-glycerophosphate pH7, 200mM NaCl, 2mM EDTA, 0.5% NP40, 0.5mM spermidine, 0.15mM spermine, 1mM DTT and 1X complete protease inhibitor and incubated with 30μl of protein G Dynabeads (ThermoFisher: 10004D) for 40 minutes on ice with occasional agitation. Bead-bound nuclei were recovered on a magnet stand followed by a 20 minute incubation on ice in 9mls of 20mM β-glycerophosphate pH7, 300mM NaCl, 1M urea, 0.5% NP40, 2mM EDTA, 0.5mM spermidine, 0.15mM spermine, 1mM DTT, 1X complete protease inhibitor, 0.075mg/ml torula RNA and 0.05U/ml Superasin (ThermoFisher: AM2696). Nuclei were then recovered on a magnet stand, resuspended in 1ml of the previous buffer, passed over a 10μm cup filter, a 5μl aliquot was withdrawn for quantitation and the remainder of the sample solubilized in Arcturus Picopure RNA extraction buffer (ThermoFisher: KIT0204).

TAPIN purification of nuclei

100-3000 frozen heads were added to 5ml of sodium acetate pH8.5, 2.5mM MgCl2, 250mM sucrose, 0.5% NP-40, 0.6mM spermidine, 0.2mM spermine, 1mM DTT, 1X complete protease inhibitor, 0.5mg/ml torula RNA, 0.6mg/ml carboxyl coated Dynabeads and 2μg anti-GFP antibody. Homogenization was carried out on ice by 50 tractions in a Dounce homogenizer using the tight pestle followed by filtration over either a 10 or 20μm cup filter (Partec: 0400422314 or 040042315). Released chromatin and broken nuclei were adsorbed to carboxyl coated magnetic beads for 30 minutes at 4°C with constant rotation. Unbound antibody was removed by incubating the sample on ice for 20 minutes with 100μl of washed UNOsphere SUPra resin (Biorad: 1560218). After the resin was removed on a 10μm cup filter and the carboxyl beads on a magnet stand, the nuclei-containing supernatant was mixed with an equal volume of 500mM sodium acetate pH8.5, 250mM sucrose, 6mM EGTA, 6mM EDTA, 0.6mM spermidine, 0.2mM spermine, 1mM DTT, 1X complete protease inhibitor, 0.25mg/ml torula yeast RNA and 30μl Protein A Dynabeads (ThermoFisher: 10002D). A 2 hour incubation on ice with occasional agitation was used to recover tagged nuclei. Bead-bound nuclei were then recovered on a magnet stand and washed twice with 250mM sodium acetate ph8.5, 250mM sucrose and 0.1% NP40. Nuclei were then released at 37°C for 1 hour by incubation in 50μl of 10mM Tris pH7.5, 2.5mM MgCl2, 0.5mM CaCl2, 250mM sucrose, 0.1% NP40, 1mg/ml torula RNA, 40 units RNAsin (Promega: N2515), 2 units DNAseI (NEB: M0303L), 320 units IdeZ protease (NEB: P0770S). The sample was diluted to 100μl with 10mM Tris pH7.5, 2.5mM MgCl2, 0.5mM CaCl2, 250mM sucrose and 0.1% NP40, EGTA was added to 1mM and the suspension was rapidly triturated 100 times. After returning the sample to a magnet stand, 90μls of buffer containing released nuclei was removed and added to 1.5μl of Protein G Dynabeads that were previously resuspended in 10μl of 10mM Tris pH7.5, 2.5mM MgCl2, 0.5mM CaCl2, 250mM sucrose and 0.1% NP40. The second binding reaction was run for 1-3 hours on ice with occasional agitation, followed by two 250μl washes in 10mM Tris pH7.5, 2.5mM MgCl2, 0.5mM CaCl2, 250mM sucrose and 0.1% NP40. Prior to the last wash a 5μl aliquot was removed for quantitation and the remainder of the sample was solubilized in Arcturus Picopure RNA extraction buffer.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

RNA-seq library construction

Nuclear RNA was DNAseI (Qiagen: 79254) treated and purified using the Arcturus PicoPure (ThermoFisher: KIT0204) system as instructed by the supplier. Purified RNA was mixed with a 1:100,000 dilution of ERCC standard RNA mix #1 (ThermoFisher: 4456740) and amplified using the Nugen Ovation v2 system (Nugen: 7102-32). cDNA was then blunted, ligated to barcoded linkers (Nugen: 0319-32, 0320-32) and sequenced on an Illumina Hiseq 2500 to 50bp read length using Rapid Run flow cells.

In total we built 266 RNA-seq libraries, including 46 INTACT-seq, 196 TAPIN-seq, 8 total RNA libraries from dissected tissues, and 16 control libraries that we used to characterize each INTACT/TAPIN-seq step (Figure 2C, Table S1).

RNA-seq data processing

We trimmed five nucleotides from the 5' end of reads using seqtk (`https://github.com/lh3/seqtk`) to remove potential contaminating adapter sequence from the NuGen Ovation kit. We estimated the abundance of annotated genes using kallisto (v0.43.1; Bray et al., 2016) to pseudo-align trimmed reads to the fly transcriptome (cDNA and ncRNA transcript sequences from ENSEMBL release 91, based on FlyBase release 2017_04), ERCC spike-ins, and the INTACT construct sequences GAL4-DBD, p65-AD, and UNC84_2XGFP. ERCC, INTACT tag constructs, and rRNA genes were removed from the abundance tables and the estimated abundances of the remaining genes were renormalized to one million total transcripts. The ERCC spike-ins and nuclear yield values allowed us to convert relative transcript abundance (in Transcripts Per Million, TPM) to absolute abundance (Figure S3C). However, we only used relative abundance for our analyses. We also aligned the trimmed reads to the genome using STAR (v2.5.3c; Dobin et al., 2013) and evaluated gene body coverage bias using Picard (v 1.9.1; `http://broadinstitute.github.io/picard`).

We used three criteria to quantify the quality of each library: the number of genes detected, the pearson correlation between transcript abundances measured in replicates, and the cDNA yield. We used only high-quality libraries (at least 8,500 genes detected, 3μg cDNA yield, and 0.85 Pearson's correlation of transcript abundances in two biological replicates) as input to the model described below.

Comparison to published single cell and FACS-seq datasets

We obtained genes reported to mark the single cell clusters in a recent scRNA-seq study of the optic lobe (Konstantinides et al., 2018). We also obtained the cluster assignments for each single cell in this dataset from the SCope database (Davie et al., 2018), using Seurat clustering resolution 4.0, as reported by the authors. We analyzed FACS-sorted RNA-seq samples reported by Konstantinides et al. by downloading the raw sequencing reads from the NCBI Sequence Read Archive (`https://www.ncbi.nlm.nih.gov/sra`) and estimating transcript abundance using kallisto and the same transcriptome index as above.

Inferring expression state from transcript abundance

We begin with a catalog of S RNA-seq samples generated from nuclei isolated from cell type cell(s) and the estimated abundance (in TPM), E_{gs} of transcripts from gene g in each sample s. We consider only protein-coding genes with at least 10 TPM abundance in at least one sample (n=12,377 of 13,931 total coding genes).

To interpret E_{gs}, we assume that all genes express in either an 'on' or an 'off' state. Our goal is to infer from these abundances the probability that each gene is expressed in each cell type, $P(z_{gc} = \text{on})$. Depending on the cell types in our catalog, we will observe some genes in both on and off states (bimodal), while others are exclusively off (unimodal-off) or on (unimodal-on). We deal with these scenarios in turn below.

Assuming that a gene is bimodal, we model its expression as arising from a mixture of two gene-specific log-normal distributions describing expression in cells where the gene is off, $P(E_g|z = \text{off})$, and those where the gene is on, $P(E_g|z = \text{on})$, combined with a mixing weight, π_g. We use the same standard deviation for both on and off distributions to ensure a monotonic relationship between transcript abundance and the posterior probability of the on state. If we use different standard deviations for each component distribution, the wider one would become more probable than the narrower one at both low and high expression levels.

$$\log E_g|z \sim \mathcal{N}(\mu_{gz}, \sigma_g)$$

We estimate the posterior probability of the on state (assuming bimodal expression):

$$P(z_{gs} = \text{on}|\text{bimodal}) = \frac{\pi_g p(E_{gs}|z = \text{on})}{\pi_g p(E_{gs}|z = \text{on}) + (1 - \pi_g)p(E_{gs}|z = \text{off})}$$

We treated each replicate sample of the same driver as an independent probe of the same underlying driver-line expression state. To combine replicates of the same driver we sum over their likelihoods:

$$P(z_{gd} = \text{on}|\text{bimodal}) = \frac{\pi_g \prod_s p(E_{gs}|z = \text{on})}{\pi_g \prod_s p(E_{gs}|z = \text{on}) + (1 - \pi_g) \prod_s p(E_{gs}|z = \text{off})}$$

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Similarly, to combine samples from the same cell type we sum over their likelihoods:

$$P(z_{gc} = \text{on}|\text{bimodal}) = \frac{\pi_g \prod_s p(E_{gs}|z = \text{on})}{\pi_g \prod_s p(E_{gs}|z = \text{on}) + (1 - \pi_g) \prod_s p(E_{gs}|z = \text{off})}$$

We estimated parameters for each gene-specific mixture model by maximizing the likelihood for the observed sample-level data:

$$\mathcal{L} = \prod_g \prod_s (\pi_g P(E_{gs}|z = \text{on}, \mu_{gz}, \sigma_g) + (1 - \pi_g) P(E_{gs}|z = \text{off}, \mu_{gz}, \sigma_g))$$

Because we assume independence of genes, we separately optimized the model parameters for each gene. To model the possibility that a gene is unimodally expressed across the cell types we analyzed, we also model the data using a single log-normal distribution, estimating the distribution parameters μ and σ and estimating the data likelihood as:

$$\mathcal{L} = \prod_g \prod_s P(E_{gs}|\mu_g, \sigma_g)$$

Deciding whether a gene is bimodally or unimodally expressed is an example of the model selection problem in statistics. To compare the quality of the unimodal and bimodal models for each gene, we used a recently developed approach to leave-one-out cross-validation that uses Pareto-smoothed importance sampling (PSIS-LOO; Vehtari et al., 2016). Specifically, we performed 10-fold cross validation, by randomly holding out 1/10 of the samples as a "test" set (requiring that at least one replicate of each driver exist in the remaining "training" set), fitting the models using only the training data, and then evaluating the likelihood of the test data using the fitted parameters. Each of the ten cross-validation fits, i, returns an ensemble of S=500 draws from the posterior distribution of the model parameters. We estimated the expected log pointwise predictive density (elpd) of each cross-validation fit by evaluating the likelihood of each held-out dataset i using each parameter draw s :

$$\widehat{elpd}_i = log(\frac{1}{draws} \sum_s^{draws} p(y_i|\theta_{s,k}))$$

We then combined the pointwise log-likelihoods for each cross-validation fit to calculate a single estimate for each model:

$$\widehat{elpd} = \sum \widehat{elpd}_i$$

To compare the unimodal (u) and bimodal (b) models, we calculated the difference in elpd as well as its standard error:

$$\Delta\widehat{elpd} = \widehat{elpd}^{\text{b}} - \widehat{elpd}^{\text{u}}$$

$$\text{se}(\Delta\widehat{elpd}) = \sqrt{nV_i^n(\widehat{elpd}_i^{\text{b}} - \widehat{elpd}_i^{\text{u}})}$$

We then picked the model with the higher elpd, unless the difference in elpd was within two multiples of its standard error ($abs(\Delta\widehat{elpd}) \leq 2 \cdot \text{se}(\Delta\widehat{elpd})$), corresponding approximately to the half-width of a 95% confidence interval in a normal sampling distribution) in which case we considered the two models' performance to be indistinguishable and chose the simpler unimodal model.

If we decide a gene is unimodal, we must still decide if it is expressed or not. To model the expression state of unimodal genes, we created two separate log-normal distributions of abundances of confidently bimodal genes ($\Delta\widehat{elpd} > 10$) using samples where they were either estimated to be on according to the bimodal model ($p(z_{gs} = \text{on}|\text{bimodal}) > 0.9$) and where they were estimated to be off ($p(z_{gs} = \text{on}|\text{bimodal}) < 0.1$), combined with a mixing weight, π, set to the fraction of datapoints that were estimated to be 'on' according to the bimodal model.

$$\log E_g|z \sim \mathcal{N}(\mu_z, \sigma_z)$$

We estimate the posterior probability of the on state assuming unimodal expression as:

$$P(z_g = \text{on}|\text{unimodal}, \mu_g) = \frac{\pi p(\mu_g|z = \text{on})}{\pi p(\mu_g|z = \text{on}) + (1 - \pi) p(\mu_g|z = \text{off})}$$

To build the final matrix of $P(z_{gs} = \text{on})$ calls, we used bimodal estimates for genes where the bimodal model was a better fit than the unimodal model, and the unimodal estimates for the remaining genes.

$$P(z_{gs} = \text{on}) = \begin{cases} P(z_{gs} = \text{on}|\text{bimodal}), & \text{if } \widehat{elpd}^{\text{b}}(g) > \widehat{elpd}^{\text{u}}(g) \\ & \text{and } \Delta\widehat{elpd} > 2 \cdot \text{se}(\Delta\widehat{elpd}), \\ P(z_g = \text{on}|\text{unimodal}), & \text{otherwise} \end{cases}$$

We did not include the transcriptomes of the dissected samples in the mixture models because we were concerned that their cellular heterogeneity would violate our assumption of binary gene expression in each sample. That is, genes expressed in a subset of the cells of a dissected sample would give rise to transcript abundance intermediate between the off and on states, and thus make it more difficult to accurately infer the component distributions. However, in some cases the dissected samples could be useful for interpreting transcript levels in the cells that we profiled, by providing examples that extend the observed dynamic range. For example, in the

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

case of a gene expressed in a dissected tissue, but not in the cells that we specifically profiled, the dissected levels would add "on" examples that would make it easier to interpret the levels in the cell types as "off". To use the dissected samples to better model dynamic range, we added two "dummy" samples to each model: the minimum and maximum observed level across both the cell catalog and the dissected samples. This choice allowed us to use the dissected levels if they in fact outflanked the cell type-specific levels, while not confusing the model with intermediate abundance levels. Once the models were fit, we could use the inferred parameters to estimate expression probabilities for samples that were not used in the model fit. For example, we estimated the probabilities of expression in the dissected samples to search for genes expressed exclusively in the dissected samples and not in the anatomically defined cell type libraries, indicating potential markers for cells that we did not specifically profile.

We implemented all models using RStan (Stan Development Team, 2016) to infer the posterior distribution of unknown parameters using hamiltonian Markov chain Monte Carlo. We used the same weak prior (N(7,5)) for the mean log-expression levels of both on and off components, allowing us to use Stan's positive_ordered data type to describe the location of the two components.

Evaluating model accuracy

To evaluate the accuracy of the mixture modeling approach we created a benchmark set of expression data extracted from FlyBase. Specifically, we queried the FlyBase website (`http://flybase.org`) for genes expressed in the optic lobe or the photoreceptor. The resulting benchmark set included 193 positive and 4 negative expression datapoints. We quantified the model's accuracy on this benchmark in two ways. First, we quantified concordance between the benchmark expression state and our model's inferred state. Second, we computed the cumulative distribution function of the inferred probabilities of expression for the positive benchmark datapoints.

Expression-based tree of cell types

To study cell relationships, we used phylogenetic tree-building to compare their expression profiles. We first selected a subset of genes with on-component means of at least exp(3) ~21 TPM and difference between on and off components of at least exp(1.5) ~4.5 fold. We then encoded the expression profile of each cell as a "sequence" of expression states, where each position represents a gene, and the character indicates the gene is expressed ('A', $P(z_{gc} = \text{on}) > 0.8$), not expressed ('C', $P(z_{gc} = \text{on}) < 0.2$), or its expression is uncertain ('N'; $0.2 < P(z_{gc} = \text{on}) < 0.8$). We computed the Hamming distance between pairs of expression 'sequences' considering only unambiguous positions, using the dist.dna() routine in the ape R package (Paradis et al., 2004). We then used the minimum evolution approach to estimate the 'phylogeny' of the cells, using the balanced weighting scheme (Desper and Gascuel, 2002), as implemented in the ape fastme.bal() routine. We then built trees from 1000 bootstrapped replicates and quantified the support for each branch on the original tree. We visualized the tree using the phytools R package (Revell, 2012).

Identifying marker genes

We identified marker genes specifically enriched in individual cell types and groups of cells (photoreceptor, glia, muscle, neuron) by searching for genes inferred to be almost exclusively expressed in a single cell type or cell group ($p(on) \geq 0.9$ for all cells within a group, and at most two cells outside a group) and with transcript abundance higher than all cells outside the group.

Evaluating expression patterns for genes with different functions

We used FlyBase Gene Groups (release 2018_02) to assign functions to genes, and considered the most terminal groups in the hierarchy that had at least 10 genes.

Mapping receptor expression onto synapses

To map receptor expression onto synaptic connectivity, we first obtained synapse pairs from Takemura et al., to identify synaptic targets of R8 (cell #111), R7 (cell #205), and C2 (cell #214) cells in the medulla (Takemura et al., 2013). When multiple instances of a cell type were available in the synaptic table, we chose the one with the greatest number of synaptic partners. For target cell types that we profiled with TAPIN/INTACT-seq, we discretized their expression as either on ($p(on) \geq 0.8$) or off ($p(on) < 0.8$). For cell types that we did not profile, we classified them as unknown receptor expression.

Data and software availability

All raw and processed transcriptome data is available from NCBI GEO (accession GSE116969). The shell scripts used to process the raw RNA-seq data, and the R and Stan programs that implement the mixture model as well as generate all figures and tables in this paper are available at github (`http://github.com/fredpdavis/opticlobe`).

Acknowledgements

This project was supported by the Howard Hughes Medical Institute. FPD was supported in part by the Intramu-

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

ral Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health. We would like to thank Teresa Tian and all members of the Janelia Fly Facility for their help in maintaining and rearing flies, Tanya Wolff and Arnim Jenett for the central complex and Heather Dionne for the l-LNv split-GAL4 lines, and the FlyLight Project Team for imaging of split-GAL4 expression patterns.

References

Amrein, H. and Axel, R. (1997). Genes expressed in neurons of adult male Drosophila. Cell *88*, 459–469.

Apitz, H. and Salecker, I. (2014). A challenge of numbers and diversity: neurogenesis in the Drosophila optic lobe. Journal of Neurogenetics *28*, 233–249.

– (2018). Spatio-temporal relays control layer identity of direction-selective neuron subtypes in Drosophila. Nature Communications *9*, 2295.

Arenz, A., Drews, M.S., Richter, F.G., Ammer, G., and Borst, A. (2017). The Temporal Tuning of the Drosophila Motion Detectors Is Determined by the Dynamics of Their Input Elements. Current biology: CB *27*, 929–944.

Aso, Y., Hattori, D., Yu, Y., Johnston, R.M., Iyer, N.A., Ngo, T.-T.B., Dionne, H., Abbott, L.F., Axel, R., Tanimoto, H., et al. (2014). The neuronal architecture of the mushroom body provides a logic for associative learning. eLife *3*, e04577.

Barnstedt, O., Owald, D., Felsenberg, J., Brain, R., Moszynski, J.-P., Talbot, C.B., Perrat, P.N., and Waddell, S. (2016). Memory-Relevant Mushroom Body Output Synapses Are Cholinergic. Neuron *89*, 1237–1247.

Bazigou, E., Apitz, H., Johansson, J., Lorén, C.E., Hirst, E.M.A., Chen, P.-L., Palmer, R.H., and Salecker, I. (2007). Anterograde Jelly belly and Alk receptor tyrosine kinase signaling mediates retinal axon targeting in Drosophila. Cell *128*, 961–975.

Belote, J.M., Handler, A.M., Wolfner, M.F., Livak, K.J., and Baker, B.S. (1985). Sex-specific regulation of yolk protein gene expression in Drosophila. Cell *40*, 339–348.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology *34*, 525–527.

Chen, Y., Akin, O., Nern, A., Tsui, C.Y.K., Pecot, M.Y., and Zipursky, S.L. (2014). Cell-type-specific labeling of synapses in vivo through synaptic tagging with recombination. Neuron *81*, 280–293.

Crocker, A., Guan, X.-J., Murphy, C.T., and Murthy, M. (2016). Cell-Type-Specific Transcriptome Analysis in the Drosophila Mushroom Body Reveals Memory-Related Changes in Gene Expression. Cell Reports *15*, 1580–1596.

Croset, V., Treiber, C.D., and Waddell, S. (2018). Cellular diversity in the Drosophila midbrain revealed by single-cell transcriptomics. eLife *7*.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., Bravo González-Blas, C., et al. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell *in press*.

Deal, R.B. and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Developmental Cell *18*, 1030–1040.

Desper, R. and Gascuel, O. (2002). Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. Journal of Computational Biology *9*, 687–705.

Diao, F., Ironfield, H., Luan, H., Diao, F., Shropshire, W.C., Ewer, J., Marr, E., Potter, C.J., Landgraf, M., and White, B.H. (2015). Plug-and-play genetic access to drosophila cell types using exchangeable exon cassettes. Cell Reports *10*, 1410–1421.

Dionne, H., Hibbard, K.L., Cavallaro, A., Kao, J.-C., and Rubin, G.M. (2018). Genetic Reagents for Making Split-GAL4 Lines in Drosophila. Genetics *209*, 31–35.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dolan, M.-J., Luan, H., Shropshire, W.C., Sutcliffe, B., Cocanougher, B., Scott, R.L., Frechter, S., Zlatic, M., Jefferis, G.S.X.E., and White, B.H. (2017). Facilitating Neuron-Specific Genetic Manipulations in Drosophila melanogaster Using a Split GAL4 Repressor. Genetics *206*, 775–784.

Ecker, J.R., Geschwind, D.H., Kriegstein, A.R., Ngai, J., Osten, P., Polioudakis, D., Regev, A., Sestan, N., Wickersham, I.R., and Zeng, H. (2017). The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. Neuron *96*, 542–557.

Enell, L., Hamasaka, Y., Kolodziejczyk, A., and Nässel, D.R. (2007). γ-Aminobutyric acid (GABA) signaling components in Drosophila: Immunocytochemical localization of GABAB receptors in relation to the GABAA receptor subunit RDL and a vesicular GABA transporter. Journal of Comparative Neurology *505*, 18–31.

Fei, H., Chow, D.M., Chen, A., Romero-Calderón, R., Ong, W.S., Ackerson, L.C., Maidment, N.T., Simpson, J.H., Frye, M.A., and Krantz, D.E. (2010). Mutation of the Drosophila vesicular GABA transporter disrupts visual figure detection. The Journal of Experimental Biology *213*, 1717–1730.

Fischbach, K.F. and Dittrich, A.P.M. (1989). The optic lobe of Drosophila melanogaster. I. A Golgi analysis of wild-type structure. Cell and tissue research *258*, 441–475.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Fortini, M.E. and Rubin, G.M. (1990). Analysis of cis-acting requirements of the Rh3 and Rh4 genes reveals a bipartite organization to rhodopsin promoters in Drosophila melanogaster. Genes & Development *4*, 444–463.

Gao, S., Takemura, S., Ting, C.-Y., Huang, S., Lu, Z., Luan, H., Rister, J., Thum, A.S., Yang, M., Hong, S.-T., et al. (2008). The Neural Substrate of Spectral Preference in Drosophila. Neuron *60*, 328–342.

Gisselmann, G., Plonka, J., Pusch, H., and Hatt, H. (2004). Drosophila melanogaster GRD and LCCH3 subunits form heteromultimeric GABA-gated cation channels. British Journal of Pharmacology *142*, 409–413.

Hardie, R.C. (1987). Is histamine a neurotransmitter in insect photoreceptors? Journal of Comparative Physiology. A, Sensory, Neural, and Behavioral Physiology *161*, 201–213.

Henry, G.L., Davis, F.P., Picard, S., and Eddy, S.R. (2012). Cell type-specific genomics of Drosophila neurons. Nucleic Acids Research *40*, 9691–9704.

Hergarden, A.C., Tayler, T.D., and Anderson, D.J. (2012). Allatostatin-A neurons inhibit feeding behavior in adult Drosophila. Proceedings of the National Academy of Sciences of the United States of America *109*, 3967–3972.

Hobert, O. (2016). Terminal Selectors of Neuronal Identity. Current Topics in Developmental Biology *116*, 455–475.

Hu, W., Wang, T., Wang, X., and Han, J. (2015). Ih Channels Control Feedback Regulation from Amacrine Cells to Photoreceptors. PLOS Biology *13*, e1002115.

Huang, Y., Ng, F.S., and Jackson, F.R. (2015). Comparison of Larval and Adult Drosophila Astrocytes Reveals Stage-Specific Gene Expression Profiles. G3: Genes|Genomes|Genetics *5*, 551–558.

Im, S.H. and Taghert, P.H. (2010). PDF receptor expression reveals direct interactions between circadian oscillators in Drosophila. The Journal of Comparative Neurology *518*, 1925–1945.

Iwai, Y., Usui, T., Hirano, S., Steward, R., Takeichi, M., and Uemura, T. (1997). Axon patterning requires DN-cadherin, a novel neuronal adhesion receptor, in the Drosophila embryonic CNS. Neuron *19*, 77–89.

Järvilehto, M. and Zettler, F. (1971). Localized intracellular potentials from pre- and postsynaptic components in the external plexiform layer of an insect retina. Zeitschrift für vergleichende Physiologie *75*, 422–440.

Jenett, A., Rubin, G.M., Ngo, T.-T.B., Shepherd, D., Murphy, C., Dionne, H., Pfeiffer, B.D., Cavallaro, A., Hall, D., Jeter, J., et al. (2012). A GAL4-driver line resource for Drosophila neurobiology. Cell Reports *2*, 991–1001.

Kebschull, J.M., Garcia da Silva, P., Reid, A.P., Peikon, I.D., Albeanu, D.F., and Zador, A.M. (2016). High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. Neuron *91*, 975–987.

Kiragasi, B., Wondolowski, J., Li, Y., and Dickman, D.K. (2017). A Presynaptic Glutamate Receptor Subunit Confers Robustness to Neurotransmission and Homeostatic Potentiation. Cell reports *19*, 2694–2706.

Klapoetke, N.C., Murata, Y., Kim, S.S., Pulver, S.R., Birdsey-Benson, A., Cho, Y.K., Morimoto, T.K., Chuong, A.S., Carpenter, E.J., Tian, Z., et al. (2014). Independent optical excitation of distinct neural populations. Nature Methods *11*, 338–346.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell *58*, 610–620.

Konstantinides, N., Kapuralin, K., Fadil, C., Barboza, L., Satija, R., and Desplan, C. (2018). Phenotypic Convergence: Distinct Transcription Factors Regulate Common Terminal Features. Cell *174*, 622–635.

Kudron, M.M., Victorsen, A., Gevirtzman, L., Hillier, L.W., Fisher, W.W., Vafeados, D., Kirkey, M., Hammonds, A.S., Gersch, J., Ammouri, H., et al. (2018). The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. Genetics *208*, 937–949.

Laughlin, S.B. and Hardie, R.C. (1978). Common strategies for light adaptation in the peripheral visual systems of fly and dragonfly. Journal of comparative physiology *128*, 319–340.

Li, H., Horns, F., Wu, B., Xie, Q., Li, J., Li, T., Luginbuhl, D.J., Quake, S.R., and Luo, L. (2017). Classifying Drosophila Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing. Cell *171*, 1206–1220.e22.

Long, X., Colonell, J., Wong, A.M., Singer, R.H., and Lionnet, T. (2017). Quantitative mRNA imaging throughout the entire Drosophila brain. Nature Methods *14*, 703–706.

Lowenstein, C.J. and Snyder, S.H. (1992). Nitric oxide, a novel biologic messenger. Cell *70*, 705–707.

Luan, H., Peabody, N.C., Vinson, C.R., and White, B.H. (2006). Refined spatial manipulation of neuronal function by combinatorial restriction of transgene expression. Neuron *52*, 425–436.

Luo, L., Callaway, E.M., and Svoboda, K. (2018). Genetic Dissection of Neural Circuits: A Decade of Progress. Neuron *98*, 256–281.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Mauss, A.S., Vlasits, A., Borst, A., and Feller, M. (2017). Visual Circuits for Direction Selectivity. Annual Review of Neuroscience *40*, 211–230.

Meinertzhagen, I.A. and O'Neil, S.D. (1991). Synaptic organization of columnar elements in the lamina of the wild type in Drosophila melanogaster. The Journal of Comparative Neurology *305*, 232–263.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Meinertzhagen, I.A. and Sorra, K.E. (2001). Synaptic organization in the fly's optic lamina: few cells, many synapses and divergent microcircuits. Progress in Brain Research *131*, 53–69.

Micheva, K.D. and Smith, S.J. (2007). Array tomography. Neuron *55*, 25–36.

Mo, A., Mukamel, E.A., Davis, F.P., Luo, C., Henry, G.L., Picard, S., Urich, M.A., Nery, J.R., Sejnowski, T.J., Lister, R., et al. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron *86*, 1369–84.

Morante, J. and Desplan, C. (2008). The color-vision circuit in the medulla of Drosophila. Current biology: CB *18*, 553–565.

Nagarkar-Jaiswal, S., Lee, P.-T., Campbell, M.E., Chen, K., Anguiano-Zarate, S., Gutierrez, M.C., Busby, T., Lin, W.-W., He, Y., Schulze, K.L., et al. (2015). A library of MiMICs allows tagging of genes and reversible, spatial and temporal knockdown of proteins in Drosophila. eLife *4*.

Nériec, N. and Desplan, C. (2016). From the Eye to the Brain: Development of the Drosophila Visual System. Current Topics in Developmental Biology *116*, 247–271.

Nern, A., Pfeiffer, B.D., and Rubin, G.M. (2015). Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. Proceedings of the National Academy of Sciences *112*, E2967–E2976.

Otsuna, H. and Ito, K. (2006). Systematic analysis of the visual projection neurons of Drosophila melanogaster. I. Lobula-specific pathways. The Journal of Comparative Neurology *497*, 928–958.

Özkan, E., Carrillo, R.A., Eastman, C.L., Weiszmann, R., Waghray, D., Johnson, K.G., Zinn, K., Celniker, S.E., and Garcia, K.C. (2013). An extracellular interactome of immunoglobulin and LRR proteins reveals receptor-ligand networks. Cell *154*, 228–239.

Pankova, K. and Borst, A. (2017). Transgenic line for the identification of cholinergic release sites in Drosophila melanogaster. The Journal of Experimental Biology *220*, 1405–1410.

Pantazis, A., Segaran, A., Liu, C.-H., Nikolaev, A., Rister, J., Thum, A.S., Roeder, T., Semenov, E., Juusola, M., and Hardie, R.C. (2008). Distinct roles for two histamine receptors (hclA and hclB) at the Drosophila photoreceptor synapse. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience *28*, 7250–7259.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

Peng, H., Ruan, Z., Long, F., Simpson, J.H., and Myers, E.W. (2010). V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. Nature Biotechnology *28*, 348–353.

Raghu, S.V. and Borst, A. (2011). Candidate glutamatergic neurons in the visual system of Drosophila. PloS One *6*, e19472.

Raghu, S.V., Claussen, J., and Borst, A. (2013). Neurons with GABAergic phenotype in the visual system of Drosophila. The Journal of Comparative Neurology *521*, 252–265.

Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution *3*, 217–223.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nature Biotechnology *17*, 1030–1032.

Rivera-Alba, M., Vitaladevuni, S.N., Mishchenko, Y., Mischenko, Y., Lu, Z., Takemura, S., Scheffer, L., Meinertzhagen, I.A., Chklovskii, D.B., and Polavieja, G.G.d. (2011). Wiring economy and volume exclusion determine neuronal placement in the Drosophila brain. Curr Biol *21*, 2000–5.

Sarthy, P.V. (1991). Histamine: a neurotransmitter candidate for Drosophila photoreceptors. Journal of Neurochemistry *57*, 1757–1768.

Schnaitmann, C., Haikala, V., Abraham, E., Oberhauser, V., Thestrup, T., Griesbeck, O., and Reiff, D.F. (2018). Color Processing in the Early Visual System of Drosophila. Cell *172*, 318–330.e18.

Siegert, S., Cabuy, E., Scherf, B.G., Kohler, H., Panda, S., Le, Y.-Z., Fehling, H.J., Gaidatzis, D., Stadler, M.B., and Roska, B. (2012). Transcriptional code and disease map for adult retinal cell types. Nature Neuroscience *15*, 487–495.

Silies, M., Gohl, D.M., and Clandinin, T.R. (2014). Motion-detecting circuits in flies: coming into view. Annual Review of Neuroscience *37*, 307–327.

Stay, B., Chan, K.K., and Woodhead, A.P. (1992). Allatostatin-immunoreactive neurons projecting to the corpora allata of adult Diploptera punctata. Cell and Tissue Research *270*, 15–23.

Strausfeld, N.J. (2009). Brain organization and the origin of insects: an assessment. Proceedings. Biological Sciences *276*, 1929–1937.

Strother, J.A., Wu, S.-T., Rogers, E.M., Eliason, J.L.M., Wong, A.M., Nern, A., and Reiser, M.B. (2018). Behavioral state modulates the ON visual motion pathway of Drosophila. Proceedings of the National Academy of Sciences of the United States of America *115*, E102–E111.

Strother, J.A., Wu, S.-T., Wong, A.M., Nern, A., Rogers, E.M., Le, J.Q., Rubin, G.M., and Reiser, M.B. (2017). The Emergence of Directional Selectivity in the Visual Motion Pathway of Drosophila. Neuron *94*, 168–182.e10.

Swanson, L.W. and Lichtman, J.W. (2016). From Cajal to Connectome and Beyond. Annual Review of Neuroscience *39*, 197–216.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Takemura, S., Bharioke, A., Lu, Z., Nern, A., Vitaladevuni, S., Rivlin, P.K., Katz, W.T., Olbris, D.J., Plaza, S.M., Winston, P., et al. (2013). A visual motion detection circuit suggested by Drosophila connectomics. Nature *500*, 175–81.

Takemura, S., Lu, Z., and Meinertzhagen, I.A. (2008). Synaptic circuits of the Drosophila optic lobe: the input terminals to the medulla. The Journal of Comparative Neurology *509*, 493–513.

Takemura, S., Nern, A., Chklovskii, D.B., Scheffer, L.K., Rubin, G.M., and Meinertzhagen, I.A. (2017). The comprehensive connectome of a neural substrate for 'ON' motion detection in Drosophila. eLife *6*.

Takemura, S., Xu, C.S., Lu, Z., Rivlin, P.K., Parag, T., Olbris, D.J., Plaza, S., Zhao, T., Katz, W.T., Umayam, L., et al. (2015). Synaptic circuits and their variations within different columns in the visual system of Drosophila. Proceedings of the National Academy of Sciences *112*, 13711–13716.

Tan, L., Zhang, K.X., Pecot, M.Y., Nagarkar-Jaiswal, S., Lee, P.-T., Takemura, S., McEwen, J.M., Nern, A., Xu, S., Tadros, W., et al. (2015). Ig Superfamily Ligand and Receptor Pairs Expressed in Synaptic Partners in Drosophila. Cell *163*, 1756–1769.

Tirian, L. and Dickson, B. (2017). The VT GAL4, LexA, and split-GAL4 driver line collections for targeted expression in the Drosophila nervous system. bioRxiv, 198648.

Tuthill, J.C., Nern, A., Holtz, S.L., Rubin, G.M., and Reiser, M.B. (2013). Contributions of the 12 neuron classes in the fly lamina to motion vision. Neuron *79*, 128–140.

Tuthill, J.C., Nern, A., Rubin, G.M., and Reiser, M.B. (2014). Wide-field feedback neurons dynamically tune early visual processing. Neuron *82*, 887–895.

Varija Raghu, S., Reiff, D.F., and Borst, A. (2011). Neurons with cholinergic phenotype in the visual system of Drosophila. The Journal of Comparative Neurology *519*, 162–176.

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing. arXiv: 1507.04544.

Wagh, D.A., Rasse, T.M., Asan, E., Hofbauer, A., Schwenkert, I., Dürrbeck, H., Buchner, S., Dabauvalle, M.-C., Schmidt, M., Qin, G., et al. (2006). Bruchpilot, a protein with homology to ELKS/CAST, is required for structural integrity and function of synaptic active zones in Drosophila. Neuron *49*, 833–844.

Wenick, A.S. and Hobert, O. (2004). Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans. Developmental Cell *6*, 757–770.

Wolff, T., Iyer, N.A., and Rubin, G.M. (2015). Neuroarchitecture and neuroanatomy of the Drosophila central complex: A GAL4-based dissection of protocerebral bridge neurons and circuits. The Journal of Comparative Neurology *523*, 997–1037.

Wolff, T. and Rubin, G.M. (2018). Neuroarchitecture of the Drosophila Central Complex: A Catalog of Nodulus and Asymmetrical Body Neurons and a Revision of the Protocerebral Bridge Catalog. The Journal of Comparative Neurology *in press*.

Wu, M., Nern, A., Williamson, W.R., Morimoto, M.M., Reiser, M.B., Card, G.M., and Rubin, G.M. (2016). Visual projection neurons in the Drosophila lobula link feature detection to distinct behavioral programs. eLife *5*.

Yasuyama, K. and Meinertzhagen, I.A. (1999). Extraretinal photoreceptors at the compound eye's posterior margin in Drosophila melanogaster. The Journal of Comparative Neurology *412*, 193–202.

Zhang, H.G., Lee, H.J., Rocheleau, T., ffrench-Constant, R.H., and Jackson, M.B. (1995). Subunit composition determines picrotoxin and bicuculline sensitivity of Drosophila gamma-aminobutyric acid receptors. Molecular Pharmacology *48*, 835–840.

Zheng, L., Nikolaev, A., Wardill, T.J., O'Kane, C.J., Polavieja, G.G.d., and Juusola, M. (2009). Network Adaptation Improves Temporal Representation of Naturalistic Stimuli in Drosophila Eye: I Dynamics. PLOS ONE *4*, e4307.

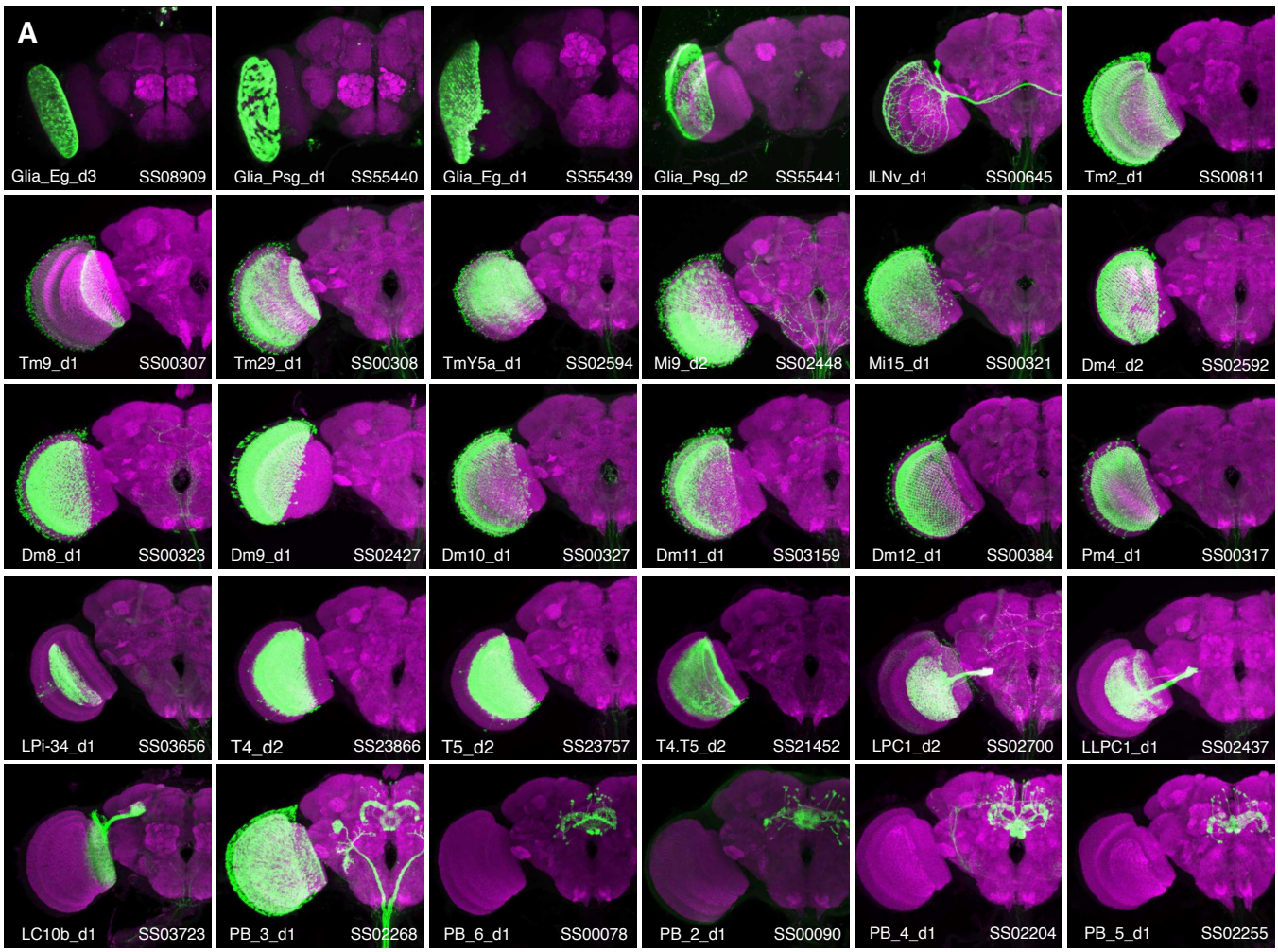
Zheng, Z., Lauritzen, J.S., Perlman, E., Robinson, C.G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C.B., Sharifi, N., et al. (2018). A Complete Electron Microscopy Volume Of The Brain Of Adult Drosophila melanogaster. Cell *174*, 730–743.

Supplemental Tables

S1. Genetic drivers used to create INTACT/TAPIN-seq libraries
S2. RNA-seq samples
S3. Benchmark expression data obtained from FlyBase
S4. Marker genes enriched in photoreceptors, glia, muscle, and pigment cells
S5. Genotype and labeling details for anatomy figures

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE S1



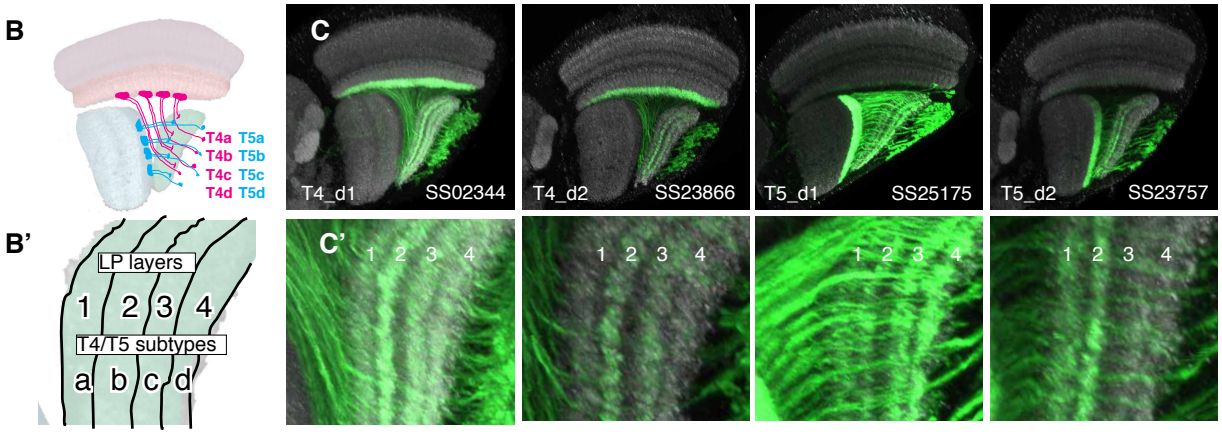


Figure S1: Related to Figure 1. Whole brain expression patterns of new driver lines generated in this study. A. Maximum intensity projection of confocal stacks taken from whole fly brains (only one optic lobe is shown). Expression patterns of the driver lines (myristoylated-GFP) are in green and magenta is a neuropil marker. Imaging parameters and brightness and contrast were individually adjusted for each sample. B, B'. T4 and T5 cells comprise four subtypes (a,b,c,d) each of which project to specific layers of the lobula plate (B'). C, C'. Individual T4 and T5 driver lines label combinations of subtypes but show preferential expression in some subtypes. Subtypes were identified by their projections to specific layers in the lobula plate (C,C'). For example, T5_d2 mainly labels lobula plate layers one and two, indicating expression in T5a and T5b. Each of the lower panels is a higher magnification view of the lobula plate region (C'). In (C) both the driver and the split identifier are indicated in the lower left and right corner respectively. Driver lines SS00078, SS00090 and S02255 are from Wolff and Rubin, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE S2

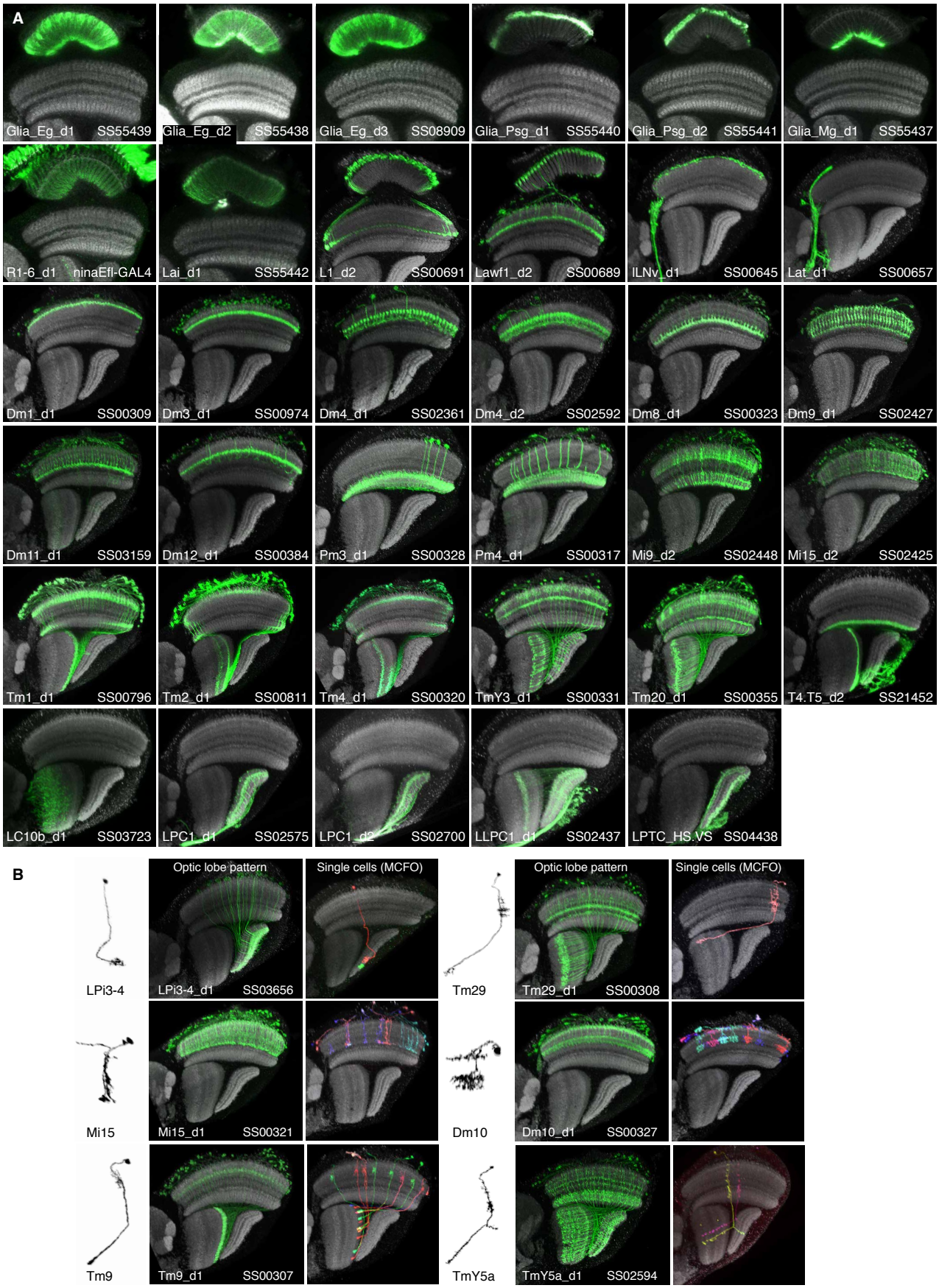
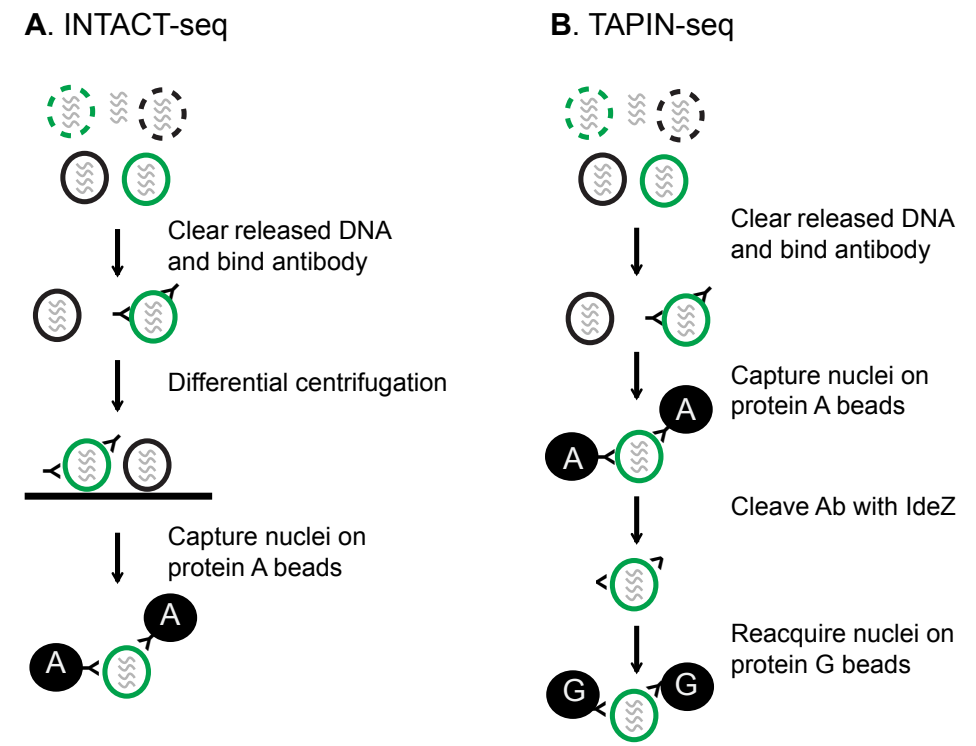
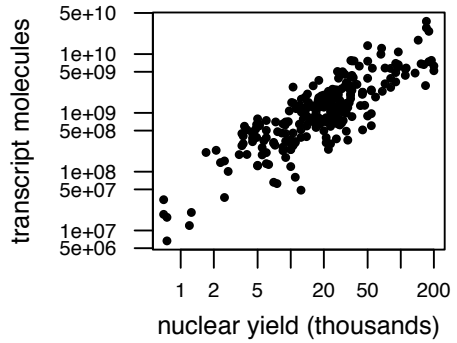


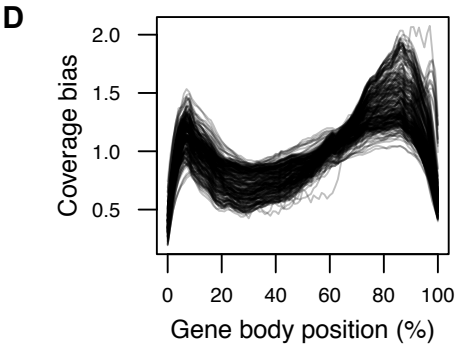
Figure S2: Related to Figure 1. Optic lobe patterns of driver lines. A. Optic lobe expression patterns of new driver lines used in this study. All images orient the mediolateral axis of the brain vertically and are resampled substack projections generated from of high resolution (63x) confocal stacks. B. Examples of segmented single cells illustrating cell morphology (left), the complete optic lobe expression pattern (middle) or individual cells labeled by MultiColor FlpOut (MCFO). In all images the neuropil marker is in gray and both the targeted cell type and driver are indicated in the lower left and right corner respectively.

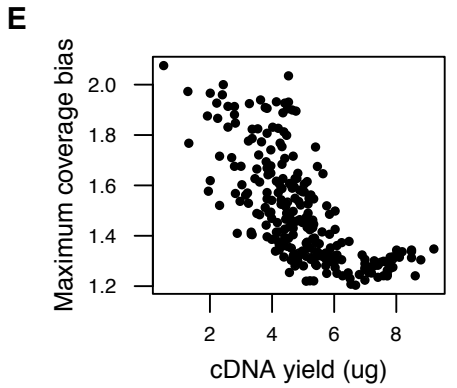
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

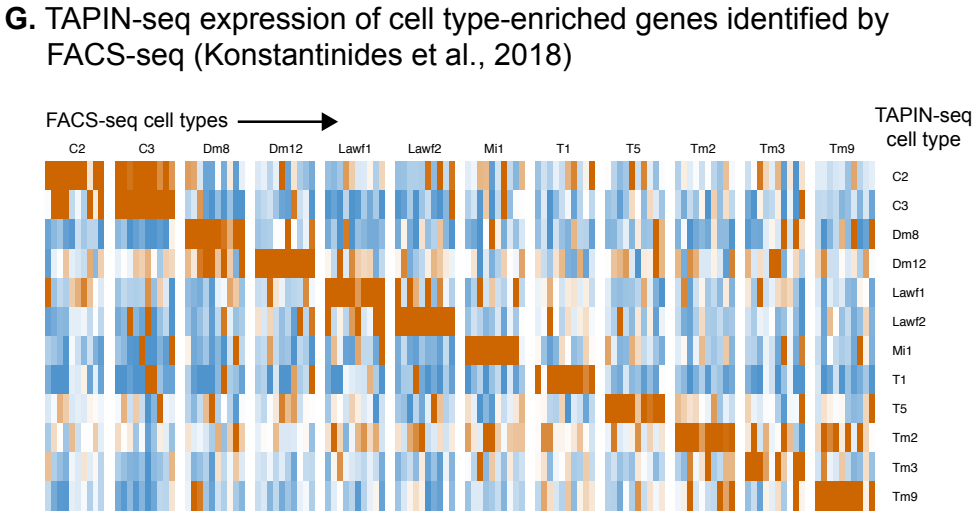
FIGURE S3











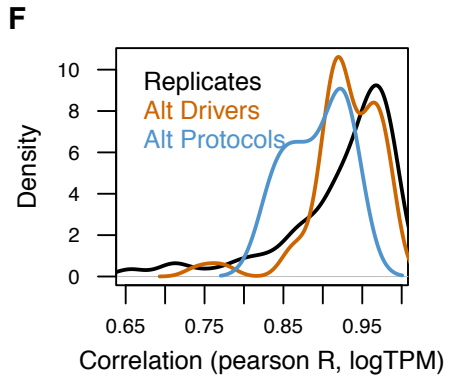
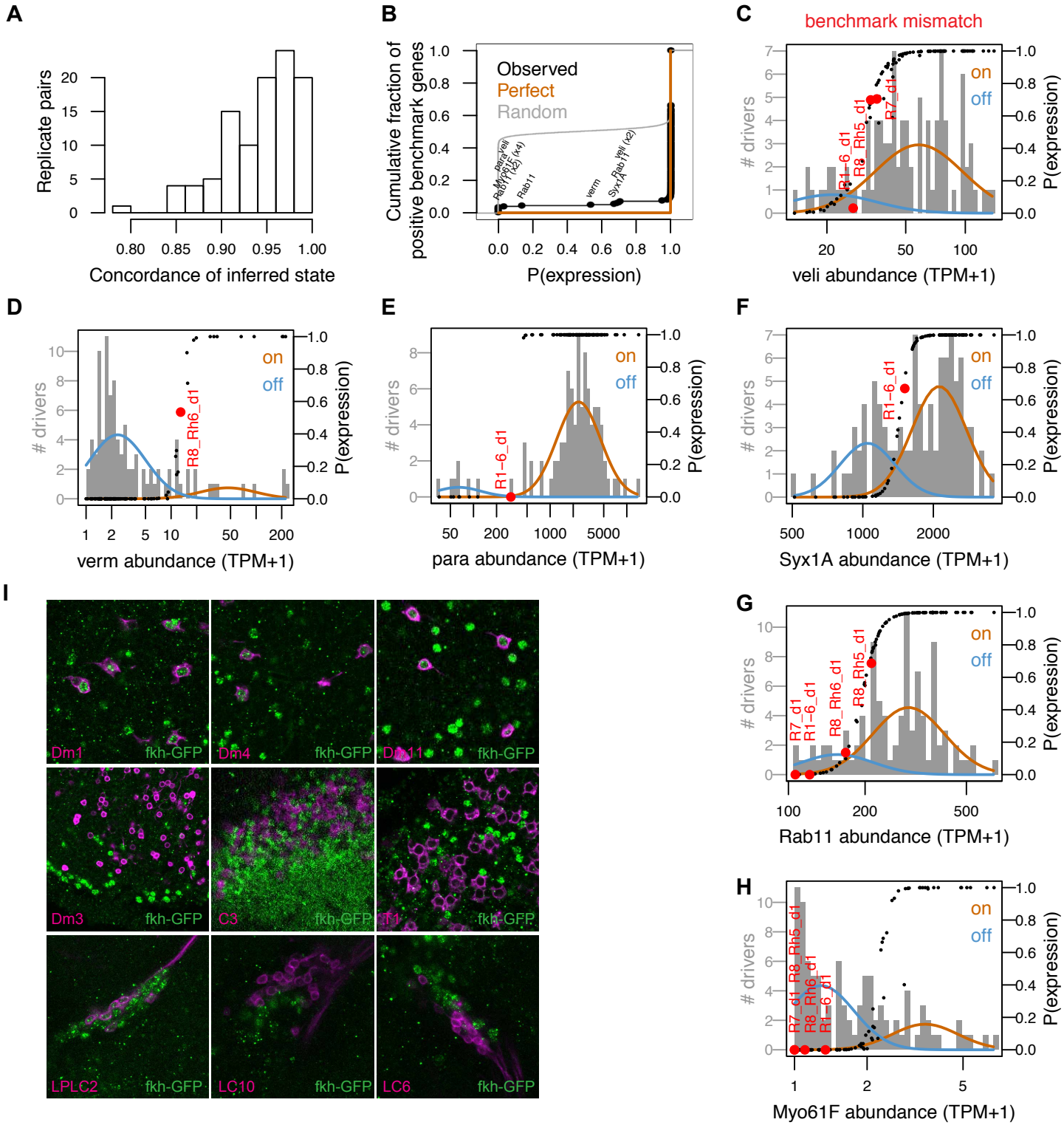


Figure S3: Related to Figure 2. We used two variants of nuclei capture, INTACT and TAPIN. A. We used an INTACT-seq variant, that we originally developed for mouse (Mo et al., 2015), that purifies nuclei by differential centrifugation. B. TAPIN-seq replaces the space- and time-intensive centrifugation with a two-step capture enabled by antibody hinge cleavage with the bacterial protease IdeZ. Both protein A and protein G bind the Fc region, while only protein G is able to bind F(ab')$_2$. C. Libraries built from more nuclei have more transcript molecules (estimated using synthetic spike-ins). D. Nearly all libraries showed relatively unbiased positional coverage across gene bodies. E. The maximum bias in positional coverage observed in each library was inversely correlated with cDNA yield, although with large variance in bias for lower yield libraries. F. Estimated transcript abundances were reproducible as evaluated by Pearson correlation (of log-transformed transcript abundance) between biological replicates (black), alternative drivers for the same cell type (orange), or comparing TAPIN to INTACT profiles (blue). G. TAPIN-seq expression of marker genes identified from an independent FACS-seq dataset covering 12 cell types we also profiled (Konstantinides et al., 2018). We defined marker genes for each cell type as the top-10 most highly expressed genes relative to the mean of all cell types, requiring at at least 4x higher abundance than the mean and a relative abundance of at least 50 TPM.

bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE S4



bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure S4: Related to Figure 3. Overview of INTACT-seq and TAPIN-seq libraries. A. Concordance of inferred expression states between replicates. Concordance was computed as the number of genes predicted to express ($p(on) \geq 0.8$) or not ($p(on) \leq 0.2$) in both replicates divided by the number of genes predicted to express or not in either replicate. B. Cumulative distribution of inferred expression probabilities for gene/cell pairs reported to express in FlyBase (n=193 positive benchmark points). Our mixture model correctly inferred expression of 179 of the 193 gene/cell pairs. The 14 discordant pairs involved six genes (labeled in black). C-H. Modeling results for the six genes with benchmark mismatches. The on and off components are represented as orange and blue curves, respectively. Black points represent the inferred probabilities of expression for all drivers. Red points highlight the drivers where the model results disagreed with the benchmark. The transcript abundance (x-axis) reflects the average of all "high quality" replicates (minimum two per cell type). I, J. To evaluate our modeling results for fkh and Ets65A we evaluated protein expression in several cell types (related to fig. 3J-L) using GFP-fusion proteins (Methods). The indicated cell types (lower left corner) were visualized with a membrane marker (magenta). The second Mi15 panel includes examples of occasional cells without detectable Ets65-GFP expression (arrows).

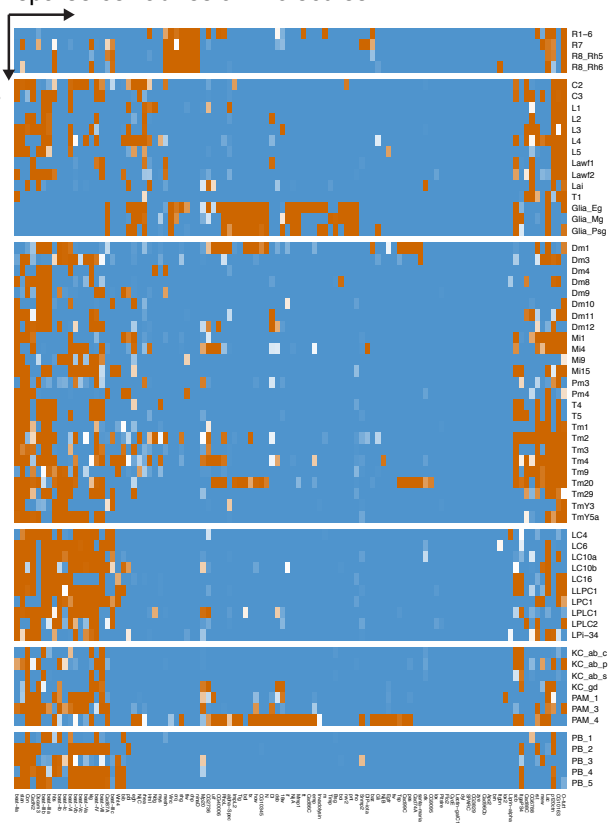
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE S5

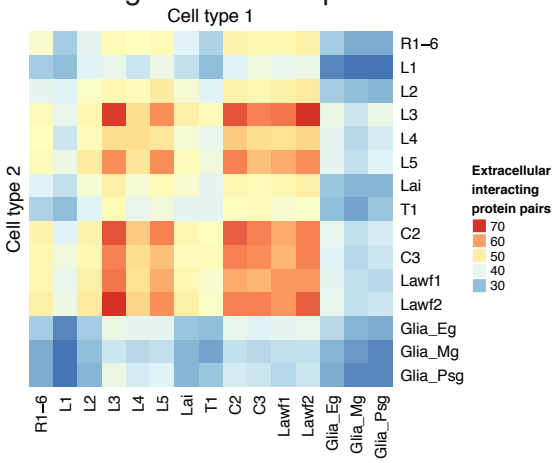
A. Cell adhesion molecules

sparse cell adhesion molecules

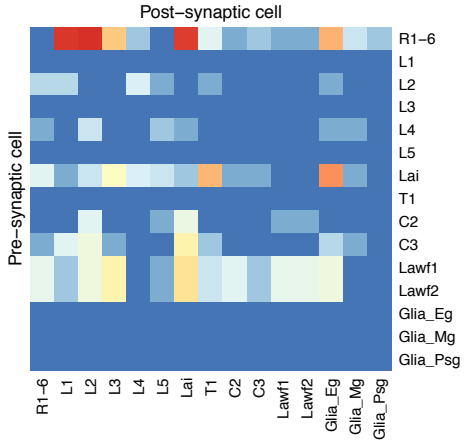
cell types

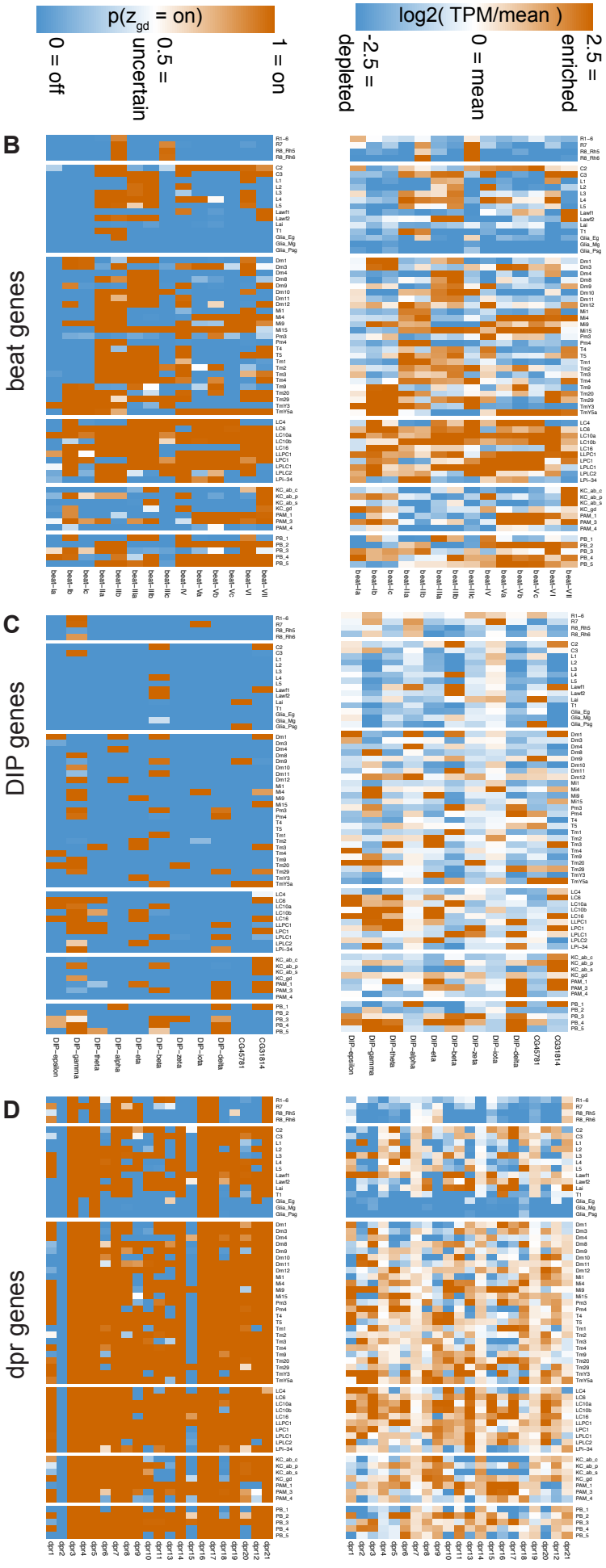


E. Interacting extracellular proteins



F. Laminar connectivity (Rivera-Alba 2011)





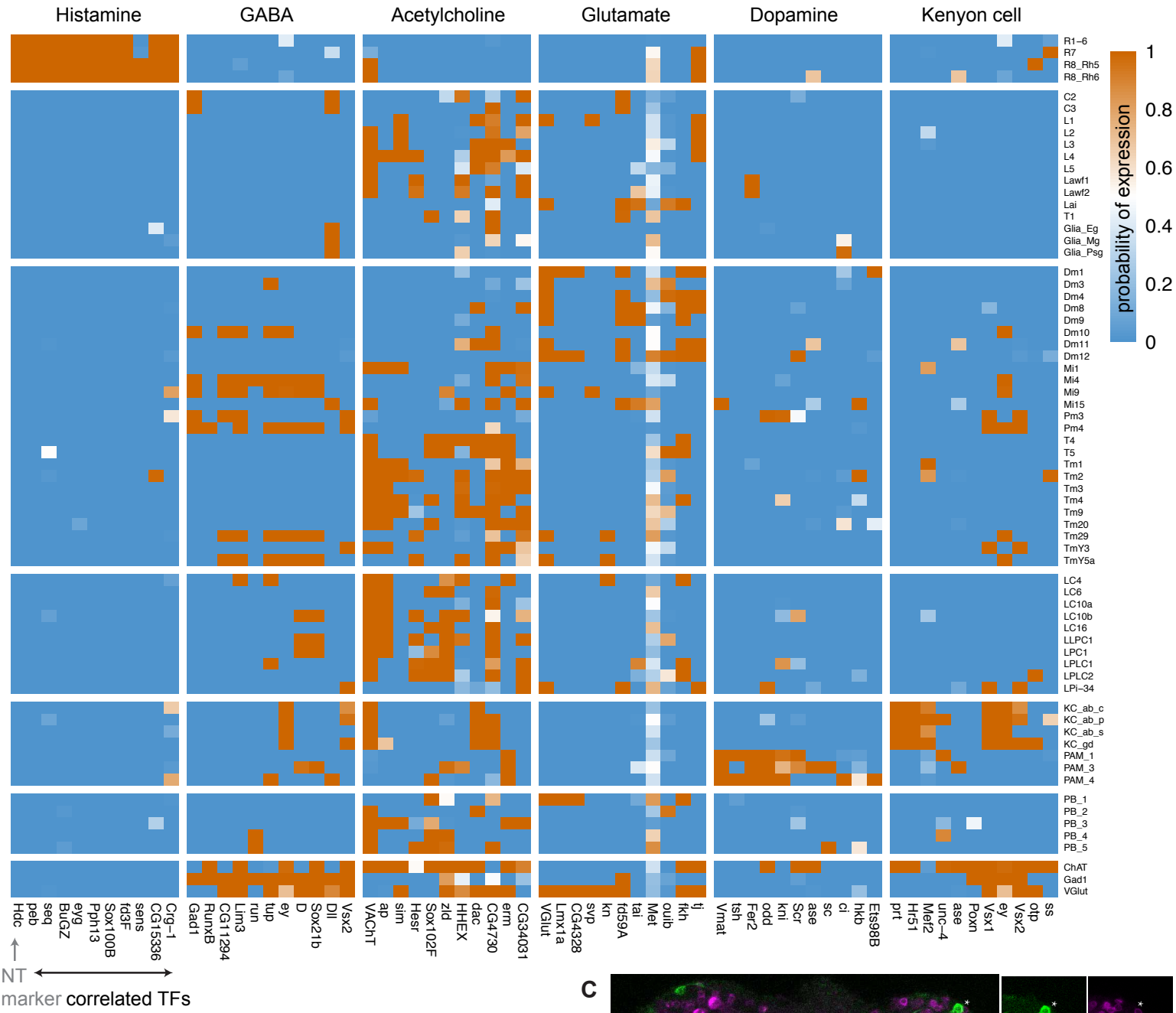
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure S5: Related to Figure 4. TAPIN-seq profiles identify genes enriched in cell types and groups. A. Cell adhesion molecules specifically expressed across our transcriptome catalog. B-D. The expression pattern of all beat, DIP, and Dpr family members depicted as either probabilities of expression (left) or relative transcript abundance (right). E,F. The number of interacting pairs of extracellular protein pairs (Özkan et al., 2013) expressed by pairs of cells in the lamina (E) is not sufficient to predict the synaptic connectivity of these cells (F. data from Rivera-Alba et al., 2011). To match our expression data, we summed the synapse counts for the individual R1-R6 photoreceptors originally reported by Rivera-Alba et al., 2011. For the same reason, we also duplicated the subtype-unidentified Lawf synapse counts as separate Lawf1 and Lawf2 entries in the connectome matrix.

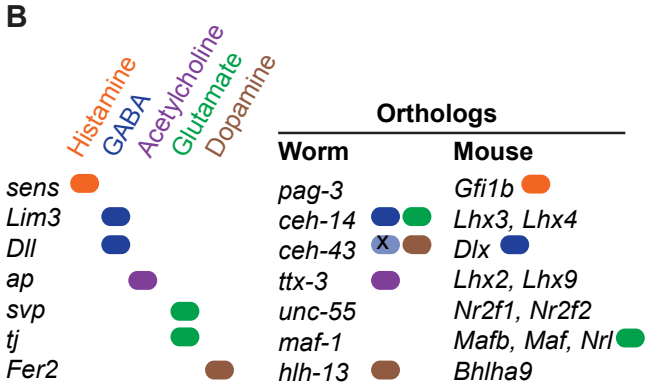
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE S6

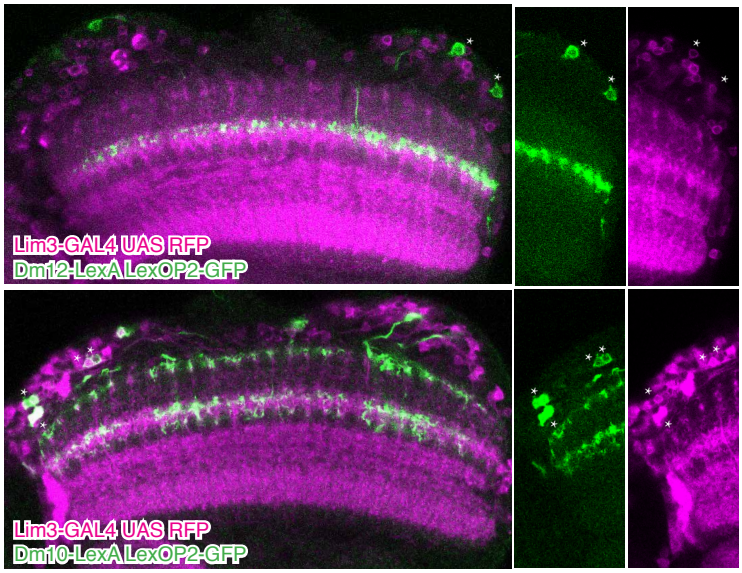
A Transcription factors correlated with neurotransmitter output



B



C



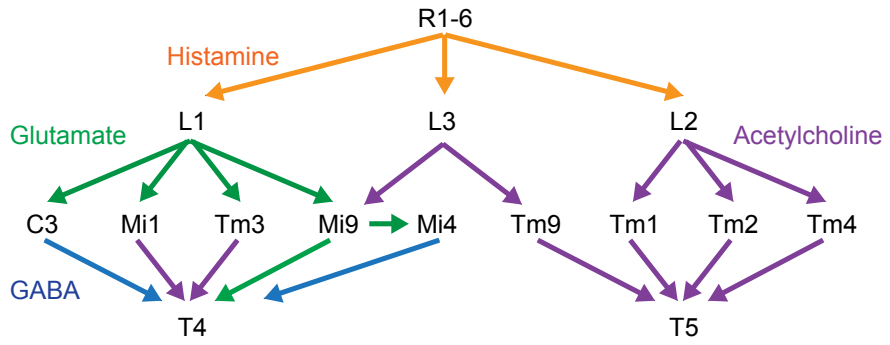
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure S6: Related to Figure 5. Transcriptional regulators of neurotransmitter identity. A. Transcription factors whose expression is predictive of neurotransmitter phenotype (i.e., high P(neurotransmitter output | transcription factor expressed)). The ten most predictive transcription factors are shown for each neurotransmitter output marker. B. Summary of orthologous transcription factors in worm and mouse and their association with specific neurotransmitter types. C. The Gad1-associated gene Lim3 does not express in cholinergic Dm12 neurons, but does in the GABA-ergic Dm10 neurons. Double labeling using LexA-markers for Dm12 and Dm10 (green) with a Lim3 protein-trap-GAL4 driving RFP (magenta).

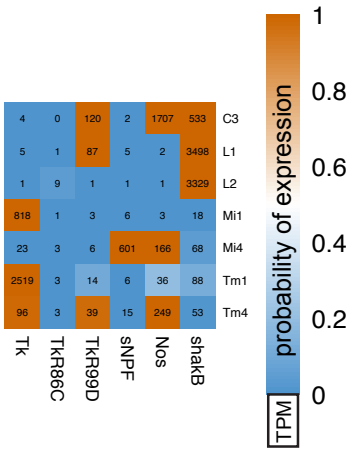
bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

FIGURE S7

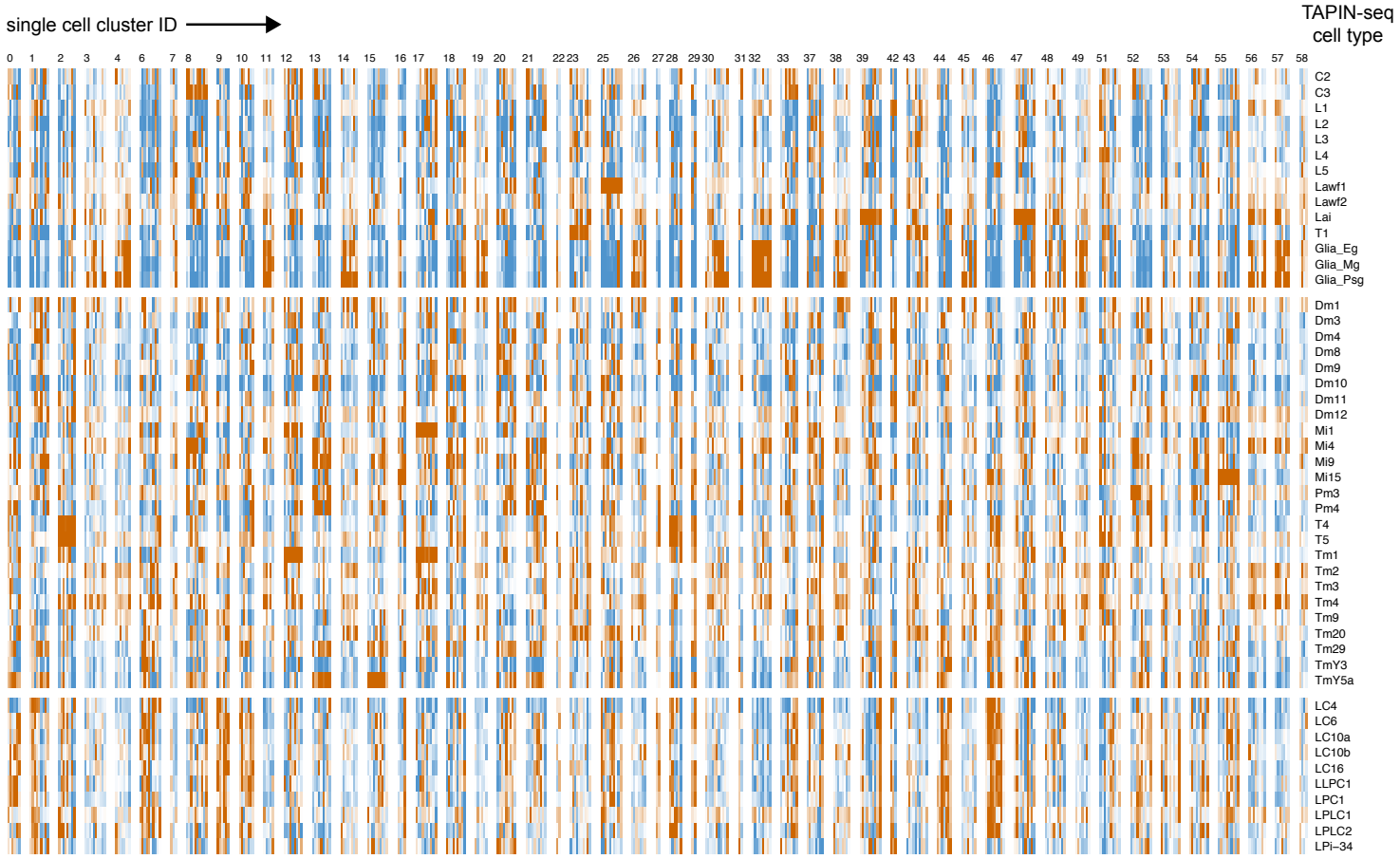
A. Motion detection pathway

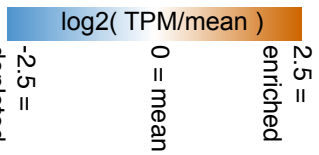


B.

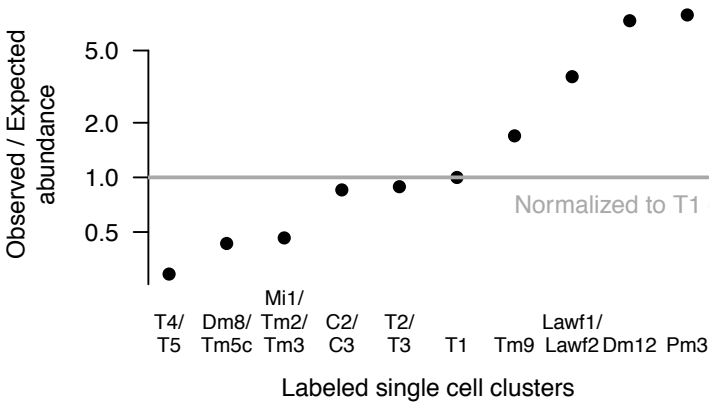


C. TAPIN-seq expression of cluster markers identified by scRNA-seq (Konstantinides et al., 2018)





D



bioRxiv preprint doi: https://doi.org/10.1101/385476; this version posted August 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure S7: Related to Figure 7. Patterns of neurotransmitter receptor expression complement connectomics. A. Transcriptomes reveal the neurotransmitters in core cell types of the ON and OFF components of the motion detection pathway. B. Examples of the expression of genes involved in neuropeptide, non-canonical small molecule (nitric oxide), or gap junction communication in the cell types in (A). C. We evaluated expression of marker genes for each single cell cluster (as reported in Konstantinides et al., 2018) in our TAPIN-seq profiles of visual system neurons. If a single cell cluster marker corresponds to one of our identified cell types, we expect to see its marker genes highly enriched in the corresponding cell type's expression. Note that some of the single cell clusters with the best apparent cell type matches (e.g. cluster 15/TmY5a, cluster 55/Mi15) were originally reported with a different annotation. D. We evaluated whether the single cell map proportionally represents cell types found in the optic lobe. By comparing the single cell cluster sizes to the true abundance of each cell type we found that the scRNA-seq map can both under- and over-estimate the abundance of each cell type (assuming accurate cell type labels), or that the cell type is incorrectly assigned (i.e., contains different or additional cell types). Observed/expected ratio = ((size of cluster labeled as cell type X / size of cluster labeled as T1) / (true abundance of cell type X / true abundance of T1)).