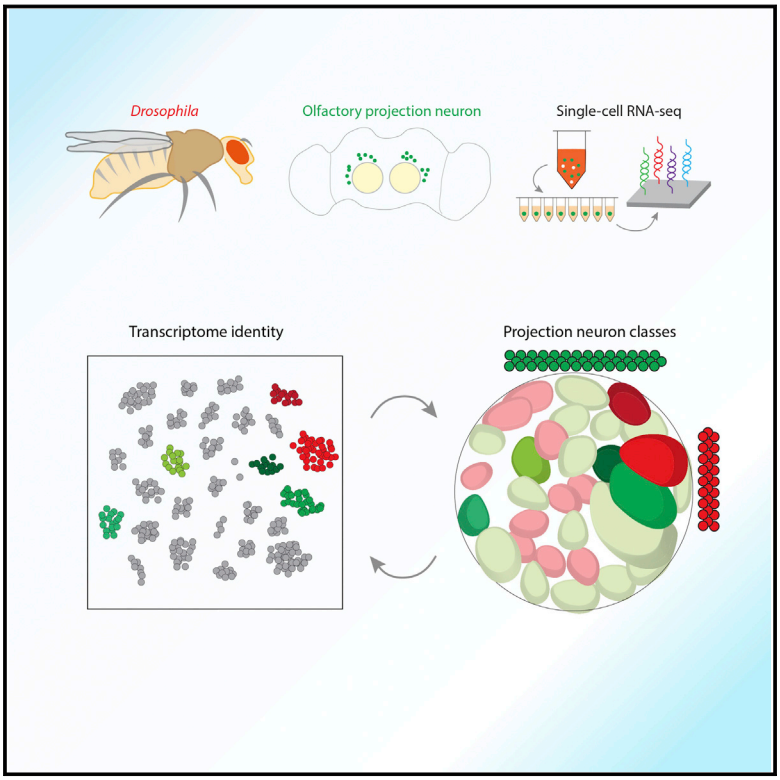
Resource

Cell

Classifying *Drosophila* Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing

Graphical Abstract



Authors

Hongjie Li, Felix Horns, Bing Wu, ..., David J. Luginbuhl, Stephen R. Quake, Liqun Luo

Correspondence

quake@stanford.edu (S.R.Q.), lluo@stanford.edu (L.L.)

In Brief

Single-cell RNA sequencing establishes that transcriptomic identity of *Drosophila* olfactory projection neurons corresponds with connectivity and function and identifies transcription factors and cell-surface molecules as highly informative in encoding neuronal identity.

Highlights

- We established a single-cell RNA-seq protocol for neurons and glia in *Drosophila*
- Transcriptome identity corresponds with olfactory projection neuron subtypes
- Neuronal transcriptome diversity peaks during circuit assembly
- Neuronal cell types are specified by a combinatorial molecular code

Li et al., 2017, Cell *171*, 1206–1220
November 16, 2017
https://doi.org/10.1016/j.cell.2017.10.019

CellPress

Cell

Resource

Classifying *Drosophila* Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing

Hongjie Li,[1,6] Felix Horns,[2,6] Bing Wu,[1] Qijing Xie,[1,3] Jiefu Li,[1] Tongchao Li,[1] David J. Luginbuhl,[1] Stephen R. Quake,[4,5,*] and Liqun Luo[1,7,*]

[1]Department of Biology and Howard Hughes Medical Institute
[2]Biophysics Graduate Program
[3]Neurosciences Graduate Program
[4]Departments of Bioengineering and Applied Physics
Stanford University, Stanford, CA 94305, USA
[5]Chan Zuckerberg Biohub, Stanford, CA 94305, USA
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: quake@stanford.edu (S.R.Q.), lluo@stanford.edu (L.L.)
https://doi.org/10.1016/j.cell.2017.10.019

SUMMARY

The definition of neuronal type and how it relates to the transcriptome are open questions. *Drosophila* olfactory projection neurons (PNs) are among the best-characterized neuronal types: different PN classes target dendrites to distinct olfactory glomeruli, while PNs of the same class exhibit indistinguishable anatomical and physiological properties. Using single-cell RNA sequencing, we comprehensively characterized the transcriptomes of most PN classes and unequivocally mapped transcriptomes to specific olfactory function for six classes. Transcriptomes of closely related PN classes exhibit the largest differences during circuit assembly but become indistinguishable in adults, suggesting that neuronal subtype diversity peaks during development. Transcription factors and cell-surface molecules are the most differentially expressed genes between classes and are highly informative in encoding cell identity, enabling us to identify a new lineage-specific transcription factor that instructs PN dendrite targeting. These findings establish that neuronal transcriptomic identity corresponds with anatomical and physiological identity defined by connectivity and function.

INTRODUCTION

The nervous system comprises many neuronal types with varied locations, input and output connections, neurotransmitters, intrinsic properties, and physiological and behavioral functions. Recent transcriptome analyses, especially from single cells, have provided important criteria to define a cell type. Indeed, single-cell RNA sequencing (RNA-seq) has been used to classify neurons in various parts of the mammalian nervous system (e.g., Darmanis et al., 2015; Johnson et al., 2015; Usoskin et al., 2015; Zeisel et al., 2015; Földy et al., 2016; Fuzik et al., 2016; Gokce et al., 2016; Shekhar et al., 2016; Tasic et al., 2016), but the extent to which it is useful to define subtypes of neurons and the relationship between cell type and connectivity is unclear in most cases. Indeed, what constitutes a neuronal type in many parts of the nervous system remains an open question (Johnson and Walsh, 2017).

The *Drosophila* olfactory circuit offers an excellent system to investigate the relationship between transcriptomes and neuronal cell types. 50 classes of olfactory receptor neurons (ORNs) form one-to-one connections with 50 classes of second-order projection neurons (PNs) in the antennal lobe in discrete glomeruli, forming 50 parallel information-processing channels (Figure 1A; Vosshall and Stocker, 2007; Wilson, 2013). Each ORN class is defined by expression of one to two unique olfactory receptor gene(s) and by the glomerulus to which their axons converge. Correspondingly, each PN class is also defined by the glomerulus within which their dendrites elaborate, which correlates strongly with the axonal arborization patterns at a higher olfactory center (Marin et al., 2002; Wong et al., 2002; Jefferis et al., 2007). Furthermore, while on average ~60 ORNs and ~3 PNs form many hundreds of synapses within a single glomerulus (Mosca and Luo, 2014), every ORN forms synapses with every PN to convey the same type of olfactory information (Kazama and Wilson, 2009; Tobin et al., 2017). Indeed, PNs that project to the same glomerulus exhibit indistinguishable electrophysiological properties and olfactory responses (Kazama and Wilson, 2009). Thus, one can define each PN class as a specific neuronal type (or subtype, if all PNs are collectively considered a cell type) with confidence that each class has unique connectivity, physiological properties, and functions, whereas PNs of the same class most likely do not differ. In other words, the ground truth of cell types for fly PNs is one of the best defined in the nervous system. We describe here a robust single-cell RNA-seq protocol for neurons and glia in the *Drosophila* brain and its application to *Drosophila* PNs to establish the relationship between transcriptome, neuronal cell identity, and development.

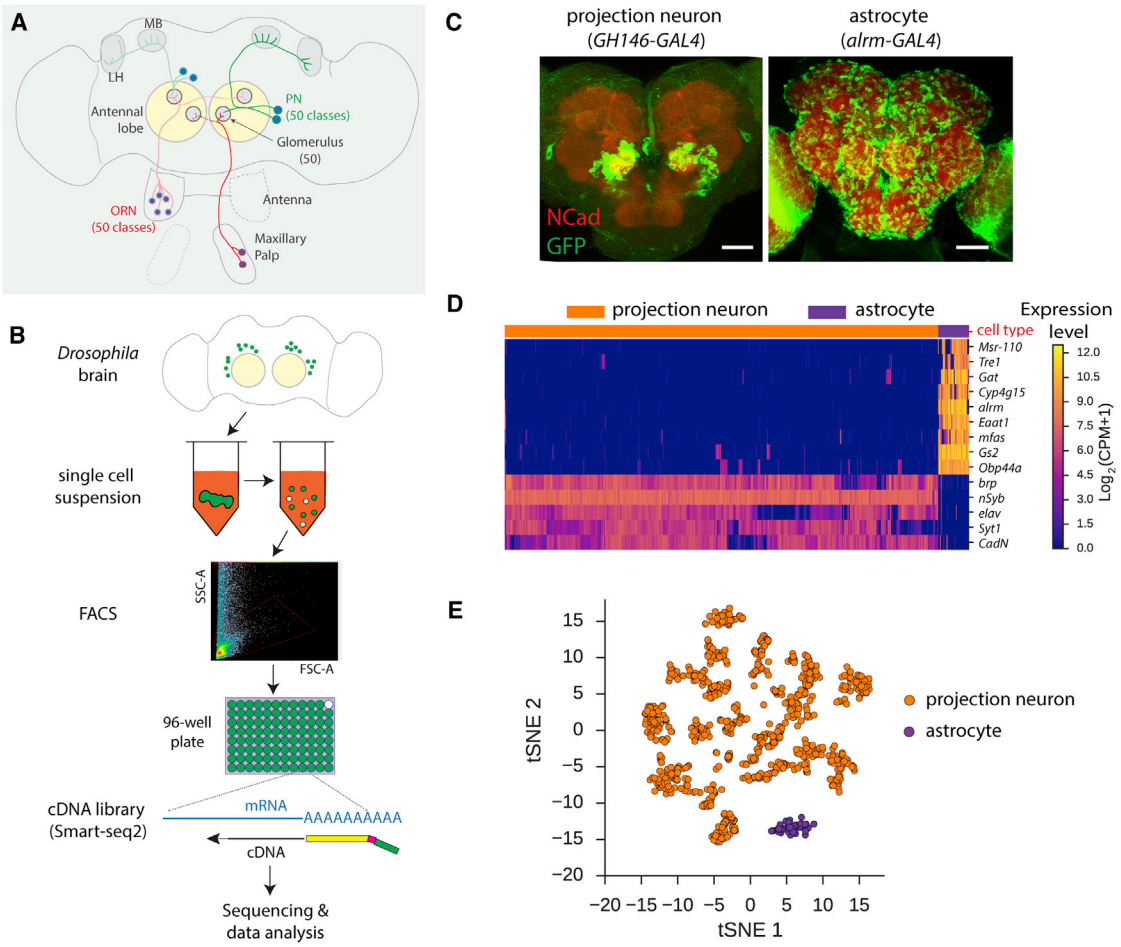


Figure 1. Single-Cell RNA-Seq Protocol for the *Drosophila* Pupal Brain

(A) Schematic of fly olfactory system organization. Olfactory receptor neurons (ORNs) expressing the same odorant receptor (same color) target their axons to the same glomerulus in the antennal lobe. Projection neuron (PN) dendrites also target single glomeruli, and their axons project to the mushroom body (MB) and lateral horn (LH).

(B) Schematic of single-cell RNA-seq protocol.

(C) Representative confocal images of *Drosophila* central brains labeled by *UAS-mCD8GFP* crossed with PN driver *GH146-GAL4* (24 hr APF) or astrocyte driver *alrm-GAL4* (72 hr APF). N-cadherin (Ncad, red) staining labels neuropil. Scale, 50 μm.

(D) Heatmap showing expression levels of genes that are specific for neurons or astrocytes. Each column is an individual cell. 67 *alrm-GAL4+* and 946 *GH146-GAL4+* cells are shown, with driver indicated by the color above. Cell-type-specific genes are enriched in astrocytes (top nine) and PNs (bottom five). Expression levels are indicated by the color bar (CPM, counts per million). Cells and genes were ordered using hierarchical clustering.

(E) Visualization of astrocyte and PN populations using t-distributed stochastic neighbor embedding (tSNE). Each dot is a cell.

See also Figure S1.

RESULTS

A Robust Single-Cell RNA-Seq Protocol for the *Drosophila* Pupal Brain

Brains containing cells labeled by mCD8GFP driven from specific GAL4 lines were manually dissected, single-cell suspensions were prepared following a method modified after Tan et al. (2015), and cDNA were sequenced using a modified SMART-seq2 protocol (Picelli et al., 2014) (Figure 1B; Figure S1A; STAR Methods). We sequenced cells from *Drosophila* pupal brains that were labeled by the astrocyte driver *alrm-GAL4* (Doherty et al., 2009) and olfactory PNs labeled by the *GH146-GAL4* driver, which is expressed in 40 of 50 PN classes (Stocker et al., 1997; Jefferis et al., 2001) (Figure 1C). About 5% of GFP-labeled cells within the brain were recovered as single cells, and 90% of PNs yielded high-quality cDNA after reverse transcription (Figures S1B and S1C). Cells were sequenced to a depth of ~1 million reads per cell, and 1,000–4,000 genes were detected per cell (Figure S1D). Data quality was evaluated by examining expression of five neuronal markers (*brp*, *nSyb*, *elav*, *Syt1*, and *CadN*) and four astrocyte markers (*alrm*, *Eaat1*, *Gat*, and *Gs2*) (Doherty et al., 2009; Sinakevitch et al., 2010; Stork et al., 2014); they were specifically expressed in the corresponding cell types (Figure 1D). We also identified five new genes (*Msr-110*, *tre1*, *Cyp4g15*, *mfas*, and *Obp44a*) that were expressed in pupal astrocytes, but not in PNs (Figure 1D). Unbiased clustering

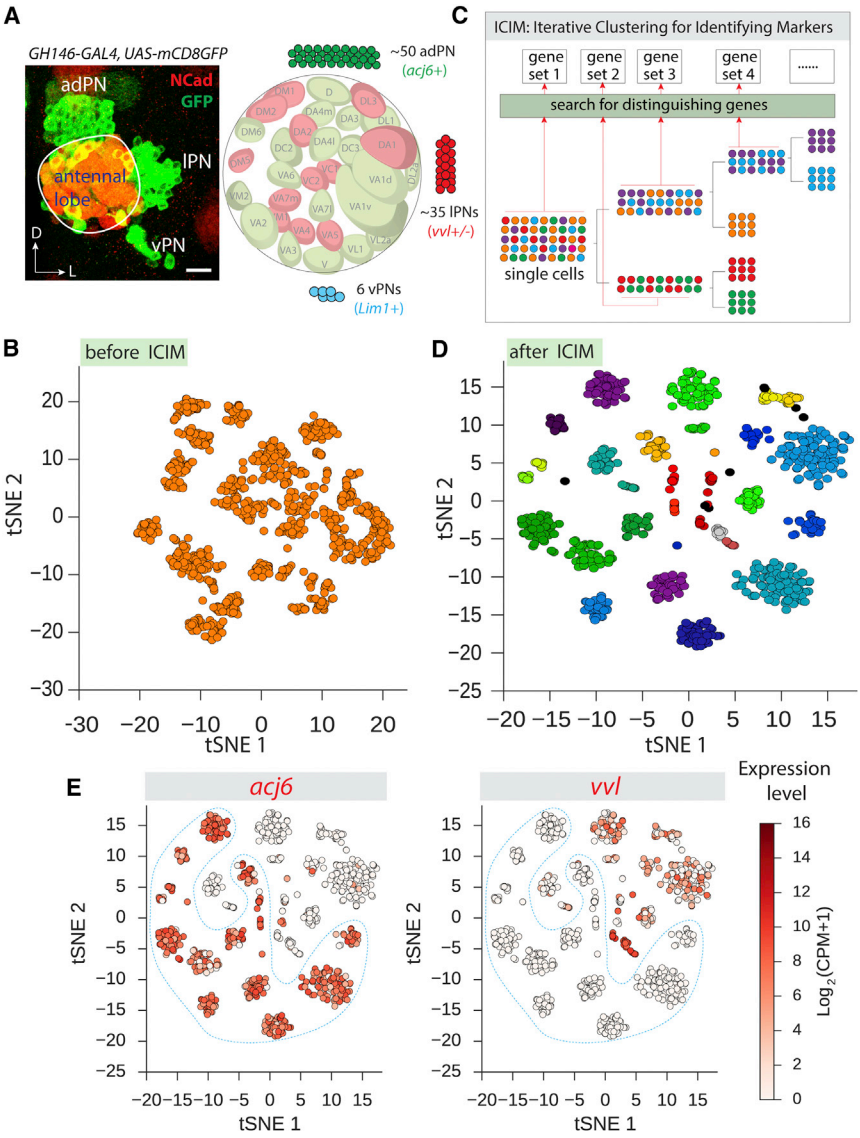


Figure 2. Single-Cell RNA-Seq Analysis of *GH146+* PNs

(A) Representative confocal projection and schematic of *GH146+* PNs, which include (per antennal lobe) 50 adPNs (*acj6+*), 35 lPNs (*vvl* expression begins to decrease from 18 hr APF; Komiyama et al., 2003), and 6 vPNs (*Lim1+*). The cell bodies of adPNs, lPNs, and vPNs are located anterodorsal, lateral, and ventral, respectively, to the antennal lobe neuropil (circled, stained by Ncad). All *GH146+* adPNs and lPNs send dendrites to a single glomerulus. The schematic shows the stereotyped locations of a large subset of glomeruli (named according to their locations; Laissue et al., 1999), color-coded according to adPNs or lPNs. Scale, 20 μm. D, dorsal; L, lateral.

(B) Visualization of *GH146+* PNs using dimensionality reduction by PCA followed by tSNE. Each dot is a cell. Cells are arranged according to transcriptome similarity.

(C) Schematic of iterative clustering for identifying markers (ICIM), an unsupervised machine-learning algorithm for identifying genes that distinguish cell types.

(D) Visualization of *GH146+* PNs using tSNE based on 561 genes identified using ICIM. *GH146+* adPNs and lPNs form 30 distinct clusters (differentially colored). Black dots are cells that could not be assigned to any cluster.

(E) Visualization of *GH146+* PNs as in (D), colored according to *acj6* and *vvl* expression level. *acj6* and *vvl* are expressed in *GH146+* PNs in a mutually exclusive manner.

See also Figure S2.

based on transcriptome profiles readily distinguished PNs and astrocytes (Figure 1E). Among PNs, housekeeping genes (e.g., *Act5C* and *α-Tub84B*) were reliably detected in all cells, and stress-related genes (e.g., *Hsp70*-family genes) were not widely induced (Figure S1E). ~50% of cells co-expressed two male-specific RNAs (Meller et al., 1997) (Figure S1F), as expected, given that we did not discriminate sex. These data demonstrate the reliability of our single-cell RNA-seq protocol for analyzing cell types and transcriptomes in *Drosophila* pupal brain.

Clustering *GH146-GAL4+* PNs Based on Single-Cell Transcriptomes

GH146-GAL4+ (*GH146+* hereafter) PNs are derived from three neuroblast lineages whose cell bodies are located anterodorsal, lateral, or ventral to the antennal lobe neuropil (Figure 2A; Jefferis et al., 2001). The anterodorsal and lateral lineages produce uniglomerular, cholinergic, and excitatory PNs (adPNs and lPNs), whereas the ventral lineage produces GABAergic inhibitory PNs (vPNs), some of which target dendrites to multiple glomeruli (Jefferis et al., 2001; Liang et al., 2013). We sequenced 1,046 single *GH146+* cells at 24–30 hr after puparium formation (APF). At this stage, PNs are refining their dendrite targeting in the antennal lobe; these dendrites also serve as targets for ORN axons that will invade the antennal lobe and establish one-to-one connections in the following 24 hr (Jefferis et al., 2004). We analyzed 946 cells that passed quality filtering (STAR Methods).

Conventional dimensionality reduction and clustering methods based on principal-component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE) (van der Maaten and Hinton, 2008) identified only ~12 distinct PN clusters (Figure 2B). The inability to resolve more distinct clusters is likely due to the limited sensitivity of these methods to distinguish cell types with highly similar transcriptomes, as we expect for the PN classes. To address this challenge, we developed an unsupervised machine-learning algorithm, iterative clustering for identifying markers (ICIM), to identify genes that distinguish PN classes. ICIM searches for genes having the highest expression variability within a cell population, partitions the cells into two

subpopulations using clustering based on these genes, then iteratively repeats the search on each subpopulation. Iteration continues until distinct subpopulations cannot be separated because gene expression patterns within the population are homogeneous (Figure 2C). Stopping criteria are defined in an unbiased manner without supervision. Genes identified using ICIM were then used for further dimensionality reduction using tSNE and clustering using HDBSCAN, a hierarchical density-based clustering algorithm (Campello et al., 2013), on the tSNE space. Applying ICIM to the transcriptomes of *GH146+* PNs, we identified 561 genes that segregate the 946 *GH146+* cells into 35 distinct clusters (Figures S2A and S2B).

Two of the 35 clusters expressed known markers for vPNs (Figure S2A): *Gad1+*, a GABA biosynthetic enzyme, and *Lim1+*, a transcription factor (TF) expressed in vPNs, but not in adPNs or lPNs (Komiyama and Luo, 2007), suggesting that they correspond to vPNs. Besides PNs, the only other cells that *GH146-GAL4* also consistently labeled at a high level at 24 hr APF were the anterior paired lateral (APL) neurons (Figure S2C). Three other clusters expressed *VGlut* (Figure S2B), which specifically marked *GH146+* APL neurons, but not PNs (Figure S2D), suggesting that this *VGlut+* population consists of APL neurons. Because we were interested primarily in excitatory adPNs and lPNs, we removed inhibitory vPNs and APL neurons from subsequent analysis. Nearly all of the remaining 902 *GH146+* cells should be adPNs and lPNs, which collectively target 40 glomeruli. Clustering analysis using ICIM and tSNE identified 30 distinct clusters (Figure 2D). Library complexity and sequencing depth did not drive clustering (Figure S2E). The number of cells belonging to each cluster varied from 5–108, likely reflecting the fact that different PN classes contain different cell numbers, ranging from 1–7 cells per antennal lobe (Yu et al., 2010; Lin et al., 2012). It is likely that we did not sample a sufficient number of cells to detect rare PN classes.

We previously showed that two TFs, abnormal chemosensory jump 6 (Acj6) and ventral veins lacking (Vvl; also known as Drifter), are expressed in adPNs and lPNs, respectively (Figure 2A), and instruct lineage-specific dendrite targeting (Komiyama et al., 2003). Indeed, our single-cell RNA-seq analysis revealed that *acj6* and *vvl* were expressed in a mutually exclusive manner (Figures 2E and S2F). Among the 30 clusters of adPNs and lPNs, 18 clusters (60%) expressed *acj6*, but not *vvl*, and thus represent adPNs. The remaining 12 clusters were likely lPNs.

In summary, single-cell RNA-seq analysis revealed distinct clusters of *GH146+* adPNs and lPNs expressing lineage markers in a manner consistent with prior knowledge. TF transcripts, whose protein levels are generally low (Ghaemmaghami et al., 2003), were reliably detected within PNs and could be used to assign lineage identity, supporting the specificity and sensitivity of this method.

Matching Clusters to PN Classes Using Known Markers

We next attempted to map the correspondence between transcriptome-based PN clusters and glomerular-target-based PN classes by leveraging drivers that label specific PN classes. We found that *91G04-GAL4* (Jenett et al., 2012) was robustly expressed in PNs at 24 hr APF. To limit expression only to PNs, we utilized an intersectional strategy by combining *91G04-GAL4* with *GH146-Flp* (Potter et al., 2010) and *UAS-FRT-STOP-FRT-mCD8GFP* (Hong et al., 2009), such that only cells that express both *91G04-GAL4* and *GH146-Flp* would express mCD8GFP (hereafter referred to as "intersecting with *GH146-Flp*"). This resulted in expression of mCD8GFP in just two adPNs per hemisphere, both of which project dendrites to the DC2 glomerulus (Figure 3A). We sequenced 23 *91G04+* PNs at 24–30 hr APF and performed clustering analysis using ICIM and tSNE together with the *GH146+* cells. We found that all *91G04+* PNs mapped to one *GH146+* cluster (Figure 3C; cluster #1). All *91G04+* cells could also be unambiguously mapped to this *GH146+* cluster using a random forest classifier (data not shown). Thus, cluster #1 corresponds to DC2 PNs.

Mz19-GAL4 is expressed from 24 hr APF to adulthood (Figure 3B; Jefferis et al., 2004). After intersecting with *GH146-Flp*, *Mz19-GAL4* labels three PN classes: adPNs that project to VA1d and DC3 (*acj6+*), and lPNs that project to DA1 (*acj6−*). We sequenced 123 *Mz19+* cells at 24–30 hr APF and mapped them to four clusters of *GH146+* cells (Figure 3C). The *Mz19+* and *acj6−* cells, corresponding to DA1 PNs, mapped to two clusters (#2 and #2′), suggesting that both correspond to DA1 PNs, a notion that we explore further below. The *Mz19+* and *acj6+* cells, corresponding to VA1d and DC3 PNs, mapped to two clusters of *GH146+* cells (#3 and #4; Figure 3C). Thus, clusters #3 and #4 correspond to VA1d and DC3 PNs. We establish a one-to-one correspondence between these clusters and PN classes below.

Matching Clusters to PN Classes Using Newly Identified Markers

To map additional PN transcriptome clusters to glomerular classes, we searched the single-cell transcriptome data for new markers. We identified *terribly reduced optic lobes* (*trols*) as predominantly expressed in a single cluster (Figure 4A). Intersecting an existing *trol-GAL4* (*NP5103-GAL4*, inserted into an intron of *trol*) with *GH146-Flp* labeled 2–3 adPNs at both 24 hr and 72 hr APF, with dendrites projecting to the VM2 glomerulus at 72 hr APF (Figure 4B). 28 sequenced *trol-GAL4+* PNs (after intersecting with *GH146-Flp*) mapped to the original *trol+* cluster #5 (Figure 4C). These data indicate that *trol-GAL4* mimics endogenous *trol* expression and that cluster #5 corresponds to VM2 PNs.

Using *Mz19-GAL4*, we mapped *Mz19+* and *acj6+* VA1d and DC3 PNs to clusters #3 and #4 (Figure 3C) but could not resolve which cluster belonged to which PN class. VA1d and DC3 PNs are closely related: VA1d PNs are born immediately after DC3 PNs from the same lineage and target dendrites to neighboring glomeruli. To establish a one-to-one mapping, we identified *CG31676* to have the strongest differential expression between the two clusters (Figure 4D; compare clusters #3 and #4). We generated *CG31676-GAL4* by inserting into the first intron of *CG31676* a cassette containing a splice acceptor (SA) sequence followed by a T2A peptide sequence and the *GAL4* coding sequence (Figure 4E). After intersecting with *GH146-Flp*, *CG31676-GAL4* labeled a similar number of PNs at 24 hr, 48 hr, and 72 hr APF, which targeted dendrites to VA1d, but not DC3 (Figure 4F). Thus, *Mz19+*, *acj6+*, and *CG31676+* cluster #3 corresponds to VA1d PNs; the *Mz19+*, *acj6+*, and *CG31676−* cluster #4 corresponds to DC3 PNs.

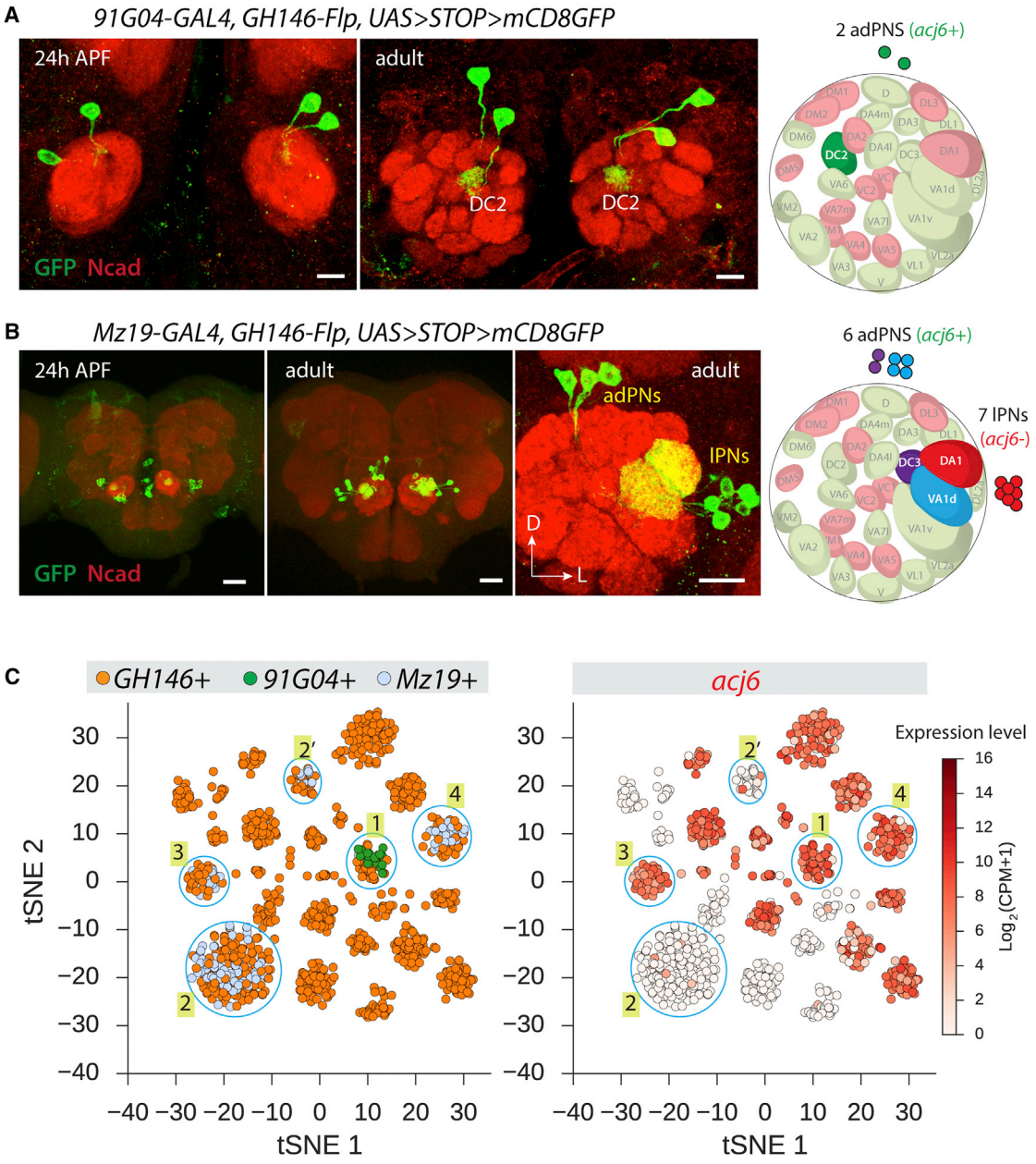


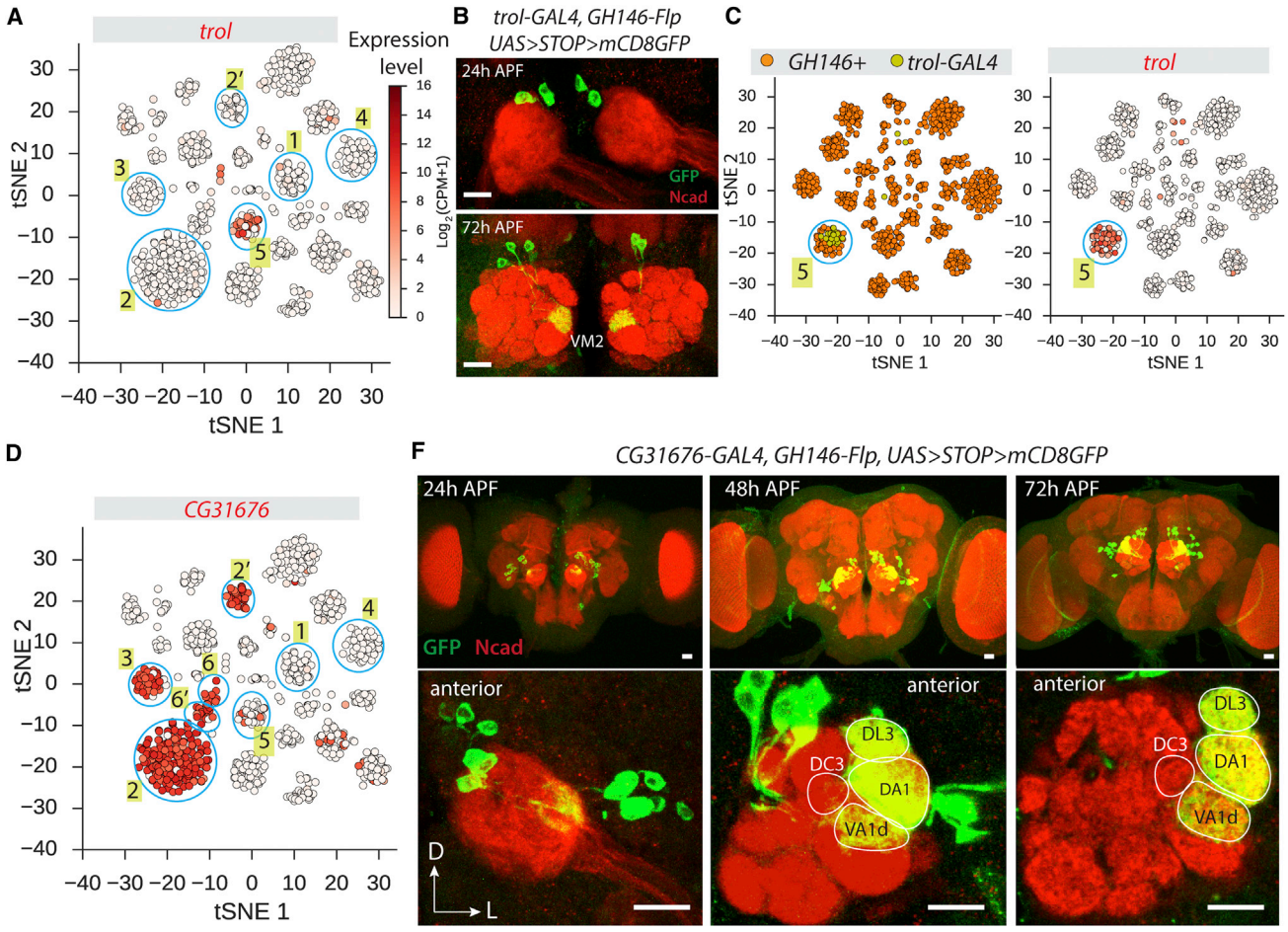
Figure 3. Mapping Clusters to PN Classes Using Known Markers

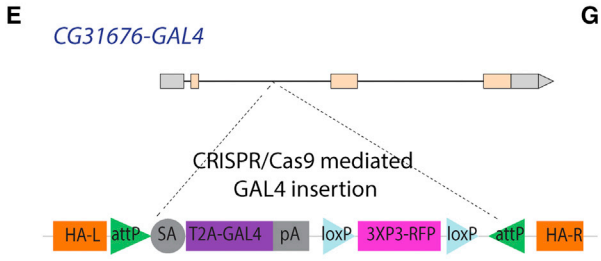
(A and B) Intersecting *GH146-Flp* with *91G04-GAL4* (A) or with *Mz19-GAL4* (B) labels only one PN class (DC2) or 3 PN classes (*acj6*+ VA1d and DC3; *acj6*− DA1), respectively, at 24 hr APF and in adults. Driver schematics are shown on the right. Scale for (A), 20 μm; for (B), 50, 50, and 20 μm.

(C) Visualization of *GH146+, 91G04+,* and *Mz19+* PNs using tSNE as in Figure 2D. Cells are colored according to driver (left) or by expression level of *acj6* (right). *91G04+* cells (green) map to a single cluster (#1) of *GH146+* cells (orange); thus, cluster #1 corresponds to DC2 PNs. DA1 PNs (*Mz19+* and *acj6*−) map to two clusters, #2 and #2′. VA1d and DC3 PNs (*Mz19+* and *acj6+*) map to two clusters, #3 and #4.

In addition to DA1 (#2 and #2′) and VA1d (#3), *CG31676-GAL4* also strongly labeled DL3, which is targeted by *acj6*− lPNs (Figures 4F and S3A) (Jefferis et al., 2001). Among the 30 clusters, only two clusters (#6 and #6′) were *CG31676+* and *acj6*− (Figures 4D and S3B); these two clusters displayed highly similar transcriptomes, as reflected in their close proximity in the tSNE plot. We therefore mapped clusters #6 and #6′ to DL3 PNs. *CG31676-GAL4* transiently labeled two other glomeruli targeted by *acj6+* adPNs (Figure S3A), but we could not unambiguously assign them to corresponding clusters.

Among the six glomerular classes we have mapped, four corresponded to a single transcriptome cluster each, but DA1 and DL3 PNs each corresponded to two clusters (Figures 4D and 4G). All PN classes are born in a stereotyped order





G

cluster	PN class	lineage	Marker expression
1	DC2	adPN	*91G04+*
2	DA1	lPN	*Mz19+/acj6-/fru+*
2'	DA1	lPN	*Mz19+/acj6-/fru-*
3	VA1d	adPN	*Mz19+/acj6+/CG31676+*
4	DC3	adPN	*Mz19+/acj6+/CG31676-*
5	VM2	adPN	*trol+*
6	DL3	lPN	*Mz19-/CG31676+/acj6-/mirr-*
6'	DL3	lPN	*Mz19-/CG31676+/acj6-/mirr+*

Figure 4. Mapping Clusters to PN Classes Using Newly Identified Markers

(A) Visualization of *GH146+* PN cells using tSNE as in Figure 2D, showing expression of *trol* enriched in one cluster (#5). Clusters #1–#4 from Figure 3 are also indicated.

(B) After intersecting with *GH146-Flp*, *trol-GAL4* labels 2–3 PNs in each hemisphere at 24 hr and 72 hr APF, which project dendrites to the VM2 glomerulus.

(C) Visualization of *GH146+* and *trol+* PNs using tSNE based on 561 genes previously identified using ICIM (Figure 2D). Cells are colored according to driver (left) or by expression level of *trol* (right; color bar in A). *trol-GAL4+* PNs map to one *GH146+* PN cluster (left), which expresses high levels of *trol* (right). Thus, cluster #5 corresponds to VM2 PNs.

(D) Visualization of *GH146+* PNs using tSNE as in (A) showing expression of *CG31676* (color bar in A). Among two *Mz19+* and *acj6+* clusters, *CG31676* is highly expressed in cluster #3, but not #4. Several *acj6−* (see Figure S3B) clusters also express *CG31676*, including #2 and #2′ (DA1) and #6 and #6′.

(E) Schematic of CRISPR/Cas9-mediated insertion of *T2A-GAL4* into the first intron of *CG31676*.

(F) *CG31676-GAL4* expression in PNs after intersecting with *GH146-Flp*. Similar numbers of PNs are labeled at 24 hr, 48 hr, and 72 hr APF. VA1d, but not DC3, is labeled, enabling us to map cluster #3 to VA1d (*CG31676+*) and cluster #4 to DC3 (*CG31676−*). In addition, DA1 and DL3 are labeled. Thus, the remaining *acj6−* and *CG31676+* cells (clusters #6 and #6′) correspond to DL3 PNs.

(legend continued on next page)

within a specific lineage, and most PN classes are born consecutively within a single time window (Jefferis et al., 2001; Yu et al., 2010; Lin et al., 2012). DA1 and DL3 PNs are the only two exceptions: they are born in two time windows separated by more than 24 and 12 hr, respectively (Figure S3D; Lin et al., 2012). This birth-timing difference may contribute to the transcriptome heterogeneity of DA1 and DL3 PNs. For DA1 PNs, we found that *fruitless* (*fru*), encoding a TF and a key regulator of male sexual behavior (Dickson, 2008), was expressed only in the large cluster (#2). This is consistent with a previous finding that *NP21-GAL4* (inserted into a *fru* intron near the sexually dimorphic splicing site) only labels DA1 PNs after intersecting with *GH146-Flp* (Potter et al., 2010). On the other hand, *CG45263* was only expressed in the small cluster (#2′) (Figure S3C). We also identified genes that were expressed only in one of the two DL3 clusters (Figure S3B). It remains to be determined whether the transcriptional differences between clusters #2 and #2′ and between clusters #6 and #6′ reflect only differences in birth timing or also potential differences in biological functions.

In summary, by using a combination of existing markers and new markers discovered using single-cell RNA-seq, we have unambiguously mapped six PN classes to corresponding transcriptome clusters (Figure 4G). Our results indicate that the combination of genetic drivers and single-cell RNA-seq offers a simple strategy for mapping transcriptome clusters to cell types.

A New Lineage-Specific TF Regulates Dendrite Targeting

Our single-cell transcriptome analysis identified many TFs that were differentially expressed in separate clusters. For example, *prospero* mRNA was expressed in a majority of PNs, including all *Mz19+* PNs, whereas *cut* mRNA was expressed in a few PNs, all of which were *Mz19−* (Figure S4A). Indeed, antibody staining validated these observations (Figure S4B), and the expression of Cut is consistent with our previous finding (Komiyama and Luo, 2007).

Our analysis also identified new lineage-specific expression for several TFs. Specifically, *C15* and *knot* mRNAs were observed only in adPNs, and *unplugged* (*unpg*) was observed only in lPNs (Figure 5A). We confirmed these results by immunostaining using antibodies against C15 and Knot and a *lacZ* reporter for *unpg* (Figure 5B). *knot* plays a critical role in controlling dendrite development of *Drosophila* sensory neurons (Jinushi-Nakao et al., 2007), and *unpg* is a marker for specific neuroblast sublineages in the *Drosophila* embryonic ventral nerve cord (Cui and Doe, 1995). *C15*, encoding a homeobox-containing protein, is a homolog of human *Hox11* critical in regulating a gene network in the developing *Drosophila* leg (Campbell, 2005), but its neural function is unknown. We tested whether *C15* plays a role in PN dendrite targeting.

In a loss-of-function experiment, we used *elav-GAL4* to knockdown *C15* in all neurons, *Mz19-QF*-driven *QUAS-mCD8GFP* to monitor dendrite targeting of VA1d and DA1 PNs (*Mz19-QF* labels DA1 and VA1d, but not DC3 PNs in wild-type [Hong et al., 2012]), and *Or88a* promoter-driven myristolated tdTomato (*Or88a-mtdT*) to monitor axon targeting of VA1d ORNs (Ward et al., 2015). Pan-neuronal knockdown of *C15* using a strong RNAi line (Figure S4C) caused a highly penetrant dorsal shift of the VA1d glomerulus without affecting DA1 dendrite targeting (Figures 5C, 5D, and S4D), concomitant with a loss of dendrites in the VA1d glomerulus. This loss could be because (1) C15 controls the expression of *Mz19-QF* in VA1d PNs, (2) VA1d neurons die or are not born, or (3) VA1d dendrites mistarget to the DA1 glomerulus.

In a gain-of-function experiment, we used the *Mz19-GAL4*-based MARCM system to misexpress *C15*. Control *Mz19+* adPNs target to the VA1d and DC3 glomeruli, and lPNs target to the DA1 glomerulus (Figures 3B and 5E, left panels). However, when *C15* was misexpressed, *Mz19+* lPNs (DA1 PNs only) sent dendrites to regions outside the DA1 glomerulus, including VA1d, DC3, DA3, and DA4l, that are all normally targeted by adPNs, while *Mz19+* adPNs targeted dendrites correctly (Figures 5E [right panels], S4E, and S4F). These data suggest that the TF C15, as with Acj6 and Vvl (Komiyama et al., 2003), instructs lineage-specific PN dendrite targeting.

Transcriptomes of Closely Related PN Classes Exhibit the Largest Differences during Circuit Assembly

How do neuronal transcriptomes change as development proceeds? By mapping clusters from single-cell RNA-seq data to specific PN classes at different developmental stages, we can address this key question at the resolution of single PN classes. We focused on the three classes of *Mz19+* adPNs and lPNs, which have been unequivocally mapped to specific transcriptome clusters (Figure 4G).

Following the coarse patterning of PN dendrites at 24 hr APF, ORN axons invade the antennal lobe to identify their PN partners beginning ~30 hr APF, until they match with cognate PNs and establish discrete glomerular compartments first visible ~48 hr APF (Jefferis et al., 2004). Following further expansion of terminal branches of ORN axons, PN dendrites, and synaptogenesis, pupae become adults at ~100 hr APF. Using the intersection of *Mz19-GAL4* and *GH146-Flp*, we sequenced and analyzed 485 single cells from five time points (~100 cells each): 24–30 hr, 36–42 hr, 48–54 hr, 72–78 hr APF, and 1–2 days adult (Figure 6A).

Clustering analysis using ICIM and tSNE revealed that *Mz19+* and *acj6+* (VA1d and DC3) and *Mz19+* and *acj6−* (DA1) PNs were clearly separable at all times (Figure 6B). Interestingly, VA1d and DC3 PNs formed distinct clusters at the four pupal stages but merged into a single cluster in the adult (Figures 6B and 6C). To confirm this observation quantitatively, we calculated cell-type identity scores using the 22 most differentially expressed genes ($p < 10^{-5}$) between VA1d and DC3 PNs across all pupal stages and found that the difference between the transcriptional states of these two PN classes was maintained from 24–48 hr APF but began to shrink at 72 hr APF and were indistinguishable in the adult (Figure 6D). Using an alternative,

(G) Summary of the mapping of six PN classes to corresponding transcriptome clusters. Markers used for unambiguous mapping (A–D and Figures 3, S3B, and S3C) are listed.
Ncad is used as a neuropil marker. Scale, 20 μm.
See also Figure S3.

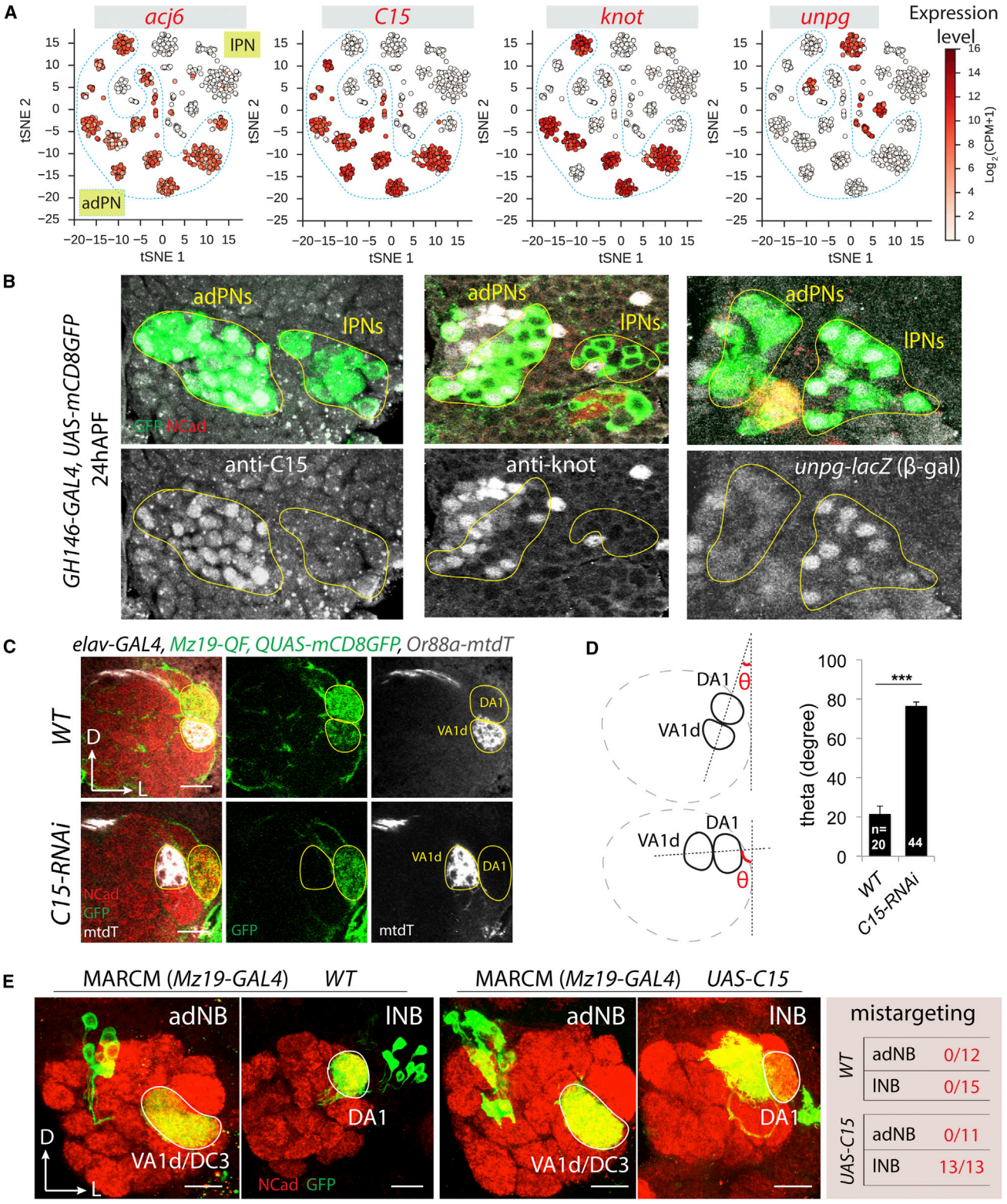


Figure 5. Identification of New Lineage-Specific Transcription Factors Using Single-Cell RNA-Seq

(A) Visualization of *GH146*+ PNs using tSNE as in Figure 2E showing expression of *acj6*, *C15*, *knot*, and *unpg*. adPNs are outlined (based on *acj6* expression) and remaining cells are lPNs.

(legend continued on next page)

unbiased genome-wide method, we calculated the Pearson correlation between the expression profiles of all pairs of cells based on 497 genes identified by ICIM. This analysis confirmed that transcriptome differences between VA1d and DC3 PNs disappeared in the adult (Figure 6E). Indeed, clustering analyses using only adult VA1d and DC3 PNs failed to find distinct populations (data not shown). Collectively, these data indicate that VA1d and DC3 PNs exhibit peak transcriptome differences during early pupal stages (24–48 hr APF) when PNs are refining their dendrite targeting and presenting cues for ORN axon targeting. These differences progressively diminish in late pupal and adult stages (Figure 6F).

These observations suggest that PN subtype identity genes, which distinguish VA1d and DC3 during the wiring stages, are downregulated once wiring specificity is established. To test this, we systematically identified differentially expressed genes at different stages in all *Mz19+* PNs. Gene ontology (GO) analysis indeed revealed that downregulated genes consisted of factors associated with development and differentiation, whereas most upregulated genes were associated with metabolic processes (Figure S5A). Clustering of genes based on their dynamic expression pattern revealed transcriptional waves consisting of genes that were coordinately turned down or up at different developmental stages (Figure S5B). Notably, many more TFs and cell-surface and secreted molecules (CSMs) were downregulated than upregulated (Figure S5B); CSMs were drawn from a database curated for relevancy to cell recognition and wiring specificity but excluding ion channels, transporters, and secreted enzymes (Kurusu et al., 2008). *CG31676*, which was expressed in VA1d, but not DC3, PNs at 24 hr APF (Figures 4D and 4G), was turned off in both PN classes in the adult while its expression in DA1 persisted (Figure S5C); this was validated with *CG31676-GAL4* expression analysis across developmental stages (Figure S5D).

Next, we asked if transcriptomes of PN classes from the same neuroblast lineage are more similar than those from different lineages. We found that the transcriptome differences between VA1d and DC3 PNs (both adPNs) were consistently smaller than that between VA1d and DA1 PNs (adPNs and lPNs, respectively) across developmental stages (Figure 6G). Similarly, the transcriptome differences at 24 hr APF between DA1 and DL3 lPNs were similar to those between VA1d and DC3 but smaller than those between VA1d and DA1 PNs. All four PN classes target to adjacent glomeruli in the dorsolateral antennal lobe (Figure 6A, right). Thus, lineage origin correlates to transcriptome similarities more than dendrite targeting position does, highlighting the important contribution of cellular ancestry to transcriptome state.

PN Subtype Identity Is Encoded by a Combinatorial Molecular Code

How is cell-type identity encoded in the transcriptome? It is possible that (1) each cell type expresses at least one unique gene, or (2) each cell type expresses a unique subset of a shared pool of genes. The strategy used for encoding cell-type identity in the nervous system remains an unresolved question. To comprehensively address how neuronal subtype identity is encoded in 24-hr-APF pupal PNs, we approximated the 30 *GH146+* transcriptome clusters as 30 subtypes and searched for marker genes that were uniquely expressed in a single subtype. We designed two criteria: (1) the gene must be robustly expressed within a cluster (> 7 counts per million [CPM], or $\log_2(\text{CPM}+1) > 3$, in > 50% of the cells of a cluster), and (2) the gene must not be expressed in any other cluster (> 7 CPM in < 10% of the cells of any other cluster). Only 6 genes fulfilled these criteria were (Figure S6A), sufficient to encode 5 of the 30 clusters. With relaxed criteria, we quickly entered a regime where identified genes were expressed in multiple clusters and hence not unique (Figure S6B). The inability to detect unique markers in most cell types was not due to transcript dropouts (Figure S6C). Thus, with few exceptions, *GH146+* PNs lack marker genes that uniquely encode subtypes.

Next, we sought to identify combinatorial molecular codes for cell type identity. We searched for a minimal set of genes that could uniquely encode PN subtypes using an information theoretic approach. We calculated the information content of each gene with respect to PN subtype identity, formally defined as the mutual information between the binarized expression state of the gene (ON/OFF) and PN-cluster identity (STAR Methods). We ranked genes by their information content and then selected a minimal set of genes by greedy search, iteratively drafting the gene carrying the most non-redundant information about identity into the set until 95% of the uncertainty of subtype identity was explained. The result of this search is a set of genes for which knowledge of their expression states (ON/OFF) alone is sufficient to classify subtype identity with high accuracy. We first applied this strategy to the three *Mz19+* PN classes. Only two genes, *C15* and *CG31676*, were sufficient to distinguish these three subtypes (Figure 7A), explaining 92% of the uncertainty of classification of individual *Mz19+* PN cells into subtypes. Both *C15* and *CG31676* were independently identified and characterized earlier in our study (Figures 4D and 5). This finding demonstrates

(B) Consistent with RNA-seq data in (A), 24 hr APF expression of C15 and Knot (antibody staining) in *GH146+* PNs (green) is restricted to adPNs, while *unpg* (anti-β-gal staining) is restricted to lPNs.

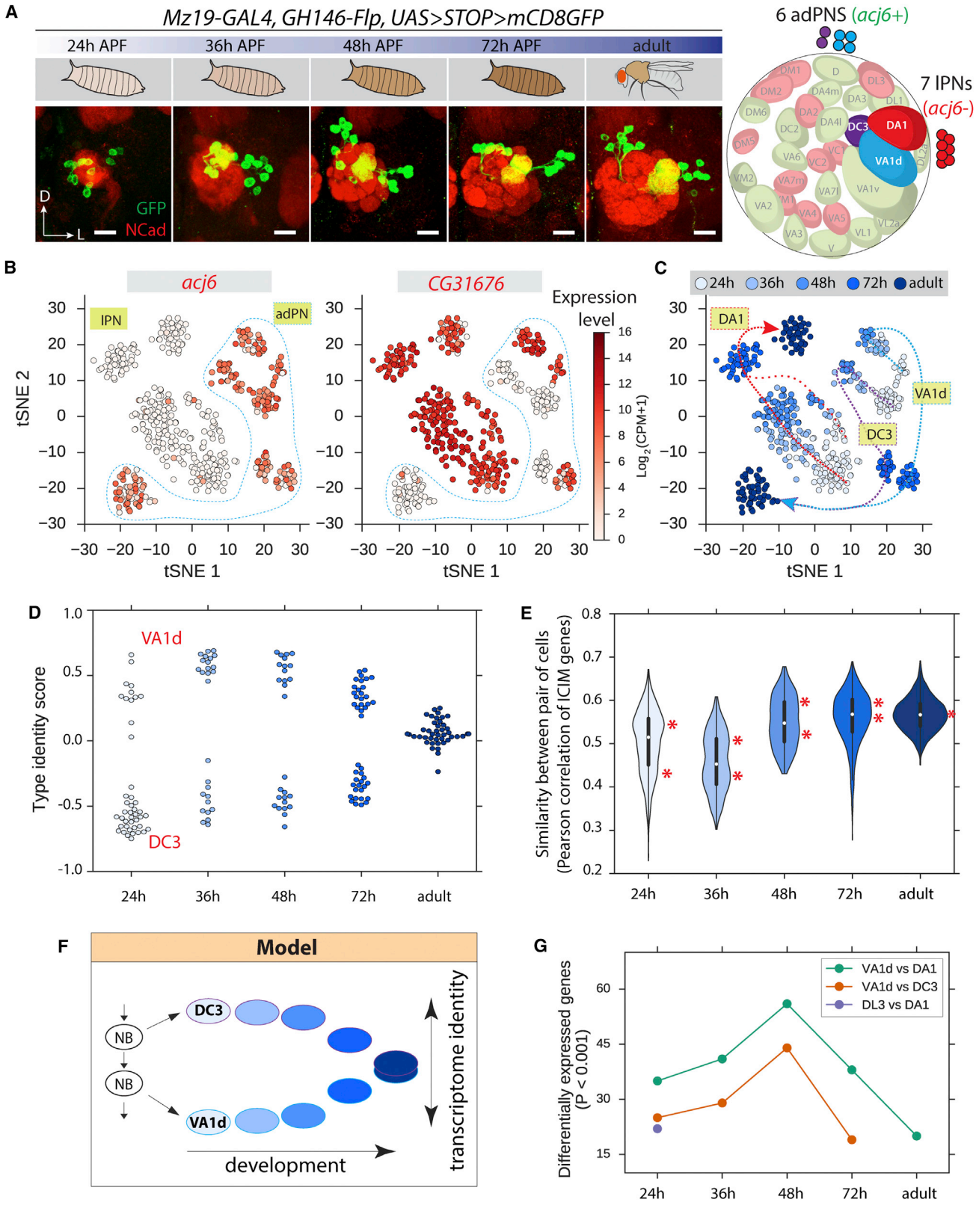
(C) Loss-of-function analysis of *C15* using *elav-GAL4* driven *UAS-C15-RNAi* (line #2; see Figure S4C). Wild-type (WT) control: *elav-GAL4* x w^{1118}. When *C15* is knocked down, the VA1d glomerulus (visualized by VA1d ORN axons labeled by *Or88a-mtdT*) displays a dorsal shift. In addition, GFP signal in VA1d PN dendrites (visualized by *Mz19-QF*-driven *QUAS-mCD8GFP*) is undetectable.

(D) Quantification of position shift of the VA1d glomerulus due to *C15* knockdown in (C). θ is the angle between the dorsoventral axis and a line drawn through the centers of the VA1d and DC3 glomeruli. Error bars are SEM. ***$p < 0.001$ (Student's t test).

(E) Gain-of-function analysis of *C15* in *Mz19-GAL4+* MARCM misexpression clones. In WT, dendrites of adPN neuroblast (adNB) clones target VA1d and DC3, and lPN neuroblast (lNB) clones target DA1. When *C15* is misexpressed, dendrite targeting of adNB clones is not affected, while dendrite targeting of lNB clones is affected with 100% penetrance.

Ncad is used as a neuropil marker. Scale, 20 μm.

See also Figure S4.



(legend on next page)

that this approach can identify gene sets that robustly encode cell-type identity in a combinatorial manner.

Applying this strategy to the 30 *GH146+* PN subtypes, we identified 11 genes whose expression states uniquely identified every PN subtype (Figures 7C and S7A). Knowledge of the expression states of these genes alone is sufficient to resolve 95% of the uncertainty in classification of individual *GH146+* PN cells into subtypes (Figure 7B, pink line). Similar results were obtained using a range of different thresholds for binarization of expression state (Figures S7B and S7C) or when we examined combinatorial codes based on gene expression levels after discretization into four states (OFF, Low, Medium, High) instead of binary states (ON/OFF) (data not shown). Indeed, a multinomial classifier using these 11 genes correctly classified 82% of individual *GH146+* PN cells into subtypes despite measurement noise (Figure S7D). Together, these analyses indicate that *GH146+* PN subtype identity can be distinguished using a combinatorial code composed of expression states of only 11 genes. This code is more compact than a code distinguishing each subtype using a unique marker (30 genes required), but substantially above the theoretical minimum of 5 genes (which can encode 2^5 or 32 binary states).

TFs and CSMs Are Highly Informative in Encoding PN Identity and Enriched among Differentially Expressed Genes

What types of genes distinguish neuronal subtypes? It is widely thought that TFs establish and maintain cell-type identity, while CSMs determine wiring specificity. But there has not been, to our knowledge, genome-wide analysis to show this in an unbiased manner. Strikingly, 8 of the 11 genes in the minimal combinatorial code identified by our information theoretic analysis above were TFs (Figure 7C), supporting a central role for TFs in specifying cell-type identity. To further explore the roles of TFs and CSMs in class identity, we searched for minimal codes for cell-type identity consisting only of TFs or CSMs, using our information theoretic approach along with previously annotated lists of TFs (FlyTF database) and CSMs (Kurusu et al., 2008), each containing ~1,000 genes. Minimal codes consisting of 13 TFs (Figures 7D and S7A) or 12 CSMs (Figures 7E and S7A) were sufficient to resolve 95% of the uncertainty in classifying *GH146+* cells into PN subtypes (Figure 7B). That is, *GH146+* PNs can be reliably classified into 30 subtypes based on the expression states of either 13 TFs alone or 12 CSMs alone. The compactness of these minimal codes was similar to that of the most compact code obtained in our genome-wide search (Figure 7C).

To evaluate whether TFs and CSMs are particularly informative with respect to subtype identity, we measured the amount of information contained within minimal combinatorial codes built from other genes (not TFs or CSMs) chosen at random from the genome (sampling 1,000 genes at random with 100 replicates). Randomly chosen genes carried significantly less information than TFs or CSMs (Figure 7B) despite having similar expression level distributions (Figure S7E). These findings indicate that, on average, TFs and CSMs carry more information about *GH146+* subtype identity than other genes.

To test this idea further, we asked whether TFs and CSMs were enriched in differentially expressed genes among PN subtypes. Among *Mz19+* adPNs and lPNs, TFs and CSMs accounted for a large proportion of differentially expressed genes (Figure 7F). Representation of CSMs peaked during the circuit assembly state (24–48 hr APF), consistent with a role for differential expression of CSMs in determining wiring specificity. We also analyzed differentially expressed genes separating every pair of 30 clusters, comprising 435 (30 × 29/2) pairs altogether. TFs and CSMs were highly enriched among differentially expressed genes, with the strongest enrichment found among the most significantly differentially expressed genes (Figure 7G). These findings support the notion that expression of TFs and CSMs plays key roles in determining PN subtype identity and wiring specificity.

DISCUSSION

Single-cell RNA-seq has recently emerged as a powerful technique to investigate cellular heterogeneity, discover new cell

Figure 6. Transcriptome Analysis of *Mz19+* PNs across Developmental Stages

(A) Representative confocal projections of *Mz19+* PNs from five stages. Schematic (right) shows cell body and glomerular targets of *Mz19+* PNs. Ncad, red. D, dorsal; L, lateral. Scale, 20 μm.

(B) Visualization of *Mz19+* PNs from all developmental stages using tSNE based on 497 genes identified using ICIM. Color shows expression of *acj6* and *CG31676* (CPM, counts per million). *acj6+* cells are adPNs (VA1d and DC3, outlined), and *acj6−* cells are lPNs (DA1). Within adPNs, *CG31676+* cells are VA1d PNs, and *CG31676−* cells are DC3 PNs. *CG31676* is turned off in all adult adPNs (see also C).

(C) Visualization of *Mz19+* PNs as in (B), with color indicating developmental stages. Expression patterns of *acj6* and *CG31676* (B) enabled unambiguous identification of three PN classes. Dashed lines indicate the developmental trajectories of these classes. VA1d and DC3 PNs are distinct at all pupal stages, but merge to form one cluster in the adult. The densely and sparsely dashed red lines represent the trajectories of cluster 2 and 2′, respectively, which become indistinguishable by 72 hr APF and remain so in the adult.

(D) Type identity score of VA1d and DC3 PNs from five developmental stages. Each dot represents a cell. Colors show developmental stages as in (C). The identity score is calculated as a scaled sum of the 22 most significantly differentially expressed genes between VA1d and DC3 PNs (STAR Methods). Scores range from −1 (high expression of the DC3 signature genes and no expression of the VA1d signature genes) to +1 (the opposite expression profile).

(E) Violin plot showing the distribution of transcriptome similarity between all pairs of *Mz19+* adPNs. Peaks are indicated by asterisks. The upper peak consists of pairs in which both cells are from the same class. The lower peak consists of pairs in which the two cells are from different classes. The adult distribution is unimodal, indicating a lack of transcriptome differences between the two classes.

(F) Schematic summary. VA1d and DC3 PNs derive from a common neuroblast (NB) lineage. Their transcriptomes are distinct at pupal stages and become indistinguishable in the adult.

(G) Differentially expressed genes between PN classes belonging either to the same lineage (VA1d and DC3; DL3 and DA1) or different lineages (VA1d and DA1). Adult data do not exist for VA1d versus DC3 PNs, as they are indistinguishable. For DL3 PNs, we only have data for 24 hr APF.

See also Figure S5 and Table S1.

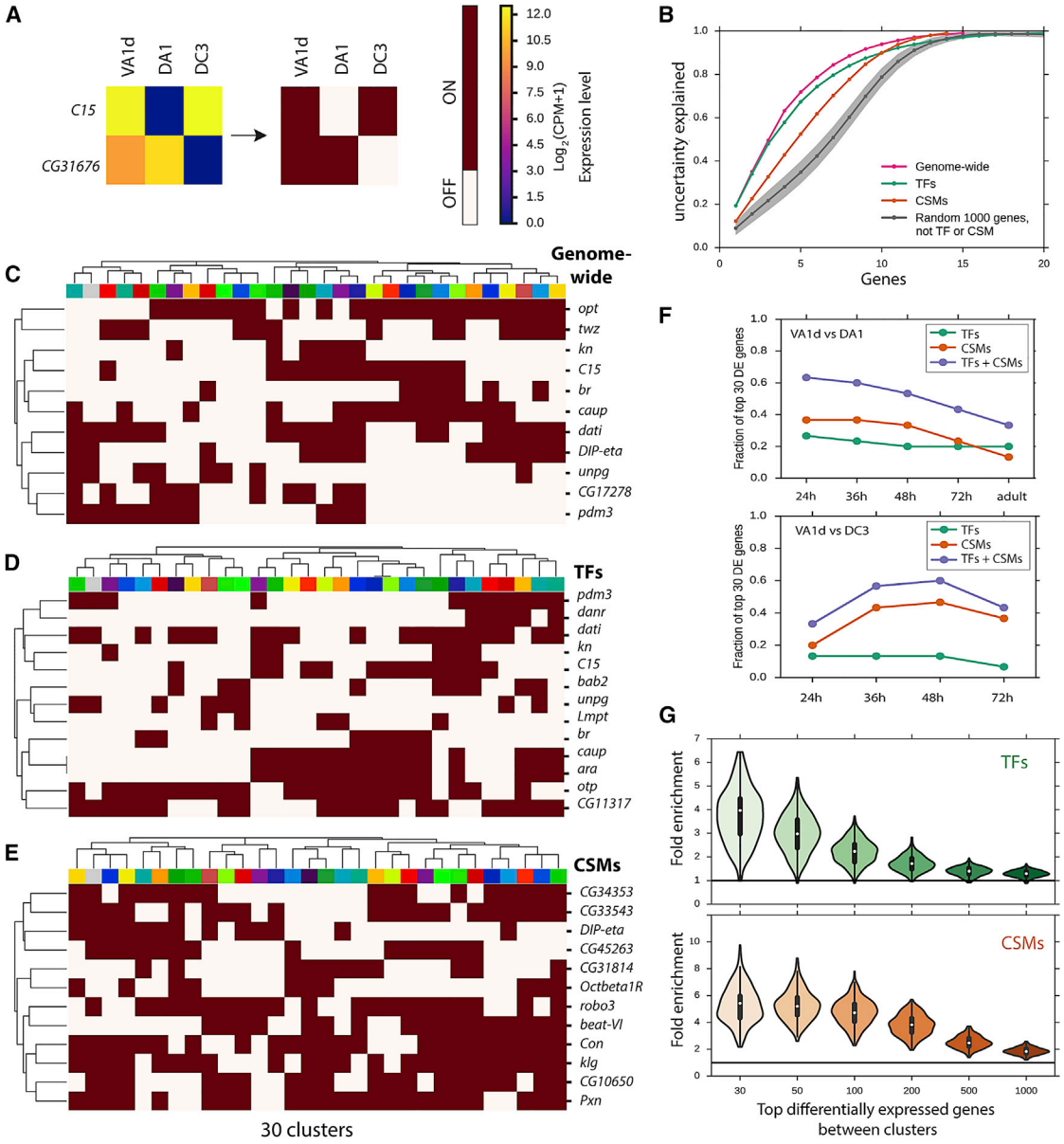


Figure 7. Combinatorial Molecular Codes of PN-Subtype Identity

(A) Minimal combinatorial code for subtype identity among *Mz19+* PNs identified using an information theoretic approach. (Left) Mean expression level of each gene among cells belonging to each *Mz19+* class. (Right) Binarized expression levels of the same genes (cutoff: $\log_2(CPM+1) = 3$). Each *Mz19+* PN class expresses a distinct combination of these two genes.

(B) Information contained in minimal combinatorial codes for *GH146+* subtype identity. x axis is the number of genes included in the code. y axis is the amount of uncertainty (entropy) of cell type classification that is explained by the code. Colors denote codes constructed from different sets of genes. The genome-wide code (pink) is constructed from all genes, while the TF (green) or CSM (orange) codes use only 1045 TF or 955 CSM genes. Gray denotes codes constructed from 1,000 randomly sampled non-TF and non-CSM genes, with the line indicating the median and the shading indicating the standard deviation across 100 replicates, respectively.

(C–E) Minimal combinatorial codes for *GH146+* subtype identity constructed from (C) all genes, (D) TFs, or (E) CSMs. Heatmap indicates the binary expression state of genes in each cluster, as in (A). Clusters and genes are arranged by hierarchical clustering.

(F) Representation of TFs and CSMs among the top 30 differentially expressed (DE) genes between pairs of *Mz19+* PN subtypes as indicated. y axis shows the fraction of the 30 most differentially expressed genes that are TFs (green), CSMs (orange), or TF + CSM (blue) at each developmental stage. Adult stage is absent from the VA1d versus DC3 comparison because their transcriptomes cannot be distinguished.

(G) Enrichment of TFs and CSMs among the top differentially expressed genes between pairs of clusters of *GH146+* cells (435 pairs). x axis shows the number of top differentially expressed genes under consideration. y axis shows the distribution of enrichment of either TFs or CSMs within these genes. Enrichment is calculated relative to the genomic representation of TFs (6.7%) and CSMs (6.2%), indicated by the horizontal line.

See also Figures S6 and S7.

types, and identify cell-type-specific markers. We established a robust single-cell RNA-seq protocol for *Drosophila* neurons and glia. By focusing on olfactory PNs, among the best-characterized cell types in all nervous systems, we established unequivocally that transcriptomic identity corresponds with the anatomical and physiological neuronal subtypes defined by connectivity and function.

Several lines of evidence support the sensitivity and reliability of our single-cell RNA-seq protocol. First, differential gene expression identified by single-cell RNA-seq is highly consistent with previous literature. We found highly correlated expression of two male-specific RNAs at the level of individual cells (Figure S1E) and mutually exclusive expression of two lineage-specific TFs (Figure S2F), as previously reported. Second, we validated five differentially expressed TFs derived from single-cell RNA-seq data (Figures 5A, 5B, S4A, and S4B). Third, sequencing of cells marked by known or newly identified PN-class-specific markers matched well with specific transcriptome clusters (Figures 3 and 4), enabling us to unequivocally match transcriptome clusters with glomerular classes. We expect that this approach can be generally applied to single-cell transcriptome analyses of many tissues and developmental stages in *Drosophila* and other organisms with small cell size, thus expanding the use of single-cell transcriptomics for addressing diverse biological questions.

We have developed a machine-learning algorithm called ICIM for unbiased identification of genes that distinguish subtypes. Because this algorithm recursively examines finer-grained subpopulations, it is capable of detecting genes that distinguish small subpopulations. ICIM is conceptually similar to previously described iterative analysis methods (Usoskin et al., 2015; Zeisel et al., 2015; Gokce et al., 2016; Tasic et al., 2016). However, ICIM may discriminate highly similar cell types with greater sensitivity than methods based on PCA would because it reduces the feature space to only those genes that are informative for distinguishing cell types. ICIM allowed us to distinguish 30 clusters for 40 *GH146+* PN classes. Our classification is limited by sampling depth because 17 classes contain only 1 cell per hemisphere (Yu et al., 2010). Sequencing of many more cells may resolve these classes into distinct clusters, resulting in a more complete description of PN transcriptome diversity.

Our analyses of transcriptome changes of identified PN classes across developmental stages demonstrate that transcriptomes of neuronal subtypes exhibit the largest difference during development, coincident with circuit assembly (Figure 6). This could be because key features of different PN classes are their input and output partners. Once PNs establish differential connectivity during development, they may use largely the same signaling machineries to convey different olfactory information in adults. This finding has important implications for using single-cell RNA-seq to classify neuronal types, since most studies have focused on adults (see Introduction). While these studies have been highly successful in classifying major neuronal types, functionally distinct subtypes may have been overlooked, resulting in an underestimate of neuronal type diversity.

TFs and CSMs are widely considered to be key determinants of cell fate and wiring specificity, respectively. Our single-cell transcriptome analyses provided objective data to support these notions. First, TFs and CSMs together account for more than 50% of the top 30 differentially expressed genes (Figures 7F and 7G). Second, information theoretic analyses revealed that among the top 11 information-rich genes for distinguishing different PN subtypes, 8 are TFs. Third, TFs or CSMs alone contain nearly as much information in distinguishing different PN subtypes as all genes contain (Figure 7B). A key readout of TFs in determining neuronal subtype may be to control differential expression of CSMs such that different subtypes differentially respond to a common extracellular environment to achieve their wiring specificity. However, we did not find a simple relationship between TFs and CSMs (Figure S7F). Supporting a role for TFs in regulating wiring specificity, we show that a newly identified lineage-specific homeobox-containing C15 can instruct lineage-specific dendrite targeting (Figure 5).

Finally, our analyses of PN transcriptomes shed light on the nature of the coding strategies that distinguish closely related neuronal subtypes. PN subtype identity is largely determined by a combinatorial code that utilizes a number of genes between the number of subtypes and the theoretical minimum for a maximally compact code, suggesting redundancy. The transcriptomes of closely related PN classes differed substantially during development (Figure 7F and Table S1), consistent with a recent report that closely related retinal cells have dozens of differentially expressed CSMs (Tan et al., 2015). A certain degree of redundancy can provide robustness to wiring precision (Hong and Luo, 2014) but creates challenges for dissecting genetic control of wiring specificity using single-gene manipulation. The transcriptomes of identified PN classes can inform design of more precise experiments in which simultaneous manipulation of multiple genes through loss- and gain-of-function approaches allows experimental testing of the combinatorial TF and CSM codes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Fly stocks
- METHOD DETAILS
  - MARCM analysis
  - Immunostaining
  - Quantitative PCR
  - Imaging and quantification procedure
  - Single-cell RNA-sequencing
  - Fly brain dissociation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Sequence alignment and preprocessing
  - Dimensionality reduction and clustering
  - Overdispersion analysis
  - Iterative Clustering for Identifying Markers
  - Differential expression analyses
  - TF and CSM lists
  - Analysis methods for figures
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental information includes seven figures and one table and can be found with this article online at https://doi.org/10.1016/j.cell.2017.10.019.

AUTHOR CONTRIBUTIONS

H.L., F.H., and L.L. designed experiments with support from S.R.Q. F.H. and H.L. established the single-cell RNA-seq protocol. H.L. performed all fly experiments with help from B.W., Q.X., J.L., T.L., and D.J.L. and prepared sequencing libraries. F.H. performed cell sorting, designed the ICIM algorithm, and analyzed data with input from H.L., L.L., and S.R.Q. H.L., F.H., S.R.Q., and L.L. wrote the paper.

ACKNOWLEDGMENTS

We thank L. Tan and S.L. Zipursky for sharing detailed protocols of cell dissociation; G. Rubin, G. Campbell, A. Moore, B. White, and Bloomington and Kyoto Stock Centers for reagents; S. Darmanis and J. Lui for discussions; N. Neff and J. Okamoto for assistance with sequencing; and T. Clandinin, J. Lui, D. Pederick, K. Shen, and A. Shuster for comments on the manuscript. H.L. is a Stanford Neuroscience Institute Interdisciplinary Postdoctoral Scholar, F.H. acknowledges support from the National Science Foundation Graduate Research Fellowship, S.R.Q. is a Chan Zuckerberg Investigator, and L.L. is an HHMI Investigator. This work was supported by NIH grant R01-DC005982 (to L.L.).

Received: June 3, 2017
Revised: August 5, 2017
Accepted: October 12, 2017
Published: November 16, 2017

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. *5*, 537–550.

Campbell, G. (2005). Regulation of gene expression in the distal region of the Drosophila leg by the Hox11 homolog, C15. Dev. Biol. *278*, 607–618.

Campello, R.J.G.B., Moulavi, D., Zimek, A., and Sander, J. (2013). A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. Data Min. Knowl. Discov. *27*, 344–371.

Cover, T.M., and Thomas, J.A. (2006). Elements of Information Theory, Second Edition (Wiley).

Cui, X., and Doe, C.Q. (1995). The role of the cell cycle and cytokinesis in regulating neuroblast sublineage gene expression in the Drosophila CNS. Development *121*, 3233–3243.

Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. USA *112*, 7285–7290.

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 Rep. *10*, 1410–1421.

Dickson, B.J. (2008). Wired for sex: the neurobiology of Drosophila mating decisions. Science *322*, 904–909.

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.

Doherty, J., Logan, M.A., Taşdemir, O.E., and Freeman, M.R. (2009). Ensheathing glia function as phagocytes in the adult Drosophila brain. J. Neurosci. *29*, 4768–4781.

Földy, C., Darmanis, S., Aoto, J., Malenka, R.C., Quake, S.R., and Südhof, T.C. (2016). Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. Proc. Natl. Acad. Sci. USA *113*, E5222–E5231.

Fuzik, J., Zeisel, A., Máté, Z., Calvigioni, D., Yanagawa, Y., Szabó, G., Linnarsson, S., and Harkany, T. (2016). Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. Nat. Biotechnol. *34*, 175–183.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gokce, O., Stanley, G.M., Treutlein, B., Neff, N.F., Camp, J.G., Malenka, R.C., Rothwell, P.E., Fuccillo, M.V., Südhof, T.C., and Quake, S.R. (2016). Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. Cell Rep. *16*, 1126–1137.

Hong, W., and Luo, L. (2014). Genetic control of wiring specificity in the fly olfactory system. Genetics *196*, 17–29.

Hong, W., Zhu, H., Potter, C.J., Barsh, G., Kurusu, M., Zinn, K., and Luo, L. (2009). Leucine-rich repeat transmembrane proteins instruct discrete dendrite targeting in an olfactory map. Nat. Neurosci. *12*, 1542–1550.

Hong, W., Mosca, T.J., and Luo, L. (2012). Teneurins instruct synaptic partner matching in an olfactory map. Nature *484*, 201–207.

Jefferis, G.S., Marin, E.C., Stocker, R.F., and Luo, L. (2001). Target neuron prespecification in the olfactory map of Drosophila. Nature *414*, 204–208.

Jefferis, G.S., Vyas, R.M., Berdnik, D., Ramaekers, A., Stocker, R.F., Tanaka, N.K., Ito, K., and Luo, L. (2004). Developmental origin of wiring specificity in the olfactory system of Drosophila. Development *131*, 117–130.

Jefferis, G.S., Potter, C.J., Chan, A.M., Marin, E.C., Rohlfing, T., Maurer, C.R., Jr., and Luo, L. (2007). Comprehensive maps of Drosophila higher olfactory centers: spatially segregated fruit and pheromone representation. Cell *128*, 1187–1203.

Jenett, A., Rubin, G.M., Ngo, T.T., 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 Rep. *2*, 991–1001.

Jinushi-Nakao, S., Arvind, R., Amikura, R., Kinameri, E., Liu, A.W., and Moore, A.W. (2007). Knot/Collier and cut control different aspects of dendrite cytoskeleton and synergize to define final arbor shape. Neuron *56*, 963–978.

Johnson, M.B., and Walsh, C.A. (2017). Cerebral cortical neuron diversity and development at single-cell resolution. Curr. Opin. Neurobiol. *42*, 9–16.

Johnson, M.B., Wang, P.P., Atabay, K.D., Murphy, E.A., Doan, R.N., Hecht, J.L., and Walsh, C.A. (2015). Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. Nat. Neurosci. *18*, 637–646.

Kazama, H., and Wilson, R.I. (2009). Origins of correlated activity in an olfactory circuit. Nat. Neurosci. *12*, 1136–1144.

Komiyama, T., and Luo, L. (2007). Intrinsic control of precise dendritic targeting by an ensemble of transcription factors. Curr. Biol. *17*, 278–285.

Komiyama, T., Johnson, W.A., Luo, L., and Jefferis, G.S. (2003). From lineage to wiring specificity. POU domain transcription factors control precise connections of Drosophila olfactory projection neurons. Cell *112*, 157–167.

Kurusu, M., Cording, A., Taniguchi, M., Menon, K., Suzuki, E., and Zinn, K. (2008). A screen of cell-surface molecules identifies leucine-rich repeat proteins as key mediators of synaptic target selection. Neuron *59*, 972–985.

Kwak, N., and Choi, C.H. (2002). Input feature selection for classification problems. IEEE Trans. Neural Netw. *13*, 143–159.

Laissue, P.P., Reiter, C., Hiesinger, P.R., Halter, S., Fischbach, K.F., and Stocker, R.F. (1999). Three-dimensional reconstruction of the antennal lobe in Drosophila melanogaster. J. Comp. Neurol. *405*, 543–552.

Lee, T., and Luo, L. (1999). Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. Neuron *22*, 451–461.

Liang, L., Li, Y., Potter, C.J., Yizhar, O., Deisseroth, K., Tsien, R.W., and Luo, L. (2013). GABAergic projection neurons route selective olfactory inputs to specific higher-order neurons. Neuron *79*, 917–931.

Lin, S., Kao, C.F., Yu, H.H., Huang, Y., and Lee, T. (2012). Lineage analysis of Drosophila lateral antennal lobe neurons reveals notch-dependent binary temporal fate decisions. PLoS Biol. *10*, e1001425.

Marin, E.C., Jefferis, G.S., Komiyama, T., Zhu, H., and Luo, L. (2002). Representation of the glomerular olfactory map in the Drosophila brain. Cell *109*, 243–255.

Meller, V.H., Wu, K.H., Roman, G., Kuroda, M.I., and Davis, R.L. (1997). roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. Cell *88*, 445–457.

Mosca, T.J., and Luo, L. (2014). Synaptic organization of the Drosophila antennal lobe and its regulation by the Teneurins. eLife *3*, e03726.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Potter, C.J., Tasic, B., Russler, E.V., Liang, L., and Luo, L. (2010). The Q system: a repressible binary system for transgene expression, lineage tracing, and mosaic analysis. Cell *141*, 536–548.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 623–656.

Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., et al. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell *166*, 1308–1323.

Sinakevitch, I., Grau, Y., Strausfeld, N.J., and Birman, S. (2010). Dynamics of glutamatergic signaling in the mushroom body of young adult Drosophila. Neural Dev. *5*, 10.

Stocker, R.F., Heimbeck, G., Gendre, N., and de Belle, J.S. (1997). Neuroblast ablation in Drosophila P[GAL4] lines reveals origins of olfactory interneurons. J. Neurobiol. *32*, 443–456.

Stork, T., Sheehan, A., Tasdemir-Yilmaz, O.E., and Freeman, M.R. (2014). Neuron-glia interactions through the Heartless FGF receptor signaling pathway mediate morphogenesis of Drosophila astrocytes. Neuron *83*, 388–403.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE *6*, e21800.

Tan, L., Zhang, K.X., Pecot, M.Y., Nagarkar-Jaiswal, S., Lee, P.T., Takemura, S.Y., 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.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346.

Tobin, W.F., Wilson, R.I., and Lee, W.A. (2017). Wiring variations that enable and constrain neural computation in a sensory microcircuit. eLife *6*.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P.V., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci. *18*, 145–153.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Vosshall, L.B., and Stocker, R.F. (2007). Molecular architecture of smell and taste in Drosophila. Annu. Rev. Neurosci. *30*, 505–533.

Ward, A., Hong, W., Favaloro, V., and Luo, L. (2015). Toll receptors instruct axon and dendrite targeting and participate in synaptic partner matching in a Drosophila olfactory circuit. Neuron *85*, 1013–1028.

Wilson, R.I. (2013). Early olfactory processing in Drosophila: mechanisms and principles. Annu. Rev. Neurosci. *36*, 217–241.

Wong, A.M., Wang, J.W., and Axel, R. (2002). Spatial representation of the glomerular map in the Drosophila protocerebrum. Cell *109*, 229–241.

Wu, J.S., and Luo, L. (2006). A protocol for dissecting Drosophila melanogaster brains for live imaging or immunostaining. Nat. Protoc. *1*, 2110–2115.

Yu, H.H., Kao, C.F., He, Y., Ding, P., Kao, J.C., and Lee, T. (2010). A complete developmental sequence of a Drosophila neuronal lineage as revealed by twin-spot MARCM. PLoS Biol. *8*.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rat anti-DNcad	Developmental Studies Hybridoma Bank	RRID: AB_528121
Chicken anti-GFP	Aves Labs	RRID: AB_10000240
Rabbit anti-DsRed	Clontech	RRID: AB_10013483
Rat anti-C15	(Campbell, 2005)	N/A
Guinea pig anti-knot	(Jinushi-Nakao et al., 2007)	N/A
Deposited Data		
Sequencing reads	This paper	GEO: GSE100058
Preprocessed sequence data	This paper	GEO: GSE100058
Experimental Models: Organisms/Strains		
D. melanogaster: *Mz19-GAL4*	Bloomington *Drosophila* Stock Center	RRID: BDSC_34497
D. melanogaster: *Alrm-GAL4*	Bloomington *Drosophila* Stock Center	RRID: BDSC_67032
D. melanogaster: *C15-RNAi #1*	Bloomington *Drosophila* Stock Center	RRID: BDSC_27649
D. melanogaster: *C15-RNAi #2*	Bloomington *Drosophila* Stock Center	RRID: BDSC_35018
D. melanogaster: *GH146-GAL4*	(Stocker et al., 1997)	RRID: BDSC_30026
D. melanogaster: *UAS-STOP-mCD8GFP*	(Potter et al., 2010)	RRID: BDSC_30125
D. melanogaster: *Mz19-QF*	(Hong et al., 2012)	RRID: BDSC_41573
D. melanogaster: *91G04-GAL4*	(Jenett et al., 2012)	RRID: BDSC_40588
D. melanogaster: *trol-GAL4*	Kyoto Stock Center	RRID: DGRC_113584
D. melanogaster: *GH146-Flp*	(Hong et al., 2009)	N/A
D. melanogaster: *unpg-lacZ*	(Cui and Doe, 1995)	N/A
D. melanogaster: *UAS-C15*	(Campbell, 2005)	N/A
D. melanogaster: *nos-Cas9*	(Diao et al., 2015)	N/A
D. melanogaster: *CG31676-GAL4*	This paper	N/A
Oligonucleotides		
Actin5C (qPCR forward primer): 5′-CTCGCCACTTGCGTTTACAGT-3′	This paper	N/A
Actin5C (qPCR reverse primer): 5′-TCCATATCGTCCCAGTTGGTC-3′	This paper	N/A
C15 (qPCR forward primer): 5′- AGCGCTTCCACAAGCAAAAG-3′	This paper	N/A
C15 (qPCR reverse primer): 5′- CCGTCTGTCGTCTCCACTTG-3′	This paper	N/A
Recombinant DNA		
Plasmid: *pT-GEM(1)*	Addgene	62893
Plasmid: *pU6-BbsI-ChiRNA*	Addgene	45946
Plasmid: *TOPO-CG31676-T2A-GAL4*	This paper	N/A
Software and Algorithms		
Custom analysis software	This paper	https://github.com/felixhorns/FlyPN
Iterative Clustering for Identifying Markers	This paper	https://github.com/felixhorns/FlyPN

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Liqun Luo (lluo@stanford.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Fly stocks

In all experiments, both male and female flies were used. The following fly lines were used in this study: *Mz19-GAL4* (BL#34497) (Jefferis et al., 2004), *C15-RNAi* (line1, BL#27649), *C15-RNAi* (line2, BL#35018), *Alrm-GAL4* (BL# 67032). *Mz19-QF* (Hong et al., 2012), *GH146-GAL4* (Stocker et al., 1997), *UAS-STOP-mCD8GFP* (Hong et al., 2009), *GH146-Flp* (Potter et al., 2010), *unpg-lacZ* (Cui and Doe, 1995), *trol-GAL4* (*NP5103-GAL4*, Kyoto Stock Center #113584), *UAS-C15* (gift from Dr. Gerard Campbell) (Campbell, 2005), and *91G04-GAL4* (gift from Gerry Rubin) (Jenett et al., 2012).

CG31676-GAL4 was generated using CRISPR/Cas9 based insertion of *SA-T2A-GAL4* into the first intron of *CG31676* gene following the method described by Diao et al. (2015). In brief, a 2.5kb DNA fragment, containing a PAM site in the middle within the first intron of the *CG31676* gene, was PCR amplified from wild-type genomic DNA, and inserted into Blunt TOPO vector (Invitrogen). Then, *SA-T2A-GAL4* was PCR amplified from the *pT-GEM(1)* plasmid (Addgene #62893) and was inserted (NEBuilder HiFi DNA assemble kit) into the TOPO-*CG31676*-intron construct three nucleotides before the PAM site of the intron. This construct and a gRNA plasmid (*pU6-BbsI-ChiRNA*, Addgene #45946) containing a 20-nt target sequence upstream of the PAM inserted into the BbsI site were co-injected to *nos-Cas9* (gift from Dr. Ben White) (Diao et al., 2015) embryos to obtain transgenic flies.

METHOD DETAILS

MARCM analysis

hsFlp based MARCM analyses were performed as previously described (Lee and Luo, 1999; Jefferis et al., 2001). Briefly, transgenic flies linked with a *FRT* chromosome were crossed with MARCM-ready flies (containing *hsFlp, UAS-CD8GFP, Mz19-GAL4, TubP-GAL80* and desired *FRT*). *Mz19-GAL4* was used to label VA1d, DC3 and DA1 PNs. Larvae (24h to 48h after hatching) from the cross were heat shocked for 1h in a 37°C water bath. Both single-cell and neuroblast clones could be observed in this fashion.

Immunostaining

Tissue dissection and immunostaining were performed following previously described methods (Wu and Luo, 2006). Briefly, fly pupal and adult brains were dissected in 1x PBS and then fixed in 4% paraformaldehyde (20% paraformaldehyde diluted in PBS with 0.015% Triton X-100) for 20 min at room temperature. Fixed brains were washed three times with PBST (PBS with 0.3% Triton X-100) and incubated in PBST twice, each time for 20 min. The samples were incubated in blocking buffer (5% normal goat serum in PBST) for 30 min at room temperature or overnight at 4°C. Then, primary antibodies diluted in blocking buffer were applied and samples were incubated for 24–48 h at 4°C. Then, samples were washed with PBST for 20 min twice, and secondary antibodies diluted in blocking buffer were applied and samples were incubated in the dark for > 24 h at 4°C. Samples were washed with PBST for 20 min twice and mounting solution (Slow Fade Gold) was added. Samples were left in mounting solution for at least 1 h before mounting them onto glass slides. All wash steps were performed at room temperature. Primary antibodies used in this study include rat anti-DNcad (DN-Ex #8; 1:40; DSHB), mouse anti-Prospero (1:200; DSHB), mouse anti-Cut (2B10; 1:50; DSHB), mouse anti-β-gal (1:500; Promega), chicken anti-GFP (1:1000; Aves Labs), rabbit anti-DsRed (1:250; Clontech), mouse anti-ratCD2 (OX-34; 1:200; AbD Serotec), rat anti-C15 (1:200; gift from Dr. Gerard Campbell) (Campbell, 2005), and guinea pig anti-knot (1:200; gift from Dr. Adrian Moore) (Jinushi-Nakao et al., 2007). Secondary antibodies were raised in goat or donkey against rabbit, mouse, rat, and chicken antisera (Jackson Immunoresearch), conjugated to Alexa 405, 488, FITC, Cy3, Cy5, or Alexa 647.

Quantitative PCR

Total RNA from 3–5 day old adult fly heads was extracted using a MiniPrep kit (Zymo Research, R1054). Complementary DNA was synthesized using an oligo-dT primer. qPCR was performed on a Bio-Rad CFX96 detection system. Relative expression was normalized to *Actin5C*. Primer sequences used for qPCR were:

Actin5C (F): 5′-CTCGCCACTTGCGTTTACAGT-3′
Actin5C (R): 5′-TCCATATCGTCCCAGTTGGTC-3′
C15 (F): 5′- AGCGCTTCCACAAGCAAAAG-3′
C15 (R): 5′- CCGTCTGTCGTCTCCACTTG-3′

Imaging and quantification procedure

Confocal images were collected with a Zeiss LSM 780 and processed with ImageJ and Adobe Illustrator. For quantification of the angles in Figure 5C, the vertical line was drawn based on the position of two antennal lobes and the intersecting line was drawn through the centers of gravity of the VA1d and DA1 glomeruli, then the intervening angle was measured using ImageJ.

Single-cell RNA-sequencing

Drosophila brains with mCD8GFP-labeled cells using specific GAL4 drivers were manually dissected, and optic lobes were removed. Single-cell suspensions were prepared following Tan et al. (2015) with several modifications (see detailed procedure below). Single labeled cells were sorted via Fluorescence Activated Cell Sorting (FACS) into individual wells of 96-well plates containing lysis buffer

using an SH800 instrument (Sony Biotechnology). Full-length poly(A)-tailed RNA was reverse-transcribed and amplified by PCR following the SMART-seq2 protocol (Picelli et al., 2014) with several modifications. To increase cDNA yield and detection efficiency, we increased the number of PCR cycles to 25. To reduce the amount of primer dimer PCR artifacts, we digested the reverse-transcribed first-strand cDNA using lambda exonuclease (New England Biolabs) (37°C for 30 min) prior to PCR amplification. Sequencing libraries were prepared from amplified cDNA using tagmentation (Nextera XT). Sequencing was performed using the Illumina Nextseq 500 platform with paired-end 75 bp reads.

Fly brain dissociation

Before the fly brain dissociation, make sure following essential reagents and supplies are readily available: Schneider's medium (Thermo Fisher, 21720024), papain (Worthington PAP2, LK003178), liberase TM (Roche, 5401119001; liberase TM was reconstituted in 1x sterile PBS on ice to get final concentration 2.5mg/ml and aliquots of 20ul were stored at –20°C), Falcon tubes with cell strainer (35um), microfuge tube shaker, and syringes with 25G 5/8 needles.

Make fresh papain solution for every dissociation experiment. Get 1 vial of papain (Worthington PAP2, LK003178), dissolve the powder using 1x sterile PBS (final concentration 100 units/ml), and re-suspend papain by gently shaking the vial (if use pipet, avoid bubbles which will decrease the enzyme activity). Then make aliquots of 300 μL papain solution for each EP tubes, and activate it in 37°C water bath for 10-30 min. Add 4.1 μL of liberase TM solution (2.5mg/ml) into 300 μL papain solution to obtain a final concentration of about 0.18 units/ml. Cool down the solution to room temperature before adding it to brain samples. Since it takes 10-30 min to activate the papain solution, coordinate this step with fly brain dissection.

To estimate how many brains are required during sample preparation, please refer to recovery rates in Figure S1C for calculation. In our current study, we dissected about 120 pupal brains from *GH146-GAL4,UAS-mCD8GFP* flies to collect about 10 plates of cells (10*96). For the rare population labeled by *91G04-GAL4* or *trol-GAL4* (2-3 cells each hemisphere), we dissected about 200 pupal brains to get half plate of cells (~50).

To prepare a single-cell suspension:

Dissect pupal/adult fly brains in ice-cold Schneider's medium. Remove the optical lobes if all desired cells are in the central brain (e.g., *GH146+* PNs). Transfer every brain using P20 pipet into the EP tube containing 500ul Schneider's medium and keep the tube on ice. The brains can be in the cold Schneider's medium for up to 2 hours before the next step.

After dissecting a sufficient number of brains, spin them down using bench-top microfuge for 10 s and remove Schneider's medium. Wash brains for 3 times at room temperature with 1x sterile PBS to completely remove Schneider's medium.

Add 300 μL papain solution (37°C activated and 4.1ul liberase added) to each sample and incubate it in a microfuge tube shaker (25°C, 1,000rpm) for 20 min in total. At 5 and 10 min time points, pipet the solution up and down 30 times (avoiding bubbles), and then continue shaking. At 15 min time point, pass solution through 25G 5/8 needles for 7X (avoiding bubbles). Shake the tube for another 5 min. To increase yield, use spare papain solution to coat the tips/needles before passing the brain sample.

Inactivate the enzyme by adding 400 μL cold Schneider's medium (total 700ul). Filter solution through cell strainer (35 μm) into a 5ml falcon tube (keep tapping the tube until the solution go through the filter). Wash the EP tube with 800 μL cold Schneider's medium and filter through cell strainer (total 1,500 μl).

Transfer the 1,500 μL solution to an EP tube and centrifuge for 7min at 4°C, 600xg. Discard supernatant, re-suspend cells with 1,000 μL (or desired volume depending on cell density) Schneider's medium, and transfer it to 5ml FACS tube. Add desired florescent dye (e.g., Ethidium homodimer-1, Initrogen L3224, as a dead cell marker) and keep the tube on ice until FACS sorting.

QUANTIFICATION AND STATISTICAL ANALYSIS

For RNA-seq data analysis, we first provide an overview of our methods, then describe how these methods were applied to create each figure. All analysis was performed in Python using Numpy, Scipy, Pandas, scikit-learn, and a custom single-cell RNA-seq module. Sequencing reads and preprocessed sequence data are freely available from the Gene Expression Omnibus (accession number GEO: GSE100058). Code is freely available from Github (https://github.com/felixhorns/FlyPN).

Sequence alignment and preprocessing

Reads were aligned to the *Drosophila melanogaster* genome (r6.10) using STAR (2.4.2a) (Dobin et al., 2013) with the ENCODE standard options, except "–outFilterScoreMinOverLread 0.4–outFilterMatchNminOverLread 0.4–outFilterMismatchNmax 999–outFilterMismatchNoverLmax 0.04." Uniquely mapped reads that overlap with genes were counted using HTSeq-count (0.7.1) (Anders et al., 2015) with default settings except "-m intersection-strict." Cells having fewer than 300,000 uniquely mapped reads were removed. To normalize for differences in sequencing depth across individual cells, we rescaled gene counts to counts per million (CPM). All analyses were performed after converting gene counts to logarithmic space via the transformation Log_2(CPM+1). Cells that were labeled with neuron-specific GAL4 drivers (*GH146+, Mz19+, 91G04+*, and *Trol+*) were filtered for expression of canonical neuronal genes (*elav, brp, Syt1, nSyb, CadN,* and *mCD8GFP*), retaining only those cells that expressed at least 4/6 genes at > 15 CPM. After filtering, 97.3% of *GH146+* PN cells express *mCD8GFP* (at > 15 CPM).

Dimensionality reduction and clustering

Single-cell RNA-seq yields high dimensional gene expression data. To visualize and interpret these data, we obtained two-dimensional projections of the cell population by first reducing the dimensionality of the gene expression matrix using principal component analysis (PCA), then further reducing the dimensionality of these components using t-distributed Stochastic Neighbor Embedding (tSNE) (van der Maaten and Hinton, 2008). We note that tSNE is a nonlinear embedding that does not preserve distances, so one cannot directly interpret the distances in the projected space directly as distances between gene expression profiles (i.e., in the pre-transformation space).

We performed PCA on a reduced gene expression matrix composed of the top 500 overdispersed genes (as described below). To identify significant principal components (PCs), we examined the distribution of eigenvalues obtained by performing PCA after shuffling the gene expression matrix (with 100 replicates). A PC was considered significant if the magnitude of its associated eigenvalue exceeded the maximum magnitude of eigenvalues observed in the shuffled data. Significant components (typically 7–12 PCs) were used for further analysis. We further reduced these components using tSNE to project them into a two-dimensional space.

ICIM (see detailed description of ICIM below) is an unsupervised machine learning algorithm that identifies a set of genes which distinguishes transcriptome clusters, which may correspond to cell types (described below). In our analysis of *GH146+* PNs, this set typically includes ~500 genes. To visualize and interpret the single-cell gene expression data, we further reduced its dimensionality using tSNE to project the reduced gene expression matrix (consisting of only the genes identified by ICIM) into a two-dimensional space.

Overdispersion analysis

Genes that are highly variable within a population often carry important information for distinguishing cell types. We were interested in identifying such genes and using them for dimensionality reduction and clustering analyses. Variability of gene expression depends strongly on the mean expression level of a gene. This motivates the use of a metric called dispersion, which measures the variability of a gene's expression level in comparison with other genes that are expressed at a similar level. Overdispersed genes are those that display higher variability than expected based on their mean expression level.

To identify overdispersed genes, we binned genes into 20 bins based on their mean expression across all cells. We then calculated a log-transformed Fano factor D(x) of each gene x

$$D(x) = \log_{10}\left[\sigma^2(x)/\mu(x)\right]$$

where $\sigma^2(x)$ is the variance and $\mu(x)$ is the mean of the expression level of the gene across cells. Finally, we calculated the dispersion d(x) as the Z-score of the Fano factor within its bin

$$d(x) = (D(x) - \mathrm{Mean}[D(x)])/\mathrm{Std}[D(x)]$$

where Mean[D(x)] is the mean log-transformed Fano factor within the bin and Std[D(x)] is the standard deviation of the log-transformed Fano factor within the bin. We then ranked genes by their dispersion and selected the top genes for downstream analysis.

Iterative Clustering for Identifying Markers

To identify subpopulations of cells corresponding to PN subtypes, we developed an unsupervised machine-learning algorithm, which we call Iterative Clustering for Identifying Markers (ICIM). We observed that standard dimensionality reduction and clustering methods using PCA and tSNE failed to discriminate subpopulations that corresponded to known PN lineages and molecular features. We attributed the failure of these methods to the high degree of similarity of transcriptional states among PN subtypes, which represent closely related neurons having similar functions. All PN subtypes are born from one of the two common progenitor cells (neuroblasts) and have similar functional roles in the adult fly. Thus, PN subtypes may be distinguished by a small number of genes.

In the language of machine learning, the performance of dimensionality reduction and clustering methods depends critically on feature selection. Selection of informative genes that vary among cell types can improve discrimination in dimensionality reduction and clustering analysis.

We developed ICIM as a strategy to identify the most informative genes for distinguishing subpopulations within a population of closely related cells in an unbiased way. Starting with a population of cells, we first identify the top 100 overdispersed genes within this population. Next we expand this set of genes by finding genes whose expression profiles are strongly correlated with the overdispersed genes (Pearson correlation > 0.5). We also filter this set of genes by (1) removing those having < 2 correlated partners, and (2) those that are expressed in > 80% of cells. Filter (1) removes noisy genes based on the idea that genes that carry information about cell type are expressed within gene modules and therefore have expression profiles that are correlated with other genes. Filter (2) removes housekeeping genes that are detected in nearly all cells, and have variation in expression levels due to biological and technical noise, but this variation is not informative for purposes of distinguishing cell types. Cells are then clustered based on their expression profiles of these genes (average-linkage clustering using correlation metric). We cut the dendrogram at the deepest branch and partition the population into two subpopulations. The same steps are then performed iteratively on each subpopulation. Iteration continues until a population cannot be split into subpopulations because it is "homogeneous." The termination condition is defined as the minimum terminal branch length (the most similar nearest-neighbor correlation distance between the expression profiles of

cells) being larger than 0.2. This condition arises when the algorithm attempts to discover genes within a homogeneous population and finds a very large number of genes (typically > 1000 genes) that vary in an incoherent manner between cells. When the algorithm terminates, we collect all genes that were identified at any stage. The result of this analysis is a set of genes that discriminate subpopulations within a population, which can be used for dimensionality reduction (as described above). We note that this algorithm identifies informative genes in an unbiased manner without knowledge of the ground truth of the number of cell types and their differences. The results of the algorithm were robust across a wide range of parameters.

Why does ICIM outperform previously used approaches, such as PCA? PCA reduces the feature space in a manner that assigns weights to genes based on their information content. This has two consequences: (1) downstream analysis uses the weighted gene expression information, which imposes assumptions about the statistical relationships between genes, and (2) while less informative genes are assigned smaller weights, they nevertheless can contribute to downstream analysis. In contrast, ICIM explicitly removes genes that are deemed uninformative from further consideration and assigns equal weights to those that are kept. These attributes make ICIM a more effective feature selection strategy for analysis of highly similar cellular subtypes.

Differential expression analyses

To find differentially expressed genes, we used the Mann-Whitney U test, a non-parametric test that detects differences in the level of gene expression between two populations. The Mann-Whitney U test is advantageous for this application because it makes very general assumptions: (1) observations from both groups are independent and (2) the gene expression levels are ordinal (i.e., can be ranked). Thus the test applies to distributions of gene expression levels across cells, which rarely follow a normal distribution. Using the Mann-Whitney U test, we compared the distributions of expression levels of every gene separately. p values were adjusted using the Bonferroni correction for multiple testing. Different significance thresholds for determining whether a gene is differentially expressed were used for various analyses in this work.

TF and CSM lists

To identify genes that are transcription factors (TFs) or cell surface molecules (CSMs), we used manually curated lists. We obtained a list of *Drosophila* TFs from the FlyTF v1 database, (http://www.mrc-lmb.cam.ac.uk/genomes/FlyTF) and CSMs from (Kurusu et al., 2008). These lists were manually curated to remove spurious annotations and redundancies according to Flybase annotation, resulting in 1045 TFs and 955 CSMs.

Analysis methods for figures

Single-cell transcriptome analyses of neurons and glia in Figure 1

We formed a population consisting of 946 *GH146-GAL4+* cells and 67 *alrm-GAL4+* cells. We performed dimensionality reduction and clustering analysis using PCA and tSNE as described above. We identified the top 500 overdispersed genes in the population. We used PCA to reduce dimensionality, retaining 7 significant PCs. Then we projected the population into a two-dimensional space using tSNE with perplexity 30 and learning rate 500 (Figure 1E). We also performed hierarchical clustering using complete linkage and a Euclidean metric based on manually selected neuronal and glial marker genes (Figure 1D).

Removal of GH146+ vPNs and APL neurons in Figure 2

We initially formed a population consisting of 946 *GH146+* cells. Using ICIM, we identified 158 genes that distinguish subtypes. We then projected the population into a two-dimensional space using tSNE. We observed several distinct subpopulations corresponding to *GH146+* neuronal types that do not belong to the adPN or lPN lineages. Specifically, two clusters were composed of ventral PNs (vPNs), which robustly express several specific markers (*Gad1, Lim1,* and *toy*). Three other clusters were composed of APL neurons, which robustly express other specific markers (*Wnt4, VGlut,* and *fd102C*), and are arranged adjacent to one another by tSNE, reflecting the similarity of expression profiles among these cells. For subsequent analyses of *GH146+* adPNs and lPNs, we removed these cells by excluding cells expressing 2/3 of these marker genes at > 15 CPM.

Single-cell transcriptome analyses of GH146+ cells in Figure 2

We initially attempted to identify distinct subpopulations representing PN subtypes using PCA and tSNE for the 946 *GH146+* PNs (including vPNs and APL neurons). We began by identifying the top 500 overdispersed genes and performing PCA to reduce the gene expression data to 10 significant PCs. Then we projected the population into a two-dimensional space using tSNE with perplexity 30 and learning rate 500. We observed that this analysis fails to separate distinct subpopulations (Figure 2B).

We next attempted to distinguish subpopulations corresponding to PN subtypes using ICIM and tSNE. Using ICIM (Figure 2C), we identified 561 genes for the 902 *GH146+* PNs (representing adPNs and lPNs, after removing vPNs and APL neurons as described above). We projected these cells into a two-dimensional space using tSNE using as a distance matrix the pairwise Pearson correlation of the expression profiles of these genes, and perplexity 10, learning rate 250, and early exaggeration 4.0 (Figure 2D). Because tSNE computes a nonlinear embedding that does not preserve distances in the original space, the distances between cells cannot be directly interpreted in terms of similarity of expression profiles. As a consequence, there are cases where cells belonging to the same cluster are separated by larger distances than cells belonging to different clusters. We classified cells into clusters in an unbiased manner using HDBSCAN with min_cluster_size = 5 and min_samples = 3 on coordinates after tSNE projection.

Mapping clusters to PN classes in Figure 3

We formed a population consisting of 902 *GH146+* cells, 123 *Mz19+* cells (at 24h APF), and 23 *91G04+* cells. Using the 561 genes identified using ICIM on *GH146+* cells, we projected this population into a two-dimensional space using tSNE with perplexity 15 and learning rate 1000. For visualization, we colored the cells according to their genotype, revealing that the *Mz19+* and *91G04+* cells belong exclusively to 5 clusters (Figure 3C).

Mapping clusters to PN classes in Figure 4

We formed a population consisting of 902 *GH146+* cells and 41 *trol-GAL4+* cells. *trol-GAL4+* cells that were not expressing *trol* (CPM < 7) were removed, leaving 28 cells for further analysis. Using the 561 genes identified using ICIM on *GH146+* cells, we projected this population into a two-dimensional space using tSNE with perplexity 15 and learning rate 1000. For visualization, we colored the cells according to their genotype, revealing that the vast majority of *trol+* cells belong exclusively to 1 cluster (Figure 4C).

Analysis of transcriptome changes during development in Figure 6

To understand how transcriptional state changes during development and maturation of PN subtypes, we collected *Mz19+* cells from flies at 5 stages of development: 24h, 36h, 48h, and 72h after puparium formation (APF), and 1–2 day adults. We formed a population consisting of 485 cells (123, 83, 92, 92, and 95 cells at each stage, respectively), after filtering to remove low quality cells and those not expressing neuronal markers (as described above). Using ICIM, we identified 497 genes that distinguish cell subtypes and developmental stages. We projected this population into a two-dimensional space based on these genes using tSNE with perplexity 10 and learning rate 500. Cells formed several distinct subpopulations corresponding to different PN subtypes and developmental stages (Figures 6B and 6C). We assigned subpopulations to subtypes (DA1, VA1d, and DC3) based on the expression of key lineage factors (Figure 6B).

To quantify transcriptome changes in the closely related PN subtypes VA1d and DC3 during development, we devised a metric called the type identity score, which is the scaled sum of expression levels of genes that distinguish VA1d and DC3 cells. We identified these genes using differential expression analysis comparing VA1d and DC3 populations at all times that the two populations are distinct as determined by ICIM and tSNE (24h, 36h, 48h, and 72h APF). 78 cells were included in the VA1d group and 64 cells were included in the DC3 group. This analysis yielded 22 genes of which 13 are highly expressed in VA1d and 9 are highly expressed in DC3 cells at a significance level of $p < 10^{-5}$ after the Bonferroni adjustment for multiple testing. We rescaled expression levels of these genes to the range 0 to 1 (by dividing each expression level by the maximum among the population), then calculated the type identity score I of each cell as the mean normalized expression level,

$$I = \frac{1}{|X_{VA1d}|} \sum_{x \in X_{VA1d}} x - \frac{1}{|X_{DC3}|} \sum_{x \in X_{DC3}} x,$$

where X_{VA1d} is the set of genes that are highly expressed in VA1d and X_{DC3} is the set of genes that are highly expressed in DC3, and $|X|$ is the cardinality of set X. We then plotted the type identity scores of each cell at each developmental stage (Figure 6D).

As an alternative method to analyze transcriptome differences, we also examined correlations in transcriptome states in an unbiased genome-wide manner. This method has the advantage that it does not require the choice of a p value cutoff for determining significance. We formed a population consisting of *Mz19+* adPNs (belonging to both subtypes VA1d and DC3) at each stage of development. Then we calculated the Pearson correlation of the expression profiles of the 497 genes identified by ICIM for every pair of cells (Figure 6E). These plots revealed a bimodal distribution containing two distinct peaks, corresponding to pairs of cells that both belong to the same subtype (more similar peak) and pairs of cells which belong to different subtypes (less similar peak). As development proceeds, the transcriptome similarity of these two subtypes diminishes until vanishing, as reflected in the merging of these two distinct peaks and the emergence of a unimodal distribution in adulthood.

To compare transcriptome differences between neuroblast lineages, we performed differential expression analysis comparing VA1d and DC3, VA1d and DA1, and DL3 and DA1 PNs across developmental stages. Because the p values of a differential expression test depend strongly on the number of cells involved in the test, we sampled cells so that all populations had the same number of cells. Specifically, for each comparison, we sampled 12 cells from each population without replacement and performed differential expression analysis, then repeated this procedure (100 replicates) and calculated the median p value across the replicates for each gene. We then counted the number of differentially expressed genes at $p < 0.001$ based on the median p value (Figure 6G). All cells used for this analysis were *Mz19+* PNs, except DL3 cells were *GH146+* PNs.

To characterize transcriptome changes distinguishing PNs in the wiring stages of development from PNs in adulthood, we performed differential expression analysis comparing the population of *Mz19+* PNs at 24h APF to the population of *Mz19+* PNs in adults. We found 1097 differentially expressed genes at significance level of $p < 10^{-5}$ after the Bonferroni adjustment for multiple testing. This included 592 genes that were highly expressed in 24h APF cells, and 478 genes that were highly expressed in adult cells. We performed Gene Ontology (GO) analysis on these genes using Flymine and removed the redundant GO terms using REVIGO (Supek et al., 2011). We report the number of genes corresponding to and the p value of enrichment of each term (Figure S5A).

To identify transcriptional waves during *Mz19+* PN development, we considered the 1097 genes that were differentially expressed between 24h APF and adult cells. We calculated the median expression of each gene at each time point. We normalized these median expression values by dividing by the maximum value across time points, such that each expression value became a relative expression level between 0 and 1. We then performed dimensional reduction on the expression profiles of the genes using TSNE with

perplexity 20, learning rate 1000, and early exaggeration 6.0. We identified clusters among the genes using HDBSCAN with min_cluster_size 25 and min_samples 5 on the projected coordinates. This resulted in the identification of 13 distinct waves, of which 6 involved upregulation and 7 involved downregulation from 24h APF to adult cells (Figure S5B). We plotted the mean relative expression level of each gene at each time point (black dots connected by black lines). The relative expression profile of each individual gene belonging to each wave was also plotted (gray lines). We calculated the fraction of genes within each wave that are transcription factors (TFs) or cell surface molecules (CSMs) using the lists of TFs and CSMs that were obtained as described above.

Characterizing genes that distinguish PN subtypes in Figure 7

To identify genes that distinguish closely related PN subtypes, we performed differential expression analysis comparing the *Mz19+* VA1d and DA1 populations and the *Mz19+* VA1d and DC3 populations at each developmental stage. These genes are presented in Table S1. Because the significance level of expression level differences depends on the number of cells involved in the comparison and the number of cells varies across developmental stages, we analyzed the top 30 differentially expressed genes regardless of their significance level. We note that the significance values of these genes were nearly all $p < 10^{-4}$. We calculated the fraction of genes within each wave that are transcription factors (TFs) or cell surface molecules (CSMs) using the lists of TFs and CSMs that were obtained as described above (Figure 7F).

We next performed a similar differential expression analysis comparing all pairs of subtypes. For each subtype, we formed a population consisting of *GH146+* cells belonging to that subtype at 24h APF. For each pair of subtypes we calculated differential expression for each gene and ranked the genes by their significance level (p value). We then calculated the fraction of TFs or CSMs among the top N genes for varying N from 30 to 1000. We calculated the enrichment of TFs or CSMs compared to their genomic representation by dividing the fraction of TFs or CSMs by the genomic fraction of TFs (6.7%) or CSMs (6.2%). We plotted the distribution of these enrichment values for various values of N (Figures 7G).

Searching for unique marker genes for PN subtypes in Figure 7

We sought to identify unique markers for each *GH146+* PN subtype. We formed populations each consisting of *GH146+* cells belonging to a cluster identified using ICIM and tSNE (Figure 2D). We then performed differential expression analysis comparing each cluster to all other *GH146+* cells. We selected genes that were differentially expressed at a significance level of $p < 0.05$ after the Bonferroni adjustment for multiple testing and having median expression within the cluster of interest of > 7 CPM, resulting in 1103 genes. We then filtered for genes that were identified as significantly enriched in only one cluster, resulting in 257 genes. This step was necessary because some genes were identified as significantly enriched in multiple clusters, which is consistent with reuse of genes as identity factors within a combinatorial code. Finally, we identified genes that were robust and unique markers for a single cluster. To do this, we calculated the fraction of cells within each cluster expressing a given gene at > 7 CPM. We then filtered for genes that were expressed in > 50% of the cells within a given cluster and in < 10% of the cells in any other cluster. We plotted the distribution of expression levels of these genes in each cluster (Figure S6A). We also attempted to search using less strict criteria. For example, Figure S6B shows the result when we required that the gene is expressed in > 50% of the cells (> 7 CPM) within a given cluster and in < 25% of the cells in any other cluster.

Technical artifacts such as dropouts can hinder the identification of unique markers. We therefore estimated the probability that our failure to identify unique markers can be accounted for by dropout. To do so, we assumed that each of the 30 molecularly distinct *GH146+* PN subtypes expresses a single unique marker gene at a low level of expression (7 CPM) in a ubiquitous fashion (i.e., in all the individual cells belonging to that subtype). In our data, genes expressed at an average level of 7 CPM are not detected in ~60% of cells. We can fail to detect a gene because (1) the cell is not expressing the gene, or (2) because of noise in gene expression (biological noise) or technical dropout (measurement noise). We therefore can estimate an upper bound at 60% on the probability of dropout of a gene that is expressed at 7 CPM on average. Our approach for identifying unique markers requires that the gene is detected in 50% of cells within a cluster. The probability of failure to detect the marker gene for a given cluster due to dropout is therefore given by the probability of dropouts in 50% of the cells in a cluster. This probability is:

$$P_{failure} = P_{dropout}^{N_{dropouts}}$$

where $P_{dropout}$ is estimated to be 60% and $N_{dropouts} = 0.5 * N_{cells}$ is the number of cells in which the gene must drop out. We calculated this probability based on the number of cells N_{cells} in every cluster, ranging from 5 to 108. Then we calculated the probability that 25 out of 30 clusters do not have a unique marker gene by multiplying the probabilities of failure in 25 randomly sampled clusters. We performed this sampling 10,000 times and report the average ($p = 10^{-139}$). This value represents the probability of failing to detect a unique marker gene for 25 out of 30 clusters given that each cluster is expressing a single unique marker at 7 CPM on average.

These calculations are conservative in several ways. First, we assumed that each cluster expresses a single marker gene. Realistically, each subtype may express multiple unique markers. This would increase the probability of detecting at least one of them. Second, unique markers may be expressed at levels higher than 7 CPM. We observed that the unique marker genes that we discovered are expressed at levels well above 7 CPM (Figure S6A) and biologically it is unlikely that a type identity factor would be expressed at extremely low levels. Thus, we estimate that the likelihood that our failure to detect marker genes can be explained by dropouts alone is very small.

Information theory-based analyses in Figure 7

We sought to identify minimal sets of genes that can encode the subtype identity of *GH146+* PNs in a combinatorial fashion. Our motivation was to determine by direct search whether such molecular combinatorial codes exist.

To address this, we devised an algorithm that finds a minimal set of genes that is sufficient to encode the subtype identity of cells in a combinatorial manner drawing upon ideas from information theory. An introduction to information theory is outside the scope of this work. Nevertheless, we provide a brief description of the basic concepts. Then we describe the algorithm and how it was applied in this work.

Entropy measures the uncertainty of a random variable (Shannon, 1948; Cover and Thomas 2006). Conditional entropy measures the uncertainty of a random variable given the knowledge of another variable (i.e., after conditioning on another variable). Conditioning on data never increases uncertainty (on average), which agrees with our intuition that additional information never hurts.

We use the notion of entropy H(C) to describe the uncertainty of cell type classification C. We use conditional entropy to describe the reduced uncertainty in classification due to knowledge of the expression state of a gene, H(C|G). The information gain due to knowledge of the expression state of the gene is the mutual information between the gene and the classification I(G;C), which can be defined as

$$I(G;C) = H(C) - H(C|G),$$

where H(C) is the entropy of cell type classification (without knowledge of the expression states of any genes) and H(C|G) is the entropy of cell type classification after conditioning on the expression state of gene G. Mutual information I(G;C) describes how much our uncertainty about classification C decreases when we observe the gene G.

Mutual information I(G;C) can also be calculated directly from the probability distributions of cell type classes and expression states. For two discrete random variables G and C with their joint probability density function (pdf) p(x,y), the mutual information of G and C is defined as

$$I(G;C) = \sum_{g}\sum_{c}\frac{p(g,c)\log p(g,c)}{p(c)p(g)}$$
$$= H(C) - H(C|G).$$

We often calculate the information content of a gene G with respect to the cell type classification C using this equation. Throughout this work, the base of the logarithm is 2 and so the unit of entropy and information is bit.

We now describe the algorithm for finding a minimal combinatorial code. This problem is closely related to the Feature Reduction n-k (FRn-k) problem in machine learning (Battiti, 1994). To solve this problem, we employ a greedy algorithm using mutual information similar to that described in (Kwak and Choi, 2002). The problem is formulated as follows:

Given an initial set F with n features and set C of all output classes, find the subset $S \subseteq F$ with k features that minimizes H(C|S), which is equivalent to maximizing the mutual information I(C;S).

Our algorithm is as follows:

(1) Initialize S as the empty set and F as the initial set of n features.
(2) For all f_i in F, compute $I(C;f_i)$.
(3) Find the feature f_i that maximizes $I(C; f_i)$. Add f_i to set S. Remove f_i from set F.
(4) Repeat until the desired number of features k is selected:
  (4.1) For all f_i in F, compute $I(C;f_i|S)$.
  (4.2) Find the feature f_i that maximizes $I(C; f_i|S)$. Add f_i to set S. Remove f_i from set F.
(5) Return the set S containing the selected features.

We repeat this computation with increasing k until the output set S explains a chosen amount of uncertainty in the classification C. Typically, we choose this termination condition as 99% of the entropy H(C) of classification C.

We note that the computation of mutual information is dramatically more efficient when G and C are discrete. We therefore binarized expression levels using a cutoff of $\log_2(CPM+1) = 3$. This cutoff was chosen based on the minimum in the distribution of expression levels across all genes and all cells (Figure S7A). We varied the cutoff value between 2 and 6 and found that our results were essentially unchanged. The compactness of the minimal codes for *GH146+* PN subtype identity and the genes included in the code were nearly identical to that obtained using the cutoff of 3 (data not shown). We also calculated the correlation between the information carried by each gene under different values of the cutoff with the information carried under the cutoff of 3 (Figure S7B). This analysis revealed that the information content of genes is not very sensitive to the precise choice of cutoff for binarization across the range of 2 to 6. We also found that other discretization schemes, such as a different number of levels of expression (e.g., OFF, Low, Medium, High), yielded similar results (data not shown).

Combinatorial coding of PN subtype identity in Figure 7

We initially applied the information theory-based algorithm to *Mz19+* PN cells to test whether it is capable of identifying a set of genes that is sufficient for a combinatorial code of cell type identity. We formed a population consisting of the 175 *GH146+* cells that belong

to the classes labeled by *Mz19-GAL4* (108 DA1, 35 VA1d, and 32 DC3 cells). We created a binary expression matrix consisting of the ON/OFF states of all 15,522 genes that remained after removing genes that were not detected in any cells. We calculated the mutual information of each gene with respect to the *GH146+* subtype classification. We then used the greedy algorithm described above to find a minimal set of genes for encoding *GH146+* subtype identity (Figure 7A). The initial set of features F was the top 30 most informative genes among the 15,522 genes in the expression matrix.

We next applied this approach to all *GH146+* PNs. We formed a population consisting of the 902 *GH146+* cells belonging to the adPN and lPN lineages (Figure 2D). We created a binary expression matrix consisting of the ON/OFF states of all 15,522 genes that remain after removing genes that are not detected in any cells. We calculated the mutual information of each gene with respect to the *GH146+* subtype classification (Figure 2D). We used the greedy algorithm described above to find minimal sets of genes for encoding *GH146+* subtype identity with k varying from 1 to 20. The initial set of features F was the top 30 most informative genes among all 15,522 genes (genome-wide), among the 1045 TFs, or among the 955 CSMs. We plotted the uncertainty explained by the minimal codes obtained with each value of k (Figure 7B). We then chose minimal codes that explained 95% of the uncertainty of *GH146+* subtype classification (Figures 7C–E). For plotting, we binarized the mean expression level of each gene in each cluster using the cutoff of $\log_2(\text{CPM}+1) = 3$ (Figures 7C–E).

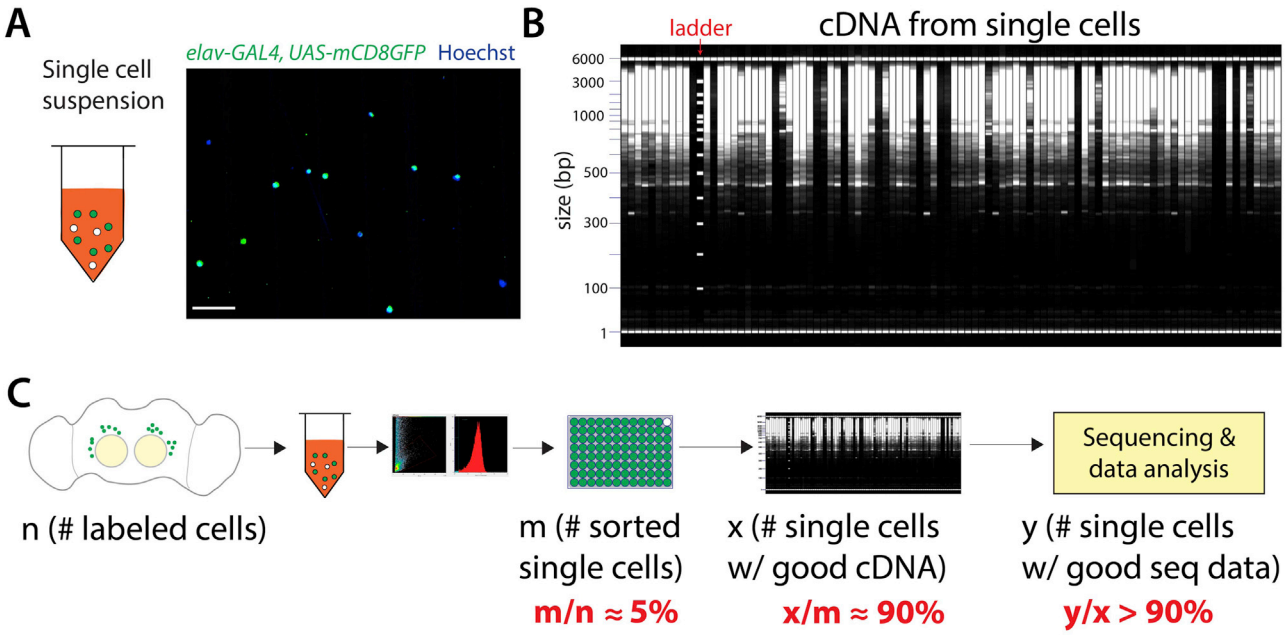
To evaluate the performance of classifiers using these minimal codes, we performed leave-one-out cross-validation. We formed a training set consisting of 901 *GH146+* cells, after leaving out a single cell. We then searched for a minimal combinatorial code for subtype identity using these cells and chose a minimal code that explained 95% of the uncertainty in subtype classification. We then performed multinomial classification of the test cell based on its expression states of the genes in the code. Specifically, the predicted class of the test cell was the class having the minimum Hamming distance to the expression state of the cell. In the event of a tie, we assumed that the classifier would make a random, uniformly weighted choice between the tied classes. Performance was plotted as a confusion matrix (Figure S7D), which depicts the fraction of cases having each true label that are classified as each predicted label.

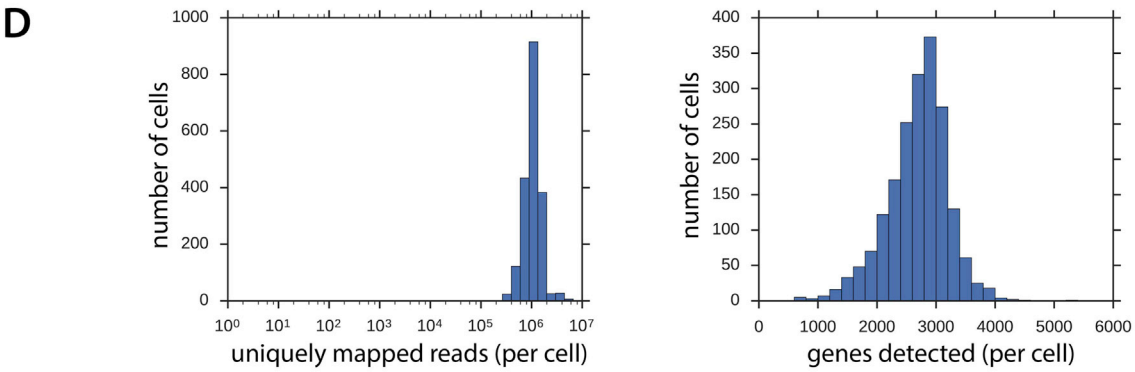
To evaluate whether TFs and CSMs carry more information than other genes, we found minimal sets of genes using an initial set of features F consisting of 1,000 genes chosen at random from among the 13,631 expressed in the genome after excluding the genes annotated as TFs and CSMs. We performed this search with 100 replicates. We plotted the mean uncertainty explained at various values of k and the standard deviation across the replicates (Figure 7B). We evaluated the distribution of expression levels of TFs, CSMs, and other genes by calculating the median expression of each gene in all *GH146+* PN cells, and plotting the distribution across all genes in each category (Figure S7E).

To identify regulatory relationships between TFs and CSMs, we performed clustering of the expression state profiles of TFs and CSMs across the clusters of *GH146+* cells. The top 30 TFs and CSMs ranked by mutual information with cell type identity were selected. The binary expression state of each gene in each cluster was calculated using the cutoff of $\log_2(\text{CPM}+1) = 3$ based on the mean expression level in the cluster. We performed average linkage clustering using the Hamming distance metric on these expression states (Figure S7F).

DATA AND SOFTWARE AVAILABILITY

Sequencing reads and preprocessed sequence data are freely available from the Gene Expression Omnibus (GEO: GSE100058). Code is freely available from Github (https://github.com/felixhorns/FlyPN).





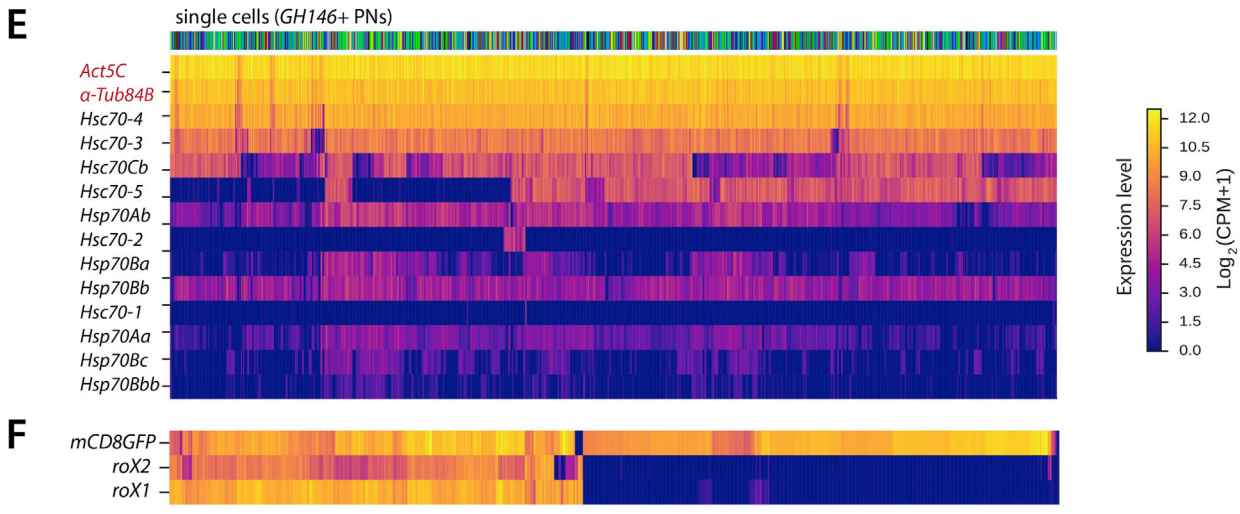


Figure S1. Single-Cell RNA-Seq Protocol for the *Drosophila* Pupal Brain, Related to Figure 1

(A) Image of single-cell suspension after brain dissociation. Pupal brains were dissected and dissociated. Sample was imaged using epifluorescence microscopy. DNA was stained using Hoescht 33342 (blue). Scale bar, 50 μm.

(legend continued on next page)

(B) Representative image of cDNA size distribution for 96 wells as measured using the Fragment Analyzer automated capillary electrophoresis system (Advanced Analytical).
(C) Summary of efficiency for key steps of the single-cell RNA-seq protocol based on PNs. We note that the reliability of generating full-length cDNA material from single glia (30%) was not as high as from single neurons (90%), suggesting that individual glia may contain fewer mRNA molecules, or dissociation procedure may cause more damage to glia than neurons.
(D) Distributions of the number of uniquely mapped reads (left) and genes detected (right) per cell. On average, 1 million reads mapped uniquely to the *Drosophila* genome per cell, and 3000 genes were detected (CPM > 3).
(E) Heatmap showing expression of housekeeping genes, *Act5c* and α-*Tub84B*, and stress-related *Hsp70* superfamily genes in individual cells. *GH146-GAL4+* cells robustly express housekeeping genes, but stress-related genes are not widely induced, supporting the faithfulness of our RNA-seq measurement. Colors above columns indicate the cluster assignment of each cell, using color code shown in Figure 2D, revealing that *Hsp70*-related genes do not drive clustering.
(F) Heatmap showing co-expression of the male-specific genes *roX1* and *roX2* in individual cells.
In (E) and (F), each column is one cell. Cells were ordered using hierarchical clustering.

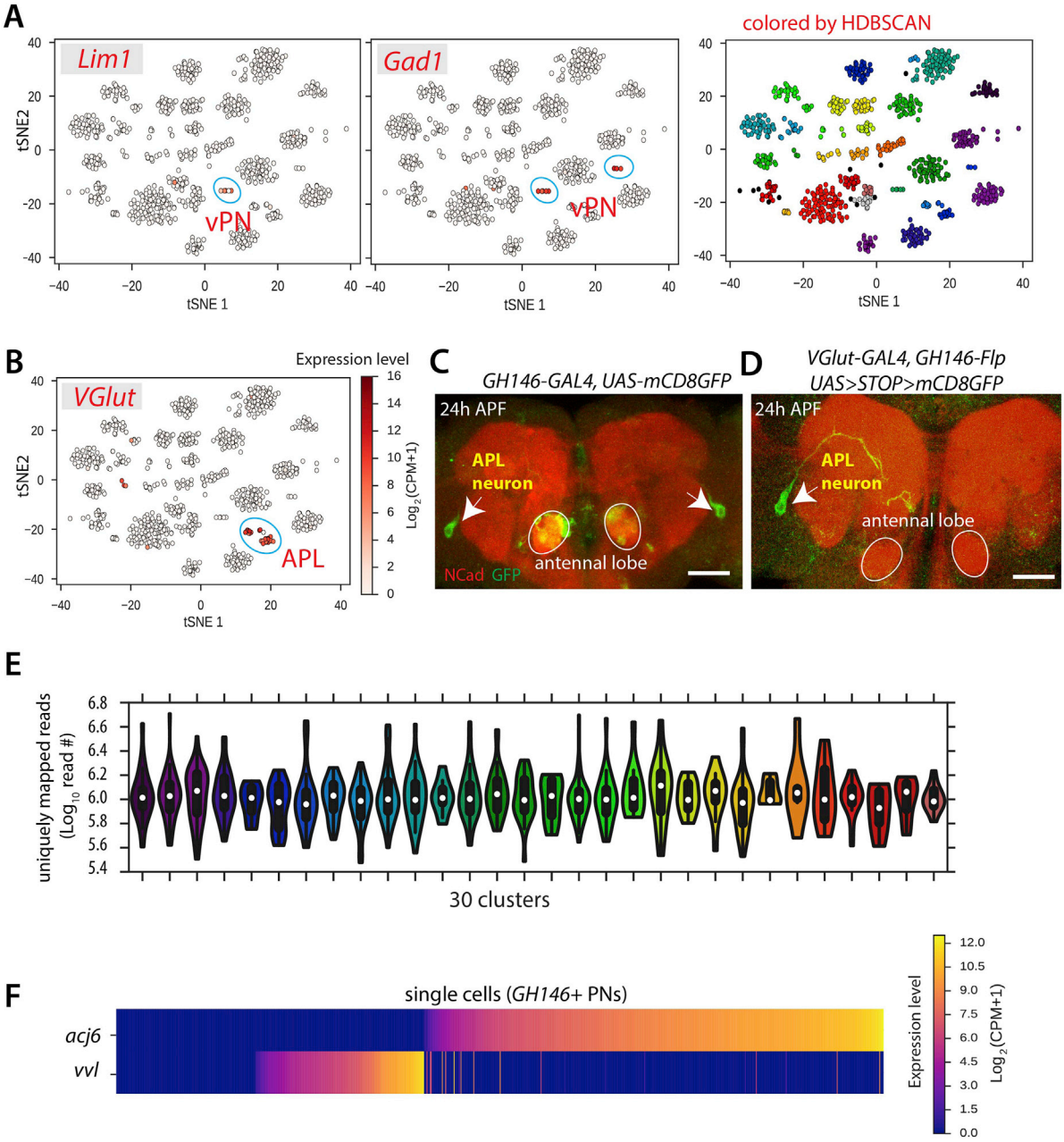


Figure S2. Single-Cell RNA-Seq Analysis of *GH146+* PNs, Related to Figure 2

(A) Visualization of *GH146+* PN cells using tSNE based on 158 genes identified using ICIM. Each dot is one cell. Cells are arranged according to similarity of expression profiles of the selected genes. Cells are colored by expression levels of *Lim1* (left), *Gad1* (middle) (see B for color bar), and by cluster identity as determined using HDBSCAN, which is a hierarchical density-based clustering algorithm (right). Two distinct clusters express *Gad1*, one of which also expresses *Lim1*; both genes are unique to vPNs, indicating that these clusters correspond to *GH146+* vPNs.

(B) Visualization of *GH146+* PN cells as in Figure S2A with cells colored according to expression of *VGlut* (CPM, counts per million). Three adjacent clusters express *VGlut* (outlined), which is a unique marker for anterior paired lateral (APL) neurons (Figure S2C and Figure S2D), indicating that these clusters correspond to APL neurons.

(C) Confocal images showing that APL neurons (indicated by arrows) are labeled by *GH146-GAL4* driven *UAS-mCD8GFP*.

(D) Confocal image showing that APL neurons (arrow) are labeled by *VGlut-GAL4* (after intersecting with *GH146-Flp*). Ncad staining labels neuropil (red), and antennal lobes are outlined.

In (C) and (D), scale bar, 50 μm.

(E) Distribution of sequencing depth across *GH146+* cells belonging to each cluster.

(F) Heatmap showing expression of the lineage-specific transcription factors *acj6* and *vvl* in *GH146+* adPN and lPN cells. Cells are ordered by *acj6* expression, then *vvl* expression.

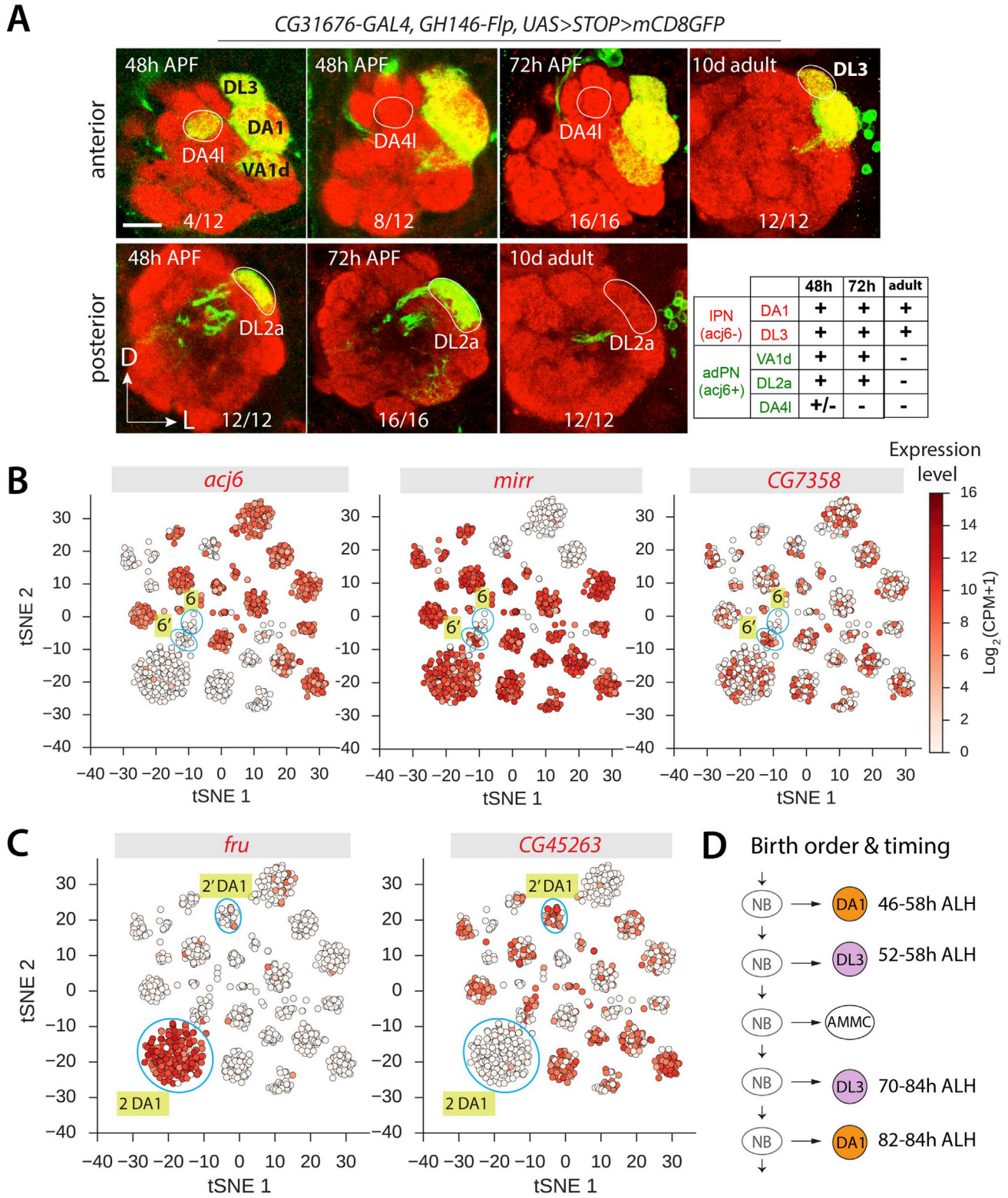


Figure S3. Mapping Clusters to PN Classes Using Newly Identified Markers, Related to Figure 4

(A) Systematic characterization of *CG31676-GAL4* expression in PNs after intersecting with *GH146-Flp* at 48h and 72h APF, and 10d adult. Expression patterns are summarized in the table: +, expressed; –, not expressed; +/–, expressed in a subset of flies. *CG31676-GAL4* stably labels DA1 and DL3 from pupa to adult.

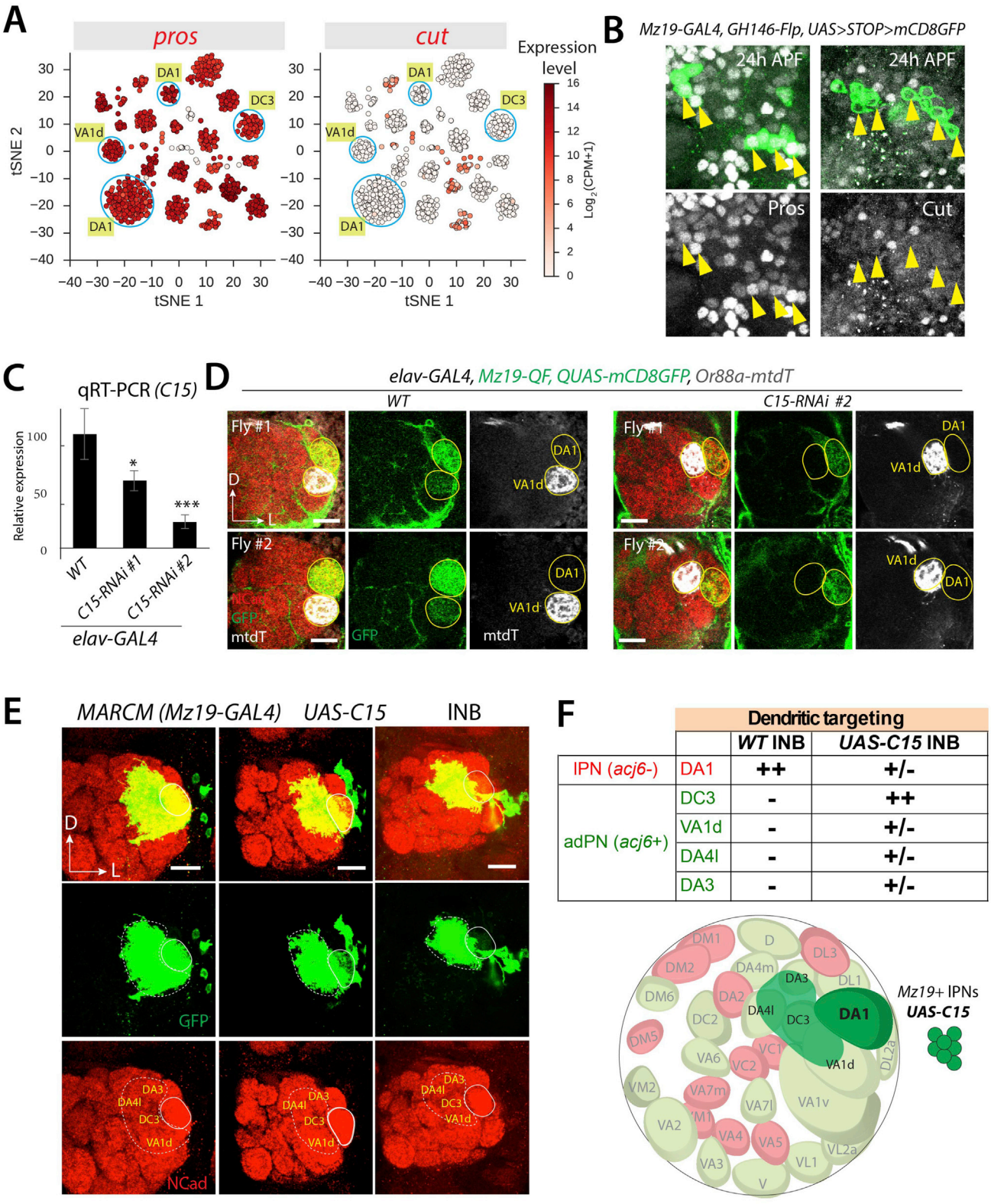
(legend continued on next page)

Note that *CG31676-GAL4* also transiently labels DL2a and DA4l adPNs, but we could not unambiguously map them to corresponding clusters. Ncad staining labels neuropil (red). Scale bar, 20 μm.

(B) Visualization of *GH146*+ PN cells using tSNE as in Figure 4A showing expression levels of *acj6*, *mirr*, and *CG7358* (CPM, counts per million). Clusters #6 and #6' are both *acj6*–. Cluster #6', but not Cluster #6, is *mirr+* and *CG7358+*.

(C) Visualization of *GH146*+ PN cells using tSNE as in Figure 4A showing expression levels of *fru* and *CG45263* (see color bar in [B]). *fru* is expressed in Cluster #2, but not Cluster #2′, while *CG45263* is expressed in Cluster #2′, but not Cluster #2. Both Clusters #2 and #2′ map to DA1 PNs.

(D) Schematic summary of birth order and timing of the lateral neuroblast (NB) lineage. Both DA1 and DL3 PNs are born in two different periods, separated by antennal mechanosensory and motor center (AMMC) neurons (Lin et al., 2012). ALH, after larval hatching.



		Dendritic targeting	
		WT INB	*UAS-C15* INB
IPN (*acj6-*)	DA1	++	+/-
adPN (*acj6+*)	DC3	-	++
	VA1d	-	+/-
	DA4l	-	+/-
	DA3	-	+/-

Figure S4. Validation of Transcription Factor Expression Patterns, Related to Figure 5

(A) Visualization of *GH146+* PN cells using tSNE as in Figure 4A showing expression of *prospero* (*pros*) and *cut* (*ct*). *pros* is expressed in *Mz19+* PNs, while *ct* is not.

(B) Antibody staining shows that *Mz19+* PNs (arrowheads) express Pros but not Cut at 24 APF, consistent with RNA-seq data, as shown in (A).

(legend continued on next page)

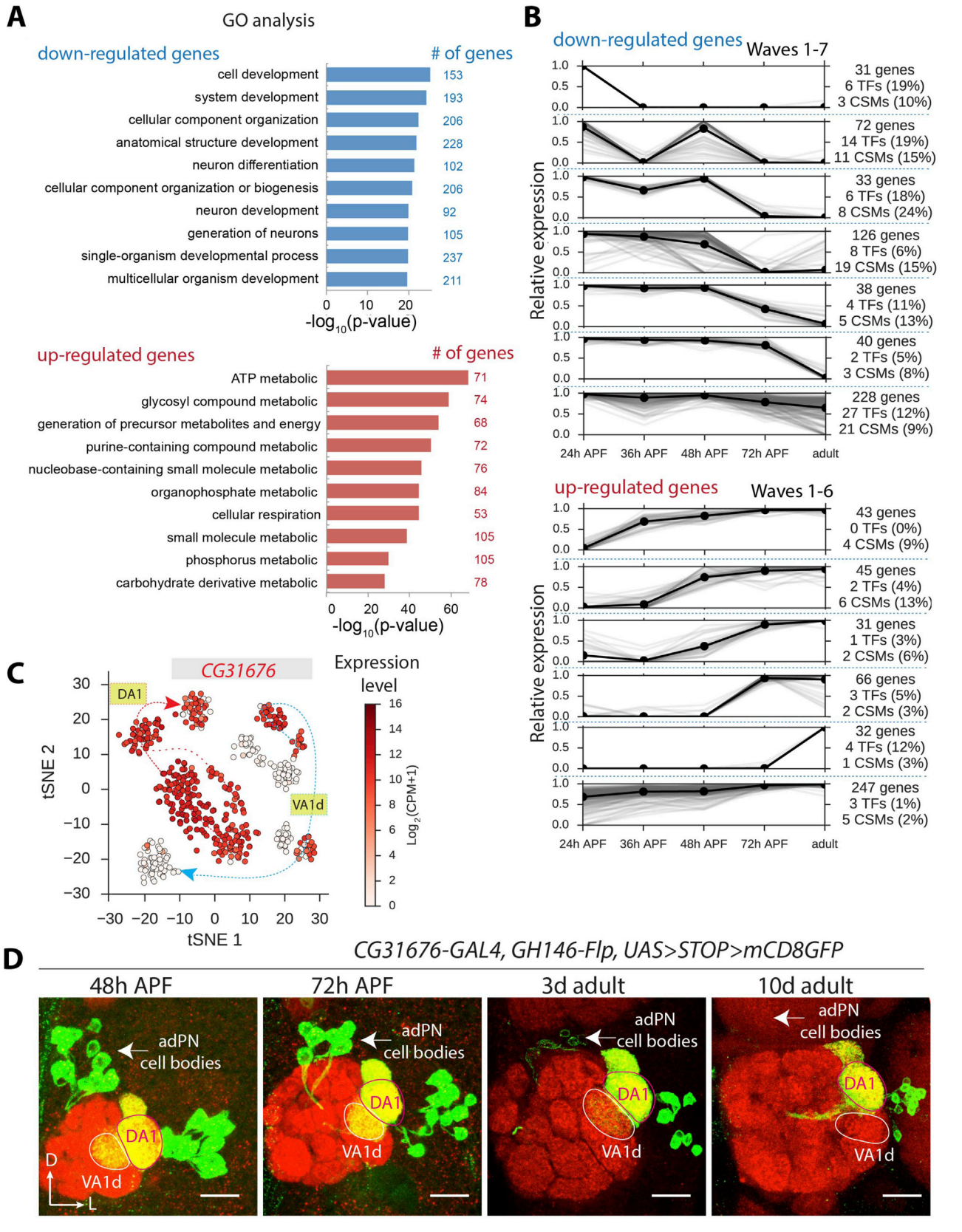
(C) Quantitative PCR (qPCR) measurement of the knockdown efficiency of two *UAS-C15-RNAi* lines. *elav-GAL4* was crossed with either w^{1118} (control) or two *C15-RNAi* lines, and mRNA was extracted from 5-day-old adult fly heads (n = 3 replicates of 10 heads pooled for each condition). Expression levels are normalized to *actin5C*. Error bars show SEM. *p < 0.05; ***p < 0.001 (t test).

(D) Two additional examples of dendrite targeting of *Mz19-QF+* PNs in *WT* and *C15* knockdown, as in Figure 5C. DA1 and VA1d glomeruli are outlined in yellow.

(E) Three additional examples of gain-of-function analysis of *C15* misexpression in *Mz19-GAL4+* MARCM lateral neuroblast (lNB) clones, as in Figure 5E. Mistargeted regions are outlined and corresponding glomeruli are indicated.

(F) Summary of mistargeting phenotypes for *UAS-C15* lNB clones. ++, fully targeted; –, not targeted; +/–, partially targeted. Lower panel is a schematic of mistargeting phenotypes. Note that all mistargeted glomeruli are normally innervated by adPNs.

Ncad staining labels neuropil (red). Scale bar, 20 μm.



(legend on next page)

Figure S5. Analysis of *Mz19+* PN Development and Maturation, Related to Figure 6

(A) Gene Ontology (GO) analysis of genes that were up- or downregulated between 24h APF and adult in *Mz19+* PNs ($p < 10^{-5}$). For the top 10 most significantly enriched GO terms, the significance of enrichment and the number of genes corresponding to each term are shown.

(B) Expression dynamics of the genes identified in (A) spanning the 24h, 36h, 48h, and 72h APF, and adult stages. Genes were classified based on their dynamical profiles (STAR Methods), revealing 7 distinct dynamical patterns of expression among the downregulated genes (waves 1–7) and 6 such patterns among up-regulated genes (waves 1–6). For individual genes, the median expression level at each developmental stage is shown in light gray (normalized to maximum expression across developmental stages). Black line shows the mean expression profile across the genes assigned to a wave. For each wave, the number and fraction of genes that were TFs and CSMs are indicated.

(C) Visualization of *Mz19+* PN cells from developmental stages ranging from 24h APF to adult as in Figure 6B showing expression of *CG31676* (CPM, counts per million). In DA1 PNs, *CG31676* is expressed at all stages. In VA1d PNs, *CG31676* is expressed at all pupal stages, but not in adult.

(D) Confocal images (anterior stacks) showing *CG31676-GAL4* expression after intersecting with *GH146-Flp*, at various pupal and adult stages. DA1 and VA1d glomeruli are outlined and adPN cell bodies are indicated (arrow). Consistent with our RNA-seq data (C), *CG31676* is expressed at all time points in DA1 PNs, while in VA1d PNs *CG31676* is expressed at all pupal stages and then turned off in 10-day-old adult flies. We observed very weak expression in 3-day-old adult flies, likely due to perdurance of mCD8GFP. Ncad staining labels neuropil (red). Scale bar, 20 μm.

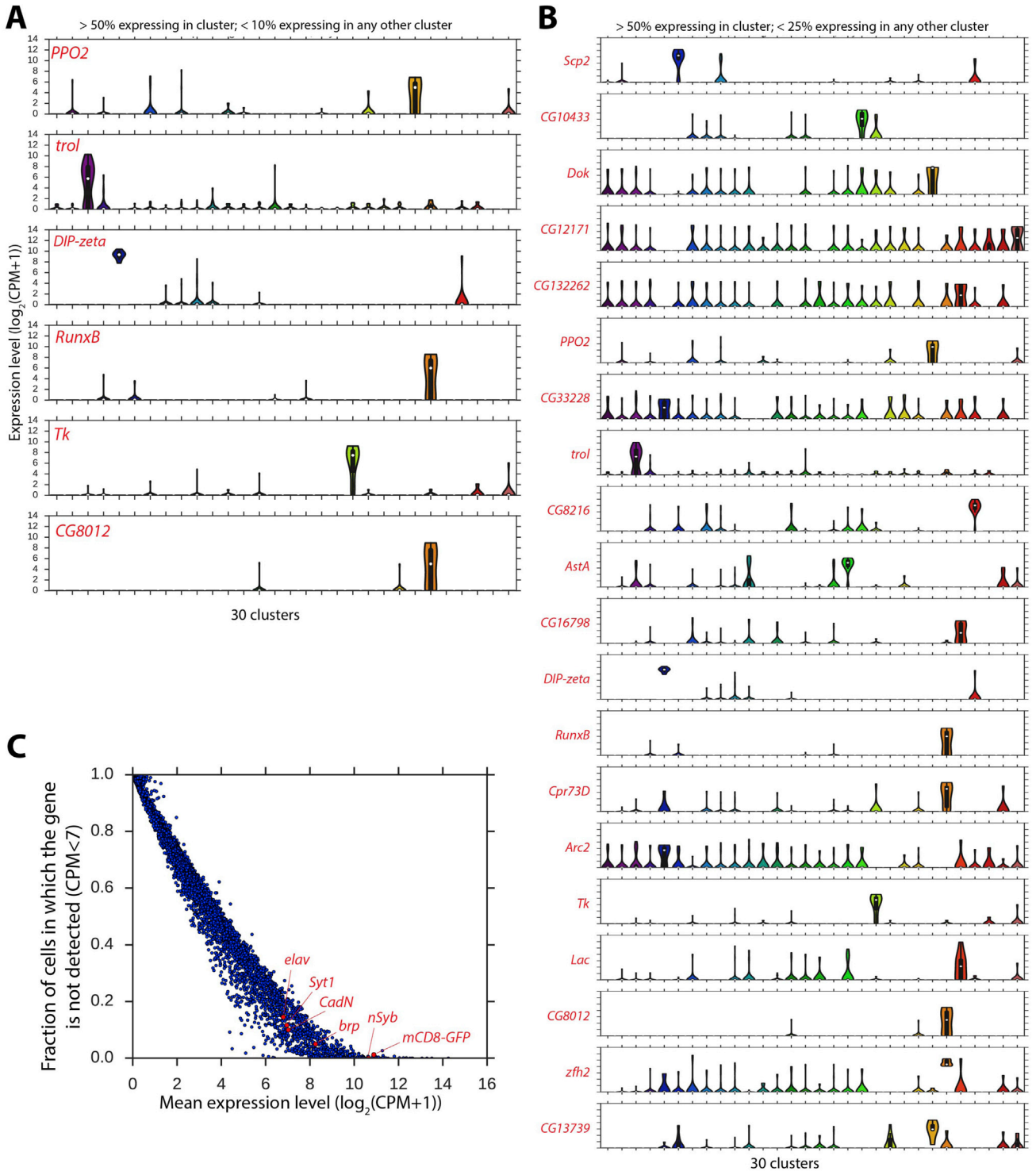


Figure S6. Searching for Unique Markers of Individual PN Subtypes, Related to Figure 7

(A) Violin plots showing expression of genes identified as unique marker genes for single *GH146+* PN clusters. These 6 genes were identified using the criteria: (1) > 50% cells within a cluster express the gene, and (2) < 10% cells in any other cluster express the gene. Expression was defined as > 7 CPM, or $\log_2$(CPM+1) > 3. These genes specify 5 distinct PN clusters.

(B) Violin plots showing expression of genes that were identified as unique markers using less stringent criteria than (A). The criteria used here were: (1) > 50% of the cells in the cluster express the gene, and (2) < 25% of the cells in any other cluster express the gene. 20 genes were found, but many of these genes are clearly

(legend continued on next page)

not unique to a single cluster. This indicates that a search with relaxed stringency yields many genes which are not in fact unique markers. See (A) for scale of expression level, which is common to all plots.

(C) Relationship between mean expression level and detection failure. Each dot is a gene. Detection was defined as > 7 CPM, or $\log_2$(CPM+1) > 3. Detection failure events can occur because either (1) the gene is not expressed in the cell, or (2) failure to detect expression of the gene despite the presence of mRNA transcripts due to technical artifact (called dropouts). Thus, the fraction of detection failure events provides an upper bound on dropout rate. We used this upper bound to calculate the probability that we failed to detect unique markers for 25 PN clusters due to dropout alone (STAR Methods). *mCD8GFP* and the 5 neuronal markers are indicated; they were used for quality filtering (STAR Methods) and shown in Figure 1D.

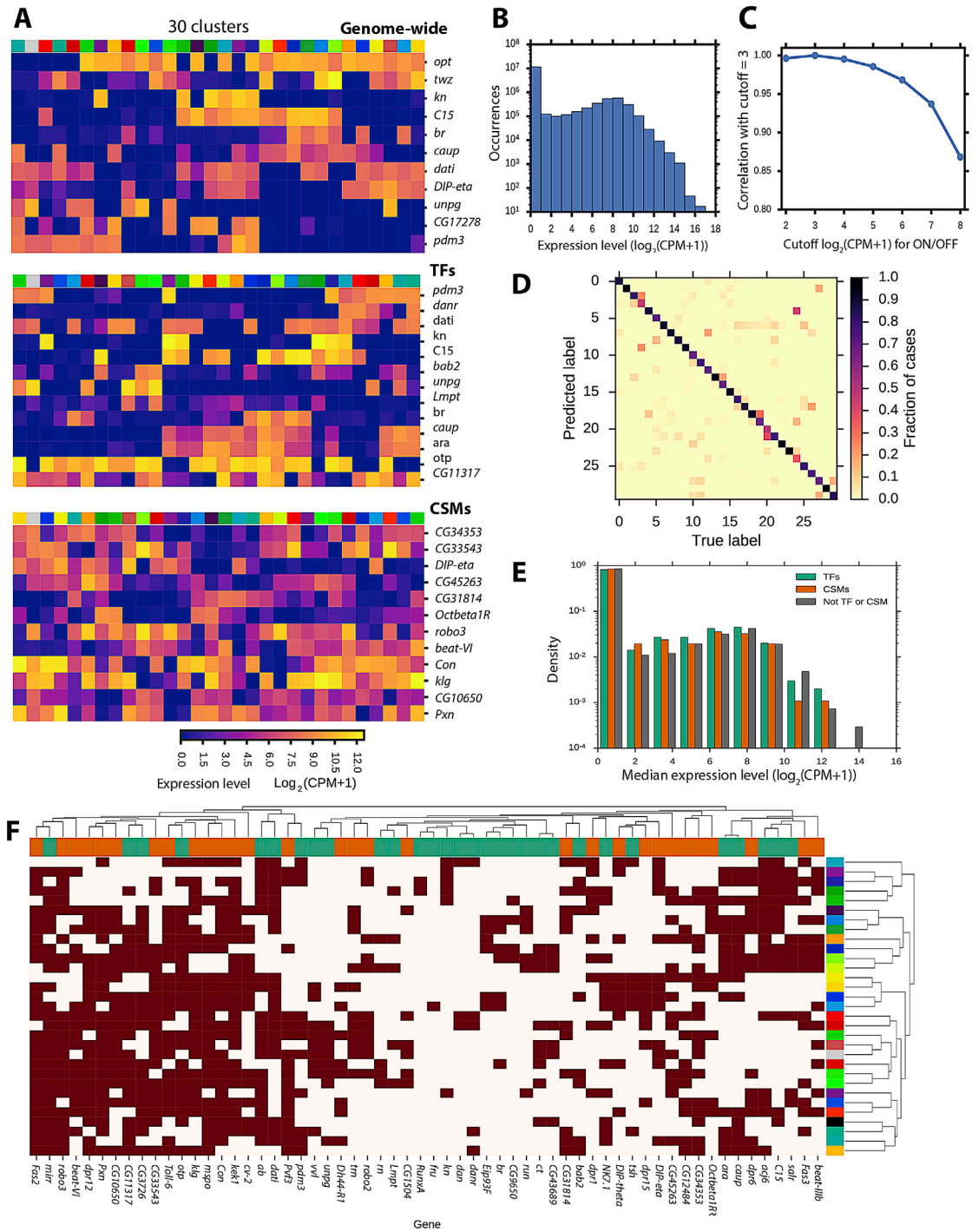


Figure S7. Combinatorial Molecular Codes of PN Subtype Identity, Related to Figure 7

(A) Expression levels of genes in the minimal combinatorial codes for *GH146*+ PN subtype identity. Each column is a cluster and each row is a gene. Color indicates mean expression of the gene among the *GH146*+ cells within the cluster (CPM, counts per million). These plots correspond to Figures 7C–7E before binarization. Cells and genes are arranged by hierarchical clustering on binarized expression states (dendrograms shown in Figures 7C–7E).

(B) Distribution of expression levels across all genes and all *GH146*+ cells. We chose to binarize expression levels using a cutoff of $\log_2(\text{CPM}+1) = 3$ (equivalent to CPM = 7) because this is a minimum of the distribution.

(C) Robustness of information content of genes to the choice of binary cutoff. We calculated the Pearson correlation between the mutual information of a gene under various binarization cutoffs and the mutual information at the binarization cutoff that we used for analysis [$\log_2(\text{CPM}+1) = 3$]. For binarization cutoffs ranging from 2 to 6, the mutual information was highly similar ($\rho > 0.95$), indicating that the precise choice of binarization threshold does not affect our results.

(legend continued on next page)

(D) Performance of multinomial classifier using 11 genes discovered in the genome-wide search for a minimal combinatorial code. Classifier performance was assessed by leave-one-out cross-validation (STAR Methods). Overall, 82% of individual *GH146+* cells were classified correctly. Errors can be attributed to measurement noise (e.g., dropouts), which gives rise to ambiguity between classes that are distinguished by expression of only one gene (see Figure 7C). Note that inclusion of additional redundant genes in the coding set would confer robustness to both measurement and biological noise.
(E) Distribution of median expression levels of TFs, CSMs, or other genes (not TF or CSM) in *GH146+* cells. No significant differences between the distributions were found ($p > 0.05$; Kolmogorov-Smirnov test, two-sample).
(F) Clustering of expression state profiles across clusters of *GH146+* cells of TFs and CSMs that are highly informative with respect to cell identity. Average linkage clustering based on Hamming distance was performed on the binarized expression profiles of the top 30 TFs and CSMs, ranked by mutual information with cell type identity. Colors next to columns indicate TFs (green) or CSMs (orange). Colors next to rows indicate cluster identity, as shown in Figure 7C–7E. No TF and CSM pairs display identical expression profiles, suggesting a lack of simple regulatory relationships.