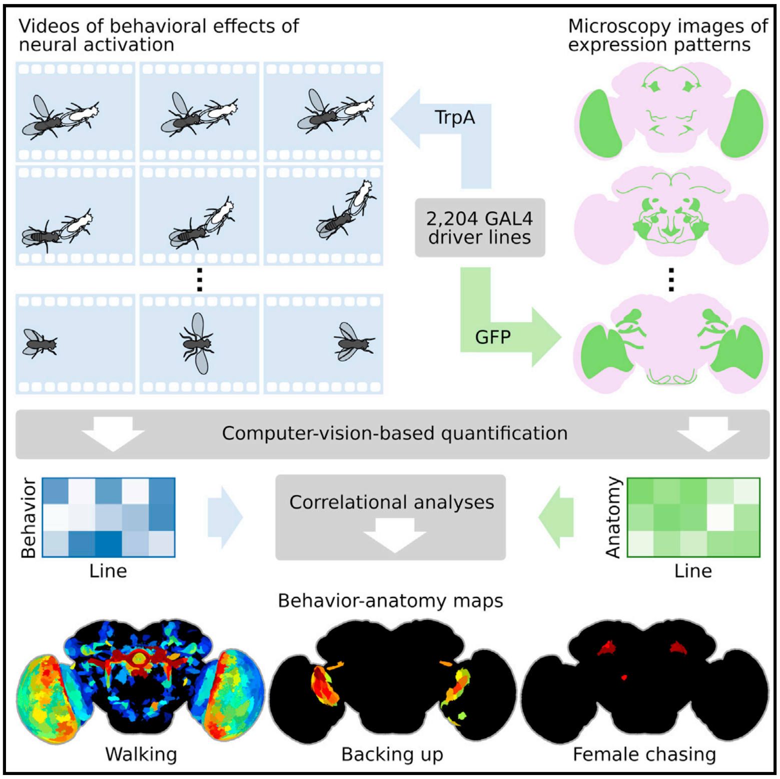
Cell

Resource

Mapping the Neural Substrates of Behavior

Graphical Abstract



Highlights

- We developed machine-vision methods to broadly and precisely quantify fly behavior
- We measured effects of activating 2,204 genetically targeted neuronal populations
- We created whole-brain maps of neural substrates of locomotor and social behaviors
- We created resources for exploring our results and enabling further investigation

Authors

Alice A. Robie, Jonathan Hirokawa, Austin W. Edwards, ..., Julie H. Simpson, Michael B. Reiser, Kristin Branson

Correspondence

bransonk@janelia.hhmi.org

In Brief

Machine-vision analyses of large behavior and neuroanatomy data reveal whole-brain maps of regions associated with numerous complex behaviors.

Robie et al., 2017, Cell *170*, 393–406
July 13, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2017.06.032

Mapping the Neural Substrates of Behavior

Alice A. Robie,[1] Jonathan Hirokawa,[1,2] Austin W. Edwards,[1] Lowell A. Umayam,[1] Allen Lee,[1] Mary L. Phillips,[1,4] Gwyneth M. Card,[1] Wyatt Korff,[1] Gerald M. Rubin,[1] Julie H. Simpson,[1,3] Michael B. Reiser,[1] and Kristin Branson[1,5,*]
[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA
[2]Laboratory of Integrative Brain Function, The Rockefeller University, New York, NY, USA
[3]Department of Molecular, Cellular, and Developmental Biology, UC Santa Barbara, Santa Barbara, CA, USA
[4]Department of Neurobiology, University of Alabama at Birmingham, Birmingham, AL, USA
[5]Lead Contact
*Correspondence: bransonk@janelia.hhmi.org

http://dx.doi.org/10.1016/j.cell.2017.06.032

SUMMARY

Assigning behavioral functions to neural structures has long been a central goal in neuroscience and is a necessary first step toward a circuit-level understanding of how the brain generates behavior. Here, we map the neural substrates of locomotion and social behaviors for *Drosophila melanogaster* using automated machine-vision and machine-learning techniques. From videos of 400,000 flies, we quantified the behavioral effects of activating 2,204 genetically targeted populations of neurons. We combined a novel quantification of anatomy with our behavioral analysis to create brain-behavior correlation maps, which are shared as browsable web pages and interactive software. Based on these maps, we generated hypotheses of regions of the brain causally related to sensory processing, locomotor control, courtship, aggression, and sleep. Our maps directly specify genetic tools to target these regions, which we used to identify a small population of neurons with a role in the control of walking.

INTRODUCTION

To address the fundamental question of how nervous systems generate behavior, we must first identify which neurons constitute the neural circuits generating these behaviors. In model organisms like *Drosophila melanogaster*, if we furthermore obtain genetic access to these neurons, we can leverage powerful genetic tools for manipulating and recording neural activity (Owald et al., 2015; Sivanantharajah and Zhang, 2015) to probe, observe, and ultimately understand the neural computations that give rise to behavior. Comprehensive functional mapping of each neuron to its behavioral roles is difficult, even in model organisms, because of the scales involved: hundreds of thousands of interconnected neurons and approaching-infinite variety of behavior. To create a map of the neural substrates of a single behavior would require the ability to monitor neural activity at the resolution of individual cells across the entire brain in freely behaving animals. Calcium imaging (Ahrens et al., 2013; Seelig and Jayaraman, 2013), calcium integrators (Fosque et al., 2015), and related approaches (Randlett et al., 2015) have recently been used to measure neural activity in large neuronal populations of behaving animals. As yet, these approaches are limited in at least one of the following ways: the need to restrain the animal or study a single behavior, spatial extent, sensitivity, and spatial and temporal resolution. A complementary method for mapping the neural substrates of behavior, with different strengths and weaknesses, is to manipulate neural activity and observe the behavioral effects. Here, we synthesized whole-brain-behavior maps by combining the results of manipulating activity in thousands of small neural populations across the brain in hundreds of thousands of animals. To make these maps of the neural substrates of behavior, we have taken an approach analogous to early stimulation experiments used to map somatosensory cortex (Penfield, 1950), using modern genetic tools to activate thousands of sparse populations of neurons, and machine-vision and machine-learning methods to extract behavior-anatomy maps from the resulting mass of data. These maps can serve as a guide for future mechanistic and functional studies of the neural substrates of behavior.

We leveraged a powerful resource available in *Drosophila*: a collection of thousands of GAL4 driver lines, each providing control over activity in the same neuronal populations across individuals (Jenett et al., 2012). These neuronal populations consist of tens to hundreds of (often) functionally unrelated cell types across the brain (Pfeiffer et al., 2008). While this feature of GAL4 lines is often viewed as a limitation, it implies that many lines will have overlapping expression, a fact we exploit to accumulate evidence to identify the neural substrates of a behavior. As these driver lines often have expression in multiple regions throughout the brain, we could efficiently test the role of all brain regions by systematically assaying the locomotor and social behavior of flies across a large collection of driver lines. To this end, we quantified the behavioral effects of neuronal activation for over 400,000 flies across 2,204 GAL4 lines, resulting in over 100 billion annotations of behavior. The size of this dataset—over 500 TB of video data—necessitated automation, and we developed computer-vision-based methods to quantitatively profile the behavioral effects of activation. We furthermore mined high-resolution, volumetric images of the GAL4 driver line expression patterns to create a new subdivision of the fly brain into genetically defined subcompartments, then automatically identified which GAL4 lines targeted each neuronal subcompartment. We developed an analytical framework for combining

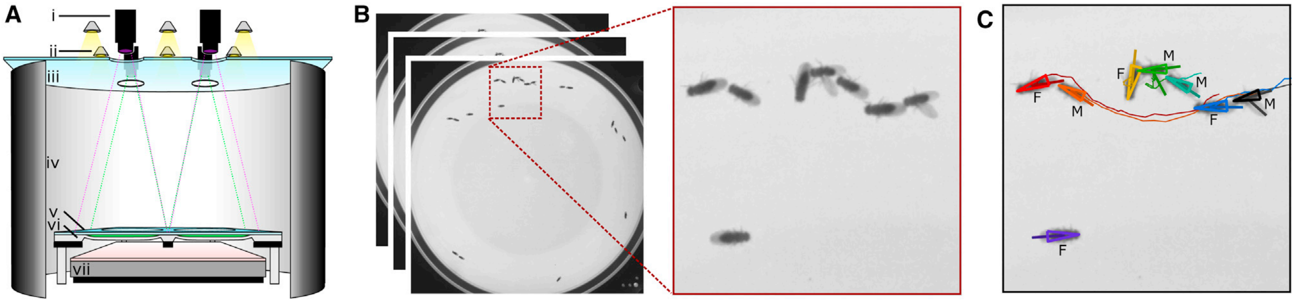


Figure 1. Behavior Video Data Collection and Tracking

(A) Schematic diagram of the Fly Bowl behavior rig. Each of our two rigs consisted of four walking arenas that can be used in parallel; diagram shows a cross-section through the two nearest arenas. *i*: A camera with an IR-pass filter was positioned above each arena. *ii*: Visible light was provided by LED lights and diffused through *iii*, a thin sheet of white acrylic. *iv*: A cylinder with white paper inside and black-out material outside provided a constant visual surround. *v*: A sheet of glass coated with Sigmacote provided a low, slippery ceiling to the arenas. *vi*: The arenas were milled into a sheet of polycarbonate, each with a diameter of 127 mm and a center height of 3.5 mm. To reduce the amount of time spent at the edges, each arena has sloped walls (Simon and Dickinson, 2010). *vii*: A panel of near-IR LEDs provided bright, constant illumination to the camera that was invisible to the flies.

(B) Collected video. Each video consisted of ~30,000 1024×1024 pixel frames of ~10 male and ~10 female flies.

(C) Example results of automatic tracking of flies' bodies and wings. We used Ctrax (Branson et al., 2009) to automatically track the flies and fit directed-ellipses to their bodies in each frame (size and orientation indicated by triangles; color indicates identity). The sex of each fly was classified automatically based on fit-ellipse area (M, male; F, female). Thin lines indicate wing angles tracked using custom MATLAB software.

these large, quantitative behavior and anatomy datasets, the results of which were brain-wide maps of the neural substrates for a broad range of behavior phenotypes.

These maps identify distinct neural substrates for the control of at least six behavior features (increased walking, jumping, backing up, wing-grooming, female aggression, and wing angle). They can contain multiple behavior-circuit components, as in the case of the increased walking map, and thus facilitate a circuit-level understanding of behavior control. Additionally, these maps allow immediate identification of genetic tools to further test and investigate the proposed structure-function relationships. We demonstrate how our maps can be used to create sparse, intersectional genetic driver lines targeting specific subpopulations of neurons within the map. To foster the systematic identification and investigation of many more behavior-anatomy relationships throughout the brain, we created searchable websites and interactive software.

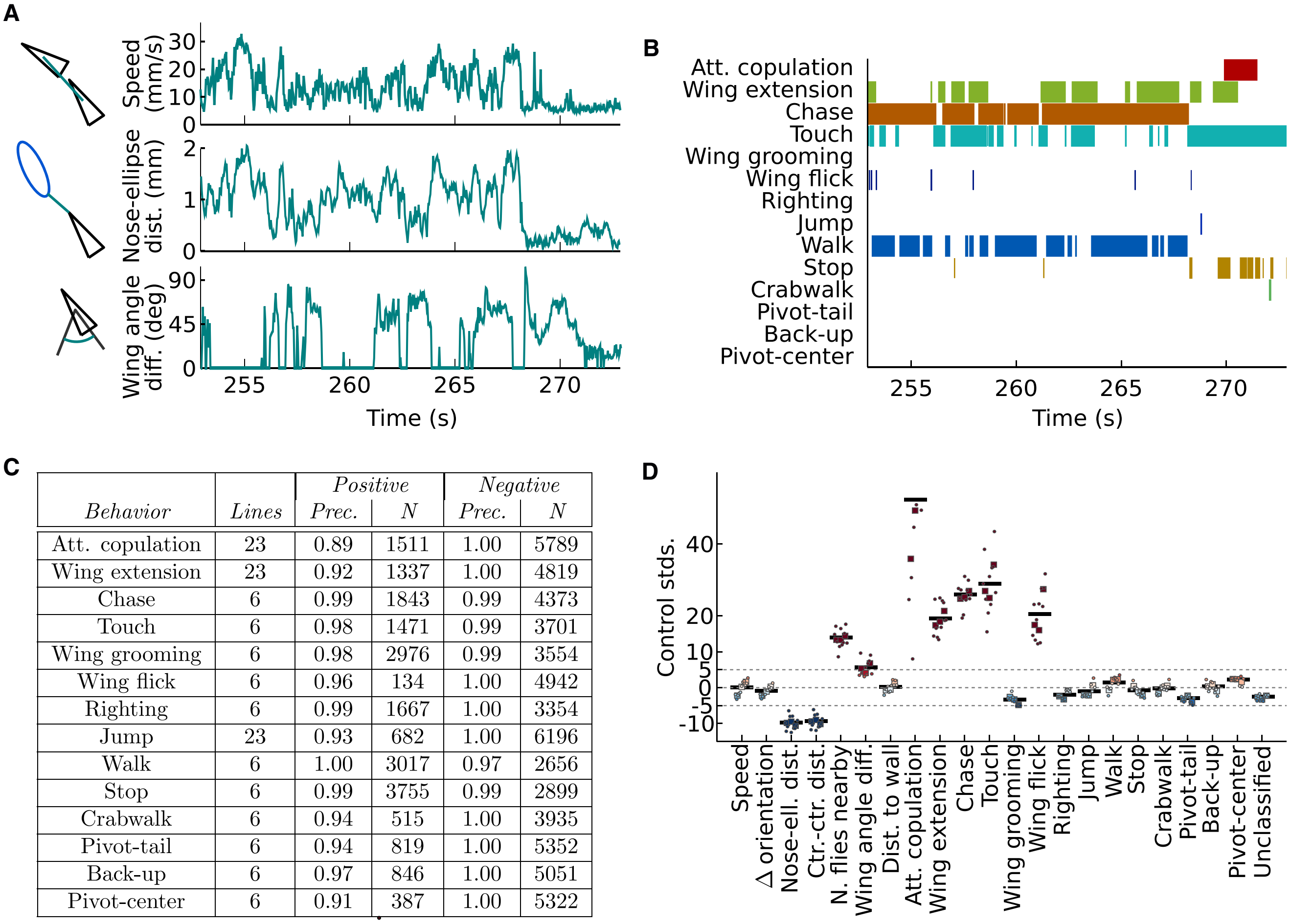
RESULTS

Computer-Vision-Based Quantification of the Effects of Neural Activation

2,204 GAL4 lines from the Janelia GAL4 collection were selected for behavioral measurement (Table S1) based on their expression patterns, as imaged by the Janelia Fly Light project (Jenett et al., 2012), prioritizing lines with sparser expression and providing coverage of the entire brain (Figure S1). To genetically activate subsets of neurons, we used the GAL4-UAS system to target expression of the temperature-sensitive dTRPA1 cation channel. For each of these lines, we recorded approximately eight videos of mixed-sex groups of ~20 flies freely behaving in a Fly Bowl (Simon and Dickinson, 2010), a shallow arena designed to facilitate automatic tracking of groups of walking flies, above the activating temperature for dTRPA1. Parameters of our behavior assay were chosen so that flies from our genetic control (an empty GAL4 line crossed to the same effector) performed both independent locomotion/foraging behaviors and social behaviors such as courtship, allowing us to see activation-induced increases and decreases in activity and social behavior. Our Fly Bowl apparatus and custom data-capture software were optimized for high-throughput, high-fidelity, consistent behavior data collection (Figures 1A, 1B, and S2A; Method Details). Using this system, we collected data at a rate of up to 116 videos from 25 GAL4 lines per day and recorded and curated in total 20,288 1,000 s videos at 30.3 fps (>500 TB of uncompressed data) over a 1.5-year period. The data collected were sufficiently stereotyped that we could use identical parameters for all analyses, and retests of each GAL4 line produced consistent behavioral measurements.

We used fully automated computer vision methods to measure a broad range of detailed statistics of the flies' locomotion and social behavior at scale. We automatically tracked the body position, orientation, and wing positions and classified the sex of each fly in each video (Figures 1C and S2B–S2D; Method Details) (Branson et al., 2009). From these trajectories, we computed a suite of 128 hand-engineered features, termed per-frame features, that captured behaviorally relevant, continuous properties of the flies' instantaneous poses and movements in each frame, such as the instantaneous speed, the distance to the closest fly, and the angle between the two wings (Figure 2A; Method Details; Table S2) (Kabra et al., 2013).

We created 14 automatic behavior classifiers for locomotion behaviors (walk, stop, back up, crabwalk, pivot-tail, pivot-center, jump, righting, wing grooming) and social behaviors (chase, touch, wing extension, attempted copulation, wing flick) (Figure 2B; Method Details; Movie S1), chosen based on our initial observations of control-line behavior and previously described components of courtship (Yamamoto and Koganezawa, 2013). While this description of fly behavior is not complete, 84% of control-line frames fell into at least one of these categories.



Behavior	*Lines*	*Positive*		*Negative*	
		Prec.	*N*	*Prec.*	*N*
Att. copulation	23	0.89	1511	1.00	5789
Wing extension	23	0.92	1337	1.00	4819
Chase	6	0.99	1843	0.99	4373
Touch	6	0.98	1471	0.99	3701
Wing grooming	6	0.98	2976	0.99	3554
Wing flick	6	0.96	134	1.00	4942
Righting	6	0.99	1667	1.00	3354
Jump	23	0.93	682	1.00	6196
Walk	6	1.00	3017	0.97	2656
Stop	6	0.99	3755	0.99	2899
Crabwalk	6	0.94	515	1.00	3935
Pivot-tail	6	0.94	819	1.00	5352
Back-up	6	0.97	846	1.00	5051
Pivot-center	6	0.91	387	1.00	5322

Figure 2. Automated Quantification of Behavior

(A) Example per-frame feature time series for one fly for 20 s. Per-frame features are simple, engineered functions of the trajectories.

(B) Automatic behavior classification results for the same fly and time interval as (A). Each row and color corresponds to a different behavior classifier. Color indicates that the classifier predicted that the behavior was occurring.

(C) Accuracy of automated behavior classifiers. We quantified the classifiers' precision for each behavior classifier—for all frames predicted by the classifier to be of a given class, what fraction were also manually labeled to be of that class? *Positive*, *Negative* class refers to frames for which the behavior is, is not occurring, resp. We report precision (*Prec.*), computed across all frames and lines annotated, and the number of frames predicted as positive and negative (*N*, number of points from which precision is computed). Per-line results are in Table S4.

(D) Example population-level behavior statistics (*x* axis) for line R82E08, computed from both per-frame features (A) and behavior classifications (B). *y* axis indicates the signed number of control standard deviations different from control mean. Circles corresponds to videos, squares to retests of the line at different times of year from different crosses, and black horizontal lines to the line-level mean. Across retests, R82E08 spends more time performing social behaviors, e.g., attempted copulation, wing extension, and chase. "Unclassified" is the fraction of time when behavior is not classified as any of our 14 behaviors. Data was clipped at 60 standard deviations.

The automatic classifiers were trained using JAABA, an interactive machine-learning tool (Kabra et al., 2013), to match our behavior definitions. Here, we manually labeled the behaviors the flies were performing in a small set of frames, and the machine-learning algorithm learned classifiers that could automatically reproduce these manual labels. To ensure that the classifiers were accurate across all the behaviorally diverse lines assayed, we used JAABA's interactive training framework to iteratively add training labels for lines selected based on several heuristics (Method Details). On average, to train each classifier, we labeled 9,000 frames from 750 contiguous bouts, nine videos, and eight lines (Table S3). To our knowledge, this is the largest, most diverse dataset to which machine-vision-based behavior classification has been applied. Averaging across behaviors and GAL4 lines, the classifiers' per-frame average accuracy was 97.6%, measured comparing the automatic classifiers' predictions to human annotations on videos from each of 5–23 representative GAL4 lines (Figures 2C and S3A; Method Details; Table S4).

By combining time series of per-frame features and behavior classifications for a given video, we computed 203 statistics to describe the behavior of the population, such as the fraction of time the flies spent walking, the average speed of the flies, and the average speed of the flies while walking

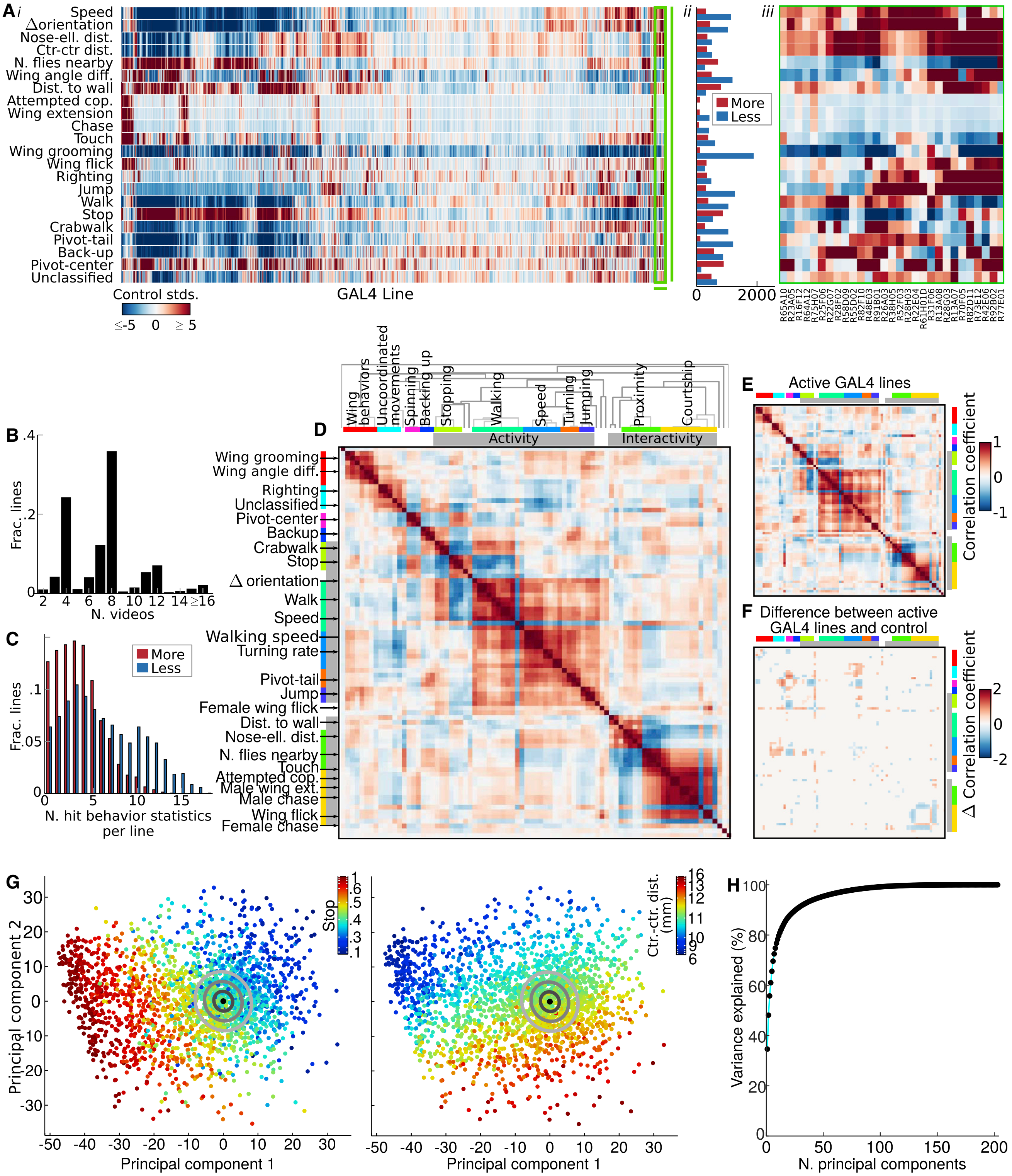


Figure 3. Behavioral Effects of Neural Activation

(A) *i:* Table of behavioral effects of neural activation for all 2,205 GAL4 lines assayed. Rows correspond to line-level behavior statistics (Table S6), columns to GAL4 lines. Color indicates how much higher (red) or lower (blue) the behavior statistic was for the line than control. Lines are sorted to highlight behavioral similarities. *ii:* Number of GAL4 lines that had significantly larger (red) and smaller (blue) values for each of the behavior statistics shown (FDR [false discovery rate] ≤ 0.1). *iii:* Zoom-in of green box in *i*.

(legend continued on next page)

(Figures 2D and 3A; Method Details; Tables S5 and S6). These statistics were hand engineered without looking at the results of the screen and included measures we hypothesized might be independently modulated. We obtained line-level statistics by averaging in a manner chosen to account for the strengths of dependencies: the approximately four videos recorded simultaneously from the same cohort were more similar than the approximately two biological replicates of the same line recorded at different times of the year (Figures 2D and 3B; Method Details). Each line-level behavior statistic was the average of ~4 million per-fly, per-frame measurements of behavior and, as measurement noise averaged out, were accurate and consistent across retests (Figures S3B and S3C). This allowed our analyses to be sensitive to smaller changes than were apparent to the human eye; for example, our analyses of walking behavior used increases in walking rates of under 5%.

To measure behavior changes due to neural activation, we compared the measurements for each GAL4 line to those from a genetic control (Figures 3A and 3C; Method Details). For each behavior statistic, we computed both the statistical significance (Method Details) and the magnitude of the effect of activation in terms of the signed number of standard deviations from control. Figure 3A summarizes the effects of neural activation for all 2,205 lines assayed and a subset of the behavior statistics.

As we measured the effects of activation for a large, unbiased sample of neuronal populations, patterns in the observed behavioral effects may be informative about how behavior is modulated by, and thus its representation in, the nervous system. Many of our behavior measurements were strongly correlated or anti-correlated (Figures 3D, 3E, and S4A). Some correlations were a result of our behavior definitions; e.g., fraction of time stopped and average speed are necessarily anti-correlated. Other correlations were a result of the structure of fly behavior; e.g., fraction of time chasing and attempting copulation are highly correlated (Figures 3D, 3E, S4A, and S4B). Both across retests of the control line and across GAL4 lines, behavior appeared to be most commonly modulated on the axes of activity level (e.g., fraction of time stopped) and how much the flies interacted (e.g., average inter-fly distance). Indeed, the first two principal components of the line-versus-behavior matrix (Figure 3A) corresponded to these two properties (Figures 3G, 3H, S4D, and S4E). Furthermore, we observed significant differences in these properties for many lines (60% of lines showed a significant difference in fraction of time stopped, 40% in inter-fly distance, Figure 3A, ii). For both properties, we observed lines with extreme values (flies that rarely moved and flies that rarely stopped, flies that clustered together and flies that actively avoided each other; Method Details), as well as for every value in between.

Groups of behavior measures that were strongly correlated across control retests (Figure 3D) were usually modulated together by neural activation as well, and the correlation structure across control line retests and across GAL4 lines was remarkably similar (Figures 3D–3F, and S4A–S4C). For example, it was the rare exception to evoke male chasing without also evoking other courtship behaviors such as attempted copulation (though such exceptions did occur; Figure S4B). Similarly, locomotion behaviors were modulated together; e.g., increases in walking were accompanied by increases in turning (Figure S4B).

We were surprised by how common it was to observe behavioral effects of activation, despite the sparsity of the expression patterns of the lines we assayed; based on our sensitive quantitative analyses, we observed a significant difference for at least one behavioral statistic for 98% of the GAL4 lines (FDR ≤ 0.1; Method Details). We observed a variety of extreme phenotypes, including lines that jumped up to 100× more than control and maintained large inter-fly distances (R68C07, R73E12), lines for which males chased up to 20× more than control (R82E08, R71G01), lines for which females chased over 20% of the time (R26F09, R26E01), lines that groomed their wings over 5× more than control (R29E04, R73E01), a single line that copulated (R34H05), and lines that stopped a large fraction of the time in close proximity to other flies and pivoted around their centers (R61C12, R66A06). The high rate of behavioral phenotypes suggests that the activation of almost any subset of neurons can generate a detectable change in behavior (due to direct as well as indirect effects of activation); thus, our behavior dataset is rich in information about the functional roles of neural structures. However, this also presents a computational challenge for discovering meaningful behavior-anatomy correlations and necessitates analyses that take advantage of automation.

As a resource for the neuroscience community, we have created a website summarizing the measured behavioral effects of activation for each GAL4 line (http://research.janelia.org/bransonlab/FlyBowl/BehaviorResults). These webpages can be searched by line name, allowing researchers who have identified anatomically interesting GAL4 lines to find behavioral

(B) Histogram across GAL4 lines of the total number of videos recorded.

(C) Histogram across GAL4 lines of the number of behavior statistics (out of the 22 shown in [A]) that a given GAL4 line has a significantly higher (red) or lower (blue) value for than control (FDR ≤ 0.1).

(D) Spearman's rank correlation coefficient between line-level behavior statistics across 762 retests of the control line (red indicates positive correlation, e.g., between attempted copulation and male chasing; blue indicates negative correlation, e.g., between stop and walk; and white indicates no correlation). Statistics are sorted according to their hierarchical clustering, shown at the top. We give descriptive names to several clusters (colored bars) and indicate a few representative behavior statistics for each (left). The same ordering is used in (E) and (F) and Figure S4A.

(E) As in (D), but for the 1646 GAL4 lines that are active (stopped less than 55% of the time). The correlation structure between control lines is similar to that of the active lines.

(F) Differences in correlations between active GAL4 lines and control retests. Red indicates that correlation is significantly higher for the GAL4 lines than control retests, blue that it is significantly lower (FDR ≤ 0.1), white that correlation differences are not significant.

(G) Principal directions of behavioral effects of neural activation. Behavior vectors for all GAL4 lines projected onto their first two principal components. Dots correspond to GAL4 lines, color indicates average fraction of time stopped (left) and inter-fly distance (D center, right). The black dot corresponds to the control line, and gray circles to 1–3 standard deviations. The first component represents activity level, while the second represents how much the flies interacted.

(H) Amount of variance of GAL4-line behavior explained by its principal components.

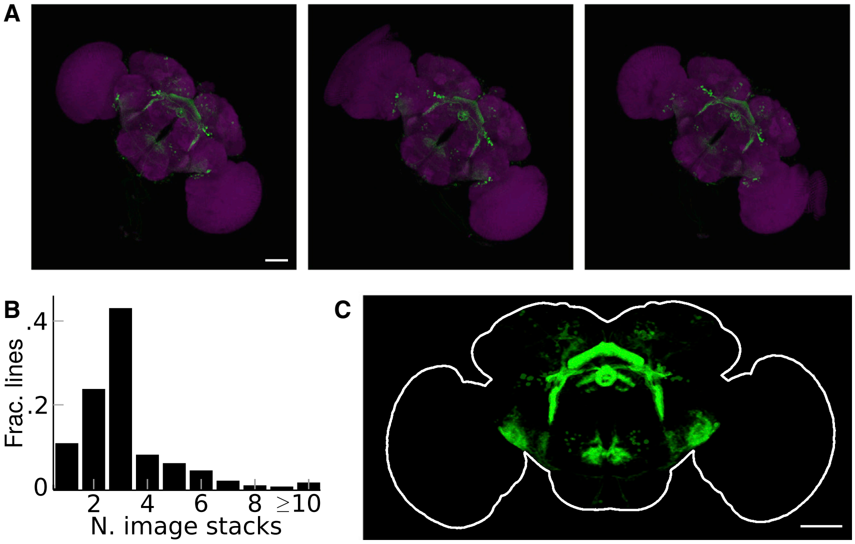


Figure 4. Expression Pattern Image Processing

(A) Maximum intensity projections (MIPs) of raw image stacks of three dissected brain samples for R26C06 (Jenett et al., 2012). GFP expression is in green, and nc82 reference stain is in magenta. All panels: scale bar, 50 μm.

(B) Histogram across GAL4 lines of the number of image stacks combined to create the per-line expression pattern.

(C) Raw image stacks are aligned, normalized, averaged, and blurred to create a single image stack for the GAL4 line.

annotations of the effects of neural activation. Alternatively, they can be indexed by any of our behavior statistics, allowing researchers interested in a particular behavior to find GAL4 lines that exhibit strong phenotypes for those behaviors upon neural activation.

Mapping Behavior to Neural Anatomy

To find regions of the brain correlated with each behavior measure, we combined our behavior dataset with a novel quantification of the neural expression pattern for each GAL4 line. As there is no cell-type annotation of the Janelia GAL4 collection, we developed a method to quantify expression pattern from 3D images of the expression pattern in the brain of each line (Jenett et al., 2012) (Figures 4A and 4B). We aligned (Peng et al., 2011) and normalized the expression pattern images so that, across images and GAL4 lines, each voxel location corresponded to approximately the same location in the brain, and voxel intensity approximated our confidence that there was expression at that location (Figures 4C and S5; Method Details). For each of the 30 million voxels in the brain images, each representing <1 μm^3 of the brain, we could test for statistically significant correlation across GAL4 lines between anatomical expression (intensity) at that voxel location and a given behavior measure. However, this amounts to performing 30 million hypothesis tests, and correcting for the effects of this large number of comparisons would require the original correlation hypothesis tests to have extremely high significance. Thus, a subdivision of the brain into larger regions was necessary. The *Drosophila* brain has been partitioned into 68 named structures, such as the mushroom body and the medulla, based on neuroanatomy (Ito et al., 2014). Thus, we could instead perform just 68 correlation hypothesis tests, one for each of these named structures, in which expression for the region is computed by averaging voxel intensities across the region. However, these regions are large and imprecise, and averaging expression across them obscured signal and resulted in incomplete, low-resolution maps.

To perform a reasonable number of statistical tests, increase signal-to-noise ratio, obtain high-resolution behavior-anatomy maps, and improve speed, we developed a novel segmentation of the entire fly brain into 7,065 supervoxels—spatially coherent clusters of voxels with similar expression across the GAL4 lines. This clustering was based on the assumption that if the expression for a pair of voxels was the same across most GAL4 lines, they might be part of the same neuronal cell type or functional unit (Figures 5A and S7; Method Details). The resulting segmentation shows correspondence to known structures in the fly brain (Figures 5B–5D; Data S1; Movie S2). For example, our segmentation consists of layers in the medulla (Nern et al., 2015) and fan-shaped body (Wolff et al., 2015) and annuli in the ellipsoid body (Wolff et al., 2015). A similar methodology has been employed to discover novel optic glomeruli (Panser et al., 2016) (Method Details).

We used this clustering of the entire fly brain to quantitatively represent the neural expression pattern for each GAL4 line by computing the average pixel intensity within each supervoxel (Figure 6). Then, for a given behavior measure (e.g., walking more than control), we tested whether, for each supervoxel, there was a positive correlation between the behavior measure and the neural expression across the GAL4 lines (Method Details). We visualized the correlated neural regions by mapping the p values of these correlation hypothesis tests back to the supervoxel locations (Figure 7A; Movie S3), creating 3D maps of the neural regions related to the behavior phenotype. Combining information from thousands of lines allowed us to find these correlations despite confounding factors—that multiple regions of the brain are likely involved in a given behavior and that the expression patterns of the majority of GAL4 lines likely contain multiple cell types that interact (Figure 7B).

For each of the behavior statistics we measured, we made such maps for unexpectedly high (e.g., wing grooming more than control) and unexpectedly low values (e.g., wing grooming less than control). In addition, we made behavior-anatomy maps for combined behavior measures (e.g., stopping more but not wing grooming more than control).

We can further analyze these maps by focusing on those lines most important in producing high correlations (lines that both have expression in some of the significantly correlated supervoxels and show the given behavior phenotype; Method Details).

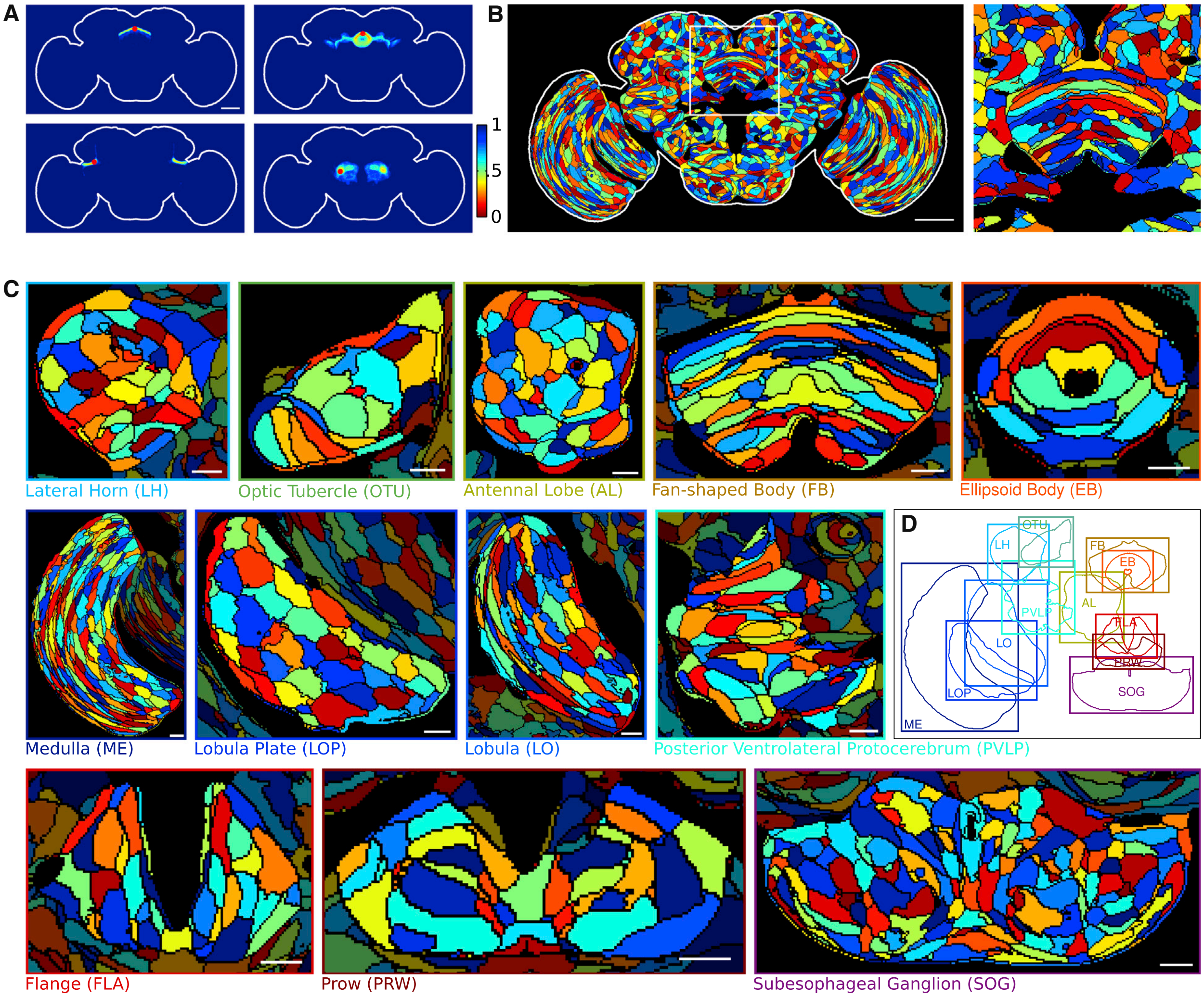


Figure 5. Clustering the Fly Brain

(A) Illustration of voxel clustering distance function. For the voxel indicated by the red dot, we show in pseudocolor the distance to each other voxel (minimum projection over *z*). Distances to voxels in related regions are smaller. Scale bar, 50 μm.

(B) Resulting clustering of brain into supervoxels. We show cross-sections of supervoxels at z = 72.5 μm; each supervoxel has a different, randomly selected color. White box indicates location of zoomed-in region in right panel. Scale bar, 50 μm.

(C) As in (B), clustering for selected z-slices through 12 compartments. Regions outside the given compartment are darkened. Scale bars, 10 μm.

(D) Locations of each of the shown compartments (C) within the brain.

We used these lines to automatically cluster the increased-walking map (Figure 7A) based on expression correlation between supervoxels (Figure S7A). The resulting clustering revealed components of a putative visual information pathway from the optic lobes, through the optic tubercles, to the bulb and ellipsoid body (Figure 7C). Such a visual pathway has been described in other insects (Pfeiffer et al., 2005; Träger et al., 2008), and neurons in the ellipsoid body of *Drosophila* have been shown to have visual response properties (Seelig and Jayaraman, 2013). Our results imply that this visual circuit feeds into the neuronal control of locomotion, and activation of any component of it increases the probability of walking.

The lines most important in producing high correlations for a selected neuronal region also provide direct and immediate genetic access for further investigation of the observed behavior-anatomy correlation. The intersectional split-GAL4 method can be used to refine the expression patterns of these GAL4 lines (Luan et al., 2006; Pfeiffer et al., 2010), creating more precise genetic tools and potentially identifying the specific cell types involved in neural control of the behavior. To demonstrate this, we used the split-GAL4 intersectional strategy with GAL4 lines identified based on our increased-walking map (Figures 7D–7F) to discover a small population of R2/R4 ring neurons in the ellipsoid body that were sufficient to elicit an increase in walking probability (Figures 7D–7F and S7; Method Details; Table S7). A role for these neurons in the control of walking is consistent with earlier studies of structural mutations and inactivation of ellipsoid body neurons, both of which caused

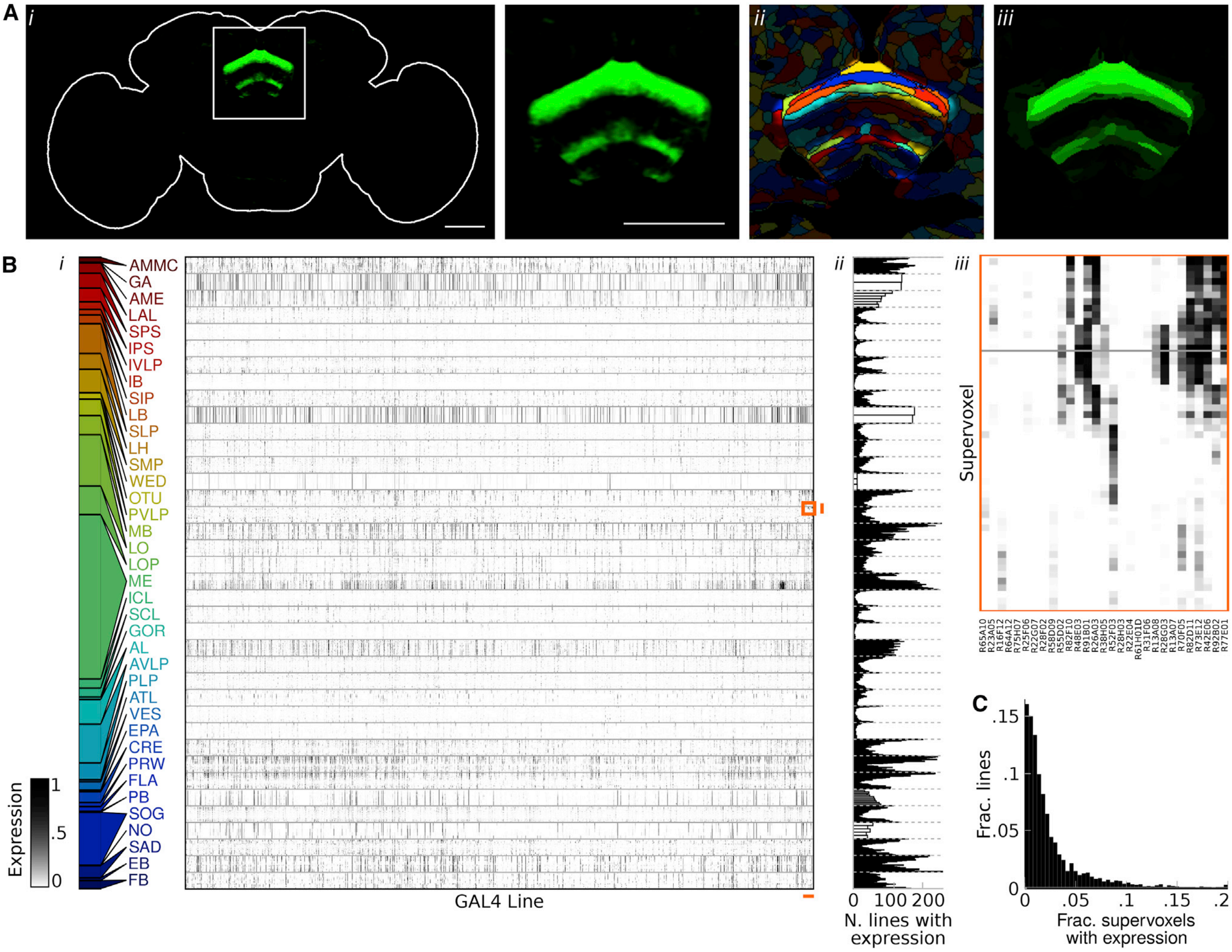


Figure 6. Anatomical Expression Data

(A) Reduced-dimensional representation of the GAL4 expression pattern for R26C06 using the supervoxel clustering. *i*: High-dimensional representation of the expression pattern for R26C06 at z = 72.5 μm. *ii*: Expression pattern overlaid on supervoxel clustering, with color indicating supervoxel and brightness indicating expression. Expression appears constant within each supervoxel. Our low-dimensional representation is the average expression within each supervoxel. *iii*: Reconstruction of original 30-million-dimensional expression pattern from 7,065-dimensional supervoxel representation, by setting the expression within an entire supervoxel to its average. Right panel of *i* and *iii* appear similar.

(B) *i:* Table of anatomical expression levels for all GAL4 lines assayed (columns, ordered as in Figure 3A). Rows correspond to supervoxels, which are grouped by compartment (Jenett et al., 2012). Row heights are set so each compartment has the same height. Darkness indicates amount of expression. Colored bar (left) indicates the relative number of supervoxels in each compartment. *ii:* Number of GAL4 lines with expression in each supervoxel. As expression level is a number between 0 and 1, we sum the total expression level values over lines, interpreting non-integer values as probabilities. These counts are mapped to supervoxel location in Figure S1. *iii:* Zoom-in of orange box in *i*.

(C) For each GAL4 line, we computed the fraction of supervoxels with expression, i.e., how dense the expression pattern is (as in [B], interpreting non-integer values as probabilities). We histogram this value across GAL4 lines.

locomotor deficits (Martín-Peña et al., 2014; Strauss and Heisenberg, 1993).

We discovered several other behavior-anatomy correlations by exploring the data collected in our unbiased activation study using our behavior-anatomy maps. We both identified previously known structure-function relationships (Figures 8A and 8B; Movie S3), further confirming the validity of our method, and discovered novel correlations (Figures 8C–8G; Movie S3), demonstrating the potential for identifying novel neural substrates of behavior. We briefly describe these maps next.

Our map of regions of the brain correlated with increased unilateral wing extension by males (a component of courtship behavior [Yamamoto and Koganezawa, 2013]; Figure 8A; Movie S3) closely resembles the *fruitless* circuit in the protocerebrum (Yu et al., 2010), neuronal regions previously demonstrated to control courtship behavior. Also confirming the validity of our approach, our map of regions of the brain correlated with an

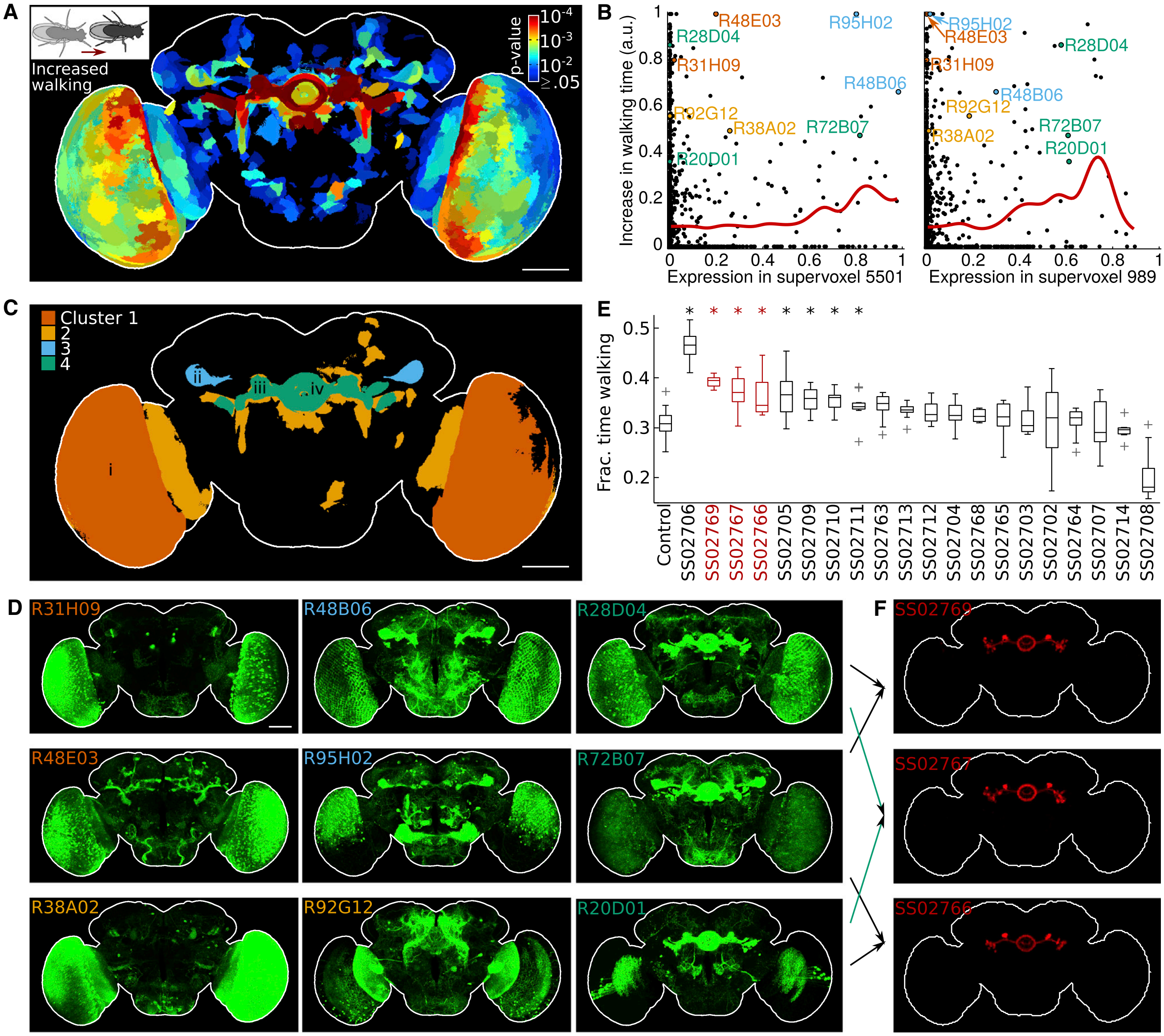


Figure 7. Neural Correlates of Walking

(A) Regions of the brain significantly correlated with an increase in fraction of time spent walking (FDR ≤ 0.25). For each supervoxel, we compute the significance of the observed correlation between expression in the supervoxel and increase in fraction of time spent walking and color by this p value. We show the minimum intensity projection. All panels: Scale bar, 50 μm.

(B) For two significantly correlated supervoxels, relationship between expression and walking behavior across GAL4 lines. Left: Supervoxel within cluster 3; right: cluster 4. We plot expression level versus normalized increased-walking statistic (Method Details) for each GAL4 line (dots). Colored circles indicate lines shown in (D). Red lines indicate average behavior score for GAL4 lines for sliding intervals of supervoxel expression level. While expression level and walking behavior are significantly correlated, we observe both lines in the upper-left of the plots for which other brain regions might be causing the behavior phenotype and lines in the bottom-right for which activation of other brain regions may be masking the behavior phenotype.

(C) Clustering of significantly correlated regions of the increased-walking map ($p \leq 0.005$) into substructures (Method Details). We indicate locations of *i*: optic lobe, *ii*: optic tubercle, *iii:* bulb, and *iv*: ellipsoid body.

(D) For each cluster in (C), we show MIPs of expression patterns for lines important in creating the behavior-anatomy correlation for the cluster (those with the highest product of behavior score and average anatomy score over all supervoxels in the cluster).

(E) Boxplot of fraction of time walking for each of the 20 split-GAL4 lines created from parental lines important in creating the behavior-anatomy correlation for cluster 4. Asterisks indicate significant increases compared to control (Mann-Whitney test, $p < 0.05$ with Bonferroni correction, number of videos ranged from 6 to 22 [Table S7]); boxes indicate 25th and 75th percentiles of data; pluses indicate points outside whiskers. Red indicates split lines shown in (F).

(F) MIPs of expression patterns for split-GAL4 lines with expression in the R2/R4 ring neurons. Arrows indicate parental lines used to create split-half lines.

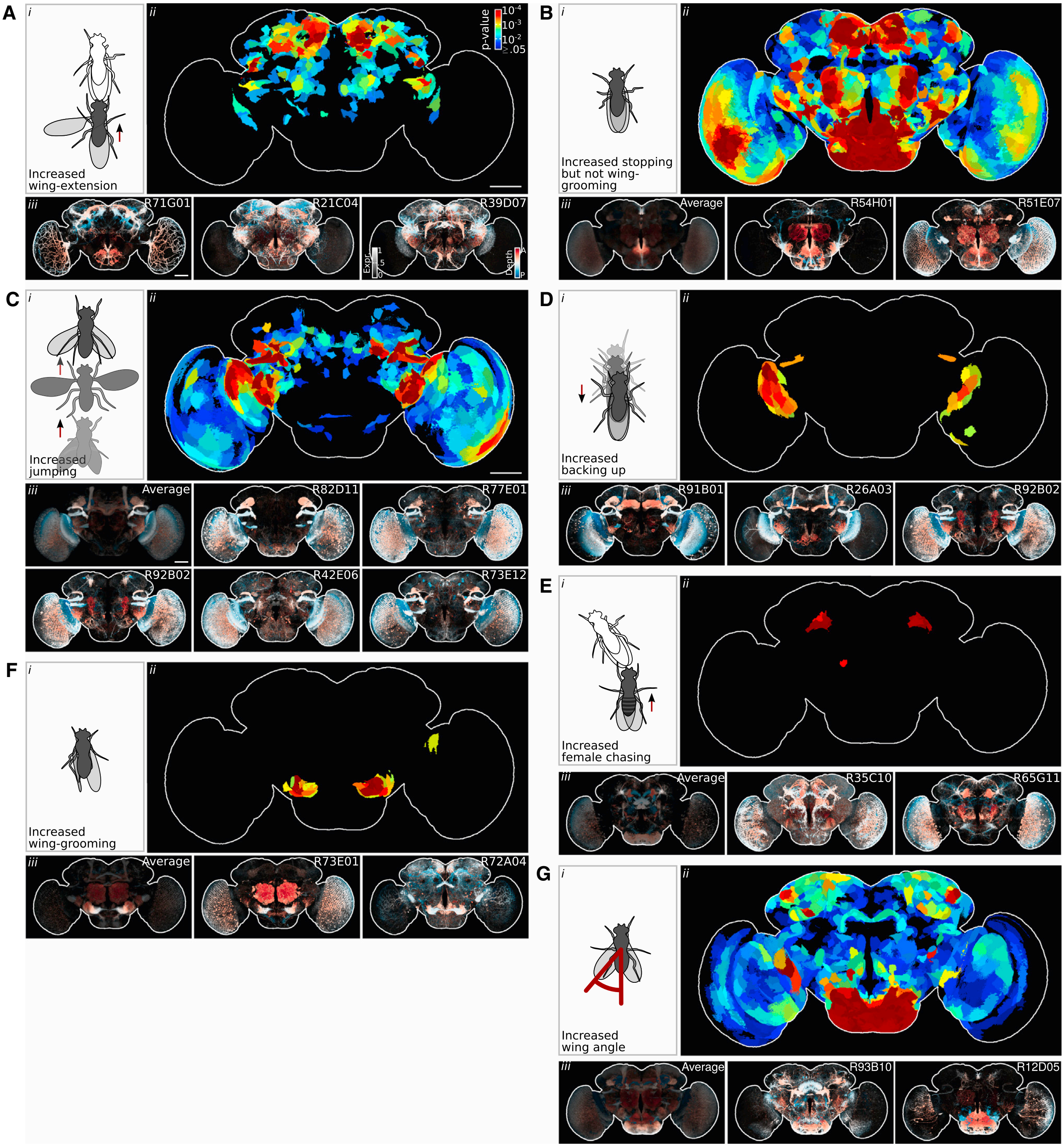


Figure 8. Behavior-Anatomy Correlation Maps

(A–G) Each panel corresponds to a different behavior phenotype. For each, we show:

i: Cartoon illustrating the behavior phenotype.

ii: A map of brain regions significantly correlated with the behavior phenotype (FDR $\leq$ 0.25, as in Figure 7A).

iii: Left: Average expression pattern for lines important in creating the significant behavior-anatomy correlations shown in *ii* ($p \leq 0.005$). Right: Selected important lines. All: MIPs for which intensity indicates maximum expression level and hue indicates depth at this maximum (anterior, red; posterior, blue). No average shown in panels (A) and (D). Average shown in (B) corresponds to a selected cluster of supervoxels.

Scale bars, 50 μm.

approximate measure of increased sleep (Figure 8B; Method Details) identified a cell type with morphology similar to a subset of a previously identified heterogeneous population of neuromodulatory neurons involved in sleep, circadian rhythms, and feeding (Cavanaugh et al., 2014; Dus et al., 2015; Foltenyi et al., 2007).

While components of the motor circuits of escape in the fly (jumping) have been well described (Hale et al., 2016), the visual inputs to these circuits are not. Our increased-jumping map (Figure 8C) suggests several such inputs: lobular columnar (LC) neurons with morphologies similar to those of types LC6, LC9, and LC10 (Otsuna and Ito, 2006). Concurrent to our study, using novel split-GAL4 lines targeting each LC type, Wu et al. independently found a jumping activation phenotype for LC6 neurons (Wu et al., 2016), confirming a portion of our increased-jumping map.

The sparse increased-backing-up map (Figure 8D) indicates that there is also a strong correlation between visual projection neurons and the backing-up behavior, which may be an avoidance or lower-intensity escape behavior. These maps contain a previously undescribed LC neuron that projects to a region close to the LC6 glomeruli and enters the ventrolateral protocerebrum (VLPR) from the posterior side. Wu et al. concurrently found a back-up activation phenotype for this cell type, which they termed LC16 (Wu et al., 2016), confirming our increased-backing-up map.

Female aggression has been described in flies (Nilsen et al., 2004), but, unlike male aggression, its neural substrates have not been well characterized. The map corresponding to an increase in chasing (a phenotype of aggression) exclusively in females is very sparse (Figure 8E)—a single bilateral region in the protocerebrum. Besides this increase in female chasing, we also observed that, for lines important for this map, females performed other aggressive behaviors such as low-posture fencing, shoving, and headbutting.

The map corresponding to increased wing grooming (Figure 8F) is similarly sparse and contains a single bilateral brain region within the antennal mechanosensory and motor center (AMMC). There are two plausible alternatives for this observed correlation. The first is causal: connectivity between antennal and wing sensory-motor circuits in the ventral nerve cord (VNC), which cause wing grooming. The second is correlational: activation of a different population of neurons, one for which GAL4 expression is genetically correlated with that in the AMMC, causes wing grooming. This second hypothesis is supported by the fact that the AMMC receives multiple types of mechanosensory input from the antenna, and proteins of mechanosensory transduction are shared across many sensory organs (Walker et al., 2000). To investigate this, we examined VNC images for lines important to the wing-grooming map (Figure S8). While there appears to be expression in the wing sensory neuropil of at least three to five lines, there does not appear to be a common neuronal cell-type throughout. Such images of VNC expression exist for many of the lines we assayed; thus, our methodology could be extended to the VNC, an interesting direction for future research.

The regions of the brain most correlated with the flies' wings being held out are within the subesophageal zone (SEZ) (Figure 8G). While interactions between feeding and locomotion circuits in the VNC have been described (Mann et al., 2013; Schoofs et al., 2014), this map suggests a novel role for circuits within the SEZ in wing motor control and is congruous with evidence that the SEZ has dense innervation by descending neurons (Hsu and Bhandawat, 2016) that connect the brain and the VNC, where wing motor control circuits are located.

These and other behavior-anatomy correlations suggested by our whole-brain maps will be a source of hypotheses for future research in *Drosophila* neurobiology. As we can create maps not only for the 203 behavior statistics we measured but also for combinations of them, the number of behavior-anatomy maps to explore is practically unlimited. To allow researchers to explore our data, we have created the Browsable Atlas of Behavior-Anatomy Maps (BABAM) software that allows users to select behavior measures and generate behavior-anatomy maps (Data S1; https://kristinbranson.github.io/BABAM/; Movie S4). BABAM allows users to explore maps in a variety of ways. Following our analysis of the increased-walking map, they can further cluster the correlated regions of the map (Figure 7C) and explore subregions of a given map. Users can find the GAL4 lines that contribute most to the correlation between a selected behavior and supervoxel or cluster of supervoxels and access the per-line web pages describing their behavior and anatomy. Finally, they can examine the supervoxel clustering of the brain. This software will facilitate testing of our structure-function hypotheses by refining expression using the intersectional split-GAL4 strategy, activating and silencing using other effectors, and/or neural recording.

DISCUSSION

In this work, we used machine-vision and machine-learning methods to quantify the behavioral effects of neural activation for thousands of GAL4 lines with overlapping expression in multiple, often unrelated, cell types. When activating multiple cell types with distinct behavioral roles, it is likely that behavioral effects will combine complexly; for example, behavioral effects of one cell type could mask another's. By using this large dataset to jointly analyze the effects of neural activation across many lines with overlapping expression, we were able to find the same, sometimes subtle, behavioral effects in multiple lines with regions of common expression and thus develop testable hypotheses of the neural substrates of a variety of locomotion and social behaviors across the entire *Drosophila* brain.

These behavior-anatomy maps both recapitulate known and suggest novel structure-function relationships. Novel relationships we discovered include a putative visual information pathway from the optic lobes to the central complex involved in production of walking behavior, visual output neurons involved in escape and avoidance behavior, a subregion of the protocerebrum involved in female aggression, a subregion of the AMMC involved in wing grooming, and a subregion of the SEZ involved in controlling wing position. Our study resulted in thousands such behavior-anatomy maps, which we share with the neuroscience community in the form of our interactive atlas software. This software allows the user to identify GAL4 lines most important in creating a selected region of a map, allowing one to gain genetic access to sparse neuronal populations within a map. Using these GAL4 lines and lines derived from them via split-intersectional strategies, one can target effectors, markers, and activity indicators to these cells, allowing researchers to manipulate and record activity, test the hypothesized relationship, and further investigate the behavior circuit. Taken together with efforts to find and identify neuronal cell types in these GAL4 lines (Chiang et al., 2011; Costa et al., 2016; Panser et al., 2016), our functional description of the fly brain will enhance the pace of research tackling neural circuit mechanisms.

Our work builds upon a long history of behavior screens in model organisms, from genetic screens elucidating the molecular and genetic bases of behavior to more recent, genetically targeted neural activation and silencing screens in *Drosophila* to identify cellular and circuit mechanisms of behavior control (Owald et al., 2015). Recent behavioral screens of large GAL4 collections have resulted in the identification of behavioral roles for a handful of neurons (Hampel et al., 2015; Hoopfer et al., 2015; Triphan et al., 2016; von Philipsborn et al., 2011). Each of these studies were focused on a single behavior and only provided careful behavior quantification for a few ''hits'' from their screens. In contrast, we have created a searchable, detailed quantification of the effects of activation for a broad range of behaviors across all 2,204 GAL4 lines assayed. Most similarly, activation phenotypes of 1,054 GAL4 lines were quantified in *Drosophila* larvae but without analysis of the neural anatomy of those lines (Vogelstein et al., 2014). Our work uniquely combines analysis of the behavioral effects of activation and quantification of the anatomical expression patterns to discover correlations between behavior and anatomy: a brain-wide atlas of behavior-anatomy maps.

We collected, to our knowledge, the largest-to-date video behavior dataset as part of this study. It describes the behavioral effects of perturbing the neural activity of thousands of sets of neurons throughout the nervous system. As such, characteristics of the space of behavioral effects may be indicative of the structure and organization of behavior and its neural representation. Despite the sparsity of the expression patterns, we observed significant behavioral effects of activation for a large fraction of GAL4 lines. However, in other ways, for the majority of lines, the behavior statistics we measured were surprisingly normal. We rarely observed abnormal correlations between our measured behavior statistics. Instead, the correlational structure across GAL4 lines was similar to that across retests of our control. This suggests that behavior modules (e.g., courtship, foraging, escape) were more commonly modulated by neural activation, as opposed to individual behavior components within these modules (e.g., chasing and wing extension, walking and turning, jumping and backing up). One explanation is that many of the effects we observed were from sensory or near-sensory neural populations, or neurons that affect the concentrations of neuromodulators, which then affect large downstream populations of neurons—entire behavior modules. Alternatively, one could speculate that the observation that GAL4 line behavior remains within the low-dimensional manifold of normal fly behavior is evidence that neural activity is also regulated to remain within a low-dimensional manifold. By exciting a small population of neurons, we push neural activity away from this manifold, and one could envision homeostatic mechanisms that pull neural activity back to the closest point on the manifold by compensatory changes in other neurons involved in the behavior module. Low-dimensional neural dynamics during locomotion behaviors have been observed previously in the *C. elegans* (Kato et al., 2015). In the fly, two of the major axes of this manifold, according to our measures of behavior, are activity and interaction levels, the most important features for describing the effects of activation. A potentially fruitful direction of future research is thus to activate sparse neuronal populations suggested by our study and observe the resulting neural dynamics across large populations via calcium imaging. For two lines with seemingly unrelated expression patterns but similar behavior, do we observe similar neural activity patterns upon activation? How does neural activity after activation compare to activity observed in control flies in different environments and states?

An important qualification to the results of this study is that they are based on potentially nonphysiological patterns of neural activity—those induced by several minutes of continuous excitation via the dTrpA effector of multiple, possibly unrelated, neurons. However, as our maps are based on activation effects across many GAL4 lines, they should be less susceptible to irregularities in any single line. Because dTrpA perturbations, in addition to direct effects, will also yield phenotypes due to nonspecific network effects, future experiments will be necessary to establish a causal role for identified brain regions. However, dTrpA activation has been used previously to successfully identify neural substrates of behavior (Hampel et al., 2015; Hoopfer et al., 2015; von Philipsborn et al., 2011). A second qualification of our results is that they are based on correlational analyses, which are not necessarily causal. For example, if there is a strong genetic correlation in expression between region A, which controls a behavior, and one region B, which does not, there will still be a correlation between expression in B and the behavior. However, we rarely observed genetic correlations in expression between regions not known or suspected to be functionally related. For example, in Figure S7A, we observe expression correlation between the selected supervoxel in the crepine and other regions in the central complex. Genetic correlations in expression can be investigated using our interactive software (Movie S4). The structure-function hypotheses suggested by our analyses should be followed up using intersectional genetic techniques, silencing and manipulation with other effectors such as temporally acute optogenetic activators, and/or neural recordings. Our interactive software and web pages will facilitate these follow-up experiments.

Further analyses of our dataset will reveal more about the neural correlates of behavior and its organization, and we make our data publicly available, in the raw video form, at the level of per-fly trajectories and finally at the level of a condensed matrix of per-line behavior statistics. From the videos and trajectories, new behavior classifiers can be trained and applied to our dataset using supervised machine learning (Kabra et al., 2013), and new behavior modes and representations can be discovered using unsupervised clustering or manifold learning techniques (Berman et al., 2014; Schwarz et al., 2015; Vogelstein et al., 2014). We believe that this data will be of interest to computational ethologists interested in understanding the vocabulary and structure of behavior, neuroanatomists interested in cataloguing cell types, and machine-vision and machine-learning researchers interested in developing behavior analysis techniques that are effective on such large, diverse datasets. From our matrices of per-line behavior statistics and per-line anatomical expression patterns, theorists could develop brain-wide models predicting behavior from neural activity, perhaps in combination with brain-wide electron-microscopy connectomics and whole-brain calcium imaging data currently being collected for *Drosophila*.

STAR★METHODS

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

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures, seven tables, four movies, and three data files and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2017.06.032.

AUTHOR CONTRIBUTIONS

Conceptualization: A.A.R., K.B., M.B.R., G.M.C., W.K., G.M.R., and J.H.S.; Methodology: K.B., A.A.R., W.K., and J.H.; Software: K.B., L.A.U., A.L., and A.A.R.; Formal Analysis: K.B. and A.A.R.; Investigation: A.A.R., J.H., A.W.E., and M.L.P.; Data Curation: A.A.R., J.H., A.W.E., and K.B.; Writing: K.B., A.A.R., and M.B.R.; Visualization: K.B. and A.A.R.; Supervision and Project Administration: A.A.R., K.B., G.M.C., W.K., M.B.R., G.M.R., and J.H.S.

ACKNOWLEDGMENTS

Behavior data was collected as part of the Janelia Fly Olympiad Team Project. We thank A. Nern, A. Jenett, and Y. Aso for selecting GAL4 driver lines based on imagery produced by the Janelia Fly Light team; J. McMahon and the Janelia Fly Core team for fly preparation; S. Roberts for help with data collection; T. Tabachnik and B. Biddle for help with rig design and fabrication; S. Branson and M. Kabra for help with software; the Janelia Fly Light team for access to anatomy images; L. Guignard and P. Keller for help rendering videos; J.-M. Knapp for genetics consulting; J. Jeter for imaging split-GAL4 lines; the Janelia Fly Olympiad Team and Scientific Computing team for support and discussions; Y. Yu, F. Long, and H. Peng for help with anatomy image processing; and P. Keller, V. Jayaraman, U. Heberlein, and A. Hantman for feedback on the manuscript.

Received: December 8, 2016
Revised: May 1, 2017
Accepted: June 19, 2017
Published: July 13, 2017

REFERENCES

Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., and Keller, P.J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat. Methods *10*, 413–420.

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

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc B Met *57*, 289–300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. *29*, 1165–1188.

Berman, G.J., Choi, D.M., Bialek, W., and Shaevitz, J.W. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. J. R. Soc. Interface *11*, 20140672.

Branson, K., Robie, A.A., Bender, J., Perona, P., and Dickinson, M.H. (2009). High-throughput ethomics in large groups of *Drosophila*. Nat. Methods *6*, 451–457.

Cavanaugh, D.J., Geratowski, J.D., Wooltorton, J.R., Spaethling, J.M., Hector, C.E., Zheng, X., Johnson, E.C., Eberwine, J.H., and Sehgal, A. (2014). Identification of a circadian output circuit for rest:activity rhythms in *Drosophila*. Cell *157*, 689–701.

Chiang, A.S., Lin, C.Y., Chuang, C.C., Chang, H.M., Hsieh, C.H., Yeh, C.W., Shih, C.T., Wu, J.J., Wang, G.T., Chen, Y.C., et al. (2011). Three-dimensional reconstruction of brain-wide wiring networks in *Drosophila* at single-cell resolution. Curr. Biol. *21*, 1–11.

Costa, M., Manton, J.D., Ostrovsky, A.D., Prohaska, S., and Jefferis, G.S. (2016). NBLAST: rapid, sensitive comparison of neuronal structure and construction of neuron family databases. Neuron *91*, 293–311.

Dasgupta, S., and Long, P.M. (2005). Performance guarantees for hierarchical clustering. J. Comput. Syst. Sci. *70*, 555–569.

Dus, M., Lai, J.S., Gunapala, K.M., Min, S., Tayler, T.D., Hergarden, A.C., Geraud, E., Joseph, C.M., and Suh, G.S. (2015). Nutrient sensor in the brain directs the action of the brain-gut axis in *Drosophila*. Neuron *87*, 139–151.

Foltenyi, K., Greenspan, R.J., and Newport, J.W. (2007). Activation of EGFR and ERK by rhomboid signaling regulates the consolidation and maintenance of sleep in *Drosophila*. Nat. Neurosci. *10*, 1160–1167.

Fosque, B.F., Sun, Y., Dana, H., Yang, C.T., Ohyama, T., Tadross, M.R., Patel, R., Zlatic, M., Kim, D.S., Ahrens, M.B., et al. (2015). Neural circuits. Labeling of active neural circuits in vivo with designed calcium integrators. Science *347*, 755–760.

Groppe, D.M., Urbach, T.P., and Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology *48*, 1711–1725.

Hale, M.E., Katz, H.R., Peek, M.Y., and Fremont, R.T. (2016). Neural circuits that drive startle behavior, with a focus on the Mauthner cells and spiral fiber neurons of fishes. J. Neurogenet. *30*, 89–100.

Hampel, S., Franconville, R., Simpson, J.H., and Seeds, A.M. (2015). A neural command circuit for grooming movement control. eLife *4*, e08758.

Hanesch, U., Fischbach, K.-F., and Heisenberg, M. (1989). Neuronal architecture of the central complex in *Drosophila* melanogaster. Cell Tissue Res. *257*, 343–366.

Hoopfer, E.D., Jung, Y., Inagaki, H.K., Rubin, G.M., and Anderson, D.J. (2015). P1 interneurons promote a persistent internal state that enhances inter-male aggression in *Drosophila*. eLife *4*, e11346.

Hsu, C.T., and Bhandawat, V. (2016). Organization of descending neurons in *Drosophila* melanogaster. Sci. Rep. *6*, 20259.

Ito, K., Shinomiya, K., Ito, M., Armstrong, J.D., Boyan, G., Hartenstein, V., Harzsch, S., Heisenberg, M., Homberg, U., Jenett, A., et al.; Insect Brain Name Working Group (2014). A systematic nomenclature for the insect brain. Neuron *81*, 755–765.

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.

Kabra, M., Robie, A.A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). JAABA: interactive machine learning for automatic annotation of animal behavior. Nat. Methods *10*, 64–67.

Kato, S., Kaplan, H.S., Schrödel, T., Skora, S., Lindsay, T.H., Yemini, E., Lockery, S., and Zimmer, M. (2015). Global brain dynamics embed the motor command sequence of *Caenorhabditis elegans*. Cell *163*, 656–669.

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

Mann, K., Gordon, M.D., and Scott, K. (2013). A pair of interneurons influences the choice between feeding and locomotion in *Drosophila*. Neuron *79*, 754–765.

Martín-Peña, A., Acebes, A., Rodríguez, J.R., Chevalier, V., Casas-Tinto, S., Triphan, T., Strauss, R., and Ferrús, A. (2014). Cell types and coincident synapses in the ellipsoid body of *Drosophila*. Eur. J. Neurosci. *39*, 1586–1601.

Nern, A., Pfeiffer, B.D., and Rubin, G.M. (2015). Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. Proc. Natl. Acad. Sci. USA *112*, E2967–E2976.

Nilsen, S.P., Chan, Y.B., Huber, R., and Kravitz, E.A. (2004). Gender-selective patterns of aggressive behavior in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *101*, 12342–12347.

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

Owald, D., Lin, S., and Waddell, S. (2015). Light, heat, action: neural control of fruit fly behaviour. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140211.

Panser, K., Tirian, L., Schulze, F., Villalba, S., Jefferis, G.S., Bühler, K., and Straw, A.D. (2016). Automatic segmentation of *Drosophila* neural compartments using GAL4 expression data reveals novel visual pathways. Curr. Biol. *26*, 1943–1954.

Penfield, W. (1950). The supplementary motor area in the cerebral cortex of man. Arch Psychiatr Nervenkr Z Gesamte Neurol Psychiatr *185*, 670–674.

Peng, H., Chung, P., Long, F., Qu, L., Jenett, A., Seeds, A.M., Myers, E.W., and Simpson, J.H. (2011). BrainAligner: 3D registration atlases of *Drosophila* brains. Nat. Methods *8*, 493–500.

Pfeiffer, B.D., Jenett, A., Hammonds, A.S., Ngo, T.T., Misra, S., Murphy, C., Scully, A., Carlson, J.W., Wan, K.H., Laverty, T.R., et al. (2008). Tools for neuroanatomy and neurogenetics in *Drosophila*. Proc. Natl. Acad. Sci. USA *105*, 9715–9720.

Pfeiffer, B.D., Ngo, T.T., Hibbard, K.L., Murphy, C., Jenett, A., Truman, J.W., and Rubin, G.M. (2010). Refinement of tools for targeted gene expression in *Drosophila*. Genetics *186*, 735–755.

Pfeiffer, K., Kinoshita, M., and Homberg, U. (2005). Polarization-sensitive and light-sensitive neurons in two parallel pathways passing through the anterior optic tubercle in the locust brain. J. Neurophysiol. *94*, 3903–3915.

Randlett, O., Wee, C.L., Naumann, E.A., Nnaemeka, O., Schoppik, D., Fitzgerald, J.E., Portugues, R., Lacoste, A.M., Riegler, C., Engert, F., and Schier, A.F. (2015). Whole-brain activity mapping onto a zebrafish brain atlas. Nat. Methods *12*, 1039–1046.

Renn, S.C., Armstrong, J.D., Yang, M., Wang, Z., An, X., Kaiser, K., and Taghert, P.H. (1999). Genetic analysis of the *Drosophila* ellipsoid body neuropil: organization and development of the central complex. J. Neurobiol. *41*, 189–207.

Schoofs, A., Hückesfeld, S., Schlegel, P., Miroschnikow, A., Peters, M., Zeymer, M., Spieß, R., Chiang, A.S., and Pankratz, M.J. (2014). Selection of motor programs for suppressing food intake and inducing locomotion in the *Drosophila* brain. PLoS Biol. *12*, e1001893.

Schwarz, R.F., Branicky, R., Grundy, L.J., Schafer, W.R., and Brown, A.E. (2015). Changes in postural syntax characterize sensory modulation and natural variation of *C. elegans* locomotion. PLoS Comput. Biol. *11*, e1004322.

Seelig, J.D., and Jayaraman, V. (2013). Feature detection and orientation tuning in the *Drosophila* central complex. Nature *503*, 262–266.

Simon, J.C., and Dickinson, M.H. (2010). A new chamber for studying the behavior of *Drosophila*. PLoS ONE *5*, e8793.

Sivanantharajah, L., and Zhang, B. (2015). Current techniques for high-resolution mapping of behavioral circuits in *Drosophila*. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. *201*, 895–909.

Strauss, R., and Heisenberg, M. (1993). A higher control center of locomotor behavior in the *Drosophila* brain. J. Neurosci. *13*, 1852–1861.

Träger, U., Wagner, R., Bausenwein, B., and Homberg, U. (2008). A novel type of microglomerular synaptic complex in the polarization vision pathway of the locust brain. J. Comp. Neurol. *506*, 288–300.

Triphan, T., Nern, A., Roberts, S.F., Korff, W., Naiman, D.Q., and Strauss, R. (2016). A screen for constituents of motor control and decision making in *Drosophila* reveals visual distance-estimation neurons. Sci. Rep. *6*, 27000.

Vogelstein, J.T., Park, Y., Ohyama, T., Kerr, R.A., Truman, J.W., Priebe, C.E., and Zlatic, M. (2014). Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. Science *344*, 386–392.

von Philipsborn, A.C., Liu, T., Yu, J.Y., Masser, C., Bidaye, S.S., and Dickson, B.J. (2011). Neuronal control of *Drosophila* courtship song. Neuron *69*, 509–522.

Walker, R.G., Willingham, A.T., and Zuker, C.S. (2000). A *Drosophila* mechanosensory transduction channel. Science *287*, 2229–2234.

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

Wu, M., Nern, A., Williamson, W.R., Morimoto, M.M., Reiser, M.B., Card, G.M., and Rubin, G.M. (2016). Visual projection neurons in the *Drosophila* lobula link feature detection to distinct behavioral programs. eLife *5*, e21022.

Yamamoto, D., and Koganezawa, M. (2013). Genes and circuits of courtship behaviour in *Drosophila* males. Nat. Rev. Neurosci. *14*, 681–692.

Young, J.M., and Armstrong, J.D. (2010). Structure of the adult central complex in *Drosophila*: organization of distinct neuronal subsets. J. Comp. Neurol. *518*, 1500–1524.

Yu, J.Y., Kanai, M.I., Demir, E., Jefferis, G.S., and Dickson, B.J. (2010). Cellular organization of the neural circuit that drives *Drosophila* courtship behavior. Curr. Biol. *20*, 1602–1614.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
rabbit anti-GFP	Thermo Fisher Scientific	Cat #: A-11122; RRID: AB_221569
Alexa Fluor 488-conjugated goat anti-rabbit	Thermo Fisher Scientific	Cat #: A-11034; RRID: AB_2576217
Alexa Fluor 568-conjugated goat anti-mouse	Thermo Fisher Scientific	Cat #: A-11031; RRID: AB_144696
Biological Samples		
Drosophila: GAL4 collection see Table S1	Pfeiffer et al., 2010	See Table S1
Drosophila: Split GAL4 lines see Table S2	this paper	http://splitgal4.janelia.org
Drosophila: *+;UAS-dTRPA1/(CyO);+*	Hoopfer et al., 2015	N/A
Drosophila: w^{1118}; ; *pBDPGAL4U (attP2)*	Pfeiffer et al., 2010	N/A
Drosophila: *w^{1118};pBPp65ADZpUw (attP40); pBPZpGAL4DBDUw (attP2)*	Hampel et al., 2015	http://splitgal4.janelia.org
Drosophila: *pJFRC2-10xUAS-IVS-mCD8::GFP (attP2)*	Pfeiffer et al., 2010	FlyBase: FBst0032185
Deposited Data		
Janelia FlyLight imagery data	Jenett et al., 2012	http://flweb.janelia.org/
Drosophila template brain	Jenett et al., 2012	JFRC2/JFRC2010, https://github.com/VirtualFlyBrain/DrosAdultBRAINdomains
Drosophila brain mask	Jenett et al., 2012	https://github.com/VirtualFlyBrain/DrosAdultBRAINdomains
Software and Algorithms		
MATLAB	MathWorks	https://www.mathworks.com/
FlyBowlDataCapture GUI	this paper	SuppData/code/FlyBowlDataCapture
Ctrax: The Caltech Multiple Walking Fly Tracker	modified from Branson et al., 2009	SuppData/code/Ctrax
Wing tracking software	this paper	SuppData/code/simplewing
Sex classification software	this paper	SuppData/code/FlyBowlAnalysis
JAABA: Janelia Automatic Animal Behavior Annotator	Kabra et al., 2013	http://jaaba.sourceforge.net/
Brain Aligner	Peng et al., 2011	https://github.com/Vaa3D
BABAM: Browsable Atlas of Behavior-Anatomy Maps	this paper	https://kristinbranson.github.io/BABAM/
Other		
Fly Bowl arena design	this paper	SuppData/design
Fly Bowl behavior results	this paper	http://research.janelia.org/bransonlab/FlyBowl/BehaviorResults/
Supervoxel clustering data	this paper	SuppData/data/SupervoxelClusteringData.mat
JAABA behavior classifiers	this paper	SuppData/params/JAABAClassifiers

CONTACTS 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, Kristin Branson (bransonk@janelia.hhmi.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

2,205 *Drosophila melanogaster* lines were selected from the Janelia GAL4 collection (Jenett et al., 2012) based on their expression patterns (Table S1). The negative control line used for behavioral comparisons was pBDPGALU (a *GAL4* insertion without a promoter in w^{1118};*attp2*) (Pfeiffer et al., 2008). The Janelia GAL4 collection was created using an injection stock that was re-isogenized every 1-2 years, but subsequent crosses were made to remove the integrase and homozygose each line (Janelia Fly Core personal communication). The *GAL4* lines were used to drive the expression of *dTRPA1* using the effector *+;UAS-dTRPA1/(CyO);+* (X and third chromosomes backcrossed to Canton-S from M. Heisenberg for six generations).

For the *GAL4* lines with high behavior-anatomy scores for the increased-walking map, we obtained AD and DBD stocks that had previously been constructed using those promoter sequences (provided by G. Rubin prior to publication). These AD (in attP40) and DBD (in attP2) stocks were constructed as previously described (Pfeiffer et al., 2010). We used combinations of these AD and DBD stocks to make the set of 20 split-GAL4 stocks listed in Table S7. The negative control line used for behavioral comparisons of the split-GAL4 lines was an empty split-GAL4 line made from AD and DBD without enhancers to drive expression (Hampel et al., 2015).

In all behavior experiments, flies were approximately 3-5 days old, and sorted into groups of 10 males and 10 females 2-4 days prior to video recording. To ensure that our genetic control flies were sufficiently active, all flies were wet-starved for 1 day prior to video recording during the flies' evening activity peak.

For our assay of the Janelia GAL4 collection, all flies used were reared in standard vials on dextrose-based food in a 16/8 light/dark cycle at 21.8°C and 55% relative humidity (RH). The parental flies were crossed on day 0 (cross date), after 4 days were transferred to fresh food vials (flip date) and then discarded after an additional 3 days of egg-laying. Egg laying periods were restricted to limit the age-range of flies collected for experiments. The first set of cross vials were used for a different assay and the post-flip set were used the Fly Bowl screen. 14 days after the flip date, the offspring were collected and sorted into groups of 10 males and 10 females under cold anesthesia. These experimental flies were maintained on dextrose media until 24 ± 2 hr prior to the experiments, at which point they were transferred to agar-based media for wet starvation. Experiments were performed 2-4 days after collection, and as no newly eclosed flies were selected, the majority of flies used were 3-5 days post-eclosion. Two sets of rearing incubators were used: AM incubators programmed set for lights off at 12:00 and PM incubators were set for lights off at 17:00. All experiments were done during the respective evening activity peak – 8:00 to 12:00 for morning experiments and 13:00 to 17:00 for afternoon experiments.

Split-GAL4 line flies for behavioral experiments were raised as above, except that the experimental flies were collected from eggs laid during the first 4 days of egg-laying.

To image the expression pattern of the split-GAL4 set, each line was crossed to *pJFRC2-10XUAS-IVS-mCD8::GFP* in attP2 (Pfeiffer et al., 2010). The dissections, immunohistochemistry, imaging and registration of the split-GAL4 set used protocols modified from Aso et al. (2014). Brains (~4 per line) and VNCs (~2 per line) were dissected from 3-7 day old flies in Schneider's insect medium and fixed in 2% paraformaldehyde (PFA) at room temperature for 55 min. Tissues were washed in PBT (0.5% Triton X-100 in phosphate buffered saline (PBS) and blocked using 5% normal goat serum). Primary antibodies (rabbit anti-GFP A-11122 from Molecular Probes at 2 μg/ml and mouse anti-Bruchpilot nc82 hybridoma supernatant from Developmental Studies Hybridoma Bank (University of Iowa) diluted 1:30) were applied for 2-3 days. After a rinse and four 15 min washes in PBT, tissues were then incubated for 2-3 days in secondary antibodies (Alexa Fluor 488-conjugated goat anti-rabbit and Alexa Fluor 568-conjugated goat anti-mouse from Molecular Probes; 2.5 μg/ml and 5 μg/ml, respectively). Each of the antibody incubations were done for 4 hr at room temperature before placing the samples at 4°C for the remainder of the incubation time. After a rinse and four 15 min washes in PBT, tissues were fixed with 4% PFA in PBS for 4 hr, followed by a rinse and four 15 min washes in PBT. Directly before mounting, tissues were rinsed and washed for 15 min in PBS to remove the Triton. The tissues were mounted on poly-L-lysine-coated coverslips and then dehydrated with 10 min ethanol baths of 30%, 50%, 75%, 95% and 3 × 100% followed by three 5 min washes in 100% xylene. Finally, mounted samples were embedded in xylene-based mounting medium (DPX; Electron Microscopy Science, Hatfield, PA) and dried for 2 days. Images were collected using an LSM710 confocal microscope (Zeiss, Germany) fitted with a Plan-Apochromat 20x/ 0.8 M27 objective. Brain images were aligned to the JFRC2013 standard brain (Aso et al., 2014) using the Janelia Brain Aligner (Peng et al., 2011).

METHOD DETAILS

Data collection

We simultaneously recorded 4 16 min videos, each of groups of 10 male and 10 female flies from the same GAL4 line at the permissive temperature for dTRPA1 of 30°C. For the majority of lines, we collected videos twice at different times of the year, thus for each GAL4 line, we collected on average 8 videos (Figure 3B). Each morning and afternoon of data collection, we also collected 8 videos of the genetic control line, pBDPGAL4U, 4 videos per rig. These data were used as indicators of issues with fly rearing and data collection. In addition, to account for small changes in the flies' behavior over time, we normalized our behavior measurements based on those of control flies in the surrounding month.

Apparatus

We developed a high-throughput behavioral assay for walking fruit flies. The arena shape for this assay was based on a design developed by Simon and Dickinson (Simon and Dickinson, 2010), with modifications for high-throughput data collection, experimental monitoring and consistency, and image quality (CAD designs included in Data S1). Each Fly Bowl arena was shaped like a shallow plate, with a total diameter of 127 mm, center depth of 3.5 mm, and edges that gently sloped up at an 11-degree angle starting at a diameter of 79 mm. Each arena was undercut to a depth of 5 mm following the contour of the arena to create a uniform floor thickness (we later found this to be an unnecessary design feature but maintained it for consistency).

Four of these arenas, arranged on a 127 mm grid, were milled into a 12.7 × 305 × 381 mm clear polycarbonate plate. This plate was secured to a 9.5 × 305 × 381 mm anodized aluminum backing-plate to ensure flatness. The ceiling of the 4 arenas was a single 6.35 × 305 × 381 mm lime-soda laminated safety glass sheet with four counter bored 3.18 mm diameter loading holes. The bottom of this glass sheet was treated with Sigmacote, Sigma-Aldrich, St Louis, MO, a siliconizing reagent that was slippery to flies and prevented the flies from walking on the ceiling of the arena. The Sigmacote was reapplied once per week at least 24 hr before data collection.

We built two video recording rigs, each constructed around a 4-arena plate with each Fly Bowl arena video-recorded by a separate camera. In each rig, the arena plate was mounted on four 54 mm standoffs, each with a set screw in the top that aligned to the outmost through-hole in the corners of the plate. This secured the lateral positioning of the arenas in the recording apparatus.

The arenas were backlit from below by a 330 × 330 mm 880nm near-IR LED panel fabricated by Advanced Illumination, Rochester, VT. This backlight was cooled in two ways: it was mounted on a 305 × 305 mm water-cooled bread board, Thorlabs, Newton, NJ, and eight 45 mm, 12VDC, 10.4 CFM computer fans, Mouser Electronics, Czech Republic, created ambient temperature airflow between the backlight and plate (four each on opposing sides but blowing the same direction). The water-cooled bread board was cooled by a radiator system with a circulating pump, Koolance, Auburn, WA. A 3.18 × 305 × 305 mm white acrylic diffuser was placed directly on top of the backlight (assembly schematic of 4-arena plate and backlight included in Data S1).

Each of the four arenas in the plate were video-recorded from above with an A622f FireWire camera, Basler, Inc., Exton, PA, using a 25 mm fixed focal length lens and an IR pass filter (Hoya R-72), both from Edmund Optics, Barrington, NJ. Each camera was mounted with an XZ 0.5-in translation stage on a dovetail rail carrier Thorlabs Newton, NJ which allowed fine-scale X, Y, Z adjustments to position each arena in the field of view of the respective camera. The cameras were mounted 472 mm above the arenas as measured from the top of the arena plate to bottom of the dovetail rail on which the cameras were mounted.

We optimized the camera settings for image quality and frame rate: the region-of-interest was set to 1024 × 1024, the brightness to 25, the gain to 170, and the shutter to 100. Images were collected in Format 7 Mode 0, which allowed the maximum frame rate for data acquisition. The frame rate for video collection was 30.3 frames per second, and we used the median frame rate for a given movie in calculating the per-frame behavior metrics.

The visual environment of the flies during experiments was carefully controlled. A visual surround, a 535 mm diameter and 415 mm tall cylinder with white interior walls and blackout material external walls, was place around the plate after the flies were loaded to block any external visual cues such as experimenters moving around. A fixed white acrylic sheet (585 × 585 mm), mounted 40 mm below the cameras with openings for each camera, provided a consistent overhead visual environment. To provide visible light at 1340-1400 lux for the flies, this sheet was backlight from above with nine evenly spaced natural daylight LED lamps, Buylighting.com, Burnsville, MN.

Environmental control and monitoring

All experiments were performed in a walk-in environmental control chamber. The 8-ft x 12-ft chamber controlled temperature and humidity, and was custom built by Bahnson Environmental Specialties, LLC (Raleigh, NC). For dTRPA1 activation, the chamber temperature and humidity were set to 29°C and 50% RH respectively. We measured an average temperature of 29.7°C and humidity of 47% RH in the chamber during all experiments included in the analysis (outliers in temperature and humidity were excluded from the analysis (see *Data curation,* below). As mentioned above, the IR backlight generated heat that caused a relative increase in the Fly Bowl plate temperature. The average temperature of the Fly Bowl plates across the screen was 29.9°C, and reflects the shifted temperature the flies experienced during experiments. We turned the lights on for each rig at least 30 min prior to each recording session to ensure the temperature stabilized before data collection began.

We automatically recorded the temperature and humidity of the environmental chamber during each experiment using a Precon HS-2000D sensor, Kele Precision Manufacturing, Memphis, TN. Additionally, the temperature of the polycarbonate plate adjacent to each walking arena was measured and logged once per second during video collection with a T-type thermocouple PT-6 Physitemp Instruments, Inc., Clifton, NJ and a TC-08 thermocouple data acquisition module, OMEGA Engineering, Inc., Norwalk, CT.

Data capture software

Videos were recorded using a Basler A622f camera for each of the 8 arenas at a temporal resolution of 30.3 frames per second (fps) and a spatial resolution of 1024x1024 pixels (7.9 pixels/mm). We created custom MATLAB software (FlyBowlDataCapture (FBDC) GUI, Figure S2A, Data S1) to collect all behavior videos and to record experiment metadata such as the genotype and room temperature.

This GUI was initialized with a parameter file that controlled the details of the experimental set-up such as camera parameters, video capture and compression parameters, temperature and humidity capture parameters, and ranges and values for metadata input. Use of the FBDC GUI, the parameter file entries, and metadata definitions and xml file are detailed in the FBDC User Guide in the Data S1 and S3.

After initializing the GUI, the experimenter selected the camera and temperature probe for a given bowl, and the FBDC GUI displayed the live video, frame rate, and temperature readings throughout the experiment (Figure S2A(i)). Next, the experimenter brought a set of four vials containing flies of one genotype into the environmental chamber, and recorded the timing of this by pressing the "Shift Fly Temp" button in the FBDC GUI. Then, the experimenter loaded the pre-sorted vials of 10 male and 10 female flies into the four bowls, one vial per bowl. Loading was facilitated by a custom designed device with a 3-D print fly vial-to-culture tube adaptor, allowing the experimenter to tap the entire group of flies from the food vial into a small chamber made from a culture tube and then mouth aspirate the group of flies into the arena through the loading hole in the glass cover plate. After loading, this hole was plugged to prevent flies escaping while the other bowls were loaded. The time point of loading each bowl, when the flies were shifted the permissive temperature, was recorded by pressing the "Flies Loaded" button in the respective FBDC GUI. This also started a count-down timer that indicated to the experimenter how much of the 225 s loading period remained. When all four bowls were loaded, the glass cover plate was slid into the recording position with the loading holes outside of the bowls, the hole plugs removed, and the visual surround put into place around the four bowls. At this point, the experimenter started video collection by pressing the "Start Recording" button.

During the recording, the experimenter could enter behavior observations as well as notes about technical issues during the data collection. After the recording finished, the experimenter scanned the barcode attached to each fly vial, automatically populating metadata information such as genotype, cross date, sorting time, and other information about the flies in the experiment.

Finally, the FBDC software showed the experimenter the recorded video and a technical summary of the experiment (Figure S2A(ii)), which allowed the user to quickly detect and flag any experiments with technical problems. When finished, the experimenter pressed the "Done" button and reset the FBDC GUI for the next experiment. To reset the rig for the next experiment, the experimenter removed the visual surround, anesthetized the flies with CO_2, and removed them from the bowls with vacuum suction. Then the experimenter cleaned the bowls with 70% ethanol and allowed them to dry for at least 1 min before the glass cover plate was replaced.

The FlyBowlDataCapture program created a single directory for each experiment. Within this directory, all files had the same name. Important files recorded into this directory were the video (movie.ufmf), an XML file containing all metadata (Metadata.xml), and a temperature log (temperature.txt).

Real-time video compression

To control the camera and record the video, we used the MATLAB Image Acquisition Toolbox. In real time, we compressed video into the micro-Fly Movie Format (ufmf) using a custom C++-based adaptor (*udcam*, Data S2). The ufmf compression is based on background subtraction, and first stores an estimate of the appearance of the video if no flies were present (the background image, Figure S2B). Then, for each frame, it stores only the locations and intensities of pixels that are very different from the background appearance (foreground pixels, Figure S2A(iii-iv)). The majority of these foreground pixels correspond to the flies in the arena.

The background image is computed as the median pixel intensity across a sample of frames:

$$B_{uv} = \text{median}\left(I_{uv}^{t_0}, I_{uv}^{t_0+\tau}, I_{uv}^{t_0+2\tau}, \ldots, I_{uv}^{t_1}\right),$$

where B_{uv} is the background image appearance at pixel location (u,v), I_{uv}^{t} is the intensity of pixel (u,v) in frame t of the video, τ is the sampling period, and t_0 and t_1 are the start and end frames of the video clip. To estimate the median efficiently, we update a histogram of the number of sampled frames with each of the 256 possible intensity values. We require a 256-bin histogram for each pixel location (u,v). Foreground pixels are those that are very different from the background intensity: $|I_{uv} - B_{uv}| \geq \lambda$.

Early in the video (e.g., at frame 2), we have not collected enough frames to estimate a good model of the background appearance (the initial background estimate, computed from just the first frame, was the first frame itself), and there is the potential that the background appearance changes over time. Thus, we output estimates of the background appearance at set intervals during the video collection. An inaccurate background appearance model primarily affects compression rate, not video quality.

Data curation

While we carefully engineered our data capture system to minimize errors, both mistakes and technical failures still occurred over the course of collecting 20,288 videos. To identify and remove these experiments from our dataset, we employed both automatic and manual methods to curate experiments.

Experiments were automatically failed by the analysis pipeline after data transfer if the: (1) video file or metadata file did not exist, (2) time the flies were in the environmental chamber before loading exceeded 300 s, (3) loading time was less than 25 s or greater than 225 s (4) video file was truncated, (5) barcode was not recorded, or (6) experimenter set the "Redo" flag during data collection. A second automatic quality control check was performed at the end of the analysis pipeline. Experiments were automatically failed if: (1) they were missing the results of any of the pipeline steps, (2) the number of flies tracked was less than 18 or more than 22, or (3) the number of male or female flies was less than 7. The automatic curation steps resulted in the failure of 601 experiments.

After passing the automatic quality control, each experiment was also manually curated. Additionally, each experiment with a review flag or a technical note were assessed for issues. In order to further detect possible errors, we developed a software GUI (FlyBowlExamineVariables, Data S1 and S2) to review the metadata and diagnostic statistics from data collection and the analysis pipeline. Outliers in these variables were identified and reviewed for problems. In total, 622 experiments were manually failed and removed from the dataset for the following reasons: 220 videos had dead or damaged flies in the arena, 167 videos had irreconcilable line name swaps or possible line contamination, 56 videos had debris in the arena, 53 videos had temperatures below 29°C, above 30.5°C, or humidity below 40% RH, 31 videos contained lines with wing-development phenotypes, 24 videos had loading procedure errors, 22 videos had data capture software errors, 16 videos had issues with the environmental chamber during the experiment, 7 videos contained flies with irregular rearing conditions, and in 1 video the number of males or females did not meet the required minimum. The thorough and complete curation of the data collected in the GAL4 screen removed any potential confounds from the data analysis due to experiments with identifiable errors.

Tracking

Tracking fly bodies using Ctrax

In each frame of each video, we fit an ellipse to the body of each of the 20 flies using a modified version of Ctrax version 0.3.1 (Branson et al., 2009). Ctrax uses background subtraction to classify pixels as either fly body (not wings or legs) or background, then connected components analysis and Expectation-Maximization for Gaussian mixture models to cluster the foreground pixels into groups corresponding to each fly. The Gaussian distribution fit to each fly's pixels is used to estimate an ellipse fit to the body. Identities are matched between frames using a dampened constant velocity model. Head-tail ambiguities in the ellipse orientation are resolved using a dynamic programming algorithm that maximized consistency in both the fly's orientation between pairs of frames and the fly's orientation and velocity direction. We made two main modifications to Ctrax to allow it to run completely automatically on our large collection of videos of behaviorally diverse animals (*Ctrax modifications*, below*).*

We ran Ctrax on each video using its command-line interface on Janelia's cluster, allowing the 100 videos collected in a day to be processed overnight. We used the same tracking parameters for all videos (with a different pooled background and foreground model for the first set of Fly Bowl plates used before August 1, 2011 than the second set of plates used after this date). All tracking parameter files are included in Data S1 and S3. Tracking parameters were initially set by observation on a handful of videos. Ctrax was then run on a larger subset of videos using these parameters. The results were examined, and several heuristics were used to select tracking parameters that would work well on this larger dataset. These parameters were used for the rest of the video collection.

Ctrax modifications

The originally published version of Ctrax uses the median of each pixel appearance over time to estimate the background appearance – the appearance of the arena if no flies were present, which is only accurate if the flies are sufficiently active. Thus, the original background estimation algorithm failed for lines that had greatly decreased activity. To ameliorate this, we learned pooled models of the background and foreground appearance from a collection of 16-50 videos of sufficiently active GAL4 lines (Figure S2B). When learning the background appearance for a new video, we estimated the probability that a pixel in a frame is foreground based on these pooled models. To estimate the background appearance for a new video, we used a weighted median of pixel appearance, in which each pixel is weighted by the probability that it is foreground (as opposed to each pixel having equal weight in the original version of Ctrax).

The learned foreground and background models are based on a combination of the pixel's location in the image, its appearance, and its distance in pixels from an object detection. The appearance of a pixel is its grayscale intensity. Object detections are based on thresholding and non-maximal suppression of the output of Laplacian of Gaussian filtering. Let l indicate the label of a given pixel (either foreground or backgroud), x the location of the pixel in the image, y the appearance of the pixel in the current image, and d the distance to the closest object detection. By assuming conditional independence between the different types of features given the pixel label, we can compute the probability that a pixel is foreground or background by multiplying marginal likelihoods:

$$P(l \mid x, y, d) \propto p(y \mid l, x) p(d \mid l) P(l \mid x)$$

To decompose the probability, we assume that the distance to an object detection d is conditionally independent to both pixel appearance y and location x given the label l: $d \perp y \mid l$, x and $d \perp x \mid l$.

We model the pixel appearance given its label and location, $p(y \mid l, x)$ as a Gaussian, with mean and variance fit as the sample mean and variance (Figure S2B(i-ii)). We estimate the distribution of the distance to an object detection $p(d \mid l)$ using a sample histogram (Figure S2B(iii)), and the label marginal given location $P(l \mid x)$ by the sample fractions.

When estimating the background appearance as the median of each pixel over time, we use the foreground probability $P(\mathit{foreground} \mid x, y, d)$ to weight each pixel in the median computation.

Python code for estimating these foreground and background models is included in Data S1 and S3.

We also modified Ctrax to output diagnostic statistics of the number of trajectory births and deaths, as well as statistics of the number of merges and splits resolved using the Ctrax "hindsight" module.

Wing tracking

We tracked the wings of the flies using custom MATLAB software which implemented an image morphology-based algorithm (Data S1). Our wing tracking algorithm operates on each frame independently, and begins by classifying pixels as body, wing, or background by thresholding pixel intensity difference between the current frame and the background appearance (estimated using Ctrax), followed by a sequence of morphological erosions and dilations to incorporate some spatial information into this classification (Figure S2D(i)).

Next, pixels classified as wing are then assigned to individual flies based on the path distance to the back of the fly's body through wing or body pixels (Figure S2D(ii-iii)). To do this, we perform connected component analysis on the foreground (either wing or body) pixels, and assign each fly identity output by Ctrax to a connected component. We assign a fly to the closest connected component using the Mahalanobis distance (according to the tracked body covariance matrix) between the tracked fly center and any pixel in the connected component (Figure S2D(ii)). Small, unassigned connected components are ignored. Large, unassigned connected components are assigned to the nearest fly, and merged with that fly's originally assigned connected component.

If a connected component is assigned to exactly one fly, then all pixels in that connected component are assigned to that fly. Otherwise, we need to segment the pixels in the connected component and assign them to individual flies. Body pixels are assigned to flies according to the Mahalanobis distance to the tracked fly center. Wing pixels are over-segmented using the watershed algorithm, and each cluster of wing pixels is assigned to a single fly based on the shortest-path distance to body pixels in the back half of the fly (Figure S2D(iii)). We then remove spurious wing pixel detections (usually corresponding to legs) that are outside of the interval $[-(3\pi/4),(3\pi/4)]$ relative to each fly's tail are ignored.

Once the wing pixels for a given fly have been identified, we histogram the angle from the center of the fly to each of its assigned wing pixels angles. We then find peaks in this histogram, and assign each wing's angle to a peak (Figure S2D(iv)). We find up to two peaks in this histogram, which will roughly correspond to the wing angles. If two peaks are found, we obtain sub-histogram-bin precision in our estimate of the wing angles by fitting quadratic functions to the histogram counts for each wing, and finding the maximum of the quadratic. These angles are then returned as the wing angles. If a single peak is found, we instead compute the median angle to all wing pixels, and return this as the wing angle for both wings. If no peaks are found, then we return that both wings are directly behind the fly (Figure S2D(iv)).

Temporal and spatial registration of trajectories

We analyze portions of the flies' trajectories corresponding to 225 to 1,015 s after the flies have been loaded into the Fly Bowl arena (the time when flies are loaded into arenas is recorded as part of the data capture).

For each video, using landmarks automatically detected in the background image output by Ctrax, we scale, translate, and rotate the trajectories so that they correspond to a common coordinate system in the world. We scale the trajectories so that a unit corresponds to a millimeter, translate them so that (0,0) is the center of the Fly Bowl arena, and rotate them so that the x- and y-axes correspond to the same directions in all arenas. Thus, small changes in *e.g.* the camera position will have little effect on the trajectories. The landmarks detected are the circular arena wall (detected using Canny edge detection and the Hough transform) and a registration mark (detected using cross correlation).

Classification of fly sex

Each video corresponded to a mixed population of approximately 10 males and 10 females. In order to obtain measures of the behaviors of males and females separately, for example to find sex-specific behavioral effects, we automatically annotate the sex of each fly in each video. As female flies are bigger than males, we do this using the area of the ellipses fit to the fly by Ctrax. We annotate the sex of each fly in each frame, in case there are identity swaps in the tracking, but assume that these are rare, thus, in most cases, the sex assigned in consecutive frames for a given fly should be the same. We combine these two heuristics (that females are larger than males and that the sex assigned in consecutive frames should be the same) via a hidden Markov model (HMM) whose parameters are fit for each video, and find the globally optimal sequence of sex classifications for each trajectory.

We do this using an HMM in which the hidden states are the sex classifications for the fly for each frame and the observations are the areas of the ellipses fit by Ctrax. Let x_t represent the unknown sex classification for the fly at frame t and y_t the area of the ellipse fit to the fly by Ctrax at frame t. Using the first-order HMM conditional independence assumptions, we can factor the joint posterior distribution of sex classification at each frame given the observed ellipse areas as

$$P(x_{1:T}\,|\,y_{1:T}) \propto \prod_{t=1}^{T} P(x_t\,|\,x_{t-1})p(y_t\,|\,x_t)$$

Here, the transition probability $P(x_t\,|\,x_{t-1})$ is fixed so that it is much more likely that $x_t = x_{t-1}$. The appearance likelihood $p(y_t\,|\,x_t)$ is modeled as a Gaussian whose parameters are learned for a specific video using the Baum-Welch algorithm. Once the parameters of the HMM have been learned, we use the Viterbi algorithm to find the globally optimal sequence of sex classifications for the given fly. Our custom MATLAB code for implementing this automatic sex classification is provided in Data S1.

To evaluate the accuracy of our sex classification algorithm, we manually annotated the sex of 1,000 randomly selected flies and frames within 7 videos from the GAL4 line R14C07 (chosen because the males showed an increase in chasing, thus behavior could be used to aid in manual sex classification). We compared the automatic classifications of our algorithm to these manual annotations. Our classification algorithm was 99% accurate (Figure S2C).

We expect sex classification to be similar across GAL4 lines, as the sizes of the flies did not vary much between lines, and the standard deviation across lines was similar to that across retests of the same line. In more detail, the mean difference in female and male area is 0.47 mm^2, with standard deviation 0.040 mm^2 across all lines screened, while the average standard deviation in area difference across retests was 0.038 mm^2. The mean area difference for R14C07 was 0.5175 mm^2 with a standard deviation of 0.0672 mm^2 across retests.

JAABA behavior classifiers

As described in Kabra et al., we created 14 behavior classifiers using our JAABA interactive machine learning software (Kabra et al., 2013). Each binary classifier corresponded to a different behavior, with positive examples being frames in which the fly is performing the behavior of interest and negative examples being frames in which the fly is not performing the behavior. These classifiers are not mutually exclusive — a fly can perform multiple behaviors at the same time, such as performing walk, chase, and wing extension simultaneously. A brief description of criteria used to train each behavior classifier is listed below (examples of each behavior are shown in Movie S1). We also created a copulation classifier but only one line, R34H05, had a strong and repeatable increase in copulation so we excluded copulation from further analysis.

Attempted copulation – the fly curls its abdomen in preparation for copulation (also called abdominal bending).

Backup – the fly moves at least two legs with a backward swing phase and the center of mass of the fly moves backward without much rotation about the body axis.

Chase – one fly follows another moving fly and maintains a small, somewhat constant distance to the fly it is following (not dependent on sex of flies, and does not require co-occurrence with wing extension).

Crabwalk – the fly translates sideways with the angle between its velocity vector and the orientation of the body axis greater than $\sim 60^\circ$.

Jump – the fly launches itself into the air. This is often accompanied by wing movement, but this classifier does not use wing-based features.

Pivot-center – the fly performs a large and rapid change in orientation in which the center of rotation is close to the centroid of the fly and little to no translational movement.

Pivot-tail – the fly performs a large and rapid change in orientation in which the center of rotation is close to the end of abdomen, followed by translational movement in the new orientation.

Righting – to return itself to an upright posture from a supine posture, the fly performs high velocity rolling to and fro before succeeding in righting itself.

Stop – the fly has almost no translational or rotational body movement. Stop is not the lack of all movement — the fly can perform grooming movements and touches while stopped.

Touch – the fly makes contact with the body, wings, or legs of another fly, using its body or legs. This behavior is non-transitive: a fly is not necessarily performing a ‘touch’ because it is being touched by another fly.

Walk – the fly moves forward using approximately alternating movements of the mesothoracic legs.

Wing extension – the fly unilaterally rotates a wing out from its body in the horizontal plane, often for extended periods of time. This behavior is likely an indication of the fly producing courtship song with the extended wing.

Wing flick – the fly rapidly and symmetrically moves its wings out and back into their rest position, performing a quick scissoring movement in the horizontal plane. This is often done multiple times in a row.

Wing grooming – the fly wipes its mesothoracic leg over the dorsal surface of a wing, folding it down against its abdomen.

Machine learning algorithms are based on the assumption that the distributions of the training and test data are the same. As the behavior of many of our GAL4 lines are extremely different from that of our genetic control, this assumption would be violated if we trained only on our genetic control and applied the resulting classifier to our GAL4 lines. Conversely, it would be impractical to train on data from all of our 2,205 GAL4 lines. To efficiently train classifiers that were accurate across all 2,205 phenotypically-diverse GAL4 lines, we used the following protocol. First, we trained an initial version of the classifier on flies from our genetic control. For behaviors such as attempted copulation which are rarely performed by control flies, to easily obtain positive examples of the behavior, we included GAL4 lines we observed performing canonical versions of the behaviors at a higher rate. We then applied this initial classifier to our entire video dataset. We visually inspected the output of this initial classifier on GAL4 lines that were outliers in line-level behavior statistics based on per-frame features, for example the average speed of the animals or the average distance to the closest other fly. When classification errors were identified, we added new training data to refine and retrain the classifier. We also visually inspected and added training data for lines for which the co-occurrence rates of pairs of behaviors were abnormal, and lines for which selected per-frame features had abnormal values within positively classified frames. As a final check for false positives, we applied the refined classifier to the entire video collection and visually confirmed the behavior for the 50 lines that were classified as performing the behavior the most.

Details of training each classifier, including parameters and training set sizes, are given in Table S3.

Accuracy of automatic behavior classifiers

To quantitatively measure the generalization error of the classifiers across the entire dataset, we manually labeled the behavior in selected GAL4 lines, and compared these manual labels with the output of the automatic behavior classifiers (Figure 2C, Table S4). To manually label data, we used the JAABA software in “groundtruthing mode.” Intervals of frames to label were selected randomly using the “balanced random” method, which set the probability that a given interval of frames was selected so that approximately the same number of frames would be selected that were classified as positive and negative. Thus, for behaviors that were performed rarely by a given line, the intervals containing positively classified examples were up-weighted. To obtain an unbiased estimate of accuracy, we then scaled by the inverse of this weight. This importance-sampling approach was necessary to estimate the false negative rate for rare behaviors from a reasonable number of manual labels.

For each behavior, we manually labeled up to 2,000 frames in the GAL4 lines that were at the 0th, 5th, 15th, 50th, and 100th-percentile for fraction of time spent performing that behavior (according to the automatic classifier), and the control line. For 3 classifiers (attempted copulation, jump, and wing extension), we performed more extensive groundtruthing, and manually labeled frames in an additional 17 (totaling 23) representative lines:

- To find lines for which the classifier might be performing poorly, we selected the 2 lines with the highest fraction of frames with low-confidence classifications for the given behavior (classifier scores between -.25 and 0.25).
- To find lines that well-represented the diversity of observed behavior phenotypes, we clustered the lines into 10 groups using k-means based on their line-level behavior statistics, and selected the line closest to the cluster mean as a representative.
- To find lines with unusual behavior, we ranked the lines by the fraction of frames which were not classified as any of our 14 behaviors, and sampled the 0th, 5th, 15th, 50th, and 100th-percentile lines.

The average accuracy of each classifier across all lines is shown in Figure 2C, and the accuracy for each GAL4 line is shown in Table S4. Note that the accuracy is high for all lines that perform the behavior a reasonable amount of the time.

We further examined the effect of behavior classification errors on the accuracy of our line-level behavior statistics (Figure S3A). For the lines with groundtruth behavior labels, we compared the line-level statistics calculated from the subset of frames we manually labeled with those from the automatic behavior classification across all frames. These comparisons show good concordance between statistics from manually and automatically classified behaviors.

Computation of per-line behavior statistics

Following the JAABA analysis (Kabra et al., 2013), we computed a suite of 128 continuous-valued *per-frame features*, describing the instantaneous locomotion of each fly, and their relative positions and motion (Table S2).

These per-frame features were used in two ways. First, they were the inputs to the 14 JAABA behavior classifiers. Second, we used both the per-frame features and the per-frame behavior annotations to compute a suite of summary behavior statistics to summarize the behavior of the flies in a video (*per-video* behavior statistics). Each of our behavior statistics is defined by a *per-frame feature* (Tables S2 and S5) and conditions on which frames and flies to include in the computation (Table S5): <perframe>_fly<flycondition>_frame<framecondition>.

For example, velmag_ctr_flymale_framechase, which corresponds to the average velmag_ctr per-frame feature (velocity magnitude of the fly center) in fly-frames which are classified as both males and chasing, is computed by pooling together all frames that satisfy the fly and frame conditions, and computing the average of the per-frame feature in those frames:

$$\frac{\sum_{i=1}^{nflies}\sum_{t=1}^{T_i} I(chase_{it}=1)I(sex_{it}=M)velmag\_ctr_{it}}{\sum_{i=1}^{nflies}\sum_{t=1}^{T_i} I(chase_{it}=1)I(sex_{it}=M)},$$

where *nflies* is the number of trajectories, T_i is the length of trajectory *i*, $chase_{it}$ is the binary output of the chase classifier for trajectory *i* and frame *t*, sex_{it} is the output of the sex classifier for trajectory *i* and frame *t*, $I(\bullet)$ is the indicator function, and $velmag\_ctr_{it}$ is the instantaneous speed of fly *i* between frames *t* and $t+1$.

Our code for computing these per-video behavior statistics is modular, and new behavior statistics can be specified using a controlled vocabulary without coding. Any of the per-frame features in Table S2 as well as a few special case keywords (Table S5, type = perframe) can be specified as a per-frame feature. We define a controlled vocabulary of conditions on the frames (Table S5, type = framecondition) and flies (Table S5, type = flycondition) whose data is pooled to compute a given behavior statistic, specified in Table S5.

Our behavior statistic computations pool together frames across flies in the video, all with equal weight. We exclude very short trajectories from this analysis: we specify a minimum number of frames within each trajectory for which both the frame and fly conditions must be satisfied and the number of frames in which just the fly conditions must be satisfied. A full list of the behavior statistics computed, and the minimum number of frames for a trajectory to be included, is in Table S6.

Per-retest (~4 videos collected simultaneously from the same cross) behavior statistics were obtained by averaging together the per-video statistics. Retests with only a single video were discarded from analysis.

As the behavior of our genetic control changed slightly over time (Figure S3B), we normalize each per-retest behavior statistic using the measured average behavior of the genetic control in the surrounding month: $x = x_0 - \mu_{month} + \mu_{total}$, where x is the normalized retest behavior statistic, x_0 is the measured, unnormalized retest behavior statistic, μ_{month} is the average of control retests recorded in the surrounding month, and μ_{total} is the average of all control retests recorded.

To obtain (*per-line*) statistics to describe the behavioral effects of neural activation for a given GAL4 line, we averaged the normalized behavior statistics for all retests of the line. All retests were given the same weight regardless of whether they consisted of 2, 3, or 4 videos. Thus, our per-line behavior statistics were computed by first pooling data from frames and flies in a single video, then across the ~4 videos recorded simultaneously from the same cross, and finally across the ~2 retests of the line.

Automatic, high-throughput behavior analysis pipeline

Behavior analysis for each video was automatically performed the night after the video was recorded. Our analysis pipeline was triggered to automatically run each night on all newly collected data using the Hudson continuous integration server v. 3.1.0. First, a script was automatically launched to search each data collection computer for newly recorded videos and transfer them to our data server. Then, each step of the behavior analysis pipeline was automatically launched on videos that had completed the previous step. The steps of the pipeline were: (0) data transfer, (1) initial automatic quality checks (see **Data curation**), (2) Ctrax-based tracking of body positions (see **Tracking**), (3) spatial and registration of data to a common coordinate system (see **Temporal and spatial registration of trajectories**), (4) sex classification (see **Classification of fly sex**), (5) tracking of flies' wings (see **Tracking**), (6) computation of per-frame feature time series (see **Computation of per-line behavior statistics**), (7) JAABA-based behavior classification (see **JAABA behavior classifiers**), (8) compilation of per-video behavior statistics (see **Computation of per-line behavior statistics**), (9) generation of short video clips showing the behavior of the flies and tracking accuracy, and (10) final automatic quality checks (see **Data curation**).

Testing for significant effects on behavior

We used a bootstrapping method to estimate the probability of observing a value at least as large as the measured per-line behavior statistic for our genetic control. Each sample consisted of retests of the genetic control with sizes matched to the dataset size for the given GAL4 line (same number of retests, each with the same number of videos as the GAL4 line). Our bootstrap distribution was constructed from n = 762 retests (2,881 videos, 56,768 flies) of the control line. We constructed 10,000 such samples, and computed the fraction of these samples with at least as large values as our measured per-line behavior statistic. This is the estimated *p*-value that the GAL4 line has a larger value for the behavior statistic than the genetic control. A similar computation was performed to estimate the *p*-value that the GAL4 line has a smaller value for the behavior statistic than the genetic control.

To account for multiple comparisons when estimating the number of lines with a significant increase in a given behavior statistic (Figure 3A(ii)), we compute the Benjamini-Hochberg (B-H) FDR correction (Benjamini and Hochberg, 1995; Groppe et al., 2011) over 2,205 tests (one for each line), and threshold at FDR = 0.1. This correction assumes independence or positive dependency between tests.

To correct for multiple comparison tests when estimating that 98% of lines have a significant change in *some* behavior statistic, we compute the Benjamini-Yekutieli (B-Y) FDR (Benjamini and Yekutieli, 2001; Groppe et al., 2011) across tests for both increases and decreases in all 203 behavior statistics and all lines (203*2*2,205 = 895,230 hypothesis tests), and threshold at 0.1. This FDR correction is valid for all forms of dependency. We say that a line has a significant behavioral difference if the null hypothesis is rejected for at least one of the 203*2 tests. For a line to be a false positive, all positive tests for that line must be erroneous. Thus, we can upper-bound the FDR over lines as the minimum FDR over the 406 tests for that line. Based on this upper bound, the expected number of false positives is 2 out of the 2,165 lines with a significant difference, a FDR of 0.001.

Behavior statistic correlation analysis

To look for interesting relationships between our behavior statistics across GAL4 lines, we computed Spearman's rank correlation coefficient between each of $\binom{83}{2} = 3{,}403$ pairs of 83 selected behavior statistics (Table S6). We computed these rank correlation coefficients both across all GAL4 lines, across only those GAL4 lines that moved at least 55% of the time, and across retests of the negative control line.

We estimated the significance of the hypothesis test that each correlation was non-zero with a bootstrap-based test in which we computed the Spearman rank correlation coefficients after disordering the lines/retests:

$$r_{0s}^{ij} = \mathrm{rankcorr}\left(x_{1:N}^{i}, x_{n_{s_1},\ldots,n_{s_N}}^{j}\right)$$

where x_n^i is the value of behavior statistic *i* for line/retest *n* and $n_{s_k} \sim \mathrm{Uniform}(\{1,\ldots,N\})$. We computed $S = 300{,}000$ such samples, and counted the number of samples with larger and smaller correlations. The two-tailed significance was then estimated as:

$$p^{ij} = \frac{2}{S}\min\left\{\sum_{s=1}^{S} I\left(r_{0s}^{ij} \geq r^{ij}\right), \sum_{s=1}^{S} I\left(r_{0s}^{ij} \leq r^{ij}\right)\right\}$$

where r^{ij} is the true measured rank correlation between behavior statistics *i* and *j*.

We also used a bootstrapping approach to determine whether the correlation coefficients were different between control retests and GAL4 lines. A sample *s* of the bootstrap distribution was created by randomly partition the 762 control retests into two (approximately) equal halves, C_s and $\overline{C_s}$, computing the Spearman rank correlation coefficients for each half, and computing the difference:

$$r_{0s}^{ij} = \mathrm{rankcorr}\left(x_{control,C_s}^{i}, x_{control,C_s}^{j}\right)$$

$$\bar{r}^{ij}_{0s} = \text{rankcorr}\left(x^{i}_{control,\overline{C_s}}, x^{j}_{control,\overline{C_s}}\right)$$

We computed $S = 30{,}000$ such samples, then counted the number of samples with bigger absolute differences between the correlations, and estimated the significance as

$$p^{ij} = \frac{1}{S}\sum_{s=1}^{S} \mathrm{I}\left(\left|r^{ij}_{0s} - \bar{r}^{ij}_{0s}\right| > \left|r^{ij}_{GAL4} - r^{ij}_{control}\right|\right)$$

where r^{ij}_{GAL4} and $r^{ij}_{control}$ are the true correlation coefficients for the GAL4 lines and control retests. To account for multiple comparisons, we compute the B-Y FDR across the 3,403 pairs of behavior statistics.

In Figure S4C, we histogram the sign-flipped rank correlation coefficients for subsets of the pairs of behavior statistics. We choose these subsets based on the control correlation coefficients. In red, we histogram behavior statistic pairs for which the magnitude of the rank correlation coefficient across control retests was both significantly nonzero (FDR $\leq$ 0.1) and large ($\left|r^{ij}_{control}\right| \geq .5$). We try to remove trivially correlated behavior statistic pairs by also requiring that there be at least one GAL4 line which did not follow the control correlation trend. We require there to be some line for which one of the behavior statistics is very different from control and the other behavior statistic is either near 0 or has the opposite sign:

$$\begin{aligned} noutliers^{ij} \doteq \sum\nolimits_{n=1}^{N} \Big[& I\left(s^{ij}z^{i}_{n} \geq \lambda_{large}\right) I\left(z^{j}_{n} \leq \lambda_{small}\right) + I\left(z^{j}_{n} \geq \lambda_{large}\right) I\left(s^{ij}z^{i}_{n} \leq \lambda_{small}\right) + I\left(s^{ij}z^{i}_{n} \leq -\lambda_{large}\right) I\left(z^{j}_{n} \geq -\lambda_{small}\right) \\ & + I\left(z^{j}_{n} \leq -\lambda_{large}\right) I\left(s^{ij}z^{i}_{n} \geq -\lambda_{small}\right) \Big] \end{aligned}$$

where $s^{ij} \in \{-1, +1\}$ is the sign of the rank correlation coefficient between behavior statistics *i* and *j* across control retests, z^{i}_{n} is the number of control standard deviations behavior statistic *i* for GAL4 line *n* is from the control mean, $z^{i}_{n} = \left(x^{i}_{n} - \mu^{i}_{control}\right)/\sigma^{i}_{control}$, $\lambda_{large} = 5$ is a threshold defining a large difference from control, and $\lambda_{small} = 1$ is a threshold defining a small difference from control. We also histogram the Spearman correlation coefficients across GAL4 lines for pairs of behavior statistics for which the magnitude of the correlation coefficient across control retests was not significantly different from 0 (FDR > 0.1) and small ($\left|r^{ij}_{control}\right| \geq .1$)

Behavioral effects of activation

To visualize the types of behavioral effects we observe as a result of neural activation across the GAL4 collection, we project the 203-dimensional vector representing the behavior statistics for a given GAL4 line down to two dimensions. We first normalize each behavior statistic based on the distribution of values observed in retests of the control line. For behavior statistic *x*, let *z* be the number of control standard deviations from the control mean:

$$z = \left(x - \mu_{control}\right)/\sigma_{control}$$

where $\mu_{control}$ and $\sigma_{control}$ are the mean and standard deviations for this behavior statistic across retests of the control line. To prevent extreme behavioral outliers from being over-represented, we use a log-transform for values more than $\lambda = 3$ standard deviations from control:

$$z_{norm} = \begin{cases} z & -\lambda \leq z \leq \lambda \\ \lambda + \log(z - \lambda + 1) & z > \lambda \\ -\lambda - \log(-z - \lambda + 1) & z < -\lambda \end{cases}$$

Using these normalized vectors $\mathbf{z}_{norm}$, we project the behavior data down to two dimensions using principal component analysis (PCA, Figures 3G and 3H) or t-distributed Stochastic Neighbor Embedding (Figure S4E). For t-SNE, we use Euclidean distance in this normalized space, and first project the data down to 50 dimensions using PCA, and set the perplexity parameter to 30.

We observed a broad range of effects on both the locomotion and social behavior of the flies. The activity level varied continuously from lines that rarely moved to lines that moved over 95% of the time (control flies moved on average 58% of the time), with 60% of lines showing a significant difference in the fraction of time stopped (FDR $\leq$ 0.1, Figure 3A(ii)). Similarly, the average center-to-center inter-fly distance varied continuously from 5 mm $\approx$2 body lengths – flies were nearly touching) to 16 mm (1.5x the average distance if uniformly distributed in the arena – flies were actively avoiding each other), with 40% of GAL4 lines showing a significant difference in inter-fly distance (FDR $\leq$ 0.1, Figure 3A(ii)).

Among lines that walked more than control, we observed a few qualitatively different phenotypes, including lines that walked in long bouts that often crossed the arena, lines that walked much faster than control, lines that chased more than control, and lines that appeared to have general increases in their activity level.

To further narrow our behavior categories, we found it useful to combine behavior measures, for example categorizing lines that walked more but did not chase more than control, or lines that stopped more but did not groom their wings more.

Janelia Fly Light imagery data

We used the Janelia FlyLight imagery data (Jenett et al., 2012) to characterize the anatomical expression of each GAL4 line. To create this dataset, the authors used the UAS-GAL4 system to express GFP in the targeted neurons. Our analyses were performed on, on average, 5.6 sample brains from female flies (Figure 4B) kindly provided by the Janelia Fly Light Team.

Processing of neural expression pattern images

Registration

We used the BrainAligner (Peng et al., 2011) to automatically register all brain images to a common template brain (JFRC2 template, (Jenett et al., 2012)), resulting in a total of 6,542 well-registered image stacks of size 1024 × 512 × 218 pixels, across 2,082/2,204 GAL4 lines for which we measured behavior (Figure 4B). Different from previously published registered images, we used 500 landmark points, including 269 landmark points in the optic lobes (all code for improved BrainAligner registration was kindly provided by Yang Yu). Besides producing a registered image stack, the BrainAligner also outputs an indicator of the quality of the alignment. As cell bodies were not well-aligned by this procedure, we limited our analysis to neuropil. This was achieved by confining analysis to the mask developed by Jenett et al. (Jenett et al., 2012)). However, as this masking is imperfect, it should be noted that cell body fluorescence may be contributing to the putative neuropil signal in some cases.

Intensity normalization

As brightness and contrast were extremely variable across images (because of genetic differences, differences in dissection, and differences in imaging parameters such as laser power), we normalized the intensity of pixels in each image. We performed the same normalization on all pixels in a given image stack:

$$x_{norm} = \max\left(0,\ \min\left(1, \frac{x_{raw} - \lambda_{low}}{\lambda_{high} - \lambda_{low}}\right)\right)$$

Here, x_{raw} is the original intensity of the pixel, and λ_{low} and λ_{high} are intensity thresholds for the current image. When appropriate, we thought of this normalized intensity as our estimate of the probability that there was expression in a given voxel.

Different thresholds λ_{low} and λ_{high} are chosen for each image stack, but are shared across all pixels in a single image. The values of these thresholds are automatically set based on the following measured statistics of a downsampled version (by $x = 5,\ y = 5,\ z = 3$) of a given image stack:

- Mean of pixels outside the brain foreground mask (background pixels).
- Standard deviation of background pixels.
- Mean of pixels inside the brain foreground mask (foreground pixels).
- Standard deviation of foreground pixels.
- Fraction of background pixels with intesities between a sequence of 20 intervals.
- Fraction of foreground pixels with intesities between a sequence of 20 intervals.
- Percentiles of background pixel intensities (0, 1, 5, 10, 25, 50, 75, 90, 95, 99, 100).
- Percentiles of foreground pixel intensities.

The thresholds λ_{low} and λ_{high} are linear functions of these image statistics. The coefficients of these linear functions are learned using lasso regression from a training set of 148 image stacks for which we manually annotated the desired λ_{low} and λ_{high}. The GUI for annotating thresholds, as well as an example annotation, are shown in Figure S5.

Combining images of the same line

For the majority of lines, the Fly Light Dataset contained several images corresponding to different dissections and flies from the same GAL4 line (Figure 4B). To create the representation of the expression pattern we used in all further analyses, we averaged together several of these image stacks. Averaging multiple images allowed us to capture both the stochasticity of expression (e.g., if one image had expression in a given neuron and another did not, then the average would have a value of 0.5) as well as imprecision in the registration and normalization algorithms. We averaged together all images of a given GAL4 line with sufficiently good registration quality. In addition, we blurred the image with a 5 × 5 x 2 pixel box filter to account for imprecision in the registration.

Clustering the brain into supervoxels

We cluster voxels in the fly brain template into supervoxels based on the expression in each voxel across GAL4 lines sampled as well as the (x,y,z) location of the voxel. Let x_v^e be the L-dimensional vector for which entry x_{vl}^e is the expression in voxel v for line l, and let x_v^s be the 3-dimensional vector indicating the location of voxel v in the image stack. Our measure of the distance between voxels u and v is the weighted average of L_1 distances:

$$d(u,v) = \sum_{l=1}^{L} \left| x_{vl}^e - x_{ul}^e \right| + w_s \sum_{i=1}^{3} \left| x_{vi}^s - x_{ui}^s \right|,$$

where w_s is the weight of the spatial term. Both for efficiency and to make use of existing knowledge of compartmentalization of the *Drosophila* brain, we clustered each compartment, as defined in (Jenett et al., 2012), separately. We use the farthest-first algorithm to construct a hierarchical clustering of the voxels into a pre-specified number of sub-regions. The farthest-first algorithm starts with one data point to represent the first cluster, then iteratively selects the data point farthest from the first $k-1$ chosen data points to represent the k^{th} cluster (Dasgupta and Long, 2005). This simple algorithm is relatively fast and memory-efficient. This clustering produces regions of varying sizes. To force the ratio of the volume of the largest to the smallest cluster to be below some threshold, we next split all large clusters, again using the farthest-first algorithm, with an increased weight given to the spatial term. Once all clusters are sufficiently small, we iteratively merge clusters that were too small. Supervoxel clustering of selected compartments are shown in Figure 5B–5D. We tried a variety of values for the three parameters of our clustering the algorithm: the number of supervoxels, the weight of the spatial component in the distance, and the maximum allowed ratio in supervoxel volumes. The results (Figure S6) were qualitatively similar across parameter values, and we chose one parameter setting that holistically and heuristically corresponded with our knowledge of *Drosophila* brain anatomy: number of supervoxels = 7,065, position weight = 10^{-5}, and volume ratio = 2.

A related clustering technique was independently developed by (Panser et al., 2016). This approach was based on a similar hypothesis, that voxels with similar expression across GAL4 lines might be part of the same functional unit. However, the algorithms differ in their details, in particular, the preprocessing of the anatomy images, the distance function, and the clustering algorithm. Potential improvements in our implementation are (1) we use two thresholds set based on a machine learning algorithm to normalize intensity in each image and (2) our clustering algorithm is relatively fast, thus we could apply it to the entire brain without downsampling, instead of a downsampled version of the central brain. Important differences include that (1) our algorithm has an expression component as well as a spatial component and (2) our clusters are much smaller. Finally, the goals of these projects are quite different. Panser et al.'s work searches for new anatomical understanding of the brain, while our work is looking for a low-dimensional representation of expression for our behavior-anatomy correlation.

Behavior-anatomy correlation

For every pair consisting of a supervoxel and a measured behavior statistic, we can compute the correlation ρ_{uv} between estimates of whether supervoxel v has expression and whether behavior statistic u has a high value, across all GAL4 lines:

$$\rho_{uv}^{+} = \sum_{l=1}^{L} a_{vl} b_{ul}^{+} = \boldsymbol{a}_{\boldsymbol{v}}^{\top} \boldsymbol{b}_{\boldsymbol{u}}^{+}.$$

Here, $a_{vl} \in [0,1]$ is the normalized and thresholded expression image intensity in supervoxel v for line l and $b_{ul}^{+} \in [0,1]$ is a normalized and thresholded version of the value of behavior statistic u for line l. The parameters of the normalization and thresholding of the expression intensity and behavior statistic are chosen so that a_{vl} and b_{ul}^{+} can be interpreted probabilistically as estimates of whether supervoxel v has expression and whether behavior statistic u has a high value for line l given our observations, and the correlation ρ_{uv}^{+} is an estimate of the probability of co-occurrence of an increase in the behavior and expression in the supervoxel.

Normalization and thresholding of anatomical expression is described above (*Processing of neural expression patterns*, *Intensity normalization*). We normalize and threshold the behavior statistic as

$$b_{ul}^{+} = \max\left\{0, \min\left\{1, \frac{x_{ul} - \mu_u}{\lambda_u^{+}}\right\}\right\}$$

where x_{ul} is the measured per-line value for behavior statistic u and line l (see *Computation of per-line behavior statistics*), μ_u is the value of behavior statistic u for the genetic control line, and λ_u^{+} is a threshold defined as the minimum of 5 standard deviations of the control line behavior (measured over retests) and the 99th-percentile over all lines that have a higher value for behavior statistic u:

$$\lambda_u^{+} = \min\left\{5\sigma_u, \text{percentile}\left(\{x_{ul} - \mu_u : x_{ul} - \mu_u > 0\}_{l=1}^{L}, 99\right)\right\}$$

A similar computation was performed to estimate the probability that behavior statistic u had a low value for line l:

$$b_{ul}^{-} = \max\left\{0, \min\left\{1, \frac{x_{ul} - \mu_u}{\lambda_u^{-}}\right\}\right\}$$

where

$$\lambda_u^{-} = \min\left\{5\sigma_u, \text{percentile}\left(\{\mu_u - x_{ul} : x_{ul} - \mu_u < 0\}_{l=1}^{L}, 99\right)\right\}$$

Using our interpretation of the value b_{ul} as a probability, we can create logical combinations of our behavior statistics such as the probability that line l has a high value for both behavior statistics u_1 and u_2 as

$$b_{(u_1=+)\wedge(u_2=+),l} = b_{u_1,l}^{+} b_{u_2,l}^{+}$$

or the probability that line l has a high value for behavior statistic u_1 but not u_2 as

$$b_{(u_1=+)\wedge(u_2\neq+),l} = b^+_{u_1,l}\left(1 - b^+_{u_2,l}\right)$$

or the probability that line line l has a high value for either u_1 or u_2 as

$$b_{(u_1=+)\vee(u_2=+),l} = \min\left\{1,\ b^+_{u_1,l} + b^+_{u_2,l}\right\}$$

We use a bootstrapping method to compute the probability of observing a correlation as high as that measured, given the null hypothesis that there is no correlation between behavior and anatomy, which we estimate by shuffling the lines. Let $\tilde{b}^+_{ul} = b^+_{ul'}$ for a randomly chosen $l' \sim Uniform\{1,\ldots,L\}$. As $\mathbf{a_v}$ and $\widetilde{\mathbf{b}^+_{\mathbf{u}}}$ are necessarily uncorrelated, we estimate the probability of observing at least as high a correlation ρ_{uv} when the data are uncorrelated by drawing $N = 100{,}000$ such samples $\tilde{\mathbf{b}}^{+\boldsymbol{i}}_{\mathbf{u}}$ and computing the fraction with higher correlation $\tilde{\rho}^{+i}_{uv} \geq \rho^+_{uv}$. We report this fraction as an estimate of the p value of there being a correlation between expression in the supervoxel v and an increase in the value for behavior statistic u. We perform a similar computation to measure whether how likely it is that there is a correlation between expression in a supervoxel and a *decrease* in the value of a given behavior statistic by using a normalized version of the negative value of the behavior statistic.

For each behavior statistic, we compute the B-H FDR over the 7,065 correlation tests for all supervoxels (these are the tests that are performed to create a single behavior-anatomy correlation map). Our maps threshold correlation *p*-values at 0.05 and B-H FDR at 0.25. We used a permissive FDR because p value is indicated in pseudocolor.

Our analysis tests for correlations between these probabilistic behavior and anatomy scores. While our model does make assumptions, we nonetheless can find significant effects of activation when there are nonlinear interactions between cell groups. To demonstrate this, we simulated data from one such extreme interaction, in which one cell group completely masks the behavioral effect of activating another, and show that we can still find significant correlations. We used our measured anatomical expression data, and generated synthetic behavior phenotypes b_l for each line l as follows. For each line l and supervoxel i, we generated $x_{li} \sim Bernoulli(a_{li})$, which represented there is truly expression, probabilistically according to the anatomical expression score a_{li}. We randomly chose one supervoxel u, and assume that activating supervoxel u alone results in the behavior phenotype $(b = 1)$. The rest of the supervoxels will either not affect this behavior measure or will completely mask it, resulting in the complete absence of the behavior phenotype $(b = 0)$. Thus, we set that the behavior phenotype for line l based on the randomly generated true expression pattern x_l as

$$b_l = \begin{cases} 0,\ x_{lu} = 0 \\ 0,\ \exists m \in M : x_{lm} = 0 \\ 1,\ x_{lu} = 1\ and\ x_{lm} = 0\ \forall\ m \in M \end{cases}$$

Here, M are the set of supervoxels that will mask the behavior phenotype. We varied the fraction of supervoxels with masking behavior. For reasonable numbers of masking supervoxels (on average 38 masking supervoxels, resulting in, on average 18 lines showing the behavior phenotype), we do indeed find a significant correlation ($p \leq 0.0001$). Figure S8D shows the results of this simulation.

Some of our maps are quite dense, while others are sparse. For example, the increased walking map is dense, while the increased female chase map is sparse. There are many possible reasons for the density of a map. One potential explanation for the density of the map is the specificity of the behavior. There are many reasons a fly might walk, for example, it might walk because it is hungry or because it is chasing another fly. For this reason, we believe combining behavior statistics using our BABAM software (Data S1) will be useful in refining maps. While it could be that a map is denser because more neurons are involved in control of the behavior, it is important to remember that lack of statistical significance for our behavior-anatomy correlation test does not imply that there is no correlation. Frequency of positive tests will depend on the signal-to-noise ratio (SNR) for both the behavior and the brain region. Behavior SNR depends on whether the behavior change is obvious when other parts of the brain are activated, and the variability of the behavior across flies from the same genotype. Anatomy SNR depends on the structure of the brain region, as large processes have higher contrast in the image stacks and are better aligned by the BrainAligner. Other important factors contributing to sensitivity of a given behavior-anatomy correlation test are whether a given behavior can be stimulated with thermogenetic activation, and the frequency of expression in a given anatomical region in the Janelia GAL4 collection.

Behavior-anatomy maps

We created a behavior-anatomy map of neuronal regions correlated with a selected behavior measure having a high value by coloring the region corresponding to each supervoxel by our estimate of the p value of the correlation hypothesis test. We compute the B-H FDR across all supervoxels, i.e., all hypothesis tests used to create a single map, and color supervoxels for which the FDR was over 0.25 as though the p value was 1. In Figures 7 and 8 and Movie S3, we show the minimum projection of this 3D map over z-slices.

The same technique was used to create behavior-anatomy maps of regions correlated with a selected behavior measure having a low value, or for any logical combination of behavior statistics.

In Figure 7, we show the behavior-anatomy map corresponding to the fraction of time walking for all flies, fractime_flyany_framewalk, having a high value.

In Figure 8A, we show the behavior-anatomy map corresponding to the fraction of time male flies perform wing extensions, fractime_flymale_framewingextension, having a high value.

In Figure 8B, we show the behavior-anatomy map corresponding to the fraction of time all flies are stopped, fractime_flyany_framestop, having a high value, and the fraction of time all flies perform wing grooming, fractime_flyany_framewinggrooming *not* having a high $\left(b_{(stop=+)\wedge(winggroom\neq+)}\right)$.

In Figure 8C, we show the behavior-anatomy map corresponding to the fraction of time all flies jump, fractime_flyany_framejump, having a high value.

In Figure 8D, we show the behavior-anatomy map corresponding to the fraction of time all flies back up, fractime_flyany_framebackup, having a high value.

In Figure 8E, we show the behavior-anatomy map corresponding to the fraction of time female flies chase, fractime_flyfemale_framechase, having a high value, and the fraction of time male flies chase, fractime_flymale_framechase *not* having a high value $\left(b_{(chase\_female=+)\wedge(chase\_male\neq+)}\right)$.

In Figure 8F, we show the behavior-anatomy map corresponding to the fraction of time all flies perform wing-grooming, fractime_flyany_framewinggrooming, having a high value.

In Figure 8G, we show the behavior-anatomy map corresponding to the average maximum angle of either wing, max_wing_angle_flyany_frameany, having a high value.

We note that our clustering and behavior-anatomy maps are not left-right symmetric. This is because the template brain to which all our images were aligned is a single fly brain, and thus is not perfectly symmetric. In addition, z-slices cut through different portions of the brain in the left and right hemispheres. Our analyses did not enforce symmetry. However, we found looking for such symmetries in both the clustering and behavior-anatomy maps to be a useful confirmation of our findings.

Clustering behavior-anatomy maps

In Figure 7C, we cluster the supervoxels with sufficiently significant correlation (p-value ≤ 0.005) according to their expression level in GAL4 lines which both show the given behavior phenotype (behavior score ≥ 0.5) and have expression in some sufficiently correlated supervoxel (expression level ≥ 0.5). We cluster using average-linkage hierarchical clustering and the L1 distance metric.

Identifying important GAL4 lines

For a given brain region (corresponding to a set of supervoxels) $V = \{v_1, \ldots v_n\}$ which is significantly correlated with a behavior phenotype u, we define the importance of a given GAL4 line l to the behavior-anatomy correlation as $s_{Vl} = a_{Vl}b_{ul}$, where the anatomy score for a region is the average of the anatomy scores for the contained supervoxels, $a_{Vl} = (1/n)\sum_{i=1}^{n} a_{il}$. To find the most important lines, we sort the lines in descending order of the behavior-anatomy product s_{Vl}. When computing the average expression pattern over the important lines as in Figures 8B–8G(iii), we selected m out of the L lines in order of importance until they make up half the total correlation: $\sum_{l=1}^{m} s_{Vl} \geq 0.5\sum_{l=1}^{L} s_{Vl}$.

Identifying ring neurons associated with walking

The increased-walking map (Figure 7A) shows correlations between this behavior phenotype and expression in components of a putative circuit involved in the control of walking (Figure 7C), consisting of the optic lobes, the optic tubercles, and the ellipsoid body. We used the split-GAL4 intersectional strategy (Luan et al., 2006; Pfeiffer et al., 2010) in conjunction with this behavior-anatomy map to identify a small population of R2/R4 ring neurons in the ellipsoid body that were sufficient to elicit an increase in walking probability. First, we created 20 novel split-GAL4 lines based on 10 of the most important GAL4 lines for the walk-ellipsoid body correlation (Figures 7D–7F). We then independently tested these split-GAL4 lines for (1) whether we saw an increase in the fraction of time the flies spent walking and (2) whether and where we saw expression in the ellipsoid body.

Of the split-GAL4 lines, 8/20 walked significantly more than the empty split-GAL4 control when tested for a neural activation phenotype (Figure 7E). We characterized the expression patterns of these split-GAL4 lines (*Split GAL4 line annotation*, below, Figure S7, Table S7), and for 15/20 lines, there was expression in the ring neurons of the ellipsoid body, the region implicated in our walking-more behavior-anatomy map. This 75% success rate of finding split-GAL4 lines with strong expression in the ellipsoid body suggests that it is beneficial to use both behavior and anatomy information to target expression to sparse neural regions (for comparison, a screen based only on anatomical expression patterns had a much lower rate of success, 16%, in finding split-GAL4 lines with strong expression in the desired region (Aso et al., 2014)).

Three of the split-GAL4 lines with increased walking, SS02769, SS02767, and SS02766 also had relatively clean expression in what appears to be the same small population of R2/R4 ring neurons (Figure 7F). These three lines are intersections of pairs of three enhancer fragments (Figures 7D and 7F), further suggesting that they are the same population of neurons. Four of the other lines that walked more contained many more ring neuron types, and may also contain this population of neurons, or other neurons sufficient to induce increases in walking probability. These results demonstrate that, using our behavior-anatomy maps, we can identify and gain genetic access to sparse neuronal populations that are involved in the production of a given behavior.

Split-GAL4 line annotation

Identification of ring neurons was based on cell body position, microglomeruli processes in the bulb, and tangential processes in the ellipsoid body (EB) (Hanesch et al., 1989). As it was well stated by Wolff et al., ‘‘without exhaustive cell type analysis of all ring-like neurons, an accurate and complete catalog of the ring neurons does not exist’’ (Wolff et al., 2015). Using the categories previously described (Hanesch et al., 1989; Renn et al., 1999; Young and Armstrong, 2010), we used the following criteria to subcategorize the EB rings neurons in the split-GAL4 lines.

1. If there were EB ring processes entering the EB peripherally from the RF track, the line was annotated as containing R4 ring neurons.
2. If (1) and a distinct gap between the distal edge of the ellipsoid body neuropil and expression, the line was annotated as containing R4m type neurons.
3. If (1), and expression in the most distal ring of the ellipsoid body, the line was annotated as containing R4d neurons.
4. If (1), but (2) versus (3) could not be clearly determined, the line was labeled simply as R4 containing.
5. If there were processes entering the EB canal from the RF track and no other CC expression, the line was annotated as containing R1, R2, or R3 neurons.
6. If (5), and only expression in the outer ring, the line was annotated as containing R2 type neurons.
7. If (5), and continuous expression throughout the anterior portion of the inner, middle and outer rings, the line was annotated as containing R3 neurons.
8. If (5), and expression in the posterior inner ring of EB, the line was annotated as containing R1 type neurons.

These criteria resulted in annotations that are inclusive of those identifiable cell types but not exclusive of those not noted. Due to the density of expression within the ellipsoid body in some of our lines it is not possible to parse all cell types present without extensive further characterization. For example, R2 neurons cannot be definitively identified in a line that also contains R4 (overlap of expression) and R1 or R3 (axons enter through the EB canal), and for this reason was left out of annotation of such lines even though R2 type neurons very well may be present in such lines. For all lines, we noted the presence of expression in the radial inner, middle, or outer rings and anterior or posterior shells. Similarly, the presence of R4d and R2 neurons could mask the presence of additional R4m type neurons. The number of EB ring neurons present in each line was estimated from cell body clusters and categorized as less than 10, 10-30, more than 30. Expression outside the central complex in the central brain is reported by number of cell bodies or as ‘optic lobe’ with an estimate of the number of cell types.

QUANTIFICATION AND STATISTICAL ANALYSIS

The Method Details section above details the methods we used for quantitatively analyzing our data. In this section, we overview the types of statistical tests we used, which parts of the analyses they were used for, and, for each of these, the dataset sizes and definitions of significance. The table below summarizes this information for each result.

The majority of our statistical tests are based on bootstrap-based estimates of the null hypothesis distribution. In all cases, multiple comparisons are accounted for by controlling the false discovery rate (FDR). We use the Benjamini-Hochberg (B-H) FDR when the assumption that tests were independent or positively dependent was satisfied, and the Benjamini-Yekutieli (B-Y) FDR when tests might have other dependencies.

Figure 3A(ii) shows the results of our tests of whether the value of a given behavior statistic for a given GAL4 line was significantly higher than that we observed for our genetic control. This process is detailed above (*Methods, Testing for significant effects of neural activation on behavior*). Here, we use bootstrapping to estimate the distribution of the per-line behavior statistic computation for sets of retests of our genetic control. The number of retests for each GAL4 line was approximately 2 (see ‘‘*N. sets*’’ of Table S1 for the exact numbers per line). Each sample of the bootstrap distribution was created by combining k retests of the control (of which there were in total 762), where k is the number of retests of the selected GAL4 line. We approximated the bootstrap distribution by sampling 10,000 points from $\binom{762+k-1}{k}$ possible combinations of our 762 independent retests. To account for multiple comparisons across the 2,205 lines tested for each behavior statistic, we used the B-H FDR, and significance was determined so that the FDR was 0.1.

Our claim that 98% of lines showed a significant change in some behavior statistic used the p value estimates from this same bootstrap test (*Methods, Testing for significant effects of neural activation on behavior*). Here, we accounted for the multiple comparisons across all 895,230 line-behavior statistic combinations using the B-Y FDR, and significance was thresholded so that the FDR was 0.1.

As shown in Figure S4C, we tested whether the Spearman rank correlation coefficient for a given pair of behavior statistics was significantly non-zero for both the GAL4 lines and the control retests. Here, we used bootstrapping to estimate the distribution of correlation coefficients for the pair of behavior statistics given that they are uncorrelated. We did this by comparing the true correlation coefficient to that measured for the same pair of behavior statistics after randomly resampling lines/control retests for the second statistic (*Methods, Behavior statistic correlation analysis*). For the GAL4 lines, we approximated the bootstrap distribution by sampling 300,000 points from $2,205^{2,205}$ possible combinations of our 2,205 independent measurements. For control retests, we approximated the bootstrap distribution by sampling 300,000 points from 762^{762} possible combinations of our 762 independent measurements. To account for multiple comparisons across the 3,403 pairs of behavior statistics tested, we used the B-Y FDR, and significance was determined so that the FDR was 0.1.

As shown in Figures 3F and S4B, we tested whether, for a given pair of behavior statistics, the Spearman correlation coefficient across GAL4 lines were significantly different from that across retests of the control line. Here, we used bootstrapping to estimate the distribution of the absolute difference between correlation coefficients if they are both computed from our control retests. We did this by comparing the absolute difference between the correlation coefficient across the GAL4 lines and that across control retests to the absolute difference in correlation coefficients across approximately equal partitions of the control retests (*Methods, Behavior statistic correlation analysis*). The number of data points used to estimate the GAL4 correlation coefficient is 2,205, and the number used to estimate the control correlation coefficient is 762. We approximated the bootstrap distribution by sampling 30,000 points from 2^{762} possible partitions of our 762 independent measurements. To account for multiple comparisons across the 3,403 pairs of behavior statistics tested, we used the B-Y FDR, and significance was determined so that the FDR was 0.1.

Each behavior-anatomy map (Figures 7 and 8, Movie S3) shows the results of our test of whether there is a positive correlation between the value of a given behavior statistic and the expression for a given supervoxel across GAL4 lines. This process is detailed above (*Methods, Behavior-anatomy correlation*). Here, we used bootstrapping to estimate the distribution of the behavior-anatomy correlation if behavior and anatomy are not correlated. We did this by comparing the true behavior-anatomy correlation to that measured for the same pair of behavior statistic and supervoxel after randomly resampling lines for the supervoxel expression. We approximated the bootstrap distribution by sampling 100,000 points from $2,082^{2,082}$ possible combinations of our 2,082 independent measurements (number of lines with both behavior and anatomy data). To account for multiple comparisons across the 7,065 supervoxels tested for each behavior statistic, we used the B-H FDR, and significance was determined so that the FDR was 0.25. We chose this less restrictive threshold because p value is shown in the map.

In Figure 7E, we show the results of testing the split-GAL4 lines for increased walking. We used a one-sided Mann-Whitney test, with each sample corresponding to a video (20 flies), and the number of videos varying between 6 and 16 for each GAL4 line (Table S7), and the number of control videos equal to 22. To correct for multiple comparisons across the 20 split-GAL4 lines, we use the Bonferroni correction, and threshold significance at 0.05.

Statistical analysis test summary: For each of the types of hypothesis tests, we summarize the following. *Result*: Description of the results this test corresponds to. *Details*: Where to find a detailed description of the hypothesis test and bootstrapping procedure. *N. drawn samples*: Number of samples we draw to approximate the bootstrap distribution. *Max samples*: Number of different (but not necessarily independent) samples that can possibly be drawn from the bootstrap distribution. *N. meas.*: Number of measurements used in combinations to generate the bootstrap distribution. *Multiple comparison:* Method for accounting for multiple comparisons. *N. tests*: Number of tests performed, for which the multiple comparison correction is done. Note that the last result is not a bootstrap-based test. Here, we list the number of samples in each dataset under *N. meas.*

Result	*Details*	*N. drawn samples*	*Max samples*	*N. meas.*	*Multiple comparison*	*N. tests*
Figure 3C	*Methods, Testing for significant effects of neural activation on behavior*	100,000	$\binom{762+k-1}{k}$	762	B-H FDR = 0.1	2,205
Results, 98% of lines	*Methods, Testing for significant effects of neural activation on behavior*	100,000	$\binom{762+k-1}{k}$	762	B-Y FDR = 0.1	895,230
Figure S4C, GAL4 lines	*Methods, Behavior statistic correlation analysis*	300,000	$2,205^{2,205}$	2,205	B-Y FDR = 0.1	3,403
Figure S4C, control retests	*Methods, Behavior statistic correlation analysis*	300,000	762^{762}	762	B-Y FDR = 0.1	3,403
Figure 3F	*Methods, Behavior statistic correlation analysis*	30,000	2^{762}	762	B-Y FDR = 0.1	3,403
Behavior-anatomy maps, Figures 7A, 8, Movie S3	*Methods, Behavior-anatomy correlation*	100,000	$2,082^{2,082}$	2,082	B-H FDR = 0.25	7,065
Figure 7E	*Quantification and statistical analysis*	N/A	N/A	6-16,20	Bonferroni, $p < .05$	20

DATA AND SOFTWARE AVAILABILITY

Software

All software for collecting and analyzing our data is within Data S1, within the directory code. Top-level functions and modules are described within the README.txt file. Parameters for this code are in Data S2 and S3. Maintained versions of some of this software is available online:

BABAM, the Browsable Atlas of Behavior-Anatomy Maps: https://kristinbranson.github.io/BABAM/

Ctrax, the Caltech Multiple Walking Fly Tracker: http://ctrax.sourceforge.net/

JAABA, the Janelia Automatic Animal Behavior Annotator: http://jaaba.sourceforge.net/

Data

Compressed, processed forms of our data is available in Data S1. These are within the directory data, and their contents are described in the README.txt file.

A browsable and searchable website with the results of our behavior analyses is at http://research.janelia.org/bransonlab/FlyBowl/BehaviorResults.

Less processed versions of the data are available upon request, please contact Kristin Branson, bransonk@janelia.hhmi.org.

Supplemental Figures

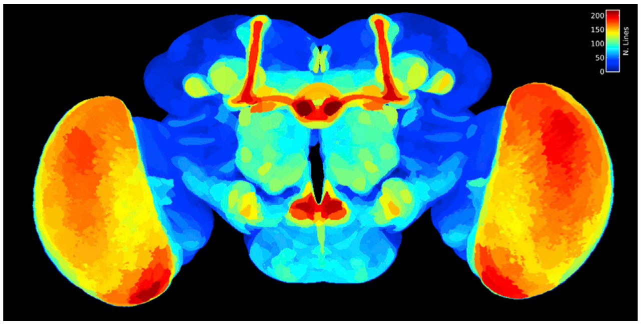
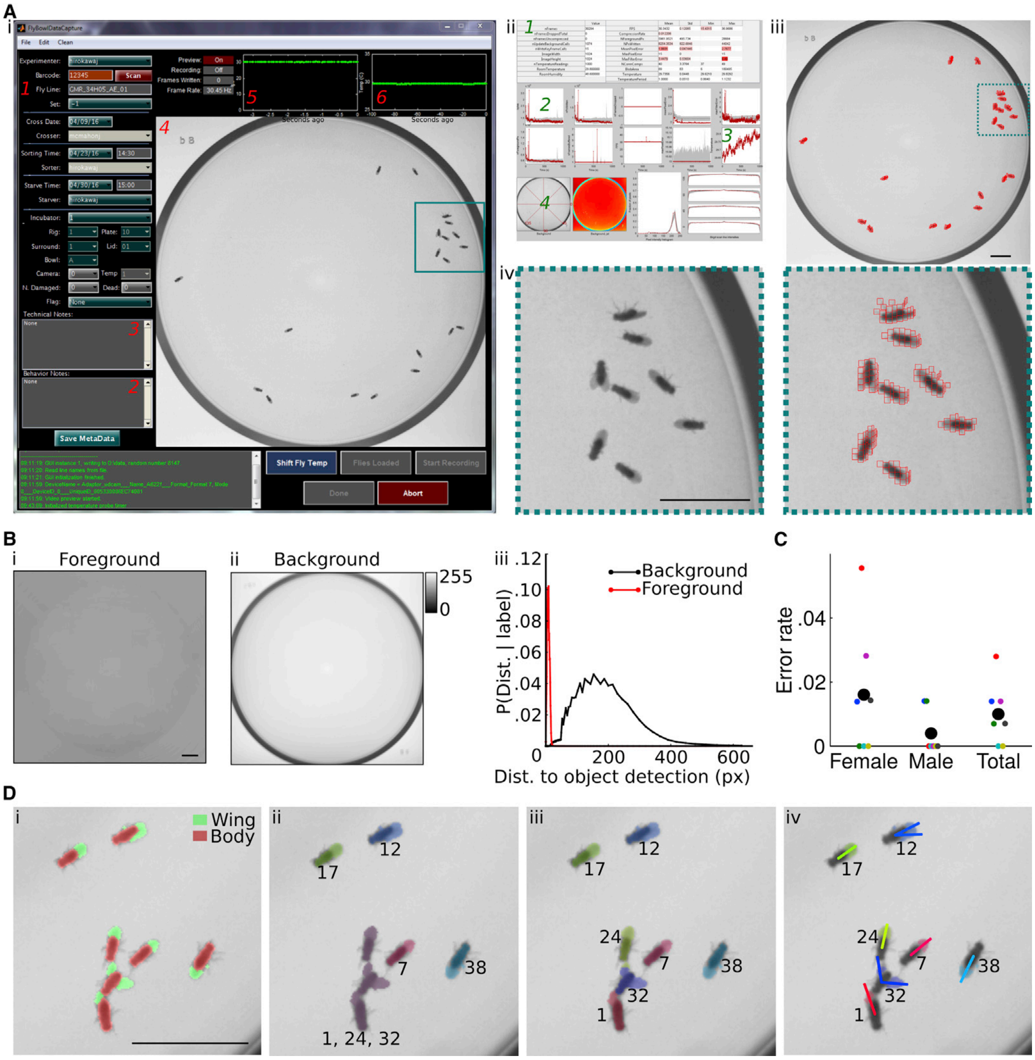


Figure S1. Max-Projection Image of the Number of Lines with Expression in Each Supervoxel, Related to Figure 6



Experimenter: hirokawaj
Barcode: 12345
Fly Line: GMR_34H05_AE_01
Set: -1
Cross Date: 04/09/16
Crosser: mcmahonj
Sorting Time: 04/23/16 14:30
Sorter: hirokawaj
Starve Time: 04/30/16 15:00
Starver: hirokawaj
Incubator: 1
Rig: 1 Plate: 10
Surround: 1 Lid: 01
Bowl: A
Camera: 0 Temp 1
N. Damaged: 0 Dead: 0
Flag: None
Technical Notes: None
Behavior Notes: None
Preview: On
Recording: Off
Frames Written: 0
Frame Rate: 30.45 Hz

```
09:11:19: GUI instance 1, writing to D:\data, random number 8147
09:11:20: Read line names from file.
09:11:21: GUI initialization finished.
09:11:59: DeviceName = Adaptor_udcam___Name_A622f___Format_Format 7, Mode
0___DeviceID_0___UniqueID_005330006EC74001
09:11:59: Video preview started.
09:43:09: Initialized temperature probe timer
```

Figure S2. Data Collection and Tracking, Related to Figure 1

(A) Video and metadata collection. (i) Screen capture of GUI used for controlling and previewing video recording, and for inputting metadata, discussed in detail in the Method Details, and included in Data S1 and S2. The left side of the GUI is used for metadata entry, such as line name (1) and notes on behavioral observations (2) or technical issues (3). The right side of the GUI shows the recorded video (4), video frame rate (5), and arena temperature (6), updated in real time. (ii) Technical feedback returned to the user immediately after the video recording has completed. This feedback includes information about the video recording (e.g., number of frames recorded, (1)), information about the video compression (e.g., compression rate, (2)), information about the environment (e.g., temperature, (3)), and information about the arena appearance (e.g., background image, (4)). In the tables at the top, the red highlights indicate the how abnormal the recorded values are (compared to a pilot dataset). In the plots at the bottom, the gray regions indicate the mean and standard devation of the pilot data, and the red lines correspond to the current experiments. (iii) Compression of a sample frame. The red boxes indicate a tiling of the foreground pixels by small 5 × 5 boxes. The locations of these boxes and the intensities of pixels inside them are stored to represent this frame. (iv) Reconstruction of the sample frame region corresponding to teal box in (i) and (iii).

(B) Pooled foreground and background appearance models learned using Ctrax. (i) Per-pixel-location means of the foreground and (ii) background appearance. (iii)Histogram-based estimate of the distribution of distances to the nearest object detection for pixels classified as background (black) and foreground (red).

(C) Error rate of automatic sex classification. We compute the error rates separately for flies labeled female and male, as well as the total error rate. Black circles indicate the error rates across all seven videos labeled for GAL4 line R14C07, while colored dots indicate error rates per video.

(D) Illustration of steps of wing tracking. (i) Segmentation of pixels into those corresponding to body (red), wing (green) and background (not colored). (ii) Assignment of fly identities output by Ctrax (indicated by numbers) to each connected component of foreground (wing or body) pixels. Each color corresponds to a different connected component. The purple connected component is assigned to three fly identities. (iii) Assignment of pixels to fly identities (indicated by numbers) output by Ctrax. Each color corresponds to a different identity. The large purple connected component in (ii) has been segmented. (iv) Wing angles fit (indicated by line segments) to wing pixels for each fly.

All: Scale bar indicates 1 cm.

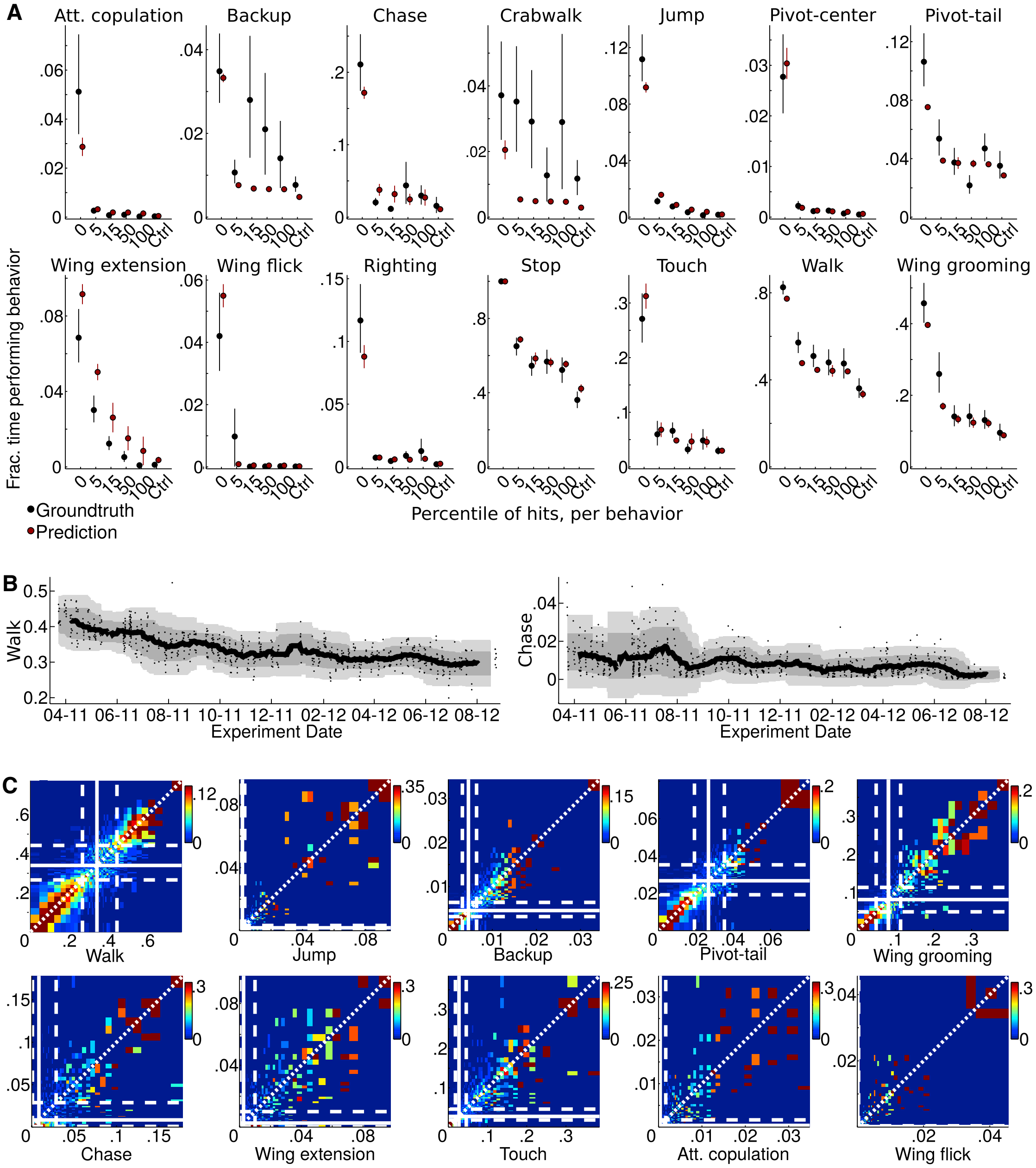


Figure S3. Behavior Analysis, Related to Figure 2

(A) Accuracy of automatically-computed line-level behavior statistics. We compare the fraction of time performing each behavior computed from the automatic behavior classifier's predictions (red), with the fraction of time computed from the sample of ≤ 2,000 manually labeled frames (black) for 5 GAL4 lines and control (Method Details). Vertical lines indicate the 10th and 90th percentiles of the estimated distributions, computed over videos.

(B) Genetic control line's behavior as a function of the experiment date, over the 1.5-year data collection period. Each black dot corresponds to a retest (a set of ≈4 simultaneously recorded videos of ≈20 flies from the same cross). The solid black line indicates the mean over the surrounding 30-day window. The darker and lighter gray regions indicate one and two standard deviations within these time windows.

(C) Behavior statistics across retests of each line. Each plot corresponds to a different behavior statistic. Each plot shows a histogram of the behavior statistic measured in the second retest (y) given the value measured in the first retest (x), for each pair of retests of each line (color indicates fraction of retests). Each column is normalized to sum to one. Bin sizes grow logarithmically from the control mean (white solid line). Dashed white lines show the 2.5th and 97.5th percentiles of retests of the genetic control. Across retests, we see that the sign and approximate magnitude of behavioral differences repeats.

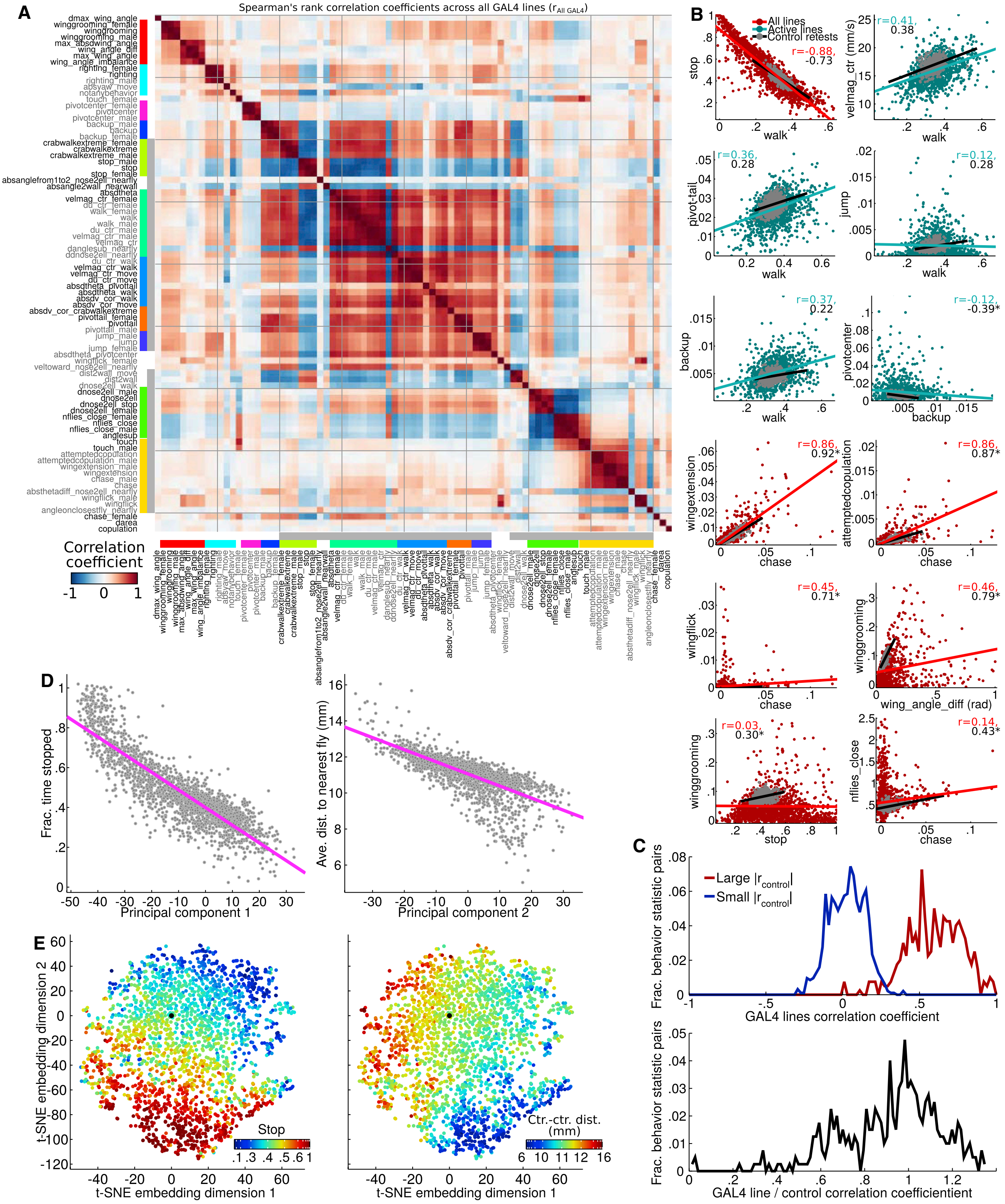


Figure S4. Behavior Data, Related to Figure 3

(A) Correlations between behavior statistics. For each pair of selected per-line behavior statistics, we compute the Spearman's rank correlation coefficient between these values across all GAL4 lines. Red indicates highly positively correlated behavior statistics (e.g., chase and chase_male are highly correlated), blue

(legend continued on next page)

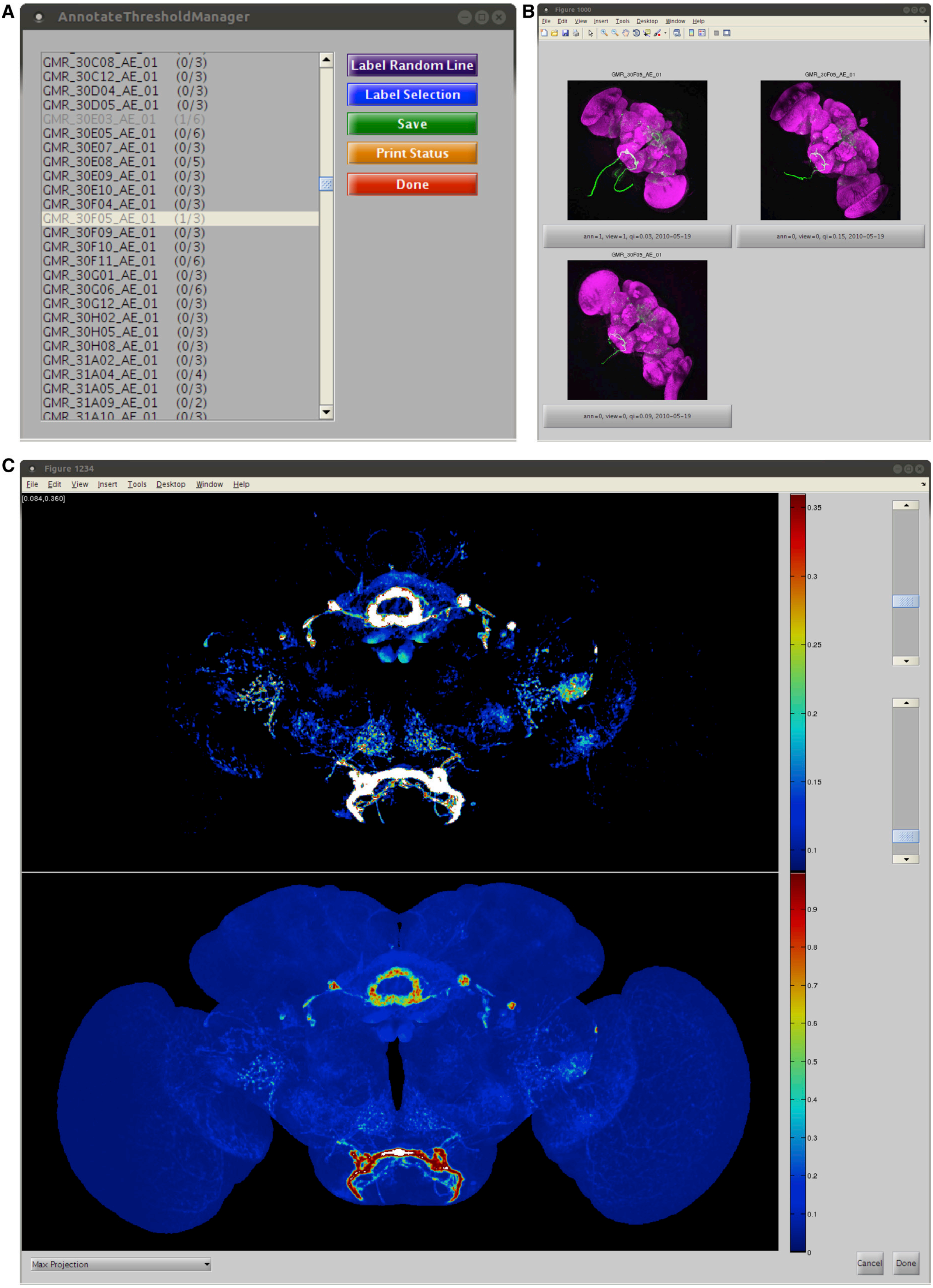
indicates highly negatively correlated statistics (e.g., stop and walk are negatively correlated), and white indicates uncorrelated statistics. Colored bars on the side correspond to the groupings of behavior statistics shown in Figure 3D. Order of behavior statistics set as in Figure 3D.

(B) Behavior statistic correlations across GAL4 lines and control line retests. Each plot corresponds to a different pair of behavior statistics. In each plot, each red or cyan point corresponds to a different GAL4 line and indicates the behavior statistics measured for that GAL4 line. Each gray point corresponds to a different retest of the control line. We plot all GAL4 lines in red and only those GAL4 lines that are active (move at least 55% of the time) in cyan. The red and cyan lines indicate the linear regression fit for the GAL4 lines and the black lines indicate the linear regression fit for the control retests. Correlation coefficients are printed at the top right of each plot, with an asterisk indicating they were significantly different between the GAL4 lines and control retests ($\text{FDR} \leq 0.1$). Correlation coefficients were significantly non-zero for all the pairs of statistics plotted except stop versus wing grooming ($\text{FDR} \leq 0.1$). Pairs of behavior statistics within the same group (Figure 3D) were selected. For many of these pairs, the correlation trends were quite similar, e.g., for the majority of lines that have increased amounts of chasing, we also see increased amounts of wing extension.

(C) Comparison between behavior-statistic correlations for GAL4 lines and control retests. *Top*: Red line: for all 798/3,403 pairs of behavior statistics for which (i) the magnitude of the control retest Spearman's rank correlation coefficient was large (≥ 0.5) and (ii) significantly non-zero ($\text{FDR} \leq 0.1$), and (iii) there was at least one GAL4 line which did not follow this correlation structure (Method Details), we histogram the sign-flipped rank correlation coefficient across the GAL4 lines ($\text{sign}(r_{\text{control}})\ r_{\text{GAL4}}$). This histogram shows that almost all the behavior statistics have high correlations across GAL4 lines of the same sign as the control. In blue, we plot the same histogram for the 1,880/3,403 pairs of behavior statistics for which (i) the magnitude of the control retest Spearman's rank correlation coefficient was small (≤ 0.1) and (ii) not significantly non-zero ($\text{FDR} > 0.1$). This histogram shows that the magnitude of correlations for these behavior statistics is small for the GAL4 lines as well. *Bottom*: Histogram of the ratio of the behavior statistic correlation coefficients across GAL4 lines to that across control retests for the 798/3,403 pairs of behavior statistics with significantly large control correlation magnitudes (red line in top plot).

(D) First (left) and second (right) principal components of GAL4-line behavior vectors plotted against average fraction of time stopped (left) and average distance to nearest other fly (right). Each gray circle corresponds to a GAL4 line.

(E) As in Figure 3G, behavior vectors of all GAL4 lines projected onto their 2-dimensional t-SNE embedding. The t-SNE embedding shows similar gradients to those in the PCA embedding.



GMR_30C08_AE_01 (0/3)
GMR_30C12_AE_01 (0/3)
GMR_30D04_AE_01 (0/3)
GMR_30D05_AE_01 (0/3)
GMR_30E03_AE_01 (1/6)
GMR_30E05_AE_01 (0/6)
GMR_30E07_AE_01 (0/3)
GMR_30E08_AE_01 (0/5)
GMR_30E09_AE_01 (0/3)
GMR_30E10_AE_01 (0/3)
GMR_30F04_AE_01 (0/3)
GMR_30F05_AE_01 (1/3)
GMR_30F09_AE_01 (0/3)
GMR_30F10_AE_01 (0/3)
GMR_30F11_AE_01 (0/6)
GMR_30G01_AE_01 (0/3)
GMR_30G06_AE_01 (0/6)
GMR_30G12_AE_01 (0/3)
GMR_30H02_AE_01 (0/3)
GMR_30H05_AE_01 (0/3)
GMR_30H08_AE_01 (0/3)
GMR_31A02_AE_01 (0/3)
GMR_31A04_AE_01 (0/4)
GMR_31A05_AE_01 (0/3)
GMR_31A09_AE_01 (0/2)
GMR_31A10_AE_01 (0/3)

ann=1, view=1, qi=0.03, 2010-05-19

ann=0, view=0, qi=0.15, 2010-05-19

ann=0, view=0, qi=0.09, 2010-05-19

(legend on next page)

Figure S5. Interface for Manual Annotation of Intensity Normalization Thresholds, Related to Figure 4

(A) Annotation manager lets the user select a line to annotate.

(B) After choosing a line, the user chooses a sample from that line to annotate. The GUI displays a max-projection image, as well as metadata such as the registration quality and the collection date to help the user choose a good image to annotate.

(C) Using the slider bars on the right, the user can adjust the high and low thresholds. The bottom image shows the raw pixel intensity, while the top image shows the normalized intensity for the selected image.

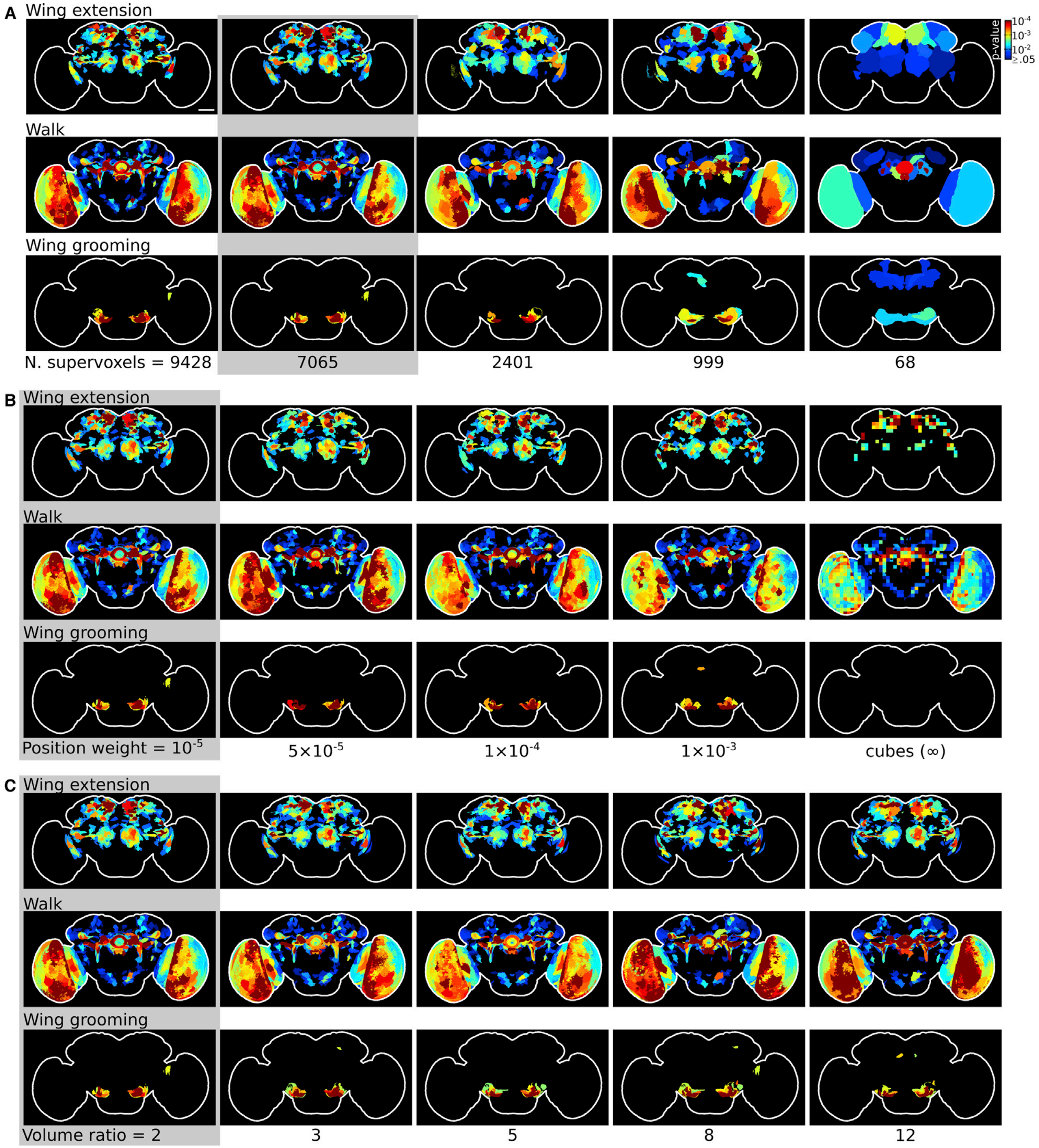


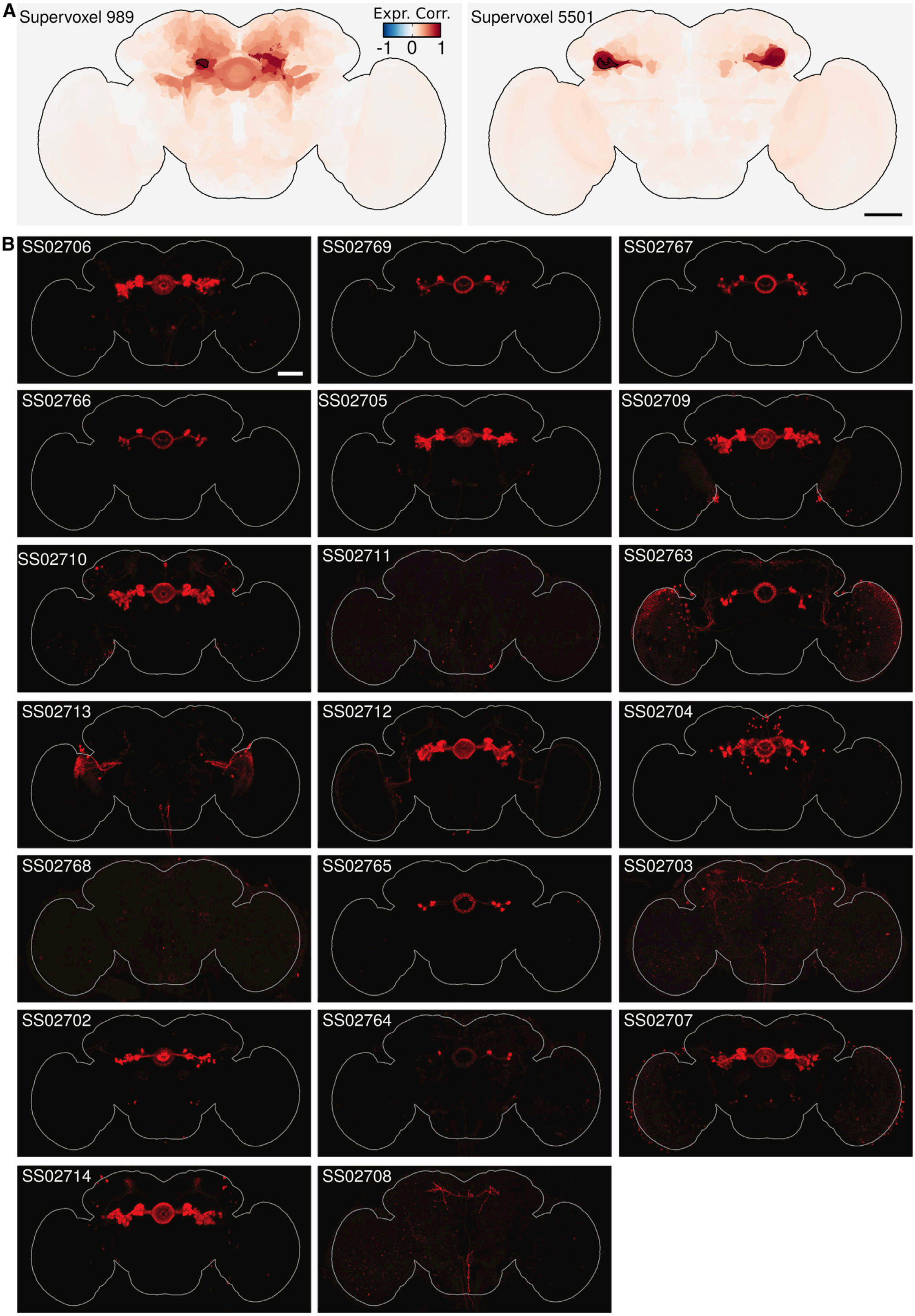
Figure S6. Comparison of Behavior-Anatomy Correlation Maps Obtained with Different Supervoxel Clusterings, Related to Figure 5

(A–C) Each row shows the behavior-anatomy correlation maps for a given behavior (specifically, performing the Wing extension (top), Walk (middle), and Wing grooming (bottom) behaviors more than control), as in Figure 8. Each column corresponds to different values of one of the parameters of the anatomy clustering algorithm. The maps corresponding to the clustering used in the rest of the paper are indicated by the gray boxes, which had 7,065 supervoxels, position weight 10^{-5}, and volume ratio 2 (gray shaded region, all panels). The scale bar indicates 50 μm.

(A) Results for different numbers of supervoxels, from 9,428 (left) to 68 (right). The rightmost column corresponds to a single supervoxel per compartment.

(B) Results of varying the weight of the voxel position versus the expression pattern, from 10^{-5} (left) to 10^{-3} (second-to-right). In the rightmost column, we show the maps resulting from defining supervoxels as 19 × 19 × 19 voxel-cubes (corresponding to an infinite weight to voxel position and ignoring expression pattern).

(C) Results of varying the allowed maximum ratio in volume between the largest and smallest supervoxels from 2 (left) to 12 (right).



(legend on next page)

Figure S7. Walking Map, Related to Figure 7

(A) Correlation in expression across GAL4 lines. We color each supervoxel according to the correlation coefficient of its expression to the expression of the selected supervoxel (outlined in black). To project over the z axis, we select the pixel with the maximum absolute correlation. Some of the supervoxels for which expression is highly correlated correspond to different neurons than those corresponding to the selected supervoxel, and are a result of genetic correlations in expression between different neuronal populations.

(B) Expression pattern of split-GAL4 lines made from walking-more map. A representative brain expression pattern is shown for each of the split-GAL4 lines created based on the walking-more map ellipsoid body region. Expression patterns were obtained with anti-GFP antibody staining and confocal imaging. All split-GAL4 lines except SS02711, SS02713, SS02768, SS02703, and SS02708 had expression in the ellipsoid body. Detailed annotation of the expression patterns is provided in Table S7.

All: Scale bars indicate 50 μm.

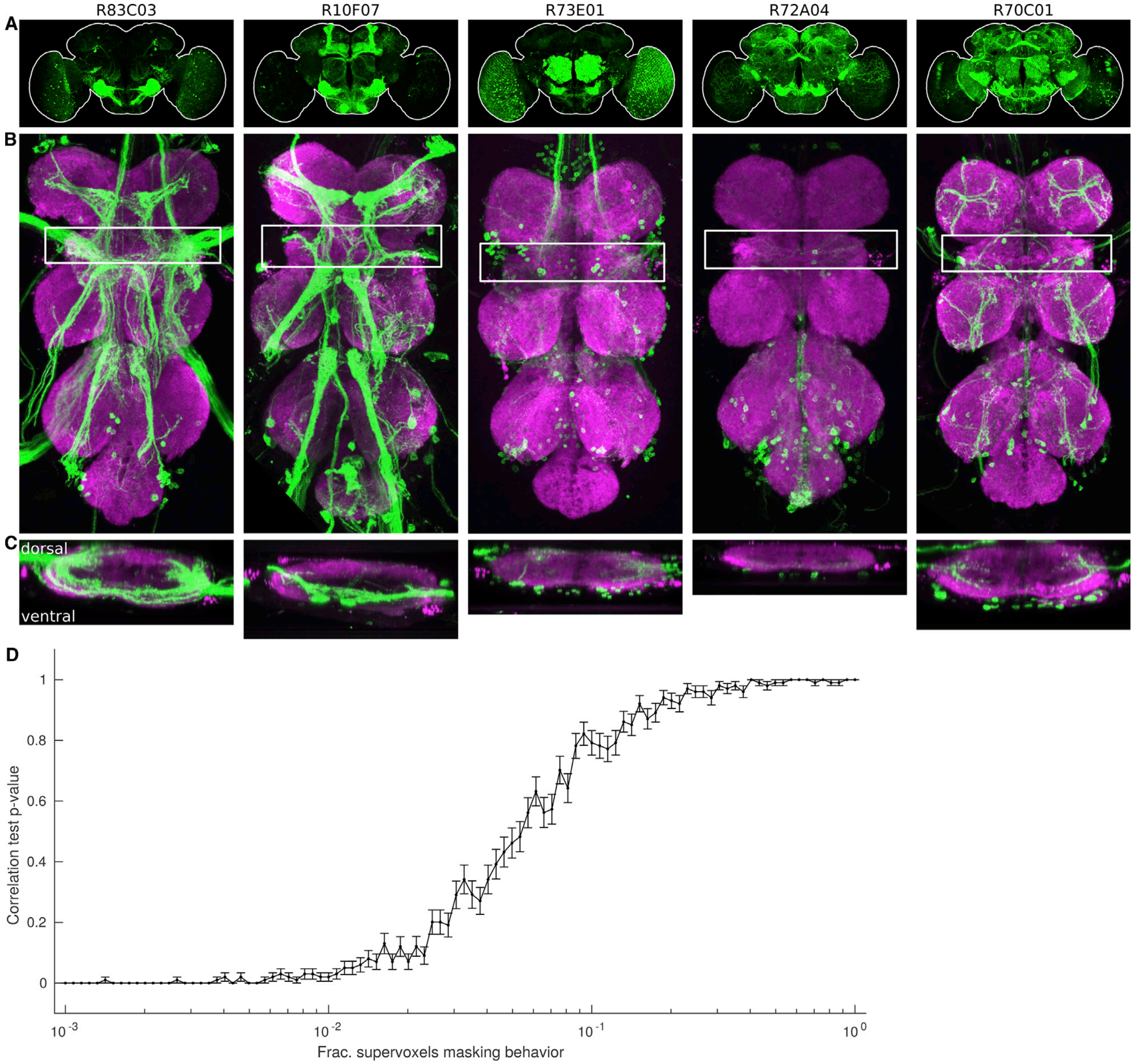


Figure S8. Behavior-Anatomy Maps, Related to Figure 8

(A–C) Maximum projection of expression patterns in (a) the brain and (b-c) the VNC for the 5 lines most important for the increased wing grooming map. VNC images are from the Janelia Fly Light project (Jenett et al., 2012). White box in (b) indicates the location of the wing neuropil, and (c) shows the maximum projection over y for this box. While we see expression in the wing neuropil in several of these lines, we do not see expression in a consistent set of neurons.

(D) Correlation hypothesis test significance for synthetized behavior data in which activation of some supervoxels can mask the behavioral effects of activation another's. The x axis corresponds to the fraction of supervoxels with masking behavior, and the y axis corresponds to the significance of the hypothesis test for the supervoxel whose activation effects can be masked. We plot the mean and standard error over 100 simulations.