bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

NBLAST: Rapid, sensitive comparison of neuronal structure and construction of neuron family databases

Marta Costa[1,2], James D. Manton[1], Aaron D. Ostrovsky[1], Steffen Prohaska[1,3], Gregory S. X. E. Jefferis[1*]

[1]Neurobiology Division, MRC Laboratory of Molecular Biology, Cambridge, CB2 0QH, UK
[2]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
[3]Zuse Institute Berlin (ZIB), 14195 Berlin-Dahlem, Germany

Please note that this preprint is our second public draft. Current versions of the open source software described in the manuscript are available by following links in the Experimental Procedures, which are also summarised at http://jefferislab.org/si/nblast. Processed data derived from raw data generously made publicly available by third parties (primarily flycircuit.tw) will be made available at least by the time this paper is accepted, hopefully rather sooner; please contact Greg for details. We welcome feedback, queries and suggestions on any aspect of the manuscript (including relevant prior art), code or data to jefferis@mrc-lmb.cam.ac.uk.

Abstract

Neural circuit mapping efforts in model organisms are generating multi-terabyte datasets of 10,000s of labelled neurons. Such data demand new computational tools to search and organize neurons. We present a general, sensitive and rapid algorithm, NBLAST, for measuring pairwise neuronal similarity. NBLAST considers both position and local geometry and works by decomposing a query and target neuron into short segments; matched segment pairs are scored using a log-likelihood ratio scoring matrix empirically defined by the statistics of real matches and non-matches.

We validated NBLAST by processing a published dataset of 16,129 single *Drosophila* neurons. NBLAST is sensitive enough to distinguish two images of the same neuron and can be used to distinguish neuronal types without *a priori* information. Detailed cluster analysis of extensively studied neuronal classes identified new neuronal types and unreported features of topographic organization. NBLAST supports diverse additional query types including matching neurite tracts with transgene expression patterns. We organize all 16,129 neurons into 1,052 clusters of highly related neurons, further organized into superclusters, simplifying exploration and identification of neuronal types including sexually dimorphic and visual interneurons.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Correlating the functional properties and behavioral relevance of neurons with their cell type is a basic activity in the study of neuronal circuits. While there is no universally accepted definition of neuron type, key descriptors include morphology, position within the nervous system, genetic markers, connectivity and intrinsic electrophysiological signatures (Migliore and Shepherd, 2005; Bota and Swanson, 2007; Rowe and Stone, 1976). Despite this ambiguity, neuron type remains a key abstraction helping to reveal organizational principles and enabling results to be compared and collated across research groups. Furthermore there is increasing appreciation that highly quantitative approaches are critical to generate the most efficient cell type catalogues in support of circuit research (Petilla Interneuron Nomenclature Group et al., 2008; Nelson et al., 2006; Kepecs and Fishell, 2014)(`http://www.nih.gov/science/brain/11252013-Interim-Report-Final.pdf`).

Since neuronal morphology and position strongly constrain (and are partially defined by) connectivity, they have been mainstays of studies of circuit organization for over a century. Classic techniques to reveal neuronal morphology include the Golgi method made famous by Cajal, microinjection, and filling of cells during intracellular recording. Recently these have been supplemented by genetic approaches to sparse and combinatorial labeling enabling increasingly large-scale characterization of single neuron morphology (Jefferis and Livet, 2012).

Classically, the position of neuronal somata or arbors was established by comparison with anatomical landmarks, often revealed by a general counterstain; this approach is especially effective in brain regions with strong laminar organization e.g. the mammalian retina (Badea and Nathans, 2004; Coombs et al., 2006; Kong et al., 2005; Sümbül et al., 2014), fly optic lobe (Fischbach and Dittrich, 1989; Morante and Desplan, 2008) or cerebellum (Cajal and Azoulay y, 1911). Recently, 3D light microscopy and image registration have enabled direct, automated image fusion to generate digital 3D atlases of brain regions or whole brains (Jefferis et al., 2007; Lin et al., 2007; El Jundi et al., 2009; Rybak et al., 2010; Cachero et al., 2010; Yu et al., 2010b; Sunkin et al., 2013; Zingg et al., 2014; Oh et al., 2014). Such atlases can generate specific, testable hypotheses about circuit organization and connectivity at large scales. The largest study to date Chiang et al. (2011) combined genetic mosaic labeling and image registration to produce an atlas of over 16,000 single cell morphologies embedded within a standard *Drosophila* brain at `http://flycircuit.tw`.

Neuronal morphologies can be represented as directed graph structures embedded in 3D space. However this is typically the (arbitrary) physical space of the imaging system used to reconstruct the neuron, rather than a brain atlas. For this reason, databases such as NeuroMorpho.org (Parekh and Ascoli, 2013) contain > 27,000 neurons, but do not include precise positional information. Data on this scale presents both an acute challenge, finding and organizing related neurons, but also an opportunity: quantitative morphological classification may help solve the problem of cell type. However, a key requirement is a tool enabling rapid and sensitive computation of neuronal similarity within and between datasets. This has clear analogies with bioinformatics: the explosion of biological sequence information from the late 80s motivated the development of sequence similarity tools such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990). These, and related algorithms, enabled pairwise similarity scoring, alignment and rapid database queries as well as hierarchically organized databases of protein families (Sonnhammer et al., 1997).

Several existing strategies for measuring neuronal similarity exist, each with distinct target applications and depending on a particular data structure. Mayerich et al. (2012) have applied general graph

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

similarity metrics (reviewed by Conte et al., 2004) to compare neuronal reconstructions represented as fully-connected graphs with a ground truth reconstruction. Basu et al. (2011) decomposed branching neuronal trees into a family of unbranched paths for which they proposed a geometric measure of similarity which could include positional information. Further simplifying the neuronal representation, Cardona et al. (2010) decomposed single unbranched neurites into sequences of vectors and used dynamic programming to find an optimal 3D alignment. Critically, they validated this approach on a database of a few hundred traced structures, achieving very high classification accuracy. However, while this algorithm could be modified for use with branched neurons, it treats each unbranched neuronal segment as a separate alignment problem and so there is no natural way to handle trees with many such segments.

The choice of data structure remains important: fully automatic 3D tracing of single neurons remains unsolved (Brown et al., 2011), while expression patterns containing multiple neurons cannot be represented as a single binary tree. We previously developed an image segmentation pipeline that represents expression patterns (consisting of up to 100 neurons) as point clouds with tangent vectors defining the local heading of the neurons. We used this simplified representation in a supervised learning approach to the challenging problem of recognizing groups of lineage-related neurons (Masse et al., 2012). Ganglberger et al. (2014) have recently applied a related approach directly to unsegmented expression pattern image data at the expense of much higher memory demands (order 100 MB per specimen), severely limiting throughput.

Combining the data representation of Masse et al. (2012) with a very large single neuron dataset (Chiang et al., 2011) allowed us to test and validate a new algorithm, NBLAST, that is flexible, extremely sensitive and very fast (pairwise search times < 5 ms). Critically, the algorithm's scoring parameters are defined statistically rather than by expert intuition.

We first describe the NBLAST algorithm, providing an open source command line implementation and a web query tool (see `jefferislab.org/si/nblast/clusters`). We validate NBLAST for applications including neuron database search and unsupervised clustering of neurons. NBLAST can identify well-studied neuronal types in *Drosophila* with sensitivity matching domain experts, in a fraction of the time of manual classification. NBLAST can also identify new neuronal types and reveal undescribed features of topographic organization. Finally, we apply our method to 16,129 neurons from the FlyCircuit dataset, reducing this to a non-redundant set of 1,052 morphological clusters. Manual evaluation of a subset of these clusters show that they closely match expert definition of cell types. These clusters, which we also organize into an online supercluster hierarchy, therefore represent a preliminary global cell type classification for the *Drosophila* brain.

Results

Algorithm

Our principal design goals were to develop a neuron similarity algorithm that included aspects of both spatial location (within a brain or brain region) and neuronal branching pattern, and that was both extremely sensitive and very fast. The applications that we had in mind were the searching of large databases of neurons (10,000–100,000 neurons), clustering of neurons into families by calculating all-against-all similarity matrices, and the efficient organization and navigation of datasets of this size. We eventually selected an approach based on direct pairwise of comparison of neurons pre-registered

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to a template brain and represented as vector clouds. Further details are provided in Supplemental Information.

The starting point for our algorithm is a representation in which neuron structures have been reduced to segments, represented as a location and an associated local tangent vector. This retains some local geometry but does not attempt to capture the topology of the neuron's branching structure. We have found that a simplified representation of this sort can be constructed for image data that would not permit automated reconstructions. In order to prepare data of this sort in quantity, we developed an image processing pipeline summarized in Figure 1A and detailed in Experimental Procedures. Briefly, brain images from the FlyCircuit dataset (Chiang et al., 2011) were subjected to non-rigid image registration (Jefferis et al., 2007) against a newly constructed intersex template brain. Neuron images were thresholded and subjected to a 3D skeletonization procedure (Lee et al., 1994) implemented in Fiji (Schindelin et al., 2012). These thresholded images were then converted to the point and tangent vector representation (Masse et al., 2012) using our R package `nat` (Jefferis and Manton, 2014); the tangent vector (i.e. the local heading) of the neuron at each point was computed as the first eigenvector of a singular value decomposition (SVD) of the point and its 5 nearest neighbors.

After pre-processing, 3D data could be visualized and analyzed in R using `nat` (Figure 1B). Neurons were represented by median 1070 points/vectors; the 16,129 neurons occupied 1.8 GB, fitting comfortably into a laptop's main memory. Since the fly brain is almost completely symmetric, but neurons were labelled randomly in both hemispheres, we mapped all neurons to the left hemisphere (defined primarily by cell body location, see Experimental Procedures and Figure 1B) using a non-rigid mirroring procedure (Manton et al., 2014).

With a database of aligned neurons in an appropriate representation, we were then able to calculate NBLAST pairwise similarity scores. One neuron is designated the query and the other the target. For each query segment (defined by a midpoint and tangent vector) the nearest neighbor (using straightforward Euclidean distance) is identified in the target neuron (Figure 1C–D). A score for the segment pair is calculated as a function of two measurements: d, the distance between the matched segments (indexed by i), and $|\vec{u}_i \cdot \vec{v}_i|$, the absolute dot product of the two tangent vectors; the absolute dot product is used because the orientation of the tangent vectors has no meaning in our data representation (Figure 1C). The scores are then summed over each segment pair to give a raw score, S:

$$S(\text{query}, \text{target}) = \sum_{i=1}^{n} f(d_i, |\vec{u}_i \cdot \vec{v}_i|). \tag{1}$$

The question then becomes: what is an appropriate function $f(d_i, |\vec{u}_i \cdot \vec{v}_i|)$? We developed an approach inspired by the scoring system of the BLAST algorithm (Altschul et al., 1990). For each segment pair we defined the score as the log probability ratio:

$$f = \log_2 \frac{p_{\text{match}}}{p_{\text{rand}}}, \tag{2}$$

i.e. the probability that the segment pair was derived from a pair of neurons of the same type, versus a pair of unrelated neurons . We could then define p_{match} empirically by finding the joint distribution of d and $|\vec{u}_i \cdot \vec{v}_i|$ for pairs of neurons of the same type (Figure 1E–G). For our default scoring matrix, we used a set of 150 olfactory projection neurons innervating the same glomerulus, unambiguously the same neuronal type (Figure 1F). p_{rand} was calculated simply by drawing 5,000 random pairs of

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neurons from the database, assuming that the large majority of such pairs are unrelated neurons. Joint distributions were calculated using 10 bins for the absolute dot product and 21 bins for the distance to give two 21 row × 10 column matrices. The 2D histograms were then normalized to convert them to probabilities and the log ratio defined the final scoring matrix (Figure 1G). Plotting the scoring matrix emphasizes the strong distance dependence of the score but also shows that for segment pairs closer than ~10 µm, the logarithm of the odds score increases markedly as the absolute dot product moves from 0 to 1 (Figure 1H).

We implemented the NBLAST algorithm as an R package (`nat.nblast`) built on top of a high-performance k nearest neighbor library (http://cran.r-project.org/web/packages/RANN/index.html), that immediately enables pairwise queries, searches of a single query neuron against a database of target neurons (Figure 2) and all-by-all searches. Runtimes on a single core laptop computer were 2 ms per comparison or 30 s for all 16,129 neurons. In order to enable clustering of neurons on the fly, we also pre-computed an all-by-all similarity matrix for all 16,129 neurons (2.6×10^8 scores, 1.0 GB). We also developed a simple web application (linked from `jefferislab.org/si/nblast`) to allow online queries for this test dataset.

NBLAST can find whole or partial matches for diverse query objects

The NBLAST algorithm is flexible, identifying both global and partial matches for multiple classes of query object (Figure 2). The only requirements are that query objects (or fragments) must be registered against a template brain and can be converted to a point and vector representation.

As a first example we query a (whole) FlyCircuit neuron against the 16,129 FlyCircuit neurons. The top hits are all very similar neurons with small differences in length and neurite position (Figure 2B). A second example uses a neurite fragment, corresponding to part of the axon; top hits all follow the same axon tract, although their variable axonal and dendritic arbors indicate that they are distinct neuron types (Figure 2C).

User tracings can also be used as queries. We traced the characteristic bundle of 20-30 primary neurites of the *fruitless* neuroblast clone pMP-e (which generates male-specific P1 neurons, (Kimura et al., 2008; Cachero et al., 2010)). Our query returned many single P1 neurons from the FlyCircuit database (Figure 2D) (for more details see Figure 7). A similar approach can be used to identify candidate neuronal types labelled by genetic driver lines where the detailed morphology of individual neurons cannot be ascertained. As an example we take the GAL4 line R18C12 (Jenett et al., 2012) (Figure 2E). The expression pattern includes an obvious bilateral dorsal tract associated with a specific cluster of cell bodies (Figure 2E). We traced the main neurites of this cell cluster; NBLAST identified three very similar FlyCircuit neurons, which completely overlapped with the expression of R18C12. These three neurons appear to be different subtypes, each varying in their terminal arborizations. Conversely, we used one tracing from a recently published projectome dataset containing >9000 neurite fibers (Peng et al., 2014) to find similar FlyCircuit neurons (Figure 2F).

NBLAST scores are sensitive and biologically meaningful

A good similarity algorithm should be sensitive enough to reveal identical neurons with certainty, while having the specificity to ensure that all high scoring results are relevant hits. We used the full FlyCircuit dataset to validate NBLAST performance.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our first example uses an auditory interneuron, fru-M-300198 as query (Figure 3A–C). Ordering search results by NBLAST score, the first returned object is the query neuron itself (since it is present in the database), followed by the top hit (fru-M-300174) which completely overlaps with the query (Figure 3A'). A histogram plot of NBLAST scores, showed that the top hit score was clearly an outlier: 96.1 % of the self-match score of the query against itself (i.e. the maximum possible score) (Figure 3C). Further investigation revealed that these "identical twins", both derived from the same raw confocal image. The next 8 hits are also very similar to the query but are clearly distinct specimens, having small differences in position, length and neurite branching that are typical of sister neurons of the same type (Figure 3A").

The score histogram shows that only a minority of hits (3 %) have a score above 0 (Figure 3B–C). A score of 0 represents a natural cutoff for NBLAST, since it means that, on average, segments from this neuron have a similarity level that is equally likely to have arisen from a random pair of neurons in the database as a pair of neurons of the same type. We divided the neurons with score>0 into 8 groups with decreasing similarity scores (Figure 3C'). Only the highest scoring real hits (group II) appear of exactly the same type, although lower scoring groups contain neurons that would be ranked as very similar.

Although raw NBLAST scores correctly identify similar neurons, they are not comparable from one query neuron to the next: the score depends on neuron size and segment number. This confounds search results for neurons of very different sizes or when the identity of query and target neurons is reversed. For example, a search with a large neuron as query and a smaller one as target (pair 1) will have a very low **forward** score, because the large neuron has many segments that are unmatched, but a high **reverse** score, since most of target will match part of the query (Figure 3D). One approach to correct for this is to normalize the scores by the size of the query neuron. Although **normalized** scores are comparable, unequal forward and reverse scores between large and small neurons remain an issue. One simple strategy is to calculate the mean of the forward and reverse scores (**mean score**). Two neurons of similar size have a higher mean score than two neurons of unequal size (Figure 3D). Repeating the analysis of Figure 3C–C' using mean scores (Figure S2) eliminated some matches due to unequal size that could be considered false positives.

During our analysis, we sporadically noticed cases where two neurons in the database actually were actually the same physical specimen (Figure S1). We tested if NBLAST could systematically reveal such instances. We collected the top hit for each neuron and analyzed the distribution of forward (Figure 3E) and reverse scores (data not shown). A small tail (~ 1 % of all top hits) have anomalously high scores (over 0.8). Given this distribution, we examined neuron pairs with both forward and reverse scores over 0.8. We classified these 72 pairs into 4 different groups. From highest to lowest predicted similarity, the groups are: same segmentation, a segmented image of a neuron has been duplicated (Figure S1A); same raw image, corresponding to a different segmentation of the same neuron (Figure 3B'); same specimen, when two images are from the same brain but not from the same confocal image (Figure S1B); and different specimen, when two neurons are actually from different brains, (therefore suggesting that they are of the same type). The distribution of NBLAST scores for these four categories matches the predicted hierarchy of similarity (Figure 3F). These results underline the high sensitivity of the NBLAST algorithm to small differences between neurons.

Taken together these results validate NBLAST as a sensitive and specific tool for finding similar neurons.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

NBLAST scores can distinguish Kenyon cell classes

We wished to investigate whether NBLAST scores can be used to cluster neurons by structure and position, potentially revealing functional classes. We decided to begin our investigation of this issue with Kenyon cells (KCs), the intrinsic neurons of the mushroom body neuropil and one of the most extensively studied category of neurons given their key role in memory formation and retrieval (reviewed in Kahsai and Zars, 2011).

There are around 2,000 KCs in each mushroom body Aso et al. 2009, and they form the medial lobe, consisting of the γ, β′ and β lobes, the vertical lobe, consisting of the α and α′ lobes, the calyx, where dendrites are found and around which cell bodies are positioned, and the peduncle, formed by the anterior projection of the axons before joining the lobes (Figure 4A). Three main classes of KCs and a few subclasses of neurons are recognized: the γ neurons are the first to be born and innervate only the γ lobe; the α′/β′ neurons are generated next and project to the α′ and β′ lobes; the α/β neurons are the last to be born, and project to both the α and β lobes. Four neuroblast clones which differ in their position in the calyx generate the KCs, with each one generating the whole repertoire of neuron types (Lee et al., 1999).

We started with a dataset of 1,664 KCs, representing 10.3 % of the FlyCircuit dataset (for details of dataset collection see Supplemental Results) and collected raw NBLAST scores of each KC against all others. An iterative hierarchical clustering approach allowed us to identify the main KC types and subsequent additional analysis for each of these distinguished several subtypes.

In the case of the γ neurons (Figure 4B'), we identified 2 subsets, one corresponding to the classical morphology (Figure 4B") (groups I and III) and another to previously described atypical neurons (γ dorsal neurons, group II) (Aso et al., 2014). Analysis of the classical γ neurons revealed that there were differences between the neurons in their medial to lateral position in the calyx (groups A-D). These differences correlated to a certain degree with differences in the dorsal/ventral position of the projections in the γ lobe, with the most medial, being also the most dorsal (Figure 4B"). These observations suggest that the relative position of the projections of classical γ neurons is maintained at the calyx and γ lobe. We experimented with clustering the classic γ neurons based only on the scores of the segments in the peduncle. The overall organization almost fully recapitulated the positioning of the neurites in the whole neuron analysis (see Figure S3 and). Thus, the stereotypical organization of the classical γ neurons is maintained throughout the neuropil.

The atypical γ neurons extended neurites posteriolaterally in the calyx and projected to the most dorsal region of the γ lobe (Figure 4B"'). We isolated a previously identified subtype –γd neurons (group a) (Aso et al., 2009)– that innervates the ventral accessory calyx (Aso et al., 2014). In addition, we identified previously uncharacterized types (Groups b-c).

Analysis of α′/β′ neurons highlighted the characterized subtypes of these neurons (Figure 4C–C') which differ in their anterior/posterior position in the peduncle and β′ lobe (Tanaka et al., 2008; Aso et al., 2014). Although we were unable to unambiguously assign a α′/β′ subtype to our neurons, there were clear trends, with a subset of neurons displaying neurites more anteriorly than others (groups ii, iii) in both the peduncle and β′ lobe.

The largest subset of KCs corresponds to α/β neurons (Figure 4D). We identified neurons from each of the four neuroblast lineages (Figure 4D') (Zhu et al., 2003) and for each of these, we distinguished morphological subtypes that correlate to their birth time (Figure 4D"). There was a clear distinction between the late born core (α/β core, α/β-c), on the inside stratum of the α lobe, and early born peripheral

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neurons (α/β surface, α/β-s), on the outside stratum of the α lobe. We also identified the earliest born α/β neurons– α/β posterior or pioneer (α/βp)– that innervate the accessory calyx and run along the surface of the posterior peduncle into the β lobe but stop before reaching the medial tip (Tanaka et al., 2008). A new clustering based on peduncle position of the neuron segments did not recapitulate the relative positions of the calyx neurites for each of the neuroblast clones observed in the whole neuron analysis suggesting that the relative position of the α/β neurons in the peduncle does not completely reflect their stereotypical organization in the calyx (see Figure S3 and).

In summary, the hierarchical clustering of KCs using the raw NBLAST scores resolved the neurons into the previously described KC types and some of the subtypes, and isolated uncharacterized subtypes in an extensively studied cell population. In addition, it revealed organizational principles that have been previously described (Tanaka et al., 2008). These observation support our claim that the NBLAST scores retain enough morphological information to accurately search for similar neurons and organize large datasets of related cells.

NBLAST identifies classic cell types at the finest level: olfactory projection neurons

We have shown that clustering based on NBLAST scores can identify the major classes and subtypes of Kenyon cells. However it is rather unclear what corresponds to an identified cell type, which we take to be the finest classification of neuron in the brain. We therefore analyzed a different neuron family, the olfactory projection neurons (PNs), which represent one of the best defined cell types in the fly brain.

PNs transmit information between antennal lobe glomeruli, which receive sensory input, and higher olfactory brain centers, including the mushroom body and the lateral horn (Masse et al., 2009). Uniglomerular PNs (uPNs), whose dendrites innervate just one glomerulus, are highly stereotyped in both morphology and developmental origin. They are classified into individual types based on the glomerulus they innervate and the axon tract they follow; these features show fixed relationships with their axonal branching patterns in higher brain centers and their parental neuroblast (Marin et al., 2002; Jefferis et al., 2001; Wong et al., 2002; Jefferis et al., 2007; Yu et al., 2010a; Tanaka et al., 2012).

We manually classified the 400 uPNs in the FlyCircuit dataset by glomerulus, and defined the manual gold standard annotations in an iterative process that took several days (for details see). We found a very large number of DL2 uPNs (145 DL2d and 37 DL2v neurons), in a total of 397 neurons. Nevertheless, our final set of uPNs broadly represents the total variability of described classes and contains neurons innervating 35 out of 56 different glomeruli (Tanaka et al., 2012), as well as examples of the three main lineage clones and tracts.

We computed mean NBLAST scores for each uPN against the remaining 16,128 neurons and checked whether the top hit was exactly the same type of uPN, another uPN or a match to another class of neuron (Figure 5A). We restricted our analysis to types with at least two examples in the dataset and to unique pairs (i.e. if PN A was the top hit for PN B and vice versa, we only counted them once) (n=327). There were only 8 cases in which the top hit did not match the class of the query. Of these, four had matches to a uPN innervating a neighboring glomerulus with identical axon projections (DL2d vs DL2v, VM5d vs VM5v) that are challenging even for experts to distinguish. There were a further four matches to neurons that were not uPNs, but corresponded to other PNs that innervated the same glomerulus.

We also compared how the top 3 hits matched the query type (Figure 5B). For uPN types with

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

more than three examples (non DL2, n=187), we collected the top three NBLAST hits for each of these neurons. We achieved very high matching rates: in 98.9 % of cases (i.e. all except two) at least one of the top hits matched the query type, and all three hits matched the query type 95.2% of cases.

Given the very high prediction accuracy, we wondered if an unsupervised clustering based on NBLAST mean scores would group uPNs by type. To test this, we clustered uPNs (non DL2, n=214) and divided the dendrogram at a height of 0.725, as this level was found to be the one at which most groups corresponded to single and unique neuron types. For types with more than one representative neuron, all neurons co-clustered, with three exceptions (Figure 5C). The cluster organization also reflects higher level features such as the axon tract / neuroblast of origin (Figure 5C'). Thus, unsupervised clustering of uPNs based on NBLAST scores gives an almost perfect neuronal classification.

In conclusion, these results demonstrate that morphological comparison by NBLAST is powerful enough to resolve differences at the finest level of neuronal classification. Furthermore, they suggest that unsupervised clustering by NBLAST scores could help to reveal new neuronal types.

NBLAST can be used to define new cell types

Visual projection neurons

Visual projection neurons (VPNs) relay information between the optic lobes and the central brain. They are a morphologically diverse group that innervate distinct optic lobe and central brain neuropils, with 44 types already described (Otsuna and Ito, 2006). We explored whether clustering of these neurons based on NBLAST scores would find previously reported neuron classes and identify new ones.

We isolated a set of VPNs including 1,793 unilateral VPNs, 72 bilateral VPNs and 2892 intrinsic optic lobe neurons. Hierarchical clustering of the unilateral VPNs (uVPNs) resulted in a dendrogram, which we divided into 21 groups (I-XXI), in order to isolate one or a few cell types by group based both directly on morphological stereotypy and on our reading of the previous literature (Otsuna and Ito, 2006)(Figure 6A–A' and Figure S4A–A'). We further investigated these groups to determine if central brain innervation was a major differentiating characteristic between classes.

Lobula-, AOTU- and PVLP-innervating uVPNs We took neuron skeletons from groups I–III uVPN and isolated only the axon arbors innervating the anterior optic tubercle (AOTU) and posterior ventrolateral protocerebrum (PVLP) (Figure 6B). A new clustering based on the scores of these partial skeletons, allowed us to identify seven different groups (1-7). A clear distinction between neurons that innervated the PVLP (groups 1, 2) and those that extensively innervated the AOTU (groups 3–6) was evident. Our analysis divides the LC10 uVPN class into 5 subgroups, four of them not previously identified (Table S1 and Figure 6C).

Lobula-, PVLP- and PLP-innervating uVPNs We performed a similar analysis with uVPN groups that had dendritic innervation restricted to the lobula and axons projecting to the PVLP and posterior lateral protocerebrum (PLP) (groups IV, VI, VII and XI) (Figure S4A'). Following the same strategy, we re-clustered neurons based on the scores calculated only for the axon arbors that overlapped with the PVLP or PLP (Figure S4B). We also obtained seven distinct types, including a new subtype (Figure S4B–C and Table S2).

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bilateral VPNs In addition to the analysis of the uVPNs, we also performed a hierarchical clustering of the bilateral VPNs (Figure 6A"). Of the resulting 8 distinct groups (i–viii), we were able to match one to the bilateral LC14 neurons (Otsuna and Ito, 2006).

VPN summary Our analysis of VPN neurons has demonstrated that similarity searches performed with only part of a neuron are useful to highlight morphological features that might be most important for defining neuron classes. We were able to match 11 of our defined groups to known VPN types, and furthermore described two new subclasses and four subtypes of uVPNs, showing that this type of analysis is able to identify new cell types even for intensively studied neuronal classes.

Auditory neurons

Auditory projection neurons (PNs) are characterized by their innervation of the primary or secondary auditory neuropils, the antennal mechanosensory and motor center (AMMC) and the inferior ventrolateral protocerebrum (IVLP or wedge). Several distinct types have been described based on anatomical and physiological features (Yorozu et al., 2009; Lai et al., 2012; Kamikouchi et al., 2006, 2009). Just as for the VPNs, we tested our ability to identify known and new cell types. In this case, we employed a two-step search strategy using a previously identified FlyCircuit neuron named by Lai et al. (2012) as the seed neuron for the first search (for details see). For each of the 5 types we analyzed, hierarchical clustering of the hits revealed new subtypes of known auditory PN types that differed mainly in their lateral arborizations (Figure S5E and Table S3).

mAL neurons

The *fruitless*-expressing mAL neurons are sexually dimorphic interneurons that are known to regulate wing extension by males during courtship song (Koganezawa et al., 2010; Kimura et al., 2005). Males have around 30 neurons, but there are only 5 in females. Although the gross neuronal morphology is similar in both sexes, both axonal and dendritic arborisations are located in distinct regions, likely altering input and output connectivity. We investigated whether clustering could distinguish male and female neurons and identify male subtypes. Clustering a set of 41 mAL neurons (for dataset collection see) cleanly separated male from female neurons (Figure 7A–B). Clustering analysis of the male neurons using partial skeletons that only contained the axonal and dendritic arbors (Figure 7C), identified 3 main types and 2 subtypes of male mAL neurons. The 3 types differed in the length of the ipsilateral ventral projection; this feature has previously been proposed as the basis of a qualitative classification of mAL neurons (Kimura et al., 2005). However all types and subtypes also showed reproducible differences in the exact location of their axon terminal arbors. Our analysis therefore suggests that the population of male neurons includes types with correlated differences in input-output connectivity.

P1 neurons

P1 neurons are the most significantly dimorphic *fruitless*-expressing neurons. Male P1 neurons are involved in the initiation of male courtship behavior while female P1 neurons degenerate during development due to the action of *doublesex* (Kimura et al., 2008). There are around 20 P1 neurons that develop from the pMP-e *fruitless* neuroblast clone. They have extensive bilateral arborizations in the

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PVLP and ring neuropil, partially overlapping with the male-specific enlarged brain regions (MER) (Kimura et al., 2008; Cachero et al., 2010) (Figure 7D).

Despite their critical role, P1 neurons have been treated as an homogeneous neuronal class. We therefore investigated whether clustering could identify anatomical subtypes. Hierarchical clustering of the P1 neuron set (for dataset collection see), distinguished 10 groups (Figure 7D'). Nine of these (1–9) contain only male *fru*-GAL4 neurons as expected. The 9 male groups have the same distinctive primary neurite and send contralateral axonal projections through the arch (Yu et al., 2010b) with extensive arborizations in the MER regions (Cachero et al., 2010). However each group shows a highly distinctive pattern of dendritic and axonal arborisations suggesting that they are likely to integrate distinct sensory inputs and to connect with distinct downstream targets.

Intriguingly, group 10 consists only of female neurons, including two female *fru*-GAL4 and two other drivers. Their morphology is clearly similar to but distinct from group 9 neurons, suggesting that a small population of neurons that share anatomical (and likely developmental) features with male P1 neurons is also present in females.

Superclusters and Exemplars to Organize Huge Data

In the previous examples we have shown that NBLAST is a powerful tool to identify known and uncover new neuron types when analyzing specific neuron superclasses within large datasets. However, subsetting the dataset in order to isolate the chosen neurons requires considerable time. We wished to establish a method that would allows us to organize large datasets, extracting the main types automatically, and retain information on the similarity between types and subtypes, while providing a quicker way to navigate datasets. We used the affinity propagation method of clustering (Frey and Dueck, 2007), combined with hierarchical clustering to achieve this. Applying affinity propagation to the 16,129 neurons in the FlyCircuit dataset resulted in 1,052 clusters (Figure 8A–B). Using hierarchical clustering of the exemplars and by manually removing a few stray neurons, we isolated the central brain neurons (groups B–C) from the optic lobe and visual projection ones (group A) (Figure 8C). Another step of hierarchical clustering of central brain exemplars revealed large superclasses of neuron types when we divided the dendrogram in 14 groups (I–XIV). Each one contained a distinct subset of neuron types including, for example, central complex neurons (I), P1 neurons (II), 2 groups of KCs (γ and α'/β' and α/β) (IV–V) and auditory neurons (VIII) (Figure 8D–D').

The affinity propagation clusters are also useful for identifying neuronal subtypes by comparing all clusters that contain a specified neuronal type (Figure 8E). We present examples for the neuronal types AMMC-IVLP projection neuron 1 (AMMC-IVLP PN1) (Lai et al., 2012), and the uVPNs LC10B and LC4. For each of these, morphological differences are clear between clusters, suggesting that each one might help to identify distinct subtypes.

We have shown that combining affinity propagation with hierarchical clustering is an effective way to organize and explore large datasets, by condensing information into a single exemplar and by retaining the ability to move up or down in the hierarchical tree, allowing the analysis of superclasses or more detailed subtypes.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

The challenge of mapping and cataloguing the full spectrum of neuronal types in the brain depends not only on the ability to recognize similar neurons, by shape and position, but also on establishing methods that facilitate unbiased identification of neuron types from pools of thousands or millions of individual neurons. The comparison of neurons relies both on morphology and position within the brain, as this is an essential determinant of their function and synaptic partners. A neuron search algorithm should therefore be: (1) accurate, with hits being biologically meaningful; (2) fast and computationally inexpensive; (3) provide an interactive search method and (4) generally applicable. Here, we have described NBLAST, a neuron search algorithm that satisfies all these criteria.

First, the algorithm correctly distinguishes closely related subtypes across a range of major neuron groups, with an accuracy of 97.6 % in the case of olfactory projection neurons. Unsupervised clustering of these neurons, based on NBLAST scores, correctly organized neurons into described types. We did find, however, that the size of a neuron influences the accuracy of algorithm, especially for smaller neurons, even when using the normalized score. One future research area will be to convert the raw scores that we have used into an expectation (E) value (cf. BLAST), that would directly account for the size of a neuron.

Second, NBLAST searches are very fast, with pairwise comparisons taking about 2 ms on a laptop computer, with queries against the whole 16,000 neuron dataset taking about 30 s. Furthermore, for defined datasets all-by-all scores can be pre-computed allowing immediate retrieval of NBLAST scores for highly interactive analysis. With the amount of data available only expected to increase, strategies to query and store these data need to be investigated. One effective approach to handle much larger number of neurons will be to compute sparse similarity matrices, storing only the top n hits for a given neuron, an approach often taken for genome-wide precomputed BLAST scores. Alternatively, queries could be computed only against the non-redundant set of neurons that collectively embody the structure of the brain, similarly to the strategy employed by UniProt (Suzek et al., 2007). At most, this set would not exceed 50,000 neurons (due to the strong bilateral symmetry in the fly brain) and we expect that it would in practice not need to exceed 5,000. Our clustering of all ~16,000 neurons of the FlyCircuit dataset identified ~1,000 exemplars providing a non-redundant data set that could be used for rapid searches.

Third, our method permits a variety of different search strategies from a variety of objects. Searches with whole or neuron fragments, or tracings can be used to distinguish closely related neuronal types by their terminal arbors or to identify candidate neurons from a GAL4 line.

Finally, one important question is obviously the extent to which our approach can be generalized. This issue largely reduces to the relationship between the length scales of neurons being examined and their absolute spatial stereotypy. Our method implicitly assumes spatial co-localization of related neurons; this is enforced in our input data by the use of image registration. Our search strategy should therefore be appropriate for any situation in which neuronal organization is highly stereotyped at the length scale of the neuron under consideration. There is already strong evidence that this is true across large parts of the brain for simple vertebrate models like the larval zebrafish: indeed Portugues et al. (2014) used exactly the same registration software to demonstrate highly spatially stereotyped visuomotor activity patterns. Preliminary analysis (GSXEJ and JDM, `https://github.com/jefferis/nat.examples/tree/master/05-miyasaka2014`) suggests that our method can be applied directly to olfactory projectome data (Miyasaka et al., 2014) from larval zebrafish. Mouse gene expression

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Lein et al., 2007) and long range connectivity also show global spatial stereotypy as evidenced by recent atlas studies combining sparse labeling and image registration (Zingg et al., 2014; Susaki et al., 2014; Oh et al., 2014)). Our method should allow simple querying and hierarchical organization of these datasets with relatively little modification beyond calculating an appropriate scoring matrix.

However there are clearly situations in which global brain registration is not an appropriate starting point. For example the vertebrate retina has both a laminar and a tangential organization. Recently Sümbül et al. (2014) have introduced a registration strategy that demonstrates that the lamination of retinal ganglion cells in mouse retina is spatially stereotyped to the nearest micron. However retinal interneurons and ganglion cells are organized in mosaics across the retinal surface (typically referred to as the XY plane). Therefore global registration is not appropriate in this axis, rather it is necessary to align neurons into a virtual column. The situation is similar for *Drosophila* columnar neurons in the outer neuropils of optic lobe, for which neurons are organized into about 800 parallel columns (reviewed in Paulk et al., 2013). There are two ways that we envisage this situation can be handled. The first would be to carry out a local re-registration, that maps each column onto a single canonical column. The second would be to amass sufficient data that neurons from neighboring columns would tile the brain, enabling their identification as a related group by standard clustering or graph theoretic approaches.

The aim of cataloguing all neuron types in the brain relies not only on an accurate algorithm to find similar neurons, but also on having an easy and unbiased method to distinguish neurons types and/or subtypes. This is a challenging problem, but morphological approaches may eventually provide unambiguous automated classification. Sümbül et al. (2014) recently explored the issue of defining the optimal cut height for morphological clustering of mouse retinal ganglion cells, establishing a reliable approach for these specific neurons. We have demonstrated the applicability of NBLAST across a very wide range of neuronal classes using hierarchical clustering and cutting the dendrogram at a specified height. This process of identification relied on extensive data exploration and iteration. We believe that the extent of morphological variability within a neuronal type precludes the existence of one unique value for dendrogram cutting height. Instead, the range of heights we found, between 0.7 and 2, can guide future exploration for other neuron types and datasets, although this process will still require iterative analysis and manual verification.

Experimental Procedures

Image Preprocessing

The flycircuit.tw team supplied 16,226 raw confocal stacks in the Zeiss LSM format on a single 2 TB hard drive in April 2011. Each LSM stack was first uncompressed, then read into Fiji/ImageJ (`http://fiji.sc/Fiji`) where the channels were split and resaved as individual gzip compressed NRRD (`http://teem.sourceforge.net/nrrd`) files. Where calibration information was missing from the LSM file metadata, we used a voxel size of (0.318427, 0.318427, 1.00935) microns as recommended by the FlyCircuit team. There were two important issues to solve before images could be used for registration: 1) identifying which image channel contained the anti-Dlg (discs large 1) counterstaining revealing overall brain structure and 2) determining whether the brains had been imaged from anterior to posterior, or posterior to anterior. The first issue could be solved by exporting the metadata associated with each LSM file using the LOCI bioformats (`http://loci.wisc.edu/`

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

software/bio-formats) plugin for Fiji and developing some heuristics to automate the identification of the channel sequence; for a minority of images this metadata was missing and the channel order was determined manually. The second issue, slice order, could not be determined automatically from the image metadata. We therefore made maximum intensity projections (using the unu tool, http://teem.sourceforge.net/unrrdu) along the Z axis of the channel corresponding to the labelled neuron for each stack. Each projection was then compared with the matching thumbnail available from the flycircuit.tw website. The correlation score between the projection and thumbnail images was calculated both with and without a mirror flip across the YZ plane; a large correlation score for only one orientation was used as evidence for a given slice ordering. A small number of ambiguous results were verified manually. We successfully preprocessed 16,204/16,226 total images i.e. a 0.14 % failure rate. 12 failures were due to mismatches that could not be resolved between the segmented neuron present in the LSM file and the thumbnail image for the neuron identifier on the flycircuit.tw website; the remaining 10 failures were due to physical offsets between the brain and GFP channels or corrupt image data.

Template Brain

The template brain (FCWB) was constructed by screening for whole brains within the FlyCircuit dataset, and manually selecting a pool of brains that appeared of good quality when the stacks were inspected. Separate average female and average male template brains were constructed from 17 and 9 brains, respectively using the CMTK (http://www.nitrc.org/projects/cmtk) `avg_adm` function which takes a single brain as a seed. After five iterations the resultant average male and average female brains were placed in an affine symmetric position within their image stacks so that a simple horizontal (x-axis) flip of either template brain resulted in an almost perfect overlap of left and right hemispheres. Finally the two sex-specific template brains were then averaged (with equal weight) to make an intersex template brain using the same procedure. Since the purpose of this template was to provide an optimal registration target for the flycircuit.tw dataset, no attempt was made to correct for the obvious disparity between the XY and Z voxel dimensions common to all the images in the dataset. The scripts used for the construction of the template are available at https://github.com/jefferislab/MakeAverageBrain.

Image Registration

Image registration of the Dlg neuropil staining used a fully automatic intensity-based (landmark free) 3D image registration implemented in the CMTK toolkit available at http://www.nitrc.org/projects/cmtk (Rohlfing and Maurer, 2003; Jefferis et al., 2007). An initial linear registration with 9 degrees of freedom (translation, rotation and scaling of each axis) was followed by a non-rigid registration that allows different brain regions to move somewhat independently, subject to a smoothness penalty (Rueckert et al., 1999). It is our experience that obtaining a satisfactory initial linear registration is crucial. All registrations were therefore manually checked by comparing the reformatted brain with the template in Amira (academic version, Zuse Institute, Berlin), using ResultViewer https://bitbucket.org/jefferis/resultviewer. This identified about 10 % of brains which did not register satisfactorily. For these images a new affine registration was calculated using a Global Hough Transform (Ballard, 1981; Khoshelham, 2007) with an Amira extension module available from 1000shapes GmbH; the re-

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sult of this affine transform was again manually inspected. In the minority of cases where this approach failed, a surface based alignment was calculated in Amira after manually aligning the two brains. Once a satisfactory initial affine registration was obtained, a non-rigid registration was calculated for all brains. Finally each registration was checked manually in Amira against the template brain. The result of this sequential procedure was that we successfully registered 16,129/16,204 preprocessed images, giving a registration failure rate of 0.46 %.

Image Postprocessing

The confocal stack for each neuron available at `http://flycircuit.tw` includes an 8 bit image containing a single (semiautomatically) segmented neuron prepared by Chiang et al. (2011). This image was downsampled by a factor of 2 in x and y, binarized with a threshold of 1 and then skeletonized using the Fiji plugin 'Skeletonize (2D/3D)' (Doube et al., 2010). Dot properties for each neuron skeleton were extracted following the method in Masse et al. (2012), using the `dotprops` function of our new nat package for R. This converted each skeleton into segments, described by its location and tangent vector. Neurons on the right side of the brain were flipped to left by applying a mirroring and a flipping registration as described in Manton et al. (2014). The decision of whether to flip a neuron depended on earlier assignment of each neuron to a brain hemisphere using a combination of automated and manual approaches. Neurons whose cell bodies were more than 20 µm away from the mid-sagittal YZ plane were automatically defined as belonging to the left or right hemisphere. Neurons whose cell bodies were inside this 40 µm central corridor were manually assigned to the left or right sides, based on the position of the cell body (right or left side), path taken by the primary neurite, location and length of first branching neurite. For example, neurons that had a cell body on the midline with significant innervation from the first branching neurite near the cell body on the left hemisphere, with the rest of the arborisation on the right, were assigned to the left side and not flipped. On the other hand, neurons with similar morphology to these but in which the first branching neurite is small, compared to the total innervation, were assigned to the right and flipped. The cell body positions used were based on those published on the `http://flycircuit.tw` website for each neuron; these positions are in the space of the FlyCircuit female and male template brains (typical_brain_female and typical_brain_male). In order to transform them into the FCWB template that we constructed, affine bridging registrations were constructed from the typical_brain_female and typical_brain_male brains to FCWB and the cell body positions were then transformed to this new space. Since these cell body positions depend on two affine registrations (one conducted by Chiang et al. (2011) to register each sample brain onto either their typical_brain_female or typical_brain_male templates and a second carried out by us to map those template brains onto our FCWB template) these positions are likely accurate only to ±5 microns in each axis.

Neuron Search

The neuron search algorithm is described in detail in Results and Figure 1. The reference implementation that we have written is the `nblast` function in the R package nat.nblast, which depends on our `nat` package (Jefferis and Manton, 2014). Fast nearest neighbor search, an essential primitive for the algorithm uses the RANN package (Jefferis, 2014), a wrapper for the Approximate Nearest Neighbor (ANN) C++ library (Mount and Arya, 2006). The scoring matrix that we used for FlyCircuit neurons

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was constructed by taking 150 DL2 projection neurons, which define a neuron type at the finest level, and calculating the joint histogram of distance and absolute dot product for the 150 × 149 combinations of neurons, resulting in 1.4×10^7measurement pairs; the number of counts in the histogram was then normalized (i.e. dividing by 1.4×10^7) to give a probability density, p_{match}. We then carried out a similar procedure for 5,000 random pairs of neurons sampled from the FlyCircuit dataset to give p_{rand}. Finally the scoring matrix was calculated as $\log_2 \frac{p_{\mathrm{match}}+\epsilon}{p_{\mathrm{rand}}+\epsilon}$ where ε (a pseudocount to avoid infinite values) was set to 1×10^{-6}.

Clustering

We employed two different methods for clustering based on normalized NBLAST scores. We used Ward's method for hierarchical clustering, using the default implementation in the R function `hclust`. This method minimizes the total within-cluster variance, and at each step the pair of entities or clusters with the minimum distance between clusters are merged (Ward Jr, 1963). The resulting dendrograms were cut at a single selected height chosen for each case to separate neuron types or subtypes. This value is shown as a dashed line in all dendrograms. By default, R plots the square of the Euclidean distance as the y axis, but in the plots shown, the height of the dendrogram corresponds to the unsquared distance.

For the analysis of the whole dataset, we used the affinity propagation method. This is an iterative method which finds exemplars which are representative members of each cluster and does not require any *a priori* input on the final number of clusters (Frey and Dueck, 2007) as implemented in the R package apcluster (Bodenhofer et al., 2011). The input preference parameter (p) can be set before running the clustering. This parameter reflects the tendency of data samples to become an exemplar, and affects the final number of clusters. In our analysis, we used $p = 0$, since this is the value where on average matched segments are equally likely to have come from matching and non-matching neurons. Empirically this parameter produced clusters that, for the most part, grouped neurons of the same type according to biological expert opinion.

Neuron Tracing

Neuron tracing was carried out in Amira (commercial version, FEI Visualization Sciences Group, Merignac, France) using the hxskeletonize plugin (Evers et al., 2005) or in Vaa3D (Peng et al., 2014) on previously registered image data. Traces were then loaded into R using the `nat` package. When necessary, they were transformed into the space of the FCWB template brain using the approach of Manton et al. (2014).

Computer Code and Data

The image processing pipeline and analysis code used two custom packages for the R statistical environment (`http://www.r-project.org`) `https://github.com/jefferis/nat` and `https://github.com/jefferis/nat.as` that coordinated processing by the low level registration (CMTK) and image processing (Fiji, unu) software mentioned above. NBLAST neuron search is implemented in a third R package available at `https://github.com/jefferislab/nat.nblast`. Analysis code specific to the flycircuit dataset is available in a dedicated R package `https://github.com/jefferis/`

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

flycircuit, with a package vignette showcasing the main tools that we have developed. Further details of these supplemental software and the associated data are presented at http://jefferislab.org/si/nblast. The registered image dataset can be viewed in the stack viewer of the http://virtualflybrain.org website and all 16,129 registered single neuron images will be available at https://jefferislab.org/si/nblast or on request to GSXEJ on a hard drive; the unregistered data remain available at http://flycircuit.tw.

Acknowledgments

We first of all acknowledge the flycircuit.tw team for generously providing the raw image data associated with Chiang et al. (2011). Images from FlyCircuit were obtained from the NCHC (National Center for High-performance Computing) and NTHU (National Tsing Hua University), Hsinchu, Taiwan. We thank members of the Jefferis lab for comments on the manuscript, Jake Grimmett and Toby Darling for assistance with the LMB's compute cluster and Torsten Rohlfing for discussions about image analysis and registration. We thank the Virtual Fly Brain project for their help in linking and incorporating some of the results of this study in the http://virtualflybrain.org website.

This study made use of the Computational Morphometry Toolkit, supported by the National Institute of Biomedical Imaging and Bioengineering. This work was supported by the Medical Research Council [MRC file reference U105188491] and a European Research Council Starting Investigator Grant to GSXEJ, who is an EMBO Young Investigator.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410.

Aso, Y., Grübel, K., Busch, S., Friedrich, A.B., Siwanowicz, I., and Tanimoto, H. (2009). The mushroom body of adult Drosophila characterized by GAL4 drivers. Journal of neurogenetics *23*, 156–172.

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

Badea, T.C., and Nathans, J. (2004). Quantitative analysis of neuronal morphologies in the mouse retina visualized by using a genetically directed reporter. Journal of Comparative Neurology *480*, 331–351.

Ballard, D.H. (1981). Generalizing the Hough transform to detect arbitrary shapes. Pattern Recognition *13*, 111–122.

Basu, S., Condron, B., and Acton, S.T. (2011). Path2Path: hierarchical Path-Based analysis for neuron matching. In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on (IEEE), pp. 996–999.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. Bioinformatics *27*, 2463–2464.

Bota, M., and Swanson, L.W. (2007). The neuron classification problem. Brain research reviews *56*, 79–88.

Brown, K.M., Barrionuevo, G., Canty, A.J., De Paola, V., Hirsch, J.A., Jefferis, G.S.X.E., Lu, J., Snippe, M., Sugihara, I., and Ascoli, G.A. (2011). The DIADEM Data Sets: Representative Light Microscopy Images of Neuronal Morphology to Advance Automation of Digital Reconstructions. Neuroinformatics .

Cachero, S., Ostrovsky, A.D., Yu, J.Y., Dickson, B.J., and Jefferis, G.S.X.E. (2010). Sexual dimorphism in the fly brain. Curr Biol *20*, 1589–601.

Cajal, S.R., and Azoulay y, L. (1911). Histologie du système nerveux de l'homme et des vertébrés (A. Maloine).

Cardona, A., Saalfeld, S., Arganda, I., Pereanu, W., Schindelin, J., and Hartenstein, V. (2010). Identifying neuronal lineages of Drosophila by sequence analysis of axon tracts. J Neurosci *30*, 7538–53.

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., Wu, C.C., Chen, G.Y., Ching, Y.T., Lee, P.C., Lin, C.Y., Lin, H.H., Wu, C.C., Hsu, H.W., Huang, Y.A., Chen, J.Y., Chiang, H.J., Lu, C.F., Ni, R.F., Yeh, C.Y., and Hwang, J.K. (2011). Three-dimensional reconstruction of brain-wide wiring networks in Drosophila at single-cell resolution. Curr Biol *21*, 1–11.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Conte, D., Foggia, P., Sansone, C., and Vento, M. (2004). Thirty years of graph matching in pattern recognition. International journal of pattern recognition and artificial intelligence *18*, 265–298.

Coombs, J., Van Der List, D., Wang, G.Y., and Chalupa, L. (2006). Morphological properties of mouse retinal ganglion cells. Neuroscience *140*, 123–136.

Doube, M., Kłosowski, M.M., Arganda-Carreras, I., Cordelières, F.P., Dougherty, R.P., Jackson, J.S., Schmid, B., Hutchinson, J.R., and Shefelbine, S.J. (2010). BoneJ: Free and extensible bone image analysis in ImageJ. Bone *47*, 1076 – 1079.

El Jundi, B., Heinze, S., Lenschow, C., Kurylas, A., Rohlfing, T., and Homberg, U. (2009). The locust standard brain: a 3D standard of the central complex as a platform for neural network analysis. Frontiers in systems neuroscience *3*.

Evers, J.F., Schmitt, S., Sibila, M., and Duch, C. (2005). Progress in functional neuroanatomy: precise automatic geometric reconstruction of neuronal morphology from confocal image stacks. J Neurophysiol *93*, 2331–42.

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

Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. science *315*, 972–976.

Ganglberger, F., Schulze, F., Tirian, L., Novikov, A., Dickson, B., Bühler, K., and Langs, G. (2014). Structure-Based Neuron Retrieval Across Drosophila Brains. Neuroinformatics .

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

Jefferis, G. (2014). RANN k nearest neighbour search v2.3.0. Zenodo .

Jefferis, G.S.X.E., and Manton, J.D. (2014). nat: NeuroAnatomy Toolbox R package. Zenodo .

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

Jefferis, G.S.X.E., and Livet, J. (2012). Sparse and combinatorial neuron labelling. Curr Opin Neurobiol *22*, 101–10.

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

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 reports *2*, 991–1001.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kahsai, L., and Zars, T. (2011). Learning and memory in Drosophila: behavior, genetics, and neural systems. Int Rev Neurobiol *99*, 139–67.

Kamikouchi, A., Inagaki, H.K., Effertz, T., Hendrich, O., Fiala, A., Göpfert, M.C., and Ito, K. (2009). The neural basis of Drosophila gravity-sensing and hearing. Nature *458*, 165–171.

Kamikouchi, A., Shimada, T., and Ito, K. (2006). Comprehensive classification of the auditory sensory projections in the brain of the fruit fly Drosophila melanogaster. Journal of Comparative Neurology *499*, 317–356.

Kepecs, A., and Fishell, G. (2014). Interneuron cell types are fit to function. Nature *505*, 318–26.

Khoshelham, K. (2007). Extending Generalized Hough Transform to Detect 3D Objects in Laser Range Data. In ISPRS Workshop on Laser Scanning, Proceedings, LS 2007. pp. 206–210.

Kimura, K.i., Hachiya, T., Koganezawa, M., Tazawa, T., and Yamamoto, D. (2008). Fruitless and doublesex coordinate to generate male-specific neurons that can initiate courtship. Neuron *59*, 759–769.

Kimura, K.I., Ote, M., Tazawa, T., and Yamamoto, D. (2005). Fruitless specifies sexually dimorphic neural circuitry in the Drosophila brain. Nature *438*, 229–233.

Koganezawa, M., Haba, D., Matsuo, T., and Yamamoto, D. (2010). The Shaping of Male Courtship Posture by Lateralized Gustatory Inputs to Male-Specific Interneurons. Current Biology *20*, 1–8.

Kong, J.H., Fish, D.R., Rockhill, R.L., and Masland, R.H. (2005). Diversity of ganglion cells in the mouse retina: unsupervised morphological classification and its limits. Journal of Comparative Neurology *489*, 293–310.

Lai, J.S.Y., Lo, S.J., Dickson, B.J., and Chiang, A.S. (2012). Auditory circuit in the Drosophila brain. Proc Natl Acad Sci U S A *109*, 2607–12.

Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., and Ng, A. (2012). Building high-level features using large scale unsupervised learning. In International Conference in Machine Learning.

Lee, T.C., Kashyap, R.L., and Chu, C.N. (1994). Building Skeleton Models via 3-D Medial Surface/Axis Thinning Algorithms. CVGIP: Graph. Models Image Process. *56*, 462–478.

Lee, T., Lee, A., and Luo, L. (1999). Development of the Drosophila mushroom bodies: sequential generation of three distinct types of neurons from a neuroblast. Development *126*, 4065–4076.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

Lin, H.H., Lai, J.S.Y., Chin, A.L., Chen, Y.C., and Chiang, A.S. (2007). A map of olfactory representation in the Drosophila mushroom body. Cell *128*, 1205–1217.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Manton, J.D., Ostrovsky, A.D., Goetz, L., Costa, M., Rohlfing, T., and Jefferis, G.S.X.E. (2014). Combining genome-scale Drosophila 3D neuroanatomical data by bridging template brains. Bioarxiv preprint .

Marin, E.C., Jefferis, G.S., Komiyama, T., Zhu, H., and Luo, L. (2002). Representation of the Glomerular Olfactory Map in the Drosophila Brain. Cell *109*, 243–255.

Masse, N.Y., Turner, G.C., and Jefferis, G.S.X.E. (2009). Olfactory information processing in Drosophila. Curr Biol *19*, R700–13.

Masse, N.Y., Cachero, S., Ostrovsky, A., and Jefferis, G.S.X.E. (2012). A mutual information approach to automate identification of neuronal clusters in Drosophila brain images. Frontiers in Neuroinformatics *6*.

Mayerich, D., Bjornsson, C., Taylor, J., and Roysam, B. (2012). NetMets: software for quantifying and visualizing errors in biological network segmentation. BMC Bioinformatics *13 Suppl 8*, S7.

Migliore, M., and Shepherd, G.M. (2005). An integrated approach to classifying neuronal phenotypes. Nature Reviews Neuroscience *6*, 810–818.

Miyasaka, N., Arganda-Carreras, I., Wakisaka, N., Masuda, M., Sümbül, U., Seung, H.S., and Yoshihara, Y. (2014). Olfactory projectome in the zebrafish forebrain revealed by genetic single-neuron labelling. Nat Commun *5*, 3639.

Morante, J., and Desplan, C. (2008). The Color-Vision Circuit in the Medulla of Drosophila. Current Biology *18*, 553–565.

Mount, D.M., and Arya, S. (2006). ANN: A Library for Approximate Nearest Neighbor Searching. Version 1.1.1.

Nelson, S.B., Sugino, K., and Hempel, C.M. (2006). The problem of neuronal cell types: a physiological genomics approach. Trends Neurosci *29*, 339–45.

Oh, S.W., Harris, J.A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M., et al. (2014). A mesoscale connectome of the mouse brain. Nature *508*, 207–214.

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.

Parekh, R., and Ascoli, G.A. (2013). Neuronal morphology goes digital: a research hub for cellular and system neuroscience. Neuron *77*, 1017–38.

Paulk, A., Millard, S.S., and van Swinderen, B. (2013). Vision in Drosophila: seeing the world through a model's eyes. Annual review of entomology *58*, 313–332.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A *85*, 2444–8.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Peng, H., Tang, J., Xiao, H., Bria, A., Zhou, J., Butler, V., Zhou, Z., Gonzalez-Bellido, P.T., Oh, S.W., Chen, J., Mitra, A., Tsien, R.W., Zeng, H., Ascoli, G.A., Iannello, G., Hawrylycz, M., Myers, E., and Long, F. (2014). Virtual finger boosts three-dimensional imaging and microsurgery as well as terabyte volume image visualization and analysis. Nat Commun *5*, 4342.

Petilla Interneuron Nomenclature Group, Ascoli, G.A., Alonso-Nanclares, L., Anderson, S.A., Barrionuevo, G., Benavides-Piccione, R., Burkhalter, A., Buzsáki, G., Cauli, B., Defelipe, J., Fairén, A., Feldmeyer, D., Fishell, G., Fregnac, Y., Freund, T.F., Gardner, D., Gardner, E.P., Goldberg, J.H., Helmstaedter, M., Hestrin, S., Karube, F., Kisvárday, Z.F., Lambolez, B., Lewis, D.A., Marin, O., Markram, H., Muñoz, A., Packer, A., Petersen, C.C.H., Rockland, K.S., Rossier, J., Rudy, B., Somogyi, P., Staiger, J.F., Tamas, G., Thomson, A.M., Toledo-Rodriguez, M., Wang, Y., West, D.C., and Yuste, R. (2008). Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat Rev Neurosci *9*, 557–68.

Portugues, R., Feierstein, C.E., Engert, F., and Orger, M.B. (2014). Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. Neuron *81*, 1328–1343.

Rohlfing, T., and Maurer, C. R., J. (2003). Nonrigid image registration in shared-memory multiprocessor environments with application to brains, breasts, and bees. IEEE Trans Inf Technol Biomed *7*, 16–25.

Rowe, M., and Stone, J. (1976). Naming of neurones. Classification and naming of cat retinal ganglion cells. Brain, behavior and evolution *14*, 185–216.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., and Hawkes, D.J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging *18*, 712–21.

Rybak, J., Kuß, A., Lamecker, H., Zachow, S., Hege, H.C., Lienhard, M., Singer, J., Neubert, K., and Menzel, R. (2010). The digital bee brain: integrating and managing neurons in a common 3D reference system. Frontiers in systems neuroscience *4*.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. Nat Methods *9*, 676–82.

Sonnhammer, E.L., Eddy, S.R., Durbin, R., et al. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins-Structure Function and Genetics *28*, 405–420.

Sümbül, U., Song, S., McCulloch, K., Becker, M., Lin, B., Sanes, J.R., Masland, R.H., and Seung, H.S. (2014). A genetic and computational approach to structurally classify neuronal types. Nat Commun *5*, 3512.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic acids research *41*, D996–D1008.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Susaki, E.A., Tainaka, K., Perrin, D., Kishino, F., Tawara, T., Watanabe, T.M., Yokoyama, C., Onoe, H., Eguchi, M., Yamaguchi, S., et al. (2014). Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis. Cell *157*, 726–739.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288.

Tanaka, N.K., Endo, K., and Ito, K. (2012). Organization of antennal lobe-associated neurons in adult Drosophila melanogaster brain. Journal of Comparative Neurology *520*, 4067–4130.

Tanaka, N.K., Tanimoto, H., and Ito, K. (2008). Neuronal assemblies of the Drosophila mushroom body. Journal of Comparative Neurology *508*, 711–755.

Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association *58*, 236–244.

Wong, A.M., Wang, J.W., and Axel, R. (2002). Spatial representation of the glomerular map in the Drosophila protocerebrum. Cell *109*, 229–41.

Yorozu, S., Wong, A., Fischer, B.J., Dankert, H., Kernan, M.J., Kamikouchi, A., Ito, K., and Anderson, D.J. (2009). Distinct sensory representations of wind and near-field sound in the Drosophila brain. Nature *458*, 201–205.

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

Yu, J.Y., Kanai, M.I., Demir, E., Jefferis, G.S.X.E., and Dickson, B.J. (2010b). Cellular organization of the neural circuit that drives Drosophila courtship behavior. Curr Biol *20*, 1602–14.

Zhu, S., Chiang, A.S., and Lee, T. (2003). Development of the Drosophila mushroom bodies: elaboration, remodeling and spatial organization of dendrites in the calyx. Development *130*, 2603–2610.

Zingg, B., Hintiryan, H., Gou, L., Song, M.Y., Bay, M., Bienkowski, M.S., Foster, N.N., Yamashita, S., Bowman, I., Toga, A.W., et al. (2014). Neural networks of the mouse neocortex. Cell *156*, 1096–1111.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

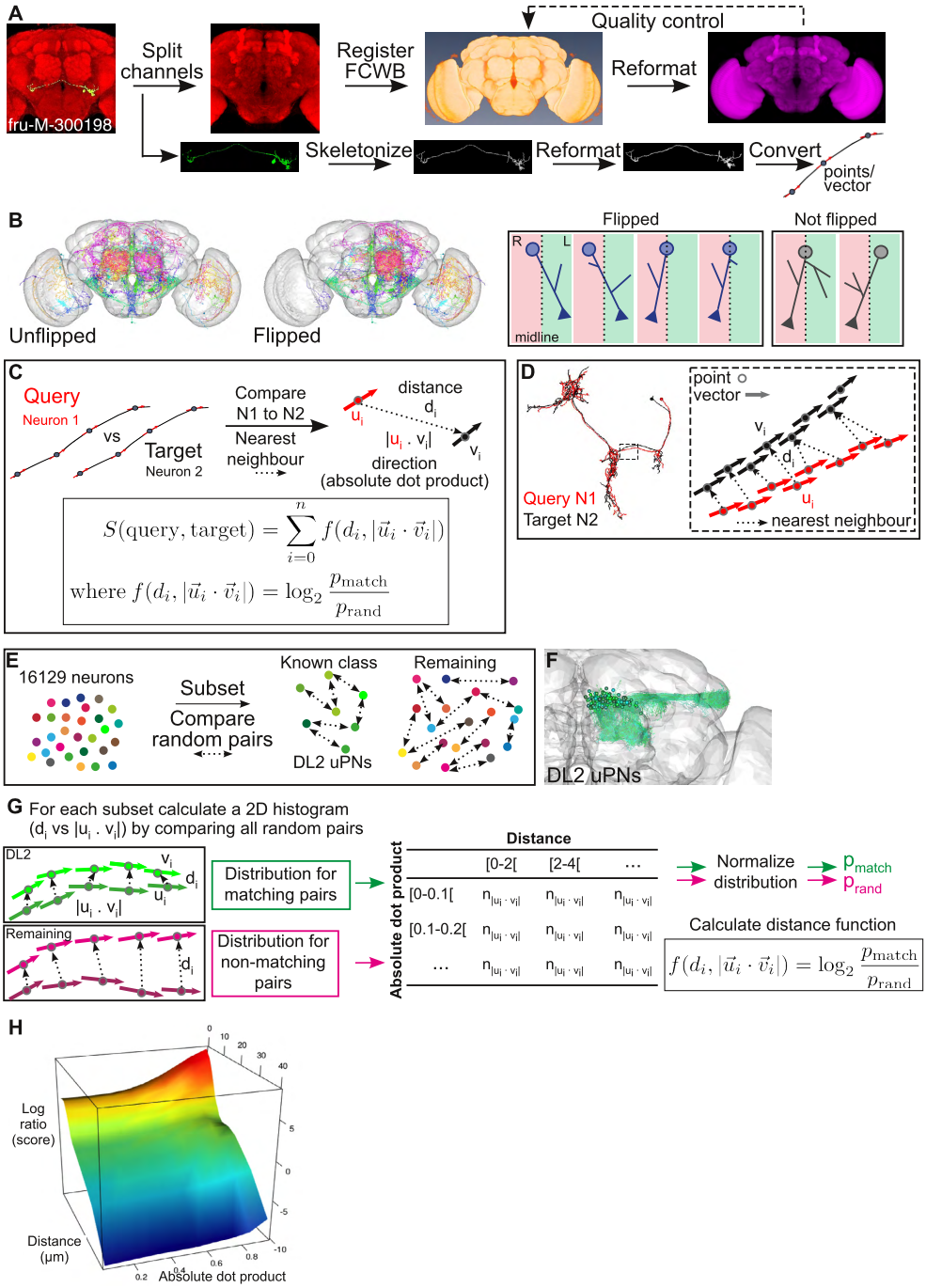


Figure 1: **Image preprocessing, registration and similarity score (NBLAST) algorithm**

(A) Flowchart describing the image preprocessing and registration procedure. FlyCircuit images were split into 3 channels. The Dlg-stained brain (discs large 1) (channel 1) images were registered against the FCWB template. Registration success was assessed by comparing the template brain with the brain images after applying the registration (reformatted). The neuron image in channel 3 was skeletonized and reformatted onto the template brain. The neuron skeleton was converted into points and vectors. **(B)** After image registration, neurons on the right side of the brain were flipped onto the left side. For most neurons this was done automatically. On the left, brain plot showing 50 random neurons before and after flipping. On the right, cases for which the neuron flipping was assessed manually. These included cases in which the cell body was on or very close to the midline, with or without small primary ipsilateral neurites. **(C)** NBLAST algorithm. The similarity of two neurons (query and target), is given by a function of the distance and absolute dot product between the nearest neighbor points of the query/target pair. This distance function reflects the probability of a match between a pair of points ($p_{\mathbf{match}}$), relative to any two random points ($p_{\mathbf{rand}}$). **(D)** Diagram illustrating how nearest neighbor points are calculated. For a query (N1)/target (N2) pair, each point of N1 (u_i) is paired to the N2 point (v_i) that minimizes the distance (d_i) between the points. **(E)** Calculating the distance function. Two groups of neurons were used to calculate the distribution probabilities of matching and non-matching pairs. The first corresponds to a known class of uniglomerular olfactory projection neurons (uPNs), DL2 uPNs that had been previously identified in the dataset. The second group corresponds to all remaining neurons. Random pairs of neurons were compared within each group. **(F)** Brain plot showing all DL2 neurons in the dataset that were used for this analysis. **(G)** Calculation of the distribution for matching and non-matching pairs of segments. For all segment pairs of all neuron pairs of each group, the distance and absolute dot product were plotted in a distance histogram. The distribution probability for matching ($p_{\mathbf{match}}$) or non-matching pairs ($p_{\mathbf{rand}}$) was calculated by normalizing the distance histogram to 1. When calculating the distance function, 1×10^{-6} was added to both $p_{\mathbf{match}}$ and $p_{\mathbf{rand}}$ to avoid a 0 denominator. **(H)** Plot showing that the similarity score depends on the spatial location of the points (distance between points) and the direction of the vectors (absolute dot product). The score is the highest for a distance of 0 µm and an absolute dot product of 1.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

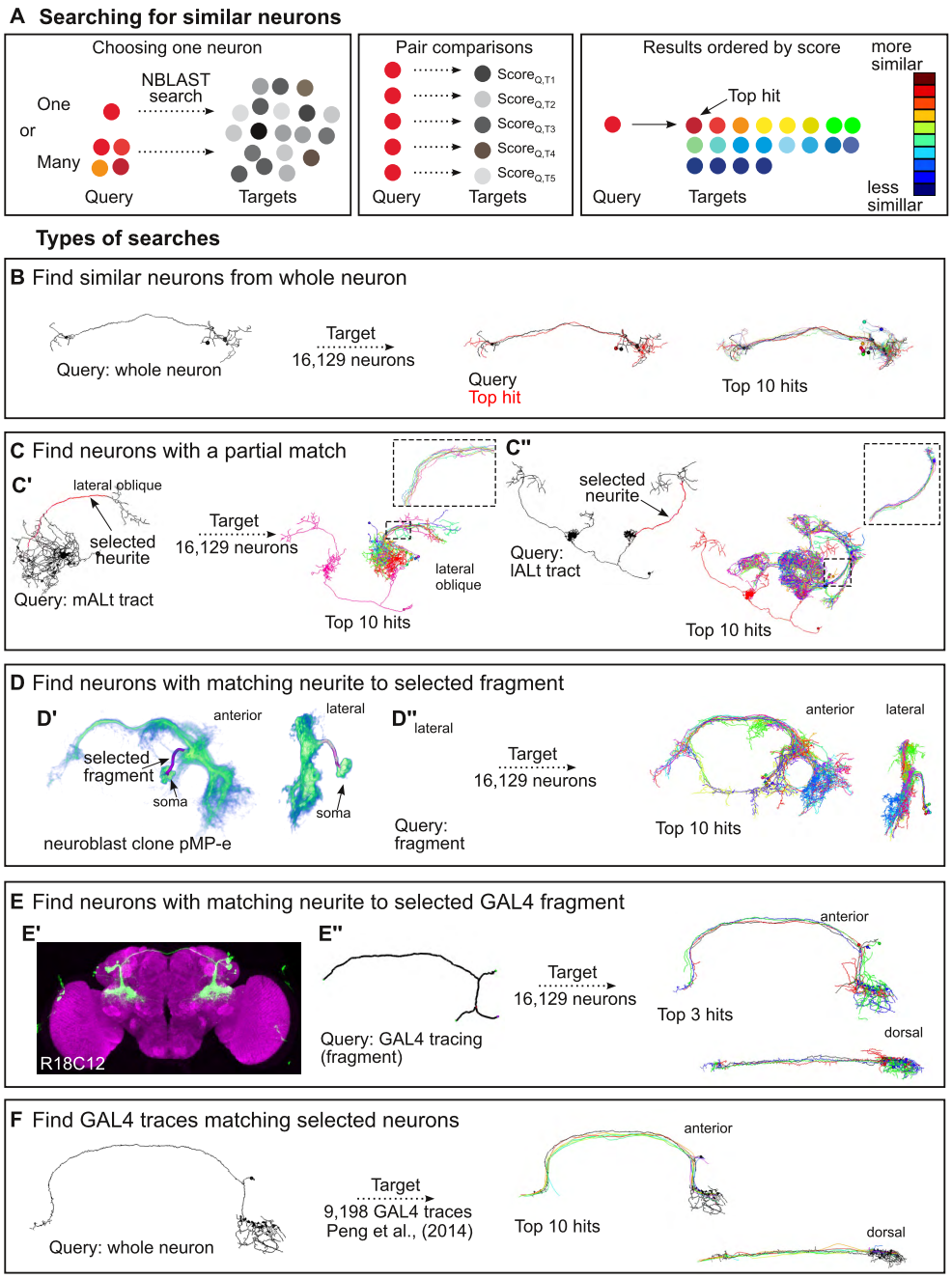


Figure 2: **NBLAST allows different types of searches**
(A) Searching for similar neurons with NBLAST. Pair comparisons between the chosen query neuron and remaining neurons in the dataset return a similarity score, allowing the results to be ordered by similarity. **(B)** NBLAST search for matching neurons using a whole neuron as a query against the FlyCircuit dataset (16,129 neurons). The query neuron, top hit and top 10 hits in anterior view are shown. A forward search is shown (returns the similarity score for the query compared to the target) but a reverse search (returns the score of the target against the query) could also be used. Query neuron: fru-M-400121. **(C)** NBLAST search for matching neurons using a neuron fragment as a query against the FlyCircuit dataset (16,129 neurons). The query neuron and top 10 hits are shown. **(C')** Search with the mALT tract from an olfactory projection neuron (Cha-F-000239). Lateral oblique view is shown. The mALT tract of the top 10 hits is shown as an inset. **(C")** Search with the lALT tract from an olfactory projection neuron (Gad1-F-200095). Anterior view is shown. The lALT tract of the top 10 hits is shown as an inset. **(D)** NBLAST search for neurons with matching neurites to a fragment from a *fruitless* neuroblast clone (pMP-e). The target is the FlyCircuit dataset (16,129 neurons). **(D')** Volume rendering of pMP-e clone with the selected fragment – the characteristic stalk of P1 neurons. An anterior and lateral view are shown. **(D")** Query fragment in lateral view. Top 10 hits in anterior and lateral view. **(E)** NBLAST search for neurons with matching neurites to a fragment traced from a GAL4 image (R18C12) ((Jenett et al., 2012)). The target is the FlyCircuit dataset (16,129 neurons). **(E')** Maximum Z projection of FlyLight line R18C12 registered to JFRC2. The image was downloaded from the Virtual Fly Brain website. **(E")** The fragment used as query (traced in Vaa3D) in anterior view. Top 3 hits in anterior and dorsal view. **(F)** NBLAST search for GAL4 traces (Peng et al., 2014) matching a selected FlyCircuit neuron. Query neuron: VGlut-F-500818. The top 10 trace hits are shown in anterior and dorsal view. Maximum Z projection of FlyLight line R18C12 registered to JFRC2. The image was downloaded from the Virtual Fly Brain website.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

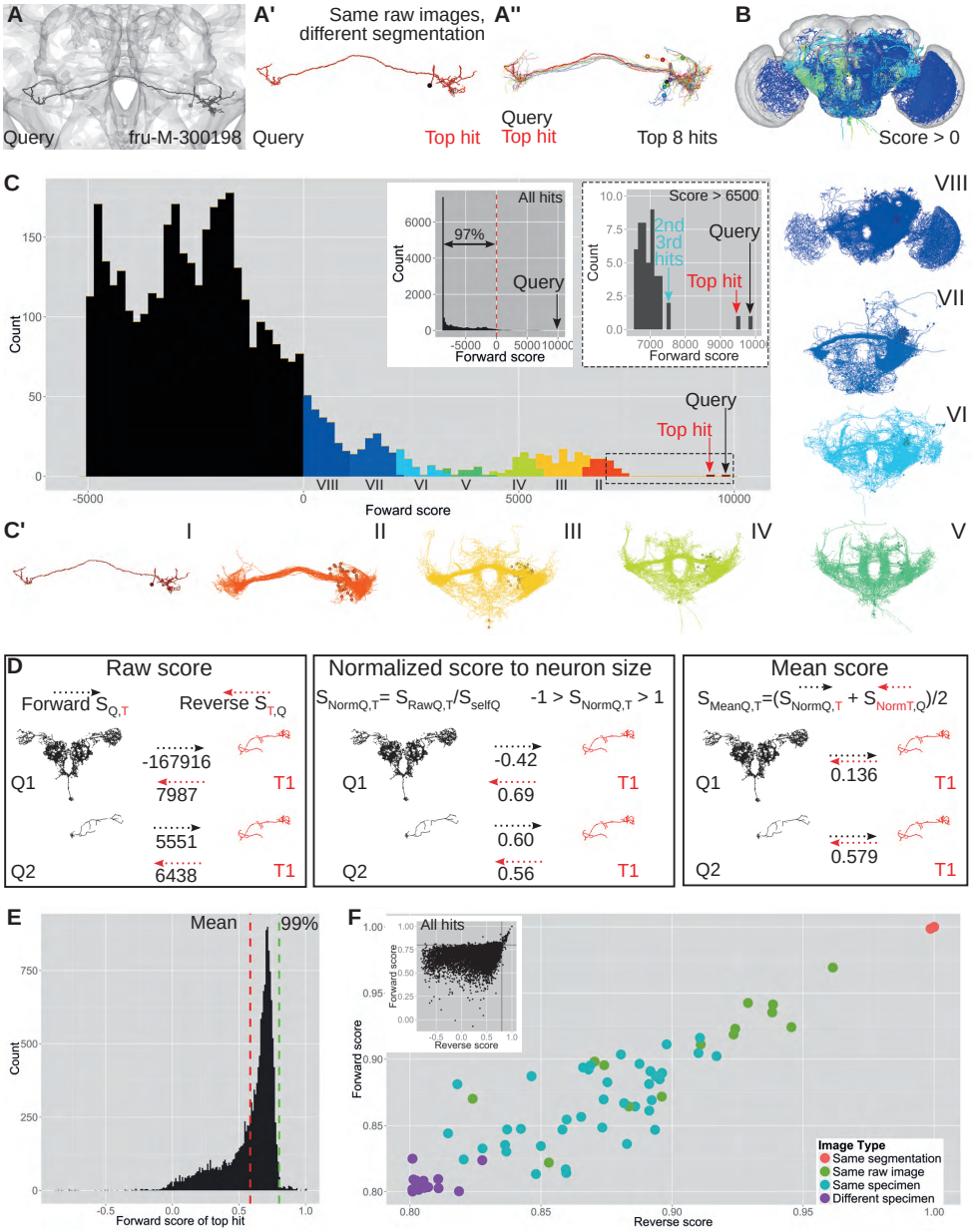


Figure 3: **NBLAST scores are accurate and meaningful**

(A) NBLAST search with fru-M-300198 as query (black). Neuron plot of the query neuron. **(A')** Neuron plot of the query (black) and top hit (red). The top hit corresponds to a different segmentation of the query neuron, from the same raw image. The differences between these two images is due to minor differences during the segmentation. **(A")** Neuron plot showing the top 8 hits. There are differences in neurite branching, length and position. **(B)** All hits with a forward score over 0, colored by score, as shown in C. **(C)** Histogram of scores for a forward search with fru-M-300198 as query. Only hits with scores over −5,000 are shown. The left inset shows the histogram of scores for all search hits. The right inset shows a zoomed view of the top hits (score > 6,500). For more examples see **S1**. **(C')** Neuron plots corresponding to the score bins in C. **(D)** Comparison of the raw, normalized and mean score, for two pairs of neurons: one of unequal (Q1, T1) and one of similar size (Q2, T1). The value of the raw score depends on the size of the neuron, whereas the normalized score corrects for it, by dividing the raw score by the query self-score (maximum score). Normalized scores are between −1 and 1. Mean scores are the average between the normalized forward and reverse score for a pair of neurons. These scores can be compared for different searches. **(E)** Histogram of the normalized score for the top hit for each neuron in the whole dataset. The mean and 99th percentile are shown as a dashed red and green lines, respectively. **(F)** Plot of reverse and forward normalized scores for 72 pairs of neurons for which both the forward and reverse scores are higher than 0.8. These pairs were classified into four categories, according to the relationship between the two images: images correspond to a segmented image that is duplicated ('Same segmentation'); images correspond to different neuron segmentations from the same raw image ('Same raw image'); images correspond to two different segmented images from the same brain ('Same specimen'); images correspond to segmented images of the same or similar neurons in different brains ('Different specimen'). The inset plot shows the normalized reverse and forward scores for all top hits. The threshold of 0.8 is indicated by two black lines.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

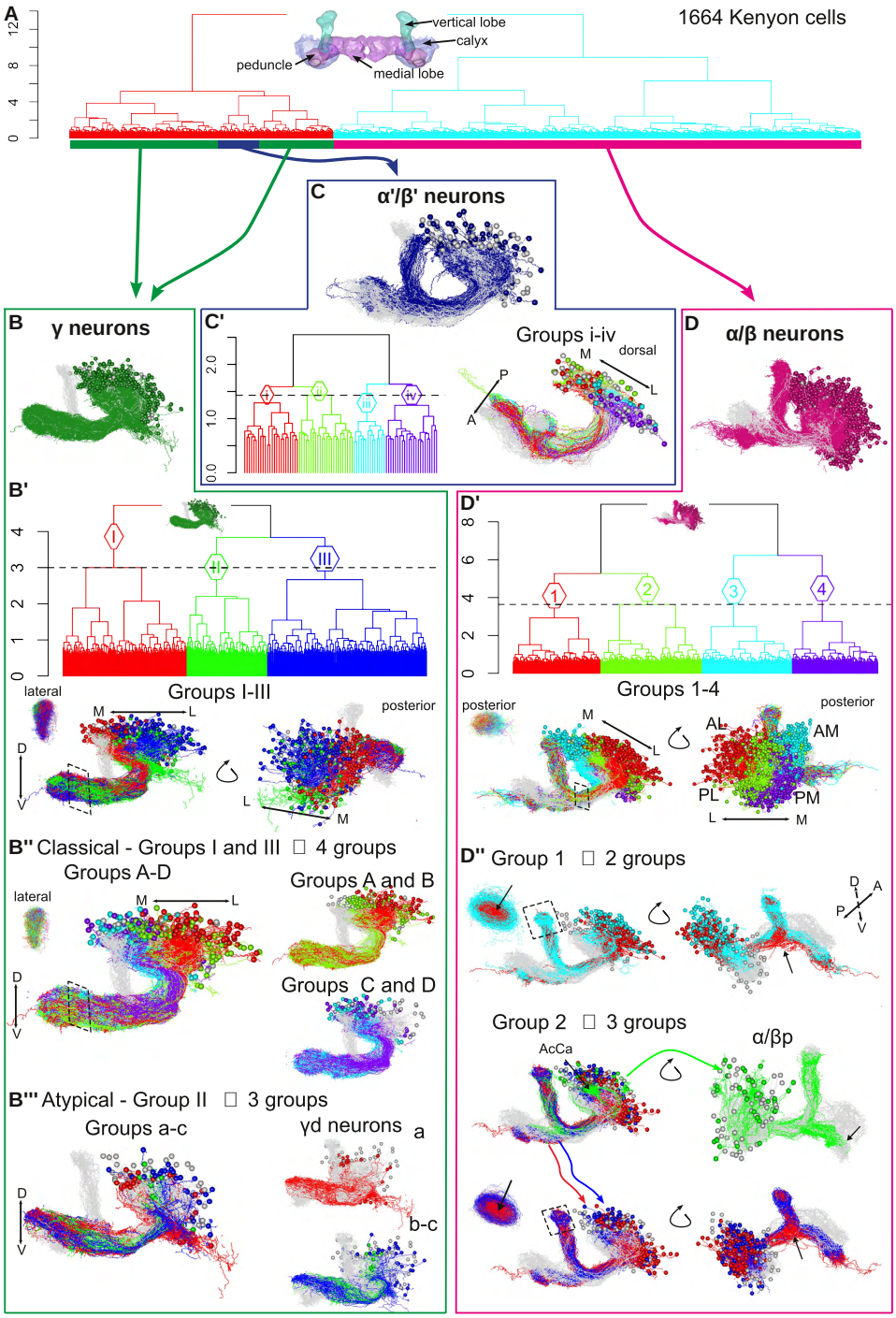
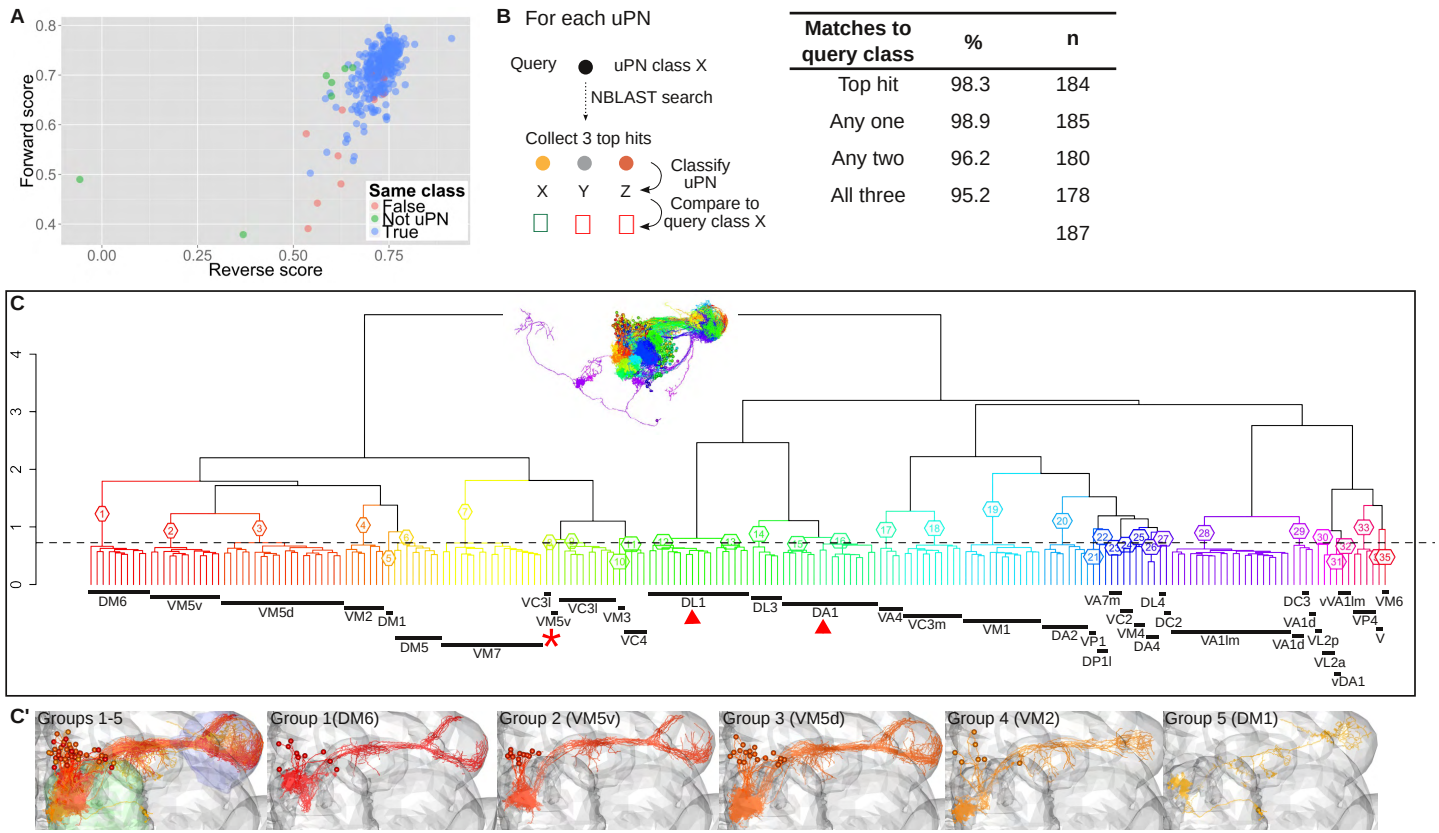


Figure 4: **NBLAST search and classification of hits reveals Kenyon cell subtypes** **(A)** Hierarchical clustering (HC) of Kenyon cells (n=1664), divided into two groups. Bars below the dendrogram indicate the neurons corresponding to a specific neuron type: γ (in green), α′/β′ (in blue) and α/β neurons (in magenta), h=8.9. Inset shows the mushroom body neuropil. **(B)** Neuron plot of the γ neurons. **(B')** HC of the γ neurons divided into three groups (I–III), h=3. Inset on the dendrogram shows the γ neurons (same as in B). Neuron plots of groups I to III. A lateral oblique and a posterior view of the neurons are shown. There are differences between the 3 groups in the calyx in the medial/lateral axis and in the dorsal/ventral axis in the γ lobe: the more medial group 1 is the most dorsal in the γ lobe. **(B")** HC of the classic γ neurons, corresponding to groups I and III in B', divided into four groups (A–D). Neuron plots of groups A–D, A–B and C–D. There are differences between the 4 groups in the calyx in the medial/lateral axis and in the dorsal/ventral axis in the γ lobe. **(B"')** HC of the atypical γ neurons corresponding to group II in B', divided into three groups (a–c). Neuron plots of groups a–c, a, and b–c. Group a corresponds to subtype γd neurons which innervate the dorsal most region of the gamma lobe and extend dendrites laterally. **(C)** Neuron plot of the α′/β′ neurons. **(C')** HC of the α′/β′ neurons, divided into four groups (i–iv), h=1.43. The groups i and iv take a more anterior route in the peduncle and β′ lobe than groups ii and iii. Dorsolateral view is shown. **(D)** Neuron plot of the α/β neurons. **(D')** HC of the α/β neurons, divided into four groups (1–4), h=3.64. Inset on the dendrogram shows the α/β neurons (same as in D). Neuron plots of groups A to D. Lateral oblique, posterior view and posterior view of a peduncle slice of these groups are shown. There are differences between the 4 groups in the calyx and in the medial/lateral axis, with each group corresponding to the indicated neuroblast clone (AM, AL, PM, PL). **(D")** HC of groups 1 and 2. Lateral oblique, posterior oblique and a dorsal view of a peduncle slice views are shown. HC of group 1 divided into 2 subgroups. This separated the neurons into peripheral (cyan) and core (red) in the α lobe. Peripheral neurons occupied a more lateral calyx position and were dorsal to core neurons in the peduncle and β lobe. Similar analysis to groups 3 and 4 is shown in Figure S3A. HC of group 2 divided into 3 subgroups. The red and blue subgroups match the core and peripheral neurons, respectively; the green subgroup the α/β posterior subtype (α/βp). These neurons innervate the accessory calyx and their axons terminate before reaching the most medial region of the β lobe. AcCa: accessory calyx. Neurons in grey: Kenyon cell exemplars.



Matches to query class	%	n
Top hit	98.3	184
Any one	98.9	185
Any two	96.2	180
All three	95.2	178
		187

Figure 5: **NBLAST search and classification of hits reveals uniglomerular olfactory projection neuronal types**
(A) Plot of the reverse and forward normalized scores for the top hit in an NBLAST search using the uniglomerular olfactory projection neurons (uPNs) as queries. Only uPN types for which we have more than one example and unique query/target pairs are included in this analysis (n=327). For each query neuron, we identified cases for which both the top hit and query were of the same class (True) (n=319); the top hit is a uPN but does not match the class of the query (False) (n=4), or the top hit is not a uPN (Not uPN) (n=4). **(B)** Percentage of neuron type matches in the top hit and top 3 hits for each uPN. The top three hits for each uPN (normalized score) were collected and the neuron type of each hit and query was compared. Only non-DL2 uPNs for which we had more than three neurons examples or a type were used (n=187). **(C)** Hierarchical clustering of uPNs (non-DL2s) (n=214) divided into 35 groups (1–35), h=0.725. Dendrogram showing the glomerulus for each neuron. The neuron plot inset shows the uPNs colored by dendrogram group. Below the leaves, the number of neurons that innervate each glomerulus is indicated by the black rectangles. Neurons that innervate DA1 and VA1lm glomeruli but originate from the ventral lineage instead of the lateral or anterodorsal, respectively, are indicated as vVA1lm and vDA1. The dendrogram groups correspond to single and unique neuron types except for DL1 and DA1 neurons which are split into 2 groups (12–13, 15–16, respectively) (red arrowhead) and the outlier neuron VM5v in group 9 (red asterisk). **(C')** Neuron plots corresponding to dendrogram groups 1–5 and to each of these individual groups, colored by dendrogram group. The antennal lobe is in green, the lateral horn in purple.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

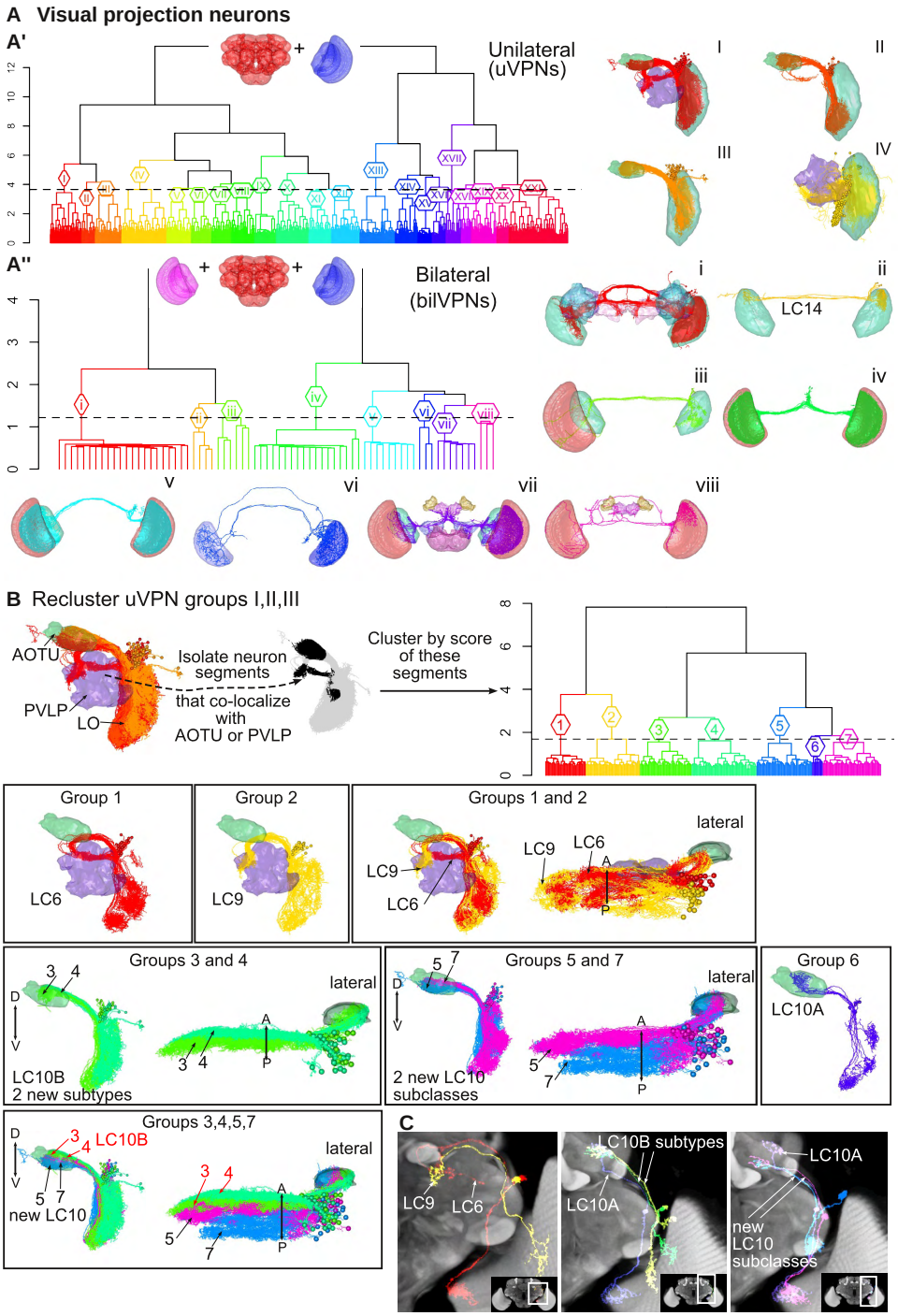


Figure 6: **NBLAST search and classification of hits uncovers visual projection neuronal types**

(A) Clustering analysis of unilateral (uVPNs) and bilateral visual projection neurons (bilVPNs), defined as neurons with segments that overlap one or two optic lobes, respectively, and some central brain neuropil. **(A')** Hierarchical clustering (HC) of unilateral visual projection neurons (uVPNs), divided into 21 groups (I–XXI), h=3.65. Inset on the dendrogram shows the neuropils considered for the overlap. To the right, neuron plots of groups I to III. The neuropils that contain the most overlap are shown. Other neuron plots are shown in Figure S4. **(A")** HC of bilVPNs, divided into 8 groups (i–viii), h=1.22. Inset on the dendrogram shows the neuropils considered for the overlap. To the right and below, neuron plots of dendrogram groups. Group ii corresponds to the LC14 neuron type that connects the 2 lobulas, with one outlier terminating in the medulla. The neuropils that contain the most overlap are shown. **(B)** Reclustering of uVPN groups I, II and III from A'. These neurons have dorsal cell bodies, arborize in the lobula (LO) and project to anterior optic tubercle (AOTU) via the anterior optic tract or to the posterior ventrolateral protocerebrum (PVLP). The neuron segments that co-localize with either the AOTU or PVLP were isolated, followed by HC of the neurons based on the NBLAST score of these neuron segments. The dendrogram was divided into seven groups (1–7), h=1.69. Neuron plots corresponding to the dendrogram groups. An anterior and a lateral view are shown. Some of dendrogram groups were matched to known uVPN types. Group 1 corresponds to LC6 neurons, group 2 to LC9. These 2 groups innervate the PVLP, and show some differences in the lobula lamination in the anterior/posterior axis. Groups 3 and 4 seem to correspond to two new subtypes of LC10B, that innervate the dorsal AOTU. They show a clear distinction in AOTU and lobula lamination, with group 4 being the dorsalmost in the AOTU and the most anterior in the lobula. Groups 5 and 7 are possible new subclasses of LC10, that innervate the ventral AOTU. They show a clear distinction in AOTU and lobula lamination, with group 7 being the dorsalmost in the AOTU (but ventral to group 3) and the most anterior in the lobula (but posterior to group 3). Group 6 corresponds to LC10A neurons that project through the ventral AOTU and turn sharply dorsally in the middle region. **(C)** Overlay of Z projections of registered image stacks of example neurons from the types identified in B on a partial Z projection of the template brain (a different one for each panel). The white rectangle on the inset shows the location of the zoomed in area. LC: lobula columnar neuron.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

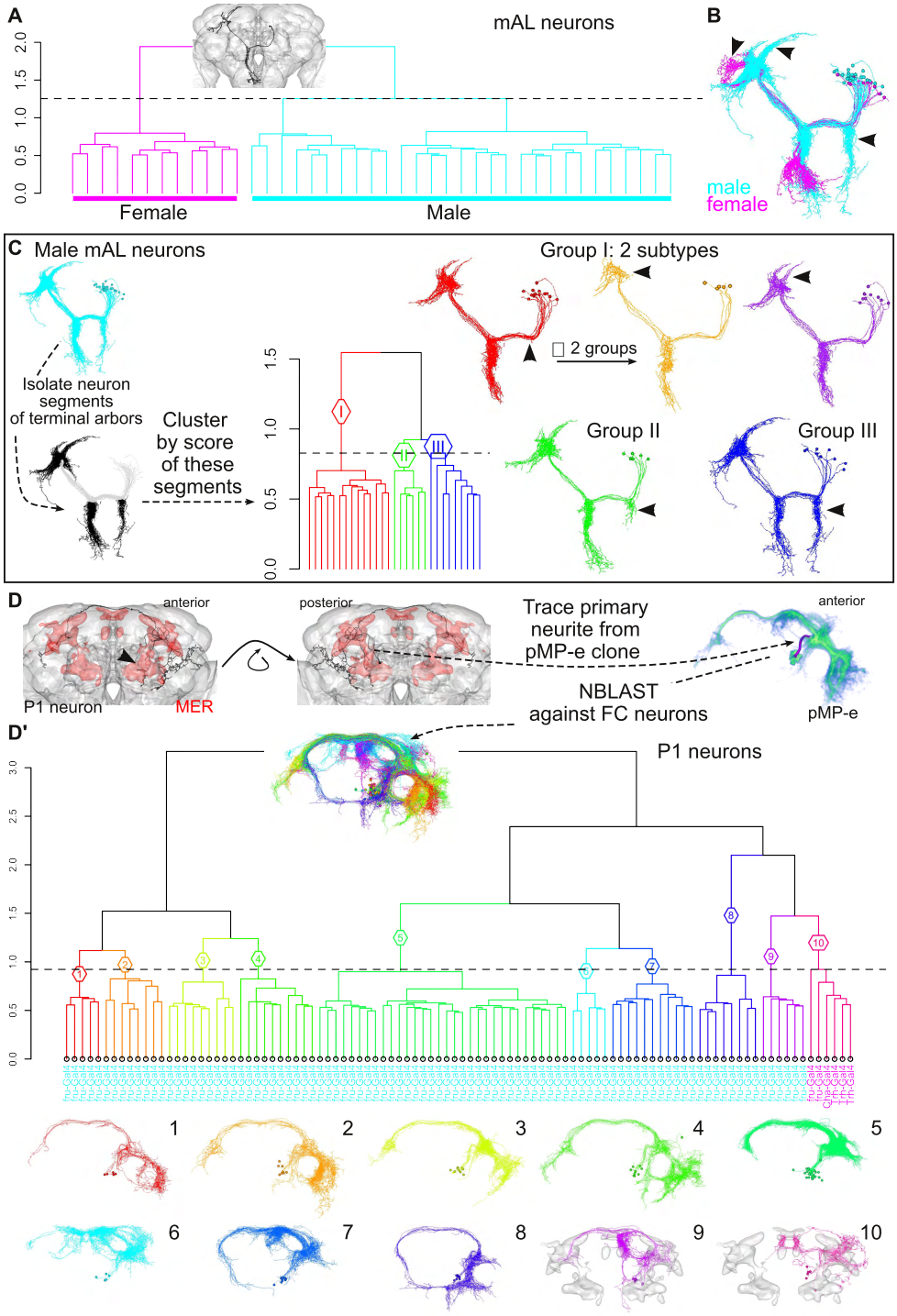


Figure 7: **NBLAST search and classification of hits reveals subtypes of *fruitless*-expressing mAL and P1 neurons**
(A) Analysis of the mAL neurons. Hierarchical clustering (HC) of the hits, divided into 2 groups (h=1.25). The mAL neuron used as the NBLAST query, fru-M-500159, is shown in the inset. Hits with a normalized score over 0.2 were collected. The leaf labels indicate the gender of the neuron: 'F' for female and 'M' for male. **(B)** Neuron plot of the 2 dendrogram groups corresponding to male (in cyan) and female (in magenta) mAL neurons. **(C)** Analysis of the male mAL neurons. The neuron segments corresponding to the terminal arbors (ipsi- and contralateral) were isolated and the neurons were clustered based on the score of these segments. HC of neurons, divided into 3 groups (groups I–III) (h=0.83), that reflect differences in the length of the ventral ipsilateral branch (arrowhead). Group I can be further subdivided into two different subtypes, which differ in the shape and extent of their dorsal contralateral arborisation (arrowhead). **(D)** Analysis of the P1 neurons. Neuron plot of a P1 neuron, fru-M-400046. The male enlarged region (MER) is shown in red. Anterior and posterior views are shown. Volume rendering of the pMP-e *fruitless* neuroblast clone, which gives rise to P1 neurons. The distinctive primary neurite was traced and used on a NBLAST search for matching neurons. **(D')** HC of hits for a search against the P1 primary neurite divided into 10 groups (1–10) (h=0.92, indicated by dashed line). This group of neurons corresponds to a subset of neurons obtained after a first HC analysis. Hits with a normalized score over 0.25 were collected and further selected. The inset shows a neuron plot with groups 1–10. The leaf labels show the GAL4 driver used to obtain that neuron; the colors follow the gender: cyan for male and magenta for female. Below the dendrogram, neuron plots of each group. The MER is shown in grey for groups 9 and 10.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

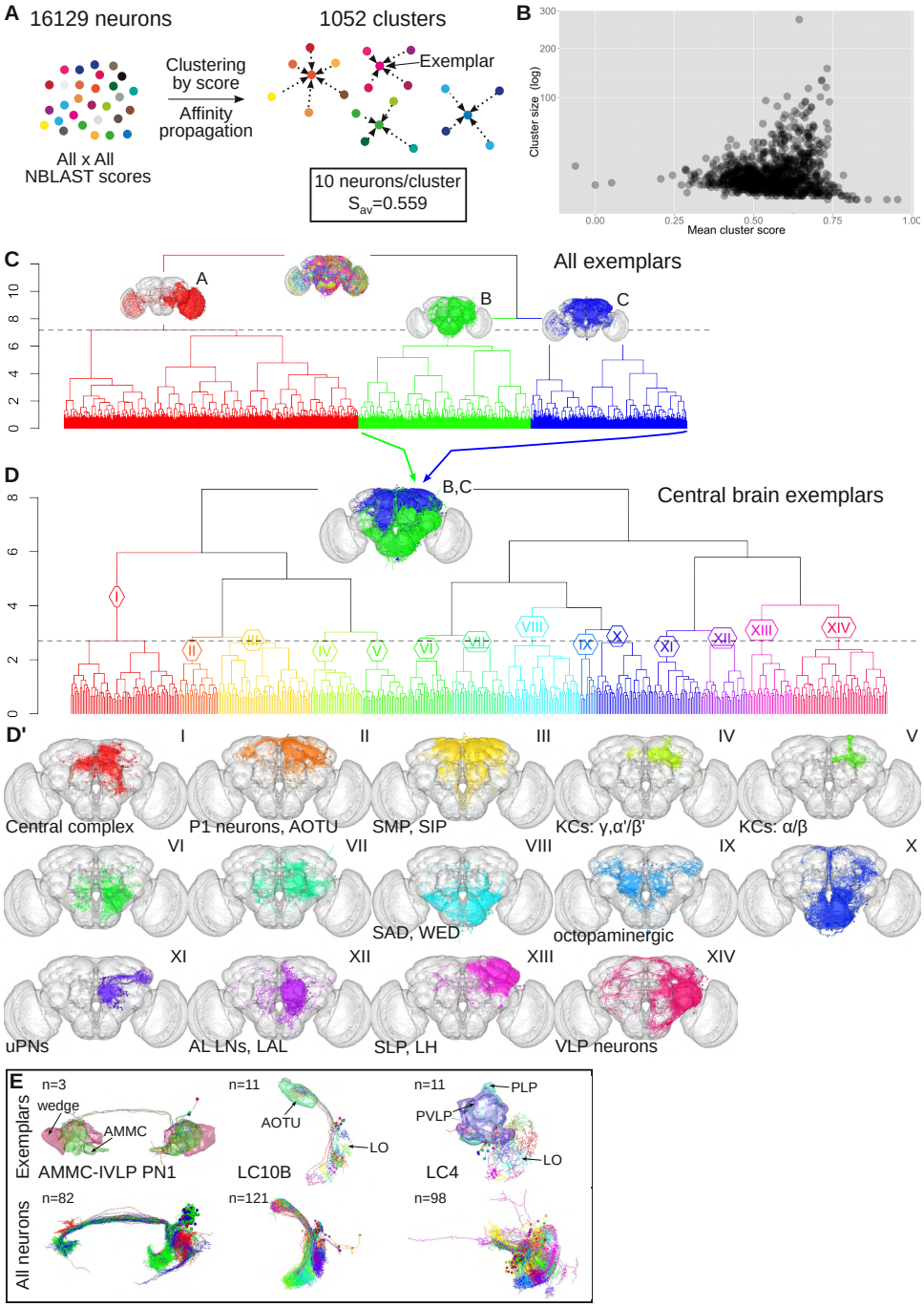


Figure 8: **Organizing NBLAST scores by affinity propagation clustering**
(A) Clustering by affinity propagation. This method uses the all-by-all matrix of NBLAST scores for the 16,129 neurons. This method defined exemplars, which are representative members of each cluster. An affinity propagation clustering of the dataset generated 1,052 clusters, with an average of 10 neurons per cluster and a similarity score of 0.559. **(B)** Plot showing the mean cluster score versus cluster size. **(C)** Hierarchical clustering (HC) of the 1,052 exemplars, dividing them into three groups (A–C). Group A corresponds mostly to optic lobe and VPN neurons; groups B and C to central brain neurons. The insets on the dendrogram show the neurons of these groups. The main neuron types or innervated neuropils are noted. **(D)** HC of central brain exemplars (groups B and C, inset on dendrogram), divided into 14 groups, h=2.7. **(D')** Neurons corresponding to the dendrogram groups in D. **(E)** Affinity propagation clusters of defined neuron types. Neuron plot of exemplars (top row) or all neurons (bottom row) for auditory AMMC-IVLP PN1 neurons (compare with Figure S5D) and VPN types LC10B (compare with Figure 6B) and LC4 (compare with Figure S4B). The number of exemplars and neurons is indicated on the top left corner for each example. The AMMC is shown in green, the wedge in magenta. AMMC: antennal mechanosensory and motor center; AOTU: anterior optic tubercle; LO: lobula; PVLP: posterior ventrolateral protocerebrum; PLP: posterior lateral protocerebrum.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Information

Supplemental Results

Algorithm Design Process

The algorithm design process was primarily motivated by a requirement for rapid and sensitive searches of neuron databases. It was necessary to consider both the data structure and the similarity algorithm jointly in the face of the design requirements.

Our first application would be data acquired in *Drosophila*, where previous studies using image registration have shown a high degree of spatial stereotypy (standard deviation of landmarks ~2.5 µm in each axis for a brain of 600 µm in its longest axis, Jefferis et al., 2007). Therefore one key design decision was to use co-registered data rather than calculating similarity using features of the neuron that are independent of absolute spatial location.

On the algorithm side, the key initial design decision was whether to develop a direct pairwise comparison algorithm or to use a form of dimensional reduction to map neuronal structure into a lower-dimensional space. The major advantage of the latter approach is that the similarity between neurons can be computed directly and almost instantaneously in the low dimensional space. However, the construction of a suitable embedding function either requires existing knowledge of neuronal similarity (likely supplied by experts in the form of large amounts of training data), huge amounts of unlabeled data that enable direct learning of features (e.g. Le et al., 2012), or a strategy based on successful extraction of key image features.

A number of considerations made us favor the approach of direct pairwise comparison. First, we suspected that it would be possible to make a more sensitive algorithm by working with the original data. Second, the amounts of image data available did not seem large enough to avoid a requirement for extensive labeled training data. Third, we reasoned that our own intuitions about neuronal similarity could be better expressed in the original physical space of the neuron than in a low dimensional embedding. Our own exploratory analysis in which we summarized each neuron in different ways as feature vectors of the same dimension and used a comparison function in the feature space (SP, GSXEJ unpublished observations) confirmed that constructing a sensitive metric of this sort is challenging.

The selection of a pairwise similarity metric meant that we had to give particularly careful consideration to performance issues in the design phase. We set two practical performance targets: 1) being able to carry out searches of a single neuron against a database of 10,000 neurons in less than a minute on a simple desktop or laptop computer. 2) Being able to complete all-against-all searches for 10,000 neurons (10^8 comparisons) in < 1 day on a powerful desktop computer. These targets meant that each elementary comparison operation should take around 5 ms or less. Image pre-processing carried out once per neuron would therefore be a good investment if it reduced the time taken for each pairwise comparison. These considerations prompted us to generate a spatially registered, compact representation of each neuron as a separate pre-processing step for each neuron, rather than develop an algorithm that simultaneously solved both the spatial alignment and similarity problem.

Algorithm Scoring

As described in the main results section, we defined NBLAST raw scores as the sum of segment pair scores:

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$S(\text{query}, \text{target}) = \sum_{i=1}^{n} f(d_i, |\vec{u}_i \cdot \vec{v}_i|). \tag{3}$$

We initially experimented with a function based on expert intuition:

$$f = \sqrt{|\vec{u}_i \cdot \vec{v}_i| \exp\left(-d_i^2/2\sigma^2\right)}. \tag{4}$$

This includes a negative exponential function of distance (related to the normal distribution), with a free parameter σ based on our previous estimates of the variability in position within the fly brain of landmarks after registration (Jefferis et al., 2007; Yu et al., 2010b) set to 3 µm. Although this provided a useful starting point, we were unhappy with a scoring system that required parameters to be specified rather than derived empirically from data. This then motivated us to investigate the statistical scoring approach described in based on log probability ratios. This single parameter approach may still be helpful in some situations where insufficient neurons are available in order to define a statistical scoring matrix (we used 150 similar neurons and 5000 random pairs). It can also enable a bootstrapping approach for new datasets in order to help idea similar pairs of neurons that can then be used to define a full scoring matrix.

Kenyon cell analysis

Dataset collection and dividing into types We collected the dataset of initial KCs by performing a forward and reverse search against all neurons using one identified KC (fru-M-500225). We then selected neurons that had both raw scores above −2,500 (2,088 neurons). We performed affinity propagation clustering (Frey and Dueck, 2007) of these neurons, obtaining 59 clusters, and manually verified each one, resulting in 1,562 neurons being identified as KCs. An additional search for high scorers against these KC exemplars uncovered an extra 102 neurons, bringing the total number of KCs used in our analysis to 1,664, representing 10.3 % of the FlyCircuit dataset.

We performed hierarchical clustering of the KCs, based on the NBLAST scores, and divided the dendrogram into two groups (Figure 4A). Contrary to expectations, one group contained both the γ and α′/β′ neurons, whereas the other group consisted exclusively of α/β neurons (Figure 4B–D), the largest subset in our sample. We separated α′/β′ from the γ neurons in a subsequent hierarchical clustering of this group. We performed additional analysis for each of the neuron types.

Analysis of γ neurons Hierarchical clustering of the 470 γ neurons resulted in a dendrogram which we divided into three groups (I–III) (Figure 4B'). The number of clusters was chosen by visual inspection in order to reveal differences in morphology and organization between the groups. Groups I and III corresponded to the classical γ neurons while group II matched atypical γ neurons. There were differences in neurite positioning in the calyx, from medial to lateral, with group I being the most medial, followed by groups II and III. There were also differences in the gamma lobe, with group II occupying the anterodorsal region, while groups I and III were mostly mixed in the rest of the lobe. A subsequent clustering analysis of the classical γ neurons divided into 4 groups (groups A-D) revealed that there were differences between the groups in their medial to lateral position in the calyx. These differences correlated to a certain degree with differences in the dorsal/ventral position of the projections in the

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

γ lobe, with groups C and D, the most medial, being also the most dorsal (Figure 4B”). These observations suggest that the relative position of the projections of classical γ neurons is maintained at the calyx and γ lobe.

In order to understand if the relative position of the classical γ neurites was maintained in between the calyx and the γ lobe, we clustered the neurons based on the scores of the segments in the peduncle. We took neuron skeletons from classical γ neurons and isolated the axon arbors that co-localized with the peduncle volume (Figure S3B’). We then carried out a new clustering based on all-by-all NBLAST scores of these partial skeletons, cutting the dendrogram at a level defined by visual inspection (4 groups). The overall organization almost fully recapitulated the positioning of the neurites in the whole neuron analysis (compare Figure 4B’–B” with Figure S3B’). A clear and expected lamination was found in the peduncle, with neurites occupying the most outer stratum. Differences in the medial to lateral positioning of neurites in the calyx followed the previously observed organization, with the most medial groups occupying the dorsal region of the gamma lobe. The overall organization almost fully recapitulated the positioning of the neurites in the whole neuron analysis (for more information see Figure S3 and Supplemental Results). Thus, the stereotypical organization of the classical γ neurons is maintained throughout the neuropil.

Group II of the γ neurons matched atypical γ neurons (γ dorsal neurons) (Aso et al., 2014) with neurons that extended neurites posteriolaterally in the calyx and projected to the most dorsal region of the γ lobe (Figure 4B”’). Hierarchical clustering of these neurons resulted in a dendrogram that we divided into 3 groups (a–c). This number of groups isolated the previously identified subtype of atypical γ neurons – γd neurons (Aso et al., 2009) – into one group (group a). These neurons extend neurites ventrolaterally at the level of the calyx (identified as ventral accessory calyx) (Aso et al., 2014). The other 2 groups (b, c) correspond to uncharacterized types. Although they project to a similar region in the γ lobe, their dendrites do not extend laterally and their calyx neurites are longer than γd neurons.

Analysis of α′/β′ neurons Hierarchical clustering analysis of the α′/β′ neurons and separation into 4 groups highlighted the characterized subtypes of α′/β′ neurons (Figure 4C–C’). They differ in their anterior/posterior position in the peduncle and β′ lobe with three types described - α′/β′ anterior and posterior (α′/β′ap) and α′/β′ medial (α′/β′m) (Tanaka et al., 2008, Y. Aso, personal communication)). Although we were unable to unambiguously assign a α′/β′ subtype to each group i–iv because our sample was too small, there were clear trends. Neurites of neurons in groups i and iv were more anterior than the other 2 groups (ii, iii) in both the peduncle and β′ lobe. These relative positions were not maintained in the calyx, with the two the anterior groups (i, iv) occupying either a medial or a lateral position.

Analysis of α/β neurons The largest subset of KCs corresponds to α/β neurons (Figure 4D). During the analysis of this group we found 18 neurons that did not correspond to α/β cells, since they innervated either only the β or the α lobes, and they were removed from the analysis. We performed hierarchical clustering on the remaining 1,091 cells and divided the resulting dendrogram into four groups (1–4) (Figure 4D’), which matched the four neuroblast lineages from which they originate (Zhu et al., 2003). The relative position of the neurites of the four groups within the calyx is somewhat maintained in the peduncle, with the lateral neuroblast clones (group 1 AL; group 2 PL) extending along the dorsolateral peduncle, while the medial clones (group 3 AM; group 4 PM) occupy a more ventromedial region.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hierarchical clustering of each group revealed an expected common organization for all neuroblast clones (Figure 4D" and Figure S3A). For all groups there was a clear distinction between the late born core (α/β core, α/β-c) and early born peripheral neurons (α/β surface, α/β-s). Core neurons are on the inside stratum of the α lobe. They are also reported to occupy the inside stratum of the peduncle and β lobe (Tanaka et al., 2008). We were unable to observe this, although the projections of α/β core neurons were ventral to α/β surface neurons in both the peduncle and β lobe. There was also a trend for α/β surface neurons to occupy a more medial position in the calyx in comparison to α/β core ones. A subgroup of group 2 corresponded to the α/β posterior or pioneers neurons (α/βp). The α/βp neurons are the earliest born α/β and they innervate the accessory calyx, run along the surface of the posterior peduncle into the β lobe but stop before reaching the medial tip (Tanaka et al., 2008). A new clustering based on peduncle position of the neuron segments did not recapitulate the relative positions of the calyx neurites for each of the neuroblast clones observed in the whole neuron analysis suggesting that the relative position of the α/β neurons in the peduncle does not completely reflect their stereotypical organization in the calyx (for more information see Figure S3 and Supplemental Results).

In order to investigate the stereotypical organization of α/β neurites, we performed a similar analysis as for the classic γ neurons, isolating the axon arbors that co-localized with the peduncle for groups 1 to 4 (Figure S3B"). The new clustering based on peduncle position of these partial neuron skeletons did not recapitulate the relative positions of the calyx neurites for each of the neuroblast clones observed in the whole neuron analysis (compare Figure 4D' with Figure S3B"). In addition, there was no clear organization of neurites in the α lobe that correlated with their position in the peduncle.

Olfactory projection neuron analysis

We started by manually classifying the 400 uPNs in the FlyCircuit dataset by glomerulus, neuroblast lineage, and axon tract, using the original image stacks. The definition of the manual gold standard annotations was an iterative process that took several days. The first round accuracy was about 95 %. Numerous discrepancies were revealed by subsequent NBLAST analysis and difficult cases were resolved by discussion between two expert annotators before finalizing assignments. We excluded 3 neurons for which no conclusion could be reached. We found a very large number of DL2 uPNs, 145 DL2d and 37 DL2v neurons, in a total of 397 neurons. Nevertheless, our final set of uPNs broadly represents the total variability of described classes and contains neurons innervating 35 out of 56 different glomeruli (Tanaka et al., 2012), examples of the three main lineage clones (adPNs, lPNs and vPNs) in addition to one bilateral uPN, and neurons that follow each of the three main tracts (medial, mediolateral and lateral antennal lobe tracts). For subsequent analysis, we removed 3 neurons for which registration failed.

Visual projection neuron analysis

We started with the 1,052 exemplars found by affinity propagation clustering of NBLAST scores (Figure 8). We then clustered those exemplars using hierarchical clustering and found that extrinsic and intrinsic optic lobe neurons together formed a distinct "optic lobe" group within this (Figure 8C). We then collected all neurons associated with those "optic lobe" exemplars and calculated the overlap of neurons with each of the standard neuropils defined by Ito et al. (2014) (see Experimental Procedures and Manton et al., 2014 for technical details). This then allowed us to separate neurons by innervation

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pattern into 3 groups: 1) ipsilateral optic lobe neuropils only (see Figure S6), 2) ipsilateral and central brain neuropils (unilateral VPNs, uVPNs) or 3) both optic lobes and central brain neuropils (bilateral VPNs). This selection procedure resulted in a set of 1,793 uVPNs and 72 bilateral VPNs.

Auditory neuron analysis

We employed a two-step search strategy. First, for 5 of the auditory types, we used the FlyCircuit neuron named by Lai et al. (2012) as the seed neuron for the first search. Candidate neurons were selected using strict anatomical criteria. A second search was then done using these candidates as query neurons and collecting all high scorers (score over 0.5). These neurons corresponded to our set for each of the types.

mAL neuron analysis

The set of maL neurons resulted from a search with a seed mAL neuron, fru-M-500159. We then collected 41 hits with a mean NBLAST score greater than 0.2.

P1 neuron analysis

The set of P1 neurons was identified by searching the FlyCircuit dataset with a tracing of the distinctive primary neurite of a pMP-e clone (Cachero et al., 2010). Hierarchical clustering of the top hits, after manual verification, identified a subset consisting solely of P1 neurons which was used in the subsequent analysis.

Online resources

The online resources provided in the paper are listed at http://jefferislab.org/si/nblast. In addition to the open source software described in the Experimental Procedures we also provide:

- Code and instructions to generate some of the figure panels used in the paper. Instructions can be found here. A video demo is also available here.
- The affinity propagation clustering of the flycircuit.tw dataset (as in Figure 8D), excluding intrinsic optic lobe neurons. This can be viewed online here, including interactive 3D rendering of clusters powered by WebGL.
- The clusters identified by affinity propagation clustering (including intrinsic optic lobe neurons) are all indexed by the Virtual Fly Brain website, which links to 3D WebGL renderings of each cluster hosted at jefferislab.org. Clusters can be identified by VFB queries for the neuropil region that they innervate. For example, search from the VFB homepage for "AMMC". From the results page, choose the query "Images of neurons with: some part here (clustered by shape)". A list of clusters, with thumbnail images is displayed; single exemplars are also displayed for each cluster, hyperlinked to the original data at flycircuit.tw. The images of the individual neurons that are part of this cluster can also be displayed in the stack browser from this page ("Show individual members"). Clicking on a cluster thumbnail links to a page which includes a snapshot and 3D rendering of the cluster, and information about the neurons that are part of this cluster, including

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

links to the appropriate Neuron ID pages at flycircuit.tw; a second table provides links to result pages for the most similar clusters. A video demo is available here

- A video demo showing how to search for FlyCircuit neurons similar to a GAL4 tracing using the R packages detailed in Experimental Procedures.
- An online web-app allowing on-the-fly NBLAST queries of FlyCircuit neurons against other FlyCircuit neurons, as well as queries of user-uploaded neurons against the FlyCircuit dataset, available at http://jefferislab.org/si/nblast/online/.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Figures and Tables

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

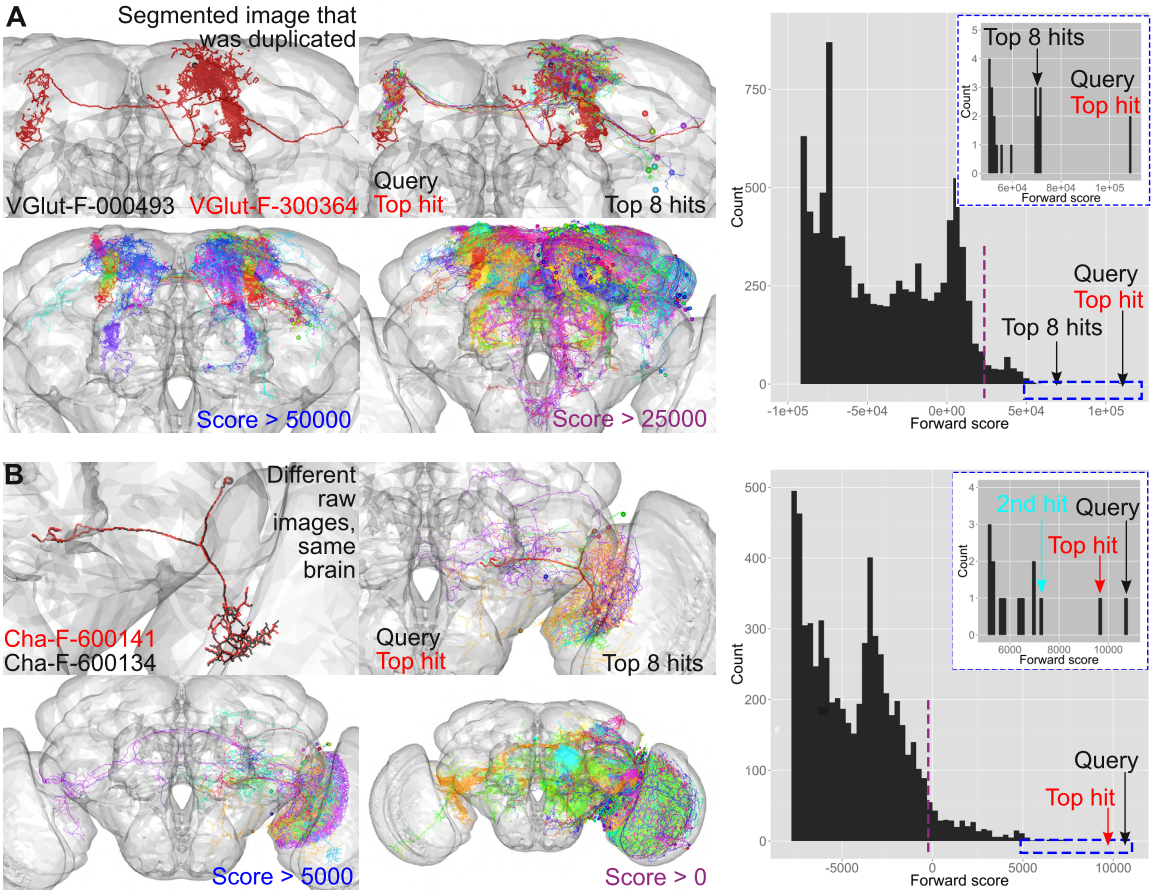


Figure S1: **Neuron search with NBLAST**

(A) NBLAST search with VGlut-F-000493 as query. Neuron plots of (from higher to lower score): the query (black) and top hit (red), top 8 hits, hits with a score over 50,000 and hits with a score over 25,000. The top hit corresponds to a segmented image that was duplicated. It perfectly overlays the query neuron. As the score decreases, so does the similarity of the hits to the query. On the right, histogram of forward scores. Only hits with scores over −100,000 are shown. The score of the query, top hit and top 8 hits are indicated. A dashed purple line marks 25,000. The left inset shows a zoomed view of the top hits (score > 50,000) (dashed blue rectangle in main plot). The score of the query, top hit and top 8 hits are indicated. **(B)** NBLAST search with Cha-F-600134 as query (black). Neuron plots of (from higher to lower score): the query and top hit, top 8 hits, hits with a score over 5,000 and hits with a score over 0. The top hit corresponds to an image of a neuron from the same brain but from a different raw image. It is very similar to the query neuron. As the score decreases, so does the similarity of the hits to the query. On the right, histogram of forward scores. Only hits with scores over −8,000 are shown. The score of the query and top hit are indicated. A dashed purple line marks 0. The left inset shows a zoomed view of the top hits (score > 5,000) (dashed blue rectangle in main plot). The score of the query, top hit and second top hits are indicated.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

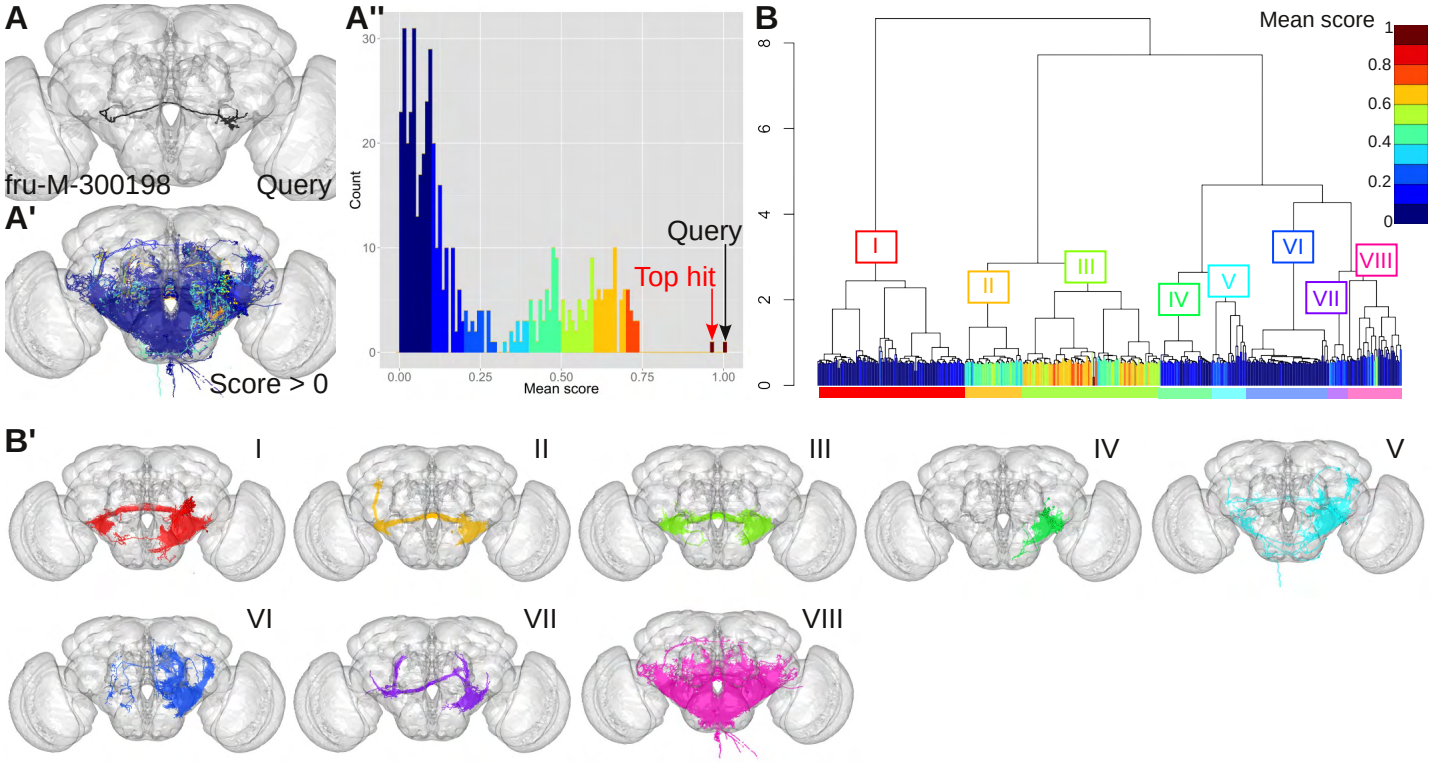


Figure S2: **NBLAST search using mean scores**
(A) The query neuron fru-M-300198 (same as inFigure 3B). **(A')** Neuron plot of the hits with a mean score over 0 for a search against the query. Hits are colored by score bin (10), as in A". **(A")** Histogram of mean scores for hits against fru-M-300198 with a score over 0 divided into 10 bins (indicated in the scale bar in B). **(B)** Hierarchical clustering of hits with a mean score over 0. The leaves of the dendrogram are colored by score (same as in A"), and as shown in the scale bar. The dendrogram was divided into eight groups (I–VIII), with each one being assigned a color, shown on the colored rectangle below the leaves. The query neuron is in group III, and the hits with the higher scores are in groups II and III. **(B')** Neuron plots corresponding to the dendrogram groups (I–VIII), following the colors assigned to each group. Groups II and III, corresponding to the highest scores, are the most similar neurons to the query.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

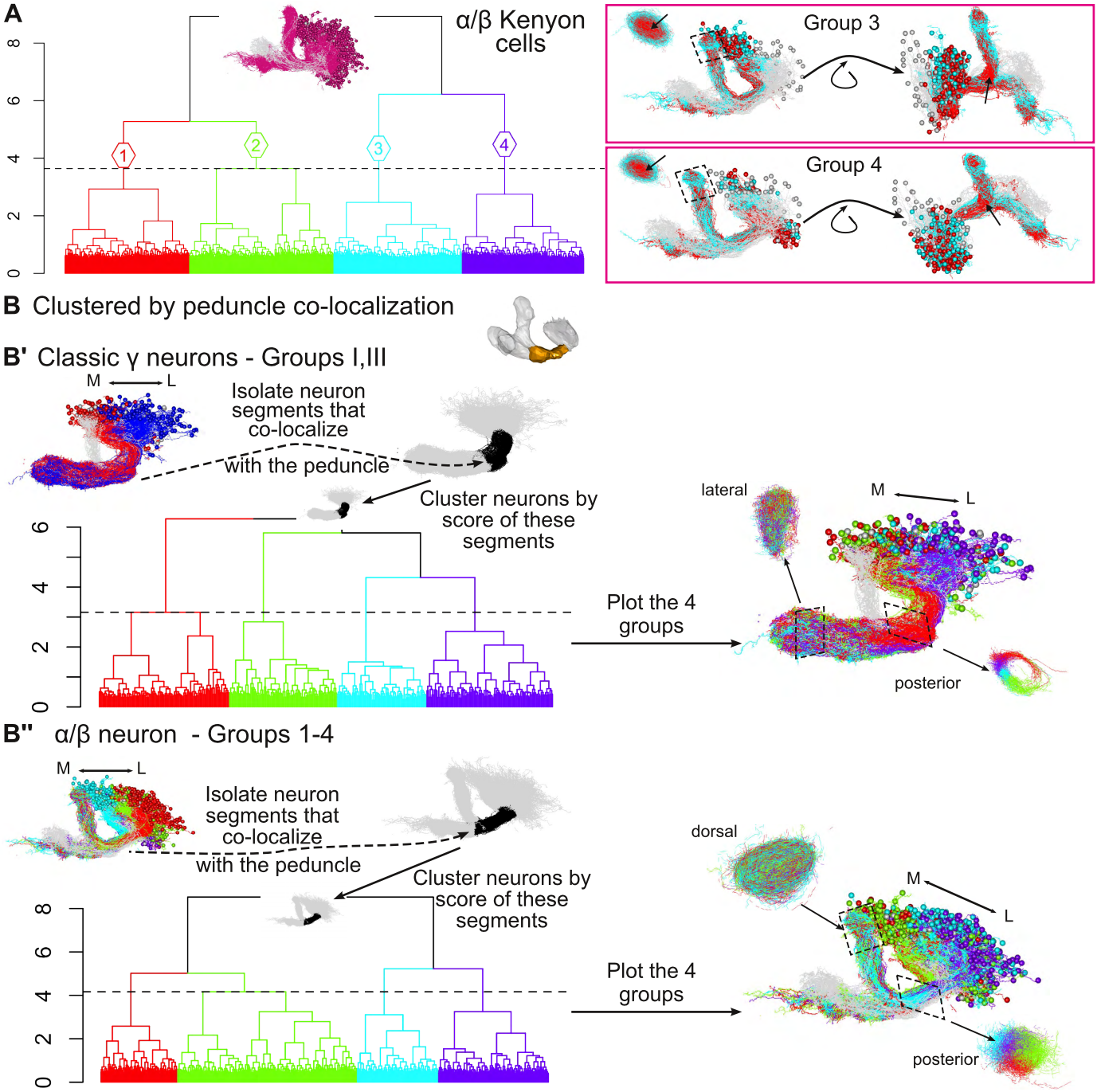


Figure S3: **NBLAST search and classification of hits reveals Kenyon cells subtypes**
(A) Hierarchical clustering (HC) of the α/β neurons, divided into four groups (1–4) (h=3.64) (same as in Figure 4D'). The inset on the dendrogram shows the α/β neurons. Groups 3 and 4 were clustered and divided into 2 groups each. This separates the neurons into peripheral (cyan) and core (red) in the α lobe. Peripheral neurons occupied a more lateral calyx position and were dorsal to core neurons in the peduncle and β lobe. Similar analysis to group 1 is shown in Figure 4D". Lateral oblique, posterior oblique and a dorsal view of a peduncle slice (position indicated by dashed rectangle) are shown. **(B)** Reclustering of Kenyon cells based on the co-localization of neurons segments in the peduncle. The neuron segments that co-localize with the peduncle were isolated, followed by HC of the neurons based on the NBLAST score of the segments. **(B')** HC of the neurons segments of classic γ neurons, groups I and III (see Figure 4B") divided into 4 groups, h=3.16. Neuron plot of the 4 groups. A posterior view of a slice of the peduncle shows an expected clear organization. It correlates to the position of the neurons in the calyx, with more medial neurons (cyan and green) being dorsal and ventral in the peduncle than more lateral neurons (red and purple). No clear structural organization is discernible in a lateral view of a slice of the γ lobe. **(B")** HC of the neurons segments of classic α/β neurons, groups 1 to 4 (see Figure 4D') divided into 4 groups, h=4.16. Neuron plot of the 4 groups. A posterior view of a peduncle slice shows an expected clear organization. It correlates to the position of the neurons in the calyx, with more medial neurons (cyan and green) being ventrolateral in the peduncle than more lateral neurons (red and purple). No structural organization is discernible in a dorsal view of a slice of the α lobe. For all neuron plots, the neurons in grey correspond to the Kenyon cell exemplars.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Dendrogram group	Neuron type	Comments
1	LC6	Lobula innervation, dorsal cell bodies, axons follow the anterior optic tract to the AOTU, turning ventrally midway to innervate the lateral PVLP.
2	LC9	As group 1, but terminating in the medial PVLP rather than extending laterally.
3	LC10B (possibly different subtype to 4)	Dorsal AOTU innervation, ventral to group 4.
4	LC10B (possibly different subtype to 3)	Dorsal AOTU innervation, dorsal to group 3.
5	New LC10 subclass	The most ventral AOTU innervation. Similar to group 7, but ventral to it in the AOTU.
6	LC10A	Axons project through ventral AOTU, turn sharply dorsally to terminate in the dorsal AOTU.
7	New LC10 subclass	Ventral AOTU innervation, dorsal to group 5.

Table S1: Correspondences between hierarchical clustering groups of AOTU- and PVLP-innervating uVPNs via NBLAST scores and previously determined neuron types (Otsuna and Ito, 2006).

Dendrogram group	Neuron type	Comments
A	LC12 (possibly different subtype to B)	Innervation in most lateral and anterior PVLP glomerulus.
B	LC12 (possibly different subtype to A)	Innervation in more posterior and medial PVLP glomerulus than group A.
C	LC4	Innervates a more medial PVLP glomerulus than LC12.
D	LT12	Tentative match. Class was identified in Otsuna and Ito (2006) based on a single neuron.
E	LC11	Innervates more dorsal PVLP glomelurus than LC12. Extends along the posterior PVLP, with a sharp anterior turn. Terminates with a blunt-stick like ending in the lateral PVLP.
F	New LT subclass	Similar to LT12, but with projections posterior to it, terminating in the lateral region of the superior posterior slope.
G	Unmatched	Do not correspond to a single type.

Table S2: Correspondences between hierarchical clustering groups of PVLP- and PLP-innervating uVPNs via NBLAST scores and previously determined neuron types (Otsuna and Ito, 2006).

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

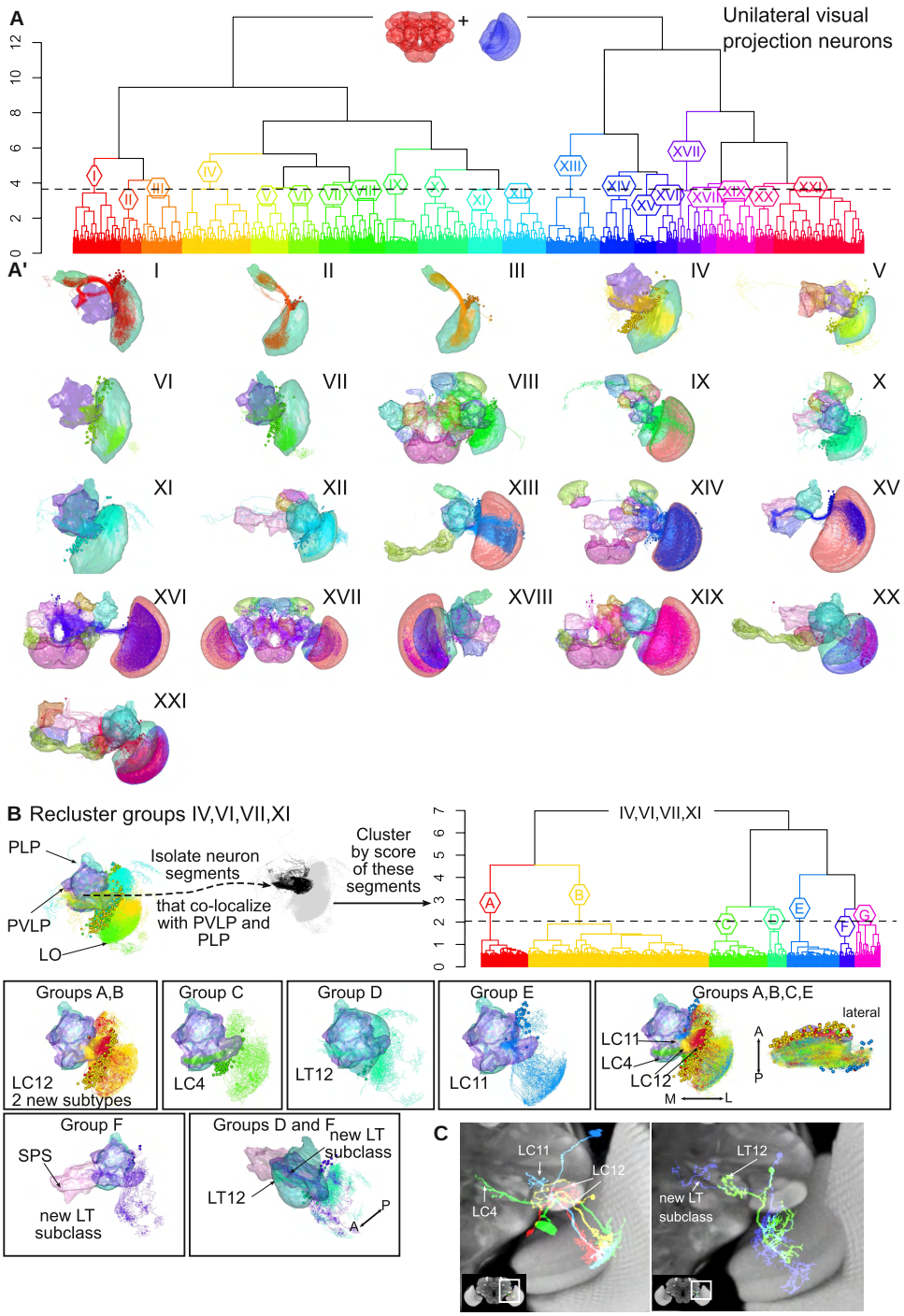


Figure S4: **NBLAST search and classification hits uncovers unilateral visual projection neurons neuronal types**
(A) Hierarchical clustering (HC) analysis of unilateral visual projection neurons (uVPNs), defined as neurons with segments that overlap one optic lobe, and some central brain neuropil. The dendrogram was divided into 21 groups (I–XXI), h=3.65. Inset on the dendrogram shows the neuropils considered for the overlap. Below, neuron plots of groups I to XXI. The neuropils that contain the most overlap are shown. **(B)** Reclustering of uVPN groups IV, VI, VII and XI from A. These neurons arborize in the lobula (LO) and project to the posterior ventrolateral protocerebrum (PVLP) or posterior lateral protocerebrum (PLP). The neuron segments that co-localize with either the PVLP or PLP were isolated, followed by HC of the neurons based on the NBLAST score of these neuron segments. The dendrogram was divided into seven groups (A–G), h=2.04. Neuron plots corresponding to the dendrogram groups. An anterior, a lateral or lateral oblique views are shown. Some of dendrogram groups were matched to known uVPN types. Groups A and B possibly correspond to two LC12 subtypes, that innervate the more lateral PVLP glomeruli. Group B innervates a more anterior and medial glomeruli than group A (see also C). Group C corresponds to LC4, that innervates a lateral PVLP glomeruli and ventral to LC12. Group D corresponds to LT12 neurons, that project from the lateral to the medial PVLP, posterior to LC4. Group E corresponds to LC11, that innervates a lateral PVLP glomeruli, dorsal to LC12. These neurons extends along the posterior PVLP and make a sharp anterior turn, terminating with a blunt-stick like ending in the lateral PVLP. Group F corresponds to a possibly new LT subclass, with neurons projecting posteriorly to LT12 in the PVLP and extending into the superior posterior slope (SPS). **(C)** Overlay of Z projections of registered image stacks of example neurons from the types identified in B on a partial Z projection of the template brain (a different one for each panel). The white rectangle on the inset shows the location of the zoomed in area. LC: lobula columnar neuron; LT: lobula tangential neuron.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

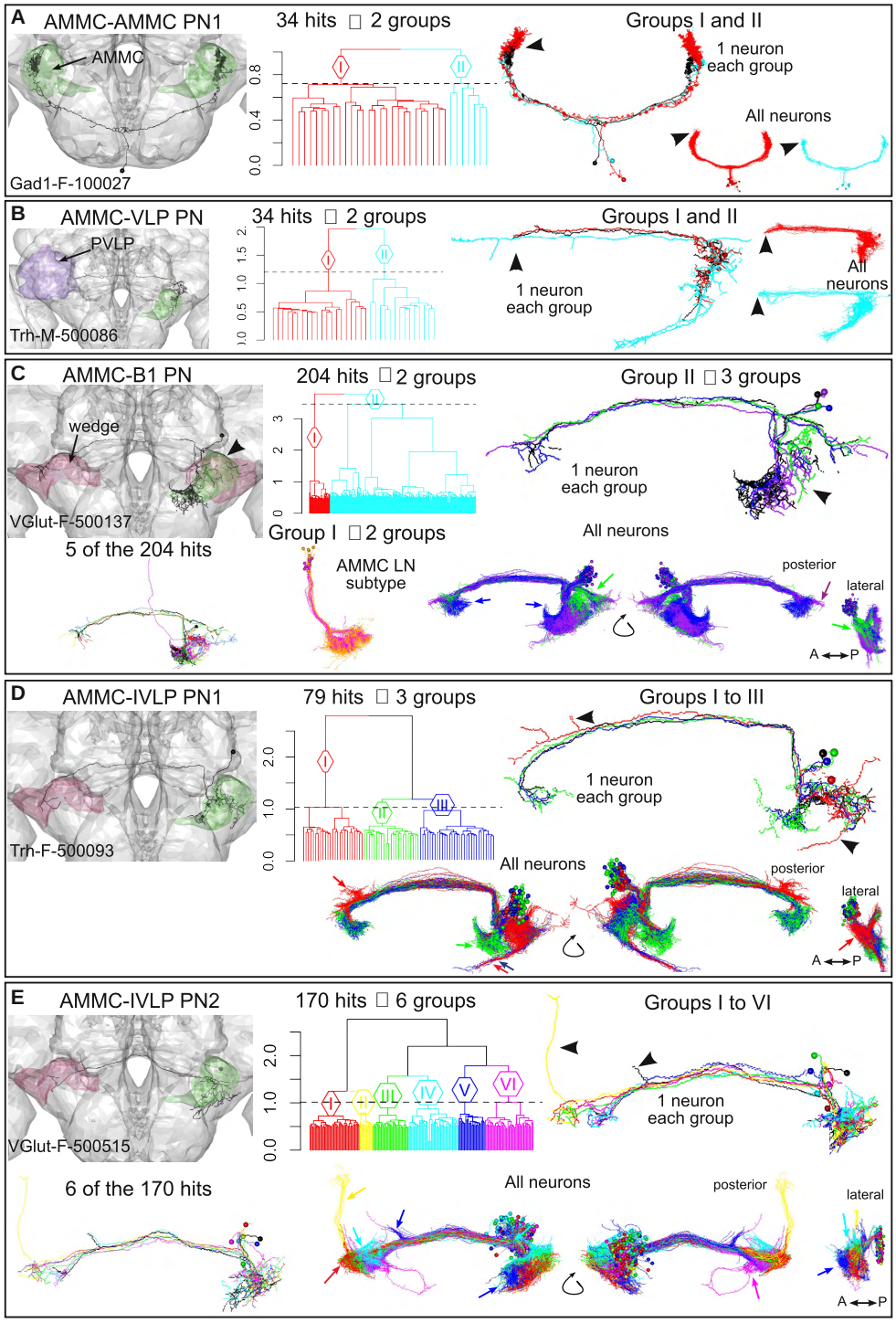


Figure S5: **NBLAST search and classification hits reveals auditory neuron types** Searches were done using types identified by Lai et al. (2012). A first search using the neuron named by Lai for each type (shown on the left panel with wedge in magenta, AMMC in green or PVLP in purple) was done to collect the possible candidates. A second search was then done using these neurons as queries and collecting all the high scorers (over 0.5). The dendrogram and neuron plots (anterior, posterior and lateral views) showing either one or all neurons from each clustering group are shown on the middle and right panels. **(A)** Hierarchical clustering (HC) of AMMC-AMMC projection neuron 1 (PN1). The 34 top scorers were clustered and divided into 2 groups, h=0.72. Neuron plot of the query neuron (black), and one neuron from each group (red and cyan). To the right, neuron plots of of each group. Differences in anterior length of the left projection are indicated by an arrowhead. **(B)** HC of AMMC-VLP PNs. The 34 top scorers were clustered and divided into 2 groups, h=1.2. Neuron plot of the query neuron (black), and one neuron from each group (red and cyan). On the left, neuron plots of each group. **(C)** HC of AMMC-B1 PNs. Five hits of the 204 top scorers are shown on the left. The 204 top scorers were clustered and divided into 2 groups, h=3.46. Group I was matched to an unidentified type of AMMC local neurons (LN). It was clustered and divided into 2 groups, h=0.77. Group II corresponds to a mix of AMMC-B1 PNs and AMMC-IVLP PN1. After selecting the AMMC-B1 PNs, the neurons were clustered and divided into 3 groups, h=1.5. Neuron plot of the query neuron (black), and one neuron from each group (purple and green). Below, neuron plots of the 3 groups. Arrows indicate differences between groups. **(D)** HC of AMMC-IVLP PN1. The 79 top scorers were clustered and divided into 3 groups, h=1.03. Neuron plot of the query neuron (black), and one neuron from each group (red, green and blue). Below, anterior, posterior and lateral view neuron plots of the 3 groups. **(E)** HC of AMMC-IVLP PN2. Six hits of the 170 top scorers are shown on the left. The 170 top scorers were clustered and divided into 6 groups, h=1.02. Neuron plot of the query neuron (black), and one neuron from each group (red, yellow, green, cyan, blue, magenta). Below, anterior, posterior and lateral view neuron plots of the 6 groups. Arrows and arrowheads indicate differences between groups.

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Panel	Neuron type	Comments
A	AMMC-AMMC PN1 (2 possible subtypes)	These neurons innervate both AMMCs, with a ventral cell body. Group I extends more dorsally than group II.
B	AMMC-VLP (2 possible subtypes)	These neurons innervate the ipsilateral AMMC and the contralateral VLP. Group II extends more laterally than group I.
C	AMMC-B1 PN (3 possible subtypes)	These neurons innervate the ipsilateral AMMC and IVLP and the contralateral IVLP. Blue group innervates more medial regions and has more extensive innervation contralaterally; purple group innervates more anterior and posterior regions ipsilaterally, and the green group innervates the dorsal regions in the ipsilateral AMMC.
C	New AMMC-LN type (2 subtypes)	A type of AMMC LN, with a dorsal cell body. Two possible subtypes: the magenta group innervates more dorsal regions than the orange group.
D	AMMC-IVLP PN1 (3 possible subtypes)	These neurons innervate the ipsilateral AMMC and the contralateral IVLP. Red group has a more dorsomedial ipsilateral innervation more medial regions, with some dorsal medial branches in the contralateral hemisphere; some neurons extend a long neurite ventrally ipsilaterally. The green group innervates the more ventromedial regions ipsilaterally. The blue group is similar to the green one, although it does not extend as ventrally in the ipsilateral side, and a few neurons extend a long neurite ventrally (similar to red group).
E	AMMC-IVLP PN2 (6 possible subtypes)	These neurons innervate the ipsilateral AMMC and the contralateral IVLP, with a posterior cell body. Group I innervates the more lateral regions in both hemispheres. Group II has a long dorsal branch in the contralateral hemisphere, at the lateral edge of the neuron. Group III and IV are very similar, with the latter innervating more dorsal regions in both hemispheres. Group V corresponds to the strict definition of the neuron type by Lai et al. (2012), showing a short dorsal branch just medial to the contralateral IVLP. Group VI are similar to group V, with a few neurons showing a short dorsal branch, and innervating a more ventral region in the contralateral IVLP.

Table S3: Correspondence between hierarchical clustering of auditory neuron via NBLAST scores and previously determined neuron types (Lai et al., 2012).

bioRxiv preprint doi: https://doi.org/10.1101/006346; this version posted February 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

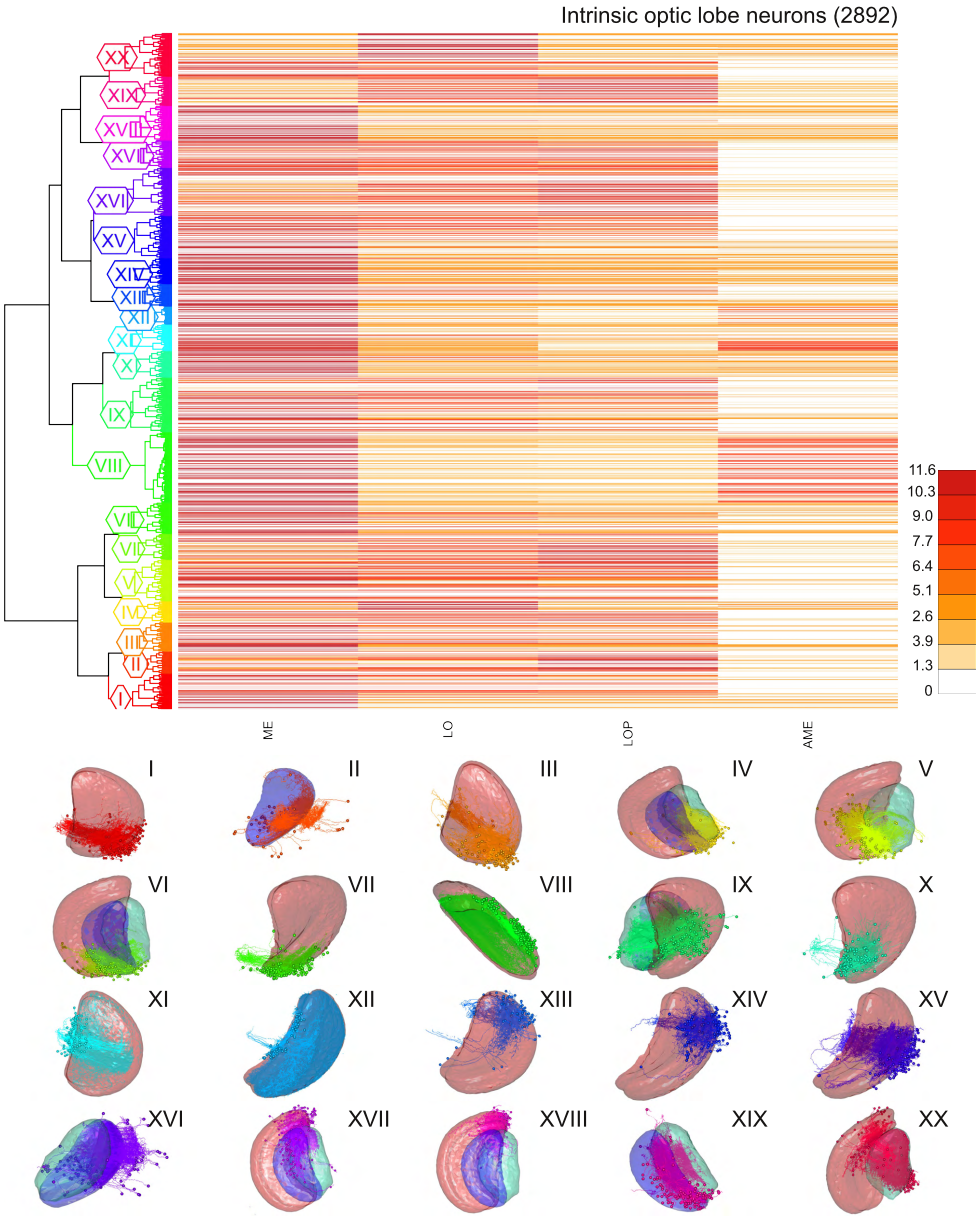


Figure S6: **NBLAST scores and classification of hits highlights neuropil organization of intrinsic optic lobe neurons**
Hierarchical clustering of intrinsic optic lobe neurons. This neuron set was defined as any neuron that overlapped only one of the optic lobes and with no arborization in the central brain neuropils. Dendrogram of the intrinsic optic lobe neurons, divided into into 20 groups (I–XX) with the corresponding heatmap calculated from the neuropil overlap in the different neuropils: medulla (ME), lobula (LO), lobula plate (LOP) and accessory medulla (AME). The values were log transformed. Neuron plots corresponding to the dendrogram groups are shown below. The neuropils for which the overlap is more significant are plotted. Although some organizational structure is seen, the dendrogram groups to do not represent unique types.