Title: Automated cell annotation in multi-cell images using an improved CRF_ID algorithm

Hyun Jee Lee[1†], Jingting Liang[3†], Shivesh Chaudhary[1], Sihoon Moon[1], Zikai Yu[2], Taihong Wu[3], He Liu[3,4], Myung-Kyu Choi[3], Yun Zhang[3,5]*, Hang Lu[1,2]*

[1]School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, United States

[2]Interdisciplinary BioEngineering Program, Georgia Institute of Technology, United States

[3]Department of Organismic and Evolutionary Biology, Harvard University, United States

[4]Present address: Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai 519087, China

[5]Center for Brain Science, Harvard University, United States

†: co-first authors

*: co-corresponding authors

Abstract

Cell identification is an important yet difficult process in data analysis of biological images. Previously, we developed an automated cell identification method called CRF_ID and demonstrated its high performance in *C. elegans* whole-brain images (Chaudhary et al, 2021). However, because the method was optimized for whole-brain imaging, comparable performance could not be guaranteed for application in commonly used *C. elegans* multi-cell images that display a subpopulation of cells. Here, we present an advance CRF_ID 2.0 that expands the generalizability of the method to multi-cell imaging beyond whole-brain imaging. To illustrate the application of the advance, we show the characterization of CRF_ID 2.0 in multi-cell imaging and cell-specific gene expression analysis in *C. elegans*. This work demonstrates that high accuracy automated cell annotation in multi-cell imaging can expedite cell identification and reduce its subjectivity in *C. elegans* and potentially other biological images of various origins.

Introduction

One of the bottlenecks in biological research is the inefficiency and inaccuracy of analyzing

bioimages, which have become essential research materials owing to the emergence of advanced microscopy and imaging modalities (Xu and Jackson, 2019). For studies with the popular model organism *Caenorhabditis elegans* (*C. elegans*), one of the most challenging image analysis processes is cell identification, annotating the cells in the image based on their anatomical or other biological features. Accurate cell identification is important in applications, such as gene expression analysis and calcium imaging in order to associate the cell-specific information with existing knowledge about the cell. Previously, we developed an automated cell identification method called CRF_ID based on graphical optimization using the Conditional Random Fields (CRF) model (Chaudhary et al., 2021). We demonstrated that for whole-brain images, CRF_ID shows higher annotation accuracy and more robustness against various sources of noises compared to conventional registration-based methods.

However, because CRF_ID was optimized for whole-brain images, it is not ideal for multi-cell imaging, which focuses on a subpopulation of cells. In fact, there are no automated methods currently designed for processing multi-cell images; only ad hoc and heuristic tracking and annotation are available. This represents an unmet demand because multi-cell imaging is still much more frequently used than whole-brain imaging despite the recent popularization of brain-wide imaging. For instance, multi-cell imaging is necessary for transcriptional or translational reporter-based gene expression analysis because the number of cells imaged is governed by the gene expression itself, with the average being around 40 neurons (Taylor et al., 2021). Another reason for the continued prominence of multi-cell imaging is that many biological questions can be effectively answered with a subset of the nervous system with specific structures and functions. For example, studies focusing on chemosensory neurons of the olfactory circuit in *C. elegans* characterized how the identity and intensity of olfactory stimuli were represented by the activity of the sensory neurons, which provided input signals for computation and integration of the downstream circuits (Lin et al., 2022). Furthermore, multi-cell imaging is more accessible than whole-brain imaging because it does not require fast-speed and multi-color volumetric microscopy techniques.

In this work, we present CRF_ID 2.0, an update of our original CRF_ID algorithm for multi-cell images. Compared with other automated whole-brain annotation methods (Toyoshima et al., 2020; Yu et al., 2021), CRF_ID is an ideal method to adapt for multi-cell images because of its demonstrated high accuracy, modularity, and efficiency for atlas-building; CRF_ID builds structured models, rather than deep-learning models, which makes it readily interpretable. In principle, the original CRF_ID algorithm should be applicable to multi-cell images; however, in practice, multi-cell specific modifications were necessary to achieve the highest accuracy. In addition to optimizing the method, we characterized its performance by comparing the accuracy of different types of atlases against

each other and against manual annotations. Such characterizations, which were not addressed in our previous work, provide a necessary reference for future users. Furthermore, we demonstrate the application of multi-cell neuron annotation in a cell-specific gene expression analysis. Thus, this follow-up work enhances the generalizability and usefulness of the CRF_ID method by enabling high performance operation regardless of the number of cells in the images.

Results

CRF_ID 2.0 for automatic cell annotation in multi-cell images

The multicell identification pipeline is a multi-step optimization algorithm (Figure 1a), built upon CRF_ID (Chaudhary et al., 2021). First, the user acquires a volumetric image set of a sample that contains multiple cells labeled with markers, such as fluorescent proteins or dyes (Figure 1a-1). Image processing begins with cell segmentation. Here we use a simple automatic method that finds the local maxima of fluorescence intensity and fits the 3D Gaussian mixture model on them (Figure 1a-2). Then, the coordinate axes of *C. elegans* body orientation are either assigned by the user or automatically predicted on the point cloud using an improved algorithm included in CRF_ID 2.0 (Fig. 1a-3). Next, based on the cell coordinates relative to the axes, the algorithm extracts the positional features of the cells, such as the pair-wise 3D positional relationships and angular relationships with other cells (Figure 1a-4). Lastly, the Conditional Random Fields (CRF) model (Lafferty et al., 2001) compares the extracted features from the dataset against a reference atlas, which could be derived from the literature or new data (Figure 1a-5). The model computes a conditional joint probability distribution over all feasible cell identification (cell ID) assignments, and the neuronal cell ID assignments are ranked for each cell based on the computed probabilities (Figure 1a-6). Additionally, to maximize the accuracy of neuron identification, the reference atlas may be constructed by the user with use-specific datasets (Figure 1b). The atlas-building process is computationally simple and fast upon the availability of ground-truth datasets, which can be manually annotated by the user.

To demonstrate the utility of the automatic cell annotation algorithm for multi-cell images, we chose as an example a *C. elegans* strain carrying a red fluorescent protein mCherry expressed by a *glr-1* promoter (see the schematic and fluorescence images in Figure 1). The fluorescent protein has been tagged with nuclear localization sequences to confine the signal to the cell nucleus and aid the separation of the labeled neurons. *glr-1* is expressed in about 27 neuronal cell types with the majority localized in the head (Hart et al., 1995; Maricq et al., 1995; Taylor et al., 2021). The set of neurons include interneurons and motor neurons implicated in various behaviors and neuronal functions, including locomotion, chemotaxis, learning and memory (Gray et al., 2005; Hart et al., 1995; Liu and Zhang, 2020; Maricq et al., 1995). This scale of expression is typical of many neuronal

genes for most of which robust and easy cell identification methods do not currently exist. In contrast, the *glr-1* promoter and its expression has been well-characterized (Hart et al., 1995; Maricq et al., 1995; Taylor et al., 2021) which enables manual cell annotation of the images. For this work, we collected images from transgenic worms expressing *glr-1::mCherry* transgene and manually annotated 26 volumes to assess the performance of our method.

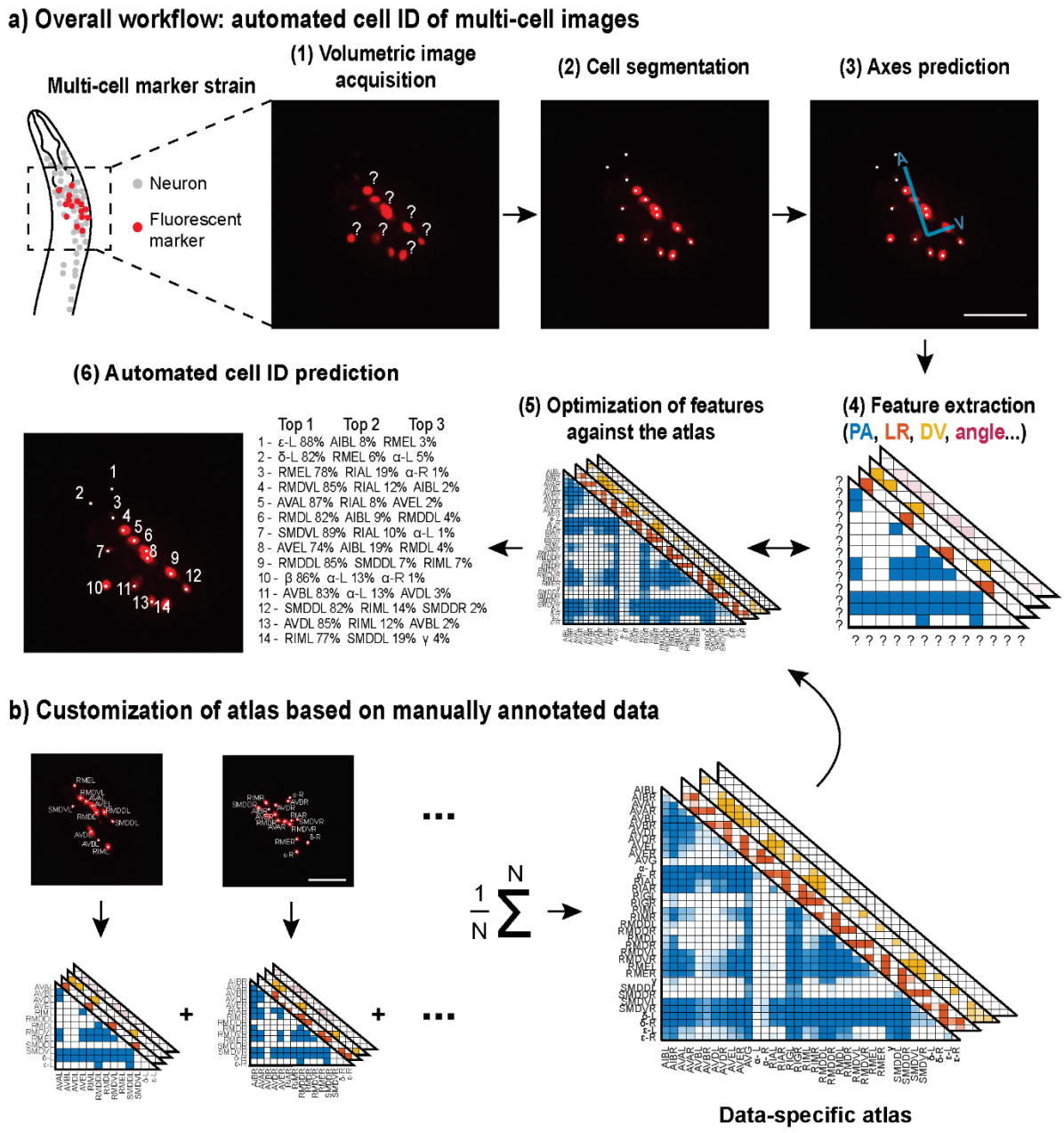

Figure 1 CRF_ID for multi-cell images. a) Computational workflow starting from image acquisition to final cell identity predictions. a-1,2,3) Image preprocessing steps include automatic cell segmentation and coordinate axes prediction. a-4) Feature variables that represent positional relationships of the cells are extracted (PA,

posterior and anterior; LR, left and right; DV, dorsal and ventral). a-5) The CRF algorithm maximizes the similarities between the extracted features from the images and those from an atlas. a-6) The final results are represented as a list of most likely neuron candidates for each cell with predicted probabilities. b) The atlas can be customized to meet the specifications of the images, and this is easily done by compiling and averaging annotated data.

New features improve prediction accuracy of body axes

While CRF_ID performs very well with whole-brain image datasets and has a great potential for generalization, multi-cell imaging poses challenges that require the algorithm to consider additional features in the data to accurately predict neuron identities. One such challenge is in the prediction of body axes from the volumetric images. It is important to correctly assign the three-dimensional coordinate axes (anterior-posterior, left-right, dorsal-ventral) for each worm because it is a method to standardize neuronal positions of worms imaged in various orientations. However, the axes assignment is not a trivial task because the worms, especially those in microfluidic devices, can deviate from its naturalistic orientation and be at an angle as large as 20 degrees from the plane (Figure 2- figure supplement).

In our previous work, principal components analysis (PCA) was employed on point clouds of head neurons segmented from fluorescence volumes (Chaudhary et al., 2021). Since PCA finds orthogonal dimensions that explain the most variance in the point cloud, the first three principal components would correspond to the coordinate axes of the worm, assuming the point cloud adequately represents the worm head shape and radial asymmetry of the nervous system. However, using PCA on cell point clouds to predict the coordinate axes is not suitable for multi-cell images. In whole-brain images, nearly all head neurons are fluorescent, which means the point cloud of the neurons is a fair representation of the worm's overall head shape. In contrast, multi-cell images have a smaller number of fluorescent cells, whose locations may not properly sample the space of the whole brain. For instance, if the fluorescent cells are concentrated near the ventral side of the head, the resulting anterior-poster axis would gravitate towards the ventral side, deviating from the ground truth.

To address the axes prediction challenge in multi-cell images, we have amended the method to be less dependent on the point cloud of cell centroids, which varies depending on which neurons are expressing the fluorophore. The new coordinate assignment method takes advantage of two common features in almost all samples - that the worm is auto-fluorescent, and that many neuronal pairs are bilaterally symmetric in their anatomical positions. It involves two correction steps (Figure 2). The first step corrects the AP axis by incorporating auto-fluorescence signals as natural landmarks to enlarge the point cloud (Figure 2a). This is easily implemented by imaging a volume in the green

channel, where the auto-fluorescence is prominent, and segmenting the fluorescent signals as points using the same cell segmentation method. The new point cloud then reflects the overall shape of the head, and the resulting AP axis from PCA aligns correctly along the head of the animal. The second step corrects the LR axis, for which we have implemented an algorithm that searches for the best plane of bilateral symmetry (Figure 2a). Using the initial LR axis as the starting point and the orthogonality to the AP axis as a constraint, the algorithm iteratively finds planes within a range and computes a symmetry score of the point cloud with respect to each plane. The plane that results in the highest symmetry score is assigned as the final LR axis. The DV axis is automatically determined by orthogonality to the first two axes.

In order to quantitatively evaluate the axes prediction performance, we manually defined "ground truth" axes for each volume and calculated the angle deviations of the predicted axes (Figure 2a). Compared to the axes predicted by PCA on cell point clouds, the new axes that have been corrected by the two-step method all showed decreased deviations from the manually defined axes. More than 90% of the corrected axes were within 10 degrees from the ground truth, which was comparable to the standard deviation of manual annotations in defining the ground truth. More importantly, the axes correction led to a significant improvement in the accuracy of the neuron ID prediction, measured using correspondence to human annotations (Figure 2c). Also, there was no significant difference between the corrected axes and the manually defined axes in terms of the resulting neuron ID correspondence; this again indicates that the automatically predicted axes are comparable to those defined by human. The details of the quantification of neuron ID accuracy are discussed in methods under the section title CRF_ID 2.0: Evaluation of accuracy.

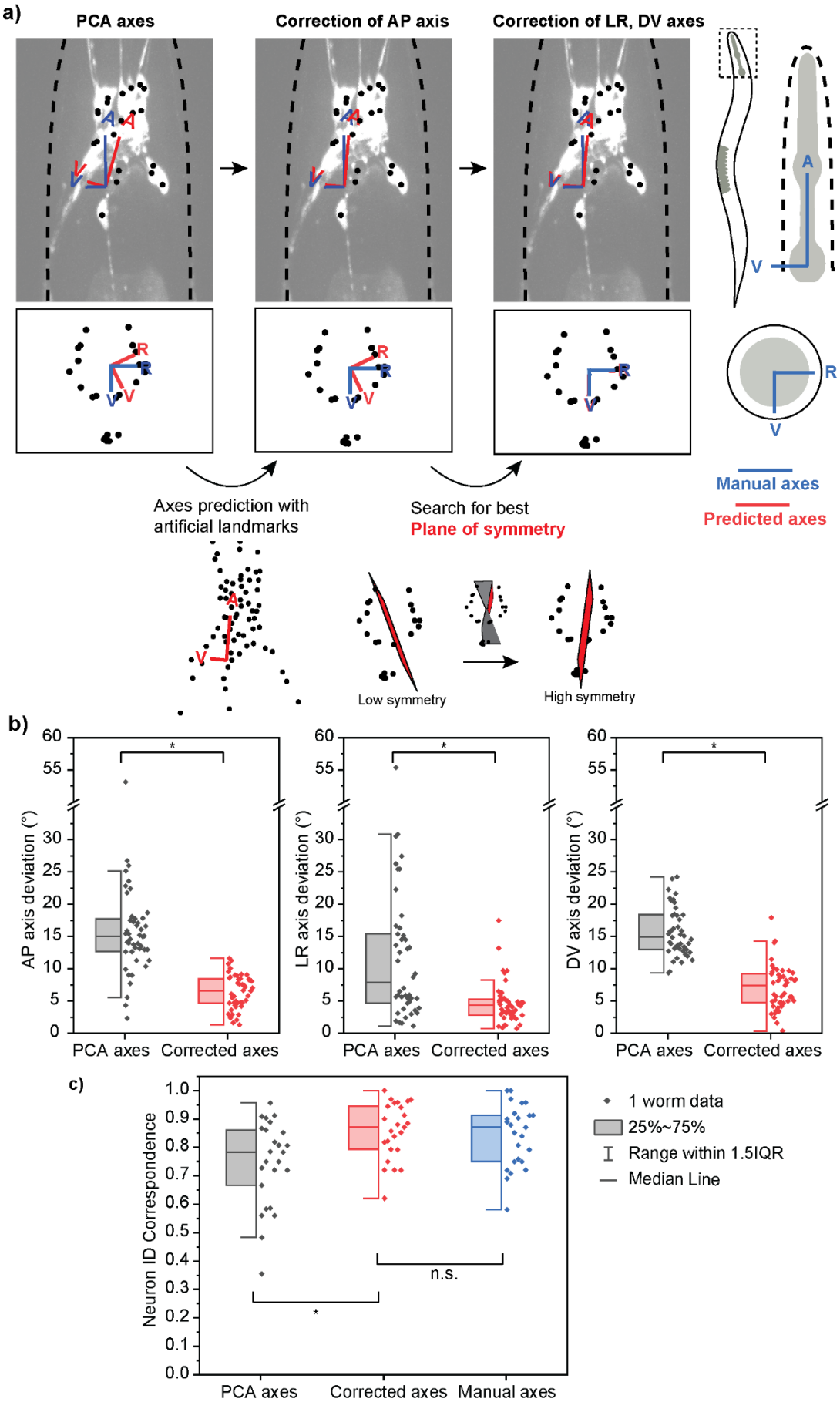
a)
PCA axes
Correction of AP axis
Correction of LR, DV axes
A
V
R
Axes prediction with artificial landmarks
Search for best
Plane of symmetry
Manual axes
Predicted axes
Low symmetry
High symmetry
b)
AP axis deviation (°)
LR axis deviation (°)
DV axis deviation (°)
PCA axes
Corrected axes
c)
Neuron ID Correspondence
PCA axes
Corrected axes
Manual axes
n.s.
1 worm data
25%~75%
Range within 1.5IQR
Median Line

Figure 2 Improved method of assigning coordinate axes. a) coordinate axes for multi-cell images generated by PCA alone are not accurate. A two-step correction process is implemented: correction of the AP axis by using natural landmarks and correction of LR, DV axes by searching for the best plane of symmetry. b) The corrected axes are more accurate than the previous axes generated by PCA alone as they show decreased angle deviations from the ground truth axes for all three coordinate axes. c) Corrected axes result in a higher and comparable neuron ID accuracy (correspondence to manual cell annotations) when compared with PCA predicted axes and ground truth axes, respectively. Best single prediction results are reported. Two sample t-tests were performed for statistical analysis. The asterisk symbol denotes a significance level of $p<0.05$.

The atlas's data-specificity is important for high neuron-identification accuracy

One of the most important requirements for accurate neuron labeling using CRF_ID 2.0 is the availability of an accurate atlas, which serves as a reference map of the stereotypical and probable positions of the neurons in the animal. To characterize the extent to which atlases influence the accuracy of neuron identification and to provide practical guidance on which atlas to use, we evaluated the performance of CRF_ID 2.0 for several possible atlases. In our previous work, we demonstrated that for whole-brain images, a data-driven atlas results in higher prediction accuracy (Chaudhary et al., 2021). We tested whether the same holds true for multi-cell applications. We characterized several different atlases for predicting neuron identities in worms expressing *glr-1p::NLS-mcherry-NLS* transgene (Figure 3A-B). The first atlas is derived from electron microscopy data in the OpenWorm project, which provides the three-dimensional coordinates of neurons in a model adult hermaphrodite *C. elegans* (Szigeti et al., 2014). The second atlas is the NeuroPAL atlas, which is the OpenWorm atlas updated with 9 imaging data of the NeuroPAL strain (Yemini et al., 2021), and this atlas was reported in our previous work (Chaudhary et al., 2021). The other atlases were derived from fluorescence imaging data of *glr-1p::NLS-mcherry-NLS* strain, the same strain of neuron ID interest. Several different versions of the *glr-1* atlas containing different numbers of data sets were created to characterize the effect of the number of datasets in the atlas.

We examined the effect of data source in atlas performance. Three factors were observed to be most important in determining an atlas's accuracy: strain specificity, mode of data acquisition or imaging conditions, and the number of datasets in the atlas. The neuron annotation accuracy was lowest with the OpenWorm atlas, which scores the poorest in all three factors (Figure 3c). This is likely due to the fact that the OpenWorm data are derived from a strain of different genotypical background from the *glr-1p::NLS-mcherry-NLS* strain, and more importantly, data acquisition from electron microscopy and the fluorescence volumetric imaging distort the anatomy differently. In addition, the OpenWorm atlas is based on a single dataset, which does not capture the variability of neuronal positions. A slightly higher accuracy was achieved by using the atlas built on NeuroPAL data (Figure 3c). Note that the genotype of the model training data (the NeuroPAL set) and that of

the test set (*glr-1p::NLS-mcherry-NLS*) are still different. The difference between the *glr-1p::NLS-mcherry-NLS* and NeuroPAL strains is significant because, in addition to possible anatomical differences from genetic make-up, there are neuron pairs in the *glr-1p::NLS-mcherry-NLS* strain that were not updated in the NeuroPAL atlas due to variable expressions of the transgenes used in the NeuroPAL strain and the difficulty of manually annotating all the neurons in the whole-brain images. In fact, only 58% of the pairwise relationships of neurons expressing the *glr-1p::NLS-mcherry-NLS* transgene were augmented with the NeuroPAL data, and the rest were assigned the default values from the OpenWorm atlas. Further, the imaging conditions for the training and test sets are not entirely comparable because the NeuroPAL images were acquired with a lower z resolution. Such a difference in the imaging condition can lead to differences in segmentation of cell centroids and affect the angular relationships between cells, and thus lowering the accuracy of using the atlas for cell ID prediction. Unlike the OpenWorm atlas, however, the NeuroPAL atlas contains a statistical distribution of neuronal positions from 9 datasets, which provides a more accurate representation of the neuronal positions than that from a single dataset (Corsi et al., 2015). Therefore, while in the absence of any data-driven reference atlas, standard atlas(es) can be used as a starting point, the more strain-specific datasets used to correct and augment the reference atlas, the more accurate the neuron identification prediction would be.

The highest neuron identification correspondence was found with the data-driven atlas, derived from the *glr-1p::NLS-mcherry-NLS* strain, the same strain that is of interest (Figure 3c). The high accuracy can be attributed to the fact that both the strain (thus presumably the anatomy) and the imaging conditions are matching the test dataset. It is notable that even the *glr-1* atlas that is derived from a single dataset performed better than the NeuroPAL atlas containing 9 datasets (Figure 3c). This implies that the matching strain type and the imaging conditions play a more important role than the sheer number of datasets in the quality of the atlas measured by the cell ID correspondence.

We also examined the effect of sample size in atlas building, in addition to the matching strain type and imaging conditions. Atlases should be derived from sufficiently large sample size to capture the variability within the dataset. Because the neuronal positions are highly variable, an accurate atlas should contain data from a sufficient number of training samples to account for the positional variability of the neurons. As seen with the OpenWorm atlas, an atlas based on only one sample does not contain any statistical information on positional variability, so it performs poorly against testing samples whose neuronal positions do not match those in the atlas well. Figure 3C demonstrates that the average correspondence increased with the number of ground-truth datasets used to construct the atlas. While the *glr-1* atlas constructed using 25 datasets had the highest

overall accuracy, the *glr-1* atlases containing 5-10 datasets are statistically indistinguishable in their performance. We observe that the saturation of information is achieved at 5-10 datasets for the *glr-1* case, given that the select datasets exhibit reliable gene expressions to provide statistically good sample sizes for all neuron candidates. This indicates that an atlas derived from 10 well-curated datasets may be sufficient for CRF_ID 2.0. In general, the performance of atlas models would depend on the natural variabilities of the neuronal anatomy and experimental noises, best determined empirically for each strain.

We also compared the differences of the atlases against the best performing atlas (the *glr-1* atlas derived from 25 datasets) as a benchmark. We found that the results correlate well with the trend observed for neuron correspondence (Figure 3c-e). The neuron ID correspondence increased with similarity to the *glr-1* atlas with 25 datasets, as defined by the smaller differences in the angular (Figure 3d) and PA, LR, DV relationships (Figure 3e). Interestingly NeuroPAL atlas displayed the highest difference in angular relationship; this is likely due to the difference in the imaging condition for NeuroPAL, in which the fluorescent images were down-sampled in the z direction (Yemini et al., Cell 2021). This would result in the neuronal locations to be more discretized along the z axis, which can distort the angular relationships more than the binary relationships. Overall, the results demonstrate that for optimal CRF_ID 2.0 accuracy, it is important to use the atlas derived from data specific to the subject of interest for neuron identification.

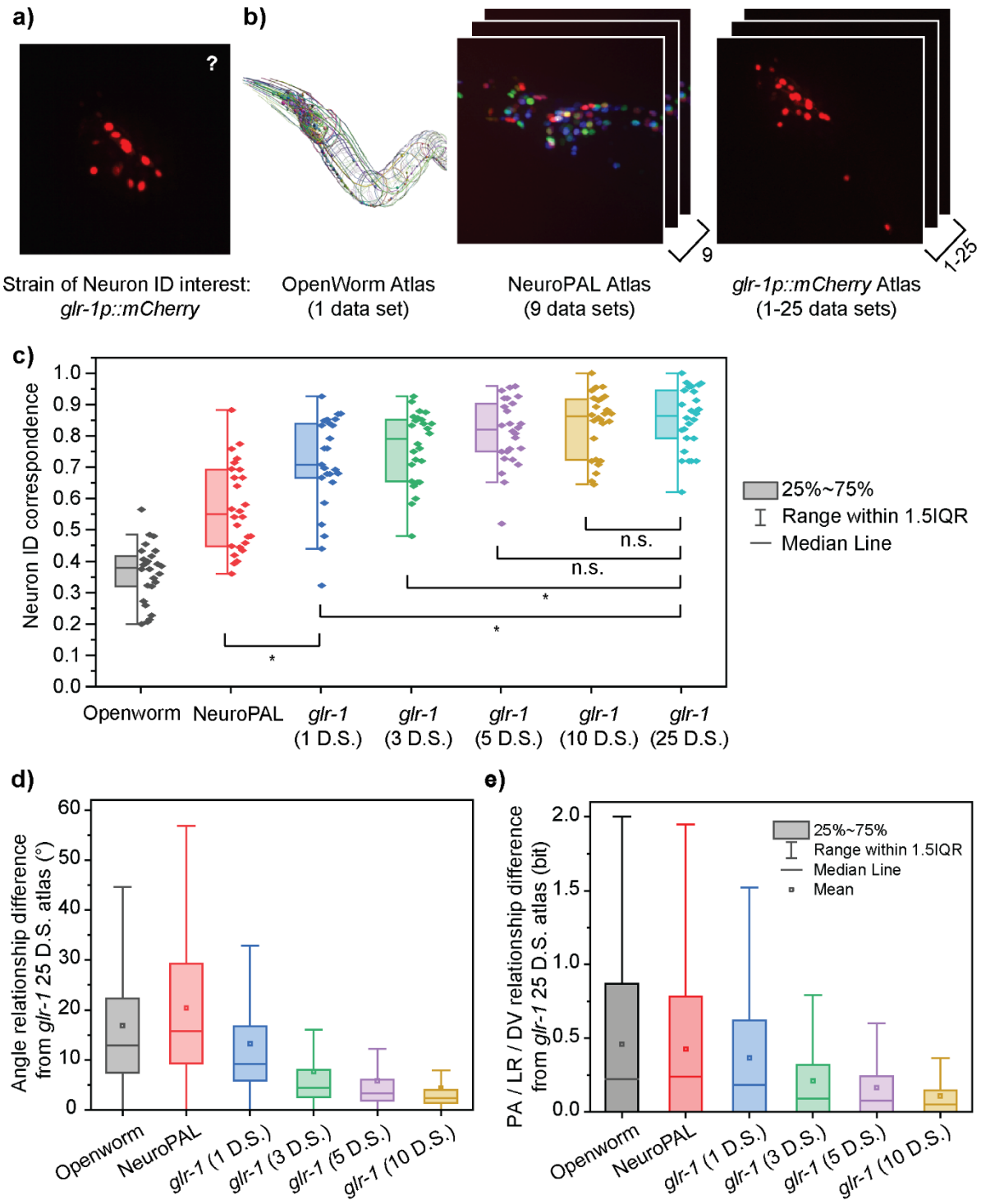

Figure 3 Characterizing the importance of data-specific atlases. a,b) Several example atlases (b) are compared on their performance in neuron ID prediction on the *glr-1p::NLS-mcherry-NLS* multi-cell images (a). c, d) Difference of each atlas from the most accurate available atlas (*glr-1* 25 datasets) in terms of pairwise angle relationships (c) and PA/ LR/ DV positional relationships (d). e) The neuron ID accuracy (correspondence to manual cell annotations) depends greatly on the atlas used. Each data point represents the cell cluster from

one animal (n=26). Best single prediction results are reported. Two sample t-tests were performed for statistical analysis. The asterisk symbol denotes a significance level of $p<0.05$.

The automated cell annotation accuracy is comparable to manual cell annotation accuracy

In evaluating the performance of CRF_ID 2.0 as an automated cell annotation method, the most important criterion is the accuracy of prediction. It should be noted that we defined accuracy as the correspondence to human annotations (which is how cell ID has been traditionally done), while being cognizant that human annotators do not always provide the absolute ground truth. For this study, the "ground truth" was established as the cell labels from the consensus of three annotators, which means that the ground truth label of a cell is one that has been agreed by at least two annotators. Although there were slight inconsistencies in labels among the three annotators, they had high degrees of correspondence with each other with an average of at least 80% (Figure 4-figure supplement). The annotations from all three annotators were used as training data, but the accuracy did not decrease to a statistically significant level even when an individual annotator's data were used as the training data (Figure 4b). It should be noted that all accuracies reported in this work represent leave-one-out cross validation, in which the test sample is excluded from the training set.

Because the absolute ground truth labels are often unknown and unattainable even for human annotators, our algorithm provides ranked multiple alternative labels in addition to the best single prediction. These top ranked labels are generated by iteratively running the annotation algorithm with a randomized set of a specific number of neurons in the candidate list removed, similar to our prior approach with building the whole-brain atlas (Chaudhary et al., 2021). "Top 3 accuracy" characterizes the fraction of cells with top three predictions that include the ground truth label. Note that the top 1 prediction may be different from the best single prediction, which is the result of running the annotation algorithm once with the full set of candidate neurons. The accuracy for the best single prediction mode was around 85%, and those for top 1, 2, 3 predictions were around 85%, 94%, and 96% respectively (Figure 4a). This indicates that in cases where there is appreciable uncertainty of the candidate neuron set, the algorithm can assist the users to decide the final label of a cell by narrowing the candidates and knowing the correct label is almost certainly among the three predictions.

To assess whether the accuracy is acceptable, we compared the aforementioned CRF_ID 2.0 accuracies against the accuracies of human annotators (Figure 4a). Annotator 1 had the lowest accuracy that is statically comparable to the best single prediction and top-1 prediction results from CRF_ID 2.0. Annotators 2 and 3 had higher accuracies, resembling the distributions for top-2 and

top-3 prediction results, respectively. Thus, the automated neuron identification using CRF_ID does not come at a loss in accuracy. Moreover, we observed that the incorrect neuron identifications are not entirely unreasonable. When the accuracies are examined per neuron basis, there was a good correlation between the automated and manual neuron ID accuracies. The neurons that were more incorrectly predicted by the CRF_ID 2.0 algorithm were more likely the ones on which human annotators disagreed with each other (Figure 4c). On the other hand, "easy" neurons for the annotators were also more likely to be easy for the CRF_ID 2.0 algorithm to predict them correctly.

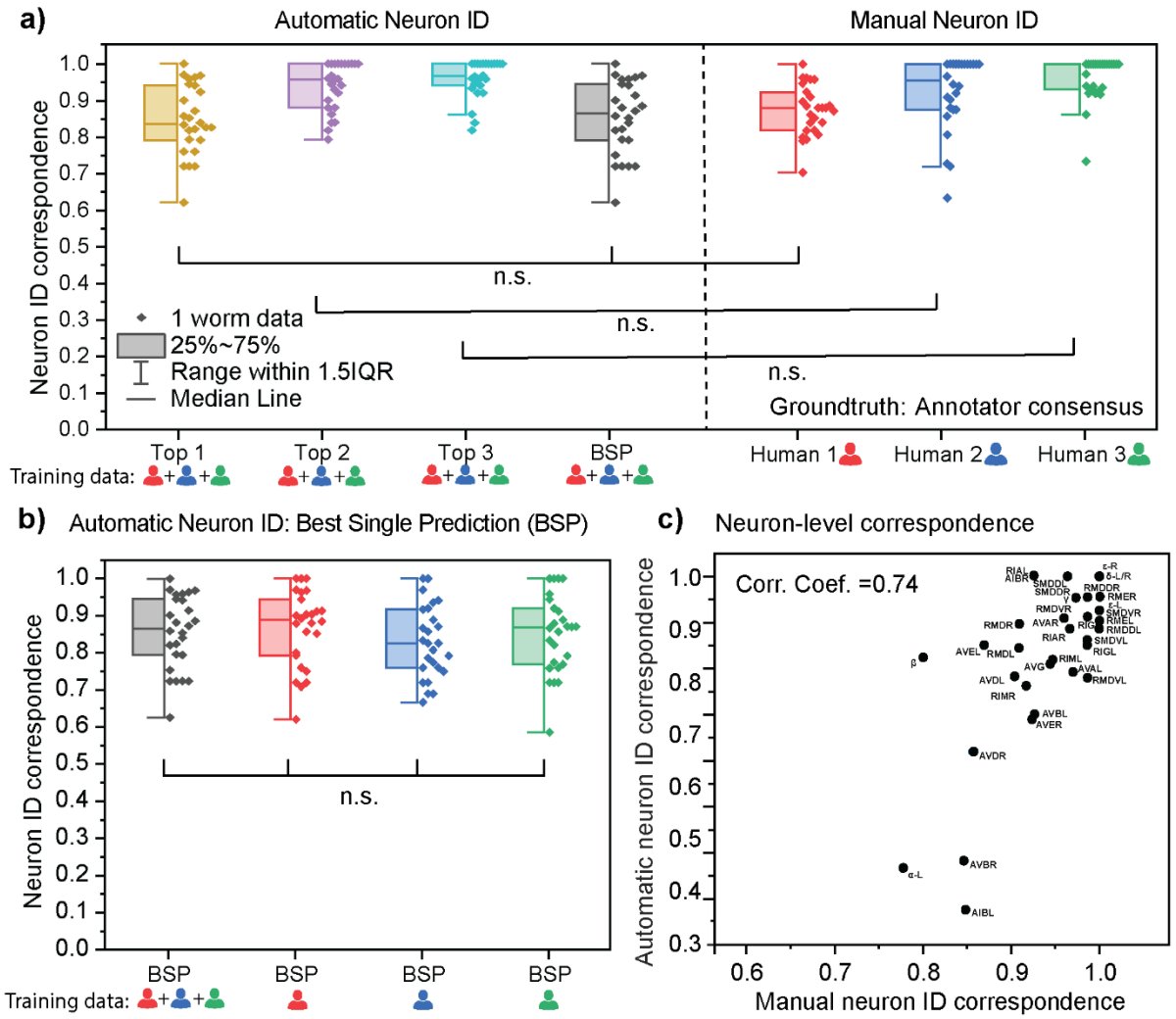

Figure 4. Characterization of neuron identification accuracy using CRF_ID. a) Side by side comparison of automated neuron ID accuracy and manual neuron ID accuracy. Each datapoint represents the cell cluster from an animal. For automated neuron ID, top 1, 2, 3 results are from an iterative operation of the CRF_ID algorithm, and the best single prediction (BSP) results are from a single run. The atlas is a compiled data from three different annotators. The ground truth labels are defined by the consensus of the three annotators. b) No significant differences in best single prediction accuracies are found when using atlases derived from data annotated by different annotators. c) There is a positive correlation between the automatic and manual neuron

ID accuracy of each neuron. Two sample t-tests were performed for statistical analysis. The asterisk symbol denotes a significance level of $p<0.05$.

The multi-cell neuron ID is useful for in-vivo gene expression analysis

To further characterize our multi-cell neuron identification method, we applied it to a problem of biological interest. Although the method has potential uses in any application that requires cells to be identified in a cell population, including calcium imaging (Ji et al., 2021) (Nguyen et al., 2016) and cell lineage tracing (Bao et al., 2006), we demonstrate its application in gene expression analysis as an example. Knowing the expression of a particular gene is important for understanding the genetic basis of neuronal function, and it can be studied by examining the expression of its mRNA (Spencer et al., 2011; Taylor et al., 2021), transcriptional or translational reporters (Choi et al., 2020; Kuroyanagi et al., 2010). While CeNGEN's mRNA expression data may be an accurate reference for gene expression (Taylor et al., 2021), fluorescent reporters allow one to monitor gene expression *in vivo*, and on individual basis. This will enable studies on changes in gene expression due to perturbations or experimental conditions.

In this type of applications, the need for neuron identification emerges when the gene of interest is expressed in multiple neurons. For such cases, manually annotating the neurons is difficult and time-consuming, and this difficulty has often prompted researchers to abandon the neuron-specificity of gene expression by measuring the collective expression of all fluorescence (Richman et al., 2018; Sánchez-Blanco and Kim, 2011). However, neuron-specific gene expression is more informative and facilitates the elucidation of neuronal functions by connecting the ongoing studies with existing knowledge on specific neurons. For this reason, we applied CRF_ID 2.0 to aid the neuron identification in multi-cell gene expression analysis.

We studied the neuron-specific expression of the *glr-1* gene. *glr-1* encodes a homolog of the mammalian AMPA type glutamate-gated ionotropic receptor subunits GluA1 and GluA2 (Brockie et al., 2001) which play important roles in neural plasticity, learning and memory (Henley and Wilkinson, 2016). *glr-1* is known to be expressed in AVA, AVE, AVD, RMDD/V, RMD, RIM, and SMDD/V among other neurons (Maricq et al., 1995). We imaged a *C. elegans* strain that has two types of transcriptional reporters, an integrated transgene (*glr-1p::gfp*) and an extrachromosomal transgene with a nuclear-localized sequence (*glr-1p::NLS-mCherry*) (Figure 5a). In general, the expression of an integrated transgene is more robust and more reflective of the physiological expression because, unlike the expression of an extrachromosomal transgene, it is more resistant to genetic mosaicism, in which a transgene is not inherited to some neurons during differentiation due to mitotic instability of the extrachromosomal DNA (Frøkjær-Jensen et al., 2008).

To quantify expression, we next segmented the neurons from the mCherry image stacks. Because of nuclear localization, the fluorophore signals are more resolved between cells. We then applied CRF_ID 2.0 to the multi-cell point cloud to assign cell identities (Figure 5a). The neuron identification results were provided as top three candidate labels for each cell for the user to decide on the final labels. There were about 20-35 fluorescent cells in each volume. The fluorescence intensities of paired neurons were pooled together and then re-grouped into two: one on the brighter side closer to the objective and one on the dimmer side farther from the objective. Figure 5b and 5c only report the fluorescent intensities of neurons on the brighter side (the side closer to the imaging objective), but the dimmer side showed a very similar profile (Figure 5- figure supplement).

The gene expression analysis revealed several insights. First, we did not observe a significant difference in relative expression trends between extrachromosomal (Figure 5b) and integrated transgenes (Figure 5c). For example, the neurons that had high expression levels for the integrated transgene, such as AVA, RMDV, RMD, SMDV, SMDD, and AVB, also exhibited high expression levels for the extrachromosomal transgene. More specifically, there was a linear correlation between GFP and mCherry intensities (Figure 5- figure supplement 2). Second, the extrachromosomal transgene expression did not have particularly more variable expression levels. This suggests that gene expression studies with extrachromosomal transcriptional reporters, which do not require the time-consuming process of gene integration, may still be able to provide a robust and meaningful picture of the gene expression. Third, gene expressions in neurons were found to be highly variable among individual samples, which was generally expected considering the dynamic nature of gene expressions. The neurons with high average expression appeared to have a wider dynamic range.

Interestingly, the mRNA profiling data generated by CeNGEN do not correlate well with the results from transgene expressions. In the plots in Figure 5, The neuron labels on the x axis are listed in the order of decreasing expression level according to CeNGEN (Taylor et al., 2021). The low correlation is probably due to several reasons. First, there is an innate difference between mRNA expression and transcriptional reporter signals, with the reporters more indicative of the expression of the fusion proteins. Second, the promoter sequence in the transgene, while carefully chosen, could still be different from the native *cis*-regulatory sequence that could include sequences residing in the intron regions or downstream regions all of which may regulate mRNA expression measured by sequencing. Lastly, the cell dissociation process needed for profiling mRNAs of individual neurons may impact gene expression.

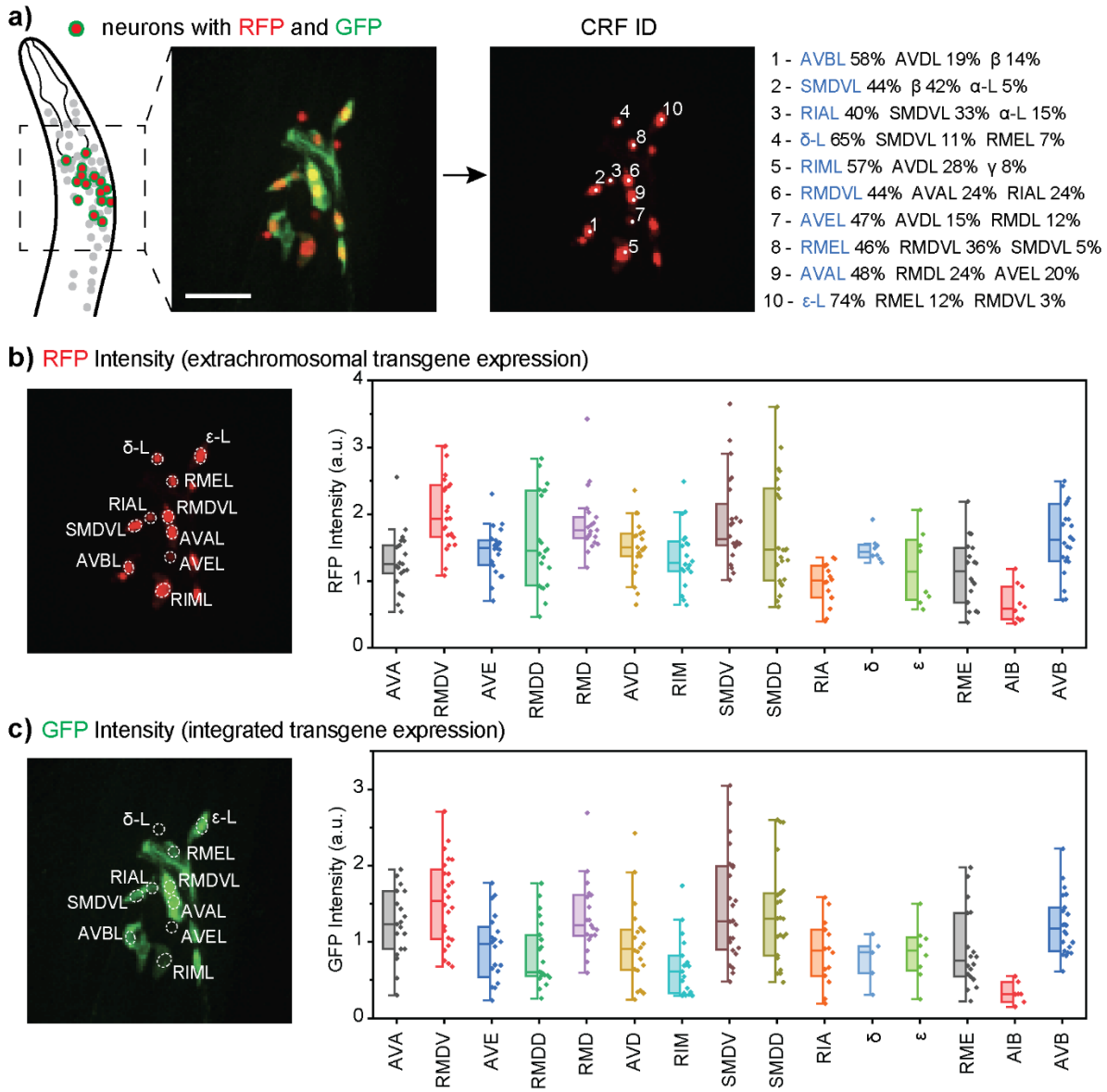

Figure 5. Multi-cell neuron identification in in-vivo gene expression analysis. a) CRF_ID 2.0 facilitates multi-cell annotation by providing top 3 most likely neuron labels for each cell, from which the user makes the final decision. b,c) The example strain contained extrachromosomal (c) and integrated transcriptional reporters for *glr-1* (c). Plotted are the neuron-specific gene expression levels displayed as the normalized fluorescence intensities of identified neurons. The neurons labels on the x axis are listed in the descending order of single-cell RNA sequencing expression levels reported by CeNGEN. Box plots indicate median, quartiles and whiskers indicate 1.5 IQR. Data points indicate signals from individual worms.

Discussion

In this work, we have upgraded and characterized the original CRF_ID method to accommodate

common needs to analyze more diverse types of biological images. While the original method demonstrated high annotation accuracy in *C. elegans* whole-brain images, its performance in multi-cell images, which include all images of neuronal groups that are not whole-brain, was not fully warranted. The modified axes assignment method accurately predicts the coordinate axes of *C. elegans* volumetric images regardless of the number of cells expressed and thereby greatly enhances cell identification success. Our characterization of the CRF_ID 2.0 performance in comparison with the atlas and human annotators offers an important benchmark for users interested in using CRF_ID 2.0 to annotate multi-cell images. In addition, we demonstrate its application in transgenic reporter-based gene expression analysis, which is multi-cell expression by nature. This work represents a practical advance, with updated features better suited for multi-cell applications, and still applicable to whole-brain images.

One of the distinguishing advantages of CRF_ID 2.0 is its flexibility, which stems from its modular architecture. Compared with registration or deep learning-based methods, in which it is difficult or impossible to adjust the optimization process, the graphical-model approach of CRF_ID allows heuristic-based deliberate selection and tuning of the features. This aspect opens the door for many other use cases. For example, for images with distinctive cellular characteristics, the user may add new unary features for characteristics, such as the cell shape and signal intensity, to be optimized. For imaging conditions for which certain features become less reliable, the user may choose to reduce or remove the weights for those features. Further, the simplicity of building the atlas is another important aspect that enables the incorporation of new images of different strains and imaging conditions into these data-driven custom atlases. Whereas the deep learning method would require high-performance computing and hours of training time, our atlases can be built or updated from existing atlases within seconds with basic arithmetical operations. As such, CRF_ID can be easily applied to biological images of various origins, not restricted to *C. elegans* that display stereotypical cellular characteristics.

Materials and Methods

Strains

Adult *C. elegans* hermaphrodites were used in this study and were cultured using standard procedures (Brenner, 1974). Two *C. elegans* transgenic strains were used. All data reported in Figures 1-4, including the training data for the *glr-1* atlases, are from images of ZC3292 *yxEx1701*[*glr-1p::GCaMP6s, glr-1p::NLS-mCherry-NLS*]. The gene expression data in Figure 5 are from images of ZC3292 and ZC3612 *lin-15B&lin-15A(n765) kyIs30[glr-1::GFP, lin-15(+)] X; yxEx1932[glr-1p::NLS-mCherry-NLS].*

Construction of transgenes and transgenic strains

glr-1p::GCaMP6s was generated by LR recombination (NEB) of a destination vector containing DNA sequence encoding GCaMP6s (Chen et al., 2013) and *unc-54* 3′UTR with an entry vector *glr-1p*-PCR8 that contained a sequence of 5331 bases upstream of *glr-1* coding region. *glr-1p::NLS-mCherry-NLS* was generated by LR recombination of the *glr-1p*-PCR8 entry vector with a destination vector that contained a mCherry sequence flanked by two NLSs (nuclear localization sequences) followed by *unc-54* 3′UTR constructed using Gibson assembly (Thermo Fisher). Each plasmid was injected at 10ng/μL to generate ZC3292, and glr-1p::NLS-mCherry-NLS was injected at 40ng/μL to generate ZC3612 using standard methods of microinjection (Mello et al., 1991).

Imaging data collection

Imaging of the ZC3292 strain was performed using an Andor Spinning Disk confocal system on a Nikon Eclipse Ti-E inverted microscope using a 40× oil-immersion objective (N.A.=1.3) and images were recorded using an ANDOR iXon Ultra EMCCD camera. Hermaphrodites at day 1 adult stage were used and were immobilized with sodiam azide in the chamber of a microfluidic device (Chronis et al., 2007) controlled by an AutoMate Scientific ValveBank perfusion system (Berkeley, CA). Volumetric images of the head region were acquired in both green (laser: 488nm; filter 525nm/50nm) and red channels (laser: 561nm; filter: 617nm/73nm) with a Z step size of 0.3μm and XY resolution of 0.4μm. The exposure time is 20ms for both green channel and red channel.

Imaging of the strain ZC3612 was performed using a Nikon W1 spinning disk confocal system on Nikon Ti2-E microscope with Hamamatsu ORCA-Fusion Gen-III sCMOS camera. A 60x objective lens with N.A. 1.4 was used. The animals were age-synchronized to day 1 adult stage and chemically immobilized in 20mM tetramisole. In order to efficiently image straight-headed animals, we loaded the animals into an array type microfluidic device (Lee et al., 2014). 3D stacks of the animal's head region in red (laser: 561nm; filter: T605/52m) and green (laser: 488nm; filter: ET525/36m) channels were acquired with a Z step size of 0.3μm. The exposure times were 10ms and 50ms for red and green channels respectively. The XY resolution of the images was 0.12μm.

Manual annotation and atlas construction

Building an atlas requires manually annotated datasets. Three participants separately segmented and annotated cells on raw 3D stacks of 26 worm data. Most of the annotations were done by visually comparing the cell point cloud against reference images. The references include the anatomical features of head neurons, including the positions of the cell bodies and the shape of the neuronal processes, on the WormAtlas website (Hall et al., 2008) as well as 3D representation

of the neuron coordinates on OpenWorm (Szigeti et al., 2014). 3D reconstruction of fluorescent neuronal images was also used to determine cell identity. The list of candidate neurons expressing *glr-1* was carefully curated based on known expression patterns [(Hart et al., 1995; Maricq et al., 1995) and Taylor et al., Cell 2021] and by analyzing the expression patterns in our own data. The list of annotated neurons is as follows: AIBL/R, AVAL/R, AVBL/R, AVDL/R, AVEL/R, AVG, α-L/R, β, RIAL/R, RIGL/R, RIML/R, RMDL/R, RMDDL/R, RMEL/R, γ, RMDVL/R, SMDDL/R, SMDVL/R, δ-L/R, ε-L/R. Neuronal expressions with low confidence have been indicated in Greek symbols; we cautiously mention α, β, γ, δ, ε may correspond to AVJ, M1, RIS, URYD, and URYV, respectively. The data-driven atlas was constructed using the atlas generation codes included in the original CRF_ID. Instead of building an atlas from the consensus labels of three annotators, the labels from different annotators were considered as separate annotations, effectively capturing the statistics of the information from the data set instead of using majority-vote single label.

CRF_ID 2.0: Improved axes prediction.

The new axes prediction method consists of two parts: AP axis correction and LR/DV axes correction. For AP axis correction, the point cloud was artificially enlarged by including naturally fluorescent landmarks in the animal's body. The autofluorescence signals were segmented by thresholding local maxima from the image after Gaussian filtering. The threshold value of 99.85 was determined experimentally and can be tuned for different imaging conditions. Then, PCA was applied to the point cloud to find the AP axis and the initial LR and DV axes.

The algorithm for LR/DV axes correction takes the initial LR and DV axes and cell point cloud as inputs and outputs the corrected LR and DV axes. It searches for the plane/axes pair that divides the point cloud into two sides with the highest symmetry across the plane. First, it finds a range of planes that satisfy the following constraints: orthogonality to the AP axis, within the angle of 20 degrees from the initial axis, and an increment of 1 degree. The initial axis is the LR axis that was generated by PCA in the previous step. By iteratively testing each plane, the one that results in highest symmetry was found. Symmetry was quantified as the inverse of the deviation from perfect symmetry. The deviation was calculated by reflecting each point with respect to the plane and calculating the distance of the reflected point to the closest neighboring point. The distance was calculated for all points in the point cloud and the average of the lowest 7 distances was used for the given plane. The threshold value of 7 was derived from the expectation of at least 7 left/right pairs of neurons in the *glr-1* strain, but this threshold can be tuned depending on the characteristics of the strain or the images. Higher symmetry with respect to a plane would result in a lower distance value. The plane that resulted in the lowest distance was assigned the final LR axis, and the DV axis

is automatically determined by orthogonality to the first two axes. The average run time for each point cloud was under 1 min.

CRF_ID 2.0: Evaluation of accuracy

Because the absolute ground truth neuron labels for the cells are not available, we defined the annotation accuracy as correspondence to the consensus labels of three human annotations. The consensus was established as the label that has been agreed by at least two annotators. Cells whose identities were differently annotated by all three annotators, which account for less than 3% of all cells, were omitted from the accuracy calculation. Also, all reported correspondence values are results of leave-one-out cross validations, meaning the atlas used to test the accuracy on a specific worm did not include that specific data set. Thus, there are 26 different versions of the *glr-1* 25 dataset-atlas, one for each dataset exclusion. For example, in Figure 3a, a correspondence value of 0.8 indicates that 80% of the cells in the dataset were assigned the neuron label that matches the consensus label, and the atlas used is driven from manual annotations of 25 other worm datasets. The correspondences reported in Figures 2 and 3 are from results of running the prediction algorithm once in the best single prediction mode. Figure 4 additionally reports top 1, 2, 3 results from 100 iterations while randomly removing neuron labels from the candidate list.

Gene expression analysis

To extract the cell-specific fluorescent signals, we first used the automatic segmentation tool to segment the cells in a total of 27 image stacks of ZC3612. The segmentation was done on the nuclear-localized mCherry signals. The quality of the segmentation was visually inspected to eliminate false positives. Then, iterative neuron ID predictions were performed on the segmented red channel images. The resulting top 3 candidates were reviewed for each cell, and the final neuron label was selected based on human judgment. Then, a series of data processing steps were necessary to best represent the cell-specific gene expression data from different animals. First, we extracted the mCherry and GFP signal intensities by averaging the intensities of the brightest 100 pixels in the segmented masks, which had around 500 pixels on average. Because mCherry and GFP expressions are driven by separate transgenes, mCherry expression does not guarantee GFP expression. Thus, GFP expressions that were deemed absent were eliminated during data curation. Second, because the expression levels are variable among animals, the fluorescence intensities were normalized for each animal by dividing by the average intensity of all neurons in the animal. Lastly, we compared the intensities of cells only on the same side of the animal because the side of the animal closer to the objective lens was generally brighter due to light scattering through the biological tissues. The reported values in Figure 5 are the intensities on the brighter side of the

animal.

Statistical analysis

Statistical analyses of the data were performed using Paired Comparisons App in OriginPro 2020. The asterisk symbol denotes a significance level of $p<0.05$. Not significantly different comparisons are denoted non-significant (n.s.).

Code and data availability

CRF_ID 2.0 can be accessed at https://github.com/lu-lab/CRF-Cell-ID-2.0. This repository contains all components of the framework and the atlases produced and compared in this work.

Acknowledgments

The authors acknowledge the funding support of the U.S. NIH (R01MH130064, R01NS115484) to HL and YZ and the U.S. NSF (1764406) to HL. Some nematode strains used in this work were provided by the Caenorhabditis Genetics Center (CGC), which is funded by the NIH (P40 OD010440), National Center for Research Resources and the International *C. elegans* Knockout Consortium.

Supplementary Figures

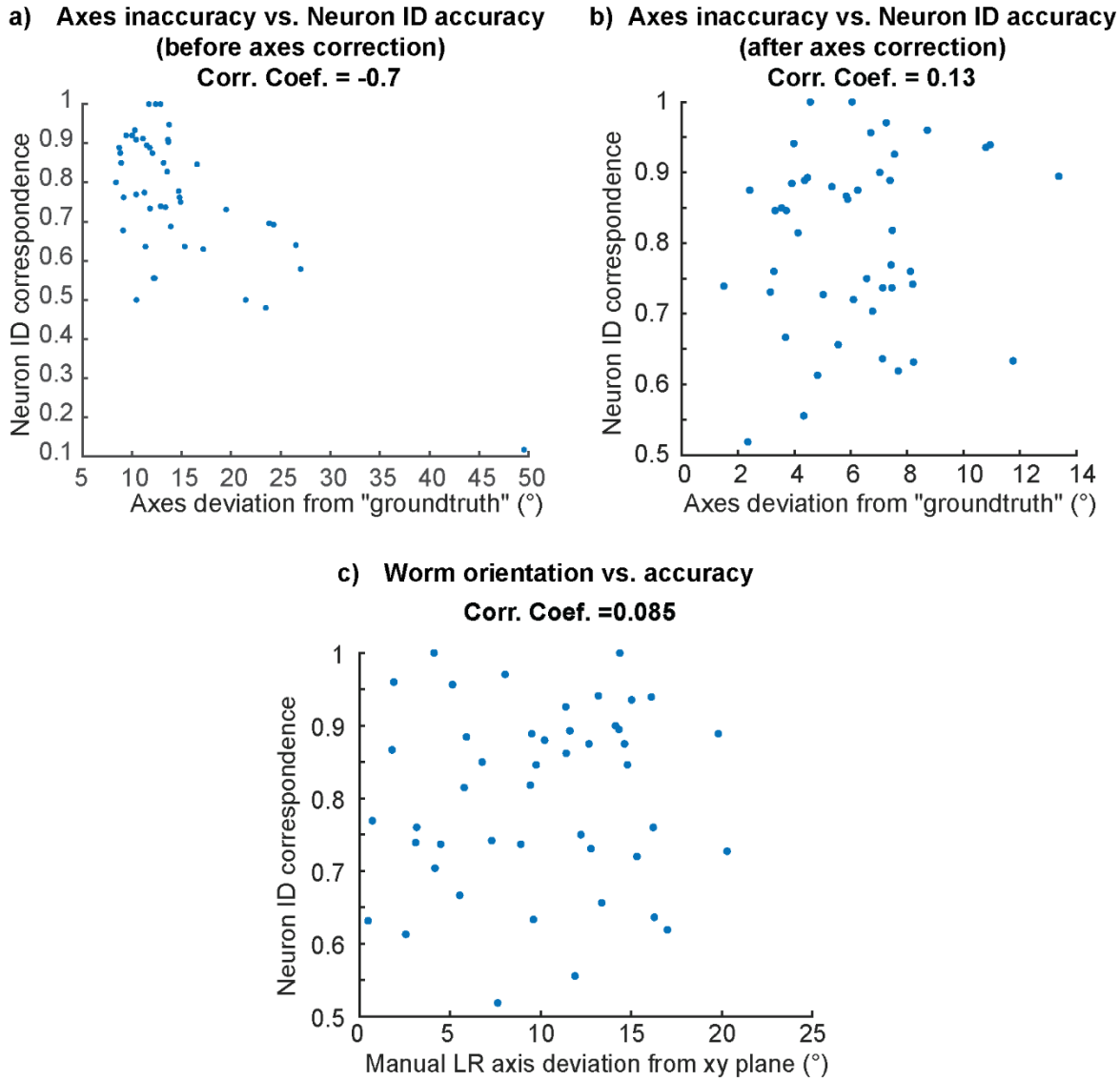

Figure 2- figure supplement. Neuron ID accuracy no longer depends on the axes inaccuracy after axes correction. a) High negative correlation between axes inaccuracy and neuron ID accuracy before axes correction. b) No correlation between axes inaccuracy and neuron ID accuracy after axes correction. c). No correlation between worm orientation and neuron ID accuracy.

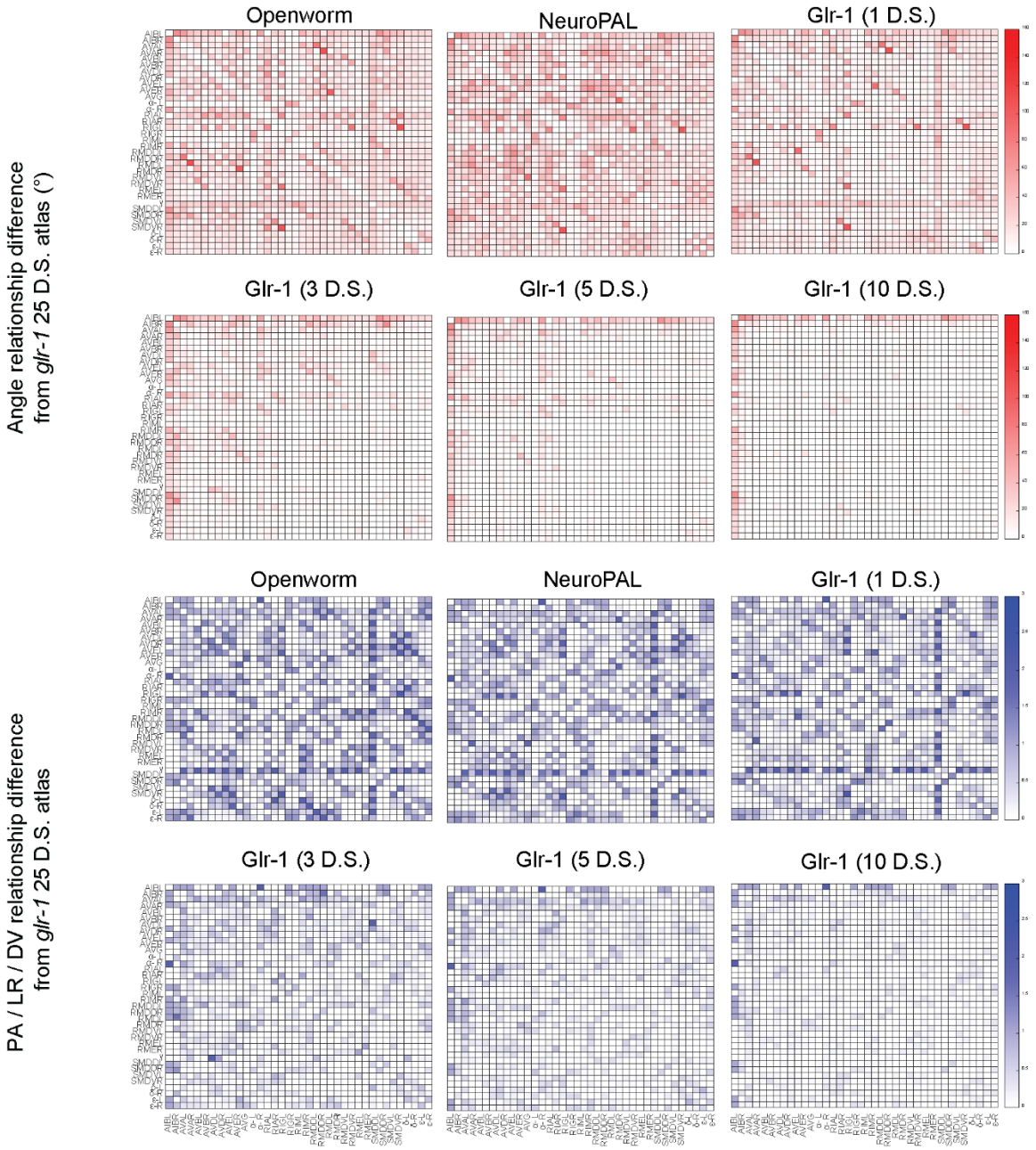

Figure 3- figure supplement. A more detailed visual representation of the difference of each atlas from the best available atlas (*glr-1* from 25 datasets). a) differences in angular relationships.b) differences in PA/LR/DV relationships.

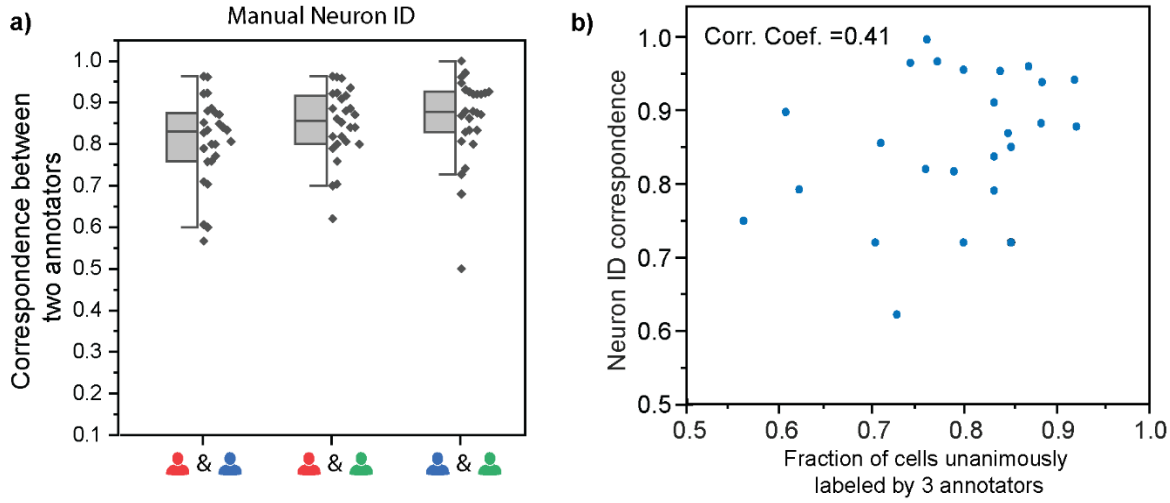

Figure 4- figure supplement. a) correspondence between any two among the three annotators. b) Slight correlation between the fraction of cells unanimously labeled by 3 annotators and neuron ID correspondence.

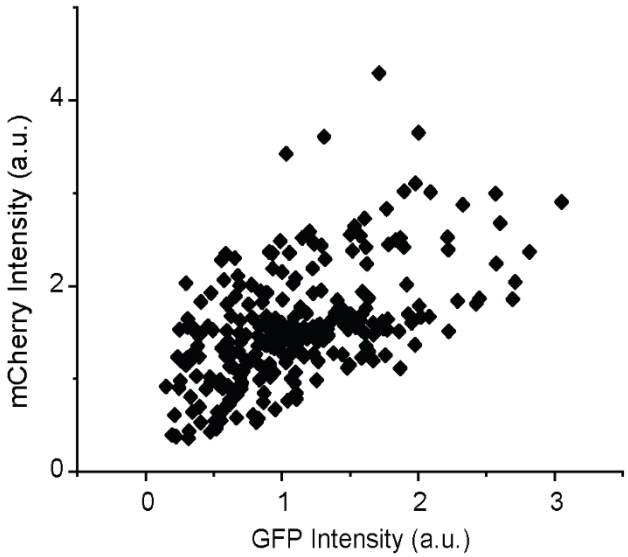

Figure 5- figure supplement 2. High correlation between extrachromosomal (mCherry) and integrated (GFP) transgene expressions.

References

Bao Z, Murray JI, Boyle T, Ooi SL, Sandel MJ, Waterston RH. 2006. Automated cell lineage tracing in Caenorhabditis elegans. *Proc National Acad Sci* 103:2707–2712. doi:10.1073/pnas.0511111103

Brenner S. 1974. THE GENETICS OF CAENORHABDITIS ELEGANS. *Genetics* 77:71–94. doi:10.1093/genetics/77.1.71

Brockie PJ, Madsen DM, Zheng Y, Mellem J, Maricq AV. 2001. Differential Expression of Glutamate Receptor Subunits in the Nervous System of Caenorhabditis elegans and Their Regulation by the Homeodomain Protein UNC-42. *J Neurosci* 21:1510–1522. doi:10.1523/jneurosci.21-05-01510.2001

Chaudhary S, Lee SA, Li Y, Patel DS, Lu H. 2021. Graphical-model framework for automated annotation of cell identities in dense cellular images. *Elife* 10:e60321. doi:10.7554/elife.60321

Chen T-W, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, Schreiter ER, Kerr RA, Orger MB, Jayaraman V, Looger LL, Svoboda K, Kim DS. 2013. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 499:295 300. doi:10.1038/nature12354

Choi M-K, Liu H, Wu T, Yang W, Zhang Y. 2020. NMDAR-mediated modulation of gap junction circuit regulates olfactory learning in C. elegans. *Nat Commun* 11:3467. doi:10.1038/s41467-020-17218-0

Chronis N, Zimmer M, Bargmann CI. 2007. Microfluidics for in vivo imaging of neuronal and behavioral activity in Caenorhabditis elegans. *Nat Methods* 4:727–731. doi:10.1038/nmeth1075

Corsi AK, Wightman B, Chalfie M. 2015. A Transparent window into biology: A primer on Caenorhabditis elegans. *Wormbook* 1–31. doi:10.1895/wormbook.1.177.1

Frøkjær-Jensen C, Davis MW, Hopkins CE, Newman BJ, Thummel JM, Olesen S-P, Grunnet M, Jorgensen EM. 2008. Single-copy insertion of transgenes in Caenorhabditis elegans. *Nat Genet* 40:1375–1383. doi:10.1038/ng.248

Gray JM, Hill JJ, Bargmann CI. 2005. A circuit for navigation in Caenorhabditis elegans. *P Natl Acad Sci Usa* 102:3184–3191. doi:10.1073/pnas.0409009101

Hall DH, Crocker C, Norris CR, Herndon LA, Lints R, Altun ZF. 2008. Teaching Nematode Anatomy Online: WormAtlas and Slidable Worm. *Faseb J* 22:769.10-769.10. doi:10.1096/fasebj.22.1_supplement.769.10

Hart AC, Sims S, Kaplan JM. 1995. Synaptic code for sensory modalities revealed by C. elegans GLR-1 glutamate receptor. *Nature* 378:82–85. doi:10.1038/378082a0

Henley JM, Wilkinson KA. 2016. Synaptic AMPA receptor composition in development, plasticity and disease. *Nat Rev Neurosci* 17:337–350. doi:10.1038/nrn.2016.37

Ji N, Madan GK, Fabre GI, Dayan A, Baker CM, Kramer TS, Nwabudike I, Flavell SW. 2021. A neural circuit for flexible control of persistent behavioral states. *Elife* 10. doi:10.7554/elife.62889

Kuroyanagi H, Ohno G, Sakane H, Maruoka H, Hagiwara M. 2010. Visualization and genetic analysis of alternative splicing regulation in vivo using fluorescence reporters in transgenic Caenorhabditis elegans. *Nat Protoc* 5:1495–1517. doi:10.1038/nprot.2010.107

Lafferty J, McCallum A, Pereira FCN. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML '01 Proc Eighteenth Int Conf Mach Learn*, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data 282–289.

Lee H, Kim SA, Coakley S, Mugno P, Hammarlund M, Hilliard MA, Lu H. 2014. A multi-channel device for high-density target-selective stimulation and long-term monitoring of cells and subcellular features in C. elegans. *Lab Chip* 14:4513–4522. doi:10.1039/c4lc00789a

Lin A, Qin S, Casademunt H, Wu M, Hung W, Cain G, Tan NZ, Valenzuela R, Lesanpezeshki L, Pehlevan C, Venkatachalam V, Zhen M, Samuel ADT. 2022. Functional imaging and quantification of multi-neuronal olfactory responses in C. elegans. *Biorxiv* 2022.05.27.493772. doi:10.1101/2022.05.27.493772

Liu H, Zhang Y. 2020. What can a worm learn in a bacteria-rich habitat? *J Neurogenet* 34:1–9. doi:10.1080/01677063.2020.1829614

Maricq AV, Peckol E, Driscoll M, Bargmann CI. 1995. Mechanosensory signalling in C. elegans mediated by the GLR-1 glutamate receptor. *Nature* 378:78 81. doi:10.1038/378078a0

Mello CC, Kramer JM, Stinchcomb D, Ambros V. 1991. Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. *Embo J* 10:3959–3970. doi:10.1002/j.1460-2075.1991.tb04966.x

Richman C, Rashid S, Prashar S, Mishra R, Selvaganapathy PR, Gupta BP. 2018. C. elegans MANF Homolog Is Necessary for the Protection of Dopaminergic Neurons and ER Unfolded Protein Response. *Front Neurosci-switz* 12:544. doi:10.3389/fnins.2018.00544

Sánchez-Blanco A, Kim SK. 2011. Variable Pathogenicity Determines Individual Lifespan in Caenorhabditis elegans. *Plos Genet* 7:e1002047. doi:10.1371/journal.pgen.1002047

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, Reinke V, Petrella L, Strome S, Stetina SEV, Katz M, Shaham S, Rätsch G, Miller DM. 2011. A spatial and temporal map of C. elegans gene expression. *Genome Res* 21:325–341. doi:10.1101/gr.114595.110

Szigeti B, Gleeson P, Vella M, Khayrulin S, Palyanov A, Hokanson J, Currie M, Cantarelli M, Idili G, Larson S. 2014. OpenWorm: an open-science approach to modeling Caenorhabditis elegans. *Front Comput Neurosc* 8:137. doi:10.3389/fncom.2014.00137

Taylor SR, Santpere G, Weinreb A, Barrett A, Reilly MB, Xu C, Varol E, Oikonomou P, Glenwinkel L, McWhirter R, Poff A, Basavaraju M, Rafi I, Yemini E, Cook SJ, Abrams A, Vidal B, Cros C, Tavazoie S, Sestan N, Hammarlund M, Hobert O, Miller DM. 2021. Molecular topography of an entire nervous system. *Cell* 184:4329-4347.e23. doi:10.1016/j.cell.2021.06.023

Toyoshima Y, Wu S, Kanamori M, Sato H, Jang MS, Oe S, Murakami Y, Teramoto T, Park C, Iwasaki Y, Ishihara T, Yoshida R, Iino Y. 2020. Neuron ID dataset facilitates neuronal annotation for whole-brain activity imaging of C. elegans. *Bmc Biol* 18:30. doi:10.1186/s12915-020-0745-2

Xu C, Jackson SA. 2019. Machine learning and complex biological data. *Genome Biol* 20:76. doi:10.1186/s13059-019-1689-0

Yemini E, Lin A, Nejatbakhsh A, Varol E, Sun R, Mena GE, Samuel ADT, Paninski L, Venkatachalam V, Hobert O. 2021. NeuroPAL: A Multicolor Atlas for Whole-Brain Neuronal Identification in C. elegans. *Cell* 184:272-288 e11. doi:10.1016/j.cell.2020.12.012

Yu X, Creamer MS, Randi F, Sharma AK, Linderman SW, Leifer AM. 2021. Fast deep neural correspondence for tracking and identifying neurons in C. elegans using semi-synthetic training. *Elife* 10:e66410. doi:10.7554/elife.66410