Please also see the latest version of the repository: https://doi.org/10.5281/zenodo.6374011 and our website: https://ilandavis.com/jcb2023-yfp

Multi-Dimensional Data Viewer (MDV) for data Exploration: "Systematic analysis of YFP traps reveals common discordance between mRNA and protein across the nervous system"

ABSTRACT

The explosion in the volume of biological imaging data challenges the available technologies for data interrogation and its intersection with related published bioinformatics data sets. Moreover, intersection of highly rich and complex datasets from different sources provided as flat csv files requires advanced informatics skills, which is time consuming and not accessible to all. Here, we provide a "user manual" to our new paradigm for systematically filtering and analysing a dataset with more than 1300 microscopy data figures using Multi-Dimensional Viewer (MDV) -link, a solution for interactive multimodal data visualisation and exploration. The primary data we use are derived from our published systematic analysis of 200 YFP traps reveals common discordance between mRNA and protein across the nervous system (eprint link). This manual provides the raw image data together with the expert annotations of the mRNA and protein distribution as well as associated bioinformatics data. We provide an explanation, with specific examples, of how to use MDV to make the multiple data types interoperable and explore them together. We also provide the open-source python code (github link) used to annotate the figures, which could be adapted to any other kind of data annotation task.

INTRODUCTION

Here, we provide instructions on how to use our microscopy image-based *Drosophila* genome collection in MDV. We explain the structure of the collection and the associated metadata, annotations describing the location of mRNA and protein, linked expression and bioinformatics data and GO annotations, including molecular function and diseases. These rich datasets underpin and complement our publication describing the biological significance of the data [1]. As the collection contains over thirteen hundred microscopy images visualising mRNA and protein expression of 200 genes across the larval nervous system of the fruit fly (*D. melanogaster*), the data cannot all be presented within the limited space of a scientific journal article, even within supplementary figures. Instead, we use MDV to display all the imaging data in the form 7 figures for each of 200 genes, representing the expression in different compartments of the nervous system. The 7 compartments are: Mushroom Body (MB), Central Brain (CB), Optic Lobe (OL), Ventral Nerve Cord (VNC), Neuromuscular Junction (NMJ), Nerve and a lower magnification Overview of the larval brain. We also include high-quality annotations of the microscopy images from three different experts that assess the protein and mRNA distribution of each gene in each compartment. We wrote and used bespoke Python software (link) to minimise errors in the annotations, allow resolution of conflicts and collate the data in a mineable format. Both the images and the expert annotations are displayed within MDV in an interoperable form along with other expression data, functional and molecular information as well as structural information, derived from FlyBase [2], Flymine [3] and ENSEMBL [4]. In this manner, the user is easily able to use the annotations and other information to filter the entire data set and display a subset of the genes and figures. MDV allows the user to browse all the classes of data simultaneously in

the context of the unique findings of our research paper (link), while considering the context of that information within known genetic and molecular data from the literature.

MDV

Multi-Dimensional Viewer is a tool for analysing, annotating and sharing multidimensional data. The cloud-based system uses a large assortment of interactive charts and widgets to enable the user to browse and interrogate large datasets, with user-friendly methods to query and filter the data. Multi-dimensional and multi-modal data can be displayed and links between related data established. All the charts and widgets can be displayed in separate windows allowing for multiple screen use. Crucially, data, charts and widgets can all be added by the user, allowing a flexibility to what data is displayed and in what manner. We have provided a number of different prearranged charts and widgets in a list available in a convenient pulldown menu (top left of screen). These focus on different aspects of the data, ranging from GO annotations of molecular functions or disease to the annotation of mRNA and protein distribution to length of the 3'UTR. The charts are interactive, so that for example, the user can choose to display only genes of a certain 3'UTR length with a slider, or only genes associated with a particular disease by clicking on a specific GO term (see details below). Finally, the user can also create their own charts and widgets, displaying whatever particular aspects of the data that are of interest to them, as they see fit.

HOW TO PREPARE AND LOAD DATA INTO MDV

The input file for MDV is a basic CSV file where each row corresponds to an image file, and each column is a data feature associated with that image. For our dataset, each of the images was acquired from a tissue specimen that had a specific gene labelled with a marker protein (see Titlow et al., 2023 for experimental details). The columns contain a combination of annotations describing key experimental analyses of the marker protein's expression in different cells, the fly line stock number, as well as gene features drawn from published studies, e.g., Gene Ontology, physical characteristics of the gene, and mRNA expression levels in the brain. We have also added columns that contain active links to databases with additional information about the gene, e.g., FlyBase, Gene2Function, and our OMERO database that provides access to the raw image data. From a simple .csv file (example .csv is provided in the Zenodo repository containing this manual), MDV can display and provide access to all of this information from a single web browser. The MDV source code can be accessed here along with more information about MDV.

Manually creating a data table with thousands of rows and hundreds of columns would be extremely time consuming and error prone. We have developed a python workflow to automatically generate a .csv file from a list of gene names or image file IDs (link). This modular workflow uses a combination of reference data files and database queries to autopopulate over 50 data features for each gene that was investigated in our study, and can also export image files directly from an OMERO database.

EXPLORING OUR MDV COLLECTION

Our collection has preloaded charts for the user to explore, separated into four 'views', each with a different theme of information displayed:

- Main – This view is centred around the interrogation of the data by filtering based on our annotations of the locations of RNA and protein within the different regions of the fly nervous system.
- Genetic Analysis – This view allows the user to browse the collection using associated Gene Ontology (GO) terms for Biological Processes, Cellular Components and Molecular Functions
- Disease Association – Here the user can explore the collection in relation to with associations 173 known disease ontologies.
- Gene Expression and Structure – The user can browse the data based on tissue expression data and gene structure such as the length of the 3'UTR extension

While these views are ready to browse immediately, MDV allows the user to create new charts and widgets to display any aspect of the imaging and other data (as defined in a CSV file associated with the collection) that is of interest to them. However, the general user does not have the authority to permanently save the views, as this is a public database.

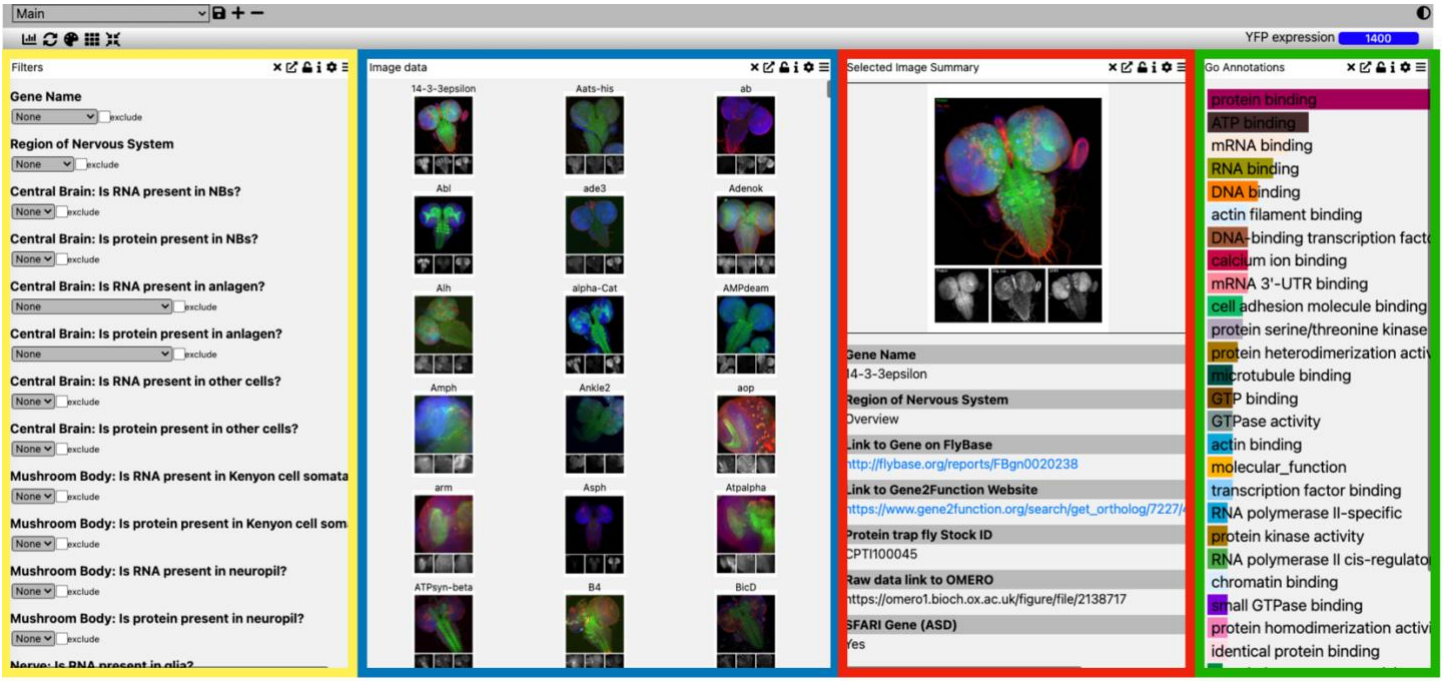
HOW TO INTERROGATE OUR DATA

You can access our publicly available collection by clicking (link). The user is presented with 'Main', the first of the four 'views' as described above. The reference screenshot (Fig 1) shows the default 'Main' view as would be seen on first opening the collection. As MDV charts and widgets are adjustable by the user it is possible to change the appearance and content of each view, but for the purposes of this guide we shall use the default layout of each view. There are four pre-loaded 'charts' in this view:

- Filters (yellow box, Fig. 1): This series of dropdown menus allows the user to filter the data according to the presence or absence of RNA and/or protein, as annotated during the project that generated the images. Selection of one or more of these filters will cause the images displayed in other charts of the view to reflect those selections
- Image Data (blue box, Fig. 1): This chart displays the image data that corresponds to selections made in the other charts and widgets on this view. The default image data displays all 1400 figures in the collection. (NOTE: As there are 7 figures for every gene screened, 1400 figures represent data from 200 genes, 70 figures represent data from 10 genes etc.). Selecting any image by clicking on it displays that data in the 'Selected Image Summary' chart. Users can highlight different categories within a dataset as well change the image size in this window by accessing the chart settings.
- Selected Image Summary (red box Fig. 1): Any image selected from the 'Image Data' chart will display here, along with selected other data relevant to that gene including links to the raw data and other databases such as FlyBase.
- GO Annotations: (green box Fig. 1): This chart displays the top 35 Gene Ontology (GO) annotations associated with the selected dataset. This information is dynamically linked, so clicking on the bars for any of these categories will filter the dataset to display only images pertaining to that GO category.

PREVIOUS VERSION OF THE DATABASE

The first version of our database (Link to Zenodo version 1) used a commercial software platform called Zegami to visualise the multimodal data. Zegami is no longer available and have now moved the data to a fully open source and supported platform called Multi-Dimensional Viewer (MDV). MDV combines images and data and is a major improvement in



the way the data can be interpreted from a biological point of view. Therefore, we recommend using MDV as resource for exploring these data.

Figure 1: the default 'Main' view of our collection. The four charts are 'Filters' (highlighted in yellow), 'Image Data' (highlighted in blue), 'Selected Image Summary' (highlighted in red) and 'GO Annotations' (highlighted in green).

The other three views of our collection display different aspects of the data from our dataset that may be of interest to the user. The 'Genetic Analysis' view (Fig. 2A) allows the user to select imaging data by filtering using three GO categories, namely GO Biological Process, GO Cellular Component and GO Molecular function. Selection of any of these GO terms on these charts causes the 'Image Data' chart to only display the images relevant to that selection. Multiple selections across these charts performs a Boolean 'AND' function, thus displaying only those images that are contained in all of the categories selected. An example of this is shown in Fig. 2B when the GO Biological Process of 'Open tracheal system' is selected. It should be noted that not only has the image data that is displayed changed to the relevant figures that intersect with this selection, but the other GO term charts have similarly changed. In the GO Cellular Component and Go Molecular Function charts, the options have been reduced to only those terms that can be selected along with 'Open tracheal system'.

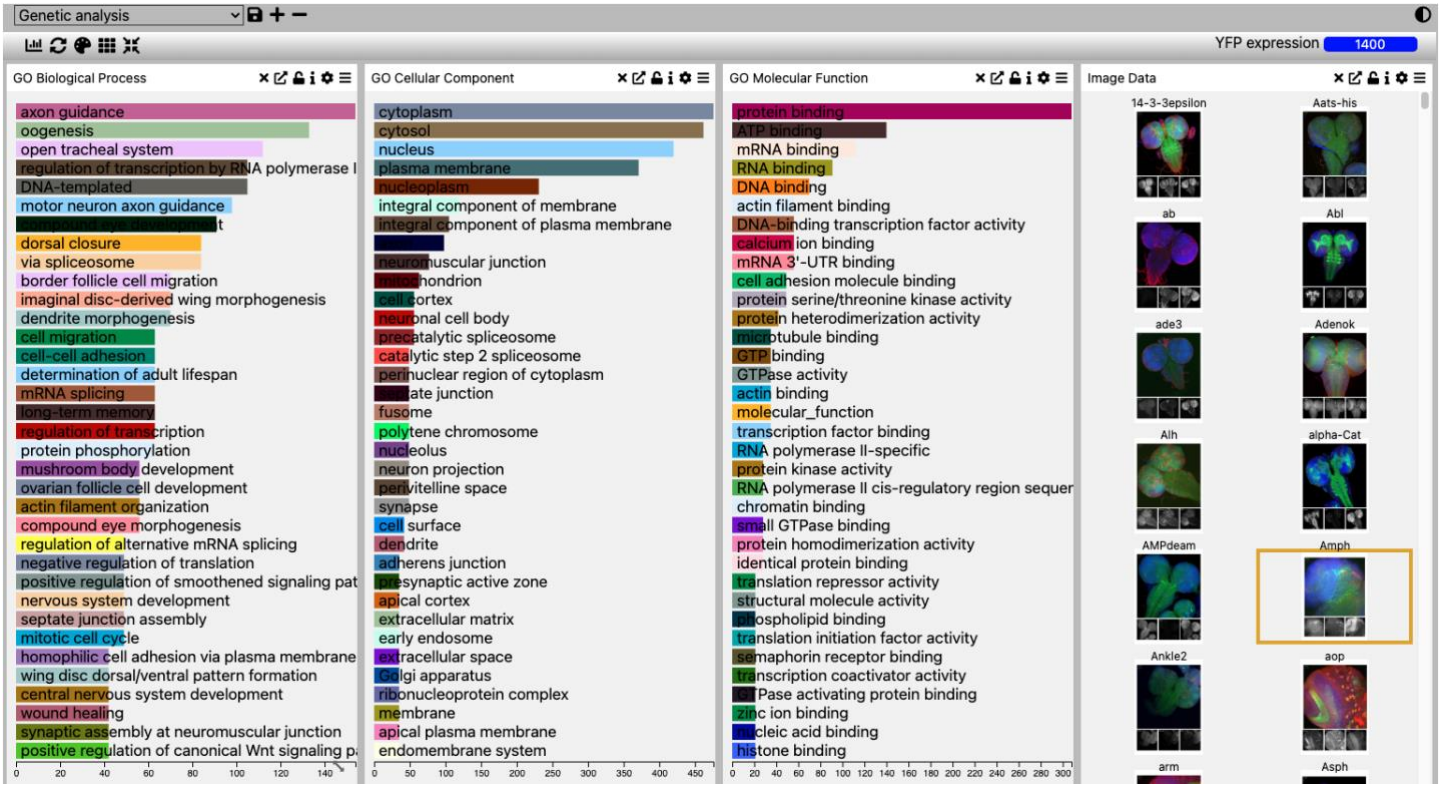


Figure 2A: The Default ‘Genetic Analysis’ view.

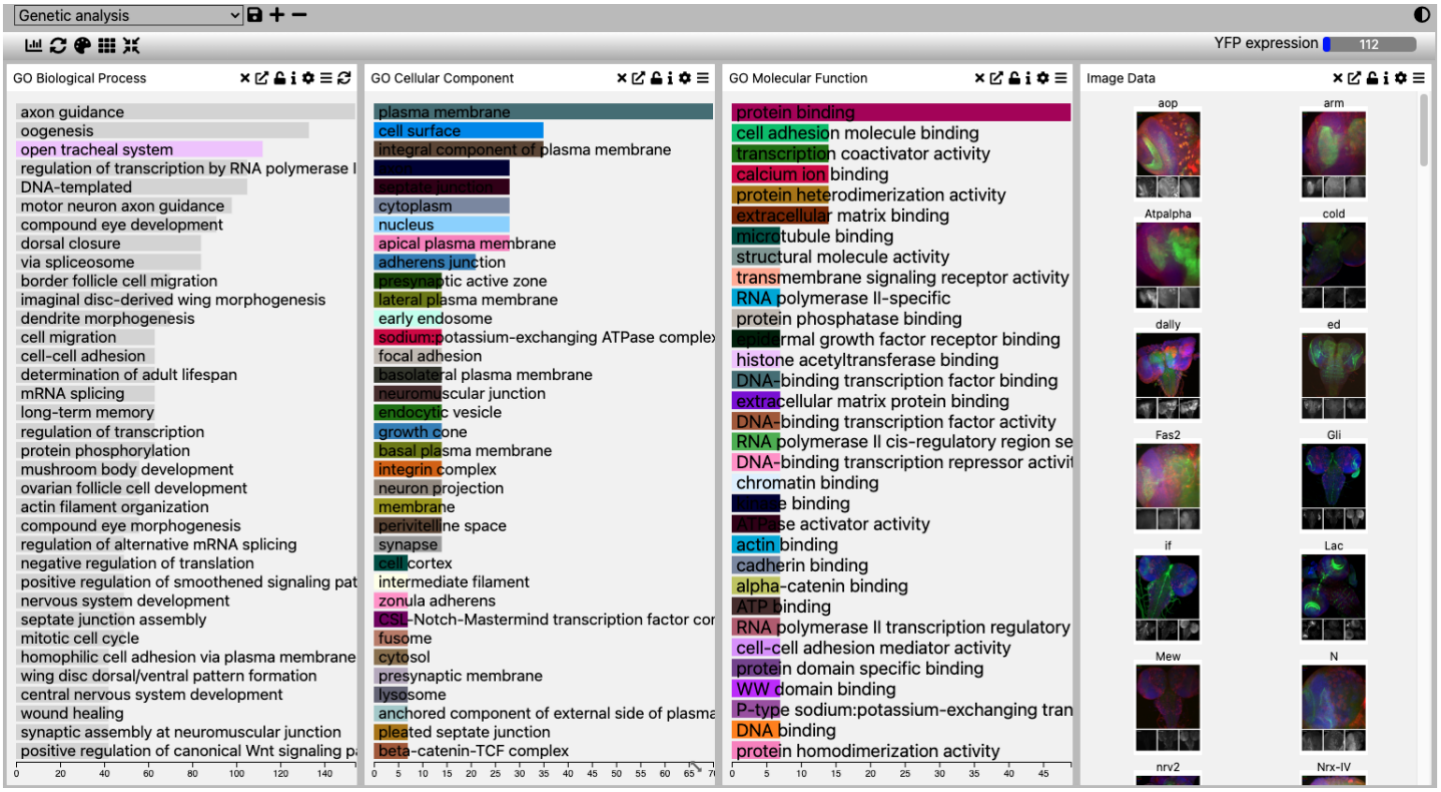


Figure 2B: ‘Genetic Analysis’ view once ‘Open tracheal system’ has been selected in the ‘GO Biological Process’ chart.

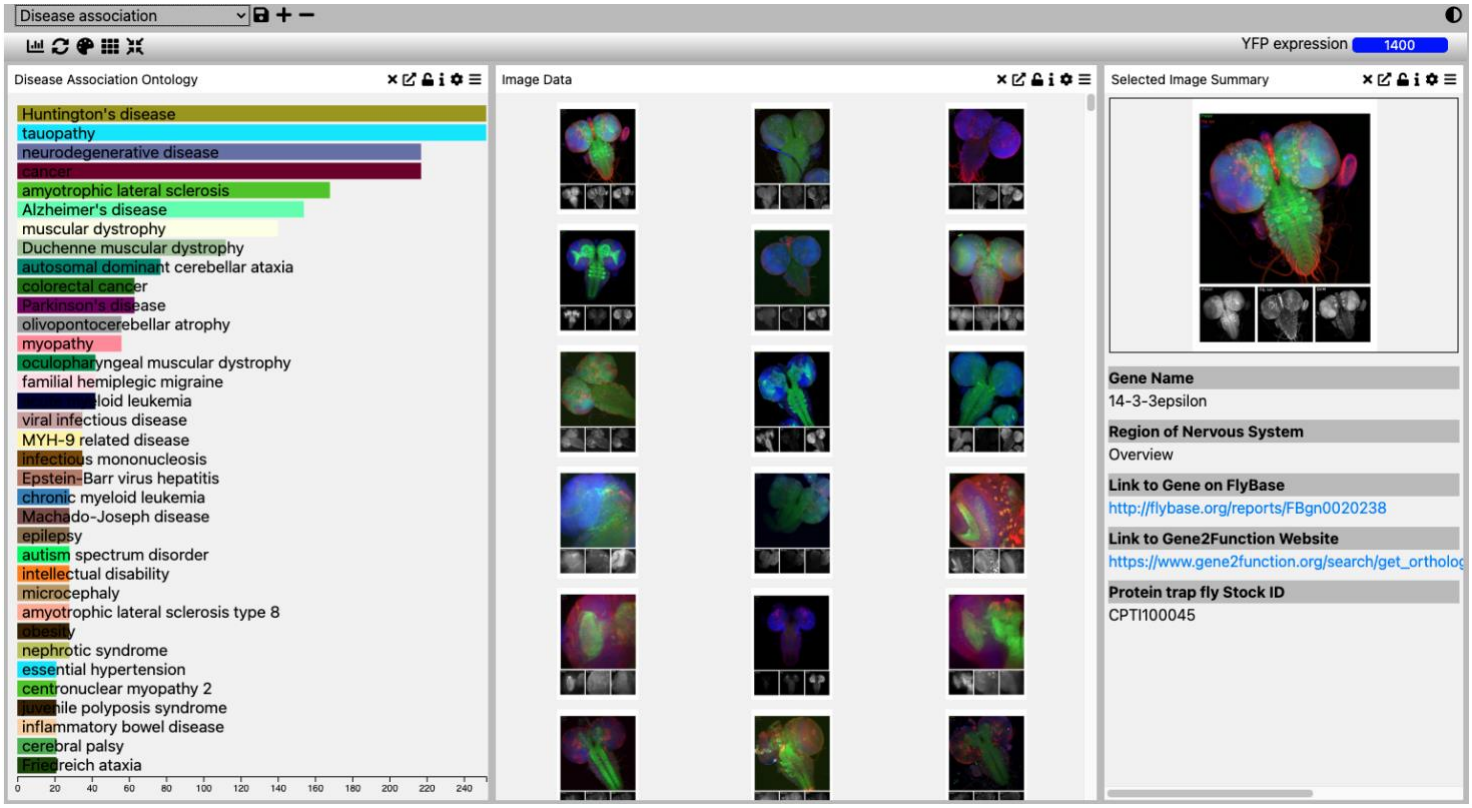


Figure 3: 'Disease Association' view of our collection.

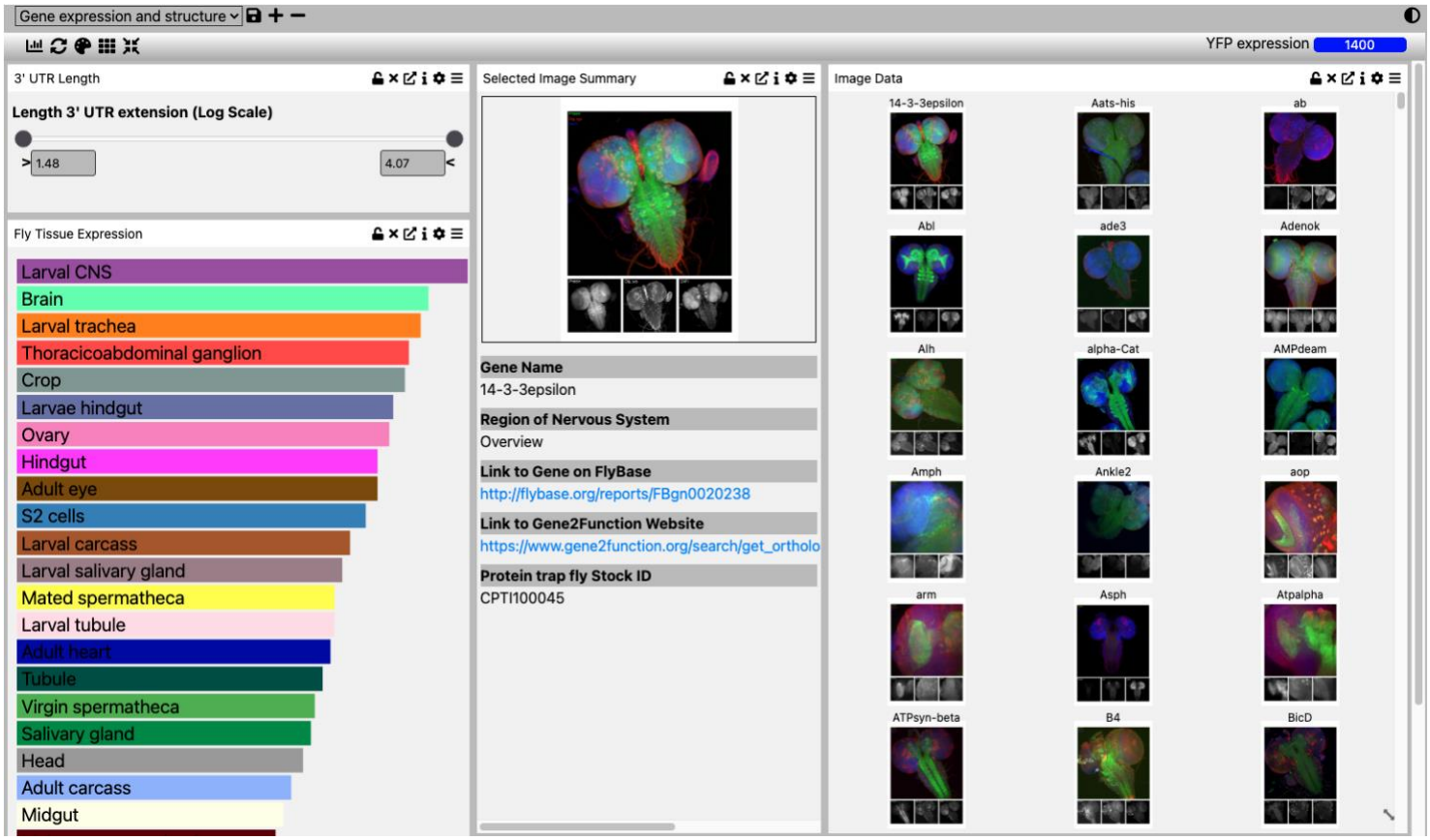


Figure 4: 'Gene Expression and Structure' view of our collection

The remaining views in our collection ‘Disease Association’ (Fig 3) and ‘Gene Expression and Structure’ (Fig 4) can be interrogated in a similar fashion. At any point while browsing the data in these views, it is possible for the user to create their own charts in order to display, filter and assess the data in a bespoke way. This flexibility enables the user to extract information that is most relevant to themselves from the collection in a straightforward manner.

HOW TO ACCESS THE RAW DATA

All imaging files are readily available. The figure diagrams used in the collection can be downloaded as a zip file within the Zenodo repository containing this manual. The original microscopy raw data files can be accessed individually from the ‘Raw Data Link to OMERO’ link in the ‘Selected Image Summary’ chart on the ‘Main’ view of the collection. This OMERO collection is part of the Open Microscopy Environment project (Allan et al., 2012) and all of the original raw data files can be explored and downloaded from (link). The non-imaging data is contained in the CSV file that can be downloaded from (link) and information on how it was populated is in the section above called ‘HOW TO PREPARE AND LOAD DATA INTO MDV’.

CONCLUSION

Using MDV we have created a new paradigm for exploring a medium throughput imaging data set in the context of the rich landscape of prior associated data in the literature and bioinformatics associated with the genes imaged. We have taken different data sets that are normally not interoperable and can only be explored using bespoke bioinformatics programming, and made them interoperable in one software platform that is intuitive to use and requires little training. Although our implementation is very much focused on our specific dataset, MDV is entirely flexible and users can follow our example by creating a multi-modal data landscape that is focused on their own interests.

REFERENCES

[1] Titlow, Joshua S, Maria Kiourlappou, Ana Palanca, Jeffrey Y Lee, Dalia S Gala, Darragh Ennis, Joyce J S Yu, et al. ‘Systematic Analysis of YFP Traps Reveals Common Discordance between MRNA and Protein across the Nervous System’. Preprint – Accepted for publication at JCB, https://doi.org/10.1083/jcb.202205129.

[2] Gramates, L Sian, Julie Agapite, Helen Attrill, Brian R Calvi, Madeline A Crosby, Gilberto dos Santos, Joshua L Goodman, et al. ‘FlyBase: A Guided Tour of Highlighted Features’. Edited by V Wood. Genetics 220, no. 4 (4 April 2022): iyac035. https://doi.org/10.1093/genetics/iyac035.

[3] Lyne, Rachel, Richard Smith, Kim Rutherford, Matthew Wakeling, Andrew Varley, Francois Guillier, Hilde Janssens, et al. ‘FlyMine: An Integrated Database for Drosophila and Anopheles Genomics’. Genome Biology 8, no. 7 (2007): R129. https://doi.org/10.1186/gb-2007-8-7-r129.

[4] Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, et al. ‘Ensembl 2022’. Nucleic Acids Research 50, no. D1 (7 January 2022): D988–95. https://doi.org/10.1093/nar/gkab1049.