Zegami user manual for data exploration: ”Systematic analysis of YFP gene traps reveals common discordance between mRNA and protein across the nervous system”

Maria Kiourlappou[1,*], Stephen Taylor[2], and Ilan Davis[1,*]

[1]Department of Biochemistry, University of Oxford, Oxford
[2]Weatherall Institute for Molecular Medicine, University of Oxford, Oxford
[*]Authors for correspondence: maria.kiourlappou@linacre.ox.ac.uk and ilan.davis@bioch.ox.ac.uk

ABSTRACT

The explosion in biological data generation challenges the available technologies and methodologies for data interrogation. Moreover, highly rich and complex datasets together with diverse linked data are difficult to explore when provided in flat files. Here we provide a way to filter and analyse in a systematic way a dataset with more than 18 thousand data points using Zegami (link), a solution for interactive data visualisation and exploration. The primary data we use are derived from a systematic analysis of 200 YFP gene traps reveals common discordance between mRNA and protein across the nervous system which is submitted elsewhere. This manual provides the raw image data together with annotations and associated data and explains how to use Zegami for exploring all these data types together by providing specific examples. We also provide the open source python code (github link) used to annotate the figures.

INTRODUCTION

This document provides instructions on how to use our microscopy image-based *Drosophila* genome collection in Zegami. It also explains the structure of the collection and the associated metadata, annotations and linked data. These rich datasets underpin our publication describing the biological significance of the data [1].

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)*. It contains images from 7 different compartments of the brain, namely Mushroom Body (MB), Central Brain (CB), Optic Lobe (OL), Ventral Nerve Cord (VNC), Neuromuscular Junction (NMJ) and Nerve.

High quality expert annotations of the microscopy images are included in this collection and available to be mined along with other expression data, functional and molecular information as well as structural information, derived from FlyBase [2], Flymine [3] and ENSEMBL [4]. The data available are not limited to the 200 genes but for the whole of the *D. melanogaster* genome.

ZEGAMI BASICS

Zegami [5] is a cloud-based solution for interactive data visualisation and exploration, with a set of machine learning algorithms that can be applied to the data for further exploration and discovery. An example includes Principal Components Analysis (PCA) clustering [6], of both image data and numerical data that are included in the collection.

A. *Creating a Zegami collection*

A collection in Zegami contains images and can contain associated data. If the collection has data, then a data file needs to be prepared that can be used as the basis of the collection.

The data needs to be arranged in rows and columns, in a csv file. The first row must contain the names for the columns and will be used by Zegami to refer to the column data.

It’s important to note that there must be a column in the data file that lists the names of the image files. In our collection this column is called ”image”. Each row in the data file corresponds to one tile in the collection. If an image name and an image exist then the tile displays the image. If not then the tile stays grey. If there are associated metadata in the row these are displayed in the panel on the right hand side.

To create the collection, the user needs to upload the data file (as a CSV file) along with the images associated with it in the Zegami interface, a typical Zegami collection can be seen in figure 1.

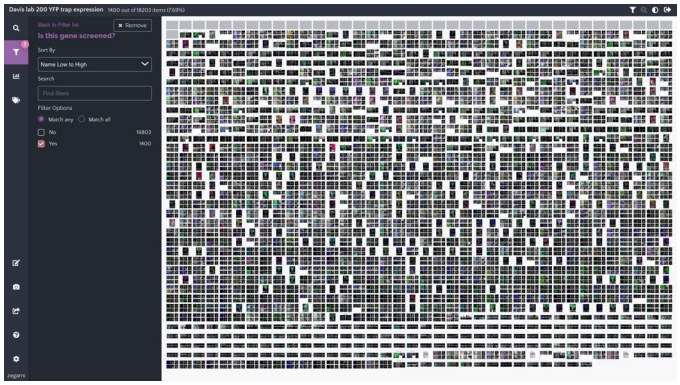


Fig. 1. A Zegami collection example. The "Filter" panel is visible on the left hand side and a selection of images in the centre. The user can zoom in in any of the images simply by clicking on them.

More details on how to create a collection in Zegami can be found in the Zegami website under the paragraph "Creating your first collection" `https://zegami.com/support/creating-your-first-collection/`.

B. *Exploring a collection*

Zegami offers multiple tools for exploring a collection. For example search and filtering. The collection search is a way of filtering the collection that should be familiar to most users. It features a text box where a metadata query can be entered `https://zegami.com/support/search-and-filtering/`. Entering a search term and clicking enter will look in all metadata fields to try and find the text. Filtering works by selecting the filter panel, which displays a list of all the columns. Text or categorical data is displayed as a list of check boxes. Selecting an item in the check box list will filter the collection to all items that contain that value. Selecting multiple values will display all items that contain any one of the selected values (if "Match any" is selected) or will display all items that contain all of the selected values (if "Match all" is selected) `https://zegami.com/support/search-and-filtering/`.

How to interrogate our data

You can access our collection by clicking here. The collection is publicly available. Once the collections loads, the user is presented with four different views:

1. Data Similarity view
2. Image Similarity view
3. Grid view
4. Table view

The Data Similarity and Image Similarity views show a clustered version of the data, helping the user understand potential relations between the data. Grid view shows all the images arranged in a grid, whereas table view shows all the associated data with the images arranged in a table.
There are different ways to navigate between views; One way is to click on the camera icon the left hand side panel. The "Snapshots" functionality allows the user to select between the

Table I
Snapshots of the collection in Zegami and their links:

Name	Description	Link
Data Similarity View	View of data clustered according to their numerical data similarity	link1
Image Similarity View	View of data clustered according to their image data similarity	link2
First-view	Only screened genes view	link3
show-all-genome view	Whole genome view including screened gnes	link4

Data Similarity view and Image similarity view. Additionally, there are two predefined Grid View views, seen in figure 2. "First-view" (figure 2 a) shows the 1361 microscopy images for the 200 genes we screened in our project. Most of the genes have images corresponding to 6 compartments of the *D. Melanogaster* brain and an overview image; a small amount of the genes have images corresponding to the compartments MB, NMJ, CB, OL, VNC and the overview image but not the nerve compartment. The user can also select the "show-all-genome" view (figure 2 b) where they can see tiles corresponding to the genes in the whole *Drosophila* genome, even for the genes that haven't been screened in our project. The names of the different views, their descriptions and the links to them are given in table I.

The user can select any one of the tiles/images. A zoomed in version of the tile/image will appear and the data of that tile/image will appear in the right-hand side panel, as shown in figure 3.

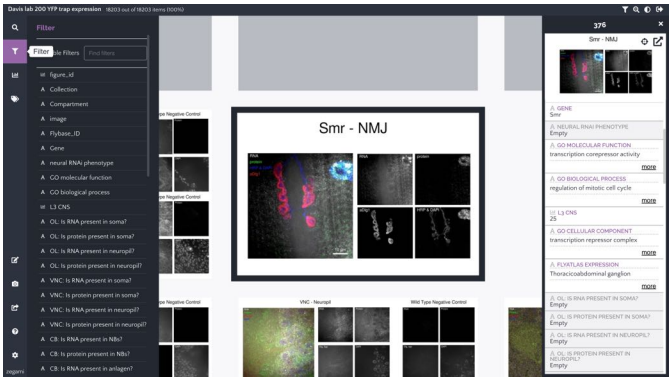
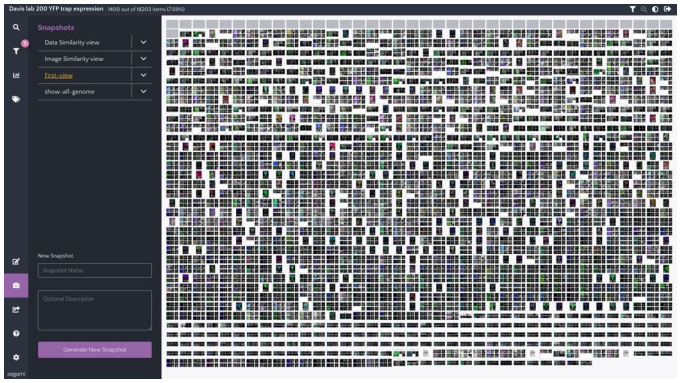
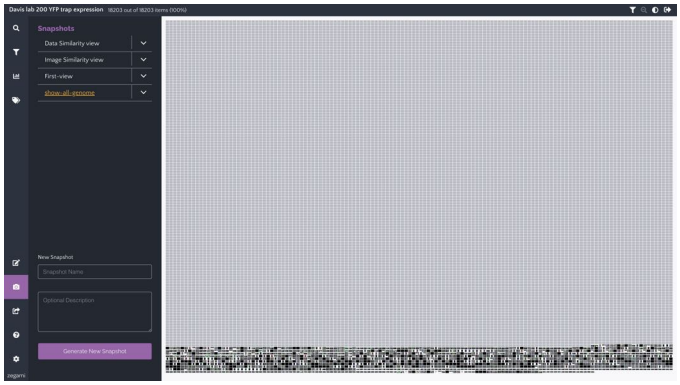


Fig. 3. Zoomed in the *Drosophila* gene Smr, the NMJ compartment. On the right hand side we can see the metadata panel with information about this gene's functions and more. On the left hand side we can see the filtering panel and can select to view other genes.

In figure 3, the gene Smr is shown, and the compartment selected is the Neuromuscular Junction (NMJ). The gene name,



(a) First-view



(b) show-all-genome view

Fig. 2. Two different views of our collection in Zegami. The user can interchange by clicking on the Snapshots panel between viewing the screened genes only or the whole genome including screened and not screened genes.

the GO Molecular function, GO Biological function, GO Cellular Component and Flyatlas expression are shown. If the user scrolls further down they can see the annotations related to that image, as scored by experts in our lab. For example for this gene there is mRNA present in the axon terminal.
The following are examples of some use cases showing how data included in the analysis in our manuscript [1] are mined and conclusions are reached.

C. Use Cases

Our microscopy-based gene screened collection in Zegami is a very-information-rich source for mRNA and protein localisation. In our manuscript [1], we reach biological conclusions and in this section we present some of the results and how the user can use our collection in Zegami to verify those.

c.1. mRNA expression in the synaptic compartments

The synaptic compartments include the mushroom body neuropil, the optic lobe neuropil, the ventral nerve cord neuropil and the neuromuscular junction axon terminal. In this use case we determine how many genes out of the 200 genes that we screened show mRNA localisation in at least one of the synaptic compartments. One can filter those genes and look the mRNA expression in each gene by selecting the filter ”Compartments with RNA expression”. Then the user should select the compartments ”VNC neuropil”, ”MB neuropil”, ”OL neuropil” and ”NMJ axon terminal” with the option ”Match any”, as seen in figure 4. When the filters are applied we can see that there are 137 genes that match those criteria (top left of figure 4).

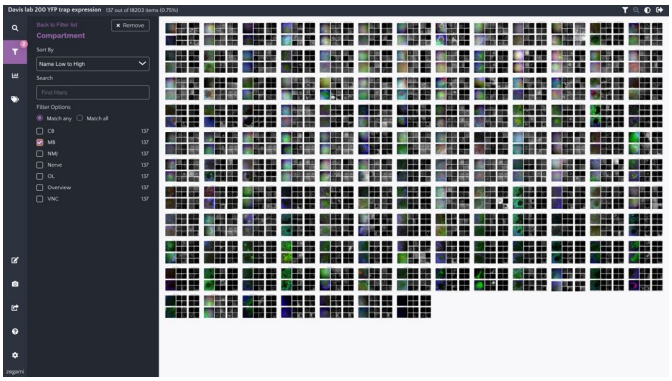


Fig. 4. The 137 genes that have mRNA expression in at least one synaptic compartment.

c.2. Mushroom Body synaptic neuropil expression

Here we examine the mRNA expression in the mushroom body synaptic neuropil. There are 67 genes that show expression as they can be seen in figure 5. To reach this result, the user can select the filter ”MB: is RNA present in neuropil” and then select yes.

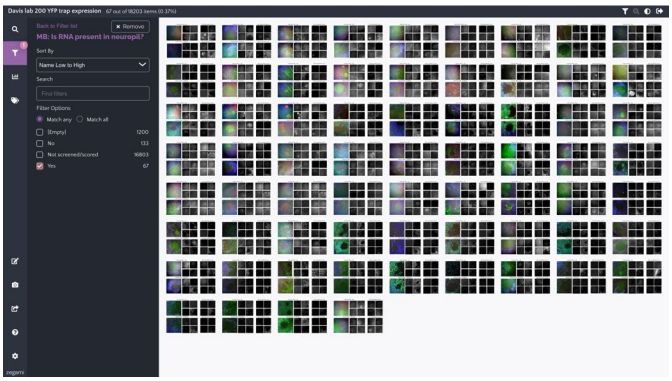


Fig. 5. mRNA expression in the mushroom body synaptic neuropil

c.3. Mushroom Body and Optic lobe synaptic neuropil expression

In this example we examine genes that express in both the MB and OL neuropils. These are 28 genes. The way to filter these and the images that correspond to those can be seen in figure 6. These are the common genes between the two compartments so the filter type that needs to be used is ”Match all”.

c.4. mRNA expression in motoneurons in the Neuromuscular Junction

There are 69 genes that show mRNA localisation in the motoneurons (figure 7). These genes can be further examined as are candidates for local translation in response to neuronal activation [1]. Being able to view the available biological information in the side panel is very important as it can guide users on how to plan and prioritise future experiments on target genes.

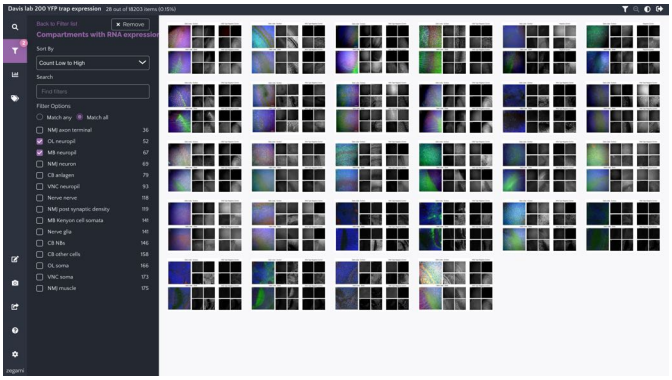


Fig. 6. mRNA expression in both the muschroom body and optic lobe neuropils.

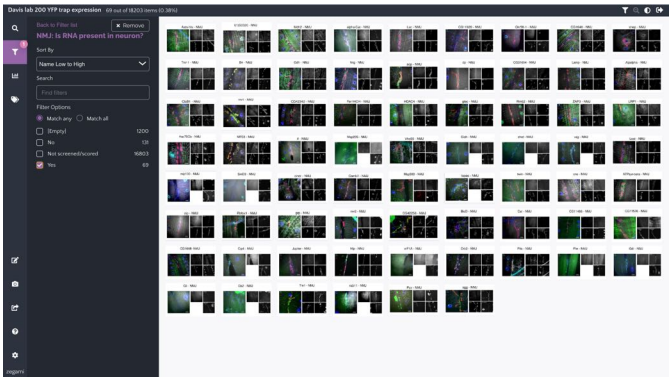


Fig. 7. Genes that show mRNA localisation in the motoneurons

D. How to access the raw data

The user can download the file containing information about each gene and each image directly from Zegami. This can be accessed from the left hand side panel by clicking on the "Export and Share" button and then click Export as shown in figure 8. We have made the file available in this Zenodo publication as well. In addition we have published all the raw images so a user can access them and download a zip folder containing them. This is to ensure availability of the data and flexibility for the users in case they want to perform additional analysis in offline pipelines.

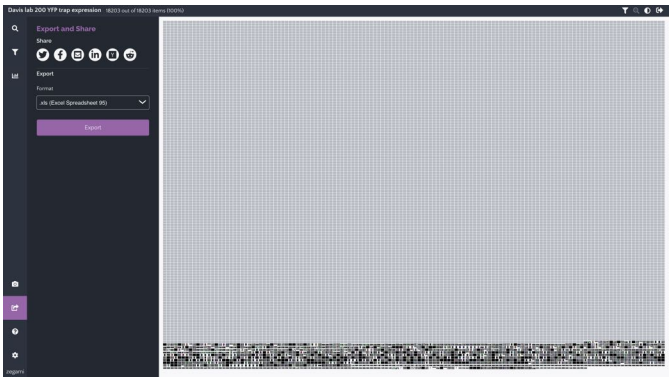


Fig. 8. The user can download the file containing information about each gene in case they want to use the information for subsequent analysis on their own pipelines.

CONCLUSIONS

We provided specific examples of filtering and interrogating image data, their associated metadata and annotations with the whole *Drosophila* genome that are flexible and can be modified very easily. Our examples of using Zegami provide a paradigm for data exploration enabling the user to discover novel insights from complex biological data collections.

REFERENCES

[1] Joshua S Titlow Maria Kiourlappou Ana Palanca Jeffrey Y Lee et al. "Systematic analysis of YFP gene traps reveals common discordance between mRNA and protein across the nervous system". In: *BioRxiv* (Mar. 2022). DOI: https://doi.org/10.1101/2022.03.21.485142.

[2] Aoife Larkin et al. "FlyBase: Updates to the Drosophila Melanogaster Knowledge Base". In: *Nucleic Acids Research* 49.10 (2021), pp. D899–907. DOI: https://doi.org/10.1093/nar/gkaa1026.

[3] Rachel Lyne et al. "FlyMine: an integrated database for Drosophila and Anopheles genomics". In: *Genome Biology* 8.R129 (2007). DOI: doi:10.1186/gb-2007-8-7-r129.

[4] Kevin L Howe et al. "Ensembl 2021". In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D884–D891. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa942. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D884/35364073/gkaa942.pdf. URL: https://doi.org/10.1093/nar/gkaa942.

[5] Stephen Taylor and Roger Noble. "HTML5 PivotViewer: high-throughput visualization and querying of image data on the web". In: *Bioinformatics* 30.2 (Sept. 2014). DOI: 10.1093/bioinformatics/btu349. URL: https://doi.org/10.1093/bioinformatics/btu349.

[6] I. T. Jolliffe. *Principal Component Analysis*. 2nd. New York, NY, USA: Springer, 2002. ISBN: 978-0-387-95442-4.