FISHtoFigure User Information

Calum Bentley-Abbot

c.bentley-abbot.1@research.gla.ac.uk

Contents

1 What is FISHtoFigure?

FISHtoFigure is a standalone software tool allowing for analysis of complex, QuPath quantified datasets in a user-defined, target-oriented manner. Capable of multi-target, multi-dataset analysis, FISHtoFigure facilitates the rapid, streamlined analysis of large datasets comprising many images in an intuitive fashion. Users can partition transcriptomic data in an automated way simply

by defining a target or series of targets for which partitioned cells must be positive allowing analysis to be conducted for specific cell types and transcriptional profiles.

2 FISHtoFigure Capabilities

FISHtoFigure can partition cellular and subcellular object information in an automated manner from datasets of quantified smFISH microscopy data, with the development focus being the analysis of quanitifed RNAScope images. Currently, FISHtoFigure is limited to the analysis of Qupath output files, however, due to the method by which FISHtoFigure carries out partitioning and analysis, the software is adaptable to other formats of quantified microscope image data.

FISHtoFigure allows users to partition transcriptomics data by desired channel, isolating cells which are positive for a particular target, or multiple targets, across multiple datasets simultaneously in order to quickly analyse specific elements of the transcriptome in datasets that would otherwise be very difficult to parse.

3 FISHtoFigure Usage

3.1 Qupath Output File

Following quantification of cellular and subcellular information, Qupath will output a text file containing image information. This file contains rows corresponding to individual cells or subcellular spots and columns containing specific information relating to cells or subcellular objects such as centroid coordinates and subcellular object intensity information. Due to the volume of information within fluorescence image files, they are generally hundreds of columns and thousands of rows long, and much of this information is not of interest for transcriptomic analysis. FISHtoFigure provides a means to quickly partition and analyse information of interest without directly working with the dataset.

4 Input Data

FISHtoFigure is currently capable of working with standard Qupath output text files of images with an arbitrary number of detection channels, however, there are certain limitations on the input data as detailed here.

Input datasets should take the form of a txt file (either a standard QuPath output file or a concatenated file comprising multiple QuPath files created using the preprocessing tool) and all datasets to be analysed should be stored within a single input directory, the path to which can be defined in the GUI, discussed below.

4.1 Naming convention

Input datasets should be named according to the following convention "***X***.txt" where *** represents a variable, non-numerical string with no spaces (eg the dataset/samplename) and X represents a numerical value unique to that dataset. Numerical variable X must be 1 in the first dataset to be analysed, and subsequent section must be labelled in ascending order.

Examples of appropriate input dataset names: Inf1.txt, Naive12.czi.txt, DatasetSection31SampleA.txt

Examples of inappropriate input dataset names: Sample12Section10.txt, Section5.czi, Dataset.txt, Inf1.txt.csv, 12Data.txt, Naive Data 1.txt

4.2 Preprocessing

In the case of datasets comprising multiple images, the Qupath outputs for said images can be concatenated in an automated manner by running the preprocessing tool on the directory containing these datasets. Running the preprocessing tool produces a "Concatenated Dataset" file in both txt and csv format. The txt format file can then be processed by the main FISHtoFigure program in the standard fashion. The "Concatenated Dataset" file is saved to the input directory specified in the preprocessing tool GUI.

4.3 Input Format

FISHtoFigure partitions cells and subcellular information by using a series of flags (some of which are defined by the user at input) to search the QuPath output file for rows and columns containing cellular and subcellular information such as cell centroids, subcellular spot intensities, and cellular object count. The flags are defined such that a standard Qupath output file will already contain the necessary flagged column and row headings, however, for clarity a description of the required flags is given below:

4.3.1 Rows:

Rows represent either cells or a single subcellular object (such as a single subcellular spot or a subcellular cluster). A row representing a cellular object will be followed by rows representing individual subcellular objects within this parent cell until all subcellular objects are exhausted at which point the subsequent row will represent a new cell. The information used to determine the feature represented by each line is contained in the "Name" column.

4.3.2 PathCellObject

Represents a cell object, this contains populated columns relating to subcellular spot and cluster count for each target channel within this cell. It does not contain populated columns for intensity of said subcellular objects.

4.3.3 Subcellular spot: Channel X object

Represents a single subcellular spot in target channel X, this contains populated columns for the intensity of each channel, though only the Channel X intensity will be non-zero.

4.3.4 Subcellular cluster: Channel X object

As above, but representing a cluster within the parent cell rather than a single spot.

4.3.5 Columns:

Column headers comprise the majority of flags used in the partitioning of datasets by FISHtoFigure. These contain much of the data of interest and the relevant flags are detailed below.

4.3.6 Centroid X Âµm

Contains a float value representing the X-axis centroid (defined in QuPath by the intensity of the nuclear staining detection channel) information for the cellular or subcellular object, this is used by FISHtoFigure to produce the scatter plot of cells coloured and sized based on transcriptomic spot information.

4.3.7 Centroid Y Âµm

Equivilant Y-axis information.

4.3.8 Subcellular: Channel X: Num single spots

For cell-type rows: Contains an integer representing the number of subcellular single spots contained in the current cell.

4.3.9 Subcellular: Channel X: Num clusters

For cell-type rows: Equivilant information regarding clusters within the current cell.

4.3.10 Subcellular cluster: Channel X: Mean channel intensity

For subcellular-type rows: Contains a float representing the intensity for Channel X for the current subcellular object.

4.3.11 Multi-dataset analysis

In the case that multiple datasets are to be compared (e.g. comparing a dataset comprising all naive controls with that of infected counterparts) datasets in text file format can simply be placed in the same directory, named following

the naming convention laid out above, and FISHtoFigure run on the directory. Datasets will be interpreted iteratively and plotted together. Note that when defining the dataset names in the GUI, dataset names should be given in the order that the datasets appear in the input directory

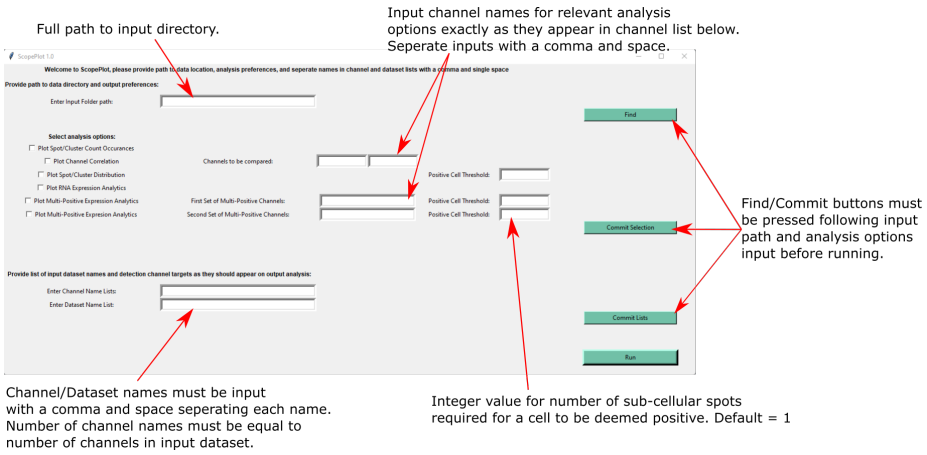

Figure 1: A simple GUI has been created such that users can use FISHtoFigure with minimal interfacing with the underlying Python program.

4.4 GUI

4.4.1 Preprocessing

For the concatenation of image data, a simple GUI has been created. The user should enter the full path to the directory containing image data and click "Find" to submit this path, clicking "Run" will then run the program and produce the concatenated file in text and csv format within the specified directory.

4.4.2 FISHtoFigure

A GUI has been created for the main FISHtoFigure partitioning and analysis program, shown annotated in Figure 1. Here, the user can should enter the full path to the directory containing data for analysis, this can be submitted as above.

Users can then define which analysis outputs they desire in the analysis options section, examples of outputs are shown in Figures 4 and 5. If users wish to use the (Multi-positive cell analysis) function of FISHtoFigure requiring the input of channel names, channel names should be input exactly as they appear in the user-defined channel list inputted in the section below. Once selected, analysis options can then be submitted via the "Commit Selection" button.

Finally, the user should provide lists of the channels within the datasets (The names will be assigned to each target channel, these will be used in plotting

and any partitioned output datasets and will replace the channel headers in the standard Qupath output [e.g. Subcellular: Channel 1: Num single spots → *Channel 1 name* Spots]). Target channel names within the list should be separated by a comma and a single space. A list containing the names associated with each dataset should be inputted below, even if the user is only analysing a single dataset. This list will be used for assigning data to each input dataset during analysis and plotting. Dataset names should be inputted such that they align with the order of datasets within the input directory (e.g., "Dataset1.txt" in the input directory will be labelled with the first name specified in the dataset name list entered in the GUI, followed by "Dataset2.txt", etc.)

5 FISHtoFigure Processing and Output

When running FISHtoFigure, a new directory is created within the input directory specified in the GUI called "F2F Output", this stores all analytical outputs created including figures, datasets associated with figures, and analytical txt format outputs.

An example of the generated "F2F Output" folder output from running all of FISHtoFigure's analysis options with two input datasets is shown in Figure 6.

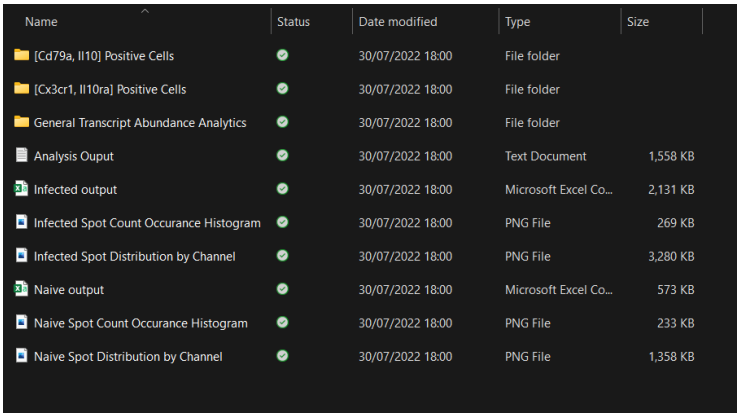

Figure 2: Example of the F2F output directory for two input datasets: "Naive" and "Infected". Here, two multi-channel expression profiles were analysed: "Cd79a, Il10" positive cells and "Cx3cr1, Il10ra" positive cells.

5.1 CSV Data

The initial process of FISHtoFigure consists of the conversion of each of the Qupath output text files in the designated directory into an equivilant csv file. This csv format is required for the rest of the data analysis carried out by

FISHtoFigure while additionally being easily interpretable by programs such as excel. Therefore, the generated csv files are saved to a newly generated sub-directory titled "csv files" within the directory specified by the user in the GUI in case users wish to conduct custom analysis. These generated csv files are named based on the dataset names in the input directory (e.g., "*Dataset name*.csv").

5.2 Non-Zero Cell Data Harvesting

Following input file format conversion, the first task carried out by FISHtoFigure is the creation of a dataframe containing only cells expressing at least one subcellular spot in at least one detection channel, referred to hereafter as non-zero cells. FISHtoFigure reads the input file in a line-by-line fashion, locates non-zero cells, and writes this cell's information to a new file. The resulting non-zero cell dataset is written to a csv file titled "*Input file name* output" and saved to the "F2F Output" directory within the user specified directory.

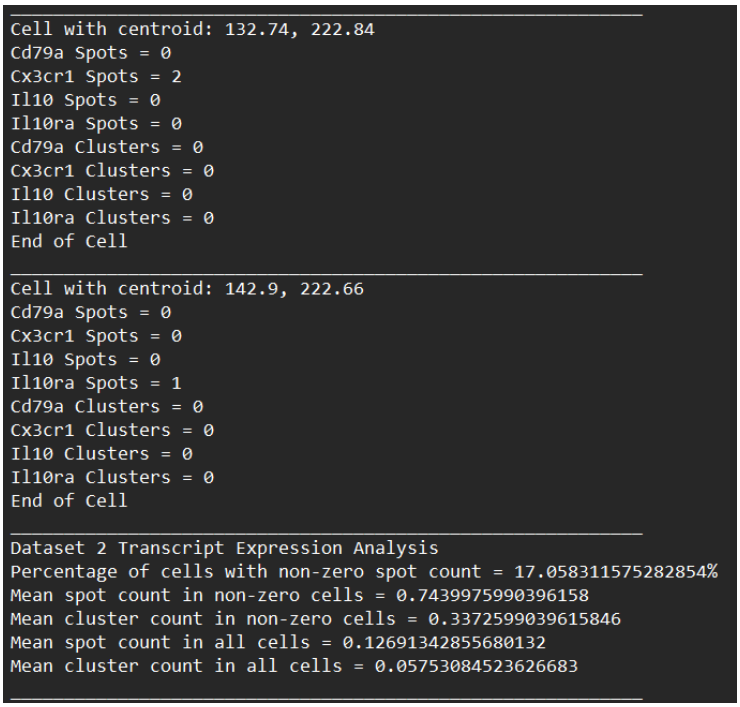

Figure 3: Short section from the pipeline output for an example RNAScope dataset. Non-zero cells are recorded by centroid information. Spot information is reported for all non-zero cells. Following cellular information, holistic analysis is reported for the dataset before automatically beginning the analysis of any subsequent datasets, reported within the same output file.

5.3 Data Harvesting and Analytics as Text File

While carrying out the non-zero cell harvesting process, FISHtoFigure harvests information such as the cell's position within the input image (centroid) and subcellular spot information for each spot within the cell. To aid with data interpretation, harvested information for each non-zero cell is written to a text file titled "Analysis Output" in the "F2F Output" directory, an example of the format for the output cellular data is shown in Figure 3.

Along with non-zero cellular information, the Analysis output file stores the list of files being analysed, input file header information (The flags being used to harvest information and the user-specified channel names associated with each flag), and some holistic analysis for each image dataset such as percentage of cells containing subcellular spots and mean spot count per cell within the image.

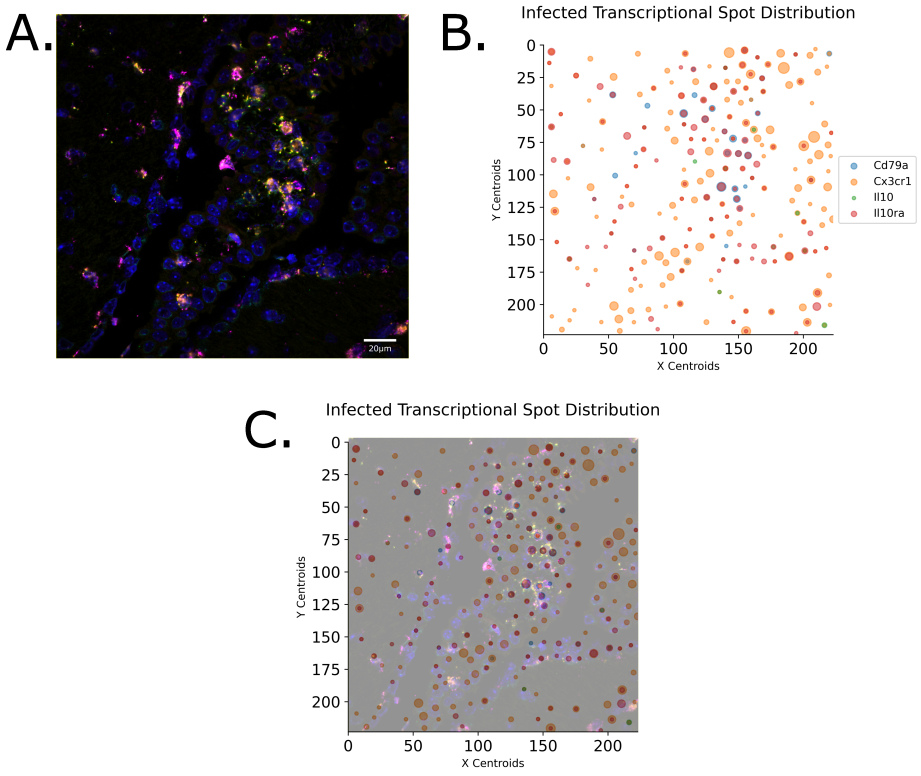

Figure 4: Spatial distribution plotting function example using a dataset comprising a single RNAScope image. A: The fluorescence image acquired via RNAScope, with fluorescent spots representing RNA transcripts. B: Spatial distribution produced by FISHtoFigure. C: An overlay of the RNAScope output image with the FISHtoFigure distribution.

5.4 Transcript Occurrences and Distribution

If the user selects either the "Plot transcript occurrences" or Plot transcript distribution" analysis options the relevant plots will be saved to the "F2F Output" directory.

The "Plot transcript occurrences" feature recreates the in-house spot occurrences analysis available in QuPath, plotting the number of cells exhibiting a given number of spots in each channel.

The "Plot transcript distribution" option will plot centroid data for cells with non-zero subcellular spot counts for a given input dataset, creating a plot analogous to the input image. As such, this option is recommend for use with single image datasets where centroid information is consistent. Use with concatenated datasets comprising multiple images will result in overlap since centroid information is recorded per image and spatial relationships between images are not conserved during concatenation.

Points within the resulting scatter plot represent individual cell centroids with the colour and size of data points representing channel and subcellular spot count respectively. This function provides not only a highly quantified equivalent of the original input microscopy data, but also visual output which can be directly compared with the original dataset to validate the harvesting of information from the dataset by FISHtoFigure, as shown in Figure 4.

5.5 Differential Transcriptomic Analysis

Following data harvesting, FISHtoFigure provides the user with a means to conduct transcript expression analysis, this can be conducted for individual images/datasets, or for an arbitrary number of datasets, enabling the comparison of transcript expression between datasets if multiple datasets are within the user specified input directory (e.g., between Naive and Infected specimens). This analysis produces cell-wise plots of spot count and intensity by target channel, these plots consist of a box and swarm-plot, where each data point in the swarm plot represents a single cell. In the case of the Cell-wise intensity plot, each data point represents the sum total intensity of all subcellular spots in that detection channel within that cell (Therefore, each cell is represented once for each detection channel). In the case of the cell-wise spot count figure, each data point simply shows the number of subcellular spots in that detection channel in that cell.

Additionally, FISHtoFigure produces plots of total subcellular spot number, intensity, and positive cell number for each detection channel as well as the total number of non-zero cells across all channels for each dataset. Examples of the analysis produced using this function are shown in Figure 5.

Along with each generated figure, the dataframe containing the information used to generate that figure is saved as a csv with the same name as the associated figure. This allows the figures generated in FISHtoFigure to be recreated using another graphing software such as Prism by GraphPad quickly and easily or for more specific analysis to be conducted on the specific subset of data used

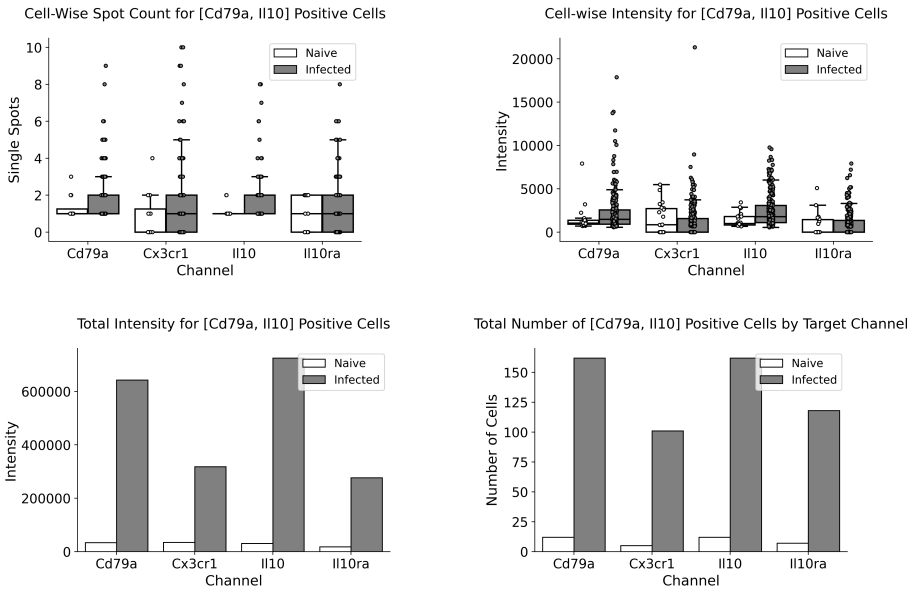

Figure 5: Differential transcript abundance analysis performed using FISHtoFigure for a pair (naive and infected) of concatenated datasets comprising multiple RNAScope images. A: Cell-wise spot count by RNA target channel, each point represents the number of spots for each detection channel for a particular cell. B: Cell-wise intensity by RNA target channel. Here, each point represents the total intensity for each channel for a particular cell. C: Total intensity by RNA target channel. Cumulative intensity count over all cells in each channel, for each dataset analysed. D: Spot-positive cell number by RNA target channel. Total number of cells expressing at least on transcript in a given detection channel, with cell-positivity being determined based on user determined threshold in GUI at input.

to create a particular figure (E.g., Cellular intensities for each target channel for non-zero cells) without need to manual harvest these data from the csv files generated earlier in the FISHtoFigure pipeline. These datasets are structured to be easily interpretable to ease further analysis, as such, columns containing data relating to each detection channel are labelled based on the user defined labels for each channel specified in the GUI, and the column representing the dataset is labelled based on the user defined dataset list. Each dataset derives it's format from the original QuPath file, as such, each row represents a cell (Other than in the cumulative analyses such as the positive cell number, which comprise only a single row containing the sum totals across all cells for each channel) and each column represents either the spot number or intensity form a particular channel, depending on the dataset.

All figures and associated datasets generated using this function are saved

to a newly generated directory titled "General Transcript Abundance Analytics" within the "F2F Output" directory. An example of a "General Transcript Abundance Analytics" directory generated using an pair of example datasets is shown in Figure .

Figure 6: Example of the generated "General Transcript Abundance Analytics" directory within the "F2F Output" directory. Each figure has an associated dataset which contains the data required to recreate the figure saved as a csv of the same name.

5.6 Custom Transcriptomic Analysis for Multi-Target Expressing Cells

Using the "Multi-target transcript abundance" function, users can define a subset of target channels on which to conduct specific transcriptomic analysis. Users should enter the names of the targets they wish to conduct analysis on as they appear in the user-defined "channel name list" within the GUI, as with the channel and dataset lists, target channel names should be separated by a comma and single space. FISHtoFigure will then search the non-zero cell data harvested from the QuPath file specifically for cells co-expressing all the targets defined by the user in the "First/Second set of multi-positive channels" input section. Cells found to fulfill the specified expression profile are stored to a separate dataframe and differential transcript expression is conducted for this new dataframe as described in the "Differential Transcriptomic Analysis" section above.

As with the "Differential Transcriptomic Analysis" section above, datasets used to generate each figure are saved as csv files with the same name as the

associated figure. All figures and associated datasets are saved to a newly generated directory with the "F2F Output" directory, the name of this directory is based on the channels specified by the user in the multi-positive analysis input (e.g., "[*Channel 1*, *Channel 2*, *Channel n*] Positive Cells").

Currently, FISHtoFigure facilitates the analysis of two multi-positive cell profiles within the input GUI, allowing for multiple multi-target positive conditions to be explored in parallel.

5.7 Cell Positivity Threshold

If users are using the "Plot transcript distribution" or the "Multi-target transcript abundance" functions, a "Positive cell threshold" can be set. This sets a minimum limit on the number of transcripts a cell must express for it to be deemed positive and thus be included in analysis. This allows for the removal of background expression, for example, if a user wishes to analyse cells which highly express a particular RNA target is expressed at low levels by all cells, users can set a threshold such that only highly expressing cells are analysed. This threshold should be input as a single integer, all cells expressing a number of transcripts greater than or equal to the threshold will be included in analysis. If no threshold is specified FISHtoFigure will include all non-zero cells by default.

6 Developer Information

FISHtoFigure is an open source software hosted on GitHub. Though currently FISHtoFigure is designed to process Qupath output files, all partitioning and analysis occurs after data is extracted from the input file and stored as various dataframes. Therefore, adapting the software to interpret input files from other software such as Fiji or ImageJ should be achievable by simply changing the flags used to identify information required for analysis. Provided cellular and subcellular information can be quantified in a tabulated format, it is likely FISHtoFigure can be adapted to work with this format.

GitHub: https://github.com/Calum-Bentley-Abbot/FISHtoFigure.git

Lead developer information:
Calum Bentley-Abbot
Wellcome Centre for Integrative Parasitology (WCIP).
Institute of Biodiversity, Animal Health, and Comparative Medicine.
Henry Wellcome building
University of Glasgow
Glasgow, UK G61 1QQ
c.bentley-abbot.1@research.gla.ac.uk