scientific reports

OPEN

An easy to use tool for the analysis of subcellular mRNA transcript colocalisation in smFISH data

Calum Bentley-Abbot[1,2,3✉], Rhiannon Heslop[1,2], Chiara Pirillo[4], Praveena Chandrasegaran[1,2], Gail McConnell[5], Ed Roberts[4], Edward Hutchinson[3] & Annette MacLeod[1,2]

Single molecule fluorescence in situ hybridisation (smFISH) has become a valuable tool to investigate the mRNA expression of single cells. However, it requires a considerable amount of programming expertise to use currently available open-source analytical software packages to extract and analyse quantitative data about transcript expression. Here, we present FISHtoFigure, a new software tool developed specifically for the analysis of mRNA abundance and co-expression in QuPath-quantified, multi-labelled smFISH data. FISHtoFigure facilitates the automated spatial analysis of transcripts of interest, allowing users to analyse populations of cells positive for specific combinations of mRNA targets without the need for computational image analysis expertise. As a proof of concept and to demonstrate the capabilities of this new research tool, we have validated FISHtoFigure in multiple biological systems. We used FISHtoFigure to identify an upregulation in the expression of *Cd4* by T-cells in the spleens of mice infected with influenza A virus, before analysing more complex data showing crosstalk between microglia and regulatory B-cells in the brains of mice infected with *Trypanosoma brucei brucei*. These analyses demonstrate the ease of analysing cell expression profiles using FISHtoFigure and the value of this new tool in the field of smFISH data analysis.

Single molecule fluorescence in situ hybridisation (smFISH) technologies such as RNAScope enable the visualisation of single mRNA molecules within single cells. mRNA transcripts are detected by fluorescence microscopy, with each transcript appearing as a single 'transcriptional spot'[1]. Quantification of these signals enables the analysis of transcriptional activity at the single cell level within the spatial context of tissues[2]. However, the large microscopy datasets produced by smFISH experiments currently require custom code in order to conduct in-depth transcriptomic analyses. QuPath is a purpose-built platform for the analysis of large images such as those acquired during smFISH experiments, and is recommended by ACDBio-Techne, the developer of the RNAScope platform (https://acdbio.com/qupath-rna-ish-analysis), for image analysis[3]. QuPath has specific in-built tools for cell segmentation and fluorescent spot detection, which can be used to quantify transcriptional spots. Furthermore, the software incorporates a batch processing feature which facilitates automated analysis of data from multiple images[3]. Following quantification, QuPath can plot quantified data, such as transcripts per cell, as a histogram[3]. However, users wishing to conduct more complex analyses, such as differential expression analysis or co-expression analysis, must develop custom pipelines to parse raw QuPath output data, thus restricting such analysis to users with extensive programming experience.

Here, we present FISHtoFigure, a standalone, open-source software tool for the in-depth analysis of transcript abundance in QuPath-quantified smFISH data by users with all levels of programming experience. FISHtoFigure can concatenate the batch processed data from QuPath, enabling the analysis of large, multi-image datasets. Notably, FISHtoFigure allows users to conduct transcript abundance analysis for cells with specific, multi-transcript expression profiles. Additionally, FISHtoFigure enables users to conduct differential expression analysis between datasets, facilitating the targeted study of differential expression in specific cell types and populations. Thus, FISHtoFigure provides a means for all users to examine mRNA expression of multiple transcripts without the need for custom analysis pipelines.

Here, we demonstrate the use of FISHtoFigure in two biological scenarios. First, we used FISHtoFigure to analyse T-cell and B-cell populations in the spleens of influenza A virus (IAV) infected mice, hereafter referred

[1]Wellcome Centre for Integrative Parasitology (WCIP), University of Glasgow, Glasgow, UK. [2]School of Biodiversity, One Health, Veterinary Medicine (SBOHVM), College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [3]MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow, UK. [4]Beatson Institute for Cancer Research, Glasgow, UK. [5]Department of Physics, University of Strathclyde, Glasgow, UK. ✉email: c.bentley-abbot.1@research.gla.ac.uk

to as the spleen dataset. Second, we demonstrate the capabilities of FISHtoFigure for the analysis of high-plex smFISH data collected from highly ramified, non-round cell types, using a dataset obtained in a recent experiment by our group investigating microglia in the brains of *Trypanosoma brucei brucei* infected mice, hereafter referred to as the brain dataset[4].

Materials and methods
Specifications and data handling
FISHtoFigure is a Python-based analytical software tool, designed to quantify cell expression profiles within smFISH data. Expression profile analysis is conducted in FISHtoFigure using the Pandas library[5]. A two-branched strategy is used to isolate cellular and subcellular data into two new datasets. Graphical outputs are generated using a combination of the Matplotlib and Seaborn Python libraries[6,7]. In addition to the graphical outputs, data from FISHtoFigure analysis are stored in CSV format for downstream statistical analysis. The statistical tests in this paper were performed with GraphPad PRISM, non-parametric tests were selected due to lack of normality.

Animal work and sample collection
All spleen samples were collected from 9 week old, male C57BL/6 mice. A mouse was intranasally infected with IAV A/Puerto Rico/8/34 (PR8; H1N1). The infected mouse was culled 6 days post infection via intraperitoneal injection of pentobarbital and whole spleens were harvested immediately. The infected mouse was weighed to monitor disease progression as per the ethical regulations of animal licence P72BA642F. The spleen from an uninfected male C57BL/6 mouse, culled by the same method, was harvested to act as a naïve control.

All brain samples were collected from 6 to 8-week-old, female C57BL/6 mice. Two mice were infected with *T. b. brucei* Antat 1.1E[4]. Mice were culled 45 days post infection via rapid decapitation following isoflurane anaesthesia and whole brains were harvested immediately. Mice were monitored for disease severity using the following clinical scoring method, score (0) normal, healthy, and explorative mouse; score (1) slow, sluggish, or displaying stary coat; score (2) animals with reduced coordination of hind limbs and/or altered gait; score (3) animals with flaccid paralysis of one hind limb. Mice with clinical scores higher than (3) were humanely killed as per the ethical regulations of animal licence PC8C3B25C. Whole brains from two uninfected female mice, culled by the same method, were harvested to act as naïve controls.

Mice were bred and housed at the Beatson Cancer Research Institute (Spleen samples) and the University of Glasgow Centre for Virus Research (Brain Samples). All animal work was carried out in line with the EU Directive 2010/63/eu and Animal (Scientific Procedures) Act 1986, under project licences P72BA642F (Spleen samples) and PC8C3B25C (Brain samples), and was approved by the University of Glasgow Animal Welfare and Ethics Review Board. This work was carried out in accordance with the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines. The reporting in this study fulfils the ARRIVE recommendations.

All samples were fixed in 4% paraformaldehyde (PFA) at room temperature for 24 h and embedded in paraffin. From paraffin blocks, sections were cut on a microtome (Thermo Scientific) and mounted on glass slides for histology.

RNAScope data collection
Commercial RNAScope control slides containing mouse NIH 3T3 cells (Advanced Cell Diagnostics, US) were used as a positive control sample for RNAScope for all samples. RNAScope was used to visualise *Cd79a* and *Cd4* transcripts in the spleens of naïve and IAV infected mice, and *Cd79a, Cx3cr1, Il10* and *Il10ra* transcripts in the brain of naïve and *T. b. brucei* infected mice. Fresh probe mixes containing the RNAScope probes were prepared for each experiment (Table 1). A single probe per channel (C1–C4) was included in each experiment. RNAScope 4-plex positive controls (for *Polr2a, Ppib,* and *Ubc*) and negative controls (for the *Bacillus subtilis* bacterial *Dapb* gene) were also included (probe details are listed at https://acdbio.com/control-slides-and-control-probes-rnascope). Slides were imaged by confocal microscopy (Zeiss LSM 880, 63 × objective for the spleen samples; Zeiss LSM 710, 63 × objective for the brain samples) within 72 h of staining.

Target	Channel	Detection dye	Peak emission wavelength (nm)	Dye dilution	Detection channel
Pol2ra	C1	Opal 520	525	1:1500	FITC
Ppib	C2	Opal 570	570	1:1500	Cy3
Ubc	C3	Opal 690	694	1:1500	Cy5.5
Dapb	C1, C2, C3	Opal 520/570/690	525, 570, 694	1:1500	FITC, Cy3, Cy5.5
Cd79a	C1	Opal 520	525	1:1500	FITC
Cd4	C4	Opal 650	650	1:1500	PE/Cy5
Cxc3r1	C2	Opal 650	650	1:1500	PE/Cy5
Il10	C3	Opal 570	570	1:750	Cy3
Il10ra	C4	Opal 540	536	1:1500	FITC/Cy3

Table 1. RNAScope targets and detection specifications.

QuPath image analysis

Once imaged, QuPath 0.3.1 Software was used to quantify the number of transcripts for each target probe[3]. Negative control images were generated by probing spleen and brain tissue sections with the RNAScope 3-plex negative control probes. Fluorescence measurements for each detection channel in the negative control images were subtracted from final experimental images to determine background fluorescence thresholds. Subtracting background fluorescence in this way ensures that all detected fluorescent spots represent true signal from RNA transcripts. Using in-built QuPath annotation tools, one large region of interest (ROI) was specified on each image such that the whole image was encompassed in a single annotation. The "Cell Detection" function was used to determine the number and position of cells in each ROI based on the DAPI nuclear stain (under the assumption that one nucleus represented one cell), and the 'Subcellular Detection' function was used to calculate the number of transcripts for each target. The accuracy of QuPath's automated annotation features have been confirmed by comparison to annotation by expert human analysts[3]. QuPath output data were then used as input data for FISHtoFigure. The analysis workflow was scripted to enable batch processing of all images within each dataset.

Results

We designed FISHtoFigure to facilitate the conversion of QuPath-quantified image data into transcript abundance analytics. We designed a simple graphical user interface and packaged the FISHtoFigure software as a standalone executable program, enabling analysis to be conducted with no interaction with the raw data or underlying Python code. Below we outline the steps involved in analysing smFISH data using FISHtoFigure, along with examples of analysis outcomes.

Step 1: Data harvesting and validation of quantified smFISH data

First, cellular boundaries and mRNA transcripts were identified using QuPath. QuPath output data were then processed using FISHtoFigure to produce differential transcript abundance analytics for different cell types or expression profiles[3]. An overview of the FISHtoFigure pipeline is given in Fig. 1a and an example of a typical image for processing is given in Fig. 1bi.

As experiments usually require numerous individual images, we created a dedicated pre-processing tool to concatenate individual QuPath-quantified image datasets into a single file comprising data from any number of smFISH images, which can then be analysed by the main FISHtoFigure program. Due to the volume of information captured during imaging, the resulting quantified files are large and include metrics not relevant for transcript expression analysis (e.g. morphometric data, such as, cell area, nucleus and cytoplasm morphology, etc.). The desired information, i.e., the number of transcripts per cell and fluorescent intensity data, which comprise only a small portion of the quantified data, are extracted by FISHtoFigure from QuPath-quantified smFISH data files and assigned to the cells from which they originate. Metrics are then calculated for each cell, i.e. the number of transcripts and total fluorescent intensities for each mRNA target. In addition to transcriptome information, cell location information is extracted in the form of the cell centroid (based on nuclear staining identified using the "Cell Detection" function in QuPath). Further information on the flags FISHtoFigure uses to harvest data are provided in the "FISHtoFigure v1.0.1 User Guide" document in the tool's GitHub repository and is recommended for developers wishing to further develop the FISHtoFigure tool. These data are then processed by FISHtoFigure using the "Plot Transcript Distribution" feature, which produces a scatter plot of points representing cell centroids, with points sized by number of mRNA transcripts within the cell and coloured by gene (Fig. 1bii). This allows users to visualise quantified data in a format analogous to the original smFISH image (Fig. 1bi) and, by overlaying this visualised data with the original smFISH image, directly validate the accuracy of data extraction by FISHtoFigure (Fig. 1biii).

Step 2: Differential target abundance analysis from smFISH data using the FISHtoFigure package

Following data extraction and assignment of transcript information to cells, differential transcript abundance analysis can be conducted using FISHtoFigure's "Transcript abundance analysis" feature.

Using our spleen dataset, we investigated T-cell and B-cell populations in the spleens of mice, either uninfected or 6 days after infection with influenza A virus. These cells are highly abundant in spleen tissue and have a classically "round" cellular morphology. Their morphology enabled easy identification of cell boundaries in QuPath, and thus generated a straightforward dataset for software validation. Spleen sections from naïve and infected mice were stained using DAPI to identify cell nuclei and probed for *Cd4* and *Cd79a* mRNA transcripts, enabling us to identify helper T-cells and B-cells, respectively[8,9]. This analysis revealed a statistically significant upregulation of *Cd4* expression within the T-cell population during infection (p < 0.01, Mann–Whitney test; Fig. 2a), while no statistically significant difference in *Cd79a* expression was observed. In addition to graphical outputs, FISHtoFigure analysis is saved in CSV format, meaning further downstream analysis can be performed using a wide variety of platforms (R, Microsoft Excel, etc.). Here, statistical analysis was performed on the FISHtoFigure output data using GraphPad PRISM.

We expanded the analysis capabilities of FISHtoFigure by adding the "Multi-target transcript abundance" feature, enabling the identification and quantification of cell types with multiplex transcriptomic profiles. This feature can be used to identify cells expressing any combination of mRNA transcripts. Here, we present an overview of the capabilities of this feature of FISHtoFigure and example analysis on the brain dataset. Comprehensive information on the underlying code and processing is available in the "FISHtoFigure v1.0.1 User Guide" document in the tool's GitHub repository. Here, we used this feature to validate the cell type quantification of our pipeline. *Cd4* and *Cd79a* are well established markers for helper T-cells and B-cells respectively[8,9]. Spleen resident B-cells do not express *Cd4*, and T-cells do not express *Cd79a*. Therefore we used the double-positive
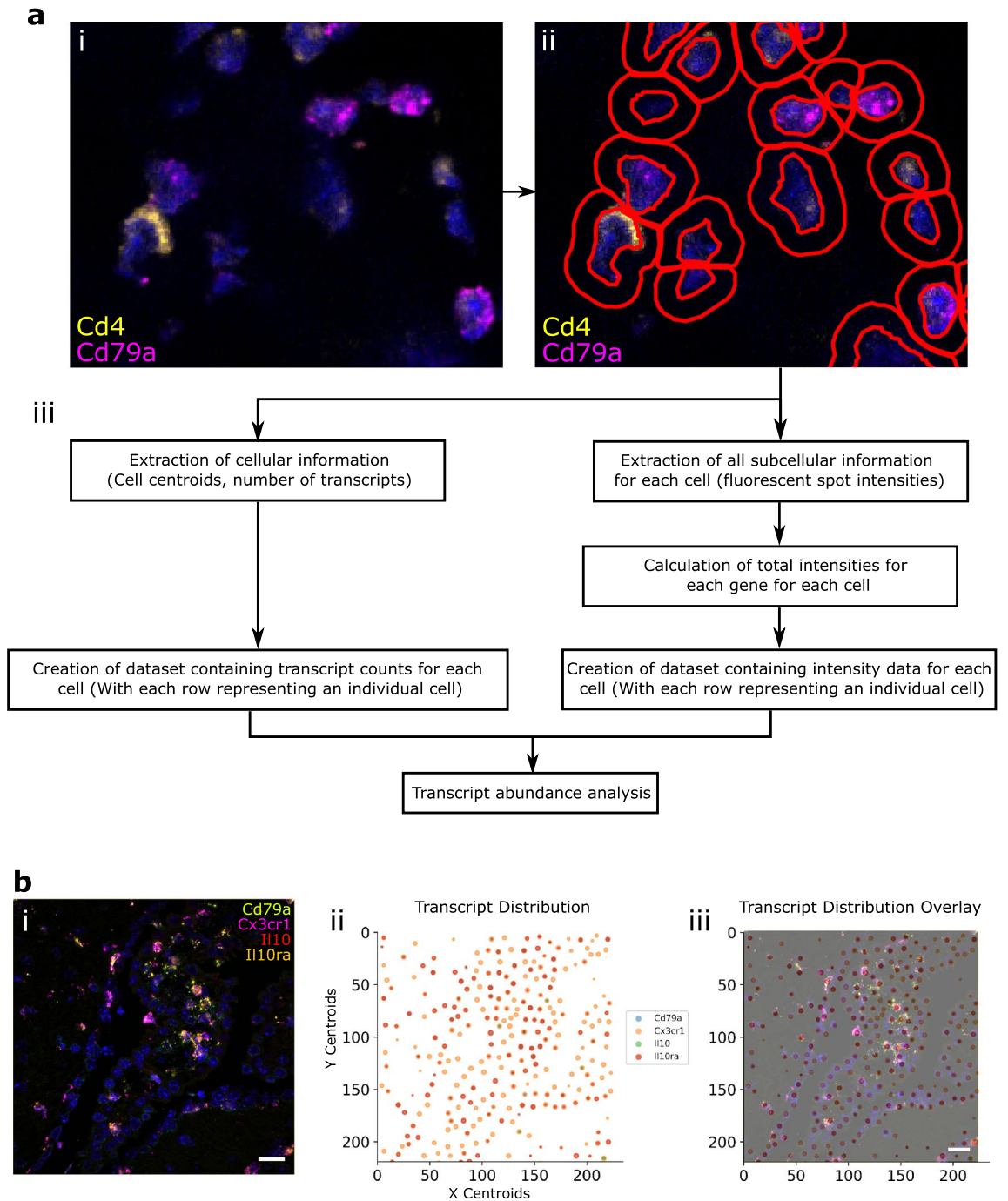
Figure 1. FISHtoFigure pipeline. (**a**) (i) An smFISH image from the spleen dataset captured via confocal microscopy (Zeiss LSM 880). (ii) QuPath's "Cell Detection" function was used to identify cell boundaries (shown in red). Cell nuclei identification is based on fluorescence above background in the channel associated with the nuclear stain (DAPI). (iii) An overview of FISHtoFigure processing of QuPath output data to generate transcript abundance outputs. (**b**) (i) An smFISH image from the brain dataset (scale bar = 20 μm), captured by confocal microscopy (Zeiss LSM 710) and (ii) processed using FISHtoFigure's "Plot transcript distribution" function, where points represent cells and are sized based on the number of transcripts being expressed by that cell. (iii) An overlay of the captured smFISH image with the plot produced by FISHtoFigure demonstrates the accuracy of the pipeline.

$Cd4^+$ $Cd79a^+$ cell population as a metric for mis-categorisation of cells by FISHtoFigure. The naïve dataset comprised a total of 1229 cells of which 273 contained transcripts of *Cd4* or *Cd79a*. A total of 18 cells were labelled as $Cd4^+$ $Cd79a^+$, representing approximately 1.5% of all cells and 6.6% of transcript-expressing cells (Fig. 2b).
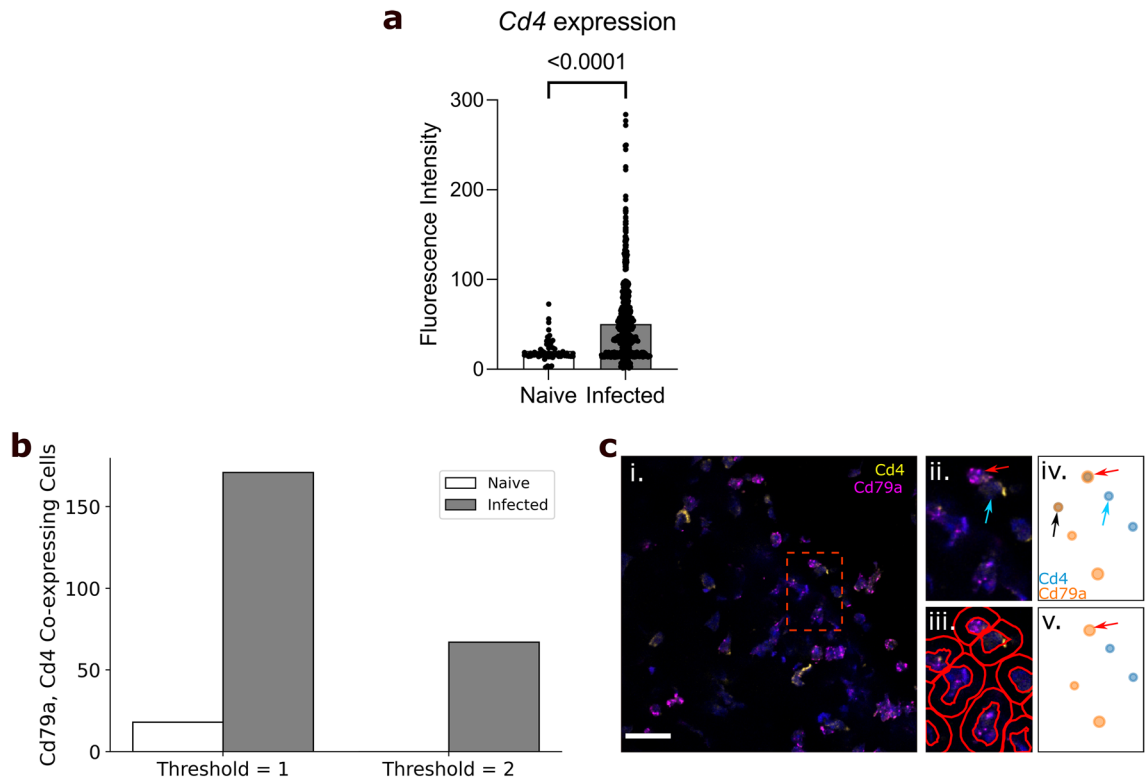
Figure 2. Analysis of spleen samples from naïve and influenza A virus infected mice. (**a**) FISHtoFigure quantification of *Cd4* expression within *Cd4*+ cells in the naïve (n = 1228) and infected (n = 1486) spleen datasets, significantly upregulated during infection (Mann–Whitney test). Bars represent mean values across all cells in the dataset, each dot represents a *Cd4*+ cell in the dataset. (**b**) Total number of cells co-expressing *Cd79a* and *Cd4*, with threshold set to 1 or 2 transcripts. (**c**) (i) An smFISH image from a naïve spleen (scale bar = 20 μm), captured by confocal microscopy (Zeiss LSM 880). (ii) A zoomed view of the region shown in the red square in (i) shows a B-cell (red arrow) and T-cell (blue arrow) in close proximity. (iii) Cell boundaries identified using QuPath. (iv) FISHtoFigure's "Plot Transcript Distribution" feature with a threshold of 1 transcript per cell, (v) with a threshold of 2 transcripts per cell. Setting a threshold of 2 transcripts per target per cell results in the B-cell being correctly categorised (red arrow)—note the removal of the ambiguous *Cd79a*+ *Cd4*+ cell expressing both transcripts as they are below threshold levels (black arrow in (iv)).

The infected dataset comprised a total of 1487 cells of which 882 contained transcripts from either marker. The infected dataset showed a higher presumed mis-categorisation rate, with 171 cells (11.5% of all cells and 19.4% of transcript-expressing cells) labelled as *Cd4*+ *Cd79a*+ (Fig. 2b).

Upon closer inspection of the quantified data, many of the apparently *Cd4*+ *Cd79a*+ cells contained a majority of transcripts from one gene, suggesting that mis-categorisation typically resulted from a small number of transcripts from the other gene. This could be plausibly explained if incorrect boundary approximations caused a small proportion of transcripts to be mis-allocated between highly localised cells. For example, a B-cell in close proximity to T-cell might appear to contain a single *Cd4* transcript due to cell boundary approximation (Fig. 2c). In such cases, it is reasonable to assume the cell identity based on the majority transcript. To address this, we introduced a thresholding feature so that users can define the minimum number of transcripts from each mRNA target required for cells to be included in analysis. By setting this threshold at 2 transcripts from each mRNA, the population of *Cd4*+ *Cd79a*+ cells was eliminated in the naïve dataset and substantially reduced (67 cells, representing 4.5% of all cells and 7.5% of transcript-expressing cells) in the infected dataset (Fig. 2b). This was consistent with the model that *Cd4*+ *Cd79a*+ cells were artefacts, and showed that thresholding allowed this source of error to be controlled.

Having demonstrated that FISHtoFigure can quantify cell types based on mRNA expression profiles, we progressed to a more challenging system containing cells with less regular boundaries. To do this we examined sections of mouse brains, which contain highly ramified cell types, using data from a study exploring the interactions between regulatory B-cells (Bregs) and microglia during infection with *T. b. brucei*[4]. The brain dataset comprised 17 images captured from brain sections of infected mice and 9 captured from uninfected (naïve) controls. Brain sections were stained using DAPI and probed for *Cd79a* (a B-cell marker), *Cx3cr1* (a microglia marker), *Il10* (an anti-inflammatory cytokine hypothesised to be involved in Breg–microglia interactions), and *Il10ra* (the receptor for *Il10*)[9,10]. These images were quantified in QuPath and concatenated into two datasets comprising naïve control data and infected data.

Cx3cr1 is a well-established microglia marker[10]. B-cells do not express *Cx3cr1* and microglia do not express *Cd79a*. Similarly to the spleen dataset, in order to examine to what extent the thresholding function could

improve cell type quantification in data containing ramified cells, presumptively mis-categorised *Cd79a⁺ Cx3cr1⁺* cells were quantified. The naïve dataset contained 1631 cells, 914 of which contained transcripts. 30 cells were labelled *Cd79a⁺ Cx3cr1⁺* double-positive (1.8% of all cells, 3.3% of transcript-expressing cells). The infected dataset contained 3907 cells, of which 3332 contained transcripts, 392 were labelled as *Cd79a⁺ Cx3cr1⁺* double-positive (10% of all cells, 11.7% of transcript-expressing cells; Fig. 3a). Applying a threshold of 2 transcripts per mRNA per cell reduced the number of *Cd79a⁺ Cx3cr1⁺* double-positive cells to 4 in the naïve dataset (0.2% of all cells, 0.4% of transcript-expressing cells), and 76 in the infected dataset (1.9% of all cells, 2.3% of transcript-expressing cells; Fig. 3a). This demonstrated that applying thresholds for transcript abundance could allow accurate allocation of transcripts to cells even for cells with complex and irregular boundaries.

Finally, as a demonstration of the application of FISHtoFigure in an experimental workflow, we re-analysed data that we had collected as part of a study of Breg-microglia crosstalk in the brains of mice infected with *T. b. brucei*[4]. Briefly, single cell and spatial transcriptomic analyses of infected mice revealed an upregulation of the anti-inflammatory cytokine *Il10*, along with Breg and microglia associated transcripts, in the brains of *T. b. brucei* infected mice. We tested the hypothesis that during infection *Il10* expression governed crosstalk between Bregs and microglia in the brain, using smFISH and FISHtoFigure to investigate the localisation of transcripts. FISHtoFigure's "Transcript abundance analysis" function revealed a statistically significant upregulation in *Cd79a* and *Il10* expression in infected specimens compared to naïve controls, in agreement with results from single cell transcriptomics[4]. Graphical outputs in the format produced by FISHtoFigure are presented in Fig. 3b (p < 0.01, Mann–Whitney test; data from Ref.[4]). We then used a variety of analyses to validate that this crosstalk was driven by two specific cell types (*Il10⁺* Bregs and *Il10ra⁺* microglia), including visualising these cell types using smFISH. Here, we expand on this analysis by using FISHtoFigure to directly quantify the abundance of two different double-positive cell types in infected and naïve mice. We used FISHtoFigure's "Multi-target transcript abundance"
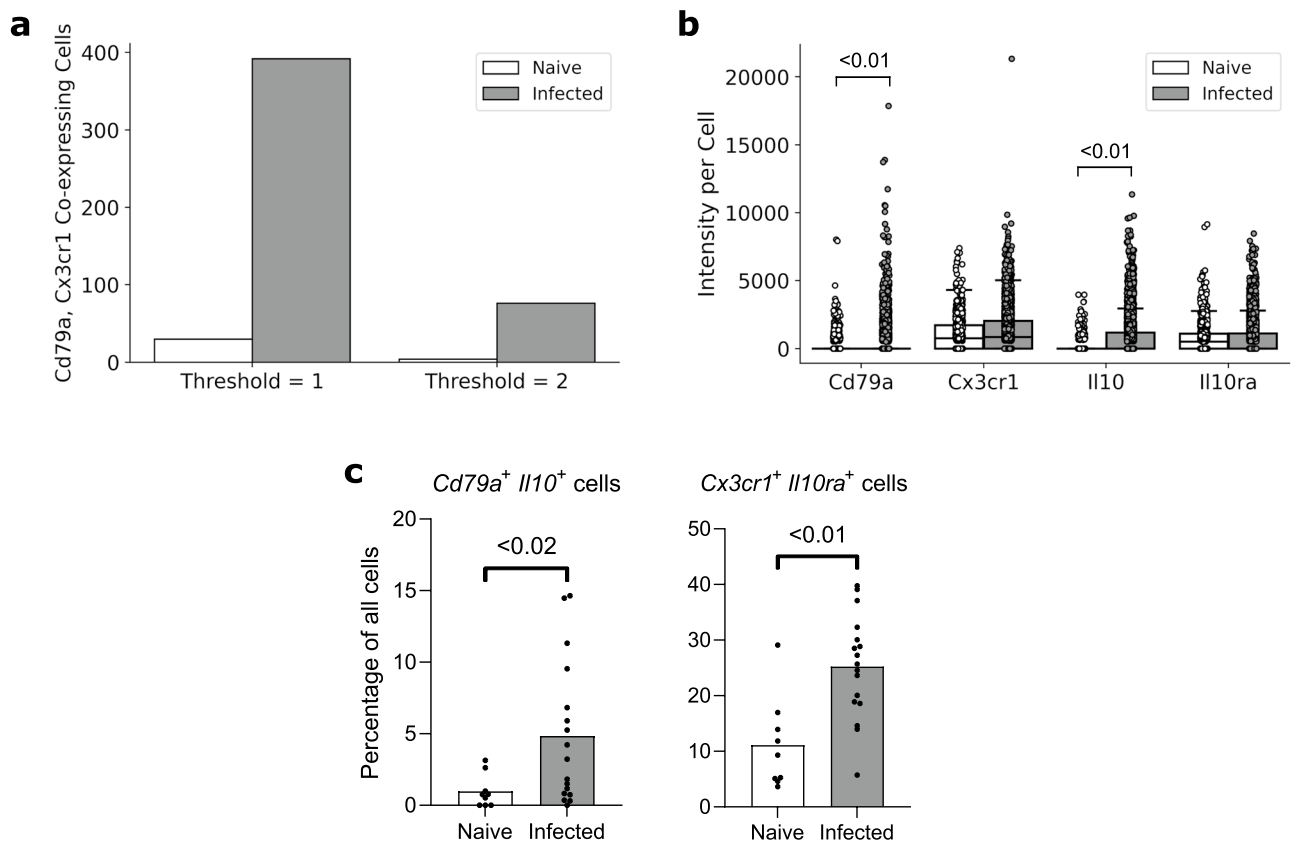
Figure 3. Examples of FISHtoFigure outputs from analysis of the brain dataset. (**a**) Total number of cells expressing both *Cd79a* and *Cx3cr1* with threshold of either 1 or 2 transcripts per mRNA per cell. (**b**) Fluorescence intensities for each marker all cells which express a transcript from at least one marker, where each point represents a cell, showing that *Cd79a* and *Il10* expression are significantly upregulated during infection (Mann–Whitney test). Box limits are defined by the interquartile range (IQR) with whiskers extending to the lowest/highest data point still within 1.5 IQR of the lower/upper quartile. This figure re-plots data originally collected in Ref.[4]. (**c**) Percentage of cells expressing both *Cd79a* and *Il10* (Bregs) and percentage expressing both *Cx3cr1* and *Il10ra* (Microglia) in smFISH images from the naïve and infected datasets, showing that both are significantly upregulated during infection (Mann–Whitney test). Bars represent mean values across all images in the dataset, each dot represents a single image from the dataset. Percentages were taken from each image in the naïve (n = 9) and infected (n = 17) datasets individually and statistical analysis was performed in GraphPad PRISM, a non-parametric test was selected due to lack of normality.

feature to analyse *Cd79a*+ *Il10*+ Breg populations and *Cx3cr1*+ *Il10ra*+ microglia populations in naïve and infected specimens. This analysis confirmed that during infection there was an upregulation of both *Cd79a*+ *Il10*+ Bregs (Fig. 3c; p < 0.02, Mann–Whitney test) and *Cx3cr1*+ *Il10ra*+ microglia (Fig. 3c; p < 0.01, Mann–Whitney test). In the context of the current paper, this demonstrates that FISHtoFigure can accurately quantify the abundance of specific cell types, including those with irregular boundaries, using multiplex expression profiles. Taken together, these findings demonstrate the value of FISHtoFigure in an experimental workflow.

Discussion

FISHtoFigure automates the extraction and processing of transcriptomic data from QuPath-quantified smFISH data, allowing users to analyse specific transcript expression profiles in datasets that would otherwise be very difficult to parse.

Our tool is capable of analysing smFISH data by any number of mRNA targets and quantifying cell types and expression profiles with a high accuracy. Furthermore, the graphical user interface allows users to specify a positivity threshold for transcript abundance analysis (i.e., the number of transcripts required for a cell to be marked as positive, and by extension, be included in analysis), allowing users to directly control the sensitivity of the FISHtoFigure platform individually for each set of analyses.

Current analysis packages for smFISH data are largely focused on quantification and labelling of transcripts and only offer limited downstream transcript abundance analysis options, which require programming experience to implement. For example, *FISH-quant* provides a means to detect transcripts in smFISH data and assign individual transcripts to cells and subcellular compartments[11]. *FISH-quant* offers downstream analysis options for mRNA expression, but this analysis is largely focused on the intracellular distributions of transcripts rather than the quantification of cells that express multiple mRNA targets. Another smFISH analysis tool, *dotdotdot*, outputs quantified cell and transcript data in a format interpretable using R or Python. However, programming experience is required to implement downstream analysis[12].

FISHtoFigure facilitates custom differential transcript and cell type abundance analyses without the need for custom code. By providing multi-transcript analysis tools in an intuitive package, FISHtoFigure significantly broadens the accessibility of smFISH analysis.

Comparison of FISHtoFigure's spatial distribution plots with the confocal microscopy images from which they were derived demonstrates high levels of concordance between raw and quantified data (Fig. 1b). We demonstrate that FISHtoFigure can accurately determine cell profiles in different biological systems, and that the in-built thresholding feature can substantially reduce mis-categorisation caused by the close proximity of different cell types (Fig. 2). It is advised that users adjust this threshold to obtain the best results from different experimental designs. In order to determine the most appropriate value for this threshold for each experiment presented here, combinations of transcripts which were not expected to be representative of any cell type in each dataset were analysed. This allowed us to adjust the false discovery rate: the threshold was adjusted until the number of cells expressing these "impossible" combinations of transcripts was below a predetermined percentage of the total number of cells in the dataset. We suggest that users conduct similar analyses of their own data in order to determine an appropriate threshold value for each experiment. In the spleen dataset, applying a threshold of 2 transcripts per mRNA target per cell completely removed all mis-categorised cells in the naïve dataset and reduced mis-categorisation by > 60% in the infected dataset (Fig. 2b). In the brain dataset, applying a threshold of 2 transcripts per mRNA target per cell reduced mis-categorisation of cell types by > 80% in both the naïve and infected datasets (Fig. 3a).

In the brain dataset, FISHtoFigure enabled rapid analysis of smFISH data which would otherwise require considerable time investment and programming experience. FISHtoFigure analysis reveals a statistically significant (p < 0.01, Mann–Whitney test) upregulation in expression of *Cd79a* and *Il10* during infection (Fig. 3b). The ability to analyse and plot cellular information for specific cell types with multiplex transcriptional profiles allowed us to identify the upregulation of *Cd79a*+ *Il10*+ Bregs and *Cx3cr1*+ *Il10ra*+ microglia in infected specimens compared with controls, a difference which would otherwise require custom code to assess (Fig. 3c).

Considerations and limitations

The labelling of transcripts is the first step of our quantification pipeline and poor sensitivity or specificity at this step will have a compounding effect on the accuracy of cell type analysis performed by FISHtoFigure. Many smFISH methods are available, and in this study we used the RNAScope assay, a well-established and widely-used method[1]. The RNAScope assay incorporates various features to ensure accurate and reproducible labelling of transcripts, such as a pair-wise probe design in which probes will not fluoresce unless adjacent probes also bind, significantly reducing the signal from non-specific binding[1]. Additionally, the RNAScope assay includes a set of positive control probes specific to the species the tissue samples are taken from. These probes target common housekeeping genes present in all cell types in the sample. Because the abundance levels of these target genes are well characterised (and by extension fluorophores will bind to these targets at known levels), processing samples with these probes prior to the final experiment provide users with a means to check that fluorophores used to visualise target probes are binding correctly and are of approximately equal brightness. Finally, the assay includes a set of negative control probes which target genes which are not expressed by any cell in the sample, and therefore any fluorescence observed in images of samples processed with these probes can be treated as background signal (generally arising from fluorophore remaining in the tissue after the wash steps). The maximum fluorescence for each fluorophore in these images can then be used to set minimum detection thresholds in experimental data (i.e. any fluorescent signal below this threshold treated as background and is removed). Using this approach, we can be confident that quantified fluorescent spots in experimental data represent true signal from transcripts.

The variety of smFISH methods available means that users may wish to analyse different formats of input data. FISHtoFigure was designed for the analysis of QuPath output files, but we have intentionally built FISHtoFigure as a modular tool, separating the data harvesting step (in which data is pulled from the QuPath output) from the analysis steps. As a result, the data harvesting section can be adjusted easily, without interfering with any downstream analysis steps. At present, to modify FISHtoFigure to work with data quantified using another platform (e.g. CellProfiler) simply requires changes to the specific flags which the program uses to identify information in the quantified data file. Information on how to make these changes to the tool are available in the "FISHtoFigure v1.0.1 User Guide" document in the FISHtoFigure GitHub repository.

Regarding identifying cell boundaries, QuPath has the capacity to quantify cell boundaries based on nuclear staining, or via a fluorescent membrane marker. Here, cell nuclei were identified via fluorescent DAPI staining, and cell boundaries were approximated by applying a set radius based on tissue cell type composition to each identified nucleus using the "Cell Detection" function in QuPath. Though we demonstrate that this can allow the accurate quantification of cells, even for cell types with irregular boundaries, further improvements in the determination of cell boundaries, and by extension cell expression profiles, could likely be achieved through adjustments in sample preparation. Though the threshold function included within FISHtoFigure can be used to eliminate the majority of cell type misclassification events, the use of a membrane marker would further improve cell type quantification. We advise the inclusion of a membrane marker if users find that they cannot sufficiently eliminate misclassification events using the inbuilt thresholding function.

Conclusion

The problem of balancing accessibility for non-specialist users and analytical scope is an important consideration in the development of software tools. Here, we present FISHtoFigure, an analytical platform for QuPath-quantified smFISH data capable of analysing specific cell types and multiplex transcriptomic profiles and of generating a variety of differential transcript abundance analytics for cells expressing a user-specified combination of mRNA transcripts. In the interest of accessibility for users with all levels of computational image analysis experience, we have created a simple graphical user interface and packaged FISHtoFigure as an executable program, thus allowing transcript expression analysis without interaction with raw quantified image data or custom analysis scripts. FISHtoFigure can therefore expand the in-house analysis capabilities of many research groups investigating transcriptomics via smFISH.

Data availability

All code involved in the production of the FISHtoFigure package and all analysis presented here is available on GitHub: https://github.com/Calum-Bentley-Abbot/FISHtoFigure.git Data are available under the terms of the MIT open access licence (https://opensource.org/license/mit/).

References

1. Wang, F. *et al.* RNAscope: A novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagn. JMD* **14**, 22–29 (2012).
2. Marx, V. Method of the year: Spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
3. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
4. Quintana, J. F. *et al.* Single cell and spatial transcriptomic analyses reveal microglia-plasma cell crosstalk in the brain during *Trypanosoma brucei* infection. *Nat. Commun.* **13**, 5752 (2022).
5. Reback, J. *et al.* pandas-dev/pandas: Pandas 1.4.3. (2022) https://doi.org/10.5281/ZENODO.3509134.
6. Caswell, T. A. *et al.* matplotlib/matplotlib: REL: v3.4.3. (2021) https://doi.org/10.5281/ZENODO.5194481.
7. Waskom, M. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
8. Luckheeram, R. V., Zhou, R., Verma, A. D. & Xia, B. CD4+T cells: Differentiation and functions. *Clin. Dev. Immunol.* **2012**, 925135 (2012).
9. Mason, D. Y. *et al.* CD79a: A novel marker for B-cell neoplasms in routinely processed tissue samples. *Blood* **86**, 1453–1459 (1995).
10. Wolf, Y., Yona, S., Kim, K.-W. & Jung, S. Microglia, seen from the CX3CR1 angle. *Front. Cell. Neurosci.* **7**, 26 (2013).
11. Imbert, A. *et al.* FISH-quant v2: A scalable and modular tool for smFISH image analysis. *RNA* **28**, 786–795 (2022).
12. Maynard, K. R. *et al.* dotdotdot: An automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues. *Nucleic Acids Res.* **48**, e66 (2020).

Acknowledgements

Author contributions

C.B.A. was responsible for the conceptualisation, experimentation and software produced in the study, as well as the writing and editing of the manuscript. R.H. was responsible for conceptualisation and experimentation. P.C., C.P. and E.R. were responsible for experimentation. E.R., E.H., G.M., and A.M.L. were responsible for the supervision of this work writing and editing of the manuscript. All authors contributed to the editing of the final manuscript draft for publication.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.B.-A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.