Documentation to image analysis

The documentation below provides a detailed explanation and pointers towards the use of the MS-ECS-2D update site. It is made available in conjunction with the publication of the paper:

Stapel, L. C., Lombardot, B., Broaddus, C., Kainmueller, D., Jug, F., Myers, E. W., & Vastenhouw, N. L. (2015). Automated detection and quantification of single RNAs at cellular resolution in zebrafish embryos. Development, dev-128918.

The documentation details:
- ***The PathFinder* Fiji plugin**
- **The *Cell annotation* plugin in Fiji**
- **The *Transcript analysis* plugin in Fiji**

Documentation to *PathFinder* Fiji plugin

This document describes the setup and usage of the *PathFinder* Fiji plugin for cell segmentation of tissue sections with a single bright membrane channel. The input to the *PathFinder* plugin are membrane and vertex classified images as obtained by the *MS-ECS-2D* KNIME workflow.

1. Installation

Open Fiji, go to *Help → Update… → Manage update sites* and make sure the following URL is added:

```
http://sites.imagej.net/MS-ECS-2D/
```
and under Host add:
```
webdav:MS-ECS-2D
```

This adds a collection of files to the ImageJ Updater site that you need for the automated detection of single RNAs at subcellular resolution. In case you are only interested in using the *PathFinder* plugin, you can choose to install only those files with the prefix Path_Finder/. You can search for Path_Finder after pressing *Advanced mode*. Select these files and click *Apply changes*, which installs them to the standard location, resulting in paths:
```
[Fiji home]/plugins/Path_Finder/*
```
Go to *Help → Refresh Menus*. You should see *Plugins → Path Finder → main*, if not try reopening Fiji.
**NOTE**: of course you can also choose to install all files of the update site. The usage of the *Cell annotation* and *Transcript analysis* plugins to analyze single RNAs at subcellular resolution is explained in their respective documentations.

2. Overview

The *PathFinder* plugin allows you to perform membrane segmentation across several images simultaneously. Before you begin, you should make sure that the storage of your images adheres to the following directory structure:

```
├── [original image folder]
│   ├── [original image 1].tif
│   ├── [original image 2].tif
│   ├── …
│   └── [original image n].tif
├── [memgrane classifier output folder]
│   ├── [classifier output 1].tif
```

```
        ├── [classifier output 2].tif
        ├── …
        └── [classifier output n].tif
└── [vertex classifier output folder]
        ├── [classifier output 1].tif
        ├── [classifier output 2].tif
        ├── …
        └── [classifier output n].tif
```
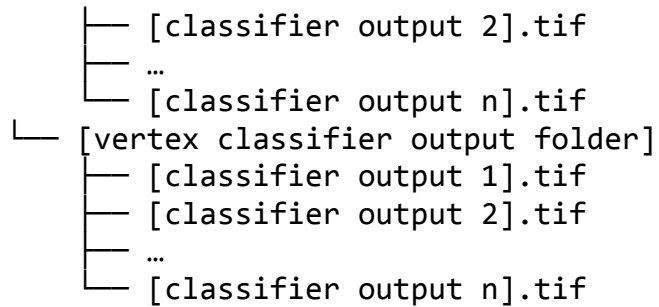
The original image folder contains the same images that were used as input to the membrane and vertex classification procedures. You are free to use whichever file-naming convention is convenient, provided that the images in these different directories match up correctly when listed in alphabetical order. By default, the classification procedure names its output files in a way that maintains this order. The plugin operates on each image in sequence, and will save the segmented version of each image to the location:

`[original image folder]/results/[original image n]_scaled.tif`

Each of these resulting images has three layers (Figure 1):

1. **segmented cells** : Each set of pixels belonging to the same cell has the same pixel value.
2. **original** : The original image from [original image folder].
3. **paths** : A binary image where all pixels belonging to the final set of paths have value 255.

The width and height of the result is equal to that of the original image that was provided, even when the output of the membrane and vertex classifiers may be downsized by a small factor for computational efficiency.
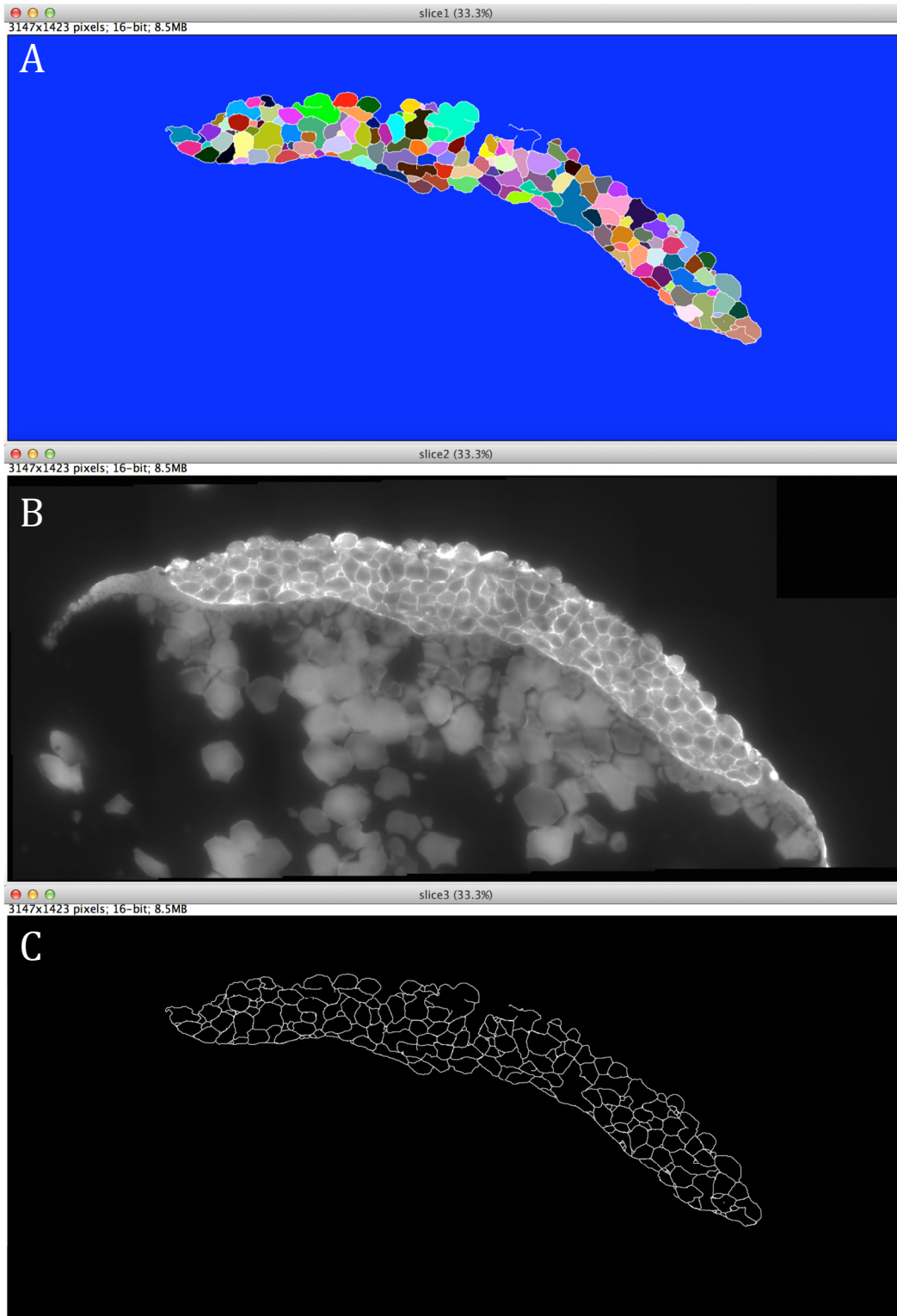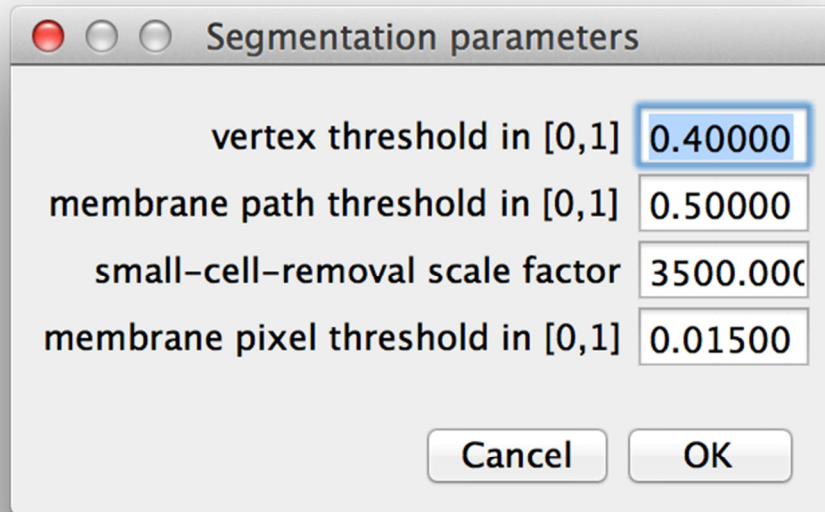
Figure 1. Example of output images from the pathfinder plugin. (A) Segmented cells. (B) Original membrane image. (C) Paths.

3. Usage Details

PathFinder

Run the plugin with *Plugins → Path Finder → main*. The following screen will appear:

The `vertex threshold` parameter determines the threshold probability above which pixels from the vertex probability map image are classified as vertices.
If the `membrane path threshold` is smaller than the average probability across all pixels in a path, then the path is removed.
The `small-cell-removal scale factor` is used after the initial path construction in the *small-cell-removal* phase where we remove bad edges that cause over segmentation. Increasing this value will cause the *small-cell-removal* step to keep higher (worse) scoring edges, and lowering it will cause more of them to be removed. Setting this parameter to zero will skip the small cell removal phase entirely.
The `membrane pixel threshold` parameter determines the threshold probability above which pixels form the membrane probability map image are classified as membrane. It is only within membrane regions that paths between vertices can be constructed.

After the parameters have been selected you will be asked to select the `original image folder`, `membrane image folder` and `vertex image` folder via a series of three folder-selection dialogs:
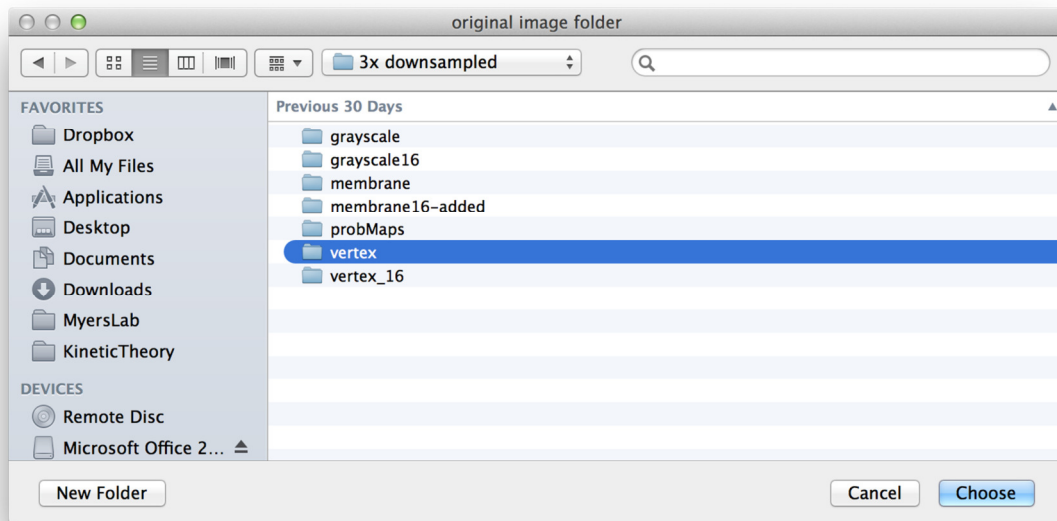
And you're done!

The plugin will briefly open new image windows in Fiji as it progresses through different stages of the algorithm, but no windows remain open when the plugin is finished running. Thus, it is currently not possible to continue to working on a separate image in Fiji simultaneously while the plugin is running, though this may change in future versions of the plugin. Finally, if there are any errors they will be written to the Fiji Console.

***

Documentation to the *Cell annotation* plugin in Fiji

The *Cell annotation* tool is an interface to manually correct a (2D) cell segmentation and to annotate cells once segmented. The tool is composed of a control window that allows choosing the annotation modes (label inspection, cell segmentation correction and cell type annotation) and a display window that shows an overlay of i. the masks to annotate and ii. the raw data that is used for the annotation. Once an annotation is finished, the annotation can be exported as a 2-channel image (cell labelling and cell annotation) that can be used as input to the *Transcript analysis* plugin (see corresponding documentation).

1. Installation

To use the plugin (assuming Fiji is installed):
- Install the plugin:
  - Start Fiji and go to the menu ***Help>Update…***
  - When the imageJ updater window is open click the ***manage update sites*** button and then check the ***MS-ECS-2D*** update site in the ***Manage update sites*** window.
  - **Close** the Manage update sites window, press **Apply changes** in the ImageJ Updater window and when prompted to, Close Fiji.
  - The plugin will be accessible in the menu ***Plugins>Cell_Transcript>Cell annotation.***

Note: these steps are not necessary if you have already activated the complete *MS-ECS-2D* update site to obtain the *Transcript analysis* or *Pathfinder* plugins.

2. Plugin usage

Before starting the plugin, load the cell segmented image (see Figure 1.A, output cell segmentation pipeline) you wish to annotate and the original raw data (see Figure 1.B). Both images should have the same number of pixels in the X and Y dimensions. Also, each cell region in the labelled image should be separated by a boundary of zeros (in white in Figure 1.A). Such a labelled image can be obtained as an output of the cell segmentation process.
You can then start the plugin by going to the menu:
Plugins>Cell_Transcript>Cell_annotation
A window will open requesting the following information (see Figure 1.C):
- **Original Image**: raw data that will guide the annotation
- **Cell Masks Image**: Masks to annotate. It can be a one-channel image showing cell segmentation or a two-channel image showing the cell segmentation in channel 1 and a previously generated cell tissue annotation in the second channel.

Cell annotation

- **Nuclei channel**: channel showing the nuclei in the raw data (first channel has value 1).
- **Membrane channel**: channel showing the nuclei in the raw data. Note that these channels will only be used for display and in practice any channel can be chosen. It is for example possible to select a channel that contains an antibody staining to facilitates cell type annotation.

When pressing OK an interface named **annotation control** opens (see Figure 1.D) as well as an image, the annotation display. The annotation display shows a maximum projection of the selected membrane and nuclei channels (see Figure 1.E), superimposed with a transparent overlay of the cell label. The name of the annotation display always starts with "[Annotation tool]".

Annotation Modes and control description:
The Annotation control window allows selecting the interaction mode and the parameters for the annotation. By changing the selection in the 'Mode' drop down menu, one can switch between three types of interactions with the annotation display.
- **Label inspection** (see Figure 2.B): This mode enables visual inspection of the cell segmentation output. In this mode the original image is overlaid with the image of labelled cells regions. When switching to this mode the image is automatically relabelled to account for recent adjustments of the cell segmentation. Besides the usual Fiji zooming (by pressing "+" or "-" keys) and panning (space bar + mouse drag) options no interaction is possible with the annotation display in that mode.
- **Correction** (see Figure 2.C and Figure 3): In this mode the cell segmentation can be corrected. The cell borders are shown in red as an overlay to the projection of the original image. On top of zooming and panning this mode allows modifying the cell regions by drawing and erasing boundaries between them. To add boundaries move the mouse while pressing the mouse left button. The same action while pressing the "Alt" key allows erasing boundaries. The size of the brush to draw or erase boundaries can be set in the field **Brush size**.
- **Annotation** (see Figure 2.D): In this mode, cells can be assigned to a specific cell or tissue type to create a cell type mask. An overlay showing how each region is associated to a cell type is shown. If no cell annotation was present with the cell labelling a default annotation where all regions are of type "DEL" is created. Cells can be annotated with the currently selected cell type when a cell is clicked on. For the purpose of our analysis we defined 4 types or regions "**EVL**", "**YSL**", "**DEL**", "**outside**". EVL and DEL correspond to cell regions while YSL is for the zebrafish yolk syncytial layer and "outside" literally is used to indicate regions outside of the embryo. The code can be adjusted to facilitate the annotation of user-selected cell types.

Cell annotation

In all modes the transparency of the overlay can be adjusted using the **Overlay alpha** slide bar.

Every time the mode is changed the masks are updated to ensure consistency between the overlays. This behaviour can also be triggered by clicking the "**Relabel**" button.

When one is satisfied with the cell labelling and the cell type annotations, the corresponding mask can be exported to the Fiji interface by pressing the "**get Masks**" button. This image can be used as input to the *Transcript analysis* plugin to obtain cell (type) specific data on gene expression.
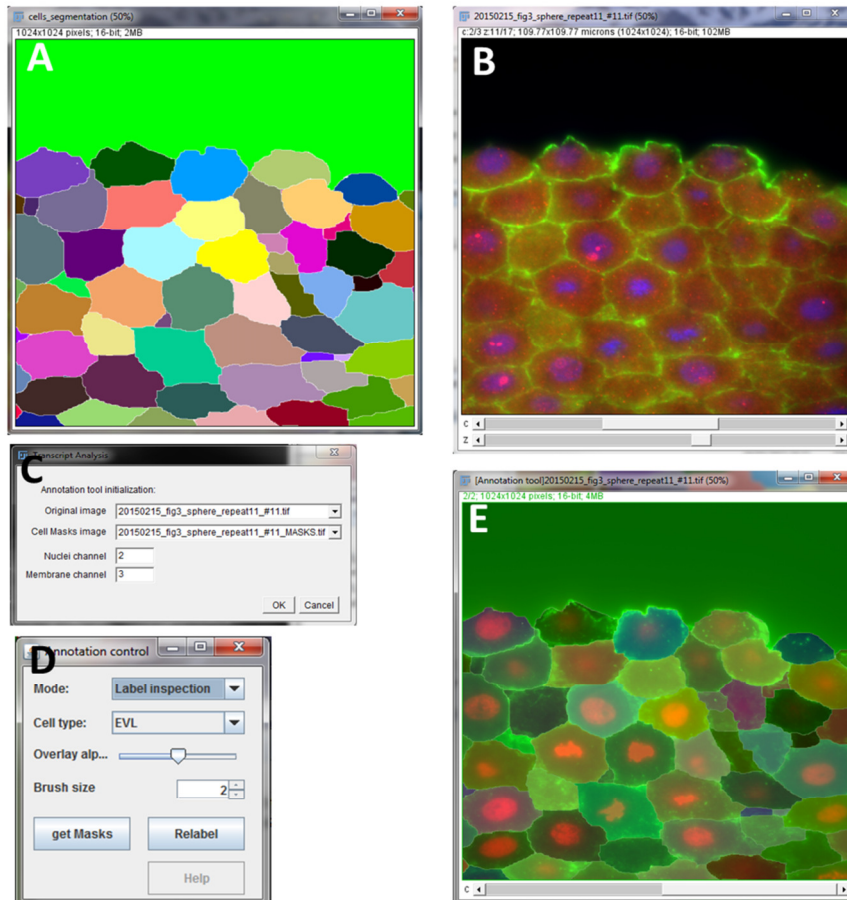

Figure 1: Typical input data and annotation tool interfaces. (A) Cell label output from the *PathFinder* tool. (B) Raw data. (C) Opening window of the Annotation tool. (D) Annotation control window. (E) Maximum projection of the raw data with an overlay of the cell labels.
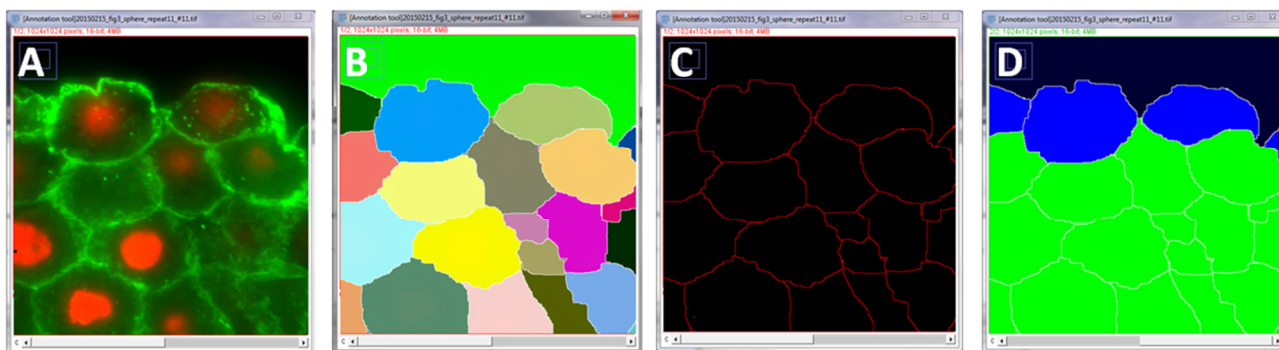
Figure 2: Illustration of the tool display overlays for the different modes of interaction. For the raw data (A) the Annotation tool overlays cell labels in "Label inspection mode" (B), red cell contours in "correction" mode (C) and cell type labels in "annotation" mode (D).
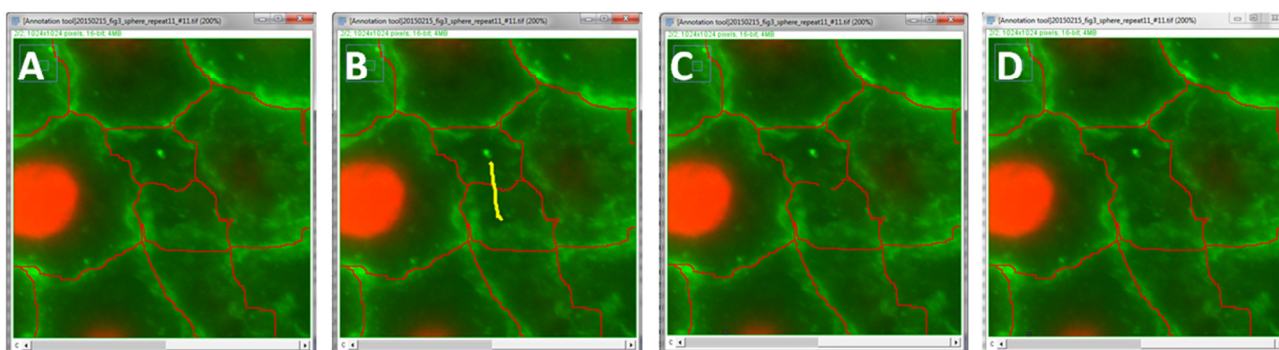

Figure 3: The correction mode allows to fuse (or split) regions by removing (or adding) boundaries. Setting the mode to "Correction" overlays regions boundaries (in red) to the raw data (A). Dragging the mouse while pressing the mouse left button and alt key draws a yellow trace (B). When releasing the mouse the piece of boundary below the trace is erased (C). When clicking the relabel button those boundaries that are not separating 2 objects are removed (D)

3. Methods description

The annotation tool relies on 3 images that are overlaid on top of the raw data and can be updated by user interaction: an image of the cell labels, an image of the cell boundaries, an image of the cell type (where images are coloured according to the type of the region). In this section we describe how these images are initialized and updated when the user interacts with the annotation tool.

- **Initialization of the overlay images:** The labelled image provided as input is used as provided while the cell boundary overlay is obtained by thresholding the labelled image such that only the zero valued pixel are selected in the mask. A cell type annotation overlay can be provided in the second channel of the input mask. If not provided, the overlay for cell type annotation is created by setting cell boundaries to zeros while the rest of the image is set to a default value corresponding to the "DEL" cell type (the most common cell type in our images). Finally, images are cleaned to remove non-separating boundaries (see details of the cleaning process in the next section).

- **Cleaning of overlay images**: Label image cleaning refers to an operation removing non-separating boundaries in the cell segmentation images (i.e. boundaries that are surrounded by a unique label). These non-separating boundaries can erroneously be generated by the *PathFinder* plugin. The process used removes all non-separating boundaries less than or equal to 2 pixels width. Boundaries thicker than 2 pixels will remain. The filter works as follows: a dilated and a closed version (with radius set to 1 pixel) of the labelled regions are subtracted. In the resulting image the non-zero values match with the cleaned boundaries of the regions (i.e. non separating boundaries do not appear anymore). From this image we create an image that is zeros in the cleaned boundaries and 1 inside the regions. This image is multiplied with the dilated version of the original labelled image. The resulting image is a cleaned version of the original labelled image.

- **Update of overlay images after boundary modification:** To create an updated label image after modification of the cell boundaries, the cell boundaries image is inverted and a connected component labelling is applied to it. The resulting image is then cleaned (see previous bullet point) to give the new cell labels image. A new boundary image is deduced from the new cell labels image by selecting the zeros-valued pixels. Regarding the cell type mask, a map from region label to region type is built and used to recolor a copy of the new cell labels according to their regions type. If a region is split, the daughter regions will have the same region type as the initial region. If regions are merged the new region type will be according to the type of most bottom right region (i.e. in the group of regions to be merged). In this case, region type might have to be adjusted manually afterwards.

- **Region type annotation**: A flood-fill algorithm with 8-connected neighbourhood is initiated in the cell annotation overlay at the position of the mouse click. The value for the flood filling corresponds to the region type selected (EVL=1, YSL=2, DEL = 3, outside regions=4). The flood fill will update the cell type value of the cell that was clicked.

***

Documentation to the *Transcript analysis* plugin in Fiji (1.0.7)

The *Transcript analysis* plugin is an image analysis pipeline for the detection of transcripts and transcription foci in 3D sections of tissue. In particular it was developed to support the analysis of transcript abundance during zebrafish early development.

The pipeline is initiated with z-stack images of smFISH data that include nuclear detection, and optionally the output of the *Cell annotation* tool (see corresponding documentation). The plugin outputs a multichannel image illustrating the detections (transcripts, nuclei, cells), and text files indicating plugin parameters, cell and transcript descriptors and a few plots illustrating the detections.

The pipeline can be started with a user interface and from a script (for batch processing for instance). These usages are described in the following sections.

1. Installation

To use the plugin, assuming Fiji is installed:
- Install the 3D imageJ Suite (the plugin relies on this library for some of the 3D image analysis):
    - Start Fiji and go to the menu ***Help>Update…***
    - When the imageJ updater window is open click the ***manage update sites*** button and then check the ***3D imageJ Suite*** update site in the ***Manage update sites*** window.
    - **Close** the Manage update sites window, press **Apply changes** in the ImageJ Updater window and when prompted to, Close Fiji.
- Install the plugin:
    - Start Fiji and go to the menu ***Help>Update…***
    - When the imageJ updater window is open click the ***manage update sites*** button and then check the ***MS-ECS-2D*** update site in the ***Manage update sites*** window.
    - **Close** the Manage update sites window, press **Apply changes** in the ImageJ Updater window and when prompted to, Close Fiji.
    - The plugin will be accessible in the menu ***Plugins>Cell_Transcript>Cell transcript analysis.***

Note: these steps are not necessary if you have already activated the complete *MS-ECS-2D* update site to obtain the *Cell annotation* or *PathFinder* plugins.

2. Plugin usage

Before starting the plugin, load the raw data and optionally the cell segmentation and annotation image (as outputted by the annotation tool for instance, see section 3). Both images should have the same X and Y dimensions.

You can then start the plugin by going to the menu: ***Plugins>Cell_Transcripts>Cell transcript analysis v3.*** A user interface will open requesting information necessary for the analysis (see Figure 1.F). The parameters are separated in 5 sections:

- **Raw data and Masks**
  - **Input image** (Figure 1.A, B): here the raw data to analyse is selected. A drop down menu lets the user choose from all images that are open in Fiji. It allows selecting the raw data to analyse. The selected image is expected to have a nuclei channel and at least one transcripts (smFISH) channel.
  - **Cell/Tissue masks** (Figure 1.C): this drop down menu can be used to select a cell segmentation/annotation image from the images that are open in Fiji. In case one does not want to use a cell/tissue mask for the analysis "None" can be selected. The selected image is expected to be a cell segmentation (labelled regions separated by zeros valued boundaries) or a 2-channel image with cell segmentation in the first channel and cell annotation in the second channel. The latter format corresponds to the output of the annotation tool presented in section 3.If no cell segmentation is provided a simple a simple one will be processed by dilating nuclei segmentation.
- **Nuclei Segmentation parameters:** these parameters can be adjusted to obtain optimal nuclear segmentation. The parameter notation (between brackets) corresponds to that used in the method description.
  - **Nuclei channel**: An integer indicating the channel with the nuclei signal (1 correspond to the first channel).
  - **Nuclei size (radius in pixel)** / (R_{nuc}): An integer value indicating the radius of the largest nuclei to be expected in the sample. This value will be used for nuclei background subtraction and other filtering related to nuclei segmentation.
  - **Nuclei threshold adjustment** / (α_{nuc}): Nuclei are differentiated from their background with to an automated threshold. This threshold will be multiplied α_{nuc}. For instance if the threshold is systematically too high, some nuclei with low intensity values are ignored and reducing α_{nuc} to a value smaller than 1 can aid the detection. If no systematic bias is observed the value should be set to 1 (i.e. the default value)
- **Transcripts segmentation parameters**: these parameters are used for transcript detection.

- o **Transcripts channel(s)**: an integer indicating the channel with the transcript signal (1 correspond to the first channel). If multiple transcript channel have to be analysed they can be indicated separated by a comma (1, 2, 3,...).
- o **Transcripts typical radius** / (R_{tx}): An integer value indicating the typical transcript radius (*i.e.* the pixel distance between a transcript maximum of intensity and the neighbouring background). This value is used to remove the background around the transcripts. A radius of 2 or 3 pixels should be used for images in which the sub-resolution transcripts are imaged with Nyquist sampling.
- o **Spot minimum intensity** / (T_{tx}): A value corresponding to a fraction of the typical transcript dynamics (*i.e.* the difference between typical transcript maxima of intensity and the neighbouring background). If this value is too low there will be false positive detections corresponding to peaks in the background noise. If it is too high transcripts with low intensities will be missed. We propose the user to first run the analysis with a coarsely chosen value here and then to use the histogram of detected maxima outputted by the script to determine an adequate threshold. In that histogram one can see a very sharp peak close to zero corresponding to the noise outside of the tissue and a second peak corresponding to the noise in the tissue. The threshold should be placed in the tail of that second peak to accurately detect the transcripts. If multiple transcripts channels are analysed multiple values separated by a comma can be set.
- **Foci segmentation parameters**: these parameters are used for the detection of transcription foci. Transcription foci are characterized by their large size (area) and intensity compared to individual transcripts. This information is used for the automated detection of transcription foci. In addition, transcription foci must fit the criterion that they are located in a nucleus.
  - o **Foci intensity threshold adjustment** / (α_{Imax}): α_{Imax}, is used to adjust an automatically determined threshold, T_{Imax}. T_{Imax} is based on the maximum intensity, I_{max}, of detected transcripts. *$T_{Imax}=median\_I_{max} +3. \alpha_{Imax} .MAD\_I_{max}$* , where median_$I_{max}$ and MAD_I_{max} are the median and the median absolute deviation of I_{max} on transcripts. to help the user to select the appropriate T_{Imax} for an image set a scatter plot of transcripts maximum intensity and volume is outputted. A value between 1.5 and 3 usually allows good focus detection in our images. If multiple transcripts channels are analysed, multiple values separated by a comma can be set.
  - o **Foci volume threshold adjustment** / (α_{vol}): See previous parameter description where maximum intensity of transcript can be replaced by volume of transcripts.

- o **Enlarge nuclei to detect foci** / (R_{ExtNuc}): transcription foci can be located at the border of nuclei and then end up just outside the nuclear segmentation. Since localization of a focus candidate in a nucleus is a prerequisite for foci detection these foci fail to be detected. To decrease the rate of false negatives the nuclear area can be dilated by a fraction of the nuclear radius. This value is expressed in pixel and can be chosen as a fraction of the parameter Nuclei size.
  - o **Correct transcript over-segmentation**: foci can be bigger than transcripts and have a tendency to get split during transcript detection when foci candidate are first identified. When checked, "Correct transcript over-segmentation" triggers a filtering step that prevents larger object from being split (see methods section below for more details).
  - o **Foci Maximum radius** / (R_{foci}): This value is used only if "Correct transcript over-segmentation" is checked. It is an integer value indicating the typical foci radius. This radius can be estimated manually from the original smFISH image.
- **Miscellaneous parameters**:
  - o **Display Graph**: when checked, a maximum spot intensity histogram and a scatter plot of the area vs intensity of all detected transcripts and foci are displayed.
  - o **Save Graph to disk**: when checked, the abovementioned plots are saved to a user-defined directory.
  - o **Display results image**: when checked a multichannel image illustrating the results of cell, nuclear and transcript detections is displayed.
  - o **Save result image**: when checked the result image is saved in tiff format to a user-defined directory
  - o **Display transcripts images**: when checked an image of transcript segmentation (colored after each transcript id) is displayed
  - o **Save directory:** Path where results files will be saved after the processing.

When the processing is finished the script will output a file with measures for each individual cell (_cell) and a file with extensive information on each of the segmented transcripts (_spot_Ch<N>):
- **<Img name>_cell.txt**: this file provides measures for each individual cell. **<Img name>** is the name of the analysed raw data without the file extension. The name of each measure in the file, and their meaning are listed below.
  - o **CellId**: Label of the cell in the mask used
  - o **X** and **Y**: average position in image units (*i.e.* as found in the image properties) of the cell along the X and Y axis. The Unit is consigned in the parameter file.

- o **AreaCell**: surface area in image units of the cell region corresponding to CellId.
- o **AreaNuc:** surface area in image units of the nuclear region corresponding to the CellId. This value is set to 0 if no nucleus is found in that particular cell.
- o **MaskLabel**: value corresponding to the cell type annotation associated with the cell (EVL=1, YSL=2, DEL = 3, outside = 4). If no region type is provided at plugin start this value is set to 0.
- o **nuclear fraction**: given that nuclei and cell segmentations originate from different sources a step was required to attribute nuclei to cells in the analysis. The value nuclear fraction denotes an assessment of this attribution and indicates the overlap of a nucleus with its parent cell. 1 corresponds to a complete overlap and zeros would correspond to a null overlap. If no nucleus is attributed to a cell the value is set to 0.
- o **NspotNuc_ch<N>**: the number of transcripts detected in the nucleus of the cell. The foci contribution is not included. Additionally, if the nucleus does no fully overlap with a cell only the transcripts in the overlapping region are counted. **<N>** is the channel number from which the analysis originates. If multiple transcript channels are analysed, the file will show multiple columns corresponding to the different channel. The transcripts that overlap with the extended nuclear region are not taken into account here.
- o **NspotCell_ch<N>**: the number of transcripts in the entire cell, nuclei included. Transcripts that are located in transcription foci are not included.
- o **Nfoci_ch<N>**: the number of foci detected in the nucleus of the cell. In the case of false positive detections this value can be larger than 2.
- o **TxDensity_ch<N>**: the ratio of NspotCell_ch<N> to the cell area in pixel.
- o **NspotFoci1_ch<N>**: the number of transcripts in the transcription focus with the largest number of transcripts. The number of transcripts is measured as the ratio of transcription focus integrated intensity to median integrated intensity of all transcripts in the image. Object integrated intensity is the sum of the pixels intensity in that object.
- o **NspotFoci2_ch<N>**: the number of transcripts in the transcription focus with the second largest number of transcripts.
- o **NtxFoci_Other_ch<N>**: the number of transcripts in remaining object that were (incorrectly) detected as foci.

- **<Img name>_spot_ch<N>.txt**: this file provides measures for each individual transcript. **<Img name>** is name of the raw data analysed without the extension. **<N>** is the number of the transcript channel analysed in the raw data. There will be

as many files as there are transcript channels analysed. The name of each measure in the file and their meaning are listed below.

- **X**, **Y** and **Z**: position of transcripts in dataset Unit
- **volume**: volume of the transcripts in dataset units
- **Iavg**: average intensity of the segmented transcript
- **maxI**: maximum intensity in the region of the segmented transcript
- **sumI**: sum of the voxel intensities in the segmented transcript
- **isFoci**: this value is 1 if the detection is considered as a focus (it belongs to a nuclei and its volume and intensity are above the defined thresholds) otherwise the value is 0.
- **cellParent**: the id of the cell, CellId, that the transcript belongs to. The value is 0 if the transcript does not belong to any cell.
- **nucParent**: the id of the nucleus that the transcript belongs to. The id of a nucleus is given after its cell parent id. nucParent is 0 if the transcript is not located in a nucleus.
- **extNucParent**: the id of the extended nucleus the transcript belongs to. Extended nucleus id is given after the cell it belongs to. This value is only relevant for transcription foci.

- **<Img name>_parameters.txt**: this file gathers information on the analysis parameters as well as the date of the processing and image resolution

- **<Img name>_MaxDistrib_ch<N>.png** and **Maxima distribution_ch<N>** window: It is possible to create a histogram plot of transcript candidate maximum intensities (Figure 1G). This plot can be used to determine an appropriate threshold for transcript detection. The histogram always has 500 bins between 0 and the candidate maximum intensity. If the "Display Graph" checkbox is checked in the input dialog then a window with the graph will open at the end of the processing. If "Save Graph to disk" is checked an 800 time 600 pixels png image will be saved to the save directory.

- **<Img name>_ScatterPlot_ch<N>.png** and **Transcript_ch<N>** window: It is possible to create a scatter plot of transcript maximum intensity versus transcript area (Figure 1H). This plot can aid the determination of appropriate thresholds for the detection of transcription foci. In the plot, cytoplasmic transcripts are in green, nuclear transcript in blue and foci in red. The thresholds that distinguish foci from nuclear transcripts are represented as red lines. If the "Display Graph" checkbox is checked in the input dialog then a window with the graph will be open. If "Save Graph to disk" is checked an 800 time 600 pixels png image will be saved to the save directory.

- **<Img name>_ResultImg.tif** file and window: It is possible to display a result image showing the cell, nuclear and transcript detections (Figure D,E). For each transcript channel analysed, a transcript detection channel is created where transcripts and foci have value 1 and 2, respectively. If more than 2 foci are detected in a nucleus all foci from that nucleus have a value of 3. An additional channel shows nuclei and cell segmentation. The 2D contour of nuclei detection are shown with value 510 and the contour of cells segmentation are set to 255. The new channels created are appended to a copy of the raw image data .If "Display results image" checkbox is checked in the input dialog then an image window will open at the end of the processing. If "Save results image to disk" is checked the image is saved to the save directory in tiff format.

3. Script and Batch Processing

3 scripts are provided on the update site to facilitate the use of the transcript analysis plugin. They can be accessed via the Fiji menu Plugins>Cell Transcript:

- **get_transcripts_maxima_distribution.py**: It creates a list of maxima intensity in the filtered transcript image. This information data are useful to build a distribution of maxima intensity and determine an adequate threshold for transcript detection. When selecting this menu entry an GUI will pop up requesting the location of the file to analyze, the location where to save the list of maxima intensity, the typical transcript radius (see description in section 2) and the channel to analyze (see description in section 2)

- **get_transcripts_maxima_distribution_batch.py**: this is a batch version of the previous script. It will measure the maxima intensity of multiple files in a folder. Once started the script will request a Source directory, the name of the image subfolder (i.e. not the full path) which is contained into the source directory, the name of the directory where result will be written and finally the analysis parameters (transcript radius and channel to analyze).

- **cell_transcript_analysis.py** : this is a batch version of the transcript analysis plugin. It creates a complete analysis of the transcript for the image in a folder. The script expects that the data are organized in 2 folders belonging to the same source folder. The first subfolder contains the raw images and the second the corresponding cell mask. If the cell mask is missing the process will run without the mask. The result folder will also be created in that source folder. Once the script is started a user interface pops up requesting the analysis parameters: the source folder, the name of the raw image subfolder, the name of the mask subfolder, the name of the result

subfolder. All the other parameters are the same as those requested in the transcript analysis plugin. One can refer to section 2 for more details.
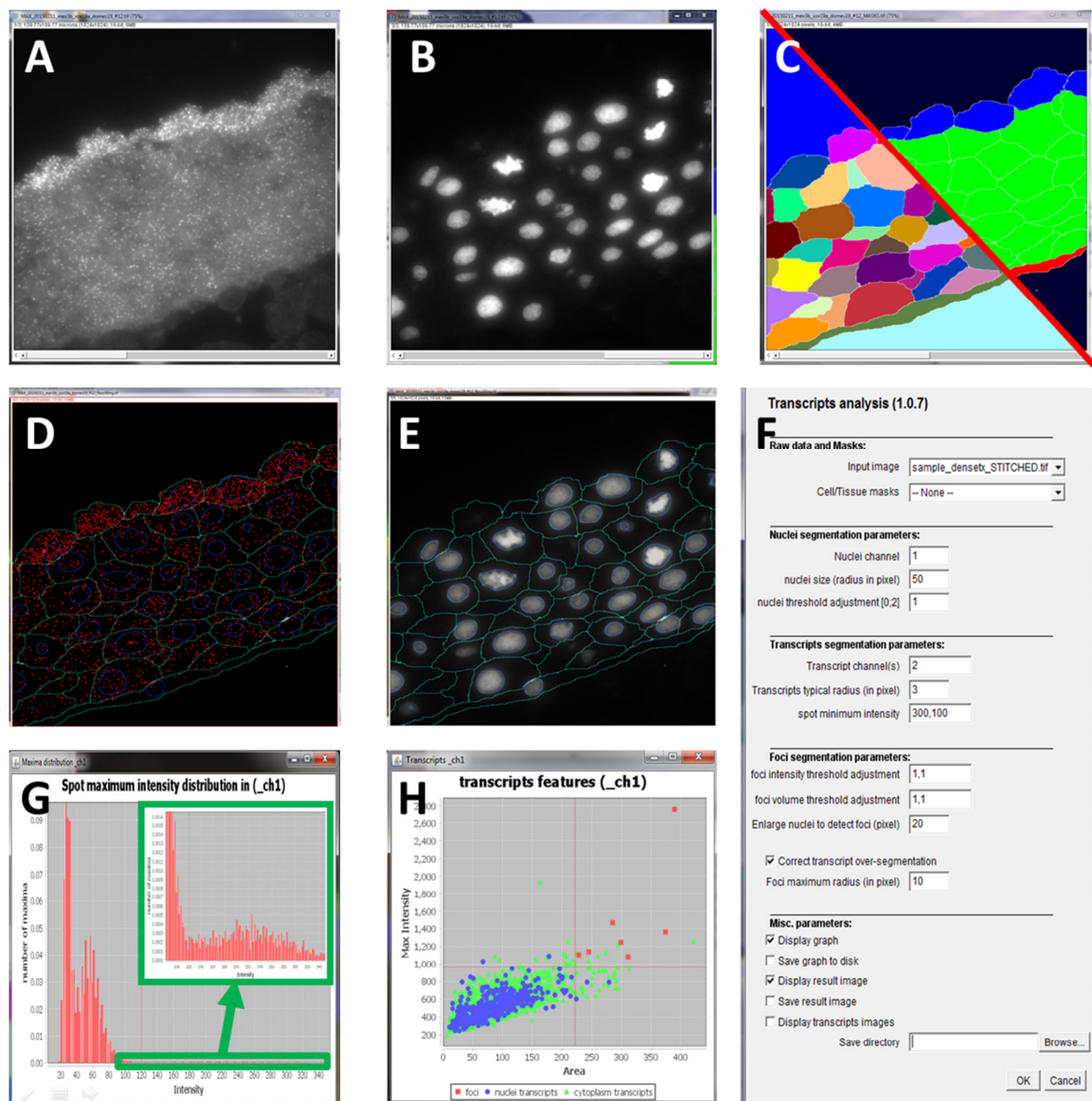

Figure 4. Illustration of input and output of the *Transcript analysis* plugin. (A) smFISH transcript signal. (B) DAPI signal for the nuclei. (C) Cell and tissue masks. (D) Detected transcripts. (E) Cell and nuclear segmentation, merged with the original DAPI signal. (F) GUI for the *Transcript analysis* plugin. (G) Intensity distribution plot of all detected maxima. The vertical red line indicates the threshold used for transcript detection. (H) Scatter plot of Area vs Maximum intensity of all detected transcripts and transcription foci. Green: cytoplasmic transcripts, blue: nuclear transcripts, red: transcription foci. Red lines depict thresholds used for the detection of transcription foci.
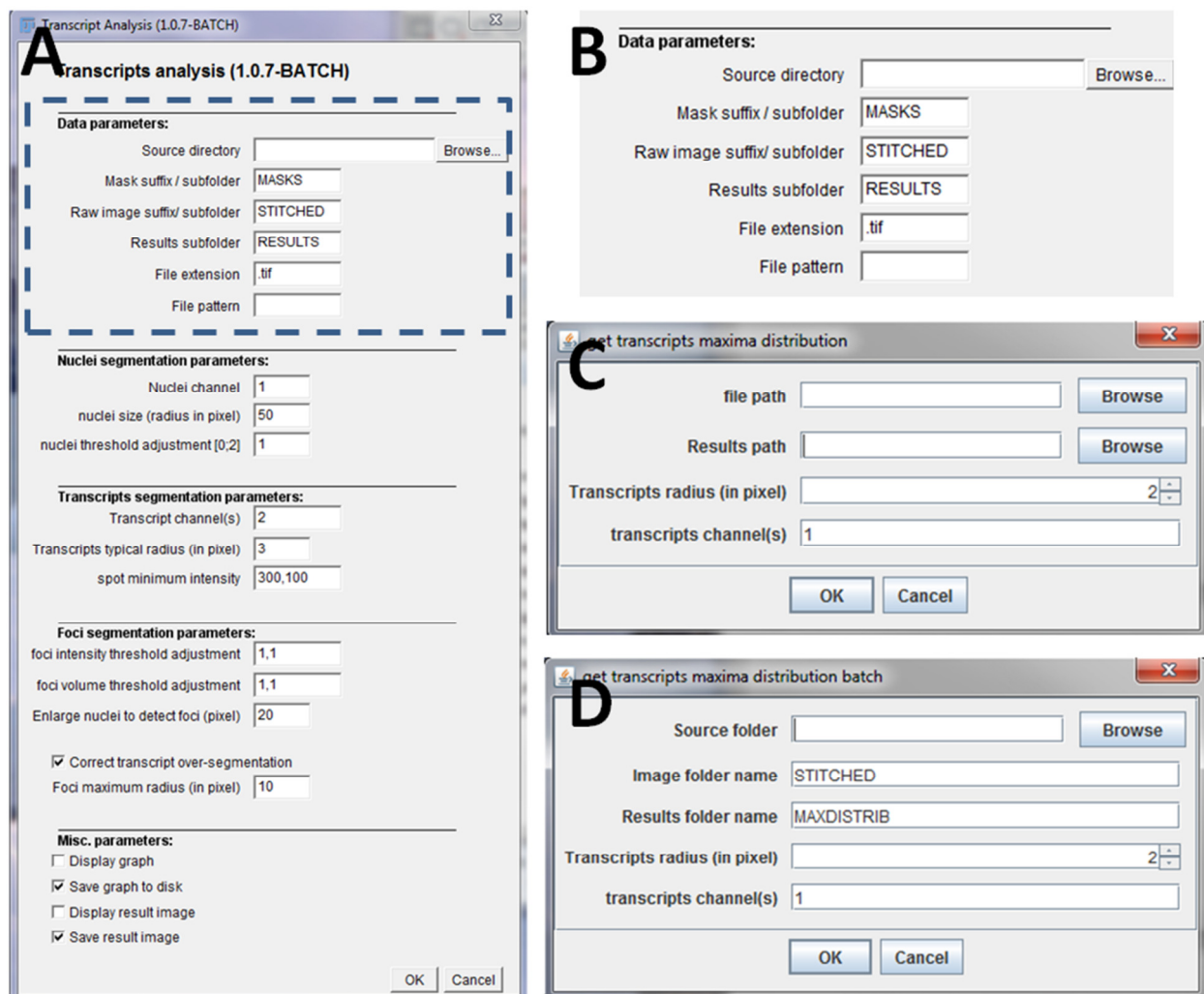
Figure 5. input interfaces of to the scripts. (A)GUI for the batch analysis of the transcript analysis. (B) Close up on the dashed square in A showing the details data parameters. (C) GUI for the transcript maxima distribution script. (D) GUI for the batch version of the transcript maxima distribution script.

4. Methods description

4.1 Analysis Pipeline overview

The analysis pipeline is composed of two parts. First, labels are created to describe the nuclei, cells and transcripts regions (image analysis). In a second step, several descriptors are measured in each of these nuclei, cells and transcript regions to characterize them (data analysis).

Image analysis:

In: nuclei raw image, transcript raw image, (cell labels).

Out: nuclei labelling, extended nuclei labelling, cell labelling, transcript labelling, filtered transcripts.

The successive steps of the image processing are as follows:

- Segment nuclei from nuclei raw image (*i.e.* create nuclei labelling).

- If a cell labelling is provided, assign nuclei to cells and relabel nuclei and cell labelling to fit the assignment. Otherwise create a cell labelling by dilating the nuclei labelling.
- Dilate the nuclei labelling to create an extended nuclei labelling for later foci detection.
- Segment the transcripts (i.e. perform transcripts labelling).
- If "Correct transcript over-segmentation" is checked: update the transcript label to prevent over-segmentation of transcription foci.
- Create a filtered transcript image for further transcript intensity measures by removing the background.

Data analysis:
In: nuclei labelling, extended nuclei labelling, cells labelling, transcript labelling, filtered transcripts, (cell annotation).
Out: cell measures, transcript measures.
The successive steps of the data analysis are as follows:
- Get cell measures.
- Get transcript measures.
- Update transcript measures with cell, nuclei and extended nuclei parenthood.
- Update cell measures with transcripts counts in nuclei and cells.
- Update transcript measures and cell measures with transcription foci information.

4.1.1 Image analysis methods details
- **Segment nuclei**

  In: nuclei raw image; **Out:** nuclei labelling.
  Method description: The segmentation of nuclei can be solved in 2D by making a maximum projection of the original nuclei image. This strategy is satisfying since the tissue sections are less than one cell layer thick, which makes it very unlikely that two nuclei overlap. A background subtraction is performed on the projection to remove slowly varying intensity in the image background such as out of focus signal. The filter radius is chosen to be similar or larger than the largest nuclear radius to make sure the nuclear signal is not affected by the filtering. For this, the input parameter R_{nuc} (i.e. "nuclei size") is used. Nuclei are then detected with a difference of Gaussian filter (DOG filter) followed by maxima detection. The DOG filter enhances roundish structures, such as nuclei, at the chosen scale[1] σ. That scale should not be too large or too small to avoid respectively fusion of close objects or splitting of large objects. We chose σ equal to $R_{nuc}/4$. The nuclei are then detected as positive maxima of intensity in the DOG filtered image. To avoid over-detection only the maxima with dynamics[2] superior to ε are kept. ε is not too sensitive. It should be superior to one and smaller than the dynamics of

---

[1] **The DOG filter** smooths the image with a Gaussian kernel of variance σ and subtract from it a copy of the image smoothed with a kernel of variance 2σ.
[2] **The dynamic of a local maximum** is the difference of intensity between its own intensity and the intensity of the saddle separating it from another local maximum. If many such saddles exist, the intensity of the highest saddle is used to calculate the difference.

objects. We chose a value of 15 but any value between 5 and 100 would give only marginally different detections. Finally this detection is used to initiate a watershed[3] on the nuclei image. The nuclei image is smoothed with a Gaussian kernel of standard deviation σ_{smooth} ($\sigma_{smooth} = R_{nuc}/10$). The watershed algorithms used invades the image from the highest to the lowest intensities and stops at a lower threshold T_{nuc}. This threshold is obtained by multiplying the results of the Otsu method, T_{Otsu}, with a user defined adjustment, α_{nuc} ($T_{nuc} = \alpha_{nuc} \cdot T_{Otsu}$). The output of the watershed is a 2D labelled image representing the segmentation of the nuclei.

- **Cell labelling and nuclei-to-cells assignments**
  In: nuclei labelling, cell labelling (optional); **Out:** nuclei labelling, cell labelling.
  Method description: If a cell labelling is provided at the plugin inputs, the cell regions it defines have to be matched with the nuclear regions. Each cell region should contain either one, or no nucleus (if the nucleus of the cell has not been captured in the image section). However, in exceptional cases a cell region may overlap with multiple nuclei. In order to assign nuclei to cells and to resolve possible ambiguities we proceed as follows: each nucleus is assigned to the cell with which it has the largest overlap assuming that this overlap is at least 50% of the nuclear surface. This last point ensures that a nucleus can be paired with only one cell. If multiple nuclei overlap (partially) with a cell then the nucleus with the largest overlap (and in case of equality the one with the largest volume) is assigned to the cell.
  Following this assignment, nuclear and cell labels are updated such that nuclei get the same label as the cell they are assigned to. Region of nuclei overlapping with other cells are set to the label 1 and will not be taken into account in further analysis.

  If no cell labelling is provided as input to the plugin, a naïve cell segmentation is generated by dilating individual nucleus regions. Each region is dilated by a distance of R_{nuc}. If two regions overlap after dilation, pixels are attributed to the closest nucleus. In the resulting image one cell region is associated with exactly one nuclear region and these regions have the same label.

- **Create an extended nuclei labelling**
  In: nuclei labelling, R_{ExtNuc}; **Out:** extended nuclei labelling.
  Method description: An extended nuclei labelling is created by dilating the nuclei labelling by a distance of R_{ExtNuc} pixels. If two regions overlap after dilation, pixels are

---

[3] **The watershed algorithm** can be seen as the flooding of a landscape of hills and valleys. The landscape altitude is determined by image signal intensity. The region to segment can be distinguished by different colors of liquid during the flooding. Sources of distinct colors are set at each seed points and fill the landscape starting first with the lower altitudes and gradually invading higher altitudes of the landscape. When liquids from two distinct sources meet, a dam is build (a watershed line) to keep the liquids separated. Additionally a threshold, T, can be set to stop the flooding when it reaches altitude T. This algorithm describes an invasion of the landscape from the lower altitudes to the highest one but in practice the principle can also be reversed. In that case the process start on hills and gradually visit lower altitudes till the bottom of the landscape or a threshold is reached.

attributed to the closest nuclei. Finally, parts of extended nuclei that overlap other cell region are set to 0.

- **Transcripts segmentation**

  In: transcript raw image, R_{tx}, T_{tx}; **Out:** transcript labelling.
  Method description: the 3D transcripts image is first filtered to remove large-scale variation with a tophat filtering (*i.e.* a background subtraction) of parameter R_{tx}. Then the image is smoothed with a Gaussian kernel of standard deviation σ_{tx} to remove the noise. This value should not be too large to avoid smoothing out of the transcripts and is fixed to 1 pixel. The transcripts are then detected as local maxima of that image. To avoid multiple maxima detections within a single transcript a minimum distance between local maxima is imposed. Its value is set to R_{tx}. All maxima superior to the user-defined threshold for transcript intensity, T_{tx} are kept as transcripts detection. Finally these detections are used to initiate a watershed segmentation of the transcripts. The watershed process is stop at 2/3 of T_{tx}. As a result a 3D labelled image is obtained where each transcript region has a distinct integer value. Additionally a list of all maxima and their intensities is generated and used -if requested- to display a distribution of maxima intensity, which can be used to choose T_{tx}.

- **Correct over-segmentation of large spots**

  In: transcript labelling,R_{tx},R_{foci},T_{tx}; **Out:** corrected transcript labelling.
  Method description: The spots corresponding to foci are first detected during transcript detection. However, foci that are much larger than transcripts can be split during that process. This method is meant to correct over-segmentation of transcription foci. This method is not a detection of transcription foci but rather a filtering of the transcript segmentation.
  For that matter a background subtraction adapted to the radius of the large foci is then performed with a tophat filter of radius R_{foci}. The maxima of the image are then detected. A threshold equal to the median plus 3 times the median absolute deviation of maxima intensity is then applied to the filtered image. All connected component smaller than a sphere of radius R_{tx} are then discarded. The remaining regions are then copied onto the image of the segmented transcript regions with unused labels value.

- **Create a filtered transcript image**

  In: transcript raw image, R_{foci}; **Out:** filtered transcripts intensities.
  Method description: The raw image of transcripts is processed with a tophat filter of radius R_{foci} to remove image background. That image will be used to do intensity-based characterisation of the transcripts and foci regions. Removing the background is important to prevent a systematic bias of the measures.

4.1.2 Data analysis methods details

- **Get cell measures.**

  In: cell labelling, nuclei labelling, cell annotation (optional); **Out:** cell measures;

Method description: cell labelling and image labels are processed to extract per cell: Id, surface area, surface area of the corresponding nucleus, average position along the image axis (i.e. X and Y) and the region type as indicated by the cell annotation masks.

- **Get transcript measures.**

  In: transcript labelling, filtered transcripts intensities; **Out:** transcripts measures;
  Method description: Transcript labelling and the filtered transcript image are read to obtain for each transcript region: Id, surface area, average intensity, maximum intensity, integrated intensity and position along x, y and z axis.

- **Update transcript measures with cell-, nuclei- and extended nuclei label they belong to.**

  In: transcript measures, transcript labelling, cell labelling, nuclei labelling, extended nuclei labelling; **Out:** localized transcript measures;
  Method description: For each transcript the nucleus (or extended nucleus, cell) it belongs to is determined by looking at the nuclear (or extended nuclear, cell) label at the position of the transcript. The z-position is not considered since cell and nuclear labelling are 2D.

- **Update cell measures with transcripts counts in nuclei and cells.**

  In: cell measures, localized transcripts measures, cell labelling, transcript labelling; **Out:** cell transcript measures;
  Method description: For each cell, the number of transcripts is counted in the whole cell and in the nuclei and a density of transcripts per pixel is measured. The transcript density per μm^3 can be calculated based on the pixel size.

- **Update transcript measures and cell measures with foci information**

  In: Cell transcript measures, localized transcript measures, foci threshold adjustments; **Out**: cell measures, transcripts measures;
  Method description: Foci are detected among the transcripts based on their maximum intensity, volume and position (extended nuclear region). Those transcripts which features significantly deviate from the median transcripts intensity and area are selected as foci. On both features we use the following threshold T_{feat}:

  $$T_{feat} = Median_{feat} + 3.\alpha_{feat}.MAD_{feat}$$

  where $Median_{feat}$ and MAD_{feat} are the median and the median absolute deviation of the feature among all transcripts detected, respectively. α_{feat} is a user-defined value that can be used to adjust the value of the threshold for the feature. In each detected transcription focus the number of transcripts is counted by dividing the integrated intensity of the focus with the median integrated intensity of the transcripts. The estimated number of transcripts per foci is then saved for each cell. If one extended nucleus contains more than 2 foci, the number of transcripts of the supernumerary (smallest) foci are summed and added to the cell information in a separate column.

***