histoCAT 1.76 manual

histoCAT – Getting started

1. Install histoCAT

histoCAT is automatically installed from the web when running the histoCATAppInstaller.exe file. Windows users must have Visual Studio installed for features like PhenoGraph to function. If Visual Studio is not already installed on your computer download it from https://www.visualstudio.com/downloads/.

2. Open histoCAT

Double click the icon generated during installation to open histoCAT.

3. Data requirements

In order for raw image data to be successfully loaded into histoCAT, all files associated with a specific image need to be stored in a separate, uniquely named folder (see example data set). Multiple folders, each containing the data corresponding to one image, can be loaded simultaneously. The different markers measured must be stored as individual (unstacked) 16-bit or 32-bit tiff files in the image folder. If there is single-cell information, the image folder can also contain a mask, such as that generated after segmentation in CellProfiler (see example data set), which identifies individual cells. The mask can be saved as a mat or tiff file in uint16 or int32 format. Naming (case sensitive) the mask with the ending "_mask.tif(f)" or ".mask.tif(f)" will enable an automatic loading without manual selection.

4. Load samples into histoCAT

After opening the histoCAT GUI, the first step is to load the image data for subsequent analysis. Click on the "Load" button on the upper left of the interface. A drop-down menu with two options will appear. Choose "Load Samples" in order to start a new histoCAT session (Figure 1).

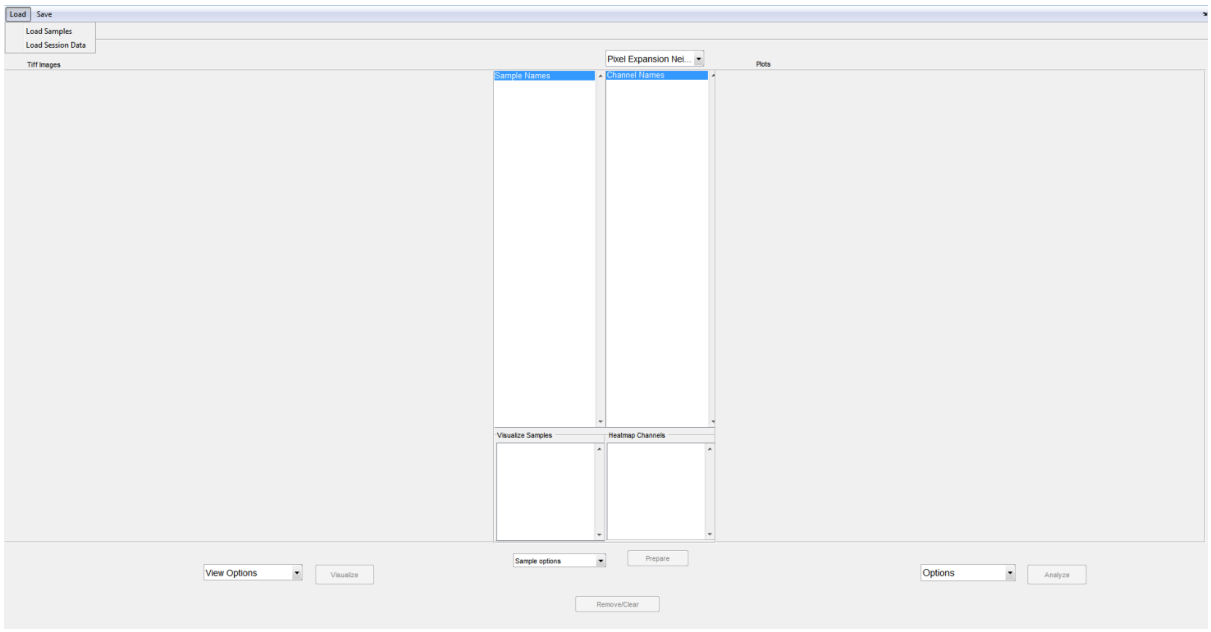

Figure 1

You will be prompted with a folder selection box (Figure 2). Navigate to the folders containing the images to be loaded. Click “Done” once you have added all the desired samples to the list on the right of the window (Figure 2).

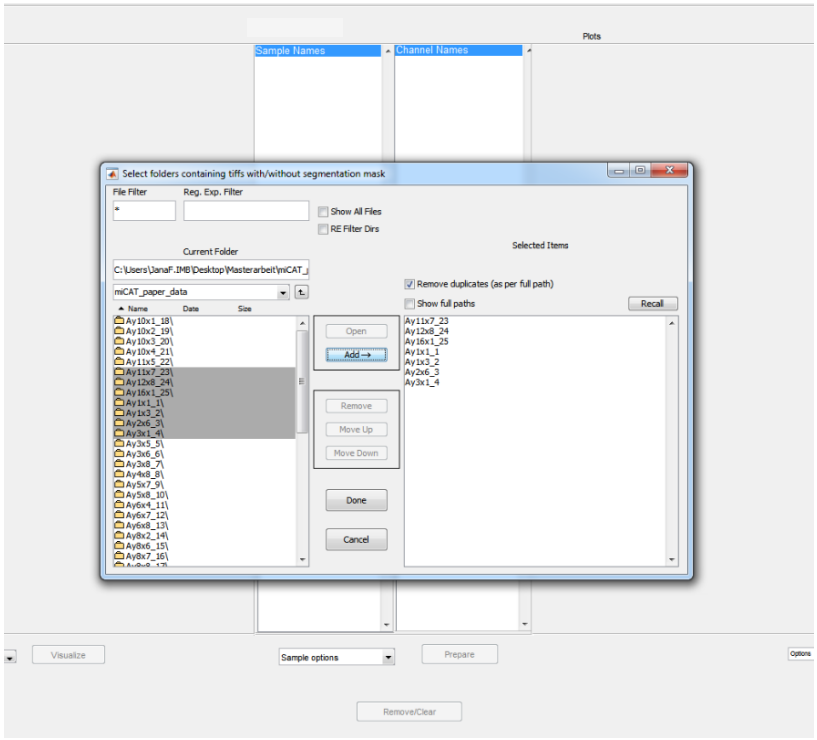

Figure 2

Once histoCAT has finished loading the image data, you will be asked to specify the number of pixels to expand from each cell in order to look for neighboring cells (Figure 3). In the top field you can choose any single integer for the calculations of the number of neighbors and percent touching. In the bottom field you can either choose a single integer or a range of pixel expansions to run through

and separately detect the neighbors for each of the values. Keep in mind that the loading will be faster if it does not run through many different pixel expansions. If a range of values was chosen, the pixel expansion drop down menu will allow you to switch between the different values for specific analyses later on. In the next step, you can apply log or arcsinh transformation to your single cell data. If arcsinh is selected, a second popup will ask for a cofactor. It may take from a couple of seconds up to one minute per image (500x500 pixels) for the single-cell information to be updated. If the images are much bigger (>2500x2500 pixels) it can take several hours to load the images including all single cell information. Once the folder selection prompt appears, browse to where you wish to store the "Custom Gates Folder". This is where any files that are automatically generated during the histoCAT session will be saved. Folder names should not include any spaces or special characters.

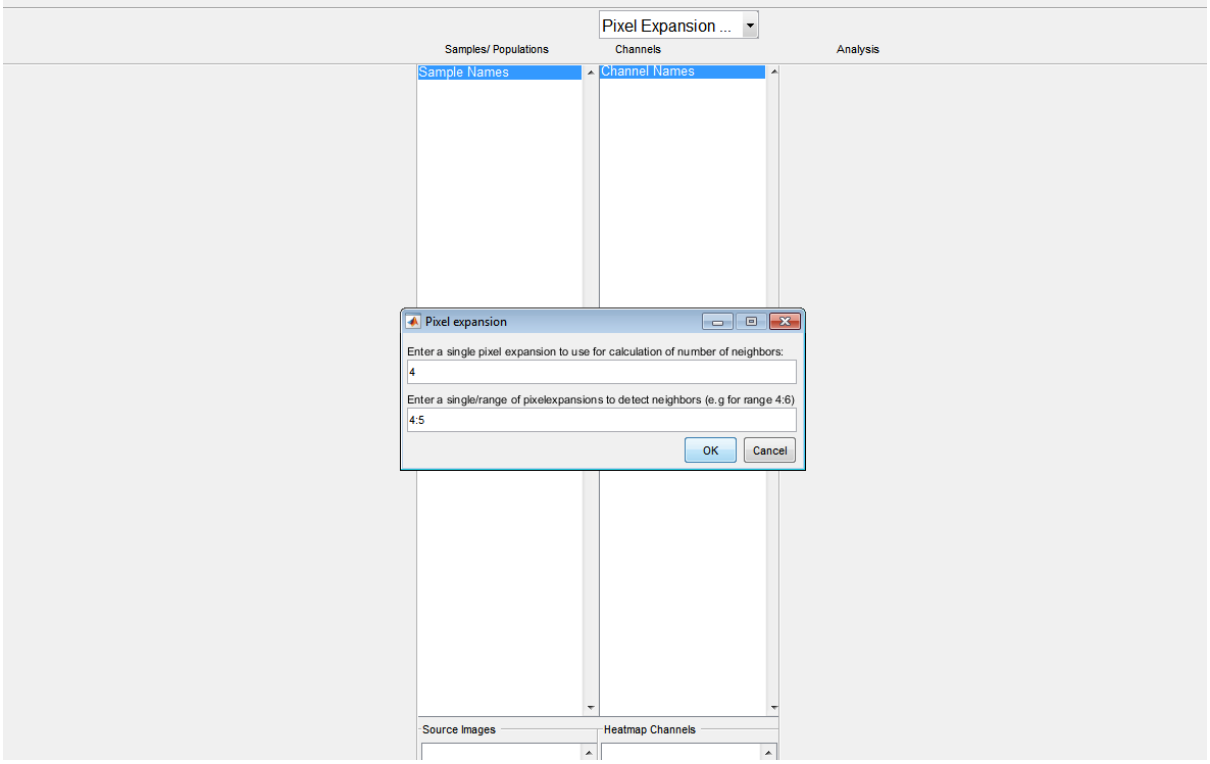

Figure 3

The loaded image data are displayed in two columns in the center of the interface (Figure 4). The box on the left hand side contains the "Samples", which at this stage in the analysis process are just the individual images. The measured markers for the currently selected image are shown in the "Channels" box on the right. Note that histoCAT also calculates image-specific cell identifiers, cell size and shape measurements, the percentage of a cell's surface touching other cells, and the number of neighboring cells. Figure 4 displays the overall arrangement of the histoCAT interface.

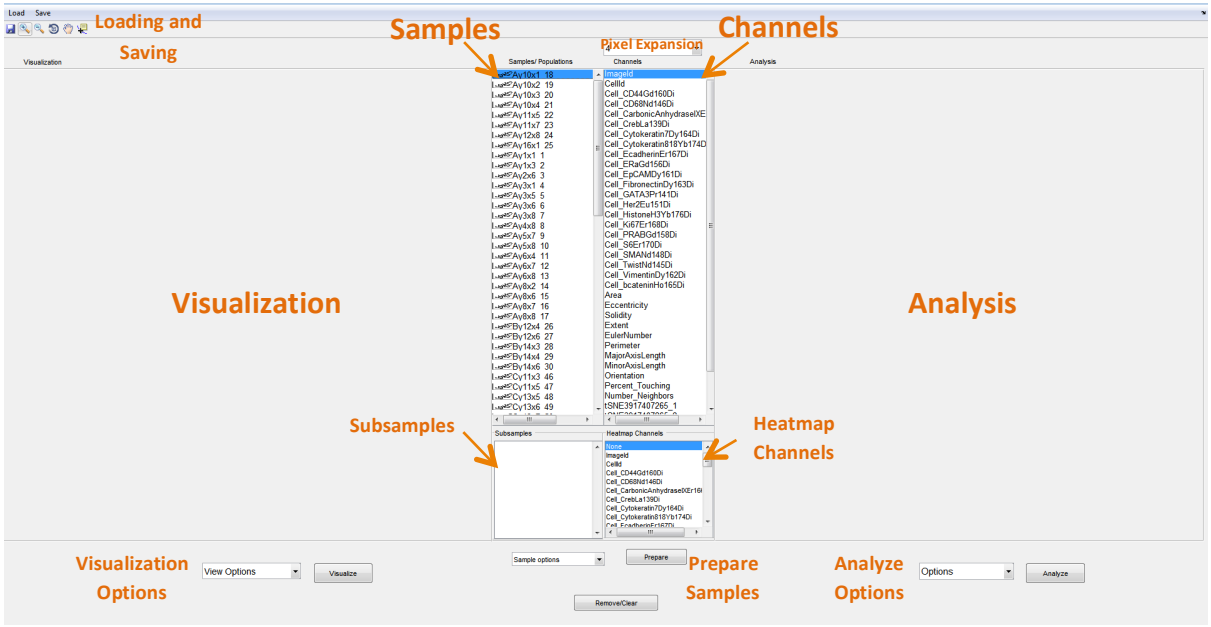

Figure 4

5. Save and reload session data

Once a histoCAT analysis has been started, the full session can be saved by clicking the save icon (do not attempt to use the save button in the drop-down menu) on the upper left of the GUI (Figure 5). The session must be saved as a mat file in order to continue the analysis at a later time. This is especially important so large data sets need only to be loaded once.

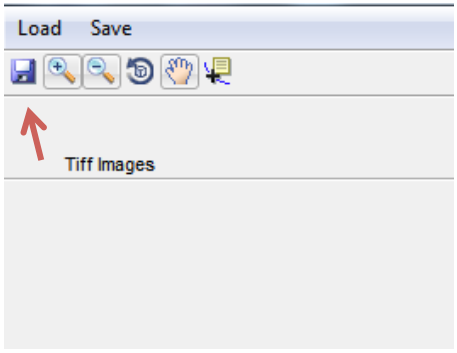

Figure 5

To reopen a previously saved session, choose the “Load Session Data” option in the drop-down menu of the “Load” button and navigate to the mat file containing the session of interest. This is much faster than reloading the images from the folder structure.

6. Visualize images

The left hand side of the GUI is dedicated to different kinds of visualizations of the selected images. The form of visualization can be determined and changed using the “View Options” drop-down

menu on the lower left (Figure 6). Having selected one of the options, the “Visualize” button must be pressed for changes to be applied. The following visualization options are supported:

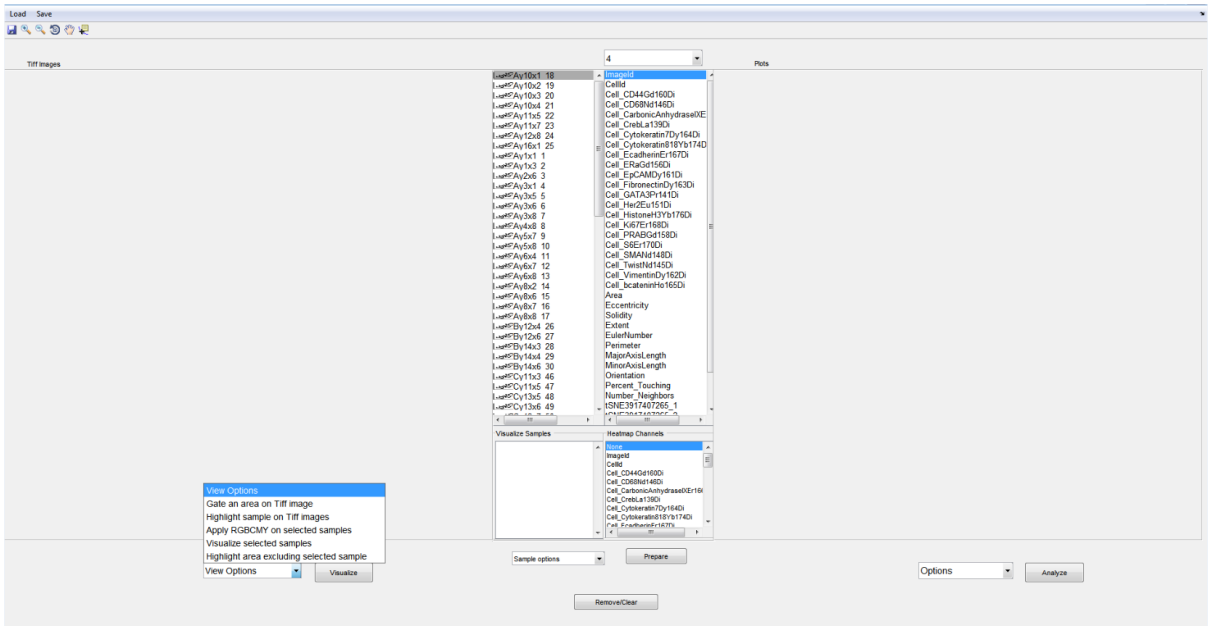

Figure 6

1. Choose “Apply RGBCMY on selected samples” in order to display the signal of the selected markers each in a different color and overlaid into one image (Figure 7a). To select multiple markers or images, please hold down the shift/CMD key. The tabs that appear above the visualization represent each of the selected images. Below the image a slider along with a checkbox for each color appears. Check a color and move the slider in order to increase the intensity of a specific signal. When the mask checkbox on the lower left is selected, the tiff image is overlaid with the segmentation mask, highlighting the individual cells (Figure 7b). The “Plot sample area XY” checkbox marks the centroid of each cell on the image.

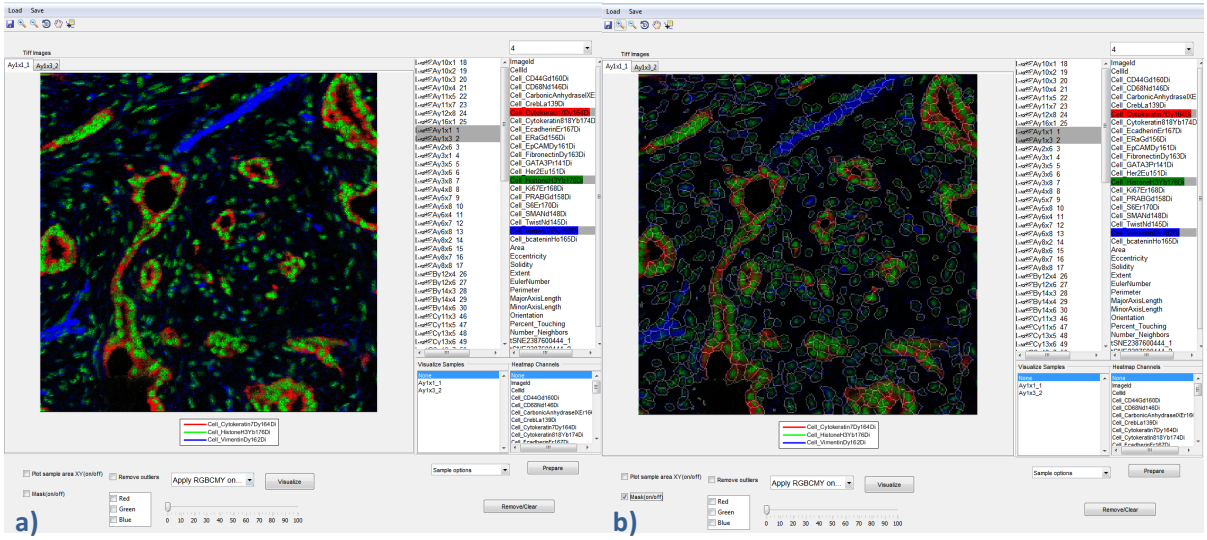

Figure 7

2. The “Heatmap channels on selected samples” option displays a heatmap of the selected channel (Figure 8a) overlaid on the single-cell mask (Figure 8b). In this case, the slider allows you to specify a “percentile cut-off” for very high values. This can be used to prevent the heatmap from being displayed as mostly blue if there are some extremely high values in very few cells. If the slider is set to one, the cut-off will be at the 99th percentile, meaning that the intensities above this value will be set to the value of the cut-off.

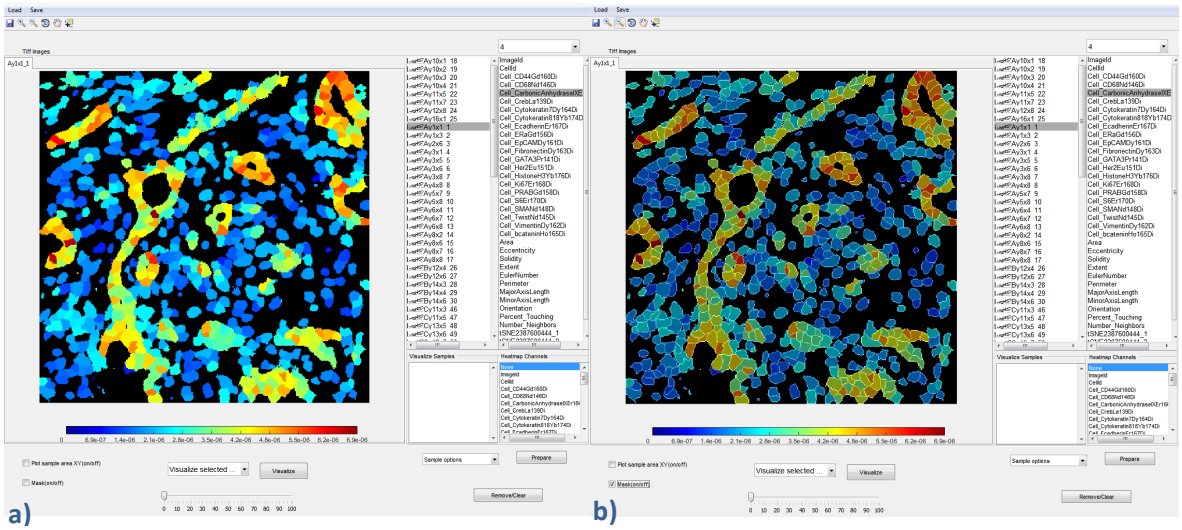

Figure 8

3. Cells of interest can be manually gated with the “Gate an area on tiff” option. Confirm the selection by pressing the “Visualization” button before gating. Afterwards, encircle a certain area with the cursor directly in the tiff image (Figure 9). Hold down the mouse key while selecting and release it when finished. These cells are then saved as a gate, which appears in the samples column bellow the loaded images.

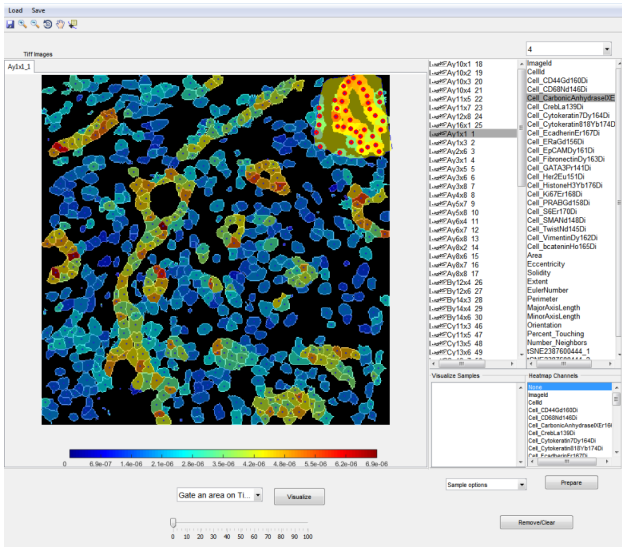

Figure 9

4. A subset of cells (such as a PhenoGraph (Levine et al., 2015) cluster or manually gated cells) can also be highlighted on the tiff image using the “Highlight samples on tiff” option. If

multiple subsets are visualized simultaneously, they appear in different colors on the image. These colors are automatically selected such that they are most distinguishable from the colors in the background image (Figure 10).

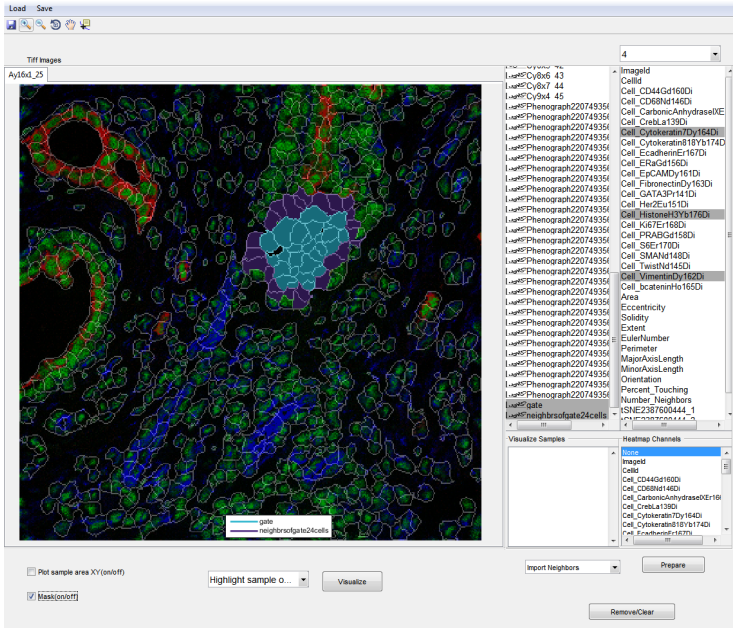

Figure 10

5. "Highlight excluding selected sample" leads to the exact opposite of "Highlight samples on tiff". All cells except for the ones in the selected gates will be marked (Figure 11).

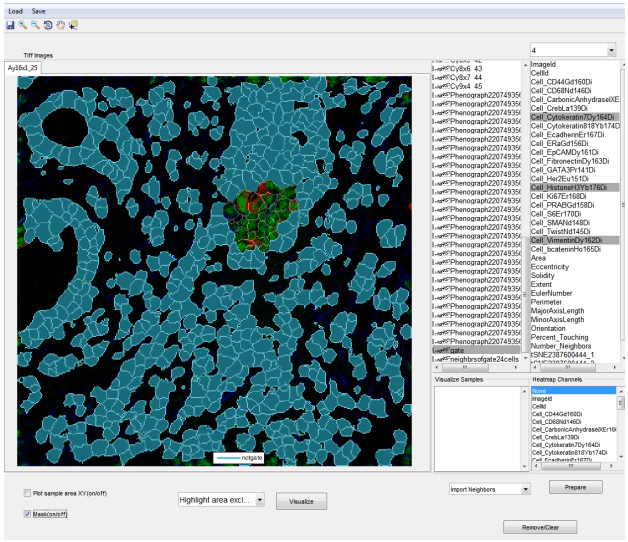

Figure 11

7. Analyze data

The right hand side of the interface is dedicated to data analysis and representation. Next to the "Analyze" button there are several options (Figure 12).

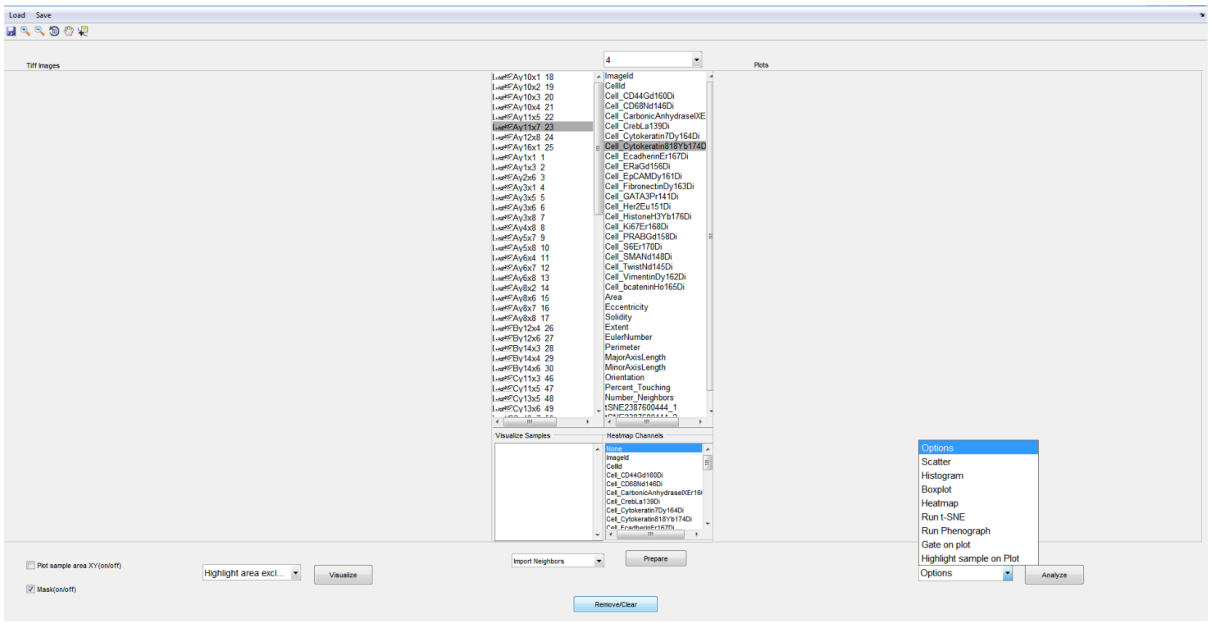

Figure 12

1. The “Scatter” option produces a scatterplot of up to three markers. If multiple images are selected, they are displayed in different colors (Figure 13a). From the channel list either two or three markers can be selected simultaneously to be displayed in a two- or three-dimensional plot. A fourth channel can be visualized in terms of color. Choose this additional channel from the “Heatmap Channels” box below the regular channels. This will overlay the dots in the scatterplot with a heatmap of the selected marker’s intensities (Figure 13b). In two-dimensional plots, a regression line can be added to the scatterplot by checking the “Regression line” box. Additionally, the R-value of Pearson’s correlation will be displayed.

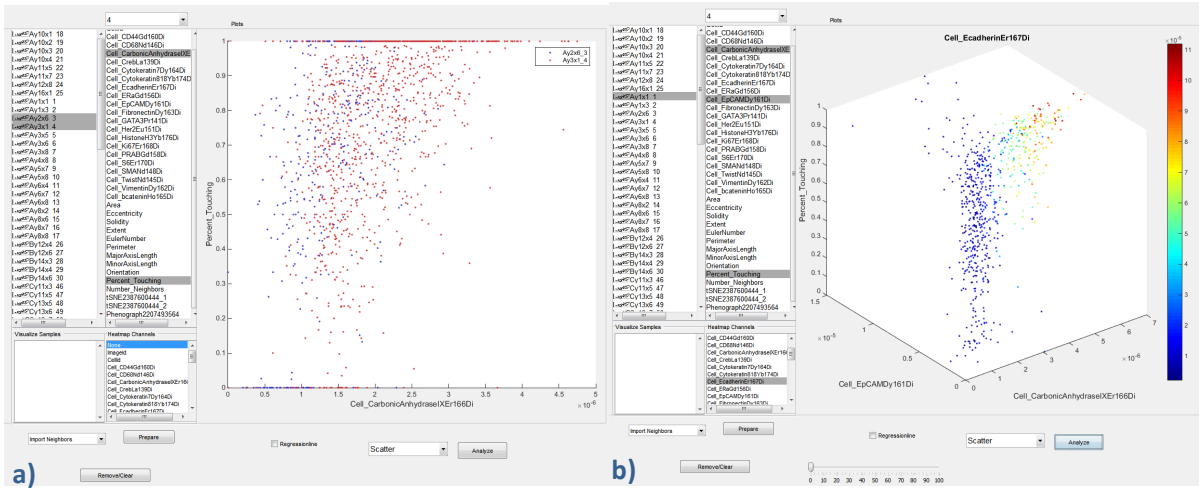

Figure 13

2. The “Histogram” option creates one histogram for each selected channel, displaying the lines for multiple selected images in different colors (Figure 14a). Alternatively, choose “Boxplot” if this representation is better suited for your purpose (Figure 14b). For optimal visualization, this option is best used with multiple images but not too many channels at the time.

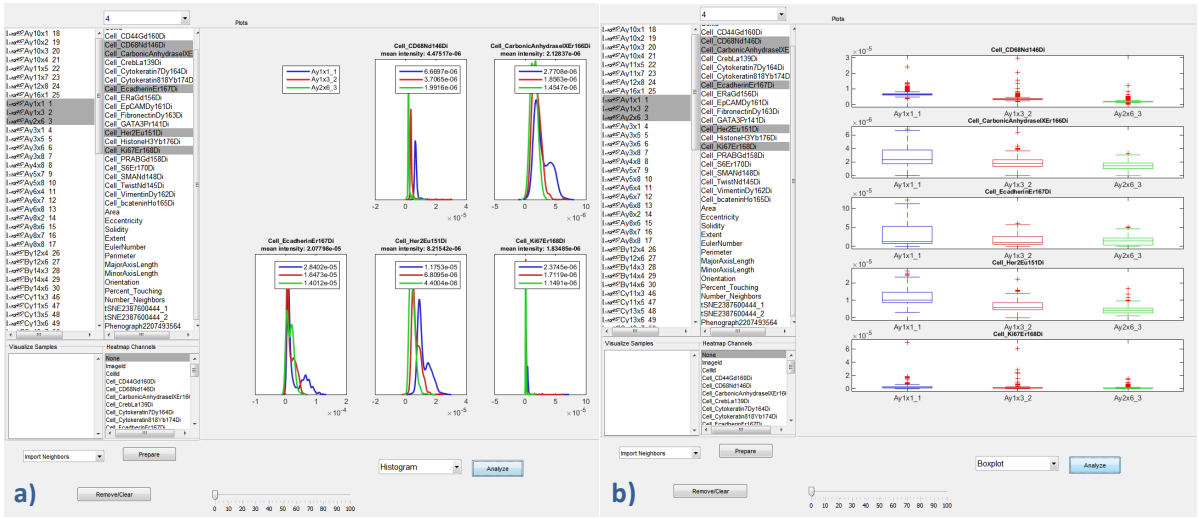

Figure 14

3. Generate a heatmap of the selected gates (y-axis) and channels (x-axis) by choosing the "Heatmap" option (Figure 15a). Checking the "b2r" checkbox will convert the regular heatmap to a more meaningful "anomaly heatmap", where white represents values close to zero and blue or red represent values below or above zero, respectively (Figure 15b). By default the heatmap displays mean values. These can be changed to medians by checking the "median instead of mean" checkbox.

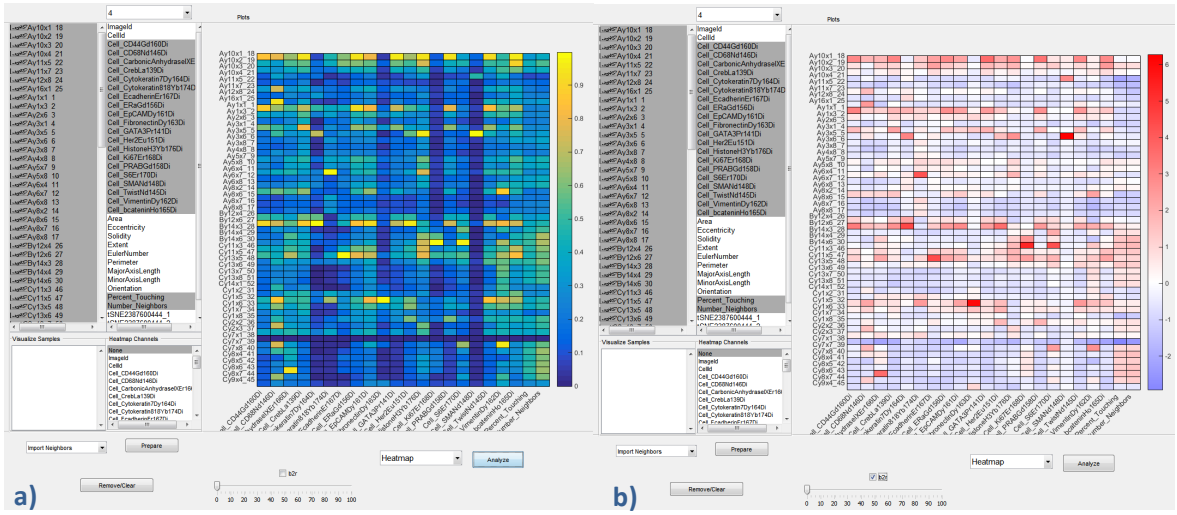

Figure 15

4. A t-SNE dimensionality reduction (Amir et al., 2013) can be run on cells in the selected gates with the selected channels by choosing the "Run t-SNE" option. A seed is set internally to ensure reproducibility. The resulting cell coordinates necessary to display the t-SNE map are saved as two additional channels at the bottom of the list. Generate a scatterplot of these two channels with the "Scatter" option of the "Analyze Options" drop-down menu in order to visualize the t-SNE map (Figure 16a). If no channel from the "Heatmap Channels" box is selected, the colors on the plot simply represent the different selected images. If a channel is selected, the t-SNE map will be overlaid with the heatmap of the marker intensities from the selected channel. When overlaying cells with a heatmap, the "Percentile cut-off" slider

allows the intensities of the highest outliers to be set to the intensity value of a given percentile (Figure 16b).

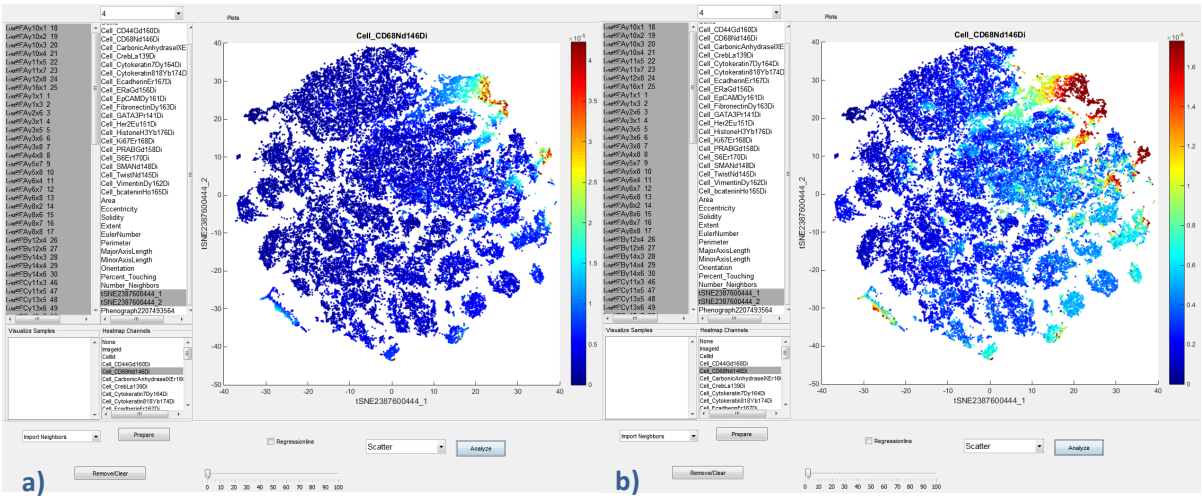

Figure 16

5. The “PhenoGraph” option will cluster the cells of the selected gates into PhenoGraph clusters (Levine et al., 2015) based on the selected channels. Each resulting cluster will be saved as an individual gate and will appear below the rest of the samples. One possibility for visualization of the PhenoGraph clusters is to overlay the t-SNE map with the differently colored clusters by choosing the PhenoGraph result from the “Heatmap Channels” box (Figure 17). PhenoGraph can be used with a random seed or with a fixed seed, which enables reproduction of the same clustering pattern.

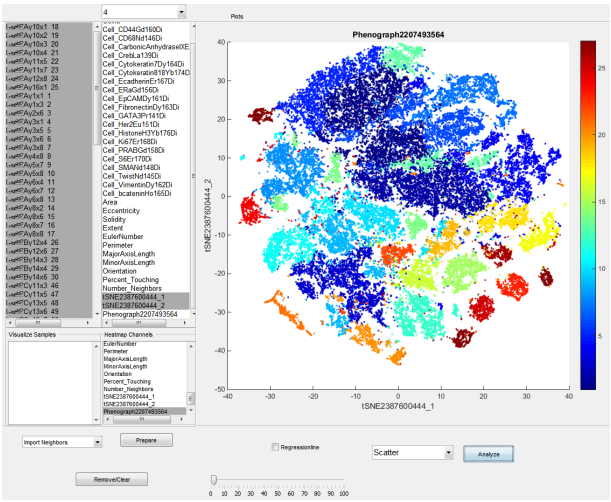

Figure 17

6. Similar to the manual gating on the tiff image, “Gate on plot” allows you to gate on certain cells of interest directly in the scatterplot (Figure 18). The cells in the gated area will be saved as a new gate with a user-specified name.

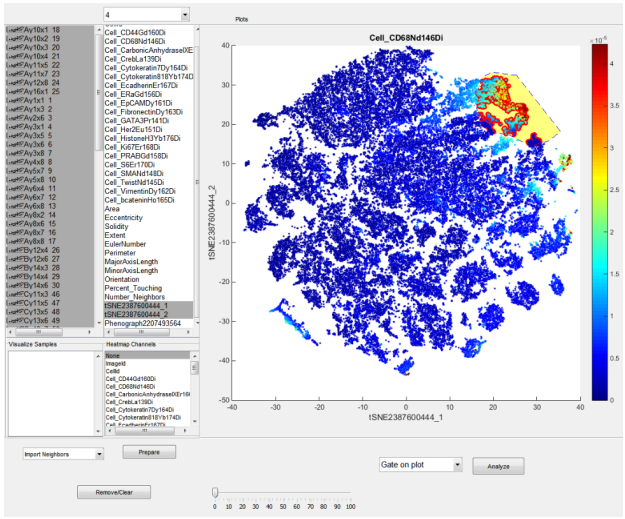

Figure 18

7. The “Highlight samples on plot” option outlines cells of the selected gates on the scatterplot or t-SNE map (Figure 19).

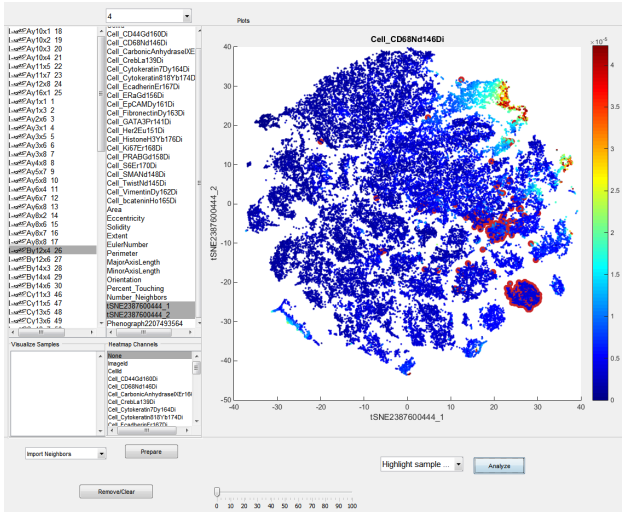

Figure 19

8. A principle component analysis can be run by selecting the “Run PCA” option. As for the t-SNE dimensionality reduction, the resulting coordinates necessary to display the first two components are saved as two new channels at the bottom of the list. Generate a scatterplot of these two channels with the “Scatter” option of the “Analyze Options” drop-down menu in order to visualize the first two principle components. If no channel from the “Heatmap Channels” box is selected, the colors on the plot simply represent the different selected images (Figure 20a). If a channel is selected, the scatter plot will be overlaid with the heatmap of the marker intensities from the selected channel (Figure 20b). When overlaying cells with a heatmap, the “Percentile cut-off” slider allows the intensities of the highest outliers to be set to the intensity value of a given percentile.

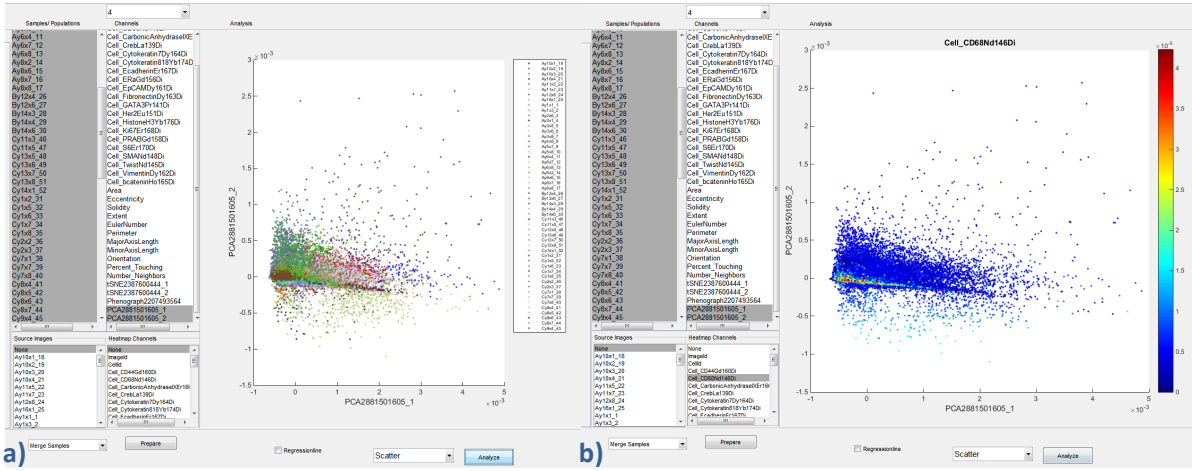

Figure 20

9. A k-means clustering can be run on any two-dimensional plot by selecting the "Run k-means" option. You will be asked to choose an amount of clusters and a number of iterations for the algorithm to run. The resulting cluster assignments for each cell are saved as a new channel at the bottom of the list. Scatterplots can then be overlaid with the color code corresponding to the k-means clusters (Figure 21).

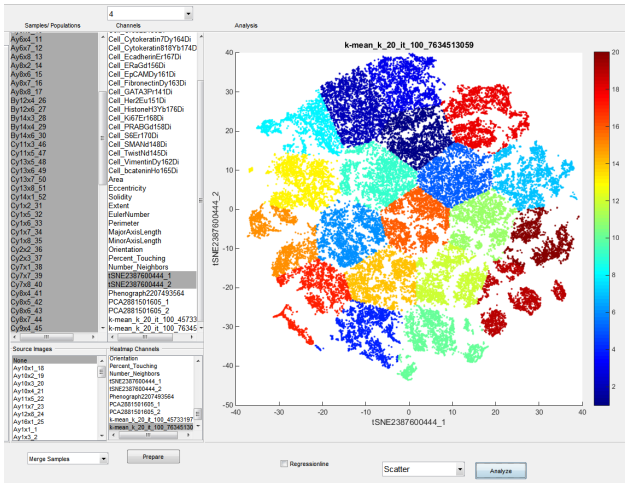

Figure 21

8. Prepare samples

Below the channel and gate selection boxes the "Sample Options" drop-down menu provides several options to prepare the loaded sample data or custom-made gates for further analysis. Click the "Prepare" button to apply the selected option (Figure 22).

1. "Import Neighbors" will create a new gate containing the neighbors of the cells in the currently selected gate.

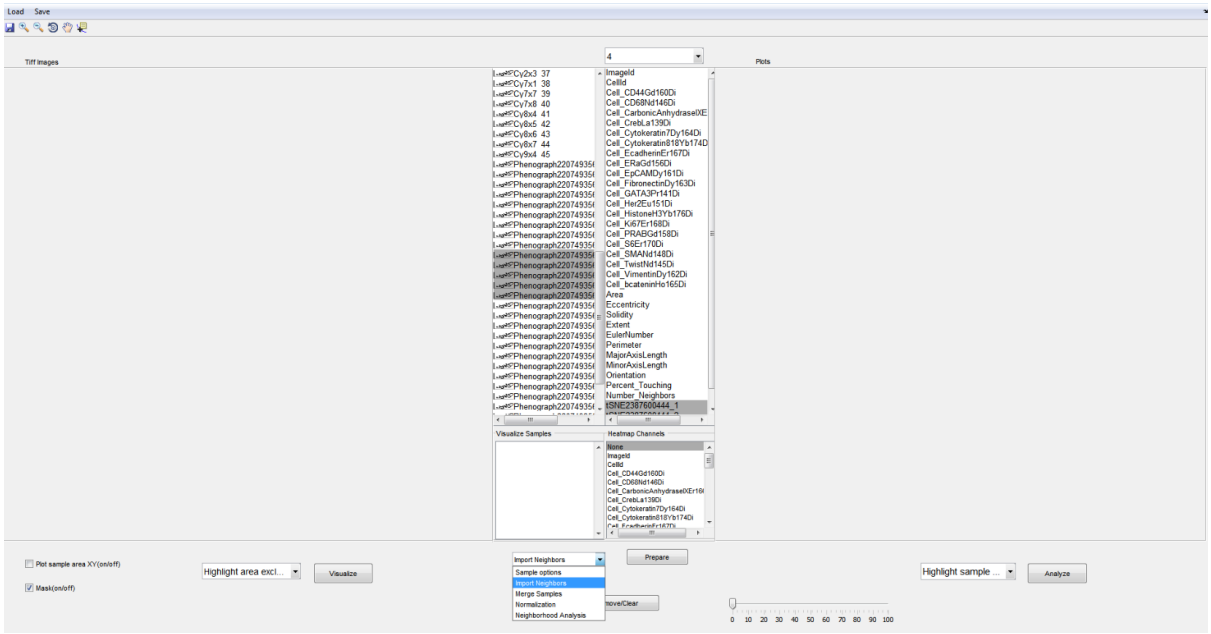

Figure 22

2. With “Merge Samples” you can pool multiple gates into one and give it a new name.
3. “Normalization” will take the Z-score of the selected marker and save it as a new channel at the bottom of the list.

9. Save options

The “Save” button drop-down menu on the upper left yields additional saving options (Figure 23). Select “Save tiff figure” in order to save the currently visualized image(s) on the left side of the interface. The following question prompt will yield the options of saving all open tabs or just the currently visible image. You will be prompted to specify in which folder and file format to save the image(s).

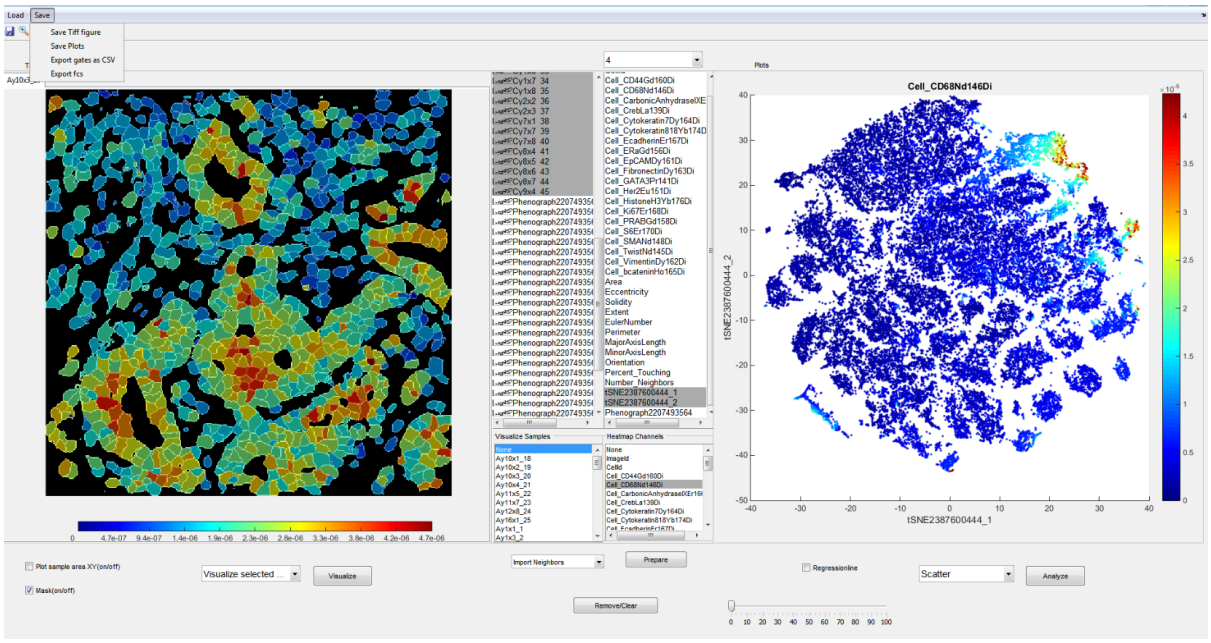

Figure 23

Similarly, the “Save plots” option will open the current plot displayed on the right half of the interface as a figure and prompt you to select a file folder in which to save it as an image (Figure 24).

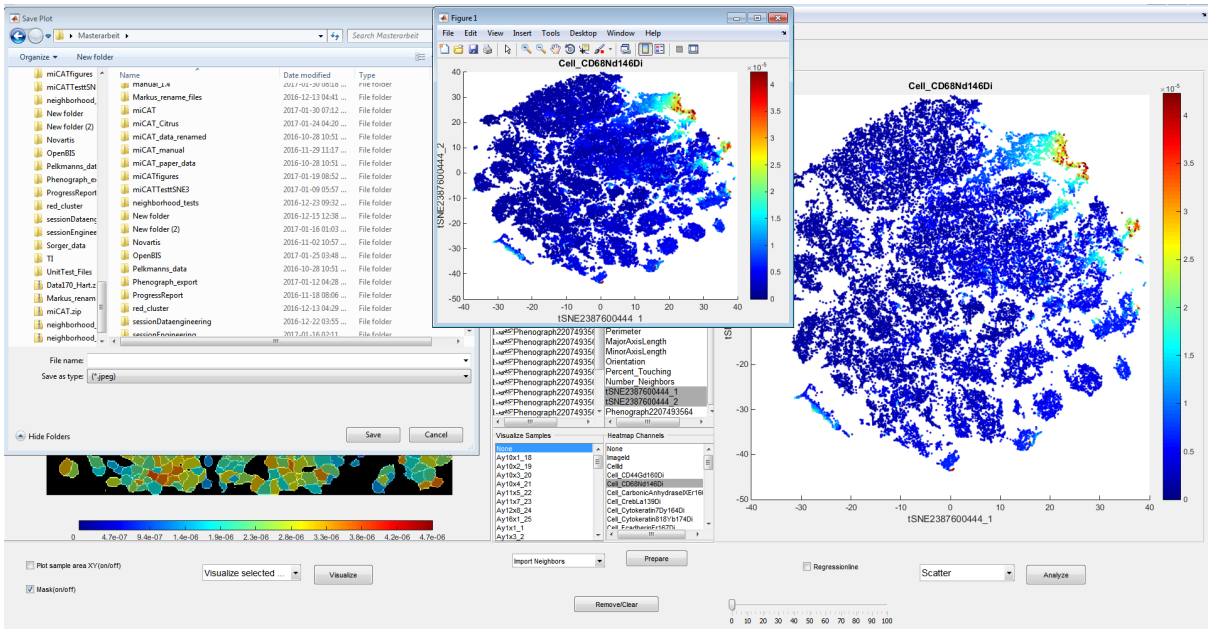

Figure 24

Click “Export gates as CSV” in order to save all of the information for the currently selected gates to your “Custom Gates” folder as CSV-files. These files contain the image and cell identifiers as the first two columns, followed by one column for each marker, each containing the measured intensities. After the marker columns the rest of the features (X / Y position, cell size, percent touching, number of neighbors, etc.) and their values are listed (Figure 25).

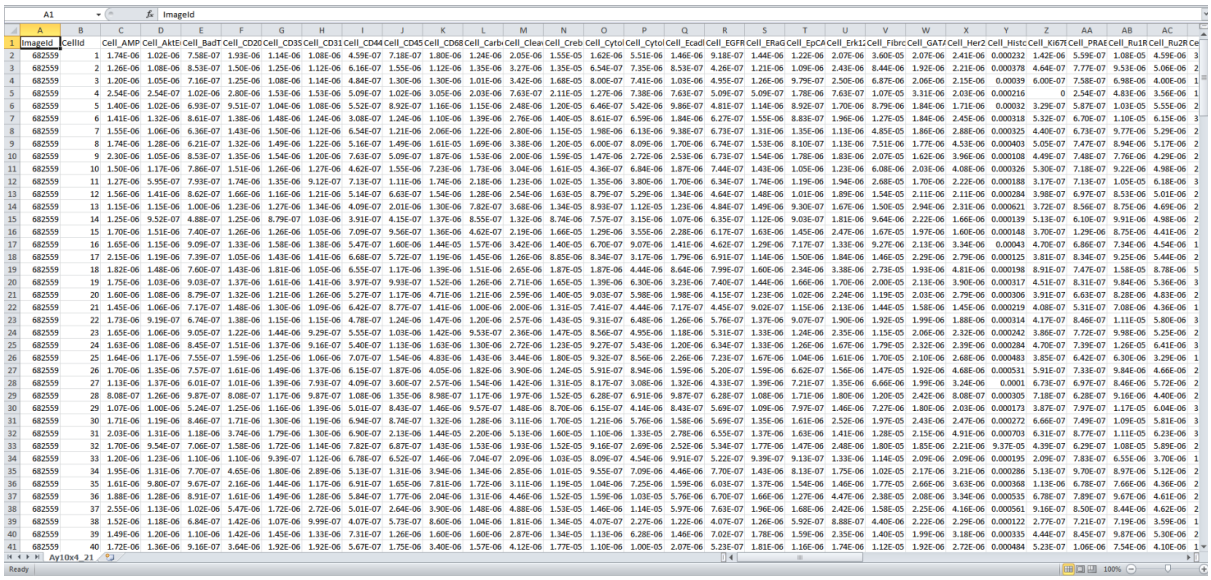

Figure 25

The “Export fcs” option will save all gates as fcs files, including all information and neighbor columns. These can be manually deleted if not needed. This format is suitable for example for a Citrus analysis (Bruggner et al., 2014). The exported fcs files are stored in the “Custom Gates” folder.

histoCAT – Advanced features

1. Neighborhood analysis

(PhenoGraph) cell clusters need to be generated before starting the “Neighborhood Analysis”. The “Neighborhood Analysis” (Schapiro et al., 2017) can be started from the “Prepare Samples” drop-down menu. Select the images (not the PhenoGraph clusters) to be considered from the samples list box. This analysis can be run across all samples at once or on a subset of selected samples.

You will be asked to specify the amount of permutations and the significance cut-off for the P-values. Furthermore you can choose a “special cluster” on which to focus in addition to displaying the results for all PhenoGraph clusters. This will generate an additional output, focused on neighborhoods involving the specified cluster. “Extra information” can be anything you know about the selected group of samples, such as “Grade1Tumors” or “allImages”. This information will appear in the file name of the figures generated during the neighborhood analysis. Additionally, the number of pixels to expand from each cell when searching for neighboring cells can be set, though it has to be within the range pixel expansions, which was set during the initial loading of the samples (Figure 3). Finally, you can specify a fraction (0-1) of samples in which the interactions have to be present in order to be displayed in the heatmap.

The heatmap across all images displays cluster-neighborhood frequencies present in at least the specified percentage (default: 10%) of the images (Figure 26a). For example, if you look at the square on the diagonal corresponding to cluster five on both axes, you are looking at how often cells of cluster five neighbor each other. The squares are displayed in different intensities of red/blue, or they remain white. Red means the cells of these clusters neighbor each other more frequently than

they would in random permutations of the cell cluster labels of each image. Blue means that the cluster cells neighbor each other less frequently than in images with randomly permuted cell labels. The same color code applies to the clustergrams (Figure 26b). Here you see the cluster "interactions" on the x-axis displayed as, for example, "5_5" for the significance of the cells of cluster five neighboring each other. On the y-axis each of the images are listed, such that cluster combinations that significantly deviate from random can be identified on the image level. A separate clustergram is generated showing only the interactions with the "special cluster".

The results of the neighborhood analysis are automatically saved to the "Custom Gates" folder.

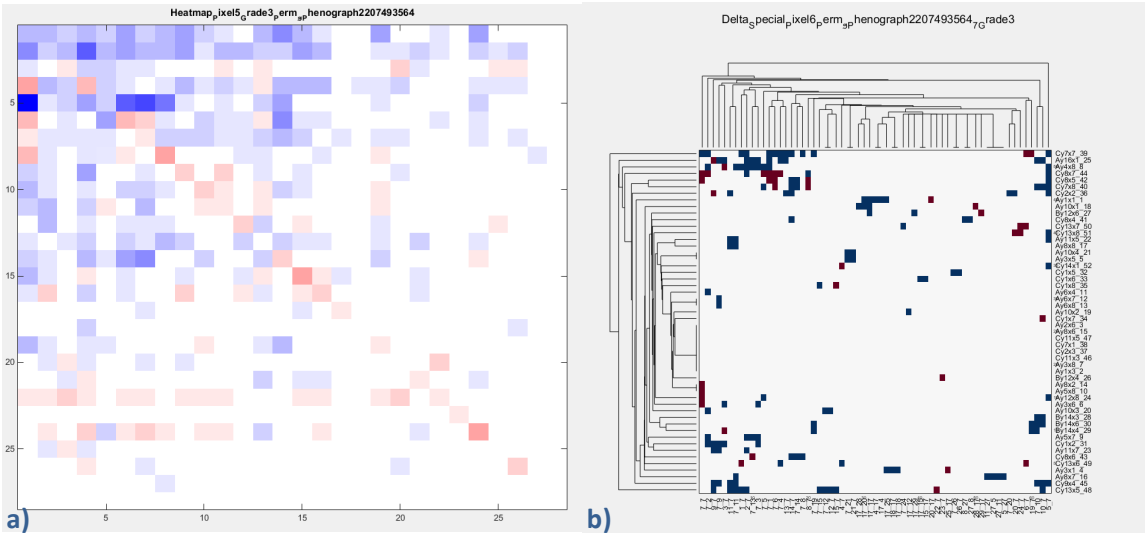

Figure 26

2. Neighborhood - Patch detection

Starting from release 1.73, a new functionality has been added to the neighborhood analysis. The patch detection allows you to specify a minimum number of neighbors that have to be involved in the interaction. If you set this number to one, the original neighborhood analysis will be run. By setting the number of involved neighbors above one, not only pairwise cell type interactions can be detected, but also significantly enriched patches of cells. Please find further details in the corresponding publication (Schulz et al., 2017).

3. Spot detection for RNA channels

The spot detection (Battich et al., 2013) was added to the loading function of histoCAT in release 1.73. This function is specific to RNA channels and is automatically initiated when a channel name contains the string "RNA". You will be able to repeatedly change the parameters for the spot detection, while inspecting the resulting spots on the image until you are satisfied with the settings. In case there are multiple RNA channels, you can click "Apply to all" to apply the current parameters to all of the present RNA channels. Clicking the "Continue" button instead will allow you to tune the parameters for every one of the RNA channels individually. Once all data is loaded, a separate

channel named “Spots_...” will appear for each of your RNA channels. When visualizing the tiff image of this channel in the RGB mode, you will see the individual detected spots. However, the single-cell data of this channel will contain the absolute count of spots per cell. To allow for comparison between absolute RNA spot count and absolute IMC count, an additional channel has been created for every channel containing the string “IMC” in its name. This “AbsCount_...” channel contains the raw IMC counts per cell instead of the mean value. Of note, usually IF images have a higher resolution than IMC images. Therefore, to overlay IF and IMC images the IMC images are scaled up to the resolution of the IF images. The “AbsCount…” intensities are thus multiples of the original IMC counts by the factor of image upscaling. Please find further details in the corresponding publication (Schulz et al., 2017). Starting with release version 1.74, “AbsCount…” channels are not required to run the RNAspot detection.

4. Custom clustering

The custom clustering enables the user to combine an unsupervised clustering method (PhenoGraph) with previous knowledge (manual gating) about the data. This function was added with version 1.73. The result of a PhenoGraph run sometimes does not separate a very specific cluster of cells, which is of interest to the user. To enforce the assignment of this group of cells to a separate cluster, you can manually gate on one or multiple cell populations, as described above. Select the resulting gate(s) from the samples list box and start the “Custom Clustering” from the “Sample Options” drop-down menu. The subgroups of cells in the selected gates will be considered individual clusters, while the rest of the cells, not contained in any of the selected clusters, will be assigned to the clusters of a previous PhenoGraph run. To distinguish the custom clusters from the PhenoGraph clusters, the user-defined clusters are named in steps of hundreds (first custom cluster = Cluster 100, second custom cluster = Cluster 200, …). Please find further details in the corresponding publication (Schulz et al., 2017).

histoCAT – For power users

The histoCAT code was written in Matlab version 2014b. For optimal performance be sure to have this version installed when running histoCAT from the source.

1. Modular build of histoCAT

histoCAT is built modularly, which yields the advantage that adding new features is easy and does not require changes in any of the existing structure. In general, features in histoCAT must include only two basic scripts: one callback from the GUI and one script executing the function. The main functions are not linked to the GUI and can be run independently.

2. Data retrieval

All data stored for the current session and necessary to perform any function can be retrieved from the GUI handles or included manually without the GUI. Throughout a session, the data are kept in an

fcs format structure. There is one main matrix containing a column for each channel and a row for each individual cell of each image. This matrix is continuously updated during the session and will therefore also contain the custom gates and channels. The corresponding channel names for each image are saved in a cell array. All individual tiff files and corresponding masks are stored in a multidimensional matrix structure.

References:

Amir, E.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. *31*, 545–552.

Battich, N., Stoeger, T., and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nat. Methods *10*, nmeth.2657.

Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., and Nolan, G.P. (2014). Automated identification of stratifying signatures in cellular subpopulations. Proc. Natl. Acad. Sci. U. S. A. *111*, E2770-2777.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*, 184–197.

Schapiro, D., Jackson, H.W., Raghuraman, S., Fischer, J.R., Zanotelli, V.R.T., Schulz, D., Giesen, C., Catena, R., Varga, Z., and Bodenmiller, B. (2017). histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat. Methods *14*, nmeth.4391.