Standard Operating Procedure

SOP details

Title	Identification and removal of TopoUnit outliers
Description	This SOP describes the steps that are required to identify, visualize, and remove TopoUnit outliers.
Author	Aliaksei Vasilevich; Miguel Faase; Tim Kuijpers
SOP number	4.4
Version number	4

1 Purpose

To identify different types of errors and outliers in the screening data and to exclude them from downstream analysis.

2 Principle

Many systematic and stochastic errors can occur in a TopoChip screen such as, image artifacts, inhomogeneous cell seeding, dust particles, miss-segmentation, or out of focus images (to name a few). This SOP guides you through the removal of those outliers.

The SOP describes the following steps:

1. Statistical calculation of outlier TopoUnits.
2. Checking the images of the proposed outliers to verify the nature of the outliers.
3. Excluding outliers from downstream analysis.

Remember that for screening we use the adagio: “Garbage in is garbage out”.

2.1 Interquartile ranges

The outliers mentioned in step 3 are calculated in step 1 based on the interquartile range (IQR). The IQR is a measure of statistical dispersion within a distribution. It is defined as the difference between 75th and 25th percentiles, or in other words, between upper and lower quartiles, IQR = Q3 − Q1. These quartiles are also typically represented in a whisker-box plot of the data. The normal data range is then defined with Q1–1.5*IQR as lower limit and Q3+1.5*IQR as upper limit. Data points outside this normal range are considered as outliers and are removed from the data.

The exact feature upon which the outlier TopoUnits are removed is decided by the user. It is important to choose a variable from the CellProfiler data that is experimentally relevant, such as cell count or integrated intensity of a cytoskeleton marker.

3 Important to know before starting

1. Screening data contains different types of erroneous data, both systematic and stochastic, due to technical variabilities in the process. These errors can be a source of outliers and can skew screening results when left uncorrected.
2. One can identify outliers by observing the raw image data or segmented image data.
3. Common sources of outliers in the TopoChip screen are
   - Out of focus images
   - Nonhomogeneous cell seeding
   - False segmentation of cells
   - Fluorescent artifacts
   - Autofluorescence of the TopoChip polymer or coating
   - Incorrect mounting
4. Representation of measurement values, derived by CellProfiler, on the TopoMap generates spatial patterns related to the location of the TopoUnits on the chip. This script creates whole TopoChip heatmaps showing effects not visible by simply observing single raw images.
5. Read the “Data-analysis strategies for image-based cell profiling” review article to learn more about typical data analysis approaches for analyzing imaging-based screening data https://www.nature.com/articles/nmeth.4397
6. Suggested free reading to get familiar with the modern statistical approach in Biology https://web.stanford.edu/class/bios221/book/introduction.html.
7. Have an understanding of the following concepts:
   - Mean
   - Standard Deviation (SD)
   - Median
   - InterquartileeRange (IQR)
8. Read the article about IQR outlier removal method https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

4 Required materials

4.1 Requirements

1. Completion of SOPs 4.1, 4.2, 4.3.
2. Jupyter notebook “identifyOutliers_20210504.ipynb” (located in your folder TopoScreen Data Analysis, SOP 4.1)

5 Procedure

5.1 Removal of TopoUnits based on the IQR

5.1.1 Load the imaging data with FeatureIdx

1. Open the Jupyter notebook “IdentifyOutliers_20210504.ipynb” located in the TopoScreen Data Analysis folder.
2. In Jupyter notebook, step 1 (“load libraries and set paths to all files”) will load the image dataset generated in SOP 4.3 from the location “/DataAnalysis/”. This image dataset contains all the numerical data of the TopoChip screen.

3. In case the program will return "*File not found, check if you finished aligning the images to the TopoMap"* there are two reasons for this error:
    a. You did not align the image to the feature idx and have to go back to SOP 4.3.
    b. You moved the file "imageWithFIdx.csv" from the default location of this workflow and should move the file back to the folder "/DataAnalysis/"

5.1.2 Visualizing measurements on TopoMap

1. Section 2.2.1 of the Jupyter notebook has the function *showOnTopoChip* ,which can display the data on TopoMap. The function requires the following input:
    a. *ScreenData* – a predefined variable you don't need to adjust.
    b. *featureOfInterest* - Measurements from the CellProfiler screen that you'd like to have displayed. You can change this value to show a heatmap of whatever feature you want to visualize. The feature of interest should be a column header in the TopoChip data frame.
    c. *zoom*- Zoom factor, which specifies the range of Rows and Columns that should be displayed, by default function shows all the TopoChip.
2. The output of this function is a heatmap (Figure 1) showing the distribution of values between the minimal and maximal value of the *featureOfInterest*. Every square on this 66x66 map corresponds to a single TopoUnit (image), and its coordinates match with the TopoMap. The function displays the colormaps of each TopoChip replicate, mean, sd, and median of all the replicas. Note: if only one or two TopoChips are being screened, no standard deviation can be computed.
3. Figure 2 demonstrates the output when the zoom is applied. In this example the row ranges from 40 to 52 (out of 66) and column from 16 to 40 (out of 66).

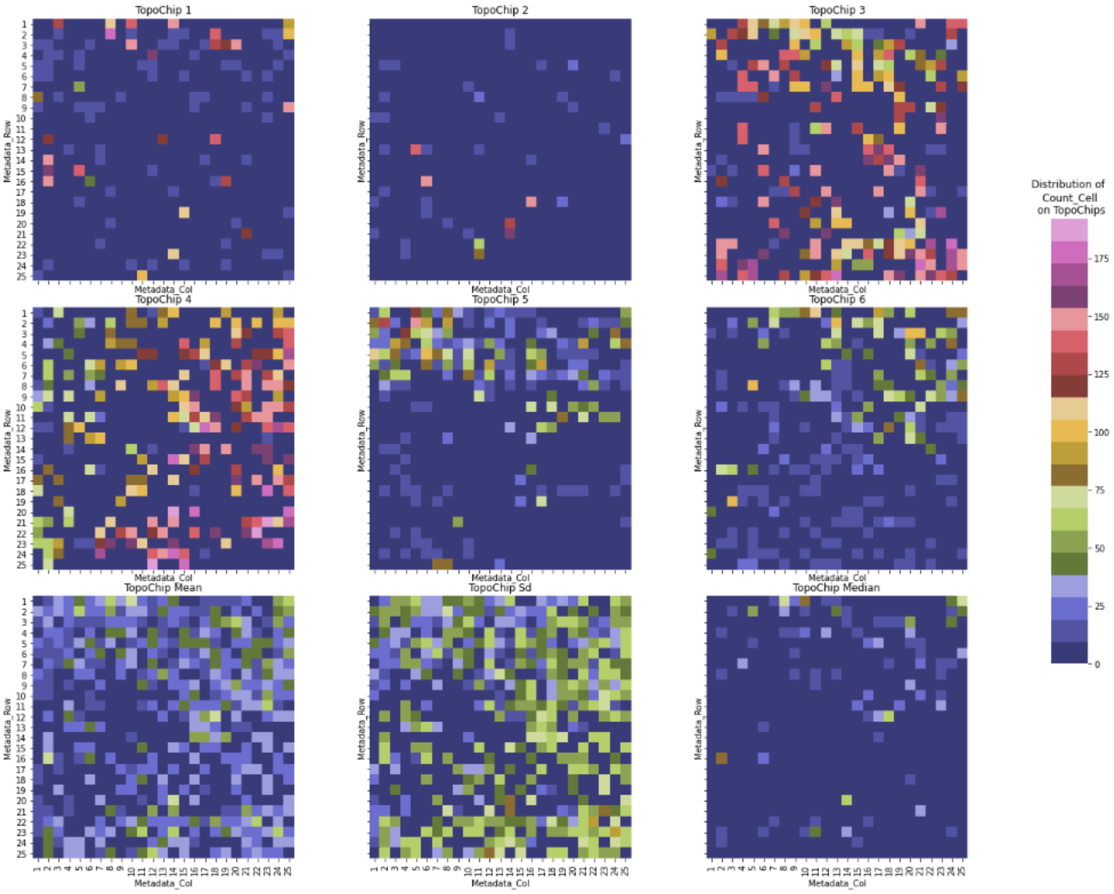

Figure 1 Heatmap showing the distribution of the values for the featureOfInterest. The legend on the right shows the meaning of the color coded values.

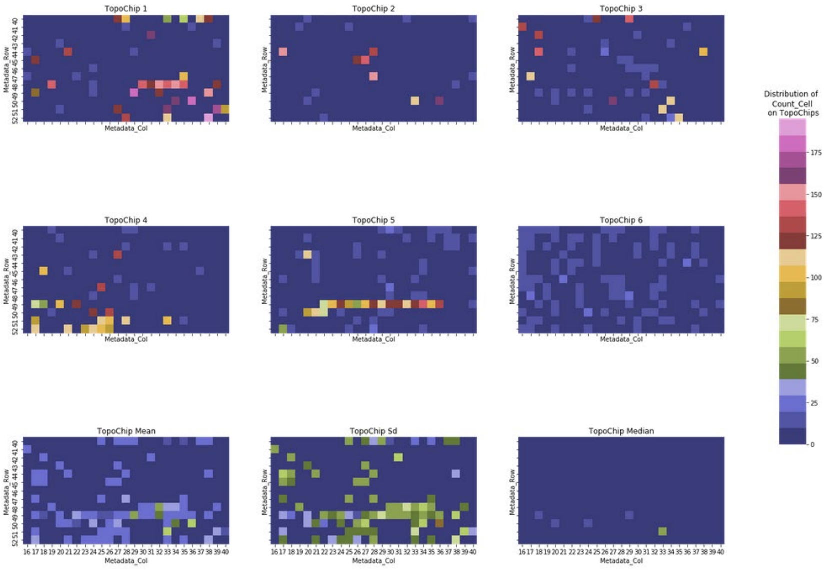

Figure 2: Application of zoom in showOnTopochip function

5.1.3 Display the distribution of Measurement values with histograms

1. In the Jupyter notebook step 2.2, the function *showHistPerTopochip* displays the histograms per replica, mean, standard deviation and the median of the replicates for the defined feature of interest (Figure 3).

2. You can change the feature of interest to display the histogram for your specific feature of interest. Change the string in variable *"featureOfInterest"*, which is *"Count_Cell"* by default to whatever you want to see.

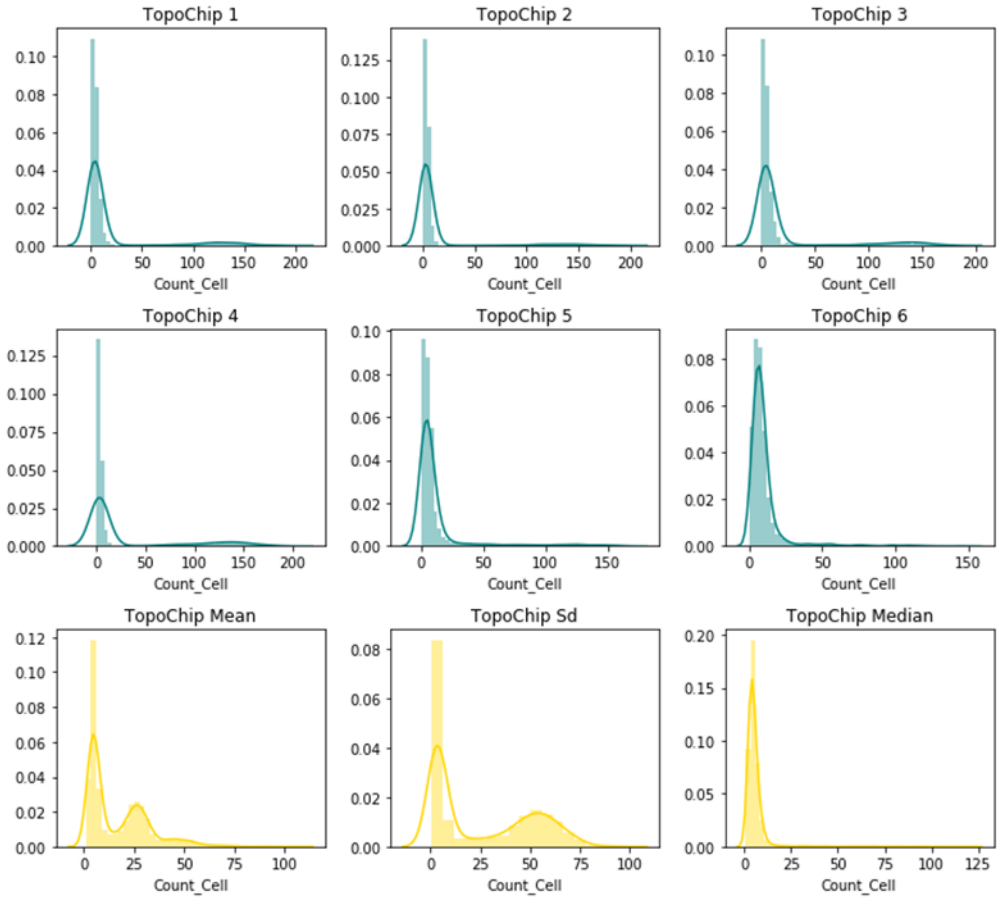

Figure 3: Output of the showHistPerTopoChip function.

5.1.4 Display the raw images for Measurements within a specified range

After observing the distribution of values for certain measurements, it is useful to observe the raw images to understand the cause of the unusually high or low values. For example, the images can be blurred, or contain staining artifacts. To help with identifying the nature of outliers, the function *showImages* can be used.

You are free to select one of the three channels you have acquired during image acquisition. To set the channel of interest, you have to specify the columns which contains the filename. This can be done in the Jupyter notebook during step 2.3. First, look for the column name you want to select and copy this name to the variable in *fileNameOfInterest.* When you specify a *fileNameOfInterest* not in the data frame, you receive a *Key Error* and should rename *fileNameOfInterest.*

The second variable you have to specify is the *valueRange*. This range selects the images which have a value in this range for the feature of interest. For example, when we select the feature of interest to be cell count and the *valueRange* as [5,9], we only select the images with a cell count between 5 and 9.

The visual inspection of the raw numerical data can be used to determine if you have to improve your CellProfiler pipeline, or if you have to remove a TopoChip.

5.2 Remove of OutlierTopoUnits using IQR

With 2174 TopoUnits per TopoChip and 6-10 TopoChips per screen, it is of course not possible to remove individual outlier TopoUnits by hand , which is why we use IQR as mechanism.

1. Select a number of parameters calculated by CellProfiler to filter the data. We advise to select at least the following parameters (now called featureOfInterest), but see for yourself what you deem important:
   a. Metrics from the MeasureImageQuality module in CellProfiler:
      i. Total intensity channel 1 (name depends on your staining).
      ii. Total intensity channel 2.
      iii. Total Intensity channel 3.

   This module collects measurements indicating possible image aberrations, e.g., blur (poor focus), intensity, saturation (i.e., the percentage of pixels in the image that are at/near the maximum possible value, and at/near the minimum possible value). The total intensity can be used to identify imaging artifacts that cause a high intensity peak.
   b. Cell count
      i. When the cell seeding is not homogeneous across the whole TopoChip, filtering based on cell count will remove TopoUnits which have a higher cell count because of this effect.

2. To remove outliers the function *removeOutliersWithIQR* is used. To run this function, you have to define two input parameters:
   a. Measurements that are used to remove outliers (*featureOfInterest*)
   b. Screening data (this variable is pre-defined and does not have to be changed)
3. The function has two modes:
   a. **Mode 1**: *perTopoFeature* is set to *False.*The outliers are estimated based on all the screening data, including other TopoChip replicates.
   b. **Mode 2**: *perTopoFeature* is set to *True*. The outliers are estimated based on the TopoFeature without taking the TopoChip replicates into account.
   Definition TopoFeatue: the feature of interest, calculated by CellProfiler.

   Removal of outliers based on mode 1 guarantees that the TopoUnits that consistently have high or low values will not be excluded from the analysis. For example, if one of the TopoUnits has a very high cell attachment, it could be removed when it is above the IQR threshold. However, when this behavior for one TopoUnit is observed over all replicates, it can be due to the surface design and therefore of relevance. . By choosing mode 1, this TopoUnit is not discarded as an outlier but will be used for further analysis. Mode 1 is advised to be used in the IQR function.
4. The output of the function is a datafile, called *rawImagedataFI_outlierRemoved*, visualized by a boxplot to show the effect of outlier removal.
5. After the removal of the outliers, data can be visualized with the *ShowonTopoChip* function discussed above. Removed data will appear as white.
6. Figure 4 compares the outlier removal by using the two distinct methods, *PerTopoFeature* set to True and to False, with data from one of our previous screens. When outliers were estimated based on all the data (*PerTopoFeature*=False), no TopoUnits with high cell counts were left, and the maximum cell number per TopoUnit was 16. When outliers were removed in the mode (*PerTopoFeature*=True), surfaces with the same design that consistently has high cell numbers

were preserved. At the same time, it is clear that outliers that had a unique spatial pattern, present only into single replicas were removed in both cases.

7. It is important to confirm that consistent high values per TopoFeature are related to the effect of the surface design, not the systematic technical error. For confirmation, raw images should be observed to identify whether these are truly erroneous.
8. In this analysis, it is important to perform several iterations between observing spatial distributions of measurements of interest on the TopoChip, relating raw images of unusually high or low values, and applying different outlier removal strategies. It is possible to combine different outlier removal steps, by applying outlier removal function sequentially, by reapplying the function on the results of the same function or by filtering using different features. It is important always to check whether outliers are truly outliers by looking at the images.

Remember that image analysis is always about going back and forth between your images and the quantified data!

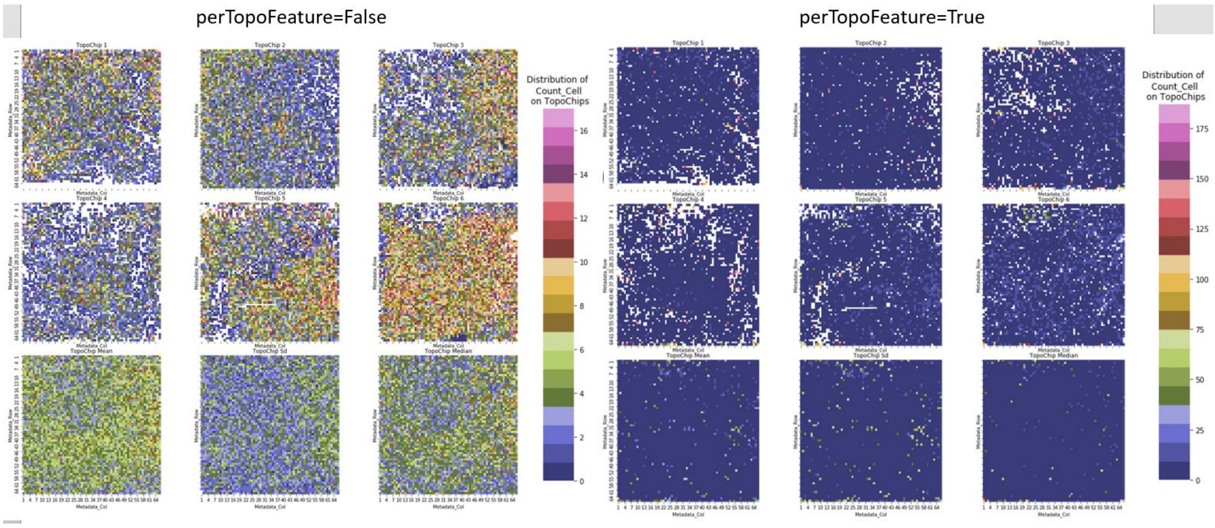

Figure 4: Comparison of outliers removal by using two modes of *PerTopoFeature* (True and False).

5.3 Save the output of your cleansed data.

1. Save the new data without the outliers by running *Step 4: save the new image object*. This will write the file *imageOutliersRemoved.csv* to the folder */DataAnalysis/*. Please do not change the location of this file since the next SOP will look for *imageOutliersRemoved.csv* in this folder.
2. Save the *Jupyter notebook* to keep a documentation of your workflow.