Standard Operating Procedure

SOP details

Title	Rank surfaces to feature of interest
Description	This SOP describes the steps that are required to rank the surfaces based on the performance in the screen.
Author	Aliaksei Vasilevich; Tim Kuijpers
SOP number	4.5
Version number	3

1 Purpose

To rank the surfaces based on screening results for downstream analysis and experimental validation.

2 Principle

This SOP guides the data analysis on ranking of the surfaces with accompanying statistics and visualizations. It explains the key components, in- and output for the functions and how to interpret them. The SOP also provides references to the statistical methods used in the tutorial. Steps that are covered in this tutorial are:

- calculate the mean value of the measurements of interest per unique surface
- calculate the variability of the selected measurement per surface
- calculate the p-value of measurement value distribution within replicas.
- filter out surfaces based on the p-value and variation
- rank surfaces based on the selected average
- plot S-curve
- plot boxplots of surfaces with the highest and lowest ranking score
- validate the results with raw images.

In SOP 4.4, we have filtered outliers based on the absolute values of a measurement. In SOP 4.5, we will look at the distribution of the TopoUnit replicates to remove those with a significant different distribution.

3 Important to know before starting

1. Every surface on the TopoChip is duplicated, and located as shown on the schematic below, where A2 is a duplicate of A1 and B2 is a duplicate of B1.

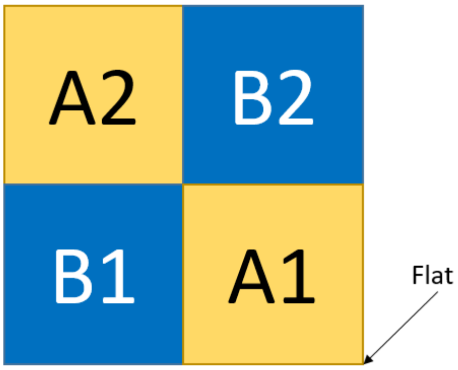


Figure 1 Schematic showing the position of the duplicates.

2. We test if the measurement values of interest are similar among the duplicates by performing Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests. The null hypothesis is that the samples are drawn from the same distribution, When the null hypothesis can be rejected, the samples (thus TopoUnits with the same feature idx) do not follow the same distribution which can be because of experimental errors. Accessible explanations of these can be found here:
    - Kolmogorov-Smirnov (KS): http://www.real-statistics.com/non-parametric-tests/goodness-of-fit-tests/two-sample-kolmogorov-smirnov-test/
    - Anderson-Darling (AD): http://www.real-statistics.com/non-parametric-tests/goodness-of-fit-tests/anderson-darling-test/

    A low p-value for either test means that the measured distributions between replicas are different, in which case we want to exclude surfaces with low p-values as no major differences between replicates are expected.
3. The variability of the measurement value between replicas is quantified by Signal to Noise Ratio (SNR), which is defined as the ratio of Mean and Standard deviation, here is a link to read more about it http://www.statistics4u.com/fundstat_eng/cc_signal_noise.html

4 Required materials

4.1 Workplace

This SOP can be performed in the office or home on your Laptop/Desktop.

4.2 Requirements

1. You should have completed the previous notebooks and obtained the file:
    - *imageOutliersRemoved.csv*
2. Have an understanding of the following concepts:
    - Difference between Mean and Median: https://www.clinfo.eu/mean-median/
    - Continuous distribution: https://machinelearningmastery.com/continuous-probability-distributions-for-machine-learning/
    - Documentation for the Kolmogorov-Smirnov (KS) test function: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html
    - Documentation for the Anderson-Darling (AD) test function: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson_ksamp.html

- Boxplot https://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

5 Procedure

5.1 Load the data to rank the surfaces

1. Load the Jupyter Notebook “RankFeatures.ipynb” and run the cells in *step 1*. By evaluating the cell in *step 1*, the data object with the outliers removed (*imageOutliersRemoved.csv*) will be loaded into the working memory.
2. When the file cannot be located, this step will produce an error. There are a few solutions to solve this error:
   a. Check if you have completed the notebook *2_IdentifyOutliers* and the folder *“/DataAnalysis/”* contains the file *“imageOutliersRemoved.csv”.* If not, complete notebook *2_IdentifyOutliers*
   b. Check if you did not change the location of the file *“imageOutliersRemoved.csv”.* When you copied it to a different location, place the file back into the folder *”/DataAnalysis/”*
   c. Check if you did not rename the folder “*/DataAnalysis/*”. If so, rename the folder to the original name.

5.2 Rank the data based on the feature of interest

1. In the Jupyter Notebook, *Step 2.1* will ask you to select your feature of interest. Note that this can be any feature present in your data and different features will results in different rankings of data.
2. The function *calculateRank* will rank the surfaces based on your feature of interest. The purpose of the function is to calculate different metrics such as statistical tests, average, and variance that quantify the response to topographies and will be used for ranking. The function requires the following input:
   a. ScreenData -The data variable with results of screening data, that also contains Metadata_Duplicate and FetaureIdx columns.
   b. featureOfInterest - Measurement that is used for ranking, that contains, for example, number of cells, expression of the protein of interest. These are variables in the CellProfiler datasheet.
3. Perform step 2.2: *rank the surfaces*. The output of this function, *surfacesRank,* is a data frame that contains p-values for KS and AD statistical tests, calculated Mean, SD, Median, and SNR. The column names of functions is constructed such that it contains prefix “Screen_” to remind that this is the results of the entire screen, followed by the type of calculations, specified above and column name that was used to make calculations.
4. To plot the metrics of the ranked surfaces, run step 2.3. Now you visually inspect the distribution of the p-values, Signal-to-Noise Ratio (SNR), as well as other metrics.

5. Run step 2.4 to display the relation between the different scoring metrics (Figure 2). By exploring the plots, you will be able to understand the relation between p-value, variability (SNR), and ranking metrics, such as mean or median.

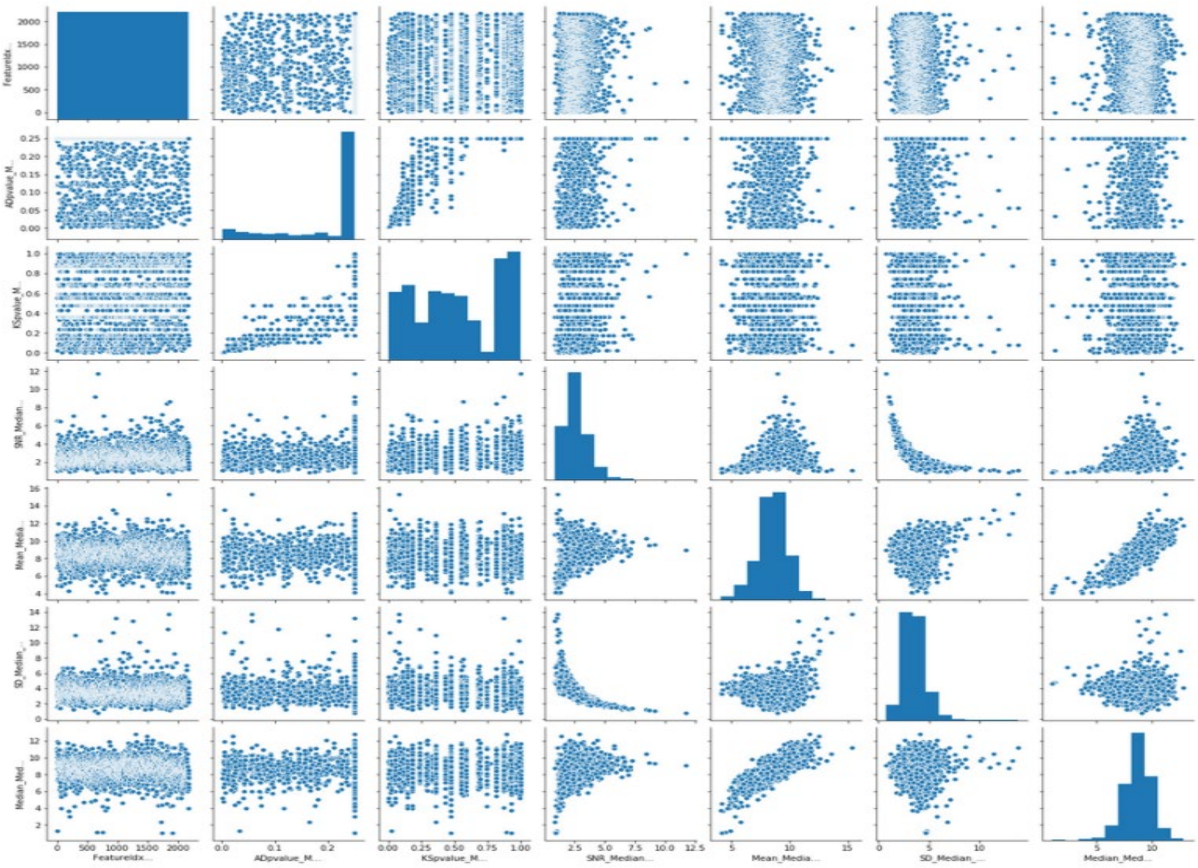
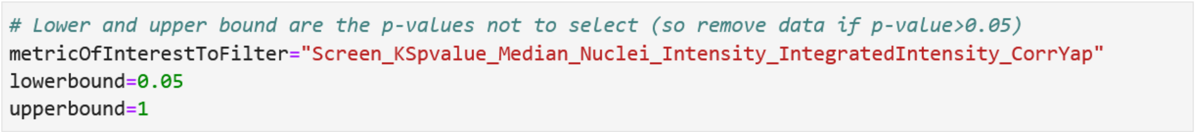


Figure 2: The output of the pairplot function on the surfacesRank data frame.

5.3 Filtering surfaces based on the metrics

1. In Step 3, you will filter the surfaces based on the statistical metrics we have calculated in step 2.
2. In step 3.1, you will filter the data based on the first statistical metric of interest. To run the function *filterRank*, you have to define the input parameters *metricOfInterestToFilter*, *lowerbound,* and *upperbound.* The input parameters *lowerbound* and *upperbound* specify the range of values you want to include. For instance, when we set the following input parameters:



```
# Lower and upper bound are the p-values not to select (so remove data if p-value>0.05)
metricOfInterestToFilter="Screen_KSpvalue_Median_Nuclei_Intensity_IntegratedIntensity_CorrYap"
lowerbound=0.05
upperbound=1
```

3. Next, in step 3.2, we perform a second filtering based on the Signal to Noise (SNR) values. The SNR values measure the desired signal to the level of noise. Images with a low ratio, thus high

noise levels, can be discarded in this step. You have to specify the lower and upper bound for the SNR by yourself. To determine these values, the distribution plot of the SNR (created in step 2.2) can be used. Note that as input, we used the output generated in step 3.1. Thus, we have applied both filtering on p-value and SNR to our data set. Note that during the filtering step, the flat surface (feature id=2177) will never be removed.

4. Step 3.4 is performed to check how relations between ranking metrics have changed after applying both the p-value and the SNR filtering step.

5.4 Plotting Ranking

1. Surfaces can be ranked based on a value of interest, called *feature of interest* in the Jupyter notebook. By ranking the data, we can correlate the TopoUnit to a certain biomarker, number of cells, or any other biologically relevant parameter.
2. In step 4.1, we plot the rank versus the feature of interest. This will result in a scatter plot with on the x-axis the rank and on the y-axis the feature of interest. The flat surface will be highlighted by a larger blue dot (Figure 3).
3. We can create a boxplot in step 4.2 to visualize the top and bottom ranked surfaces with respect to the feature of interest. This feature of interest can be any output generated by CellProfiler but should be relevant to your screen. For example, we can rank the data based on cell count if we are interested to find surfaces with high/low cell attachment.
4. In step 4.3, we can visualize the raw images of the flat surface.
5. Finally, in step 4.4, we can visualize the raw images of the top and bottom scoring surfaces.

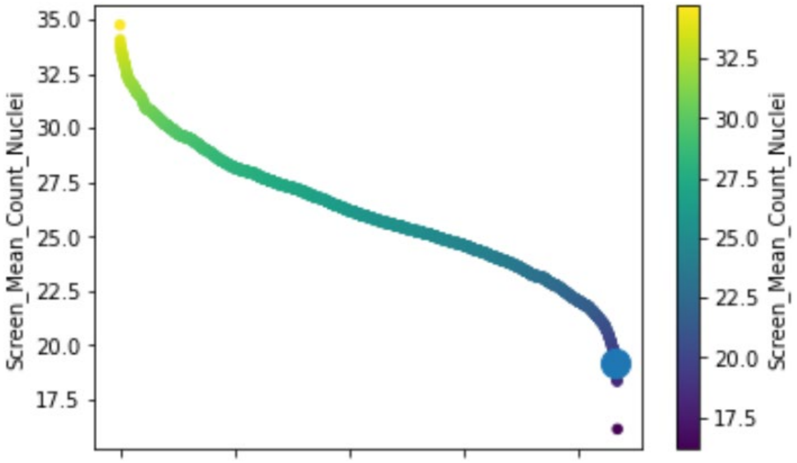


Figure 3 Ranking plot: rank number on the x-axis and the feature of interest on the y-axis (Screen_Mean_Count_Nuclei). Flat surface is visualized by the large blue dot.

5.5 Saving Ranking results

1. Save the ranking results running the cells in step 5. Because you might have investigated multiple features of interest to rank the data, you have to select your final feature of interest to rank the data.
2. Save the *Jupyter notebook* to keep a documentation of your workflow.