Data analysis work-flow for morphological profiling

Sundaramurthy Lab

May 8, 2016

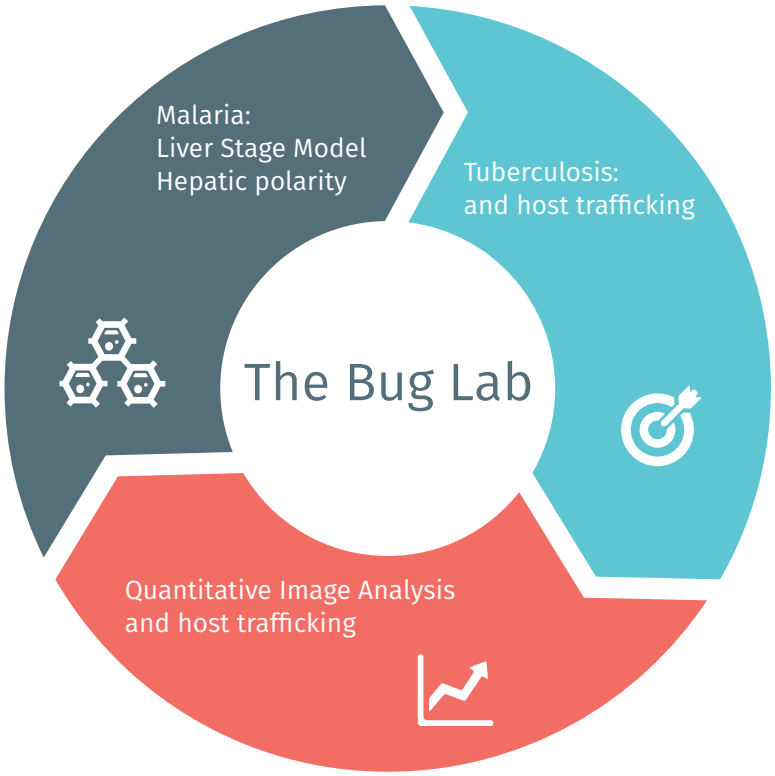
1 Biological motivations

We at the bug lab in NCBS are interested in understanding host pathogen interactions and developing new tools to understand the fundamentals. We are a group of 13 people working on different projects.

1. The projects are broadly divided into 3 categories. First being the malaria group that aims to develop the in vitro liver stage model of plasmodium infection to further the understanding of plasmodium pathogenesis and have an assay for the drug screening. People are also involved in understanding the role of hepatic polarity in the pathogenesis of the parasite.

2. Another group of people are working to understand the tuberculosis pathogenesis and its exploitation of the host trafficking pathways of endocytosis and autophagy. They induce and track the pathways and notice how infection has changed, and vice versa.

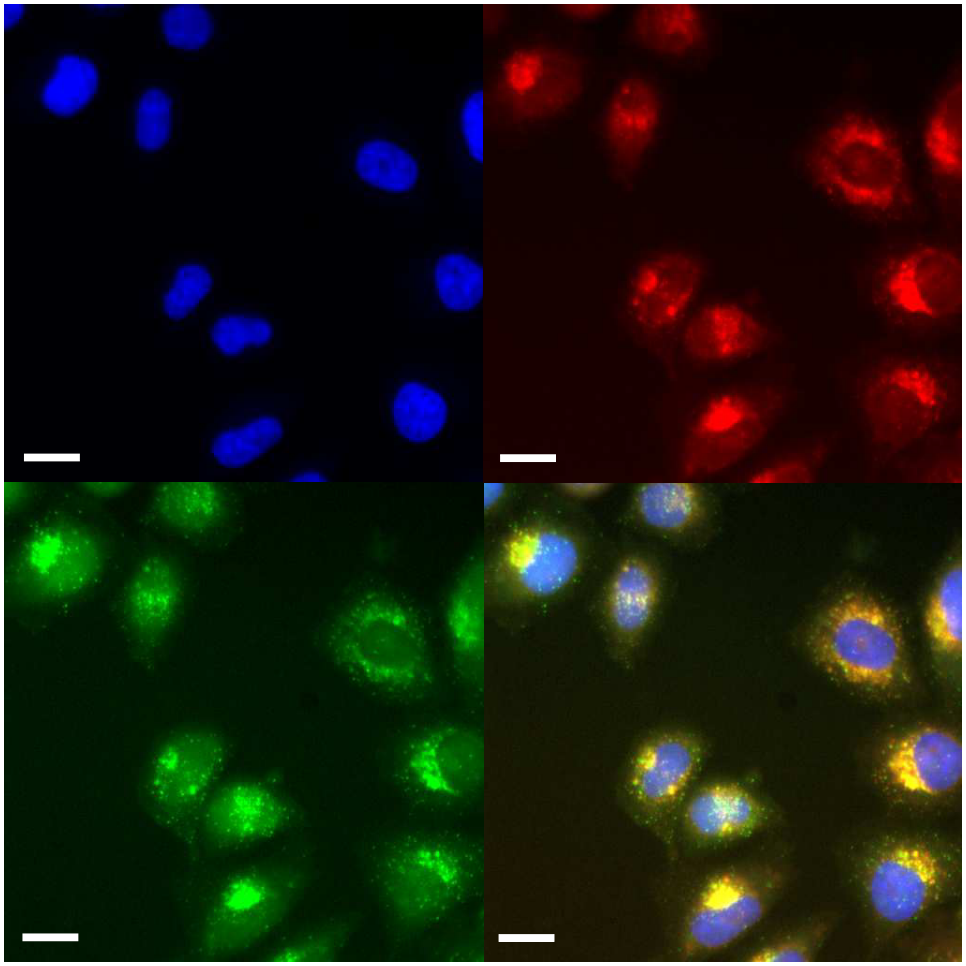
3. And there is another group of people who do quantitative image analysis optimizing segmentation, writing macros for 3D segmentation, developing new data analysis techniques and presenting new ways to look at the data.



In our lab we use widefield, confocal, and EM microscopies. The images can be single slice images, Z-stacks, and high throughput. For image processing CellProfiler is the most common software but MotionTracking, FIJI are also regularly used. For data analysis we primarily use R and KNIME.

2 Image analysis and feature extraction

The primary software used for image analysis and segmentation of the fluorescent microscopy images is CellProfiler.



The image analysis pipeline typically involves the following steps:

1. Rescaling Intensities
2. Illumination Correction: Calculate and Apply
3. Enhance/ Suppress features for easier segmentation of the small puncta
4. Identification of primary object (typically nucleus), secondary object (typically cell), other cellular organs (like endosomes, autophagosomes, lysosomes, infectious agents, etc.)
5. Identify colocalization of different objects
6. Relate objects
7. Measure features (Area/Shape, intensities, spatial spread, etc) of the objects identified. Also calculate per cell mean values for all the object features
8. Export the images of objects identified
9. Export the measurements

MotionTracking, that is built on Pluk platform (http://pluk.mpi-cbg.de/projects/motiontracking), is also regularly used. FIJI is used for 3D segmentation.

3 Image quality control

1. Flag images based on focus score to remove out of focus images
2. Plot plate maps to see if the plate variability exists

4 Data cleaning

1. Experiments conducted at different times can have novel nomenclature used (like dapi, DAPI, Dapi; identify these differences and replace with consistent nomenclature)
2. Different image analysis software will result in their own nomenclature of the feature names, like in MotionTracking the feature names have special symbols like $!. = (), etc$ which are not handled well by R. So those column names are renamed to get rid of special symbols
3. Sometimes additional metadata features have to be added to the data files like the infection status in the infection study that is added based on the bacterial count. Also if binning analysis has to be done, the new Metadata_Bin column is created at this stage
4. There are lot of NAs in the data. We need to get rid of them

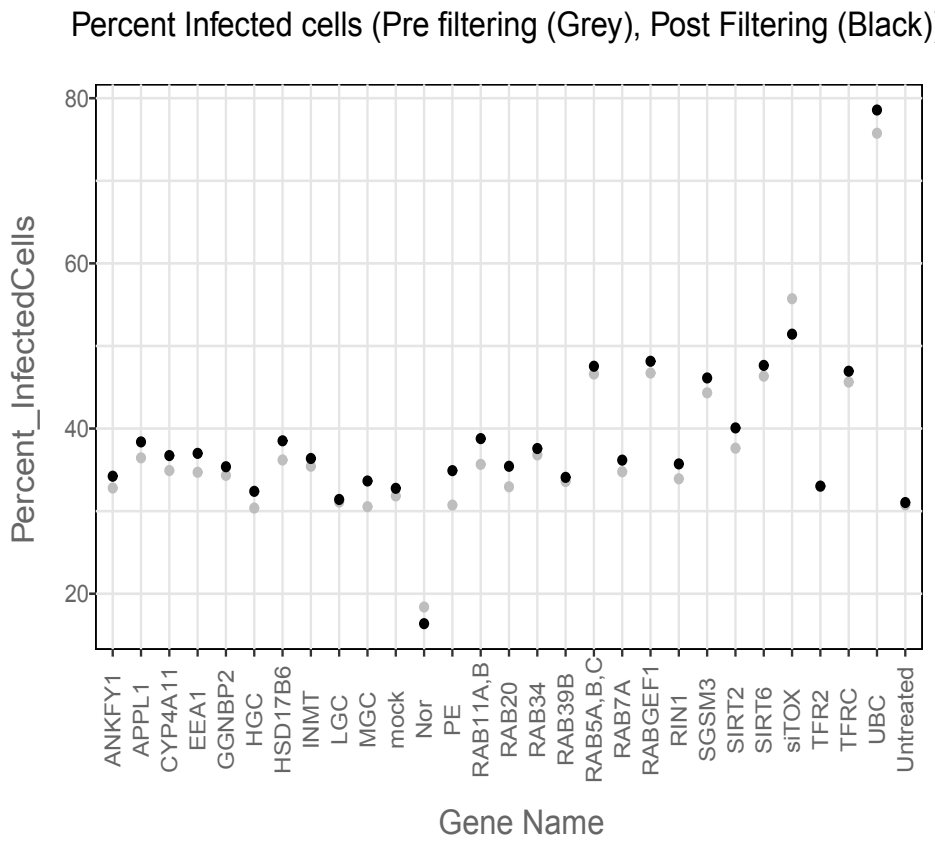
   Get rid of those columns which contain just NA

   Get rid of those rows with more than some percentage of NAs, lets say 50%. This is uncommon.

   There are some NAs that are converted to 0 (like bacterial features in non-infected cells)
5. Features with 0 standard deviation are removed

5 Data filtering

1. There are apoptotic cells/ improperly segmented cells (that are detected based on cell area/ eccentricity (artifacts)),
2. Number of nuclei/cell must be 1.
3. Then we check that the cell filtering has removed roughly equal amounts of cells from each treatment conditions, else there might be an inherent bias in that well, which might or might not be a phenotype, and will need further scrutiny.



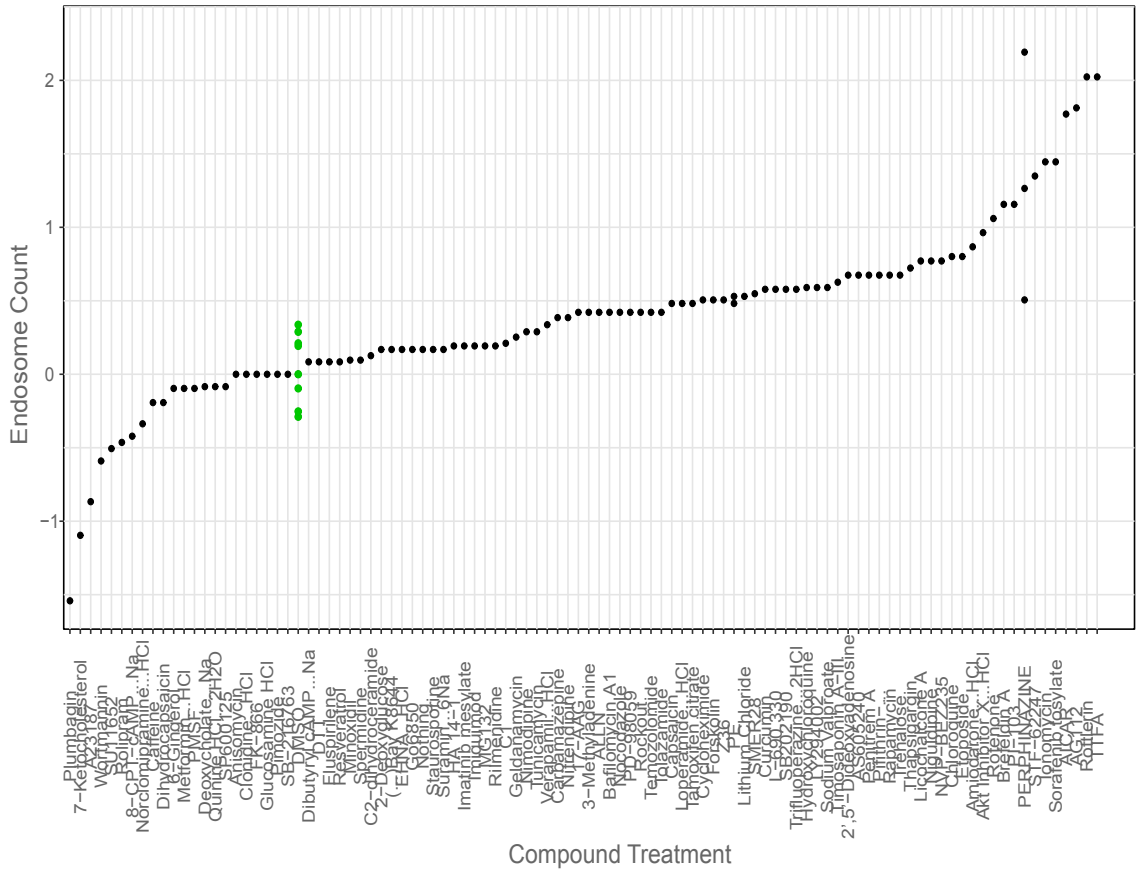
6 Normalize features

Normalization is important if we are to compare the strength of phenotype in different features, and also across assays.

1. Robust Normalization with the formula
   z-score$_{\mathrm{NC}} = \frac{Value - Median[Values_{\mathrm{NC}}]}{MAD_{\mathrm{NC}}}$
   is done typically. It is more resistant to outliers.

2. Classical Z-scoring is required in conditions when the robust normalization results in NAs (which regularly happen in the conditions where median value of a feature is zero, like bacterial counts). In those times we resort to the following formula
   z-score$_{\mathrm{NC}} = \frac{Value - Mean[Values_{\mathrm{NC}}]}{\sigma_{\mathrm{NC}}}$

The per treatments Zscore is shown in this plot and we can see that we get z-scores of ¿ 2 in case of the Positive controls, and multiple negative controls occur in the narrow range.

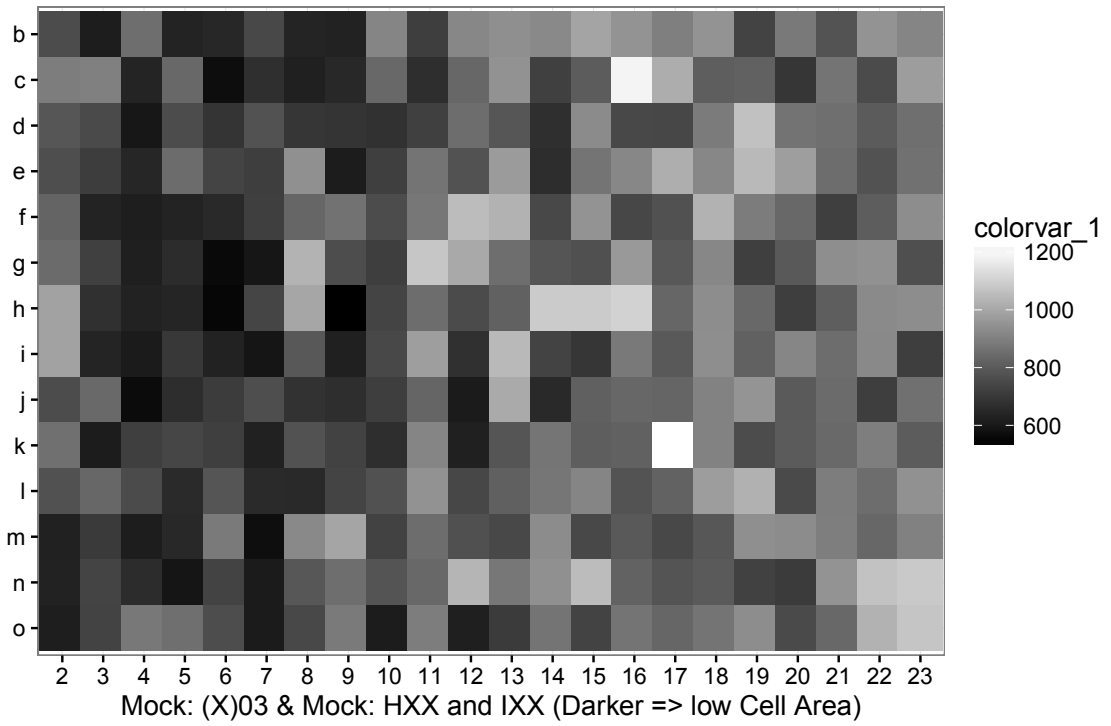


7 Transform features

The feature Transformations like principle component analysis and factor analysis have been tried and did not yield favorable results.
The normalization is also a type of feature transformation that is regularly employed. It is easy to make out the differences in different treatment conditions after transforming the data to Z-scores.
The density plots of all observations are made to identify the parent distribution. If the distribution is flat for some features, those features are removed from subsequent analysis, these features introduce significant noise when feature reduction techniques like PCA are applied. Also these features can not be removed using the redundant feature removal.

8 Correct for systematic effects

When ever a high throughput experiment is conducted, there are redundant conditions typically negative controls that are spread all across the plate in both horizontal and vertical direction that is used to access if there are any systemic plate bias. Plot plate layouts for different known features and visualize the well variability in the plate, if high throughput. We have rarely found a systemic bias in plates.



9 Select features

1. The features are selected based on non-redundancy. Only one feature out of the set of features with redundancy more than 0.90 (pearson correlation coefficient) is selected. The feature that minimizes redundancy with every other feature with correlation coefficient less than 0.90 is selected from a set of redundant features.

2. The features are tagged as those belonging to either of the two categories, metadata space and features space. So that vectorized mathematical operations can be performed on the feature space while grouping the data based on one or multiple metadata columns.

3. Since we have to regularly deal with different plates with different readouts in the same experiment, we mark the common features as "shared features" and others as "unshared features". Shared features are compared across different plates, and compared for reproducibility. If reproducibility is confirmed, then the unshared features can be compared. These comparisons are typically visualized in the form of boxplCots.

10 Create per-well profiles

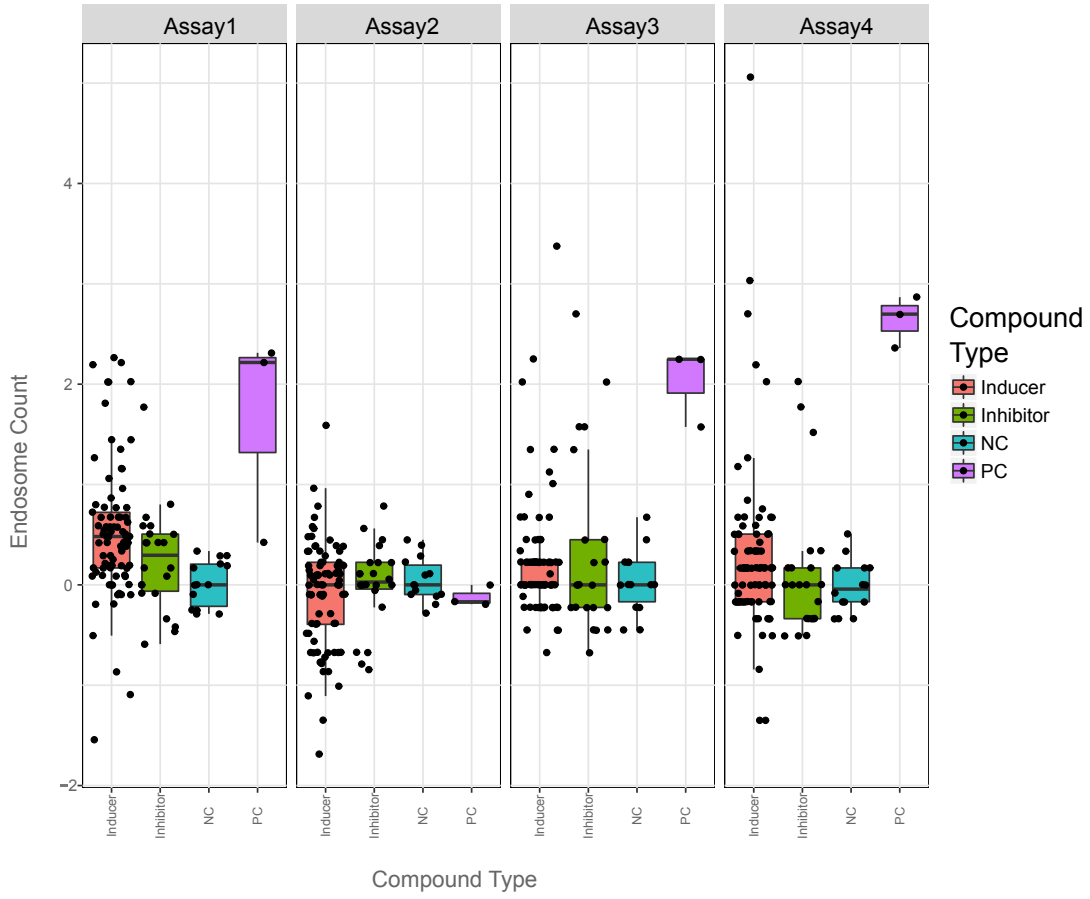
1. The individual sub-cellular object is collapsed to per cell values by taking median or sum as appropriate (for instance, median in case of MeanIntensity, sum in case of IntegralIntensity)

2. Next the per-cell values are calculated by taking the median of all the values in a specific condition. If the data set has multiple redundant conditions like different siRNAs, then they are collapsed per siRNA treatment and accessed for per siRNA changes in phenotypes.

3. If a metadata column for binning analysis was made, then the collapsing is done using median value per bin per condition.

11 Measure similarity between profiles

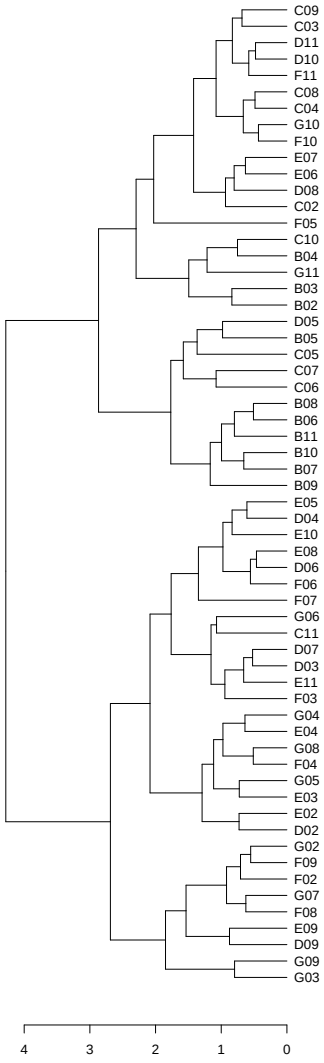
1. The magnitude of phenotype is visualized in box plots, and density plots (or heat maps if conditions to be analyzed are a lot).



2. Similarity between profiles is measured using Pearson correlation coefficient

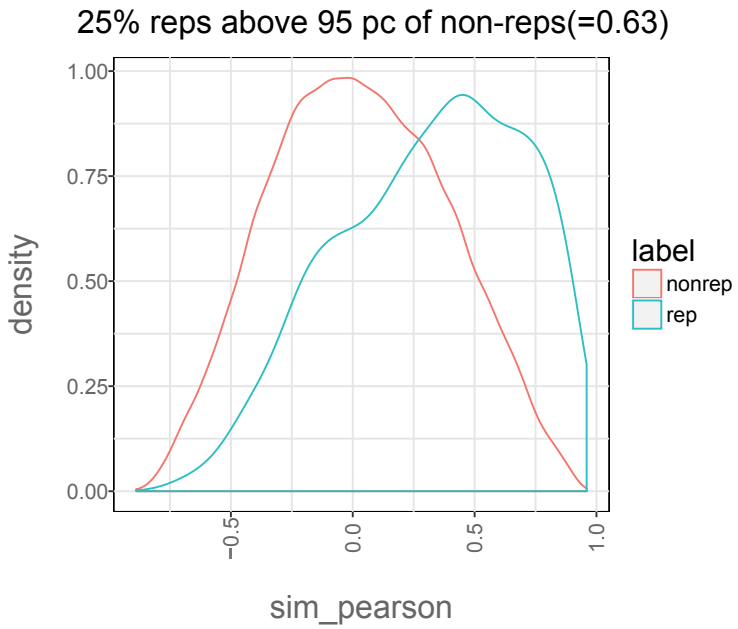
3. The quantitative significance between different conditions and control is done by tests of significance (mostly students T test, and Kolmogorov-Smirnov test). This step puts number to the significance that we already see in the box plots. Inferences of an increase or decrease are made from the box plots plotted previously.

4. Another way to assess the similarity between different phenotypes is hierarchical clustering, that clusters the phenotypes based in similarity amongst themselves, one instance is shown in the following cluster dendrogram.



12 Integrate different plates

Spectral limitations allow for only few channels to be visualized at a time in the biological sample, greatly limiting simultaneous visualization of the different sub-cellular organelles. In order to overcome this problem at the post-imaging stage, data from different plates/ assay but that are a part of same experiment need to be plotted in one plot.
In our lab, permutation testing is being tried to integrate the data from different assays. We define the null and alternate distribution based on the similarity matrix between the reps and non-reps (by reps we mean the correlation coefficient of the same treatments across different plates and by non-rep we mean the different treatments in the same/different plates). Once the significance of phenotype is established, then exact p-value is calculated as shown in the plot.



This p-value for different pair of assays is amenable for comparison and is plotted on the scatter plot. This plot explains when the two pairs of assays are behaving in same way, in other words, if same treatment results in similar changes in two different cellular organelles.