Annotate Cell-Level MEP-LINCs Data

Mark Dane

2015-07-28

Summary

This script prepares cell-level data and metadata for the Pilot MEP LINCs Analysis Pipeline.

In the code, the variable ss determines which staining set (SS1, SS2 or SS3) to merge and the variable cellLine determines the cell line. All .txt data files in the "./Raw Data" folder will be merged with the well (xlsx) and log (XML) data from the "./Metadata" folder.

The merging assumes that the actual, physical B row wells (B01-B04) have been printed upside-down. That is, rotated 180 degrees resulting in the spot 1, 1 being in the lower right corner instead of the upper left corner. The metadata is matched to the actual printed orientation.

The well metadata describes the cell line, ligands and staining endpoints that are all added on a per well basis. There is one mutlisheet .xlsx file for each plate. Each filename is the plate's barcode.

The raw data files are stored in a "Raw Data" folder inside a folder for each staining set.There is a main raw data file for each well. Staining sets with cytoplasmic data include raw data files with the word "Cyto" inplace of the word "Main".

The raw data from all wells in all plates in the dataset are read in and merged with their spot and well metadata. The number of nuclei at each spot are counted and a loess model of the spot cell count is added. Then all intensity values are normalized through dividing them by the median intensity value of the control well in the same plate. Next, the data is filtered to remove objects with a nuclear area less than 1000 pixels.

After merging the metadata with the cell-level data, several types of derived parameters are added. These include:

The origin of coordinate system is placed at the median X and Y of each spot and the local cartesian and polar coordinates are added to the dataset.

Each spot is divided into wedge-shaped bins and the wedge bin value for each cell is added to the dataset. The density around each cell is calculated from the number of nuclear centers within a radius around each nuclei. The Density value is thresholded to classify each cell as Sparse or not.The distance from the local origin is used to classify each cell as an OuterCell or not. The Sparse, OutCell and Wedge classifications are used to classify each cell as a Perimeter cell or not.

For staining set 2, each cell is classified as EdU+ or EdU-. The threshold for EdU+ is based on kmeans threshold of the mean EdU intensity from the control well of each plate.

The intensity values are normalized at each spot so that spot-level variations can be analyzed.

The cell level raw data and metadata is saved as Level 1 data. The plate and spot normalized values and the metadata is saved as Level 2 data.

The cell-level data is median summarized to the spot level and coefficients of variations on the replicates are calculated. The spot level data, CVs and metadata are saved as Level 3 data.

The spot level data is median summarized to the replicate level is stored as Level 4 data and metadata.