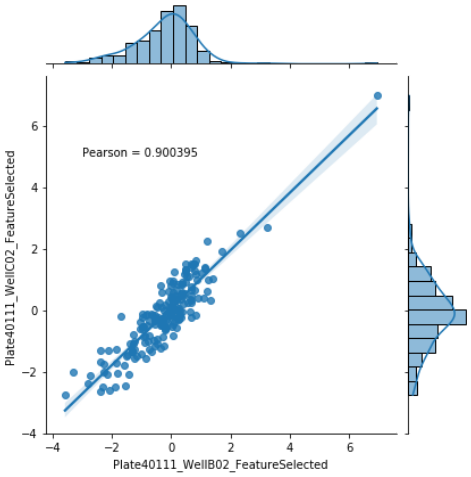
Microscopy images are powerful data

Introduction to the data set

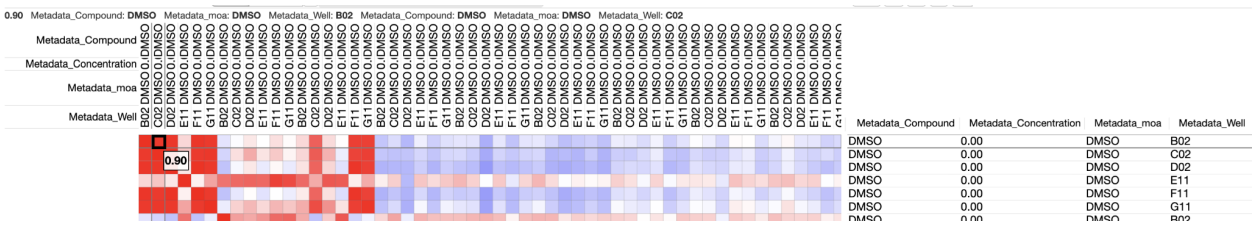
This experiment uses data from the Broad Bioimage Benchmark Collection, a collection of data sets that we host that contains sets of microscopy image with at least some sort of ground truth - the available ground truth for each image set varies but is typically either the object segmentations for each image or the relationships between various treatments.

Specifically, we'll be using the BBBC021 data set, which is a set of MCF7 cancer cells treated with compounds and stained for stains for DNA, actin, and tubulin. The entire data set contains 113 compounds at 8 concentrations, but we'll just be using a subset of compounds and concentrations at which the mechanism of action (MOA) is known plus DMSO, a negative control compound.

There are many tools and many ways to look at morphological profiling data, but this exercise will introduce you to a common one that we typically use first - examining per-well similarity matrices in Morpheus . A similarity matrix is a way to assess the co-variance in features between all pairs of wells - for each pair, one is set as X and the other is set as Y, and a Pearson correlation coefficient is calculated for all features in the data set.



The squares at the intersection of those two wells are set as the value of that correlation coefficient, and so on for each pair of wells. This allows us to see at a high level how similar the *overall phenotype* is between any pairs of wells in our experiment, and therefore how well our treatments correlate with our phenotypes.



Introduction to the exercise

The CSV and GCT files you'll need for this exercise are all located at https://github.com/broadinstitute/BBBC021_Morpheus_Exercise in the 'csvs' and 'gcts' folders. You've been provided with CSVs and GCTs for three sets of data -

- The per-well mean data with compound and MOA data added on (BBBC021_annotated)
- The previous data set, with all features Robust-Z Scored (BBBC021_normalized_mad_robustize)
- The previous data set, after performing feature selection to remove poorly behaved and/or redundant features (BBBC021_feature_selected_mad_robustize)

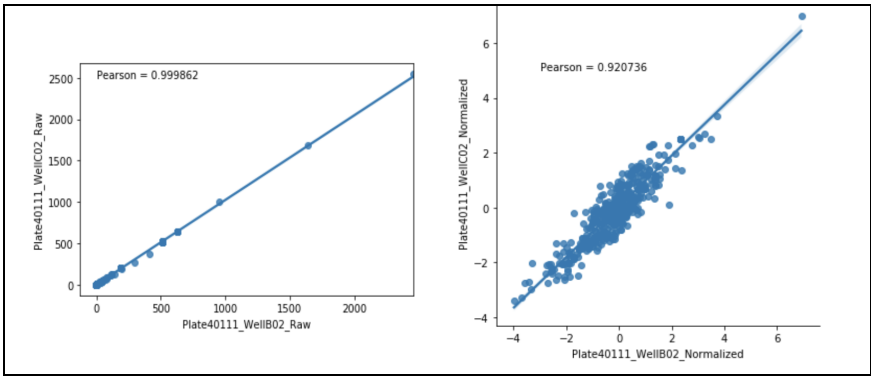
You are recommended to use the GCTs for this exercise, but CSVs are provided in case you want to explore the data further. You CAN also use the CSVs in Morpheus, but make sure to transpose them (with Tools->Transpose) before proceeding, as most other steps will assume that wells are columns and features are rows.

If you're not already familiar with Morpheus, you'll find the linked set of Google Slides helpful in finding the exact settings we use here, though you should definitely check out its written documentation and tutorials if you want to learn more- it's useful for a lot of data types, not just imaging data!

Things to try during the exercise

It's fine if you finish early or go over! The goal here is to explore. Except as indicated, no parts depend on any other parts. If you're short on time, do only section 3 (you can therefore ignore 3a)

1. Load in the un-normalized data (_annotated) and:
    a. Fix the column labels (slide step 1), unchecking "id", checking all the others, and adding colors to each label.
    b. Create a similarity matrix for the columns. (Slide step 3). How does it compare to what you expected?
    c. Do the two graphs below, of the same two wells compared either before (left) or after (right) normalization give you an idea as to why that might be?
    d. Go back to the feature values and look at the values in each row. What do you notice about the values in different rows?



2. Load in the normalized data (_normalized_mad_robustize) and:
    a. Fix the column labels (slide step 1).
    b. Check the data for unusual values (slide step 2). Note the names of any unusual values you find.
    c. Create similarity matrices and look at how they sort by MOA (slide steps 3 and 4). Does it look the way you expected?
    d. If you like, save a PNG or PDF image with File -> Save Image, OR save this exact session to reopen in Morpheus with File -> Save Session (slide step 9)
    e. Leaving the existing similarity matrix open, return to the previous tab (which has your normalized data) and filter out the weird values (slide step 6).
    f. Create a new similarity matrix on the filtered data and sort it by MOA. How does it compare to the one with the bad features included? Optionally save it if you want.
    g. Return again to the tab with the filtered normalized data and perform a similarity matrix on the *rows*. This will allow you to assess how similar the filters are to one another.
    h. Hierarchically cluster (slide step 5) the similarity matrix of the rows. What do you notice about the feature names of features that share clusters? Can you find any clusters of perfectly correlated features? Optionally save it if you want.

3. Load in the feature selected data (_feature_selected_mad_robustize) and run steps a-d and g-h from the previous section.
    a. If you examined both the normalized and feature selected data, how does the similarity matrix for columns compare to those you generated on the normalized data? The similarity matrix for rows?
    b. Hierarchically cluster the similarity matrix of the columns. Are there any MOAs that you are surprised are clustering together? That you are surprised are NOT clustering together?
    c. Return to the feature data, collapse the replicates (slide step 7), meaning you only take the median value for each feature all the set of wells that share a common compound and dose (you need to also include moa in your “collapse” settings to be able to look at it after collapse, but it will not change the outcomes). Repeat making and sorting and/or clustering a column similarity matrix. Does it differ in any significant ways from the un-collapsed data?
    d. Pick your favorite MOA or MOAs (we suggest DNA damage and DNA replication) and return to the un-collapsed feature data. Run a Marker Selection (slide step 8) to see which features are most different in this MOA(s) from DMSO. Are they what you expected or not?

Extra credit later

This exercise was based off of the cytominer-gallery github repository, specifically the "Predict compounds mechanism-of-action by morphological profiling" exercise. You can follow the instructions there to use R to start with the per-well measured features, post process them, and then quantitatively evaluate the ability to group compounds by mechanism of action.