Methods: Gene Expression Analysis

The primary data are taken from Broad's ConnectivityMap project (http://www.broadinstitute.org/cmap/). These data are processed in several steps to produce the subset of data used for analysis in the paper.

There are several files which contain the data used:

- qnorm_n855034x22268.gctx (which is in a specialized binary format for the gene expression data)
- Broad LINCS Compound Perturbagens by Plate.csv

The first file is very large (72Gb) and must be obtained from the Broad by registering with them.

The primary data are processed in five steps (scripts) in order to manage memory to be under 2Gb (necessary for python 2.7 limitations).

1. step1_setup_lincs_analysis.py
   a. This file identifies those cell lines and perturbagens in the original data which are a complete submatrix, and records these in a file called 'kosher_redone.csv'
2. step2_setup_lincs_analysis.py
   a. This file computes the mean and standard deviation of gene expression per cell line (i.e. marginalized out perturbations), using 'kosher_redone.csv' to identify the relevant data and records these in 'rawdata_mean_redone.npy' and 'rawdata_std_redone.npy' which are Numpy (numerical python) binary data formatted.
3. step3_setup_lincs_analysis.py and step4_setup_lincs_analysis.py
   a. This file uses the files produced above to perform a z-score and compute mean cell expression and treatment expressions, respectively, which are saved in 'bucket_cell_means_redone.npy' and 'bucket_treatment_means_redone.npy'
4. step4_setup_lincs_analysis.py
   a. Performs a z-score transformation using the results from step 3.
5. step5_setup_lincs_analysis.py
   a. This file uses 'kosher_redone.csv' and the 'bucket_*' files to identify the 50 genes described in the paper and form a data file 'thedata.npy' which is arranged as a 280x50 (treatments x genes), 48 cell line matrix. Treatments and gene pairs are such that the first entry is (treatment 1, gene 1), the second entry is (treatment 1, gene 2) and so on (genes cycle through per treatment).