Machine Learning method

We used a classification random forest model to predict putative tumor suppressors on chromosome 7, based on publicly available genome-wide screening data in human cancer cell lines. For our training data, we compiled a list of genome wide-screening data, and retained experiments which are suited for detecting tumor suppressor genes using the criteria below: 1) only retained the screens using AML and CML cell lines, in order to be disease specific. 2) filtered out two screens with too many 0 values (94%) and missing values (66%).

This led to 20 screening data with 17 unique cell lines. We then used k-nearest-neighbor algorithm to impute the remaining missing values. For the “ground truth” column used in training our model, we used annotation from Cancer Gene Census, which labeled 315 canonical tumor suppressors genome-wide. We then split the training and testing data. Our testing data are all the protein-coding genes on chromosome 7, and the training data are all the protein-coding genes on all other chromosomes. To reduce bias in training data and achieve training-testing data balance, we did 1000 times of bootstrapping on training data to randomly sample from the non-chromosome 7 genes. For each bootstrap, we performed hyperparameter tuning of the random forest model and select the combination of hyper parameters that gives the smallest out-of-bag (OOB) error. We then performed prediction on the testing data using the 1000 tuned models and obtained a binary result for each gene (1 for putative tumor suppressor and 0 for non-tumor suppressor). We then ranked the chromosome 7 genes based on the frequency of being labeled as tumor suppressors and obtained our result list. We achieved on average ~64.17% of accuracy.

**Combined Ery/Prolif method**

For erythrocyte differentiation and proliferation scores, we normalized each gene’s readouts in each well by subtracting from AAVS1’s readouts. Erythroid signs were inverted so that a higher score is associated with increased proliferation and decreased erythroid differentiation. Then we obtained average erythrocyte differentiation and proliferation score by taking the average of each gene’s readout across all replicate plates. We then combined the average erythrocyte differentiation and proliferation scores into one table and performed min-max normalization onto 0-1 range, in order to get rid of the effect of directionality and unify the two scores onto the same scale.

To test whether each gene’s proliferation and erythrocyte differentiation readouts are significantly different from those of the AAVS1 control, we performed non-parametric Mann-Whitney test of each gene’s readout across all replicate wells comparing to AAVS1’s readout. The p values were corrected for multiple hypothesis using FDR.(FDR < 0.12)