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 CRISPRa and overexpression experiments which are suited for detecting tumor suppressor genes. Subsequently, we applied the following filters to fine-tune our training data for the machine learning model: 1) only retained the screens related to hematological malignancies including AML, CML and Burkitt’s Lymphoma. 2) filtered out two screens with too many 0 values (94%) and missing values (66%).

This led to screening data from 8 experiments and 26 cell lines. Furthermore, we added mutational signature data from *Davoli.2013* study which were shown to have the best performance in predicting tumor suppressors using a LASSO regression model. 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. For the training data we did 100 times of bootstrapping to randomly sample from the non-chromosome 7 genes, in order to achieve training-testing data balance, and to remove the effect of randomness. 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 100 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 ~65% of accuracy.