**Pipeline Scripts**

After downloading UQ-FPKM file for patients, first must organize them into directory hierarchy: res / non -> cancer type (The first script will use this path as header in output)

1) This script normalizes a patient’s gene expression using TCGA population (-mean/sd) for their cancer type. Must be run using *Rscript,* in the parent directory of /res/.

**norm\_file.R**

**Input:** *cancer*\_norm, */res/cancer/*UQ-FPKM.gz

**Output:** *filename*\_norm.tsv

2) This script combines a directory of patients (named /all/) into one file (rows are genes, cols are *res/cancer/fileID*). Must be run in the parent directory of /all/.

**join\_models.sh**

**Input:** *group\_name*

**Output:** combined\_*group\_name*

3) This script filters the combined files by removing rows (genes) that have < 25% variance (across population from many cancers). Subset top 5000 genes based on Clayton-Pujol method (Unique vs SD).

**filter\_combined\_gene\_exp.R**

**Input:** *drug\_*drop\_cols, combined\_*group\_name*

**Output:** *group\_name*\_Filtered.tsv

4) Clustering

#Run optcluster

**optCluster\_fix.R**

#Save cluster assignments

/Users/eclayton3/mlb\_final/optcluster/**save\_clara\_clusters.R**

* Models:
* Internal Validation
* Number of clusters: 180-220

5) This script calculates cluster means for each patient by averaging expression values for all genes in a cluster

**calculateClusterMeans.R**

**Input:** combined\_*group\_name*\_log\_filtered\_7500.tsv *group\_name\_clusterMethod*\_cluster\_assignments.txt ...

**Output:** *group\_name\_clusterMethod*\_cluster\_means.txt ...

6) This script runs random forest feature selection to rank variables, stops once mean of ranks for all 200 clusters does not change 3x

**rf\_feature\_selection\_converge.R**

**Input:** *group\_name\_clusterMethod*\_cluster\_means.txt

**Output:** converge\_top200\_*group\_name*\_*clusterMethod*\_cluster\_means.txt

7) This script runs random forest cross validation to plot accuracy vs number features for determining optimal set

**rf\_optNumber\_vars\_final.R**

**Input:** *group\_name\_clusterMethod*\_cluster\_means.txt

converge\_top200\_*group\_name*\_*clusterMethod*\_cluster\_means.txt

**Output:** accuracy and plot pdf

8) This script runs random forest on validation data set using model built on all training data

**rf\_*drug*\_external\_validation\_030218.R**

Input: Training data matrix, converge file containing variable ranking, validation data matrix

Output: accuracy and contingency table

9) This script produces rf mean predicted probabilities for use in ROC curves and violin plots

**rf\_*drug*\_roc\_probabilities\_030218.R**

**Input:** Training data matrix, converge file containing variable ranking, validation data matrix

**Output**: *drug*\_probabilities.tsv and *drug*\_val\_probabilities.tsv