PRISM Biomarker Analysis Overview

Introduction

There are three primary analyses performed in the standard MTS pipeline:

1. *t*-tests of differences between discrete features (lineage and mutation).
2. Correlations between sensitivity and continuous features (gene expression, CRISPR dependency etc.).
3. Multivariate models predicting sensitivity based on multi-omics feature sets.

Discrete analysis (*t* tests)

Our discrete analyses test the relationship between discrete features, such as mutation and lineage, and sensitivity to a compound. These features are called discrete because they can only take on a finite number of values (either a cell line has a mutation or does not). To determine these discrete relationships we do a *t* test comparing cell lines that have a certain feature (e.g. a BRAF mutation) and those that do not.
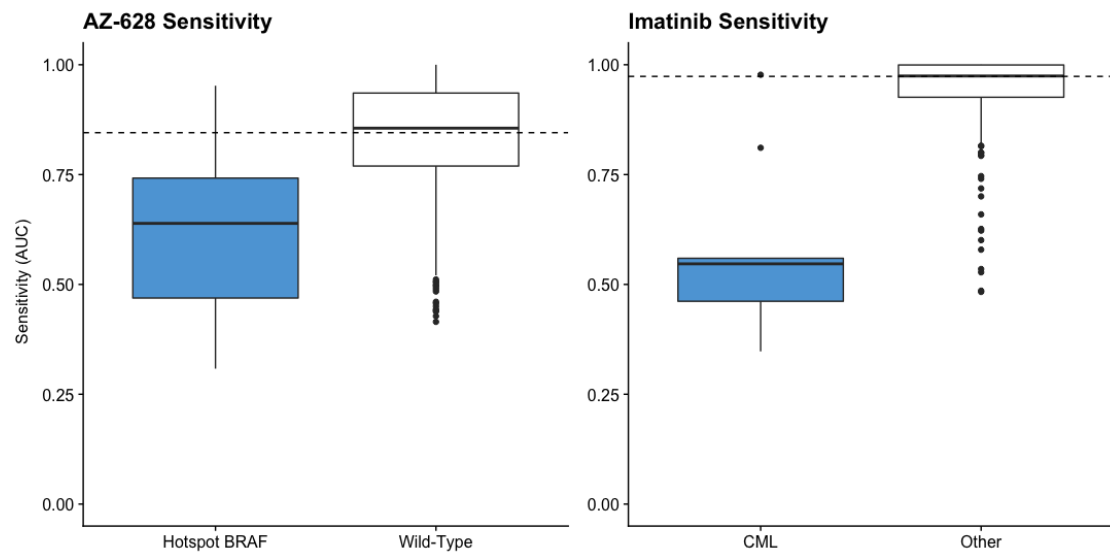

Figure 1: AZ-628 mutation and imatinib lineage

This generates an effect size, measuring the difference in the mean senstivity between the two groups, and a p value, measuring the significance of that difference. We then use the Benjamini Hochberg algorithm to correct the p values to q values, to account for false discovery rate.

Continuous analysis (correlations)

Our continuous analyses test the relationship between continuous features, such as gene expression, and sensitivity to a compound. To determine these relationships we calculate the Pearson correlation between a feature and sensitivity. We also calculate the p values associated with the correlations and correct them to q values.
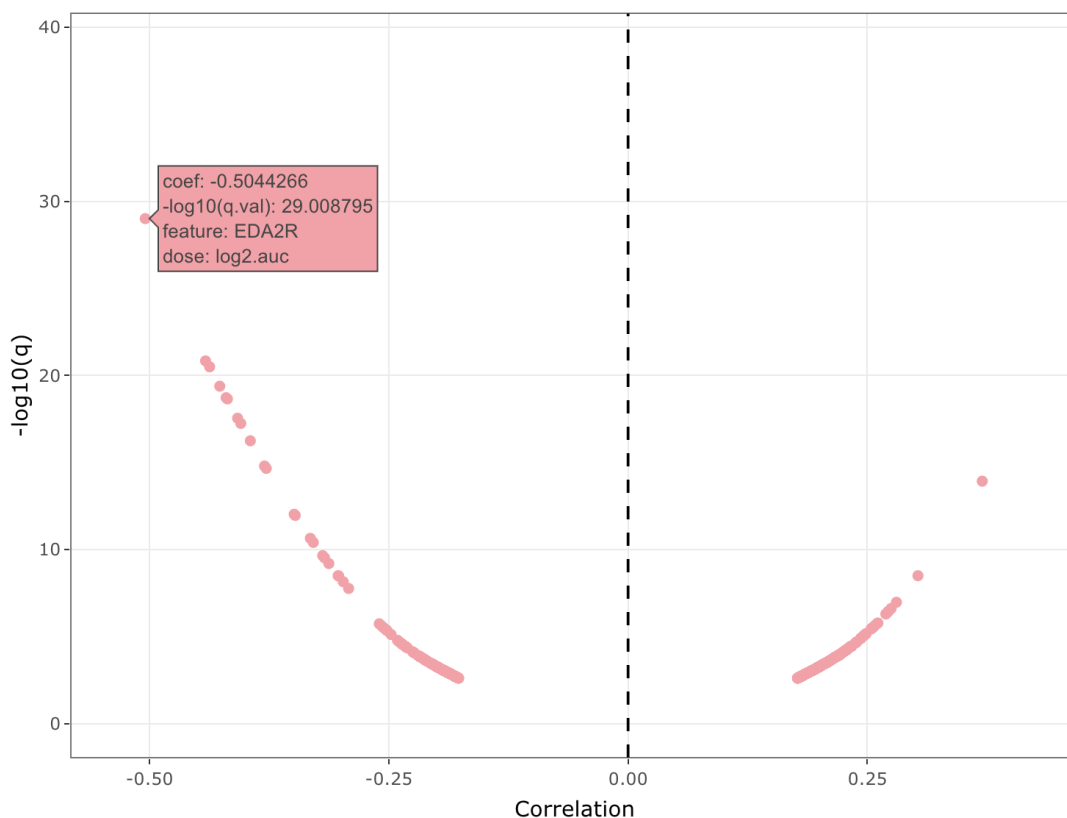

Figure 2: Nutlin-3 gene expression correlation

Shown above is the correlations between **Nutlin-3** area under the curve (AUC) and gene expression. Each point represents a gene. The highlighted point is one of the biomarkers for Nutlin-3a and shows as the top correlated. The negative correlation indicates that increased gene expression tends to coincide with decreased AUC (meaning more sensitivity).

Multivariate models

Our multivariate models use a combination of -omics datasets to generate predictions about compound sensitivity. In particular we use a cross-validated random forest model

to predict sensitivity based on -omics features. We then compare the predictions of our model with actual results to determine model accuracy (reported as R^2 and Pearson Score). We can also extract the estimated importance of each feature of the model in generated the predictions in order to pick out potential biomarkers. Features that are important in well-performing models are said to be of interest.
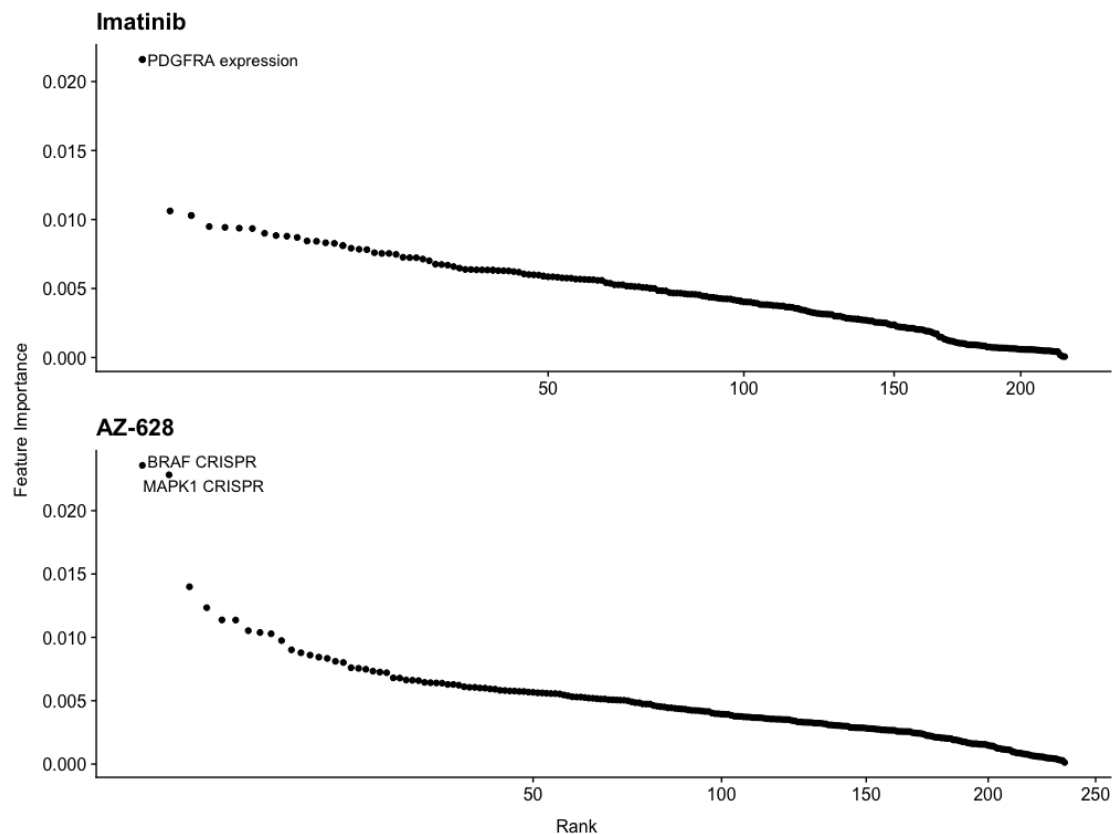

Figure 3: Random forest feature importances for AZ-628 and imatinib

Shown abovea are the importances of the top 250 features in models predicting imatinib and AZ-628 sensitivity. In both cases, targets of the compounds show up as the most important. It is important to not that feature importance is a directionless measure, so other analyses must be compared to determine whether a feature is associated with increased or decreased sensitivity.