Explaining the Sensitivity of DepMap Cell Lines to CRISPR Knockdown

Oliver Vipond, Danny Miller

ovipond@evaristetechnologies.com, dmiller@evaristetechnologies.com

EVARISTE TECHNOLOGIES

1. Abstract

The Cancer Dependency Map (DepMap) portal provides a wealth of cell line data enabling data-driven research into cancer vulnerabilities. In this work we develop an analysis pipeline to identify biomarkers which are predictive of a cell line being sensitive to the CRISPR knockdown of a candidate gene.

The pipeline has been developed to be sufficiently complex as to identify non-linear multivariate dependencies, whilst balancing this complexity with computational efficiency such that analyses can be performed in real time by end-users.

We illustrate the output of such an analyses for the selective gene epidermal growth factor receptor.

2. Work Flow

Identify a selective gene and the biomarker types of interest

↓

Annotate cell lines as Sensitive or Insensitive to knockout of the gene of interest

↓

Reduce set of $\sim 30,000$ biomarkers to predictive subset typically ~ 100

↓

Assemble interpretable output including: decision trees to stratify sensitive cell lines, and implicated pathways for sensitivity

↓

Plot decision boundaries to identify conditional importance of biomarkers

3. Implementation Details

Selective Gene Identification

We identify genes with heavily negatively-skewed distributions of CRISPR effect using Groeneveld and Meeden's coefficient: $\textsc{Skew}_{\text{GM}}(X) = \frac{\mu - \nu}{\mathbb{E}[|X - \nu|]}$ (where μ, ν denote the mean and median of X respectively).

Data Set Assembly

For the set of cell lines classified as either Sensitive or Insensitive we load descriptors for mRNA expression, gene mutation, cell line lineage, age and sex.

Descriptor Selection

We use the Boruta algorithm [KR10] with a shallow Random Forest classifier to determine which features are predictive of sensitivity. The shallowness of the decision trees reduces the computational burden and ensures that the predictor-sensitivity relationships are interpretable.

Pathway Enrichment

With the software XGR [FKBK16] we pass the genes identified in the descriptor selection step to the function `xEnricherGenes`, to identify enriched pathways from a variety of ontologies.

Model Performance

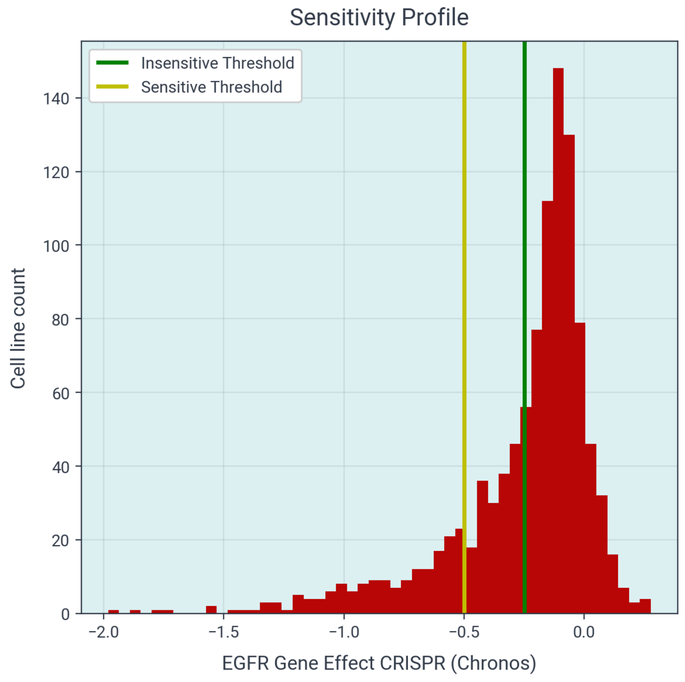
K-fold stratified cross-validation (where K is taken such that at least 10 samples of each class are present in each fold) is employed to estimate the precision and recall of cell-line sensitivity to the queried gene given the identified biomarkers.

Decision Boundary Visualisation

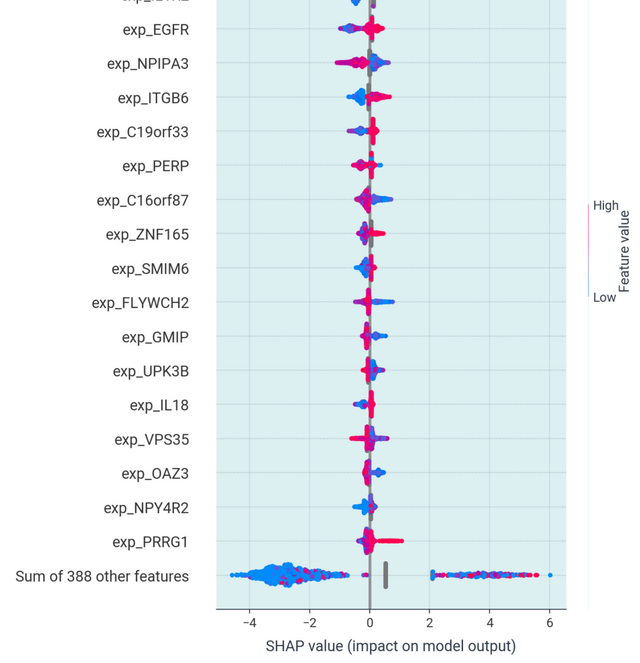
A shallow Gradient Boosting Classifier model is fit on the reduced data set of selected descriptors. The decision boundary of this classifier for pairs of predictive descriptors are plotted.

4. Example

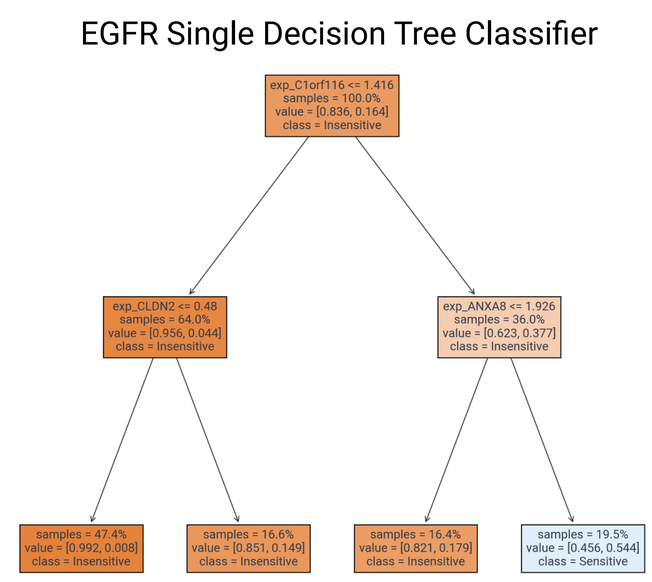
Epidermal growth factor receptor (EGFR) appears as the sixth most selective gene in the DepMap portal (as ranked by the Groeneveld-Meeden coefficient). To perform the analysis a user is only required to specify CRISPR gene effect thresholds in order to annotate the cell lines, and the full analysis completes in less than a couple of minutes.



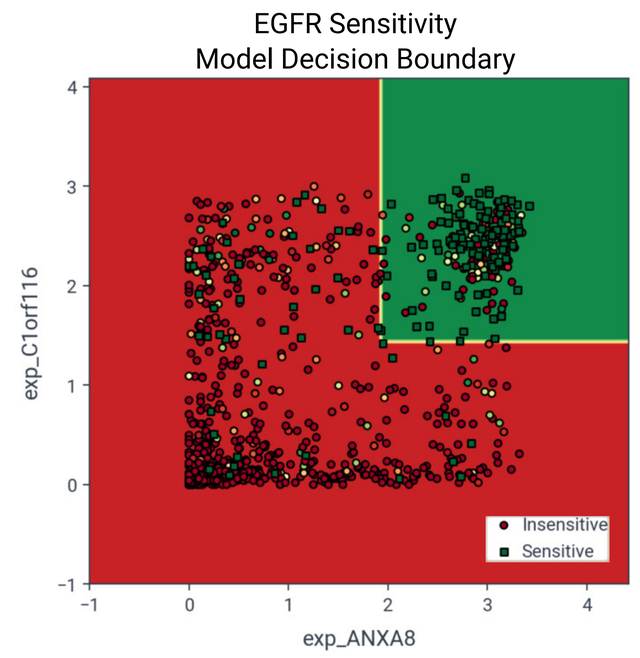
(a) Histogram of CRISPR effects on DepMap cell lines. 173 cell lines with gene effect less than -0.5 are labelled sensitive, and 687 cell lines with gene effect greater than -0.25 are labelled insensitive.



(b) 408 descriptors are selected by Boruta as *predictive* or *weakly predictive* of cell line sensitivity to EGFR knockdown. We plot Shapley values of the most predictive of these descriptors.



(c) The single decision tree most predictive of EGFR sensitivity. Each node displays: a condition (samples passing the condition fall to the left child), the percentage of samples surviving to that node, the proportion of samples of each class in that node (insensitive, sensitive), and the majority class.



(d) Decision boundary of the most predictive descriptors identified by Shapley values: expression of C1orf116 and expression of ANXA8. Cell line scatter points are coloured by cell line sensitivity. 54.4% of cell lines with high expression of ANXA8 and C1orf116 are sensitive to knockdown of EGFR.

EGFR Sensitivity Model Performance

	precision	recall	f1-score	support
Insensitive	0.8944	0.9614	0.9267	881
Sensitive	0.6822	0.4220	0.5214	173
accuracy	0.8729	0.8729	0.8729	0.8729
macro avg	0.7883	0.6917	0.7241	1054
weighted avg	0.8596	0.8729	0.8602	1054

(e) Stratified cross-validation performance of a shallow Gradient Boosting Classifier model predicting EGFR sensitivity.

References

[FKBK16] Hai Fang, Bogdan Knezevic, Katie L. Burnham, and Julian C. Knight. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Medicine*, 8(1), December 2016.

[KR10] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.