Logistic Regression Approach to Identifying Crohn’s Disease from Exome Data in CAGI2015

To construct a suitable feature set for classification, an initial assessment of each variant sequence element was carried out against the CAGI2013 data. The ability of each variant element to distinguish healthy control from Crohn’s patient exomes was evaluated by calculating log-odds scores. In this case only negative log-odds scoring features were used i.e. features which appear to anti-correlate with Crohn’s. A simple logistic regression model was used as the classifier in this case.

For onset prediction, a simple k-NN classifier was used (k=2).

Data handling was done using Perl scripts and machine learning using Python/scikit-learn.