Meta-method for Identifying Crohn’s Disease from Exome Data in CAGI2015

This submission uses a crude boosting method to combine predictions from M. Giollo and D. Jones. In total 9 predictions were combined to make the Crohn’s predictions, and probabilities are based on the fraction of predictions for which each particular sample is in the top 2/3 of the ranked list of predictions.

To construct a suitable feature set for onset prediction, an initial assessment of each variant sequence element was carried out against the CAGI2013 data. The ability of each variant element to predict onset was evaluated using a point biserial test, giving 57 SNP features (p-value threshold of 5x10-6 ). as applied, giving a total of 102 features for classification.

Onset age was predicted using a linear model trained with Lasso regression.

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