Supplementary Materials

Predicting the functional effects of human non-coding variants based on stacking ensemble learning

This file includes:

Figure S1 to S2

Supplementary Figures

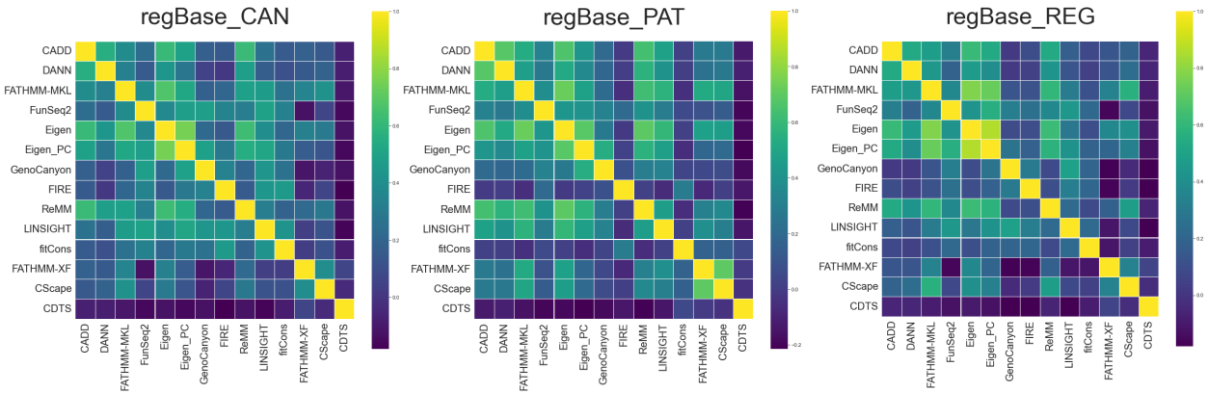

Figure S1. Spearman correlation coefficient between 14 annotation scores derived from the regBase_REG, regBase_PAT and regBase_CAN training datasets.

We investigated correlations between annotation scores using Spearman's correlation coefficient on the three training sets to observe consistency throughout different annotation scores, as well as hierarchical clustering to observe potential relationships between the scores. A small number of annotation scores, such as FATHMM-MKL and EIGEN, Eigen PC and Eigen, showed pairwise correlations (R>0.7), and there were also significant correlations between Eigen and CADD, ReMM and Eigen, ReMM and Eigen PC, etc.Most scores, however, still had weak correlations(R<0.4)

.

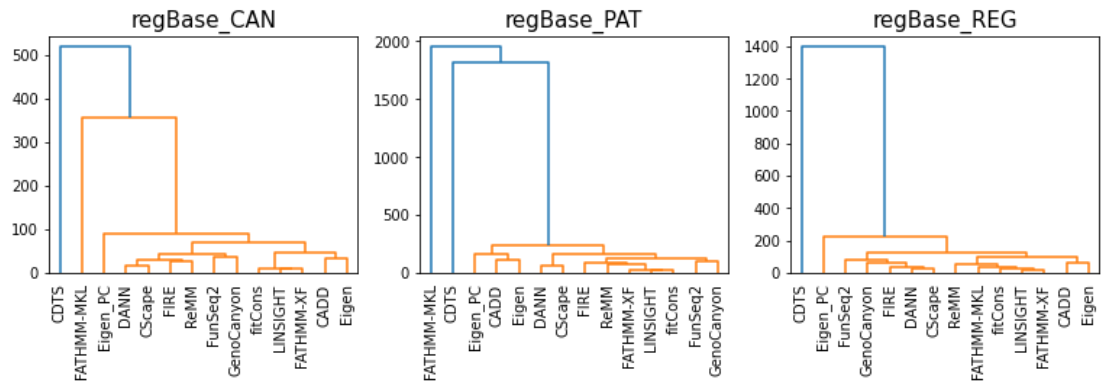

Figure S2. Hierarchical clustering between annotation scores on the regBase_REG, regBase_PAT and regBase_CAN training datasets.

We found that these scoring methods can be roughly classified into two groups: the methods in the first group are practically unrelated to others, but the methods in the second group exhibit clear correlations. This demonstrates that several of the annotation scores we employed are capable of capturing unique features that are useful for classification on three distinct types of datasets, especially CDTS and FATHMM-MKL. This may be due to the fact that CDTS focuses more on finding important genes among restricted genes, while FATHMM-MKL identifies more predictors by learning functional annotations and nucleotide-based sequence conservation measures. Concurrently, we noticed that certain approaches, like CADD and Eigen, tend to group together when clustering. This could be because Eigen primarily employs a fraction of the annotations that are available in CADD. DANN and CScape seem to be clustered together quite frequently; this could be due to their shared concern in evolutionary conservation.