A model averaging approach for improved *in silico* variant prediction

Yuan Tian, Shuwei Li, Mary Helen Black, Aaron Elliott, Hsiao-Mei Lu, Dajun Qian

A variety of *in silico* scores are currently utilized in missense variant assessment as evidence of pathogenicity. A Bayesian Integrated Evaluation method was developed to leverage the information from these scores, in which *in silico* predictors are initially selected manually or by stepwise regression. However, the models produced often contain uninformative predictors and do not quantify their gene-specific importance. A robust and stable gene-specific model is needed to calibrate information from *in silico* tools. Using data from 129,691 patients who underwent genetic testing at a single diagnostic laboratory in 2012-2016, we assembled a collection of 1,292 missense variants in 8 genes (*BRCA1*, *BRCA2*, *CDH1*, *TP53*, *MLH1*, *MSH2*, *MSH6* and *PMS2*), each classified according to a five-tier classification system per ACMG guidelines. Nine *in silico* scores (Polyphen2, SIFT, Grantham, CADD, AGVGD, MutationAssessor, phastCons, phyloP and REVEL) were collected for all variants. Gene-specific Model Averaging Logistic Regression (GMALR) models were trained on known classified variants using the nine *in silico* scores as predictors. We assessed each predictor using relative variable importance (RVI) defined as the sum of Akaike weights over all candidate models (range 0 to 1). The predicted rate of variants of unknown significance (VUS), positive predictive value (PPV), negative predictive value (NPV), and coefficient mean squared error (MSE) with 10-fold cross validation, were compared for GMALR vs. Gene-specific Stepwise Logistic Regression (GSLR) models. GMALR results indicated that the relative importance of each *in silico* predictor varied by gene. For example, AGVGD, CADD, phyloP and REVEL had the highest RVI in *BRCA1*, while MutationAssessor, phastCons and phyloP had the highest RVI in *BRCA2*. Predicted VUS rates were 39.5% and 45.2% for GMALR and GSLR models respectively; PPV and NPV were ~99% for both models. Thus, GMALR reduced the predicted VUS rate and had similarly low misclassification rates compared to GSLR. GMALR models also improved model robustness by a meaningful margin compared to GSLR models; mean ± SD of MSE across all genes was 1.7 ± 1.2 and 5.0 ± 4.9 for GMALR and GSLR models, respectively. Our results underscore the improved prediction performance and model robustness of the GMALR approach when evaluated in the context of *in silico* variant prediction and as compared to stepwise regression methods such as GSLR.

Characters with spaces=2,457