Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies

Chengliang Dong[1,2,*], Peng Wei[4,6,*], Xueqiu Jian[5], Richard Gibbs[7], Eric Boerwinkle[4,5,7], Kai Wang[1,2,3] and Xiaoming Liu[4,5]

[1]Zilkha Neurogenetic Institute, [2]Biostatistics Division, Department of Preventive Medicine, [3]Department of Psychiatry, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [4]Human Genetics Center, [5]Division of Epidemiology, Human Genetics and Environmental Sciences and [6]Division of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, 77030, USA; [7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, 77030, USA.

Correspond to:

Xiaoming Liu, Human Genetics Center, University of Texas at Houston, PO Box 20186, Houston, Texas 77225; E-mail: Xiaoming.Liu@uth.tmc.edu; Fax: (713)500-0900
And
Kai Wang, 1501 San Pablo Street, ZNI 221, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089; E-mail: Kaiwang@usc.edu; Fax: (323)442-2145

*Contributed equally. The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

Abstract

Accurate deleteriousness prediction for nonsynonymous variants is crucial for distinguishing pathogenic mutations from background polymorphisms in whole exome sequencing (WES) studies. Although many deleteriousness prediction methods have been developed, their prediction results are sometimes inconsistent with each other and their relative merits are still unclear in practical applications. To address these issues, we comprehensively evaluated the predictive performance of eighteen current deleteriousness-scoring methods, including eleven function prediction scores (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, PANTHER, PhD-SNP, SNAP, SNPs&GO, and MutPred), three conservation scores (GERP++, SiPhy and PhyloP) and four ensemble scores (CADD, PON-P, KGGSeq and CONDEL). We found that FATHMM and KGGSeq had the highest discriminative power among independent scores and ensemble scores, respectively. Moreover, to ensure unbiased performance evaluation of these prediction scores, we manually collected three distinct testing datasets, on which no current prediction scores were tuned. In addition, we developed two new ensemble scores that integrate nine independent scores and allele frequency. Our scores achieved the highest discriminative power compared with all the deleteriousness prediction scores tested and showed low false positive prediction rate for benign yet rare nonsynonymous variants, which demonstrated the value of combining information from multiple orthologous approaches. Finally, to facilitate variant prioritization in WES studies, we have pre-computed our ensemble scores for 87,347,044 possible variants in the whole-exome and made them publicly available through the ANNOVAR software and the dbNSFP database.

Introduction

One of the greatest challenges in whole exome sequencing (WES) studies is the ability to distinguish pathogenic mutations from a large number of background variations. A common strategy is to filter for novel nonsynonymous single nucleotide variants (nsSNVs) that are found in patients and computationally predicted to be deleterious, which relies on the accuracy of the prediction methods (1). Multiple algorithms were developed for predicting such deleteriousness based on different information of the variant, such as its sequence homology (2), protein structure (3, 4) and evolutionary conservation (5). These methods demonstrated themselves to be useful, however, they are not directly comparable with each other due to the difference in information and algorithms they used for predicting deleteriousness. Therefore, it is still unclear which one(s) to use for prioritizing nsSNVs in WES-based studies of human diseases to minimize both false positive and false negative prediction rates. While a few comparison studies have been done (6-10), the range of methods under comparison and/or the size of benchmark dataset often limited their scope and generalizability for WES studies.

One of the keys to a fair comparison lies in unbiased testing datasets. To ensure such fair comparison, we manually curated three distinct testing datasets, on which no prediction methods compared were tuned

(Table 1). On these three testing datasets, we performed a comprehensive comparative study of eighteen deleteriousness prediction scoring methods, including eleven function prediction scores (PolyPhen-2 (v2.2.2, released in Feb, 2013) (11), SIFT (Human_db_37_ensembl_63, released in August, 2011) (12), MutationTaster (data retrieved in 2013) (8), Mutation Assessor (release 2) (2), FATHMM (v2.3) (13), LRT (November, 2009) (6), PANTHER (version 6.1) (14), PhD-SNP (version 2.0.6) (15), SNAP (version 1.0.8)(16), SNPs&GO (17), and MutPred (version 1.0) (18)), four ensemble scores (CADD (19), PON-P (20), KGGSeq (version 0.2) (21) and CONDEL (22)), and three conservation scores (GERP++ (23), SiPhy (24, 25) and PhyloP (phyloP46way_placental) (26, 27)) (Table 2) on their performance to prioritize nsSNVs in WES studies. Here function prediction scores refer to scores that predict the likelihood of a given nsSNV causing deleterious functional change of the protein; conservation scores refer to scores that measure the conservativeness of a given nucleotide site across multiple species; and ensemble scores refer to scores that combine information of multiple component scores. To facilitate more accurate variant prediction, we also developed and evaluated two ensemble-based approaches, Support Vector Machine (SVM) (28) and Logistic Regression (LR) (29), that integrate multiple scoring methods for which whole-exome data are available.

Results

Curation of training and testing data sets

To compare the performance of the prediction methods, we constructed four datasets; one for training our SVM and LR model and three for testing their performance against all the deleteriousness prediction methods. Quality of machine learning models, such as SVM and LR, can be influenced by selection of component scores as well as the selection of parameters. To optimize the selection of component scores and parameters for our SVM and LR model, we collected training dataset, on which we performed feature selection and parameter tuning for our models. Training dataset is composed of 14,191 deleterious mutations as True Positive (TP) observations and 22,001 neutral mutations as True Negative (TN) observations, all based on the Uniprot database (30, 31). Note that the TN observations contain both common (maximum Minor Allele Frequency (MMAF)>1% in diverse populations of the 1000 Genomes project (32)) and rare (MMAF <=1%) variants to ensure the generalizability of our model. To reduce potential bias in our

comparison, we manually collected testing datasets from three distinct contexts. Testing dataset I consists of 120 TP observations that are deleterious variants recently reported to cause Mendelian diseases, diseases caused by single-gene defects, with experimental support in 57 recent publications (after January 1st, 2011) from the journal *Nature Genetics* and 124 TN observations that are common neutral variants (MMAF >1%) newly discovered from participants free of Mendelian disease from the Atherosclerosis Risk in Communities Study (ARIC) study[32] via the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) sequencing project (33, 34). Testing dataset II is derived from the VariBench (35) dataset II, a benchmark dataset used for performance evaluation of nsSNV scoring methods. Because VariBench dataset II contains mutations that overlap our training dataset, we removed these mutations and curated 6,279 deleterious variants as our TP observations and 13,240 common neutral variants (MMAF >1%) as our TN observations. As rare variants may have different features compared with common variants in the populations (36), a prediction method that puts too much weight on those features may not perform well in separating rare deleterious variants from rare neutral variants. In order to evaluate the performance of prediction methods on rare variants, we also prepared testing dataset III that contains 10,164 rare neutral mutations (singletons) from 824 European Americans from the cohort random samples of the ARIC study[32] via the CHARGE sequencing project (33, 34)). Note that as testing dataset I and III were collected from recently available projects and publications and that testing dataset II was used solely for benchmark in performance comparison studies, they are unlikely to be used for tuning current prediction methods evaluated in our study. Even though prediction methods, such as FATHMM, applied Varibench datasets (from which our testing dataset II was derived), they only used the data for performance evaluation purpose, not for model tuning (10). Therefore, to our knowledge, our testing datasets are unbiased benchmark datasets to ensure a fair comparison between different prediction methods. A summary of the datasets is presented in Table 1, and the complete datasets are available in Table S8.

Prediction scores

We obtained nine deleteriousness prediction scores, including six function prediction scores (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM and LRT) and three conservation scores (GERP++, SiPhy and PhyloP), for all variants in all datasets. Additionally, we obtained nine more deleteriousness

prediction scores (PANTHER, PhD-SNP, SNAP, SNPs&GO, MutPred, PON-P, KGGSeq, CONDEL and CADD) for testing dataset I and three more scores for testing dataset II and III (KGGSeq, CONDEL and CADD) (Table 2). Note that we were not able to harvest all eighteen deleteriousness prediction scores for all four datasets, mostly due to the accessibility of most of scores from web interface and the speed of obtaining these scores for relatively large datasets. Nevertheless, we still successfully obtained nine deleteriousness prediction scores for all datasets that we used, which is already the largest in scale for similar comparison studies.

To construct our own ensemble scores, we used these nine scores (SIFT, PolyPhen-2, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy and PhyloP) for all potential human nsSNVs collected in the dbNSFP database, along with the MMAF observed in diverse populations of the 1000 Genomes project (32), as our component scores. After imputing missing values for observations in training dataset, we employed SVM and LR algorithms for computing these two ensemble scores and evaluated three kernel functions (linear, radial and polynomial kernel) for the SVM approach, all based on default parameters (Table 3).

From univariate analysis of nine deleteriousness prediction scores and MMAF of training dataset, we found that all of the ten predictors showed significant difference in median scores between TP groups and TN groups, suggesting that all ten predictors work well in separating TP observation from TN observations ($P<0.01$, Wilcox rank sum test with Bonferroni adjustment) (Table S1, Fig. S1). Moreover, all deleteriousness prediction scores showed positive association with deleteriousness while MMAF showed negative association with deleteriousness (Table S1). This is consistent with previous findings. Higher prediction score from deleteriousness prediction tools indicates higher risk of deleteriousness, while higher MMAF means that the mutation is common and therefore is oftentimes less likely to be deleterious.

We next evaluated the extent of collinearity in the training data, because high levels of collinearity may affect the stability of prediction models and inflate variance of parameter estimation. From the results, only PolyPhen-2 HVAR and PolyPhen-2 HDIV, two scores from PolyPhen-2 trained on different datasets, raised our concern, because they have the highest collinearity and therefore may affect the stability of our prediction models (Pearson Correlation Coefficient=0.946). To further examine their effect on model stability, we applied stepwise model selection. Results suggested that we include both PolyPhen-2 HVAR and HDIV into

final prediction models. As for LR model, reduced model generated from the model selection procedure included both scores (Table S6). As for SVM model, to increase stability of our final models, we applied regularization on linear, radial and polynomial SVM models. From five-fold cross validation results, we chose parameter cost=1 to penalize L2 norm of weights of all predictors to ensure robustness of our final models (Fig. S5).

Comparison of quantitative predictions

We compared the quantitative predictions of the deleteriousness prediction methods from five different aspects. First, to estimate the relative contribution of each prediction score in our SVM and LR model, we applied multiple LR on these nine prediction scores and MMAF. Results showed that when evaluated by the estimated coefficients in our LR model, all deleteriousness prediction scores (except for SIFT and MutationTaster), conservation scores and MMAF were statistically significant predictors for estimating deleteriousness of the mutation in our training dataset at $\alpha = 0.05$ significance level, after adjusting for remaining scores in the multiple LR model. This suggests that although most of the function prediction scores have already taken certain conservation measures into account, independent conservation scores can still bring additional information to the ensemble score. Among all significant deleteriousness prediction scores, FATHMM contributed the most to estimating deleteriousness. The adjusted odds of a mutation being deleterious is 4.6 times higher than being neutral for 0.1 increase of FATHMM score when remaining scores in the multiple LR model remain constant ($P < 0.001$, Wald-test) (Table S2).

Second, we compared the predictive performance of individual deleteriousness prediction scores altogether, including function prediction scores and conservation scores, using Area Under the Curve (AUC) score from Receiver Operating Characteristic (ROC) plots and True Negative Rate (TNR, or specificity) as measurements. We found that FATHMM performed the best in all three testing datasets, as expected. The probability of a TP observation having a higher predicted deleterious score than a TN observation is the highest for FATHMM. In testing dataset I, the probability of a TP observation having a higher predicted FATHMM score than a TN observation is 0.87 (AUC = 0.87, 95% confidence interval (CI): 0.82-0.92), which was statistically significantly better than the worst performing prediction score, PANTHER (AUC = 0.65, 95% CI: 0.58-0.72). FATHMM also achieved the highest performance in testing dataset II (AUC = 0.91 for FATHMM,

95% CI: 0.9-0.91) (Fig. 1). In testing dataset III that contains only rare neutral variants, normalized FATHMM median score is the second lowest, indicating a relatively high tendency of correctly classifying TN variants. Note that even though PolyPhen-2 HVAR has a lower normalized median score, it did not achieve as high specificity as FATHMM (shown below). Therefore, FATHMM demonstrated itself to be consistently the best performing individual deleteriousness prediction method for separating deleterious variants from neutral variants in its quantitative predictions.

Third, within all individual deleteriousness prediction methods, we compared function prediction methods with conservation methods. We found that function prediction methods tend to perform better than conservation scores, especially in separating common neutral mutations from deleterious mutations. From ROC plots that grouped function prediction scores and conservation scores, we found that the average probability of a TP variant having a higher predicted deleterious score than a TN variant is statistically significantly higher for function prediction scores than for conservation scores. Indeed, the average probability of a TP variant having a higher predicted deleterious score than a TN variant is 0.73 for function prediction scores (AUC = 0.73, 95% CI: 0.72-0.75) in testing dataset I and is 0.71 (AUC = 0.71, 95% CI: 0.7-0.71) in testing dataset II, while for conservation prediction scores the probabilities are only 0.66 (AUC = 0.66, 95% CI: 0.62-0.71) and 0.6 (AUC = 0.6, 95% CI: 0.6-0.61) in these two datasets (Fig. S3).

Fourth, we compared existing ensemble methods, both with themselves and with their individual component scores. We found that KGGSeq performed the best and it was the only existing ensemble scores that outperformed all of its component scores. The probability of a TP variant having a higher KGGSeq score than a TN variant is 0.85 in testing dataset I (AUC = 0.85, 95% CI: 0.81-0.90) and is 0.89 in testing dataset II (AUC = 0.89, 95% CI: 0.89-0.90)**.** Not only did it achieve the highest performance among all existing ensemble scores, it also outperformed all its component scores (SIFT, PolyPhen-2, LRT, MutationTaster and PhyloP) for separating deleterious variants from common neutral variants. For example, in testing dataset II, the probability of a TP variant having a higher predicted deleterious score than a TN variant is 0.89 (AUC = 0.89, 95% CI: 0.89-0.90) for KGGSeq, which is statistically significantly higher than all its component scores (AUC = 0.78, 95% CI: 0.77-0.79 for SIFT; AUC = 0.76, 95% CI: 0.76-0.77 for PolyPhen-2 HDIV; AUC = 0.77, 95% CI: 0.78-0.8 for PolyPhen-2 HVAR; AUC = 0.67, 95% CI: 0.66-0.67 for LRT; AUC = 0.71, 95% CI: 0.7-0.71 for

MutationTaster and AUC = 0.67, 95% CI: 0.66-0.68 for PhyloP). In testing dataset I, KGGSeq also performed better than all its component scores (AUC = 0.85, 95% CI: 0.81-0.9), but some of the difference was not statistically significant. On the other hand, other existing ensemble scores achieved only intermediate performance compared with its component scores. For example, the probability of a TP variant having a higher predicted deleterious score than a TN variant is 0.77 for CONDEL (AUC = 0.77, 95% CI: 0.76-0.78) in testing dataset II, which was statistically lower than its component score Mutation Assessor (AUC = 0.8, 95% CI: 0.8-0.81). In testing dataset I, CONDEL also achieved intermediate performance (AUC = 0.79, 95% CI: 0.73-0.85), which is worse than two of its component scores but the difference was not statistically significant (Fig. 1).

Moreover, we have found that ensemble scores that integrate component scores that are specific to protein features may perform better than ensemble scores that integrate a large amount of unfocused component scores. For example, KGGSeq, which integrated five component scores (SIFT, PolyPhen-2, LRT, MutationTaster and PhyloP), performed better than CADD, which integrated more than forty component scores (including deleteriousness prediction scores such as SIFT, PolyPhen-2, GERP++ and many more whole genome annotation information such as reference allele, GC content and methylation level, which are beyond the scope of our comparison) in both testing datasets. Nevertheless, our results show that integration of several deleteriousness prediction tools may be helpful in identifying deleterious mutations from common neutral mutations. Therefore, we developed our own ensemble scores LR and SVM that integrated ten component scores to test this hypothesis and the results are discussed below.

Fifth, we compared our two ensemble methods, both with each other and with all other existing deleteriousness prediction methods. Our two ensemble scores both integrated nine deleteriousness prediction scores including SIFT, PolyPhen-2, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy and PhyloP as well as the MMAF (to allow for the contribution of allele frequency to deleterious mutation discovery), but they used slightly different algorithms (LR for LR score and SVM with three different kernel functions for three SVM scores) to integrate their component scores. From our results, our ensemble scores not only outperformed all its components, but also outperformed all other deleteriousness prediction methods evaluated. Indeed, the probability of a TP variant having a higher predicted LR score than a TN

variant is 0.92 in testing dataset I (AUC = 0.92, 95% CI: 0.88-0.96) and in testing dataset II (AUC = 0.94, 95% CI: 0.93-0.94), both followed by radial SVM (AUC = 0.91 with 95% CI: 0.87-0.95 for testing dataset I and AUC = 0.93 with 95% CI: 0.92-0.93 for testing dataset II), indicating excellent separation of TP and TN variants. Both LR score and radial SVM score have significantly larger AUCs compared with FATHMM (the best individual deleteriousness prediction score tested) and KGGSeq (the best existing ensemble score tested) in testing dataset I (P-values<0.02, one-sided test with 2000 bootstrap) and testing dataset II (P-values of 2.20×10^{-16}, one-sided test with 2000 bootstrap). In testing dataset III with only neutral variants, LR achieved the lowest median score (0.195) among all normalized deleteriousness scores, indicating the highest tendency of correctly classifying TN observations (Fig. S2, S4).

Comparison of qualitative predictions

For many function prediction tools, such as SIFT, MutationTaster and PolyPhen-2, categorical predictions are available in addition to quantitative function scores. For conservation scores, we determined their optimal cutoffs for their qualitative predictions based on their ROC curves for the training data (see Materials and Methods section). Therefore, we dichotomized such categorical prediction outcomes and compared their performance using a series of different measurements, such as Matthews Correlation Coefficient (MCC), True Positive Rate (TPR, or sensitivity) and TNR, for testing datasets I and II, which contained both TP and TN observations. As testing dataset III contained only TN observations, we applied TNR as the sole measurement for qualitative prediction performance. Our results showed that existing deleteriousness prediction methods tend to have unstable performance in different testing datasets and on different measurements of qualitative outcomes. For example, the best agreement between binary prediction from deleteriousness prediction tools and the actual observation was achieved by MutationTaster in testing dataset I (MCC = 0.7), which did not retain in testing dataset II (MCC = 0.4 for MutationTaster, which is the 10th among the 17 scores compared). Similar results were also observed when using TPR and TNR as measurement. For example, while KGGSeq correctly classified the most TP variants in testing datasets I (TPR = 0.96) and II (TPR = 0.98), it failed to maintain its predictive power in correctly classifying TN variants (TNR = 0.52 and 0.35 for these two datasets, respectively) (Fig. 2). Indeed, its TNR became even lower when evaluated on testing dataset III (0.22) (Fig. S4).

On the other hand, our ensemble scores, remained to be high performing methods by various performance measurements in all three testing datasets. For example, LR achieved the second best agreement between binary prediction from deleteriousness prediction tools and the actual observation in testing dataset I (MCC = 0.68) and the best agreement in testing dataset II (MCC = 0.71), all followed by radial SVM (MCC = 0.66 for testing dataset I, MCC = 0.7 for testing dataset II). Similar results were also observed when using TPR and TNR as measurements. For example, in testing dataset I and II, LR correctly classified more than 80% of TP variants (TPR = 0.8, 95% CI: 0.76-0.84 for testing dataset I, TPR = 0.86, 95% CI: 0.855-0.865) and more than 85% of TN variants (TNR = 0.89, 95% CI: 0.86-0.92 for testing dataset I, TNR = 0.85, 95% CI: 0.845-0.855 for testing dataset II) (Fig. 2). In classifying rare neutral variants, LR and SVM with radial kernel also achieved excellent performance. They correctly classified the most TN variants in testing dataset III (TNR = 0.84 for LR, 95% CI: 0.83-0.85, TNR = 0.85 for SVM, 95% CI: 0.84-0.86), statistically significantly higher than the third best tool, FATHMM (TNR = 0.79, 95% CI: 0.78-0.80) (Fig. S4). This implies that large-scale ensemble methods may be able to take advantage of different deleteriousness prediction tools and achieve a more balanced qualitative prediction performance than individual tools.

Parameter tuning and feature selection for ensemble approaches

All the analyses described above on ensemble approaches (LR and SVM) used default parameters. To assess whether these parameters can be optimized to achieve substantially better performance, we performed five-fold cross validation on training dataset. Based on the AUC values of ROC curves for different parameter cocktails (Fig. S5), we found that the performance of default parameter for SVM with linear, radial and polynomial kernel was rather similar to that of other parameter settings, indicating that LR and SVM have already reached optimal performance using default settings so that the use of default settings for these two models were justified. To examine whether our model can be further enhanced using different combination of prediction scores, we applied step-wise feature selection and generated a reduced model, which lacks MutationTaster score. From step-wise feature selection, all prediction scores except for MutationTaster (SIFT score, PolyPhen-2 HDIV, PolyPhen-2, LRT, Mutation Assessor, FATHMM, GERP++, PhyloP, SiPhy and MMAF) were chosen to be the optimal feature combination with Akaike Information Criterion (AIC) of 16483.91 (Table S6). To compare the performance of this reduced model with our final model, which contains all

prediction scores, we calculated the accuracy and AUC value of the ROC curve of both models. Results showed that these two models share the same accuracy (Accuracy = 0.91 for both reduced model and our final model) and that there is no statistical significant difference between the reduced model and our final model (AUC=0.97, 95% CI: 0.968-0.971 for both models) (Table S7). Thus, the only difference in predictive performance between our final model and the reduced model with the optimal feature combination lies in the inclusion of MutationTaster. MutationTaster comprises unique features, such as splice site analysis, protein length analysis and direct or indirect protein feature analysis, which could potentially enhance our model in the future when more training data are available. Therefore, to allow for information of unique feature analysis from MutationTaster, we included it into our final model, which achieved the optimal performance as with the reduced model with the optimal combination of prediction scores.

Discussion

In this study, we evaluated the predictive performance of eighteen existing prediction methods and two new ensemble prediction methods on three manually curated testing datasets. Among the existing methods, FATHMM achieved the highest discriminative power as evaluated by AUC values of ROC curve. However, our two ensemble methods that incorporated nine deleteriousness prediction scores and MMAF, achieved the highest discriminative power and outperformed popular tools such as SIFT, GERP++ and PolyPhen-2. Our results demonstrated the value of combining information from multiple deleteriousness prediction methods. To facilitate variant prioritization in exome sequencing studies, we also merged mutations all but testing dataset III, generated our whole-exome ensemble-based prediction scores using LR and SVM algorithms and made the prediction scores for 87,347,044 possible variants in the whole-exome publicly available in ANNOVAR (37) software and dbNSFP database (38, 39).

Collinearity is an indicator of redundancy of prediction method pairs. To estimate such redundancy and avoid numerical instability of our model, we performed collinearity analysis on current eighteen scoring methods and MMAF. Results show that PolyPhen-2 HDIV score and PolyPhen-2 HVAR score have the highest linear correlation. PolyPhen-2 HDIV and PolyPhen-2 HVAR are positive correlated with each other; every 1

increase of PolyPhen-2 HDIV score is associated with 0.946 increase of PolyPhen-2 HVAR score (Pearson correlation coefficient=0.946 evaluated on testing dataset I, Fig. S10). Such strong association potentially owes to the fact that they share the same algorithm. Indeed, the only difference between these two scores lies in their training datasets. While PolyPhen-2 HDIV uses alleles encoding human proteins and their closely related mammalian homologs as TN observations, PolyPhen-2 HVAR applies common human nsSNVs as TN observations. Therefore, algorithm redundancy exists between these two prediction scores and was captured by Pearson Correlation Coefficient statistic. On the other hand, the least linear correlation was observed between MMAF and SIFT. SIFT score and MMAF are negative correlated with each other; every 1 increase of SIFT is associated with 0.030 decrease of MMAF, indicating little redundancy within this pair of scoring methods (Pearson Correlation Coefficient = -0.030 evaluated on testing dataset I, Fig. S10).

Besides predictive performance, ease of use and speed are also important aspects for application of a function prediction method. For large-scale WES on Mendelian diseases, it is necessary to perform queries on tens of thousands of variants in a relatively short period of time. Some authors of deleteriousness prediction methods provide software tools or web servers, on which batch queries are performed. Based on our experience, currently SIFT, PolyPhen-2, MutationTaster, Mutation Assessor and KGGSeq have the highest ease of use and speed among all methods tested, which allows for direct batch queries using genome coordinates. To facilitate querying multiple predictions for large number of nsSNVs, Liu et al. developed a database compiling prediction scores from six prediction algorithms (SIFT, PolyPhen-2, LRT, MutationTaster, Mutation Assessor and FATHMM) and three conservation scores (PhyloP, GERP++ and SiPhy) for all possible nsSNVs in human genome. With the dbNSFP tool, users no longer need to use individual software tools or web servers; instead they can obtain all scores easily from dbNSFP. We have also made all dbNSFP scores available directly in the ANNOVAR software, so that users can perform automated queries on all function prediction scores rapidly (less than one minute for a typical exome). We have now incorporated the SVM and LR scores into the ANNOVAR tool and dbNSFP, and we believe that these data sources will benefit researchers working on exome sequencing studies.

Missing values is another concern when applying a prediction method to large-scale WES data. Some methods tend to restrict their predictions to well-annotated proteins or transcripts, which may improve the prediction accuracy for non-missing scores, but they suffer from higher rate of missing values. For example, PON-P has a higher missing rate than their component scores because it depends on the availability of all component scores. Based on the datasets we used for training and testing, we observed some scores (PANTHER, SNPs&GO, PON-P and LRT) with relatively high (>10%) rate of missing values (Table 3). To reduce such bias caused by missing values, we used imputed scores for computing our SVM and LR scores if some component scores are missing.

Our study also suggested that we increase the predictive power of the prediction tools by integrating them through machine learning algorithms, such as LR and SVM. Both LR and SVM scores have enhanced performance than all existing prediction tools, indicating that current prediction tools may provide orthogonal information that can be integrated using such ensemble-based approaches for better performance. Such performance may be further enhanced if we could update our SVM and LR model dynamically, with more component scores and more accurate TP/TN observations, as other machine learning applications. However, due to the computation cost and the technical challenge of dynamically implementing many of the existing deleteriousness prediction methods online or locally for large numbers of mutations, we were not able to realize this function. Nevertheless, we were still optimistic that in the future, when most of the existing deleteriousness prediction tools come up with whole exome prediction scores, like the CADD team, or when they come up with an interface/software package that can handle large dataset easily, there may be possibilities of realizing such dynamic update of our ensemble scores. Moreover, even though LR performed better than SVM in our study, we caution that this does not necessarily suggest that LR is superior for this type of tasks. LR is relatively a simpler algorithm and therefore may be better when the data are simpler and do not require kernel transformation. More complex machine learning algorithms, such as SVM, Random Forest and Neural Network, can better handle data in large-scale and with multi-collinearity, which already appeared to some extend in our intermediate level training data (Fig. S6). In the future when the input scores are more complex and more correlated, more complex machine learning algorithms may be preferred over LR. For example, if we would like to integrate information such as phenotypic information (40,

41), pathway information (42) and protein interaction (43), we may consider those more complex machine learning algorithms.

A problem associated with every comparison study is the quality of the datasets. We used various ways to make sure that our four datasets are of high quality. For training dataset, we separated the deleterious mutations in Uniprot database according to the genetic model of the underlying disease (dominant, recessive, or unspecified) and conducted independent analysis, with the initial hypothesis that the mutations identified from recessive disease have higher reliability. With no significant difference observed among the three groups, we combined them as a single deleterious mutation set to increase estimation accuracy. We also limited the neutral mutations in testing dataset I to the nsSNVs that were observed more than ten times (as reported in dbSNP) with 1% or higher MMAF in the population. As for neutral mutation set for testing dataset II, we require mutations to have MMAF>1% or being observed more than 22 times in dbSNP. This MMAF cutoff may help to eliminate variants that were due to sequencing error or potentially deleterious variants with low MMAF (36). However, as rare and common variants may have different features, a method that performs well with relatively common (MMAF>1%) neutral mutations may not perform well with rare neural mutations. Therefore, we collected singleton nsSNVs from 824 exomes with strict control of the mapping quality and individual phenotypes as a test set for rare neural variants. Our results showed that our ensemble scores outperformed all other scores evaluated with higher TNR (i.e. lower false positive rate). Another issue associated with dataset is the overlap between our testing datasets and the original training datasets of the prediction scoring methods compared. If a prediction method is tested using a dataset, which overlaps its training dataset, then it is likely that its performance evaluation will be biased. In order to avoid this issue, we manually collected all three testing datasets, on which no current prediction scoring method is likely to be trained. For example, testing dataset I has TP observations manually collected from recent publications of *Nature Genetics* and TN observations from recently available data from CHARGE project. Applying these testing datasets helps us provide a relatively unbiased evaluation of performance of current prediction scoring methods.

Thusberg et al. previously published a nice comparison of nine deleteriousness prediction methods. Although seven out of nine of their compared methods (PolyPhen-2, SIFT, PANTHER, PhD-SNP, SNAP, SNPs&GO and MutPred) overlap our study, there are major differences between their study and ours. First, we included many more up-to-date methods into comparison. For example, we included recently published methods such as Mutation Assessor, FATHMM and CADD. Second, our comparison has a larger diversity. While they focused exclusively on the function prediction methods that use information of amino acid substitutions, we not only included methods based on such amino acid substitutions and but also methods that are based on DNA sequence information (such as conservation scores) and ensemble scores. Third, while Thusberg et al. only compared binary predictions, we compared both the continuous prediction scores and the binary predictions. Finally, we constructed new ensemble-based methods (i.e. LR and SVM) that take advantage of the characteristics of multiple algorithms in a single application. Nevertheless, we also evaluated the same datasets provided by Thusberg and presented the results in Fig. S7.

In summary, we have performed a comparative study of several popular prediction algorithms for nsSNVs in response to the demand of developing a practical guideline for their use in WES-based human disease studies. Using three independent testing datasets, we evaluated eighteen existing deleteriousness prediction methods on their quantitative and qualitative prediction performance and demonstrated the potential of ensemble approaches (LR and SVM). We recommended the regular use of such ensemble methods in prioritizing disease-causing nsSNVs to boost the predictive power.

Materials and Methods

Training and testing data sets

We manually collected four datasets of nsSNVs based on the Uniprot database, previous publication in Nature Genetics, CHARGE sequencing project and VariBench dataset (Table 1). Training dataset included 14,191 deleterious mutations, which were annotated as causing Mendelian disease and 22,001 neutral mutations, which were annotated as not known to be associated with any phenotypes, all based on Uniprot annotation. Considering that some algorithms may have used a part of the nsSNVs in the Uniprot set as training data, we constructed our first independent testing dataset, testing dataset I, to reduce potential bias

associated with the Uniprot data. It included 120 deleterious mutations recently reported in the journal *Nature Genetics* that belong to 54 different genes and cause 49 diverse diseases, and 124 neutral mutations newly discovered from the CHARGE sequencing project. To ensure the quality of these deleterious mutations, we used mutations that were reported to cause Mendelian diseases with experimental supports and were published after January 1st, 2011. To ensure the quality of the neutral mutations, we applied the following criteria when selecting them: (i) with a MMAF >1% in 2,269 exomes from the Atherosclerosis Risk in Communities Study (ARIC) study via the CHARGE sequencing project (34); (ii) not reported in the 1000 Genomes Project or Uniprot and (iii) with Hardy-Weinberg exact test P-value > 0.01. The 1% threshold for MMAF was chosen to minimize the contamination of potential Mendelian disease-causing mutations in the neutral control set. To provide high quality benchmark datasets, on which novel variant scoring methods can be evaluated, Thusberg et al. curated VariBench datasets, seven different datasets of high quality variants, whose effects were experimentally verified. These seven datasets cover variants with different properties, such as variants affecting protein tolerance, variants affecting transcription factor binding sites and variants affecting mRNA splice site. Because our study is mostly related to variants affecting protein tolerance, we chose this dataset as benchmark dataset. Original TN observations from this dataset contain 17,393 nsSNV extracted from dbSNP database build 131. Original TP observations from this dataset contain 14,610 missense variations obtained by manual curation from the PhenCode database downloaded in June, 2009, IDBases and from 16 individual LSDBs. Because 57.02% of TP observations and 23.87% of TN observations overlap our training dataset, we removed these variants from VariBench dataset and constructed our second independent testing dataset, testing dataset II, and curated 6,279 deleterious variants as our TP observations and 13,240 common neutral variants (MMAF>1%) as our TN observations. In addition to the first two testing datasets that contain only common variants as TN observations, we also prepared testing dataset III that contains only 10,164 singleton neutral mutations (MMAF<0.1%). We collected the rare neutral nsSNVs from 824 European Americans from cohort random samples of the ARIC study[32] via the CHARGE sequencing project (34). We retained only nsSNVs that have only one alternative allele observed in the sample (i.e. singletons). We further removed nsSNVs that have been reported in the 1000 Genomes Project or the NHLBI's Exome Sequencing Project, to make sure those nsSNVs are truly rare and novel. To reduce the artifacts that are due to mapping errors, we further filtered out any nsSNVs that reside outside the strict mask of the 1000

Genome Project and that with an ENCODE (44) mappability score smaller than 1. To reduce interpretation complication, we removed nsSNVs that correspond to multiple transcripts as to RefSeq, Ensembl and Uniprot. After the above filtering steps, we retained 10,164 rare nsSNVs. As this number is too large to interrogate all the prediction methods, we only compared the methods that are relatively easy to scale up for large number of queries, which include SIFT, PolyPhen-2, LRT, MutationTaster, Mutation Assessor, FATHMM, CONDEL, KGGSeq and our two ensemble scores.

Training dataset was used for modeling LR and SVM and testing datasets I and II were used for evaluating the performance of all deleteriousness prediction methods on separating deleterious mutations from common neutral mutations and testing dataset III was used for evaluating the performance on distinguishing rare neutral mutations.

Additionally, to compare the performance of all deleteriousness prediction tools with the result from Thusberg et al, we derived two additional testing dataset, additional testing dataset I and additional testing dataset II, from benchmark dataset from his team (Table S4). The results are shown in Fig. S7 and S8.

Deleteriousness prediction methods

We compared the predictive performance of all prediction methods (Table 2). Despite their differences in the use of prediction features, training data, and statistical models, the premise of most algorithms was that nsSNVs, which are evolutionarily conserved and/or lead to critical protein structure changes are likely to be deleterious. Each method assigns a quantitative prediction score measuring the likelihood of an nsSNV being deleterious as well as qualitative prediction such as “benign” or “damaging” based on some algorithm-specific threshold for the prediction score, except for SiPhy, GERP++ and PhyloP, which only provide a prediction score.

In order to compare the performance of the quantitative scores of these prediction methods, we obtained all the prediction scores (Table S5). Among them, PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, GERP++ and PhyloP were obtained from the dbNSFP database version 2.1; PON-P, PANTHER, PhD-SNP, SNAP and SNPs&GO were obtained via the PON-P webserver; MutPred, Condel (ensemble score of SIFT, PolyPhen-2 and Mutation Assessor) and KGGSeq were obtained manually through

their own webservers or software package. Due to the limitation of the PON-P webserver that each batch query can only submit up to ten variants, we only obtained PON-P, PANTHER, PhD-SNP, SNAP and SNPs&GO scores for testing dataset I. It is noted that the prediction scores obtained from dbNFSP database underwent transformation from original prediction scores (Table 2). We plotted the ROC curve for each of these prediction scores, together with MMAF and ensemble scores, and computed their AUC value of the ROC curve as cumulative statistics to evaluate their performances. As for testing dataset III that contains only TN observation, we were not able to plot the ROC curve. Therefore, we rescaled all the transformed prediction scores with suggested binary cutoff from their continuous prediction score into 0-1 scale with 0.5 being the binary cutoff of deleteriousness. We calculated the median of all prediction scores and demonstrated our result in Fig. S2.

We also compared the performance of qualitative categorization of these prediction methods. We dichotomized prediction scores according to their current dichotomizing threshold recommended in the literature(38, 39) for those providing dichotomizing thresholds. Note that there are three categorical outcomes for LRT, D (Deleterious), N (Neutral) and U (Unknown). In order to rule out the influence of unknown prediction outcomes and more accurately assess its performance, we discarded the mutations with U prediction outcome and calculated its qualitative prediction performance measurements such as MCC, TPR and TNR. Note that MCC is a balanced measurement of qualitative prediction outcome and it ranges from -1 to 1(45). For the remaining prediction methods with no available dichotomizing thresholds, such as SiPhy, GERP++ and PhyloP, we calculated the thresholds as the points that were closest to the left-upper corner in their ROC curves with training dataset and used these thresholds for dichotomization with all testing datasets. We compared their sensitivity and specificity at the dichotomous thresholds for testing datasets I and II and demonstrated the results in Fig. 2. As for testing dataset III that contains only TN observations, we calculated TNR for all deleteriousness prediction scores and demonstrated the results in Fig. S4. We also dichotomized our ensemble-based scores according to their standard dichotomizing threshold (0 for SVM and 0.5 for LR) and incorporated them into qualitative comparison with other prediction methods.

Missing scores

As for mutations whose prediction scores contained certain missing values (Table 3), we used BPCAfill program, a tool designed to impute large dataset with correlated columns or rows with good performance in previous studies, to impute these missing values for each of such mutation, by borrowing information from its available scores. The imputation was conducted for PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, GERP++ and PhyloP using all non-missing scores for all potential nsSNVs in human exome collected in the dbNSFP database v2.1. We reported the overall missing value percentage before and after imputation in Table S3. Because there was less than 3% missing rate for all the prediction tools after imputation, the small percentage of mutations with missing values was discarded in our comparison analysis without risking selection bias. These imputed scores were used to compute two ensemble-based scores using SVM and LR algorithms. Note that to avoid the effect of imputed score on performance evaluation in testing phase, mutations with any missing values were excluded to ensure that comparison of all deleteriousness prediction scores was done using only output scores from the prediction methods.

SVM and LR

In order to analyze the advantage of incorporating various prediction scores, we used two models, SVM and LR, to compute the ensemble-based scores. We harvested output scores from all of the prediction methods for all mutations in all of our datasets, combined them with MMAF from various populations and integrated them into input files for constructing LR and SVM, with linear kernel, radial kernel and polynomial kernel using R package e1071 (46). Performance for each model under each specific setting was tested on testing datasets I and II and was evaluated using R package ROCR (47). Because testing dataset III contains only TN observations, we applied manually calculated TNR for evaluating its performance.

To assess whether the parameters of our SVM model can be optimized to generate substantially better performance, we performed five-fold cross validation on training set for different parameter cocktails. For linear SVM, cost of 0.01, 0.1, 1 (default), 10 and 100 were evaluated for performance. For polynomial SVM, degree of 2, 3 (default), 4, 5 and 6 with default parameter for cost were analyzed. For radial SVM, different combinations of gamma of 0.01, 0.1, 1, 10, 100 and 1/11 (default) and cost of 0.01, 0.1, 1 (default), 10 and 100 were evaluated. Default parameters for all SVMs were chosen as a result. Moreover, in order to assess the relative contribution of each prediction score to the performance of LR and SVM, we tested several modified

SVM and LR models with one prediction score deleted from the original models and plotted average ROC curve and AUC value, as shown in Fig. S9. In addition, in order to test whether our model can be further improved by using different combinations of prediction scores, we applied step-wise model selection using Akaike Information Criterion (AIC) statistic as a criterion. The resulting model from step-wise model selection is shown in Table S6 and is compared with our final model in Table S7.

To assess the model assumption and evaluate pair-wise redundancy of prediction scores, we checked the multi-collinearity between all pairs of predictors using R and demonstrated the results in Fig. S6.

ROC curves

We used ROC curves and their AUC values to compare the performance of the quantitative individual prediction scores and the ensemble-based scores. For each prediction/ensemble score we evaluated, we varied the threshold for calling deleterious nsSNVs from the minimum value to the maximum value of this score. And for each threshold, we computed corresponding sensitivity (TP/(TP+FN)), and specificity (1-FP/(FP+TN)) for this score, with respect to its proportions of true positive and true negative nsSNVs at this threshold, where TP, FP, FN and TN corresponds to true positive, false positive, false negative and true negative predictions, respectively. Having values for each point value for each score, the corresponding ROC curve was therefore obtained by plotting sensitivity against 1-specificity at each threshold for this score. On the other hand, for qualitative prediction evaluation, we computed the point estimates of sensitivity and specificity instead, using dichotomized prediction/ensemble score. The ROC plot and sensitivity against specificity plot were generated using R package ROCR and the 95% Confidence Interval (CI) using 2000 bootstrap was generated using R package pROC (48).

Funding

This work was supported by National Institute of Health [R01-HL116720, R01-CA169122, R01-HL106034 to P.W., RC2-HL02419, U54HG003273 to X.J., E.B., R.G. and X.L. and R01-HG006465 to C.D. and K.W].

Acknowledgements

We thank Sara Barton for her help on polishing the writing.

Author Contributions

X.L., K.W. and P.W. designed the study. X.J. and X.L. compiled the data. C.D. and P.W. conducted the analysis. E.B. provided critical comments. R.G. and E.B. provided ARIC sequence data. C.D., P.W., K.W. and X.L. wrote the manuscript. X.L. and K.W. coordinated all steps of the study.

Competing Interests

K.W. is board member and stock holder of Tute Genomics, a bioinformatics consulting company for next-generation sequencing data.

Supplementary Data

Supplementary Materials include eight figures and six tables.

Reference

1 Ng, S.B., Nickerson, D.A., Bamshad, M.J. and Shendure, J. (2010) Massively parallel sequencing and rare disease. *Hum. Mol. Genet.*, **19**, R119-124.

2 Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

3 Ng, P.C. and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genom. Hum. G.*, **7**, 61-80.

4 Thusberg, J. and Vihinen, M. (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.*, **30**, 703-714.

5 Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J. and Nickerson, D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250-251.

6 Chun, S. and Fay, J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553-1561.

7 Flanagan, S.E., Patch, A.M. and Ellard, S. (2010) Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomarkers.*, **14**, 533-537.

8 Schwarz, J.M., Rodelsperger, C., Schuelke, M. and Seelow, D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575-576.

9 Wei, P., Liu, X. and Fu, Y.X. (2011) Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study. *BMC Proc.*, **5 Suppl 9**, S20.

10 Thusberg, J., Olatubosun, A. and Vihinen, M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358-368.

11 Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248-249.

12 Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073-1081.

13 Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N. and Gaunt, T.R. (2013) Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, **29**, 1504-1510.

14 Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129-2141.

15 Capriotti, E., Calabrese, R. and Casadio, R. (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, **22**, 2729-2734.

16 Bromberg, Y. and Rost, B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823-3835.

17 Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. and Casadio, R. (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237-1244.

18 Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D. and Radivojac, P. (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, **25**, 2744-2750.

19 Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310-315.

20 Olatubosun, A., Valiaho, J., Harkonen, J., Thusberg, J. and Vihinen, M. (2012) PON-P: integrated predictor for pathogenicity of missense variants. *Hum. Mutat.*, **33**, 1166-1174.

21 Li, M.X., Gui, H.S., Kwan, J.S., Bao, S.Y. and Sham, P.C. (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.*, **40**, e53.

22 Gonzalez-Perez, A. and Lopez-Bigas, N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440-449.

23 Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.

24 Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N. and Xie, X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54-62.

25 URL http://www.broadinstitute.org/mammals/2x/siphy_hg19/.

26 Cooper, G.M., Stone, E.A., Asimenos, G., Program, N.C.S., Green, E.D., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901-913.

27 URL http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals/.

28 Corinna Cortes, V.V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273-297.

29 Agresti, A. (2002) *Categorical Data Analysis*. New York: Wiley-Interscience.

30 UniProt, C. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214-219.

31 Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187-191.

32 The 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

33 The ARIC investigators. (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am. J. Epidemiol.*, **129**, 687-702.

34 Morrison, A.C., Voorman, A., Johnson, A.D., Liu, X., Yu, J., Li, A., Muzny, D., Yu, F., Rice, K., Zhu, C. *et al.* (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat. Genet.*, **45**, 899-901.

35 Sasidharan Nair, P. and Vihinen, M. (2013) VariBench: a benchmark database for variations. *Hum. Mutat.*, **34**, 42-49.

36 The 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.

37 Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

38 Liu, X., Jian, X. and Boerwinkle, E. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894-899.

39 Liu, X., Jian, X. and Boerwinkle, E. (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393-2402.

40 Robinson, P.N., Kohler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D. *et al.* (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, **24**, 340-348.

41 Sifrim, A., Popovic, D., Tranchevent, L.C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J.R., Aerts, J., De Moor, B. and Moreau, Y. (2013) eXtasy: variant prioritization by genomic data fusion. *Nat. Methods*, **10**, 1083-1084.

42 Carbonetto, P. and Stephens, M. (2013) Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. *PLoS Genet.*, **9**, e1003770.

43 Guney, E. and Oliva, B. (2012) Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PloS One*, **7**, e43557.

44 ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799-816.

45 Vihinen, M. (2012) How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, **13 Suppl 4**, S2.

46 Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, A. (2014) e1071: misc functions of the Department of Statistics (e1071), TU Wien. URL http://cran.r-project.org/web/packages/e1071/index.html

47 Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940-3941.

48 Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., Muller, M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12:77.

Legends to Figures

Figure. 1. ROC curves for existing prediction scores and our ensemble scores

These two plots illustrated performance of quantitative prediction outcomes for existing prediction scores and our ensemble prediction scores evaluated by the Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC) score for the ROC curve. Higher AUC score indicates better performance. Top plot used testing dataset I as benchmark dataset and bottom plot used testing dataset II as benchmark dataset (see Table 1). 95% CI indicates 95% confidence interval computed with 2000 stratified bootstrap replicates.

Figure 2. Sensitivity and specificity plots for existing prediction scores and our ensemble scores

These two plots illustrated the performance of qualitative prediction outcomes of existing prediction scores and our ensemble prediction scores, evaluated by sensitivity and specificity. Higher sensitivity/specificity score indicates better performance. Top plot used testing dataset I as benchmark dataset and bottom plot used testing dataset II as benchmark dataset (see Table 1). The legend table showed various qualitative prediction performance measurements for each prediction tool. Matthews Correlation Coefficient (MCC) is a correlation coefficient between the observed and predicted binary classification, ranging from -1 to 1, where 1 indicates perfect prediction, -1 indicates total disagreement between prediction and observation. $\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$, where TP, TN, FP and FN denotes true positive, true negative, false positive and false negative respectively. ACC denotes accuracy. $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$. TPR denotes true positive rate, or sensitivity. $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$. TNR denotes true negative rate, or specificity. $\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$. FPR denotes false positive rate. $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$. FNR denotes false negative rate. $\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}$. PPV denotes positive predictive value. $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$. NPV denotes negative predictive value. $\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$. FDR denotes false discovery rate. $\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}$. For each qualitative prediction performance measurement, top three performance scores were highlighted. The brighter the highlight color, the better the performance.

Table 1. Description of the four datasets used in our study

Dataset	Training dataset	Testing dataset I	Testing dataset II	Testing dataset III
TP	14,191	120	6,279	0
TN	22,001	124	13,240	10,164
Total	36,192	244	19,519	10,164
Source	Uniprot database (30, 31)	Recent Nature Genetics publications for TP variants CHARGE database (34) for TN variants	VariBench dataset II (10, 35) without mutations in training dataset	CHARGE database (34)

TP: True Positive, number of deleterious mutations that were treated as true positive observations in modeling. TN: True Negative, number of non-deleterious mutations that were treated as true negative observations in modeling. Total: Total number of mutations for each dataset. (Total = TP+TN)

Table 2: Summary of deleteriousness prediction methods analyzed in our study

Name	Category	Score used for analysis	Deleterious Threshold	Information used
SIFT	Function prediction	1 - Score	>0.95	Protein sequence conservation among homologs
PolyPhen-2	Function prediction	Score	>0.5	8 protein sequence features, 3 protein structure features
LRT	Function prediction	Score*0.5 (if Omega ≥1) or 1-Score*0.5 (if Omega<1)	P	DNA sequence evolutionary model
MutationTaster	Function prediction	Score (if A or D) or 1-Score (if N or P)	>0.5	DNA sequence conservation, splice site prediction, mRNA stability prediction, protein feature annotations
Mutation Assessor	Function prediction	(Score-Min)/(Max-Min)	>0.65	Sequence homology of protein families and sub-families within and between species
FATHMM	Function prediction	1 - (Score-Min)/(Max-Min)	>=0.45	Sequence homology
GERP++ RS	Conservation Score	Score	>4.4	DNA sequence conservation
PhyloP	Conservation Score	Score	>1.6	DNA sequence conservation
SiPhy	Conservation Score	Score	>12.17	Inferred nucleotide substitution pattern per site
PON-P	Ensemble Score	Score	P	Random forest methodology-based pipeline integrating five predictors
PANTHER	Function prediction	Score	P	Phylogenetic trees based on protein sequences
PhD-SNP	Function prediction	Score	P	SVM-based method using protein sequence and profile information
SNAP	Function prediction	Score	P	Neural network-based method using DNA sequence information as well as functional and structural annotations
SNPs&GO	Function prediction	Score	P	SVM-based method using information from protein sequence, protein sequence profile, and protein function
MutPred	Function prediction	Score	>0.5	Protein sequence-based model using SIFT and a gain/loss of 14 different structural and functional properties
KGGSeq	Ensemble Score	Score	P	Filtration and prioritization framework using information from three levels: genetic level, variant-gene level and knowledge level
CONDEL	Ensemble Score	Score	>0.49	Weighted average of the normalized scores of five methods
CADD	Ensemble Score	Score	>15	63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), data from the ENCODE project and information from UCSC genome browser tracks

Score indicates raw score for the corresponding function prediction/conservation score output. Max/Min indicates the max/min prediction score. Classification outcome for deleterious threshold was used for qualitative prediction score analysis. P means that categorical prediction outcome given by the prediction method was used for dichotomizing prediction outcomes.

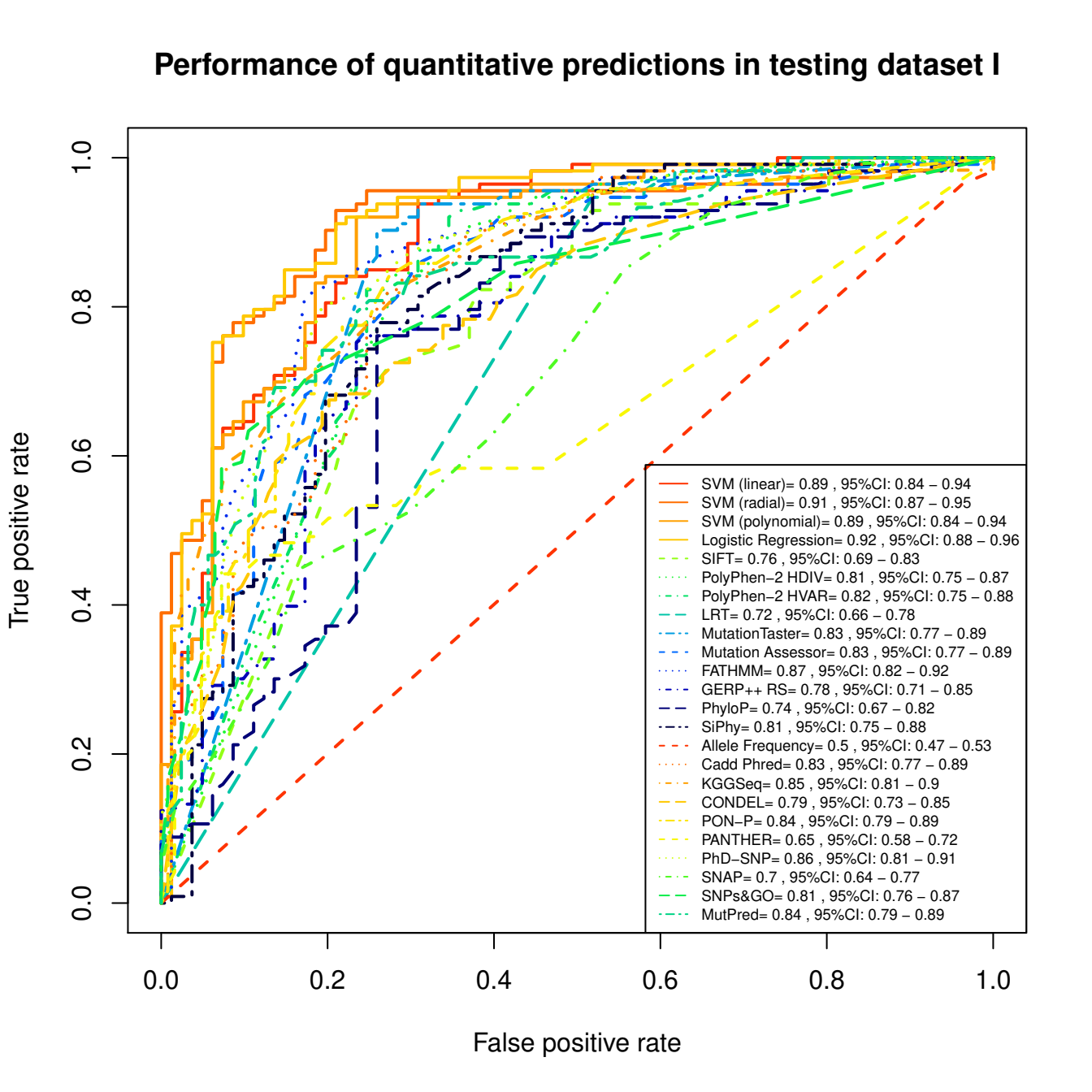
Table 3. Missing values for four datasets (%)

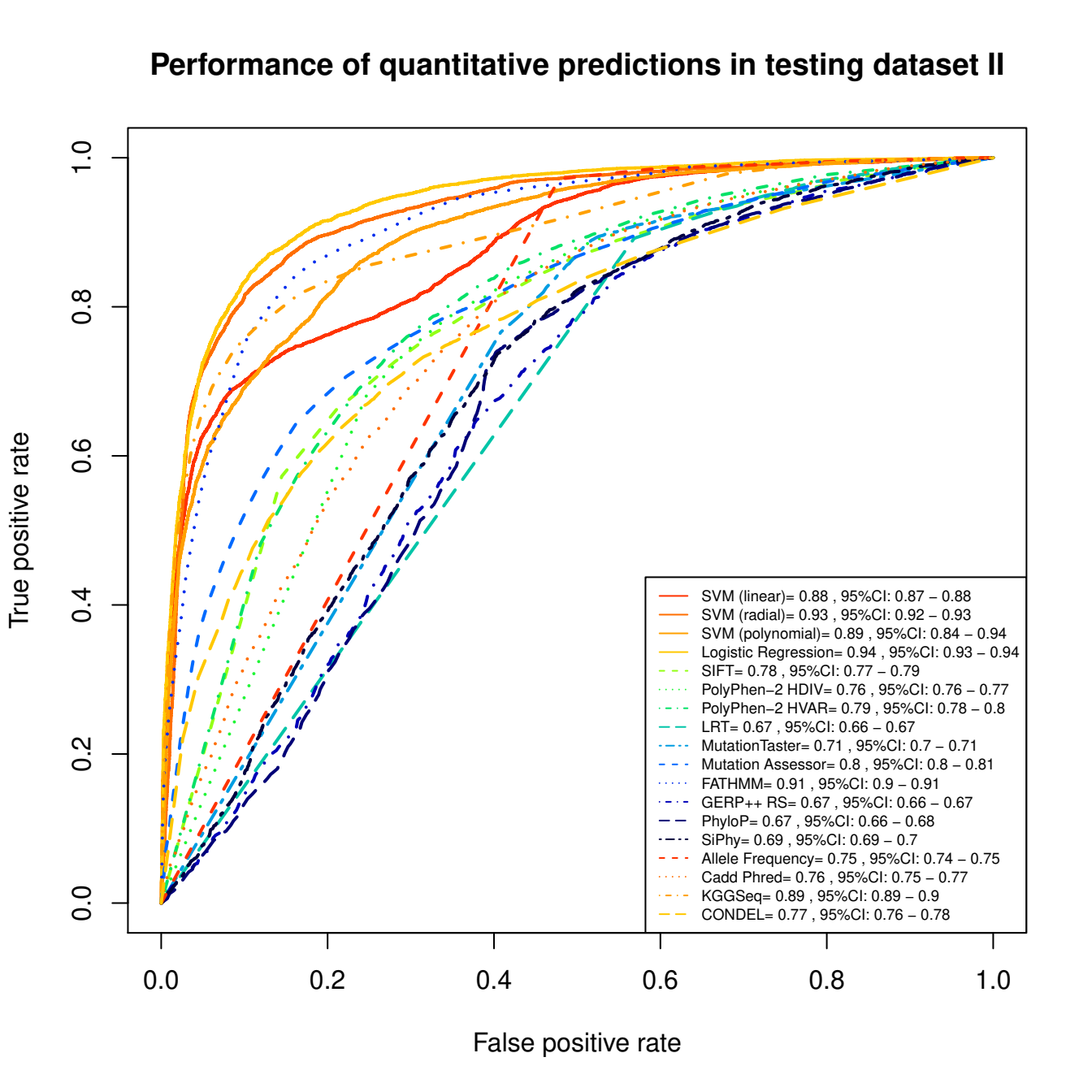
Dataset	Training dataset (%)	Testing dataset I (%)	Testing dataset II (%)	Testing dataset III (%)
SIFT	6.91	2.87	3.83	3.96
PolyPhen-2	3.79	2.87	0.55	0.02
LRT	10.49	13.93	7.97	11.33
MutationTaster	0.04	0.41	0.10	0.11
Mutation Assessor	1.51	3.69	2.23	2.67
FATHMM	4.05	5.73	3.48	6.69
GERP++	0	0	0	0.01
PhyloP	0	0	0	0
SiPhy	0	0	0	0.285
PON-P	NA	13.93	NA	NA
PANTHER	NA	47.95	NA	NA
PhD-SNP	NA	6.56	NA	NA
SNAP	NA	9.02	NA	NA
SNPs&GO	NA	15.98	NA	NA
MutPred	NA	7.37	NA	NA
KGGSeq	NA	0.82	0.05	0.82
CONDEL	NA	0.41	0.0003	0.32
CADD	0	0	0	0

(%): the percentage of missing values for each prediction scores for the corresponding dataset. NA: not available.

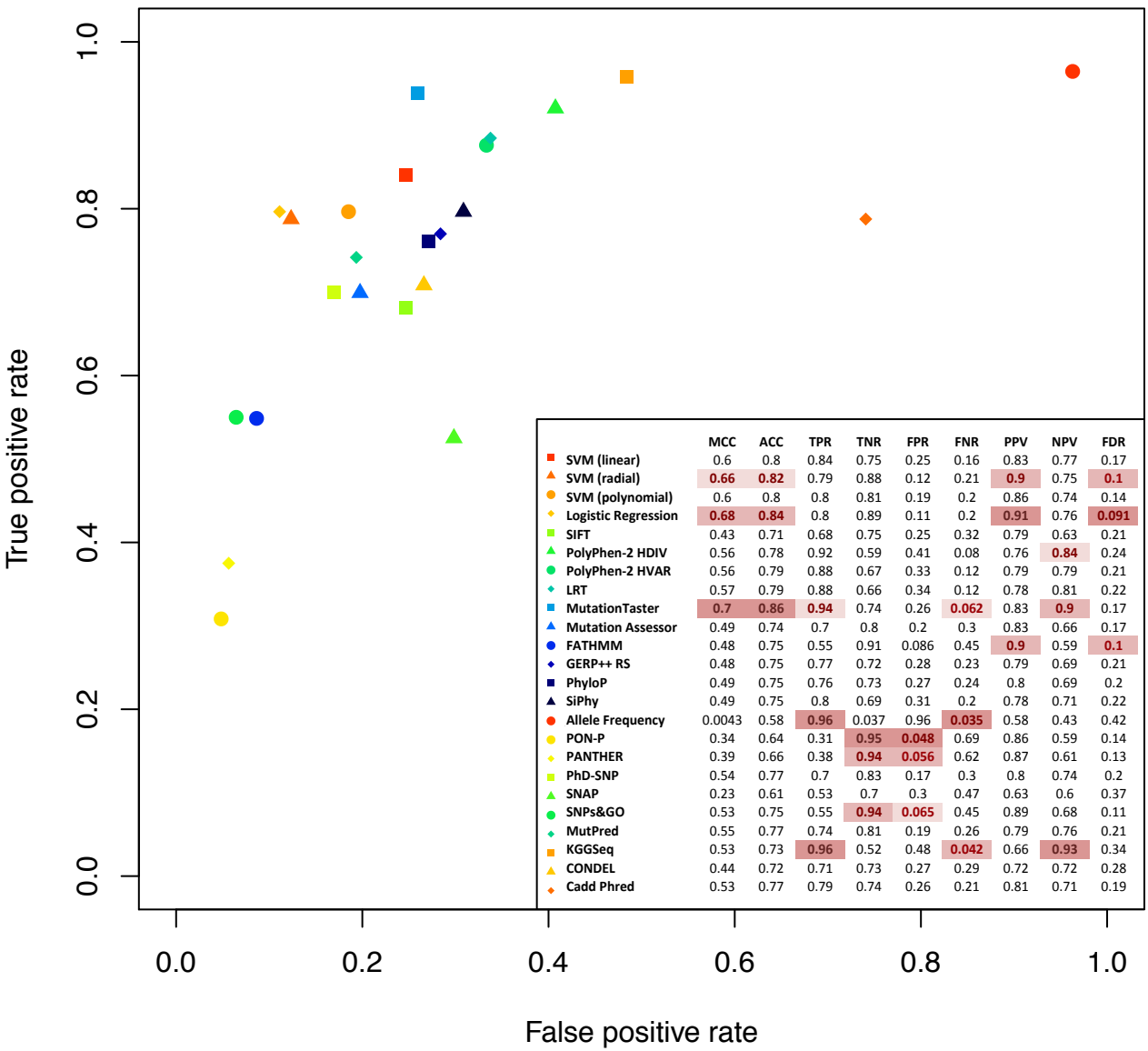
Abbreviations:

Whole Exome Sequencing (WES), nonsynonymous SNPs (nsSNVs), Support Vector Machine (SVM), Logistic Regression (LR), True Positive (TP), True Negative (TN), Maximum Minor Allele Frequency (MMAF), Atherosclerosis Risk in Communities Study (ARIC), The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), Area Under the Curve (AUC), Receiver Operating Characteristic (ROC), True Negative Rate (TNR), Matthews Correlation Coefficient (MCC), True Positive Rate (TPR).


Performance of quantitative predictions in testing dataset I
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (linear)= 0.89 , 95%CI: 0.84 – 0.94
SVM (radial)= 0.91 , 95%CI: 0.87 – 0.95
SVM (polynomial)= 0.89 , 95%CI: 0.84 – 0.94
Logistic Regression= 0.92 , 95%CI: 0.88 – 0.96
SIFT= 0.76 , 95%CI: 0.69 – 0.83
PolyPhen−2 HDIV= 0.81 , 95%CI: 0.75 – 0.87
PolyPhen−2 HVAR= 0.82 , 95%CI: 0.75 – 0.88
LRT= 0.72 , 95%CI: 0.66 – 0.78
MutationTaster= 0.83 , 95%CI: 0.77 – 0.89
Mutation Assessor= 0.83 , 95%CI: 0.77 – 0.89
FATHMM= 0.87 , 95%CI: 0.82 – 0.92
GERP++ RS= 0.78 , 95%CI: 0.71 – 0.85
PhyloP= 0.74 , 95%CI: 0.67 – 0.82
SiPhy= 0.81 , 95%CI: 0.75 – 0.88
Allele Frequency= 0.5 , 95%CI: 0.47 – 0.53
Cadd Phred= 0.83 , 95%CI: 0.77 – 0.89
KGGSeq= 0.85 , 95%CI: 0.81 – 0.9
CONDEL= 0.79 , 95%CI: 0.73 – 0.85
PON−P= 0.84 , 95%CI: 0.79 – 0.89
PANTHER= 0.65 , 95%CI: 0.58 – 0.72
PhD−SNP= 0.86 , 95%CI: 0.81 – 0.91
SNAP= 0.7 , 95%CI: 0.64 – 0.77
SNPs&GO= 0.81 , 95%CI: 0.76 – 0.87
MutPred= 0.84 , 95%CI: 0.79 – 0.89

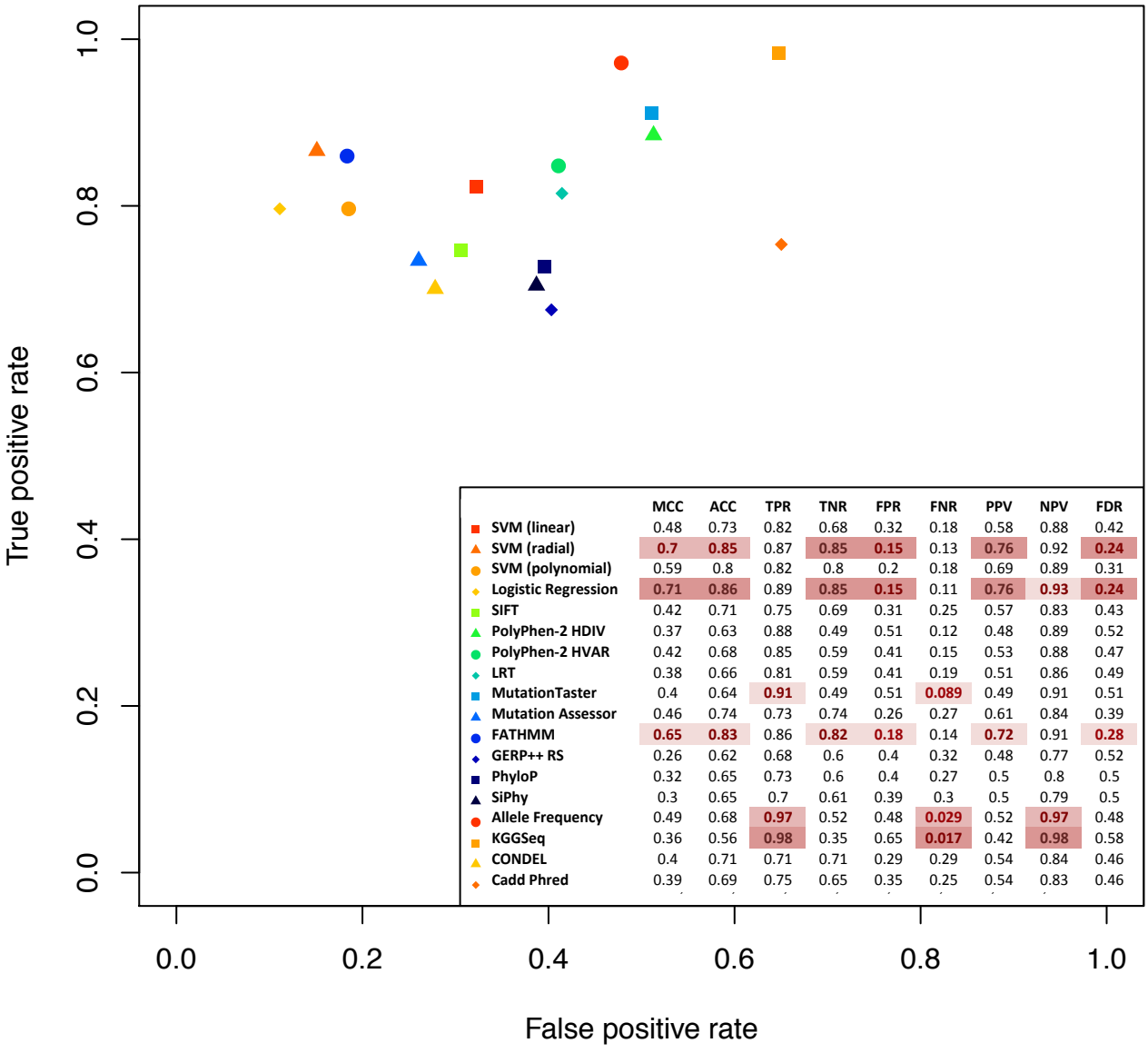

Performance of quantitative predictions in testing dataset II
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (linear)= 0.88 , 95%CI: 0.87 – 0.88
SVM (radial)= 0.93 , 95%CI: 0.92 – 0.93
SVM (polynomial)= 0.89 , 95%CI: 0.84 – 0.94
Logistic Regression= 0.94 , 95%CI: 0.93 – 0.94
SIFT= 0.78 , 95%CI: 0.77 – 0.79
PolyPhen−2 HDIV= 0.76 , 95%CI: 0.76 – 0.77
PolyPhen−2 HVAR= 0.79 , 95%CI: 0.78 – 0.8
LRT= 0.67 , 95%CI: 0.66 – 0.67
MutationTaster= 0.71 , 95%CI: 0.7 – 0.71
Mutation Assessor= 0.8 , 95%CI: 0.8 – 0.81
FATHMM= 0.91 , 95%CI: 0.9 – 0.91
GERP++ RS= 0.67 , 95%CI: 0.66 – 0.67
PhyloP= 0.67 , 95%CI: 0.66 – 0.68
SiPhy= 0.69 , 95%CI: 0.69 – 0.7
Allele Frequency= 0.75 , 95%CI: 0.74 – 0.75
Cadd Phred= 0.76 , 95%CI: 0.75 – 0.77
KGGSeq= 0.89 , 95%CI: 0.89 – 0.9
CONDEL= 0.77 , 95%CI: 0.76 – 0.78

Performance of qualitative predictions in testing dataset I



	MCC	ACC	TPR	TNR	FPR	FNR	PPV	NPV	FDR
SVM (linear)	0.6	0.8	0.84	0.75	0.25	0.16	0.83	0.77	0.17
SVM (radial)	0.66	0.82	0.79	0.88	0.12	0.21	0.9	0.75	0.1
SVM (polynomial)	0.6	0.8	0.8	0.81	0.19	0.2	0.86	0.74	0.14
Logistic Regression	0.68	0.84	0.8	0.89	0.11	0.2	0.91	0.76	0.091
SIFT	0.43	0.71	0.68	0.75	0.25	0.32	0.79	0.63	0.21
PolyPhen-2 HDIV	0.56	0.78	0.92	0.59	0.41	0.08	0.76	0.84	0.24
PolyPhen-2 HVAR	0.56	0.79	0.88	0.67	0.33	0.12	0.79	0.79	0.21
LRT	0.57	0.79	0.88	0.66	0.34	0.12	0.78	0.81	0.22
MutationTaster	0.7	0.86	0.94	0.74	0.26	0.062	0.83	0.9	0.17
Mutation Assessor	0.49	0.74	0.7	0.8	0.2	0.3	0.83	0.66	0.17
FATHMM	0.48	0.75	0.55	0.91	0.086	0.45	0.9	0.59	0.1
GERP++ RS	0.48	0.75	0.77	0.72	0.28	0.23	0.79	0.69	0.21
PhyloP	0.49	0.75	0.76	0.73	0.27	0.24	0.8	0.69	0.2
SiPhy	0.49	0.75	0.8	0.69	0.31	0.2	0.78	0.71	0.22
Allele Frequency	0.0043	0.58	0.96	0.037	0.96	0.035	0.58	0.43	0.42
PON-P	0.34	0.64	0.31	0.95	0.048	0.69	0.86	0.59	0.14
PANTHER	0.39	0.66	0.38	0.94	0.056	0.62	0.87	0.61	0.13
PhD-SNP	0.54	0.77	0.7	0.83	0.17	0.3	0.8	0.74	0.2
SNAP	0.23	0.61	0.53	0.7	0.3	0.47	0.63	0.6	0.37
SNPs&GO	0.53	0.75	0.55	0.94	0.065	0.45	0.89	0.68	0.11
MutPred	0.55	0.77	0.74	0.81	0.19	0.26	0.79	0.76	0.21
KGGSeq	0.53	0.73	0.96	0.52	0.48	0.042	0.66	0.93	0.34
CONDEL	0.44	0.72	0.71	0.73	0.27	0.29	0.72	0.72	0.28
Cadd Phred	0.53	0.77	0.79	0.74	0.26	0.21	0.81	0.71	0.19

Performance of qualitative predictions in testing dataset II



	MCC	ACC	TPR	TNR	FPR	FNR	PPV	NPV	FDR
SVM (linear)	0.48	0.73	0.82	0.68	0.32	0.18	0.58	0.88	0.42
SVM (radial)	**0.7**	**0.85**	0.87	**0.85**	**0.15**	0.13	**0.76**	0.92	**0.24**
SVM (polynomial)	0.59	0.8	0.82	0.8	0.2	0.18	0.69	0.89	0.31
Logistic Regression	**0.71**	**0.86**	0.89	**0.85**	**0.15**	0.11	**0.76**	**0.93**	**0.24**
SIFT	0.42	0.71	0.75	0.69	0.31	0.25	0.57	0.83	0.43
PolyPhen-2 HDIV	0.37	0.63	0.88	0.49	0.51	0.12	0.48	0.89	0.52
PolyPhen-2 HVAR	0.42	0.68	0.85	0.59	0.41	0.15	0.53	0.88	0.47
LRT	0.38	0.66	0.81	0.59	0.41	0.19	0.51	0.86	0.49
MutationTaster	0.4	0.64	**0.91**	0.49	0.51	**0.089**	0.49	0.91	0.51
Mutation Assessor	0.46	0.74	0.73	0.74	0.26	0.27	0.61	0.84	0.39
FATHMM	**0.65**	**0.83**	0.86	**0.82**	**0.18**	0.14	**0.72**	0.91	**0.28**
GERP++ RS	0.26	0.62	0.68	0.6	0.4	0.32	0.48	0.77	0.52
PhyloP	0.32	0.65	0.73	0.6	0.4	0.27	0.5	0.8	0.5
SiPhy	0.3	0.65	0.7	0.61	0.39	0.3	0.5	0.79	0.5
Allele Frequency	0.49	0.68	**0.97**	0.52	0.48	**0.029**	0.52	**0.97**	0.48
KGGSeq	0.36	0.56	**0.98**	0.35	0.65	**0.017**	0.42	**0.98**	0.58
CONDEL	0.4	0.71	0.71	0.71	0.29	0.29	0.54	0.84	0.46
Cadd Phred	0.39	0.69	0.75	0.65	0.35	0.25	0.54	0.83	0.46

Table S1. Description of deleteriousness prediction scores and MMAF in training dataset

	Range		Total		TP		TN		p-value	ρ
	min	max	mean	sd	mean	sd	mean	sd		
SIFT	0	1	0.76	0.31	0.93	0.16	0.64	0.33	< 0.01	0.59
PolyPhen-2 HDIV	0	1	0.59	0.45	0.9	0.25	0.39	0.43	< 0.01	0.61
PolyPhen-2 HVAR	0	1	0.51	0.44	0.85	0.29	0.3	0.38	< 0.01	0.60
LRT	0	1	0.87	0.26	0.98	0.11	0.81	0.3	< 0.01	0.45
MutationTaster	0	1	0.64	0.41	0.92	0.24	0.46	0.41	< 0.01	0.54
Mutation Assessor	0	0.95	0.62	0.14	0.72	0.1	0.55	0.12	< 0.01	0.61
FATHMM	0	1	0.43	0.12	0.53	0.09	0.36	0.08	< 0.01	0.72
GERP++ RS	-12.3	6.17	2.9	3.42	4.6	1.84	1.81	3.74	< 0.01	0.45
PhyloP	-9.28	2.94	1.22	1.31	1.96	0.84	0.74	1.35	< 0.01	0.44
SiPhy	0	27.69	11.34	5.36	14.75	3.9	9.14	5.02	< 0.01	0.50
MMAF	0	1	0.12	0.24	0	0.02	0.2	0.28	< 0.01	-0.65

Range: range of all the predictors in training dataset. TP: summary statistics for True Positive observations in training dataset. TN: summary statistics for True Negative observations in training dataset. Total: Total number of mutations in each dataset (Total=TP+TN). sd: standard deviation. p-value: Bonferroni adjusted p-value for Wilcox rank sum test between TP group and TN group. ρ: Spearman correlation coefficient of the corresponding predictor with deleteriousness status (1: deleterious or TP, 0: neutral or TN). MMAF: maximum Minor Allele Frequency in various populations.

Table S2. Summary of logistic regression model built using training dataset

	Estimate	se	z value	p-value
Intercept	-11.36	0.20	-55.98	< 0.001
SIFT	0.19	0.11	1.71	0.09
Polyphen2_HDIV	-0.32	0.13	-2.40	0.02
Polyphen2_HVAR	1.31	0.14	9.64	< 0.001
LRT	-0.32	0.13	-2.43	0.02
MutationTaster	0.10	0.08	1.26	0.21
Mutation Assessor	4.30	0.26	16.52	< 0.001
FATHMM	17.23	0.27	63.87	< 0.001
GERP	0.08	0.02	4.72	< 0.001
PhyloP	-0.14	0.05	-2.93	0.003
SiPhy	0.05	0.01	7.07	< 0.001
MMAF	-36.41	1.35	-27.06	< 0.001

Estimate: estimated value of the corresponding regression coefficient. se: Standard error of the estimate of the coefficient. Intercept: intercept in the linear form of logistic regression model. Z value: Wald-test statistics that test the null hypothesis of corresponding estimated coefficient equals zero. P-value: p-value from Wald-test. MMAF: maximum Minor Allele Frequency in various population.

Table S3. Missing values for the whole exome mutations before and after BPCAfill imputation (%)

Name	Before(%)	After(%)
SIFT	2.53	2.53
Polyphen-2	18.94	2.53
LRT	24.03	2.53
MutationTaster	22.75	2.53
Mutation Assessor	16.67	2.47
FATHMM	2.53	2.53
GERP++	0.01	0.01
PhyloP	0.001	0.001
SiPhy	0.04	0.04

Before: the percentage of missing values for each prediction scores before imputation for whole exome mutations.
After: the percentage of missing values for each prediction scores after imputation for whole exome mutations.

Table S4. Additional testing datasets curated for comparing with Thusberg team

Dataset	Additional testing dataset I	Additional testing dataset I
TP	18,900	14,221
TN	20,812	17,150
Total	39,712	31,371
Source	VariBench dataset II[9,30]	Additional training dataset I without somatic cancer mutations[30]

TP: True Positive, number of deleterious mutations that were treated as true positive observations in modeling. TN: True Negative, number of non-deleterious mutations that were treated as true negative observations in modeling. Total: Total number of mutations for each dataset. (Total = TP+TN)

Table S5: Web Resources

Resource	URL
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/
SIFT	http://sift.jcvi.org/
MutationTaster	http://www.mutationtaster.org/
LRT	http://www.genetics.wustl.edu/jflab/lrt_query.html
PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp
PhD-SNP	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
SNAP	http://rostlab.org/services/snap/
MutPred	http://mutpred.mutdb.org/
SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go/
GERP++	http://mendel.stanford.edu/SidowLab/downloads/gerp/
PhyloP	http://compgen.bscb.cornell.edu/phast/; http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phyloP44way/
PON-P	http://bioinf.uta.fi/PON-P/
KGGSeq	http://statgenpro.psychiatry.hku.hk/limx/kggseq/
CONDEL	http://bg.upf.edu/condel/home
CADD	http://cadd.gs.washington.edu/
dbNSFP	http://sites.google.com/site/jpopgen/dbNSFP
ANNOVAR	http://www.openbioinformatics.org/annovar/
Uniprot	http://www.uniprot.org/
1000 Genomes Project	http://www.1000genomes.org/

Table S6: Stepwise model selection result

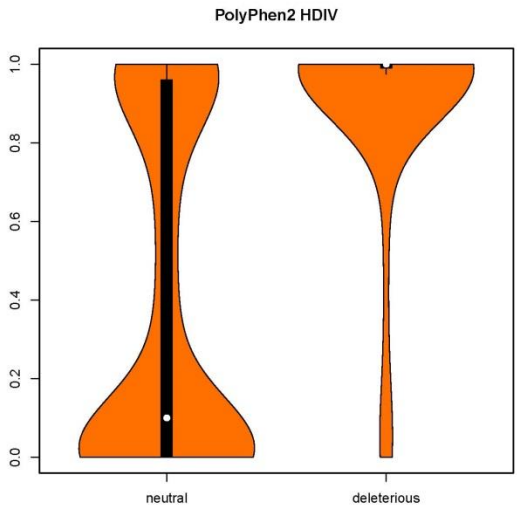
	Estimate	se	z value	p-value
Intercept	-11.40	0.20	-57.00	< 0.001
SIFT	0.19	0.11	1.71	0.09
Polyphen2_HDIV	-0.32	0.13	-2.43	0.02
Polyphen2_HVAR	1.31	0.13	9.81	< 0.001
LRT	-0.32	0.13	-2.17	0.03
Mutation Assessor	4.30	0.26	16.70	< 0.001
FATHMM	17.23	0.27	64.07	< 0.001
GERP	0.08	0.02	4.90	< 0.001
PhyloP	-0.14	0.05	-2.93	0.003
SiPhy	0.05	0.01	7.66	< 0.001
MMAF	-36.41	1.35	-27.13	< 0.001

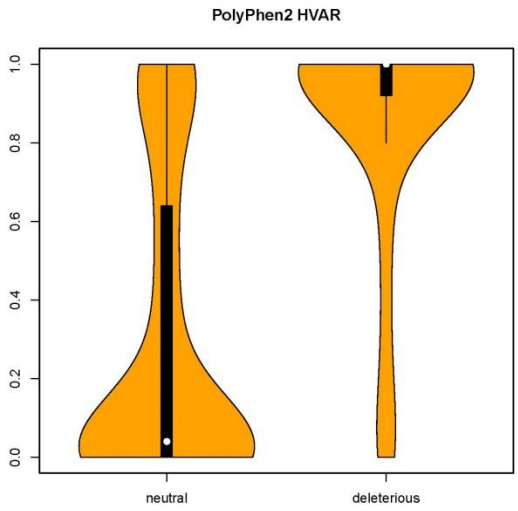
Estimate: estimated value of the corresponding regression coefficient. se: Standard error of the estimate of the coefficient. Intercept: intercept in the linear form of logistic regression model. Z value: Wald-test statistics that test the null hypothesis of corresponding estimated coefficient equals zero. P-value: p-value from Wald-test. MMAF: maximum Minor Allele Frequency in various population.

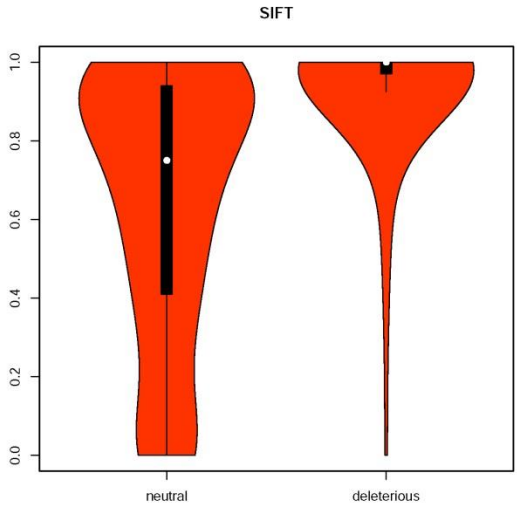
Table S7: Performance comparison between model generated from step-wise model selection and full model (final model) chosen in the current study

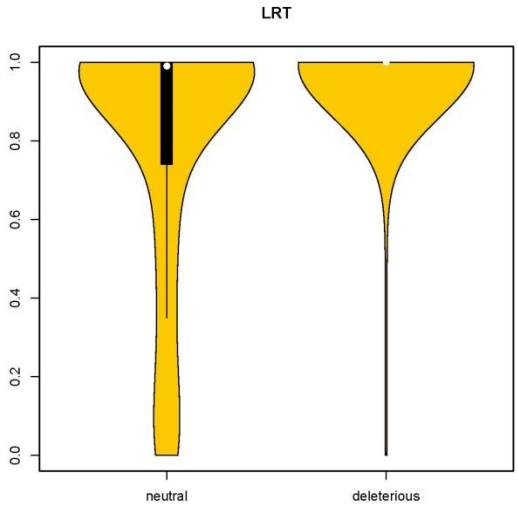
	AIC	Accuracy	AUC	95% CI for AUC
Full model	16484.51	0.91	0.97	0.968-0.971
Reduced model	16483.91	0.91	0.97	0.968-0.971

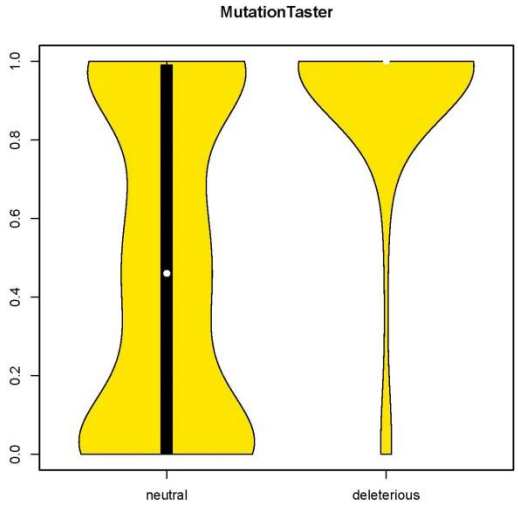
AIC: Akaike information criterion statistic. AUC: Area Under Curve value for Receiver Operating Characteristic (ROC) curve. 95% CI: 95% Confidence interval generated using bootstrap algorithm with 2000 bootstrap.

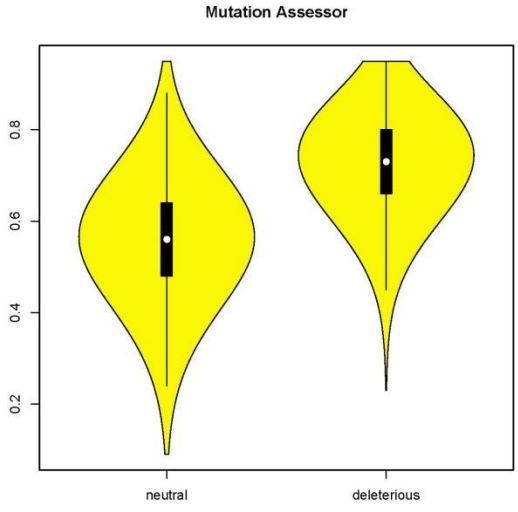

PolyPhen2 HDIV
0.0
0.2
0.4
0.6
0.8
1.0
neutral
deleterious

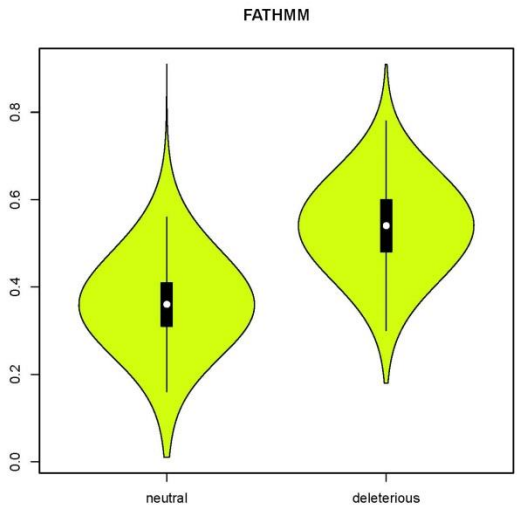

PolyPhen2 HVAR
0.0
0.2
0.4
0.6
0.8
1.0
neutral
deleterious

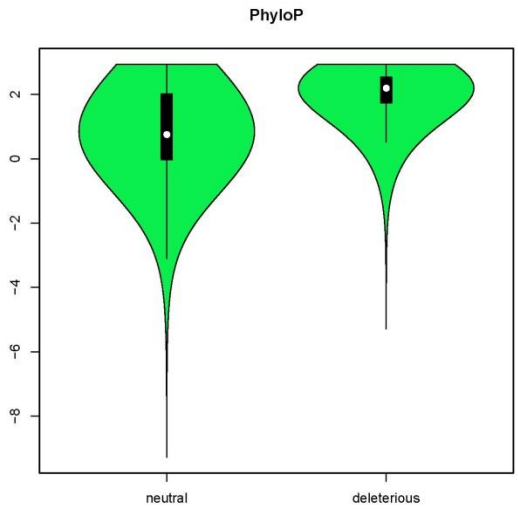

SIFT
0.0
0.2
0.4
0.6
0.8
1.0
neutral
deleterious

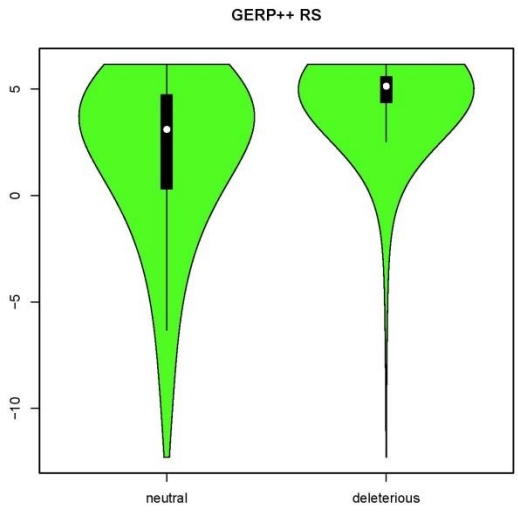

LRT
0.0
0.2
0.4
0.6
0.8
1.0
neutral
deleterious

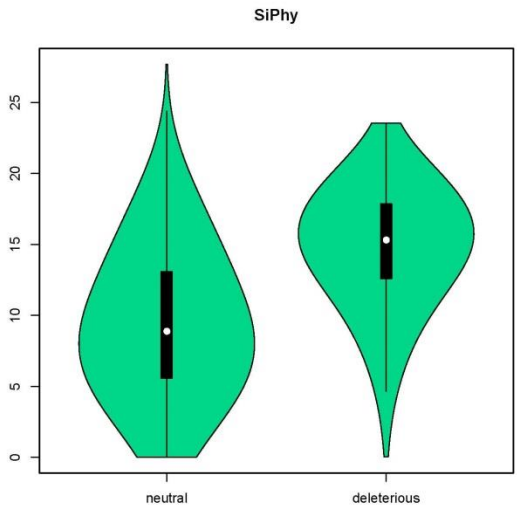

MutationTaster
0.0
0.2
0.4
0.6
0.8
1.0
neutral
deleterious


Mutation Assessor
0.2
0.4
0.6
0.8
neutral
deleterious


FATHMM
0.0
0.2
0.4
0.6
0.8
neutral
deleterious


PhyloP
−8
−6
−4
−2
0
2
neutral
deleterious


GERP++ RS
−10
−5
0
5
neutral
deleterious


SiPhy
0
5
10
15
20
25
neutral
deleterious

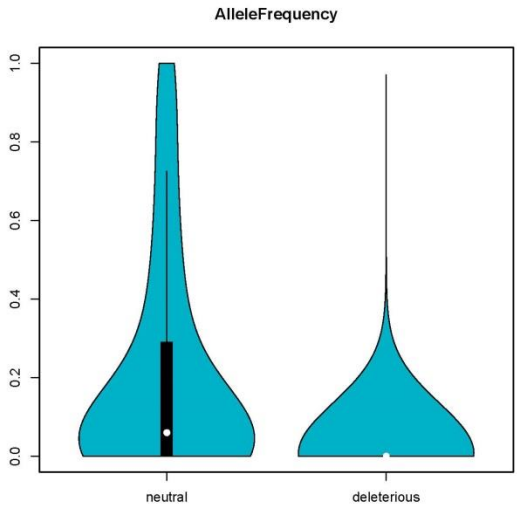

AlleleFrequency
0.0
0.2
0.4
0.6
0.8
1.0
neutral
deleterious

Figure S1. Violin plots of existing deleteriousness prediction scores in training dataset.
These violin plots illustrated the distribution of the deleteriousness prediction scores as well as MMAF in TP (deleterious) groups and TN (neutral) groups. The plot shows the median (indicated by the small white dot), the first through the third interquartile range (the thick, solid vertical band), and estimator of the density (shades with different colors) of the deleteriousness prediction scores/MMAF in each group.

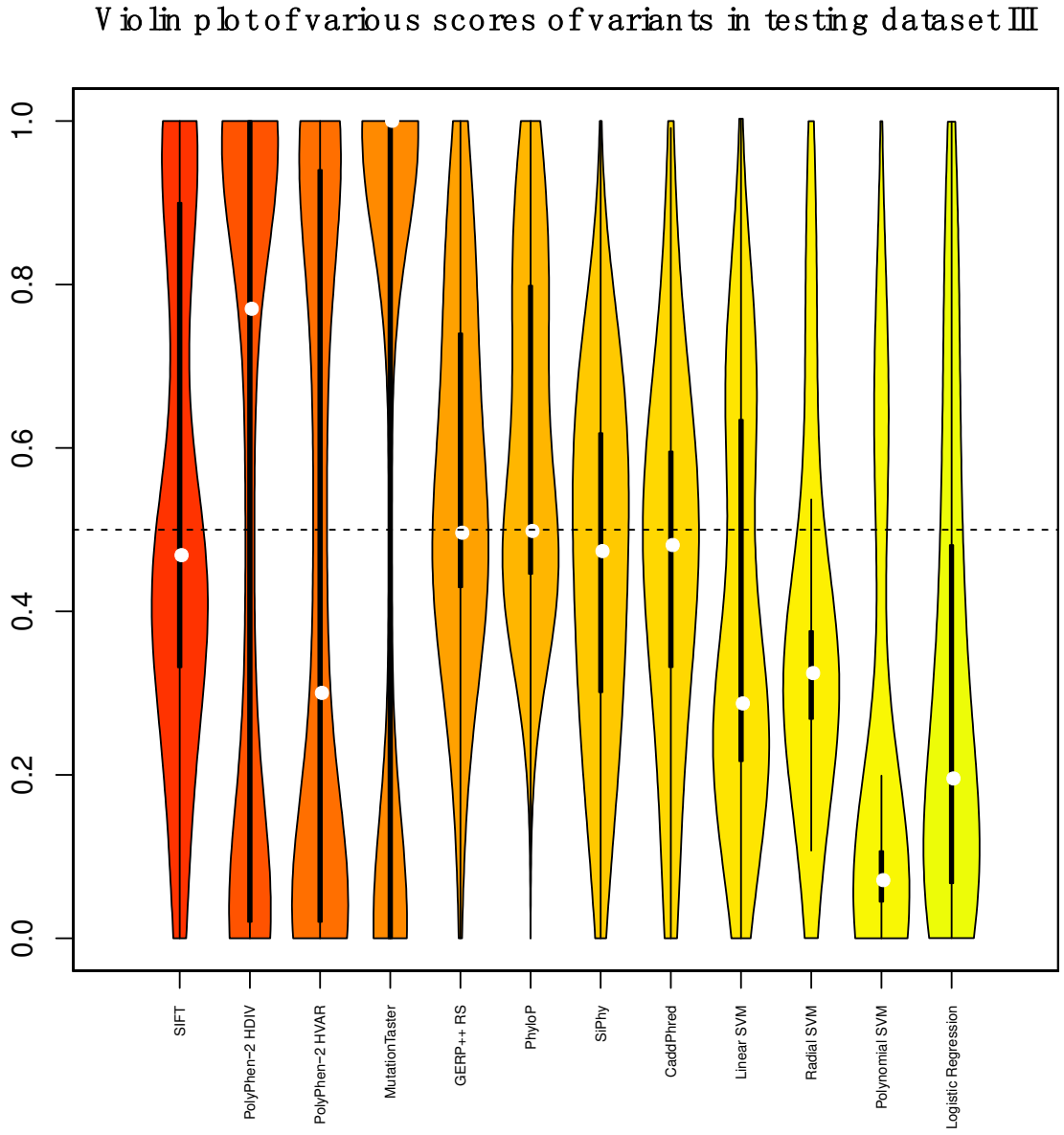
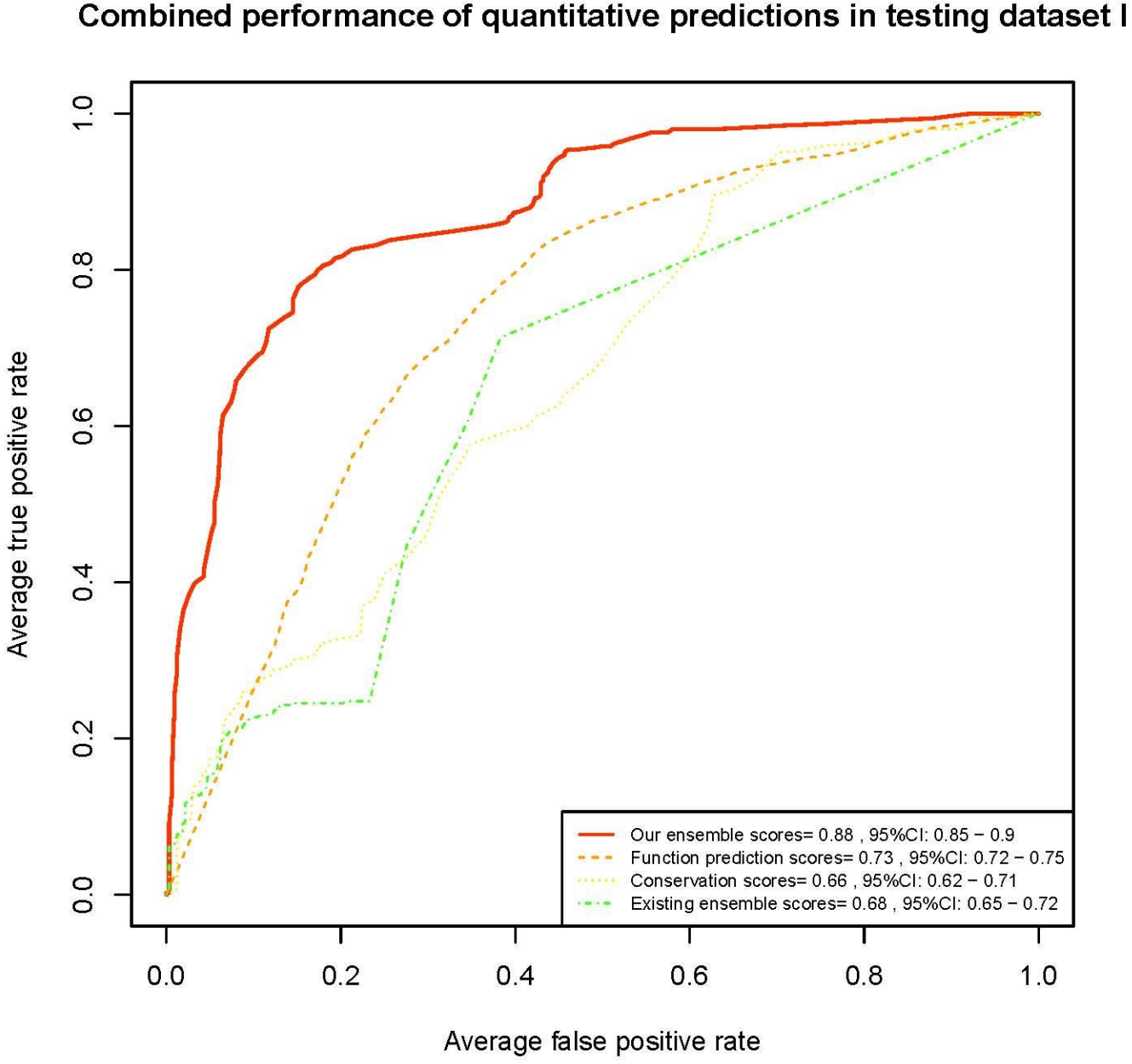


Figure S2. Violin plot of various normalized deleteriousness prediction scores in testing dataset III
This violin plot illustrated the distribution of the rescaled deleteriousness prediction scores from different deleteriousness prediction tools. Note that all prediction scores were rescaled into 0-1 scale with 0 being neutral, 1 being deleterious. Dotted black line at prediction score of 0.5 indicates the binary cutoff of deleteriousness. The plot shows the median (indicated by the small white dot), the first through the third interquartile range (the thick, solid vertical band), and estimator of the density (shades with different colors) of the deleteriousness prediction scores from each prediction tool.


Combined performance of quantitative predictions in testing dataset I
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Our ensemble scores= 0.88 , 95%CI: 0.85 − 0.9
Function prediction scores= 0.73 , 95%CI: 0.72 − 0.75
Conservation scores= 0.66 , 95%CI: 0.62 − 0.71
Existing ensemble scores= 0.68 , 95%CI: 0.65 − 0.72

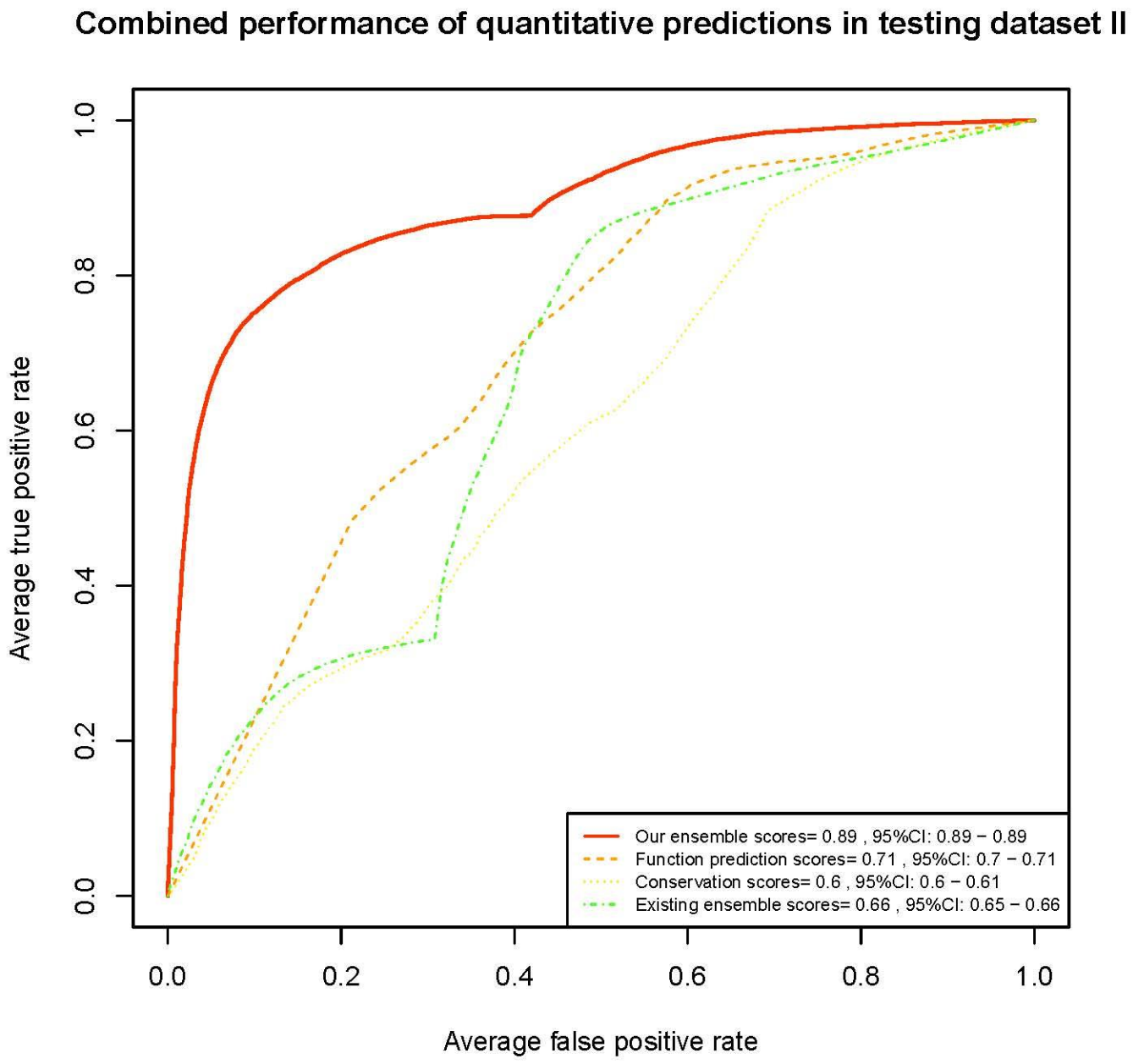


Figure S3. Combined ROC plot of three categories of deleteriousness prediction scores in testing dataset I and II

These two plots illustrated performance of quantitative prediction outcomes for combined deleteriousness prediction scores evaluated by receiver operating characteristics (ROC) curve and area under curve (AUC) score for the ROC curve. Deleteriousness prediction scores were combined into four categories and ROC curves were averaged among all prediction scores from the same category Higher AUC score indicates better performance. Top used testing dataset I as benchmark dataset and bottom plot used testing dataset II as benchmark dataset (see Table 1). 95% CI indicates 95% confidence interval computed with 2000 stratified bootstrap replicates for each category.

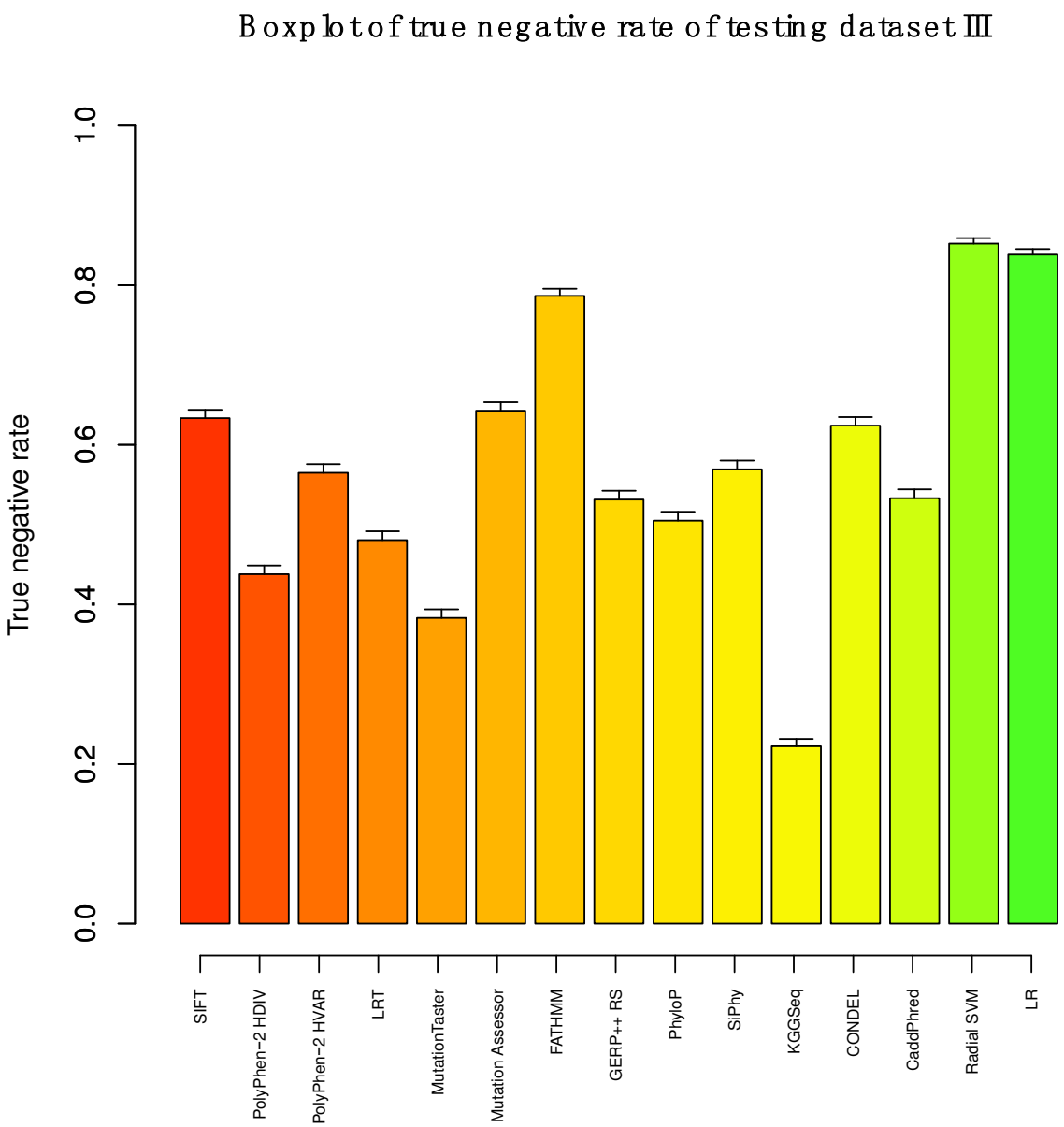
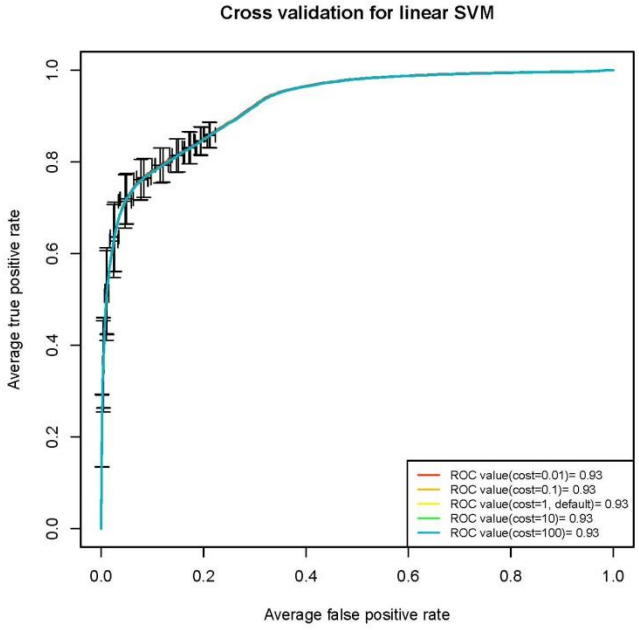
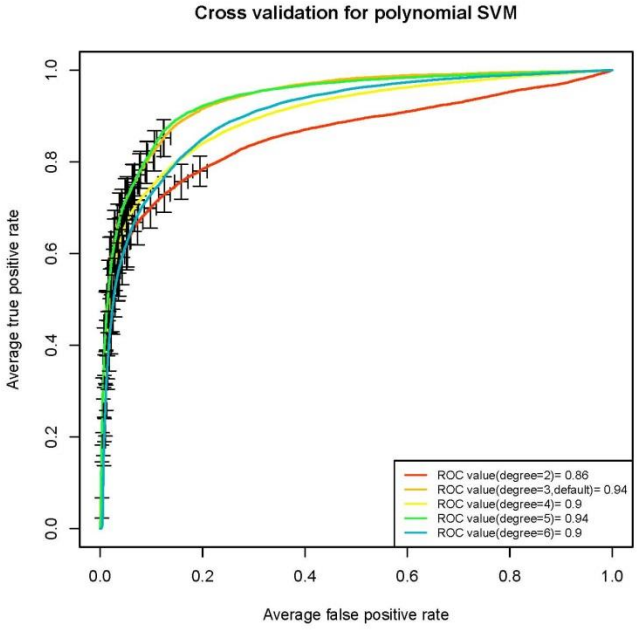
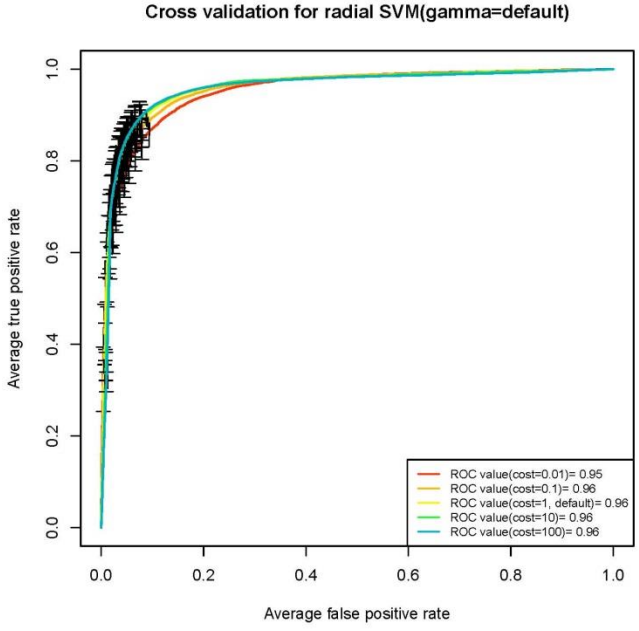


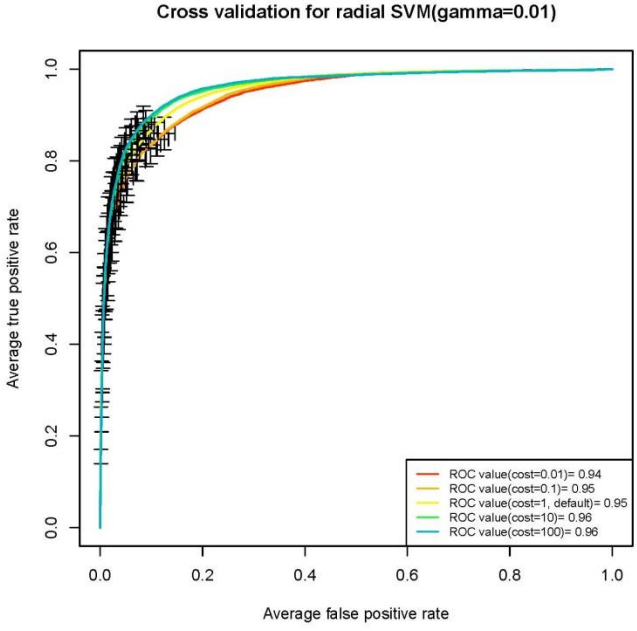
Figure S4. Boxplot of TNR of various deleteriousness prediction scores for testing dataset III

This plot illustrated the different TNR of various deleteriousness prediction scores. Different colors indicated different deleteriousness prediction scores. Error bar indicates standard error.


Cross validation for linear SVM
Average true positive rate
Average false positive rate
ROC value(cost=0.01)= 0.93
ROC value(cost=0.1)= 0.93
ROC value(cost=1, default)= 0.93
ROC value(cost=10)= 0.93
ROC value(cost=100)= 0.93


Cross validation for polynomial SVM
Average true positive rate
Average false positive rate
ROC value(degree=2)= 0.86
ROC value(degree=3,default)= 0.94
ROC value(degree=4)= 0.9
ROC value(degree=5)= 0.94
ROC value(degree=6)= 0.9


Cross validation for radial SVM(gamma=default)
Average true positive rate
Average false positive rate
ROC value(cost=0.01)= 0.95
ROC value(cost=0.1)= 0.96
ROC value(cost=1, default)= 0.96
ROC value(cost=10)= 0.96
ROC value(cost=100)= 0.96


Cross validation for radial SVM(gamma=0.01)
Average true positive rate
Average false positive rate
ROC value(cost=0.01)= 0.94
ROC value(cost=0.1)= 0.95
ROC value(cost=1, default)= 0.95
ROC value(cost=10)= 0.96
ROC value(cost=100)= 0.96

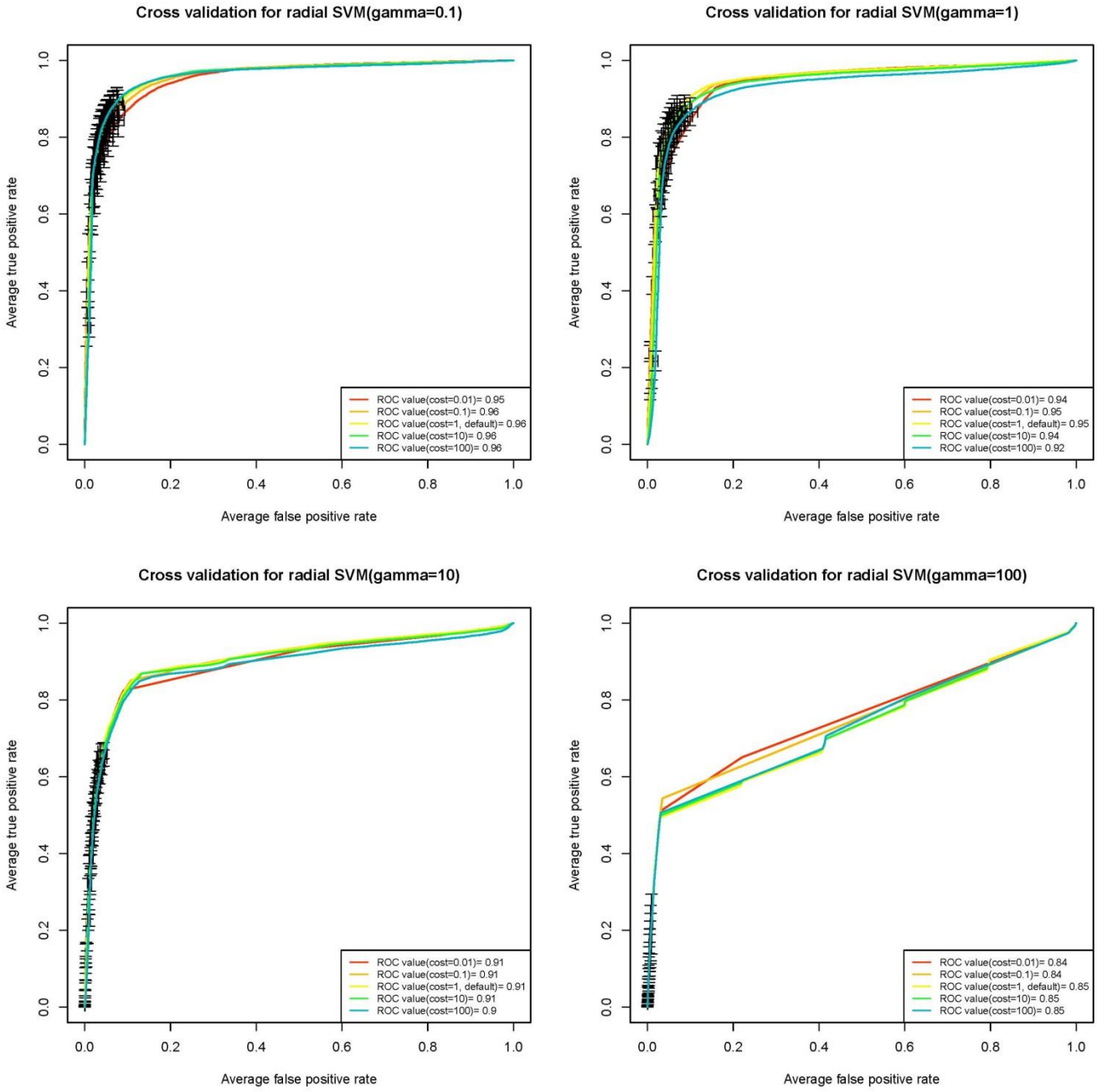


Figure S5. Five-fold cross validation for evaluating different SVM parameters cocktails on training dataset

These eight plots demonstrated that performance of SVM and LR models with different parameter cocktails was as good as models with default parameter settings. For linear SVM, cost of 0.01, 0.1, 1 (default), 10 and 100 were evaluated for performance (top left plot). For polynomial SVM, degree of 2, 3 (default), 4, 5 and 6 with default parameter for cost was analyzed (top right plot). For radial SVM, different combinations of gamma of 0.01, 0.1, 1, 10, 100 and 1/11 (default) and cost of 0.01, 0.1, 1 (default), 10 and 100 were evaluated (remaining plots). Error bar indicates horizontal and vertical variance from five-fold cross validation. ROC curve and AUC values were averaged from five-fold cross validation.

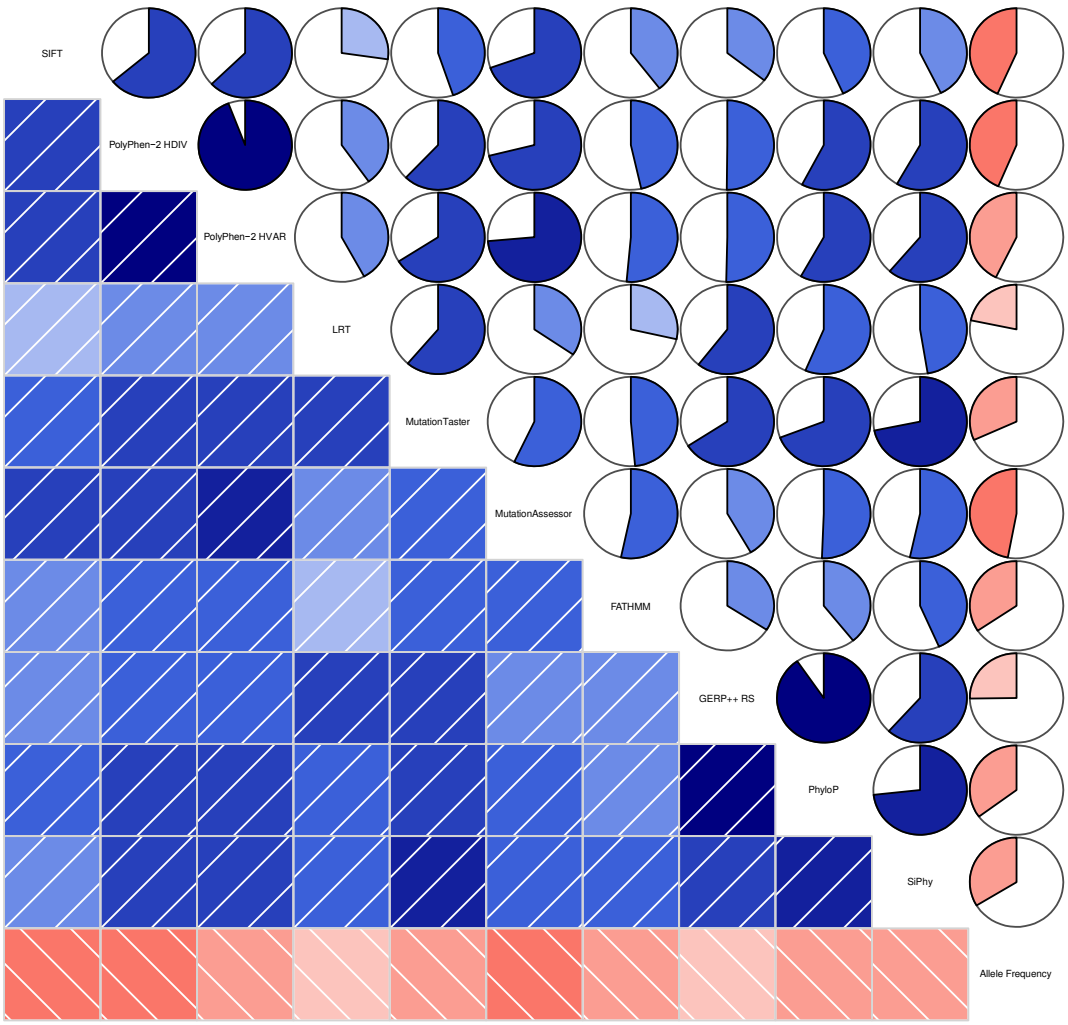
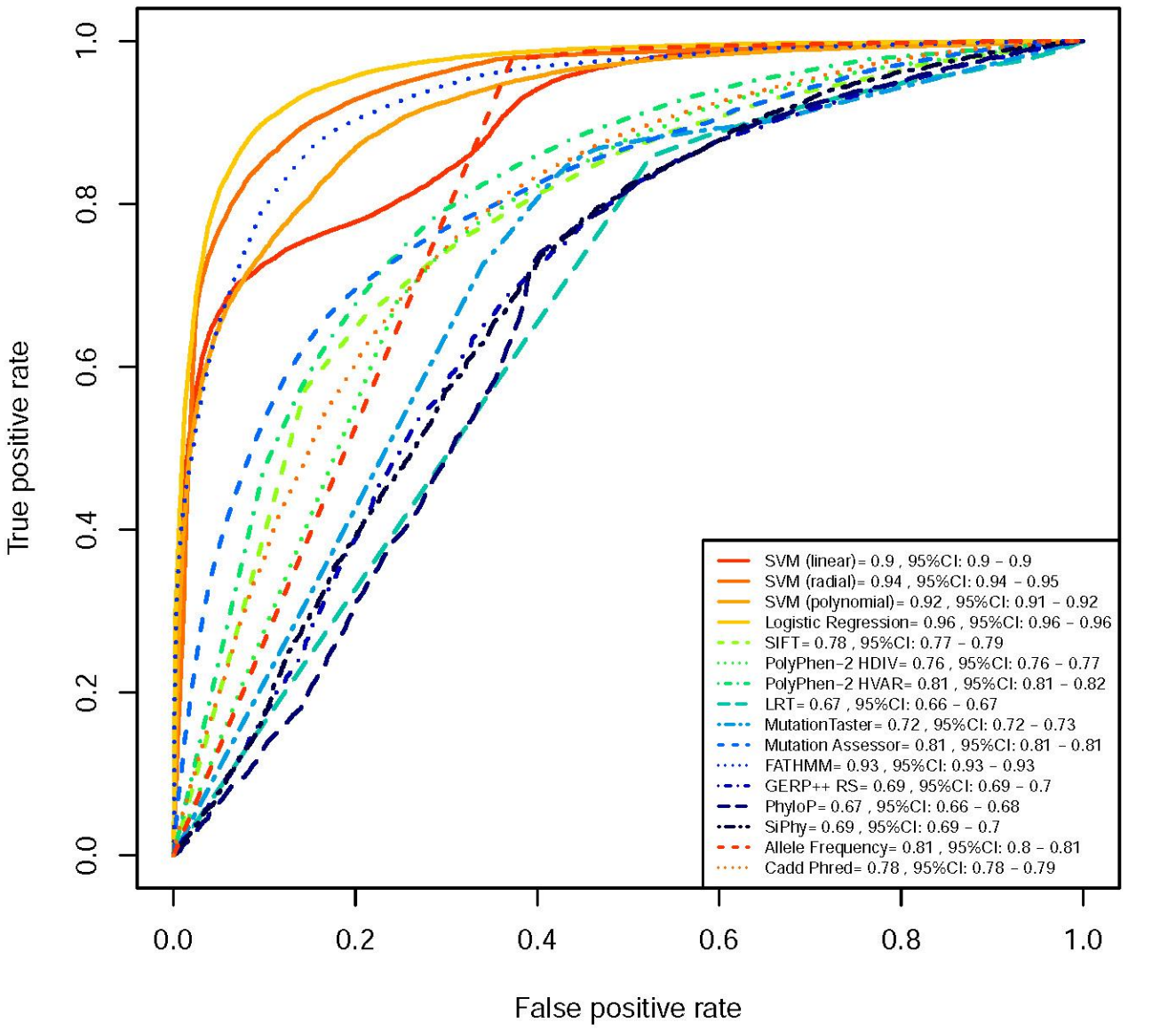


Figure S6. Correlation plot of nine existing deleteriousness prediction scores and MMAF in training dataset

This plot illustrated the pair-wise Pearson correlation of nine deleteriousness prediction scores and MMAF in training dataset. Blue color indicated positive correlation between two scores. Red color indicated negative correlation between two scores. Colored pie charts on the upper right indicated the level correlation. The larger the shade, the higher the correlation.


Performance of quantitative predictions in additional testing dataset I
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (linear)= 0.9 , 95%CI: 0.9 – 0.9
SVM (radial)= 0.94 , 95%CI: 0.94 – 0.95
SVM (polynomial)= 0.92 , 95%CI: 0.91 – 0.92
Logistic Regression= 0.96 , 95%CI: 0.96 – 0.96
SIFT= 0.78 , 95%CI: 0.77 – 0.79
PolyPhen-2 HDIV= 0.76 , 95%CI: 0.76 – 0.77
PolyPhen-2 HVAR= 0.81 , 95%CI: 0.81 – 0.82
LRT= 0.67 , 95%CI: 0.66 – 0.67
MutationTaster= 0.72 , 95%CI: 0.72 – 0.73
Mutation Assessor= 0.81 , 95%CI: 0.81 – 0.81
FATHMM= 0.93 , 95%CI: 0.93 – 0.93
GERP++ RS= 0.69 , 95%CI: 0.69 – 0.7
PhyloP= 0.67 , 95%CI: 0.66 – 0.68
SiPhy= 0.69 , 95%CI: 0.69 – 0.7
Allele Frequency= 0.81 , 95%CI: 0.8 – 0.81
Cadd Phred= 0.78 , 95%CI: 0.78 – 0.79

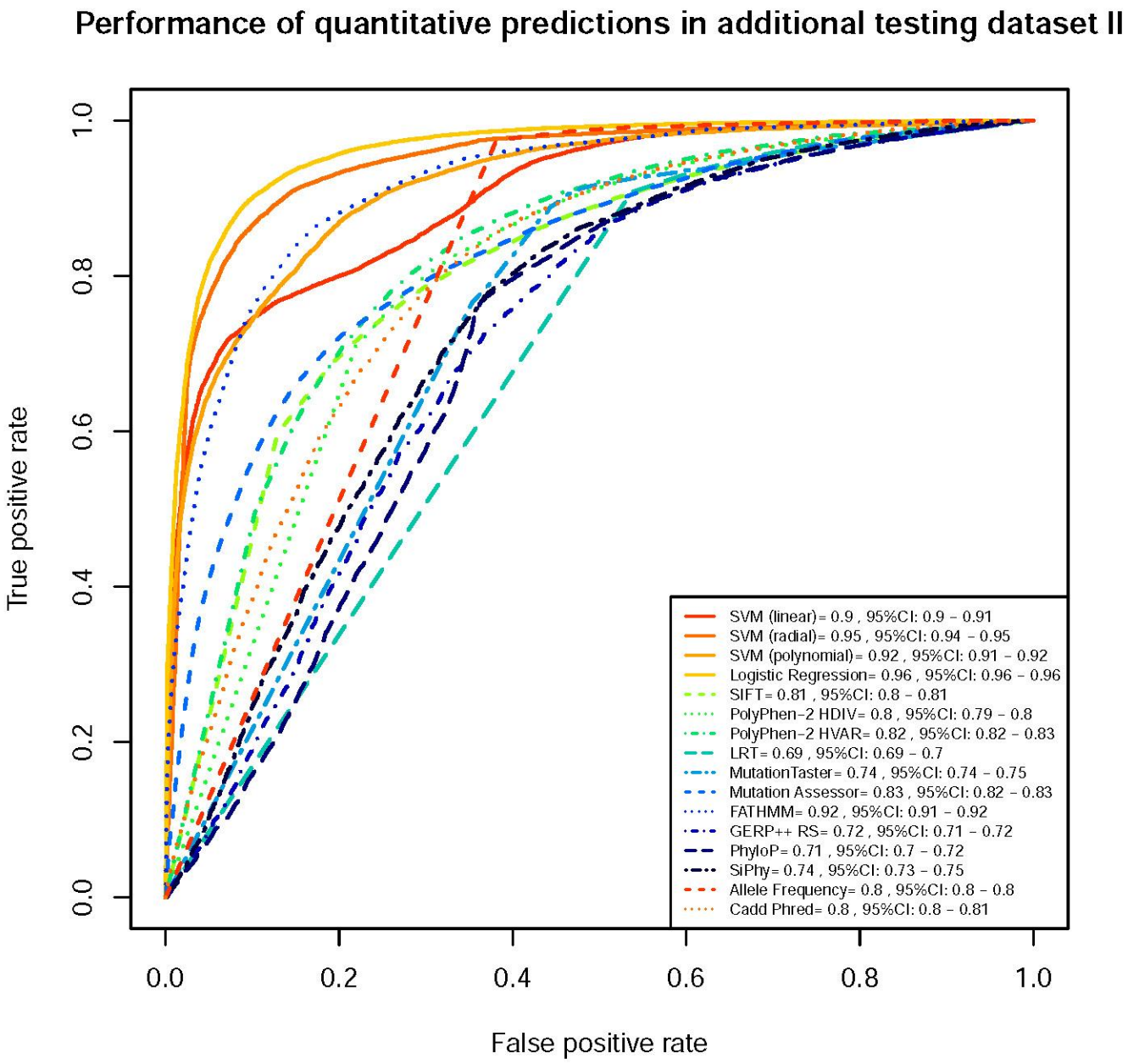
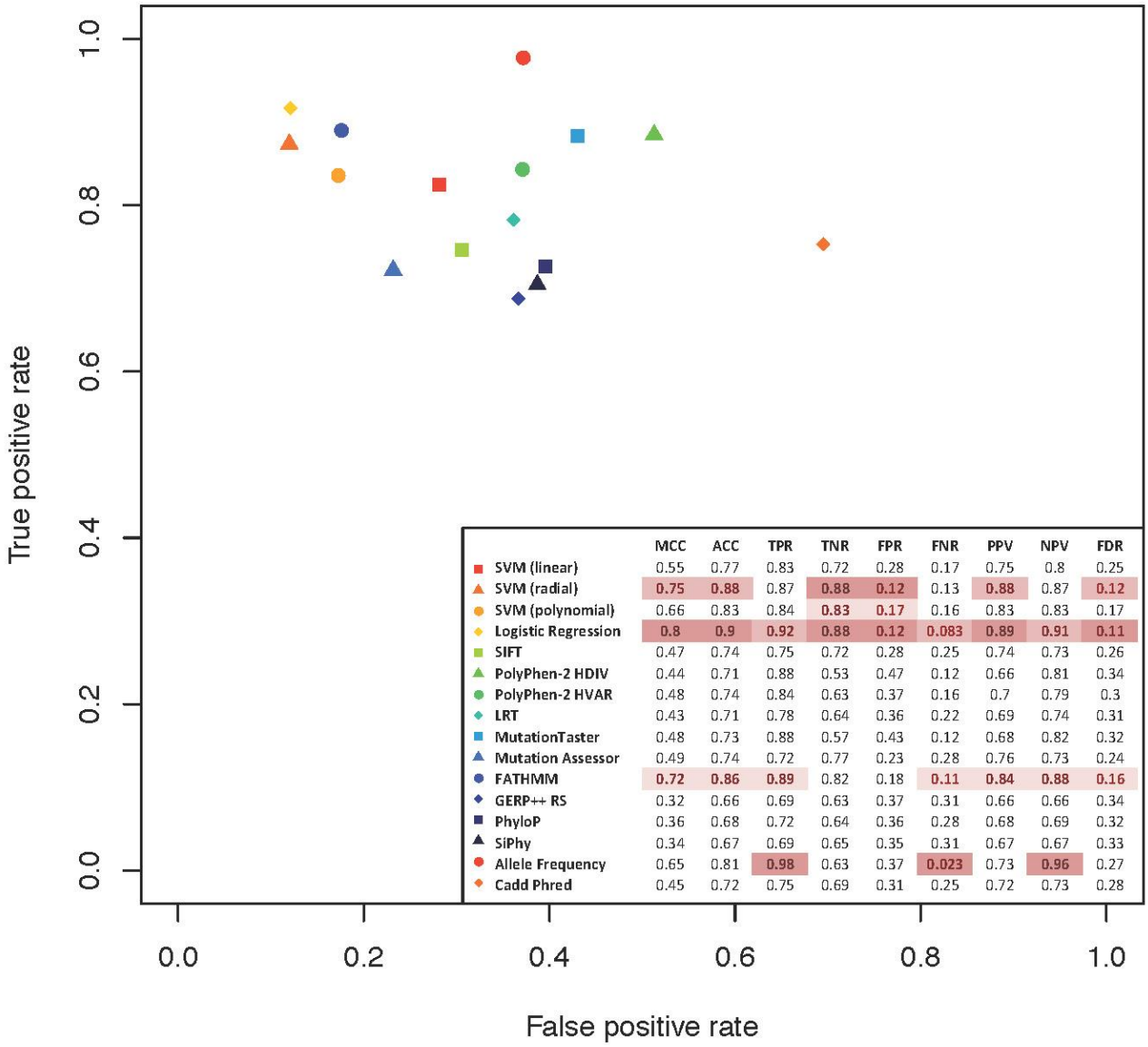


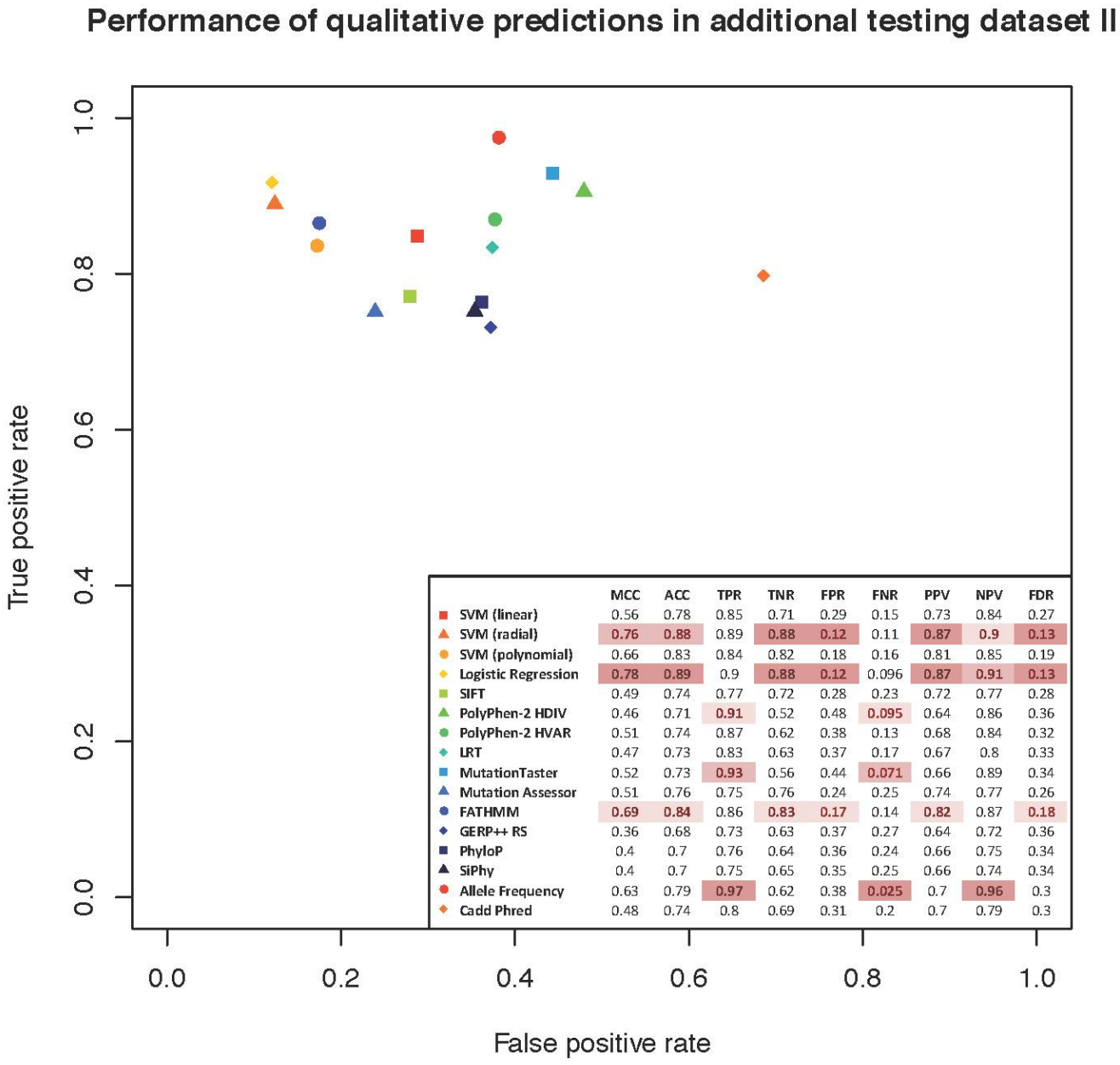
Figure S7. ROC curves for individual prediction scores and ensemble scores generated using training dataset (evaluated on additional testing dataset I and II)

These plots illustrated performance of quantitative prediction outcomes evaluated on two datasets (additional testing dataset I and II). Performance was evaluated by ROC curve and by AUC score for the ROC curve. Higher score indicates better performance. For the top plot, additional testing dataset I was used for performance evaluation; for the bottom plot, testing dataset II was used for evaluation.

Performance of qualitative predictions in additional testing dataset I



	MCC	ACC	TPR	TNR	FPR	FNR	PPV	NPV	FDR
SVM (linear)	0.55	0.77	0.83	0.72	0.28	0.17	0.75	0.8	0.25
SVM (radial)	0.75	0.88	0.87	0.88	0.12	0.13	0.88	0.87	0.12
SVM (polynomial)	0.66	0.83	0.84	0.83	0.17	0.16	0.83	0.83	0.17
Logistic Regression	0.8	0.9	0.92	0.88	0.12	0.083	0.89	0.91	0.11
SIFT	0.47	0.74	0.75	0.72	0.28	0.25	0.74	0.73	0.26
PolyPhen-2 HDIV	0.44	0.71	0.88	0.53	0.47	0.12	0.66	0.81	0.34
PolyPhen-2 HVAR	0.48	0.74	0.84	0.63	0.37	0.16	0.7	0.79	0.3
LRT	0.43	0.71	0.78	0.64	0.36	0.22	0.69	0.74	0.31
MutationTaster	0.48	0.73	0.88	0.57	0.43	0.12	0.68	0.82	0.32
Mutation Assessor	0.49	0.74	0.72	0.77	0.23	0.28	0.76	0.73	0.24
FATHMM	0.72	0.86	0.89	0.82	0.18	0.11	0.84	0.88	0.16
GERP++ RS	0.32	0.66	0.69	0.63	0.37	0.31	0.66	0.66	0.34
PhyloP	0.36	0.68	0.72	0.64	0.36	0.28	0.68	0.69	0.32
SiPhy	0.34	0.67	0.69	0.65	0.35	0.31	0.67	0.67	0.33
Allele Frequency	0.65	0.81	0.98	0.63	0.37	0.023	0.73	0.96	0.27
Cadd Phred	0.45	0.72	0.75	0.69	0.31	0.25	0.72	0.73	0.28

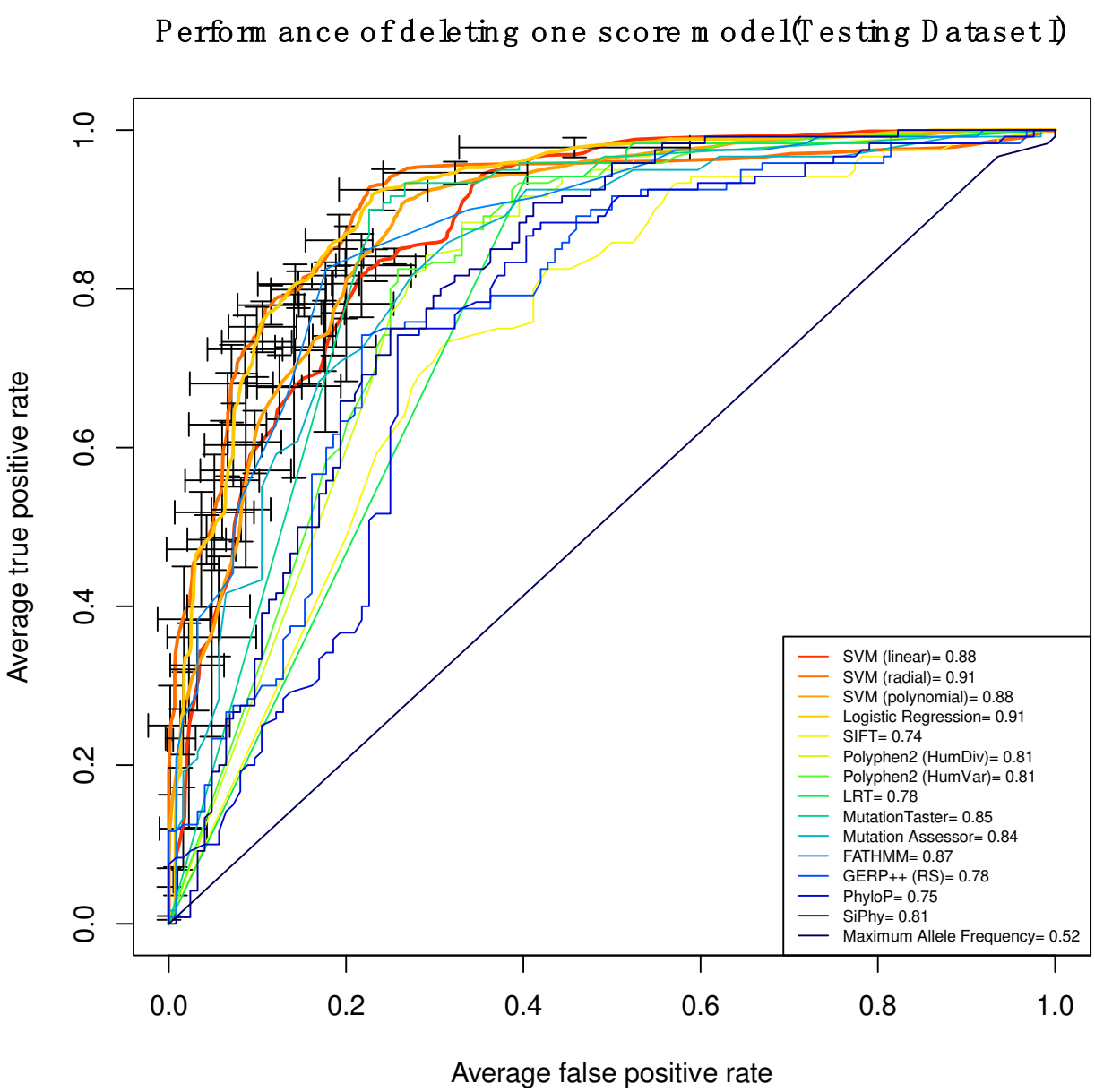


	MCC	ACC	TPR	TNR	FPR	FNR	PPV	NPV	FDR
SVM (linear)	0.56	0.78	0.85	0.71	0.29	0.15	0.73	0.84	0.27
SVM (radial)	0.76	0.88	0.89	0.88	0.12	0.11	0.87	0.9	0.13
SVM (polynomial)	0.66	0.83	0.84	0.82	0.18	0.16	0.81	0.85	0.19
Logistic Regression	0.78	0.89	0.9	0.88	0.12	0.096	0.87	0.91	0.13
SIFT	0.49	0.74	0.77	0.72	0.28	0.23	0.72	0.77	0.28
PolyPhen-2 HDIV	0.46	0.71	0.91	0.52	0.48	0.095	0.64	0.86	0.36
PolyPhen-2 HVAR	0.51	0.74	0.87	0.62	0.38	0.13	0.68	0.84	0.32
LRT	0.47	0.73	0.83	0.63	0.37	0.17	0.67	0.8	0.33
MutationTaster	0.52	0.73	0.93	0.56	0.44	0.071	0.66	0.89	0.34
Mutation Assessor	0.51	0.76	0.75	0.76	0.24	0.25	0.74	0.77	0.26
FATHMM	0.69	0.84	0.86	0.83	0.17	0.14	0.82	0.87	0.18
GERP++ RS	0.36	0.68	0.73	0.63	0.37	0.27	0.64	0.72	0.36
PhyloP	0.4	0.7	0.76	0.64	0.36	0.24	0.66	0.75	0.34
SiPhy	0.4	0.7	0.75	0.65	0.35	0.25	0.66	0.74	0.34
Allele Frequency	0.63	0.79	0.97	0.62	0.38	0.025	0.7	0.96	0.3
Cadd Phred	0.48	0.74	0.8	0.69	0.31	0.2	0.7	0.79	0.3

Figure S8. Sensitivity and specificity plots for existing prediction scores and our ensemble scores

These two plots illustrated the performance of qualitative prediction outcomes of existing prediction scores and our ensemble prediction scores, evaluated by sensitivity and specificity. Higher sensitivity/specificity score indicates better performance. Top plot used additional testing dataset I as benchmark dataset and bottom plot used additional testing dataset II as benchmark dataset (see Table 1). Legend table showed various qualitative prediction performance measurements for each prediction tool. Matthews Correlation Coefficient (MCC) is a correlation coefficient between the observed and predicted binary classification, ranging from -1 to 1, where 1 indicates perfect prediction, -1 indicates total disagreement between prediction and observation. $\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$, where TP, TN, FP, FN denotes true positive, true negative, false positive and false negative respectively. ACC denotes accuracy. $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$. TPR

denotes true positive rate, or sensitivity. $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. TNR denotes true negative rate, or specificity. $\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$. FPR denotes false positive rate. $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}}$. FNR denotes false negative rate. $\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}}$. PPV denotes positive predictive value. $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$. NPV denotes negative predictive value. $\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN+FN}}$. FDR denotes false discovery rate. $\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP+TP}}$. For each qualitative prediction performance measurement, top three performance scores were highlighted. The brighter the highlight color, the better the performance.


Performance of deleting one score model (Testing Dataset I)
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (linear)= 0.88
SVM (radial)= 0.91
SVM (polynomial)= 0.88
Logistic Regression= 0.91
SIFT= 0.74
Polyphen2 (HumDiv)= 0.81
Polyphen2 (HumVar)= 0.81
LRT= 0.78
MutationTaster= 0.85
Mutation Assessor= 0.84
FATHMM= 0.87
GERP++ (RS)= 0.78
PhyloP= 0.75
SiPhy= 0.81
Maximum Allele Frequency= 0.52

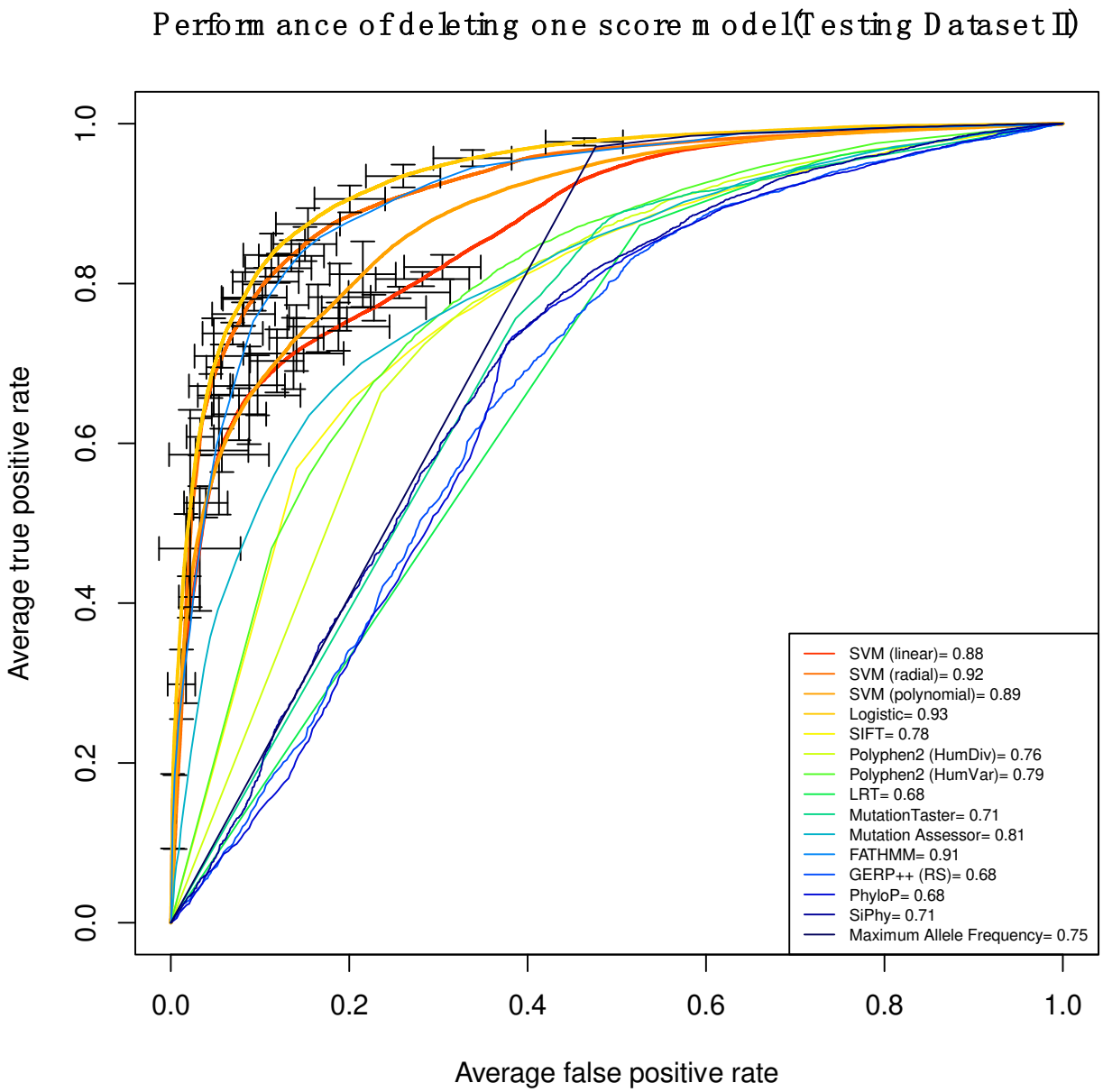


Figure S9. Performance evaluation of delete-one-score model on testing dataset I and II

These two plots demonstrated that performance of SVM and LR models with each prediction score deleted from the full model was as good as the full models with all prediction scores. Top plot used testing dataset I as benchmark dataset and bottom plot used testing dataset II as benchmark dataset (see Table 1). Error bar indicates horizontal and vertical variance from deleting-one-score models. ROC curve and AUC values were averaged from deleting every prediction score.

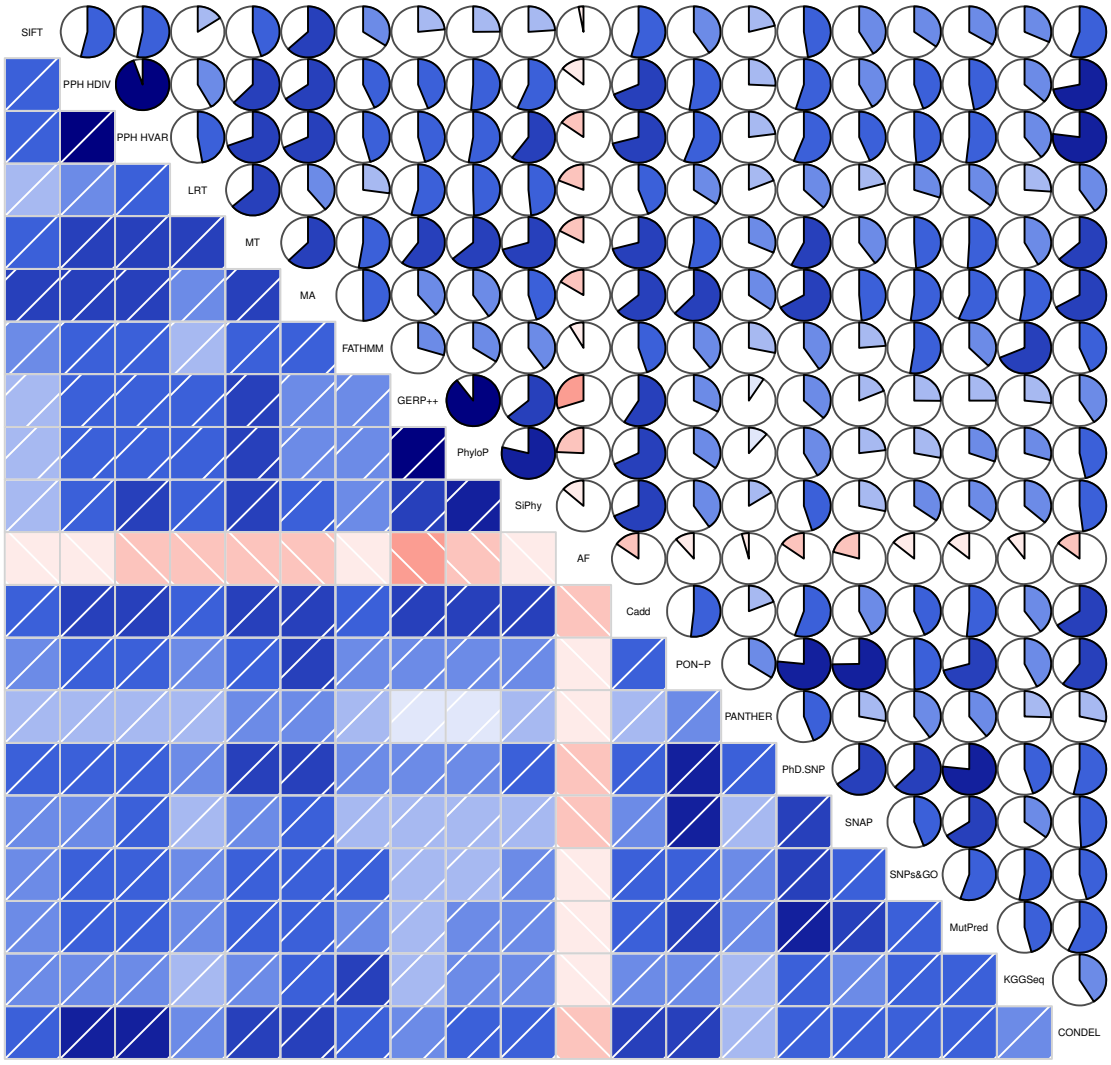


Figure S10. Correlation plot of nine existing deleteriousness prediction scores and MMAF in testing dataset I

This plot illustrated the pair-wise Pearson correlation of nine deleteriousness prediction scores and MMAF in testing dataset I. Blue color indicated positive correlation between two scores. Red color indicated negative correlation between two scores. Colored pie charts on the upper right indicated the level correlation. The larger the shade, the higher the correlation.