*Briefings in Bioinformatics*, 00(00), 2019, 1–12

**doi: 10.1093/bib/bbz047**
Advance Access Publication Date: 03 June 2019
Review article



# Comparison and integration of computational methods for deleterious synonymous mutation prediction

Na Cheng, Menglu Li, Le Zhao, Bo Zhang, Yuhua Yang, Chun-Hou Zheng and Junfeng Xia

Corresponding author: Junfeng Xia, Institutes of Physical Science and Information Technology, Anhui University, Hefei, Anhui 230601, China. Tel.: +86-551-63861990; E-mail: jfxia@ahu.edu.cn

## Abstract

Synonymous mutations do not change the encoded amino acids but may alter the structure or function of an mRNA in ways that impact gene function. Advances in next generation sequencing technologies have detected numerous synonymous mutations in the human genome. Several computational models have been proposed to predict deleterious synonymous mutations, which have greatly facilitated the development of this important field. Consequently, there is an urgent need to assess the state-of-the-art computational methods for deleterious synonymous mutation prediction to further advance the existing methodologies and to improve performance. In this regard, we systematically compared a total of 10 computational methods (including specific method for deleterious synonymous mutation and general method for single nucleotide mutation) in terms of the algorithms used, calculated features, performance evaluation and software usability. In addition, we constructed two carefully curated independent test datasets and accordingly assessed the robustness and scalability of these different computational methods for the identification of deleterious synonymous mutations. In an effort to improve predictive performance, we established an ensemble model, named Prediction of Deleterious Synonymous Mutation (PrDSM), which averages the ratings generated by the three most accurate predictors. Our benchmark tests demonstrated that the ensemble model PrDSM outperformed the reviewed tools for the prediction of deleterious synonymous mutations. Using the ensemble model, we developed an accessible online predictor, PrDSM, available at http://bioinfo.ahu.edu.cn:8080/PrDSM/. We hope that this comprehensive survey and the proposed strategy for building more accurate models can serve as a useful guide for inspiring future developments of computational methods for deleterious synonymous mutation prediction.

**Key words:** deleterious synonymous mutation; prediction model; ensemble learning; machine learning

**Na Cheng** is a PhD candidate at the Institutes of Physical Science and Information Technology, Anhui University. Her research interests include development of databases and bioinformatics tools and cancer genomics.
**Menglu Li** is a Master's student at the School of Computer Science and Technology, Anhui University.
**Le Zhao** is a Master's student at the School of Computer Science and Technology, Anhui University.
**Bo Zhang** is a Master's student at the School of Computer Science and Technology, Anhui University.
**Yuhua Yang** is an undergraduate student at the School of Computer Science and Technology, Anhui University.
**Chun-Hou Zheng** is a professor at the School of Computer Science and Technology, Anhui University. His research interests include development of bioinformatics algorithms and cancer genomics.
**Junfeng Xia** is a professor at the Institutes of Physical Science and Information Technology, Anhui University. His research interests include development of databases and bioinformatics algorithms and cancer genomics.
**Submitted:** 30 January 2019; **Received (in revised form):** 28 March 2019

## Introduction

Advances in next generation sequencing technologies have detected numerous synonymous mutations in the human genome that do not alter amino acids. In recent years, much work has pointed out the important role of synonymous mutations in many human diseases [1–7], including psychiatric disease, congenital heart disease and cancer [8–16]. However, it is difficult to distinguish disease associated synonymous mutations from benign ones. Experimental characterization of all identified synonymous mutations is not practical and is usually time-consuming, costly and labour intensive. To meet this need, several excellent bioinformatics platforms and tools have been proposed to support prioritization of synonymous mutations [17–27]. These tools can be divided into two categories, i.e. specific tools for synonymous mutation pathogenicity prediction and general tools for single nucleotide variant pathogenicity prediction. The former is specifically designed for predicting the functional consequences of synonymous mutations, including SilVA [17], DDIG-SN [18], regSNPs-splicing [19], Syntool [20] and TraP [21], while the latter commonly used methods can predict the effect of single nucleotide mutations, including but not limited to, synonymous mutations, such as CADD [22], DANN [23], FATHMM-MKL [24], PredictSNP2 [25] and PhD-SNP$^{g}$ [26]. While all serve the purpose of synonymous mutation pathogenicity prediction in general, the increasing method variations among these tools, in combination with the emerging new types of experimental data, render it necessary to rationally select the best approach, especially for the potential impact on genomics-based precision medicine.

Given the growing number of studies on computational prediction of single nucleotide mutation pathogenicity, several review papers have been published to comprehensively investigate these methods [25, 28–31]. However, no studies have focused on synonymous mutation pathogenicity prediction. In this article, we provide a comprehensive survey of the most up-to-date progress in large-scale computational studies on predicting the functional impact of synonymous mutation. In total, 10 computational methods published to date (including 5 specific tools for synonymous mutation and 5 general tools for single nucleotide mutation) were critically assessed, systematically benchmarked and thoroughly discussed in terms of algorithm construction, heterogeneous features extracted, performance evaluation strategy and software utility. Most importantly, we performed extensive independent tests to objectively assess the prediction performance of the reviewed computational approaches based on two newly constructed independent test datasets. Based on the performance evaluation of the current methods for deleterious synonymous mutation prediction, three methods, including TraP, SilVA and FATHMM-MKL, showed the best overall performance and low correlations with each other. Therefore, we proposed an ensemble model, **Pr**ediction of **D**eleterious **S**ynonymous **M**utation (PrDSM), by integrating the output of these three methods in this study. Experiments on the two benchmark datasets illustrated that our ensemble predictor PrDSM could outperform all the individual tools. We anticipate that our review will aid the future development of computational methods for efficient and accurate synonymous mutation pathogenicity prediction and that the proposed ensemble model will complement existing methods.

## Materials and methods

Figure 1 illustrates the framework of the proposed predictor methodology for predicting deleterious synonymous mutations: collection of deleterious and benign mutation set, performance evaluation and calculation of consensus score (design of ensemble predictor). The major procedures are described in the following sections.

### Deleteriousness prediction methods

We compared 10 deleteriousness predictive methods (Table 1) that can be classified into two types, including 5 specific tools for synonymous mutations: SilVA [17], DDIG-SN [18], regSNPs-splicing (regSNP) [19], TraP [21] and Syntool [20], and 5 general tools that can evaluate the pathogenicity of single nucleotide mutations: CADD [22], DANN [23], FATHMM-MKL [24], PredictSNP2 [25] and PhD-SNP$^{g}$ [26]. Notably, the cut-offs used to distinguish pathogenic or benign mutations were sourced from the original research and the study reported by Li *et al.* [30]. We obtained the deleterious score for each synonymous mutation for 10 methods by running their stand-alone programs or publicly available web servers.

### Independent test datasets construction

To evaluate the performance of the 10 methods, it is essential to construct independent test datasets in which overlapping mutations with the training data are removed in the compared methods as much as possible. We reviewed the articles or websites of 10 methods and discovered that the vast majority of the algorithms were trained on datasets from HGMD, ClinVar and 1000 Genome Project. In this study, the deleterious synonymous mutations (positive dataset) were derived from HGMD Professional version 2018.3 [32], and the putatively benign synonymous mutations (negative dataset) were retrieved from the VariSNP database version 2017-02-16 [33]. For the positive dataset, we removed 1404 mutations that appeared in HGMD before 2017 and employed only the mutations labelled as 'DM' (disease-causing mutations) or 'DM?' (likely disease-causing mutations). For the negative dataset, we excluded 315 336 overlapping mutations between the VariSNP database version 2016-06-09 and 1000 Genomes database. In addition, any mutations of putatively benign overlapping with positive mutations were removed. According to the above criteria, we obtained 2603 synonymous mutations, of which 254 were deleterious synonymous mutations (positive dataset) and 2349 were putatively benign synonymous mutations (negative dataset). We defined these dataset as the 'full dataset'. To avoid a biased performance evaluation, we constructed a fully balanced subset of the 'full dataset' in which for every positive mutation, one putatively benign mutation that was located as close as possible to the positive one in the genome was selected. In total, the balanced dataset included 494 synonymous mutations, of which half were from the positive dataset and half from the negative dataset. We refer to these benchmark dataset as the 'close-by dataset' [34].

### Performance evaluation

As mentioned in many articles [35, 36], several approaches are used to evaluate the performance of predictive methods, including cross-validation such as n-fold, leave-family-out and leave-one-out. Furthermore, the independent test and case study are also the choices of the vast majority for evaluation. Here, based on two independent test datasets, we assessed the performance of the 10 algorithms using six measures, including sensitivity,

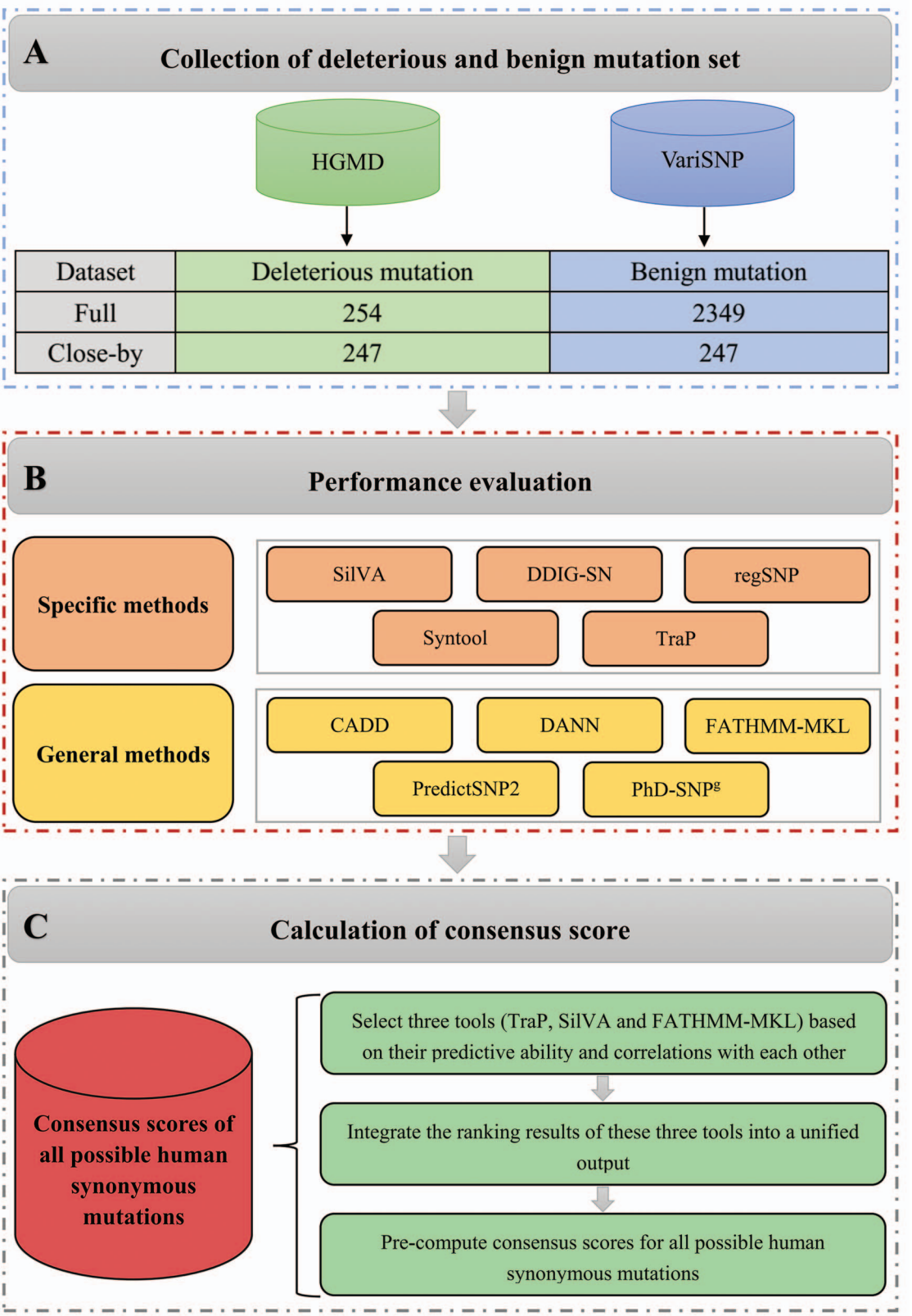


**Figure 1**. Overview of the proposed methodology for predicting deleterious synonymous mutations. **(A)** Process and statistics for the full and close-by independent test datasets. **(B)** 10 methods, including 5 specific methods for synonymous variants and 5 general methods for single nucleotide mutations evaluated in this work. **(C)** A flowchart of the calculation procedures for consensus scores of all possible human synonymous mutations.

**Table 1.** Comprehensive comparison of bioinformatics methods for the prediction of deleterious synonymous mutation

| Tool[a] (reference, year) | Predictive model | Feature representation | Performance evaluation strategy | Predictive threshold[b] | Training data | Availability |
|---|---|---|---|---|---|---|
| **Specific method** | | | | | | |
| SilVA ([17], 2013) | RF | Conservation, codon usage, sequence features, RNA splicing, pre-mRNA folding energy | Leave-one-out cross-validation and independent validation | >0.278 | Literature and 1000 Genomes Project | Software |
| DDIG-SN ([18], 2017) | SVM | Conservation, RNA splicing | 10-fold cross-validation and independent validation | >0.5 | HGMD and 1000 Genomes Project | Web server |
| regSNPs-splicing ([19], 2017) | RF | Conservation, RNA splicing, protein structural features | 10-fold cross-validation and independent validation | <0.5 | HGMD and 1000 Genomes Project | Web server |
| Syntool ([20], 2017) | RM | NA | Independent validation | ≤0 | Genome Aggregation Database | Software |
| TraP ([21], 2017) | RF | Conservation, sequence features, RNA splicing | 10-fold cross-validation | ≥0.459 | Literature and OMIM database | Software, web server |
| **General method** | | | | | | |
| CADD ([22], 2019) | LR | Conservation, regulatory information, transcript information, prote in-level scores | Independent validation | >5 | Ensembl EPO whole genome alignments | Software, web server |
| DANN ([23], 2015) | DNN | Conservation, regulatory information, transcript information, protein-level scores | Independent validation | ≥0.99 | Ensembl EPO whole genome alignments | Software |
| FATHMM-MKL([24], 2015) | MKL | Conservation, sequence features, regulatory information from ENCODE | 5-fold cross-validation and independent validation | >0.5 | HGMD and 1000 Genomes Project | Software, web server |
| PredictSNP2 ([25], 2016) | CP | NA | Independent validation | >0 | ClinVar, NHGRI GWAS catalog, COSMIC and VariSNP database | Web server |
| PhD-SNP[g] ([26], 2017) | GBA | Conservation, sequence features | 10-fold cross-validation and independent validation | >0.5 | ClinVar | Software, web server |

RF, random forest; SVM, support vector machine; RM, regression model; LR, logistic regression; DNN, deep neural network; MKL, multiple kernel learning; CP, consensus prediction; GBA, gradient boosting algorithm; NA, not applicable; ENCODE, Encyclopedia of DNA Elements; HGMD, Human Gene Mutation Database; OMIM, Online Mendelian Inheritance in Man; EPO, Enredo-Pecan-Ortheus; ClinVar, public archive of interpretations of clinically relevant variants; NHGRI, National Human Genome Research Institute; GWAS, genome wide association studies; COSMIC, Catalogue Of Somatic Mutations In Cancer; VariSNP, a benchmark database for neutral variations from single nucleotide polymorphism database.

[a]The URL addresses for the listed tools are as follows: SilVA, http://compbio.cs.toronto.edu/silva/; DDIG-SN, http://sparks-lab.org/ddig; reg-SNP, http://watson.compbio.iupui.edu/regSNP-splicing; Syntool, https://github.com/zhangtongda/Syntool; TraP, http://trap-score.org; CADD, https://cadd.gs.washington.edu/snv; DANN, https://cbcl.ics.uci.edu/public_data/DANN; FATHMM-MKL, http://fathmm.biocompute.org.uk/fathmmMKL.htm; PredictSNP2, https://loschmidt.chemi.muni.cz/predictsnp2; PhD-SNP[g], http://snps.biofold.org/phd-snpg.

[b]For regSNPs-splicing, it suggests using false discovery rate (FDR) as cut-off: FDR < 0.05 as deleterious, 0.05 ≤ FDR < 0.5 as potential deleterious and 0.5 ≤ FDR ≤ 1 as benign; For Syntool, if a Syntool scores ≤0 (or score percentile ≤0.5), the corresponding mutation is predicted as deleterious mutation.

specificity, accuracy, precision, Mathews correlation coefficient (MCC) and F1-score. These measures are defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TN}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TP}+\text{FN})\times(\text{TP}+\text{FP})}}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}},$$

where TP, FN, TN and FP represent the numbers of true positives, false negatives, true negatives and false positives, respectively.

In addition, we also assessed the overall performance of deleterious synonymous mutation prediction by the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC). When the AUC value of a predictor is larger than the area of other ROC curves, the predictor is regarded as a better predictor.

## Ensemble predictor design

In this work, to improve the distinguishable performance between deleterious and benign synonymous mutations, we developed an ensemble predictor named PrDSM that integrated the results of the TraP, SilVA and FATHMM-MKL methods. The

reasons that we selected the three tools were as follows. First, the three methods showed better performance, with an AUC >0.7 for two benchmark datasets compared with the other methods. Second, good diversity was determined among the three methods due to their low correlations with each other.

The most critical step in building an ensemble predictor is how to integrate obviously different outputs of the three algorithms. To integrate the three predictive methods, we precomputed the predictive scores for three algorithms of synonymous mutations in the whole genome extracted from CADD. Furthermore, we ranked the predictive scores of all synonymous mutations in each algorithm, respectively. The following step was the calculation of the percentile value of every synonymous mutation in each tool, with the purpose of making the different scales of all combined methods uniform. For example, let us assume that there are *n* synonymous mutations, and the mutation with the lowest score was ranked first after the above three processing steps. The final new score or percentile value of the top ranked synonymous mutation was then *1/n*. Finally, we calculated the mean percentile values of three combined methods for each synonymous mutation, i.e. the 'PrDSM' consensus score, using the following equation:

$$PrDSM\ score = \frac{\sum_{i=1}^{N} V_i}{N},$$

where $N$ is the number of combined tools, and $V_i$ represents the percentile value of each tool. The consensus scores range from 0 to 1, with the highest score indicating the most deleterious synonymous variant and the lowest score the most benign synonymous variant. The cut-off of the 'PrDSM' score was set when the sum of the accuracy, sensitivity and specificity was maximum [37].

## Results and discussion

### Overview of methods for deleterious synonymous mutation prediction

Table 1 shows the key aspects of the 10 methods, including the predictive model employed, input features used, performance evaluation strategy, threshold of the damaging score, training data used to train these tools and software/web server availability. We categorized these 10 methods into two types: specific method for deleterious synonymous mutation prediction and general method for the effect of single nucleotide mutation prediction.

In comparison to general methods, most synonymous mutation-specific methods are build based on the random forest, which is a widely used machine learning method in bioinformatics [38]. Machine learning-based methods usually need to calculate the sequence and/or structure-based features for model training. A variety of features have been used in the reviewed deleterious synonymous mutation prediction methods. Table 1 shows that splicing and conservation features are the most commonly used features. We speculated that these two features played an important role in the identification of deleterious synonymous mutations. For general methods, conservation features are still the most commonly used. In addition, epigenetic modification information, such as Histone ChIP-Seq and TFBS PeakSeq, is also used to construct general models. Most of the 10 tools are available as web servers, which require only an internet connection and a web browser. A few tools, including SilVA, Syntool and DANN, are provided in the form of a source code. Overall, most of these bioinformatics tools can be readily used by non-bioinformaticians.

### Evaluation of different prediction methods based on the independent test datasets

We used the independent test datasets constructed in this study to conduct a performance comparison among the tools listed in Table 1. We assembled the independent datasets by using the up-to-date datasets and removing the synonymous mutations from earlier versions of the HGMD and VariSNP database, thereby minimizing the overlap between our independent test datasets and the training datasets of the compared tools. We then submitted the synonymous mutations from our independent test datasets to the web servers or local tools listed in Table 1 to obtain corresponding prediction results. The command lines for stand-alone tools and running parameters used in this study are provided in Supplementary Table S1. Figure 2A illustrates the ROC curves of 10 methods on the full dataset. Specific tools such as SilVA, DDIG-SN, regSNP and TraP achieved better performance on the full dataset (with an AUC value of 0.770 for SilVA, 0.763 for DDIG-SN, 0.747 for regSNP and 0.740 for TraP), while general tools, with the exception of FATHMM-MKL, showed poor in ROC performance.

We also evaluated the performance of these tools in a binary classification based on several popular measures: sensitivity, specificity, precision, accuracy, MCC and F1 score. As shown in Table 2, the MCC values of the 10 tools ranged between 0.051 and 0.518 on the full dataset. The top-ranking tools were TraP, SilVA and DDIG-SN. Similarly, the arguably most representative single measure of the binary prediction, the F1-score, ranged between 2.31% and 50.5% on the full dataset. The top three tools were TraP, SilVA and CADD. We noted that SilVA and TraP showed the best performance in MCC, F1-score and accuracy. For other measures, we found that the sensitivities of 10 methods were, in general, <50%, excluding regSNP, Syntool and FATHMM-MKL. The specificities of the 10 methods were, in general, >80%, excluded regSNP, Syntool and FATHMM-MKL. The precisions of the 10 methods in general were <30%, excluding DANN, PredictSNP2, DDIG-SN, SilVA and TraP. In general, synonymous mutation-specific methods achieved better performance than general methods. We speculate that one reason why synonymous mutation-specific tools show best performance is that the training data and most features such as splicing for those tools are characteristically synonymous mutations [3, 39, 40].

In addition to predictive performance, missing values are a concern when applying a prediction method to large-scale synonymous mutation data generated from sequencing studies. 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. Based on the full datasets used for testing, we found that one of the methods (Syntool) showed a relatively high rate of missing values (22.67%). Another method (regSNP) yielded 1172 missing predictions for our datasets, possibly because of its limitation in the prediction of functional effects for mutations based on their impact on mRNA splicing and protein structure information. Compared with general methods, synonymous mutation-specific methods had more missing values. We decided to retain these mutations on our test datasets after determining that the overall ranking of the compared methods was not affected.

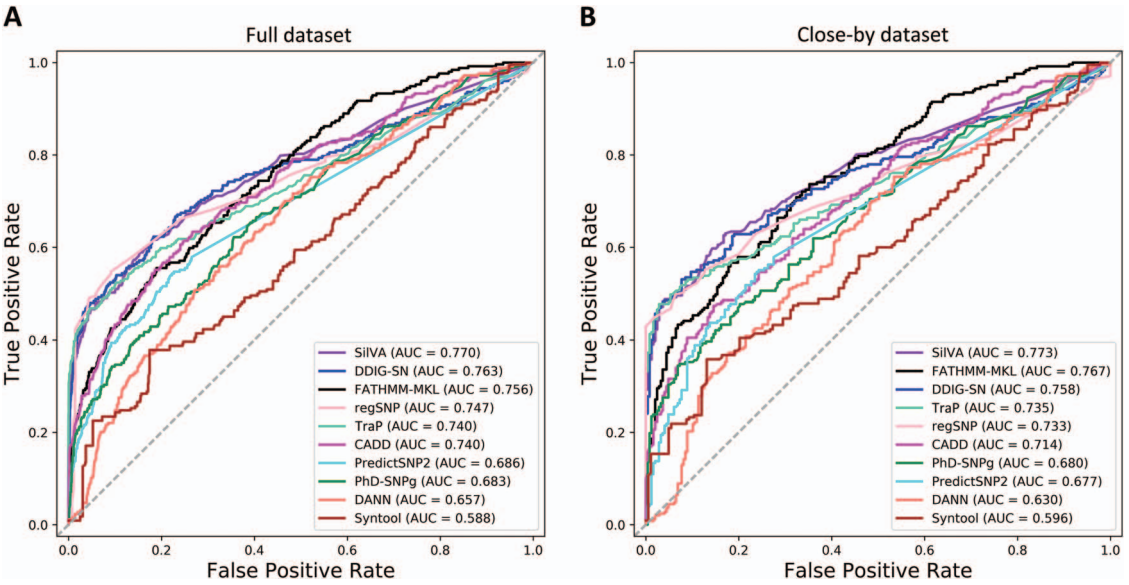


**Figure 2**. ROC curves of 10 methods on two independent test datasets. **(A)** full dataset; **(B)** close-by dataset.

**Table 2.** Performance evaluation of 10 methods based on the full dataset

| Tool | Missing (%) | Sen (%) | Spe (%) | Pre (%) | F1 (%) | ACC (%) | MCC | AUC | BIAS |
|---|---|---|---|---|---|---|---|---|---|
| SilVA | 1.1 | 34.0 | 99.3 | 83.5 | 48.3 | 92.9 | 0.505 | 0.770 | 0.658 |
| DDIG-SN | 0.1 | 19.1 | 99.8 | 90.6 | 31.5 | 92.0 | 0.394 | 0.763 | 0.809 |
| regSNP | 45.0 | 78.6 | 44.7 | 19.6 | 31.4 | 49.7 | 0.167 | 0.747 | −0.431 |
| Syntool | 22.7 | 71.2 | 36.6 | 12.2 | 20.8 | 40.4 | 0.051 | 0.588 | −0.486 |
| TraP | 0.1 | 36.6 | 99.1 | 81.6 | 50.5 | 93.0 | 0.518 | 0.740 | 0.631 |
| CADD | 0 | 46.5 | 87.6 | 28.9 | 35.6 | 83.6 | 0.278 | 0.740 | 0.469 |
| DANN | 0.1 | 1.18 | 99.9 | 50.0 | 2.31 | 90.2 | 0.065 | 0.657 | 0.988 |
| FATHMM-MKL | 0 | 59.8 | 74.0 | 19.9 | 29.9 | 72.6 | 0.220 | 0.756 | 0.192 |
| PredictSNP2 | 0.5 | 19.7 | 97.3 | 43.9 | 27.2 | 89.7 | 0.246 | 0.686 | 0.798 |
| PhD-SNP$^g$ | 0 | 34.7 | 89.9 | 27.1 | 30.4 | 84.5 | 0.220 | 0.683 | 0.614 |

**Sen**, sensitivity; **Spe**, specificity; **Pre**, precision; **F1**, F1-score; **ACC**, accuracy; **MCC**, Mathew correlation coefficient; **AUC**, area under the curve. **BIAS**, normalized value of difference between sensitivity and specificity.

Integrating the evaluation results in Table 2 and Figure 2A, we could reach the following conclusions. First, based on the full benchmark dataset, we determined that the method of SilVA achieves the best performance. Although SilVA possesses such advantages, it yields no predictions for mutations located on the Y chromosome. Taking these results into consideration, SilVA is still an outstanding tool for synonymous mutations. It is recommended that users employ SilVA to predict the pathogenicity of synonymous mutations, excluding those located on the Y chromosome. Second, FATHMM-MKL, DDIG-SN, regSNP, TraP and CADD also provide better performance for the full benchmark dataset. FATHMM-MKL, TraP and CADD provide local packages, while the other two methods can only be used online. If users have large amount of data, FATHMM-MKL, TraP and CADD are excellent choices. If not, DDIG-SN is also a good choice. Third, we noticed that Syntool provides poor performance for the full dataset. Although the method is specifically designed for synonymous mutations, it provides a region-based intolerance score for synonymous mutations. Thus, synonymous mutations within one region have the same predictive scores. This feature may explain why Syntool performed the worst.

Based on the close-by benchmark dataset, the predictive results for seven measures and the missing values for each method are shown in Table 3. Compared with the results of the full dataset, there were no large difference between these methods. Figure 2B shows the ROC curves for 10 methods. We found that SilVA (AUC = 0.773) and FATHMM-MKL (AUC = 0.767) showed the best performance, and the following four methods, including DDIG-SN, TraP, regSNP and CADD, had AUCs >0.7 (DDIG-SN = 0.758, TraP = 0.735, regSNP = 0.733 and CADD = 0.714). The remaining four methods had AUCs <0.69 (PhD-SNP$^g$ = 0.680, PredictSNP2 = 0.677, DANN = 0.630 and Syntool = 0.596).

Previous studies have reported that many mutations associated with purported diseases in HGMD are erroneous findings or misclassifications [41, 42], which might result in a relatively poor sensitivity. We then performed the analyses for the DM and DM? taken from HGMD, with the purpose of determining the association between the clinical significance of mutations and the prediction results. If the numbers of deleterious synonymous mutations labelled with 'DM' are higher than 'DM?' in the correct prediction by each tool while there is an opposite result for the incorrectly predicted mutations, this may suggest that the low sensitivity of the predictive tools is related to the clinical significance of the mutations extracted from HGMD. To test this hypothesis, we calculated the numbers of synonymous mutations labelled as 'DM' and 'DM?' that were

**Table 3.** Performance evaluation of 10 methods based on the close-by dataset

| Tool | Missing (%) | Sen (%) | Spe (%) | Pre (%) | F1 (%) | ACC (%) | MCC | AUC | BIAS |
|---|---|---|---|---|---|---|---|---|---|
| SilVA | 0.6 | 34.6 | 99.2 | 97.7 | 51.1 | 66.8 | 0.442 | 0.773 | 0.651 |
| DDIG-SN | 0.4 | 19.6 | 100 | 100 | 32.8 | 60.0 | 0.330 | 0.758 | 0.804 |
| regSNP | 36.4 | 78.5 | 41.8 | 71.6 | 74.9 | 65.7 | 0.215 | 0.733 | −0.468 |
| Syntool | 19.4 | 70.2 | 36.6 | 56.6 | 62.7 | 54.8 | 0.073 | 0.596 | −0.479 |
| TraP | 0.4 | 37.3 | 99.2 | 97.9 | 54.0 | 68.1 | 0.463 | 0.735 | 0.624 |
| CADD | 0 | 46.6 | 85.4 | 76.2 | 57.8 | 66.0 | 0.347 | 0.714 | 0.454 |
| DANN | 0 | 0.81 | 100 | 100 | 1.61 | 50.4 | 0.064 | 0.630 | 0.992 |
| FATHMM-MKL | 0 | 60.7 | 75.7 | 71.4 | 65.7 | 68.2 | 0.369 | 0.767 | 0.198 |
| PredictSNP2 | 0 | 19.8 | 96.4 | 84.5 | 32.1 | 58.1 | 0.252 | 0.677 | 0.795 |
| PhD-SNP$^g$ | 0 | 35.2 | 91.5 | 80.6 | 49.0 | 63.4 | 0.323 | 0.680 | 0.615 |

**Sen**, sensitivity; **Spe**, specificity; **Pre**, precision; **F1**, F1-score; **ACC**, accuracy; **MCC**, Mathew correlation coefficient; **AUC**, area under the curve. **BIAS**, normalized value of difference between sensitivity and specificity.

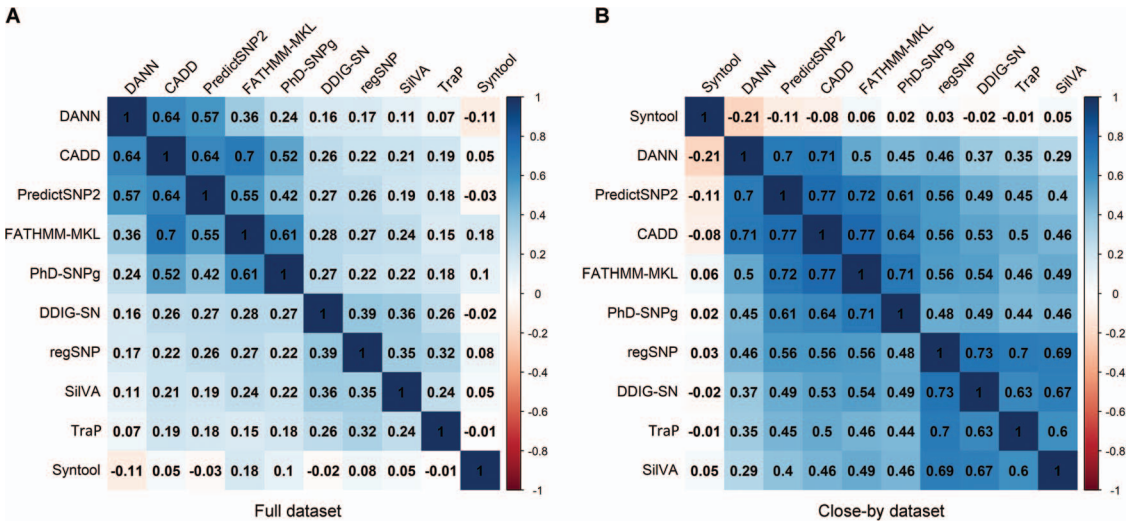


**Figure 3**. Correlations between the 10 methods. **(A)** full dataset; **(B)** close-by dataset.

predicted correctly or incorrectly by four tools (FATHMM-MKL, CADD, TraP and SilVA) with an AUC >0.7, a sensitivity >30% and ratio of missing values <20% for the full dataset, respectively. The results showed that the number of mutations with the label 'DM' was greater than 'DM?' among those correctly predicted by the FATHMM-MKL, CADD, TraP and SilVA tools, respectively (*P*-value = 0.002882, 0.0005421, 2.459e-07 and 4.727e-08; Fisher's exact test). For the false prediction, the number of mutations with the label 'DM?' was greater than 'DM' among those incorrectly predicted by FATHMM-MKL, CADD, TraP and SilVA tools (*P*-value = 3.276e-05, 0.001053, 0.000251 and 0.000444; Fisher's exact test). Our results suggested that the low sensitivity for synonymous mutations of predictive tools is likely due to a large number of prediction errors for mutations labelled as 'DM?' in HGMD.

We also tested whether the low sensitivities of many evaluated methods are associated with the disease types for synonymous mutations taken from HGMD. To verify how many of these mutations are associated with Mendelian diseases, multifactorial diseases or cancer, we manually queried the disease-related database [43–46] to find out which disease type is associated with each synonymous mutation in HGMD. The results showed that the DMs from Mendelian diseases generally tend to be correctly predicted in comparison to those from multifactorial diseases and cancer by FATHMM-MKL, CADD, TraP and SilVA tools (Supplementary Figure S1). This may be due to the relatively simple genetics model of Mendelian diseases. However, for the DM?, we found that the synonymous mutations from multifactorial diseases generally tend to be correctly classified as deleterious mutations in comparison to those from Mendelian diseases and cancer.

## Correlation of prediction methods

To evaluate the correlation of predictive results between any two computational methods, we calculated the Spearman correlation coefficient based on the full and close-by datasets. For the full dataset (Figure 3A), we observed that DANN, CADD, PredictSNP2, FATHMM-MKL and PhD-SNP$^g$ were almost all highly (Spearman correlation coefficient, $R > 0.6$) to moderately correlated (Spearman correlation coefficient, $0.6 > R > 0.4$). The highest correlation was found between CADD and FATHMM-MKL ($R = 0.7$). The rest of methods were lowly correlated with other methods. For the close-by dataset, we also observed that DANN, CADD, PredictSNP2, FATHMM-MKL and PhD-SNP$^g$ were all highly to moderately correlated with each other. In contrast to the correlation observed on the full dataset, the rest of methods were lowly to moderately correlated with other methods based on the close-by dataset (Figure 3B). Together, we obtained similar patterns for the two benchmark datasets.

To further investigate the correlation between the computational methods, we analysed the extent of concordance among

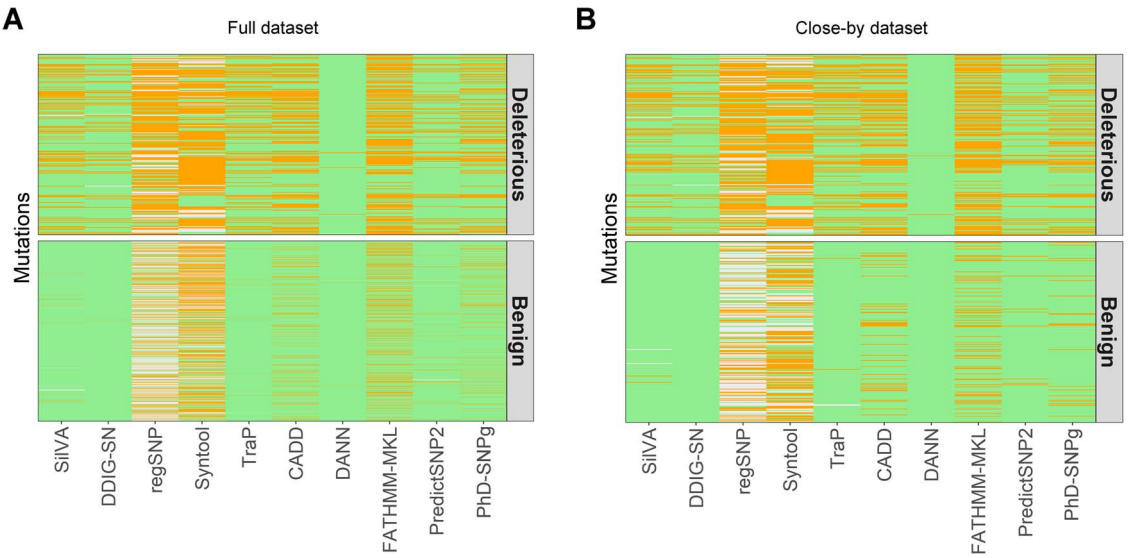


**Figure 4**. Concordance among predictions of 10 methods. Binary predictions made by 10 methods for each pathogenic or benign mutations shown in the figure. Each mutation is presented along a row, and an orange, green or white tile represents a pathogenic, benign or missing data, respectively, based on the corresponding methods. **(A)** full dataset; **(B)** close-by dataset.

**Table 4.** Concordance rate of different combinations of methods

| Functional category | Dataset | Methods | Total mutations | Concordance [n (%)] | False concordance [n (%)] |
|---|---|---|---|---|---|
| Benign | Full | All 10 | 2349 | 646 (27.50) | 0 (0) |
| Deleterious | Full | All 10 | 254 | 0 (0) | 17 (6.69) |
| Benign | Close-by | All 10 | 247 | 70 (28.34) | 0 (0) |
| Deleterious | Close-by | All 10 | 247 | 0 (0) | 17 (6.88) |
| Benign | Full | SilVA, TraP, CADD, FATHMM-MKL | 2349 | 1593 (67.82) | 0 (0) |
| Deleterious | Full | SilVA, TraP, CADD, FATHMM-MKL | 254 | 72 (28.35) | 82 (32.28) |

10 predictive methods based on the full benchmark dataset. The concordance of the 10 methods was calculated according to a previous proposed method [29]. Figure 4A and Table 4 show the results for the concordance across 10 methods. We found that the concordance of benign mutations was 27.50% and deleterious synonymous mutations was 0%. We also calculated the 'false concordance' [29], for which 10 methods provided concordant assertions that were opposite to the true labels of the mutations in VariSNP and HGMD. For the mutations classified as deleterious in HGMD, 6.69% of them were predicted as benign by all 10 methods and conversely no benign mutation in VariSNP was predicted deleterious across 10 methods. To ensure that the above results were not influenced by the imbalanced full dataset, we also analysed the concordance with the close-by dataset among 10 methods (Figure 4B and Table 4). As shown in Table 4, 28.34% of the benign and 0% of the deleterious mutations had concordance across 10 predictive results. In addition, 6.88% of the deleterious and 0% of the benign mutations showed false concordance among 10 predictive results. The results based on the close-by dataset suggested that the imbalance of the full dataset contributed little to the concordance among the 10 tools.

Based on the results for the full and close-by datasets, we found that synonymous prediction methods generally showed low concordance, potentially due to the commonly low sensitivity among the evaluated methods. We then computed the level of concordance for four tools (FATHMM-MKL, CADD, TraP and SilVA), which had an AUC >0.7, a sensitivity >30% and rate of missing values <20% for the full dataset. The results showed that 67.82% of the benign and 28.35% of the deleterious mutations had concordance across four predictive results. In addition, 32.28% of the deleterious and 0% of the benign mutations had false concordance among four predictive results (Table 4). The results suggested that the low sensitivity of the prediction tools greatly contributed to the concordance computing.

## Additional analysis of prediction methods

We found that the predictions were towards functionally neutral variants of most tools based on the full and close-by datasets. To verify that the biased prediction of tools was not the reason for the imbalanced full dataset, we extended the tool evaluation by generating five additional 'close-by' independent datasets with the same ratio of pathogenic to neutral mutations. We performed the same operations on additional five 'close-by' datasets as the first 'close-by' dataset, but with the 2nd nearest, 3rd nearest, 4th nearest, 5th nearest and 6th nearest distance on the genome (named close-by1, close-by2, close-by3, close-by4 and close-by5, respectively). We calculated the criteria, including the sensitivity, specificity, precision, accuracy, F1, MCC and AUC, based on these additional datasets. The results on these five datasets are similar to those obtained on the close-by dataset (Supplementary Table S2A–E). We also provided a criterion called BIAS ($-1 \leq$ BIAS $\leq 1$), which is the normalized value of the difference in specificity and sensitivity [(specificity-sensitivity)/maximum (specificity, sensitivity)], to measure the predictive bias of tools. If BIAS >0.25, the prediction is

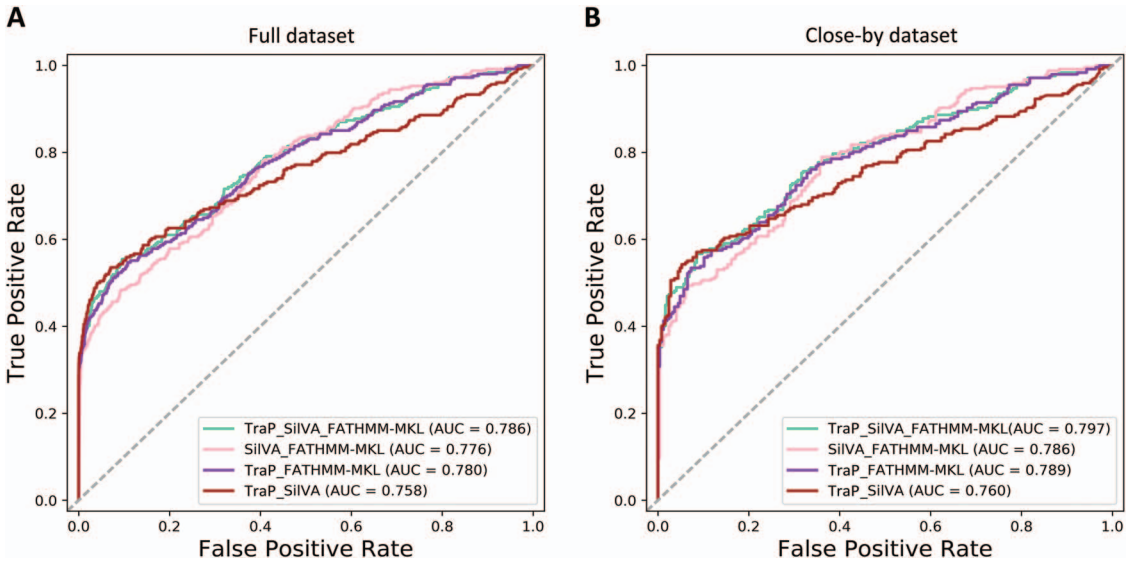


**Figure 5**. ROC curves of ensemble predictors on two independent test datasets. **(A)** full dataset; **(B)** close-by dataset.

considered to be biased to functionally neutral variants, which suggests that the majority of mutations predicted to be neutral are actually deleterious synonymous mutations. If BIAS $\leq -0.25$, the prediction is considered to be biased to functionally deleterious mutations, which means that many actually benign mutations are predicted to be deleterious synonymous mutations. We compared BIAS among seven independent datasets (full, close-by and additional close-by 1–5 datasets) (Tables 2 and 3; Supplementary Table S2A–E). Using this parameter, the best-performing method was FATHMM-MKL. In comparison, we found that seven tools (DANN, CADD, PhD-SNP$^g$, PredictSNP2, DDIG-SN, SilVA and TraP) had BIASs $>0.25$, where regSNP and Syntool had BIASs $\leq -0.25$ on all seven independent datasets, which indicates that the prediction was still biased for the extended five close-by datasets. Based on the above experiments, we demonstrated that the bias of the prediction results of the tools towards neutral synonymous mutations was not due to an imbalance of the full dataset. In order to further extend the tools evaluation, we also utilized the test dataset from SilVA (4 deleterious and 10 benign synonymous mutations). We only evaluated four tools (FATHMM-MKL, CADD, TraP and SilVA), which had an AUC $>0.7$, a sensitivity $>30\%$ and missing values for the full dataset $<20\%$. The results showed that the BIAS of the four tools was $<0.25$ (Supplementary Table S2F). The results on the SilVA test dataset did not match the results for the above seven datasets, which might be due to the smaller test dataset of SilVA (with only 4 deleterious and 10 benign synonymous mutations). Regardless, we could still find that the performance of synonymous mutation-specific methods is higher than general methods for single nucleotide mutations.

To further investigate whether the bias of the predictive results towards negative samples was due to the methods themselves, we reviewed the literature to obtain the original reported results. Only three tools, DDIG-SN, PredictSNP2 and PhD-SNP$^g$, provided the results of sensitivity and specificity. The values obtained for the sensitivity and specificity of PredictSNP2 and PhD-SNP$^g$ were very close, while DDIG-SN had a BIAS $>0.25$. Thus, these findings also partly support our speculation. In summary, the predictive results of the tools were biased towards benign synonymous mutations, which was not due to the imbalance of the full dataset but more likely because the methods themselves were biased towards the prediction of synonymous mutations.

## The ensemble method improves the performance of deleterious synonymous mutation prediction

Given the consideration of the predictive performances and the correlations between any two methods based on the full and close-by benchmark datasets, three methods including TraP, SilVA and FATHMM-MKL were combined to construct an ensemble predictor, with the purpose of improving the performance of predicting deleterious synonymous mutations. For the three methods, we generated all possible combinations of two (TraP_SilVA, TraP_FATHMM-MKL, SilVA_FATHMM-MKL) or three (TraP_SilVA_FATHMM-MKL) algorithms. For the full benchmark dataset, we found that the combination of three algorithms achieved the best performance, with an AUC value of 0.786. The consensus predictors of the combination of two algorithms (SilVA_FATHMM-MKL and TraP_FATHMM-MKL) achieved an AUC of 0.776 and 0.780, respectively (Figure 5A). We observed sensitivities with a scope from 42.5 to 50.0%, specificities with a scope from 93.2% to 96.6%, precisions with a scope from 44.4% to 60.5%, F1-scores with a scope from 45.9% to 54.0%, accuracies with a scope from 89.0% to 91.9% and MCCs with a scope from 0.407 to 0.500 (Supplementary Table S3A). For the close-by benchmark dataset, we found that the combination of the three algorithms also achieved the best performance, with an AUC of 0.797, followed by SilVA_FATHMM-MKL and TraP_FATHMM-MKL with AUCs of 0.786 and 0.789, respectively (Figure 5B). Supplementary Table S3B shows the seven measures with the close-by dataset. To select the best combination of these four integration tools, we also utilized five additional close-by datasets (close-by1, close-by2, close-by3, close-by4 and close-by5) and a test dataset from SilVA to extend our testing experiments. The results showed that the sensitivity and AUCs were higher for TraP_SilVA_FATHMM-MKL compared with other three combination tools (Supplementary Table S3C–H). The ensemble tool of TraP_SilVA showed the best performance on the SilVA test dataset. Considering that the SilVA test dataset is very small, the results may not be representative and robust, and the combination of TraP_SilVA was not considered.

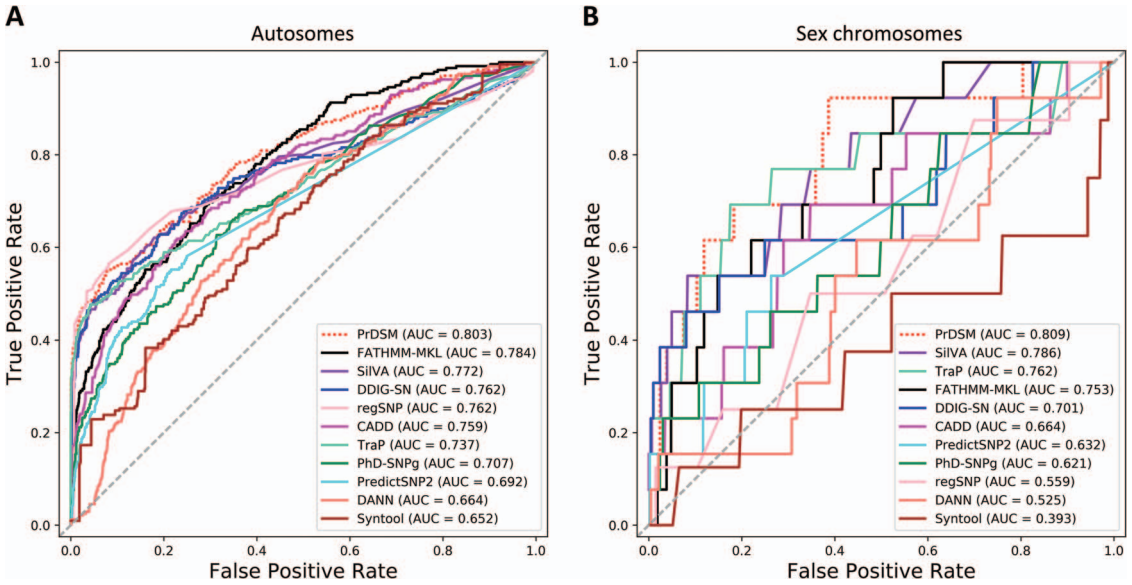


**Figure 6**. ROC curves of 10 methods based on mutations located on different chromosomes. **(A)** autosomes; **(B)** sex chromosomes.

Because the features of splicing and conservation play an important role in the prediction of deleterious synonymous mutations, we conducted additional tests to determine whether these two features can improve our consensus predictor. We selected two commonly used methods, SPANR (based on splice site) [7] and GERP++ (based on conservation features) [47], and evaluated their performance on the full, six close-by datasets and the test dataset from SilVA. We found that GERP++ had an AUC >0.65 and SPANR had an AUC <0.40 (Supplementary Figure S2). Although GERP++ showed good predictive performance compared with SPANR, the AUC values of these two methods were generally lower than 0.7. Thus, we did not integrate these two tools into our ensemble model. Given a comprehensive consideration of the results among eight benchmark datasets, we finally selected the combination of TraP_SilVA_FATHMM-MKL as our ensemble predictor and designated it Prediction of Deleterious Synonymous Mutation (PrDSM).

### Evaluation of prediction methods based on mutations of autosomes and sex chromosomes

As sex chromosomes have a different evolutionary rate than autosomes [48], several methods, such as FATHMM-XF [49], are restricted methodologically to mutations located on the autosomes. Given the importance of the discovery by Charlesworth *et al.* [48], we assessed the predictive performance among 11 algorithms including PrDSM for mutations on autosomes and sex chromosomes, respectively. We found that the ensemble predictor PrDSM achieved the best performance, with AUCs of 0.803 and 0.809 for the two types of chromosomes, respectively. The three methods of TraP, SilVA and FATHMM-MKL that we selected to construct the ensemble predictor also showed relatively stable results with AUCs of 0.737, 0.772 and 0.784 on the dataset of mutations on the autosomes, and 0.762, 0.786 and 0.753 for mutations on the sex chromosomes, respectively. The other seven algorithms, CADD, DDIG-SN, regSNP, PhD-SNP$^g$, PredictSNP2, DANN and Syntool, showed large differences in predictive performance between the two categories of chromosomes. We observed that the difference ranged from 6% to 26%. We found that compared with the mutations on sex chromosomes, predictive methods showed better performance for the autosomal mutations (Figure 6). In summary, the results highlight the differences in predictive performance between mutations located on autosomes and sex chromosomes for most prediction methods, which suggest that, to better distinguish deleterious synonymous mutations from benign synonymous mutations, chromosome-specific prediction algorithms should be developed to build more reliable computational models.

### Web server of the ensemble predictor

To facilitate access to our ensemble predictor, PrDSM, we constructed a user-friendly web server. Users can query the predictor by entering tab separated string, or upload large batches of mutations as a tab delimited file or variant call format (VCF) file, which at a minimum must include chromosome, position, identity, reference and altered mutation alleles. For the output page, users can observe the predictive and percentile values generated by the ensemble predictor PrDSM and the component predictors TraP, SilVA and FATHMM-MKL. The output is available in the format of VCF file. To obtain results in a time-efficient manner, we also pre-calculated the prediction of 23 206 778 synonymous mutations for the GRCh37/hg19 version of the human genome, and users can access this information online in 'Download' section on the 'home' page of the PrDSM website (http://bioinfo.ahu.edu.cn:8080/PrDSM).

## Conclusion

Owing to the functional significance of deleterious synonymous mutations in diverse human diseases, several computational tools are available for the evaluation of synonymous mutations. Given the significant achievement in this area, we provided a comprehensive evaluation of 10 state-of-the-art predictors based on two independent test datasets. The results showed that specific tools for deleterious synonymous mutation generally outperformed general tools for single nucleotide mutations. Among the evaluated 10 methods, we found that

SilVA is the most robust method for predicting the pathogenicity of synonymous mutations. To improve the predictive performance, we selected three methods with high outperformance and low correlations, TraP, SilVA and FATHMM-MKL, to construct an ensemble predictor named PrDSM. Based on two independent test datasets, the results demonstrated that PrDSM has the best performance compared with the 10 evaluated algorithms. Although the ensemble tool outperformed the existing methods, it still exhibited a low-sensitivity performance similar with other synonymous prediction tools. Our conclusion is that, on the one hand, synonymous mutations with more reliable clinical significance are required for model construction and performance evaluation. On the other hand, there are larger differences in the genetic patterns of different disease types. For example, multifactorial diseases are related to genetic and environmental factors; therefore, deleterious synonymous mutations in individual genes may not play a decisive role in the occurrence and development of those diseases. We recommend that the construction of individualized predictive models is needed for different disease types to improve the predictive performance of synonymous mutations. Taking the different rates of evolution between autosomes and sex chromosomes into consideration, we also provided some insight into further investigations of predictive performance among 11 methods, including PrDSM, for identifying mutations on autosomes and sex chromosomes. We noted that there are large deviations in some methods, which suggests that chromosome-specific prediction algorithms should be developed to build more reliable computational models. Concerning the future development, we also plan to assess new tools for predicting the effect of synonymous mutations as they emerge, such as IDSV [27], and to consider integrating any tools that would boost the predictive power of our ensemble model. Together, we anticipate that this survey will serve as a useful guide for interested readers to facilitate selection of existing tools and to inspire the future development of new tools for deleterious synonymous mutation prediction.

## Key Points

- We provide a comprehensive survey of 10 computational methods (including 5 specific methods for deleterious synonymous mutation and 5 general methods for single nucleotide mutation) developed for predicting deleterious synonymous mutations. Our survey analyses these 10 computational methods in terms of underlying algorithms used, input features, performance evaluation strategy and practical utility.
- We construct two independent test datasets and accordingly perform comprehensive tests to assess the performance of existing tools for predicting deleterious synonymous mutations.
- We propose and built a new ensemble model, Prediction of Deleterious Synonymous Mutation (PrDSM), to further improve the performance in predicting the functional effects of synonymous mutation. Independent tests demonstrate that the ensemble model outperforms the current predictors. We make PrDSM publicly accessible at http://bioinfo.ahu.edu.cn:8080/PrDSM.
- This article provides a useful guide for interested readers to facilitate selection of existing tools and development of new tools for deleterious synonymous mutation prediction.

## Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

## Funding

National Natural Science Foundation of China (61672037, 11835014 and 61873001); the Anhui Provincial Outstanding Young Talent Support Plan (gxyqZD2017005); the Young Wanjiang Scholar Program of Anhui Province; and the Recruitment Program for Leading Talent Team of Anhui Province.

## References

1. Hunt RC, Simhadri VL, Iandoli M, *et al*. Exposing synonymous mutations. *Trends Genet* 2014;**30**:308–21.
2. Parkes M, Barrett JC, Prescott NJ, *et al*. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* 2007;**39**:830–2.
3. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* 2011;**12**:683–91.
4. Brest P, Lapaquette P, Souidi M, *et al*. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* 2011;**43**:242–5.
5. Chen R, Davydov EV, Sirota M, *et al*. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One* 2010;**5**:e13574.
6. Solis AS, Shariat N, Patton JG. Splicing fidelity, enhancers, and disease. *Front Biosci* 2008;**13**:1926–42.
7. Xiong HY, Alipanahi B, Lee LJ, *et al*. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 2015;**347**:1254806.
8. Takata A, Ionita-Laza I, Gogos JA, *et al*. De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia. *Neuron* 2016;**89**:940–7.
9. Zheng S, Kim H, Verhaak RGW. Silent mutations make some noise. *Cell* 2014;**156**:1129–31.
10. Supek F, Minana B, Valcarcel J, *et al*. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 2014;**156**:1324–35.
11. Diederichs S, Bartsch L, Berkmann JC, *et al*. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol Med* 2016;**8**:442–57.
12. Schutz FAB, Pomerantz MM, Gray KP, *et al*. Single nucleotide polymorphisms and risk of recurrence of renal-cell carcinoma: a cohort study. *Lancet Oncol* 2013;**14**:81–7.
13. Kandoth C, McLellan MD, Vandin F, *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 2013;**502**:333–9.
14. Gotea V, Gartner JJ, Qutob N, *et al*. The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment Cell Melanoma Res* 2015;**28**:673–84.
15. Dixit R, Kumar A, Mohapatra B. Implication of GATA4 synonymous variants in congenital heart disease: a comprehensive in-silico approach. *Mutat Res* 2018;**813**:31–8.
16. Reitz C, Felsky D, Santa-Maria I, *et al*. Rare, synonymous variants in Cdh23, Slc9a3r1, Rhbdd2 and Itih2 are associated

with Alzheimer's disease in multiplex Caribbean Hispanic families. *Alzheimers Dement* 2018;**14**:P339.
17. Buske OJ, Manickaraj A, Mital S, *et al*. Identification of deleterious synonymous variants in human genomes. *Bioinformatics* 2013;**29**:1843–50.
18. Livingstone M, Folkman L, Yang Y, *et al*. Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Hum Mutat* 2017;**38**:1336–47.
19. Zhang X, Li M, Lin H, *et al*. regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution. *Hum Genet* 2017;**136**:1279–89.
20. Zhang T, Wu Y, Lan Z, *et al*. Syntool: a novel region-based intolerance score to single nucleotide substitution for synonymous mutations predictions based on 123,136 individuals. *Biomed Res Int* 2017;**2017**:5096208.
21. Gelfman S, Wang Q, McSweeney KM, *et al*. Annotating pathogenic non-coding variants in genic regions. *Nat Commun* 2017;**8**:236.
22. Rentzsch P, Witten D, Cooper GM, *et al*. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94.
23. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–3.
24. Shihab HA, Rogers MF, Gough J, *et al*. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**:1536–43.
25. Bendl J, Musil M, Stourac J, *et al*. PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Comput Biol* 2016;**12**:e1004962.
26. Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res* 2017;**45**:W247–52.
27. Shi F, Yao Y, Bin Y, *et al*. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. *BMC Med Genomics* 2019;**12**(Suppl 1):12.
28. Olatubosun A, Valiaho J, Harkonen J, *et al*. PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat* 2012;**33**:1166–74.
29. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol* 2017;**18**:225.
30. Li J, Zhao T, Zhang Y, *et al*. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res* 2018;**46**:7793–804.
31. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics* 2013;**14**(Suppl 3):S2.
32. Stenson PD, Mort M, Ball EV, *et al*. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;**136**:665–77.
33. Schaafsma GC, Vihinen M. VariSNP, a benchmark database for variations from dbSNP. *Hum Mutat* 2015;**36**:161–6.
34. Ritchie GR, Dunham I, Zeggini E, *et al*. Functional annotation of noncoding sequence variants. *Nat Methods* 2014;**11**:294–6.
35. Li F, Wang Y, Li C, *et al*. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Brief Bioinform* 2018. doi: 10.1093/bib/bby077.
36. Bao Y, Marini S, Tamura T, *et al*. Toward more accurate prediction of caspase cleavage sites: a comprehensive review of current methods, tools and features. *Brief Bioinform* 2018. doi: 10.1093/bib/bby041.
37. Pan Y, Wang Z, Zhan W, *et al*. Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 2018;**34**:1473–80.
38. Boulesteix A-L, Janitza S, Kruppa J, *et al*. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov* 2012;**2**:493–507.
39. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002;**3**:285–98.
40. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 2006;**7**:98–108.
41. Cassa CA, Tong MY, Jordan DM. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum Mutat* 2013;**34**:1216–20.
42. McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, *et al*. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. *BMC Med Genet* 2014;**15**:134.
43. U.S. National Institutes of Health, National Library of Medicine. *Genetics Home Reference*. http://ghr.nlm.nih.gov (26 February 2019, date last accessed).
44. U.S. National Institutes of Health, National Library of Medicine. *MEDLINEplus*. http://www.medlineplus.gov (27 February 2019, date last accessed).
45. U.S. National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, Johns Hopkins University. *Online Mendelian Inheritance in Man (OMIM)*. http://www.ncbi.nlm.nih.gov/Omim/ (28 February 2019, date last accessed).
46. Kibbe WA, Arze C, Felix V, *et al*. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;**43**:D1071–8.
47. Davydov EV, Goode DL, Sirota M, *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**:e1001025.
48. Charlesworth B, Coyne JA, Barton NH. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat* 1987;**130**:113–46.
49. Rogers MF, Shihab HA, Mort M, *et al*. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* 2018;**34**:511–3.