Supplementary Information

Contents

1. Supplementary Methods
   Dataset curation
   Feature extraction
   Feature standardization
   Feature selection
   Performance evaluation

2. Supplementary Tables
   Table S1 The 175 features for feature selection
   Table S2 The best hyper parameters in each prediction model
   Table S3 The average performance of G10FCV on the training dataset
   Table S4 Summary of the independent testing datasets before and after removing homologs
   Table S5 Performance on the independent testing dataset before and after removing homologs
   Table S6 Performance comparison with MetaSVM, MetaLR, and M-CAP
   Table S7 Performance comparison on the dataset whose proteins contains both disease and neutral variants
   Table S8 Versions of python packages
3. Supplementary Figures

**Fig. S1** Flowchart of feature selection.

**Fig. S2** Performance comparison based on four metrics (A, C) and the ROC curves (B, D) on the independent testing dataset.

**Fig. S3** Performance comparison based on four metrics (A, C) and the ROC curves (B, D) on the third-party dataset.

**Fig. S4** Rank of the feature importance in the XGBoost-based model (A) and the LightGBM-based model (B).

**Fig. S5** The distributions of standardized Z-scores of the 17 selected features.
1. Supplementary Methods

Dataset curation

The nsSNV dataset was derived from the “disease” and “polymorphism” natural variants recorded in the “humsavar.txt” file from the UniProt/Swiss-Prot database (Release 2019_01 of January 16 2019) (The UniProt Consortium, 2019). It is worth noting that the “Polymorphism” label here does not necessarily mean large population frequency. The nsSNVs occurring multiple times with consistent labels were retained with only one occurrence, while those with inconsistent labels were discarded. We further removed all nsSNVs in TITIN (UniProt accession number: Q8WZ42) due to its extreme length that makes it difficult to extract alignment-based features.

Feature extraction

The full list of 175 feature candidates is provided in Table S1. The detailed descriptions are as follows.

Substitution matrix scores

Although they do not contain information discriminating different proteins and different substitution positions, substitution matrix scores were widely used in the features for predicting disease-associated nsSNV by providing general similarity or evolutionary distance between different amino acids. We used the Miyata matrix (Miyata, et al., 1979) and BLOSUM62 matrix (Henikoff and Henikoff, 1992) to extract scores of the substitution between mutant and wildtype residues as two features.

Sequence alignment features

The sequence alignment features we attempted are similar to those used in PMut2017 (Lopez-Ferrando, et al., 2017). In detail, for each of the protein sequences, we ran PSI-BLAST 2.6.0+ (Altschul, et al., 1997) against UniRef100 and UniRef90 (Release of 2018_07) clusters (Suzek, et al., 2015) with default E-values, default H-values, and 3 iterations to search for similar sequences. Two sequence alignments were built accordingly. Based on each of them, we constructed 3 additional alignments by applying 3 different filtering strategies: (1) keeping only the human sequences, (2) excluding all the human sequences, and (3) keeping the sequences under a stricter E-value threshold (less than 10E-45 and 10E-75 for UniRef90 and UniRef100 alignments, respectively). Taking together, we constructed 8 alignments. For each of them, a set of 10 features was extracted, including:

1. Number of sequences (N_align),
2. Number of residues at the substitution site \( i \) (\( N_{\text{all,aa,i}} \)),

3. Number of wildtype residues at the substitution site \( i \) (\( N_{\text{wt, i}} \)),

4. Number of mutant residues at the substitution site \( i \) (\( N_{\text{mt, i}} \)),

5. Proportion of wildtype residues at the substitution site \( i \) (\( p_{\text{wt, i}} \)),

6. Proportion of mutant residues at the substitution site \( i \) (\( p_{\text{mt, i}} \)),

7. Position Weight Matrix (PWM) score at the substitution site \( i \) (\( \text{PWM}_i \)),

8. PWM score multiplied by BLOSUM62 score (\( \text{PWM}_i \ast \text{BLOSUM}_{\text{wt, mt}} \)),

9. PWM score multiplied by Miyata score (\( \text{PWM}_i \ast \text{Miyata}_{\text{wt, mt}} \)),

10. Relative Entropy (\( \text{RE}_i \)) at the substitution site \( i \).

Among them, the Position Weight Matrix (PWM) score is defined as

\[
\text{PWM}_i = \ln \left( \frac{\frac{N_{\text{wt,i}}}{Q_{\text{wt}}} \ln \left( \frac{\frac{N_{\text{wt,i}}}{Q_{\text{wt}}} \right)} - \ln \left( \frac{\frac{N_{\text{mt,i}}}{Q_{\text{mt}}} \right)} \right)
\]

, where \( Q_{\text{wt}} \) and \( Q_{\text{mt}} \) represent the frequency of wildtype and mutant residue in the nature (based on the statistics of Swiss-Prot database, scaled with 100 times, i.e., the sum of all the Q values is equal to 100), respectively. Relative Entropy (RE) is a measure of the uncertainty, and was adopted here to quantify the conservation of the substitution site. The RE at the substitution site \( i \) is defined as

\[
\text{RE}_i = -\sum_j p_{j,i} \log_2 \left( \frac{p_{j,i}}{Q_j} \right)
\]

, where \( j \) transverses all the residue types, \( p_{j,i} \) represents the frequency of residue type \( j \) at the substitution position \( i \) in the alignment (gaps were not considered in this calculation), and the \( Q_j \) represents the frequency of residue type \( j \) in the nature, as described above.

For features from 1 to 9 described above, we further calculated a weighted version. In brief, we used the BLAST scores as weights when we count the number of sequences or residues. The detailed calculations are described as follows.

Weighted version of number of sequences (\( \text{wN}_{\text{all}} \)) is defined as

\[
\text{wN}_{\text{all}} = \sum_n S_n
\]

, where \( S_n \) represents BLAST score of sequence \( n \) in the alignment. Weighted version of number of residues at the substitution site \( i \) (\( \text{wN}_{\text{all,aa,i}} \)) is defined as
\[ wN_{all,aa,i} = \sum_{n_i} S_{n_i} \] (4)

where \( n_i \) represents each sequence in the alignment without gap at the substitution site \( i \), and \( S_{n_i} \) represents BLAST score of \( n_i \). Weighted version of number of wildtype residues at the substitution site \( i \) (\( wN_{wt,i} \)) is defined as

\[ wN_{wt,i} = \sum_{n_{wt,i}} S_{n_{wt,i}} \] (5)

where \( n_{wt,i} \) represents each sequence in the alignment with the wildtype residue in substitution site \( i \), and \( S_{n_{wt,i}} \) represents BLAST score of \( n_{wt,i} \). Weighted version of number of mutant residues at the substitution site \( i \) (\( wN_{mt,i} \)) is defined as

\[ wN_{mt,i} = \sum_{n_{mt,i}} S_{n_{mt,i}} \] (6)

where \( n_{mt,i} \) represents each sequence in the alignment with the mutant residue in substitution site \( i \), and \( S_{n_{mt,i}} \) represents BLAST score of \( n_{mt,i} \). Weighted version of proportion of wildtype residues at the substitution site \( i \) (\( wp_{wt,i} \)) is defined as

\[ wp_{wt,i} = \frac{wN_{wt,i}}{N_{all,aa,i}} \] (7)

where \( wN_{wt,i} \) represents weighted version of number of wildtype residues at the substitution site \( i \) and \( N_{all,aa,i} \) represents number of residues at the substitution site \( i \). Weighted version of proportion of mutant residues at the substitution site \( i \) (\( wp_{mt,i} \)) is defined as

\[ wp_{mt,i} = \frac{wN_{mt,i}}{N_{all,aa,i}} \] (8)

where \( wN_{mt,i} \) represents weighted version of number of mutant residues at the substitution site \( i \) and \( N_{all,aa,i} \) represents number of residues at the substitution site \( i \). Weighted version of position weight matrix (PWM) score at the substitution site \( i \) (\( wPWM_i \)) is defined as

\[ wPWM_i = \ln \left( \frac{wN_{mt,i}}{Q_{mt}} \right) - \ln \left( \frac{wN_{wt,i}}{Q_{wt}} \right) \] (9)

where \( Q_{wt} \) and \( Q_{mt} \) represent the frequency of wildtype and mutant residue in the nature (based on the statistics of Swiss-Prot database), respectively. \( wN_{wt,i} \) and \( wN_{mt,i} \) represent weighted version of number of wildtype and mutant residues at the substitution site \( i \), respectively. Weighted version of PWM score multiplied by BLOSUM62 score (wPWMB62) is defined as

\[ wPWMB62 = wPWM_i \times BLOSUM_{wt,mt} \] (10)

where \( wPWM_i \) represents weighted version of position weight matrix (PWM) score at the substitution site \( i \). \( BLOSUM_{wt,mt} \) represents BLOSUM62 matrix scores of
wildtype residue and mutant residue. Weighted version of PWM score multiplied by Miyata score \((wPWMM)\) is defined as

\[
wPWMM = wPWM_i \cdot Miyata_{wt,mt}
\]

where \(wPWM_i\) represents weighted version of position weight matrix (PWM) score at the substitution site \(i\). \(Miyata_{wt,mt}\) represents Miyata matrix scores of wildtype residue and mutant residue.

Taking together, we extracted 19 features for each of the 8 alignments, resulting in 152 alignment-based features in total. Among them, many are sequence conservation or are related to sequence conservation at the substitution site, such as RE values and PWM scores.

**Hydrophobicity features**

We adopted Wimley-White hydropathy index (Wimley and White, 1996) and octanol-water free energy transfer index (Eisenberg and McLachlan, 1986) of each residue type. With window sizes of 3 and 9 around the substitution site, the index values were summed to describe the hydrophobicity microenvironment. In addition, difference between mutant and wildtype residue was also calculated. Taken together, 6 features were extracted.

**Protein/gene-level annotations**

Protein/gene-level functional annotations were extracted from dbNSFP v3.5a (Liu, et al., 2016). These features include estimated probability of haploinsufficiency of a gene (Huang, et al., 2010), score of predicting the gene haploinsufficiency (Steinberg, et al., 2015), estimated probability that a gene is a recessive disease gene (MacArthur, et al., 2012), residual variation intolerance score (RVIS) (Petrovski, et al., 2013), ExAC-based RVIS (Petrovski, et al., 2013), FDR value for preferential LoF depletion among the ExAC population (Petrovski, et al., 2015), gene damage index (GDI) (Itan, et al., 2015), Phred-scaled GDI (Itan, et al., 2015), and gene essentiality (1 for essential gene, -1 for non-essential gene, and 0 for genes with no essentiality annotations) (Georgi, et al., 2013). Notably, some of the gene-level features were calculated based on the allele frequencies of its variants in the population, but these gene features are not the same as variant allele frequencies. Taken together, 9 features were extracted.

**Disorder scores**

SPOT-Disorder (Hanson, et al., 2016) was applied to the original protein sequence and mutant sequence. For the substitution site, we extracted disorder scores of wildtype residues and the score difference between wildtype and mutant residue. Considering that IDRs often expose short linear peptide motifs of about 3–10 amino acids to perform their functions (Van Der Lee, et al., 2014), we also calculated values describing the disorder-related microenvironment, \(i.e.,\) sum of disorder scores of
residues with window sizes of 3 and 9, in the original and mutant protein sequence respectively. Taken together, 6 disorder-related features were extracted.

**Feature standardization**

Feature values may range from small to extremely large numbers. However, if a feature has values in a large range, it would be more likely to have a larger impact on the machine-learning model (Althauser and Wigler, 1972). Therefore, all feature values should be standardized to improve the model performance.

Here, we used the following formula (Z-score transformation) to standardize all feature values:

\[ X' = \frac{X - \mu}{\sigma} \]

(12)

, where \( \mu \) is the mean of values of a feature in all training samples, and \( \sigma \) is the standard deviation. \( X \) and \( X' \) denote the original and standardized feature value of a variant.

**Feature selection**

In this work, we implemented a novel feature selection strategy by combining forward selection and backward elimination, as described as follows (Fig. S1).

(1) All the N features were divided into selected feature subset with n features and candidate feature subset with \( N - n \) features. Initially, the selected feature subset was empty (\( n = 0 \)) and the candidate feature subset contained all the features. A machine learning algorithm (LightGBM in this work) was pre-determined as feature selection backend engine and a model evaluation metric (AUC in this work) was also pre-set for evaluating the goodness of the selected feature subset.

(2) Each feature in the candidate feature subset was iteratively added into the selected feature subset, resulting in \( N - n \) temporary feature subsets with \( n + 1 \) features. For each of them, the backend machine learning engine tried different hyper parameter combinations to train the model in a grouped 10-fold cross-validation (G10FCV) manner. We randomly tried the hyper parameter combinations for a pre-set number (\( i.e., 20 \) in our work) of rounds, and obtained the optimal evaluation metric (AUC) in the cross-validation. After this, we had an AUC for each of the \( N - n \) temporary feature subset, and the subset with the highest AUC was chosen (\( n + 1 \) features).

(3) Step 2 was repeated once again to do another step of forward selection, resulting in a temporary feature subset with \( n + 2 \) features.
(4) The previous step also gave out a ranking of importance in the temporary feature subset. The feature with the least importance was eliminated from the selected feature subset. After this step, the temporary feature subset contain $n + 1$ features.

(5) The AUCs were compared between the selected feature subset with $n$ features (at the beginning of step 2) and the temporary feature subset with $n + 1$ features (at the end of step 4).

a. If the AUC of the model based on the temporary feature subset was not less than that on the selected feature subset, update the selected feature subset with the temporary feature subset with $n + 1$ features and go to step 2 for the next iteration; else go to (b).

b. Re-train the model based on the temporary feature subset with trying more hyper parameter combinations, i.e., 40 (double of that in step 2). If the AUC was not less than that on the selected feature subset, update the selected feature subset with the temporary feature subset and go to step 2 for the next iteration; else go to (c).

c. Add one more feature as described in step 2 so that we get a feature subset with $n+2$ features and one AUC value accordingly. Next add one more feature again so that we get a feature subset with $n+3$ features and another AUC value accordingly. If both of these two AUC values were not less than the AUC of the selected feature subset, update the selected feature subset with the temporary feature subset with $n + 1$ features, and go to step 2 for the next iteration; else stop the feature selection process, resulting in the selected feature subset with $n$ features.

Performance evaluation

The performance of IDRMutPred and other predictors was evaluated by four metrics: accuracy (ACC), Matthew’s Correlation Coefficient (MCC), F1 score, Area Under the receiver operating characteristic Curve (AUC). The definitions of first three metrics are given in the following formulas:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN}$$

(13)

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

(14)
\[
\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}
\]

\[
\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \tag{16}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{17}
\]

In these formulas, TP (true positives) and TN (true negatives) represent the number of correctly predicted disease-associated nsSNVs and neutral ones respectively, and FP (false positives) and FN (false negatives) are the numbers of incorrectly predicted as disease-associated nsSNVs and as neutral ones respectively.

Additionally, the receiver operating characteristic (ROC) curve is a plot of the sensitivity in Y-axis and 1- specificity in X-axis. The area under the ROC curve (AUC) is widely used to visually measure the comprehensive performance in binary classification. AUC ranges from 0 to 1. A random predictor will show its curve as the diagonal line and have an AUC of 0.5. A perfect predictor will show its curve as close as possible to Y axis and the line Y=1, and have an AUC of nearly 1.
## 2. Supplementary Tables

### Table S1 The 175 features for feature selection

| # | Feature name | Description | Database | Filter | Weighted |
|---|-------------|-------------|----------|--------|----------|
| 1 | b9_all_nal | Number of sequences in the alignment | UniRef90 | All | No |
| 2 | b9_all_nal_w | Number of sequences in the alignment | UniRef90 | All | BLAST score |
| 3 | b9_hum_nal | Number of sequences in the alignment | UniRef90 | Human | No |
| 4 | b9_hum_nal_w | Number of sequences in the alignment | UniRef90 | Human | BLAST score |
| 5 | b9_nhu_nal | Number of sequences in the alignment | UniRef90 | No Human | No |
| 6 | b9_nhu_nal_w | Number of sequences in the alignment | UniRef90 | No Human | BLAST score |
| 7 | b9_eva_nal | Number of sequences in the alignment | UniRef90 | E-value < 10E-45 | No |
| 8 | b9_eva_nal_w | Number of sequences in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 9 | b9_all_naa | Number of amino acids at nsSNV site in the alignment | UniRef90 | All | No |
| 10 | b9_all_naa_w | Number of amino acids at nsSNV site in the alignment | UniRef90 | All | BLAST score |
| 11 | b9_hum_naa | Number of amino acids at nsSNV site in the alignment | UniRef90 | Human | No |
| 12 | b9_hum_naa_w | Number of amino acids at nsSNV site in the alignment | UniRef90 | Human | BLAST score |
| 13 | b9_nhu_naa | Number of amino acids at nsSNV site in the alignment | UniRef90 | No Human | No |
| 14 | b9_nhu_naa_w | Number of amino acids at nsSNV site in the alignment | UniRef90 | No Human | BLAST score |
| 15 | b9_eva_naa | Number of amino acids at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | No |
| 16 | b9_eva_naa_w | Number of amino acids at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 17 | b9_all_nrt | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | All | No |
| 18 | b9_all_nrt_w | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | All | BLAST score |
| 19 | b9_hum_nrt | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | Human | No |
| 20 | b9_hum_nrt_w | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | Human | BLAST score |
| 21 | b9_nhu_nrt | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | No Human | No |
| 22 | b9_nhu_nrt_w | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | No Human | BLAST score |
| 23 | b9_eva_nrt | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | No |
| 24 | b9_eva_nrt_w | Number of wildtype residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 25 | b9_all_rnt | Number of mutant residues at nsSNV site in the alignment | UniRef90 | All | No |
| 26 | b9_all_rnt_w | Number of mutant residues at nsSNV site in the alignment | UniRef90 | All | BLAST score |
| 27 | b9_hum_rnt | Number of mutant residues at nsSNV site in the alignment | UniRef90 | Human | No |
| 28 | b9_hum_rnt_w | Number of mutant residues at nsSNV site in the alignment | UniRef90 | Human | BLAST score |
| 29 | b9_nhu_rnt | Number of mutant residues at nsSNV site in the alignment | UniRef90 | No Human | No |
| 30 | b9_nhu_rnt_w | Number of mutant residues at nsSNV site in the alignment | UniRef90 | No Human | BLAST score |
| 31 | b9_eva_rnt | Number of mutant residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | No |
| 32 | b9_eva_rnt_w | Number of mutant residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 33 | b9_all_nwt | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | All | No |
| 34 | b9_all_nwt_w | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | All | BLAST score |
| 35 | b9_hum_nwt | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | Human | No |
| 36 | b9_hum_nwt_w | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | Human | BLAST score |
|   |   |   |   |   |
|---|---|---|---|---|
|   | b9_rnu_rwt | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | No Human | No |
| 37 | b9_rnu_rwt_w | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | No Human | BLAST score |
| 38 | b9_eva_rwt | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | No |
| 39 | b9_eva_rwt_w | Proportion of wildtype residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 40 | b9_all_rmt | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | All | No |
| 41 | b9_all_rmt_w | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | All | BLAST score |
| 42 | b9_rhum_rmt | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | Human | No |
| 43 | b9_rhum_rmt_w | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | Human | BLAST score |
| 44 | b9_all_rmt | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | No Human | No |
| 45 | b9_rnu_rmt | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | No Human | BLAST score |
| 46 | b9_rnu_rmt_w | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | No |
| 47 | b9_eva_rmt | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 48 | b9_eva_rmt_w | Proportion of mutant residues at nsSNV site in the alignment | UniRef90 | E-value < 10E-45 | BLAST score |
| 49 | b9_all_pwm | Position Weight Matrix score | UniRef90 | All | No |
| 50 | b9_all_pwm_w | Position Weight Matrix score | UniRef90 | All | BLAST score |
| 51 | b9_hrum_pwm | Position Weight Matrix score | UniRef90 | Human | No |
| 52 | b9_hrum_pwm_w | Position Weight Matrix score | UniRef90 | Human | BLAST score |
| 53 | b9_rnh_pwm | Position Weight Matrix score | UniRef90 | No Human | No |
| 54 | b9_rnh_pwm_w | Position Weight Matrix score | UniRef90 | No Human | BLAST score |
| 55 | b9_eva_pwm | Position Weight Matrix score | UniRef90 | E-value < 10E-45 | No |
| 56 | b9_eva_pwm_w | Position Weight Matrix score | UniRef90 | E-value < 10E-45 | BLAST score |
| 57 | b9_all_p62 | Position Weight Matrix score multiplied by BLOSUM62 score. | UniRef90 | All | No |
| 58 | b9_all_p62_w | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | All | BLAST score |
| 59 | b9_hrum_p62 | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | Human | No |
| 60 | b9_hrum_p62_w | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | Human | BLAST score |
| 61 | b9_rnh_p62 | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | No Human | No |
| 62 | b9_rnh_p62_w | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | No Human | BLAST score |
| 63 | b9_eva_p62 | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | E-value < 10E-45 | No |
| 64 | b9_eva_p62_w | Position Weight Matrix score multiplied by BLOSUM62 score | UniRef90 | E-value < 10E-45 | BLAST score |
| 65 | b9_all_pmi | Position Weight Matrix multiplied by Miyata score | UniRef90 | All | No |
| 66 | b9_all_pmi_w | Position Weight Matrix multiplied by Miyata score | UniRef90 | All | BLAST score |
| 67 | b9_hrum_pmi | Position Weight Matrix multiplied by Miyata score | UniRef90 | Human | No |
| 68 | b9_hrum_pmi_w | Position Weight Matrix multiplied by Miyata score | UniRef90 | Human | BLAST score |
| 69 | b9_rnh_pmi | Position Weight Matrix multiplied by Miyata score | UniRef90 | No Human | No |
| 70 | b9_rnh_pmi_w | Position Weight Matrix multiplied by Miyata score | UniRef90 | No Human | BLAST score |
| 71 | b9_eva_pmi | Position Weight Matrix multiplied by Miyata score | UniRef90 | E-value < 10E-45 | No |
| 72 | b9_eva_pmi_w | Position Weight Matrix multiplied by Miyata score | UniRef90 | E-value < 10E-45 | BLAST score |
| 73 | b9_all_ree | Relative Entropy | UniRef90 | All | No |
| 74 | b9_hrum_ree | Relative Entropy | UniRef90 | Human | No |
| 75 | b9_rnh_ree | Relative Entropy | UniRef90 | No Human | No |
| 76 | b9_eva_ree | Relative Entropy | UniRef90 | E-value < 10E-45 | No |
| 77 | b1_all_nal | Number of sequences in the alignment | UniRef100 | All | No |
| 78 | b1_all_nal_w | Number of sequences in the alignment | UniRef100 | All | BLAST score |
|    |                |                                    |          |            |          |
|----|----------------|------------------------------------|----------|------------|----------|
| 79 | b1_hum_rmt     | Number of sequences in the alignment | UniRef100 | Human      | No       |
| 80 | b1_hum_rmt_w   | Number of sequences in the alignment | UniRef100 | Human      | BLAST score |
| 81 | b1_rhu_rmt     | Number of sequences in the alignment | UniRef100 | No Human   | No       |
| 82 | b1_rhu_rmt_w   | Number of sequences in the alignment | UniRef100 | No Human   | BLAST score |
| 83 | b1_eva_rmt     | Number of sequences in the alignment | UniRef100 | E-value < 10E-75 | No |
| 84 | b1_eva_rmt_w   | Number of sequences in the alignment | UniRef100 | E-value < 10E-75 | BLAST score |
| 85 | b1_all_rmt     | Number of amino acids at nsSNV site in the alignment | UniRef100 | All        | No       |
| 86 | b1_all_rmt_w   | Number of amino acids at nsSNV site in the alignment | UniRef100 | All        | BLAST score |
| 87 | b1_rhu_rmt     | Number of amino acids at nsSNV site in the alignment | UniRef100 | Human      | No       |
| 88 | b1_rhu_rmt_w   | Number of amino acids at nsSNV site in the alignment | UniRef100 | Human      | BLAST score |
| 89 | b1_eva_rmt     | Number of amino acids at nsSNV site in the alignment | UniRef100 | No Human   | No       |
| 90 | b1_eva_rmt_w   | Number of amino acids at nsSNV site in the alignment | UniRef100 | No Human   | BLAST score |
| 91 | b1_all_rmt     | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | No |
| 92 | b1_all_rmt_w   | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | BLAST score |
| 93 | b1_all_rmt     | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | All        | No       |
| 94 | b1_all_rmt_w   | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | All        | BLAST score |
| 95 | b1_hum_rmt     | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | Human      | No       |
| 96 | b1_hum_rmt_w   | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | Human      | BLAST score |
| 97 | b1_rhu_rmt     | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | No Human   | No       |
| 98 | b1_rhu_rmt_w   | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | No Human   | BLAST score |
| 99 | b1_eva_rmt     | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | No |
|100 | b1_eva_rmt_w   | Number of wildtype residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | BLAST score |
|101 | b1_all_rmt     | Number of mutant residues at nsSNV site in the alignment | UniRef100 | All        | No       |
|102 | b1_all_rmt_w   | Number of mutant residues at nsSNV site in the alignment | UniRef100 | All        | BLAST score |
|103 | b1_hum_rmt     | Number of mutant residues at nsSNV site in the alignment | UniRef100 | Human      | No       |
|104 | b1_hum_rmt_w   | Number of mutant residues at nsSNV site in the alignment | UniRef100 | Human      | BLAST score |
|105 | b1_rhu_rmt     | Number of mutant residues at nsSNV site in the alignment | UniRef100 | No Human   | No       |
|106 | b1_rhu_rmt_w   | Number of mutant residues at nsSNV site in the alignment | UniRef100 | No Human   | BLAST score |
|107 | b1_eva_rmt     | Number of mutant residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | No |
|108 | b1_eva_rmt_w   | Number of mutant residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | BLAST score |
|109 | b1_all_rmt     | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | All        | No       |
|110 | b1_all_rmt_w   | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | All        | BLAST score |
|111 | b1_hum_rmt     | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | Human      | No       |
|112 | b1_hum_rmt_w   | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | Human      | BLAST score |
|113 | b1_rhu_rmt     | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | No Human   | No       |
|114 | b1_rhu_rmt_w   | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | No Human   | BLAST score |
|115 | b1_eva_rmt     | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | No |
|116 | b1_eva_rmt_w   | Proportion of wildtype residues at nsSNV site in the alignment | UniRef100 | E-value < 10E-75 | BLAST score |
|117 | b1_all_rmt     | Proportion of mutant residues at nsSNV site in the alignment | UniRef100 | All        | No       |
|118 | b1_all_rmt_w   | Proportion of mutant residues at nsSNV site in the alignment | UniRef100 | All        | BLAST score |
|119 | b1_hum_rmt     | Proportion of mutant residues at nsSNV site in the alignment | UniRef100 | Human      | No       |
|120 | b1_hum_rmt_w   | Proportion of mutant residues at nsSNV site in the alignment | UniRef100 | Human      | BLAST score |
| Row | Symbol       | Description                                                                 | UniRef100       | No Human | BLAST score |
|-----|--------------|------------------------------------------------------------------------------|-----------------|----------|-------------|
| 121 | b1_rhu_rmt   | Proportion of mutant residues at nsSNV site in the alignment                  | UniRef100       | No       |             |
| 122 | b1_rhu_rmt_w | Proportion of mutant residues at nsSNV site in the alignment                  | UniRef100       | No       | BLAST score |
| 123 | b1_eva_rmt   | Proportion of mutant residues at nsSNV site in the alignment                  | UniRef100       | E-value < 10E-75 | No         |
| 124 | b1_eva_rmt_w | Proportion of mutant residues at nsSNV site in the alignment                  | UniRef100       | E-value < 10E-75 | BLAST score |
| 125 | b1_all_pwm   | Position Weight Matrix score.                                               | UniRef100       | No       |             |
| 126 | b1_all_pwm_w | Position Weight Matrix score.                                               | UniRef100       | No       |             |
| 127 | b1_hum_pwm   | Position Weight Matrix score.                                               | UniRef100       | Human    |             |
| 128 | b1_hum_pwm_w | Position Weight Matrix score.                                               | UniRef100       | Human    |             |
| 129 | b1_rhu_pwm   | Position Weight Matrix score.                                               | UniRef100       | No       |             |
| 130 | b1_rhu_pwm_w | Position Weight Matrix score.                                               | UniRef100       | No       | BLAST score |
| 131 | b1_eva_pwm   | Position Weight Matrix score.                                               | UniRef100       | E-value < 10E-75 | No         |
| 132 | b1_eva_pwm_w | Position Weight Matrix score.                                               | UniRef100       | E-value < 10E-75 | BLAST score |
| 133 | b1_all_p62   | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | No       |             |
| 134 | b1_all_p62_w | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | No       |             |
| 135 | b1_hum_p62   | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | No       |             |
| 136 | b1_hum_p62_w | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | No       |             |
| 137 | b1_rhu_p62   | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | No       |             |
| 138 | b1_rhu_p62_w | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | No       |             |
| 139 | b1_eva_p62   | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | E-value < 10E-75 | No         |
| 140 | b1_eva_p62_w | Position Weight Matrix multiplied by BLOSUM62 score.                        | UniRef100       | E-value < 10E-75 | No         |
| 141 | b1_all_pmi   | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | No       |             |
| 142 | b1_all_pmi_w | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | No       |             |
| 143 | b1_hum_pmi   | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | No       |             |
| 144 | b1_hum_pmi_w | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | No       |             |
| 145 | b1_rhu_pmi   | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | No       |             |
| 146 | b1_rhu_pmi_w | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | No       |             |
| 147 | b1_eva_pmi   | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | E-value < 10E-75 | No         |
| 148 | b1_eva_pmi_w | Position Weight Matrix multiplied by Miyata score                           | UniRef100       | E-value < 10E-75 | BLAST score |
| 149 | b1_all_re    | Relative Entropy                                                           | UniRef100       | No       |             |
| 150 | b1_hum_re    | Relative Entropy                                                           | UniRef100       | No       |             |
| 151 | b1_rhu_re    | Relative Entropy                                                           | UniRef100       | No       |             |
| 152 | b1_eva_re    | Relative Entropy                                                           | UniRef100       | E-value < 10E-75 | No         |
| 153 | pos_spo      | SPOT-Disorder score of the wildtype residue at the nsSNV site               |                 | No       |             |
| 154 | pos_spo_3    | Sum of SPOT-Disorder scores of neighboring residues with a window size of 3 |                 | No       |             |
| 155 | pos_spo_9    | Sum of SPOT-Disorder scores of neighboring residues with a window size of 9 |                 | No       |             |
| 156 | dis_spo_d    | Difference of SPOT-Disorder scores between mutant and wildtype residues    |                 | No       |             |
| 157 | dis_spo_3_d  | Difference of SPOT-Disorder scores between mutant and wildtype sequences containing neighboring residues with a window of 3 |                 | No       |             |
| 158 | dis_spo_9_d  | Difference of SPOT-Disorder scores between mutant and wildtype sequences containing neighboring residues with a window of 9 |                 | No       |             |
| 159 | pro_Phi      | Estimated probability of haploinsufficiency of the gene                     |                 | No       |             |
| 160 | pro_Prec     | Estimated probability that the gene is a recessive disease gene              |                 | No       |             |
### Document:

- **pro_RVIS_EVS**: Residual Variation Intolerance Score, a measure of intolerance of mutational burden of the gene (the higher the score the more tolerant to mutational burden the gene is).
- **pro_LoF_FDR_ExAC**: A gene's corresponding FDR value for preferential LoF depletion among the ExAC population (lower FDR corresponds with genes that are increasingly depleted of LoF variants).
- **pro_RVIS_ExAC**: ExAC-based RVIS, with setting 'common' MAF filter at 0.05% in at least one of the six individual ethnic strata from ExAC.
- **pro_GHIS**: A score predicting the gene haploinsufficiency (the higher the score, the more likely the gene is haploinsufficient).
- **pro_GDI**: Gene damage index score, a genome-wide, gene-level metric of the mutational damage that has accumulated in the general population.
- **pro_GDI_Phred**: Phred-scaled GDI score of the gene.
- **pro_Essential_gene**: Essential ("E") or Non-essential phenotype-changing ("N") to indicate the essentiality of a gene.
- **hww_d**: Difference between mutant and wildtype residues in Wimley-White hydrophathy index.
- **hww_3**: Sum of Wimley-White hydrophathy index of neighboring residues with a window of 3.
- **hww_9**: Sum of Wimley-White hydrophathy index of neighboring residues with a window of 9.
- **hwo_d**: Difference of octanol-water free energy transfer index between mutant and wildtype residue.
- **hwo_3**: Sum of octanol-water free energy transfer index of neighboring residues with a window of 3.
- **hwo_9**: Sum of octanol-water free energy transfer index of neighboring residues with a window of 9.
- **blosum62**: Score of the amino acid substitution in the BLOSUM62 matrix.
- **miyata**: Score of the amino acid substitution in the Miyata matrix.

*a#1-152 are the sequence alignment features, #153-158 are disorder features, #159-167 are protein/gene-level annotations feature, #168-173 are hydrophobicity features, and #174-175 are substitution matrix scores.*

### Table S2: The best hyper parameters in each prediction model

| Algorithm  | Hyper parameter                        | Value   |
|------------|----------------------------------------|---------|
| RF         | Maximum tree depth (max_depth)         | 9       |
|            | Fraction of features to try in individual tree (max_features) | 0.183   |
|            | Number of trees to fit (n_estimators)  | 293     |
|            | Random number seed (random_state)      | 58      |
| XGBoost    | Maximum tree depth (max_depth)         | 3       |
|            | Number of trees to fit (n_estimators)  | 141     |
|            | Subsample ratio of columns when constructing each tree (colsample_bytree) | 0.198   |
|            | Random number seed (random_state)      | 40      |
| LightGBM   | Subsample ratio of columns when constructing each tree (colsample_bytree) | 0.35   |
|            | Number of boosted trees to fit (n_estimators) | 87      |
|            | Maximum tree leaves (num_leaves)       | 11      |
|            | Random number seed (random_state)      | 115     |
Table S3 The average performance of G10FCV on the training dataset

| Method               | ACC   | MCC   | F1 score | AUC  |
|----------------------|-------|-------|----------|------|
| RF-based model       | 0.837 | 0.678 | 0.826    | 0.920|
| XGBoost-based model  | 0.847 | 0.695 | 0.841    | 0.929|
| LightGBM-based model | 0.844 | 0.690 | 0.836    | 0.930|

Table S4 Summary of the independent testing datasets before and after removing homologs

| Dataset                     | # of disease nsSNVs | # of neutral nsSNVs | # of IDRs | # of proteins |
|-----------------------------|---------------------|---------------------|-----------|---------------|
| Before removing homologs    | 297                 | 262                 | 313       | 262           |
| After removing homologs a   | 245                 | 204                 | 239       | 206           |

aProteins in the independent testing dataset whose sequences clustered with sequences in the training dataset with $\geq 30\%$ identity by the cd-hit web server were removed.

Table S5 Performance on the independent testing dataset before and after removing homologs

| Testing dataset            | Method               | ACC   | MCC   | F1 score | AUC  |
|----------------------------|----------------------|-------|-------|----------|------|
| Before removing homologs   | RF-based model       | 0.857 | 0.716 | 0.861    | 0.927|
|                           | XGBoost-based model  | 0.859 | 0.719 | 0.863    | 0.934|
|                           | LightGBM-based model | 0.868 | 0.737 | 0.872    | 0.931|
| After removing homologs    | RF-based model       | 0.855 | 0.716 | 0.860    | 0.934|
|                           | XGBoost-based model  | 0.864 | 0.735 | 0.868    | 0.946|
|                           | LightGBM-based model | 0.871 | 0.748 | 0.874    | 0.939|
Table S6 Performance comparison with MetaSVM, MetaLR, and M-CAP

| Method         | Independent testing dataset | Third-party dataset |
|----------------|------------------------------|---------------------|
|                | ACC  | MCC  | F1 score | AUC  | ACC  | MCC  | F1 score | AUC  |
| MetaSVM        | 0.778| 0.526| 0.608    | 0.839| 0.849| 0.720| 0.828    | 0.902|
| MetaLR         | 0.820| 0.661| 0.820    | 0.939| 0.853| 0.722| 0.841    | 0.941|
| M-CAP          | 0.792| 0.560| 0.834    | 0.866| 0.828| 0.548| 0.885a   | 0.884|
| RF-based modelb| 0.880| 0.759| 0.888    | 0.951| 0.879| 0.759| 0.879    | 0.957a|
| XGBoost-based modelb| 0.891b | 0.781b | 0.898b   | 0.954b | 0.877| 0.755| 0.878    | 0.955|
| LightGBM-based modelb| 0.882 | 0.763| 0.889    | 0.952| 0.881b| 0.763b| 0.881    | 0.956|

a The best value in each metric (column) is underlined.
b Our models were re-trained by adding allele frequency (AF) into the optimal feature subset for the fair comparison, as MetaSVM and MetaLR used AF as one of their features, and M-CAP used the prediction scores of MetaSVM and MetaLR as their features.

Table S7 Performance comparison on the dataset whose proteins contains both disease and neutral variants

| Method         | ACC  | MCC  | F1 score | AUC  |
|----------------|------|------|----------|------|
| SIFT           | 0.634| 0.257| 0.714    | 0.709|
| pph2-HumDiv    | 0.610| 0.254| 0.686    | 0.687|
| pph2-HumVar    | 0.604| 0.311| 0.666    | 0.745|
| PhD-SNP        | 0.570| 0.361| 0.599    | .a   |
| MutationAssessor| 0.618| 0.311| 0.681    | 0.768|
| fathmm-W       | 0.832| 0.597| 0.882    | 0.900|
| fathmm-U       | 0.529| 0.223| 0.573    | 0.676|
| PON-P2         | 0.830| 0.529| 0.889    | 0.864|
| PROVEAN        | 0.537| 0.278| 0.570    | 0.676|
| PANTHER-PSEP   | 0.443| 0.001| 0.530    | 0.609|
| Eigen          | 0.588| 0.272| 0.651    | 0.709|
| REVEL          | 0.643| 0.439| 0.687    | 0.897|
| PMut2017       | 0.682| 0.387| 0.525    | 0.753|
| MutPred2       | 0.481| 0.263| 0.475    | 0.738|
| CADD           | 0.647| 0.285| 0.725    | 0.718|
| LIST           | 0.781| 0.380| 0.859    | 0.777|
| RF-based modelb| 0.846b| 0.625b| 0.892    | 0.910|
| XGBoost-based modelb| 0.846b| 0.617| 0.894b   | 0.915b|
| LightGBM-based modelb| 0.846b| 0.616| 0.894b   | 0.914|

a No AUC was calculated for PhD-SNP due to lack of continuous prediction scores.
b The best value in each metric (column) is underlined.
**Table S8** Versions of python packages

| Package name    | Version | Package name     | Version |
|-----------------|---------|------------------|---------|
| numpy           | 1.15.4  | imbalanced-learn | 0.4.3   |
| scipy           | 1.1.0   | xgboost          | 0.82    |
| pandas          | 0.23.4  | lightgbm         | 2.2.1   |
| scikit-learn    | 0.20.1  | biopython        | 1.72    |
3. Supplementary Figures

Start with \( n = 0 \) selected features

**Step 1:** selected feature subset (n)

**Step 2:** add each feature in candidate feature subset (N-n) to selected feature subset separately in order to form N-n temporary feature subsets and select the one with the best AUC by hyper-parameter tuning and G10FCV

**Step 3:** repeat above step once again

\( n+2 \) selected features

**Step 4:** remove the feature with least importance according to the feature importance of lightGBM

\( n+1 \) selected features

Yes: \( n <= n+1 \)

**Step 5:** keep the \( n+1 \) features (criteria in texts)?

No

Terminate with the optimal feature subset (n)

**Fig. S1** Flowchart of feature selection.
Fig. S2 Performance comparison based on four metrics (A, C) and the ROC curves (B, D) on the independent testing dataset.
Fig. S3 Performance comparison based on four metrics (A, C) and the ROC curves (B, D) on the third-party dataset.
Fig. S4 Rank of the feature importance in the XGBoost-based model (A) and the LightGBM-based model (B).
The importance is defined as the average gain of splits that use the feature (A) or the number of times that the feature is used in the model (B).
Fig. S5 The distributions of standardized Z-scores of the 17 selected features.
References

Althauser, R.P. and Wigler, M. Standardization and component analysis. *Sociological Methods & Research* 1972;1(1):97-135.

Altschul, S.F., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-3402.

Eisenberg, D. and McLachlan, A.D. Solvation energy in protein folding and binding. *Nature* 1986;319(6050):199-203.

Georgi, B., Voight, B.F. and Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* 2013;9(5):e1003484.

Hanson, J., et al. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* 2016;33(5):685-692.

Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* 1992;89(22):10915-10919.

Huang, N., et al. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* 2010;6(10):e1001154.

Itan, Y., et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. U. S. A.* 2015;112(44):13615-13620.

Liu, X., et al. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* 2016;37(3):235-241.

Lopez-Ferrando, V., et al. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res.* 2017;45(W1):W222-W228.

MacArthur, D.G., et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012;335(6070):823-828.

Miyata, T., Miyazawa, S. and Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* 1979;12(3):219-236.

Petrovski, S., et al. The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *Plos Genetics* 2015;11(9).

Petrovski, S., et al. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 2013;9(8):e1003709.

Steinberg, J., et al. Haploinsufficiency predictions without study bias. *Nucleic Acids Res.* 2015;43(15):e101-e101.

Suzek, B.E., et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31(6):926-932.

The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47(D1):D506-D515.

Van Der Lee, R., et al. Classification of intrinsically disordered regions and proteins. *Chemical reviews* 2014;114(13):6589-6631.

Wimley, W.C. and White, S.H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* 1996;3(10):842-848.