OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–14
https://doi.org/10.1093/bib/bbac014
Problem Solving Protocol

SVPath: an accurate pipeline for predicting the pathogenicity of human exon structural variants

Yaning Yang, Xiaoqi Wang, Deshan Zhou, Dong-Qing Wei and Shaoliang Peng
Corresponding author. Shaoliang Peng, College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; School of Computer Science, National University of Defense Technology, Changsha 410073, China; Peng Cheng Lab, Shenzhen 518000, China. Tel: +86 0731 88822273; Fax: +86 0731 88664153; E-mail: slpeng@hnu.edu.cn

Abstract

Although there are a large number of structural variations in the chromosomes of each individual, there is a lack of more accurate methods for identifying clinical pathogenic variants. Here, we proposed SVPath, a machine learning-based method to predict the pathogenicity of deletions, insertions and duplications structural variations that occur in exons. We constructed three types of annotation features for each structural variation event in the ClinVar database. First, we treated complex structural variations as multiple consecutive single nucleotide polymorphisms events, and annotated them with correlation scores based on single nucleic acid substitutions, such as the impact on protein function. Second, we determined which genes the variation occurred in, and constructed gene-based annotation features for each structural variation. Third, we also calculated related features based on the transcriptome, such as histone signal, the overlap ratio of variation and genomic element definitions, etc. Finally, we employed a gradient boosting decision tree machine learning method, and used the deletions, insertions and duplications in the ClinVar database to train a structural variation pathogenicity prediction model SVPath. These structural variations are clearly indicated as pathogenic or benign. Experimental results show that our SVPath has achieved excellent predictive performance and outperforms existing state-of-the-art tools. SVPath is very promising in evaluating the clinical pathogenicity of structural variants. SVPath can be used in clinical research to predict the clinical significance of unknown pathogenicity and new structural variation, so as to explore the relationship between diseases and structural variations in a computational way.

Keywords: structural variation, SNP, clinical pathogenic, machine learning, exome

Introduction

Compared with single nucleotide polymorphisms (SNPs), structural variation (SV) has a greater impact on living organisms [1]. Recent studies have further shown that many diseases and phenotypic differences are related to the structural variation of the genome [2]. In the genome of each individual, the number of nucleotides affected by structural variations may be as many as millions [3]. But determining whether a structural variant event is pathogenic is very challenging. The first premise is to compare the structural variation of the diseased population with the healthy population. The diseased population may carry too many pathogenic structural variations [4, 5]. Another strategy is to determine the inherited structural variation through the affected family [6]. However, these two strategies may require decades of effort to determine the pathogenic structural variations and the mechanism of action [7].

Another way to identify the pathogenicity of a variant event is to predict it by calculation. For example, Kircher et al. [8] proposed Combined Annotation-Dependent Depletion (CADD), which pre-calculated the *C* scores of all 8.6 billion possible SNPs by using a support vector machine (SVM) method. The *C* scores are related to allelic diversity, pathogenicity and complex trait associations. Ioannidis et al. [9] developed rare exome variant ensemble learner (REVEL), a comprehensive method for predicting the pathogenicity of missense variants based on multiple tools, providing pre-calculated REVEL scores for all possible human missense variants. Jagadeesh et al. [10] used the gradient boosting trees method to classify the pathogenicity of rare missense variants by integrating related pathogenicity scores (including SIFT [11], Polyphen-2 [12] and CADD [8]) and new feature values. Another part of the related research is based on the gene level rather than the mutation level. They explore the correlation scores between genes and diseases by introducing the correlation between proteins, genes and diseases, so as to predict pathogenic genes, such as [13–17]. Alyousfi et al. [18] established a model

Yaning Yang is a PhD candidate in the College of Computer Science and Electronic Engineering, Hunan University. His research interests include bioinformatics, machine learning algorithms, and high performance computing.
Xiaoqi Wang is a PhD candidate in the College of Computer Science and Electronic Engineering, Hunan University. His research interests include bioinformatics, deep learning algorithms for biomedical network data.
Deshan Zhou is a master student in the College of Computer Science and Electronic Engineering, Hunan University. He works on computational biology and deep learning. Dong-Qing Wei is a tenured professor at the School of Life Science and Technology, Shanghai Jiao Tong University. He has long been engaged in research in the fields of bioinformatics and artificial intelligence- assisted drug design.
Shaoliang Peng is a professor in the College of Computer Science and Electronic Engineering, Hunan University. His research interests include biomedical big data, computer aided drug discovery and high- performance computing.
Received: September 24, 2021. **Revised:** January 11, 2022. **Accepted:** January 12, 2022

to prioritize the identification of single-gene disease genes by integrating gene-level predictors and combining essentiality-specific pathogenicity prioritization (ESPP) scores. Several variant annotation tools are designed to predict the impact of SNP mutations, such as snpEFF [19] and VEP [20].

However, most of the current methods like these are used to predict the pathogenicity of SNPs, there are few tools for structural variations. In fact, due to the large variation length, structural variation plays a vital role in the occurrence and development of various diseases [21]. SVScore [22] uses the precomputed SNP scores from CADD and applies an operation (such as maximun, sum, mean and mean of the top N scores) to predict structural variation impact. One limitation is that SVScore does not consider other types of important information on the genome, and it cannot predict the pathogenicity probability of variants like the SNP pathogenicity prediction method. SVFX [23] integrates a variety of features, such as average histone marker signal, CTCF signal, methylation level and other data, and uses random forest machine learning methods to evaluate a pathogenicity score for deletions and duplications structural variations. But SVFX only considers epigenome feature data, and ignores the influence of structural variation on other biological features, such as protein structure or function.

Structural variations in exons are likely to occur in the position of the protein encoded by the gene. Such variations are likely to cause gene function inactivation and lead to diseases. Here, we propose SVPath to predict the pathogenicity of deletions, insertions and duplications structural variations in human exomes. We collected feature data related to mutation from various aspects to train an ensemble supervised machine learning framework. These feature data mainly include mutation scores based on single nucleic acid substitution, gene-level scores and transcriptome-based related features. We hypothesized that a structural variation is caused by a series of SNP events to take into account the impact of every base substitution in the structural variation on cell activity. Experimental results show that SVPath exhibits excellent prediction and generalization capabilities. SVPath still shows stable prediction performance on two independent test sets, which is very important for solving the pathogenicity prediction of structural variants on exons.

Materials and methods

Data sets

The structural variation data used to train the pathogenicity prediction model in this paper are all from the ClinVar database [24] and dbVar [25]. The ClinVar integrated dbSNP [26], OMIM [27] and other databases of genomic variants and related phenotypes, as well as their clinical data and information. ClinVar is a standard, accurate and reliable database of mutation information and clinical information. We eliminated the variant data whose review status was no assertion, single submitter and conflicting interpretations from ClinVar to make clinical variant data more reliable. Because of the large difference between the number of pathogenic and benign variants in ClinVar, we screened out three structural variant events that were clearly marked as benign from dbVar, making the sample size more balanced. dbVar contains individual instances of structural variation observed in the study, based on the output of raw data analysis. The variation data used in this paper are based on the GRCh37(hg19) reference genome. The types of genomic variation in ClinVar are divided into multiple types according to clinical information, including:

- Uncertain_significance
- Benign
- Benign/Likely_benign
- Likely_benign
- Pathogenic
- Pathogenic/Likely_pathogenic
- Likely_pathogenic

Table 1. The number of pathogenic and benign variations

	Pathogenic from ClinVar	Benign from ClinVar	Benign from dbVar
Deletion	6291	80	4396
Insertion	514	127	9
Duplication	2519	32	968

First, we use the ANNOVAR [28] variant annotation tool to annotate the ClinVar and dbVar variant data to filter out the variants that occur on the exons. Second, we filter out the structural variations of deletions, insertions and duplications. Finally, we further filter out the above three structural variation events with clinical signals Benign, Benign/Likely_benign, Pathogenic and Pathogenic/Likely_pathogenic. And, we regard Benign/-Likely_benign as Benign, Pathogenic/Likely_pathogenic as Pathogenic. We have performed a de-redundancy operation on these variations, the number of variations finally selected is shown in Table 1.

Since the imbalance between the number of pathogenic and benign samples, we employed an oversampling method to expand the benign variants. Although a database DGV (The Database of Genomic Variants) [29] only contains variant events in healthy people, a variant in DGV does not mean that it will not cause disease in patient samples. Therefore, we did not use the variation in DGV as a benign control group. Two independent mutation data sets, gnomAD (Genome Aggregation Database) [30] and DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources) [31], are both used to test the prediction performance of SVPath. (These two independent mutation data sets were retrieved manually, see the supplementary materials for specific retrieval methods).

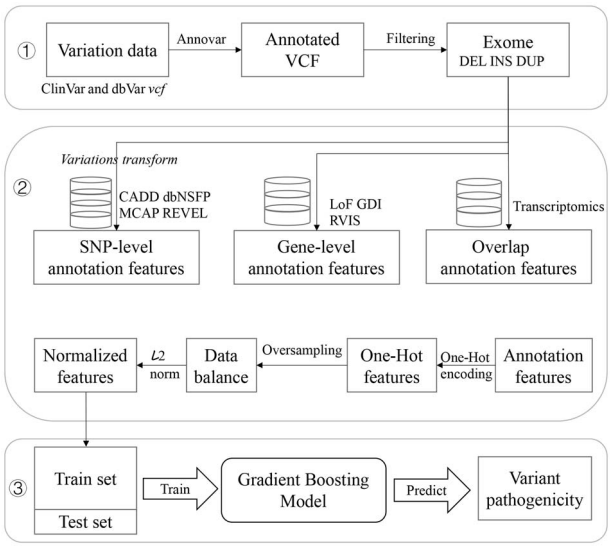


Figure 1. SVPath pipeline. In the first step, we obtained the structural variation of deletion, insertion and repeat type that occurred in the exon from ClinVar. In the second step, we construct features for each structural variation from three aspects. These features are derived from the existing database on the effects of variation. In this process, we speed up feature annotation in a multi-process parallel manner. And one-hot encoding and normalization of the annotated features are performed. In the third step, we employ a machine learning method called gradient boosting decision tree to train a structural variant pathogenicity prediction model.

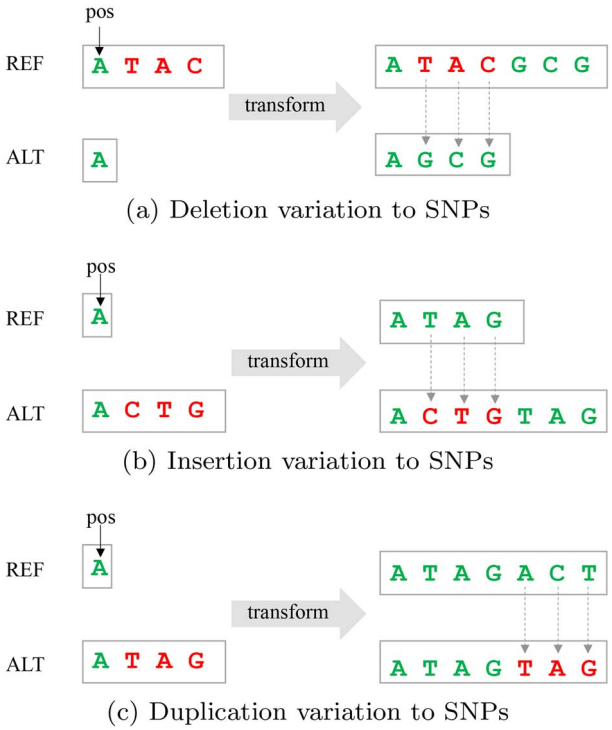


Figure 2. Convert structural variation to multiple consecutive SNPs.

We excluded the mutations shared with ClinVar and dbVar from these two data sets. However, the mutation data in DECIPHER is based on GRCh38. During the test, we used the Remap tool provided by NCBI to map the variation sites to the hg19 reference genome.

SVPath pipeline

SVPath is implemented with the Gradient Boosting machine learning classifier model, the specific process can be seen in Figure 1. First, we obtain the original disease-related human genome variations data from ClinVar, and then use Annovar to annotate it and filter out the pathogenic and benign (clinically verified) deletion, insertion and duplication type structural variations that occur on the exons. Second, we introduced multiple features for these three types of structural variations, mainly including SNP-level features data (after converting structural variations into multiple consecutive SNP events), gene-level features and transcriptomics related features. Then post-process the annotated features data, including one-hot encoding, oversampling and normalization. Finally, we use the processed structural variation data to train a Gradient Boosting machine learning classifier model to predict the pathogenicity of unknown deletions, insertions and repetitive structural variations that occur in exons.

Variations transform

Similar to the SVScore [22] method, the calculation of part of the features of our SVPath is also based on the existing scores of the pathogenicity of a single base or the impact on the protein structure. It's just that SVScore is only based on CADD scores, and it uses the tabix [32] method to calculate a variant's interval scores. In this paper, we treat a structural variation event as multiple consecutive SNPs.

The structure variation transform method is shown in Figure 2. REF is the reference genome segment (or a base) on the *pos* coordinate, and ALT is alternate base(s), which means the allele of the variant. Suppose the length of the structural variation is *len*, which is the absolute value of the difference between the length of REF and ALT. For a deletion variation, we need to obtain the genome fragment from *pos+len+1* coordinates to *pos+2*len* from the reference genome, which we call snpREF. As shown in Figure 2a, the missing sequence is TAC, which we call snpALT, and snpREF is GCG. Here, the length of snpREF must be the same as the length of snpALT. Then the deletion structural variation (*pos:ATAC-A*) can be converted to (*pos+1:T-A*, *pos+2:A-C*, *pos+3:C-G*). Similarly, the insertion structure variation (*pos:A-ACTG*) in Figure 2b is converted to (*pos+1:T-C*, *pos+2:A-T*, *pos+3:G-G*). But the snpREF in insertion is the sequence from *pos+1* to *pos+len* in the reference genome. The duplication variation in the vcf format file is represented as *pos:A-ATAG* as shown in Figure 2c, but the actual variation is *pos:ATAG-ATAGTAG*, and the duplication fragment is TAG. Therefore, the snpREF in the duplication variation is the sequence from *pos+len+1* to *pos+2*len* in the reference genome. So the mutation in Figure 2c can be transformed into (*pos+len+1:A-T*, *pos+len+2:C-A*, *pos+len+3:T-G*).

The transform of structural variation is to fully consider the impact of each base site on the function of the genome after the structural variation occurs. SVScore

Table 2. Annotation features description based on SNP variant sites

Database	Field	Data type	Aggregation method
CADD [8]	CADD13_RawScore	float	avg
	CADD13_PHRED	float	avg
dbNSFP [33]	SIFT_pred [34]	char	min
	SIFT4G_pred [11]	char	min
	Polyphen2_HDIV_pred [35]	char	max
	Polyphen2_HVAR_pred [35]	char	max
	VEST4_score [36]	float	avg
	MVP_score [37]	float	avg
	MPC_score [38]	float	avg
	DANN_score [39]	float	avg
	GenoCanyon_score [40]	float	avg
	integrated_fitCons_score [41]	float	avg
	GM12878_fitCons_score [41]	float	avg
	H1-hESC_fitCons_score [41]	float	avg
	HUVEC_fitCons_score [41]	float	avg
	LINSIGHT [42]	float	avg
	GERP++_NR [43]	float	avg
	GERP++_RS [43]	float	avg
	phyloP100way_vertebrate [44]	float	avg
	phyloP30way_mammalian [44]	float	avg
	phyloP17way_primate [44]	float	avg
	phastCons100way_vertebrate [45]	float	avg
	phastCons30way_mammalian [45]	float	avg
	phastCons17way_primate [45]	float	avg
M-CAP [10]	MCAP13	float	avg
REVEL [9]	REVEL	float	avg

only considers the CADD scores of the left and right breakpoints of structural variation as features.

Features construction

Different feature values indicate different biological meanings. For example, the main features are the influence of SNPs on protein sequence and function calculated by multiple algorithms, as well as the feature scores of gene function loss and the feature scores of histone signals. The data types of these characteristics are not uniform. Roughly speaking, the feature values constructed for each structural mutation event can be divided into three types: annotation information based on single base, annotation information based on gene function and annotation information based on variation region. For different feature data, we use different construction methods. All feature calculations are based on the hg19 reference genome.

The calculation method based on the annotation feature of the SNP base site is to convert the structural variation event into multiple consecutive SNP events, and then calculate the average value or the maximum or minimum values of the multiple SNP events. The scores of these original single-base site mutations come from CADD v1.3 [8], dbNSFP v4.1a [33], M-CAP v1.3 [10] and REVEL [9] databases, respectively. For the description of the field values used in this paper in these four databases, see Table 2.

We introduced two fields in the CADD database as two feature values, namely *raw* and *PHRED* fields. The *raw* field is the C-score obtained by the combined SVM score. The higher the C-score value, the more harmful the impact of this single-base substitution. The *PHRED* field is the C-score scale, which ranks all possible substitutions (8.6 billion) variants relative to the human genome. For a structural variation, we calculate the average raw and PHRED values of multiple SNPs separately.

The dbNSFP v4.1a contains scores or discrete values calculated by 51 prediction algorithms for functional prediction and annotation of possible non-synonymous SNP events in the human genome. We selected 22 of the scores used to calculate part of the feature value of each structural variation, including 14 functional prediction scores (including SIFT, SIFT4G, Polyphen2_HDIV, Polyphen2_HVAR, VEST, MVP, MPC, DANN, GenoCanyon, four fitCons and LINSIGHT) and eight conservation scores (including GERP_NR, GERP_RS, three phyloP100way scores and three phastCons100way scores). Among the scores of functional prediction, SIFT (Sort Intolerated From Tolerated) and SIFT4G (SIFT for Genomes) are the predicted impact scores of an amino acid substitution on protein function, and their values are *D* (Deleterious, value $\leqslant$ 0.05) and *T* (Tolerated, value $>$ 0.05). After a structural variation undergoes variation transform, if the value of one or more mutation sites is *D*, the feature value of the SIFT (or SIFT4G) of this structural variation is *D*, that is, take the smaller value of *D* and T ($D < T$). The calculation methods of Polyphen2_HDIV (Polymorphism phenotyping v2 based on HumDiv [46]) and Polyphen2_HVAR (Polyphen2 based on HumVar [46]) are similar to those of SIFT. The difference is that the higher the score of PolyPhen2, the more harmful it is.

Table 3. Gene-based annotation features

Gene data	Field	Description	Data type
LoFtool	LoFtool_percentile	Gene loss-of-function score percentiles	float
GDI	GDI	Gene damage index	float
	GDI-Phred	Phred-scaled gene damage index scores	float
	All disease-causing genes	Damage prediction for all disease-causing genes	char
	All Mendelian disease-causing genes	Damage prediction for all Mendelian disease-causing genes	char
	Mendelian AD disease-causing genes	Damage prediction for Mendelian autosomal dominant (AD) disease-causing genes	char
	Mendelian AR disease-causing genes	Damage prediction for Mendelian autosomal recessive (AR) disease-causing genes	char
	All PID disease-causing genes	Damage prediction for all primary immunodeficiency (PID) disease-causing genes	char
	PID AD disease-causing genes	Damage prediction for PID AD disease-causing genes	char
	PID AR disease-causing genes	Damage prediction for PID AR disease-causing genes	char
	All cancer disease-causing genes	Damage prediction for all cancer disease-causing genes	char
	Cancer recessive disease-causing genes	Damage prediction for cancer recessive disease-causing genes	char
	Cancer dominant disease-causing genes	Damage prediction for cancer dominant disease-causing genes	char
RVIS-ESV	RVIS_ExAC_0.05%	The Residual Variation Intolerance Score (RVIS) based on the Exome Aggregation Consortium (ExAC) database with 0.05% Minor Allele Frequency (MAF) from any population	float
	%RVIS_ExAC_0.05%	RVIS percentile values that reflect the relative rank of the gene	float

Its value has three discrete values, namely *D* (Probably damaging, value $\geqslant$0.957), *P* (Possibly damaging, 0.453 $\leqslant$ value $\leqslant$ 0.956) and *B* (Benign, value $\leqslant$ 0.452). The aggregation method for the two features of Polyphen2 is to take the maximum value, where the maximum value refers to the maximum value of D, P and B ($D > P > B$). The aggregation method for other functional scores and conservation scores is to take the average of all amino acid substitutions.

Both M-CAP (Mendelian Clinically Applicable Pathogenicity) and REVEL integrate previous pathogenicity scores and use machine learning models to predict the pathogenicity scores of rare missense SNP variants. The integration method of these two features is similar to the previous one.

If a feature in Table 2 is absent during integration, for the CADD features, we use the average of the three possible amino acid substitutions at the left end of the structural variation (pos position) to fill in. For the features in SIFT and Polyphen2, we use *T* and *B* instead of blanks. For the absence of other features, we treat it as 0.

The second type of annotation features is based on gene function, including gene Loss of function (Lof score), gene damage index (GDI) and genetic intolerance of genes to functional mutations (RVIS-ESV). The features information about the gene is shown in Table 3.

First, we need to locate which gene(s) an exon structural variation occurs, and secondly, we use these scores of the current gene as the feature values of the structural variation event. We use a novel gene intolerance ranking system (LoFtool) proposed by Fadista et al. [47] to construct the feature of gene function loss score (named this feature after *Lof_score*). If a structural variation occurs in a gene, the smaller the *Lof_score* of this gene, the less tolerance the variation has to the loss of gene function. The other is the GDI data set [48], which defines the mutation damage accumulated by each human gene encoding protein in the general population at the gene level, including the score of the GDI and the general damage prediction of different disease types. These disease types include all, Mendelian AD (autosomal dominant) and AR (autosomal recessive), all primary immunodeficiency (PID), PID AD and AR, all cancer, cancer recessive and dominant. The predictive values of disease-related gene damage include *Low*, *Medium* and *High*. The third gene level data are RVIS (The Residual Variation Intolerance Score), and the version used in this paper is RVIS based on the Exome Aggregation Consortium (ExAC) database with 0.05% Minor Allele Frequency (MAF) from any population. The score of intolerance is based on the neutral variation found in the gene to assess whether the gene has more or less functional genetic variation than expected. For each gene, the intolerance score and tolerance percentile are included.

Inspired by SVFX [23], we added more annotation features (21 in total) to each mutation event, such as functional genomics data, annotation metrics data and conservation score, see Table 4. Functional genomics data such as the histone marker signal of the H1 human embryonic stem cell line (E003) (including H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3), DNase hypersensitive sites (DHSs) signal, fractional methylation and whole genome methylation data obtained from the Roadmap Epigenomics project [49] and replication timing data and CCCTC-binding factor (CTCF) data obtained from ENCODE project [50]. Annotate measurement data such as protein coding domains (CDS), promoters, 5'UTR and 3'UTR definitions, splice sites, heterochromatin regions, topologically associated domains (TADs) obtained from ENCODE and IHEC [51]. Conservative feature scores

Table 4. Annotation features based on transcriptomics

Annotation type	Data	Description	Data format
Functional genomics	E003-H3K4me1	The mark of monomethylation at the fourth lysine residue of histone H3 protein is usually associated with gene enhancers	BigWig
	E003-H3K4me3	It is a marker that indicates the trimethylation of the fourth lysine residue of histone H3 protein and is often involved in the regulation of gene expression	BigWig
	E003-H3K9me3	Indicates the trimethylation mark at the ninth lysine residue of histone H3 protein, usually related to heterochromatin	BigWig
	E003-H3K27ac	Indicates a marker for acetylation of the 27th lysine residue at the N-terminal of histone H3 protein and it is associated with higher activation of transcription	BigWig
	E003-H3K27me3	It is a label indicating that lysine 27 on histone H3 protein is trimethylated. This trimethylation is related to the down-regulation of nearby genes through the formation of heterochromatin regions	BigWig
	E003-H3K36me3	Indicates the trimethylation mark at the 36th lysine residue of histone H3 protein, usually related to the genomic body	BigWig
	DHSs signal	DNase hypersensitive sites	BigWig
	Fractional methylation	Fractional methylation calls at CpG	BigWig
	Whole genome methylation	Single CpG site resolution whole-genome DNA methylation map generated by whole-genome shotgun bisulfite sequencing (WGBS)	BigWig
	Replication timing	Refers to the order in which DNA fragments are copied along the length of a chromosome	BigWig
	CTCF	CCCTC-binding factor, a transcription factor encoded by the CTCF gene in humans	BigWig
Annotate metrics	CDS	Protein coding domains	bed
	Promoters	A DNA sequence to which a protein binds, and a single RNA can be transcribed from the downstream DNA	bed
	5'UTR	The 5' untranslated region	bed
	3'UTR	The 3' untranslated region	bed
	Splice sites	Recognizable sequence and linker site of intron and exon junction boundary in RNA precursor	bed
	Heterochromatin regions	Refers to regions of chromatin that are condensed and transcriptionally inactive during the interphase	bed
	TADs	Topologically associated domains	bed
Conservative scores	Ultra-conserved regions	Refers to a nucleotide fragment in a DNA molecule or an amino acid fragment in a protein thatremains basically unchanged during evolution	bed
	Sensitive regions	Sensitive regions across the genome, contrary to the meaning of ultra-conserved regions	bed
	phyloP100way	Conservation scoring by phyloP (phylogenetic p-values) for multiple alignments of 99 vertebrate genomes to the human genome	BigWig

such as the annotation of ultra-conserved and sensitive regions of the whole genome, and the scores of 100 cross-species PhyloP.

For each feature of functional genomics and *phyloP100way* in Table 4, we first calculate the corresponding feature value of a structural variation left and right breakpoint. If the variation site does not match the relevant region, then the feature is represented by 0, otherwise we use the weighted average of this feature in the variation interval as the value of this feature. For other features (*bed* format), we use the overlap of the region where the variation occurs with the given region as the value of the feature, that is, calculate the percentage of a certain region that the variation occupies.

In this paper, we use the functional genomics data of the H1 cells to construct some features. For constructing tissue-specific features data, it can be adjusted according to the description in the Roadmap. For details, please refer to https://egg2.wustl.edu/roadmap.

We also take the subtype of variation as its features. After annotating the variation data through ANNOVAR [28], we obtain the subtypes of the variation, which mainly include *frameshift*, *noframeshift*, *startloss*, *stoploss*, *stopgain* variations and other types.

On the one hand, for the feature annotations of the three structural variants of deletion, insertion, and duplication, due to the differences in SNP conversion, we carried out features construction for them separately. On the other hand, when constructing features of a certain type of structural variation, since there is no correlation between the variation events, we execute the construction process in a multi-process parallel manner to speed up the generation of the features matrix.

Data post-processing

After calculating all the features of each structural variation, there are still some problems, such as the data format is not uniform, the positive and negative samples are

not balanced, etc. Therefore, we need to further process it to make the prediction of the pathogenicity of the variant more accurate.

First, we performed One-Hot encoding [52] on the features of the character type, so that the feature value of the character type is converted to a numeric value of 0 or 1. Specifically, for the character-type features, taking the variation subtype as an example, we use all the values of the original variation subtype of a structural variation event as the features of each variation. If the variation belongs to the *frameshift* type, the corresponding *frameshift* feature is 1, and the rest are 0.

Second, we can see from Table 1 that there is a big difference in the number of pathogenic and benign samples, and the data of benign variants are far less than that of pathogenic ones. If such unbalanced data are used to train a machine learning model, it will result in too little feature information provided by the category of benign variants, and the prediction will be biased toward pathogenic variants, which will greatly reduce the generalization ability of the model. In this paper, we use an oversampling method called Borderline SMOTE [53] to sample the benign variant data, so that the data of benign variants and pathogenic variants reach a balance, which makes the prediction model better fit each feature of the two types of variation data.

Third, the features of each structural variation have a large difference in the numerical interval. If the sample is directly used for training, it may cause many iterations to converge, or it may not converge. Therefore, after the above two steps are processed, we normalize each feature element of each variant in the sample to the interval [0,1] with the ℓ^2-norm of the sample.

Classification models

Through the above processing of structural variation data, we can convert the problem of predicting the pathogenicity of unknown structural variation into a problem of binary classification based on prior knowledge. We applied five machine learning models to compare the prediction performance on this feature-annotated structural mutation data set, namely SVM, logistic regression (LR), decision tree (DT), K-nearest neighbors (KNN) and gradient boosting decision tree (GBDT). GBDT [54] is an iterative decision tree algorithm, which consists of multiple decision trees, and the conclusions of all trees are added up to output the final result.

Results

Training and testing SVPath

Before training the model, we performed a de-redundancy operation to eliminate data contamination. In order to accurately assess the predictive performance of SVPath, we used leave-four-chromosomes-out (for deletions and duplications) and leave-eight-chromosomes-out (for insertions) cross-validation methods. In each round of training and verificating, we selected four (about 20%, for deletions and duplications) or eight (about 36%, for insertions, because there are fewer records of insertion variations) chromosomes variations as the test set, the rest as the training set. The variation on the chromosomes reserved in this way can be regarded as a completely independent blind test set, so as to avoid the contamination of the train set and test set. See the supplementary materials for the specific data set division method. During the training process, we adopted an oversampling strategy to balance the number of pathogenic variants and benign variants in the train set. All variations in the test set come from the ClinVar and dbVar databases, not oversampling. The training sets are used to train five machine learning models (SVM, LR, DT, KNN and GBDT), and the test sets are used to assess the prediction performance of the model. At the same time, we also used two existing models for predicting the pathogenicity of structural variants for comparisons, namely SVFX [23] and SVScore [22] with mean operation. In each comparison, the structural variation data used by these two models are the same as that used by the previous machine learning models.

Due to the unbalanced data of pathogenic variants and benign variants, we use relevant indicators based on confusion matrix to evaluate the performance of pathogenicity prediction models, including accuracy (Eq.1), precision (Eq.2), recall (Eq.3), specificity (Eq.4), F1-score (Eq.5), G-mean (Eq.6), Matthews correlation coefficient (MCC)(Eq.7). In the confusion matrix, we treat pathogenic variants as positive cases and benign variants as negative cases. TP represents the predicted result and the actual value are both positive, FP represents the predicted result is positive but the actual value is negative, TN represents the result is negative and FN represents the predicted result is negative but the actual value is positive.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{5}$$

Table 5. Evaluation of several methods on the deletion, insertion and duplication variations

SV type	Model(Method)	Accuracy	Precision	Recall	Specificity	F1-score	G-mean	MCC
Deletion	GBDT	0.979	0.987	0.969	0.987	0.978	0.978	0.956
	SVM	0.806	0.882	0.777	0.874	0.803	0.811	0.651
	Logistic Regression	0.732	0.669	0.911	0.522	0.767	0.689	0.473
	Decision Tree	0.936	0.948	0.920	0.954	0.932	0.936	0.869
	KNeighbors	0.816	0.910	0.705	0.936	0.793	0.812	0.640
	SVScore	0.586	0.616	0.573	0.615	0.584	0.593	0.181
	SVFX	0.806	0.829	0.780	0.838	0.802	0.808	0.602
Insertion	GBDT	0.883	0.858	0.955	0.756	0.897	0.848	0.704
	SVM	0.367	0.658	0.335	0.637	0.370	0.424	-0.019
	Logistic Regression	0.542	0.675	0.612	0.342	0.602	0.429	-0.011
	Decision Tree	0.713	0.823	0.743	0.679	0.767	0.708	0.351
	KNeighbors	0.534	0.684	0.369	0.750	0.420	0.437	0.100
	SVScore	0.511	0.711	0.574	0.440	0.605	0.500	0.018
Duplication	GBDT	0.978	0.988	0.979	0.973	0.983	0.976	0.948
	SVM	0.768	0.963	0.694	0.929	0.796	0.796	0.594
	Logistic Regression	0.709	0.800	0.782	0.600	0.774	0.671	0.389
	Decision Tree	0.948	0.975	0.949	0.952	0.961	0.950	0.877
	KNeighbors	0.744	0.930	0.670	0.894	0.777	0.773	0.524
	SVScore	0.632	0.764	0.660	0.576	0.705	0.616	0.220
	SVFX	0.833	0.882	0.869	0.762	0.874	0.813	0.614

$$G-mean=\sqrt{Recall\times Specificity} \tag{6}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{7}$$

We used the average of the results obtained by cross-validation as the metric when evaluating the performance of each model. The results are shown in Table 5. By comparison, we can see that in the five machine learning models we used, GBDT has the best effect in these seven indicators. First, even if the categories of pathogenic and benign variants are not balanced, the two indicators of GBDT used to evaluate the category imbalanced data sets, G-mean and Matthews correlation coefficient, both achieved high scores and were both higher than other models. Second, even in the case of a small number of samples (insertion variation), the classification effect of GBDT is still considerable. Compared with deletion and duplication, although the MCC value of insertion is significantly lower, only 0.704, it still shows great advantages compared with other models, especially SVM and logistic regression. Therefore, based on the above comprehensive analysis, we choose GBDT machine learning algorithm to build our structural variant pathogenicity prediction model SVPath.

In addition, during each round of training and testing, we saved the test set of each of the previous five machine learning models to test the two existing pathogenicity prediction models, SVScore and SVFX. The experimental results are shown in Table 5. It is obvious that by introducing various annotation features, the indicators in Table 5 of GBDT machine learning models are higher than the ones of SVScore and SVFX. SVScore uses too few features, only considers the calculation of CADD scores, which causes under-fitting to a certain extent. Structural variation is critical to protein synthesis and function, and protein structure and function are closely related to the occurrence of diseases and even cancer, but SVFX does not consider the features of the impact of structural variation on protein function. The lack of SVFX prediction algorithm in the performance evaluation of insertion type in Table 5 is because SVFX only predicts the pathogenicity of deletion and duplication type structural variations. In short, in the two existing methods and the five machine learning models we used, the GBDT method is the best among the three structural variation types. Even in the case of unbalanced labels and a small sample size, the Matthews correlation coefficients of deletion, insertion and duplication variants obtained by GBDT reached 0.956, 0.704, and 0.948, respectively. The reason why GBDT can achieve such high scores is because, first, we have introduced enough features for each structural variation (of course not that more features are better), then the model can learn more relevant information. Second, the GBDT model is a boosting method, which can promote a weak learner to a strong learning algorithm, and build the model in a step-by-step iterative manner.

Statistical significance

Taking SVPath and SVFX testing deletion variations as examples, we performed McNemar test on the statistical significance of the two models. Assume that the number of SVFX classified correctly but SVPath classified incorrectly is e_{01}, and the number of SVPath classified correctly but SVFX classified incorrectly is e_{10}. Give a null hypothesis: SVPath and SVFX classifiers have the same performance in predicting the pathogenicity of deletion

varitions. Then $e_{01} = e_{10}$, the variables $|e_{01} - e_{10}|$ should obey the normal distribution, and the mean value is 1, the variance is $e_{01}+e_{10}$. So the variable

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \tag{8}$$

obeys the χ^2 distribution with one degree of freedom. Take the deletion variations on chromosomes 1, 7, 13, 19 as the test set, and the remaining deletions as the training set (the two are mutually exclusive). The experimental result is that the result of Eq.8 is 550.359. However, when the significance degree $\alpha = 0.05$ is given, the critical value of the χ^2 test with one degree of freedom is 3.84, which is much smaller than the calculated result. And the error rate (1-Accuracy) of SVPath is 0.016, and that of SVFX is 0.213. And from Table 5, the average error rate of SVPath is less than that of SVFX. Through the above analysis, we can reject the null hypothesis and conclude that SVPath is better than SVFX in predicting the pathogenicity of deletion structural variants. Similarly, the statistical significance of SVScore is tested on the same test set, and the value of Eq.8 is 1151.403, the error rate is 0.402. It can be seen that our model SVPath is significantly better than other existing methods in terms of statistical significance. Please refer to the supplementary materials for the McNemar test results of each round of cross-validation.

Features contribution

In order to evaluate the method we put forward in Section 2.3 to transform structural variants into multiple consecutive SNP events, and the impact of these three types of features on the pathogenicity prediction of structural variants, based on the GBDT model, we separately train and test the three types of structural variations with the features in Tables 2, 3 and 4. The previous indicators are also used as the evaluation standard, and the experimental results are shown in Table 6. We can see that when the number of samples is large enough (deletion and duplication type variations), the effect of this feature construction method based on variations transforming is better than the other two types of features. Our variations transforming method takes into account the features value of the nucleic acid substitution in each variation site in the structural variations.

In addition, we explored the contribution of all the features we constructed to the pathogenicity prediction model. Figure 3 shows the top 25 features that contribute to the pathogenicity model (Measured in deletion type variation). We can see that the variation subtype is the most important feature, as well as Methylation signal, LINSIGHT (Protein function score) and so on. For example, the subtype value *stopgain*, this type of variation will cause the chromosome to introduce a stop codon when encoding a protein, which will cause the encoding of the protein to terminate prematurely or the protein cannot be translated. On the contrary, *stoploss* may cause the loss of a stop codon, which will cause the final protein to be longer than the original protein, thereby affecting normal cell activities.

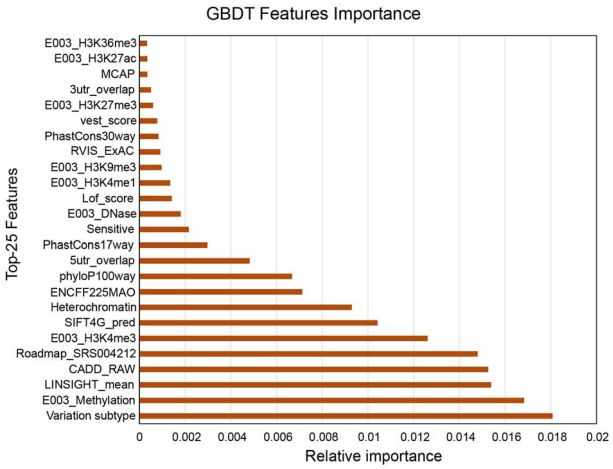


Figure 3. Top 25 relatively important features.

SVPath performance on independent data sets

In the above experiment, some benign variations in the training set are obtained through Borderline SMOTE oversampling, one possible shortcoming of this oversampling technique is that it cannot overcome the data distribution problem of some features, and it is easy to cause the problem of distribution marginalization. Therefore, we tested the performance of SVPath in two separate variations databases, DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources) v11.6 [31] and gnomAD (Genome Aggregation Database) v2.1.1 [55]. These two databases provide clinical pathogenicity explanations for part of the variants data (online version). The variations in DECIPHER are based on GRCh38, so we first converted the coordinates of the structural variations using the Remap tool provided by NCBI. We only selected a part of the three types of structural variants that are pathogenic and benign in DECIPHER. Most of the variations in gnomAD comes from ClinVar, so we manually searched and sorted out some of the pathogenic and benign deletions, insertions and duplications that occur in the exons from gnomAD that are not in ClinVar. The numbers of deletions, insertions and duplications are 934, 356 and 512 in gnomAD, and 628, 152 and 451 in DECIPHER, respectively.

We used all the variants in Table 1 to train the pathogenicity prediction models of these three structural variants. Then use the pre-trained models to test on these two independent data sets. The test results are shown in Table 7. It is obvious from Table 7 that SVPath still performs outstandingly on the two independent test sets. The Matthews correlation coefficients of deletions and duplications variations are still maintained

Table 6. The performance of different types of features based on GBDT model

SV type	Features	Accuracy	Precision	Recall	Specificity	F1-score	G-mean	MCC
Deletion	Table 2	0.913	0.937	0.880	0.945	0.907	0.912	0.819
	Table 3	0.738	0.832	0.533	0.914	0.641	0.684	0.464
	Table 4	0.967	0.976	0.954	0.978	0.965	0.965	0.929
Insertion	Table 2	0.589	0.819	0.560	0.735	0.640	0.636	0.251
	Table 3	0.264	0.470	0.076	0.769	0.112	0.226	-0.199
	Table 4	0.747	0.892	0.671	0.805	0.754	0.722	0.460
Duplication	Table 2	0.911	0.961	0.902	0.922	0.930	0.911	0.796
	Table 3	0.632	0.854	0.514	0.850	0.627	0.648	0.346
	Table 4	0.962	0.979	0.962	0.961	0.970	0.961	0.910

Table 7. The performance on gnomAD and DECIPHER databases

Databases	SV type	Accuracy	Precision	Recall	Specificity	F1-score	G-mean	MCC
gnomAD	Deletion	0.971	0.958	0.988	0.953	0.972	0.970	0.942
	Insertion	0.860	0.969	0.808	0.953	0.881	0.877	0.731
	Duplication	0.955	0.933	0.985	0.922	0.958	0.953	0.911
DECIPHER	Deletion	0.955	0.975	0.955	0.955	0.965	0.955	0.904
	Insertion	0.934	0.976	0.946	0.870	0.961	0.907	0.764
	Duplication	0.953	0.937	0.967	0.941	0.952	0.954	0.907

above 0.9. Due to less insertion variation, the results of insertion variation are difficult to measure. However, compared to the test set reserved in ClinVar and dbVar, the prediction performance of SVPath on the new data set has declined. One of the main reasons is that, in order to achieve a balance between positive and negative samples when training SVPath, some benign variant data are obtained by Borderline SMOTE oversampling based on the existing variation features. The oversampling method divides the benign samples into safe samples, boundary samples and noise samples, and performs nearest neighbor interpolation on the boundary samples. Although the uneven distribution of benign sample data is considered, the difference between boundary samples is not considered. Therefore, the fitting degree of the characteristic data distribution of the benign variation is not good, so that the model cannot better learn the data distribution characteristics of such samples. Of course, a better solution is also our strategy for continuous improvement in future work, which is to mine more and experimentally proven benign structural variations data to train SVPath to achieve better prediction performance of the models.

Predicting unknown pathogenicity of SV in ClinVar

In the ClinVar variant database, there are still a large number of structural variants of unknown pathogenicity. Based on the pathogenicity prediction performance of SVPath verified by the above experiments, here, we use the pre-trained SVPath to predict the pathogenicity of three types of structural variants of unknown pathogenicity. First, we filter out deletion, insertion and duplication structural variants from ClinVar that are of uncertain significance, likely benign, and likely pathogenic. Firstly, we use the ANNOVAR variation annotation tool to annotate the selected variation data and filter out the variations that occur on the exons. Secondly, we used the pre-trained SVPath to predict the pathogenicity of these variants of unknown clinical significance. The prediction results are shown in Table 8. See the supplementary materials for the pathogenicity prediction results of each variant whose pathogenicity is unknown.

We investigated the structural variants that were originally labeled as likely benign and that SVPath predicted to be pathogenic in ClinVar. We found that most of these structural variants that were predicted to be pathogenic were related to certain clinical phenotypes. For example, a deletion variation on chromosome 1, coordinate 55512243 (chr1:55512243 CTT>C) is related to Familiar hypercholesterolemia [56]; the variation of chr3:137484347 from TA to T is related to Anophthalmia-microphthalmia syndrome [57]. Although it cannot be determined from clinical experiments that these structural variants are directly related to specific diseases, our SVPath structural variant pathogenicity prediction model can provide a strong evidence for the correlation between these variants and diseases. SVPath can be used as a reference in clinical experiments exploring the relationship between structural variation and disease.

Running time

First, due to the large number of feature annotation files, and the larger files based on single nucleic acid replacement such as CADD, we have to divide these files into multiple chunks according to the chromosome number and fixed step size (1Mb) to reduce the running

Table 8. Prediction of pathogenicity of variants of unknown clinical importance in ClinVar

Type	Clinical significance	number of variation	SVPath prediction
DEL	Uncertain significance	6484	5396P+1088B
	Likely benign	394	249P+145B
	Likely pathogenic	6442	6014P+428B
INS	Uncertain significance	532	224P+308B
	Likely benign	66	16P+50B
	Likely pathogenic	362	279P+83B
DUP	Uncertain significance	2708	2081P+627B
	Likely benign	160	111P+49B
	Likely pathogenic	2352	2263P+89B

Where, *P* stands for disease-causing and *B* stands for benign.

time of features annotations. Second, since each structural mutation event is uncorrelated, we use a multi-process-based parallel approach to annotate multiple variation events at the same time. The total running time of SVPath depends on two aspects, namely the total number of variations and the length of the variations. We roughly merged the variations in Table 1 several times (a total of 60 765 deletions, 33 450 insertions and 17 600 duplications) to increase the number of variations, and the SV length distribution of the merged variants is shown in . The running time of the three variant annotations is shown in Figure 4a. If the variation length is evenly distributed, the speedup of parallel acceleration is ideal, such as insertion variation. However, there are more large-scale variations in deletions and duplications, and these variations determine the final running time.

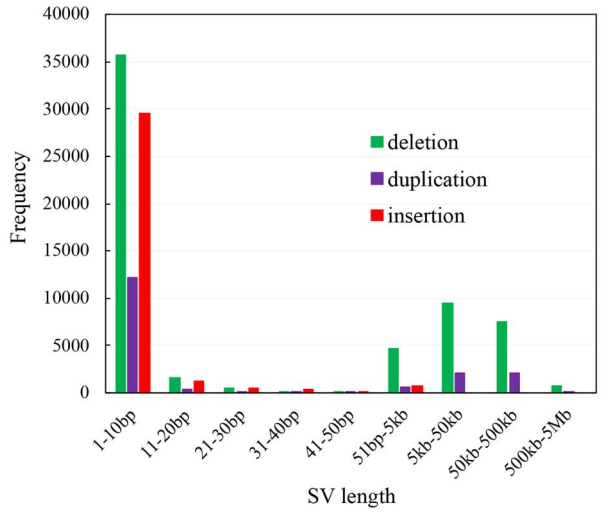
Discussion

Although there have been many studies in the past to try to reveal the relationship between variants and diseases or phenotypes in a computational manner, most of the tools are based on the scores of the impact on protein function caused by single nucleic acid substitutions. There are few attempts to quantify or qualitatively explain the pathogenicity of unbalanced variants such as deletions, insertions and duplications. Therefore, in this paper, we propose SVPath, a pathogenicity prediction model based on machine learning algorithms for structural variations of deletions, insertions and duplications that occur in exons.

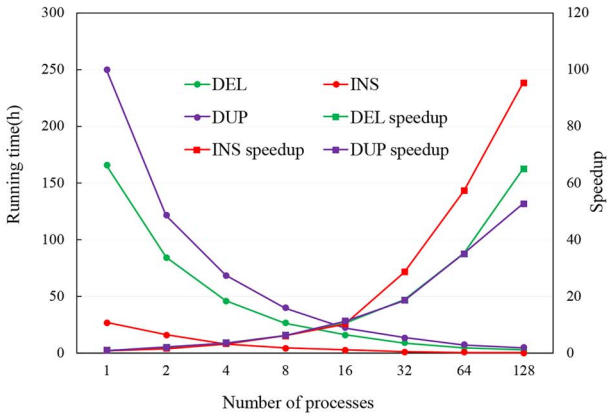
When constructing the features required by the machine learning model, we collected information about the biological features caused by the mutation from multiple angles. These features can generally be divided into three categories, namely (1) The correlation score based on the replacement of a single nucleic acid, mainly about the impact score of the SNP on the structure and function of the protein, etc.; (2) Feature scores based on the gene level, mainly gene function loss score and GDI, etc.; (3) Based on the features of transcriptomics, it mainly includes the overlap ratio of histone marker signals and variations with specific regions of genes. After constructing features for the deletion, insertion and duplication structural variation events on each exon in the ClinVar clinical variation database, a part was selected to train a GBDT machine learning model, and a reserved test set was used to verify the SVPath prediction model. In addition, we used two independent variation databases, DECIPHER and gnomAD, to further verify the performance of SVPath. The experimental results show that SVPath achieves quite excellent prediction effects whether it is on the test set or on two independent datasets.

Data contamination issues due to use of scores from other methods are less likely. The work in this paper focuses on the pathogenicity prediction of SVs, while the scores in Table 2 are all based on SNPs and the scores in Table 3 are based on genes. Table 4 is based on transcriptomic features and defined regions of chromosomes, so the dataset we used is completely independent of those used by other methods. In addition, the score of one of the other methods is only a feature of a structural variant, and the pathogenicity of the structural variant predicted in this paper is determined by all the features.

In order to speed up the process of feature annotation, we optimized it from two aspects. First of all, because any two structural mutation events are unrelated, we adopt a multi-process level parallel approach to shorten the overall time, such as using four processes to perform feature annotations for four structural mutations at the same time. Secondly, SVPath involves many annotation features and requires reading a large number of files. The largest file is the CADD score (323GB), which contains the scores of 8.6 billion possible nucleic acid substitutions. Therefore, we split the four larger files (CADD, dbNSFP, MCAP and REVEL) according to chromosome numbers to shorten the time to read the files. For the values of the features SIFT, SIFT4G, Polyphen2_HDIV and Polyphen2_HVAR, we follow the value method in the dbNSFP database, that is, the discretized value, rather than the original continuity value. On the one hand, the conversion from continuous to discrete is in dbNSFP. Its purpose is to make these features monotonous in the same direction. Higher scores may mean higher destructiveness. If a non-synonymous single-nucleotide variant (nsSNV) has multiple scores, dbNSFP uses the most harmful one. On the other hand, if these four databases



(a) SV length distribution



(b) Running time and speedup

Figure 4. Variations length distribution and running time. Because we split large files, when using multi-process parallelism, feature annotations with small structural variations hardly affect the total running time. Since the annotation of long variation events and small-length variations may be carried out at the same time, although the variations within 50bp in Figure 4b accounts for half, the final time-consuming is determined by the large-scale variations (500kb~5Mb). Figure 4b shows the running time and speedup when using different numbers of processes. Assuming that the running time of a single process is T_s, the running time of multiple processes is T_p, and the number of processes is p, the speedup when using p processes is T_s/T_p.

(files) are used separately, it will greatly increase the running time of SVPath in constructing features.

In addition, the histone marker signal and methylation signal characteristics in our pre-trained model are based on human H1 embryonic stem cells (E003); germline mutations and somatic mutations have different causes, genetics, functions and occurrence periods, so theoretically pre-trained models are only suitable for the pathogenicity prediction of germline mutations. Because there are not enough data that clearly indicate the structural variation of germline mutations and somatic mutations with clinical pathogenicity labels, it is difficult to distinguish the pathogenicity predictions of germline mutations and somatic mutations. For the prediction of the pathogenicity of somatic mutations, if there are enough data on somatic mutations with known pathogenicity, such as Muscle Satellite Cultured Cells (E052), Lung (E096), Liver (E066), Ovary (E097), etc., we can use these tissue-specific histone and methylation signals to replace the H1 embryonic cell data we used to train each somatic mutation pathogenicity prediction model.

In future work, first, we will dig out more data about the pathogenicity and benignity of the inversion structural variation, so as to improve the data fitting ability of the model. Second, regarding the pathogenicity prediction of inverted structural variants, because the current mainstream databases on clinical variants (ClinVar, gnomAD and DECIPHER, etc.) do not have enough data about inversion variants, we intend to manually sort out the clinical importance data of inverted variants from related clinical literature to realize the pathogenicity prediction of inverted variants in the next work.

Conclusions

Genome variation is one of the causes of most diseases, and it is also an important factor affecting phenotypic diversity. However, most of the existing tools and algorithms try to reveal the causal relationship between SNPs and diseases, and there are few related algorithms to explore the pathogenicity of structural variations. Therefore, in this paper, we propose the SVPath to predict the pathogenicity of deletion, insertion and duplication structural variations that occur in exons. First, in order to make the most of the relevant information of each variation site, we convert each structural variation into multiple consecutive SNP events, thereby introducing SNP-based feature information, such as the effect of SNP on protein structure and function. Second, we have introduced gene-level feature data for each structural variation, such as loss of function and GDI. Third, we also introduced related features based on transcriptomics, such as histone signal, the overlap ratio of variation and genomic element definitions, etc. Finally, based on the clinical structural variation data in the ClinVar database, we employed a GBDT machine learning method to train a structural variation pathogenicity prediction model SVPath. Experimental results show that our SVPath has achieved excellent predictive performance no matter on the reserved test set or on two independent data sets. On the test sets, after cross-validating, we get the average scores of Matthews correlation coefficients for deletions, insertions, and duplications structural variations of 0.956, 0.704 and 0.948, respectively.

Key Points

- This paper proposes SVPath which is based on a machine learning model to predict the pathogenicity of deletions, insertions and duplications structural variations that occur in exons with higher performance.
- SVPath constructs the annotation features of each structural mutation event from multiple perspectives.
- Designed and implemented a method to convert structural variation into multiple consecutive SNP events,

thereby introducing relevant features based on single nucleic acid substitution.
- Experimental results prove that SVPath achieved more accurate prediction and generalization capabilities. On independent data sets, SVPath's predictive ability is still excellent.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics*

Data availability

SVPath is implemented in Python, and the source code can be downloaded from https://github.com/pengsl-lab/SVPath. The clinical variation data of ClinVar can be obtained at https://www.ncbi.nlm.nih.gov/clinvar, dbVar data are available at https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/GRCh37/vcf/. The gnomAD is at https://gnomad.broadinstitute.org. The DECIPHER is available at https://www.deciphergenomics.org.

The features data used to annotate variations are as follows:

- The CADD annotation information is obtained from https://cadd.gs.washington.edu.
- The dbNSFP database containing various feature values is at https://sites.google.com/site/jpopgen/dbNSFP.
- The Mendelian clinically applicable pathogenicity (M-CAP) score is obtained from http://bejerano.stanford.edu/mcap.
- The REVEL pathogenicity score is obtained from https://sites.google.com/site/revelgenomics.
- The GDI gene damage index data can be obtained from https://lab.rockefeller.edu/casanova/GDI.
- The RVIS genetic intolerance data is downloaded from http://genic-intolerance.org
- The functional genomics data are downloaded from the ENCODE Roadmap (https://egg2.wustl.edu/roadmap).
- The genomic element definitions data can be downloaded from http://pcawg.gersteinlab.org.

Acknowledgments

This work was supported by National Key R&D Program of China 2017YFB0202602, 2018YFC0910405, 2017YFC1311003, 2016YFC1302500, 2016YFB0200400, 2017YFB0202104; NSFC Grants U19A2067, 61772543, U1435222, 61625202, 61272056; Science Foundation for Distinguished Young Scholars of Hunan Province (2020JJ2009); Science Foundation of Changsha kq2004010; JZ20195242029, JH201991 42034,Z202069420652; The Funds of Peng Cheng Lab, State Key Laboratory of Chemo/Biosensing and Chemometrics; the Fundamental Research Funds for the Central Universities, and Guangdong Provincial Department of Science and Technology under grant No. 2016B090918122.

References

1. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**(5):363–76.
2. 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* 2010;**467**(7319):1061.
3. Sudmant PH, Rausch T, Gardner EJ, *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**(7571):75–81.
4. Lek M, Karczewski KJ, Minikel EV, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**(7616):285–91.
5. Natarajan P, Peloso GM, Zekavat SM, *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun* 2018;**9**(1):1–12.
6. Terasaki A, Nakamura M, Urata Y, *et al.* Dna analysis of benign adult familial myoclonic epilepsy reveals associations between the pathogenic tttca repeat insertion in samd12 and the nonpathogenic tttta repeat expansion in tnrc6a. *J Hum Genet* 2021;**66**(4):419–29.
7. Eichler EE. Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine* 2019;**381**(1):64–74.
8. Kircher M, Witten DM, Jain P, *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**(3):310–5.
9. Ioannidis NM, Rothstein JH, Pejaver V, *et al.* Revel: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* 2016;**99**(4):877–85.
10. Jagadeesh KA, Wenger AM, Berger MJ, *et al.* M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**(12):1581–6.
11. Vaser R, Adusumalli S, Leng SN, *et al.* Sift missense predictions for genomes. *Nat Protoc* 2016;**11**(1):1–9.
12. Adzhubei I, Schmidt S, Peshkin L, *et al.* Polyphen-2: prediction of functional effects of human nssnps. *Nat Methods* 2010.
13. Yang L, Guo Y, Liu X, *et al.* Pathogenic gene prediction based on network embedding. *Brief Bioinform* 2021;**22**(4):bbaa353.
14. Onisiforou A, Spyrou GM. Identification of viral-mediated pathogenic mechanisms in neurodegenerative diseases using network-based approaches. *Brief Bioinform* 2021;**05**:bbab141.
15. Ata SK, Wu M, Fang Y, *et al.* Recent advances in network-based methods for disease gene prediction. *Brief Bioinform* 2021;**22**(4):bbaa303.
16. Xiang J, Zhang J, Zheng R, *et al.* Nidm: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Brief Bioinform* 2021.
17. Peng J, Xue H, Wei Z, *et al.* Integrating multi-network topology for gene function prediction using deep neural networks. *Brief Bioinform* 2021;**22**(2):2096–105.
18. Alyousfi D, Baralle D, Collins A. Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data. *Brief Bioinform* 2021;**22**(2):1782–9.
19. Cingolani P, Platts A, Wang LL, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms,

snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012;**6**(2):80–92.
20. McLaren W, Gil L, Hunt SE, *et al.* The ensembl variant effect predictor. *Genome Biol* 2016;**17**(1):1–14.
21. Weischenfeldt J, Symmons O, Spitz F, *et al.* Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 2013;**14**(2):125–38.
22. Ganel L, Abel HJ, Consortium FMS, *et al.* Svscore: an impact prediction tool for structural variation. *Bioinformatics* 2017;**33**(7):1083–5.
23. Kumar S, Harmanci A, Vytheeswaran J, *et al.* Svfx: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol* 2020;**21**(1):1–21.
24. Landrum MJ, Chitipiralla S, Brown GR, *et al.* Clinvar: improvements to accessing data. *Nucleic Acids Res* 2020;**48**(D1):D835–44.
25. Lappalainen I, Lopez J, Skipper L, *et al.* Dbvar and dgva: public archives for genomic structural variation. *Nucleic Acids Res* 2012;**41**(D1):D936–41.
26. Sherry ST, Ward M-H, Kholodov M, *et al.* dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res* 2001;**29**(1): 308–11.
27. Amberger JS, Hamosh A. Searching online mendelian inheritance in man (omim): a knowledgebase of human genes and genetic phenotypes. *Curr Protoc Bioinformatics* 2017;**58**(1): 1–2.
28. Wang K, Li M, Hakonarson H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**(16):e164–4.
29. MacDonald JR, Ziman R, Yuen RKC, *et al.* The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;**42**(D1): D986–92.
30. Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**(7809):434–43.
31. Firth HV, Richards SM, Bevan AP, *et al.* Decipher: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics* 2009;**84**(4):524–33.
32. Li H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;**27**(21):2987–93.
33. Liu X, Li C, Mou C, *et al.* dbnsfp v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. *Genome Med* 2020;**12**(1):1–8.
34. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;**11**(5):863–74.
35. Adzhubei IA, Schmidt S, Peshkin L, *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4):248–9.
36. Carter H, Douville C, Stenson PD, *et al.* Identifying mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 2013;**14**(3):1–16.
37. Qi H, Chen C, Zhang H, *et al.* Mvp: predicting pathogenicity of missense variants by deep learning bioRxiv. 2018; 259390.
38. Samocha KE, Kosmicki JA, Karczewski KJ, *et al.* Regional missense constraint improves variant deleteriousness prediction BioRxiv. 2017;148353.
39. Quang D, Chen Y, Xie X. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**(5):761–3.
40. Qiongshi L, Yiming H, Sun J, *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 2015;**5**(1):1–13.
41. Gulko B, Hubisz MJ, Gronau I, *et al.* A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 2015;**47**(3):276–83.
42. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 2017;**49**(4):618–24.
43. 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**(12):e1001025.
44. Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. In: *Annual International Conference on Research in Computational Molecular Biology*. Springer, 2006, 190–205.
45. Siepel A, Bejerano G, Pedersen JS, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**(8):1034–50.
46. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 2006;**22**(22):2729–34.
47. Fadista J, Oskolkov N, Hansson O, *et al.* Loftool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* 2017;**33**(4):471–4.
48. Yuval Itan, Lei Shang, Bertrand Boisson, Etienne Patin, Alexandre Bolze, Marcela Moncada-Vélez, Eric Scott, Michael J Ciancanelli, Fabien G Lafaille, Janet G Markle, The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences*, **112**(44):13615–20, 2015.
49. Kundaje A, Meuleman W, Ernst J, *et al.* *Nature* 2015;**518**(7539):317–30.
50. ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature* 2012;**489**(7414):57.
51. Stunnenberg HG, Abrignani S, Adams D, *et al.* The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell* 2016;**167**(5):1145–9.
52. Harris D, Harris S. *Digital design and computer architecture*. Morgan Kaufmann, 2010.
53. Han H, Wang W-Y, Mao B-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: *International conference on intelligent computing*. Springer, 2005, 878–87.
54. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 2001;1189–232.
55. Wang Q, Pierce-Hoffman E, Cummings BB, *et al.* Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat Commun* 2020;**11**(1):1–13.
56. Brænne I, Kleinecke M, Reiz B, *et al.* Systematic analysis of variants related to familial hypercholesterolemia in families with premature myocardial infarction. *Eur J Hum Genet* 2016;**24**(2):191–7.
57. Chassaing N, Davis EE, McKnight KL, *et al.* Targeted resequencing identifies ptch1 as a major contributor to ocular developmental anomalies and extends the sox2 regulatory network. *Genome Res* 2016;**26**(4):474–85.