ACS OMEGA

Review of Computational Methods and Database Sources for Predicting the Effects of Coding Frameshift Small Insertion and Deletion Variations

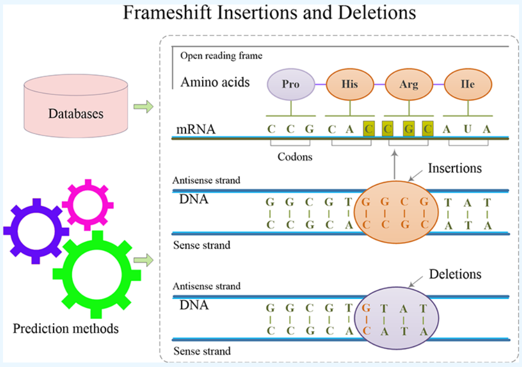
Fang Ge,* Muhammad Arif, Zihao Yan, Hanin Alahmadi, Apilak Worachartcheewan, and Watshara Shoombuatong*

Cite This: *ACS Omega* 2024, 9, 2032−2047

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Genetic variations (including substitutions, insertions, and deletions) exert a profound influence on DNA sequences. These variations are systematically classified as synonymous, nonsynonymous, and nonsense, each manifesting distinct effects on proteins. The implementation of high-throughput sequencing has significantly augmented our comprehension of the intricate interplay between gene variations and protein structure and function, as well as their ramifications in the context of diseases. Frameshift variations, particularly small insertions and deletions (indels), disrupt protein coding and are instrumental in disease pathogenesis. This review presents a succinct review of computational methods, databases, current challenges, and future directions in predicting the consequences of coding frameshift small indels variations. We analyzed the predictive efficacy, reliability, and utilization of computational methods and variant account, reliability, and utilization of database. Besides, we also compared the prediction methodologies on GOF/LOF pathogenic variation data. Addressing the challenges pertaining to prediction accuracy and cross-species generalizability, nascent technologies such as AI and deep learning harbor immense potential to enhance predictive capabilities. The importance of interdisciplinary research and collaboration cannot be overstated for devising effective diagnosis, treatment, and prevention strategies concerning diseases associated with coding frameshift indels variations.



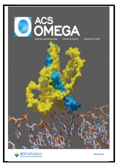
1. INTRODUCTION

Genetic variations, expressed as nucleotide substitutions, insertions, or deletions within DNA sequences, are systematically classified as follows: synonymous, nonsynonymous, and nonsense. While synonymous variations do not alter the AA sequence, both nonsynonymous and nonsense variations bring about changes in the AA sequence, consequently modifying the resulting protein structure.[1] These modifications can give rise to a diverse range of functional consequences, spanning from LOF (Loss of Function) or GOF (Gain of Function), thereby potentially contributing to genetic disorders or oncogenesis.[2,3]

The advent and widespread utilization of advanced sequencing techniques have catalyzed the production and dissemination of vast gene variation data, consequently enriching our comprehension of protein structure, functionality, and disease predisposition.[4,5] Some variations have the potential to initiate significant changes in protein structures, thereby affecting their functions. The precise location of a variation within a gene can significantly influence these outcomes, especially if it resides within structural domains or protein−protein interaction interfaces.[6−8]

Frameshift variations, particularly small indels, represent a critical category of gene variations. These variations occur when an insertion or deletion event disrupts the DNA sequence reading frame, causing a shift in the frame. During translation, if these events do not occur in multiples of three,[9] the entire protein amino acid sequence may be affected. This disruption could lead to LOF, introduce nonsense variations, or generate structurally defective proteins, all of which have implications in diseases such as Duchenne muscular dystrophy, cystic fibrosis, and hereditary breast cancer.[10] To predict the impacts of these small frameshift indels variations, numerous computational methodologies and databases have been developed. Notable examples include MutationTaster2,[11] PROVEAN,[12] SIFT Indel,[13] CADD,[14] DDIG-IN,[15] VEST-Indel,[16] MutPred-LOF,[17] PredCID,[18] SPD_Pred,[19] PRO-FOUND,[20] ClinVar,[21] dbSNP,[22] 1000GP,[23] ExAC,[4] gnomAD,[24] COSMIC,[25] HGMD,[26] dbVar,[27] DGVa,[27] OMIM,[28]

Received: October 3, 2023
Revised: November 30, 2023
Accepted: December 4, 2023
Published: January 3, 2024



ACS Publications

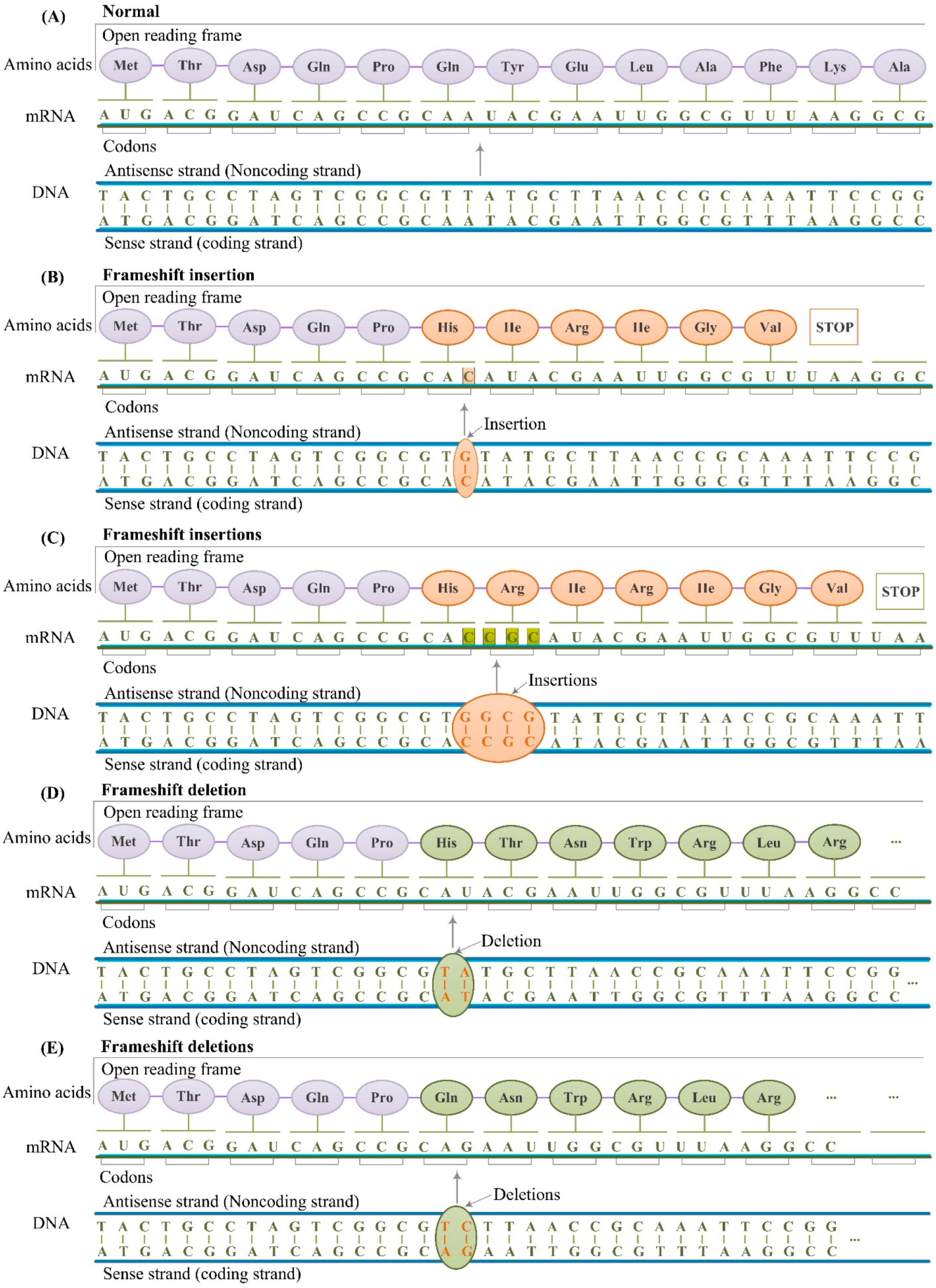

(A)
Normal
Open reading frame
Amino acids
Met Thr Asp Gln Pro Gln Tyr Glu Leu Ala Phe Lys Ala
mRNA
A U G A C G G A U C A G C C G C A A U A C G A A U U G G C G U U U A A G G C G
Codons
Antisense strand (Noncoding strand)
DNA
T A C T G C C T A G T C G G C G T T A T G C T T A A C C G C A A A T T C C G G
A T G A C G G A T C A G C C G C A A T A C G A A T T G G C G T T T A A G G C C
Sense strand (coding strand)
(B)
Frameshift insertion
Open reading frame
Amino acids
Met Thr Asp Gln Pro His Ile Arg Ile Gly Val STOP
mRNA
A U G A C G G A U C A G C C G C A C A U A C G A A U U G G C G U U U A A G G C
Codons
Antisense strand (Noncoding strand)
Insertion
DNA
T A C T G C C T A G T C G G C G T G T A T G C T T A A C C G C A A A T T C C G
A T G A C G G A T C A G C C G C A C A T A C G A A T T G G C G T T T A A G G C
Sense strand (coding strand)
(C)
Frameshift insertions
Open reading frame
Amino acids
Met Thr Asp Gln Pro His Arg Ile Arg Ile Gly Val STOP
mRNA
A U G A C G G A U C A G C C G C A C C G C A U A C G A A U U G G C G U U U A A
Codons
Antisense strand (Noncoding strand)
Insertions
DNA
T A C T G C C T A G T C G G C G T G G C G T A T G C T T A A C C G C A A A T T
A T G A C G G A T C A G C C G C A C C G C A T A C G A A T T G G C G T T T A A
Sense strand (coding strand)
(D)
Frameshift deletion
Open reading frame
Amino acids
Met Thr Asp Gln Pro His Thr Asn Trp Arg Leu Arg ...
mRNA
A U G A C G G A U C A G C C G C A U A C G A A U U G G C G U U U A A G G C C
Codons
Antisense strand (Noncoding strand)
Deletion
DNA
T A C T G C C T A G T C G G C G T A T G C T T A A C C G C A A A T T C C G G ...
A T G A C G G A T C A G C C G C A T A C G A A T T G G C G T T T A A G G C C
Sense strand (coding strand)
(E)
Frameshift deletions
Open reading frame
Amino acids
Met Thr Asp Gln Pro Gln Asn Trp Arg Leu Arg ... ...
mRNA
A U G A C G G A U C A G C C G C A G A A U U G G C G U U U A A G G C C
Codons
Antisense strand (Noncoding strand)
Deletions
DNA
T A C T G C C T A G T C G G C G T C T T A A C C G C A A A T T C C G G ...
A T G A C G G A T C A G C C G C A G A A T T G G C G T T T A A G G C C
Sense strand (coding strand)

Figure 1. Coding frameshift small insertion and deletion variations. (A) Normal, (B) frameshift insertion, (C) frameshift insertions, (D) frameshift deletion, (E) frameshift deletions.

LOVD,[29] DECIPHER,[30] VarSome,[31] and Ensembl.[32] Each of these resources possesses unique strengths, including capabilities for large-scale data handling and the provision of comprehensive variant annotations. However, they also face common challenges, primarily associated with prediction accuracy, false positives and negatives, and cross-species prediction transferability. Furthermore, limitations may arise due to the availability and quality of data, as well as inherent biases and assumptions within prediction algorithms.[33]

In this comprehensive review, we present a succinct overview of computational methodologies, databases, ongoing challenges, and prospective directions in predicting the consequences of coding frameshift small indels variations. Moreover, a comparative analysis of prediction methodologies was conducted on the GOF/LOF pathogenic variation data. The field envisions the seamless integration of emerging technologies such as deep learning and artificial intelligence into predictive models. These cutting-edge techniques hold immense promise for the development of next-generation tools that enhance our comprehension and prognostication of gene variation impacts, with particular emphasis on frameshift indels.[34]

2. CODING FRAMESHIFT SMALL INSERTION AND DELETION VARIATIONS

2.1. Definitions and Occurrence of Variations in DNA. Frameshift variations, which encompass small indels in the DNA sequence, disrupt the conventional reading frame of three nucleotides responsible for encoding an AA. These variations commonly occur during critical DNA processes such as replication, repair, or recombination events, often caused by DNA polymerase slippage resulting in the insertion or deletion of bases[35] (depicted in Figure 1). In Figure 1A, the normal reading frame is depicted, while in Figure 1B,C, frameshift insertions are illustrated, leading to the generation of premature stop codons and consequently shortening the encoded protein sequence. Simultaneously, in Figure 1D,E, frameshift deletions are exemplified, altering the protein sequence and resulting in different amino acid sequences.

Contributing factors to these variations encompass the following aspects:

(1) Mistakes in DNA replication and repair processes: Errors or external factors (e.g., radiation) during DNA replication and repair can lead to frameshift variations with erroneous base insertions or deletions.[36]

(2) Influence of repetitive sequences and DNA crossovers: DNA regions abundant in repetitive sequences are susceptible to base insertions or deletions. Nearby similar base pairs in these locations may induce DNA polymerase slippage, resulting in frameshift variations.[37] Additionally, unequal recombination between chromatids or chromosomes can contribute to frameshift variations.[38]

(3) Trans-acting elements instigating frameshift variations: Transposons and lncRNAs can trigger frameshift variations. Random insertion or rearrangement of transposons within the genome,[39] along with Alu sequences associated with genetic diseases due to their insertion or deletion events,[40] plays a significant role. LncRNAs can provoke local frameshifts by binding to target mRNA, promoting degradation.[39]

(4) Frameshift variations originating from point variations: Base pair replacements can disrupt splice signals, leading to incorrect splicing of introns or exons, causing frameshift variations if the insertions or deletions are not multiples of three.[41]

(5) Environmental triggers: Factors (such as chemicals, ultraviolet light, and radiation) can induce frameshift variations during DNA damage repair. Carcinogens like polycyclic aromatic hydrocarbons found in cigarette smoke interfere with base pairing, elevating the risk of base insertions or deletions.[42] Moreover, exposure to ultraviolet light leads to the formation of thymine dimers, indirectly contributing to frameshift variations through effects on DNA replication and repair.[43] These primary factors contribute to the aforementioned variations and emphasize the importance of understanding the mechanisms behind variations in genetic and environmental contexts.

2.2. Potential Impact on Protein Structure and Function. Frameshift variations, especially small indels, play a significant role in generating phenotypic diversity among individuals with identical gene variations. These variations are influenced by various factors, such as genetic predisposition, environment, and stochastic events, and can lead to diverse clinical manifestations. An excellent illustration of this phenomenon is observed in cystic fibrosis patients who carry CFTR frameshift variations, wherein symptoms can vary from lung dysfunction to digestive complications.[44] The impact of small indels in frameshift variations on protein structure and its function can be categorized into three distinct types: (1) LOF: critical functional domains or protein folding can be compromised due to inserted or deleted AAs, resulting in nonfunctional proteins. For example, frameshift-induced indels in Duchenne muscular dystrophy disrupt the normal structure and function of muscle cells due to dystrophin deficiency or abnormality.[45] (2) Reduced function: newly synthesized proteins may retain some level of functionality; However, their efficiency or stability might be diminished. Structural variations caused by small indels can impair interactions with other proteins or ligands, leading to a reduction in overall function.[46] (3) GOF: small indels can facilitate the emergence of novel protein functions. The introduction of new AA sequences alters protein structures, resulting in novel functions that enhance gene diversity,[47] and understanding GOF is crucial for identifying therapeutic and diagnostic targets.[48]

Comprehending the profound impact of small indels variation on protein function is pivotal for decoding disease mechanisms and exploring functional protein diversity. Therefore, further insightful research is necessitated to illuminate these specific mechanisms and quantify their effects.

2.3. Association with Genetic Diseases. Frameshift variations, frequently instigated by small indels, play a momentous role in a myriad of genetic disorders, including Duchenne muscular dystrophy (DMD), cystic fibrosis, and hereditary breast cancer. First, in the context of DMD, around two-thirds of cases can be attributed to frameshift small indels variations in the DMD gene, resulting in the synthesis of dystrophin proteins with impaired functionality, consequently triggering the onset of the disease.[49] Second, cystic fibrosis is correlated with frameshift variations in the CFTR gene, disrupting protein production, impairing chloride ion channel function, and subsequently leading to organ damage, particularly affecting the pancreas, lungs, and liver.[44] Similarly, frameshift variations in the BRCA1 and BRCA2 genes significantly increase the susceptibility to hereditary breast and ovarian cancers. These variations alter protein structure and thereby influence its function, thereby compromising DNA repair mechanisms.[50] Notably, frameshift variations have also been associated with complex diseases such as autism and cancer.[51,52] Consequently, understanding the molecular mechanisms underlying frameshift small indels becomes imperative, representing a crucial step toward developing effective therapeutic strategies for managing these genetic disorders.

3. COMPUTATIONAL TOOLS AND DATABASE RESOURCES

3.1. Principles Central to Genome Variation Effect Prediction. There are six core principles central to genome variation prediction approaches, as outlined below: (1) Multiple Sequence Alignment: Advanced tools, such as HHalign,[53] MUSCLE,[54] and ECA,[55] enhance the prediction of effects by prioritizing coding frameshift small indels during sequence alignment. (2) Nucleic Acid and Amino Acid Sequence Evolution Information: Analyzing evolutionary features considers the conservation, selection pressure, and evolutionary rate within frameshift variations. Tools like PAML and FUBAR[56] combine HMMs, PSSMs, and DCA to improve the prediction of variation impacts.[55,57] (3) Functional Domain Analysis: Utilization of resources like GO[58] and the Reactome pathway database[59] improves the prediction accuracy of frameshift variations' effects. Integrative analytical methods and machine learning techniques, such as VarCoPP[60] and PrismNet,[61] further refine predictions. (4) Network Analysis: Estimating the impact of genetic variations is achieved through biological network analysis targeting protein interactions. Networks involving domain interaction, protein complex topology, and metabolic pathways provide critical insights for accurate predictions.[62] The utilization of GNNs helps elucidate network modifications, crucial for understanding variation impacts.[63] (5) Machine Learning and Ensemble Methods: Employing advanced deep learning strategies, such as ResNet[64] and Transformer,[65] significantly forecasts the consequences of frameshift variations. Models like MutaBind2[66] and GraphDTA[67] capture intricate features for precise predictions. Ensemble Methods combine multiple prediction approaches, enhancing performance[68] with CNNs, RNNs, and exhibiting robust performance on sequence data.[69,70] Techniques like meta-learning and transfer learning optimize models, further improving accuracy.[71] (6) Multiomics Data for Variation Impact Evaluation: the integration of multiomics data enables a comprehensive evaluation of perturbations in biological systems. Incorporating gene expression, evolutionary features, and network analysis reveals potential influences on transcription factor binding sites and gene regulatory networks.[72] DNA methylation and ChIP-seq data will offer insights into epigenetic impacts.[73]

In conclusion, the accurate prediction of the impacts of coding frameshift small indels necessitates an integrative approach that incorporates evolutionary features, network analysis, and ensemble techniques. The potential for future technological advancements, enriched data resources, and enhanced computational capabilities holds promise for fortifying these endeavors. Consequently, such advancements can significantly contribute to and advance biomedical research and precision medicine, particularly in the study of coding frameshift variations. The following two subsequent sections will present a comprehensive overview of the computational tools and database resources (Figure 2 illustrates the timeline).

3.2. Computational Tools for Coding Frameshift Indels Variations Prediction. *3.2.1. Predictors.* This section introduces methods for analyzing coding frameshift small indels, which are crucial for comprehending their functional outcomes. For further details, refer to Tables 1, 2, and 3.

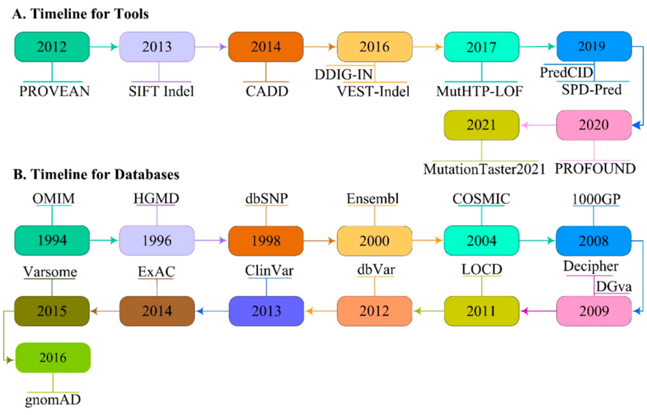


Figure 2. Timeline for coding frameshift indicator databases and effect prediction tools. (A) Timeline for Tools, (B) Timeline for Databases.

(1) **PROVEAN**[74] is an exclusive tool that predicts the consequences of small indels (from UniProt[75]) on protein functionality and phenotype, which employs machine learning and statistical models to generate a score indicating the functional impacts of these variations. PROVEAN[74] enables researchers to grasp the importance of coding frameshift small indels in protein function and phenotype, and notably its balanced accuracy was reported at 71.73% and 75.10% for deletions and insertions variations, respectively.

(2) **SIFT Indel**,[13] implemented with a refined algorithm, proficiently predicts the ramifications of coding frameshift variations, taking into account sequence alignment, conservation, and protein architecture. By providing the SIFT scores and *p*-values, it effectively quantifies disruptions caused by small indels. Its user-friendly interface and compatibility with multiple species facilitate the utilization of pre-existing homologous data sets. SIFT Indel was demonstrated noteworthy performance metrics, achieving an accuracy, sensitivity, specificity, precision, MCC, and AUC of 82%, 81%, 82%, 82%, 0.63, and 0.87, respectively.[13]

(3) **CADD**[14] provides a comprehensive approach to evaluate the effect of SNPs and small indels on genomic functionality at the functional level. By synergizing multiple biological features and machine learning algorithms, CADD[14] integrates rich features for optimal predictions, suiting various species and coding variation classes. The primary data source for CADD is primarily derived from ClinVar[21] and 1000GP.[23] Significantly, CADD has extensively annotated the entire set of 8.6 billion potential substitutions within the human genome and subsequently generates precise prediction scores for each substitution possibility.[14]

(4) **DDIG-IN**[15] presents an exhaustive evaluation of small indels' influence. By employing the AdaBoost algorithm and considering various factors, such as sequence conservation, structural conservation, and AA free energy changes, it demonstrates exceptional proficiency in predicting small indels. DDIG-IN[15] offers several notable advantages: first, it enables a thorough assessment of multiple features, leading to improved prediction accuracy; second, machine learning is employed for automatic weight adjustment, thereby

Table 1. Pred-Technology, Year, and Used Features of Ten Coding Frameshift Small Indels Predictors

tool	year	Pred-Technology[a]	used features
PROVEAN	2012	Scoring method	Clustering score of protein sequences after AA substitutions
SIFT Indel	2013	Decision Tree	Conservation score, sequence homology
CADD	2014	SVM	Multiple biological features, such as conservation, functional domains, splice signals
DDIG-IN	2015	SVM	Gene-level, transcript-level, nucleotide-level, and protein-level features
VEST-Indel	2016	RF	Conservation scores, indel length, protein disordered regions and predicted local structural characteristics
MutPred-LOF	2017	neural networks	Sequence and evolutionary features, structural and functional property features
PredCID	2021	XGBoost	Information on genes, DNA sequences, transcript expression, protein levels
SPD_Pred	2019	RF	Structural and evolutionary features, chemical interaction and spatial arrangement features, dynamic and flexibility features
PROFOUND	2020	positive-unlabeled	Fold-level attributes, environment-specific properties, and deletion site-specific properties
MutationTaster2021	2021	RF	Conservation at DNA-level and protein-level, protein-domains, splice site

[a] Note: Pred-Technology: Prediction Technology; PROVEAN, https://www.jcvi.org/research/provean; SIFT Indel, http://sift-dna.org/; https://sift.bii.a-star.edu.sg/www/code.html; CADD, https://cadd.gs.washington.edu/, https://cadd.gs.washington.edu/api; DDIG-IN, https://sparks-lab.org/server/ddig/; VEST-Indel, http://cravat.us; MutPred-LOF, http://mutpred2.mutdb.org/mutpredlof/; PredCID, http://www.xialab.info:8080/PredCID/; SPD_Pred, http://cse.iitkgp.ac.in/~pralay/resources/SPD_Pred/; PROFOUND, http://cse.iitkgp.ac.in/~pralay/resources/PROFOUND/; MutationTaster2021, https://www.genecascade.org/MutationTaster2021/.

Table 2. Max Data Upload, Software, Web Server, and Variation Types of Ten Coding Frameshift Small Indels Predictors

tool	max data upload[a]	software avail[a]	webserver avail[a]	type of variations
PROVEAN	web-based version has been retired	YES	NO	SNV, insertions and deletions (<6AA)
SIFT Indel	<100 raw	YES	YES	Insertions and deletions (<5 bp)
CADD	<2 MB	YES	YES	SNV, insertions and deletions (<50 bp)
DDIG-IN	Multiple Variations	NO	YES	NFS/FS indels, nonsense, nonymous
VEST-Indel	100−1000000 variations	NO	YES	In-frame, frameshift
MutPred-LOF	<100 variations	YES	YES	Frameshift stop-loss
PredCID	<500 indels	YES	YES	Insertions and deletions
SPD_Pred	N.M.	NO	YES	Single point deletions
PROFOUND	N.A.	YES	N.A.	deletions
MutationTaster2021	One sample in VCF file	N.A.	YES	noncoding variants, short insertions and deletions (<40 bp), frameshift, premature stop codons

[a] Note: Software avail: Software availability; Webserver avail: Webserver availability; YES: the corresponding tool has software or Web server; N.A., not applicable; N.M., not mentioned.

Table 3. Evaluation Strategy, Data Set, Genome Reference, and Last Updated of Coding Frameshift Small Indels Predictors[a]

predictors	evaluation strategy	data set	email for result	genome reference	last updated	total site	IF
PROVEAN	N.A.	Deletions: 729 Insertions: 171	NO	Not provided	7-May-14	2957	3.7
SIFT Indel	10-fold CV	Indels: 10,184 (neutral: 9,710)	YES	Genome 37 Genome 38	31-Mar-22	137	3.7
CADD	N.M.	Insertions: 627,071 Deletions: 926,968	YES	GRCh37/hg19 RCh38/hg38	31-Jul-18	5780	30.8
DDIG-IN	10-fold CV independent test	Training: 1,240 Testing: 4,016	YES	GRCh37/hg19	13-Sep-20	57	5.8
VEST-Indel	10-fold CV independent test	Deletions: 17,606 Insertions: 8,265	YES	GRCh38	29-Oct-18	125	3.9
MutPred-LOF	10-fold CV	FS-disease: 18,116 FS-neutral: 90,135	YES	N.A.	N.A.	62	5.8
PredCID	10-fold CV	Training: 4054 Independent test: 813	YES	N.A.	N.A.	17	9.5
SPD_Pred	3-fold CV	Positives: 132 Negatives: 30	N.M.	N.A.	N.A.	8	4.4
PROFOUND	10-fold CV	MPD: 153 unlabeled MPD: 7650	N.M.	N.A.	N.A.	3	5.6
MutationTaster2021	3-fold CV	Benign: 11,168,768 Deleterious: 236 400	Yes	GRCh37 Ensembl 102	24-Mar-21	108	14.9

[a] Note: YES: the corresponding tool has software or Web server; N.A., not applicable; N.M., not mentioned. Total site: is the total account of citations for the corresponding article, with the count recorded up to November 7, 2023. IF: stands for the latest Impact Factor of the journal in which the article was published.

enhancing classification accuracy; third, it surpasses other existing methods in its ability to predict coding frameshift small indels. The DDIG-IN data are sourced from PDB[76] and UniProt[75] and DDIG-in (FS) achieved impressive MCC, sensitivity, specificity of 0.59, 86%, and 72% for FS indels, respectively.[15]

(5) **VEST-Indel**[16] specializes in the precise prediction of pathogenicity for small indels with a remarkable level of accuracy, and its achievement is attained through a comprehensive integration of diverse features, such as AA conservation, structural domains, functional regions, and contextual information from variation sites, all skillfully incorporated into statistical models. The primary source of data for VEST-Indel is ClinVar,[21] ensuring a reliable and robust data set. Notably, VEST-Indel attains an impressive balanced accuracy and sensitivity of 0.87 and 0.89, respectively, on the test data.[16]

(6) **MutPred-LOF**[17] is designed to discern the potential impact of coding frameshift small indels on protein function, which leverages random forest, along with a diverse array of features encompassing AA properties, structural conservation, sequence conservation, and domain annotations. Such a comprehensive approach aims to augment prediction accuracy and effectively assess LOF variations. The data used for analysis were meticulously curated from reputable sources, including ClinVar,[21] HGMD,[26] ExAC.[4] Results from the evaluation of all LOF variants using MutPred-LOF revealed an impressive AUC of 0.85.[17]

(7) **PredCID**[18] demonstrates exceptional proficiency in predicting cancer-related driver small indels, utilizing data from reliable sources such as HGMD[26] and 1000GP.[23] By integrating carefully selected biological features across various levels, including gene, DNA, transcript, and protein, and employing the XGBoost classification algorithm, PredCID[18] effectively manages missing values that may arise from different transcript selections. As a result, it surpasses noncancer-specific methods in predictive accuracy. Notably, PredCID accurately identified 150 (94.9%) out of 158 putatively passenger frameshift indels, exhibiting a performance level close to that observed on the test data.[18]

(8) **SPD_Pred**[19] examines the influence of single-point deletions on protein stability, with a particular focus on small indels (using data from PDB[76] and UniProt[75]), which employs computational simulations and stability prediction tools to elucidate the underlying mechanisms that govern protein stability, thereby presenting new opportunities in protein engineering and drug design. By employing a rigorous cross-validation technique and utilizing RF, SPD_Pred demonstrated the exceptional accuracy of 99.4%.

(9) **PROFOUND**[20] assesses variations in folding capacity caused by MPDs (indels from UniProt[75]). By integrating folding-level, context-specific, and site-specific attributes, it effectively discerns detrimental MPDs from beneficial deletions. Moreover, the inclusion of evolutionary attributes substantially improves prediction accuracy. Through the implementation of 10-fold cross-validation, PROFOUND reported recall and fallout rate values of 82.2% (86.6%) and 14.2% (20.6%) for MPDs in protein loop regions, respectively.[20]

(10) **MutationTaster2**[11] is founded upon the Naive Bayes classifier and represents an advancement over its precursor, MutationTaster, with the primary objective of proficiently predicting the impacts of gene mutations on protein function, notably small indels. This method incorporates sequence conservation, comprehensive annotations, and genomic context while effectively discerning diverse mutation types through the use of pathogenicity scores and annotations, thus revealing the extent of the mutations' impact.[11] The data employed in MutationTaster2 were meticulously gathered from reputable sources, including 1000GP,[23] ClinVar,[21] and HGMD.[26] In 2021, **MutationTaster2021**[77] was developed using the RF classifier. The data set used for MutationTaster2021 was collected from gnomAD,[24] ClinVar,[21] and HGMD,[26] comprising more than 11 million benign variants and 236 thousand deleterious variants. Through comparative analysis with MutationTaster2, it was demonstrated that MutationTaster2021 achieved a significantly enhanced balanced accuracy of 93.3%, as opposed to 90.7% for MutationTaster2, specifically concerning small deletions, frameshifts, or premature stop codons.[77]

3.2.2. Predictive Efficacy, Reliability, and Utilization of Computational Methods. In a bid to furnish a comparative evaluation of an array of computational methodologies, we delve meticulously into their predictive efficacy, reliability, and application spectrum. Each predictor under appraisal will be individually scrutinized for its merits and constraints within these parameters.

(1) **Predictive Efficacy**: PROVEAN,[74] in operation since 2012, has amassed an impressive publication count, underscored by a total site score of 2957. Its unique clustering algorithm analyzing protein sequences post amino acid substitutions markedly enhances its predictive proficiency. SIFT Indel,[13] introduced in 2013, exhibits admirable predictive accuracy for insertions and deletions due to its application of sequence homology and conservation scores. Despite possessing a lower site count relative to other predictors, its performance does not falter. CADD,[14] established in 2014, published in a high-impact journal (IF: 30.8), records a robust Total site score. The amalgamation of a linear regression/SVM approach with various biological features bolsters its predictive prowess. DDIG-IN,[15] notwithstanding a lesser site score, applies SVM, accounting for diverse variations, thereby ensuring a solid predictive performance. VEST-Indel,[16] operational since 2016, targets indels and frameshift mutations specifically. Its application of the random forest method enhances prediction accuracy. MutPred-LOF,[17] created in 2017, excels in predicting frameshift mutations, utilizing random forests, making it tailor-made for specific categories of variants. PredCID,[18] launched in 2021, shows immense potential through its use of XGBoost combined with gene, DNA sequences, and protein level data. SPD_Pred,[19] despite a limited site score, incorporates characteristics from diverse sources and applies the random forest method, making it a reliable contender for single point deletions. PROFOUND,[20] introduced in 2020, is predicated on a positive-unlabeled technique with a focus on structural

stability and folding speed. MutationTaster2021,[77] incepted in 2021 with a high site score, utilizes the random forest algorithm, promising significant utility for predicting the effects of sequence and structural features.

(2) **Reliability**: Long-standing tools such as PROVEAN and SIFT Indel validate their reliability. However, the discontinuation of PROVEAN's web version may impact its accessibility. CADD, with its substantial publication record in a reputable journal, underscores its reliability. Its multifaceted biological feature functionality enhances its credibility further. DDIG-IN, VEST-Indel, and MutPred-LOF have proven their reliability, though they might be more narrowly focused than some established contemporaries. Emerging tools like PredCID and MutationTaster2021 are promising prospects but necessitate further validation to establish their reliability credibly.

(3) **Utilization**: PROVEAN and SIFT Indel, owing to their well-established usage over several years, have been widely adopted. With its myriad features, CADD offers adaptive utilization for diverse variant types. While DDIG-IN, VEST-Indel, and MutPred-LOF may be specialized, they still hold substantial utility for particular mutation scenarios. Novel tools like PredCID and MutationTaster2021 demonstrate potential that may resonate with increased popularity pending further exploration and validation.

To summarize, the selection of a computational method should be dictated by the specifics of the research question and variant categories. Tools with proven track records, such as PROVEAN and SIFT Indel offer reliable performance, albeit with a limited variant spectrum. CADD distinguishes itself with a comprehensive approach. Emerging tools like PredCID and MutationTaster2021, while promising, mandate rigorous validation, preceding broader adoption. Consequently, researchers should strike a judicious balance between predictive performance and reliability, while selecting the most suitable tool for their analyses.

3.3. Comparison of Methods for Predicting Frameshift Indels Variation.

We downloaded the data (based on HGMD) classified as pathogenic variants from the GOF/LOF databases, which comprise 9618 pathogenic variations (GRCh37).[78] It is essential to note that the GOF/LOF data set encompasses various types of genomic variations, including frameshift variations, inframe_deletion, inframe_insertion, start_loss, stop_gained, stop_loss, and synonymous variation.[3]

In this section, we conducted comparative experiments using prediction tools, as presented in Section 3.2 to analyze these genomic variation data. However, due to existing technical constraints, not all tools could be evaluated. Specifically, the web servers of DDIG-IN,[15] VEST-indel,[16] PROFOUND,[20] PredCID,[18] SIFT Indel[13] were unavailable. Additionally, access to the Web site of SPD_Pred[19] was forbidden. Furthermore, MutPred-LOF[17] allows the input of only a single protein at a time. Considering these limitations, our evaluation focused solely on three tools: PROVEAN, CADD, and MutationTaster2021. The results derived from a comparison of their performance are comprehensively documented in Table 4.

Upon thorough data assessment, a meticulous comparison was conducted between PROVEAN, CADD, and MutationTaster2021 from various perspectives. Four key metrics (i.e., TP, Accuracy$^+$, Sensitivity$^+$, and Recall$^+$) were considered to evaluate their performance. From the data in Table 4, it is apparent that MutationTaster2021 outperforms both PROVEAN and CADD across all metrics. First, MutationTaster2021 boasts the highest TP value (8920, 97.37%), indicating its superior precision in identifying true positive results. Additionally, MutationTaster2021 surpasses the other two methodologies in terms of Accuracy$^+$, Sensitivity$^+$, and Recall$^+$, further cementing its excellent capabilities in predictive accuracy and identification of positive instances.

Table 4. Comparison Results of PROVEAN, CADD, and MutPred2021 on the GOF/LOF Variation Data[a]

evaluation values	PROVEAN	CADD	MutationTaster2021
TP (percent)	4600 (80.25%)	7172 (97.02%)	8920 (97.37%)
FN (percent)	1132 (19.75%)	220 (2.98%)	241 (2.63%)
BLANK	3886	2226	810
False Negative Rate$^+$	0.1975	0.0298	0.0263
Accuracy$^+$	0.8025	0.9702	0.9737
Sensitivity$^+$	0.8025	0.9702	0.9737
Recall$^+$	0.8025	0.9702	0.9737
F_1^+	0.8904	0.9849	0.9867

[a]Note: The False Negative Rate$^+$, Accuracy$^+$, Sensitivity$^+$, Recall$^+$, and F_1^+ are all calculated based on the pathogenic variation class (+). Demonstrated in the formula provided in Section 4.2, when TN and FP are equal to 0, the values of Accuracy$^+$, Sensitivity$^+$, and Recall$^+$ are the same.

However, when examining FN, a significant drawback was identified with PROVEAN, which registered the highest FN percentage (19.75%). This suggests a propensity to incorrectly classify negative results, representing a notable weakness compared with CADD and MutationTaster2021. On the other hand, MutationTaster2021 consistently demonstrates a balance between precision and recall, evidenced by the highest F_1^+ score (a critical metric for model performance) of 0.9867.

In conclusion, based on this detailed comparative analysis, it becomes clear that, among the three methodologies, MutationTaster2021 emerges as the most efficient predictor. It not only exhibits higher accuracy in predicting positive results but also has a lower likelihood of falsely identifying negative outcomes. By attainment of a commendable balance between precision and recall, MutationTaster2021 proves to be a reliable choice for studies centered around genome variation.

3.4. Database Resources.

3.4.1. Databases. In the realm of genomic variation research, specialized databases are instrumental in collating and organizing data pertinent to coding frameshift indels, as indicated in Tables 5−6.

(1) Johns Hopkins University-curated **OMIM**[28] is an all-encompassing genetic database for human diseases and genes. With over 15,000 genes and disease-associated variations including coding frameshift indels, OMIM[28] provides exhaustive annotations for variation types, disease phenotypes, inheritance patterns, and literature references, which serves as a valuable resource for evaluating prediction methods' accuracy in disease-related genetic variations. The analysis of documented pathogenic coding frameshift variations in OMIM provides profound insights into the biological mechanisms underlying these disease-causing variations.

(2) **HGMD**[26] comprises over 250,000 disease-associated genetic variations, encompassing small coding frameshift

Table 5. Database Resources of Coding Frameshift Small Indels

database name	abbreviations	year	organizer
OMIM	Online Mendelian Inheritance in Man	1995	Johns Hopkins University School of Medicine
HGMD	Human Gene Mutation Database	1996	Cardiff University
dbSNP	Database of Single Nucleotide Polymorphisms	1998	National Center for Biotechnology Information
Ensembl	Ensemble Genome Browser	2000	EMBL-EBI
COSMIC	Catalogue of Somatic Mutations in Cancer	2004	Wellcome Sanger Institute
1000GP	1000 Genomes Project	2008	International consortium of researchers
DGVa	Database of Genomic Variants Archive	2009	EMBL-EBI
Decipher	Deciphering Developmental Disorders	2009	Wellcome Sanger Institute
LOVD	Leiden Open Variation Database	2004	Leiden University Medical Center
dbVar	Database of Genomic Structural Variation	2010	National Center for Biotechnology Information
ClinVar	Clinical Variation	2013	National Center for Biotechnology Information
ExAC	Exome Aggregation Consortium	2014	Broad Institute of MIT and Harvard
VarSome		2015	Saphetor SA
gnomAD	Genome Aggregation Database	2016	Broad Institute of MIT and Harvard

indels within well-known genes. The inclusion of extensive annotations offers crucial insights into nucleotide and amino acid alterations, disease phenotypes, and literature references, rendering it an indispensable resource for the accurate evaluation of prediction methodologies.

(3) **dbSNP**,[22] managed by NCBI, is a vital repository for genetic variation, which has emerged as an indispensable resource for genomic variation studies in humans since 1998. With over 1.5 billion records of genetic variants, it encompasses diverse populations and regions globally, including various variation types, including SNPs and small indels. Importantly, dbSNP[22] provides thorough annotations and curation for indels, furnishing detailed information about genomic positions, gene structures, and AA alterations across several biological scales. These rich data inputs facilitate the examination of distribution patterns of coding frameshift variations.

(4) **Ensembl**[32] is a comprehensive genomic database and analysis platform that includes various types of genetic variations, with a specific focus on coding frameshift indels. This valuable resource offers researchers a wide range of user-friendly tools to facilitate data querying and analysis. Validating and optimizing prediction algorithms using Ensembl enhances both accuracy and applicability; additionally, its integration with other bioinformatics resources significantly boosts its ability to predict variant effects.[32]

(5) **COSMIC**,[25] a distinguished database of genetic variation, accumulates diverse genomic data from international cancer research centers. Its contents encompass SNPs, indels, CNVs, and fusion genes. Detailed annotations for coding frameshift indels provide valuable insights facilitating extensive investigations into their role in tumor development. COSMIC[25] also equips researchers with comprehensive contextual information, which they can efficiently retrieve via the database's versatile query options. Integration with bioinformatics tools and databases (e.g., TCGA[79] and ICGC[80]) enhances the interpretation of COSMIC data.

Table 6. Updated Frequency, Variant Types, and Web Sites of Coding Frameshift Database Name[a]

database name	variant account	updated frequency	variant types
OMIM	Not Given	Regular updates	PM, DI, CNVs, SV, TRE
HGMD	410,743	Regular updates	PM (SNVs, indel), SV, small indels, CNVs, RSV, MH
dbSNP	1.1 billion	Regular updates	SNVs, frameshift, SV, CNVs, RV, MNVs, US, NC, SNPs
Ensembl	714 million	Regular updates	SNV, indel, SV, CNV, Inversion, TR,
COSMIC	2.39 million	Regular updates	PM (SNVs, indel), SV, GR, CV
1000GP	0.87 million indels	Regular updates	SNVs, frameshift, SV, CNVs, US, MNVs
DGVa	37,852	Regular updates	CNVs, SV, MD, SNVs, SNPs, D/IP
Decipher	12,870	Regular updates	frameshift, small indels, PM
LOVD	23,030	Regular updates	PM, DI, CNVs, SV, TRE
dbVar	37.7 million	Regular updates	SV, CV, MD, CNV, UT, CR, H/DV
ClinVar	2 million	Regular updates	SNVs, frameshift, SV, CNVs, RV, MNVs, US
ExAC	600,000 indels	Into gnomAD	SNVs, frameshift, SV, CNVs, US
VarSome	Can annotate arbitrary variants	Regular updates	Frameshift, small indels, PM, SNVs, SV
gnomAD	frameshift: 1,186,588indels: 122,583,462	Regular updates	SNVs, frameshift, SV, CNVs, US

[a]Note: OMIM, https://www.omim.org/; HGMD, http://www.hgmd.cf.ac.uk/ac/index.php; dbSNP, https://www.ncbi.nlm.nih.gov/snp/; Ensembl, https://www.ensembl.org/; COSMIC, https://cancer.sanger.ac.uk/cosmic; 1000GP, https://www.internationalgenome.org/; DGVa, https://www.ebi.ac.uk/dgva/; Decipher, https://decipher.sanger.ac.uk/; LOVD, https://databases.lovd.nl/shared/genes/; dbVar, https://www.ncbi.nlm.nih.gov/dbvar/ ; ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/; ExAC, https://gnomad.broadinstitute.org/; VarSome, https://varsome.com/; gnomAD, https://gnomad.broadinstitute.org/.

(6) **1000GP**[23] seeks to reveal human genetic variations through whole-genome sequencing, which includes sequences of more than 2,500 individuals from 26 disparate populations, comprising SNPs, indels, CNVs, and SVs and offering insights into coding frameshift variations across populations. The project employs cutting-edge sequencing technologies and rigorous data processing to ensure high-quality output.[23] Users can retrieve data via FTP, AWS, and Google Cloud Storage, with API interfaces available for efficient data extraction and querying.

(7) **DGVa**,[27] EBI's comprehensive global database for genomic SVs includes a wide range of SVs such as

small coding frameshift indels, CNVs, and repeat sequences, which amasses samples from more than 100 research projects, making it a trusted source of reliable SV data. DGVa[27] has a central role in evaluating the accuracy of prediction algorithms and seamlessly integrates with Ensembl,[32] EGA,[81] and ENA.[82] This integration facilitates cross-database querying and collaborative analysis, contributing to enhanced predictions of variation effects.

(8) **Decipher**[30] is a genomic database for SVs, which catalogs disease-associated coding frameshift indels and provides bioinformatics tools for variant effect prediction, gene expression queries, and protein domain analysis, thereby facilitating the study of variation mechanisms and pathogenicity. Background information and analytical features enrich our understanding of disease-causing coding frameshift variations.

(9) **LOVD**[29] maintains an extensive web-based genetic variation database, providing a versatile platform for diverse variation data, including coding frameshift indels. With a curated collection of over 500,000 variant records spanning known genes, LOVD[29] offers practical tools for the assessment of precise prediction methods and algorithm validation. Its seamless integration with OMIM[28] and HGVS[83] enhances LOVD's effectiveness in predicting variant effects.

(10) **dbVar**,[27] an NCBI-maintained robust database for human genome structural variations, provides a multifaceted platform for querying and analyzing genetic variations such as coding frameshift indels, CNVs, and repeat sequences. Backed by reputable sources like the 1000 Genomes Project,[23] ClinGen,[84] DECIPHER,[30] dbVar[27] houses over 5.7 million variant records that span known genes. Researchers are afforded substantial background information and experimental data, facilitating prediction method evaluation. The seamless integration with NCBI resources including dbSNP,[22] ClinVar,[21] and OMIM[28] allows dbVar to offer comprehensive services for querying and analyzing genetic variations.

(11) **ClinVar**[21] is a globally accessible database that aggregates genetic variation data from laboratories, genetic testing enterprises, and research institutions worldwide. It presents invaluable evidence relating human genetic variations to health and disease conditions. ClinVar has particular utility for predicting coding frameshift indels and is routinely updated to maintain accuracy. Users can query specific genes, diseases, or types of variations and access expert evaluations on variant pathogenicity.

(12) The **ExAC** consortium[4] utilizes global exome sequencing data to chart genetic variations comprehensively, which incorporates information from 60,000 individuals free from disease, providing extensive coverage. Besides, backdrop information, such as haplotype structure and population frequencies, assists in understanding variation origins and evolution. Integration with resources including gnomAD,[24] ClinVar,[21] and OMIM,[28] allows for comprehensive querying and analysis of genetic variations via multiple methods and API interfaces.

(13) **VarSome**[31] is a robust genetic variation database and annotation platform that seamlessly integrates various public resources, delivering comprehensive variant information including coding frameshift indels, gene function, phenotype features, and literature references. VarSome[31] provides online tools for variant effect prediction, frequency queries, and protein domain analysis, optimizing data mining and result interpretation. It serves as a robust platform for evaluating prediction methods, empowering researchers to expedite algorithm validation and optimization.

(14) **gnomAD**[24] comprehensively maps human genetic variations with sequencing data from over 140,000 individuals, which provides in-depth annotations of coding frameshift indels allowing thorough investigation into their distribution, genetic background, and functional consequences across various populations. gnomAD[24] offers advanced tools for efficient data access and processing while adhering to stringent quality control measures. Integration with reputable resources (such as ClinVar,[21] OMIM,[28] and HGMD[85]) augments the platform's comprehensive analysis capacity of genetic variations.

3.4.2. Database Account, Reliability, and Utilization. From Table 5−6, we conducted a comparative analysis of various database sources in terms of the number of data sets, reliability, and utilization.

First, in terms of the number of data sets, VarSome[31] (3.3 billion) and dbSNP[22] (1.1 billion) far exceed the variant accounts of other databases, followed by Ensembl (714 million), while Decipher,[30] LOVD,[29] and DGVa[27] have relatively small scales with less than 50,000 each. Please note: this statistic is based on the data publicly provided by each database.

Second, regarding the updating frequency, apart from 1000GP[23] and ExAC[4] which have finished their updates, all other databases are updated annually. This demonstrates their commitment and capability to continually collect and collate genetic variant information.

Third, onto the reliability aspect, the organizers of these databases are world-renowned research institutions or universities, such as Johns Hopkins University School of Medicine, QIAGEN Bioinformatics, National Center for Biotechnology Information, European Bioinformatics Institute, etc. Therefore, we can consider that the data from these databases have a high scientific reliability.

Furthermore, in terms of the types of variants, these databases all provide multiple types of variation data, such as SNVs, indels, SV, CNVs, etc., meeting the demands of different studies. Additionally, some databases like COSMIC,[25] ClinVar,[21] etc., particularly focus on including PM, facilitating clinical related research. Finally, in terms of utilization, databases like HGMD,[85] COSMIC,[25] and ClinVar[21] have been widely used in disease research fields due to their high association with diseases. Large-scale databases like dbSNP,[22] Ensembl,[32] etc., with their comprehensiveness, hold significant value in several bioinformatics research domains. Moreover, emerging databases like gnomAD,[24] with its large scale and comprehensive types of data, is gradually gaining broader usage.

The information presented above is our comparative analysis of various database sources. We hope it can be helpful to readers. If there are any other questions or suggestions, please feel free to ask them at any time.

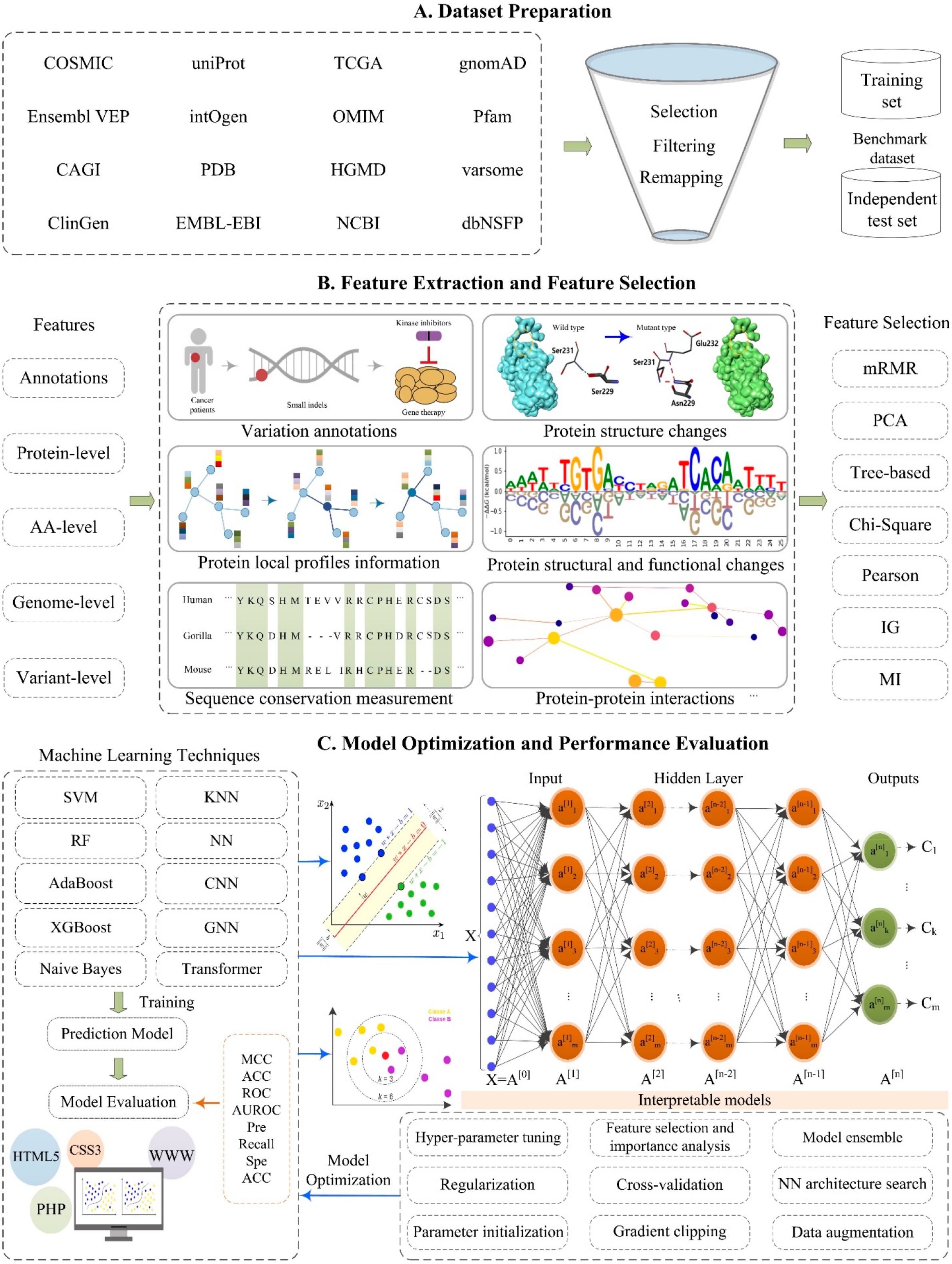


Figure 3. Systematic Evaluation for Coding Frameshift Indels. (A) Data set Preparation, (B) Feature Extraction and Feature Selection, and (C) Model Optimization and Performance Evaluation.

4. EVALUATION: COMMON STEPS AND EVALUATION METRICS

4.1. Common Steps for Evaluation. The primary focus of genomic variation research is to assess the impact on gene functionality, with particular emphasis on studying the effects of coding frameshift indels. To achieve a thorough and accurate evaluation of the predictive results, a systematic approach that integrates valid standards and criteria is essential. The following steps (as depicted in Figure 3) are employed:

(1) **Data set Preparation**: Empirical data from indels variation and protein notation databases (such as ClinVar[21] and HumVar[86]) are collected, providing details on known protein function changes due to variations. This benchmark data set is then divided into training and independent test (0.8:0.2 or 0.7:0.3 ratio). The training subset is utilized for developing the predictor, whereas testing subset assesses the efficacy.

(2) **Feature Extraction and Model Development**: Multiple features are extracted for each variation,[3] encompassing protein-level attributes (e.g., changes in protein structure, functional changes in proteins, local profile information, and protein−protein interactions), AA-level characteristics (such as physicochemical properties of wild-type AA, mutant AA, and neighboring AAs), genome-level attributes (e.g., ExAC_AMR_AF), variant-level features (e.g., variant position and class), and other annotations (including splice signals and binding sites). After extracting features, prediction models are built using methods such as traditional machine learning methods (such as KNN) as well as deep learning models (e.g., CNN, GNN, Transformer). These models are then applied to the independent test group to generate prediction outcomes.

(3) **Model Optimization and Performance Evaluation**: Employ various strategies to enhance the model, encompassing hyper-parameter tuning, feature selection, importance analysis, data augmentation, model ensemble, regularization, cross-validation, neural network architecture search, parameter initialization, and gradient clipping. Furthermore, calculate evaluation metrics (e.g., TPR, ACC) based on predicted and true label outcomes to comprehensively evaluate model's performance.

To ensure a prudent selection of the most suitable prediction method for frameshift effects, it is essential to compare evaluative metrics derived from various tools and methods applied to the same data set. Depending on the volume, quality, and specific attributes of the data, consideration of combined approaches may be warranted. Preferential treatment should be given to methods that analyze functional domains when dealing with variations within these domains. For variations in highly conserved areas, sequence-alignment-based methods are strongly recommended.

4.2. Evaluation Metrics.

A confusion matrix has been implemented to illustrate the relationship between predicted and actual results, encompassing TP, FP, TN, and FN. These components enable the computation of multiple evaluation metrics, including Sensitivity (Sen)/Recall, Specificity (Spe), FNR, FPR, Precision, NPV, Accuracy (ACC), FDR, and MCC (as defined in Formulas 1 to 10). Additionally, the AUROC provides a holistic measure of a model's classification efficacy, capturing the sensitivity-specificity dynamic across diverse thresholds. Higher AUC values signify superior predictive performance.[87] By calculating evaluation metrics and plotting curves, we can assess the precision of prediction methods for coding frameshift indels.

$$\mathrm{Pre} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{1}$$

$$\mathrm{Spe} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \tag{2}$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP}) \tag{3}$$

$$\mathrm{FNR} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN}) \tag{4}$$

$$\mathrm{NPV} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FN}) \tag{5}$$

$$\mathrm{ER} = \mathrm{FP}/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{6}$$

$$F_1 = 2 \times \mathrm{TP}/(2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}) \tag{7}$$

$$\mathrm{Recall/Sensitivity} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{8}$$

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{9}$$

$$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) / \sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(TN + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})} \tag{10}$$

5. CURRENT CHALLENGES AND FUTURE DIRECTIONS

5.1. Current Challenges.

Technological advancements have significantly contributed to the field of genomics, yet the accurate prediction of coding frameshift indices remains a challenging endeavor. The occurrence of false positive and negative results not only hinders biological comprehension but also potentially jeopardizes clinical decision-making processes.[88]

The initial hurdle is the exact forecasting of coding frameshift indels coupled with the predicament of false positives/negatives. New-age technology and methods have certainly propelled the progression in this field; however, attaining superior predictive accuracy remains an unfulfilled goal. In response to this persistent predicament, several remedial steps warrant consideration: (1) Rigorous verification protocols ought to be integrated during the model's developmental phase to ensure the detection of authentic genetic variations, thereby diminishing the likelihood of misdiagnosis.[89] (2) Incorporation of refined machine learning methodologies alongside extensive variant data sets can enhance the predictive models, curtailing inaccuracies and bolstering precision.[90] Additionally, (3) the establishment of novel statistical methods for estimating confidence intervals of predictive models could aid clinicians in result interpretation, subsequently reducing the potential for diagnostic errors.[91]

Nonetheless, the applicability of the present prediction instruments extends primarily to human genomic data, consequently constraining their usefulness across different species. To overcome this limitation, it becomes essential to (1) construct extensive multispecies databases that facilitate the prediction of frameshift indels across diverse species;[92] (2) develop adaptable algorithm frameworks that are attuned to the distinct genetic characteristics intrinsic to specific species[93] and, finally, (3) promote cross-species research endeavors to illuminate the biological implications of frameshift indels across various organisms.[94]

5.2. Future Directions.

The field of genomics is currently witnessing rapid advancements, posing the necessity for comprehensive exploration in several critical areas:

(1) Embracing Cutting-edge Computational Techniques for Genomic Analysis: Bioinformatics has experienced significant progress with the introduction of artificial intelligence (AI) and deep learning, thereby refining prediction accuracy. These sophisticated techniques detect sequence and structural patterns through rigorous data training, leveraging contemporary strategies such as

convolutional neural networks (CNNs) and recurrent neural networks (RNNs).[95] Deep learning has been efficaciously employed to various genomic tasks, including DNA sequence classification: Predicting the functional or regulatory role of DNA sequences.[96] Gene expression prediction: Identifying genes that are likely to be expressed under specific conditions.[97] Variant calling: Identifying genetic variants from sequencing data.[98] Genome assembly: Reconstructing the complete genome sequence from fragmented DNA reads.[99] Additionally, methods like transfer learning and reinforcement learning utilize existing models and data sets to enhance predictive capabilities, thereby facilitating meticulous and efficient estimation of genetic variation effects.[100]

(2) Implementing Systematic Approaches for Investigating Coding Frameshift Variations: Modern genomic tools enable complex analyses such as genome-wide association studies (GWAS) and whole-exome sequencing (WES). These mechanisms aid in identifying coding frameshift variations in genetic diseases by comparing data between healthy individuals and patients.[101] For instance, GWAS has identified a frameshift variant in the BRCA1 gene associated with increased breast cancer risk.[102] Similarly, WES has detected frameshift variants in several genes involved in diseases like cystic fibrosis, sickle cell anemia, and Li-Fraumeni syndrome.[44,103] Moreover, integrating multiomics data could offer a comprehensive perspective on the influence of coding frameshift variations on gene regulation and protein function.[104] This integrated approach helps identify novel regulatory mechanisms and gene interactions impacting the development of genetic diseases.[105]

(3) Utilizing Gene Editing Tools for Correcting Coding Frameshift Variations: Several gene editing technologies, such as CRISPR/Cas9[106] and homologous directed repair (HDR),[107] have emerged to correct coding frameshift variations. By precisely introducing or eliminating genes, they rectify the reading frame, introducing potential therapeutic avenues for genetic disorders.[108] For example, CRISPR/Cas9 has successfully corrected a frameshift mutation causing Duchenne muscular dystrophy.[109] In parallel, HDR has demonstrated its capacity to correct a frameshift mutation causing sickle cell anemia.[110]

(4) Providing Genetic Counseling and Diagnosis for Diseases Associated with Coding Frameshift Variations: Genetic counseling and diagnostic efforts are critical in preventing and managing diseases linked with coding frameshift variations.[111] Comprehensive assessments involving family history, genetic mapping, and biomarker evaluations assist in creating personalized preventative measures and treatment plans, thereby reducing disease incidence and mortality rates. Genetic counseling can help individuals at risk make informed decisions regarding family planning and reproductive choices, while genetic testing identifies carriers of genetic mutations, allowing proactive healthcare decisions and potentially preventing the transmission of genetic disorders.[112,113]

Through leveraging computational models, gene editing technologies, and innovative therapeutics, scientific and technological advancements enhance our understanding of coding frameshift variations, enabling effective prevention and treatment strategies for genetic diseases. Interdisciplinary collaborative efforts continue to fuel breakthroughs in this domain.

6. CONCLUSIONS

Coding frameshift variations, shaped by DNA replication, repair mechanisms, and environmental factors, are common genetic contributors to diseases. Gaining insight into their molecular mechanisms is pivotal to understanding genetic diseases and their implications. Methodological strategies, gene editing technologies, and genetic counseling aid in the research and treatment of genetic diseases linked to frameshift variations. Interdisciplinary collaborations coupled with technological innovations foster progress. This review focuses on predicting effects of frameshift variations, delivering an overview of methodologies, metrics, databases, and resources. We analyzed the predictive efficacy, reliability, and utilization of computational methods, and variant account, reliability, and utilization of database. Besides, we also compared the prediction methodologies on GOF/LOF pathogenic variation data. This review highlights challenges and underscores research and collaboration for algorithm enhancement and experimental validation. Newfound technologies such as deep learning and AI contribute to increased prediction accuracy while considering generalizability across different species. Experimental validation propels research and provides reliable evidence.

■ AUTHOR INFORMATION

Corresponding Authors

Watshara Shoombuatong − *Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand;* orcid.org/0000-0002-3394-8709; Email: watshara.sho@mahidol.ac.th

Fang Ge − *State Key Laboratory of Organic Electronics and Information Displays & Institute of Advanced Materials (IAM), Nanjing University of Posts & Telecommunications, Nanjing 210023, China; Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand;* Email: gfang0616@njupt.edu.cn

Authors

Muhammad Arif − *College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar*

Zihao Yan − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*

Hanin Alahmadi − *College of Computer Science and Engineering, Taibah University, Madinah 344, Saudi Arabia*

Apilak Worachartcheewan − *Department of Community Medical Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c07662

Author Contributions

F.G. and W.S.: Conceptualization, project administration, supervision, investigation. M.F. and Z.H.Y.: Manuscript revision. F.G., M.F., Z.H.Y., A.W., and H.A.: Manuscript preparation. All authors reviewed and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY223062), the College of Science and Engineering, Hamad Bin Khalifa University (HBKU), National Research Council of Thailand and Mahidol University (N42A660380), and the Specific League Funds from Mahidol University.

■ ABBREVIATIONS

AA, Amino Acid; 1000GP, 1000 Genome Project; AUROC, Area under the Receiver Operating Characteristic Curve; CADD, Combined Annotation Dependent Depletion; CNNs, Convolutional Neural Networks; CNVs, Copy Number Variants; COSMIC, Catalogue Of Somatic Mutations In Cancer; CR, Complex Rearrangements; CV, Complex Variants; D/IP, Deletions/Insertions Polymorphisms; DCA, Direct-Coupling Analysis; DDIG-IN, Detecting Disease-Causing Genetic Variation in Noncoding Regions; DI, Deletions and Insertions; DNA, Deoxyribonucleic Acid; DGVa, Database of Genomic Variants Archive; ECA, Evolutionary Coupling Analysis; EBI, European Bioinformatics Institute; ExAC, Exome Aggregation Consortium; FDR, False Discovery Rate; FN, False Negatives; FNR, False Negative Rate; FP, False Positives; GNNs, Graph Neural Networks; GNVs, Gene Rearrangements; GO, Gene Ontology; GOF, Gain-Of-Function; GR, Gene Rearrangements; GraphDTA, Graph Neural Network for Drug-Target Binding Affinity Prediction; HHalign, Homology Extension by Iterative Alignment; H/DV, Haploid/Diploid Variants; HGMD, Human Gene Mutation Database; HMMs, Hidden Markov Models; ICGC, International Cancer Genome Consortium; Indel, Insertion/Deletion; IG, Information Gain; KNN, K-Nearest Neighbors; KLOF, Loss-Of-Function; LOVD, Leiden Open Variation Database; LncRNAs, Long Noncoding RNAs; MCC, Matthews' Correlation Coefficient; MD, Microduplications; MH, Mutation Heterozygosity; MI, Information Gain; MNVs, Multi-Nucleotide Variants; MPDs, Multi-Point Deletions; MutaBind2, Protein Mutation Impact Analysis and Prediction Tool 2; MUSCLE, Multiple Sequence Comparison by Log-Expectation; NCBI, National Center for Biotechnology Information; NPV, Negative Predictive Value; NV, Noncoding Variants; NN, Neural Networks; OMIM, Online Mendelian Inheritance in Man; PM, Point Mutations; PredCID, Predictor for Cancnnner Driver Frameshift Indels; PPV, Positive Predictive Value; PrismNet, Protein-RNA Interaction by Structure-informed Modeling using Deep Neural Network; PROFOUND, Predicting Protein Foldability owing to Multipoint Deletions; PROVEAN, Protein Variation Effect Analyzer; PSSMs, Position-Specific Scoring Matrices; RF, Random Forest; RNNs, Recurrent Neural Networks; RSV, Repeat Sequence Variations; RV, Repeat Variants; SIFT Indel, Sorting Intolerant from Tolerant Indel; Small indels, Small Insertions/Deletions; SNPs, Single Nucleotide Polymorphisms; SNVs, Single Nucleotide Variants; SVM, Susvmpport Vector Machine; SV, Structural Variants; TRE, Triplet Repeat Expansions; TN, True Negatives; TP, True Positives; TR, Tandem Repeat; TRE, Triplet Repeat Expansions; UT, Unbalanced Translocations; US, Uncertain Significance; VarSome, Variant Knowledge Systems; XGBoost, eXtreme Gradient Boosting; EBI, European Bioinformatics Institute

■ REFERENCES

(1) Li, B.; Krishnan, V. G.; Mort, M. E.; Xin, F.; Kamati, K. K.; Cooper, D. N.; Mooney, S. D.; Radivojac, P. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **2009**, *25* (21), 2744−2750.

(2) Stratton, M. R.; Campbell, P. J.; Futreal, P. A. The cancer genome. *Nature* **2009**, *458* (7239), 719−724.

(3) Ge, F.; Li, C.; Iqbal, S.; Muhammad, A.; Li, F.; Thafar, M. A; Yan, Z.; Worachartcheewan, A.; Xu, X.; Song, J.; Yu, D.-J. VPatho: a deep learning-based two-stage approach for accurate prediction of gain-of-function and loss-of-function variants. *Brief Bioinform* **2023**, *24* (1), bbac535.

(4) Lek, M.; Karczewski, K. J.; Minikel, E. V.; Samocha, K. E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A. H.; Ware, J. S.; Hill, A. J.; Cummings, B. B. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536* (7616), 285−291.

(5) Ge, F.; Zhang, Y.; Xu, J.; Muhammad, A.; Song, J.; Yu, D.-J. Prediction of disease-associated nsSNPs by integrating multi-scale ResNet models with deep feature fusion. *Briefings in Bioinformatics* **2022**, *23* (1), bbab530.

(6) Fragoza, R.; Das, J.; Wierbowski, S. D.; Liang, J.; Tran, T. N.; Liang, S.; Beltran, J. F.; Rivera-Erick, C. A.; Ye, K.; Wang, T.-Y.; Yao, L.; Mort, M.; Stenson, P. D.; Cooper, D. N.; Wei, X.; Keinan, A.; Schimenti, J. C.; Clark, A. G.; Yu, H. Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat. Commun.* **2019**, *10* (1), 4141.

(7) David, A.; Sternberg, M. J. The contribution of missense mutations in core and rim residues of protein-protein interfaces to human disease. *Journal of molecular biology* **2015**, *427* (17), 2886−2898.

(8) Ge, F.; Zhu, Y.-H.; Xu, J.; Muhammad, A.; Song, J.; Yu, D.-J. MutTMPredictor: robust and accurate cascade XGBoost classifier for prediction of mutations in transmembrane proteins. *Computational and Structural Biotechnology Journal* **2021**, *19*, 6400−6416.

(9) Gemayel, R.; Cho, J.; Boeynaems, S.; Verstrepen, K. J. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes* **2012**, *3* (3), 461−480.

(10) Ceyhan-Birsoy, O.; Murry, J. B.; Machini, K.; Lebo, M. S.; Timothy, W. Y.; Fayer, S.; Genetti, C. A.; Schwartz, T. S.; Agrawal, P. B.; Parad, R. B. Interpretation of genomic sequencing results in healthy and ill newborns: results from the BabySeq Project. *Am. J. Hum. Genet.* **2019**, *104* (1), 76−93.

(11) Schwarz, J. M.; Cooper, D. N.; Schuelke, M.; Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **2014**, *11* (4), 361−362.

(12) Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **2012**, *7*, e46688.

(13) Hu, J.; Ng, P. C. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PloS one* **2013**, *8* (10), No. e77940.

(14) Kircher, M.; Witten, D. M.; Jain, P.; O'roak, B. J.; Cooper, G. M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **2014**, *46* (3), 310−315.

(15) Folkman, L.; Yang, Y.; Li, Z.; Stantic, B.; Sattar, A.; Mort, M.; Cooper, D. N.; Liu, Y.; Zhou, Y. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics* **2015**, *31* (10), 1599−1606.

(16) Douville, C.; Masica, D. L.; Stenson, P. D.; Cooper, D. N.; Gygax, D. M.; Kim, R.; Ryan, M.; Karchin, R. Assessing the

pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel). *Human mutation* **2016**, *37* (1), 28−35.

(17) Pagel, K. A.; Pejaver, V.; Lin, G. N.; Nam, H.-J.; Mort, M.; Cooper, D. N.; Sebat, J.; Iakoucheva, L. M.; Mooney, S. D.; Radivojac, P. When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics* **2017**, *33* (14), i389−i398.

(18) Yue, Z.; Chu, X.; Xia, J. PredCID: prediction of driver frameshift indels in human cancer. *Briefings in Bioinformatics* **2021**, *22* (3), bbaa119.

(19) Banerjee, A.; Levy, Y.; Mitra, P. Analyzing change in protein stability associated with single point deletions in a newly defined protein structure database. *J. Proteome Res.* **2019**, *18* (3), 1402−1410.

(20) Banerjee, A.; Kumar, A.; Ghosh, K. K.; Mitra, P. Estimating change in foldability due to multipoint deletions in protein structures. *J. Chem. Inf. Model.* **2020**, *60* (12), 6679−6690.

(21) Landrum, M. J.; Lee, J. M.; Benson, M.; Brown, G.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Hoover, J.; Jang, W.; Katz, K.; Ovetsky, M.; Riley, G.; Sethi, A.; Tully, R.; Villamarin-Salomon, R.; Rubinstein, W.; Maglott, D. R. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **2016**, *44* (D1), D862−D868.

(22) Sherry, S. T.; Ward, M.-H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29* (1), 308−311.

(23) The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526* (7571), 68.

(24) Karczewski, K. J.; Francioli, L. C.; Tiao, G.; Cummings, B. B.; Alföldi, J.; Wang, Q.; Collins, R. L.; Laricchia, K. M.; Ganna, A.; Birnbaum, D. P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581* (7809), 434−443.

(25) Forbes, S. A.; Bindal, N.; Bamford, S.; Cole, C.; Kok, C. Y.; Beare, D.; Jia, M.; Shepherd, R.; Leung, K.; Menzies, A.; Teague, J. W.; Campbell, P. J.; Stratton, M. R.; Futreal, P. A. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **2011**, *39* (suppl_1), D945−D950.

(26) Stenson, P. D.; Mort, M.; Ball, E. V.; Evans, K.; Hayden, M.; Heywood, S.; Hussain, M.; Phillips, A. D.; Cooper, D. N. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human genetics* **2017**, *136*, 665−677.

(27) Lappalainen, I.; Lopez, J.; Skipper, L.; Hefferon, T.; Spalding, J. D.; Garner, J.; Chen, C.; Maguire, M.; Corbett, M.; Zhou, G.; Paschall, J.; Ananiev, V.; Flicek, P.; Church, D. M. dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Research* **2012**, *41* (D1), D936−D941.

(28) Hamosh, A.; Scott, A. F.; Amberger, J. S.; Bocchini, C. A.; McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **2005**, *33* (suppl_1), D514−D517.

(29) Fokkema, I. F.; Taschner, P. E.; Schaafsma, G. C.; Celli, J.; Laros, J. F.; den Dunnen, J. T. LOVD v. 2.0: the next generation in gene variant databases. *Human mutation* **2011**, *32* (5), 557−563.

(30) Firth, H. V.; Richards, S. M.; Bevan, A. P.; Clayton, S.; Corpas, M.; Rajan, D.; Van Vooren, S.; Moreau, Y.; Pettett, R. M.; Carter, N. P. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *American Journal of Human Genetics* **2009**, *84* (4), 524−533.

(31) Kopanos, C.; Tsiolkas, V.; Kouris, A.; Chapple, C. E.; Aguilera, M. A.; Meyer, R.; Massouras, A. VarSome: the human genomic variant search engine. *Bioinformatics* **2019**, *35* (11), 1978.

(32) Aken, B. L.; Ayling, S.; Barrell, D.; Clarke, L.; Curwen, V.; Fairley, S.; Fernandez Banet, J.; Billis, K.; García Girón, C.; Hourlier, T. The Ensembl gene annotation system. *Database* **2016**, *2016*, 2016:baw093.

(33) Stenson, P. D.; Mort, M.; Ball, E. V.; Shaw, K.; Phillips, A. D.; Cooper, D. N. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics* **2014**, *133*, 1−9.

(34) Ray, P. P. A survey on Internet of Things architectures. *Journal of King Saud University-Computer and Information Sciences* **2018**, *30* (3), 291−319.

(35) Petrov, D.; Hartl, D. Pseudogene evolution and natural selection for a compact genome. *Journal of Heredity* **2000**, *91* (3), 221−227.

(36) Chen, Z.; Liu, Y.; He, A.; Li, J.; Chen, M.; Zhan, Y.; Lin, J.; Zhuang, C.; Liu, L.; Zhao, G.; Huang, W.; Cai, Z. Theophylline controllable RNAi-based genetic switches regulate expression of lncRNA TINCR and malignant phenotypes in bladder cancer cells. *Sci. Rep* **2016**, *6* (1), 30798.

(37) Zhang, Y.; Li, T.; Preissl, S.; Amaral, M. L.; Grinstein, J. D.; Farah, E. N.; Destici, E.; Qiu, Y.; Hu, R.; Lee, A. Y.; Chee, S.; Ma, K.; Ye, Z.; Zhu, Q.; Huang, H.; Fang, R.; Yu, L.; Izpisua Belmonte, J. C.; Wu, J.; Evans, S. M.; Chi, N. C.; Ren, B. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* **2019**, *51* (9), 1380−1388.

(38) Fikus, M. U.; Mieczkowski, P. A.; Koprowski, P.; Rytka, J.; Śledziewska-Gójska, E.; Cieśla, Z. The product of the DNA damage-inducible gene of Saccharomyces cerevisiae, DIN7, specifically functions in mitochondria. *Genetics* **2000**, *154* (1), 73−81.

(39) Mi, S.; Lee, X.; Li, X.-p.; Veldman, G. M.; Finnerty, H.; Racie, L.; LaVallie, E.; Tang, X.-Y.; Edouard, P.; Howes, S.; Keith, J. C.; McCoy, J. M. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **2000**, *403* (6771), 785−789.

(40) Bennett, E. A.; Keller, H.; Mills, R. E.; Schmidt, S.; Moran, J. V.; Weichenrieder, O.; Devine, S. E. Active Alu retrotransposons in the human genome. *Genome Res.* **2008**, *18* (12), 1875−1883.

(41) Sterne-Weiler, T.; Sanford, J. R. Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome biology* **2014**, *15* (1), 201.

(42) Gorrini, C.; Harris, I. S.; Mak, T. W. Modulation of oxidative stress as an anticancer strategy. *Nat. Rev. Drug Discovery* **2013**, *12* (12), 931−947.

(43) Cadet, J.; Wagner, J. R. DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *Cold Spring Harbor perspectives in biology* **2013**, *5* (2), a012559.

(44) Cutting, G. R. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat. Rev. Genet.* **2015**, *16* (1), 45−56.

(45) Pires, D. E.; Ascher, D. B.; Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research* **2014**, *42* (W1), W314−W319.

(46) Vihinen, M. Functional effects of protein variants. *Biochimie* **2021**, *180*, 104−120.

(47) Romero, P. A.; Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10* (12), 866−876.

(48) Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet* **2016**, *17* (4), 239−250.

(49) Aartsma-Rus, A.; Van Ommen, G-JB Antisense-mediated exon skipping: a versatile tool with therapeutic and research applications. *Rna* **2007**, *13* (10), 1609−1624.

(50) Miki, Y.; Swensen, J.; Shattuck-Eidens, D.; Futreal, P. A.; Harshman, K.; Tavtigian, S.; Liu, Q.; Cochran, C.; Bennett, L. M.; Ding, W. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **1994**, *266* (5182), 66−71.

(51) De Rubeis, S.; He, X.; Goldberg, A. P.; Poultney, C. S.; Samocha, K.; Ercument Cicek, A.; Kou, Y.; Liu, L.; Fromer, M.; Walker, S. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **2014**, *515* (7526), 209−215.

(52) Hanahan, D.; Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144* (5), 646−674.

(53) Söding, J.; Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **2005**, *33* (suppl_2), W244−W248.

(54) Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **2004**, *32* (5), 1792−1797.

(55) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (49), E1293−E1301.

(56) Murrell, B.; Moola, S.; Mabona, A.; Weighill, T.; Sheward, D.; Kosakovsky Pond, S. L.; Scheffler, K. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution* **2013**, *30* (5), 1196−1205.

(57) Jones, D. T.; Buchan, D. W.; Cozzetto, D.; Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **2012**, *28* (2), 184−190.

(58) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T. Gene ontology: tool for the unification of biology. *Nature genetics* **2000**, *25* (1), 25−29.

(59) Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B. The reactome pathway knowledgebase. *Nucleic acids research* **2018**, *46* (D1), D649−D655.

(60) Papadimitriou, S.; Gazzo, A.; Versbraegen, N.; Nachtegael, C.; Aerts, J.; Moreau, Y.; Van Dooren, S.; Nowé, A.; Smits, G.; Lenaerts, T. Predicting disease-causing variant combinations. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (24), 11878−11887.

(61) Kalinin, A. A.; Higgins, G. A.; Reamaroon, N.; Soroushmehr, S.; Allyn-Feuer, A.; Dinov, I. D.; Najarian, K.; Athey, B. D. Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics* **2018**, *19* (7), 629−650.

(62) Zhu, X.; Gerstein, M.; Snyder, M. Getting connected: analysis and principles of biological networks. *Genes & development* **2007**, *21* (9), 1010−1024.

(63) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* **2021**, *32* (1), 4−24.

(64) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE, 2016; pp 770−778.

(65) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N., Kaiser, Ł; Polosukhin, I.Attention is all you need. *Advances in Neural Information Processing Systems* **2017**, *30*.

(66) Zhang, N.; Chen, Y.; Lu, H.; Zhao, F.; Alvarez, R. V.; Goncearenco, A.; Panchenko, A. R.; Li, M. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience* **2020**, *23* (3), 100939.

(67) Nguyen, D.-T.; Mathias, S.; Bologa, C.; Brunak, S.; Fernandez, N.; Gaulton, A.; Hersey, A.; Holmes, J.; Jensen, L. J.; Karlsson, A. Pharos: collating protein information to shed light on the druggable genome. *Nucleic acids research* **2017**, *45* (D1), D995−D1002.

(68) Shihab, H. A.; Rogers, M. F.; Gough, J.; Mort, M.; Cooper, D. N.; Day, I. N.; Gaunt, T. R.; Campbell, C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **2015**, *31* (10), 1536−1543.

(69) Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **2018**, *15* (10), 816−822.

(70) AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **2019**, *35* (22), 4862−4865.

(71) Zhang, W.; Chien, J.; Yong, J.; Kuang, R. Network-based machine learning and graph theory algorithms for precision oncology. *NPJ. precision oncology* **2017**, *1* (1), 25.

(72) Zhang, S.; Zhou, J.; Hu, H.; Gong, H.; Chen, L.; Cheng, C.; Zeng, J. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic acids research* **2016**, *44* (4), No. e32.

(73) Li, Y.; Shi, W.; Wasserman, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC bioinformatics* **2018**, *19* (1), 1−14.

(74) Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **2012**, *7* (10), No. e46688.

(75) Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506−D515.

(76) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(77) Steinhaus, R.; Proft, S.; Schuelke, M.; Cooper, D. N.; Schwarz, J. M.; Seelow, D. MutationTaster2021. *Nucleic Acids Res.* **2021**, *49* (W1), W446−W451.

(78) Sevim Bayrak, C.; Stein, D.; Jain, A.; Chaudhary, K.; Nadkarni, G. N.; Van Vleck, T. T.; Puel, A.; Boisson-Dupuis, S.; Okada, S.; Stenson, P. D.; Cooper, D. N.; Schlessinger, A.; Itan, Y. Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants. *Am. J. Hum. Genet.* **2021**, *108* (12), 2301−2318.

(79) Weinstein, J. N; Collisson, E. A; Mills, G. B; Shaw, K. R M.; Ozenberger, B. A; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M The cancer genome atlas pan-cancer analysis project. *Nature genetics* **2013**, *45* (10), 1113−1120.

(80) The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **2010**, *464* (7291), 993−998.

(81) Lappalainen, I.; Almeida-King, J.; Kumanduri, V.; Senf, A.; Spalding, J. D.; Ur-Rehman, S.; Saunders, G.; Kandasamy, J.; Caccamo, M.; Leinonen, R. The European Genome-phenome Archive of human data consented for biomedical research. *Nature genetics* **2015**, *47* (7), 692−695.

(82) Amid, C.; Alako, B. T.; Balavenkataraman Kadhirvelu, V.; Burdett, T.; Burgin, J.; Fan, J.; Harrison, P. W.; Holt, S.; Hussein, A.; Ivanov, E. The European nucleotide archive in 2019. *Nucleic acids research* **2020**, *48* (D1), D70−D76.

(83) den Dunnen, J. T.; Dalgleish, R.; Maglott, D. R.; Hart, R. K.; Greenblatt, M. S.; McGowan-Jordan, J.; Roux, A. F.; Smith, T.; Antonarakis, S. E.; Taschner, P. E. HGVS recommendations for the description of sequence variants: 2016 update. *Human mutation* **2016**, *37* (6), 564−569.

(84) Rehm, H. L.; Berg, J. S.; Brooks, L. D.; Bustamante, C. D.; Evans, J. P.; Landrum, M. J.; Ledbetter, D. H.; Maglott, D. R.; Martin, C. L.; Nussbaum, R. L.; Plon, S. E.; Ramos, E. M.; Sherry, S. T.; Watson, M. S. ClinGen—the clinical genome resource. *New England Journal of Medicine* **2015**, *372* (23), 2235−2242.

(85) Stenson, P. D.; Ball, E. V.; Mort, M.; Phillips, A. D.; Shaw, K.; Cooper, D. N. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current protocols in bioinformatics* **2012**, *39* (1), 1.13.11−11.13.20.

(86) Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat. Methods* **2010**, *7* (4), 248−249.

(87) Kremic, E.; Subasi, A. Performance of random forest and SVM in face recognition. *Int. Arab J. Inf. Technol.* **2016**, *13* (2), 287−293.

(88) Cingolani, P.; Platts, A.; Wang, L. L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S. J.; Lu, X.; Ruden, D. M. 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.

(89) Li, Z.; Li, X.; Zhou, H.; Gaynor, S. M.; Selvaraj, M. S.; Arapoglou, T.; Quick, C.; Liu, Y.; Chen, H.; Sun, R. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat. Methods* **2022**, *19* (12), 1599−1611.

(90) Whitford, W.; Lehnert, K.; Snell, R. G.; Jacobsen, J. C. Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. *Journal of biomedical informatics* **2019**, *94*, 103174.

(91) Lin, X. Genomic variation prediction: a summary from different views. *Frontiers in Cell and Developmental Biology* **2021**, *9*, 795883.

(92) Maurano, M. T.; Humbert, R.; Rynes, E.; Thurman, R. E.; Haugen, E.; Wang, H.; Reynolds, A. P.; Sandstrom, R.; Qu, H.; Brody, J. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **2012**, *337* (6099), 1190−1195.

(93) Yue, F.; Cheng, Y.; Breschi, A.; Vierstra, J.; Wu, W.; Ryba, T.; Sandstrom, R.; Ma, Z.; Davis, C.; Pope, B. D. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **2014**, *515* (7527), 355−364.

(94) Klann, T. S.; Barrera, A.; Ettyreddy, A. R.; Rickels, R. A.; Bryois, J.; Jiang, S.; Adkar, S. S.; Iglesias, N.; Sullivan, P. F.; Reddy, T. E. Genome-wide annotation of gene regulatory elements linked to cell fitness. *bioRxiv* **2021**, 2021.2003. 2008.434470.

(95) Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nature genetics* **2019**, *51* (1), 12−18.

(96) Rizzo, R.; Fiannaca, A.; La Rosa, M.; Urso, A. A deep learning approach to DNA sequence classification. In: *Computational Intelligence Methods for Bioinformatics and Biostatistics: 12th International Meeting*; CIBB 2015, Naples, Italy, September 10−12, 2015, Revised Selected Papers 12; Springer, 2016; pp 129−140.

(97) Toneyan, S.; Tang, Z.; Koo, P. K. Evaluating deep learning for predicting epigenomic profiles. *Nature machine intelligence* **2022**, *4* (12), 1088−1100.

(98) Ainscough, B. J.; Barnell, E. K.; Ronning, P.; Campbell, K. M.; Wagner, A. H.; Fehniger, T. A.; Dunn, G. P.; Uppaluri, R.; Govindan, R.; Rohan, T. E.; Griffith, M.; Mardis, E. R.; Swamidass, S. J.; Griffith, O. L. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature genetics* **2018**, *50* (12), 1735−1743.

(99) Llinares-López, F.; Berthet, Q.; Blondel, M.; Teboul, O.; Vert, J.-P. Deep embedding and alignment of protein sequences. *Nat. Methods* **2023**, *20* (1), 104−111.

(100) Liu, L.; Meng, Q.; Weng, C.; Lu, Q.; Wang, T.; Wen, Y. Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data. *PLOS Computational Biology* **2022**, *18* (7), No. e1010328.

(101) Wang, W. Y.; Barratt, B. J.; Clayton, D. G.; Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **2005**, *6* (2), 109−118.

(102) Rebbeck, T. R.; Mitra, N.; Wan, F.; Sinilnikova, O. M.; Healey, S.; McGuffog, L.; Mazoyer, S.; Chenevix-Trench, G.; Easton, D. F.; Antoniou, A. C. Association of type and location of BRCA1 and BRCA2 mutations with risk of breast and ovarian cancer. *Jama* **2015**, *313* (13), 1347−1361.

(103) Olivier, M.; Hollstein, M.; Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology* **2010**, *2* (1), a001008.

(104) Single-cell multi-omics allows functional characterization of structural variants. *Nat. Biotechnol.* **2023**, *41*, 771−772.

(105) Cao, Z.-J.; Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **2022**, *40* (10), 1458−1466.

(106) Akram, F.; Haq, I. u.; Sahreen, S.; Nasir, N.; Naseem, W.; Imitaz, M.; Aqeel, A. CRISPR/Cas9: A revolutionary genome editing tool for human cancers treatment. *Technology in Cancer Research & Treatment* **2022**, *21*, 15330338221132078.

(107) Fu, Y.; Reyon, D.; Joung, J. K. Targeted genome editing in human cells using CRISPR/Cas nucleases and truncated guide RNAs. In *Methods in Enzymology*; Elsevier; 2014; Vol. *546*, pp 21−45.

(108) Lalonde, S.; Stone, O. A.; Lessard, S.; Lavertu, A.; Desjardins, J.; Beaudoin, M.; Rivas, M.; Stainier, D. Y.; Lettre, G. Frameshift indels introduced by genome editing can lead to in-frame exon skipping. *PloS one* **2017**, *12* (6), No. e0178700.

(109) Xiang, X.; Zhao, X.; Pan, X.; Dong, Z.; Yu, J.; Li, S.; Liang, X.; Han, P.; Qu, K.; Jensen, J. B.; Farup, J.; Wang, F.; Petersen, T. S.; Bolund, L.; Teng, H.; Lin, L.; Luo, Y. Efficient correction of Duchenne muscular dystrophy mutations by SpCas9 and dual gRNAs. *Molecular Therapy-Nucleic Acids* **2021**, *24*, 403−415.

(110) Tasan, I.; Jain, S.; Zhao, H. Use of genome-editing tools to treat sickle cell disease. *Human genetics* **2016**, *135*, 1011−1028.

(111) Kim, H. J. Genetic testing and genetic counseling. *Journal of the Korean Medical Association* **2006**, *49* (7), 603−611.

(112) Savanevich, A.; Vasilkevich, M.; Abdrashitov, V.; Stepuro, T. BRCA1 and BRCA2 genes mutations among women with clinical signs of hereditary breast cancer in western Belarus. *The Journal of V. N. Karazin Kharkiv National University, series Medicine* **2021**, No. 42, 68−76.

(113) Chan, C. L.; Ode, K. L.; Granados, A.; Moheet, A.; Moran, A.; Hameed, S. Continuous glucose monitoring in cystic fibrosis-A practical guide. *Journal of cystic fibrosis* **2019**, *18*, S25−S31.