Genome Biology

REVIEW

Open Access

In silico methods for predicting functional synonymous variants

Brian C. Lin[1], Upendra Katneni[1], Katarzyna I. Jankowska[1], Douglas Meyer[1] and Chava Kimchi-Sarfaty[1*]

*Correspondence: Chava.Kimchi-Sarfaty@fda.hhs.gov

[1] Hemostasis Branch 1, Division of Hemostasis, Office of Plasma Protein Therapeutics CMC, Office of Therapeutic Products, Center for Biologics Evaluation and Research, US FDA, Silver Spring, MD, USA

Abstract

Single nucleotide variants (SNVs) contribute to human genomic diversity. Synonymous SNVs are previously considered to be "silent," but mounting evidence has revealed that these variants can cause RNA and protein changes and are implicated in over 85 human diseases and cancers. Recent improvements in computational platforms have led to the development of numerous machine-learning tools, which can be used to advance synonymous SNV research. In this review, we discuss tools that should be used to investigate synonymous variants. We provide supportive examples from seminal studies that demonstrate how these tools have driven new discoveries of functional synonymous SNVs.

Background

The primary source for evolutionary diversity is genetic variation [1, 2]. Single nucleotide variants (SNVs) make up only ~ 0.1% of the entire human genome but are responsible for differences in the human population, including disease susceptibility and response to drugs [3]. SNVs can be divided into nonsynonymous variants, which alter the encoded amino acids, or synonymous variants that alter the codon sequence, but preserve the native amino acid structure. While the effects of nonsynonymous variants are evident, synonymous variants have been assumed to be neutral and yield minimal functional consequences. Compelling evidence over the last decade has disputed this view, and both in silico and experimental studies have revealed a variety of effects of synonymous variants, spanning from alterations to RNA structure to changes in protein expression and function to engendering adaptive evolution [4–7]. In fact, synonymous variants have now been implicated in cancers [8] and over 85 genetic diseases [9] and are responsible for many cellular disruptions at both the RNA and protein levels [7, 10]. The most prominent effects include changes to RNA structure/stability [11], splicing [12, 13], and miRNA binding [14, 15]. As these mechanisms mostly result from direct changes to the nucleotide sequence, in silico tools have been applied in both the discovery of pathogenic synonymous variants and in their characterization [16, 17]. To date, many notable studies on synonymous variants have implemented a dual strategy: first, using in silico tools to screen and predict for functional variants, and second, applying sensitive experimental techniques to validate these in silico predictions [7, 9, 11, 18–20]. Undoubtedly, the rising incorporation of computational approaches in biological research has driven a significant increase in discoveries of functional and pathogenic synonymous variants [21]. Though still in its infancy, many in silico variant predictors represent promising methods to distinguish between pathogenic and benign synonymous variants [22–24].

In addition, the computational field has also undergone a significant transformation. Through machine-learning (ML) and deep learning (DL) platforms, in silico tools have evolved to better integrate biological factors and experimental data into their algorithms [25]. Many tools use publicly available genetic datasets to train the ML systems to better predict functional variants [26–28]. New tools continue to be developed with unprecedented improvements in predictability and accuracy, and in many cases, substantial updates have been released, which have refined many popular existing tools. As researchers continue to acknowledge the importance of sequence properties, such as codon usage and GC content, in determining protein characteristics and new metrics and resources have been adopted for their evaluation [29–33], these dimensions have further enriched prediction models. Currently, well over a hundred tools have been used to characterize variants, each with their own specific predictive algorithms, but also with limitations that must be accounted for. While in silico tools have advanced research, their rapid development has also posed a conundrum of whether a single tool is preeminent or if multiple tools should be used. To realize the full potential of these in silico tools in synonymous variant research, further integration of these tools into a consistent workflow and substantiation of the predicted results through experimental data are required.

In this review, we highlight the process by which in silico tools should be used to effectively characterize synonymous variants (Fig. 1), while providing numerous examples from studies that have successfully implemented these methods. We characterize the differences among in silico tools by sorting them into sections based on their intended functions and provide a framework for how these tools should be optimally used to investigate various effects of synonymous variants. This review will discuss the most commonly utilized tools and introduce many that were more recently developed to provide a thorough resource for applying in silico tools in the study of synonymous variants.

In silico resources for assessing codon usage and sequence properties of synonymous variants

Genomes of most organisms are degenerate with multiple different codons translated into the same amino acid. However, synonymous codons are not used in a uniform fashion and genomes are biased to favor particular codons. Sharp and Li characterized this codon usage bias (CUB) in *Escherichia coli and Salmonella typhimurium* genes by introducing two metrics, Codon Adaptation Index (CAI) and Relative Synonymous Codon Usage (RSCU) [34]. Around the same time, another popular measure of CUB was devised called the expected number of codons (ENC), quantifying how far a gene's codon usage deviates from equal usage of synonymous codons [35]. These metrics formed the original systems to score gene level CUB, computing the difference between scores assigned to wildtype sequences and sequences containing synonymous variants.

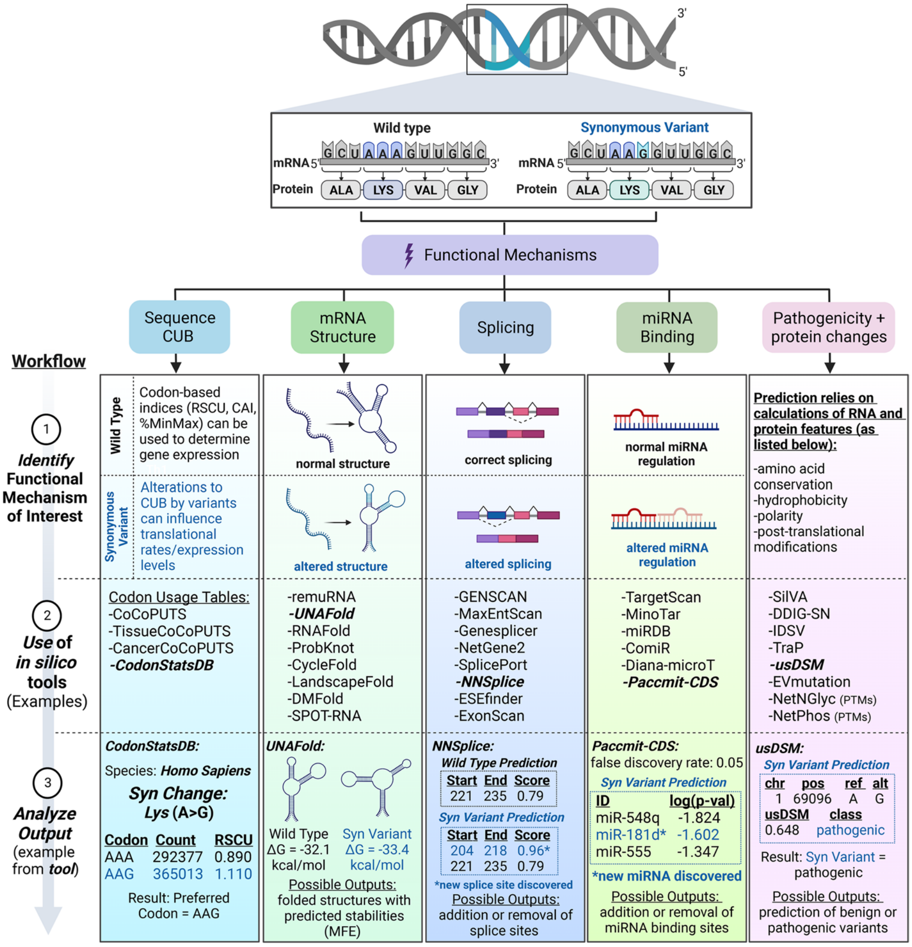

Fig. 1 Workflow schematic for how to optimally use in silico tools to investigate synonymous variants. Genetic sequences containing synonymous variants can cause many different functional effects, including alterations to codon usage biases, mRNA structure, splicing, miRNA binding, disease pathogenesis, and protein characteristics. After (1) identifying a functional mechanism of interest, (2) a variety of different in silico tools can be chosen and applied to evaluate the sequence containing synonymous variants. After the sequence has been processed, (3) outputs of these tools can be analyzed to form predictions. For proper evaluation, most tools will require input of a short nucleotide sequence containing the synonymous variant. The wild-type sequence for the identical region encompassing the synonymous variant should be processed for comparison. Examples of potential outputs for tools highlighted in row 2 are shown in row 3. CodonStatsDB determines codon preferences based on RSCU values. UNAFold can generate predicted mRNA structures and calculate differences in mRNA stability. NNSplice will reveal any new or lost splice sites. Paccmit-CDS is able to capture changes to miRNA binding sites. usDSM is able to predict the pathogenicity of the variant. Outputs may vary depending on the algorithms and structure of the tools. It is highly beneficial to analyze the sequence through multiple tools and to validate the results through experimental methods

Today, while these methods continue to be used extensively, new insights into translational processes have led to the creation of additional methods to quantify CUB. Commonly used codons are thought to correlate with more abundant tRNAs [36–38], leading to the development of the tRNA Adaptation Index (tAI) based on tRNA usages [39] and a species-specific tAI calculator (stAI$_{calc}$) that infers organism-specific tAI wobble weights for 100 different species [40]. In addition, non-random codon biases have been found to impact translation kinetics and co-translational folding [31–33, 41–44]. Moura et al. reported that both missense and synonymous mutations are under selective pressure to maintain usage of codon multiples in bacteria, archaea, and eukaryotes [45]. Codon pairs, two adjacent codons (i.e., bicodon), also exhibit usage biases that have been found to impact translational efficiencies [46]. Others have reported that codon pair frequencies provide no additional information towards predicting expression than single codon frequencies in S. cerevisiae [47] and that viral codon pair usage bias is dictated primarily by avoiding certain dinucleotides [48]. By distinguishing rare or optimal codons, many metrics can be used to identify synonymous variants that impact protein properties through disrupting translational kinetics and co-translational folding [49]. For this purpose, Rodriguez et al. developed the %MinMax tool to calculate synonymous codon usage with a focus on measuring deviations in optimal cotranslational folding patterns [29].

Furthermore, in multicellular organisms, CUB can vary across different tissue contexts. Plotkin et al. reported tissue-specific codon usage patterns by comparing groups of human genes previously reported to be expressed in specific tissues [50]. Similarly, Qingpo Liu found differences in codon usage between tissue-specific genes in rice [51]. tRNA expression differs among human tissues [52]. Therefore, CUB metrics should incorporate tissue-specific contexts into its calculations. In recent years, two databases have been assembled to aid in these tissue-specific calculations: TissueCoCoPUTs, which uses transcriptomic data from different tissue contexts to compute a weighted average codon usage in several different tissue contexts [32] and CancerCoCoPUTs, which reports differences in codon usage across several different solid tumor types [33]. These resources, along with large databases, such as the Codon Statistics Database [53], have made it remarkably effortless to evaluate CUB and sequence properties of synonymous variants.

In silico tools for assessing the effect of synonymous variants on mRNA structure and stability

Synonymous variants can have functional and disease consequences through altering mRNA secondary structure and stability. Encoded within the primary mRNA sequence is the information to establish local mRNA secondary structure motifs and dictate RNA stability of individual regions, which can determine the accessibility of ribosome binding sites and speed of local translation [54–56]. One seminal discovery in the field of synonymous variants was the observation that in the mutated *CFTR* gene (c.1520_1522delTCT), which causes cystic fibrosis, a single synonymous variant (c.507 T > A) [18, 57, 58] caused the formation of two enlarged loops in the mRNA structure [18]. This deviation correlated with a reduction in translational rate and reduced expression of the CFTR protein [18]. While this finding was validated experimentally through RNA folding assays and circular dichroism analysis, like many other studies, its initial discovery was uncovered through molecular modeling.

In essence, RNA structure and its folding process have been found to be deeply rooted in a couple of principles, which has inspired the development of RNA structure prediction tools. First, RNA secondary structure evolutionarily favors stability, except for select situations where unstable areas in the transcript, such as at the 5′ end, supports translational initiation [59–61]. Stable RNA provides many benefits, including increased half-life, fine-tuning of translational speed, and establishing favorable binding sites for

RNA-binding proteins and miRNAs [62, 63]. mRNA conforms to structures that more easily maintain its structural integrity, which in most cases, the realized structure is one that possesses the lowest free energy [64, 65]. However, although a single structure may be the most stable and dominant, multiple structures co-exist within the dynamic cellular environment. RNA populates a heterogeneous ensemble of conformations, and the goal of most prediction tools is to differentiate the native structure from its numerous subpopulations [66]. Second, across species, coding regions contain many structurally conserved elements [59, 67–69], which can be used to infer both function and structure. Based on these assumptions, many tools have been established with algorithms designed to identify the minimum free energy (MFE) structure with consideration of conserved motifs, temperature, ion concentrations, and sequence-based properties.

In silico tools, such as mFold [70] (recently updated and renamed to UNAFold [71]), remuRNA [72], Kinefold [73], CoFold [74], and RNAfold [75], are examples of tools that predict structures based on algorithms to minimize free energy. These tools require input of RNA sequences with recommended length limit of < 1500 nucleotides as longer sequences significantly increase folding complexities and software run-time. These tools are extensively used to generate predicted mRNA structures due to their reputable accuracy and fast computing speed. For example, mFold was used in the CFTR study to reveal structural loop elements in the mutated CFTR structure [18]. Likewise, Duan and colleagues [11] used mFold to show that one synonymous mutation (c.957C > T) in human *DRD2* (dopamine receptor D2) led to decreased mRNA stability and decreased expression. In a separate study, mFold, Kinefold and NUPACK [76] were used collectively by Simhadri and colleagues to highlight how a *F9* (Factor IX) synonymous variant (c.459G > A) alters mRNA structure to facilitate changes in protein expression [77].

As applied in these aforementioned studies, prediction tools can be used to simulate folding of both the wild type and mutant sequences and to calculate the free energy of the best candidate structures. A single synonymous variant can perturb the conformational ensemble and shift folding dynamics, thereby forming misfolded or non-native structures of higher or lower free energy (ΔG). Any observed difference in predicted minimum free energies (ΔΔG) between wild type and mutated structures may suggest a change in mRNA structure (example workflow is shown in Fig. 1). The significance of a change in MFE may vary among RNA structures and can be affected by various input parameters. Wayment-Steele and colleagues found that increasing the simulated folding temperature can improve the correlation of predicted structures to experimental data [78]. In addition, sequence length is another factor that can alter the magnitude of MFE differences due to the added complexity of folding larger structures and should be a variable closely considered [16]. Due to these potential factors, these tools provide an effective method to screen for potential RNA structural changes, but results do require further validation through experimental methods.

Additionally, while RNA prediction tools based on MFE are effective at accurately rendering RNA structures that are composed of a high number of canonical Watson–Crick base pairs, RNA folding is dynamic and complex. New insights into the structural topology of RNA has revealed special base pairing configurations, such as pseudoknots and noncanonical intramolecular base pairing patterns that support specific structural contexts (i.e., geometric motifs, higher-order multiplexes) and tertiary interactions [79].

Noncanonical base pairs are base interactions that deviate from the standard Watson–crick base pairings, such as G-A pairs, and pseudoknots are non-nested structures that form from two stem-loops. In consideration of these features, ProbKnot [80], IPKnot [81], Knotty [82], and LandscapeFold [83] are dependable tools used for pseudoknot predictions and MC-Fold-DP [84] and CycleFold [85] are equipped with special features to handle noncanonical base pairs. These are powerful tools that employ sophisticated algorithms to include special base pairings and improve prediction performance but can only consider small nucleotide sequences due to computation times. Nevertheless, shorter sequences can provide significant information about the effects of synonymous variants on mRNA structure, in which subtle changes may occur locally.

New machine-learning approaches are able to circumvent computational time issues because these techniques are data-driven approaches rather than score-dependent. Two ML tools, DMfold [86] and SPOT-RNA [87, 88], have been generated with accuracies that supersede existing tools. These multivariate tools are able to consider free energy parameters, sequence characteristics, and other properties while having the unique advantage of using genetic databases and RNA structure datasets for model training. However, because of their novelty, these ML approaches remain relatively enigmatic, and there remain concerns of potential issues with overfitting and inaccuracies in predicting structures that are more dissimilar to structures that appeared in training sets. Nevertheless, these ML techniques represent the most promising methods for predicting RNA structures and the performance of these tools will likely continue to improve as more publicly available RNA data is collected. Similar to the state of ML RNA prediction tools, computational 3D modeling of complex RNA structures remains a significant challenge but has undergone significant improvements in recent years as more RNA structures have been revealed experimentally and computationally [89]. Eterna (https://eternagame.org/), a crowdsourcing initiative, has rapidly accelerated discoveries in the RNA field and has stimulated improvements in the design of RNA structures for RNA-based therapeutics [78, 90]. Current 3D modeling can be separated into 3 approaches: (i) comparative modeling, in which RNA structures are predicted based on homologous structures (e.g., ModeRNA [91], RNABuilder [92]); (ii) fragment assembly, whereby RNA structures are decomposed into fragments and compared to the target sequence for assembling a predicted structure (e.g., RNAComposer [93], VfoldLA [94]); and (iii) de novo modeling, which relies on coarse grained molecular dynamics and knowledge-based force-field principles to generate structures (e.g., SimRNA [95], iFoldRNA [96]). Many recent reviews and methodology articles provide a thorough overview of the applications of RNA 3D modeling tools [89, 97]. For synonymous variant research, 3D RNA modeling tools have not yet been implemented, but with rapid advancements in this growing field, these tools may be applicable in the near future.

Ultimately, assessing RNA structure with a combination of tools that employ various algorithms and parameters is the most optimal approach to evaluate synonymous variants. Agreement between prediction tools increases confidence in predicted structures, while disagreement suggests that the RNA structure is complex. Recently, computational tools, such as SSRTool [98], have been generated with the goal to distinguish the most likely native structure after assessing predictions from a large class of selected prediction tools. However, when tested against known RNA structures from various different species, the tool was unable to guarantee an optimal structure prediction. Therefore, we recommend the use of multiple tools to evaluate synonymous variants and to complement these in silico studies with experimental approaches. A comprehensive list of tools used for assessing synonymous variants is shown in Table 1.

In silico tools for determining effects of synonymous variants on RNA splicing

Pre-mRNA splicing is the co-transcriptional process of excising non-coding introns and joining protein-coding exons. Splicing is mediated by the spliceosome complex, composed of five small nuclear ribonuclear proteins (snRNPs) and more than 150 proteins, and involves recognition of *cis*-acting elements, including 5′ and 3′ splice sites (donor and acceptor sites, respectively), branch point sequences, and polypyrimidine tract (PPT) [103]. A majority of the splice sites (> 98%) have invariant GT and AG as the first and last two intronic nucleotides, respectively, and less conserved sequences in the remaining splice site sequence [104]. Furthermore, there are *cis*-acting splicing regulatory elements (SREs) in both exons and introns that regulate splicing. The SREs are 6 to 8 nucleotides long and can positively (enhancers) or negatively (suppressors) affect splicing through recruiting trans-acting serine/arginine-rich (SR) proteins or heterogeneous nuclear ribonucleoproteins (hnRNPs), respectively.

Synonymous variants can either disrupt native splice sites, create de novo splice sites, activate cryptic splice sites, or affect SREs (those located in exons are called exonic splicing enhancers (ESEs) or exonic splicing silencers (ESSs)) and result in variable outcomes, including exon skipping and partial exon deletions [105]. Splicing dysregulation is arguably the best studied mechanism by which synonymous variants affect phenotypes and thus far has been implicated as the primary underlying mechanism for a majority of diseases caused by these variants [7]. A plethora of in silico tools have been developed for predicting the effects of genetic variants on splicing (Table 2).

These tools can be broadly categorized as motif-based or ML- and DL-based algorithms [117]. Splice Site Finder-like (SSF-like, embedded in other platforms referenced below), Genscan [106], Genesplicer [108] and MaxEntScan (MES) [107] are examples of tools employing motif-based algorithms. Specifically, Spliceview and SSF-like employ position weight matrices (PWM) [118] to derive potential splice-site strength estimates for a sequence. Genscan uses a maximal dependence decomposition (MDD) model, which is a decision tree-based method that attempts to capture dependencies between both adjacent and non-adjacent positions. Genesplicer combines MDD with Markov models (MM) to capture additional dependencies between neighboring positions. MES uses maximum entropy principle (MEP) for modeling short sequence motifs found in splice sites while also accounting for higher-order dependencies between adjacent and non-adjacent positions. Some tools combine multiple algorithms or tools for their SS predictions. For example, Human Splicing Finder (HSF) [119] uses both PWM and algorithms from MES. On the other hand, SPiCE (Splicing Prediction in Consensus Elements) [120] uses logistic regression to combine MES and SSF-like tool predictions.

Increasingly, tools employing ML-based algorithms are being developed for SS prediction. NetGene2 [109], NNSplice [121], Alternative Splice Site Predictor (ASSP) [122], Spliceport [110], SpliceAI [123], MMSplice [111], and SpliceRover [124] are some examples in this category. Of these, NNSplice, NetGene2, and ASSP employ neural networks

Table 1 In silico tools for predicting effects of synonymous variants on mRNA structure

Prediction Algorithms	Tool	Input	Output	Special features	Notes	URL	Ref
Free energy minimization	CoFold	Single nucleotide sequence: limit of 50 kb; 1500 nt sequence has a run-time of approximately ~ 15 s	Structure diagram + predicted MFE (visualized as an arc plot and supports many other output formats)	Two different thermodynamic parameter options + scaling choices	Algorithm considers co-transcriptional folding to improve accuracy of predicting structure of longer sequences	https://e-rna.org/cofold/	[74]
	remuRNA	Wild-type and mutant sequence (no upper limit)	Structure diagram + predicted MFE; relative entropy plot		Algorithm incorporates relative entropy between Boltzmann ensembles of wild-type and mutant secondary structures	https://github.com/bgruening/galaxytools/tree/master/tools/rna_tools/remurna	[72]
	RNAfold	Single nucleotide sequence: 7500 nt limit for partition function calculations; 10,000 nt limit for free energy minimization prediction	Interactive RNA secondary structure plot; mountain plot	Parameter options to avoid isolated base pairs, to use partition function, and/or exclude GU pairs at end of helices	Algorithm employs partition function calculations in addition to free energy minimization	http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi	[75]
	UNAFold	Single nucleotide sequence or multiple short sequences (no upper limit)	Predicted MFE; circular structure plot; energy dot plot	Parameter constraint options to optimize loop types, base numbering frequencies, regularization angles	Uses free energy minimization program with folding temperature fixed at 37 °C	http://www.unafold.org/	[71]
Pseudoknots	IPKnot	Single or multiple sequences (FASTA format or multiple sequence alignments)	2D diagram using VARNA program and structure as an arc plot	Options between multiple scoring models and prediction complexity levels	Uses integer programming to compute the maximum expected accuracy structure (MEA)	http://rtips.dna.bio.keio.ac.jp/ipknot/	[81]
	Kinefold	Single sequence (no upper limit)	Lowest free energy structure diagram + predicted MFE; folding path movie; helix tracing graph	Stochastic simulation—co-transcriptional folding or renaturation folding	Stochastic folding simulations using folding dynamic algorithms [99] and physical constraint modeling for pseudoknot prediction	http://kinefold.curie.fr/	[73]
	Knotty	Single sequence (no upper limit)	Structure diagram + predicted MFE; provides information on all candidate structures		Predicts complex pseudoknot structures with optimization of run-time through sparsification technique and a CCJ-type algorithm	https://github.com/HosnaJabbari/Knotty	[82]
	Landscape-Fold	A list of sequences (up to 2), option to consider intramolecular pseudoknots and define minimum number of nucleotides within each hairpin	Identifies all possible structures and provides indexing/sorting via MFE and equilibrium probabilities	Multiple sequence structural analysis for predicting base interactions with option to assess equilibrium concentrations	Polymer physical model based on entropy calculations of arbitrary pseudoknotted structures	https://github.com/ofer-kimchi/RNA-FE-Landscape	[83]
	ProbKnot	Single sequence (no upper limit)	Base pair probability plot	Optimization of iterations and minimum helix length	Predicts for presence of pseudoknots in sequence	https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/ProbKnot/ProbKnot.html	[80]

Table 1 (continued)

Prediction Algorithms	Tool	Input	Output	Special features	Notes	URL	Ref
Noncanonical base pairings	CycleFold	Single or multiple sequences (no upper limit), can apply maximum expected accuracy (MEA) or ProbKnot to generate structures	Lowest MFE structure, matrix of pairing probabilities between each nucleotide sequence	TurboFold mode can be engaged to process multiple sequences, considers evolutionary conservation	Uses nucleotide cyclic motifs to predict noncanonical base pairings and minimizes free energy	http://rna.urmc.rochester.edu	[85]
	MC-Fold-DP	No sequence limit, but runtime scales polynomially	Returns all structures within energy band above the ground state	Cannot currently consider pseudoknots	Prediction based on combining small nucleotide cyclic motifs	https://hackage.haskell.org/package/MC-Fold-DP	[84]
Machine-learning	DMfold	Single sequences (no upper limit)	Folded RNA structure and energy model	Folding parameters automatically determined based on deep learning	Deep-learning and improved base pair maximation principles; trained with 3948 known RNA primary sequences [100]	https://github.com/linyuwangPHD/RNA-Secondary-Structure-Database	[86]
	SPOT-RNA	Single sequence (maximum—2000 nts); can run longer sequences or batch sequences locally	2D plots of structure through VARNA visualization tool, output of secondary structure motifs can be seen through Vienna format		Deep contextual neural network implemented with model training and transfer learning from high quality datasets of > 10,000 RNA structures; trained with bpRNA [101] and PDB [102] databases	https://sparks-lab.org/server/spot-rna/	[87, 88]

Table 2 Select list of in silico tools for predicting mRNA splicing effects

Tool	Algorithm/prediction method	Input	Output	URL/comments	Ref
Splice site prediction tools					
GENSCAN	Motif-based, maximal dependence decomposition (MDD)	Sequences up to 1 million nucleotides can be analyzed	Predicted exons and/or peptides in the sequence	http://hollywood.mit.edu/GENSCAN.html A copy of the program is available on request	[106]
MaxEntScan	Motif-based, based on maximum entropy principle (MEP)	9 and 23 nucleotide long sequences for donor and acceptor site predictions respectively	Tool provides a score for the sequence indicating its strength as splice site	http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html Offers multiple scoring models as options Perl scripts to run algorithm are available for download	[107]
Genesplicer	Motif-based, MDD with Markov models	Sequences of up to 200,000 nucleotides	Predicted acceptor and donor sites in the sequence with scores	https://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml Program is available for download	[108]
NetGene2	Machine-learning (ML) based, neural networks	One sequence between 200 and 100,000 nucleotides	Program provides predicted acceptor and donor sites with a confidence score	https://services.healthtech.dtu.dk/service.php?NetGene2-2.42 Program is available for download. The training dataset curated from NCBI GenBank included 65 human genes with 331 donor and acceptor splice sites	[109]
Spliceport	ML-based, support vector machine	Sequence of up to 30,000 nucleotides	Spliceport provides a list a predictions for donor and acceptor sites along with a score	https://spliceport.cbcb.umd.edu/SplicingAnalyser2.html Training dataset included a collection of 4000 pre-mRNA human RefSeq sequences	[110]
MMSplice	ML(deep learning)-based, modular neural networks	Predictive analysis of variants is performed on any exon with 50 and 13 nucleotides upstream and downstream respectively	Predicts effects of variants on exon skipping, splice site choice, splicing efficiency, and pathogenicity	Models are available in the Kipoi repository Trained on distinct large-scale genomics datasets. Please refer to Table 1 of citation for a detailed summary of trained modules and models	[111]
SRE prediction tools					
ESEfinder	Functional SELEX, PWM	Sequences of up to 5000 nucleotides	Tool provides predicted splice sites and SREs along with a score	http://krainer01.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home	[112]
EX-SKIP	Predictions are based on a ratio of ESE/ESSs identified from 5 different models including RECUE-ESE, FAS-ESSs	Two exonic sequences strictly in uppercase in FASTA format up to a total length of 4000 nucleotides	Program compares the ESE/ESS profile of a native and variant sequence and predicts the probability of exon skipping	https://ex-skip.img.cas.cz/	[113]

Table 2 (continued)

Tool	Algorithm/prediction method	Input	Output	URL/comments	Ref
FAS-ESS	Experimental verification of random decanucleotide (10-nucleotides) sequences and weight matrices of aligned sequences	Single or multiple sequences in FASTA format. Length limit not specified	Program shows predicted ESS motifs. Users can search for hex2 or hex3 sets with relatively higher sensitivity and specificity respectively	http://hollywood.mit.edu/fas-ess/	[114]
ESRSeq	Experimental assessment of the SRE properties of all possible hexamer motifs	6-hexamer sequences overlapping the variants	Calculate the net ESRseq score. Net score change could indicate prediction direction	Scores of ESE/ESS hexamers are available as supplementary materials	[115]
Combination analysis tools					
SROOGLE	Predictions based on 9 different models including MEP, PWM, ESEFinder, RESCUE-ESE, FAS-ESS, etc	Target exon along with flanking introns	Provides predictions for both splice sites and SREs	http://sroogle.tau.ac.il/	[116]
ExonScan	Splice site predictions based on MEP and SRE predictions based on RESCUE-ESE and FAS-ESS	DNA sequence with exons with at least 20 bases upstream of the exon it can predict and 60 bases of downstream of the last	Provides predictions for both splice sites and SREs	http://hollywood.mit.edu/exonscan/	[114]

algorithms, while Spliceport employs a support vector machine algorithm. Similarly, tools based on other ML algorithms like random forest, naïve Bayes, and decision trees have been developed. More recently, DL technique-based tools employing deep/convoluted neural networks were developed, including SpliceAI, MMSplice, and SpliceRover [125]. These tools have exhibited promising results and are touted for freeing algorithms from the constraints of human intervention, while enabling the use of novel methods and parameters to identify splice sites and classify nucleotide variants [117].

Similar to splice site prediction, a variety of tools for predicting a genetic variant's effect on SREs have been developed. ESEFinder [112], RESCUE-ESE [126], and FAS-ESS [114] were among the earliest developed SRE prediction tools. ESEFinder employs PWMs supported by functional SELEX ((Systematic Evolution of Ligands by EXponential enrichment) screen data to predict ESEs in the targeted sequence. RESCUE-ESE (Relative Enhancer and Silencer Classification by Unanimous Enrichment) employed a hybrid computational-experimental approach where putative ESEs were first predicted computationally and then experimentally verified by minigene assays. FAS-ESS employed experimental procedures (similar to functional SELEX) to screen random decanucleotide sequences and identify ESSs in the exon sequences. ESRSeq [115] and HEXplorer [127] are more recently developed tools for SRE prediction in exons. Of these, ESRseq analyzed the effects of all possible (4096) hexamer sequences on splicing using a minigene assay and categorized them as either ESEs or ESSs and assigned a score depending on the strength of effect. HEXplorer on the other hand employs a RESCUE-type in silico approach to categorize and assign scores for hexamer sequences. Additionally, tools like EX-SKIP [113] combine predictions of ESE/ESSs from multiple methods, including RESUCE-ESE and FAS-ESS and assign a score based on their relative density to indicate their ability to induce exon skipping.

A select list of tools performs predictions for both splice sites and SREs. For example, SROOGLE [116] provides predictions for both splice sites and SREs along with branch point sequences and PPT using 9 different algorithms. HSF provides splice site, SRE and BP predictions employing multiple algorithms. Similarly, ExonScan [114] provides splice site predictions using maximum entropy model and SRE predictions using RESCUE-ESE and FAS-ESS approaches.

The above discussed in silico tools were successfully used for the evaluation of splicing effects of genetic variants by multiple studies [12, 128–131]. Zhou et al. employed HSF and ESEFinder for the evaluation of naturally occurring synonymous variants in the *ATP7B* Gene [129]. Zhang et al. used SpliceSiteFinder-like, MaxEntScan, NNsplice, GeneSplicer, and HSF for the assessment of *F9* synonymous variants [130]. Overall, users have access to a large variety of tools. A majority of the tools provide scores indicating the strength of the splice site or SREs in a sequence of interest. A measure of change in score between native and variant sequences generally indicates the effect of the variant on splicing. While higher score changes generally indicate greater impact on splicing, there is no consensus on a threshold/cut-off score. Several studies were conducted to compare the performance of tools [132]; however, they are incomparable as they varied in both tools studied and test datasets and consequently differed in their conclusions. A recent comparative study with tools based on both motif-based and ML-based algorithms showed variable tool performances depending on the context of the test dataset

[117]. Generally, predictions for variants located within consensus splice sites tend to be more accurate than for deep exonic variants [12]. For optimal use, the user needs to understand the features and limitations of individual tools. For example, the length of consensus SSs used in training varies between tools and not all tools were trained to identify noncanonical SSs (e.g., GC-AG and AT-AC). The presence and/or lack of tissue-specific splicing events in the training datasets could also influence predictions [117]. The type of input sequence required by tools, ability to perform batch analysis and the availability of source code will also influence tool choices. Use of a combination of tools predicting both SSs and SREs and employing different algorithms is recommended to overcome potential deficiencies of a single tool and is expected to improve predictive values [12, 132, 133].

In silico tools for predicting the effect of synonymous variants on miRNA binding

miRNAs, short (17–22 nucleotides) single, non-coding RNAs, bind to the complementary sequences of target proteins and regulate their expression [134]. miRNA genes are located either in intergenic regions or within introns of protein coding genes. miRNA expression is cell-type and cell-state specific [135], and genetic variants can affect the gene regulation network. Numerous studies have demonstrated that single nucleotide variants within the miRNA or mRNA untranslated regions (UTR) can affect mRNA-miRNA interactions [136, 137], dysregulate protein expression by causing the gain or loss of miRNA binding sites within the gene's coding sequence (CDS) [138], and may lead to disease pathogenesis [136]. In fact, recent studies estimated that nearly half of sSNVs can affect miRNA binding, disturb protein functions, and increase disease risk [15]. For example, a synonymous variant (c.313C > T) in *IRGM* disturbs the miR-196 binding site and dysregulates IRGM-dependent xenophagy in Crohn's disease [14], and a synonymous variant (c.51C > T) in *BCL2L12*, identified in melanoma tumors, causes loss of the miR-671-5p binding site that stimulates protein expression [139].

The mechanism underlying miRNA association is complex and not fully understood, but the main interaction occurs via the 5′ seed region (nucleotides 2–8). Additional pairing at the 3′ end stabilizes the miRNA interaction [134]. Due to a non-perfect complementarity, miRNA can bind and regulate multiple genes through multiple binding sites either in the UTR or CDS regions [140].

As miRNAs regulate gene expression mainly by binding to their target sequence within 3′ untranslated region (3'UTR), most in silico tools have predominantly focused on miRNA target site predictions within the UTR [141]. Nevertheless, a few tools are currently available to identify miRNA target sites within the CDS and to study the effect of synonymous variants (Table 3). A large list of miRNA target prediction tools can be found on the Tools4miRs platform, which has amassed over 170 methods for broadly defined miRNA analysis (https://tools4mirs.org/). Here, we focused on tools that can be used to investigate genetic variants within the coding region.

TargetScan predicts biological targets of miRNAs by searching for the presence of conserved motifs (mer sites) within the gene that matches the miRNA seed region [142]. The online version of the tool is limited to the reference gene and is not specifically

Table 3 In silico tools for assessing effects of synonymous variants on miRNA binding

Tool	Algorithm/prediction method	Output/score	Year	URL	Ref
TargetScan and Target Scan S	Sequence alignment	Weighted context + + score (from -1 to 1). The scores with a lower negative value indicate a greater prediction of repression	2005	https://www.targetscan.org/vert_80/	[142, 143]
MinoTar	Sequence alignment and conservations scoring	Probability Conserved Targeting	2010	https://www.flyrnai.org/cgi-bin/DRSC_MinoTar.pl	[144]
miRDB (MirTarget)	Machine Learning (Support vector machine [SVM])	Target prediction scores between 50 and 100. A predicted target with prediction score > 80 is most likely to be real	2020	http://mirdb.org/	[145, 146]
ComiR	Machine learning (support vector machines)	Ranked vector of scores; therefore, each gene is associated with a reliability of being a target of the set of miRNAs given in input	2015 (updated in 2020 to include coding regions)	http://www.benoslab.pitt.edu/comir/help.html	[147, 148]
Diana-microT	microT-CDS algorithm	miTG score (from 0 to 1). The closer to 1, the greater the confidence	2009 (updated in 2013)	https://dianalab.e-ce.uth.gr/html/dianauniverse/index.php?r=microT_CDS	[149]
Paccmit-CDS	Ranking based on Markov model and sequence alignment	The predictions are ranked according to the *P*-value that the observed number of conserved and/or accessible seed matches would appear in the target sequence by chance	2015	https://paccmit.epfl.ch/	[150]
miRanda	Ranking based on seed match, conservation and free energy (G:U pairs allowed in the seed)	mirSVR score (< 0) is an estimate of the miRNA effect on the mRNA expression level. PhastCons score (0–1) measures the conservation of nucleotide positions across multiple vertebrates	2005 (updated in 2010)	https://cbio.mskcc.org/miRNA2003/miranda.html	[151, 152]
PITA	Ranking based on seed match, free energy, site accessibility and target-site abundance (G:U pairs allowed in the seed)	The predictions are ranked based on having a full match 7- or 8-mer seed and a conservation score of 0.9 or higher	2007	https://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html	[153]

designed to predict miRNA binding site within the CDS. To analyze custom sequences, TargetScan provides a downloadable version of the code.

Another tool, MinoTar (miRNA ORF Target), predicts miRNA binding sites within the CDS by identifying highly conserved regulatory motifs [144]. However, the current version of the tool limits the prediction to reference sequences.

miRNA database (miRDB) searches for miRNA target sites through a support vector machines (SVMs) algorithm and is trained with high-throughput experimental datasets. The database can perform predictions in the CDS but is limited to native gene sequences. The tool allows for analyzing any customer mRNA sequence using the 3′ UTR region model [145]. In addition, the database was recently updated with cell-specific miRNA targets [146, 154].

ComiR (Combinatorial miRNA targeting) uses predictions from four common algorithms (PITA [153], miRanda [151], TargetScan [142], miRSVR [155]) and converts the results into a single probabilistic score using ensemble learning to predict whether a given mRNA is targeted by a set of miRNAs [147, 156]. This tool can accommodate custom mRNA sequences. The current version focuses on prediction within the 3′ UTR region, but the database may soon be upgraded to include CDS binding sites along with miRNA expression data. Preliminary studies have shown that information contained in the CDS significantly improves the accuracy of ComiR predictions [148].

DIANA-microT-CDS can identify miRNA targets in the 3′ untranslated region (3′ UTR) and in the CDS [149]. This algorithm uses miRNA-recognition elements (MREs) for the miRNA:mRNA base pairing. The software provides an automatic pipeline as well as plug-ins that allow the user to access the target prediction server and incorporate advanced miRNA analysis into custom pipelines.

Paccmit-CDS (Prediction of Accessible and/or Conserved MIcroRNA Targets) searches for potential microRNA targets within CDS by identifying conserved complementary motifs to the microRNA seed region and ranking them with respect to a random background that preserves both codon usage and amino acid sequence [150]. The tool presented on the website allows for evaluation of reference genes, but the program written in C++can be used to evaluate the effect of synonymous variants. Paccmit-CDS, TargetScan, and miRDB prediction tools have been recently used to evaluate for the effect of synonymous variants in ADAMTS13 [157].

MiRanda, which is accessible online, allows searches for miRNA binding sites within the 3′ UTR region of specific genes, by inputting gene names. Installing the miRanda package allows for the detection of potential microRNA target sites in genomic sequences and can be used to evaluate the effect of synonymous variants [151, 152].

The online miRNA prediction tool, PITA, can process UTR sequences. While it is not designed to study miRNA binding sites within the CDS, it was previously used in concert with miRanda to identify miRNA target sites, encompassing the C51T variant site in BCL2L12 [139].

For validation of miRNA binding sites within the protein coding region, these prediction software require input of the gene sequence, which is then aligned with miRNA sequences derived from miRbase [158]. By comparing the outcome of the WT sequence, which is defined by a list of predicted miRNAs and with associated scores generated by specific prediction tools, with the list of miRNAs predicted to bind the variant sequence, the gain or loss of miRNA binding can be determined.

The main limitations of some current prediction algorithms are that they are based on conservation and are not fully adapted for processing the CDS. Many tools neglect consideration of cell-type specific miRNA expression levels, do not consider target site availabilities due to protein folding, and limit the analysis to a reference gene sequence. Since mRNA-miRNA association is based on non-perfect complementarity, the outcome data contains hundreds of predicted miRNAs, and it is advisable to validate miRNA predictions by comparing the output data from three or more prediction tools. As synonymous variant prediction outcomes within the CDS have not been extensively validated, and variants that have been experimentally assessed do not always support the prediction algorithms [159], it is difficult to recommend a specific tool that is best for forming SNV miRNA predictions. Nevertheless, many tools have recently evolved to include CDS analysis and the development of more robust bioinformatic and experimental methods to evaluate miRNA alterations by synonymous variants remains an ongoing pursuit.

In silico tools for predicting pathogenicity of synonymous variants

As more synonymous variants are being implemented in the development of genetic therapies and drugs, the creation of more powerful tools to predict functional synonymous variants has become even more important. Many discovered synonymous variants have been linked to increased risks for developing diseases and cancers [9]. For example, synonymous variants have been found to underlie Hemophilia [77, 160] and in cancer, about 6–8% of pathogenic single nucleotide substitutions identify as synonymous variants [161]. As a result, there is growing interest in the development of in silico tools that can reliably predict the pathogenicity of synonymous variants.

Currently, methods to predict rare coding variants, mostly targeting pathogenic missense variants, have proven to be quite effective, such as REVEL [162] and CADD [24]. However, progress towards predicting pathogenic synonymous variants remains far behind. While creating pathogenic synonymous variant prediction tools is complicated and challenging, recent progress towards this objective has come on the heels of advancements in ML platforms and greater insight on the importance of a variety of sequence properties in influencing disease. mRNA metrics and protein-associated variables, such as amino acid conservation, have been considered in algorithms to predict pathogenicity [21, 163]. In addition, generation of robust prediction tools is highly dependent on the availability of disease-associated genetic data that can be used to train ML systems. Numerous data sets have been curated with information on disease-related variants, such as Human Gene mutation database (HGMD) [164] and VariSNP [165], and there are numerous resources for curating neutral synonymous variants, including the 1000 Genomes Project (1000G) [166, 167]. But, while these are the most extensive datasets and have been used to train ML prediction tools, these datasets require further improvements. Unfortunately, as many have noted [168], there are inconsistencies in characterizations, nomenclature, and disease annotations in these databases, which have encouraged many recent efforts to correct these annotation flaws [169]. However, these factors have made it exceedingly difficult to generate accurate disease predictions.

Nevertheless, many ML tools based on supervised algorithms, such as random forests (RFs), deep neural networks, or support vector machine (SVMs), have been generated with reasonable proficiencies at predicting pathogenic synonymous variants. Some examples of such tools include SilVA (Silent Variant Analyzer) [22], DDIG-SN (Detecting Disease-causing Genetic SynoNymous variants) [23], IDSV (Identification of Deleterious Synonymous Variants) [163], and TraP (Transcript-inferred Pathogenicity) [170]. Each of these tools utilize a different assortment of features to predict pathogenicity of synonymous variants, but the most common implemented features include conservation, splicing, and RNA folding metrics. Most of these tools require a list of variants, formatted as VCF or tag-like files, and will rank synonymous variants based on their predicted pathogenicity. While it seems unreasonable to compare the accuracies of prediction tools due to the lack of an ideal standardized testing set, Zeng and colleagues found that when tested with a mock dataset, SilVA, DDIG-SN, and TraP were highly correlated in their predictive capacities but were not effective at large-scale variant predictions [171].

Ultimately, improvements in variant predictors will only occur with enhancements to genetic data sets. usDSM (Deleterious Synonymous Mutation Prediction using Undersampling Scheme) [172] and synVep (Synonymous Variant Effect Predictor) [21] are newer tools that have demonstrated improved proficiencies by implementing undersampling methods and positive-unlabeled learning, respectively, to circumvent the lack of robust training sets. In addition, concerted efforts have been made to create artificial datasets to train prediction models [171]. Alternatively, transitioning from a supervised ML system to unsupervised or semi-supervised methodologies may help to overcome the scarcity of available data. These methods are advantageous as they eliminate biases by removing the need for predefined labels like “pathogenic or benign” in training sets. One example of an unsupervised prediction tool is ParsSNP [173], which has outperformed existing tools in identifying driver mutations of cancer. However, specific application of unsupervised methods for synonymous variant prediction has not been adopted.

Importance of in vitro validation of in silico tool predictions in synonymous variant research

While computational tools for evaluating synonymous variants have improved significantly in recent years, in silico tools are still fundamentally imperfect systems. In many cases, predicted disease variants do not mirror the actual biological outcomes due to unknown biological complexities or deficiencies in the number of reliable and comprehensive genomic data sets. Therefore, it is increasingly important that in silico tool predictions be performed by multiple prediction tools with a variety of algorithms and parameters and validated through in vitro experiments. Currently, examples of experimentally corroborated synonymous variants are still quite low, which can be partially attributed to the necessity for more sensitive, standardized experimental assays. Detected protein or RNA alterations are usually significant, but small in magnitude. Many seminal works began as studies that leveraged the power of synonymous variant prediction tools to identify potential candidates and followed up these findings with experimental confirmation (see Table 4 for examples from highly cited studies that employed a combination of in silico and in vitro experiments to effectively investigate sSNV mechanisms). For a thorough review of experimental

Table 4 Examples of studies that effectively used prediction tools to study disease-causing synonymous variants

Disease association	Variant	Prediction tool	Description	Ref
Crohn's disease	***IRGM*** (c.313C > T)	SnipMir, RegRNA, and Patrocles (miRNA)	Synonymous variant predicted to delete a miRNA binding site, leading to increased risk for Crohn's disease (validated to be the causal mechanism through experiments assessing IRGM regulation)	[175]
Cystic fibrosis	***ΔF508 CFTR*** (c.1520_1522delTCT)	mFold (mRNA structure)	Synonymous site within the ΔF508 CFTR predicted to alter mRNA structure and stability and found to responsible for altered expression of the mutant protein	[18, 176]
Hemophilia B	***FIX*** [Factor IX] (c.459G > A)	mFold, Kinefold, NUPACK (mRNA structure), RSCU, CAI (codon usage indices)	mRNA structure prediction tools indicated a moderate reduction in mRNA stability, which coincided with diminished FIX expression through decreased translational speed	[77]
Hereditary cardiac arrhythmia	***hERG*** (codon-modified)	RNAfold (mRNA structure)	Codon modified hERG was predicted to have increased mRNA stability, resulting in altered translation of the ion channel	[19]
Pain sensitivity	***COMT*** (3 haplotypes with synonymous variations [c.198A > G, c.186C > T, c.408C > G])	mFold (mRNA structure)	COMT haplotype with predicted highest mRNA stability correlated with the lowest activity and expression levels. Other haplotypes with different thermodynamic stabilities elicited different pain sensitivities	[147, 177]
Phenylketonuria	***PAH*** (c.30C > G)	ESE Finder 3.0 (splicing)	An exonic splicing silencer was identified through splicing predictions and validated experimentally to be the main mechanism underlying the PKA-causing variant	[178]
Tuberculosis	***mabA*** (c.609G > A)	GENETYX-MAC (promoter prediction)	Synonymous variant predicted to cause the formation of an alternative promoter site, next to the mutation position, which was validated and found to increase transcription of inhA, leading to increased isoniazid resistance	[155, 179]

methods and discussion of studies that have investigated synonymous variants, we recommend reviewing Chapter 7 of a recently published book on *Single Nucleotide Polymorphisms* [174]. With the incessant rise in accumulations of genetic data and improving landscape of computational tools, the number of functional synonymous variants should dramatically increase over the next decade.

Concluding remarks and future perspectives

While overlooked in the past, synonymous variants are now recognized for their numerous functional effects and contribution to diseases. While this change in perspective was certainly precipitated by the rapid expansion of genetic testing and improvements in sequencing technologies, it must also be ascribed to recent significant advancements in bioinformatic AI and ML platforms. As highlighted in this review, in silico tools, especially those rooted in machine-learning algorithms, have been used to enhance our understanding of mechanisms underlying synonymous variants, while giving rise to additional inventive ideas, such as leveraging synonymous variants in genomic engineering strategies (e.g., codon optimization) to develop therapeutics [180]. In addition, the identification of recurrent disease mechanisms among synonymous variants, such as splicing or disrupted mRNA structure, has facilitated the discovery of new synonymous variants in other disease states, such as cancers [159]. The extended application of these technologies will be dependent on whether continued progress can be made in developing accurate synonymous variant computational predictors as these tools represent the most efficient means to process large-scale variant datasets. In the short term, the shortage of reliable genetic datasets on synonymous variants remains a significant obstacle for their rapid improvement, but as sequencing continues to become affordable and commonly used, this issue may be resolved naturally over time.

Thus, in the near future, promising improvements in these prediction tools may originate from enhanced understanding of codon, RNA, and sequence properties that correlate with functional synonymous variants. Future studies will need to address many outstanding questions in this field, including determining whether an array of sequence features can accurately discriminate functional or pathogenic synonymous variants. In addition, it will be important to develop refined models, specifically intended for synonymous variants, as many existing methods rely on adapting generic tools for synonymous variant assessment. This is suboptimal, as certain tools may place greater emphasis on particular variables and may not be able to sensitively detect functional variants. Fortunately, our understanding of biological relationships between codon usage, mRNA structure, and other protein sequence features continues to improve, and once intractable questions, such as how synonymous variants can alter the specific activity of proteins, have now been described [181]. The incorporation of these new variables into the design of in silico tools and the expanding use of these tools by the broad research community will only help to expedite novel discoveries in synonymous variant research.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02966-1.

Additional file 1. Review history.

Acknowledgements
This work was supported by funds from the Center for Biologics Evaluation and Research, US Food and Drug Administration.

Peer review information
Anahita Bishop and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 1.

Authors' contributions
All authors conceived and contributed to the writing, review, and approval of the manuscript.

Declarations

Ethics approval and consent to participate
Ethical approval was not required for this study.

Competing interests
The authors declare that they have no competing interests.

Received: 3 August 2022 Accepted: 10 May 2023
Published online: 22 May 2023

References

1. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci. 2010;107:961–8.
2. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337:64–9.
3. Shastry BS. SNP alleles in human disease and evolution. J Hum Genet. 2002;47:561–6.
4. Bailey SF, Hinz A, Kassen R. Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. Nat Commun. 2014;5:4076.
5. Cuevas JM, Domingo-Calap P, Sanjuán R. The fitness effects of synonymous mutations in DNA and RNA viruses. Mol Biol Evol. 2012;29:17–20.
6. Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. The distribution of fitness effects among synonymous mutations in a gene under directional selection. Elife. 2019;8:e45952. https://doi.org/10.7554/eLife.45952.
7. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. Exposing synonymous mutations. Trends Genet. 2014;30:308–21.
8. Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, Backofen R, Diederichs S. A pan-cancer analysis of synonymous mutations. Nat Commun. 2019;10:2569.
9. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011;12:683–91.
10. Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013;41:2073–94.
11. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet. 2003;12:205–16.
12. Katneni UK, Liss A, Holcomb D, Katagiri NH, Hunt R, Bar H, Ismail A, Komar AA, Kimchi-Sarfaty C. Splicing dysregulation contributes to the pathogenicity of several F9 exonic point variants. Mol Genet Genomic Med. 2019;7:e840.
13. Savisaar R, Hurst LD. Exonic splice regulation imposes strong selection at synonymous sites. Genome Res. 2018;28:1442–54.
14. Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier JF, Hébuterne X, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. 2011;43:242–5.
15. Wang Y, Qiu C, Cui Q. A large-scale analysis of the relationship of synonymous SNPs changing microRNA regulation with functionality and disease. Int J Mol Sci. 2015;16:23545–55.
16. Hamasaki-Katagiri N, Lin BC, Simon J, Hunt RC, Schiller T, Russek-Cohen E, Komar AA, Bar H, Kimchi-Sarfaty C. The importance of mRNA structure in determining the pathogenicity of synonymous and non-synonymous mutations in haemophilia. Haemophilia. 2017;23:e8–17.
17. Holcomb D, Hamasaki-Katagiri N, Laurie K, Katneni U, Kames J, Alexaki A, Bar H, Kimchi-Sarfaty C. New approaches to predict the effect of co-occurring variants on protein characteristics. Am J Hum Genet. 2021;108:1502–11.
18. Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, Kappes J, Collawn JF, Bebok Z. A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. J Biol Chem. 2010;285:28741–8.
19. Bertalovitz AC, Badhey MLO, McDonald TV. Synonymous nucleotide modification of the KCNH2 gene affects both mRNA characteristics and translation of the encoded hERG ion channel. J Biol Chem. 2018;293:12120–36.
20. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science. 2007;315:525–8.

21. Zeng Z, Aptekmann AA, Bromberg Y. Decoding the effects of synonymous variants. Nucleic Acids Res. 2021;49:12673–91.
22. Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. Identification of deleterious synonymous variants in human genomes. Bioinformatics. 2013;29:1843–50.
23. Livingstone M, Folkman L, Yang Y, Zhang P, Mort M, Cooper DN, Liu Y, Stantic B, Zhou Y. Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. Hum Mutat. 2017;38:1336–47.
24. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
25. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.
26. Bandyopadhyay S, Ghosh D, Mitra R, Zhao Z. MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci Rep. 2015;5:8004.
27. Calonaci N, Jones A, Cuturello F, Sattler M, Bussi G. Machine learning a model for RNA structure prediction. NAR Genom Bioinform. 2020;2:lqaa090.
28. Zhao Q, Zhao Z, Fan X, Yuan Z, Mao Q, Yao Y. Review of machine learning methods for RNA secondary structure prediction. PLoS Comput Biol. 2021;17:e1009291.
29. Rodriguez A, Wright G, Emrich S, Clark PL. %MinMax: a versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. Protein Sci. 2018;27:356–62.
30. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
31. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN, Hamasaki-Katagiri N, Osipova E, Simonyan V, Bar H, et al. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. J Mol Biol. 2019;431:2434–41.
32. Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, Katneni U, Golikov A, Ibla JC, Bar H, Kimchi-Sarfaty C. TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. J Mol Biol. 2020;432:3369–78.
33. Meyer D, Kames J, Bar H, Komar AA, Alexaki A, Ibla J, Hunt RC, Santana-Quintero LV, Golikov A, DiCuccio M, Kimchi-Sarfaty C. Distinct signatures of codon and codon pair usage in 32 primary tumor types in the novel database CancerCoCoPUTs for cancer-specific codon usage. Genome Med. 2021;13:122.
34. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 1986;24:28–38.
35. Wright F. The 'effective number of codons' used in a gene. Gene. 1990;87:23–9.
36. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981;151:389–409.
37. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol. 2001;53:290–8.
38. Liu C, Yuan J, Zhang X, Jin S, Li F, Xiang J. tRNA copy number and codon usage in the sea cucumber genome provide insights into adaptive translation for saponin biosynthesis. Open Biol. 2021;11:210190.
39. Reis Md. Savva R, Wernisch L: Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004;32:5036–44.
40. Sabi R, Volvovitch Daniel R, Tuller T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. Bioinformatics. 2016;33:589–91.
41. Gutman GA, Hatfield GW. Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci. 1989;86:3699–703.
42. Irwin B, Heck JD, Hatfield GW. Codon pair utilization biases influence translational elongation step times (∗). J Biol Chem. 1995;270:22801–6.
43. Wang F-P, Li H. Codon-pair usage and genome evolution. Gene. 2009;433:8–15.
44. Tats A, Tenson T, Remm M. Preferred and avoided codon pairs in three domains of life. BMC Genomics. 2008;9:463.
45. Moura GR, Pinheiro M, Freitas A, Oliveira JL, Frommlet JC, Carreto L, Soares AR, Bezerra AR, Santos MAS. Species-specific codon context rules unveil non-neutrality effects of synonymous mutations. PLoS ONE. 2011;6:e26817.
46. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. Adjacent codons act in concert to modulate translation efficiency in yeast. Cell. 2016;166:679–90.
47. Friberg M, von Rohr P, Gonnet G. Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae. Yeast. 2004;21:1083–93.
48. Kunec D, Osterrieder N. Codon pair bias is a direct consequence of dinucleotide bias. Cell Rep. 2016;14:55–67.
49. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell. 2015;59:149–61.
50. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci. 2004;101:12588–91.
51. Liu Q. Mutational bias and translational selection shaping the codon usage pattern of tissue-specific genes in rice. PLoS ONE. 2012;7:e48295.
52. Dittmar KA, Goodenbour JM, Pan T. Tissue-specific differences in human transfer RNA expression. PLoS Genet. 2006;2:e221.
53. Subramanian K, Payne B, Feyertag F, Alvarez-Ponce D: The Codon Statistics Database: a database of codon usage bias. Mol Biol Evol. 2022;39(8):msac157. https://doi.org/10.1093/molbev/msac157.
54. Samatova E, Daberger J, Liutkute M, Rodnina MV. Translational control by ribosome pausing in bacteria: how a non-uniform pace of translation affects protein production and folding. Front Microbiol. 2021;11:619430. https://doi.org/10.3389/fmicb.2020.619430.

55. Faure G, Ogurtsov AY, Shabalina SA, Koonin EV. Role of mRNA structure in the control of protein folding. Nucleic Acids Res. 2016;44:10898–911.
56. Espah Borujeni A, Salis HM. Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. J Am Chem Soc. 2016;138:7016–23.
57. Faa V, Coiana A, Incani F, Costantino L, Cao A, Rosatelli MC. A synonymous mutation in the CFTR gene causes aberrant splicing in an italian patient affected by a mild form of cystic fibrosis. J Mol Diagnost: JMD. 2010;12:380–3.
58. Pagani F, Raponi M, Baralle FE. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci. 2005;102:6368–72.
59. Peeri M, Tuller T. High-resolution modeling of the selection on local mRNA folding strength in coding sequences across the tree of life. Genome Biol. 2020;21:63.
60. Babendure JR, Babendure JL, Ding J-H, Tsien RY. Control of mammalian translation by mRNA structure near caps. RNA (New York, NY). 2006;12:851–61.
61. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 2010;6:e1000664.
62. Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 2005;6:R75.
63. Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, Link K, Khatwani N, Reynders J, Moore MJ, McFadyen IJ. mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci U S A. 2019;116:24075–83.
64. Zuker M, Mathews DH, Turner DH: Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In RNA biochemistry and biotechnology. Springer; 1999: 11–43.
65. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9:133–48.
66. Ding J, Lee Y-T, Bhandari Y, Schwieters CD, Fan L, Yu P, Tarosov SG, Stagno JR, Ma B, Nussinov R, et al. Visualizing RNA conformational and architectural heterogeneity in solution. Nat Commun. 2023;14:714.
67. Kirsch R, Seemann SE, Ruzzo WL, Cohen SM, Stadler PF, Gorodkin J. Identification and characterization of novel conserved RNA structures in Drosophila. BMC Genomics. 2018;19:899.
68. Seemann SE, Mirza AH, Hansen C, Bang-Berthelsen CH, Garde C, Christensen-Dalsgaard M, Torarinsson E, Yao Z, Workman CT, Pociot F, et al. The identification and functional annotation of RNA structures conserved in vertebrates. Genome Res. 2017;27:1371–83.
69. Meyer IM, Miklos I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acids Res. 2005;33:6338–48.
70. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31:3406–15.
71. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol. 2008;453:3–31.
72. Salari R, Kimchi-Sarfaty C, Gottesman MM, Przytycka TM. Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. Nucleic Acids Res. 2013;41:44–53.
73. Xayaphoummine A, Bucher T, Isambert H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. Nucleic Acids Res. 2005;33:W605-610.
74. Proctor JR, Meyer IM. COFOLD: an RNA secondary structure prediction method that takes co-transcriptional folding into account. Nucleic Acids Res. 2013;41:e102.
75. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res. 2008;36:W70–4.
76. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA. NUPACK: analysis and design of nucleic acid systems. J Comput Chem. 2011;32:170–3.
77. Simhadri VL, Hamasaki-Katagiri N, Lin BC, Hunt R, Jha S, Tseng SC, Wu A, Bentley AA, Zichel R, Lu Q, et al. Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. J Med Genet. 2017;54:338–45.
78. Wayment-Steele HK, Kladwang W, Strom AI, Lee J, Treuille A, Becka A, Das R, Eterna P. RNA secondary structure packages evaluated and improved by high-throughput experiments. Nat Methods. 2022;19:1234–42.
79. Zhao Y, Wang J, Zeng C, Xiao Y. Evaluation of RNA secondary structure prediction for both base-pairing and topology. Biophysics Reports. 2018;4:123–32.
80. Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA. 2010;16:1870–80.
81. Sato K, Kato Y. Prediction of RNA secondary structure including pseudoknots for long sequences. Brief Bioinform. 2021;23(1):bbab395. https://doi.org/10.1093/bib/bbab395.
82. Jabbari H, Wark I, Montemagno C, Will S. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. Bioinformatics. 2018;34:3849–56.
83. Kimchi O, Cragnolini T, Brenner MP, Colwell LJ. A polymer physics framework for the entropy of arbitrary pseudoknots. Biophys J. 2019;117:520–32.
84. zu Siederdissen CH. Bernhart SH, Stadler PF, Hofacker IL: A folding algorithm for extended RNA secondary structures. Bioinformatics. 2011;27:i129–36.
85. Sloma MF, Mathews DH. Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs. PLoS Comput Biol. 2017;13:e1005827.
86. Wang L, Liu Y, Zhong X, Liu H, Lu C, Li C, Zhang H. DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. Front Genet. 2019;10:143.
87. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10:5407.

88. Singh J, Paliwal K, Zhang T, Singh J, Litfin T, Zhou Y. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. Bioinformatics. 2021;37(17):2589–600. https://doi.org/10.1093/bioinformatics/btab165.
89. Miao Z, Westhof E. RNA structure: advances and assessment of 3D structure prediction. Annu Rev Biophys. 2017;46:483–503.
90. Wayment-Steele HK, Kladwang W, Watkins AM, Kim DS, Tunguz B, Reade W, Demkin M, Romano J, Wellington-Oguri R, Nicol JJ, et al. Deep learning models for predicting RNA degradation via dual crowdsourcing. Nat Machine Intell. 2022;4:1174–84.
91. Rother M, Rother K, Puton T, Bujnicki JM. ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Res. 2011;39:4007–22.
92. Flores SC, Wan Y, Russell R, Altman RB: Predicting RNA structure by multiple template homology modeling. In Biocomputing 2010. World Scientific; 2010: 216–227.
93. Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW. Automated 3D structure composition for large RNAs. Nucleic Acids Res. 2012;40:e112.
94. Xu X, Zhao C, Chen SJ. VfoldLA: A web server for loop assembly-based prediction of putative 3D RNA structures. J Struct Biol. 2019;207:235–40.
95. Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, Rother KM, Bujnicki JM. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res. 2016;44:e63.
96. Krokhotin A, Houlihan K, Dokholyan NV. iFoldRNA v2: folding RNA with constraints. Bioinformatics. 2015;31:2891–3.
97. Magnus M, Miao Z. RNA 3D structure comparison using RNA-Puzzles toolkit. Methods Mol Biol. 2023;2586:263–85.
98. Yang T-H, Lin Y-C, Hsia M, Liao Z-Y. SSRTool: a web tool for evaluating RNA secondary structure predictions based on species-specific functional interpretability. Comput Struct Biotechnol J. 2022;20:2473–83.
99. Isambert H, Siggia ED. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. Proc Natl Acad Sci. 2000;97:6515–20.
100. Ward M, Datta A, Wise M, Mathews DH. Advanced multi-loop algorithms for RNA secondary structure prediction reveal that the simplest model is best. Nucleic Acids Res. 2017;45:8541–50.
101. Danaee P, Rouches M, Wiley M, Deng D, Huang L, Hendrix D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic Acids Res. 2018;46:5381–94.
102. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.
103. Herzel L, Ottoz DSM, Alpert T, Neugebauer KM. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat Rev Mol Cell Biol. 2017;18:637–50.
104. Roca X, Olson AJ, Rao AR, Enerly E, Kristensen VN, Børresen-Dale AL, Andresen BS, Krainer AR, Sachidanandam R. Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genome Res. 2008;18:77–87.
105. Riolo G, Cantara S, Ricci C. What's wrong in a jump? Prediction and validation of splice site variants. Methods Protoc. 2021;4:62.
106. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. J Mol Biol. 1997;268:78–94.
107. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11:377–94.
108. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 2001;29:1185–90.
109. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouzé P, Brunak S. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res. 1996;24:3439–52.
110. Dogan RI, Getoor L, Wilbur WJ, Mount SM. SplicePort—an interactive splice-site analysis tool. Nucleic Acids Res. 2007;35:W285–91.
111. Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG. Avsec ž, Gagneur J: MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20:48.
112. Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet. 2006;15:2490–508.
113. Raponi M, Kralovicova J, Copson E, Divina P, Eccles D, Johnson P, Baralle D, Vorechovsky I. Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. Hum Mutat. 2011;32:436–44.
114. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. Cell. 2004;119:831–45.
115. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 2011;21:1360–74.
116. Schwartz S, Hall E, Ast G. SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. Nucleic Acids Res. 2009;37:W189–92.
117. Riepe TV, Khan M, Roosing S, Cremers FPM. 't Hoen PAC: Benchmarking deep learning splice prediction tools using functional splice assays. Hum Mutat. 2021;42:799–810.
118. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 1987;15:7155–74.
119. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37:e67–e67.
120. Leman R, Gaildrat P, Le Gac G, Ka C, Fichou Y, Audrezet M-P, Caux-Moncoutier V, Caputo SM, Boutry-Kryza N, Léone M, et al. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. Nucleic Acids Res. 2018;46:7913–23.
121. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. J Comput Biol. 1997;4:311–23.
122. Wang M, Marín A. Characterization and prediction of alternative splice sites. Gene. 2006;366:219–27.

123. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176:535-548.e524.
124. Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. Bioinformatics. 2018;34:4180–8.
125. Albaradei S, Magana-Mora A, Thafar M, Uludag M, Bajic VB, Gojobori T, Essack M, Jankovic BR. Splice2Deep: an ensemble of deep convolutional neural networks for improved splice site prediction in genomic DNA. Gene. 2020;763:100035.
126. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002;297:1007–13.
127. Erkelenz S, Theiss S, Otte M, Widera M, Peter JO, Schaal H. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res. 2014;42:10681–97.
128. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42:13534–44.
129. Zhou X, Zhou W, Wang C, Wang L, Jin Y, Jia Z, Liu Z, Zheng B. A comprehensive analysis and splicing characterization of naturally occurring synonymous variants in the ATP7B gene. Front Genet. 2021;11:592611. https://doi.org/10.3389/fgene.2020.592611.
130. Zhang H, Chen C, Wu X, Lou C, Liang Q, Wu W, Wang X, Ding Q. Effects of 14 F9 synonymous codon variants on hemophilia B expression: alteration of splicing along with protein expression. Hum Mutat. 2022;43:928–39.
131. Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frébourg T, Tosi M, Martins A. Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. PLoS Genet. 2016;12:e1005756.
132. Moles-Fernández A, Duran-Lozano L, Montalban G, Bonache S, López-Perolio I, Menéndez M, Santamariña M, Behar R, Blanco A, Carrasco E, et al. Computational tools for splicing defect prediction in breast/ovarian cancer genes: how efficient are they at predicting RNA alterations? Front Genet. 2018;9:366. https://doi.org/10.3389/fgene.2018.00366.
133. Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, Bronner M, Buisson M, Coulet F, Gaildrat P, et al. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. Hum Mutat. 2012;33:1228–38.
134. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol. 2018;9:402. https://doi.org/10.3389/fendo.2018.00402.
135. Guo Z, Maki M, Ding R, Yang Y. zhang B, Xiong L: Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. Sci Rep. 2014;4:5150.
136. Moszyńska A, Gebert M, Collawn JF, Bartoszewski R. SNPs in microRNA target sites and their potential role in human disease. Open Biol. 2017;7:170019.
137. Landi D, Gemignani F, Landi S. Role of variations within microRNA-binding sites in cancer. Mutagenesis. 2012;27:205–10.
138. Bali V, Bebok Z. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. Int J Biochem Cell Biol. 2015;64:58–74.
139. Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N, et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proc Natl Acad Sci. 2013;110:13481–6.
140. Fang Z, Rajewsky N. The impact of miRNA target sites in coding sequences and in 3′UTRs. PLoS ONE. 2011;6:e18067.
141. Riffo-Campos ÁL, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: what to choose? Int J Mol Sci. 1987;2016:17.
142. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.
143. Lewis BP, Shih Ih, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian MicroRNA targets. Cell. 2003;115:787–98.
144. Schnall-Levin M, Zhao Y, Perrimon N, Berger B. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3′UTRs. Proc Natl Acad Sci. 2010;107:15751–6.
145. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res. 2014;43:D146–52.
146. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2019;48:D127–31.
147. Coronnello C, Benos PV. ComiR: combinatorial microRNA target prediction tool. Nucleic Acids Res. 2013;41:W159–64.
148. Bertolazzi G, Benos PV, Tumminello M, Coronnello C. An improvement of ComiR algorithm for microRNA target prediction by exploiting coding region sequences of mRNAs. BMC Bioinformatics. 2020;21:201.
149. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res. 2013;41:W169-173.
150. Marín RM, Sulc M, Vanícek J. Searching the coding region for microRNA targets. RNA. 2013;19:467–74.
151. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5:R1.
152. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol. 2004;2:e363.
153. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39:1278–84.
154. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. Genome Biol. 2019;20:18.

155. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 2010;11:R90.
156. Coronnello C, Hartmaier R, Arora A, Huleihel L, Pandit KV, Bais AS, Butterworth M, Kaminski N, Stormo GD, Oesterreich S, Benos PV. Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density. PLoS Comput Biol. 2012;8:e1002830.
157. Jankowska KI, Meyer D, Holcomb DDF, Kames J, Hamasaki-Katagiri N, Katneni UK, Hunt RC, Ibla JC, Kimchi-Sarfaty C. Synonymous ADAMTS13 variants impact molecular characteristics and contribute to variability in active protein abundance. Blood Adv. 2022;6(18):5364–78. https://doi.org/10.1182/bloodadvances.2022007065.
158. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2018;47:D155–62.
159. Kaissarian NM, Meyer D, Kimchi-Sarfaty C. Synonymous variants: necessary nuance in our understanding of cancer drivers and treatment outcomes. JNCI: J Natl Cancer Institute. 2022;114(8):1072–94. https://doi.org/10.1093/jnci/djac090.
160. Kimchi-Sarfaty C, Simhadri VL, Kopelman D, Friedman A, Edwards N, Javaid A, Okunji C, Komar A, Sauna Z, Katagiri N. The synonymous V107V mutation in factor IX is not so silent and may cause hemophilia B in patients. Blood. 2010;116:2197–2197.
161. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014;156:1324–35.
162. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99:877–85.
163. Shi F, Yao Y, Bin Y, Zheng CH, Xia J. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. BMC Med Genomics. 2019;12:12.
164. Cooper DN, Ball EV, Krawczak M. The human gene mutation database. Nucleic Acids Res. 1998;26:285–7.
165. Schaafsma GC, Vihinen M. V ari SNP, a benchmark database for variations from db SNP. Hum Mutat. 2015;36:161–6.
166. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
167. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
168. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. Hum Mutat. 2018;39:1623–30.
169. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10:1556–66.
170. Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, Schoch K, Ratzon F, Heinzen EL, Boland MJ, et al. Annotating pathogenic non-coding variants in genic regions. Nat Commun. 2017;8:236–236.
171. Zeng Z, Bromberg Y. Predicting functional effects of synonymous variants: a systematic review and perspectives. Front Genet. 2019;10:914. https://doi.org/10.3389/fgene.2019.00914.
172. Tang X, Zhang T, Cheng N, Wang H, Zheng C-H, Xia J, Zhang T. usDSM: a novel method for deleterious synonymous mutation prediction using undersampling scheme. Brief Bioinform. 2021;22(5):bbab123. https://doi.org/10.1093/bib/bbab123.
173. Kumar RD, Swamidass SJ, Bose R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. Nat Genet. 2016;48:1288–94.
174. Lin BC, Jankowska KI, Meyer D, Katneni UK. Methods to evaluate the effects of synonymous variants. In: Sauna ZE, Kimchi-Sarfaty C, editors. Single Nucleotide Polymorphisms: Human Variation and a Coming Revolution in Biology and Medicine. Cham: Springer International Publishing; 2022. p. 133–68.
175. Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics. 2009;25:2625–31.
176. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N. Combinatorial microRNA target predictions. Nat Genet. 2005;37:495–500.
177. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science. 2006;314:1930–3.
178. Dobrowolski SF, Andersen HS, Doktor TK, Andresen BS. The phenylalanine hydroxylase c.30C>G synonymous variation (p.G10G) creates a common exonic splicing silencer. Mol Genet Metab. 2010;100:316–23.
179. Ando H, Miyoshi-Akiyama T, Watanabe S, Kirikae T. A silent mutation in mabA confers isoniazid resistance on Mycobacterium tuberculosis. Mol Microbiol. 2014;91:538–47.
180. Lin BC, Kaissarian NM, Kimchi-Sarfaty C. Implementing computational methods in tandem with synonymous gene recoding for therapeutic development. Trends Pharmacol Sci. 2022;44(2):73–84. https://doi.org/10.1016/j.tips.2022.09.008.
181. Jiang Y, Neti SS, Sitarik I, Pradhan P, To P, Xia Y, Fried SD, Booker SJ, O'Brien EP. How synonymous mutations alter enzyme structure and function over long timescales. Nat Chem. 2022;15:308–18. https://doi.org/10.1038/s41557-022-01091-z.

Publisher's Note