Investigating DNA, RNA and protein-based features as a means to discriminate pathogenic synonymous variants

Mark Livingstone[1, #], Lukas Folkman[1, #], Yuedong Yang[1, 2], Ping Zhang[3], Matthew Mort[4], David N. Cooper[4], Yunlong Liu[5], Bela Stantic[1], Yaoqi Zhou[1, 2, *]

[1]School of Information and Communication Technology, Griffith University, Southport, 4222 Queensland, Australia

[2]Institute for Glycomics, Griffith University, Southport, 4222 Queensland, Australia

[3]Menzies Health Institute, Griffith University, Southport, 4222 Queensland, Australia

[4]Institute of Medical Genetics, Cardiff University, Cardiff CF144XN, United Kingdom

[5]Department of Medical and Molecular Genetics, Indiana University, Indianapolis, 46202 Indiana, USA

# These authors contributed equally.

* Corresponding author, phone: +61 7 555 28349, fax: + 61 7 555 28066, e-mail: yaoqi.zhou@griffith.edu.au

Grant sponsor: National Health and Medical Research Council of Australia (grant numbers 1059775 and 1083450)

ABSTRACT

Synonymous single nucleotide variants (SNVs), although they do not alter the encoded protein sequences, have been implicated in many genetic diseases. Experimental studies indicate that synonymous SNVs can lead to changes in the secondary and tertiary structures of DNA and RNA, thereby impacting translational efficiency, co-translational protein folding as well as the binding of DNA/RNA-binding proteins. However, the importance of these various features in disease phenotypes is not clearly understood. Here we have built a support vector machine model (termed DDIG-SN) as a means to discriminate disease-causing synonymous variants. The model was trained and evaluated on nearly 900 disease-causing variants. The method achieves robust performance with the area under the receiver operating characteristic curve (AUC) of 0.84 and 0.85 for protein-stratified 10-fold cross-

This is the author's manuscript of the article published in final edited form as:
Livingstone, M., Folkman, L., Yang, Y., Zhang, P., Mort, M., Cooper, D. N., Liu, Y., Stantic, B. and Zhou, Y. (2017), Investigating DNA, RNA and protein-based features as a means to discriminate pathogenic synonymous variants. Human Mutation. Accepted Author Manuscript. http://dx.doi.org/10.1002/humu.23283

validation and independent testing, respectively. We were able to show that the disease-causing effects in the immediate proximity to exon-intron junctions (1−3 bp) are driven by the loss of splicing motif strength, whereas the gain of splicing motif strength is the primary cause in regions further away from the splice site (4−69 bp). The method is available as a part of the DDIG server at http://sparks-lab.org/ddig.

KEY WORDS

synonymous SNV, same-sense variant, silent mutation, bioinformatics, machine learning

INTRODUCTION

Synonymous (SN) variants (also called silent or same-sense mutations) are single nucleotide variants (SNVs) which result in synonymous codon substitutions (a codon encoding a specific amino acid residue is replaced by another codon encoding the same amino acid). Despite not altering the translated protein product, more and more studies have suggested that both germline and somatic SN variants may be deleterious and can lead to a number of genetic diseases (Hunt, et al., 2014; Niroula and Vihinen, 2016; Sauna and Kimchi-Sarfaty, 2011; Shabalina, et al., 2013) including cancer (Supek, et al., 2014), autism spectrum disorders (Neale, et al., 2012; Samocha, et al., 2014), asthma, and osteoporosis (Macaya, et al., 2009), as well as increasing disease susceptibility to idiopathic dilated cardiomyopathy (Stark, et al., 2010), altering disease outcome in paediatric acute myeloid leukaemia (Ho, et al., 2011), and creating individualised responses to drugs by affecting the function of drug transporters (Kimchi-Sarfaty, et al., 2007). The number of known disease-causing SN variants has been steadily increasing in recent years, as is evidenced by a 16% increase in the number of pathogenic SN variants recorded in the Human Gene Mutation Database (HGMD) (Stenson, et al., 2017) over the last 15 months (from 1,181 to 1,368 in the HGMD versions 2015.3 and 2016.4, respectively).

SN variants alter the sequences of DNA and their corresponding messenger RNA (mRNA) transcripts. One direct consequence of the sequence change is the altered codon. Codon usage (termed 'codon bias') has long been known to affect gene expression, translation velocity, and folding efficiency (Hershberg and Petrov, 2008; Plotkin and Kudla, 2011) although whether or not such codon bias in mammals is caused by evolutionary selection or large variation in GC content (isochores) remains an active subject of debate (Kirchner and Ignatova, 2015; Rudolph, et al., 2016). Sequence change could also modify cytosine-phosphate-guanine (CpG) sites thereby influencing local chromatin structure and DNA methylation patterns (Deaton and Bird, 2011). Moreover, coding regions have been found to directly interact with microRNA (Bentwich, et al., 2005; Brest, et al., 2011; Gartner, et al., 2013; Hurst, 2006) as well as with DNA and RNA-binding proteins (Dreyfuss, et al., 2002; Stergachis, et al., 2013). Altering these interactions through sequence variation will affect the regulatory control and outcome of transcription and translation. For example, exonic splicing elements (enhancers and silencers) regulate splicing by interacting with SR-proteins and heterogeneous ribonucleoprotein particles, respectively (Zhu, et al., 2001). At least 4% of SN variants have been shown to be deleterious to splicing enhancers (Cáceres and Hurst, 2013; Carlini and Genut, 2006; Fairbrother, et al., 2004; Parmley, et al., 2006; Savisaar and Hurst, 2017; Wu and Hurst, 2016). Variant-induced changes to mRNA secondary structure have been found to alter mRNA stability (Chamary and Hurst, 2005; Duan, et al., 2013), hamper protein expression [at least in bacteria (Kudla, et al., 2009)], and affect mRNA splicing (Buratti and Baralle, 2004), whereas analysis of the solvent-accessible surface area of RNA indicates that rare alleles are more likely to occur in the buried, structured regions of coding RNA (Yang, et al., 2017). In addition, changing the nucleotide sequence may lead to the formation or disruption of a square-planar structure formed by guanine (G-quadruplex), which has been implicated in both positive and negative transcriptional regulation (Rhodes and Lipps, 2015; Simone, et al., 2015).

Thus, it is clear that an SN variant can lead to changes in the structure and function of both the DNA sequence and the mRNA transcript. Consequently, this disrupts regulatory controls and affects the

structure and function of the final protein product. Although mRNA splicing accounts for a large proportion (20–40%) of disease-associated SNVs (Wu and Hurst, 2016), what is not yet clear is the role that other factors might play in mediating the pathogenic influence of disease-causing SN variants. SilVA (Buske, et al., 2013), the first method dedicated to the discrimination of disease-causing SN variants, employed a range of features including evolutionary conservation at the DNA level, codon usage, CpG sites, splicing site motifs, and mRNA folding energy. However, it was trained and tested using a dataset of only 41 disease-causing SN variants. Alternatively, methods for the prediction of defects in RNA splicing, including SPANR or SPIDEX (a pre-computed index of SPANR scores for the human genome) (Xiong, et al., 2015), MutPred Splice (Mort, et al., 2014) and ExonImpact (Li, et al., 2017), can also be used to study SN variants. Other techniques, such as CADD (Kircher, et al., 2014), MutationTaster (Schwarz, et al., 2014), and FATHMM-MKL (Shihab, et al., 2015), provide general frameworks for predicting all types of pathogenic genetic variation (missense, nonsense, synonymous, microinsertions and microdeletions – 'indels') in both coding and non-coding regions. However, owing to their generality, it is difficult to dissect the relative importance of various features specifically for SN variants.

In this work, we investigated the ability of DNA, RNA and protein-based features to discriminate pathogenic from putatively benign SN variants, with or without filtering variants with low minor allele frequencies. We employed a non-redundant training dataset of 318 genes with at least one disease-causing and one putatively benign variant per gene from the HGMD (Stenson, et al., 2017) and 1000 Genomes Projects (1000 Genomes Project Consortium, 2015), respectively. Analysis at a single feature level as well as cross-validation and independent testing using a support vector machine model indicated that the impact on RNA splicing is the dominant factor for disease-causing SN variants, even if evolutionary conservation had been excluded from feature selection.

MATERIALS AND METHODS

Datasets

We compiled two SN variants datasets: one for design, cross-validation, and training of our method and one for independent testing. The putatively benign variants were derived from the 1000 Genomes Project (1kGP), phase 3, version 5b, 20130502 (1000 Genomes Project Consortium, 2015). The disease-causing variants were retrieved from the Human Gene Mutation Database (HGMD) Professional, version 2015.3 (Stenson, et al., 2017), utilising only the variants labelled as disease-causing ('DM' and 'DM?' labels). As a reference transcript set, we used the Consensus CoDing Sequence (CCDS) project, version 20131129 (Pruitt, et al., 2009).

We first compiled the disease-causing datasets in a mutually independent fashion. To this end, we split the HGMD variants with a ratio of 2:1 for training and testing. We ensured that the split was protein-stratified with a sequence identity threshold of 30%. That is, we ensured low sequence similarity between the genes from the training and test sets. Finally, for every gene, we added the putatively benign 1kGP variants. If there were no benign variants for a given gene, we discarded all disease-causing variants from this gene. Ensuring that every gene had at least one disease-causing and one benign variant has been suggested in order to allow correct learning of variant-specific features rather than gene-related properties (Grimm, et al., 2015).

In a comparison with related work, we found that one of the methods (SilVA) yielded no predictions for two variants (located on the Y chromosome) from our design/training dataset. To facilitate a fair comparison of related work, we removed these two variants from our dataset. Another method (SPIDEX) yielded 2,604 missing predictions for our datasets, mostly because of its limitation to predict splicing effects for variants located only up to 300 nucleotides from splice junctions. We decided to *retain* these variants in our datasets after checking that the overall ranking of the compared methods was not affected.

As a result of the described procedure, we obtained two mutually independent and protein-stratified datasets. The design/training dataset containing 318 genes with 592 disease-causing and 10,925 putatively benign SN variants. This dataset was utilised for design, training, and 10-fold cross-validation of our method. The test set comprised 143 genes with 279 disease-causing and 4,945 benign variants. Finally, we also compiled fully balanced subsets of these datasets where for every disease-causing variant, one benign variant was selected so that the genomic distance between the two was as small as possible. We refer to these datasets as to 'close-by' datasets since the subsampling procedure ensured that every selected benign variant was located as close as possible to some disease-causing variant. Table 1 summarises the four different datasets.

Predictive features

We investigated 54 features for discriminating disease-causing and benign SN variants at three different levels: DNA, RNA, and protein levels. Out of these 54 features, 26 features were also employed for developing SilVA (Buske, et al., 2013). Supp. Table S1 briefly summarises all predictive features.

DNA-based features

We derived six new DNA-based features. The feature $\text{phyloP}_{\text{cons46way}}$ was derived from phylogenetic *p*-values of the multiple sequence alignment of 45 vertebrates to the human genome (cons46way) as calculated with the phyloP program (Pollard, et al., 2010). Similarly, using an alignment of 99 vertebrates to the human genome (cons100way) (Karolchik, et al., 2004; Miller, et al., 2007), we calculated the allele frequency (AF) of the reference allele ($\text{rAF}_{\text{cons100way}}$) and the difference in the frequencies of the alternative and reference alleles ($\Delta\text{AF}_{\text{cons100way}}$). Next, we used DeepBind (Alipanahi, et al., 2015) to predict variant-induced binding affinity changes of 515 DNA-binding proteins. To this end, DeepBind was trained using 137 ChIP-seq and 378 SELEX human models

downloaded from the DeepBind website. From DeepBind predictions, we derived two features calculated as the maximum ($DeepBind_{max}$) and mean ($DeepBind_{mean}$) of the 515 binding affinity changes. Another DNA-based feature (referred to as G4) tests whether the variant lies within a G-quadruplex sequence pattern (Todd, et al., 2005). In addition, we employed another five features from SilVA including DNA conservation score $GERP{+}{+}_{cons34}$ (Davydov, et al., 2010), relative synonymous codon usage (RSCU), variant-induced change in codon usage ($|\Delta RSCU|$) (Sharp and Li, 1987), variant-induced change in CpG (CpG?), and the ratio of the observed and expected CpG content of the exon (CpG_{exon}).

RNA-based features

We implemented 11 new RNA-based features. We predicted RNA solvent-accessible surface area (ASA_{RNA}) of the variant site using RNAsnap trained on experimentally determined protein-RNA complex structures (Yang, et al., 2017). We predicted the change between the binding affinities (ΔRBP_{aff}) of the reference and alternative sequences for a set of 53 RNA-binding proteins (RBPs) (Zhang, et al., 2014). Another feature, $SPIDEX_{\Delta\Psi}$, was extracted from the SPIDEX database (Xiong, et al., 2015) and is based on the maximum $\Delta\Psi$ scores across all predicted tissues. $\Delta\Psi$ expresses the difference between the predicted exon-inclusion probabilities of the reference and alternative sequences. As suggested in (Wu and Hurst, 2016), we also examined the location of the exon within the given gene, implemented as the relative exon number feature (the exon number in the 5'-3' direction divided by the number of exons), and the distance to the nearest 5', 3' or closest splice site (exon-intron junction). Further, we defined a feature termed fraction of unaffected transcripts where an unaffected transcript (splice isoform) is one for which the variant is located within an intron (whereas the variant is located in an exon for some other splice isoforms of the same gene). Finally, we implemented three features based on the consensus dataset of 84 exonic splicing enhancer (ESE) motifs, INT3 (Cáceres and Hurst, 2013). These three binary features were the loss (ESE_{loss}), gain (ESE_{gain}), and loss or gain ($ESE_{loss/gain}$) of an INT3 ESE motif due to the introduced variant.

In addition to the features listed above, we implemented 21 features utilized in SilVA. These features include the changes in the secondary structure folding energy of pre-mRNA and mature mRNA calculated with UNAFold (Markham and Zuker, 2008), changes in the diversity of the structural ensemble of pre-mRNA and mature mRNA obtained using ViennaRNA (Lorenz, et al., 2011), relative distances to the end of pre-mRNA and mature mRNA, various features pertaining to the splice site motif strength calculated with MaxEntScan (Yeo and Burge, 2004), and gain and loss of ESE and exonic splicing silencer (ESS) motifs based on ESE Finder (Smith, et al., 2006), PESX (Zhang and Chasin, 2004), and FAS-hex3 (Wang, et al., 2004) datasets.

Protein-based features

We implemented 11 protein-based features that were found to be useful for predicting disease-causing indels (Folkman, et al., 2015; Zhao, et al., 2013) and stability changes induced by single amino acid substitutions (Folkman, et al., 2016). Among those, three were related to protein conservation, five to protein structure, and three to global sequence properties. We examined protein conservation scores calculated as Jensen-Shannon divergence (JS_{cons}) using the available implementation (Capra and Singh, 2007). We calculated these conservation scores from the multiple sequence alignment (MSA) generated with PSI-BLAST (NCBI non-redundant database, three iterations, e-value threshold 0.001) (Altschul, et al., 1997). Another protein conservation feature, $HHblits_{cons}$, was calculated using HHblits (Remmert, et al., 2012), which searches for similar sequences from the UniProt database (UniProt Consortium, 2015) using hidden Markov model sequence profiles. In addition, we considered the $HHblits_{neff}$ feature, which expresses the number of sequences aligned to the variant site in the HHblits' MSA. Other protein-based features described the relative variant location as a distance to the N-terminal ($N\text{-}term_{dist}$), C-terminal ($C\text{-}term_{dist}$), and the centre of the sequence ($centre_{dist}$), divided by the length of the protein sequence. Finally, we considered five protein-level structural features predicted from the protein sequence: relative accessible surface area and helix, sheet, coil and disorder probabilities. The accessible surface area and secondary structure probabilities were

predicted with SPIDER2 (Heffernan, et al., 2015), and the disorder probability was calculated using SPINE-D (Zhang, et al., 2012).

Single feature analysis

To analyse the potency of a single feature to discriminate disease-causing SN variants, we randomly sub-sampled the design/training dataset into 100 balanced samples so that there were exactly the same number of disease-causing and putatively benign variants for each gene (2×591 variants in total). Then, we performed protein-stratified 10-fold cross-validation using a single feature. This procedure ensured that the single feature analysis would not be adversely affected by gene-level global properties.

Support vector machines

We implemented our method with the support vector machine (SVM) algorithm (Cortes and Vapnik, 1995), which classify examples (variants) as positive (disease-causing) or negative (putatively benign) based on the optimal separating hyperline, which is learned from the training data. We used the radial basis function (RBF) kernel to perform a non-linear transformation of the feature space. To set optimal values of hyper-parameters (namely, the regularisation parameter C and RBF width parameter γ), we performed a grid search which spanned across all combinations of $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{11}\}$ and $\gamma \in \{2^{-11}, 2^{-9}, \ldots, 2^{-1}\}$. Because our training dataset was extremely unbalanced with a ratio of 1:18, we set the SVM weight penalty (w) for misclassifying a positive (disease-causing) example to $w = 18$.

Feature selection

We used the stability selection algorithm (Meinshausen and Buhlmann, 2010) to select a robust combination of predictive features. In our implementation, we exploited the unbalanced nature of our

training dataset. We randomly sub-sampled the dataset 100-times so that there were exactly the same number of disease-causing and putatively benign variants per gene (2×591 variants in total). Next, we applied the sequential forward selection (SFS) algorithm (Whitney, 1971) to select a feature set for each of the data samples. Finally, the stability selection creates a stable set of features by selecting those features which SFS repeatedly selected (≥ 25 times) across the 100 samples.

The SFS proceeds as follows. For each sample, the SFS starts with an empty set of features S_0 and iteratively selects a new feature f such that $S_i = S_{i-1} \cup \{f\}$ yields the best prediction performance. We adopted the area under the receiver operating characteristic curve (AUC, see the next section for definition) to assess prediction performance during feature selection. We let the SFS run until the AUC improvement yielded by increasing the number of features was < 0.005 (so called early-stop criterion which helps to avoid overtraining).

Evaluation

We employed protein-stratified 10-fold cross-validation to design our method and approximate the prediction performance on the design/training dataset. The 10-fold cross-validation procedure works by dividing the dataset into ten roughly equally-sized folds. Then, nine folds are merged and used for training, whilst the remaining fold is used for testing. This is repeated ten times, each time with a different test fold. Finally, the ten predictions are pooled together. In a protein-stratified cross-validation, it is ensured that the sequence identity of any two proteins from two distinct folds is $< 30\%$. This in turn ensures that a test fold will never comprise a protein similar to the proteins from the folds used for training.

We adopted the receiver operating characteristic (ROC) curve as the primary evaluation measure in this work. From the ROC curve, we calculated the area under the curve (AUC). A ROC curve plots the true positive rate (sensitivity) as a function of the false positive rate (1−specificity) at different

prediction thresholds. The AUC value was used to select the best performing features during feature selection, optimise the SVM hyper-parameters C and γ, as well as for a comparison with related work. We used DeLong's test to compare AUC scores and calculate statistically sound *p*-values (DeLong, et al., 1988).

For the sake of completeness (Vihinen, 2013), we also evaluated prediction performance in terms of Matthews correlation coefficient (MCC), classification accuracy (Q_2), sensitivity (Se, also referred to as recall), specificity (Sp), positive predictive value (PPV, also referred to as precision), and negative predictive value (NPV). Because these metrics, unlike AUC, are threshold-dependent, we set the prediction threshold for each method so that the MCC of the given method was maximised. The definitions of these metrics can be found in the Supporting Material.

RESULTS

Discriminating using a single feature

Our design/training dataset contained 318 unique genes with 592 disease-causing (HGMD) and 10,925 putatively benign (1kGP) variants. Such a highly unbalanced dataset simply reflects the real-world situation where a few possible disease-causing variants have to be prioritised for experimental validation among many neutral variants. The 318 genes were carefully selected so that the dataset comprised at least one disease-causing and one benign variant per gene. Having positive and negative variants from the same gene is necessary to avoid the potential bias for specific genes (Grimm, et al., 2015). This in turn allows us to search for features and develop methods for discriminating disease-causing variants rather than disease-associated genes.

Using this design/training dataset, we evaluated 54 diverse features at DNA, RNA and protein levels as shown in Supp. Table S1. The utility of each feature for SN variant discrimination was measured

using the AUC (area under the ROC curve). The AUC is 1 for perfect discrimination and 0.5 for random discrimination. As shown in Supp. Table S1, many of the investigated features had a very weak, statistically non-significant, discrimination capability with $0.50 \leq \text{AUC} < 0.55$ [DeLong's test (DeLong, et al., 1988), significance level 0.05, power 0.80]. These features were DNA structural features [G-quadruplex pattern prediction (Todd, et al., 2005)], DNA functional features [variant-induced changes to protein-DNA binding affinities by DeepBind (Alipanahi, et al., 2015)], RNA structural features [RNA secondary structure folding by UNAFold (Markham and Zuker, 2008), structural ensemble diversity by ViennaRNA (Lorenz, et al., 2011), and predicted RNA solvent accessibility by RNAsnap (Yang, et al., 2017)], RNA functional features [variant-induced changes to protein-RNA binding affinities (Zhang, et al., 2014)], protein conservation features [protein sequence conservation by PSI-BLAST (Altschul, et al., 1997) and HHblits (Remmert, et al., 2012)], and protein structural features [protein intrinsic disorder by SPINE-D (Zhang, et al., 2012), and protein secondary structure and solvent accessibility by SPIDER2 (Heffernan, et al., 2015)]. Given that SN variants do not alter protein sequences, the lack of strong discriminatory power was expected for the protein-based features but not for some of the DNA and RNA-based features.

Table 2 lists the top ten predictive features ranked by their AUC values. Of the top ten, six are related to splicing, including various encodings of splice site motif strength (MES, |ΔMES|, and MES-KM) calculated with MaxEntScan (Yeo and Burge, 2004), the difference in the predicted exon-inclusion probabilities of the reference and alternative sequences using SPIDEX (Xiong, et al., 2015), and distances from the 3' and nearest splicing sites (exon-intron junctions). The other four features are related to DNA conservation calculated from different multiple-species genomes alignments (Miller, et al., 2007): the AF of the reference allele ($\text{rAF}_{\text{cons100}}$), difference in the AFs of the variant and that of the reference ($\Delta\text{AF}_{\text{cons100}}$), DNA conservation score ($\text{phyloP}_{\text{cons46}}$) derived from phylogenetic *p*-values calculated with

the phyloP program (Pollard, et al., 2010), and evolutionary rates estimated with a maximum likelihood algorithm (GERP++$_{cons34}$) (Davydov, et al., 2010).

Figures 1A and 1B depict the distributions of the disease-causing and benign variants for the best RNA-based (MES) and DNA-based ($\Delta AF_{cons100}$) features, respectively. Disease-causing variants occurred more frequently when the predicted splice site motif strength (MES) was high, implying that the variant falls within a splicing motif with a high probability. Regarding $\Delta AF_{cons100}$, disease-causing variants were characterised by negative values (close to –1), denoting a high AF for the reference allele and a low AF for the variant. Finally, Figure 1C shows the distributions of disease-causing and benign variants in the immediate proximity of the exon-intron junction (1–3 bp), in the larger region commonly considered to be implicated in RNA splicing (4–69 bp), and in the exon core (≥70 bp from the exon-intron junction). The clear separation shown in this plot explains the high predictive power of naïve features such as the distance to the closest exon-intron junction, which ranked as the third best RNA-based feature with the AUC of 0.705 (Table 2). The results summarised in Table 2 and Figure 1 demonstrate that features related to RNA splicing and DNA conservation provide the most discriminative information about the nature of the disease-causing SN variants in our dataset.

Combining multiple features with feature selection

Most strongly discriminative single features were related to either DNA conservation or RNA splicing. It is of interest to know if the weakly discriminative features (when evaluated as single features) can potentially complement the strongly discriminative features when combined into a predictive non-linear SVM model (Cortes and Vapnik, 1995). We employed the stability selection algorithm (Meinshausen and Buhlmann, 2010) to select the most relevant and stable feature combination, avoiding the redundant and most-correlated features. To this end, we created 100 samples from our design/training dataset so that each sample had an equal number of disease-causing and benign variants (thereby ensuring that variant-specific, rather than gene-specific, features would be selected) and performed sequential forward selection (SFS) (Whitney, 1971) on each of the 100

samples. The *stability* metric of a feature was then measured as the fraction of times the feature was selected amongst the 100 feature combinations.

The feature selection resulted in a combination of six features with *stability* ≥ 0.25, yielding an AUC of 0.84 based on protein-stratified 10-fold cross-validation. For brevity, we refer to this model as DDIG-SN, which stands for 'Detecting DIsease-causing Genetic variations with a SyNonymous model'. Figure 2 shows how the performance varied across the 100 samples depending upon the number of features selected using the SFS algorithm. The plot demonstrates that the prediction performance improvements became smaller as more features were added; the median AUC improvement was < 0.005 after six features had been combined. Thus, the saturation of the number of features for the internal SFS algorithm was in agreement with the number of features with *stability* ≥ 0.25, suggesting that a combination of six features was indeed optimal.

Figure 3 shows the most stable ten features. The top three features, $\text{phyloP}_{\text{cons46}}$, $|\Delta\text{MES}|$, and $\text{SPIDEX}_{\Delta\Psi}$, were selected 99, 87, and 73 out of 100 times, respectively. The figure also indicates that five of the six most 'stable' features also yielded an AUC around 0.7 in the single feature analysis. That is, they are among the top nine features ranked in Table 2: MES, $|\Delta\text{MES}|$, $\text{phyloP}_{\text{cons46}}$, $\text{rAF}_{\text{cons100}}$, and $\text{SPIDEX}_{\Delta\Psi}$. The remaining feature with *stability* > 0.5 but with low AUC of 0.52 was MES-CS, which signifies whether the variant causes a cryptic splice site to become a site with the strongest splicing motif according to MaxEntScan. Thus, while this feature is discriminative only for a subset of all variants (hence the low overall AUC), the feature selection was able to identify its contribution when combined with other predictive features (it was selected 42 out of 100 times). In summary, all selected features were related to either DNA conservation or RNA splicing, in agreement with our single feature analysis.

Comparison with other methods and independent test performance

The ROC curves in Figure 4 show the performance comparison of DDIG-SN with other available methods: a synonymous-specific model – SilVA (Buske, et al., 2013), conservation score – phyloP (Pollard, et al., 2010), general approaches for all types of SNVs – CADD (Kircher, et al., 2014), MutationTaster (Schwarz, et al., 2014), and FATHMM-MKL (Shihab, et al., 2015), and an approach to predict how variants affect splicing – SPIDEX (Xiong, et al., 2015).

To further confirm the generality of the DDIG-SN model, we compiled an independent test dataset. This test set comprised a hold-out portion of the 279 disease-causing (HGMD) and 4,945 putatively benign (1kGP) variants located in 143 genes whose protein sequence similarity was < 30% compared to any protein in our design/training dataset. On this independent test set, DDIG-SN achieved the highest AUC of 0.85 (Table 3, Figure 4B), which represents a relative improvement of 2% (p = 0.229, DeLong's test) over the second best method, SilVA, which yielded the AUC of 0.83. Importantly, this independent test performance was close to the DDIG-SN's cross-validation performance (Figure 4A, AUC of 0.84, a 4% relative improvement as compared to SilVA, $p = 5.7\times10^{-5}$) on the dataset used to design our method and optimise all parameters. Hence, DDIG-SN performed robustly and this evidence suggests that overtraining was avoided. The AUC scores of the other five methods were in the range of 0.74–0.57 and 0.76–0.59 for the design/training and test datasets, respectively. Supp. Table S2 lists also the Q_2, Se, Sp, PPV, and NPV scores yielded by all compared methods for both design/training and test datasets.

Both the design/training and test datasets were highly imbalanced with many more benign than disease-causing variants per gene (the overall ratio of 18:1). We were interested in whether DDIG-SN could retain its performance when tested on a fully balanced, more challenging, dataset where for every disease-causing variant, one benign variant was selected so that the genomic distance between the two was as small as possible. We refer to these datasets as to 'close-by' datasets. As shown in Supp. Figure S1 and Figure 4C, DDIG-SN retained most of its prediction performance with the AUC of 0.83 and 0.84 in cross-validation and independent testing, respectively. Compared to the second

best method, SilVA, DDIG-SN yielded larger relative (as well as absolute) improvements (5%, $p = 1.3\times10^{-4}$, for cross-validation and 5%, $p = 0.008$, for the test) than for the general case using the full datasets.

Another way of reducing the imbalance in our datasets is by filtering based on the AF for the common (AF ≥ 1%) or rare (AF < 1%) putatively benign variants. While the first helps to reduce the potential false negatives (non-benign 1kGP variants), the latter shows how accurate the method is for rare variants which cannot be prioritised by comparing with variants from the general healthy population. DDIG-SN retained its prediction performance and relative improvements compared to SilVA for discriminating both common and rare benign variants (Supp. Table S3).

Finally, we evaluated how robustly DDIG-SN performed upon randomly masking 30% of the test data. Figure 5 shows the distributions of DDIG-SN's and SilVA's AUC scores for the cross-validation, independent test, and 'close-by' test datasets. Both DDIG-SN and SilVA performed reasonably robustly with standard deviations ranging 0.007–0.009 and 0.008–0.010, respectively. The slightly smaller standard deviations yielded by DDIG-SN show that its performance was somewhat more robust.

Evaluating predictions based on variant's distance to the exon-intron splice junction

Four of the six features in DDIG-SN are related to alternative splicing. Could this be caused by possible dominance of diseases caused by splicing-related events in our dataset? To examine this possibility, we defined three distinct exonic regions based on the distance to the nearest exon-intron junction: 1) the immediate proximity to the exon-intron junction (1–3 bp), 2) the larger region commonly considered as being implicated in splicing (4–69 bp), and 3) the exon core (≥70 bp from the junction). We selected these three regions as they had been used previously for the analysis of the distribution of human pathogenic variants (Wu and Hurst, 2016).

Table 4 shows the performance of DDIG-SN and the six compared methods for the three different regions (1–3, 4–69, and ≥70 bp away from an exon-intron junction). All methods yielded comparable or improved prediction performance for the 1–3 bp region when compared to the full dataset. For instance, on the cross-validation dataset, DDIG-SN and SilVA yielded AUC scores of 0.85 and 0.89 (compared to 0.84 and 0.81 for the full dataset). Thus, SilVA outperformed DDIG-SN in the immediate neighbourhood of the splice site junction with a relative improvement of 5% ($p = 0.003$). SilVA yielded the same improvement (5%) to DDIG-SN also on the independent test set ($p = 0.009$). Regarding the two regions further away from the exon-intron junction (4–69 and ≥70 bp), prediction accuracy dropped considerably for all compared methods as opposed to the full dataset. For instance, DDIG-SN and SilVA yielded the AUC of 0.76 and 0.69 for the 4–69 bp region, respectively, and 0.64 and 0.62, respectively, for the ≥70 bp region (cross-validation dataset). Thus, DDIG-SN yielded a relative improvement of 10% to SilVA for the 4–69 bp region ($p = 1.3\times10^{-5}$). Even though DDIG-SN retained most of its prediction performance also for the independent test (AUC of 0.75), its improvement to SilVA (AUC of 0.73) was not significant ($p = 0.333$).

Apart from comparing DDIG-SN with related work, we wanted to explore if it was possible to build a more accurate method, dedicated to each of the three exonic regions. Hence, we developed another method, called 'Region-Specific Model' (RSM), which comprised three SVM models ($RSM_{1–3}$, $RSM_{4–69}$, and $RSM_{\geq70}$), each optimised for the different region of an exon (1–3, 4–69, and ≥70 bp, respectively). The optimisation was performed in terms of splitting the design/training dataset based on the variant's distance the exon-intron junction and then running the stability feature selection and grid search for optimal SVM hyper-parameters on each subset individually. When predicting a test set variant using the RSM method, exactly one of the three models was selected depending upon the distance of the test variant to the exon-intron junction. As shown in Figure 1, more than 40% of all disease-causing variants in our design/training dataset were found within 3 bp of an exon-intron junction, indicating that the disease-causing SN variants rarely occur further away or in an exon core.

Figures 6A and 6B show the comparison of DDIG-SN and RSM (in terms of AUC) for the training and test datasets, respectively. The region-specific design of RSM resulted in a significantly improved performance as compared to DDIG-SN for the 1–3 bp region ($p = 1.3\times10^{-5}$ and 0.023 for cross-validation and test set, respectively). However, RSM yielded only comparable performance for the 4–69 bp region ($p = 0.689$ and 0.622) and its performance was worse than random (AUC < 0.5) for the ≥70 bp region. The latter can be attributed to 1) the small size of the dataset (only 75 disease-causing variants present as opposed to 261 and 256 disease-causing variants in the 1–3 and 4–69 bp regions, respectively); and 2) lack of relevant features (the *stability* metric of the most often selected feature was only 0.24 as opposed to 1.00 and 0.92 for the top features selected for the 1–3 and 4–69 bp regions, respectively). From RSM's performance, we concluded that employing several individually-trained models may increase the risk of overtraining, yet it does not offer considerable performance improvements over DDIG-SN.

It is of interest to know if the region-specific approach (RSM) selected some interesting features that were specific for the three distinct regions. Table 5 lists the features selected for $RSM_{1–3}$, $RSM_{4–69}$, and $RSM_{\geq70}$, alongside the six features of DDIG-SN, ranked by the *stability* metric. As mentioned in the previous paragraph, there were no features with *stability* ≥ 0.25 for $RSM_{\geq70}$. The selected features (*stability* ≥ 0.25) for either of the two other models ($RSM_{1–3}$ and $RSM_{4–69}$) were all related to RNA splicing and evolutionary conservation. Specifically, the features with the highest *stability* were ΔMES– and ΔMES+ for $RSM_{1–3}$ and $RSM_{4–69}$, respectively. Thus, the disease-causing effects in the immediate proximity to exon-intron junctions are driven by the loss of splicing motif strength, whereas the gain of splicing motif strength is the primary cause in regions adjacent to the splice site.

DISCUSSION

We have developed a new method, termed DDIG-SN, for discriminating disease-causing synonymous variants. The method was trained and evaluated using two different datasets with protein sequence identity < 30%. At the same time, each gene had at least one disease-causing and one neutral variant present. This allows for protein-stratified cross-validation and independent testing while ensuring that DDIG-SN was trained for discriminating variants rather than genes, meaning that its performance is a realistic estimate for variants in previously unseen genes. We also minimised overtraining by selecting the most stable six features using the stability selection algorithm (Meinshausen and Buhlmann, 2010). To this end, we performed 100 sequential forward selections (Whitney, 1971) using 100 randomly created data samples that had an equal number of disease-causing and benign variants per gene, compared to 1:18 ratio of disease-causing to benign variants in the full 'real-world' training dataset. The robustness of DDIG-SN is evident from the similar performance between 10-fold cross-validation and the independent test (Table 3), stable performance upon randomly masking 30% of the test data (Figure 5) as well as using only the common (AF ≥ 1%), rare (AF < 1%), or 'close-by' putatively benign 1kGP variants (Supp. Table S3 and Figure 4C). More importantly, DDIG-SN had the best performance among all the methods compared across different datasets.

SN variants have a direct effect only on the DNA and RNA sequences. Thus, it is expected that the protein-based features would not be useful for discrimination of SN variants. However, it has been observed that SN variants can impact both structure and function of the protein (Kimchi-Sarfaty, et al., 2007; Montera, et al., 2001; Zhou, et al., 2013). Thus, it is necessary to examine some of the protein-based features that have been found to be important in discriminating non-synonymous variants that result in single amino acid substitutions (Adzhubei, et al., 2010; Folkman, et al., 2016), nonsense variants that introduce premature termination codons (Folkman, et al., 2015), and indels that lead to the addition or removal of a short segment of amino acid residues (non-frameshifting indels) (Bermejo-Das-Neves, et al., 2014; Zhao, et al., 2013) or randomization of the protein sequence (frameshifting indels) (Douville, et al., 2016; Folkman, et al., 2015; Hu and Ng, 2012). To this end,

we analysed the sequence and structural features related to protein sequence conservation derived from PSI-BLAST (Altschul, et al., 1997) and HHblits (Remmert, et al., 2012), protein intrinsic disorder predicted with SPINE-D (Zhang, et al., 2012), and protein secondary structure and solvent accessibility predicted with SPIDER2 (Heffernan, et al., 2015). All these features yielded a marginal discriminative power (AUC < 0.54) when evaluated individually. Moreover, they were not selected during the feature selection for the final SVM model. This result supports the notion that changes in the protein structure and function are likely the secondary effects of the changes in the RNA/DNA sequence.

When analysing the selected DNA and RNA-level features, we found that DDIG-SN model was dominated by features related to RNA splicing (four out of six features in total). This confirms the damaging role of SN variants in RNA splicing found in many studies (Buske, et al., 2013; Li, et al., 2017; Xiong, et al., 2015; Yeo and Burge, 2004). Here, we were able to show that whilst the loss of splicing motif strength is the most robust predictive feature in the immediate neighbourhood of the exon-intron junction (1–3 bp), it is the gain of splicing motif strength that is important for the region adjacent to the splice site (4–69 bp).

The key features other than splicing relate to DNA sequence conservation. The evolutionary conservation of DNA sequence can be due to a structural or functional requirement at either the DNA or RNA level. To clear up the ambiguity, we removed all features related to evolutionary conservation and performed the feature selection again. Then, all selected features were related to splicing (Supp. Table S4).

The lack of mRNA structure-related features [secondary structure folding energy (Markham and Zuker, 2008), ensemble diversity (Lorenz, et al., 2011), and solvent-accessible surface area (Yang, et al., 2017)], despite ample experimental evidence for their role in translation and transcription

(Mortimer, et al., 2014; Wan, et al., 2014), could reflect the challenge in predicting RNA structural properties accurately, in particular, in the absence of their interacting partners, irrespective of whether it is at the secondary or tertiary level (Hajdin, et al., 2013; Miao, et al., 2015; Seetin and Mathews, 2012). Another feature that was not selected was the possible formation or disruption of G-quadruplexes due to SN variants. It had a near random AUC score as a single feature (AUC = 0.502) despite the occurrence of G-quadruplexes that have been implicated in neurodegenerative diseases and the non-coding transcriptome (Rhodes and Lipps, 2015; Simone, et al., 2015). In addition, variant-induced changes in binding affinities of DNA and RNA-binding proteins were not selected, despite ample evidence of their regulatory roles (Dreyfuss, et al., 2002; Stergachis, et al., 2013). Possible reasons for this could be: firstly, inaccuracy of the employed (state-of-the-art) prediction methods, secondly, limited size of the disease-causing SN variants dataset from the HGMD, and thirdly, binding disrupted at one site may be compensated for at another site.

The limited prediction performance of all evaluated methods for SN variants in exon cores (≥70 bp away from exon-intron junction) suggests that there are unknown features, not considered by any of the methods. Possibly, a more sensitive prediction of RNA and protein structural features, which could predict not only the wild-type but also the structural changes induced by genetic variations, could help improve predictions in these regions. Additionally, more accurate models of the functional impact of genetic variants, such as 'compensation-aware' models of RNA and DNA-binding proteins affinity changes, are needed.

ACKNOWLEDGEMENTS

We are grateful to Orion Buske for discussions regarding SilVA.

This research was supported by National Health and Medical Research Council of Australia (grant numbers 1059775 and 1083450) to Y.Z. The authors also gratefully acknowledge the support of the

Griffith University eResearch Services Team and the use of the High Performance Computing Cluster 'Gowonda' to complete this research. D.N.C. and M.M. acknowledge the financial support of Qiagen Inc through a Licence Agreement with Cardiff University.

Conflict of interest statement: None declared.

REFERENCES

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nature Methods 7(4):248-9.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology 33:831–838.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17):3389-3402.

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E and others. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37(7):766-70.

Bermejo-Das-Neves C, Nguyen HN, Poch O, Thompson JD. 2014. A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). BMC Bioinformatics 15:111.

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier J-F, Hébuterne X and others. 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nature Genetics 43:242-245.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 24(24):10505-14.

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. 2013. Identification of deleterious synonymous variants in human genomes. Bioinformatics 29(15):1843-50.

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. Genome Biology 14:R143.

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. Bioinformatics 23(15):1875-82.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. Journal of Molecular Evolution 62:89-98.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biology 6:R75.

Cortes C, Vapnik V. 1995. Support-Vector Networks. Machine Learning 20(3):273-297.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP plus. Plos Computational Biology 6(12).

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. Genes & Development 25(10):1010-1022.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837-845.

Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, Ryan M, Karchin R. 2016. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat 37(1):28-35.

Dreyfuss G, Kim VN, Kataoka N. 2002. Messenger-RNA-binding proteins and the messages they carry. Nature Reviews Molecular Cell Biology 3(3):195-205.

Duan J, Shi J, Ge X, Dolken L, Moy W, He D, Shi S, Sanders AR, Ross J, Gejman PV. 2013. Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. Sci Rep 3:1318.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS biology 2:E268.

Folkman L, Stantic B, Sattar A, Zhou Y. 2016. EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. Journal of Molecular Biology 428:1394–1405.

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

Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N and others. 2013. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proceedings of the National Academy of Sciences of the United States of America 110:13481-13486.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW and others. 2015. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. Human Mutation 36(5):513-523.

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc Natl Acad Sci U S A 110(14):5498-5503.

Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y. 2015. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific Reports 5:11476.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet 42:287-99.

Ho PA, Kuhn J, Gerbing RB, Pollard JA, Zeng R, Miller KL, Heerema NA, Raimondi SC, Hirsch BA, Franklin JL and others. 2011. WT1 synonymous single nucleotide polymorphism rs16754 correlates with higher mRNA expression and predicts significantly improved outcome in favorable-risk pediatric acute myeloid leukemia: a report from the children's oncology group. J Clin Oncol 29(6):704-11.

Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome biology 13(2):R9.

Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. Trends Genet 30(7):308-21.

Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. Journal of Molecular Evolution 63:174-182.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Research 32:D493-D496.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315(5811):525-528.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310-5.

Kirchner S, Ignatova Z. 2015. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nature Reviews. Genetics 16:98-112.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324(5924):255-8.

Li M, Feng W, Zhang X, Yang Y, Wang K, Mort M, Cooper DN, Wang Y, Zhou Y, Liu Y. 2017. ExonImpact: Prioritizing Pathogenic Alternative Splicing Events. Hum Mutat 38:16-24.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol Biol 6:26.

Macaya D, Katsanis SH, Hefferon TW, Audlin S, Mendelsohn NJ, Roggenbuck J, Cutting GR. 2009. A synonymous mutation in TCOF1 causes Treacher Collins syndrome due to mis-splicing of a constitutive exon. Am J Med Genet A 149A(8):1624-7.

Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. In: Keith J, editor. Methods in Molecular Biology (Bioinformatics, Volume II.). Totowa, NJ: Humana Press. p 3-31.

Meinshausen N, Buhlmann P. 2010. Stability selection. Journal of the Royal Statistical Society Series B-Statistical Methodology 72:417-473.

Miao Z, Adamiak RW, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cheng C, Chojnowski G, Chou FC, Cordero P and others. 2015. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. RNA 21(6):1066-84.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D and others. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Research 17(12):1797-1808.

Montera M, Piaggio F, Marchese C, Gismondi V, Stella A, Resta N, Varesco L, Guanti G, Mareni C. 2001. A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. J Med Genet 38(12):863-867.

Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, Sanford JR, Mooney SD. 2014. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol 15(1):R19.

Mortimer SA, Kidwell MA, Doudna JA. 2014. Insights into RNA structure and function from genome-wide studies. Nature Reviews Genetics 15(7):469-479.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V and others. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485(7397):242-5.

Niroula A, Vihinen M. 2016. Variation interpretation predictors: principles, types, performance, and choice. Human Mutation 37(6):579–597.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Molecular Biology and Evolution 23:301-309.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1):32-42.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20(1):110-21.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ and others. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19(7):1316-23.

Remmert M, Biegert A, Hauser A, Soding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 9(2):173-5.

Rhodes D, Lipps HJ. 2015. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Research 43(18):8627-8637.

Rudolph KLM, Schmitt BM, Villar D, White RJ, Marioni JC, Kutter C, Odom DT. 2016. Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. PLoS Genetics 12.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A and others. 2014. A framework for the interpretation of de novo mutation in human disease. Nat Genet 46(9):944-50.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12(10):683-91.

Savisaar R, Hurst LD. 2017. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. Molecular Biology and Evolution 34:1110-1126.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. Nature Methods 11(4):361-362.

Seetin MG, Mathews DH. 2012. RNA structure prediction: an overview of methods. Methods Mol Biol 905:99-122.

Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res 41(4):2073-94.

Sharp PM, Li WH. 1987. The Codon Adaptation Index - a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications. Nucleic Acids Research 15(3):1281-1295.

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31(10):1536-1543.

Simone R, Fratta P, Neidle S, Parkinson GN, Isaacs AM. 2015. G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome. FEBS Lett 589(14):1653-68.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Human Molecular Genetics 15:2490-2508.

Stark K, Esslinger UB, Reinhard W, Petrov G, Winkler T, Komajda M, Isnard R, Charron P, Villard E, Cambien F and others. 2010. Genetic association study identifies HSPB7 as a risk gene for idiopathic dilated cardiomyopathy. PLoS Genet 6(10):e1001167.

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Human Genetics 136(6):665-677.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM and others. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. Science 342(6164):1367-72.

Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. 2014. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell 156(6):1324-1335.

Todd AK, Johnston M, Neidle S. 2005. Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Research 33:2901-2907.

UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43(Database issue):D204-12.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Human Mutation 34(2):275–282.

Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E and others. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505(7485):706-9.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. Cell 119:831-845.

Whitney AW. 1971. A direct method of nonparametric measurement selection. IEEE Transactions on Computers 100:1100-1103.

Wu X, Hurst LD. 2016. Determinants of the Usage of Splice-Associated cis-Motifs Predict the Distribution of Human Pathogenic SNPs. Molecular Biology and Evolution 33:518-529.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua YM, Gueroussov S, Najafabadi HS, Hughes TR and others. 2015. The human splicing code reveals new insights into the genetic determinants of disease. Science 347(6218).

Yang Y, Li X, Zhao H, Zhan J, Wang J, Zhou Y. 2017. Genome-scale characterization of RNA tertiary structures and their functional impact by RNA solvent accessibility prediction. RNA 23:14-22.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11(2-3):377-94.

Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. 2012. SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. Journal of Biomolecular Structure and Dynamics 29:799–813.

Zhang X, Lin H, Zhao H, Hao Y, Mort M, Cooper DN, Zhou Y, Liu Y. 2014. Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation. Human Molecular Genetics 23:3024–3034.

Zhang XH-F, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. Genes & Development 18:1241-1250.

Zhao H, Yang Y, Lin H, Zhang X, Mort M, Cooper DN, Liu Y, Zhou Y. 2013. DDIG-in: Discriminating between disease-causing and neutral non-frameshifting micro-INDELs by support vector machines by means of integrated sequence- and structure-based features. Genome Biol 14:R43.

Zhou M, Guo JH, Cha J, Chae M, Chen S, Barral JM, Sachs MS, Liu Y. 2013. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 495(7439):111-115.

Zhu J, Mayeda A, Krainer AR. 2001. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Molecular Cell 8(6):1351-1361.

FIGURE LEGENDS

Figure 1. The distributions of the disease-causing and benign SN variants for **(A)** the top RNA-based feature MES (the splicing motif strength predicted with MaxEntScan); **(B)** the top DNA-based feature $\Delta AF_{cons100}$ [the difference between the alternative (alt.) and reference (ref.) allele frequencies]; and **(C)** three exonic regions defined based on the distance to the nearest exon-intron junction [immediate proximity of the splice junction (1–3 bp), larger region commonly considered as implicated in splicing (4–69 bp), and exon core (≥70 bp)].

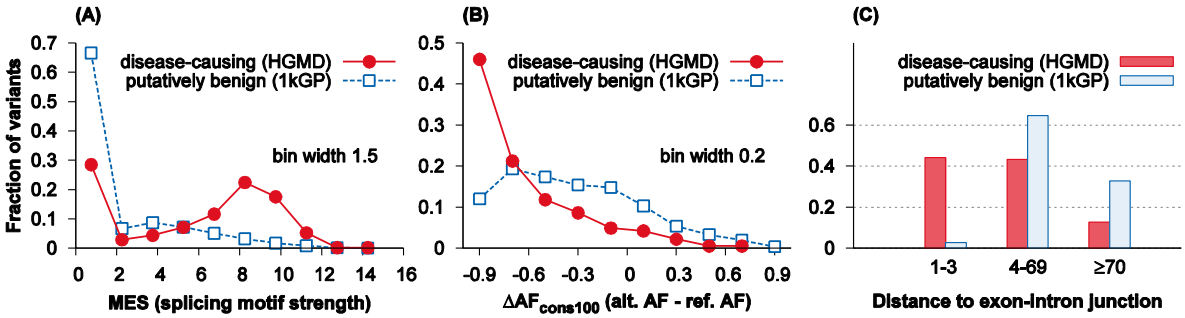

Figure 2. The area under the ROC curve (AUC) as a function of the number of features selected with the sequential floating selection (SFS) algorithm. The boxplots show the distributions of AUC values for the 100 random balanced sampled datasets used for feature selection.

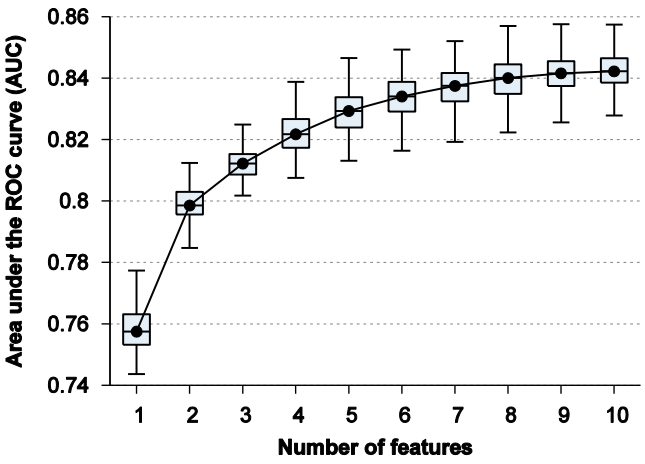

Figure 3. The *stability* metric (reported by stability selection) is compared to the single feature area under the ROC curve (AUC) for the ten most stable features. The first six features were above the *stability* threshold of 0.25 and were therefore selected for the DDIG-SN model.

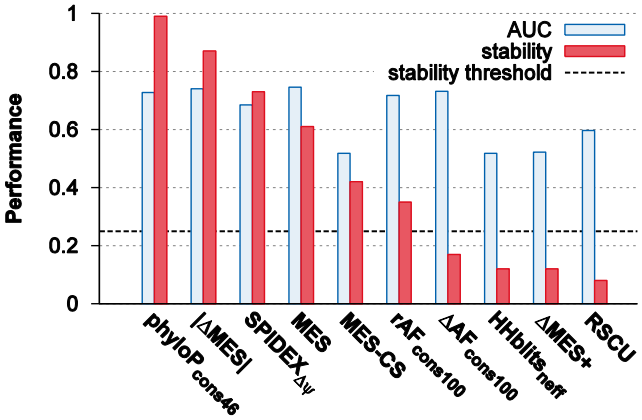

Figure 4. The receiver operating characteristic (ROC) curves of DDIG-SN compared to other available methods: a synonymous-specific model – SilVA, conservation score – phyloP, general approaches for all types of SNVs – FATHMM-MKL, CADD and MutationTaster, and an approach to predict how a variant affects splicing – SPIDEX. **(A)** DDIG-SN was evaluated using protein-stratified 10-fold cross-validation (CV) on the same dataset as used for its design. **(B)** A comparison using an independent test set with < 30% sequence similarity to the design/training dataset. **(C)** A balanced subset of the independent test set in which the closest benign (1kGP) variant was selected for every disease-causing (HGMD) variant (the 'close-by' dataset).

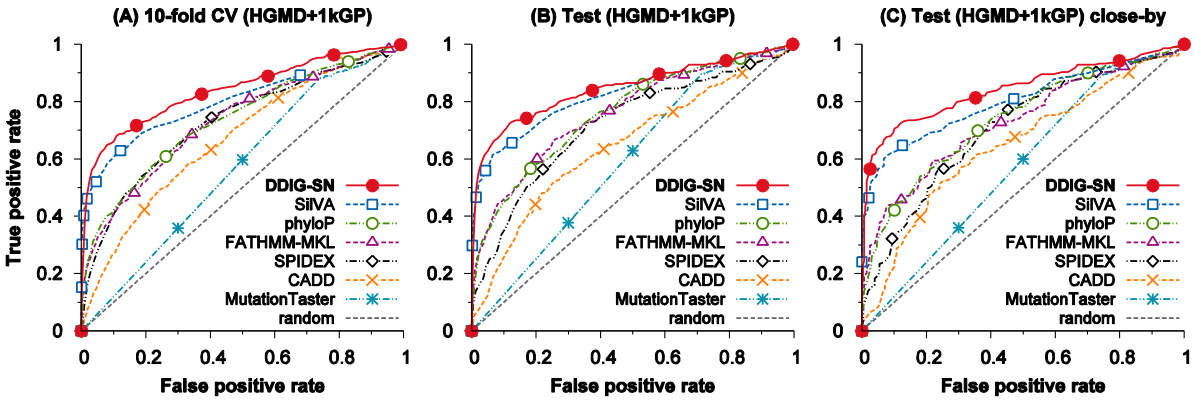

Figure 5. Distributions of AUC scores for DDIG-SN and SilVA after randomly masking 30% of the test data. Although both methods show relatively robust performance, DDIG-SN is somewhat more robust as shown by the smaller size of the boxes (each box represents the interquartile of the AUC distribution).

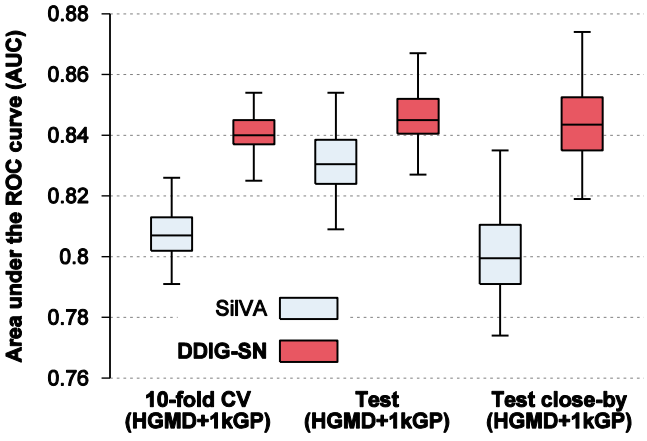

Figure 6. Area under the ROC curve (AUC) of DDIG-SN and RSM (Region-Specific Model) for the three distinct exonic regions defined based on the distance to the nearest exon-intron junction. RSM's AUC scores which were significantly different ($p < 0.05$, DeLong's test) from DDIG-SN are highlighted using diagonal stripes. The plot demonstrates that the prediction ability of both methods drops with the increased distance to the exon-intron junction.

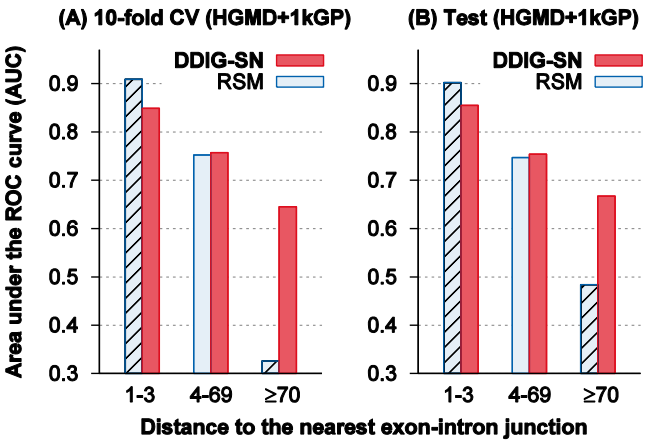

Table 1. Synonymous (SN) variants datasets used for the design and evaluation of DDIG-SN.

Dataset	Sampling	Genes	Variants	
			HGMD	1kGP
design/ training	full	318	592	10,925
	close-by		591	591
test	full	143	279	4,945
	close-by		278	278

Table 2. Ten most discriminative single features sorted by their areas under the ROC curve (AUC).

Feature name	AUC[a]	MCC[b]
MES	0.746 [+]	0.494
|ΔMES|	0.740 [+]	0.526
$\Delta AF_{cons100}$	0.732 [–]	0.390
$phyloP_{cons46}$	0.727 [+]	0.381
$GERP{+}{+}_{cons34}$	0.724 [+]	0.379
$rAF_{cons100}$	0.718 [+]	0.384
$splice_{closest}$	0.705 [–]	0.426
$splice_{3'}$	0.697 [–]	0.367
$SPIDEX_{\Delta\Psi}$	0.685 [–]	0.300
MES-KM	0.670 [+]	0.464

[a] Area under the ROC curve; the [+]/[–] symbol determines positive/negative correlation (i.e. [+] means that the disease causality is positively correlated with the particular feature).

[b] Matthews correlation coefficient

Table 3. Comparison of DDIG-SN's prediction performance with six other available methods using cross-validation and independent test datasets.

Method	10-fold CV (HGMD+1kGP)				Test (HGMD+1kGP)			
	full dataset[a]		close-by[b]		full dataset[a]		close-by[b]	
	AUC[c]	MCC[c]	AUC[c]	MCC[c]	AUC[c]	MCC[c]	AUC[c]	MCC[c]
MutationTaster 2	0.57	0.07	0.57	0.18	0.59	0.09	0.57	0.19
CADD 1.3	0.67	0.13	0.65	0.25	0.65	0.14	0.65	0.29
SPIDEX 1.0	0.73	0.22	0.69	0.32	0.72	0.22	0.71	0.34
FATHMM-MKL 2.3	0.73	0.29	0.70	0.34	0.76	0.33	0.73	0.41
phyloP (cons46way)	0.74	0.30	0.71	0.35	0.76	0.31	0.74	0.38
SilVA 1.1.1	0.81	**0.54**	0.79	0.52	0.83	0.55	0.80	0.57
DDIG-SN	**0.84**	0.51	**0.83**	**0.58**	**0.85**	**0.56**	**0.84**	**0.62**

[a] The full unbalanced dataset with a ratio of 1:18 of disease-causing (HGMD) to putatively benign (1kGP) variants.

[b] A balanced subset of the full dataset in which the closest benign variant was selected for each disease-causing variant.

[c] AUC, area under the ROC curve; MCC, Matthews correlation coefficient

Table 4. Comparison of DDIG-SN's prediction performance with six other available methods with the dataset partitioned based on variant's distance to the exon-intron junction.

Method	AUC[a]					
	10-fold CV (HGMD+1kGP)			Test (HGMD+1kGP)		
	1–3[b]	4–69[b]	≥ 70[b]	1–3[b]	4–69[b]	≥ 70[b]
MutationTaster 2	0.54	0.53	0.52	0.54	0.54	0.55
CADD 1.3	0.76	0.56	0.60	0.77	0.53	0.48
SPIDEX 1.0	0.79	0.61	0.42	0.80	0.62	0.49
FATHMM-MKL 2.3	0.83	0.62	0.56	0.85	0.62	0.64
phyloP (cons46way)	0.85	0.62	0.59	0.86	0.62	0.65
SilVA 1.1.1	**0.89**	0.69	0.62	**0.90**	0.73	0.65
DDIG-SN	0.85	**0.76**	**0.64**	0.86	**0.75**	**0.67**

[a] AUC, area under the ROC curve.

[b] Dataset subset comprising variants in the given range of base pairs to the nearest exon-intron junction.

Table 5. Comparison of different features selected (*stability* ≥ 0.25) for DDIG-SN and the three models of the RSM method.

DDIG-SN		**Region specific model (RSM)**					
		1–3[a]		**4–69**[a]		**≥ 70**[a]	
feature	***stability***[b]	**feature**	***stability***[b]	**feature**	***stability***[b]	**feature**	***stability***[b]
$phyloP_{cons46}$	0.99	ΔMES–	1.00	ΔMES+	0.92	$C\text{-}term_{dist}$	0.24[c]
|ΔMES|	0.87	$phyloP_{cons46}$	0.89	$\Delta AF_{cons100}$	0.56	PESE–	0.19[c]
$SPIDEX_{\Delta\Psi}$	0.73	FAS6+	0.30	$SPIDEX_{\Delta\Psi}$	0.47	ASA_{prot}	0.17[c]
MES	0.61	$ESE_{loss/gain}$	0.28	$rAF_{cons100}$	0.41	$centre_{dist}$	0.17[c]
MES-CS	0.42			$phyloP_{cons46}$	0.34	RSCU	0.16[c]
$rAF_{cons100}$	0.35			MES	0.34	$DeepBind_{max}$	0.15[c]

[a] Nearest exon-intron junction distance of variants used for training the given model of the RSM

[b] The *stability* metric equals to the fraction of times the particular feature was selected using sequential feature selection (SFS) out of 100 runs on random balanced samples of the training/design dataset.

[c] Stability selection for RSM model comprising variants ≥ 70 bp from an exon-intron junction resulted in no features with *stability* ≥ 0.25. Thus, we employed a threshold of 0.15 to be able to train a model for these variants.