Hum Genet
DOI 10.1007/s00439-017-1783-x

ORIGINAL INVESTIGATION

regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution

Xinjun Zhang[1,2] · Meng Li[2,3] · Hai Lin[2,4] · Xi Rao[2] · Weixing Feng[3] · Yuedong Yang[5] · Matthew Mort[6] · David N. Cooper[6] · Yue Wang[7] · Yadong Wang[8] · Clark Wells[9] · Yaoqi Zhou[10] · Yunlong Liu[2,7,11]

Received: 30 August 2016 / Accepted: 27 February 2017

Abstract While synonymous single-nucleotide variants (sSNVs) have largely been unstudied, since they do not alter protein sequence, mounting evidence suggests that they may affect RNA conformation, splicing, and the stability of nascent-mRNAs to promote various diseases. Accurately prioritizing deleterious sSNVs from a pool of neutral ones can significantly improve our ability of selecting functional genetic variants identified from various genome-sequencing projects, and, therefore, advance our understanding of disease etiology. In this study, we develop a computational algorithm to prioritize sSNVs based on their impact on mRNA splicing and protein function. In addition to genomic features that potentially affect splicing regulation, our proposed algorithm also includes dozens structural features that characterize the functions of alternatively spliced exons on protein function. Our systematical evaluation on thousands of sSNVs suggests that several structural features, including intrinsic disorder protein scores, solvent accessible surface areas, protein secondary structures, and known and predicted protein family domains, show significant differences between disease-causing and neutral sSNVs. Our result suggests that the protein structure features offer an added dimension of information while distinguishing disease-causing and neutral synonymous variants.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-017-1783-x) contains supplementary material, which is available to authorized users.

✉ Yunlong Liu
yunliu@iu.edu

Xinjun Zhang
zhangxin@indiana.edu

Meng Li
limenghrbeu@foxmail.com

Hai Lin
linhai@iupui.edu

Xi Rao
raox@indiana.edu

Weixing Feng
fengweixing@hrbeu.edu.cn

Yuedong Yang
yuedong.yang@griffith.edu.au

Matthew Mort
mortm@cardiff.ac.uk

David N. Cooper
cooperdn@cardiff.ac.uk

Yue Wang
yuewang@iu.edu

Yadong Wang
ydwang@hit.edu.cn

Clark Wells
wells4@iu.edu

Yaoqi Zhou
yaoqi.zhou@griffith.edu.au

1 School of Informatics and Computing, Indiana University, Bloomington, IN 47408, USA

2 Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 w 10th st, HS 5000, Indianapolis, IN 46202, USA

3 Pattern Recognition and Intelligent System Institute, Automation College, Harbin Engineering University, Harbin 150001, Heilongjiang, China

4 School of Informatics and Computing, Indiana University Purdue University Indianapolis, Indianapolis, IN 46202, USA

5 School of Information and Communication Technology, Gold Coast Campus, Griffith University, Southport, QLD 4222, Australia

Published online: 08 April 2017

The inclusion of structural features increases the predictive accuracy for functional sSNV prioritization.

Introduction

While single-nucleotide variants (SNVs) underlie a myriad of diseases, synonymous SNVs (sSNVs) that do not alter which amino acid is encoded have traditionally been assumed to have little or no biological impact. However, recent work suggests that sSNVs may contribute to disease pathogenesis by affecting the affinity of RNA-binding proteins to disrupt RNA processing and/or translational control (Wan et al. 2014). The importance of synonymous point mutations in cancer has been further demonstrated by a recent survey based on roughly 4000 cancer exomes from 19 cancer types, which showed a significant enrichment of synonymous mutations in oncogenes, as compared to non-cancer genes with matched genomic features (Supek et al. 2014; Li et al. 2016; Xiong et al. 2015; Cartegni et al. 2002; Sauna and Kimchi-Sarfaty 2011; Duan et al. 2003; Macaya et al. 2009; Chamary and Hurst 2005).

Current bioinformatics tools in prioritizing deleterious sSNVs mainly focus on the potential impacts of individual variants on splicing outcome. Such methods often derive a series of genomic features describing how a candidate variant can potentially affect splicing regulation, and attempt to use these features to predict either disease relevance, or splicing outcome, as measured by large-scale RNA-seq experiments. Despite the positive prediction power in prioritizing disease-causing sSNVs, such methods, however, do not consider whether the affected splicing events will result in major protein function changes (Mort et al. 2014). As demonstrated in our previous analysis on non-frame shifting micro-insertions/deletions (INDELs), inclusion or exclusion of a stretch amino-acid sequences does not guarantee the functional changes of affected protein, unless they occur within key structural elements of the protein (Zhao et al. 2013). In addition, recent surveys also suggest that many splicing variations are crucial to the protein functions and organismal phenotypes (Xiong et al. 2015; Kelemen et al. 2013; Rivas et al. 2015; Zheng and Black 2013; Faustino and Cooper 2003).

In this study, we hypothesize that considering the exon-specific protein structure features will significantly increase the accuracy of the prediction. Using potential disease-causing and neutral data sets derived from the human gene mutation database (HGMD), ClinVar, and 1000 Genomes projects, we systematically evaluated hundreds of genomics and protein structure features that are associated to individual synonymous SNVs. Our results suggest that including protein structure features dramatically increases our ability for identifying disease-causing synonymous SNVs.

Results

Training data set

We constructed a training data set that includes both disease-causing and neutral sSNVs. The disease-causing sSNVs were selected from the human gene mutation database (HGMD) (Stenson et al. 2014), and the neutral sSNVs were selected from the 1000 Genomes database (Genomes Project C et al. 2012). As of September 2014, the HGMD database contains 1111 deleterious synonymous mutations that affect splicing, of which 697 locate on the splice sites (+1/+2/+3 loci in donor site and −1 locus in acceptor site), and 414 reside inside the exon but off the splice sites. These two types of sSNVs are referred as variants on splice site consensus (VSS) and variants in internal exons (VIE), respectively.

Most of the variants in our training data set are in the DM (disease-causing) category with direct evidence of being disease-causing mutations. Specifically, out of 697 VSS HGMD variants, 656 (94.1%) are in the DM (disease-causing) category, and out of 414 VIE HGMD variants, 344 (83.1%) are from DM (disease-causing) category. The overall distribution of categories is shown in Fig. S1. Since these two types of variants may affect splicing regulation with different mechanisms, with VSS variants more likely to directly interfere with the formation of the splicesome, while VIE variants playing more roles in affecting RNA-binding protein (RBP) binding, their impacts on splicing regulation were evaluated separately. To avoid inflating the over-representation of certain genomic features due to the occurrences of multiple variants in the same affected exon, we randomly select only one variant per exon in the further analysis. This process results in a total of 980 deleterious sSNVs in the HGMD database, of which 651 and 329 locate on and off splice sites, respectively. Similar as our earlier

[6] Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

[7] Departments of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[8] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China

[9] Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[10] Institute for Glycomics and School of Informatics and Communication Technology, Griffith University, Parklands Dr., Southport, QLD 4215, Australia

[11] Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

study on INDELs (Zhang et al. 2014), the neutral sSNVs were selected from the 1000 Genomes Project, in which genotyped individuals did not exhibit any apparent disease phenotypes. The 1000 Genomes data contain 2582 VSS and 66,900 VIE variants, respectively. To minimize false positives in the neutral group of the training set, we only selected those sSNVs with a minor allele frequency (MAF) greater than a threshold (3% for VSS and 10% for VIE variants). The overall gold standard data set includes 651 disease-causing and 399 neutral VSS variants, and 329 disease-causing and 7231 neutral VIE variants, respectively. To make a balanced training set, we randomly selected the same number of negative data set as positive data set to train and test our models. To evaluate features that are associated with disease-causing and neutral sSNVs, and build computational model for novel variant prioritization, we used 2/3 of our data set as training data, and the remaining 1/3 as independent test data.

Disease-causing variants tend to impact splicing regulation

We evaluated a broad array of features that can be classified into three major categories: genomic features characterizing how individual sSNVs affect splicing regulation, the structural features evaluating how the inclusion/exclusion of alternatively spliced exons affect protein function, and others (such as conservation). A detailed list of features and how they are derived can be found in the supplementary materials and online methods. As reported in the previous studies, features characterizing how sSNVs affect splicing regulation play important roles in distinguishing disease-causing and neutral variants (Barash et al. 2010). For instance, among the 201 RNA-binding proteins (RBPs) with known position weight matrices (PWMs) (Ray et al. 2013), disease-causing sSNVs showed greater alteration on RBP binding in terms of specific diseases, comparing with neutral variants (Fig. S2). This is consistent on both the magnitude of matching score changes, and the probability that an sSNV changes RBP binding (detailed calculation methods can be found in online methods, Fig. S3). Similarly, other features associated with individual variants, such as the RNA secondary structure features on the variant loci, the inherent strength of 5′- and 3′-splicing sites of exons containing candidate variant, the distance between the variant loci and splicing junction, and the ability of the variants disrupting the cluster of exonic splicing enhancers and silencers, all have statistically significant prediction power for distinguishing disease-causing and neutral variants, as evaluated by the Matthew's correlation coefficient (MCC), and Kolmogorov–Smirnov (K–S) test (Supplement Table 1; Fig. S4).

Disease-causing variants tend to locate within the exons of key structural regions

In addition to the genomic features related to splicing regulation, we have observed strong prediction power for the measurements characterizing the protein structure features of the exons containing the putative variants. Disease-causing sSNVs tend to locate in the exons with lower average solvent accessible surface areas (ASA), indicating that they are more likely to be in buried core protein regions (K–S test p value $= 2.6 \times 10^{-7}$, Fig. 1a, b, Fig. S5). In addition, comparing to neutral variants, disease-causing ones are also under-represented in the exons in intrinsic disorder regions (K–S test p value $= 7.0 \times 10^{-10}$, Fig. 1c, d, Fig. S6), suggesting that they are more likely to be in the structural regions. Consistent with these observations, disease-causing variants tend to reside in the exons with higher percentage of overlapping with known or predicted protein family domains (K–S test p value $= 5.9 \times 10^{-6}$, Fig. 1g, h, Fig. S7). As for the protein secondary structures, exons containing disease-causing sSNVs are enriched for alpha-helix (K–S test p value $= 0.004$), and random coil (K–S test, p value $= 0.001$) (Fig. 1e, f, Fig. S8). All these observations strongly suggest that, in addition to features related to splicing regulation, protein structure features on the variant-containing exons can provide additional layer of information in distinguishing disease-causing and neutral variants.

Prioritizing sSNVs based on their impact of splicing regulation and protein structure

Based on the aforementioned evaluation, a random forest algorithm was employed for building a prediction model for distinguishing disease-causing and neutral sSNVs. We evaluated the model prediction using an independent test data set that is not used in model training. The test data sets for on- and off-splicing site sSNVs include 232 and 100 pairs of disease-causing and neutral sSNVs, respectively. For VSS variants (on-splicing site variants), the MCC and AUC values using the whole feature set were 0.67 and 0.91, respectively. For VIE variants (off-splicing site variants), the MCC and AUC values were 0.47 and 0.82, respectively. To compare our algorithm with available tools focusing only on the effects of *sSNVs* on splicing outcome, but not on the structural features of alternatively spliced exons, we applied SPANR (Splicing-based Analysis of Variants), a tool for evaluating how SNVs cause splicing mis-regulation, on our independent test data set (Xiong et al. 2015). We also compared with a previously published

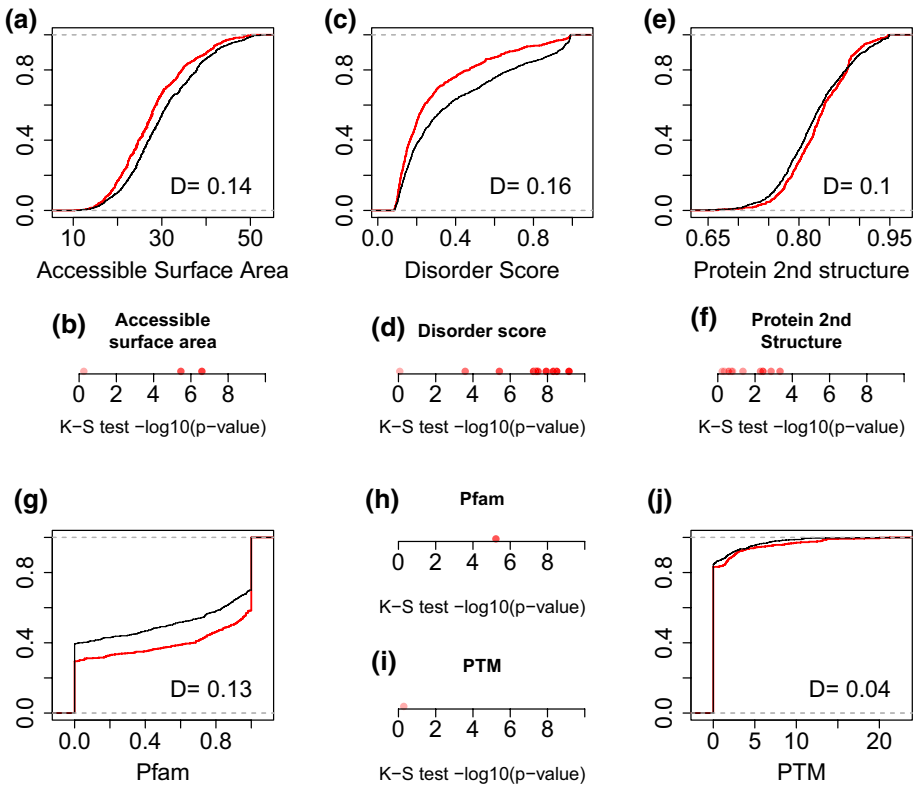

Fig. 1 Cumulative probability density function (CDF) curves and Kolmogorov–Smirnov (K–S) test *p* values on various protein structure features for the exons containing disease-causing (*red*) and neutral (*black*) sSNVs. **a** CDF of the average solvent accessible surface area (ASA) of all the amino-acid residuals in the exon. **b** KS–S test *p* values for the average, minimum and maximum ASA values of all the amino-acid residuals in the exon. **c** CDF of the average disorder score of all the residuals in the affected exon. **d** K–S test *p* values for 12 disorder score-derived features (Supplementary Table 1). **e** CDF of the average probability of the most likely protein secondary structure (alpha-helix, beta sheet, or random coil) on all the residuals in the affected exon. **f** K–S test *p* value for 12 protein secondary structure-derived features (Supplementary Table 1). **g**, **h** CDF and K–S *p* values of the percentage of the exon overlapping with known/predicted Pfam domain. **i**, **j** CDF and K–S *p* values of the normalized PTM counts in the affected exon

tool mutPred Splice (Mort et al. 2014). In our study, we used the maximum mutation-induced change in PSI across 16 tissues which is reported by SPANR by default. In addition, we used the general score reported by mutPred as an indicator of disease-causing probability. In both cases (on- and off-splicing sites), our algorithm significantly out-performed SPANR and mutPred Splice in distinguishing disease-causing and neutral variants (Fig. 2). The areas under curve (AUCs) for regSNPs-splicing and SPANR are 0.91 and 0.68 for VSS variants, and 0.82 and 0.67 for VIE variants, respectively. In addition, mutPred Splice has AUC as 0.65 for VSS variants and 0.59 for VIE variants. We have also tested our algorithm on the pathogenic and benign synonymous SNVs documented in the ClinVar database. For VSS variants, similar to the test in our independent test data set, regSNPs-splicing demonstrated significantly improved performance than SPANR; AUCs for these two algorithms are 0.85 and 0.73, respectively. For the variants that are not on the splice site (VIE variants), however, the performances of the two algorithms are similar (AUCs for regSNPs and SPANR are 0.70 and 0.68, respectively). One possible explanation for this is that most benign sSNVs in the ClinVar database do not change splicing outcome; based on SPANR prediction, among 3703 benign sSNVs, only 3 (0.08%) can cause more than 20% change of splicing inclusion ($|dPSI| \geq 0.2$). Based on the rationale of the model design, regSNPs-splicing works more effective if the pathogenic variants contain substantial amount of variants that do cause splicing change, while the resultant splicing change does not cause protein function changes.

Minor allele frequency is reversely correlated with disease-causing probability

We obtained allele frequencies for all the synonymous variants from the 1000 Genomes Project data. The allele frequency in the population should, in general, reflect the occurrences of that allele with respect to its putative biological function. As expected, there was a strong negative correlation between the predicted disease-causing

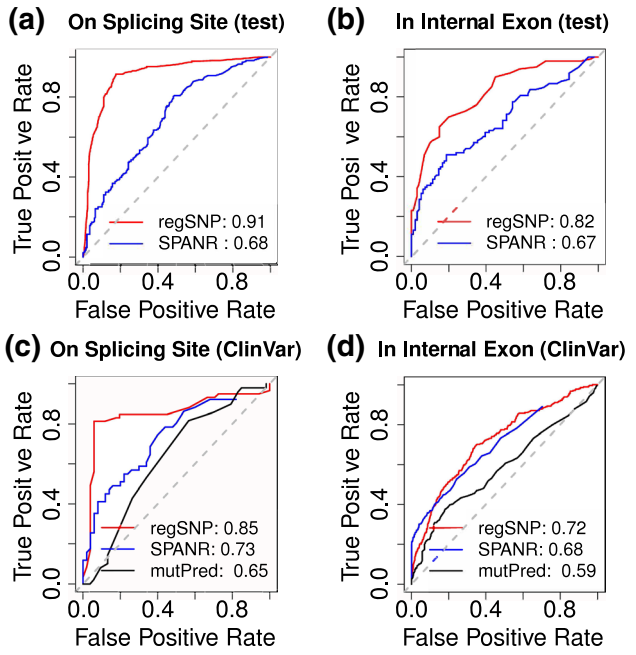

Fig. 2 Comparison between regSNP-splicing and SPANR on independent test variant data set and ClinVar variant data set. **a**, **b** ROC curves showing the performance of regSNP-splicing (*red curve*) and SPARN (*blue curve*) on an independent test data sets for VSS, and VIE variants, respectively. **c**, **d** ROC curves showing the performance of regSNP-splicing (*red curve*), SPARN (*blue curve*) and mutPred Splice (*black curve*) for VSS and VIE variants documented in the ClinVar database, respectively

probability and allele frequency for both VSS and VIE variants with correlation coefficient = 0.32 and 0.88, respectively (Fig. 3a, b).

Web-based analysis portal

We provide a Web access to our tool (http://watson.compbio.iupui.edu/regSNP-splicing/) and also a download link of the source code and database annotation files.

Discussion

The genetic code error arising from a single silent nucleotide variation can be populated through transcriptional regulation, post-transcriptional regulation, and translation process. The mis-splicing of exon in mRNA eventually will change protein structure and affect protein's function. Many disease phenotypes can be traced back to that "nucleotide switch" which triggered dramatic alteration of biological processes (Milenkovic et al. 2010; Akiyama et al. 2007; Neveling et al. 2012; Banerjee et al. 2011; Leontiou et al. 2008). To elucidate the subsequent abnormal biological process which are implemented with aberrant genetic information, several studies have investigated the genetic context of nucleotide sequence around sSNV and further extend to a comprehensive profile of affected splicing regulation elements (Kurmangaliyev and Gelfand 2008; Ward and Cooper 2010; Baralle and Baralle 2005). Despite of these efforts, the examination of protein function and structural integrity is largely dismissed yet providing a direct interpretation of disease mechanism. In addition to the genomic sequence including length of exon and intron, proximity to exon boundary, conservation, exon junction site strength, and splicing regulation motifs, we extend our scope to protein structure and function features:

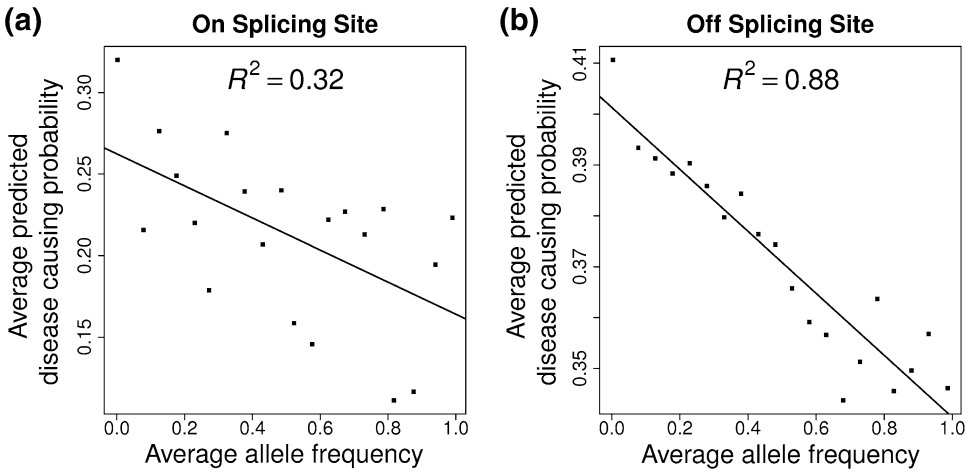

Fig. 3 Reverse correlation between average minor allele frequency (MAF) and average predicted disease-causing probability for **a** on-splicing site and **b** off-splicing site 1000 Genomes variants, respectively. Minor allele frequency, ranging between 0 and 1, is divided into 20 equal bins, and each bin represents 0.05 increment of MAF. For all the variants with MAF falling into each bin, we calculated their average MAF and average disease-causing probability values. *One dot* represents a pair of average MAF and average DCP. A linear model was fitted for the 20 dots and R^2 value is calculated

solvent accessible surface area, protein secondary structure, intrinsic disorder score, Pfam domains, and post-translation modification sites. Our analyses confirmed the conclusions from the previous studies (Scott et al. 2012; Ward and Kellis 2012; Pagani et al. 2005; Woolfe et al. 2010) that synonymous SNVs can have great influence on the splicing regulation. We have observed significant differences on a broad array of genomic features that are associated with disease-causing and neutral sSNVs, respectively. Importantly, our results strongly suggest that the protein structure features offer an added dimension of information while distinguishing disease-causing and neutral synonymous variants. The inclusion of structural features increases the predictive accuracy for functional *sSNV* prioritization.

In our research, we specifically split our data set into two parts: one set of sSNVs defined as "on consensus splicing site" and it contains sSNVs which are very close (within 3 nucleotides to donor site or one nucleotide to acceptor site). In addition, the other set of sSNVs are defined as "variants in internal exon" which contains sSNVs that are >3 nucleotides from donor site and >1 nucleotide from acceptor site. For the data set "on consensus splicing site", we did not use the proximity to exon boundary as a feature for model training, since it is a very strong indicator to distinguish whether a synonymous mutation is disease-causing or neutral. Therefore, our model for the variant on splicing site does not learn the position of sSNVs. For the data set "in internal exon", we examined the proximity from the location of sSNV to the closer splicing junctions and we can tell that proximity is a useful feature but not overwhelming (plot h, Fig. S4). Therefore, for the data set "in interval exon", our model does not just learn the position of sSNV either.

To evaluate the performance of our tool in predicting disease-causing mutations, we have compared our tool with SPANR and mutPred Splice (Mort et al. 2014). SPANR and mutPred Splice were primarily designed to quantify splicing level change of one exon in the presence of single-nucleotide variation. Although SPANR and mutPred Splice do not have a specific disease focus, the splicing level change is a strong indicator of disease relevance (Ward and Cooper 2010). Therefore, both tools are capable of predicting disease-causing mutations based on same rationale.

As a direct evaluation of the importance of protein feature, we evaluated the power of protein level information separately. We divided all the features into three categories: DNA evolution, splicing regulation features, and protein function and structure features. For every category, we trained and tested one single model based on tenfold cross validation. For the model built using nucleotide evolution, we plotted ROC curve, and the AUC is 0.56 for VSS variants and 0.59 for VIE variants. For the model built using splicing regulation features, the AUC is 0.91 for VSS variants and 0.81 for VIE variants. For the model built using protein features, the AUC is 0.67 for VSS variants and 0.71 for VIE variants. Therefore, for both VIE and VSS variants, protein function and structure features have been demonstrated of strong classification power (Fig. S9).

To make our model independent of gene, we have strictly controlled sequence similarity in our model training and testing process. In the original HGMD and 1000 Genomes Project data set, there are scenarios that multiple variants originate from the same exon. To avoid over-representation of such exons and genes in our models, we kept only one variant per exon in both training and testing data. Furthermore, we performed a more strict control of sequence similarity in our training and testing data set. For each gene family, we only selected the gene with most number of exons and then we only keep variants in this gene. And then, we trained our models using the remaining data. Using tenfold cross-validation strategy, the model for VSS has achieved AUC as 0.93 and the model for VIE has AUC as 0.84 (Fig. S10). We demonstrated that our models were not significantly affected by sequence similarity.

In our current model, the structural information on the potential disease-causing isoform is not calculated. The first reason is that we believe that evaluating the structural information on the naturally occurring splicing events has provided enough sensitivity to our approach. In addition, more importantly, for practical purpose, calculation for structural information based on amino-acid sequences is very time-consuming. To make the tool usable to general public, most of the protein features are pre-calculate based on the current gene annotation.

However, our models also have some limitations. One of the limitations is that our training data size of HGMD is not large enough. This can be improved with the growth of the databases of HGMD and ClinVar in the future. Another limitation is that the protein structure features are all prediction based. This is reasonable, otherwise using known protein structures information in PDB database will limit the training data set size and impose difficulty on implementation of both model training and testing. However, this would add another level of inaccuracy.

Methods

Training data sets

A "gold standard" data set for the development of machine learning-based prediction algorithms was obtained from the human gene mutation database (HGMD) and the 1000 Genomes Project. The positive training set was acquired from the HGMD database, which contains 1373 disease-causing synonymous SNVs (sSNVs) that have

experimentally been verified to cause disease through affecting the processes of alternative splicing. We further removed the variants reside in the first exon and last exon of a gene, whose inclusion/exclusion status is often regulated through mechanisms other than splicing regulation, such as alternative promoter, or alternative polyadenylation. The positive and negative data sets may be biased by repetitively appeared genes and exons which can introduce highly homologous sequences into our training data sets. To avoid the over-representation of certain exons due to the occurrences of multiple disease-causing variants, we only keep one variant per exon. Therefore, one exon is used only once for training and testing purposes. The remaining 980 disease-causing sSNVs were further classified into two groups, the one locating at +1/+2/+3 nucleotides on the donor side, and −1 nucleotide on the acceptor side is considered as on consensus splice site variants (VSS), while the other variants are considered as in internal exon (VIE). We have also removed the variants appearing in the ClinVar database for training purpose. This classification resulted in 651 VSS and 329 VIE disease-causing sSNVs, respectively.

The negative training set, i.e., "neutral" synonymous variants, was acquired from the 1000 Genomes Project, in which genotyped individuals did not exhibit any apparent disease phenotypes. To minimize false positives, we selected only those sSNVs with a minor allele frequency (MAF) greater than 10% for VIE variants, and 3% for VSS variants. The reduced MAF cutoff for VSS variants is implemented due to the limited available number of on splicing site sSNVs in the 1000 Genomes database. This selection criterion resulted in 7231 neutral VIE variants, and 329 VSS variants, respectively.

In addition to the HGMD database, the synonymous variants documented in the ClinVar database were used as test data set for model evaluation. The current ClinVar database contains 4765 synonymous variants, of which 230 and 4535 are pathogenic and benign, respectively. To avoid potential evaluation bias due to the overlapping records between HGMD and ClinVar, we have removed the overlapping variants from the training data set. Although the total number of usable sSNVs in ClinVar database is limited, it offers the opportunity to validate the prediction accuracy from an independent test data set.

Feature description

We evaluated a broad array of features that can be classified into three major categories: genomic features characterizing how individual SNVs affect splicing regulation, structural features evaluating how the inclusion/exclusion of alternatively spliced exons affect protein function, and others (such as conservation). A detailed list of features can be found in Supplementary Table 1.

Genomic features

Potential impacts of sSNVs on the binding affinities of RNA-binding proteins (RBPs) For a given sSNV, its effect on the binding of a particular RBP (RNA-binding protein) will be evaluated by the differences in the RBP-binding scores between reference and alternative alleles. The RBP-binding score was calculated based on the RNA sequence and the RBP position weight matrix (PWM) documented in the RBPDB and cisBP-RNA databases; collectively, these two databases contain the PSSMs of 201 RNA-binding proteins (Ray et al. 2013; Cook et al. 2011). A PWM is a matrix of values that gives the count of each nucleotide at each locus of the binding site. The binding affinity between the n-nt RNA sequence, and the PWM is described by a matching score S as follows:

$$S = \sum_{i=1}^{k} \sum_{j \in \{A,T,G,C\}} \log_2 \frac{(n_{ij} + c_{ij})/(N + \sum_{j=1}^{4} c_{ij})}{d_j}, \quad (1)$$

where $n_{i,j}$ is the count of the jth nucleotide on the ith position in one PWM, $c_{i,j}$ is the pseudocount for the jth nucleotide on the ith position in the PWM, and d_j is the prior base frequency for the jth nucleotide ($d_j = 0.25$ for j = A, T, G, C). N is the total number of experimentally validated binding sites for one RBP, and k is the width of the binding site.

In Eq. (1), a high or low matching score indicates that the putative sequence has, respectively, a high or low likelihood to be a potential binding site. Each position of a binding site is assumed to be independent of the other. The matching score distributions for binding and non-binding events were both estimated based on PSSM of an individual RBP. We assume that the matching score follows a Gaussian distribution, with mean as M_s and variance as V_s. The mean and variance of the binding scores for specific RBP-binding events are defined as follows:

$$s_{ij} = \log_2 \frac{(n_{ij} + c_{ij})/(N + \sum_{j=1}^{4} c_{ij})}{d_j}, \quad (2)$$

$$M_s = \sum_{i=1}^{k} \sum_{j \in (A,T,G,C)} f_{ij} \times s_{ij}, \quad (3)$$

$$V_s = \sum_{i=1}^{k} \sum_{j \in \{A,T,G,C\}} f_{ij} \times s_{ij}^2 - (f_{ij} \times s_{ij})^2. \quad (4)$$

In Eq. (2), the score $s_{i,j}$ is the value of the ith column and the jth row of the position specific score matrix (PSSM), which is defined as the logarithmic ratio of the percentage of the jth nucleotide (A, C, G, or U) in column i of the binding sites to the percentage in random

sequence. In this equation, $n_{i,j}$ is the count of the jth nucleotide on the ith position in the PWM, $c_{i,j}$ is the pseudocount for the jth nucleotide on the ith position in the PWM. N is the total number of experimentally validated binding sites for each RBP. d_j is the prior base frequency for the jth nucleotide ($d_j = 0.25$ for j = A, T, G, C).

In Eqs. (3) and (4), $f_{i,j}$ is the approximation of the true frequency of each nucleotide at each binding locus. For binding events,

$$f_{ij} = \frac{2^{s_{ij}}}{4}, \tag{5}$$

and for non-binding events, $f_{i,j} = 0.25$.

As defined in our previous study on transcription factors and micro-INDELs, the magnitude (M) of a sSNV affecting the binding of an RBP is defined as a likelihood ratio of the sSNV affected loci being a binding event as opposed to it being a non-binding event in reference and alternative forms, respectively:

$$\begin{aligned}
P &= P\left(R = B, A = NB | S_R, S_A\right) + P\left(R = NB, A = B | S_R, S_A\right) \\
&= \frac{P(R = B, A = NB) P\left(S_R, S_A | R = B, A = NB\right)}{P\left(S_R, S_A\right)} \\
&\quad + \frac{P(R = NB, A = B) P\left(S_R, S_A | R = NB, A = B\right)}{P\left(S_R, S_A\right)} \\
&= \frac{P(R = B) P(A = NB) P(S_R | R = B) P\left(S_A | A = NB\right)}{P(S_R) P\left(S_A\right)} \\
&\quad + \frac{P(R = NB) P(A = B) P(S_R | R = NB) P\left(S_A | A = B\right)}{P(S_R) P\left(S_A\right)} \\
&= \frac{P(B)(1 - P(B))[P(S_R | R = B) P\left(S_A | A = NB\right) + P(S_R | R = NB) P\left(S_A | A = B\right)]}{P(S_R) P(S_A)} \\
&= \int_0^1 [P(B)(1 - P(B))(P(S_R | R = B) P\left(S_A | A = NB\right) \\
&\quad + P(S_R | R = NB) P\left(S_A | A = B\right)) / P(S_R) P\left(S_A\right) d(B)
\end{aligned} \tag{7}$$

where R and A indicates whether a genetic locus is in reference or alternative form, B means "binding event" and NB means "non-binding event". Therefore, $R = B$ means that the genetic locus in its reference status is a binding site of RBP, and vice versa. Random variables B and NB

$$M = \log_2 \frac{P(S_A|B)/P(S_A|\mathrm{NB})}{P(S_R|B)/P(S_R|\mathrm{NB})} = \log_2 \left(\frac{\int_{-\infty}^{S_A} \frac{1}{\sqrt{2\pi} V_S} e^{-\frac{1}{2}\left(\frac{x - M_S}{V_S}\right)^2} d(x) \Big/ \left(1 - \int_{-\infty}^{S_A} \frac{1}{\sqrt{2\pi} V'_S} e^{-\frac{1}{2}\left(\frac{x - M'_S}{V'_S}\right)^2} d(x)\right)}{\int_{-\infty}^{S_R} \frac{1}{\sqrt{2\pi} V_S} e^{-\frac{1}{2}\left(\frac{x - M_S}{V_S}\right)^2} d(x) \Big/ \left(1 - \int_{-\infty}^{S_R} \frac{1}{\sqrt{2\pi} V'_S} e^{-\frac{1}{2}\left(\frac{x - M'_S}{V'_S}\right)^2} d(x)\right)} \right) \tag{6}$$

where R and A indicates the reference and mutated sites, respectively; B and NB denote binding and non-binding events, respectively. S_R and S_A each represents the matching scores of the reference and mutated sites. $P(S_A|B)$ is the probability of matching score S_A of mutated site when it is a binding event, and $P(S_A|\mathrm{NB})$ is the probability of S_A when it is a non-binding event. Similarly, $P(S_R|B)$ is the probability of matching score S_R of reference site when it is a binding event and $P(S_R|\mathrm{NB})$ is the probability of matching score S_R for non-binding event. M_S and V_S are, respectively, the mean and variance of the matching score for binding events, and M'_S and V'_S are the mean and variance of the matching score of non-binding events. A positive M score indicates a gain of an RBP-binding site, whereas a negative M score indicates the loss of an RBP-binding site.

We further calculate a Bayesian-based posterior probability for RBP-binding-site gain/loss, defined as the probability that a genetic locus could switch between binding and non-binding status, with and without the synonymous variant:

are both assumed to follow beta distribution. We assume that R and A are independent of each other and S_R and S_A are also independent of each other. $P(B)$ is the prior probability that a genomic region is a binding site for a RNA-binding protein. We also assume that the distribution of random variable B has mode value as 0.05. S_A denotes the matching score for alternative form and S_R denotes the matching score for reference form. We integrated over B to get the overall probability that one sSNV has changed the status of a genomic region from binding status to non-binding status, or from non-binding status to binding status.

RNA secondary structure features on the variant loci RNA-binding protein binding has well-established preference on specific RNA secondary structures; some proteins tends to bind on single-stranded regions, others double-stranded regions. Such preference may provide additional specificity for RBP binding. In addition, single-nucleotide changes may disrupt the overall RNA secondary structure on the RBP-

binding sites, and further affect RBP-binding affinity. For a specific sSNV, we calculated the average single-strandness probability for the nucleotides upstream and downstream 7 bases of variant locus (putative-binding sites), which is calculated using RNAfold (Lorenz et al. 2011). The changes on the RNA secondary structure caused by sSNV on the putative binding sites are calculated using RNAdistance (Lorenz et al. 2011).

Inherent strength of 5′- and 3′-splicing sites of exons containing candidate sSNVs We previously reported that the sSNVs residing in the AS exons are more likely to have phenotypic consequences (Teng et al. 2011). We therefore evaluated the inherent strength of 5′- and 3′-splicing sites of exons containing candidate sSNVs. This measurement may serve as an important feature for quantifying whether the candidate splicing events require additional assistance from other RNA-binding proteins; more assistance from RBPs may be needed for an exon with weaker junction strength. The inherent splicing strength of 5′- and 3′-splicing sites are calculated based on the position weight matrices (PWMs) describing the sequence features on/around canonical splicing sites (Itoh et al. 2004).

ESE/ESS cluster scores Exonic splicing motifs which consist of 6 nucleotides within an exon are categorized as exon spicing enhancer (ESE) or exon splicing silencer (ESS) based on whether they promote or prohibit splicing process, respectively. We scanned the affected exons and search for occurrences of known or predicted ESE and ESS motifs. For this purpose, we have collected 76 known motifs, 2298 predicted ESE motifs, and 1195 predicted ESS motifs (Barash et al. 2010; Fairbrother et al. 2002; Zhang and Chasin 2004). Overlapping motifs are combined and further defined as a 'motif set'. Multiple motif sets which are located within 6 bp apart are defined as a 'motif cluster'. Within a motif cluster, a gap $\leq$3 bp is denoted as a short interval—and a gap larger than 3 bp and less than or equal to 6 bp is denoted as a long interval. The total number of occurrence of short intervals is denoted as I_s and the total count of long intervals is denoted as I_l. Then, a motif set is scored as $S_{set} = 2^{N_m - 1}$, where N_m is the number of overlapping motifs. A motif cluster is scored based on I_s, I_l and number of motif sets within a motif cluster, where $S_{cluster} = (2 \cdot I_s + I_l + \#\text{motif set}) + S_{set}$. S_{set} is defined to measure the local density of splicing motifs within an exon, and $S_{cluster}$ is measuring the aggregation of motif sets. Finally, the enrichment of ESE and ESS within an exon using a 'cluster score' is defined as follows:

$$\text{Cluster score} = \frac{\log^{S_{cluster}}}{\text{exon length}} \quad (8)$$

where exon length's unit is per 100 base pairs. The effect of a specific SNV on ESE/ESS clustering is evaluated based on the differences of cluster scores for the reference and alternative alleles, respectively.

Proximity to the 5′- and 3′-splicing junction The proximity is defined as the distance from a variant and exon boundaries. Here, we separate our SNVs into two different categories: on splicing site (within 3 bp of donor sites or 1 bp of acceptor site) and off-splicing site (the other regions of exon). SNVs on splicing site mainly interferes with various molecular and affect formation of spliceosome, and the off-splicing site SNVs mainly affect the binding of splicing regulators. This feature is not used for on-splicing site variants.

Protein structure/function features

Disruption of protein secondary or tertiary structures is one possible reason for deleterious alternative splicing events. We, therefore, evaluated several features describing the effects of affected splicing pattern on protein structures. Such features include protein structure/intrinsic disorder scores, solvent accessible surface areas (ASA), protein secondary structures, and known and predicted protein family domains (pfam). In addition, the known post-translational modification status on the affected splicing event is also evaluated.

Intrinsic disordered regions Intrinsically disordered regions are defined as a stretch of amino-acid sequences that lack the ordered tertiary and/or secondary structures. We have previously reported applying disorder score of affected protein regions in distinguishing disease-causing and neutral micro-insertion and -deletions (Zhang et al. 2014). Similarly, we measured the disorder property of affected protein regions that result from the mis-spliced exon in transcript. Disorder property of the affected region is quantified through calculating the disorder score of each involved amino acid using spine-D (Zhang et al. 2012).

Solvent accessible surface areas (ASA) Solvent accessible surface area has been used as an important feature for variant prioritization (Folkman et al. 2015; Zhao et al. 2013). Based on ASA value, an amino acid can be classified as buried inside or on the surface of a protein. To some degree, ASA can be used to infer the flexibility and predict binding induced structure conformational change of monomeric proteins (Marsh and Teichmann 2011). The ASA value for the affected exon is calculated using Spline-X with default parameters (Faraggi et al. 2009).

Protein secondary structure The most probable secondary structure (alpha-helix, beta sheet, or random coil) on the affected exons are calculated using Spline-X (Faraggi et al. 2009) using default parameters.

Overlapping with known or predicted protein family domains (Pfam) The functional regions of proteins are generally termed as domains. The direct consequence of abnormal splicing is loss or gain of one or more protein domains due to missing or addition of a fragment of protein sequence. The integrity of protein function is determined by the combination of domains and therefore abnormal splicing directly affects protein's function. We have collected in total 86,748 high quality Pfam-A protein families (49,991 domains, 28,062 families, 703 Motifs, and 7992 repeats) from Pfam database (Finn et al. 2014). As a measurement of the importance of affected exon on protein domains, we calculated a percentage value as the proportion of affected protein region which overlaps with documented Pfam domains.

Post-translational modification sites (PTMs) Post-translational modifications on amino acids play an important role determining the function and activities of a protein. To evaluate the potential PTM status of the exons containing functional sSNVs, We downloaded 372,456 experimentally verified PTM sites from dbPTM 3.0 database (Lu et al. 2013). Among those PTMs, most common modifications are phosphorylation, ubiquitylation, and acetylation. As a comparison on the density of PTM sites, we calculated the normalized PTM site amount per 100 amino acids on the affected protein region.

Machine learning model

We discovered the excellent classification capability of random forest in our previous study. In addition, in this study, we continued to use random forest as the tool to learn the distinct genomic and protein structural and functional features between disease-causing and neutral variants. Random forest is composed a certain number of decision trees and the final prediction is the polled vote of each tree's prediction result. For training purpose, random forest algorithm randomly selects (bootstrap) a proportion of training samples for growing each node. The feature for each node is selected from a subset of features bootstrapped from the total set of features, based on a certain split criterion such as information gain or Gini index. In our study, we used a software package called Weka to build our random forest model (Witten et al. 2016). We did not implement feature selection before training our model due to the bootstrap step in both selecting training sample and selecting features when building each node. We tuned the number of trees to grow for random forest as 51 and the number of features subset for building each node as 35. Two different models are trained independently for variants on splicing sites and off-splicing sites, respectively.

Acknowledgements This work was supported in part by the US National Institutes of Health R01CA21346 (to YL), R01GM11847 (to YL), the National High-Tech Research and Development Program (863) of China 2015AA020101 (to YL and YW), and the National Health and Medical Research Council (1059775 and 1083450) of Australia (to YZ). In addition, MM and DNC acknowledge the financial support of Qiagen Inc through a License Agreement with Cardiff University.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Akiyama M, Titeux M, Sakai K, McMillan JR, Tonasso L, Calvas P, Jossic F, Hovnanian A, Shimizu H (2007) DNA-based prenatal diagnosis of harlequin ichthyosis and characterization of ABCA12 mutation consequences. J Invest Dermatol 127:568–573

Banerjee I, Skae M, Flanagan SE, Rigby L, Patel L, Didi M, Blair J, Ehtisham S, Ellard S, Cosgrove KE et al (2011) The contribution of rapid KATP channel gene mutation analysis to the clinical management of children with congenital hyperinsulinism. Eur J Endocrinol 164:733–740

Baralle D, Baralle M (2005) Splicing in action: assessing disease causing sequence changes. J Med Genet 42:737–748

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. Nature 465:53–59

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

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

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR (2011) RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 39:D301–D308

Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet 12:205–216

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297:1007–1013

Faraggi E, Yang Y, Zhang S, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17:1515–1527

Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. Genes Dev 17:419–437

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

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

Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Itoh H, Washio T, Tomita M (2004) Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. RNA 10:1005–1018

Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S (2013) Function of alternative splicing. Gene 514:1–30

Kurmangaliyev YZ, Gelfand MS (2008) Computational analysis of splicing errors and mutations in human transcripts. BMC Genomics 9:13

Leontiou CA, Gueorguiev M, van der Spuy J, Quinton R, Lolli F, Hassan S, Chahal HS, Igreja SC, Jordan S, Rowe J et al (2008) The role of the aryl hydrocarbon receptor-interacting protein gene in familial and sporadic pituitary adenomas. J Clin Endocrinol Metab 93:2390–2401

Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK (2016) RNA splicing is a primary link between genetic variation and disease. Science 352:600–604

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:26. doi:10.1186/1748-7188-6-26

Lu CT, Huang KY, Su MG, Lee TY, Bretana NA, Chang WC, Chen YJ, Chen YJ, Huang HD (2013) DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res 41:D295–D305

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:1624–1627

Marsh JA, Teichmann SA (2011) Relative solvent accessible surface area predicts protein conformational changes upon binding. Structure 19:859–867

Milenkovic T, Zdravkovic D, Savic N, Todorovic S, Mitrovic K, Koehler K, Huebner A (2010) Triple A syndrome: 32 years experience of a single centre (1977–2008). Eur J Pediatr 169:1323–1328

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:R19

Neveling K, Collin RW, Gilissen C, van Huet RA, Visser L, Kwint MP, Gijsen SJ, Zonneveld MN, Wieskamp N, de Ligt J et al (2012) Next-generation genetic testing for retinitis pigmentosa. Hum Mutat 33:963–972

Pagani F, Raponi M, Baralle FE (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci USA 102:6368–6372

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499:172–177

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M et al (2015) Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348:666–669

Sauna ZE, Kimchi-Sarfaty C (2011) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12:683–691

Scott A, Petrykowska HM, Hefferon T, Gotea V, Elnitski L (2012) Functional analysis of synonymous substitutions predicted to affect splicing of the CFTR gene. J Cyst Fibros 11:511–517

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133:1–9

Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B (2014) Synonymous mutations frequently act as driver mutations in human cancers. Cell 156:1324–1335

Teng M, Wang Y, Wang G, Jung J, Edenberg HJ, Sanford JR, Liu Y (2011) Prioritizing single-nucleotide variations that potentially regulate alternative splicing. BMC Proc 5(Suppl 9):S40

Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E et al (2014) Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505:706–709

Ward AJ, Cooper TA (2010) The pathobiology of splicing. J Pathol 220:152–163

Ward LD, Kellis M (2012) Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol 30:1095–1106

Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques, 4th edn. Morgan Kaufmann

Woolfe A, Mullikin JC, Elnitski L (2010) Genomic features defining exonic variants that modulate splicing. Genome Biol 11:R20

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR et al (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347:1254806

Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18:1241–1250

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. J Biomol Struct Dyn 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. Hum Mol Genet 23:3024–3034

Zhao H, Yang Y, Lin H, Zhang X, Mort M, Cooper DN, Liu Y, Zhou Y (2013) DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol 14:R23

Zheng S, Black DL (2013) Alternative pre-mRNA splicing in neurons: growing up and extending its reach. Trends Genet 29:442–448