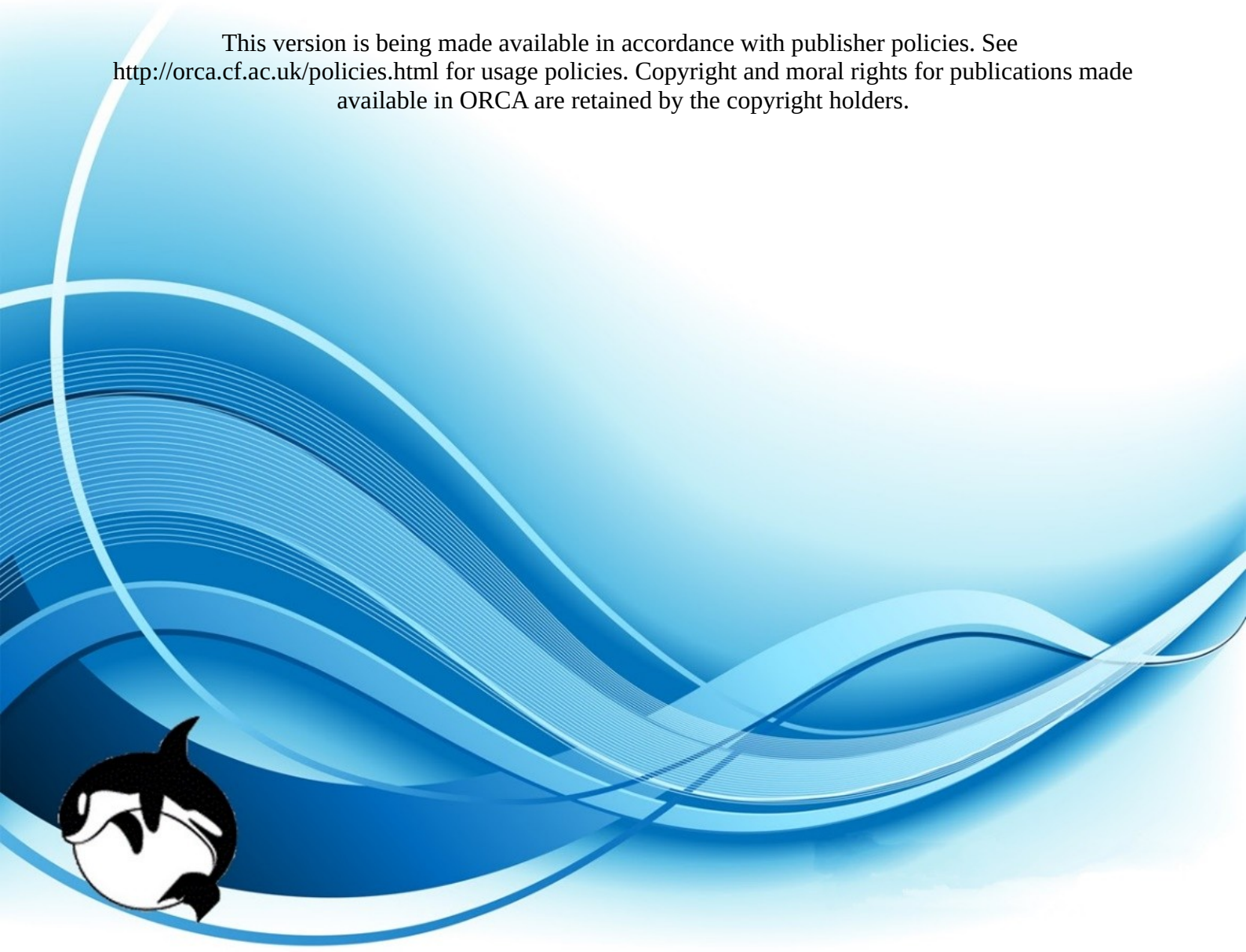

ExonImpact: Prioritizing pathogenic alternative splicing events

Meng Li[1], Weixing Feng[1*], Xinjun Zhang[2,3], Yuedong Yang[4], Kejun Wang[1], Matthew Mort[5], David N Cooper[5], Yue Wang[6], Yaoqi Zhou[4], and Yunlong Liu[3,6*]

[1]Institute of Intelligent System and Bioinformatics, College of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, China.

[2]School of Informatics and Computing, Indiana University, Bloomington, IN 47408, USA

[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

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

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

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

*Corresponding authors

Email addresses:

ML: li487@outlook.com

WF: fengweixing@hrbeu.edu.cn

XZ: zhangxin@indiana.edu

Formatted

Field Code Changed

Field Code Changed

YY: yuedong.yang@griffith.edu.au

KW: wangkejun@hrbeu.edu.cn

MM: mortm@cardiff.ac.uk

DNC: cooperdn@cardiff.ac.uk

YW: yuewang@iu.edu

YZ: yaoqi.zhou@griffith.edu.au

YL: yunliu@iu.edu

Field Code Changed

Field Code Changed

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

ABSTRACT

Alternative splicing (AS) is a closely regulated process that allows a single gene to encode multiple protein isoforms, thereby contributing to the diversity of the proteome. Dysregulation of the splicing process has been noted to be associated with many diseases. However, in amongst the pathogenic AS events there are numerous "passenger" events whose inclusion or exclusion does not lead to significant changes with respect to protein function. In this study, we evaluate the ~~protein~~ secondary and tertiary structural features of proteins associated with disease-causing and neutral AS events, and show that several ~~protein~~ structur~~ale~~ features are strongly associated with the pathological ~~relevance~~impact of exon inclusion. We further develop a machine learning-based computational model, ExonImpact, for prioritizing and evaluating the functional ~~impact~~ consequences of hitherto uncharacterized AS events. We evaluated our model using several strategies including cross-validation, and the data from the Gene-Tissue Expression (GTEx) and ClinVar databases. ExonImpact is freely available at http://watson.compbio.iupui.edu/ExonImpact

Introduction

Alternative splicing is a tightly regulated process by which exons in the pre-mRNA transcripts are differentially joined or skipped, thereby allowing a single gene to encode multiple protein isoforms. As an important level of gene regulation that greatly contributes to proteome diversity, alternative splicing is a widespread phenomenon in eukaryotic cells, and plays key roles in a variety of biological processes, including cell division, cell fate decisions, tissue maturation and cellular responses to the changes of extracellular environment, such as responses to stress [YUNLONG: Review reference?]. Importantly, many studies have suggested that dysregulation of the alternative splicing process is associated with a wide variety of human genetic diseases[1]. Therefore, investigating the functional effects of individual alternative splicing events is crucial to understanding the complexity of biological systems and the molecular mechanisms of human disease.

Recent advances in the development and implementation of high-throughput sequencing technology have enabled the profiling of pre-mRNA splicing patterns on a genome-wide scale. Many alternative splicing events can be identified from RNA-seq data using existing bioinformatics tools such as Mix-of-Isoforms (MISO)[2], Multivariate Analysis of Transcript Splicing (rMATS)[3], SpliceTrap[4], and ASprofile[5]. Despite these advances, very few tools are actually available for evaluating the functional impact~~s~~ of individual alternative splicing events on protein function. Indeed, several studies suggest that only a subset of alternative splicing events have an influence on protein function[6]. In addition, our previous work on non-frameshifting INDELs (insertions /deletions) showed that the addition or omission of a limited number of amino acid residues does not guarantee a functional change, unless the residues occur within key structural elements of the protein[7]. Therefore, a systematic strategy to prioritize functional alternative splicing

events will be essential for biologists to identify important candidates for further mechanistic investigation.

Several studies have attempted to select the most influential events based on the genomic and protein structure features in the vicinity of the alternatively spliced exons. For instance, Lu et al.[6] designed a prediction method by combining multi-genome alignment data with RNA selection pressure calculations, which were based on the number of synonymous variants within the alternatively spliced exons and flanking intronic regions. This relieds upon the assumption that such regions are enriched with binding sites for splicing regulatory factors, and therefore should be depleted of variants responsible for functionally important events. Other studies have evaluated the potential impact of individual alternatively spliced exons on protein function by examining their overlap with known protein functional domains, post-translational modification sites, and binding domains that facilitate protein-protein interactions. It has been observed that the proteins containing tissue-specific alternative exons tend to have more a greater number of interactions in the protein-protein interaction (PPI) networks, suggesting their widespread role in controlling the tissue-specific dynamics of the protein interactions[8].

Prompted by these studies, we systematically evaluated the protein structural features that are related to the exons whose splicing outcomes are associated with human genetic disease. By comparing the exons containing disease-causing variants documented in the Human Gene Mutation Database[9] and the exons containing putatively non-harmful insertions/deletions in the 1000 Genomes Database[10], we confirmed that several features relating to protein structure are strongly associated with the disease-relevance or otherwise of the exon inclusion status. We further developed an algorithm, ExonImpact, to predict the disease relevance of the alternatively spliced

exons that have not been previously investigated. The model was further evaluated by a subset ~~from~~ of an independent test dataset from our own data set [YUNLONG: can we simplify this sentence by removing at least one mention of the word 'set'?], splicing patterns derived from the GTEx (Genotype-Tissue Expression) database[11] and the ClinVar[12] database, which document~~s~~ both the benign and pathogenic mutations from clinical testing, literature curation and population studies. Our results suggest that ExonImpact can reliably identify disease-causing events, derived from various genome-wide studies, located within hundreds of alternatively spliced exons. The schematic of the overall study is illustrated in Figure 1.

Results

Disease-causing AS events have distinct protein structure features

In order to examine what differentiates the pathogenic and neutral AS events, we systematically evaluated dozens of protein structure features that are potentially associated with ~~the~~ alternatively spliced exons. Features were clustered into five categories, each with slightly different measurements (Supplementary Table 1). The five categories included solvent accessible surface area (ASA), protein secondary structure, probability of intrinsic disordered region (disorder score), relationship with known protein family domains, and the presence of known post-translational modification (PTM). In addition, the evolutionary conservation scores and the length of the alternatively spliced region were also included.

Disease-causing AS events tend to locate in structured protein regions

Intrinsically disordered protein (IDP) regions are stretches of amino acid residues that lack a fixed or ordered three-dimensional structure. Compared to structured proteins, IDPs tend to have distinct properties in terms of function, structure, sequence, interactions, evolution and regulation. The disorder measurement is calculated using SPINE-D[13], an algorithm that we previously developed to predict disorder probability based on a protein's amino acid sequence. For each amino acid residue, SPINE-D reports a disorder score ranging from 0 to 1, where 0 and 1 indicate that the residue in question locates fully within structured and disordered regions, respectively. We define 12 features to characterize the disorder status of the AS exon (Supplementary Table 1); one example of the 12 features can be found in Supplementary Figure 1. Each feature is evaluated based on its power to discriminate the AS events in the positive (disease-causing) and negative (neutral) training sets. All the 12 disorder-related features showed different distributions between the positive and negative training sets (Kolmogorov-

Smirnov test) with most features exhibiting p-values < $1x10^{-16}$ (Figure 2C). The probability density function (PDF) and cumulative distribution function (CDF) of one representative disorder feature, the maximum disorder score of all the amino acids, are shown in Figures 2A and 2B. As with our earlier observation on small exo-INDELs[7] , pathogenic AS events showed significantly lower disorder scores, which suggests that they tend to be located within structured protein regions.

Disease-causing AS events tend to be buried inside core regions of protein structures

Solvent accessible surface area (ASA) is a measurement of the surface area of a protein that is accessible to a solvent. A smaller ASA value usually indicates that an amino acid residue is buried within the core regions of the protein structure, whereas a larger ASA value suggests that it is exposed on the protein surface. For each amino acid residue, we calculated its ASA value using SPINE-X[14] . For each exon, three ASA-related measures, average, minimum and maximum ASA values across all residues in the exon, were used to assess the differences between disease-causing and neutral AS events. As with protein disorder measurements, all three features showed statistically significant differences ~~on~~ in terms of the distribution between the two groups of exons (K-S test), while the average ASA of an exon's translated amino acids yields the largest K-S statistic (D-value=0.258, p-value<$1x10^{-16}$). Figure 2B clearly demonstrates that the pathogenic AS events have lower average ASA values, and therefore tend to be buried inside the core protein structures.

Protein secondary structures

In addition to protein tertiary structure, secondary structures serve to define the distinct characteristic local structural conformation of the proteins, and ~~therefore~~ hence also play

important roles in protein function. In order to evaluate how the secondary structure states of the AS exons affect their disease-causing potential, we used SPINE-X to predict the three predominating states, helix, sheet and random coil, for all the disease-causing and neutral AS events documented in our training data set. As with protein disorder and ASA measurement, for each amino acid residue in the exon, three scores were calculated indicating the probability of each of the three states (helix/sheet/coil) being adopted. For each AS event, 12 features were derived from the three probability scores for every amino acid residue encoded by the exon, including the maximum, average and minimum probabilities of all the residues associated with helix, sheet or coil states. In addition, the maximum, average and minimum probabilities of all the residues associated with the most probable secondary structure state are also included. We observed strong relationships between an exon's secondary structure state and its disease association. All 12 secondary structure-derived features exhibited different distributions between disease-causing and neutral AS events, with significant p-values (K-S test, Figure 2C), but moderate D-values, ranging from 0.05 to 0.145. This suggests that secondary structures alone provide insufficiently strong discriminant power to predict the disease relevance of specific exons.

We further examined whether protein secondary structures might provide additional information in separating pathogenic and neutral AS events in combination with tertiary structural features. We found that random coil structures on the protein surface area (with high ASA scores) are strongly associated with deleterious AS events (Figure 3A). Similar results were observed with the intrinsic disordered regions. Since random coil structures are often present at key protein-protein and protein-DNA/RNA interaction domains[15], disrupting such regions can lead to detrimental phenotypic consequences. Taking the *SPAST* (spastin) gene as an example (Figure 3B), in the HGMD database,

the mutation at chr2:32339706 locus is listed as altering the regulation of splicing ~~regulation~~ of the fifth exon (NM_014946:227:290), and leads to a disease called spastic paraplegia[16]. According to UniProtKB (Uniprot ID: Q9UBP0,NM_014946), this region of the protein encodes a random coil structure, and serves as an interaction site with microtubules. Such an observation strongly suggests that protein secondary structures complement the tertiary structure features, and should play a key role in prioritizing functional AS events.

Disease-causing AS events are enriched for known protein family domains

Aberrant alternatively spliced exons can disrupt protein function by breaking up protein family domains. For each AS event in the training set, we evaluated whether the exon overlapped with a putative protein family domain, predicted by searching its sequence against the profiles of hidden Markov models characterizing the documented protein families in the Pfam library[17]. For each exon, we calculated the proportion of predicted domains covered by the AS exon, and the proportion of the AS exon that overlapped with the predicted domains. Both features showed different distributions between the positive and negative datasets (Supplementary Figure 2), with disease-causing events displaying enriched overlap with predicted protein family domains.

Disease-causing AS exons exhibit different PTM density comparing to neutral events

Post-translational modifications (PTMs) are chemical changes on specific amino acid residues, which provide key molecular mechanisms for both diversifying and regulating the functions of proteins. Splicing variants of the exons containing the key PTM sites have the potential to influence the function and signaling of the protein. We hypothesized that pathogenic AS exons ~~will~~ would tend to contain a higher density of PTM sites since

they may be enriched in signaling elements. For each AS exon, we define PTM density as the proportion of the amino acid residues that have documented an experimentally validated PTM site. Perhaps surprisingly, we observed a higher PTM density among the neutral AS events, the opposite of what we expected. This is likely to be due to the fact that disease-causing events tend to be buried ~~inside~~ within the protein structures (with lower ASA values), whereas PTM sites tend to be located on the surface of the protein structure. ~~[YUNLONG: Have you asked whether, among the surface-located residues, those residues which are known PTM sites are more likely to be disease-associated than non-PTM sites? I would expect this to be so.]~~ We therefore performed logistic regression analysis for modeling the group of AS exon (disease-causing or neutral) by using both the average ASA and the PTM density as independent variables. The PTM density received a positive coefficient (beta=0.037) with a p-value=0.02. This suggests that higher PTM density is associated with disease-causing potential when correcting the effect of the geometric loci of the exon, i.e. among the surface-located residues, those exons containing known PTM sites are more likely to be disease-associated than non-PTM sites. Consistent with earlier results (Figure 2A), we observed a negative coefficient (beta=-0.078, p-value<$1x10^{-16}$) for the average ASA, indicating that disease-causing events tend to have smaller ASA values.

Relationships among various structural features

Many of the features being evaluated are related to the tertiary structures of the proteins, and are therefore not independent. In order to examine the correlation structures of all the features, we performed principal component analysis (PCA) on the training set containing both deleterious and neutral AS events. As shown in the biplot (Figure 4), the first two principal components account for 47.6% of the total variations. The first principal component clearly aligns with most features derived from protein tertiary structures,

including disorder scores, ASA scores, and protein family domains. Protein secondary structure-related features, however, contribute more to the second principal component. Clearly, the first principal component offers makes the largest effects incontribution toward separating the two groups of events, whereasile minimum random coil score and maximum ASA scores are the two features that point to the small group of disease-causing events in the unstructured regions.

Construction of a machine-learning model for prioritizing AS events on in relation to their disease relevance

We built a predictive model, ExonImpact, for prioritizing disease-causing AS events using the Random Forest algorithm, which composes comprises of a collection of decision trees that vote on the output (deleterious or neutral). Two thirds of the events (1,776 disease-causing and 1,776 neutral events), randomly selected from the gold-standard data set, were used to train the model, whereas the other one third (882 disease-causing and 882 neutral events) were used for model validation. Based on the prediction results from the validation data set, ExonImpact yielded 0.83 for the AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristic). Similar performance was achieved when all the experiments were replicated 100 times with random sampling from the gold-standard data set (Supplementary Figures 3A and 3B). To test if the resultant AUC could have been due to homologous genes in the training set, we used the HGNC gene family[18] to cluster the disease-causing and neutral exons respectively retaining one gene in each family. This reduced the total number of positive and negative exons to 1,914 and 651 exons, respectively. Ten-fold cross-validation based on this reduced training data set yielded an AUC of 0.835, which suggests that homologous genes did not artificially boost the model performance.

For each AS exon, the Random Forest model outputs a functional impact score (FIS) ranging from 0 to 1 that is equivalent to the posterior probability of the event falling into the disease-causing category. In order to select the cutoff FIS score with the desired statistical stringency, we calculated TPR (True Positive Rate), FPR (False Positive Rate), F1 score, and MCC (Matthews Correlation Coefficient) corresponding to different cutoffs (Figure 5C). When controlling the FPR at 0.1, the cutoff score from Random Forest model was 0.82. At this FIS cutoff score, the corresponding True Positive Rate (TPR, or sensitivity), F1 score, and Matthews Correlation Coefficient (MCC) were 0.502, 0.63 and 0.44, respectively. The FIS cutoff score for FPR at 0.05 was 0.91 with TPR=0.25, F1=0.32, and MCC=0.365.

Exons containing 1,000 Genomes variants that alter splicing regulation tend to have smaller functional impact scores

We evaluated all the common coding variants in the 1,000 Genomes Project in relation to their potential to disrupt splicing regulation using SPANR (Splicing-based Analysis of Variants)[19], a bioinformatics tool that provides a score (ΔΨ) which characterizes the variant-induced changes in the proportion of transcripts with the exon spliced in. We hypothesized that the variants with the highest impact on exon inclusion (higher ΔΨ value) should reside within the exons with lowest functional impact scores (FIS), since the genotyped individuals of the 1,000 Genomes Project did not exhibit any apparent clinical phenotypes. We focussed our analysis on the common coding variants with MAF (Minor Allele Frequency) ≥ 5% that reside in the ±20bp exon region around the splice site. As shown in Figure 5, we observed sizable differences in the FIS scores for the exons containing variants with weak (|ΔΨ|<10%), intermediate (10%≤|ΔΨ|<20%) and strong (|ΔΨ|≥20%) impacts on splicing regulation. For the exons containing variants with weak regulation impact, 16.7% of the exons have FIS scores larger than 0.91 (cutoff for

FPR≤0.05). As expected, the~~is~~ ~~percentage~~ proportion decrease~~s~~ to 8.3% and 5.4% for the exons harboring variants with intermediate and strong impact on splicing regulation, respectively. The FIS scores of the events with strongly (|ΔΨ|≥20%) and weakly (|ΔΨ|<10%) impacting~~ful~~ variants showed significant differences with p-value=0.009 (two-tailed Wilcoxon test).

Independent test with ClinVar database

We further tested the effectiveness of our algorithms on an independent data set documented in the ClinVar database[12], a public archive of reports of relationship among medically important variants and phenotypes. Among the 1,032 exons containing deleterious splicing-altering variants, 498 (48%) were predicted to be pathogenic when using the FIS=0.91 cutoff (FPR ≤ 0.05). We further examined the exons that contain benign variants at splicing junction sites. These variants are very likely to change splicing outcome, but did not lead to pathogenic phenotypes. Among the 11 exons that fitted into this category, only 2 (18%) were predicted to be pathogenic (FIS ≥ 0.91, FPR ≤ 0.05). The FIS scores for the pathogenic events are significantly higher than those for benign events (two-tailed Wilcoxon test p-value<0.02).

We further examined the differences between exons containing pathogenic and neutral INDELs that are documented in ClinVar. The current ClinVar database contains 308 and 5,554 benign and pathogenic INDELs, located in 36 and 1,652 exons respectively. In order to avoid ~~the~~ an evaluation bias due to the overlapping entries in the HGMD and ClinVar databases, we re-trained our model using the HGMD entries that were not included in the ClinVar database, and tested the prediction results on the pathogenic events derived from the ClinVar database. We found that the new model's TPR was similar to the cross-validation results using HGMD only. When using cutoff FIS=0.82,

which corresponds to a 10% False Positive Rate (FPR), 662 out of 1,652 pathogenic events (40.1%) were predicted to be functionally important. Similarly, the FIS scores on 478 events (28.9%) were larger than 0.91, which is equivalent to 5% FPR. This proportion contrasts with the exons containing benign INDELs. None of the 36 benign exons are predicted to be functionally important when selecting the FIS cutoffs based on 10% FPR. The overall distribution of the FIS scores for pathogenic and benign exons also showed significant differences (Supplementary Figure 4, two-tailed Wilcoxon test p-value<1.8×10^{-14}), with the scores for the benign group skewing to the left (lower FIS scores), whereas the pathogenic ones skew to the right (higher FIS scores).

Exons with strong FIS scores have higher inclusion ratios in human brains

In order to further evaluate the biological relevance of the predicted FIS scores, we examined whether there were differences in the exon inclusion ratios for the exons with high and low FIS scores. We hypothesized that higher inclusion ratios would be observed for the exons with high FIS scores since their functions are more likely to be essential. Using the MISO (Mixture of Isoforms) algorithm, we evaluated the inclusion ratios (percent-spliced-in, ψ) of 42,485 previously documented skipped exon (SE) events on 310 RNA-seq samples from 11 brain regions of 44 individuals, collected in the Genotype-Tissue Expression (GTEx) database. Among the 42,485 SE events, measurements of the ψ values on 1,909 events were reliably identified (the confidence interval for ψ value, CI$\leqslant$0.1) in no fewer than 10 RNA-seq samples. Among these events, 1,852 events located to within the middle exons of protein coding genes. FIS scores for these events were calculated using the ExonImpact model. All events were categorized into three groups based on their FIS scores, FIS<0.82 (FPR>0.05) 0.82$\leqslant$ FIS<0.91 (0.01<FPR$\leqslant$0.05), and FIS$\geqslant$0.91 (FPR$\leqslant$0.01). We found that in the FIS$\geqslant$

0.91 group, 93.8% of the events had a very high inclusion ratio (ψ>0.8), whereas this percentage dropped to 91.3% in the 0.82≤FIS<0.91 group, and to 71% in the FIS<0.82 group (two-tailed Wilcoxon test between FIS<0.82 and FIS≥0.91, p-value =0.004, Figure 6). This observation strongly supports our hypothesis that the events with higher functional impact scores are required by the organism, and therefore have much higher inclusion ratios.

Discussion

As an essential molecular mechanism that significantly contributes to proteome diversity, alternative pre-mRNA splicing selectively includes or excludes certain protein coding elements, thereby allowing the coding of proteins with distinct functionalities. With the increasing popularity of high-throughput sequencing technology in transcriptome-wide profiling, our ability to identify splicing variants has been greatly enhanced. Usually, hundreds or even thousands of alternatively spliced events can be identified when comparing different tissues[5], and dozens to hundreds of events will be discovered when comparing the same tissue at different developmental stages[20]. Despite the large number of alternatively spliced exons with different cellular status, not all the changes in the exon inclusion ratio have obvious biological consequences. The variations in many of the splicing events simply reflect the differences in cellular conditions, and will not affect the functions of the proteins produced. Our earlier study on coding exonic INDELs[7,21] also indicates that the addition or deletion of a short stretch of amino acid sequence does not guarantee pathological consequences.

In this study, we systematically examined the protein structure features that distinguish pathogenic ~~and~~ from non-pathogenic AS events. Our results suggest that most protein structure-related features, including disorder score, ASA values, and Pfam prediction, differ significantly between the two groups of events. In addition, protein secondary structures, random coil in particular, also offer additional predictive power when the tertiary structure features fail to recognize the differences. The positions of the pathogenic events in the protein may reflect the locations of the interaction domains with other molecular components, such as DNA, RNA or other proteins.

The proposed ExonImpact model uses features describing the secondary and tertiary structures of the candidate splicing events. Although such features can be acquired from various biochemical assays, such as X-ray crystallography or nuclear magnetic resonance technologies, they are only available for a small proportion of protein regions. In order to broaden the utility of our algorithm, we adopted prediction-based methods for deriving the protein structure-related features directly from a series of neural-network-based SPINE techniques that make amino acid sequence-based predictions for protein secondary structures, ASA[14], and disorder probability[13] . Such a strategy was used in our early work for prioritizing the functions of exo-INDELs[7,21]. This allows us to prioritize the events whose experimentally determined structures are not available.

For machine-learning-based prediction algorithms, the selection of training data set is critical, and may have a major influence on prediction accuracy. In this study, the positive (deleterious) and negative (neutral) events were selected from two resources. The positive events were selected from the exons containing disease-causing variants that have been determined to disrupt splicing outcome of the exon, as documented in the Human Gene Mutation Database. Selection of “negative” training sets (neutral exons) was more challenging, since there is no existing database documenting such events. We therefore selected exons containing micro-INDELs from the 1000 Genomes database as our putatively neutral data set. Since these INDELs were identified in apparently healthy individuals, the addition or removal of a few amino acids in these exons should not have generated deleterious phenotypes [YUNLONG: There is a major difference between being deleterious to health and being deleterious to protein function. The latter does not imply the former. Is it worth mentioning that we are nevertheless aware of this distinction?]. Since INDELs are in general shorter than the exons, we extended the “neutral” region to the entire exon. We consider that this exon is located

within a functionally dispensable region. It is conceivable that the sequences outside of the INDEL regions (non-frameshift INDELs in particular) in these exons are functionally important. This will generate false negatives in our training set, thereby reducing the accuracy of the model's prediction. This implies that our negative training set may contain data that should be part of the positive group, and if this is so, it may reduce the discriminant capacity of some key features between the two groups. Such inaccuracy, however, will potentially decrease the discriminant powers of the features, and therefore make the prediction more conservative.

In order to increase the usability of our model, we provided two means for using ExonImpact, both through a web server (http://watson.compbio.iupui.edu/ExonImpact/), and a downloadable version (1.0, https://github.com/regSNPs/ExonImpact). Our tool accepts both BED format defining the alternatively spliced exon region and MISO event formats (examples can be found on the website).

Methods

Disease-causing and neutral alternatively spliced exons

In order to systematically evaluate the impact of protein structure features that are associated with pathogenic and neutral AS events, we first constructed a database that contains a group of exons whose splicing outcomes have strong implications for various diseases (positive dataset), and others that are considered to be neutral (negative dataset). For a positive training set, we extracted 4,667 exons containing 7,639 deleterious single nucleotide variants (SNVs) documented in the Human Gene Mutation Database (HGMD) that are responsible for causing human inherited disease by disrupting splicing regulation[9]. The SNV-induced splicing abnormality associated with these exonic events has been previously demonstrated to cause documented phenotypic consequences. The negative training set, i.e. the alternatively spliced events that are presumed to be functionally neutral, were derived from the 1000 Genomes database, in which genotyped individuals do not exhibit any apparent clinical phenotypes. Since genotyping information from the 1000 Genomes Project does not provide the profiling of exonic splicing patterns, we turned our attention to the small exonic insertions/deletions (exo-INDELs, the length of the INDELs is less than xx~~100~~21 nts), whose outcome at the protein level is the inclusion or exclusion of a stretch of amino acid residues. This consequence is similar to the results of alternative splicing. Although small exo-INDELs usually only represent a proportion of the entire exon, we constructed our negative training set by compiling all the INDEL-containing exons, assuming the fitness of the organisms are less sensitive to their status of being included or excluded from the protein product. It should be noted that this strategy might result in the inappropriate inclusion of deleterious events into the negative training data set, especially when the functional domains of the protein are encoded by part of the exon

that lies outside of the exo-INDEL regions. Our overall training dataset contained 4,211 and 2,664 exons in the positive and negative training groups, respectively.

ClinVar test dataset

An independent test dataset was collated from the NCBI ClinVar database. The test data set included 1,032 and 11 exons containing pathogenic and benign variants in the splicing sites (see Supplementary Table 2). In addition, the ClinVar database also documented 1,652 and 36 exons containing pathogenic and benign exo-INDELs, respectively.

SPANR prediction

Splicing-based Analysis of Variants (SPANR) was used to evaluate the roles of 1,000 Genomes variants in disrupting splicing regulation. We used the maximum mutation-induced change in PSI (percent-spliced-in) across 16 tissues that is reported by SPANR by default.

MISO prediction

In order to examine the inclusion ratios (ψ, percent-spliced-in) of the exons with high FIS (Functional Impact Scores), we ran MISO (mixture of isoforms) analysis on RNA-seq data from 310 brain samples across 11 brain regions, as documented in the GTEx (Genotype-Tissue Expression project)[11], and downloaded from dbGaP. Without losing generalizability, we only analyzed the 42,485 skipped exons (SE) that were documented in the MISO annotation. For each event, MISO provides maximum likelihood estimation on the ψ for each sample, as well as its 95% confidence interval (CI). A wide CI indicates less accurate prediction. We therefore removed all the events with no more than 10 samples containing CI≤0.1. After this filtering step, 1,909 events remained. After

removing the first and last exons, and the exons in the non-coding genes, ExonImpact calculated FIS scores for 1,852 events. Default parameters were used for MISO calculation.

Features

Disorder, ASA, and secondary structure predictions

For all the exon regions to be evaluated, a series of SPINE algorithms were used to derive its structure-related features. SPINE-D[13] was used to predict the disorder probability, and SPINE-X[14] was used to predict soluble accessible surface areas (ASA) and protein secondary structures (helix, sheet or coil). Default parameters were used for both algorithms.

Pfam domains

In order to quantify the level of overlap between the alternatively spliced exons and known protein family domains, Pfam (version 28.0) was used to predict the putative domains derived from amino acid sequences. We used two measurements as input features: the proportion of the exons that overlap with predicted protein family domains, and the proportion of predicted protein family domain that overlap with the candidate exon. If more than one domain was overlapped, we used the maximum percentage across all the domains.

Post-Ttranslational Modification (PTM)

PTM feature were derived from the dbPTM database[22] and we only considered the experimentally-validated PTMs. The PTM feature was defined as the number of documented PTM sites in the exon, normalized by the exon length.

Machine learning algorithm

RandomForest algorithm[23] was used for model training and prediction. Performance evaluation was performed using 10-fold cross-validation and bootstrapping (100 trees, mtry = 12). We found the performance of the RandomForest model was not sensitive to the selection of parameters. In addition, an independent test data set that hads not been used in training the model has beenwas used for to evaluateing the model performance.

References

1 Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev* **17**, 419-437, doi:10.1101/gad.1048803 (2003).

2 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7**, 1009-1115. doi:10.1038/nmeth.1528 (2010).

3 Shen, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc Natl Acad Sci USA*, E5593-E5601, doi:10.1073/pnas.1419161111 (2014).

4 Wu, J. *et al.* SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics* **27**, 3010-3016, doi:10.1093/bioinformatics/btr508 (2011).

5 Florea, L., Song, L. & Salzberg, S. L. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Research* **2**, 188. doi:10.12688/f1000research.2-188.v2 (2013).

6 Lu, H. *et al.* Predicting fFunctional aAlternative sSplicing by mMeasuring RNA sSelection pPressure from mMultigenome aAlignments. *PLoS Comput Biol* **5**, e1000608. doi:10.1371/journal.pcbi.1000608 (2009).

7 Zhao, H. *et al.* DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biology* **14**, R23 (2013).

8 Ellis, J. D. *et al.* Tissue-sSpecific aAlternative sSplicing rRemodels pProtein-pProtein iInteraction nNetworks. *Molecular Cell* **46**, 884-892, doi:10.1016/j.molcel.2012.05.037 (2012).

9 Stenson, P. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Mutation Genet* **133**, 577-5811-9, doi: "doi">10.1007/s00439-013-1358-4 (2014). 10.1002/humu.10212 (2014).

10 Consortium, T. G. P. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2012).

11 Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585. doi:10.1038/ng.2653 (2013).

12 Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* **42**, (Database issue):D980-985. doi:10.1093/nar/gkt1113 (2014).

13 Zhang, T. *et al.* SPINE-D: Accurate pPrediction of sShort and lLong dDisordered rRegions by a sSingle nNeural-nNetwork bBased mMethod. *J Biomol Struct Dyn* **29**, 799-813, doi:10.1080/073911012010525022 (2012).

14 Faraggi, E. *et al.* SPINE X: Improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* **33**, 259-267, doi:10.1002/jcc.21968 (2012).

15 Ghoorah, A. W. *et al.* A sStructure-bBased cClassification and aAnalysis of pProtein dDomain fFamily binding sSites and tTheir iInteractions. *Biology* **4**, 327-343, doi:10.3390/biology4020327 (2015).

16 N, L. *et al.* Spastin gene mutations in Bulgarian patients with hereditary spastic paraplegia. *Clin~~ical~~ Genet~~ics~~* **70**, 490-495, doi:10.1111/j.1399-0004.2006.00705.x (2006).
17 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res~~earch~~* **44**, D279-D285, doi:10.1093/nar/gkv1344 (2016).
18 KA, G. *et al.* genenames.org: the HGNC resources in 2015. *Nucleic Acids* **43**, D1079-D1085, doi:10.1093/nar/gku1071 (2015).
19 Xiong, H. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**,~~.~~ 1254806. doi:10.1126/science.1254806 (2015).
20 Wang, H. *et al.* Alternative splicing during *Arabidopsis* flower development results in constitutive and stage-regulated isoforms. *Frontiers in GENETICS* **5**, 25. doi:10.3389/fgene.2014.00025 (2014).
21 Folkman, L. *et al.* 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, doi:10.1093/bioinformatics/btu862 (2015).
22 Huang, K.-Y. *et al.* dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res~~earch~~* **44**, D435-D446, doi:10.1093/nar/gkv1240 (2016).
23 Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **Vol. 2/3**, 18-22 (2002).

Acknowledgements

This work was supported in part by grants from National Natural Science Foundation of China (61471139, 61403092). This work was also supported by National Health and Medical Research Council (1059775 and 1083450) of Australia to YZ. DNC and MM receive financial support from Qiagen Inc through a License Agreement with Cardiff University.

Author contributions

YL and ML designed the work program and ~~wrote most of~~drafted the manuscript. ML and XZ wrote the code and implemented the analysis. MM and DNC provided training data~~set~~ from the HGMD database. YY and YZ ~~conducted~~ performed the protein structure analysis. MM, DNC, YW~~,~~ and YZ participated in the writing of the manuscript and the interpretation of the results.

Competing financial interests

The authors are unaware of any competing interests.

Figures and Tables:

Figure 1: The workflow of the study, which contains four major components: data collection, feature extraction, model training and model evaluation.

Figure 2: Feature evaluation. (A) Probability density of each feature in pathogenic and neutral groups, respectively. (B) Probability cumulative density of each feature in HGMD and neutral groups, respectively. (C) Scatter plot of K-S test's p-values and D-values for each category of features. X-axis shows the $-\log_{10}$(p-value) and Y-axis is the D-value.

Figure 3: (A) Scatter plot between the average ASA score and the minimum probability of random coil. (B) An example (NM_014946) demonstrating the relationship between protein secondary and tertiary structures.

Figure 4: PCA biplot of all ~~the~~ features. Red and green dots represent pathogenic and neutral events, respectively. Each arrow line demonstrates one feature, and the color of the line indicates its category.

Figure 5: (A) ROC curve on an independent test data set. (B) Employing the 1,000 Genomes data set, the percentage of predicted high impact events (FIS ≥ 0.91) among the events with weak, intermediate and strong variants that disrupt ~~the~~ splicing are shown~~ed~~. (C) Relationship between True Positive Rate, False Positive Rate, F1 Score, MCC with cutoff, y=0.1 and x=0.82 is plotted to show the corresponding cutoff for the False Positive Rate = 0.1.

Figure 6: Proportion (%) of events with different levels of inclusion ratio in human brains that have low, intermediate and high FIS scores.

Supplementary Figure 1: The demonstration of disorder feature calculation. The example given is exon 11 of the *ACC* gene. The meanings of the twelve kinds of feature are given on the right of the Figure.

Supplementary Figure 2: Probability density for each feature. Feature's K-S test statistic is given at the top of each panel.

Supplementary Figure 3: 10-fold cross-validation and 100 times bootstrap validation.

Supplementary Figure 4: FIS's probability density for exons contained in benign INDELs and pathogenic INDELs respectively.

Supplementary Table 1: A brief description of each feature.

Supplementary Table 2: Exons' IDs that contain respectively benign and pathogenic SNPs in the splice junction site respectively.