Title

S-CAP extends clinical-grade pathogenicity prediction to genetic variants that affect RNA splicing

Authors

Karthik A. Jagadeesh*[1], Joseph M. Paggi*[1], James S. Ye[2], Peter D. Stenson[3], David N. Cooper[3], Jonathan A. Bernstein[4], Gill Bejerano[1,4,5,6]

Affiliations

[1] Department of Computer Science, Stanford University, Stanford, California 94305, USA
[2] Department of Biology, Stanford University, Stanford, California, 94305, USA
[3] Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff, UK
[4] Department of Pediatrics, Stanford University, Stanford, California 94305, USA
[5] Department of Developmental Biology, Stanford University, Stanford, California 94305, USA
[6] Corresponding Author: bejerano@stanford.edu
* denotes equal contribution

Abstract

There are over 15,000 known variants that cause human inherited disease by disrupting RNA splicing. While several *in silico* methods such as CADD, EIGEN and LINSIGHT are commonly used to predict the pathogenicity of noncoding variants, we introduce S-CAP, a tool developed specially for splicing which is better able to effectively distinguish pathogenic splicing-relevant variants from benign variants. S-CAP is a novel splicing pathogenicity predictor that reduces the number of splicing-relevant variants of uncertain significance in patient exomes by 41%, a nearly 3-fold improvement over existing noncoding pathogenicity measures while correctly classifying known pathogenic splicing-relevant variants with a clinical-grade 95% sensitivity.

Introduction

Genomic sequencing, and in particular exome sequencing, is revolutionizing the diagnosis of Mendelian disease[1–4], with over 5,000 genetic diseases already successfully mapped to over 3,000 genes[5]. Sifting through patient exomes in search of a causal variant is a time consuming process that often focuses on the coding sequence (CDS) of genes[6,7]. Powerful pathogenicity meta-predictors such as M-CAP[8] integrate multiple primary predictors such as SIFT[9], Polyphen-2[10] and CADD[11] with cross-species sequence conservation features to offer accurate clinical grade predictions capable of missing only a tiny fraction of known coding pathogenic mutations[12]. Achieving high sensitivity is important because reducing the size of the candidate list of variants of uncertain significance (VUS) is futile if the pathogenic variant itself is incorrectly classified as benign[13]. M-CAP is commonly used by clinicians to prioritize nonsynonymous variants[14,15], so that they can effectively consider first the variants that are most likely to yield a diagnosis.

However, with estimates that 70% of patient cases remain undiagnosed[16,17], there is value in looking beyond the CDS itself. Splicing is a complex and crucial step of gene expression, wherein vast sections of RNA, called introns, are removed from a pre-messenger RNA and the remaining RNA, called exons, are joined together to form the mature messenger RNA (mRNA). Changes in splicing induced by genetic variants can have severe impacts on the protein coding potential of an mRNA, such as the exclusion of an entire exon[18] or a frameshift[19,20] induced by creation of a new splice site[21], among other effects[22–24][22–24]. Exome sequencing typically captures sequence information up to 50 base pairs past exon boundaries into each adjacent intron[25]. This region covers a broad class of splicing-relevant variants: those that disrupt existing splice sites or exonic and proximal intronic splicing regulators, such as the branch point, and some that create new splice sites.

Indeed, there are over 15,000 known Mendelian disease causing variants that impact the gene product through RNA splicing[26,27]. At the same time, a typical singleton patient's exome contains over 500 splicing-relevant variants of uncertain significance (VUS)[28]. Existing tools such as CADD[11], EIGEN[29] and LINSIGHT[30] tackle a broad spectrum of non-coding variants, but in doing so they do not effectively predict splicing-relevant variant pathogenicity or provide clinical grade assurances of minimizing the false prediction of known pathogenic splicing-relevant variants as benign. Generic methods ignore the rich literature characterizing mRNA splicing and predicting associated molecular phenotypes, such as the percentage-spliced-in of exons[31–34]. These findings and methods provide invaluable insight into the potential of a variant to disrupt splicing. However, the splicing literature does not tell the whole story either, as predicting molecular phenotypes is a fundamentally different task from predicting if a variant will cause a disease. For a variant to cause a disease, the variant must disrupt normal splicing in one or more relevant tissues *and* the induced change in mRNA phenotype must be pathogenic. Similar obstacles are met in the few cases where clinicians attempt to identify patient genetic variants that disrupt splicing by performing a costly and time consuming RNA-seq experiment in an accessible, but not necessarily disease relevant, cell population[35,36].

We introduce S-CAP (**S**plicing **C**linically **A**pplicable **P**athogenicity), a machine learning tool that integrates knowledge of splicing with measures of variant, exon and gene importance into a splicing-specific pathogenicity score. We evaluate S-CAP at the high sensitivity required in clinical settings and show that it far outperforms existing non-coding pathogenicity scores, as well as tools focused solely on identifying synonymous variants that disrupt splicing. S-CAP will allow clinicians to consider a broad class of splicing-relevant variants, resulting in the diagnosis of more patients suffering from Mendelian disease.

Results

We developed a machine-learning framework to model and evaluate splicing-relevant variant pathogenicity. We analyzed the positional distribution and potential functional effects of variants near splice sites and defined 6 genomic regions that display distinct mutation rates and functional effects. We trained 6 models (1 per region) to predict splicing variant pathogenicity. This involved building (i) a labeled dataset of known pathogenic and benign splicing-relevant variants, (ii) a set of features to help discriminate between the pathogenic and benign variants in each region and (iii) a learning algorithm to identify patterns in the features and to distinguish between variants in the two classes. Finally, we evaluated our models on a set of known pathogenic and benign variants, as well as on real patients with various diseases caused by splicing-relevant variants.

The landscape of splice region variation

To build a dataset of labeled splicing-relevant variants, we considered variant pathogenicity, semantic effect and population frequency. We start by taking the union of 109,279 pathogenic single nucleotide variants (SNVs) from the Human Gene Mutation Database[26] (HGMD) and 25,793 pathogenic SNVs from ClinVar[37] to get a total of 114,382 unique pathogenic variants. We curated 15,833,389 benign SNVs observed in controls from the gnomAD database[38] who generally do not suffer from a Mendelian disease. To identify a subset of variants with a likely splicing semantic effect, these sets were filtered to the 'splicing region', all synonymous or intronic variants within 50 base pairs[25,39] of an exon boundary (we justify the choice of 'splicing region' below). Removing nonsynonymous and loss of function (stop gain or stop loss) variants ensured that our model was trained and evaluated nearly exclusively on splicing-relevant variants. To avoid any ambiguities, the benign SNVs were further filtered to remove any variant labelled pathogenic in HGMD or ClinVar. This resulted in 14,938 splicing-relevant pathogenic variants and 7,027,609 splicing-relevant benign variants. Then a frequency filter, based on the ACMG guidelines[13] that suggest clinicians consider common (> 1% frequency) variants as definitively benign, was applied to both sets yielding 14,838 rare splicing-relevant pathogenic variants and 6,760,450 rare splicing-relevant benign variants. In support of this ACMG guideline, we note that only 100 of 14,938 (0.67%) known pathogenic variants in the splicing region are common in the general population (see Methods).

To substantiate that nearly all pathogenic 'splicing region' variants do, in fact, disrupt splicing, we developed a simple model to assign a putative effect of variants on splicing. Variants can affect splicing by (1) create a cryptic splice site (2) disrupt an existing splice site or (3) disrupt an existing branchpoint, a mechanistically essential sequence motif generally located 18 to 45 base pairs upstream of each 3'SS[40]. We used the existing tools MaxEntScan[41] and LaBranchoR[42], which predict the strength of splice site sequences and branchpoint sequences, respectively. We denoted a variant as creating a cryptic splice site if the variant creates a splice site with a MaxEntScan score at least as high as the score for the reference splice site (with the variant included), disrupting a splice site if it greatly reduces the MaxEntScan score of the reference splice site, and disrupting a branchpoint if it has a low LaBranchoR *in silico* mutagenesis score. In cases where a variant had multiple putative effects, e.g. creating a cryptic splice site and disrupting an existing splice site, we resolved to the most extreme effect (cryptic splice site creation > splice site disruption > branchpoint disruption). We found that 97% of pathogenic splicing-region variants are predicted to have an effect on splicing, as opposed to only 18% of likely benign variants (**Fig. 1a-b**). We found that pathogenic variants are enriched at positions where the mechanistically essential U2 snRNA, U2AF, and U1 snRNA bind[43] (**Fig. 1b**). Conversely, variants occurring in the general population are biased away from these high information content positions[44] (**Fig. 1a**).

Region specific models to increase performance and alleviate ascertainment bias

In order to effectively capture these position specific patterns, we separated the variants into 6 regions relative to the splice sites with generally homogenous function and built a separate model for each region (**Supplementary Fig. 1**). Specifically, we grouped variants occurring in the obligate 5' GT (5' core) and 3' AG (3' core) dinucleotides, intronic variants upstream of a 3'ss (3' intronic), variants lying in the canonical U1 snRNA binding site, excluding the core 5'ss (5' extended), intronic variants downstream of a 5'ss (5' intronic), and synonymous variants within the protein coding gene (exonic) (**Fig. 1c**).

Core splice site variants are well known as having a large functional effect and are readily identified, a fact that has likely led to an overrepresentation of core splice site variants in pathogenic variant databases. Around 73% of known splicing region pathogenic variants occur within the core splice sites (**Fig. 1c**). Although this is consistent with the mechanistic importance of these positions, in unbiased studies of splicing quantitative trait loci, it is generally found that fewer than 1% of splicing-relevant variants are found to be located in core splice sites[45,46] and in a recent study employing RNA-seq data to identify splicing-relevant variants resulting in Mendelian disease only 2 of 6 (33%) causal variants in the splicing region were in core splice sites[35]. If left unaddressed, this bias would allow a classifier to have strong test set performance (by calling most core splice site variants pathogenic and others benign), but would often miss non-core splice site pathogenic variants. Separating variants by position allows us to *guarantee* that pathogenic variants are rarely misclassified as benign regardless of genomic region, thereby ensuring an overall low false negative rate irrespective of ascertainment biases present in annotated data.

S-CAP features

We curated existing metrics and developed several novel features to help distinguish between pathogenic and benign variants within the splicing region (**Supplementary Table 1**). The set consists of chromosome, gene, exon, and variant level features. At the **chromosome level**, 3 binary features distinguish between variants found on chromosome X, chromosome Y and the autosomes. Variants on the X chromosome present an important subset since in males, a hemizygous X chromosome variant inducing loss of function results in no viable gene product. Consistent with this intuition, pathogenic variants are highly enriched on the X chromosome as compared to the autosomes (7.11 fold enrichment, $p < 10^{-140}$ by two-sided Fisher's Exact Test). At the **gene level**, pLI[44], RVIS[47], and a haploinsufficiency score[48] help to measure the likelihood that a given gene is a 'disease gene'. At the **exon level**, exon length, exon length modulo 3, reference splice site strengths, an existing regional constraint score[49] and the exon sequence similarity between hg19 and 99 species from the UCSC 100way alignment serve to assist in distinguishing critical exons from those that may be safely excluded. Additionally, we developed a novel splice site constraint score to measure the fragility and tolerance of each exon to splice site mutations (see Online Methods). At the **variant level**, CADD[11] measures pathogenicity based on functional data annotations, LINSIGHT[30] measures variants' fitness effect through functional data and molecular evolution and SPIDEX[33] was incorporated to measure the impact of a variant on exon inclusion. PhyloP[50] and PhastCons[51] scores from the multiz46way and multiz100way alignments[52] measure the evolutionary importance of the affected base across primate species, placental mammals and all vertebrates. We also included a feature to capture the change in 3-mer content induced by a variant[53]. Additionally, we included region-specific features, such as a branchpoint disruption term for the 3' intronic region[42] and a 5' cryptic splice site creation term for the 5' intronic region (see Online Methods for a complete description of features).

The machine learning algorithm

Similar to the M-CAP classifier[8] for nonsynonymous SNVs, S-CAP is built using a gradient boosting tree classifier, a highly effective machine learning model[54]. This model iteratively builds decision trees, where each tree is picked to correct the most cases that were misclassified in the previous step. The final classifier is a linear combination of each of the previously derived decision trees (see Online Methods for details).

S-CAP consistently outperforms existing pathogenicity scores

Each of the 6 regions (described above) contains a set of pathogenic and benign variants (**Fig. 1c**), which were used to train 6 separate models. We performed 5-fold cross-validation and selected the median performing model as the final model for each region. S-CAP was evaluated against the most popular existing methods that score splicing-relevant variants: CADD, SPIDEX, LINSIGHT and EIGEN. The 6 S-CAP models outperformed all existing methods in all regions resulting in up to a 26.6% improvement in the AUC over the next best performing model (**Fig. 2**). S-CAP performance ranged from achieving an AUC of 0.804 in the 3' core region (**Fig. 2b**) to achieving an AUC of 0.953 in the 3' intronic region (**Fig. 2a**). No existing method consistently outperforms the other existing metrics across all splicing regions (**Fig. 2**).

To ensure a fair comparison, S-CAP performance on exonic variants was also independently compared against MutPred Splice[53], a tool focused on scoring only splicing-relevant synonymous variants, and found superior in both AUC and especially hsr-AUC (see Online Methods and **Supplementary Fig. 2**).

Clinically relevant threshold maintains high sensitivity

As previously shown in M-CAP[8], it is important to tune thresholds for clinical settings so that fewer than 5% of known pathogenic variants are misclassified. Neither LINSIGHT, EIGEN or SPIDEX provides a default threshold to consider a variant as pathogenic. As a result, it is difficult to use these methods for variant pathogenicity classification. CADD provides a threshold but, at the author-recommended default threshold, virtually all pathogenic splicing mutations outside the two core splice sites are discarded and incorrectly classified as benign (**Table 1**). Similarly, MutPred Splice provides a default threshold, but over 48% of known pathogenic exonic variants are misclassified at this threshold. We generated a high sensitivity threshold for all metrics in each region by finding the lowest threshold that results in the correct classification of at least 95% of a test set of pathogenic variants from that region (**Table 2**).

S-CAP is the best performer at clinically relevant thresholds

Moving into the high sensitivity domain, at or above a 95% true positive rate, S-CAP's performance (**Supplementary Fig. 3a-f**) ranged from achieving an hsr-AUC of 0.186 in the exonic region (**Supplementary Fig. 3c**) to an hsr-AUC of 0.549 in the 3' intronic region (**Supplementary Fig. 3a**). Similarly, S-CAP outperforms MutPred Splice on an independent test set of exonic variants (**Supplementary Fig. 2b**). Overall, S-CAP improves on existing metrics in all 6 splicing regions when focused on the high sensitivity domain, and no existing method consistently outperforms the others .

Different patterns observed for recessive and dominant variants

Recessive and dominant diseases are associated with different selective pressure on alleles, which we hypothesized would result in different feature importance and thresholds for determining variant pathogenicity. To address this complexity, we developed separate classifiers for dominant (heterozygous in patient) and recessive (homozygous or compound heterozygous in patient) at the gene level. In patients, we are given whether each variant appears in the heterozygous or homozygous state. However, our training data did not provide dominant and recessive labels, so we developed a framework for predicting this information from a control population for the sake of model training (see Methods).

Tagging 3’ and 5’ core variants as dominant or recessive resulted in improved performance **Supplementary Fig. 4a-h**) compared to models without these tags (**Fig. 2b,d**). For 3’ core variants, we built a model that improved upon an AUC of 0.804 and hsr-AUC of 0.257 from the original 3’ Core model to an AUC of 0.805 and 0.890 (**Supplementary Fig. 4c,d**) and hsr-AUC of 0.296 and 0.454 (**Supplementary Fig. 4g,h**) when tested on just dominant or recessive variants, respectively. Similarly, for 5’ core variants, we built a model that improves upon an AUC of 0.805 and hsr-AUC of 0.291 from the original 5’ core model to an AUC of 0.779 and 0.880 (**Supplementary Fig. 4a,b**) and hsr-AUC of 0.222 and 0.518 (**Supplementary Fig. 4e,f**) when testing on just dominant or recessive variants, respectively. The S-CAP model to be used on patients takes advantage of these split dominant and recessive models in the core splice site regions.

S-CAP eliminates the most VUS in patient exomes

Resources like S-CAP are developed on large sets of benign and pathogenic variants but ultimately are used to help with the interpretation of VUS in individual patients (**Fig. 1**c). To demonstrate the practical utility of S-CAP, we evaluated S-CAP and each of the comparison methods on 14 patients with Mendelian diseases caused by splice-altering mutations. After applying the standard allele frequency filter of ≤ 1%, a typical individual has on average a total of 533 rare variants within the splicing region (**Fig. 1c**). Typically, ~32% of variants were observed in each of the 5’ intronic, exonic and 3’ intronic regions. 3% were in the 5’ extended regions and 0-3 variants were in each of the 3’ core and 5’ core regions (**Fig. 1c**). Existing methods, CADD, LINSIGHT, EIGEN and SPIDEX, using the 95% sensitivity thresholds (**Table 2**), perform comparably when applied to all VUS in the splicing region of an individual patient and on average reduce the number of VUS by 4% - 15%. By contrast, S-CAP is more powerful, reducing the number of VUS in the splicing region for an individual patient by 31-46% (**Fig. 3c**), while confidently retaining the pathogenic variant for further detailed analysis (**Table 3**). For a larger sample size, we also evaluated each method on all (n=2054) individuals in the 1000 Genomes Project[28], which can conceptually be thought of as Mendelian disease patients with their pathogenic variants removed (see **Fig. 1**c). The observed fraction of VUS reduced on average per 1000 Genomes Project individual is consistent with the performance observed when applied to patients (**Fig. 3c-d**). Specifically, S-CAP is nearly three times as powerful as existing methods and on average results in a 41% reduction of the VUS within the splicing region of a given individual.

Discussion

Variants affecting splicing comprise the second largest category of known pathogenic mutations[27]. A broad class of potential splicing-relevant variants are already being captured by exome sequencing, yet clinicians are less able to interpret these variants for lack of proper interpretation tools. Here we address this problem by developing S-CAP, the first clinically applicable pathogenicity predictor dedicated exclusively to splicing-relevant variants.

In order for an *in silico* pathogenicity predictor to be useful in the clinic, it needs to be easy to use, carefully evaluated at high sensitivity and confidently remove a substantial fraction of benign variants. Existing noncoding variant tools are not easy to use for splicing variants because their performance is strongly dependent upon position relative to splice sites and no single method consistently outperforms the others. This means that employing existing methods, a clinician would have to consult multiple pathogenicity scores for variants in different regions. Furthermore, none of the existing methods have been carefully evaluated at clinical-grade sensitivity and most methods give no guidance about what cutoff should be used or result in the misclassification of an unacceptably high number of pathogenic variants (**Table 1**). After carefully evaluating the existing methods, we found that none of them confidently removed a considerable fraction of benign variants (**Fig. 3**). For example, after we retuned it (**Table 2**), SPIDEX performed well on variants in the intronic bins, but was close to random at predicting the pathogenicity of core splice site variants. CADD, only after we retune it, (**Table 1,2**) performed well on core and exonic variants but poorly on intronic variants (**Fig. 2**). S-CAP, addresses these important issues, as it consistently outperforms existing methods across all the regions, its performance has been carefully evaluated at clinical-grade sensitivity, and it removes close to three times as many benign variants as any other method (**Fig. 3**).

Central to the design of S-CAP is the use of region-specific models to alleviate the effects of ascertainment biases in curated pathogenic variant databases. Curated pathogenic variant databases contain invaluable information about the properties of pathogenic variants, but they also over-represent variants in known disease genes and in association with easily identifiable features. Of particular concern in splicing pathogenicity prediction is the inflated number of variants in core splice sites, which exists because they are easily recognized and have well established molecular consequences. If left unaddressed, this bias in the labeled pathogenic data would lead to unrealistic model performance, as a model could achieve relatively high test set performance simply by predicting that all core splice site variants are pathogenic and all others are benign. However, in the clinic, such a model would incorrectly classify pathogenic, non-core splice site variants as benign at an unacceptably high rate. Introducing separate models for each region alleviates this concern since each model's performance is evaluated using data from the same region, thereby assuring high sensitivity irrespective of the underlying positional distribution of pathogenic variants. Additionally, we accounted for the over-representation of pathogenic variants in known disease-associated genes by ensuring that variants from the same gene were never split between the training and evaluation sets (see Methods). This guaranteed that no gene level information was shared by features across folds.

The use of patient RNA-sequencing (RNA-seq) data to identify pathogenic variants that disrupt splicing is a growing and promising field[35,36,55] to which we believe S-CAP is complementary. In fact, our model already includes scores from SPIDEX[33], a deep learning model that was trained on tissue-specific RNA-seq data to predict the change in percent spliced-in ($\Delta\Psi$) of an exon given a variant. It would only be a small step to supplement these predicted $\Delta\Psi$s with experimentally measured $\Delta\Psi$s from RNA-seq experiments. Observing $\Delta\Psi$ through RNA-seq bypasses the difficult problem of explicitly or implicitly predicting the effect of a given variant on splicing in a particular cell type, but this is only part of the problem. Whether predicted or measured, gene expression and Ψ values vary between cell contexts and time points and in many cases the most relevant cell population to sequence would not be clear[36,56]. Perhaps even more importantly, molecular phenotypes are multiple steps away from real phenotypic change, making it difficult to predict organismal pathogenicity from molecular phenotype[57]. These factors limit the direct applicability of observed or predicted $\Delta\Psi$s to cases where there is a good understanding of the relationship between a disease, the cell population it affects, and the set of potentially causative genes. Many of the features used by S-CAP, such as evolutionary conservation, implicitly integrate an allele's importance over all cell populations and time points, complementing the strengths of RNA-seq based methods. Another direction for the integration of S-CAP and experimental methods is the use of cheap and fast site-directed sequencing in a diverse array of cell types to validate putatively pathogenic sites identified by S-CAP[58].

Eventually, comprehensive analysis of the splicing region will become commonplace in clinical settings. This is currently a difficult task given the complexity of splicing and the difficulty in predicting whether a change in splicing will result in disease. There are over five hundred rare variants of uncertain significance per individual in the splicing region with no clear semantic effect outside of the core splice sites and most are not observed in control populations. S-CAP represents an important step towards effectively interpreting splicing variation, but we will have to continue to improve on these methods, learning more from RNA-seq experiments, to render this problem more tractable.

Online Methods

Variant Processing

Dataset of pathogenic and benign variants

Pathogenic variants were obtained from two manually curated databases: the Human Gene Mutation Database (HGMD) Professional version 2017.1 and ClinVar release 20170406. Only HGMD variants tagged as Disease Mutation (DM) and ClinVar variants with Pathogenic Clinical Significance were included in the final set of pathogenic variants. Benign variants were obtained by identifying variants observed in individuals from gnomAD[38] r2.0.2.

Variant Annotation

ANNOVAR[59] v527 was used to annotate variants with predicted effect on protein-coding genes using gene isoforms from Ensembl[60] gene set version 75 for the hg19/GRCh37 assembly of the human genome. All coding isoforms were used where the transcript start and end sites were marked as complete and the coding span was a multiple of three.

Variant Filtering

All variants were filtered so as to only include rare variants in the splicing region that do not directly affect the protein coding sequence. Rare variants are defined to be variants with an allele frequency of 1% or less in all control populations and subpopulations in KGP phase 3, ExAC v0.3.1 and gnomAD r2.0.2. Variants in the splicing region that do not affect protein coding sequence are those determined to be synonymous or in the core splicing, extended splicing or intronic regions.

Positional Subsetting of Variants

Variants at different positions relative to splice sites have different properties, such as distributions of sequence conservation and ratio of pathogenic variants to benign variants. To better understand the relationship between position relative to splice site and the mechanistic effects on splicing, we developed a simple model to assign a putative effect of variants on splicing. To identify variants that create cryptic splice sites, we scanned a window the width of the MaxEntScan[41] input for the highest scoring splice site that is not the reference and reported the variant as creating a cryptic splice site if this highest scoring site was stronger than the reference (with the alternative allele included) and is made stronger by the variant. Variants that result in a decrease the MaxEntScan score of the reference splice site by 1 or more were determined to disrupt an existing splice sites. Finally, variants that have a LaBranchoR[42] *in silico* mutagenesis score of less than -0.1 were determined to disrupt a branchpoint.

This led us to allocate variants into subsets based upon their position relative to splice sites and train separate models for each. Subsets were selected so as to group together functionally related positions. In total, we constructed 6 subsets (**Fig. 1**). We created subsets of 3’ and 5’ core (1 or 2 bases from the splice site), extended (1-2 bases into the exon on the 5’ side and 3 to 6 bases from the 5’ splice site), 3’ intronic (2 to 50 bases from the splice site), 5’ intronic (7-50 bases from the 5’ splice site) and exonic (inside the exon but outside the extended region) regions.

We associated each variant with a neighboring exon. For variants close to multiple exons, we attempted to assign the variant to the exon considered as having the highest chance of a pathogenic effect, which we inferred from the density of pathogenic variants in each bin. Specifically, we favored associations in the following order: core, 5' extended, intronic, and exonic.

Features

Our models utilized a diverse library of previously described and novel features (fig. S1). These can be divided into gene level, exon level, and variant level features as well as a few features that are specific to individual regions. Below, we introduce our novel features and describe how we curated previously described features.

Gene Level

We obtained **RVIS**[47] scores from, the appropriately named, genic-intolerance.org (see URLs). We used the data in column "RVIS[pop_maf_0.05%(any)]" as a feature in our models. We obtained **pLI**[44] scores from supplementary table 13 of the original publication on 7 April 2017. We used as a feature the column denoted as pLI. We obtained a **haploinsufficiency score**[48] which is a probability of a gene being haploinsufficient directly from the publication page on the PLOS website in May 2017. We downloaded MPC, a recently proposed regional constraint score[49] (see URLs).

Exon Level

For each exon, we created **splice site features** which measure the number of rare and common variants observed in gnomAD in the 5' and 3' core (0-2) and extended (2-6) regions. This was motivated by the desire to share information between functionally similar positions. We take this feature to represent the constraint on the exon's splicing region in the human population. To avoid data leakage, when constructing this feature for a particular variant, we masked the effect of the variant itself on this score. Specifically, if a given 5'ss had one associated core variant, it was assigned a count of 0. If it had 2, both were assigned a count of 1. Additionally, we measured **exon identity** across vertebrates and found that the exon identities in many organisms were highly correlated. Principal components analysis (PCA) of the identity scores for all exons showed that 5 components explain the vast majority of variation in the data. To prevent overfitting, we included the original exon identities projected onto these first 5 principal components as features. We also included exon length and exon length mod 3 as features associated with each variant.

Variant Level

In order to consider the local sequence context of variants, we included **spectrum kernel features** representing the change in trinucleotide content induced by the variant[53]. For all 64 possible trinucleotides, we created a vector counting the number of occurrences in the alternative sequence and subtracted an equivalent vector for the reference sequence. The **MaxEntScan reference score, alternative score and difference** between the two were all used as features to quantify the strength of each exons' reference and alternative 5' and 3' splice site. **SPIDEX** scores[33] were downloaded directly from the Deep Genomics website on 7 April 2017. These scores can now be downloaded from Annovar (see URLs). Any variant not assigned a SPIDEX score was assigned a value of 0. We included 8 scores measuring evolutionary conservation, specifically, **PhyloP 46way vertebrates, placental mammals, primates, PhyloP 100way vertebrates, PhastCons 46way vertebrates, placental mammals, primates and the PhastCons 100way vertebrates**. Any base not annotated with a conservation score was assigned a value of 0. We also included the signed **distance to the 5'ss and 3'ss splice site of the associated exon** as features to measure base-pair importance.

Region Specific features

The 3' intronic S-CAP model includes a **branchpoint feature** modeled using LaBranchoR, a bi-directional LSTM (long short-term memory) model trained on the genome sequence surrounding experimentally validated branchpoint sites[42]. Specifically, we used the *in silico* mutagenesis scores available online (see URLs) as a feature.

We explicitly represented the strength of **cryptic sites** created by each variant using MaxEntScan[41]. In 3' intronic, 3' core, and exonic bins, we included a 3' cryptic splice site creation term and, in the exonic, 5' extended, 5' core, and 5' intronic bins, we included a 5' cryptic splice site creation term. For each variant, we scanned for the highest scoring splice site motif that overlaps the variant, excluding reference splice sites. We used as features the strength of the cryptic site, the change in the strength of the cryptic site induced by the variant, the distance from the reference splice site to the cryptic site and the difference in strength between the cryptic site and the reference splice site.

Model Training and Testing

We performed 5-fold cross-validation to train and identify a generalizable S-CAP model. 5-fold cross-validation refers to splitting the data into 5 roughly equally sized parts (folds). All variants found in a single gene were included in the same fold to ensure that there was no leakage of feature information across the training and test sets. We then merged 4 of 5 sets to form a training dataset, trained the model on this training dataset and evaluated on the remaining fold to obtain an expected accuracy. We performed this process 5 times (each combination of 4 folds was merged together to form the training dataset) testing on the fold that was not included in the training dataset.

To train the S-CAP model, we used a Gradient Boosting Tree model implemented in the python 2.7.13 sklearn version 0.18.1 library and used the default parameters to reduce the chance of overfitting the model. After training 5 models during the cross-validation phase, we picked the median performing model as the final classifier. The ROC curves were built based on performance on the test set for this specific median model.

Comparison Metrics

We sought to compare our performance to those of other methods used to infer the importance / pathogenicity of noncoding variants. We evaluated the performance of each of the methods below by using the output score directly as a pathogenicity score. For SPIDEX, we negated the score, as is consistent with large negative scores having a larger impact on function and constraint, respectively. We report performances for all subsets where the method reported scores for at least 50% of variants. Variants where a method did not report a score are excluded from evaluation.

CADD[11] v1.3 scores were downloaded from the CADD website (see URLs). **SPIDEX**[33] scores were downloaded directly from the Deep Genomics website on 7 April 2017. Any variant not assigned a SPIDEX score was assigned a default value of 0. **LINSIGHT**[30] scores were downloaded directly from the LINISGHT website (see URLs) on 7 April 2017. Any variants not assigned a LINSIGHT score was defaulted to be 0. **Eigen v1.0 coding** and **Eigen v1.0 noncoding**[29] Phred scores were downloaded from the Eigen website (see URLs). **MutPredSplice**[53] score were downloaded from the MutPredSplice website (see URLs).

As MutPred Splice was trained using some of the same data used to build S-CAP, a random train and test split of the pathogenic data would have resulted in the inclusion of variants in the test set that were used to train the MutPred Splice classifier. To ensure zero information leakage between the training and testing datasets, we carefully built a test set that excluded all MutPred Splice training data.

Recessive v. Dominant Classifiers

We developed separate classifiers for recessive and dominant acting variants in the 3' and 5' core splice site regions. We opted not to include dominant and recessive classifiers for the other 4 regions as we did not have a sufficient number of pathogenic variants to train and evaluate multiple models and, in our exploration, they made less of an impact. Intuitively, for core variants the molecular phenotype is obvious, a loss of splicing. This places the full burden on predicting if this change will result in disease, a task heavily dependent upon whether the variant acts via a recessive or dominant mechanism. Whereas in the other regions, the majority of possible variants have little impact on splicing, making predictions as to whether or not the variant will have an effect on splicing is the primary challenge, a task unrelated to inheritance mode.

When evaluating a patient, core variants observed as heterozygous are routed to the dominant classifier and variants observed to be homozygous are routed to the recessive classifier. Since there is no genotype information available for the labeled pathogenic and benign variants, we developed a framework to label variants as dominant or recessive based on their occurrence in a control population. We labeled pathogenic variants observed as heterozygous in the control population as likely recessive, because a single copy can be harbored with no major issues, and pathogenic variants never observed in the control population as likely dominant. Benign heterozygous variants that are never observed as homozygous in the healthy control population provide little information regarding their potential to cause a recessive acting disease. In this case, we tagged benign heterozygous variants never observed to be homozygous as dominant, and benign variants observed as homozygous as recessive. Additionally, as any variant on the X chromosome resembles a homozygous autosomal variant in males, all X chromosome variants were labelled as recessive.

We encoded whether each variant was considered dominant or recessive as a binary feature and trained a single gradient boosting tree model for each region. Then, we found two 95% true positive rate thresholds separately based on test sets of only dominant-tagged and only recessive-tagged variants.

Patient Datasets

Sequencing and diagnosis for all patients was performed by other laboratories. All patient data were downloaded by requesting access to the European Genome-Phenome Archive (EGA) and the database of Genotype and Phenotype (dbGaP) databases. Variant call files (VCFs) for 11 patients were submitted by the Deciphering Developmental Disorders (DDD) study in the European Genome-Phenome Archive (EGA) study EGAS00001000775. An additional 3 patient VCFs were submitted to the database of Genotype and Phenotype (dbGaP) study phs000655.v3.p1.

Data availability

S-CAP scores for all rare variants in the predefined splicing region in the human genome, along with the source code repository and final trained models for the S-CAP classifier, will be made available via the S-CAP website upon publication.

URLs

S-CAP website, http://bejerano.stanford.edu/scap (upon publication)
S-CAP codebase, https://bitbucket.org/bejerano/splicing_classifier (upon publication)
RVIS, http://genic-intolerance.org/data/RVIS_Unpublished_ExACv2_March2017.txt
SPIDEX, http://www.openbioinformatics.org/annovar/spidex_download_form.php
LINSIGHT, http://compgen.cshl.edu/~yihuang/LINSIGHT/
Haploinsufficiency, https://doi.org/10.1371/journal.pgen.1001154
LabranchOR, http://bejerano.stanford.edu/labranchor/
ExAC, ftp.broadinstitute.org/pub/ExAC_release/release1/regional_missense_constraint/

Acknowledgements

We thank Ron Dror, Aaron Wenger, Mark Berger, Johannes Birgmeier and all members of the Bejerano Lab for useful discussions, feedback and advice. P.D.S and D.N.C receive financial support from Qiagen through a license agreement with Cardiff University. This work was funded in part by Stanford Graduate Fellowship and the Computational and Evolutionary Genomics Fellowship to K.A.J. and the Stanford Pediatrics Department, DARPA, a Packard Foundation Fellowship and a Microsoft Faculty Fellowship to G.B.

Author Contributions

K.A.J., J.M.P. and G.B. designed the study. K.A.J., J.M.P. and J.S.Y. developed the features, trained the model and evaluated the results. P.D.S. and D.N.C. curated the HGMD data. J.A.B. provided patient exome cases and feedback. K.A.J, J.M.P. and G.B. wrote the manuscript. All authors reviewed the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

2. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).

3. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).
4. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42,** 790–793 (2010).
5. Hamosh, A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30,** 52–55 (2002).
6. Dewey, F. E. *et al.* Clinical Interpretation and Implications of Whole-Genome Sequencing. *JAMA* **311,** 1035 (2014).
7. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47,** 717–726 (2015).
8. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48,** 1581–1586 (2016).
9. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
10. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
11. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
12. Anderson, D. & Lassmann, T. A phenotype centric benchmark of variant prioritisation tools. *NPJ Genomic Med.* **3,** 5 (2018).
13. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15,** 733–747 (2013).

14. Einarsdottir, E. *et al.* Identification of NCAN as a candidate gene for developmental dyslexia. *Sci. Rep.* **7,** (2017).
15. Bodian, D. L. *et al.* Genomic analysis of an infant with intractable diarrhea and dilated cardiomyopathy. *Mol. Case Stud.* **3,** (2017).
16. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369,** 1502–1511 (2013).
17. Iglesias, A. *et al.* The usefulness of whole-exome sequencing in routine clinical practice. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **16,** 922–931 (2014).
18. Cuajungco, M. P. *et al.* Tissue-specific reduction in splicing efficiency of IKBKAP due to the major mutation associated with familial dysautonomia. *Am. J. Hum. Genet.* **72,** 749–758 (2003).
19. Wong, J. J.-L. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154,** 583–595 (2013).
20. Marquez, Y., Höpfler, M., Ayatollahi, Z., Barta, A. & Kalyna, M. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.* 995–1007 (2015). doi:10.1101/gr.186585.114
21. Eom, T. *et al.* NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *eLife* **2,** e00178 (2013).
22. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17,** 419–437 (2003).
23. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17,** 19–32 (2016).

24. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17,** 407–421 (2016).
25. Samuels, D. C. *et al.* Finding the lost treasures in exome sequencing data. *Trends Genet. TIG* **29,** 593–599 (2013).
26. Stenson, P. D. *et al.* 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 (2014).
27. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr. Protoc. Bioinforma.* **Chapter 1,** Unit1.13 (2012).
28. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
29. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48,** 214–220 (2016).
30. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49,** 618–624 (2017).
31. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
32. Zhang, C. *et al.* Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls. *Science* **329,** 439–443 (2010).
33. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347,** (2015).

34. Barbosa-Morais, N. L. *et al.* The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* **338,** 1587–1593 (2012).
35. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9,** (2017).
36. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
37. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862-868 (2016).
38. Karczewski, K. The genome Aggregation Database (gnomAD). *MacArthur Lab* (2017).
39. Meienberg, J. *et al.* New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res.* **43,** e76 (2015).
40. Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. & Fairbrother, W. G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.* **19,** 719–721 (2012).
41. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **11,** 377–394 (2004).
42. Paggi, J. M. & Bejerano, G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *bioRxiv* 185868 (2017). doi:10.1101/185868
43. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15,** 108–121 (2014).

44. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
45. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* **8,** (2017).
46. Zhang, X. *et al.* Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat. Genet.* **47,** 345–352 (2015).
47. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).
48. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and Predicting Haploinsufficiency in the Human Genome. *PLOS Genet.* **6,** e1001154 (2010).
49. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353 (2017). doi:10.1101/148353
50. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
51. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
52. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14,** 144–161 (2013).
53. Mort, M. *et al.* MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* **15,** R19 (2014).

54. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29,** 1189–1232 (2001).
55. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8,** (2017).
56. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
57. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
58. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6,** 986–994 (1996).
59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).
60. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662-669 (2015).

Figures and Tables

Table 1

	Method		
Region	CADD	LINSIGHT,EIGEN,SPIDEX	Mutpred Splice
	Author's Thresholds		
	≥ 20	N/A	≥ 0.6
3' intronic	99%	N/A	N/A
3' core	4%	N/A	N/A
exonic	98%	N/A	48%
5'core	2%	N/A	N/A
5' extended	96%	N/A	N/A
5' intronic	97%	N/A	N/A

Table 1. Misclassification rate (low is good) of existing metrics at author-recommended thresholds. The two tools that provide classification thresholds, misclassify 48%-99% of known pathogenic variants in different regions. Other tools do not provide a threshold.

Table 2

region	CADD	LINSIGHT	EIGEN	SPIDEX	MutPred Splice	S-CAP
3’ intronic	≥ 0.709	≥ 0.048	≥ 3.57	≤ -1.45	N/A	≥ 0.005
3’ core	≥ 21.50	≥ 0.767	≥ 7.25	≤ 1.14	N/A	Dom.: ≥ 0.031
						Rec.: ≥ 0.144
Exonic	≥ 0.061	≥ 0.086	≥ 3.78	≤ -2.31	≥ 0.090	≥ 0.012
5’ core	≥ 22.6	≥ 0.795	≥ 8.381	≤ 1.65	N/A	Dom.: ≥ 0.032
						Rec.: ≥ 0.357
5’ extended	≥ 7.423	≥ 0.211	≥ 13.80	≤ -0.910	N/A	≥ 0.003
5’ intronic	≥ 0.852	≥ 0.047	≥ 0.847	≤ -1.636	N/A	≥ 0.004

Table 2. High sensitivity thresholds after recalibrating each method. We retuned each existing method for each of Figure 1’s six regions by finding the smallest threshold that resulted in the correct classification of 95% of pathogenic variants from that region. S-CAP thresholds for all regions (including the dominant and recessive modes) are included in the last column. By definition with the high sensitivity thresholds each method will misclassify at most 5% of known pathogenic mutations, but their effectiveness at correctly classifying benign variants and reducing the number of patient VUS will vary greatly.

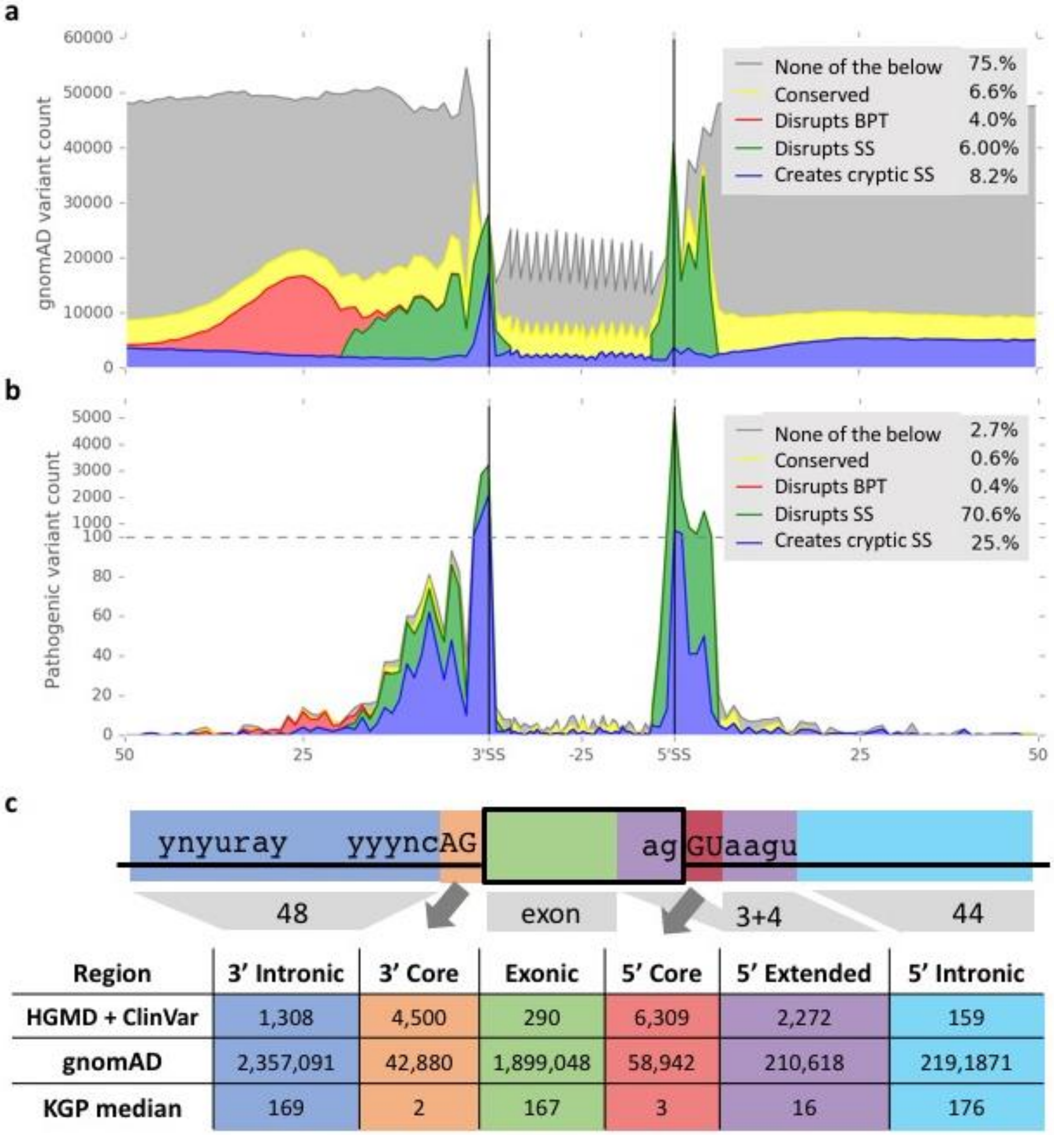
Table 3

Patient_ID	Disease	chr:pos (hg19)	Zygosity	Region	SCAP (%-ile)	SPIDEX (%-ile)	CADD (%-ile)	LINSIGHT (%-ile)	EIGEN (%-ile)
DDDP102313	**Koolen-de-vries syndrome**	**17:44144914 C>T**	**Heterozygous**	**5' Core**	0.081 (63.3)	-14.532(52.6)	25.8 (77.5)	0.978 (65.6)	15.932 (49.0)
DDDP100243	**Claes-Jensen type mental retardation**	**X:53245380 C>A**	**Hemizygous**	**3' Core**	0.919 (99.1)	-1.305 (22.1)	24.8 (65.02)	0.822 (23.2)	N/A
DDDP110794	**Mental Retardation**	**3:71037144 C>T**	**Heterozygous**	**5' Core**	0.217 (88.2)	-23.241(69.6)	26.4 (86.9)	0.988 (90.8)	20.234 (73.9)
DDDP111322	**Nephrotic Syndrome**	**19:36333453 C>T**	**Compound Heterozygous**	**3' Core**	0.629 (38.5)	-1.67 (24.2)	25 (69.8)	0.97 (54.8)	14.702 (40.2)
DDDP102111	**Epileptic encephalopathy**	**2:166165305 G>A**	**Heterozygous**	**5' Core**	0.348 (93.6)	-32.609 (82.8)	25 (59.8)	0.988 (90.8)	26.022 (87.7)
DDDP108825	**Cohen Syndrome**	**8:100729602 G>A**	**Compound Heterozygous**	**5' Core**	0.879 (94.2)	-7.189 (35.5)	25.1 (62.2)	0.982 (79.2)	25.405 (85.5)
DDDP100128	**Sotos Syndrome**	**5:176673677 A>G**	**Heterozygous**	**3' Core**	0.137 (81.4)	-12.574 (61.1)	22.9 (24.8)	0.961 (94.5)	9.876 (18.6)
DDDP111152	**Ehlers-Danlos Syndrome**	**2:189873948 G>A**	**Heterozygous**	**5' Core**	0.777 (98.6)	-5.594 (31.1)	27.9 (97.5)	0.989 (92.9)	29.666 (90.5)
DDDP102594	**Mental Retardation**	**X:135095506 A>G**	**Hemizygous**	**3' Core**	0.758 (98.6)	-2.955 (31.0)	24.3 (52.9)	0.874 (29.6)	N/A
DDDP111486	**Noonan Syndrome**	**11:119148874 A>T**	**Heterozygous**	**3' Core**	0.539 (97.6)	-28.352 (89.7)	24.9 (67.4)	0.982 (79.3)	22.809 (80.2)
DDDP100281	**Epileptic encephalopathy**	**16:56370773 G>A**	**Heterozygous**	**5' Core**	0.151 (82.1)	-7.739 (37.0)	26.6 (89.2)	0.987 (88.5)	24.055 (84.9)
C11	**Congenital fiber-type disproportion**	**19:38958362 C>T**	**Heterozygous**	**Exonic**	0.657 (97.8)	-0.366 (89.6)	19.430 (97.6)	N/A	7.879 (19.8)
C1	**Dystroglycanopathy**	**1:46655129 C>A**	**Compound Heterozygous**	**5' Core**	0.817 (77.03)	-52.028 (95.9)	21.500 (14.3)	0.992 (96.8)	4.976 (7.63)
		1:46660532 G>A		**Exonic**	0.095 (83.0)	-1.556 (96.4)	12.050 (79.2)	N/A	21.658 (76.5)
E2	**Nemaline myopathy**	**2:152520057 C>T**	**Compound Heterozygous**	**5' Extended**	0.147 (98.9)	-11.190 (85.4)	19.680 (97.9)	0.981 (79.2)	20.272 (65.0)
		2:152544805 C>T		**5' Core**	0.915 (99.1)	-2.970 (22.6)	27.200 (94.1)	0.990 (94.7)	20.384 (74.5)

Table 3. Causative variants in each patient and pathogenicity predictions. Each row describes a single patient, their underlying disease, causative variant and its zygosity, the region of the genome in which it is located, and the score and percentile each method assigns to the variant. The percentile was computed by measuring the fraction of variants in the same region with a score less than the score assigned to the causative variant. The 99th percentile denotes that 99% of variants in that region have a score less-pathogenic than the one we are observing

thereby indicating that the variant is considered to be highly pathogenic. Dark and light green filled entries have the highest and second highest percentile scores for the given variant, respectively. Red entries highlight patients where the causative variant would have been classified as benign using the author-recommended thresholds.

Figure 1



Region	3' Intronic	3' Core	Exonic	5' Core	5' Extended	5' Intronic
HGMD + ClinVar	1,308	4,500	290	6,309	2,272	159
gnomAD	2,357,091	42,880	1,899,048	58,942	210,618	219,1871
KGP median	169	2	167	3	16	176

Figure 1. Distribution of rare, noncoding variants in the splicing region. We built two sets of rare variants near exon boundaries that have no effect on the annotated coding sequence: (A) a set of likely benign variants from the Genome Aggregation Database (gnomAD) and (B) a set of putatively pathogenic variants from the Human Gene Mutation Database (HGMD) and ClinVar. We developed a simple model to ascertain the effect of these variants on splicing, using MaxEntScan to assess if a variant created a cryptic splice site (SS) or disrupted the reference splice site, and LaBranchoR *in silico* mutagenesis scores to assess if a variant disrupted a branchpoint (BPT). In (A) and (B), we plot the aggregate counts of each variant set as a function of position relative to the nearest splice site, colored by their putative effect. Nearly 97% of pathogenic variants in the splicing region (as defined in the text) are predicted to have a putative deleterious effect on splicing, as compared to only 18% of benign variants. (C) We split the variants into different regions with largely homogenous function. The majority of known pathogenic variants (and the easiest ones to detect) are found in the 3' and 5' core splice site

regions, whereas the majority of benign variants are found in the 3' intronic, exonic and 5' intronic regions. In a typical (median) individual from the 1000 genomes project (KGP), the distribution of variants is similar to the distribution of gnomAD benign variants.

Figure 2

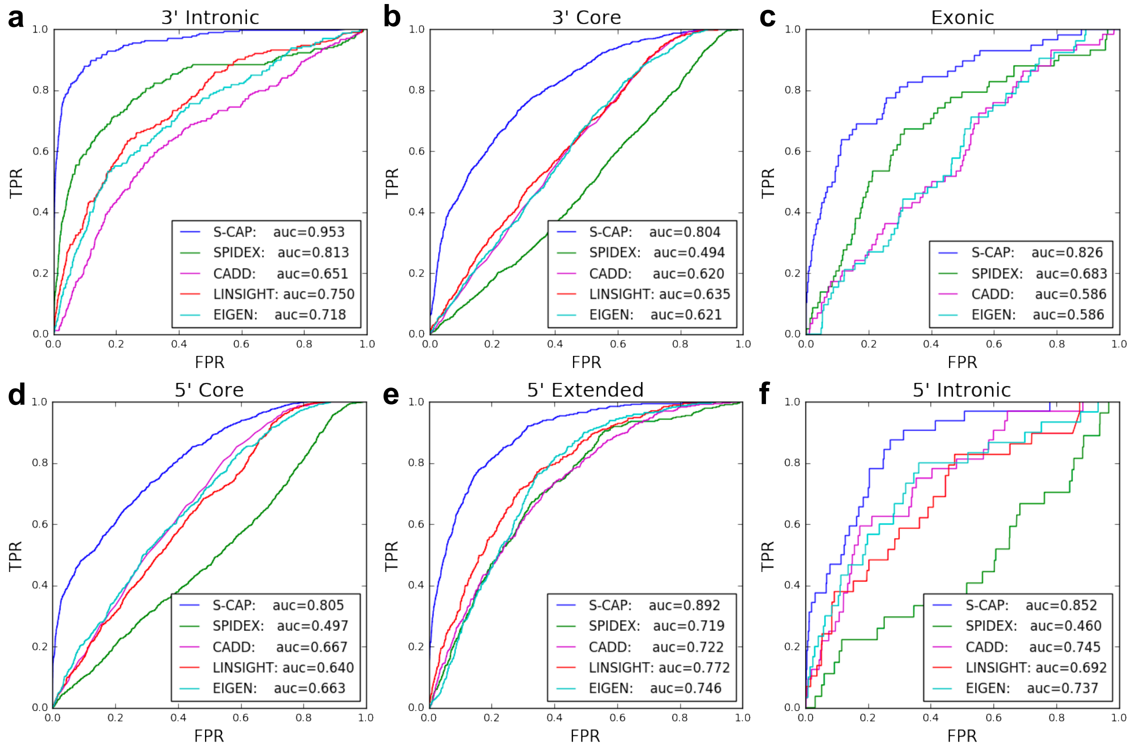


Figure 2. Overall performance per region for splicing pathogenicity classification. For each method, 1000 threshold points were determined by evenly spanning the range from the minimum to the maximum score observed for the method. A true positive rate and false positive rate were determined for each threshold value and used to build the receiver operating characteristic (ROC) curve. We compared to the next best method in each region and found that S-CAP performs 17.2% better in the 3’ intronic region (A), 26.6% better in the 3’ core sites (B), 20.9% better in the exonic region (C), 20.7% better in the 5’ core sites (D), 15.5% better in the 5’ extended region (E) and 14.4% better in the 5’ intronic region (F). S-CAP outperforms existing metrics in all regions whereas none of the existing methods consistently outperforms the others.

Figure 3

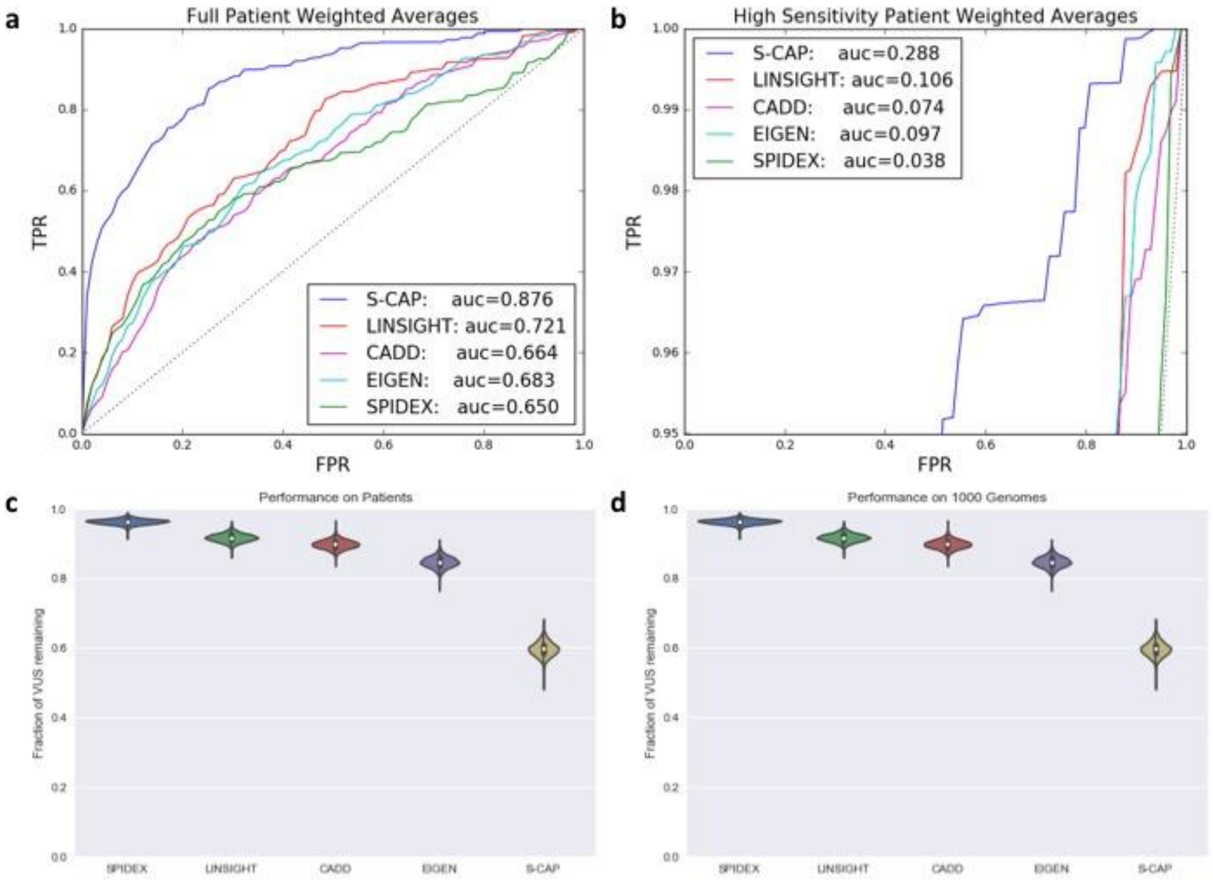


Figure 3. Overall performance on patient data. We take a weighted sum of the AUC from each of the 6 regions based on the distribution of variants seen in a typical individual to form an overall receiver operating characteristic (ROC) curve representative of the overall performance expected on patients. (A) S-CAP improves on the next best method by 21.5% in overall AUC and (B) and by 172% in the hsr-AUC. (C) S-CAP reduces the number of splicing related variants of uncertain significance (VUS) from patient exomes by 40% while maintaining the pathogenic variants with 95% sensitivity. At the same sensitivity requirement, existing methods reduce the VUS by only 4% - 15%. (D) We observe a similar reduction in VUS over all (n=2054) Thousand Genomes Project individuals, which conceptually only differ from Mendelian disease patients by up to 2 mutations.

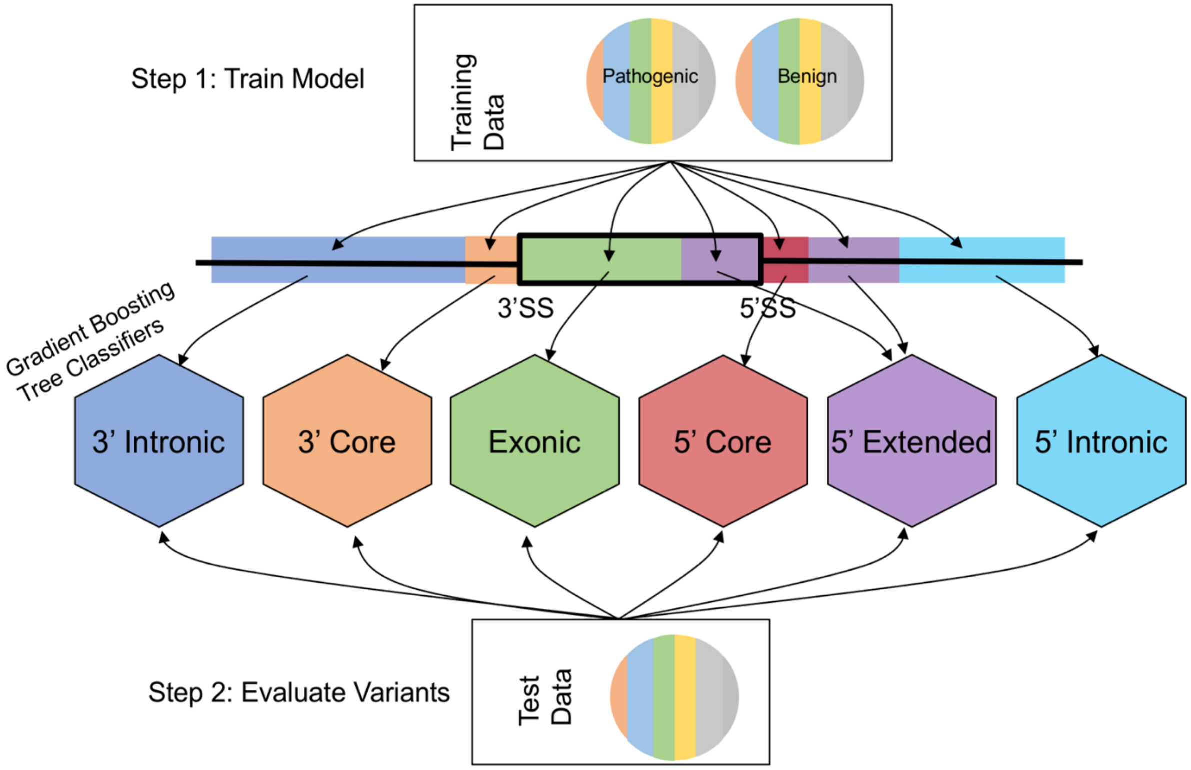
Supplementary Figures

Supplementary Table 1

Feature Category	Features	Description
Chromosome Level	X, Y or not XY	3 binary features to indicate whether a variant is found on the autosomes or the X or Y chromosome
Gene Level	RVIS	Measure of gene mutability. Observed v. Expected variant abundance
	pLI	Measure of gene loss of function tolerance.
	haploinsufficiency score	Recessive or dominant gene inheritance
Exon Level	# rare 3’ core variants	rare variants observed in 3’ core region
	# rare 5’ core variants	rare variants observed in 5’ core region
	# common 3’ core variants	common variants observed in 3’ core region
	# common 5’ core variants	common variants observed in 5’ core region
	# rare 3’ extended variants	rare variants observed in 3’ extended region
	# rare 5’ extended variants	rare variants observed in 5’ extended region
	# common 3’ extended variants	common variants observed in 3’ extended region
	# common 5’ extended variants	common variants observed in 5’ extended region
	exon % identity	Measure the % identity of hg19 exon with each of the 99 species in the multiz100way. Take top 6 PCA components after fitting a PCA to all coding exons.
	exon length	Number of bases in the exon
	exon length % 3	Number of bases modulo 3
	MPC	Regional mutational constraint score
Variant Level	Spectrum Kernel	Count of all 3-mers introduced and removed by mutation
	MaxEntScan	The difference in motif match
	SPIDEX	Predicted $\Delta\Psi$ (change in exon expression)
	distance to 5’ SS	Number of bases to the 5’ splice site
	distance to 3’ SS	Number of bases to the 3’ splice site
	CADD	Functional data SVM-based classifier
	LINSIGHT	Conservation based model to identify variants under negative selection.
	PhyloP	Base-pair conservation across primates, mammals and vertebrates
	PhastCons	Regional conservation across primates, mammals and vertebrates
3’ Intronic	LaBranchoR	Sequence based deep learning branchpoint prediction
3’ Intronic, 3’ Core, Exonic	3’ cryptic splice site creation terms	MaxEntScan based feature to measure cryptic splice creation near the 3’ side
Exonic, 5’ Core, 5’ Extended, 5’ Intronic	5’ cryptic splice site creation terms	MaxEntScan based feature to measure cryptic splice creation near the 5’ side
3’ Core, 5’ Core	Zygosity	Indicate whether the variant is seen in a heterozygous or homozygous state.

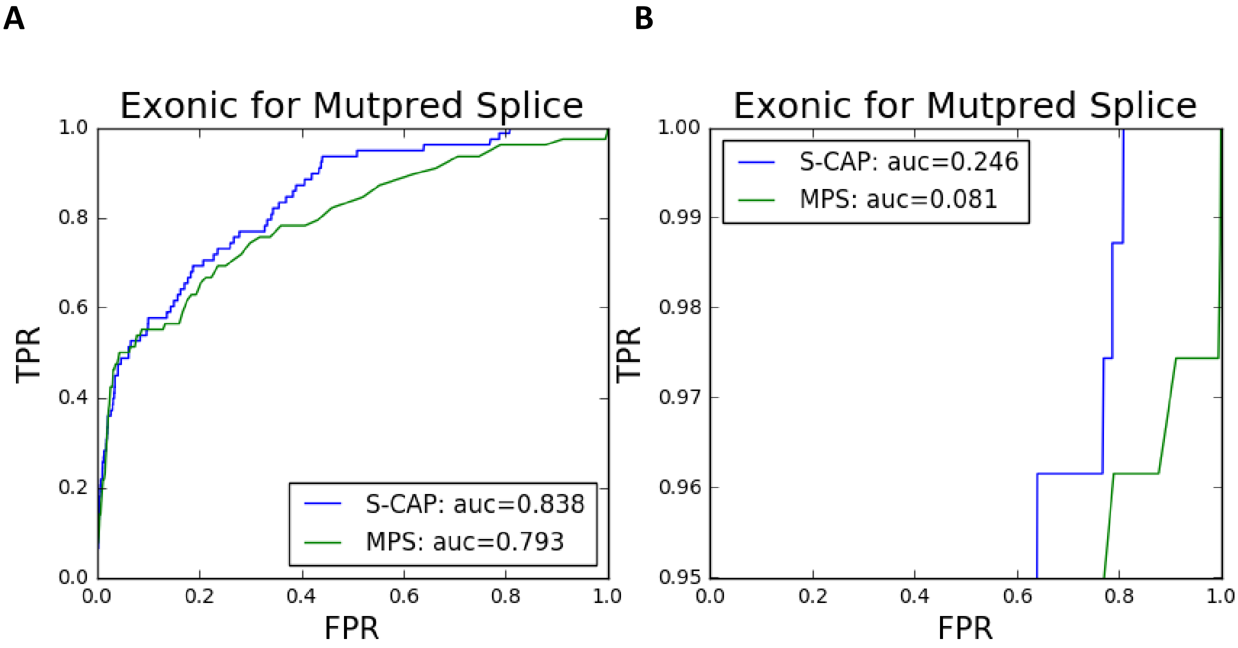
Supplementary Table 1. Description of features used to build S-CAP. The chromosome, gene, exon and variant level features were used in all models for all regions. Additional features were specific to certain regions. These are enumerated in the rows below the variant level features section.

Supplementary Figure 1


Step 1: Train Model
Training Data
Pathogenic
Benign
3'SS
5'SS
Gradient Boosting Tree Classifiers
3' Intronic
3' Core
Exonic
5' Core
5' Extended
5' Intronic
Step 2: Evaluate Variants
Test Data

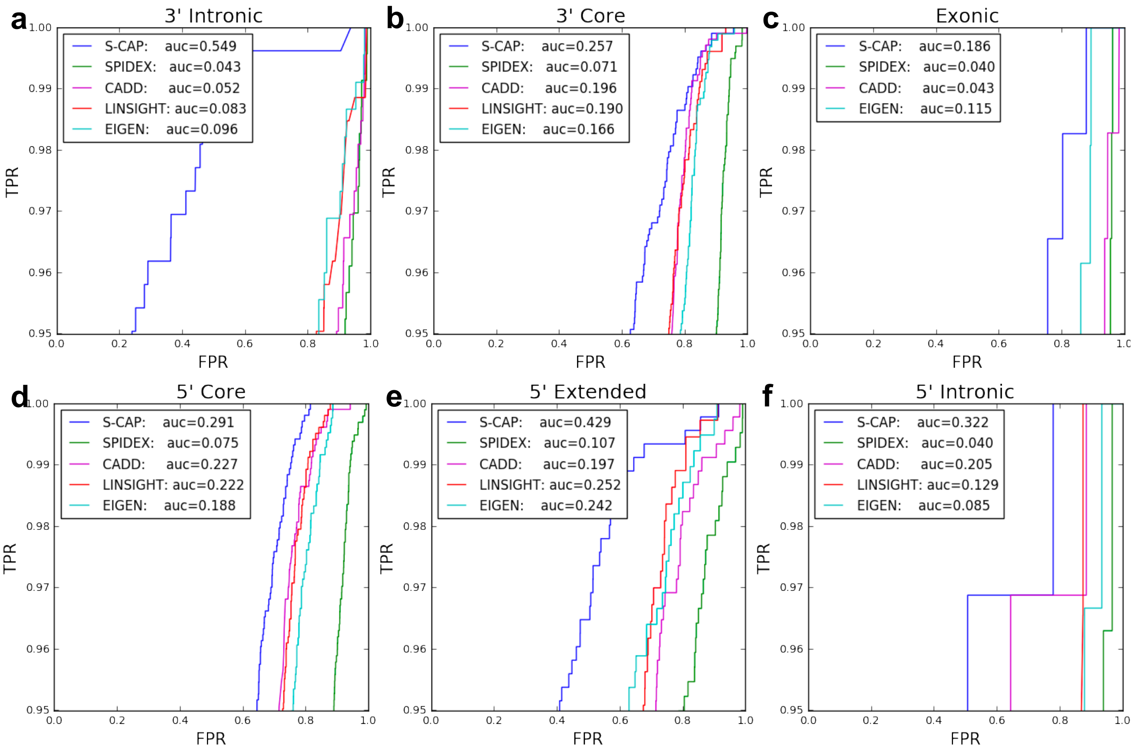
Supplementary Figure 1. Framework for training and evaluating 6 pathogenicity models. The splicing region is split into 6 independent regions as defined in **Fig. 1c**, and a separate model is trained for variants residing in each region. Given a set of variants to be scored, we calculate the S-CAP score for each variant by using the corresponding model associated with the region where the variant is found.

Supplementary Figure 2



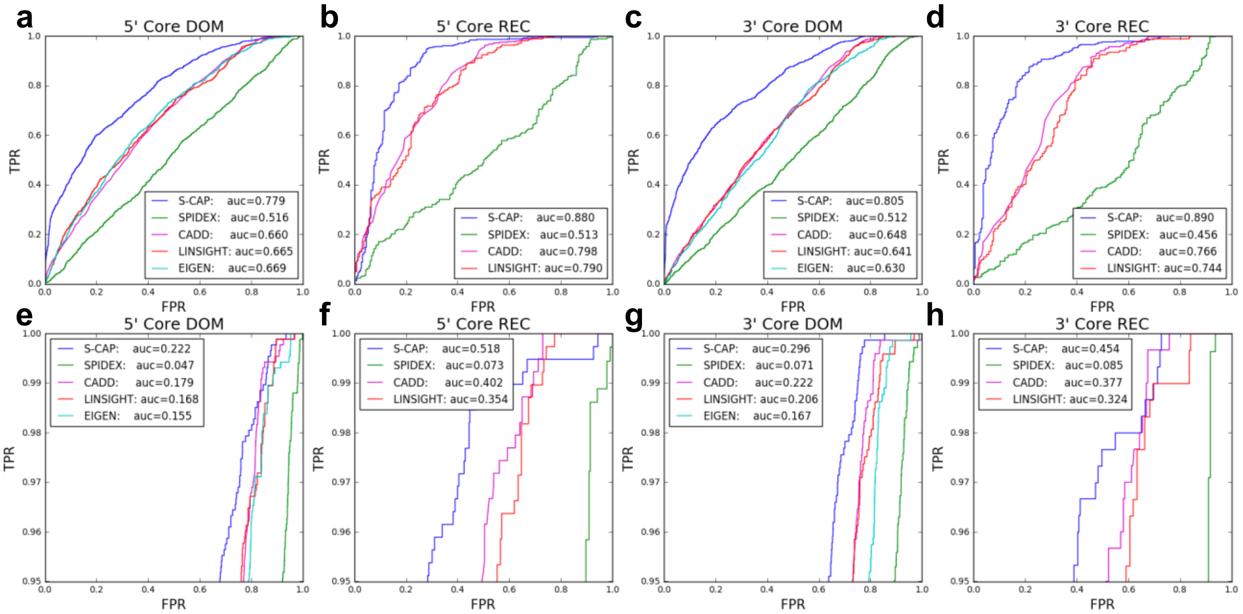
Supplementary Figure 2. Performance of S-CAP compared to MutPred Splice. MutPred Splice is a computational method for predicting the pathogenicity of exonic synonymous variants. MutPred Splice was trained by its authors using a subset of the pathogenic data used to train/test S-CAP. As a result, we need to independently test MutPred Splice on a set of variants that was not used in its training. This test set comprises rare synonymous variants from HGMD added to the database in 2013 or later. On this set S-CAP improves on MutPred Splice by 5.7% when comparing the overall AUC. S-CAP performs especially well in the high sensitivity domain and improves on the MutPred Splice hsr-AUC by 204%.

Supplementary Figure 3



Supplementary Figure 3. S-CAP performance in the high sensitivity region. The hsr-AUC curve is formed by subsetting the overall AUC to just the region where pathogenic variants are correctly classified over 95% of the time. An hsr-AUC curve is calculated for each of the regions as defined in **Fig. 1c**. S-CAP improves on the next best method's hsr-AUC by 472% in the 3' intronic region (A), 31.1% in the 3' core sites (B), 61.7% in the exonic region (C), 0.28.2% in the 5' core sites (D), 70.2% in the 5' extended region (E) and 0.57.1% in the 5' intronic region (F). None of the existing methods consistently outperforms the others.

Supplementary Figure 4



Supplementary Figure 4. Performance on recessive and dominant classes. The distribution of the underlying features is dramatically different for dominant and recessive variants. This results in a big difference in performance when classifying recessive and benign variants in the core splice site regions. S-CAP achieves an AUC (A) of 0.779 on dominant tagged variants and (B) of 0.880 on recessive tagged variants in the 5' core region. There is a similar performance difference in the 3' core region where S-CAP achieves an AUC (C) of 0.805 on dominant tagged variants (D) and of 0.890 on recessive tagged variants. In the high sensitivity region, S-CAP achieves an hsr-AUC (E, F) of 0.222 on dominant tagged variants and of 0.518 on recessive tagged variants in the 5' core region (G, H) and of 0.296 on dominant tagged variants and of 0.454 of recessive tagged variants in the 3' core region. S-CAP AUC and hsr-AUC are consistently better than those of all other tools, and no existing method consistently outperforms the others.