Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders

Charlie Rowlands¹, Huw Thomas², Jenny Lord³, Htoo Wai³, Gavin Arno⁴, Glenda Beaman¹, Panagiotis Sergouniotis⁵, Beatriz Gomes-Silva¹, Christopher Campbell⁵, Nicole Gossan⁵, Claire Hardcastle⁵, Kevin Webb⁵, Christopher O'Callaghan⁴, Robert Hirst⁶, Simon Ramsden⁷, Elizabeth Jones⁵, Jill Clayton-Smith⁷, Andrew Webster⁴, Andrew Douglas⁸, Raymond T O'Keefe², William Newman¹, Diana Baralle³, Graeme Black², and Jamie Ellingford¹

¹University of Manchester
²The University of Manchester
³University of Southampton Faculty of Medicine
⁴University College London
⁵Central Manchester University Hospitals NHS Foundation Trust
⁶University of Leicester
⁷Manchester University NHS Foundation Trust
⁸University Hospital Southampton NHS Foundation Trust

October 1, 2020

Abstract

The development of computational methods to assess pathogenicity of pre-messenger RNA splicing variants is critical for diagnosis of human disease. We assessed the capability of eight algorithms, and a consensus approach, to prioritize 250 variants of uncertain significance (VUS) that underwent splicing functional analyses. It is the capability of algorithms to differentiate VUSs away from the immediate splice site as 'pathogenic' or 'benign' that is likely to have the most substantial impact on diagnostic testing. We show that SpliceAI is the best single strategy in this regard, but that combined usage of tools using a weighted approach can increase accuracy further. We incorporated prioritization strategies alongside diagnostic testing for rare disorders. We show that 15% of 2783 referred individuals carry rare variants expected to impact splicing that were not initially identified as 'pathogenic' or 'likely pathogenic'; 1 in 5 of these cases could lead to new or refined diagnoses.

Abstract:

The development of computational methods to assess pathogenicity of pre-messenger RNA splicing variants is critical for diagnosis of human disease. We assessed the capability of eight algorithms, and a consensus approach, to prioritize 250 variants of uncertain significance (VUS) that underwent splicing functional analyses. It is the capability of algorithms to differentiate VUSs away from the immediate splice site as 'pathogenic' or 'benign' that is likely to have the most substantial impact on diagnostic testing. We show that SpliceAI is the best single strategy in this regard, but that combined usage of tools using a weighted approach can increase accuracy further. We incorporated prioritization strategies alongside diagnostic testing for rare disorders. We show that 15% of 2783 referred individuals carry rare variants expected to impact splicing that were not initially identified as 'pathogenic' or 'likely pathogenic'; 1 in 5 of these cases could lead to new or refined diagnoses.

Keywords: splicing; rare disease; RNA; Mendelian disorders; variant interpretation

A number of computational tools have been developed to assist in the interpretation of genomic variation impacting splicing (Rowlands, Baralle, & Ellingford, 2019). These tools have been expanded recently to include an array of machine learning tools that have been trained to prioritize splice-disrupting variation through diverse means (Cheng et al., 2019; Jagadeesh et al., 2019; Jaganathan et al., 2019; Lee et al., 2017; Xiong et al., 2015). Here we compare the accuracy of nine *in silico* strategies, including eight state-of-the art algorithms and a consensus approach, to prioritize variants impacting splicing.

First, we ascertained and performed functional analyses for 250 VUSs to observe their impact on splicing (Table S1). To the best of our knowledge, this is the largest set of VUSs that have been functionally interrogated for impact on splicing as part of diagnostic services for individuals with rare disease. Variants had been identified in individuals undergoing genome sequencing and targeted gene panel analysis, with diverse phenotypic presentations including familial susceptibility to breast cancer (MIM #604370), syndromic disorders such as Marfan syndrome (MIM #154700) and isolated inherited retinal disorders such as retinitis pigmentosa (MIM #300029). The approaches for VUS functional analysis are described elsewhere (Wai et al., 2020) and in the Supporting Information. We observed that 80/250 (32%) of the VUSs significantly impacted splicing, and as a result could be reclassified as 'likely pathogenic' according to ACMG guidelines for variant interpretation (Richards et al., 2015). This reclassification resulted in new molecular diagnoses for individuals carrying these variants. All VUSs impacted regions outside of canonical splice acceptor and donor sites, and included examples of deeply intronic cryptic splice sites, exonic cryptic splice sites and branchpoint variants. In some cases, functional investigations demonstrated a range of consequences on mRNA splicing (Figure 1), reinforcing the concept that the precise effect of splicing variants is an important piece of evidence for consideration during clinical variant interpretation that, in the future, may enable refinements in appropriate targeted treatments (Bauwens et al., 2019; Shen & Corey, 2018).

We obtained in silico prediction scores for each of the 250 functionally assessed variants using eight in silicoprioritization algorithms (Table S1) and calculated sensitivity, specificity and receiver operating characteristic area under the curve (AUC), observing significantly variable performances (Figure 2). Pairwise statistical comparisons of AUC for the 250 functionally assessed VUSs, after Bonferroni correction for multiple testing, demonstrated that SpliceAI outperformed other single algorithm approaches (Figure 2: Table S2). The AUC analysis for single algorithms calculated the optimal score for each of the algorithms to distinguish between true positives (80 variants shown to impact splicing in our functional assays) and true negatives (170 variants shown not to impact splicing in our functional assays) in this dataset. We acknowledge that splicing machinery may be influenced by cell-/tissue-specific factors which are outside the scope of assays performed here (Aicher, Jewell, Vaquero-Garcia, Barash, & Bhoj, 2020; Cummings et al., 2020; Vig et al., 2020), and variants may have pathogenic impacts on gene expression and/or regulation without any detrimental impact on splicing (Castel et al., 2018; Evans et al., 2018; Short et al., 2018; Zhang, Wakeling, Ware, & Whiffin, 2020). Such factors will influence comparative metrics between algorithms, and future investigations may uncover pathogenic roles for variants reported here. However, the optimal thresholds calculated in light of these limitations for the 250 functionally assessed VUSs in this study are reported in Table S3.

Global approaches to variant analysis, as assessed through the AUC, may fail to capture region-specific intricacies in splicing disruption (Jagadeesh et al., 2019). For example, variants could be sub-divided by their pathogenic mechanism, their effect on pre-mRNA splicing, their predicted molecular consequence or the location of the variant with respect to known splicing motifs, and each of these sub-groups may require different approaches or thresholds for accurate prioritization of pathogenic variation. We therefore predicted variants to be 'disruptive' or 'undisruptive' according to pre-defined thresholds, utilizing region-specific thresholds where appropriate (Table S4), and compared accuracy of each of the prioritization strategies across 2000 iterations of sampling with replacement. We utilized a single score threshold for tools where region-specific thresholds have not been previously identified (Table S4). This analysis highlighted differences across the tools and significantly differentiated their ability to accurately predict pathogenicity (Kruskal Wallis, df=8, p<0.0001; Figure 2c-d). Similar to the AUC analysis, SpliceAI (using a threshold of 0.2) was significantly the best performing strategy across all assessed single algorithms for our set of analysed VUSs (Kruskal Wallis, p<0.0001 for all pairwise comparisons of accuracy between SpliceAI and other tools; Figure 2c-d).

To determine if combining one or more of these metrics could achieve greater accuracy than prioritization scores in isolation, we developed a consensus score for each variant which considered the region-specific thresholds for each tool (Table S4, consensus score range = 0.8 tools predicting splicing disruption). We observed that the consensus approach performed similarly to SpliceAI when assessed through the receiver operating characteristic AUC (Figure 2; Tables S2-S3). The consensus approach (using a threshold of 4/8 algorithms supporting splicing disruption) also performed more similarly to SpliceAI than other strategies when measuring accuracy across sampling iterations (Figure 2c), but was less frequently the best performing approach (Figure 2d). To understand if the relative scores from each algorithm could assist interpretation we developed a novel metric which incorporates weighted scores from the prioritization strategies (Supporting Information). This analysis considered the actual score of the variant relative to the maximum score possible from each prediction algorithm (Supporting Information). Of note, the weighted approach considering scores from SpliceAI and a consensus approach performs better than these two approaches in isolation (Figure 2b; Table S3). Although not mutually exclusive and underpowered to detect significant statistical differences in the AUC from this combined analysis - due to marginal gains in accuracy and sample size - this demonstrates the potential utility of combined approaches utilizing combinations of scores to improve accuracy for the identification of variants impacting splicing.

Next, we sought to examine the impact of these approaches on clinical variant analysis. Therefore we integrated region-specific prioritization strategies (Table S4) into an accredited diagnostic service for 2783 individuals with rare diseases (Ellingford et al., 2016). All individuals included in this analysis have received genetic testing for rare disease within the UK national healthcare service through a clinically accredited laboratory. We calculated in silico scores for 20,617 variants (18,013 of which were rare), observed a total of 1,346,744 times in the cohort. We observed substantial variability in the number of rare variants prioritized by each *in-silico* tool (Figure 3a; Table S5) and in the specific variants prioritized by the most correlated *in*silico splicing tools (Figure 3b). We observed that while variants which show the highest consensus between in silico splicing tools impact the canonical splice site (Figure 3c; Table S6), 99% (n = 17,871) of variants analysed impact exonic or intronic regions of genes outside of the canonical splice sites. Splicing variants are often considered as a single class of variants and canonical splice site variants are therefore highly susceptible to over-prioritization by in silico tools, as such variant represents the majority ($^{70\%}$) of known splicing pathogenic variants (Krawczak et al., 2007; Stenson et al., 2014; Xiong et al., 2015). Our data further underline the need to develop effective and unbiased strategies for prioritizing variants impacting splicing outside of the canonical splice sites, and this will be especially important for VUS in known disease genes. Overall, these data demonstrate that different *in silico* strategies for splicing variant prioritization will alter the burden of variant analysis for clinical scientists. This is an important consideration for the analytical specificity and the throughput of diagnostic testing.

To assess the clinical impact of such strategies, we integrated a single prioritization strategy, SpliceAI, in parallel to outcomes from routine diagnostic testing. This analysis involved extensive curation of genomic findings for the 2783 referred individuals, all of which were classified in accordance with ACMG guidelines by clinically accredited scientists. We added SpliceAI predictions alongside these analyses and observed that this approach influenced analysis for 420 (15%) individuals receiving genomic testing for rare disease, and could result in new or refined molecular diagnoses in 81 (3%) cases. Overall, we prioritized 758 variants (528 unique variants) in 646 individuals (23% of cohort) with a range of predicted molecular consequences. Most (99.6%, 526/528) variants were prioritized by at least one other *in silico* tool (Table S7). The strength of the score from SpliceAI correlated highly with prioritization from other *in silico* tools (Figure 3d) and differed between regions of genome that were impacted (Table S8). We defined prioritized variants as:

• New , variant not previously highlighted or reported through diagnostic testing;

- *Clarified*, variant previously reported through diagnostic testing but pathogenicity or pathogenic mechanism was unclear;
- *Reported*, variant already described or established as 'pathogenic' or 'likely pathogenic' through diagnostic testing.

In this regard, we identified 379 *new* variants in 337 individuals, 87 *clarified* variants in 83 individuals and 292*reported* variants in 274 individuals. We found most (91%, 697/758) variants to be in genes known as a recessive cause of genetic disease. To understand if these variants impacted normal splicing, we interrogated the GTEx datasets (GTEx, 2013) for individuals carrying variants in a heterozygous state, identifying 40 carriers of variants prioritized by this analysis. Of these, 21 had suitable RNAseq datasets available for evaluation, and we were able to clearly observe significant alterations to splicing in four cases (Table 1). Whilst most variants will require bespoke functional investigations to establish precise effects on splicing and protein synthesis, leveraging the use of publicly available datasets for individuals carrying potentially pathogenic rare variants in the GTEx dataset can quickly increase certainty of variant impact and refine clinical variant analysis.

The incorporation of the prioritization and functional strategies described in this study for variants impacting splicing significantly improved molecular diagnostic services. However, we expect that the true impact of such analysis strategies will be more profound. Targeted next generation sequencing approaches employed within this large cohort ignore deeply intronic regions of genes, which, as shown here (Box 1, Case Example) and in other studies (den Hollander et al., 2006; Montalban et al., 2019; Sangermano et al., 2019), can harbor variants which result in aberrant splicing through the production of novel cryptic exons. The recent availability of genomic datasets within healthcare amplifies the current limitations in interpreting variation within the non-coding genome, particularly in large genome sequencing cohorts. Our findings demonstrate the opportunity to expand bioinformatics analysis to the pre-mRNA regions of known disease genes and provide immediate increases to diagnostic yield. Moreover, we demonstrate a requirement to functionally assess variant impact on pre-mRNA splicing as the delineation of the precise effects may be important in considerations for variant pathogenicity. The prioritization and identification of pathogenic variants impacting splicing is therefore an important consideration for diagnostic services and for the development of new targeted treatments.

Figure Legends

Figure 1. Results from *in-vitro* minigene assays demonstrating multiple consequences as a result of variants proximal to the canonical splice site. Left, gel electrophoresis snapshots of cDNA products amplified from primers designed for control exons within the minigene (exon 1 & exon 2). All prominent bands were cut out and Sanger sequenced. Right, solid red blocks illustrate alignment of sequenced cDNA transcripts to features within the minigene vector: control exons (grey boxes) and inserted exons (*purple boxes*). (a) SCN2A c.2919+3A>G, showing complete exon exclusion and exon truncation in minigene vectors containing the c.2919+3A>G variant (top two alignments) and normal splicing in minigene vectors containing the WT sequence (bottom alignment). The first resulted in a transcript with a truncated exon, NM_001040142.1:r.2563_2710del, and the second resulted in a complete exon skip, NM_001040142.1:r.2563_2919del. While we interpreted both events as 'likely pathogenic' it is noteworthy that these events were considered differently using ACMG criteria; the exon truncation event resulted in a frameshift and introduction of a premature stop codon (PVS1), whereas the complete exon skipping event resulted in the inframe removal of 119 amino acids from the transcript (PM4).(b) MERTK c.2486+6T > A, showing a shifting of the exon included in the reading frame in minigene vectors containing the c.2486+6T>A variant (top alignment) and normal splicing in minigene vectors containing the WT sequence (*bottom*alignment). This novel variant is present in two individuals with severe rod-cone dystrophy, and resulted in the simultaneous usage of a cryptic exonic splice acceptor site and a cryptic intronic splice donor site creating a novel exon (chr2: 112,779,939-112,780,082, GRCh37), and a premature stop codon in the penultimate exon, p.(Trp784Valfs*10).

Figure 2. Comparison of in silico strategies to prioritize 250 variants of uncertain signifi-

cance with functional investigations performed. (a) Receiver operating characteristics area under the curve (AUC) comparisons for nine *in silico*prioritization strategies demonstrating that SpliceAI (AUC=0.95, 95%CI=0.93-0.98) and a consensus approach (AUC=0.94, 95%CI=0.91-0.96) outperform other strategies for prioritization. (b) AUC comparisons between SpliceAI, a consensus approach and a novel metric, demonstrates that a weighted approach slightly increases accuracy of prioritization over single approaches alone (AUC=0.96, 95%CI=0.94-0.98). (c-d) Accuracy comparisons of each *insilico* prioritization approach across 2000 bootstraps utilizing region-specific pre-defined thresholds: (c) violin plot demonstrating the calculated accuracy of each*in-silico* prioritization approach; (d) frequency that each strategy is the best or joint best performing.

Figure 3. Summary of the overlap and correlations observed between the scores from *in silico* splicing prediction algorithms for 18,013 unique rare variants identified in a large cohort of 2783 individuals with rare disease undergoing genetic testing, specifically for syndromic and non-syndromic inherited retinal disorders. (a) Bar chart showing overall count of unique variants prioritized using pre-defined thresholds for each *in silico* prediction algorithm. (b) Overlap between the unique variants prioritized by the five most correlated *in silico* prediction tools. (c) Grouped bar chart the variant impacts, as defined by Jagadeesh *et al*. (d) Correlation between SpliceAI score and the number of additional tools also prioritizing the variant for the 528 unique rare variants prioritized by SpliceAI.

Bibliography

Aicher, J. K., Jewell, P., Vaquero-Garcia, J., Barash, Y., & Bhoj, E. J. (2020). Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet Med*, 22 (7), 1181-1190. doi:10.1038/s41436-020-0780-y

Bauwens, M., Garanto, A., Sangermano, R., Naessens, S., Weisschuh, N., De Zaeytijd, J., . . . De Baere, E. (2019). ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants. *Genet Med, 21* (8), 1761-1771. doi:10.1038/s41436-018-0420-y

Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., . . . Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet*, 50 (9), 1327-1334. doi:10.1038/s41588-018-0192-y

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Celik, M. H., Fairbrother, W. G., Avsec, Z., & Gagneur, J. (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol*, 20 (1), 48. doi:10.1186/s13059-019-1653-z

Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., Seaby, E. G., Watts, N. A., Singer-Berk, M., . . . Consortium, G. A. D. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581 (7809), 452-458. doi:10.1038/s41586-020-2329-2

den Hollander, A. I., Koenekoop, R. K., Yzer, S., Lopez, I., Arends, M. L., Voesenek, K. E., . . . Cremers, F. P. (2006). Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet*, 79 (3), 556-561. doi:10.1086/507318

Ellingford, J. M., Barton, S., Bhaskar, S., O'Sullivan, J., Williams, S. G., Lamb, J. A., . . . Black, G. C. M. (2016). Molecular findings from 537 individuals with inherited retinal disease. *J Med Genet*, 53 (11), 761-767. doi:10.1136/jmedgenet-2016-103837

Evans, D. G. R., van Veen, E. M., Byers, H. J., Wallace, A. J., Ellingford, J. M., Beaman, G., . . . Newman, W. G. (2018). A Dominantly Inherited 5' UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer. Am J Hum Genet, 103 (2), 213-220. doi:10.1016/j.ajhg.2018.07.002

GTEx. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet, 45 (6), 580-585. doi:10.1038/ng.2653

Jagadeesh, K. A., Paggi, J. M., Ye, J. S., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2019). S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet*, 51 (4), 755-763. doi:10.1038/s41588-019-0348-4

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., . . . Farh, K. K. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176 (3), 535-548 e524. doi:10.1016/j.cell.2018.12.015

Krawczak, M., Thomas, N. S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., & Cooper, D. N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat*, 28 (2), 150-158. doi:10.1002/humu.20400

Lee, M., Roos, P., Sharma, N., Atalar, M., Evans, T. A., Pellicore, M. J., . . . Cutting, G. R. (2017). Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites. *Am J Hum Genet*, 100 (5), 751-765. doi:10.1016/j.ajhg.2017.04.001

Montalban, G., Bonache, S., Moles-Fernandez, A., Gisbert-Beamud, A., Tenes, A., Bach, V., . . . Gutierrez-Enriquez, S. (2019). Screening of BRCA1/2 deep intronic regions by targeted gene sequencing identifies the first germline BRCA1 variant causing pseudoexon activation in a patient with breast/ovarian cancer. *J Med Genet*, 56 (2), 63-74. doi:10.1136/jmedgenet-2018-105606

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, 17 (5), 405-423. doi:10.1038/gim.2015.30

Rowlands, C. F., Baralle, D., & Ellingford, J. M. (2019). Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing. *Cells*, 8 (12). doi:10.3390/cells8121513

Sangermano, R., Garanto, A., Khan, M., Runhart, E. H., Bauwens, M., Bax, N. M., . . . Cremers, F. P. M. (2019). Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides. *Genet Med*, 21 (8), 1751-1760. doi:10.1038/s41436-018-0414-9

Shen, X., & Corey, D. R. (2018). Chemistry, mechanism and clinical status of antisense oligonucleotides and duplex RNAs. *Nucleic Acids Res*, 46 (4), 1584-1600. doi:10.1093/nar/gkx1239

Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., . . . Hurles, M. E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, 555 (7698), 611-616. doi:10.1038/nature25983

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D., & Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133 (1), 1-9. doi:10.1007/s00439-013-1358-4

Vig, A., Poulter, J. A., Ottaviani, D., Tavares, E., Toropova, K., Tracewska, A. M., . . . Consortium, G. E. R. (2020). DYNC2H1 hypomorphic or retina-predominant variants cause nonsyndromic retinal degeneration. *Genet Med*. doi:10.1038/s41436-020-0915-1

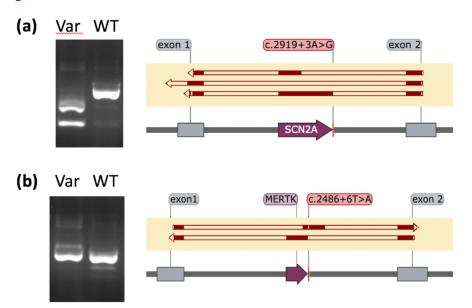
Wai, H. A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., . . . group. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med*. doi:10.1038/s41436-020-0766-9

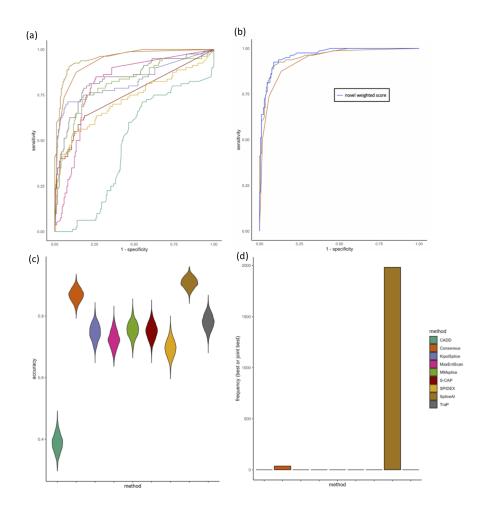
Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., . . . Frey, B. J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347 (6218), 1254806. doi:10.1126/science.1254806

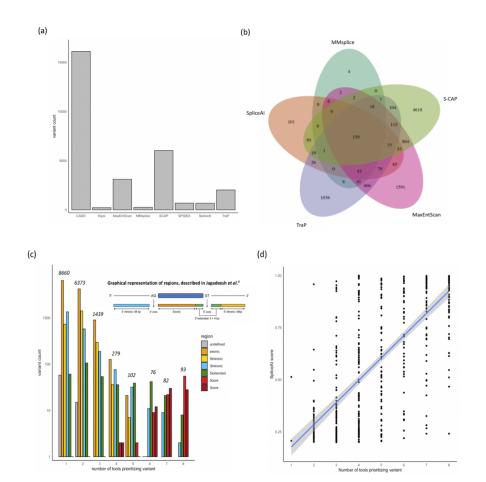
Zhang, X., Wakeling, M., Ware, J., & Whiffin, N. (2020). Annotating high-impact 5'untranslated region variants with the UTRannotator. *Bioinformatics*. doi:10.1093/bioinformatics/btaa783

Hosted file

Table 1.pdf available at https://authorea.com/users/363411/articles/484210-comparison-ofin-silico-strategies-to-prioritize-rare-genomic-variants-impacting-rna-splicing-for-thediagnosis-of-genomic-disorders







https://doi.org/10.22541/au.160157595.59675486 — This a preprint and has not been peer Posted on Authorea 1 Oct 2020 -The copyright holder is the author/funder. All rights reserved. No reuse