A systematic assessment of the impact of rare canonical splice site variants on splicing using functional and *in silico* methods

Authors: Rachel Y. Oh[1,2,*], Ali AlMail[2,3,*], David Cheerie[3,4], George Guirguis[3,4], Huayun Hou[3], Kyoko E. Yuki[3,5], Bushra Haque[3,4], Bhooma Thiruvahindrapuram[6], Christian R. Marshall[5,7], Roberto Mendoza-Londono[1,3,8], Adam Shlien[3,4,5,7], Lianna G Kyriakopoulou[5,7], Susan Walker[6], James J. Dowling[3,4,8,9], Michael D. Wilson[3,4], Gregory Costain[1,3,4,8,#]

*Denotes equal contribution
[#]Author for correspondence

Affiliations:
[1]Division of Clinical and Metabolic Genetics, Hospital for Sick Children, Toronto, Canada
[2]Temerty Faculty of Medicine, University of Toronto, Toronto, Canada
[3]Program in Genetics and Genome Biology, SickKids Research Institute, Toronto, Canada
[4]Department of Molecular Genetics, University of Toronto, Toronto, Canada
[5]Division of Genome Diagnostics, Hospital for Sick Children, Toronto, Canada
[6]The Centre for Applied Genomics, SickKids Research Institute, Toronto, Canada
[7]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada
[8]Department of Paediatrics, University of Toronto, Toronto, Canada
[9]Division of Neurology, Hospital for Sick Children, Toronto, Canada

Contact information:
Gregory Costain, MD, PhD
The Hospital for Sick Children 555 University Avenue Toronto, ON M5G1X8, CANADA
Phone: 416-813-7654 ext. 301480
E-mail: gregory.costain@sickkids.ca

Main text word count: 2826

Abstract

Background/Objectives: Canonical splice site variants (CSSVs) are often presumed to cause loss-of-function (LoF) and are assigned very strong evidence of pathogenicity (according to ACMG criterion PVS1). However, the exact nature and predictability of splicing effects of unselected rare CSSVs in blood-expressed genes is poorly understood.

Methods: A total of 184 rare CSSVs in unselected blood-expressed genes were identified by genome sequencing in 121 individuals, and their impact on splicing was interrogated manually in RNA sequencing (RNA-seq) data. Blind to these RNA-seq data, we attempted to predict the precise impact of CSSVs by applying *in silico* tools and the ClinGen Sequence Variant Interpretation Working Group 2018 guidelines for applying PVS1 criterion.

Results: There was no evidence of a frameshift nor of reduced expression consistent with nonsense-mediated decay (NMD) for 24% of CSSVs: 17% had wildtype splicing only and normal junction depths, 3.25% resulted in cryptic splice site usage and in-frame indels, 3.25% resulted in full exon skipping (in-frame), and 0.5% resulted in full intron inclusion (in-frame). Misclassification rates for splicing outcome (frameshift/NMD vs. no frameshift/no NMD) using (i) SpliceAI, (ii) MaxEntScan, and (iii) AutoPVS1 ranged from 30-41%, with none outperforming a simple “zero rule” classifier.

Conclusion: Nearly 1 in 4 CSSVs may not cause LoF based on analysis of RNA-seq data. Predictions from *in silico* methods were often discordant with findings from RNA-seq. More caution may be warranted in applying PVS1-level evidence to CSSVs in the absence of functional data.

Introduction

Canonical splice site variants (CSSVs) are DNA variants affecting splicing donor (+1 and +2) and acceptor (-1 and -2) sites defining exon-intron boundaries[1,2]. The consensus nucleotide sequences at splicing donor and acceptor sites are GT and AG, respectively, and are essential in interacting with the U2 spliceosome to result in normal splicing and generation of wildtype (WT) transcripts[2–5]. CSSVs may modify the interactions between the precursor messenger RNA (pre-mRNA) and spliceosome complex[5-7]. The resulting splice disruption events can include exon skipping, full intron inclusion, and alternative use of nearby cryptic splice sites resulting in insertions or deletions (indels) of nucleotides[6–8]. These effects may or may not induce a frameshift and premature termination codon (PTC), which can then trigger nonsense-mediated RNA decay (NMD) and result in a loss-of-function (LoF) of the gene[9,10].

Accurate variant interpretation is foundational to both genome diagnostics and screening[11–13]. Rare CSSVs are typically considered under the "null variant" code and assigned very strong evidence for pathogenicity (PVS1)[12]. The PVS1 guidelines were refined by the ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) in 2018 and again recently in 2023[11,14]. ClinGen SVI initially recommended assigning PVS1 at varying evidence strengths (i.e., supporting, moderate, strong, very strong, or not at all) after directly inspecting the genomic region to predict the impact of the CSSV on splicing and the overall reading frame[11]. SpliceAI subsequently emerged as a widely used and powerful method for annotating genetic variants with their predicted effect on splicing[15,16]. However, *in silico* tools continue to be recognized as a valuable but imperfect adjunct method for predicting the impact of CSSVs[14].

To our knowledge, there have been no systematic attempts to catalogue the precise consequences of rare CSSVs on splicing using functional data. The degree to which the impact(s) of CSSVs are predictable via inspection of the genomic region and application of *in silico* tools is also unclear. Here, we analyzed all rare CSSVs identified by genome sequencing (GS) from children and parents, in blood-expressed but otherwise unselected genes, using genome-wide RNA sequencing (RNA-seq). We hypothesized that a significant minority of CSSVs would show no evidence of a frameshift nor of reduced expression consistent with nonsense-mediated decay, recognizing that ~1/3 of inserted or deleted DNA segments would be expected to have a size divisible by 3[17]. We determined the proportions of various splicing outcomes, and we compared the results to outcomes predicted in a blinded fashion by three *in silico* tools: SpliceAI, MaxEntScan, and AutoPVS1[12,13,15]. Our results revealed a previously underappreciated complexity in CSSV impact prediction, and underscore the value of functional data in the interpretation of rare CSSVs.

Materials and Methods:

Identification of rare CSSVs expressed in blood

In this study, we analyzed blood-derived GS and RNA-seq data from a convenience sample of 121 total individuals[18,19]. The GS and RNA-seq studies were approved by the Research Ethics Board at The Hospital for Sick Children. Parents and/or guardians provided written consent on their child's behalf. Where appropriate, children provided written and/or oral assent. Demographic details for the cohort are described in Table S1. Detailed GS methods, and a subset of the GS data, were published previously[18–22]. DNA variants from GS files were filtered according to the following criteria: (I) single nucleotide substitution in a canonical splice site identified in a MANE or Ensembl Canonical transcript, (II) allele frequency of <0.05 (per 1000 Genomes, NHLBI-ESP, and ExAC), (III) >=99% Genotype Quality Score, and (IV) gene possibly detected in whole blood (according to GTEx, V8, transcripts per million (TPM) >0.05) and detected in our internal cohorts[20,23]. Variants in untranslated regions (UTRs) were excluded[23]. All 184 CSSVs included in this study are listed in Table S2. Two of the identified CSSVs were considered diagnostic for the phenotype(s) that prompted GS, and all CSSVs were heterozygous[18–22,25].

Analysis of splicing impact of CSSVs using RNA-seq data

We analyzed the impact of CSSVs on splicing in the canonical transcript using the accompanying short-read blood-derived RNA-seq data. RNA extraction, sequencing, and data processing methods were previously described in full[20,26]. Each splicing junction was manually inspected using the Integrated Genome Visualizer (IGV) by two independent evaluators (R.Y.O. and A.A.). For every CSSV, an average of five random, age-range matched "controls" (i.e., with

normal DNA sequence at the affected CSS) from this cohort were used to identify the WT splicing event(s) and provide a reference on junction read depths to account for any possible fluctuations. Sex-matched "controls" were specifically used for one CSSV located on the X chromosome. Only junctions with >5 uniquely mapped reads, which is a low cut-off, were considered in the analysis. The junction with the highest read depth was considered the predominant splicing outcome. Splicing outcome categories are as follows:

(A) presumed NMD (if the raw WT splicing junction coverage was >=20% less than control individuals but no aberrant splicing events were captured)[26],

(B) NMD not detected [if there was comparable (<=20% difference) junction depth between individuals and no aberrant splicing events were captured],

(C) exon skipping leading to frameshift deletion,

(D) exon skipping leading to in-frame deletion,

(E) full intron inclusion leading to frameshift insertion,

(F) full intron inclusion leading to in-frame insertion,

(G) activation of a cryptic splice site leading to frameshift indel,

(H) activation of a cryptic splice site leading to in-frame indel.

Prediction of splicing outcomes using in silico tools

Blind to the RNA-seq data, two independent assessors (D.C. and G.G.) attempted to predict the precise impact of CSSVs using *in silico* tools and ClinGen SVI recommendations[11]. Standard score thresholds were used for MaxEntScan (>3) and for SpliceAI (delta score >0.2). For each CSSV, the splice junction was manually inspected using Alamut Visual Plus (version v1.7, ©2022 SOPHiA GENETICS) with a +/-20 bp window for a cryptic splice site. In the case of two

or more cryptic splice sites in the same region with comparable *in silico* scores, if the use of any one site was predicted to result in an in-frame indel then NMD was not predicted as the outcome. In the absence of a cryptic splice site in the neighboring region, for acceptor losses we predicted exon skipping and for donor losses we predicted full intron inclusion. The length of the exon or intron, respectively, was then used to predict whether the impact would be in-frame or out-of-frame. While the ClinGen SVI recommendations specify use of a +/-20 bp window, we conducted additional exploratory analyses using an expanded SpliceAI window of +/-5000 bp ("SpliceAI_expanded"). Concordance of outcomes was then calculated if the frameshift/non-frameshift outcome predicted by the *in silico* tools matched the results from RNA-seq.

AutoPVS1, an automatic classification tool for PVS1 interpretation of null variants, was also used to predict the impact of CSSVs on splicing[27]. This algorithm uses Variant Effect Predictor for annotation of variants and MaxEntScan to predict the use of cryptic splice sites and resulting impact on splicing[27]. The three possible outcomes of AutoPVS1 are (I) Exon skipping or cryptic splice site usage that leads to a frameshift and NMD, (II) Exon skipping or cryptic splice site usage that leads to a frameshift without NMD, and (III) Exon skipping or cryptic splice site usage that preserves the reading frame.

Results

Comparison of frameshift outcomes of CSSVs using RNA-seq vs. in silico predictions

We assessed a total of 184 rare, blood expressed CSSVs in 180 otherwise unselected genes. By RNA-seq, 24% of these CSSVs did not result in a frameshift nor in reduced expression consistent with NMD (Figure 1a). There was no apparent difference in the patterns of variant location and specific nucleotide substitution of CSSVs by frameshift/non-frameshift outcome (Figures S1 and S2). Considering the n=18 CSSVs that showed only WT splicing and with comparable read depth to controls (outcome category B), 14 CSSVs were in the donor splice site including only 1 GT>GC variant[28] (Figure S3).

For the 184 CSSVs, blinded predictions from *in silico* methods for non-frameshift outcomes are 28% (SpliceAI), 29% (SpliceAI_expanded), to 43% (AutoPVS1) of all CSSVs (Figure 1a). For CSSVs resulting in a frameshift per RNA-seq, SpliceAI_expanded had the highest pairwise concordance (77%), followed by SpliceAI (74%), MaxENTScan (69%), and AutoPVS1 (61%; Figure 1b). For CSSVs not resulting in a frameshift per RNA-seq, SpliceAI was concordant in 36%, MaxENTScan in 41%, SpliceAI_expanded in 48%, and AutoPVS1 in 55% (Figure 1b). Across all 184 variants, SpliceAI_expanded had the highest pairwise concordance to RNA-seq with respect to the frameshift vs. no-frameshift outcome, at 71%. In order to assess the performance of each *in silico* method, we calculated misclassification rates from each technique, which ranged from 30-41% (Figure 1c). In comparison, a zero-rule classifier (which would predict that every CSSV causes frameshift/NMD, the predominant outcome category in this study) had a misclassification rate of only 24% (Figure 1c).

Comparison of specific splicing outcomes of CSSVs using RNA-seq vs. in silico predictions

Next, we compared the specific splicing outcome of CSSVs (cryptic splice site usage, exon skipping, or intron inclusion; categories C-H above) between RNA-seq and *in silico* predictions, for the n=23 variants where this could be determined from RNA-seq (Figure 2a). Total pairwise concordance was 70% (16/23) for SpliceAI_expanded (improved from 35% (8/23) for SpliceAI) and 26% (6/23) for MaxENTScan; AutoPVS1 does not provide such predictions. The performance of both *in silico* methods appeared to vary by outcome category, e.g., with SpliceAI_expanded correctly predicting all usage of cyptic splice sites (10/10), some of the exon skipping events (5/10) and none of the intron inclusion events (0/3) (Figure 2b). Two selected donor CSSVs are used to illustrate that *in silico* were often correct in predicting frameshift vs. no frameshift outcomes, however, for incorrect and/or incomplete mechanisms of abnormal splicing (Figure 3a-d).

Discussion

Challenging common assumptions and interpretation heuristics for CSSVs

As demostrated in this study cohort, most human genomes contain one or more rare CSSVs in blood-expressed genes. Although there is widespread recognition that a rare CSSV need not necessarily result in LoF, experiences from decades of targeted diagnostics-focused clinical genetic testing (with a resulting ascertainment bias) may have contributed to a misconception that CSSVs are comparable to nonsense and frameshift variants. We present the first systematic assessment of the impact of rare "unselected" CSSVs in blood-expressed genes, using RNA-seq. We found that nearly 1 in 4 CSSVs may not cause LoF, and that *in silico* predictions using established tools and published guidelines were often discordant with RNA-seq data.

In recent years, there have been numerous computational tools developed to predict the location of novel splice sites and thus the impact of DNA variants on splicing[12,15]. These tools were validated using few if any CSSVs (e.g., near intronic variants at least 3 nucleotides from a canonical exon boundary in SpliceAI-10k; n=55 in 300K-RNA Top-4) and were not created to predict specific CSSV outcomes like exon skipping or intron inclusion[15,16,29,30]. A prior report found that intron inclusion was poorly predicted using SpliceAI when compared with transcriptome sequencing data[31]. AutoPVS1 does not list intron inclusion as a specific outcome, nor does it distinguish which variants result in exon skipping versus cryptic splice site usage[27]. For now, no *in silico* methods appear to predict the precise impact of CSSVs with the sensitivity and specificity needed for clinical diagnostics. A recent study has proposed using the ACMG/AMP PM4 (moderate evidence of pathogenicity) criterion for CSSVs that are predicted to result in intron inclusion and 3 or more in-frame events as predicted by 300K-RNA Top-4[16].

Relative weighting of biological function or evolutionary conservation of the affected gene region(s) may need to be considered in addition to the length of the in-frame disruption[16]. We propose, in addition, that future updates to published guidelines on the use of PVS1 should consider the use of SpliceAI with an expanded window of +/-5000 bp.

The impact of CSSVs on splicing can be complex. Multiple and/or partial effects on splicing have been observed in the past (e.g., with some aberrantly spliced transcripts resulting in LoF and others showing no apparent impact or producing a functional transcript)[24]. Surprisingly, some CSSVs in our data showed no direct (aberrant/non-WT splice junctions) or indirect (reduced read depth, compatible with NMD) impact on splicing. The ACMG/AMP rule, BS3, may be applied to CSSVs with evidence of normal splicing patterns demonstrated by RNA-seq fulfilling specific criteria[24,32]. A recent study using cell culture-based full-length gene splicing assays demonstrated that specific nucleotide substitutions (GT>GC in the donor splice site) can generate WT transcript levels in an estimated 15-18% of cases; no other nucleotide substitutions in the +2 donor splice site were able to generate WT transcripts[28]. Moreover, WT splicing as a result of the 5' splice site GT>GC substitutions was not accurately predicted by *in silico* tools[28]. Our results showed that WT transcripts can be generated with diverse nucleotide substitutions, in no consistent ranking order in the 5' donor splice site as recently described (though this study performed RNA-seq in fibroblast samples and also included common variants), affecting +/- 1 and 2 canonical splice sites, highlighting *in vitro* assays' inability to capture the full complexity of splicing in human cells (Figure S3)[33]. Of note, sensitivity of our RNA-seq methods for detecting evidence of NMD will be <100%. Long-read RNA-seq might have facilitated more

robust quantitation of transcript isoforms and revealed additional splicing outcomes. Testing additional tissues beyond blood was also outside the scope of this initial study.

Our study has several additional limitations. First, we recognize that in-frame insertions and deletions can still result in non-functional proteins (e.g., through disruption of an essential protein domain) and that protein function cannot be inferred completely from RNA-seq. In the absence of any evidence-based guidance, we assumed for all *in silico* predictions that exon skipping would be the typical impact of an acceptor site loss and that intron retention/inclusion would be the typical impact of a donor site loss, whenever an alternative/cryptic splice site was not present. There is growing appreciation that this reasoning is overly simplistic[31,34]. The landscape of impacts of rare CSSVs may change based on tissue and/or age. Last, we were underpowered in this study to identify substitution- and location/motif-specific predictors of splicing outcomes.

Implications for diagnostics, predictive testing, and screening

These data reinforce prior expert consensus recommendations that cautioned against applying PVS1 to CSSVs in the absence of additional supportive evidence[24]. Recent ClinGen SVI recommendations emphasize the importance of annotating physiologically occurring alternative splicing events (or "leaky" splicing events) as candidate rescue transcripts that may not result in LoF, consideration of gene-specific details (e.g., the impact on critical regions), and generating more data to guide the integration of various lines of evidence (i.e., computational and *in vitro* assays to assess impact on splicing), with the end goal of improved navigation of equivocal clinical scenarios requiring interpretation of rare CSSVs[34]. The advantages and limitations of

RNA-seq, which can be done high-throughput, should be weighed against a targeted approach like RT-PCR; the latter may have greater sensitivity in detecting mis-spliced reads of low read depth as a result of low gene expression in whole blood, and being able to distinguish partial from complete abnormal splicing[33]. Although a majority of CSSVs result in LoF, this assumption should be questioned when genome-wide sequencing identifies novel rare CSSVs in genes associated with ultra-rare, poorly characterized conditions with non-specific phenotypes (e.g., developmental delays)[35]. The nuances of CSSV interpretation take on added importance when the pre-test probability is low and phenotypes are absent, as is the case with most secondary findings and in newborn genomic screening programs. Last, confirmation of genetic diagnosis and pathomechanism will be foundational to the coming wave of clinical trials of genetic therapies[36].

Declaration of interest

S.W. is currently an employee of Genomics England Limited. The other authors declare no competing interests.

Acknowledgements

We gratefully acknowledge all the individuals and their families who participated in this study. We thank the many healthcare providers involved in the diagnosis and care of the study participants. Special thanks to all staff affiliated with the Complex Care Program and The Centre for Applied Genomics. M.D.W. was supported by the Canada Research Chairs Program. Funding was provided by Genome Canada (OGI-158; M.D.W., A.S., and J.J.D.), the SickKids Centre for Genetic Medicine and Translational Genomics Node, the Sickkids Research Institute, the Canadian Institutes of Health Research (Funding Reference Number: PJT186240), and the University of Toronto McLaughlin Centre.

Web resources

SpliceAI: https://spliceailookup.broadinstitute.org

Data availability

GS and RNA-seq data were published and shared previously. The *in silico* predictions can be shared upon request to the corresponding author.

Author contributions

Conceptualization: R.Y.O., A.A., S.W., G.C.; Data curation: R.Y.O., A.A., D.C., G.G., B.H., G.C.; Formal analysis: R.Y.O., A.A., D.C., G.G., G.C.; Funding acquisition: C.R.M., R.M-L., A.S., L.G.K., J.J.D., M.D.W., G.C.; Investigation: R.Y.O., A.A., D.C., G.C.; Methodology: R.Y.O., A.A., D.C., G.G., H.H., K.Y., S.W., G.C.; Project administration: R.Y.O., A.A., T.K., B.T., G.C.; Resources: C.R.M., R.M-L., A.S., L.G.K., J.J.D., M.D.W., G.C.; Software: R.Y.O., A.A., D.C., G.G., H.H., K.Y.; Visualization: R.Y.O., A.A., D.C., H.H.; Writing-original draft: R.Y.O., A.A., G.C.; Writing-review & editing: all other authors

References

1. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet*. 1992;90(1-2):41-54. doi:10.1007/BF00210743

2. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*. 2000;28(21):4364-4375. doi:10.1093/nar/28.21.4364

3. Hastings ML, Krainer AR. Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol*. 2001;13(3):302-309. doi:10.1016/s0955-0674(00)00212-x

4. Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol*. 2014;15(2):108-121. doi:10.1038/nrm3742

5. Rogalska ME, Vivori C, Valcárcel J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat Rev Genet*. 2023;24(4):251-269. doi:10.1038/s41576-022-00556-8

6. Krawczak M, Thomas NST, Hundrieser B, et al. Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Human Mutation*. 2007;28(2):150-158. doi:10.1002/humu.20400

7. Ward AJ, Cooper TA. The pathobiology of splicing. *J Pathol*. 2010;220(2):152-163. doi:10.1002/path.2649

8. Dufner-Almeida LG, do Carmo RT, Masotti C, Haddad LA. Chapter Two - Understanding human DNA variants affecting pre-mRNA splicing in the NGS era. In: Kumar D, ed. *Advances in Genetics*. Vol 103. Academic Press; 2019:39-90. doi:10.1016/bs.adgen.2018.09.002

9. Fatscher T, Boehm V, Gehring NH. Mechanism, factors, and physiological role of nonsense-mediated mRNA decay. *Cell Mol Life Sci*. 2015;72(23):4523-4544. doi:10.1007/s00018-015-2017-9

10. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res*. 2016;44(4):1483-1495. doi:10.1093/nar/gkw010

11. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat*. 2018;39(11):1517-1524. doi:10.1002/humu.23626

12. Richards S, Aziz N, Bale S, et al. 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*. 2015;17(5):405-424. doi:10.1038/gim.2015.30

13. Costain G, Cohn RD, Scherer SW, Marshall CR. Genome sequencing as a diagnostic test. *CMAJ*. 2021;193(42):E1626-E1629. doi:10.1503/cmaj.210549

14. Walker LC, Hoya M de la, Wiggins GAR, Lindy A, Vincent LM, Parsons MT, et al. Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. *The American Journal of Human Genetics* [Internet]. https://doi.org/10.1016/j.ajhg.2023.06.002

15. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. 2019;176(3):535-548.e24. doi:10.1016/j.cell.2018.12.015

16. Dawes R, Bournazos AM, Bryen SJ, et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet*. 2023;55(2):324-332. doi:10.1038/s41588-022-01293-8

17. Resch A, Xing Y, Alekseyenko A, Modrek B, Lee C. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res*. 2004;32(4):1261-1269. doi:10.1093/nar/gkh284

18. Costain G, Walker S, Marano M, et al. Genome Sequencing as a Diagnostic Test in Children With Unexplained Medical Complexity. *JAMA Network Open*. 2020;3(9):e2018109. doi:10.1001/jamanetworkopen.2020.18109

19. Stavropoulos DJ, Merico D, Jobling R, et al. Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. *NPJ Genom Med*. 2016;1:15012-. doi:10.1038/npjgenmed.2015.12

20. Deshwar A, Yuki K, Hou H, et al. Trio RNA sequencing in a cohort of medically complex children. *The American Journal of Human Genetics*. Published online March 1, 2023. doi:10.1016/j.ajhg.2023.03.006

21. Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine*. 2018;20(4):435-443. doi:10.1038/gim.2017.119

22. Walker S, Lamoureux S, Khan T, et al. Genome sequencing for detection of pathogenic deep intronic variation: A clinical case report illustrating opportunities and challenges. *Am J Med Genet A*. 2021;185(10):3129-3135. doi:10.1002/ajmg.a.62389

23. GTEx Portal. Accessed May 14, 2023. https://gtexportal.org/home/

24. Ellingford JM, Ahn JW, Bagnall RD, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Medicine*. 2022;14(1):73. doi:10.1186/s13073-022-01073-3

25. Costain G, Jobling R, Walker S, et al. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. *Eur J Hum Genet*. 2018;26(5):740-744. doi:10.1038/s41431-018-0114-6

26. Abstracts from the 54th European Society of Human Genetics (ESHG) Conference: Oral Presentations. *Eur J Hum Genet*. 2022;30(1):3-87. doi:10.1038/s41431-021-01025-2

27. Xiang J, Peng J, Baxter S, Peng Z. AutoPVS1: An automatic classification tool for PVS1 interpretation of null variants. *Hum Mutat*. 2020;41(9):1488-1498. doi:10.1002/humu.24051

28. Lin JH, Tang XY, Boulling A, et al. First estimate of the scale of canonical 5' splice site GT>GC variants capable of generating wild-type transcripts. *Hum Mutat*. 2019;40(10):1856-1873. doi:10.1002/humu.23821

29. de Sainte Agathe JM, Filser M, Isidor B, et al. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Human Genomics*. 2023;17(1):7. doi:10.1186/s40246-023-00451-1

30. Palmisano A, Vural S, Zhao Y, Sonkin D. MutSpliceDB: A database of splice sites variants with RNA-seq based evidence on effects on splicing. *Hum Mutat*. 2021;42(4):342-345. doi:10.1002/humu.24185

31. Shiraishi Y, Okada A, Chiba K, et al. Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data. *Nat Commun*. 2022;13(1):5357. doi:10.1038/s41467-022-32887-9

32. Bournazos AM, Riley LG, Bommireddipalli S, et al. Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. *Genet Med*. 2022;24(1):130-145. doi:10.1016/j.gim.2021.09.001

33. Erkelenz S, Theiss S, Kaisers W, et al. Ranking noncanonical 5' splice site usage by genome-wide RNA-seq analysis and splicing reporter assays. *Genome Res*. 2018;28(12):1826-1840. doi:10.1101/gr.235861.118

34. Rivas MA, Pirinen M, Conrad DF, et al. Impact of predicted protein-truncating genetic variants on the human transcriptome. *Science*. 2015;348(6235):666-669. doi:10.1126/science.1261877

35. Sanders SJ, Schwartz GB, Farh KKH. Clinical impact of splicing in neurodevelopmental disorders. *Genome Medicine*. 2020;12(1):36. doi:10.1186/s13073-020-00737-2

36. Peña LDM, et. al. Contributions from medical geneticists in clinical trials of genetic therapies: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*. 2023 Jun;25(6):100831. doi: 10.1016/j.gim.2023.100831

Figures

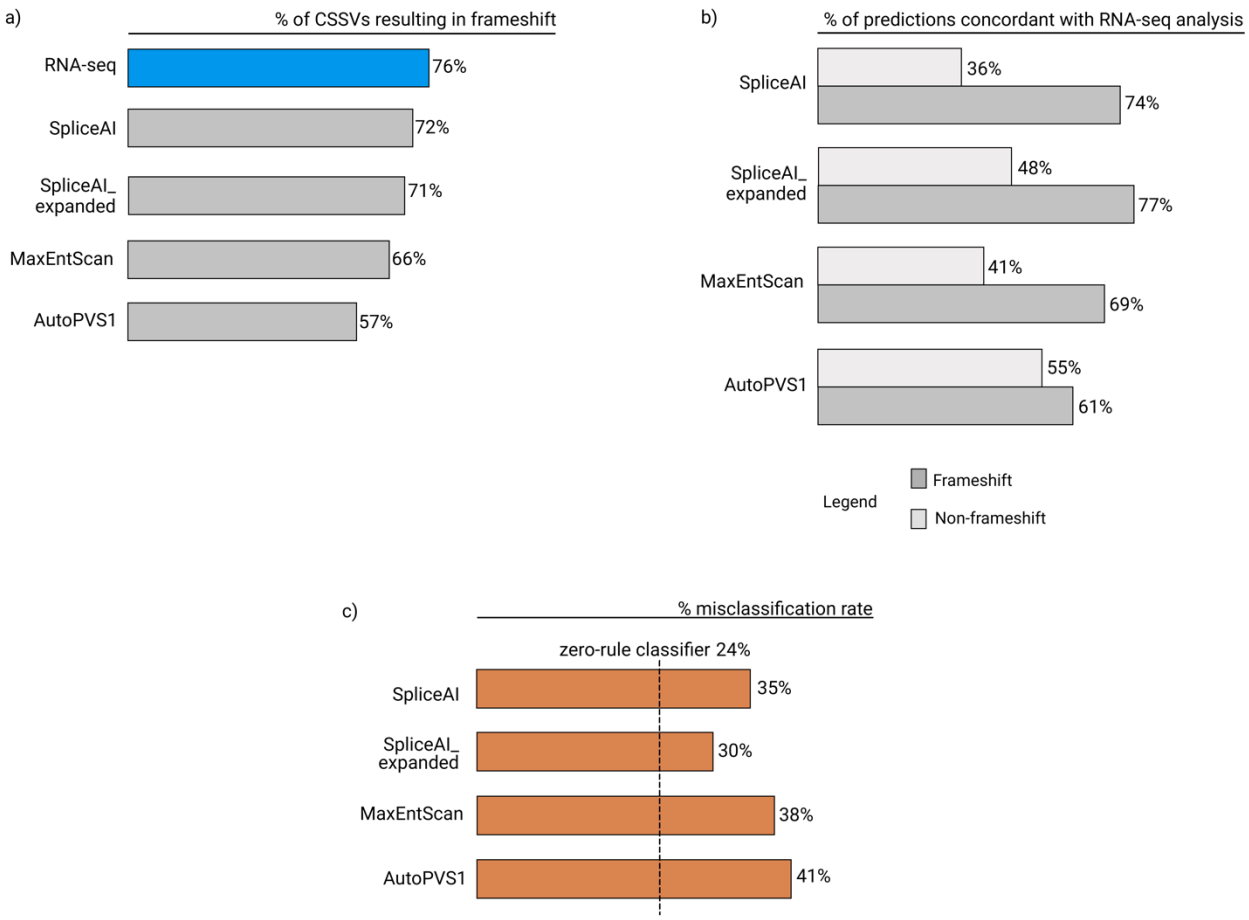
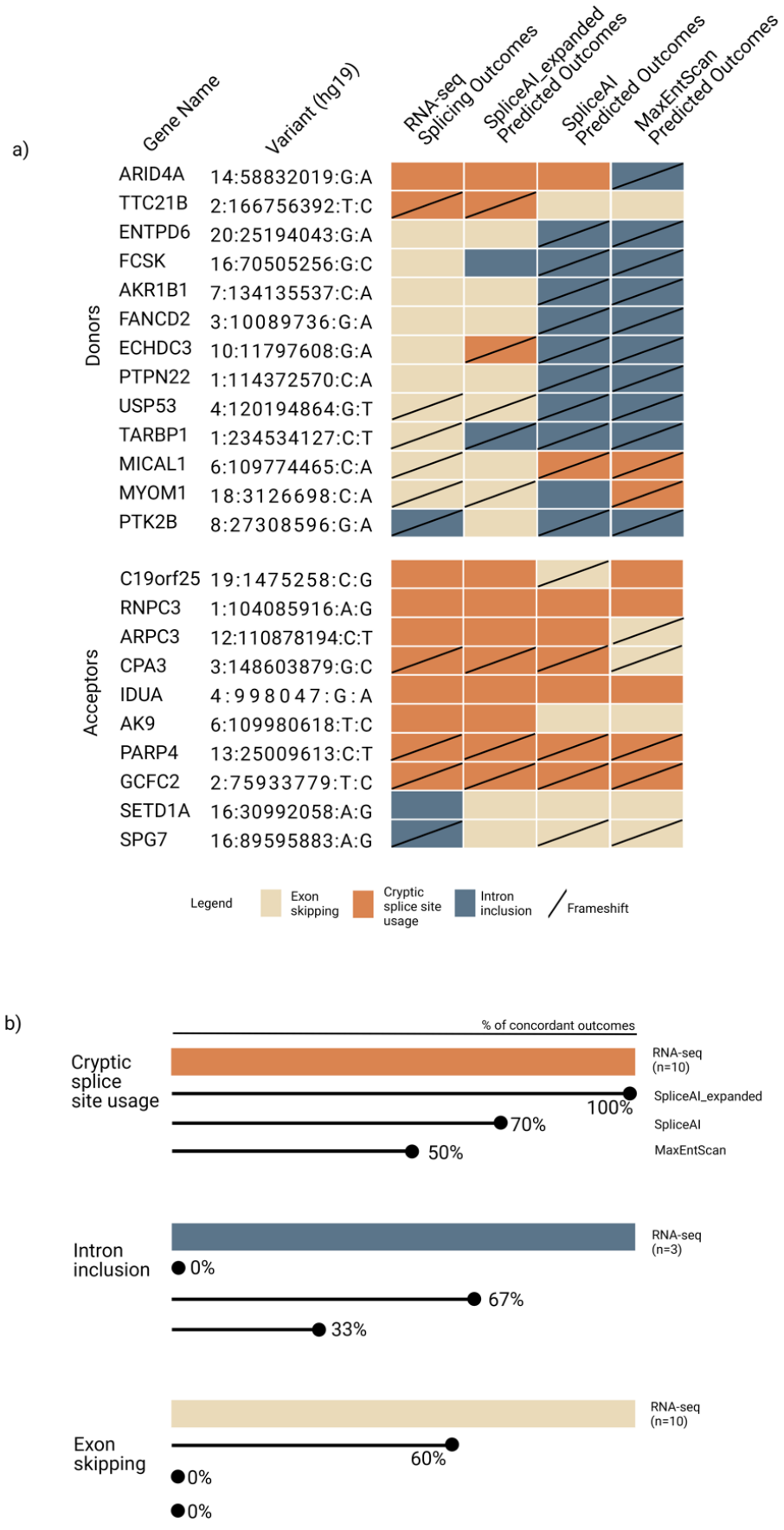
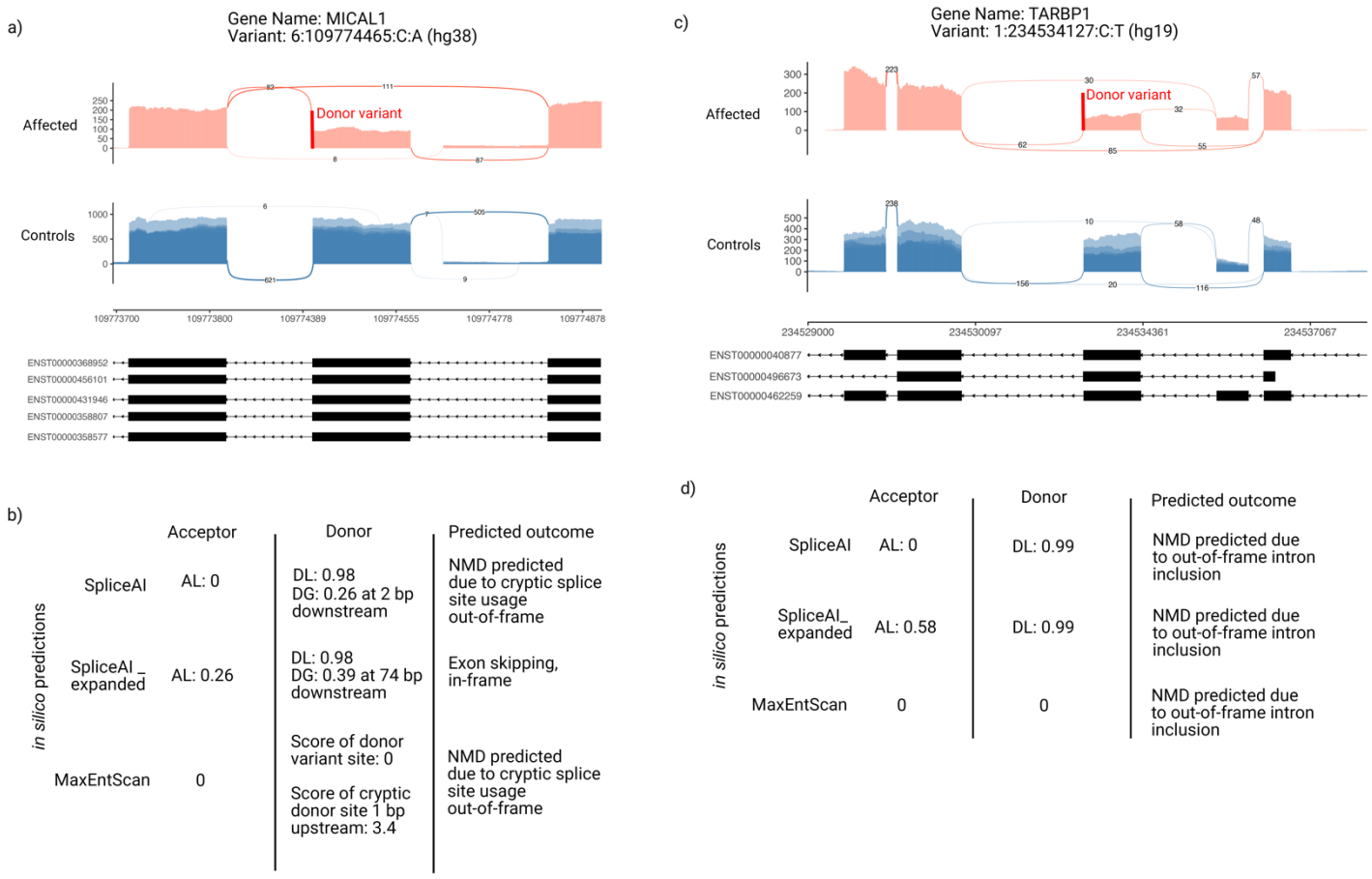


Figure 1 a) A comparison of the proportion of CSSVs resulting in a frameshift according to RNA-seq analysis vs. *in silico* predictions. **b)** Concordance between RNA-seq and *in silico* predictions of the impact of CSSVs on splicing. Concordant outcomes are defined as the RNA-seq and respective *in silico* tool identifying the same outcome (frameshift or non-frameshift). **c)** Misclassification rates of *in silico* tools compared to a zero-rule classifier.

Figure 2 a) Selected variants (n=23) with specific splicing outcomes in RNA-seq (exon skipping, cryptic splice site usage, and intron inclusion) compared with splicing outcome predictions from *in silico* tools. Total pairwise concordance in specific splicing outcomes was 70% (16/23) for SpliceAI_expanded, 35% (8/23) for SpliceAI and 26% (6/23) for MaxENTScan.

b) Concordance of specific splicing outcomes in RNA-seq vs. *in silico* predictions.





Figures 3) Two examples of donor CSSVs showing discordant splicing events in blood RNA-seq vs. *in silico* predictions but overall correct outcome (frameshift vs. no frameshift). Splicing junctions with reads >5 are shown. First transcripts on the list are the canonical transcripts indicated by Ensembl. **a)** CSSV in *MICAL1* results in exon skipping leading to a frameshift in RNA-seq. **b)** SpliceAI, SpliceAI_expanded and MaxEntScan all predicted activation of a cryptic splice site, resulting in frameshift. **c)** CSSV in *TARBP1* results in exon skipping leading to a frameshift in RNA-seq. **d)** SpliceAI, SpliceAI_expanded and MaxEntScan all predicted intron inclusion, resulting in frameshift.