SpliceAI-10k calculator for the prediction of pseudoexonization, intron retention, and exon deletion

Daffodil M. Canson[1,2*], Olga Kondrashova[3], Miguel de la Hoya[4], Michael T. Parsons[1], Dylan M. Glubb[3], Amanda B. Spurdle[1,2]

[1]Population Health Program, QIMR Berghofer Medical Research Institute, Herston, QLD 4006, Australia
[2]Faculty of Medicine, The University of Queensland, Herston, QLD 4006, Australia
[3]Cancer Research Program, QIMR Berghofer Medical Research Institute, Herston, QLD 4006, Australia
[4]Molecular Oncology Laboratory CIBERONC, IdISSC (Instituto de Investigación Sanitaria del Hospital Clínico San Carlos), Hospital Clinico San Carlos, 28040 Madrid, Spain

***Correspondence:** Daffodil.Canson@qimrberghofer.edu.au

ABSTRACT

Summary

The SpliceAI deep learning-based tool is widely used by clinical diagnostic laboratories and research groups to predict probability of variant impact on RNA splicing. SpliceAI generates four delta score predictions for loss or gain of acceptor and donor splice sites, and current application relies on the maximum delta score to assign variant impact. We developed the SpliceAI-10k calculator (SAI-10k-calc) to extend use of this tool to predict aberration type and size of inserted or deleted sequence, using an analysis window of 10,000 nucleotides.

From analysis of a control dataset of 1212 single nucleotide variants (SNVs) with curated splicing assay results, we established thresholds and devised a series of formulae to automatically predict specific types of aberrations: pseudoexonization, whole/partial intron retention, partial exon deletion, and exon skipping. Using our default thresholds, SAI-10k-calc has an overall sensitivity of 96% and specificity of 87%. Notably, it has high performance (95% accuracy, 96% sensitivity, 94% specificity) for SNVs >50 bp from the exon-intron junction.

Availability and implementation

The SAI-10k-calc is implemented in R (https://github.com/okon/SpliceAI_parser) and Microsoft Excel spreadsheet (Supplementary File, can process up to 1000 SNVs). Users can adjust the default thresholds to suit their target performance values.

Supplementary information

Supplementary data are available online.

Introduction

SpliceAI is a neural network that predicts splicing from a pre-mRNA sequence (Jaganathan et al., 2019). Previous evaluations have identified SpliceAI as the best predictor of variants that impact splicing, here termed spliceogenic variants (Ha, Kim, & Jang, 2021; Moles-Fernández et al., 2021; Riepe, Khan, Roosing, Cremers, & 't Hoen, 2021; Wai et al., 2020). These studies assessed variants across multiple locations (i.e. splice site motifs, exonic, and deep intronic), using the maximum delta score (of the four possible output scores) that passed the study-designated thresholds to predict variant spliceogenicity, but not the splicing aberration type.

SpliceAI sensitivity for detecting spliceogenic intronic variants >50 bp from exons was originally reported to be 41% using a 0.5 delta score threshold (Jaganathan et al., 2019), but improved sensitivity values of ≥90% were observed by lowering the threshold to 0.2 (Riepe et al., 2021) or 0.05 (Moles-Fernández et al., 2021). Paired donor-acceptor splice site scores were observed to be relevant for validated pseudoexonization (Moles-Fernández et al., 2021). Additional manual checking of donor-acceptor splice site pairing was incorporated into a scheme to prioritize likely spliceogenic deep intronic variants (Qian et al., 2021), but such an approach is inefficient when analyzing large datasets.

We developed the SAI-10k-calc for the prediction of different splicing aberration types and altered transcript sizes due to SNVs, with a focus on improved prediction of aberrations due to intronic variation located >50 bp from exon-intron junction.

Methods

SAI-10k-calc was designed to predict specific types of splicing aberrations, namely pseudoexonization, partial intron retention, partial exon deletion, exon skipping, and whole intron retention. Its features were derived from the application of all four delta scores and their corresponding delta positions generated by the SpliceAI tool (Jaganathan et al., 2019) using the maximum distance of +/- 4999 bp flanking the variant of interest. SAI-10k-calc can process SpliceAI scores resulting from SNVs at any exonic or intronic position, but not scores resulting from indels. The decision flowchart is shown in the Supplementary Figure.

We established the default thresholds for SpliceAI raw delta scores (0.02–0.05 for gained exon and 0–0.05 for lost exon) and the exon size range of 25–500 bp based on two training sets derived from published splicing data (Supplementary Table S1): (1) SNVs in *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, and *PMS2* genes from Shamsani et al. (2019) and (2) deep intronic SNVs in various Mendelian disease genes from Moles-Fernández et al. (2021). The maximum delta score threshold of 0.05 is also supported by previous findings (Moles-Fernández et al., 2021). For cryptic splice site activation, we used the 0.2 threshold based on the lower limit set by SpliceAI developers (Jaganathan et al., 2019) and the recommended threshold from the benchmarking study of Riepe et al. (2021).

Usage and features

SAI-10k-calc (Supplementary File and https://github.com/okon/SpliceAI_parser) requires input of raw scores generated by either SpliceAI Lookup (https://spliceailookup.broadinstitute.org/) or SpliceAI v1.3.1 run from the command line (https://github.com/Illumina/SpliceAI). This includes delta score predictions (and corresponding delta positions) for acceptor gain, acceptor loss,

donor gain, and donor loss. For the lightweight spreadsheet SAI-10k-calc implementation, users can process up to 1000 SNVs at a time, and must indicate the strand orientation of the transcript, and specify if intronic SNVs are located >250 bp (here set as exclusive region for pseudoexonization) or >50 bp from the nearby splice site. For the R SAI-10k-calc implementation, there is no limit on the number of SNVs that can be processed at a time, and users need to provide two input files: SpliceAI output VCF file and a tab-separated file with gene names and RefSeq transcript IDs used in SpliceAI calculations.

The SAI-10k-calc output is a summary of splicing predictions indicating the type of aberration, possible combinations of aberrations (e.g. one SNV producing exon skipping plus partial intron retention), and the exact size of inserted and/or deleted sequences. The latter is critical to predict the pathogenicity of the splicing alteration, and to design laboratory validation experiments. Examples are provided in the SAI-10k-calc spreadsheet and illustrated in Figure 1. However, we note that, due to SpliceAI limitations, the calculator cannot be designed to predict two specific combinations of aberrant transcripts: exon skipping plus whole intron retention, and partial exon deletion plus partial intron retention. Moreover, SAI-10k-calc cannot process indels because the sizes of genomic insertions/deletions cannot be known from SpliceAI delta positions alone; thus it cannot calculate the sizes of the resulting aberrant transcripts.

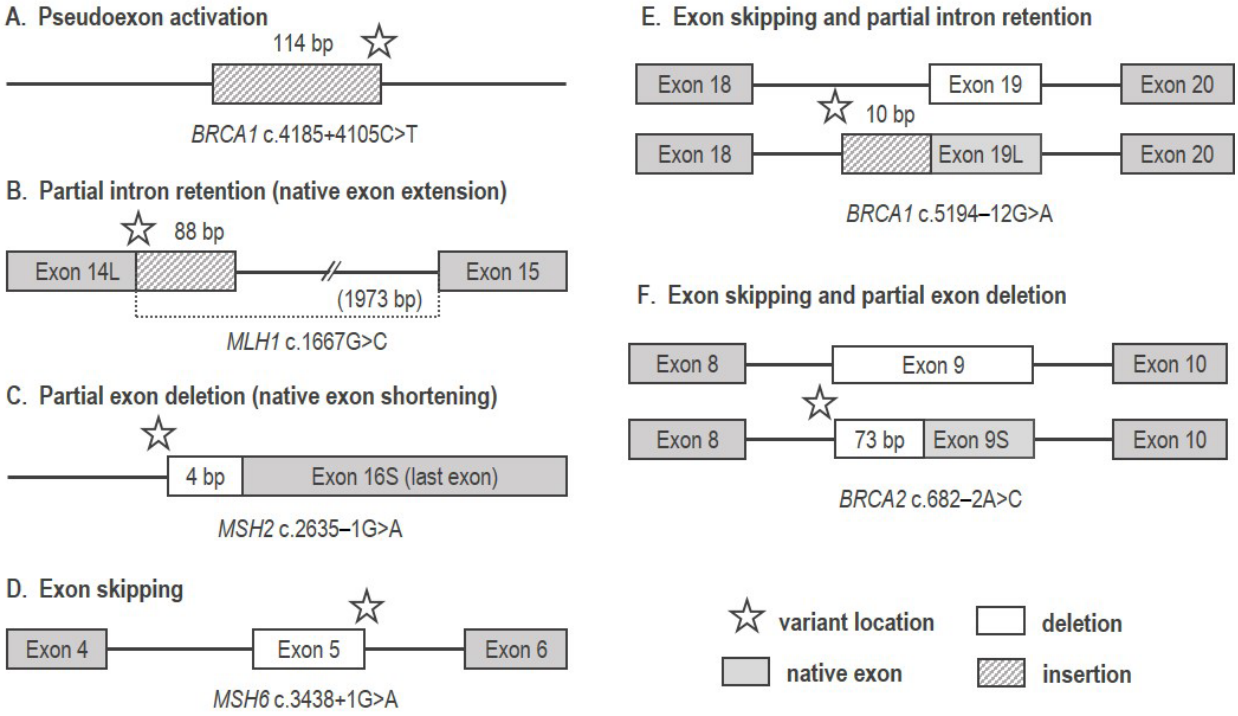


Figure 1. Types of splicing aberrations predicted by the SpliceAI-10k calculator. Six single nucleotide variants that were experimentally confirmed to alter splicing are provided as built-in examples of positive predictions in the SAI-10k-calc spreadsheet. **A.** *BRCA1* c.4185+4105C>T (NC_000017.10:g.41238856G>A) led to a 114-bp pseudoexon activation (Montalban et al., 2019). **B.** *MLH1* c.1667G>C (NC_000003.11:g.37081785G>C) led to an 88-bp partial intron

retention (reviewed in Thompson et al. (2020)); we predicted this variant to also result in a 1973-bp intron 14 retention, which might not have been detected owing to the pCAS minigene construct assay design. **C.** *MSH2* c.2635–1G>A (NC_000002.11:g.47709917G>A) led to a 4-bp partial exon deletion (Thompson et al., 2020). **D.** *MSH6* c.3438+1G>A (NC_000002.11:g.48030825G>A) caused exon skipping (van der Klift et al., 2015). Aberration type combinations can also be predicted by the calculator, e.g. **E.** Exon skipping and 10-bp partial intron retention caused by *BRCA1* c.5194–12G>A (NC_000017.10:g.41209164C>T) (Casadei et al., 2019; Théry et al., 2011), **F.** Exon skipping and 73-bp partial exon deletion caused by *BRCA2* c.682–2A>C (NC_000013.10:g.32905054A>C) (Santos et al., 2014). Extended exons resulting from partial intron retention are denoted with L after the exon number. Shortened exons resulting from partial exon deletion are denoted with S after the exon number.

Performance

Input and output data are shown in Supplementary Tables S2 and S3, and performance values are summarized in Supplementary Table S4. SAI-10k-calc has an overall sensitivity of 96% (447/464 confirmed spliceogenic SNVs) and specificity of 87% (654/748 non-spliceogenic SNVs) using default thresholds; specificity was greater (94%) for predicting spliceogenicity for SNVs >50 bp from the exon-intron junction. High accuracy in type and size predictions for pseudoexonization (85%), exon skipping (83%), and partial intron retention (80%) demonstrates applicability of SAI-10k-calc for second-tier annotation of variants prioritized from whole genome sequencing datasets.

Funding: D.M.C. has been supported by QIMR Berghofer Ailsa Zinns PhD Scholarship, QIMR Berghofer HDC Top Up Scholarship, and UQ Research Training Tuition Fee Offset. O.K. is supported by a NHMRC Emerging Leader 1 Investigator Grant (APP2008631). M.d.l.H. is supported by a grant from the Spanish Ministry of Science and Innovation, Plan Nacional de I+D+I 2013-2016, ISCIII (PI20/00110) co-funded by FEDER from Regional Development European Funds (European Union). A.B.S is supported by a NHMRC Investigator Fellowship funding (APP1177524).

Acknowledgement: Members of the QIMR Berghofer Molecular Cancer Epidemiology Laboratory updated the *BRCA1* and *BRCA2* splicing table that partly provided curated splicing data for training set 1.

Conflict of interest: None declared.

References

Casadei, S., Gulsuner, S., Shirts, B. H., Mandell, J. B., Kortbawi, H. M., Norquist, B. S., . . . Walsh, T. (2019). Characterization of splice-altering mutations in inherited predisposition to cancer. *Proceedings of the National Academy of Sciences, 116*, 26798-26807. doi:10.1073/pnas.1915608116

Ha, C., Kim, J.-W., & Jang, J.-H. (2021). Performance Evaluation of SpliceAI for the Prediction of Splicing of NF1 Variants. *Genes, 12*(9), 1308. doi:10.3390/genes12091308

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

Moles-Fernández, A., Domènech-Vivó, J., Tenés, A., Balmaña, J., Diez, O., & Gutiérrez-Enríquez, S. (2021). Role of Splicing Regulatory Elements and In Silico Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes. *Cancers, 13*(13), 3341. doi:10.3390/cancers13133341

Montalban, G., Bonache, S., Moles-Fernández, A., Gisbert-Beamud, A., Tenés, A., Bach, V., . . . Gutiérrez-Enríquez, 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. *Journal of Medical Genetics, 56*(2), 63. doi:10.1136/jmedgenet-2018-105606

Qian, X., Wang, J., Wang, M., Igelman, A. D., Jones, K. D., Li, Y., . . . Chen, R. (2021). Identification of Deep-Intronic Splice Mutations in a Large Cohort of Patients With Inherited Retinal Diseases. *Frontiers in Genetics, 12*(276). doi:10.3389/fgene.2021.647400

Riepe, T. V., Khan, M., Roosing, S., Cremers, F. P. M., & 't Hoen, P. A. C. (2021). Benchmarking deep learning splice prediction tools using functional splice assays. *Human Mutation, 42*(7), 799-810. doi:10.1002/humu.24212

Santos, C., Peixoto, A., Rocha, P., Pinto, P., Bizarro, S., Pinheiro, M., . . . Teixeira, M. R. (2014). Pathogenicity Evaluation of BRCA1 and BRCA2 Unclassified Variants Identified in Portuguese Breast/Ovarian Cancer Families. *The Journal of Molecular Diagnostics, 16*(3), 324-334. doi:10.1016/j.jmoldx.2014.01.005

Shamsani, J., Kazakoff, S. H., Armean, I. M., McLaren, W., Parsons, M. T., Thompson, B. A., . . . Spurdle, A. B. (2019). A plugin for the Ensembl Variant Effect Predictor that uses MaxEntScan to predict variant spliceogenicity. *Bioinformatics, 35*(13), 2315-2317. doi:10.1093/bioinformatics/bty960

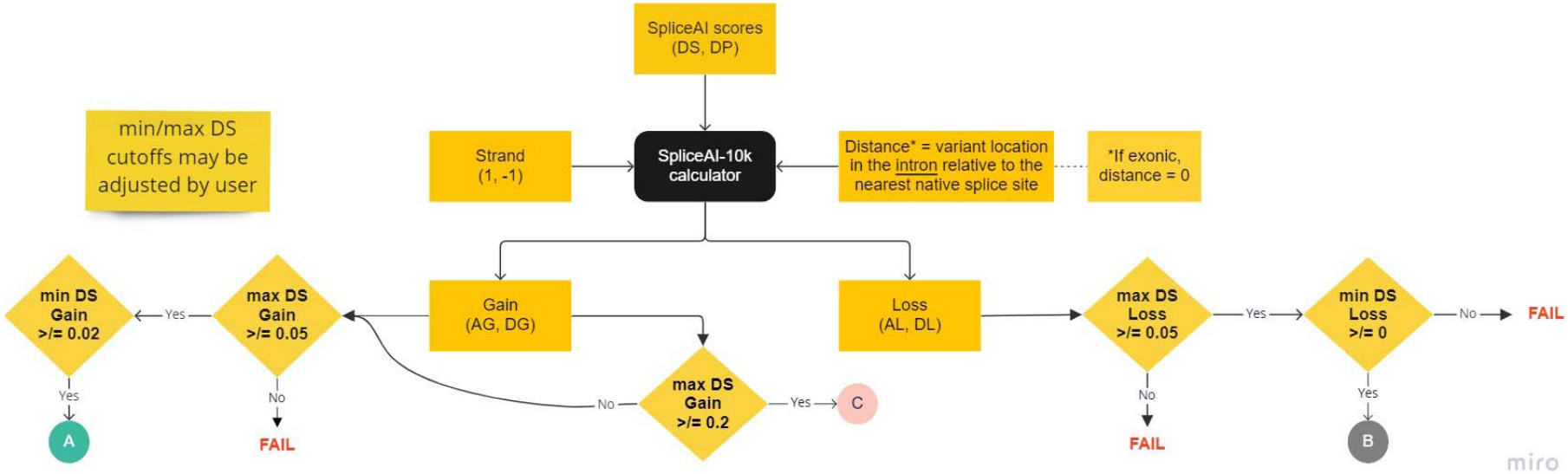
Théry, J. C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M.-P., Killian, A., . . . Tosi, M. (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *European Journal Of Human Genetics, 19*(10), 1052-1058. doi:10.1038/ejhg.2011.100

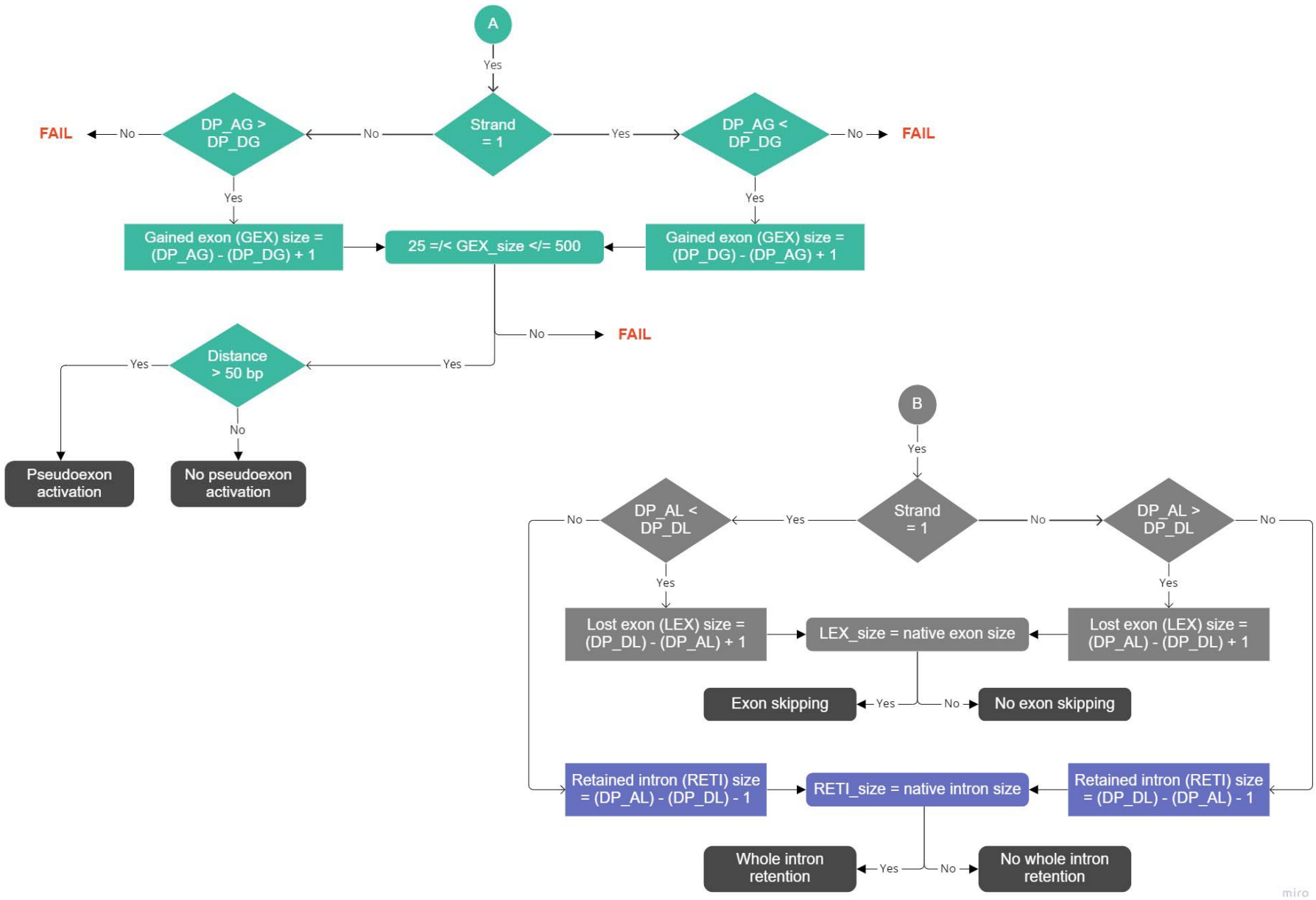
Thompson, B. A., Walters, R., Parsons, M. T., Dumenil, T., Drost, M., Tiersma, Y., . . . Tops, C. (2020). Contribution of mRNA Splicing to Mismatch Repair Gene Sequence Variant Interpretation. *Frontiers in Genetics, 11*, 798. doi:10.3389/fgene.2020.00798

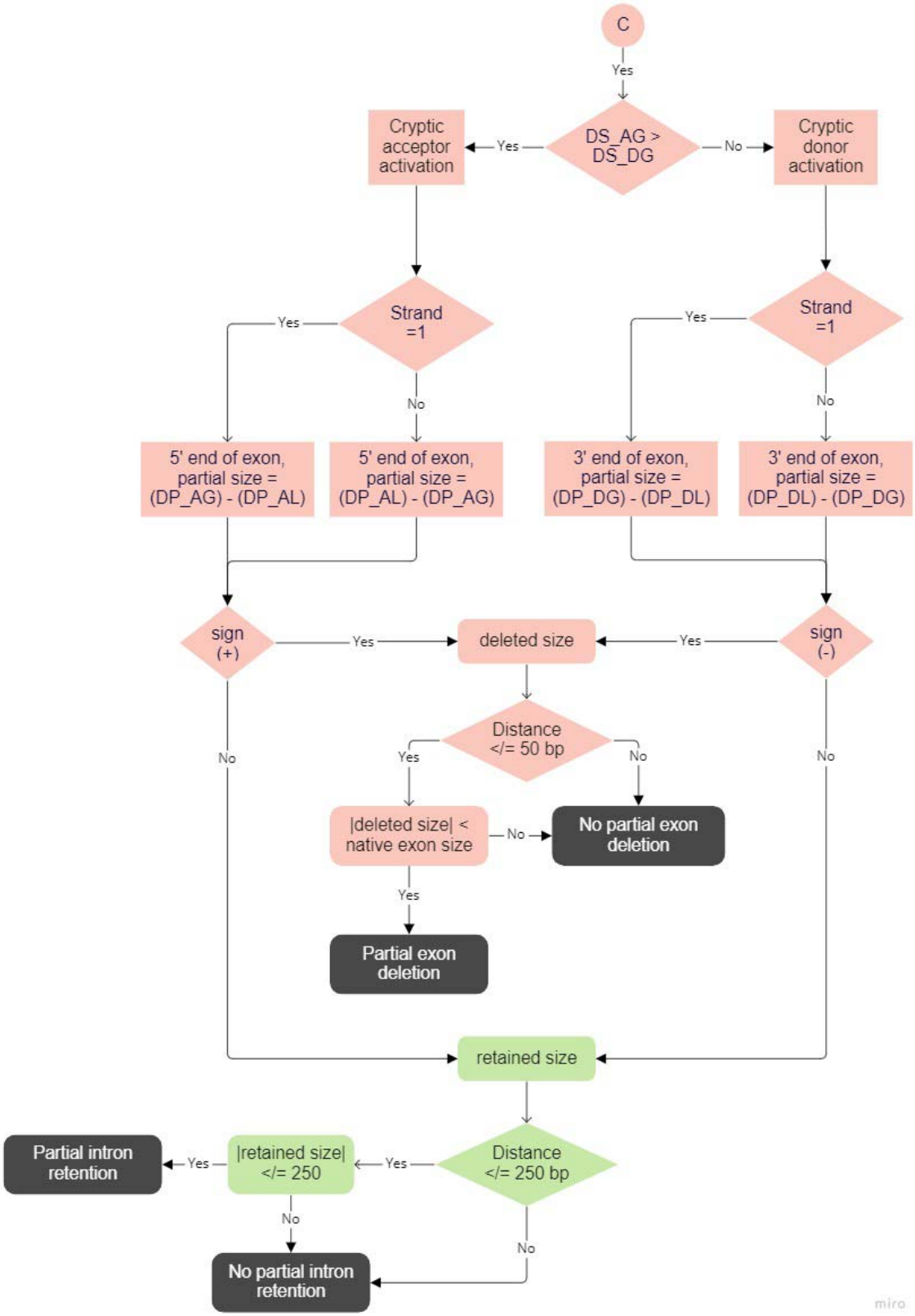
van der Klift, H. M., Jansen, A. M. L., Steenstraten, N., Bik, E. C., Tops, C. M. J., Devilee, P., & Wijnen, J. T. (2015). Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. *Molecular Genetics & Genomic Medicine, 3*(4), 327-345. doi:10.1002/mgg3.145

Wai, H. A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., . . . disease, w. (2020). Blood RNA analysis can increase clinical diagnostic rate andresolve variants of uncertain significance. *Genetics in Medicine, 22*, 1005-1014. doi:10.1038/s41436-020-0766-9

Supplementary Figure. SpliceAI-10k calculator decision flowchart. Image generated by Miro (https://miro.com/).




A
Yes
Strand = 1
No
DP_AG > DP_DG
No
FAIL
Yes
DP_AG < DP_DG
No
FAIL
Yes
Gained exon (GEX) size = (DP_AG) - (DP_DG) + 1
25 =/< GEX_size </= 500
Gained exon (GEX) size = (DP_DG) - (DP_AG) + 1
Yes
No
FAIL
Distance > 50 bp
Yes
Yes
No
Pseudoexon activation
No pseudoexon activation
B
Yes
Strand = 1
No
DP_AL < DP_DL
Yes
DP_AL > DP_DL
No
No
Yes
Yes
Lost exon (LEX) size = (DP_DL) - (DP_AL) + 1
LEX_size = native exon size
Lost exon (LEX) size = (DP_AL) - (DP_DL) + 1
Exon skipping
Yes
No
No exon skipping
Retained intron (RETI) size = (DP_AL) - (DP_DL) - 1
RETI_size = native intron size
Retained intron (RETI) size = (DP_DL) - (DP_AL) - 1
Whole intron retention
Yes
No
No whole intron retention
miro


C
Yes
DS_AG > DS_DG
Yes
No
Cryptic acceptor activation
Cryptic donor activation
Strand =1
Yes
No
Strand =1
Yes
No
5' end of exon, partial size = (DP_AG) - (DP_AL)
5' end of exon, partial size = (DP_AL) - (DP_AG)
3' end of exon, partial size = (DP_DG) - (DP_DL)
3' end of exon, partial size = (DP_DL) - (DP_DG)
sign (+)
Yes
deleted size
Yes
sign (-)
No
Distance </= 50 bp
Yes
No
No
|deleted size| < native exon size
No
No partial exon deletion
Yes
Partial exon deletion
retained size
Distance </= 250 bp
Yes
|retained size| </= 250
Yes
Partial intron retention
No
No
No partial intron retention
miro