Read Me

varSEAK Online

Splice Site Prediction

Version 2.1

Valid from 2022-02-23

JSI medical systems GmbH
Tullastr. 18
77955 Ettenheim
GERMANY
phone: +49-7822/440150-21
fax: +49-7822/440150-20
email: support@varseak.bio
web: www.jsi-medisys.de

RUO

Contents

1 About varSEAK

varSEAK is a sophisticated variant interpretation software. This bundle of tools and public DB information streamlines finding and classifying the variants relevant to you.

After intuitive and flexible filtering, you can classify your variants by using your in-house rules and/or ACMG guidelines.

Using varSEAK's own database, you will be connected to a global network of scientists. You can see if your variant has already been found and classified. Questions about it? Contact the submitter!

To import your data, we recommend variant calling with SEQUENCE Pilot for the most comfortable, automatic transfer to your local varSEAK installation. In case you use a third-party software, simply import VCF files.

Of course, Splice Site Prediction is part of varSEAK with a public access on varSEAK Online.
To enjoy all benefits, use varSEAK Local where SSP results will be automatically calculated upon import into varSEAK and can be used for filtering as well.

Share your Experience and Knowledge: Join the varSEAK community!

2 Introduction

The JSI splice site prediction tool calculates splicing effects for genetic variants. It is available on **varSEAK Online** as well as in the **varSEAK** software. The software requires canonical splice sites (core motif GT for 5' donor splice sites or AG for 3' acceptor splice sites). Non-canonical splice sites, e.g. core motif GC for 5' donor splice sites, are not taken into consideration. Such splice sites will automatically be listed as class 3 (unknown splicing effect) if they might be affected by a variant, and as class 1 if they are unaffected by the rules valid for GT 5' donor splice sites.

We would like to thank Gene Yeo for his friendly approval to integrate the MaxEntScan scoring algorithm (Yeo & Burge, 2003) into our software.

Version 2.1 was developed to optimize the algorithm as well as the usability. The re-engineered algorithm has an accuracy of 96.41% (for more details, please see Development and validation, page 5). The application is more intuitive and comprehensible due to a clearer design that focuses on the relevant scores and interpretations.

2.1 Disclaimer

The information generated by the JSI splice site prediction tool is not intended for direct diagnostic use or medical decisionmaking without consulting a genetics professional. Users must be careful and well positioned to judge and verify the information made available before relying on it.
The information shown are results based on a trained prediction algorithm (see also Development and validation, page 5). Although the predictions show a very good hit rate when validated on sample data, they should under no circumstances be used without further confirmation. The predictions are generated with the latest algorithms, but can by no means replace evidence-based data from further investigations (wet lab).

3 Development and validation

The JSI splice site prediction algorithm was trained on approximately 200 000 real splice sites taken from GRCh37 and 300 000 false splice sites from the HS3D dataset (P Pollastro & Rampone, 2003; Pasquale Pollastro & Rampone, 2002). The resulting algorithm was tested for accuracy using a dataset by Leman *et al.*, consisting of 395 variants with known splicing effect (Leman et al., 2018).
For 5' splice sites, only GT-splice sites were considered. The much rarer GC-splice sites will be listed but always receive the class 3 (unknown splicing effect) if they might be affected by a variant, and class 1 if they are unaffected by the rules valid for GT 5' donor splice sites.
Accuracy is determined as (Number of correct assessments)/(Number of all assessments). With this dataset, the JSI splice site prediction algorithm has an accuracy of 96.41%.
For more details about the evaluation, please refer to our validation document.

4 Instructions for use

To display this web page we recommend the web browsers FireFox or Chrome. Note: Internet Explorer is not supported.

4.1 Analyze a single variant

To analyze a single variant, you can either enter the required information on the main page or you can change the input at the top of the results page.

4.1.1 Enter information

The information required is:

- a gene (type to narrow down the options available from the list)
- a transcript (a list of transcripts will be offered once a gene has been selected. Identical transcripts from different sources will be listed together)
  - if no transcript is given, the longest available transcript for this gene will be used automatically
- a variant, which can either be
  - a c.-HGVS nomenclature (such as c.1585-8G>A)
  - a sequence (20 - 150 bases, we recommend at least 50 bases, the variant should approximately be at the center)

Click the *[Search]* or *[OK]* button to start the analysis.

To analyze another variant in the same gene, you can simply change the HGVS or sequence. You can also change the transcript using the arrow button behind it. If you wish to analyze a variant on another gene, click the list button behind the gene name. A modular window will open where you can change the input.

Once you have selected another gene, you can again select your desired transcript. *[OK]* starts the analysis, *[Cancel]* lets you return without any changes.

4.1.2 Results

4.1.2.1 General information

The input box presents details concerning the gene and transcript, such as chromosome, strand, start and end positions as well as exon number and cDNA length.

Below the input field for HGVS or sequence, details concerning the position of the variant (exon/intron, cDNA position, genomic position) are displayed.

4.1.2.2 Sequence graph and legend

At the top of the results, a sequence graph depicts the reference and variant sequence and important features. A legend is given below on the right-hand side in the "INFO" box.

The variant is always placed at the center of the shown sequence. It is highlighted in red, the HGVS nomenclature is given above.

Important authentic or potential splice sites are marked with numbered triangles corresponding to the table on the left-hand side below the sequence graph. Authentic splice sites are also highlighted in purple both in the graph as well as in the table on the left. Cryptic splice sites predicted to be activated are highlighted in pink.

If no relevant authentic or cryptic splice sites are within 30 bp of the variant, the next splice sites up- and/or downstream will be shown and listed as well. Underscores (__) mark the skipped sequences to allow displaying of the neighbouring splice sites. The amount of bases from the beginning of the skipped sequence to the splice site will be listed above the reference (e.g. "3564 bp ->").

4.1.2.3 Overall predicted class

The overall prediction is given as a splice site prediction class in the center of the page. Possible classes are:

- Class 1 (dark green): No splicing effect
- Class 2 (light green): Likely no splicing effect
- Class 3 (yellow): Unknown splicing effect
- Class 4 (orange): Likely splicing effect
- Class 5 (red): Splicing effect

The overall predicted class is always the highest occuring class of all listed results for both 3' and 5' splice sites, should both be influenced.

4.1.2.4 Table with relevant splice site positions

To the left of the overall predicted class and below the sequence graph is a table listing relevant splice site positions with respective SSP classes and scores.

Above the table are two buttons labeled *[SSP 3']* or *[SSP 5']* which allow displaying of the respective splice site scores. If either is greyed out and inactive, there is no splice site and scores of this type available.

If there is an authentic or an activated cryptic splice site shown in the sequence graph, the corresponding row of the table will be marked in the respective colour.

For each listed position the following information is shown:

- splice site prediction class: 1-5 (see Overall predicted class, page 6, for more details)
- Score: Predicted likelihood of this splice site being functional (positive values) or non-functional (negative values), reaching from -100% to +100%. For splice sites with unknown functionality, the score is 0 %.
- ΔScore (DeltaScore): the difference between the Score of the splice site on the reference sequence and the Score of the splice site on the variant sequence.
- MaxEntScan: The ENT score from MaxEntScan (Yeo & Burge, 2003) is displayed alongside for comparison.
- ΔMaxEntScan (DeltaMaxEntScan): the difference between the MaxEntScan score of the splice site on the reference sequence and the MaxEntScan score of the splice site on the variant sequence.

For 3' splice sites, each row in the table can be selected with a bullet point, changing for which 3' splice site the details (selected 3' ss, AG Exclusion Zone and Branch Point/ U2) should be displayed.

4.1.2.5 Prediction details

To the right, details for the corresponding prediction are displayed. If there are results for both 5' and 3' splice sites, the details will be displayed for the selected splice site type. The overall prediction class will not change, since it always is the worst occuring class for both types of splice sites.

4.1.2.6 Public DB Info

Below the prediction details, the most important information from the public databases is displayed in a small table.
You can view (if available for the corresponding variant): the rs Number, the varSEAK Classification, the ClinVar Clinical Significance and the gnomAD AF. Each available information is linked out auf derin the same way as on the Variant Table.
If you wish to find more details, click the link below the table or at the top of the page, leading you to the varSEAK Online Variant Table with all results for your selected gene. Use "Refine Results" at the top of the page to search.
If none of the four fields from public databases is available, the table will not be displayed.

4.1.3 Export result

To export the prediction for the given variant, click the button *[Export to PDF]* on the upper right-hand side above the sequence graph. A PDF file will be generated and offered to you for download.
The report will contain the gene, transcript and variant als well as the overall and detailed predictions for the positions and scores of interest.

5 Examples

#	Gene	HGVS	Class	Prediction details	Actual effect	Reference
1	ABCA1	NM_005502.3 c.1195-27G>A	5	3' Acceptor Site: Use of de novo Splice Site 25 nt upstream of 3' ss Pos 1195-27 : De novo AG in AG-Exclusion-Zone.	25bp of intron 10 included in transcript	(Fasano et al., 2012)
2	BRCA1	NM_007294.3 c.4484G>A	4	5' Donor Site: Likely loss of function for authentic Splice Site. Exon Skipping Pos 4484+1 : Decrease of Score.	Exon 14 skipped	(Leman et al., 2018)
3	BRCA1	NM_007294.3 c.5153-1G>A	5	3' Acceptor Site: Use of cryptic site 1 nt downstream of 3' ss Pos 5153-2 : No AG.	Use of a cryptic site 1 nt downstream from 3' ss	(Leman et al., 2018)
4	BRCA2	NM_000059.3 c.7807G>C	1	3' Acceptor Site: No splicing effect	No change	(Leman et al., 2018)
5	CERS3	NM_001290343 c.609+1G>T	5	5' Donor Site: Loss of function for authentic Splice Site. Exon Skipping Pos 609+1 : No GT.	ss abolished	(Radner et al., 2013)
6	CFTR	NM_000492.3 c.1585-9T>A	5	3' Acceptor Site: Use of de novo Splice Site 7 nt upstream of 3' ss Pos 1585-9 : De novo AG in AG-Exclusion-Zone. Pos 1585-2 : Loss of function for authentic Splice Site.	exon skipping and 7bp retention due to cryptic ss	(Sharma et al., 2014)
7	RHD	NM_016124.4 c.1074-2A>C	5	3' Acceptor Site: Use of cryptic site 13 nt downstream of 3' ss Use of cryptic site 13 nt downstream of 3' ss Pos 1074-2 : No AG.	exon 8 skipping of strong intensity and use of cryptic splice site at 13 nt downstream from 3' ss	(Leman et al., 2018)
8	SLC40A1	NM_014585.5 c.387C>T	2	5' Donor Site: Likely no splicing effect.	no change	(Leman et al., 2018)

6 References

Baralle, M., & Baralle, F. E. (2018). The splicing code. *BioSystems*, *164*, 39–48. https://doi.org/10.1016/j.biosystems.2017.11.002

Corvelo, A., Hallegger, M., Smith, C. W. J., & Eyras, E. (2010). Genome-wide association between branch point properties and alternative splicing. *PLoS Computational Biology*, *6*(11), 12–15. https://doi.org/10.1371/journal.pcbi.1001016

Fasano, T., Pisciotta, L., Bocchi, L., Guardamagna, O., Assandro, P., Rabacchi, C., Zanoni, P., Filocamo, M., Bertolini, S., & Calandra, S. (2012). Lysosomal lipase deficiency: Molecular characterization of eleven patients with Wolman or cholesteryl ester storage disease. *Molecular Genetics and Metabolism*, *105*(3), 450–456. https://doi.org/10.1016/j.ymgme.2011.12.008

Lee, M., Roos, P., Sharma, N., Atalar, M., Evans, T. A., Pellicore, M. J., Davis, E., Lam, A. T. N., Stanley, S. E., Khalil, S. E., Solomon, G. M., Walker, D., Raraigh, K. S., Vecchio-Pagan, B., Armanios, M., & Cutting, G. R. (2017). Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites. *American Journal of Human Genetics*, *100*(5), 751–765. https://doi.org/10.1016/j.ajhg.2017.04.001

Leman, R., Gaildrat, P., Gac, G. L., Ka, C., Fichou, Y., Audrezet, M. P., Caux-Moncoutier, V., Caputo, S. M., Boutry-Kryza, N., Léone, M., Mazoyer, S., Bonnet-Dorion, F., Sevenet, N., Guillaud-Bataille, M., Rouleau, E., Paillerets, B. B. De, Wappenschmidt, B., Rossing, M., Muller, D., … Houdayer, C. (2018). Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: An international collaborative effort. *Nucleic Acids Research*, *46*(15), 7913–7923. https://doi.org/10.1093/nar/gky372

Nabih, A., Sobotka, J. A., Wu, M. Z., Wedeles, C. J., & Claycomb, J. M. (2017). Examining the intersection between splicing, nuclear export and small RNA pathways. *Biochimica et Biophysica Acta - General Subjects*, *1861*(11), 2948–2955. https://doi.org/10.1016/j.bbagen.2017.05.027

Pohl, M., Bortfeldt, R. H., Grützmann, K., & Schuster, S. (2013). Alternative splicing of mutually exclusive exons-A review. *BioSystems*, *114*(1), 31–38. https://doi.org/10.1016/j.biosystems.2013.07.003

Pollastro, P, & Rampone, S. (2003). *HS3D: Homo Sapiens Splice Site Data Set*. https://iris.unisannio.it/handle/20.500.12070/1267#.XisN6eDF-HY.mendeley

Pollastro, Pasquale, & Rampone, S. (2002). HS3D, A dataset of homo sapiens splice regions, and its extraction procedure from a major public database. *International Journal of Modern Physics C*, *13*(8), 1105–1117. https://doi.org/10.1142/S0129183102003796

Radner, F. P. W., Marrakchi, S., Kirchmeier, P., Kim, G. J., Ribierre, F., Kamoun, B., Abid, L., Leipoldt, M., Turki, H., Schempp, W., Heilig, R., Lathrop, M., & Fischer, J. (2013). Mutations in CERS3 Cause Autosomal Recessive Congenital Ichthyosis in Humans. *PLoS Genetics*, *9*(6). https://doi.org/10.1371/journal.pgen.1003536

Ratnadiwakara, M., Mohenska, M., & Änkö, M. L. (2018). Splicing factors as regulators of miRNA biogenesis – links to human disease. *Seminars in Cell and Developmental Biology*, *79*, 113–122. https://doi.org/10.1016/j.semcdb.2017.10.008

Rogan, P. K., Caminsky, N., & Mucaki, E. J. (2014). Interpretation of mRNA splicing mutations in genetic disease: Review of the literature and guidelines for information-theoretical analysis. *F1000Research*, *3*(May). https://doi.org/10.12688/f1000research.5654.1

Sharma, N., Sosnay, P. R., Ramalho, A. S., Douville, C., Franca, A., Gottschalk, L. B., Park, J., Lee, M., Vecchio-Pagan, B., Raraigh, K. S., Amaral, M. D., Karchin, R., & Cutting, G. R. (2014). Experimental Assessment of Splicing Variants Using Expression Minigenes and Comparison with In Silico Predictions. *Human Mutation*, *35*(10), 1249–1259. https://doi.org/10.1002/humu.22624

Yeo, G., & Burge, C. B. (2003). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*, 322–331. https://doi.org/10.1145/640075.640118

7 Appendix

7.1 Version History

Table 1: Version history

Version	Date *)	Author	Status **)	Changes
V02	2022-02-23	Eva Noeske	released	Corrected format errors
V01	2021-12-27	Eva Noeske	released	Changed format of document

*) Format: YYYY-MON-DD **) Status: Draft / released