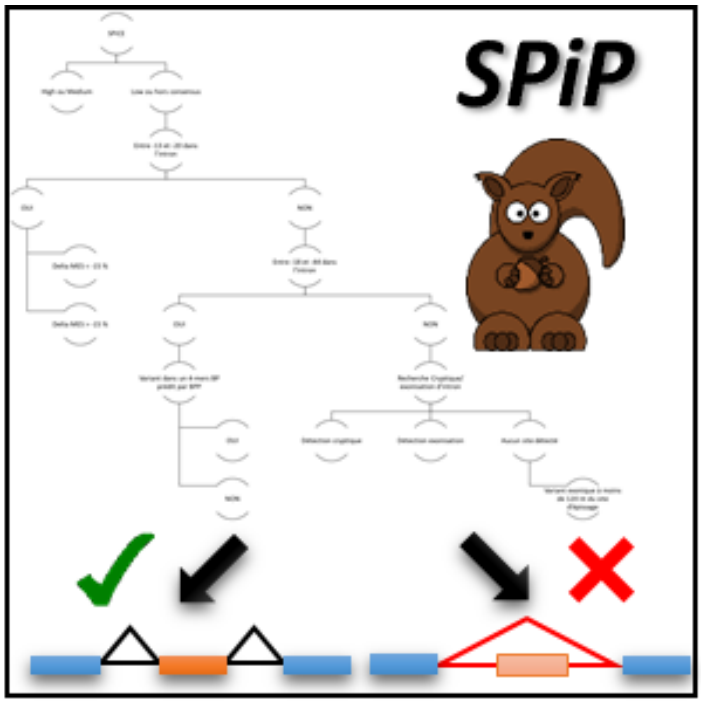
SPiPv0.3

(Splicing Prediction Pipeline)

Handbook



Raphaël LEMAN[1,2,3], Sophie KRIEGER[1,2,3] and Claude HOUDAYER[4,5,6]

[1]Laboratoire de Biologie Clinique et Oncologique, Centre François Baclesse, France, [2]Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders, UNIROUEN Normandie Université, Rouen, France, [3]Université Caen-Normandie, France, [4]Inserm U830, Centre de Recherches, Paris, France, [5]Université Paris Descartes, Sorbonne Paris Cité, Paris, France, [6]Service de Génétique, Institut Curie, Paris, France

Contents

INTRODUCTION

SPiP is a decisional tree running a cascade of bioinformatics tools. Briefly, SPiP uses SPiCE tool for the consensus splice sites (donor and acceptor sites), MES for polypyrimidine tract between -13 and -20, BPP for branch point area between -18 and -44, a homemade score to research cryptic/de novo activation and ΔtESRseq for exonic splicing regulatory element until to 120 nt in exon (see figure 1).

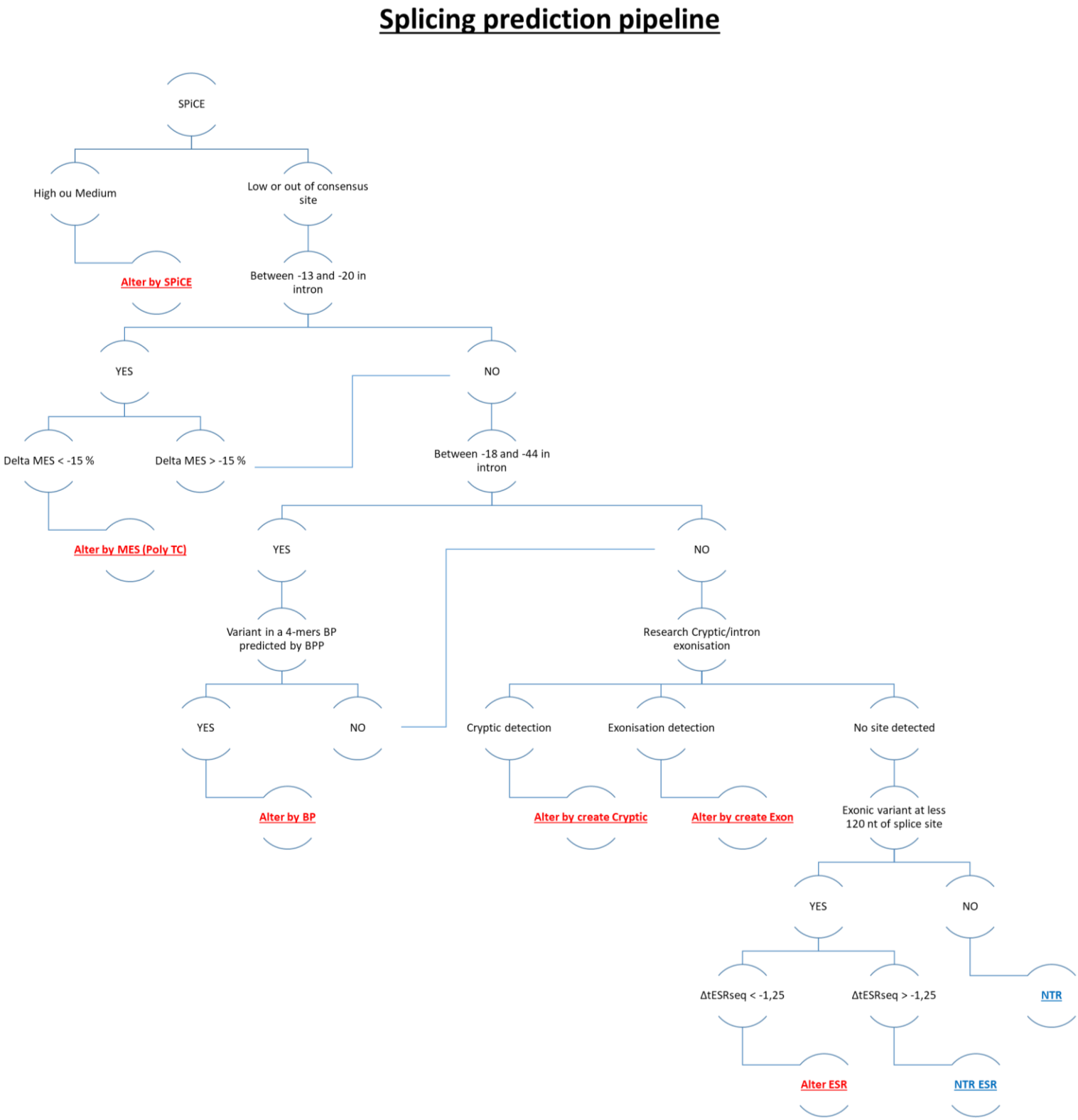


Figure 1: decisional tree of SPiP

For more information on SPiP development and rationale, please refer to the article: “SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations, validated on a curated diagnostic set of 2,784 exonic and intronic variants.” (Raphaël LEMAN *et.al*., article in progress).

SPiP has been developed in R langage (file “Rscript.R”), has embedded all score and database that it needs, and is freely available at: https://sourceforge.net/projects/splicing-prediction-pipeline/. To run variant, SPiP needs only the position and nucleotidic change either in text file or VCF file.

The main of SPiP is to prioritize the RNA studies and **NOT** to predict the pathogenicity of this variant.

SPiP FOR WINDOWS

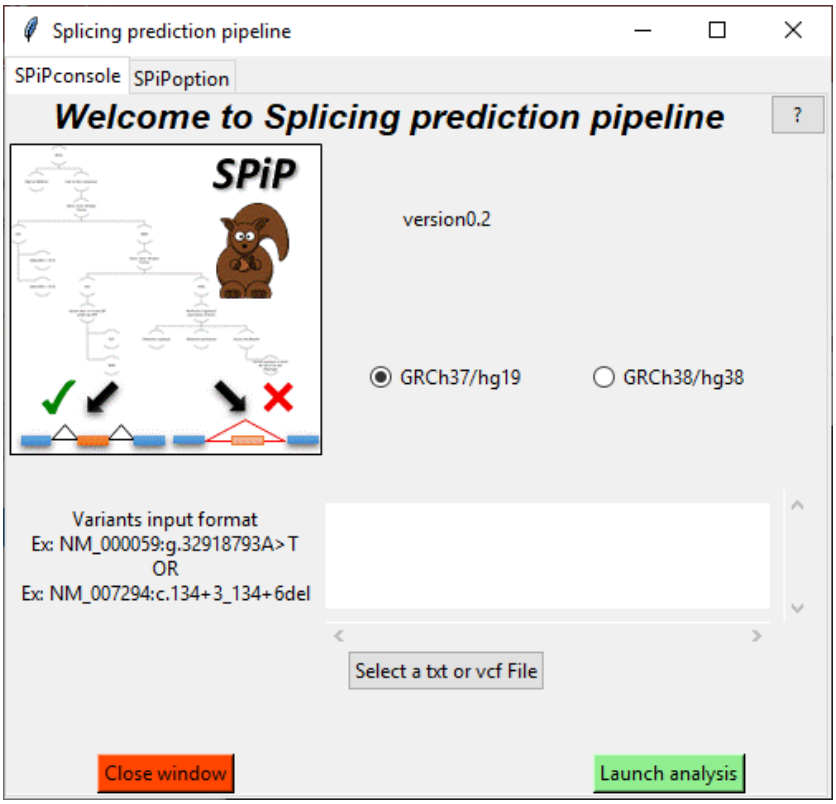
SPiP installation

SPiP is a portable software system. It is available in zip file or in executable format following your preferences. For zip file, after download, extract the files and open “SPiP.bat” to launch SPiP. For executable version, you have it in 32 and 64 bit windows versions. After download, double-click on the installer and follow the instruction. In this version a shortcut will be automatically created in your desktop.

A web connection is mandatory to request the sequences on the Ensembl database. If SPiP can’t connect to the EnsemblAPI, the program will be kill and the appropriate error message will appear in the prompt of windows console.

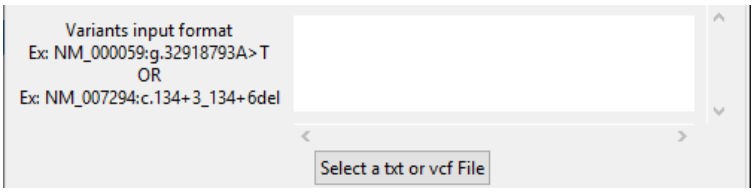
SPiP running

The SPiP console permits to select the genome version assembly, the import of variant and to set the options of SPiP.



In SPiP option, you can define what tools would be integrated in SPiP analysis, by default all tools are selected.

Import of variants



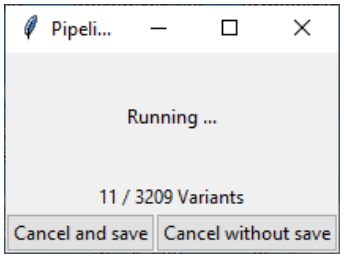
You can directly write the variations in text dialog or import a file (txt or VCF format) for a batch analysis. Excepted for VCF files, the variation must have the syntax "Transcrit:mutation". The transcripts are named according to the RefSeq database (NCBI). You can enter your variant by its position either by gDNA or cDNA coordinates. See examples below:

Mutation type	Examples
Substitution	NM_000059:c.68-4A>G NM_007294:g.41251855:G>T
Deletion	NM_007294:c.134+3_134+6del NM_133509:g.68353824_68353826del
Duplication	NM_007294:c.211dup NM_007294:g.41258474dup NM_000059:c.9501_9501+1dup
Insertion	NM_058216:c.835+5insAAC NM_024675:g.23653392insCGT
Deletion/Insertion	NM_032043:c.-30-3_-29delinsTTC NM_002878:g.33445643_33445640delinsTT

To import a variant in a txt file, the table must be tabulate separated, and you have to indicate 'varID' as the name of the column with the variant positions. A file example is providing at 'testCrypt.txt'.

For VCF file format, SPiP supports version 4 or later as shown in the example file 'testVar.vcf'. For each line, SPiP will check all transcript aligned on the mutation position. If SPiP didn't find a transcript matching with the position of the mutation, this mutation will be not analyzed. A list of mutation exclude to this issue is display at the begin of SPiP running in the prompt of windows ("I find no transcript for the mutation(s): …. ").

Runtime of SPiP
SPiP progression:

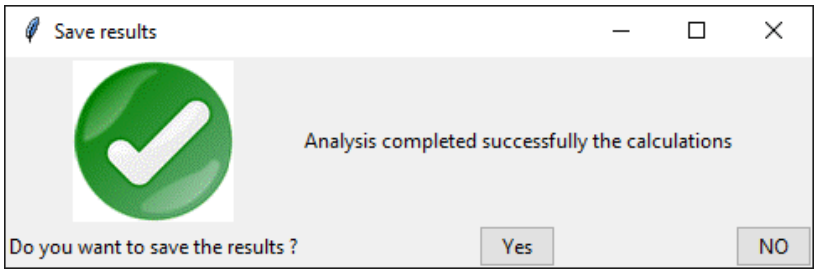


The “Cancel and save button” permits to stop the program and automatically save the current result in a text file with random name in the folder of import data.

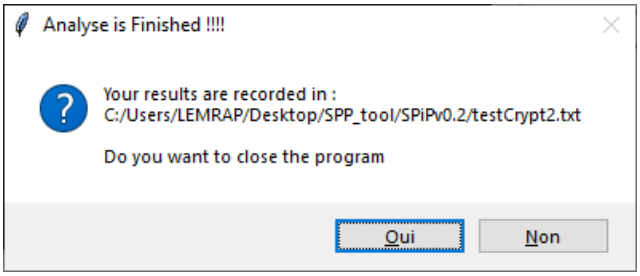
With an AMD Ryzen 7 PRO 1700 Eight-Core processor 3.00 GHz and 16 Go of RAM, SPiP toke 30-35 minutes to analyze 3,200 variants.

SPiP Output
Output depends of the input format *i.e.* either by files or by dialog boxes.

When sequences are imported as txt or VCF files, SPiP opens automatically this window:

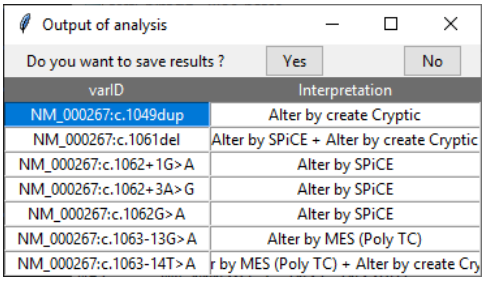


The output file is in txt format. The final window of SPiP is:



If you click on Yes (Oui) on this final windows, the program will be close and if you click on No (Non) the program will restart for another analysis.

When the box dialog was use, SPiP shows the results in a table window and allows the user to save the results as txt file. See screenshot below for the output format:



SPiP FOR LINUX

SPiP was also developed for a Linux environment. To install this version of SPiP besides the download of the SPiP package dedicated to Linux. SPiP needs an R environment with the libraries “Rcurl” and “parallel”. The installation of samtools with the human genome is preconize to improve the runtime of SPiP. Indeed, SPiP will use samtools to get the DNA sequences and no more an internet connection to get this sequence for each variant.

Install SPiP

The command line to install and to launch SPiP is:

```
$ git clone https://github.com/raphaelleman/SPiP
$ cd ./SPiP
```

Install R libraries, from R console:

```
> install.packages(“Rcurl”)
> install.packages(“parallel”)
```

(Optional) Install samtools

```
$ wget https://github.com/samtools/samtools/releases/download/1.9/samtools-1.9.tar.bz2
$ tar xfvj samtools-1.9.tar.bz2
$ cd ./samtools-1.9
$ ./configure --prefix=/where/to/install
$ make
$ make install
$ # get the human genome (here from GENCODE website: https://www.gencodegenes.org/)
$ ## Warning: download the same version of genome that the genome version of your genomic
coordinates
$ wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_29/GRCh37_mapping/GRCh37.pr
imary_assembly.genome.fa.gz -O genomehg19.fa.gz # GRCh37/hg19 example
$ gunzip genomehg19.fa.gz
$ samtools faidx genomehg19.fa
```

Run SPiP

Try the installation of SPiP with data sample:

```
$ cd /path/to/SPiP/
$ #without samtools
$ Rscript ./SPiPv0.3.r -I testCrypt.txt -O trySPiP.txt
$ #with samtools (+ hg19 assembly version)
$ Rscript ./SPiPv0.3.r -I testCrypt.txt -O trySPiP.txt -g hg19 -s /path/to/samtools -f
/path/to/fastaGenome
```

The detailed options of SPiP are:

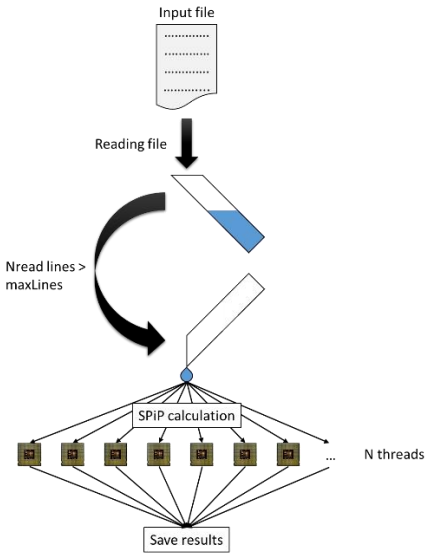
```
Usage: SPiPv0.3.r

    Mandatory

        -I, --input /path/to/inputFile              list of variants file (.txt or .vcf)
        -O, --output /path/to/outputFile            Name of ouput file (.txt)

    Genome options

        -g, --GenomeAssenbly hg19           Genome assembly version (hg19 or hg38) [default=
hg19]
        -s, --SamPath /path/to/samtools]            Path to samtools, if you want to use
Ensembl api keep this argument empty
        -f, --fastaGenome /path/to/fastaGenome              fasta file of genome used by
samtools

    Parallel options

        -t, --threads N              Number of threads used for the calculation [default= 1]
        -l, --maxLines N             Number of lines read in each time [default= 1000]

    Tools options

        -c, --SPiCE Yes              Use of consensus prediction (SPiCE) (Yes/No) [default=
Yes]
        -m, --MES Yes         Use of MES prediction in {-20; -13} (Yes/No) [default= Yes]
        -b, --BPP Yes         Use of Branch point prediction (Yes/No) [default= Yes]
        -d, --Cryptic Yes            Use of de Novo/cryptic prediction (Yes/No) [default=
Yes]
        -e, --ESR Yes         Use of ESR prediction (Yes/No) [default= Yes]

    -h, --help        Print this help message and exit

   You could : Rscript SPiPv0.3.r -I ./testCrypt.txt -O ./outTestCrypt.txt
```

SPiP and parallel analysis

RunTime of SPiP

With an AMD Ryzen 7 PRO 1700 Eight-Core processor 3.00 GHz and 16 Go of RAM, SPiP toke 8-9 minutes to analyze 3,200 variants by using samtools. Without samtools, SPiP toke 30-35 minutes to analyze the number of variants on the same machine.

During the parallel analysis, SPiP will read a part file, defined by the parameter `-l, --maxLines`, in the aim to load all file in the RAM memory. For each step, the read lines are analyzed by SPiP using the number of CPUs set by the parameter `-t, --threads`. See the illustration below:



RESULTS OF SPiP

Column header	Explanation
Interpretation	Global interpretation of SPiP
InterConfident	Confident of interpretation, ie the risk of splicing alteration
chr	Chromosome of the mutation
strand	Strand of the mutation
gNomen	Genomic position of the mutation
seqPhysio	A, C, G, T sequence of wild-type DNA
seqMutated	A, C, G, T sequence of mutated DNA
NearestSS	The nearest natural splice site type (5'/3' ss) of the mutation
distSS	Relative distance between mutation and splice site
RegType	Type of region where mutation occurring
SPiCEproba	Score of SPiCE
SPiCEinter_2thr	Interpretation class of SPiCE
deltaMES	MES variation score between wild-type and mutated sequences
mutInBParea	Mutation in Branch point
deltaESRscore	ESRseq variation score between wild-type and mutated sequences
posCryptMut	Position of the strongest cryptic site reinforced by the mutation
sstypeCryptMut	Splice type of the cryptic site
probaCryptMut	Cryptic score
classProbaCryptMut	Class of cryptic site
nearestSStoCrypt	Nearest natural splice site type to the cryptic site
nearestPosSStoCrypt	Position of the nearest natural splice site to the cryptic site
nearestDistSStoCrypt	Distance between the nearest natural splice site and the cryptic site
posCryptWT	Position of the strongest cryptic site before the mutation
probaCryptWT	Score of the strongest cryptic site before the mutation
classProbaCryptWT	Class of the strongest cryptic site before the mutation
posSSPhysio	Position of the natural splice site
probaSSPhysio	Score of the natural splice site
classProbaSSPhysio	Class of the natural splice site
probaSSPhysioMut	Score of the natural splice site after the mutation
classProbaSSPhysioMut	Class of the natural splice site after the mutation

SPiP global interpretation (columns: Interpretation, InterConfident)

In the column Interpretation, SPiP summarize the alteration detection of the mutations. The different ways of splicing alteration are annotated as follow:

Annotation	Signification
Alter by SPiCE	The mutation alters the consensus splice sites
Alter by MES (Poly TC)	The mutation alters the polypyrimidine tract (-20 to -18)
Alter BP	The mutation alters a branch point
Alter by create Exon	The mutation creates a pseudo-exon
Alter by create Cryptic	The mutation creates a cryptic/de novo splice site
Alter ESR	The mutation alters the exonic splicing regulatory elements
NTR	The mutation has no effect on splicing

When a mutation affects splicing by several ways, these ways are display together. For example, a mutation that disturbs the consenus splice site and create a cryptic site, the column Interpretation will contain “Alter by SPiCE + Alter by create Cryptic”.

In the column InterConfident, the predictions are weighted by the risk of altering splicing accordingly the predictions and the position of mutation. These risk levels were estimated from a collection of 2,784 variants with their *in vitro* RNA study and 63,171 SNPs as control data. these different levels are illustrated in the next figure. In the case where a mutation a mutation is reported as two different effects, the risk displayed is the higher among the different way of alterations.

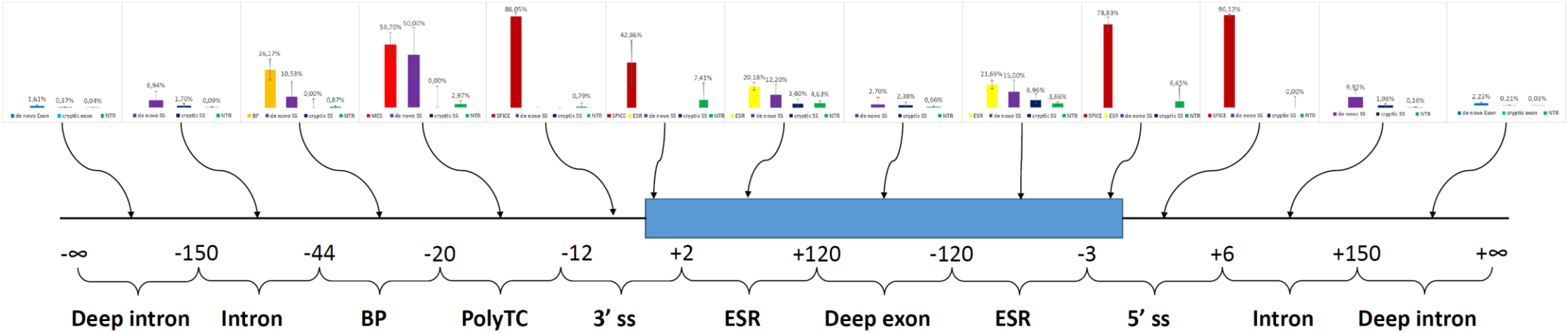
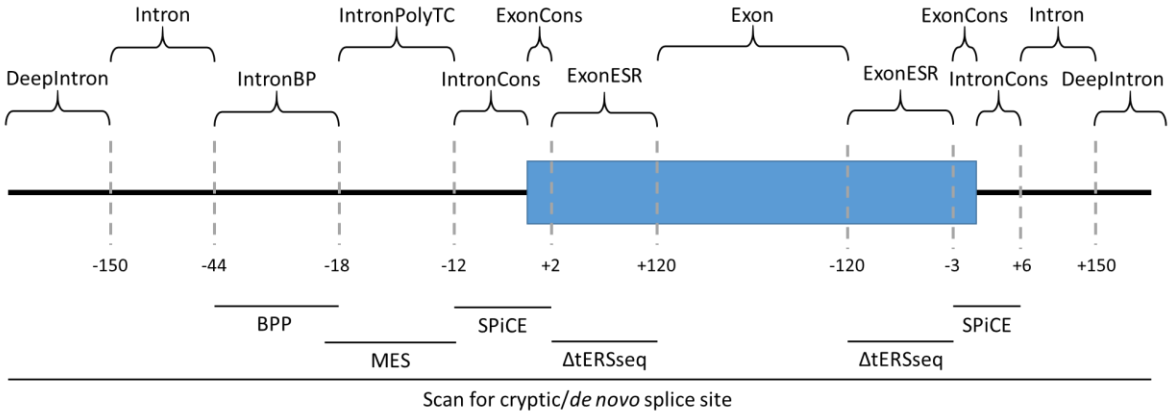


Figure 2: Probability of splicing alteration according the prediction and the localization of the mutation.

General information (columns: chr, strand, gNomen, seqPhysio, seqMutated, NearestSS, distSS, RegType)

The chromosome, strand and genomic coordinates are given accordingly the hg19 or hg38 assembly genome following your option choice at the begin of SPiP. The seqPhyio and seqMutated column contain the DNA sequences 150 nt around the position of mutation for the wild-type and mutated sequences respectively. The NearestSS and distSS columns contain the information about the nearest natural splice site. The RegType column gives information about the localization of mutation in the transcript.



SPiCE tool (columns: SPiCEproba and SPiCEinter_2thr)

SPiCE tool was used to score the consensus splice site acceptor/donor (3'ss/5'ss), defined as -12; +2 for 3'ss and -3; +6 for 5'ss. SPiCE displayed a score between 0 and 1. This score was used to class variant in three categories: low, medium and high. The categories medium and high are considered as impacting splicing and low-category as no impact on splicing. Any variant outside the consensus splice site are annotated with a score at 0 and class as Outside SPiCE interpretation.

MES tool (column: deltaMES)

MES tool was used for variant occurring in the polypyrimidine tract (defined from -20 to -12). A decreasing of score below -15 % of natural score is considered as an alteration of this tract and then an alteration of splicing.

BPP tool (column: mutInBParea)

A comprehensive collection of BPP-predicted branch points was implemented in SPiP. We set one branch point (with maximal score) for each intron in transcripts described by RefSeq database. The variants located in branch point area (from -44 to -18) are aligned to this collection of branch points. If a variant occurs in the 4-mer of branch point (motif: TRAY), the variant is annotated as altering the splicing by branch point alteration. In this case the output is 'Yes g.29554209 (-27): 6.05': Yes: means that the the variant occurs in the 4-mer of branch point, 'g.29554209': the genomic coordinate of branch point, '(-27)' the relative distance to the natural 3'ss and '6.05': the score of BPP for this branch point.

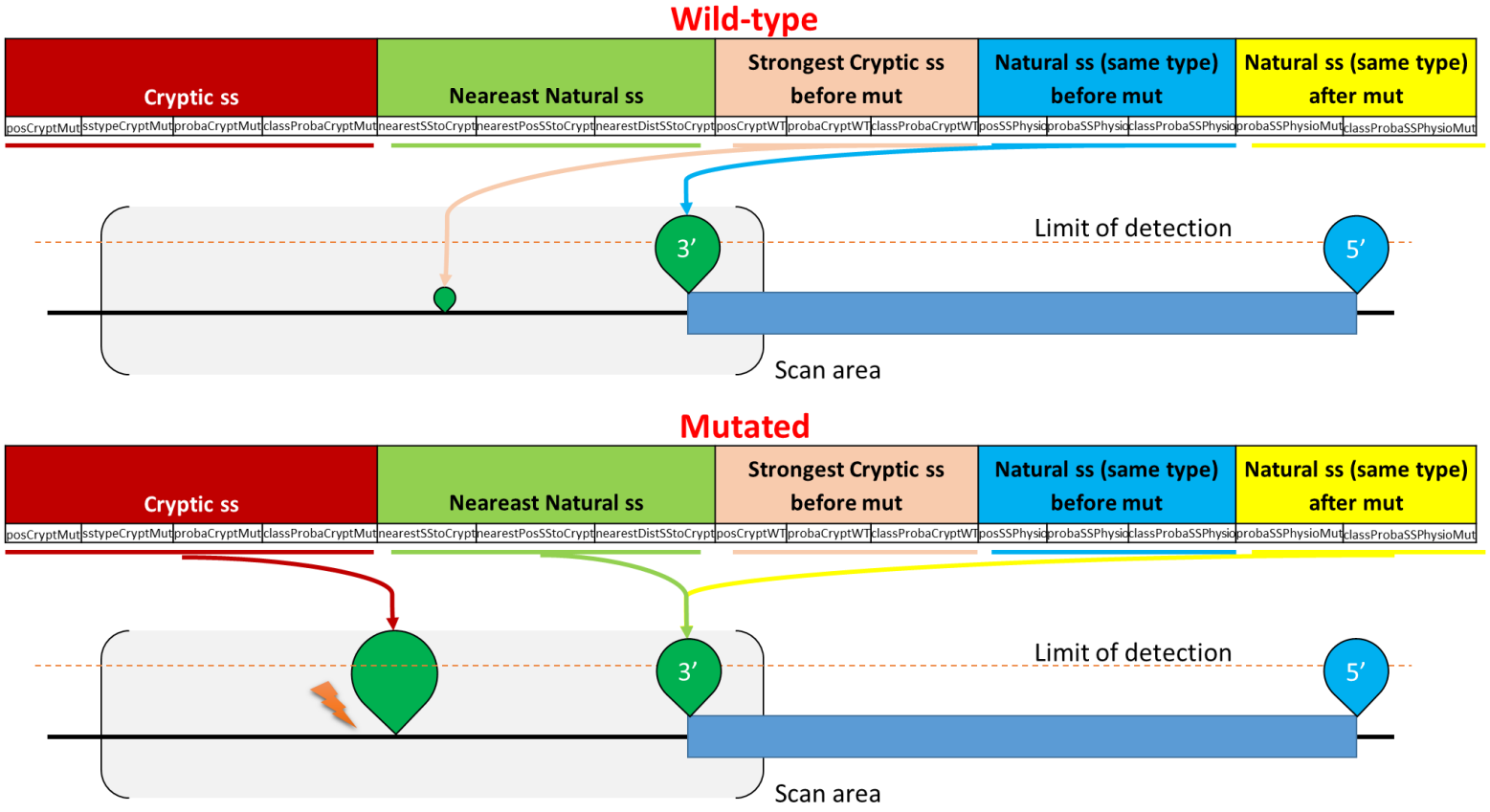
ΔtESRseq (column: deltaESRseq)

For variants in exon and at least 120 nt of natural slice site are scored by ΔtESRseq. If variant decreases the score above -1.10, we consider it as disrupting the ESR motifs and then impacting the splicing.

Scan for cryptic/*de novo* splice site (columns: posCryptMut to classProbaSSPhysioMut)

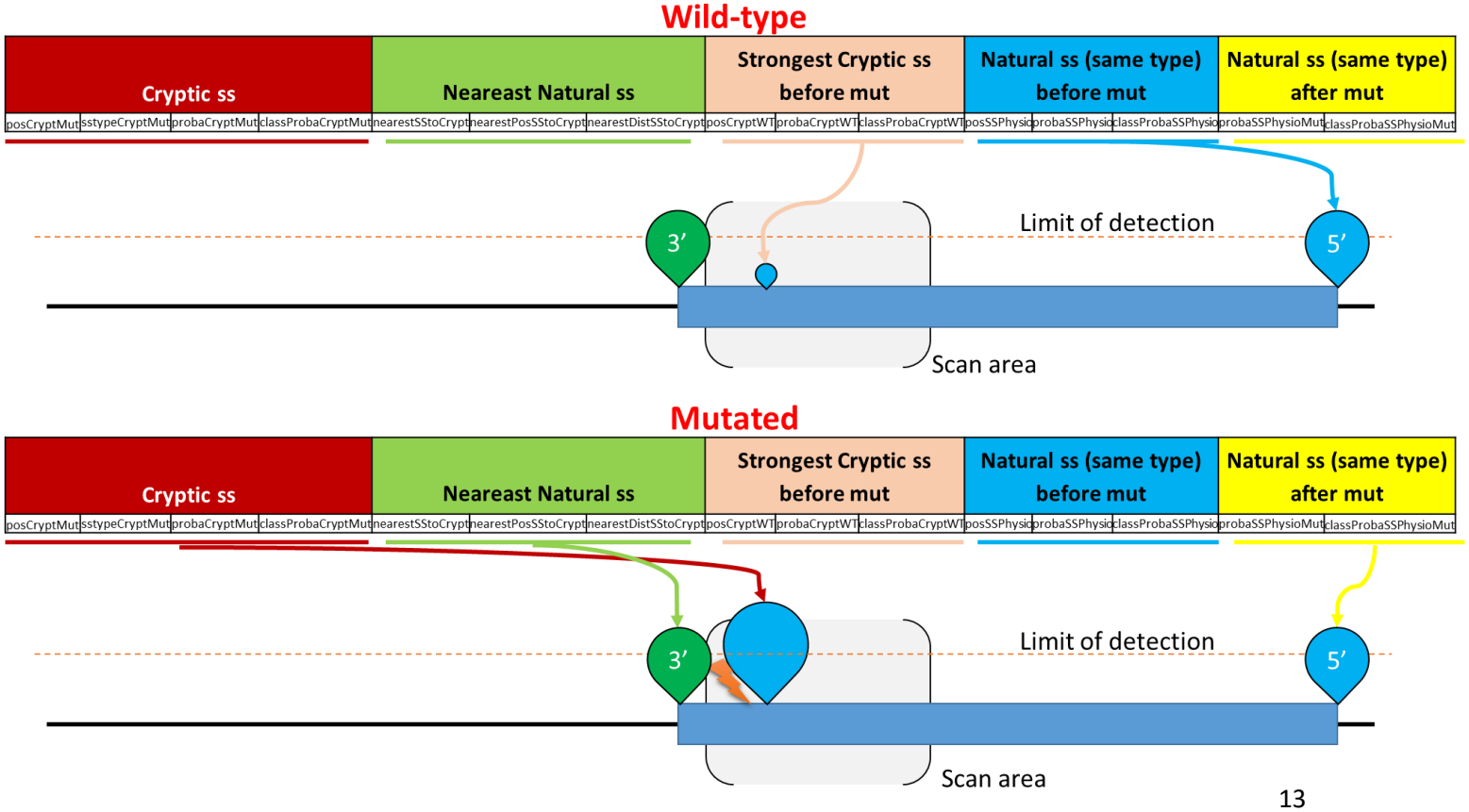
Due to the complexity of cryptic/de novo splice site prediction, SPiP displays not only the score of cryptic splice site but also the score of the splice sites around the cryptic site. Briefly SPiP gives the score of cryptic sites reinforced by the mutation, the nearest natural splice site to this cryptic site, the score of strongest cryptic sites before the mutation and the score of nearest natural splice site of the same type than cryptic site before and after the mutation. To illustrate this, some examples were shown:

Simple cryptic activation



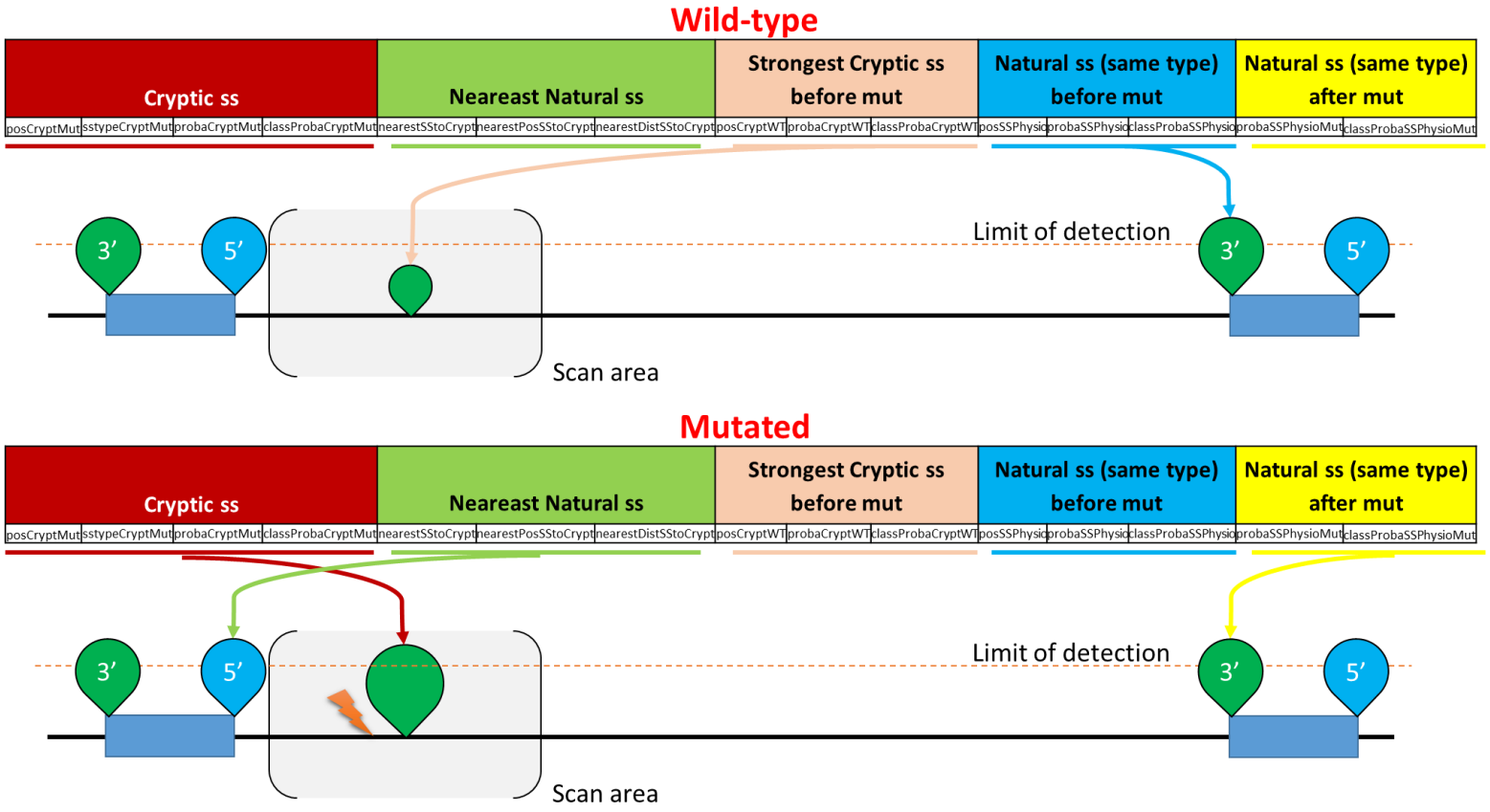
This is the simplest case, with a variant reinforcing a cryptic side beyond the threshold of cryptic detection.

Cryptic activation but two different nearest splice sites



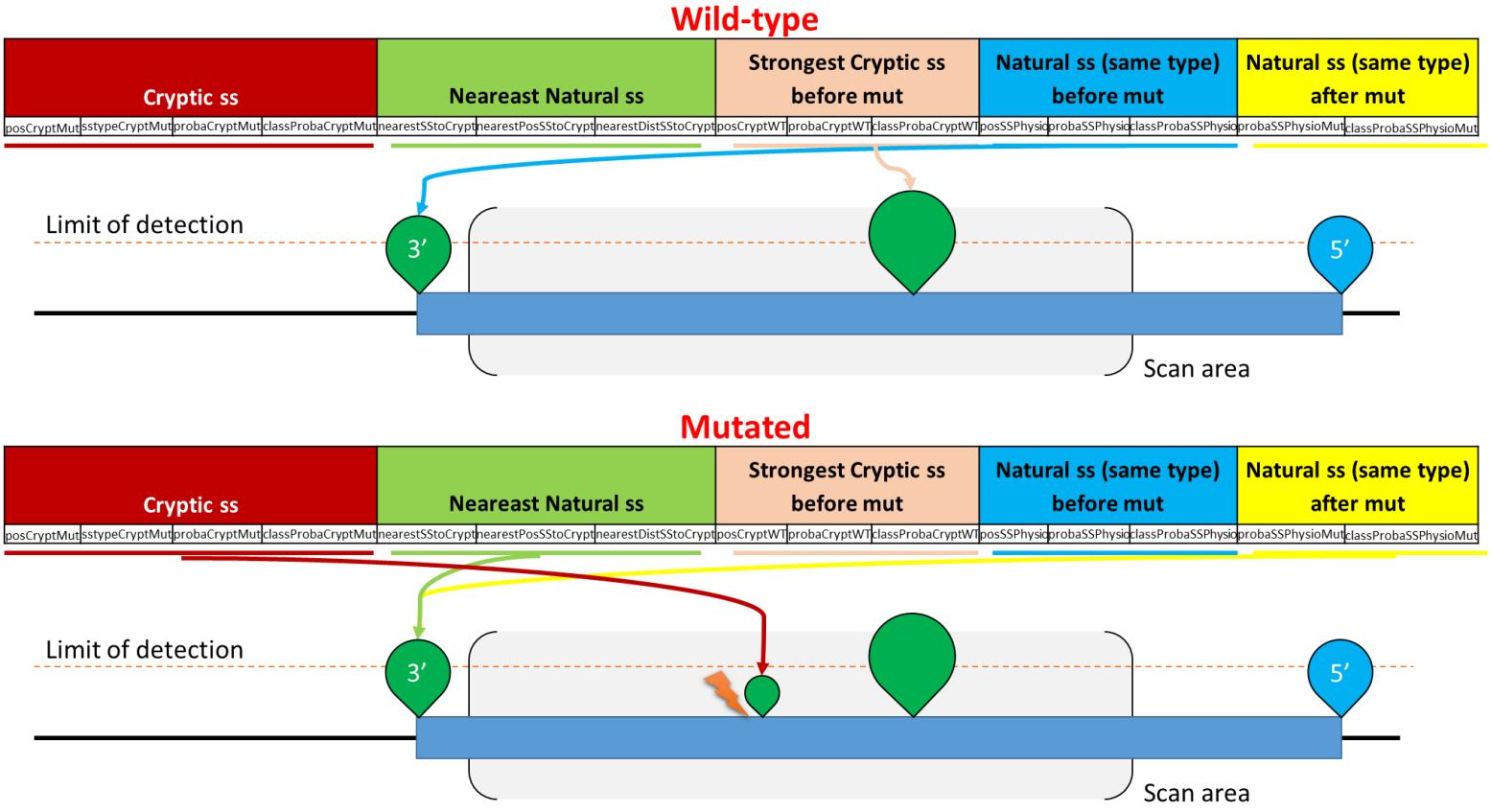
The mutation active a donor cryptic site near to the natural acceptor splice site. So the nearest splice site was this acceptor site but the score displays by SPiP of the natural splice site was the score of the nearest donor splice site, to be coherent.

SPiP selects the relevant spilce site



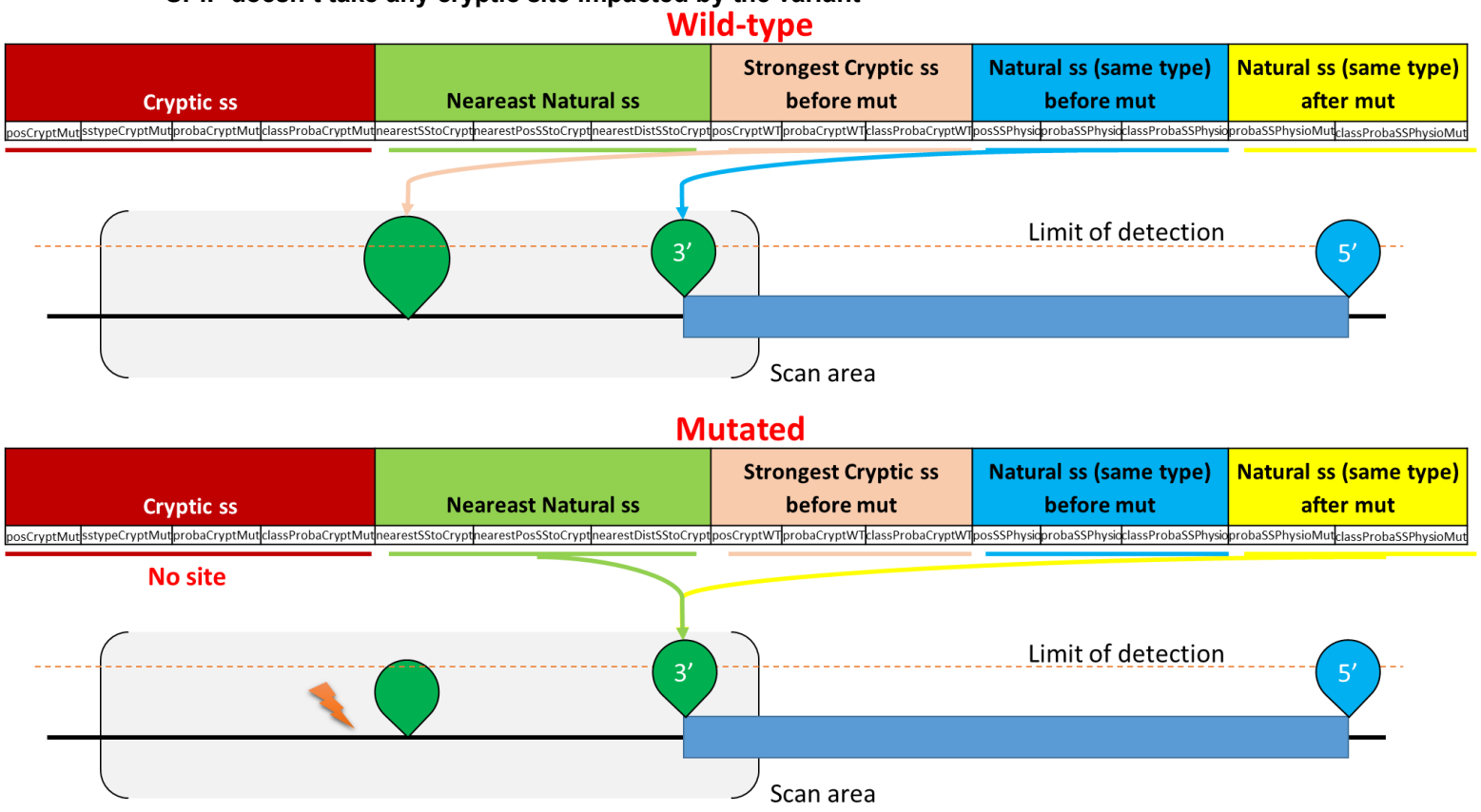
Here we have a similar configuration that in the precedent case. However, in this case, the cryptic site was an acceptor site and the nearest acceptor site was upstream of the nearest donor site. Also the use of this acceptor score to compare with the cryptic site score was not relevant. Thus SPiP toke the nearest acceptor site downstream of nearest donor site to take a relevant acceptor site as reference, even if this last was more far than the initial acceptor site.

SPiP doesn't takes systematically the stronger cryptic site



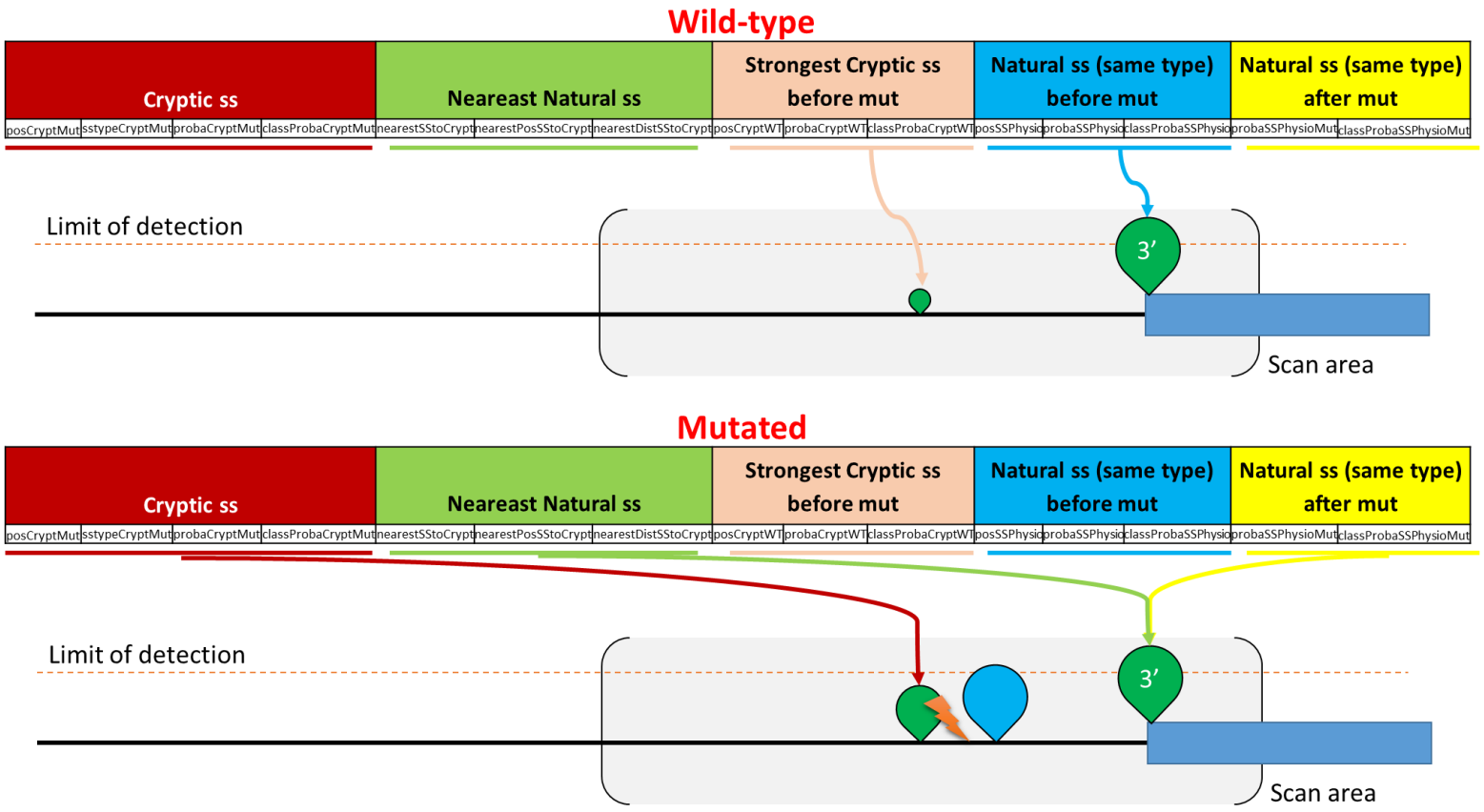
The mutated sequence shown two acceptor cryptic sites, the first was created by the mutation but below the detection threshold, the second was above the detection threshold but not impact by the mutation. SPiP displayed only the cryptic splice site impacted by the mutation even if the presence of stronger cryptic site and even if the mutation-impacted cryptic site was below the detection threshold.

SPiP doesn't take any cryptic site impacted by the variant



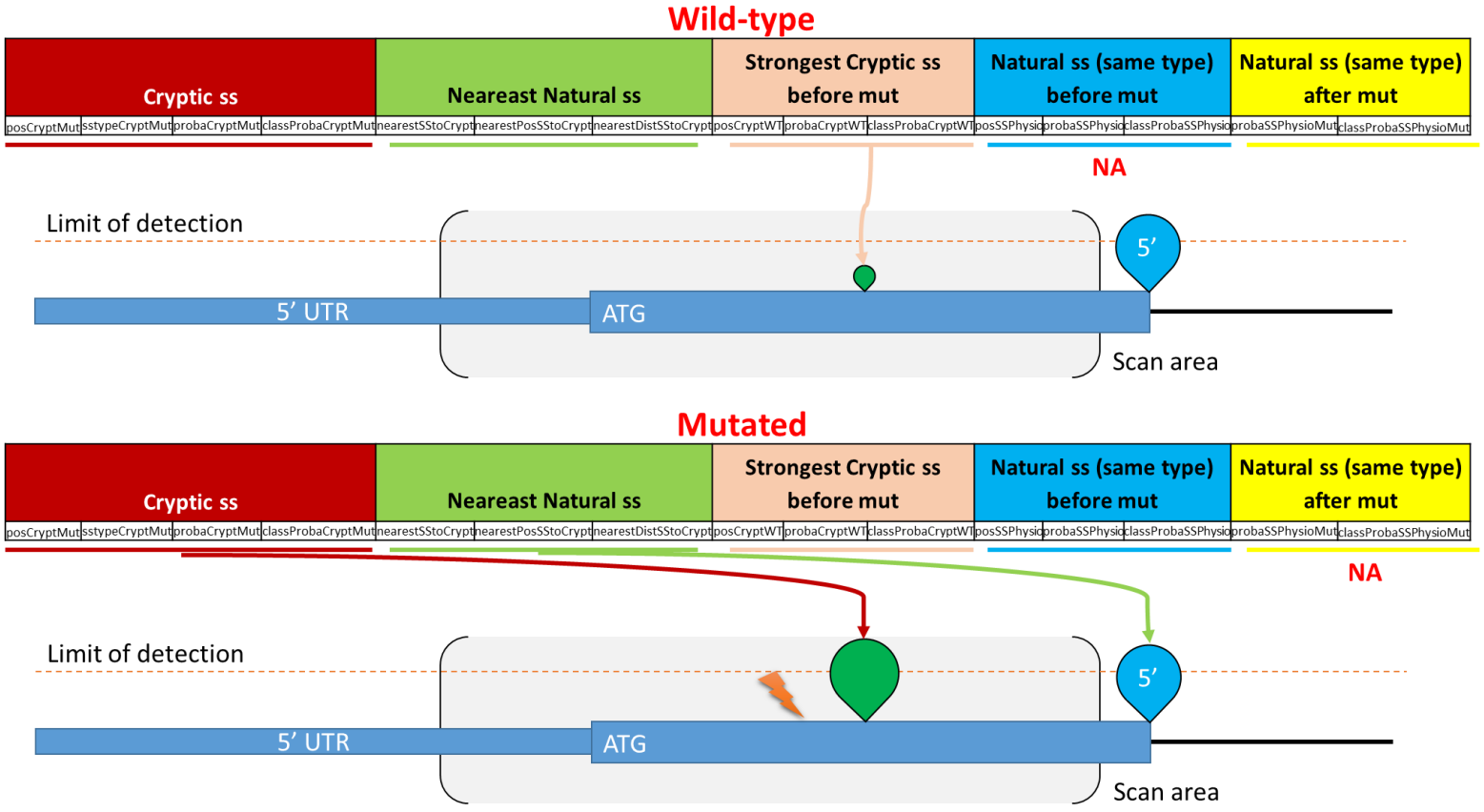
In this case, the mutation decreased the score of the acceptor cryptic site but this last remained above the limit of detection. SPiP didn't display this site even if the score of this site was above the detection threshold.

SPiP doesn't take any cryptic site impacted by the variant 2



The mutation activated two cryptic splice sites, an acceptor and donor splice site. In the area of scan, SPiP found also a natural acceptor site. So SPiP toke the acceptor cryptic splice site even if this site had a lower score than the donor cryptic splice site. This filter works only if we have an natural splice site in the area of scan.

The non-cassette exon case



The mutation activated an acceptor cryptic splice near to the 5' UTR. In this case cryptic site was displayed but not the natural splice site to compare the score. SPiP shown instead a non-available data.

ERROR MESSAGES AND RESOLUTIONS

Error message	Description	Resolution
No file was selected!	No file selected while you clicked on button "Select a txt or csv File"	Select file in txt or in csv format
Incorrect format of input, please try again!	SPiCE cannot open your file	Check if your file is in txt or in csv format
I don't find the varID column, please try again!	SPiCE cannot find column with variant positions in your file	Name of variant position column must be "varID"
No input recording, please try again!	SPiCE was launched without sequences of consensus splice site	Save sequences of consensus splice sites either with dialog box or with file in txt or csv format
You must import the variant as: Transcrit:position nucleotidic change	There is an error in the variant position input format	You must enter variant position as Transcrit:position nucleotidic change e.g.: NM_000059:c.68-4:A>G

REFERENCES

1. Burge: 20 Splicing of Precursors to mRNAs by the Spliceosomes [Internet]. [cited 2017 Sep 27]. Available from: https://cshmonographs.org/index.php/monographs/article/view/5123

2. Yeo G, Burge CB. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. Journal of Computational Biology. 2004 Mar 1;11(2–3):377–94.

3. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 1987 Sep 11;15(17):7155–74.

4. Burset M, Seledtsov IA, Solovyev VV. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. 2001 Jan 1;29(1):255–9.