SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effect on mRNA splicing.

Raphaël Leman[1], Béatrice Parfait[2], Dominique Vidaud[2], Emmanuelle Girodon[2], Laurence Pacot[2], Gérald LE GAC[3], Chandran Ka[3], Claude Ferec[3], Yann Fichou[3], Céline Quesnelle[1], Camille Aucouturier[1], Etienne Muller[1], Dominique Vaur[1], Laurent Castera[1], Flavie Boulouard[1], Agathe Ricou[1], Hélène Tubeuf[4], Omar Soukarieh[4], Pascaline Gaildrat[4], Florence Riant[5], Marine Guillaud-Bataille[6], Sandrine Caputo[7], Virginie Moncoutier[7], Nadia Boutry-Kryza[8], Françoise Bonnet-Dorion[9], Ines Schultz[10], Maria Rossing[11], Olivier Quenez[4], Louis Goldenberg[4], Valentin Harter[1], Michael Parsons[12], Amanda Spurdle[12], Thierry Frébourg[4], Alexandra Martins[4], Claude Houdayer[4], and Sophie Krieger[1]

[1]Centre Francois Baclesse Centre de Lutte Contre le Cancer
[2]Hopital Cochin Service de Radiologie
[3]Universite de Bretagne Occidentale
[4]Universite de Rouen
[5]Groupe Hospitalier Saint-Louis Lariboisiere et Fernand-Widal
[6]Gustave Roussy
[7]Institut Curie Departement d'Oncologie Medicale
[8]Hospices Civils de Lyon
[9]Institut Bergonie
[10]Centre Paul Strauss
[11]Rigshospitalet
[12]QIMR Berghofer Department of Genetics and Computational Biology

February 21, 2022

Abstract

Modeling splicing is essential for tackling the challenge of variant interpretation as each nucleotide variation can be pathogenic by affecting pre-mRNA splicing *via* disruption/creation of splicing motifs such as 5'/3' splice sites, branch sites or splicing regulatory elements. Unfortunately, most *in silico* tools focus on a specific type of splicing motif, which is why we developed the Splicing Prediction Pipeline (SPiP) to perform, in one single bioinformatic analysis based on machine learning approach, comprehensive assessment of variant effect on different splicing motifs. We gathered a curated set of 4,616 variants scattered all along the sequence of 227 genes, with their corresponding splicing studies. Bayesian analysis provided us the number of control variants, *i.e.* variants without impact on splicing, to mimic the deluge of variants from high throughput sequencing data. Results show that SPiP can deal with the diversity of splicing alterations, with 83.13% sensitivity and 99% specificity to detect spliceogenic variants. Overall performance as measured by area under the receiving operator curve was 0.986, significantly better than 0.965 spliceAI for the same dataset. SPiP lends itself to a unique suite for comprehensive prediction of spliceogenicity in the genomic medicine era. SPiP is available at: https://sourceforge.net/projects/splicing-prediction-pipeline/

Hosted file

`article_SPiP_manuscript_vf.docx` available at https://authorea.com/users/461595/articles/557263-spip-splicing-prediction-pipeline-a-machine-learning-tool-for-massive-detection-of-exonic-and-intronic-variant-effect-on-mrna-splicing