TRIFID: determining functional isoforms

Fernando Pozo, Laura Martínez-Gómez, Michael L. Tress

Bioinformatics Unit, Spanish National Cancer Research Center (CNIO)

Introduction

Background

Alternative Splicing (AS) of messenger RNA can generate a wide variety of mature RNA transcripts and this expression is confirmed by experimental transcript evidence. However, it is not clear how many alternative transcripts will code for functional proteins. (Tress et al. 2017a, Blencowe et al. 2017).

Proteomics analyses have shown that most coding genes have a single main splice isoform (Ezkurdia et al. 2015).

Human population variation data indicate that most alternative transcripts are evolving neutrally (Tress et al. 2017).

Ever more splice isoforms are annotated every day, but we do not know their function.

Objectives

Grading what proportion of AS is functional.

Developing a machine learning based tool for predicting splice isoform functional importance.

Methods

`TRIFID` (Tool to Reliable Identification of Functional Isoform Data) is a Random Forest based **predictor of the relative functional importance of splice isoforms**.

It has been trained on reliable peptide evidence from 497 genes from the largest tissue-based proteomics analysis to date (79 experiments) and 47 features categorized in 5 groups (*genome annotation*, *structural*, *splicing impact*, *cross-species conservation* and *RNA-seq expression*).

Code repository available on github.com/fpozoc/trifid.

Results

SHAP feature importances and model interpretation

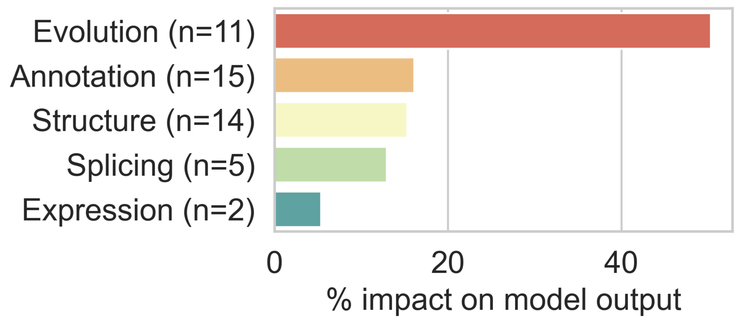

Figure 1: SHAP values (Lundberg et al. 2020) added by category. Features that best distinguish functional isoforms in the training set are conservation-based. Other important features include the length difference between the alternative isoform and the longest isoform, whether or not the transcript has a CCDS (Pruitt et al. 2009), and the conservation of Pfam functional domains. The SHAP scores can also provide clues to the influence of features on individual predictions (see example: Fibroblast growth factor receptor 1 (TRIFID Scores)

Functional importance in the human genome

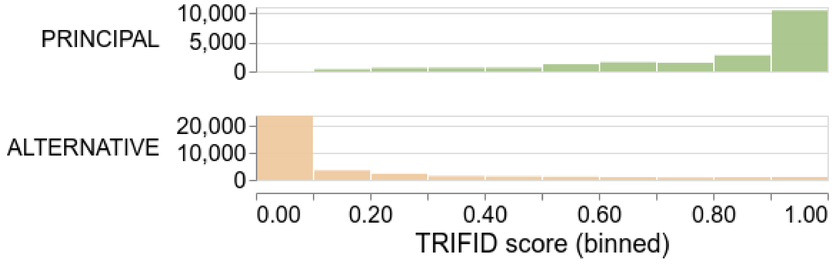

Figure 2: Non-redundant isoforms divided into PRINCIPAL or ALTERNATIVE according to their annotation in APPRIS (Rodriguez et al. 2017). Most ALTERNATIVE isoforms have TRIFID scores below 0.05. Most PRINCIPAL isoforms have predictor scores above 0.9.

This work has been funded by the US National Institutes of Health grant 2 U41 HG007234.

Validating the model against an external source of information

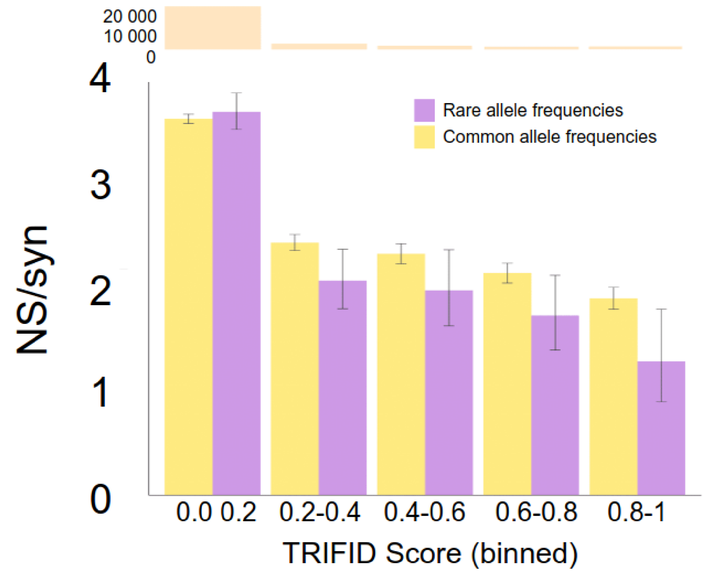

Figure 3: Non-synonymous to synonymous (NS/Syn) ratios for alternative exons (exons that are exclusively present in alternative isoforms). Exons under selective pressure should have significantly lower NS/Syn ratios for common than for rare allele frequencies. The vast majority of alternative exons are not under selective pressure.

Conclusions

`TRIFID` discriminates functionally important isoforms with high confidence (MCC=0.89, AUPRC=0.98 over 5-folds CV of the training set). Thanks to TRIFID, we can now list the most biologically relevant alternative isoforms. Our model can be successfully exported to different genome species and genome annotation databases.

`TRIFID` predicts that a large majority of splice variants (85-90%) in the human genome are likely to not be functionally important at the protein level.

NS/Syn ratios show that exons from the highest scoring of alternative transcripts are under selective pressure, while low scoring exons have little or no evidence of selection.