APPRIS Principal and MANE Select Define Reference Splice Variants

Fernando Pozo, Laura Martínez Gómez, Jose Manuel Rodríguez, Jesús Vázquez, Michael L. Tress

Bioinformatics Unit, Spanish National Cancer Research Center (CNIO)

Introduction

Motivation

Selecting the splice variant that best represents a coding gene is a crucial first step in many experimental analyses, and vital for mapping clinically relevant variants.

Objectives

To determine which method is best (**APPRIS principal, MANE Select transcript, longest isoform or transcirpt expression**) for selecting biological important reference splice variants for large-scale analyses.

Availability of Data

Now appris.bioinfo.cnio.es contains the list of splice variants where APPRIS and MANE agree

Methods

GENCODE v37 gene set: Provided APPRIS principal isoforms, MANE (Matched Annotation from NCBI) Select transcripts, and longest isoforms/CDS

RNA-seq from Human Protein Atlas: data from 36 dierent human tissues was leveraged with QSplice (our in-house method to quantify splice-junctions) and RSEM transcript counts.

Large-scale proteomics: Five datasets covering 52 distinct tissue types were analysed with Comet and Percolator. We used peptide-spectrum match counts to determine main proteomics isoforms.

Germline variants: for all sets of exons we calculated the NS/Syn ratios for both rare and common allele frequencies using variants from the 2504 individuals in the 1000 Genomes Project phase 3.

Results

Main protein isoforms and reference predictions agreement

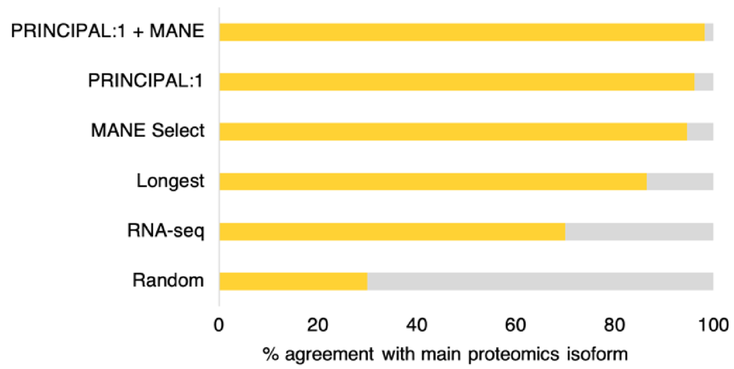

Figure 1: The percentage of genes in which predicted reference isoforms coincided with proteomics main isoforms.

Reference selected by different methods for RAB7A

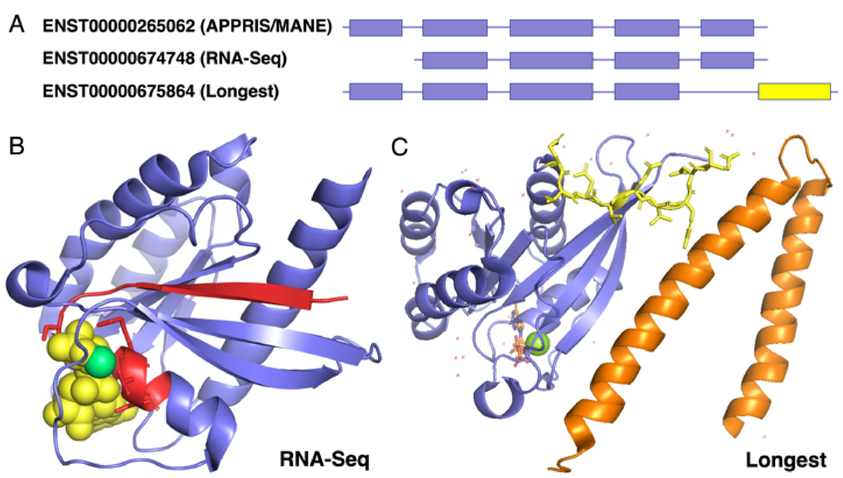

Figure 2: APPRIS and MANE chose the same transcript. The RNA-Seq transcript (ENST00000674748) is missing the first coding exon; the longest isoform has a different 3' CDS (in yellow). APPRIS and MANE both select the highly conserved 5' exon transcript as the most important splice variant.

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

MANE and APPRIS main exons are under purifying selection

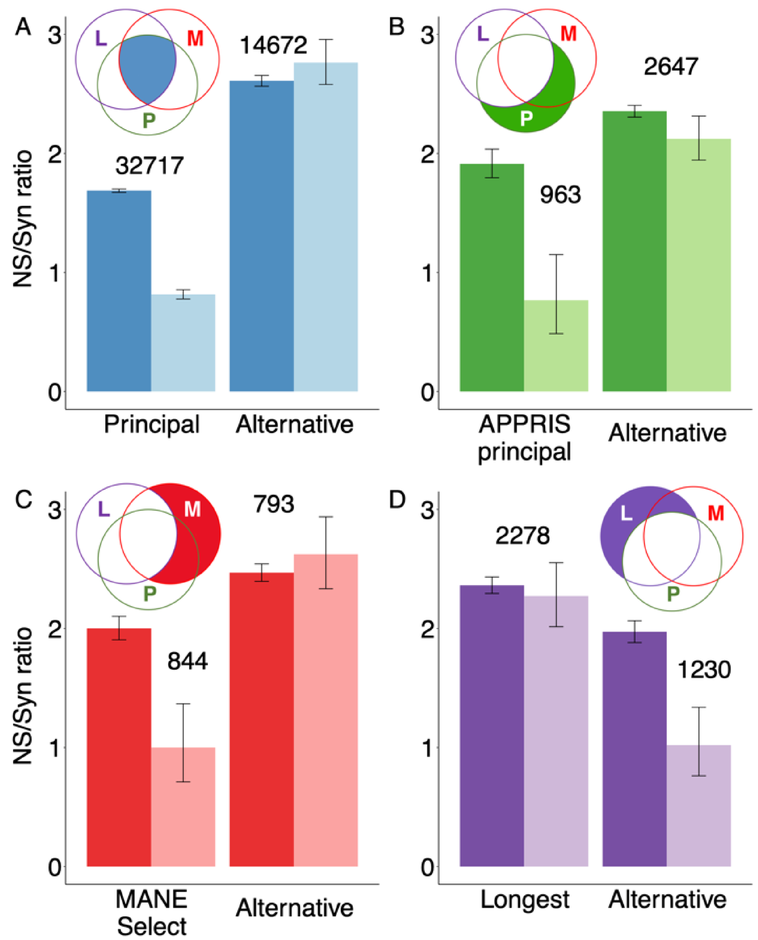

Figure 3: Ns/Syn ratios for rare (dark bars) and common (light bars) variants (95% CI) for reference and alternative exons. The number of exons in each set is indicated above. Exons "L" are those that produce the longest isoform, "M" are present in MANE Select transcripts, and "P" represent APPRIS principal isoforms.

Conclusions

The main cellular isoform is best described by APPRIS principal isoforms and MANE Select transcripts, and these methods are particularly powerful when they agree.

Researchers should use these 2 sets of reference transcripts, rather than the longest isoform, in all biomedical research.