{APPRIS}

SELECTS THE DOMINANT CELLULAR PROTEIN ISOFORM

http://appris.bioinfo.cnio.es

Jose Manuel Rodríguez[1], Angel Carro[2], Michael Tress[2], and Alfonso Valencia[1,2].
[1]Spanish National Bioinformatics Institute (INB) and [2]Structural Biology and Biocomputing Programme.
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.

ABSTRACT

The cellular role of alternative protein isoforms is a topic of growing interest. We have developed the APPRIS database (1) and **APPRIS Webserver and Web Services** (2) to annotate splice variants with information relating to **protein structure, function and cross-species conservation**.

APPRIS makes use of the conservation of protein features to identify a **single dominant** (3) **isoform for each gene**. These **principal isoforms** are confirmed by orthogonal theoretical analyses (4) and by the results of multiple large-scale mass spectrometry experiments and databases (5,6).

APPRIS is stable and is implemented as part of the **GENCODE/Ensembl human genome annotation** (7), and it has been also applied for **RefSeq** (10) and **UniProt** (11).

APPRIS has recently been expanded to ***mouse***, ***rat***, ***pig***, ***chimpanzee***, ***zebra fish***, and also to ***Drosophila*** and ***C. elegans***.

METHODS

firestar[8]

Functionally Important Residues

Functional residues are highly conserved, even across large evolutionary distances.

Since these residues are **unlikely to have arisen by chance** we can use this to help determine the principal isoform.

http://firedb.bioinfo.cnio.es/Php/FireStar.php

Matador3D

Protein structural information

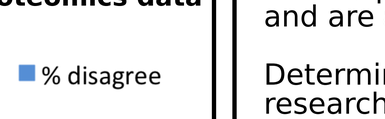
Since protein structure is much more conserved than sequence variants with large **inserts or deletions relative to their crystal structures** are also not likely to be the principal isoform.

CORSAIR

BLAST against vertebrates

Transcripts **conserved** over greater evolutionary distances are **more likely to be the principal variant**. Good alignments with more distant relatives (Danio, Xenopus, Chicken) are regarded as more valuable.

The more species that align correctly and without gaps, the better.

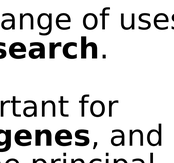


SPADE

Conservation of protein functional domains

Identifying the functional domains present in a variant can **provide insights into the function**.

The presence of protein domains is analysed with **Pfamscan** (9).



SELECTION OF PRINCIPAL ISOFORM FOR *DNAJC5G*

Seq. id	Seq. name	Length (aa)	Biotype	Codons not found	CCDS	Principal Isoform
ENST00000296097	DNAJC5G-001	189	protein_coding	-	CCDS1744.1	MINOR
ENST00000402462	DNAJC5G-002	189	protein_coding	-	CCDS1744.1	MINOR
ENST00000404433	DNAJC5G-004	173	protein_coding	-	-	PRINCIPAL:1
ENST00000406962	DNAJC5G-003	104	protein_coding	-	-	MINOR
ENST00000420191	DNAJC5G-007	62	protein_coding	stop	-	MINOR

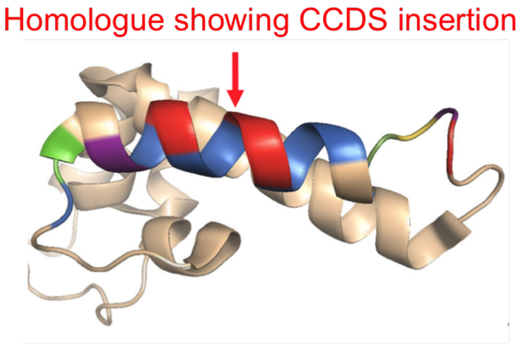
Seq. id	Seq. name	Length (aa)	No. Functional Residues	3D Structure Score	Whole Domains
ENST00000296097	DNAJC5G-001	189	0	1.8	0
ENST00000402462	DNAJC5G-002	189	0	1.8	0
ENST00000404433	DNAJC5G-004	173	0	1.8	1
ENST00000406962	DNAJC5G-003	104	2	0.8	0
ENST00000420191	DNAJC5G-007	62	0	0.8	0

Snapshot of the APPRIS web page, showing the five protein-coding transcripts annotated by GENCODE/Ensembl and the selection of the **principal isoform** by APPRIS (green). The variant selected by APPRIS (*DNAJC5G-004*) has a **conserved Pfam domain**.

The highlighted methods SPADE and Matador3D map Pfam functional domains, protein structure to the splice isoforms.

The principal isoform for *DNAJC5G* has 16 fewer residues than the **longest isoform**, **which has an inserted exon that would compromise Pfam domains and 3D structure**

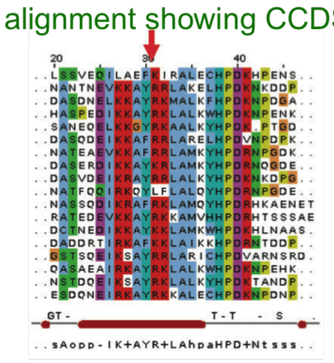
Homologue showing CCDS insertion



The 3D structure of mouse DNAJ subfamily C2 member 5 (PDB:2CTW), to DNAJC5G-004 has 56% identity with no gaps.

The large red arrow shows that the 16 extra residues in the alternative isoform would insert into an important helix.

Pfam alignment showing CCDS insertion

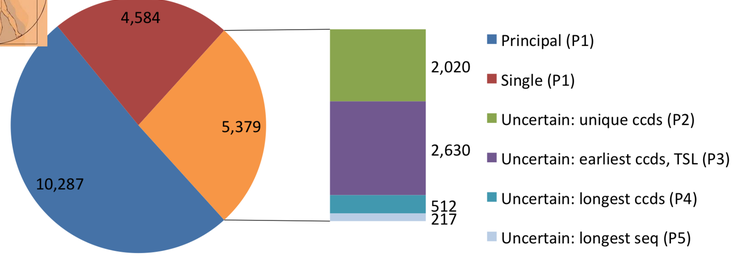


The multiple alignment for a section of the Pfam DNAJ family of sequences.

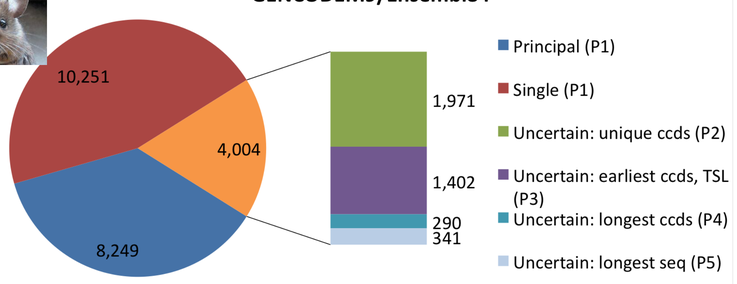
The red arrow shows that the 16 extra residues in the alternative isoform would insert into a critical region of the functional domain of DNAJC5G.

GENOME COVERAGE

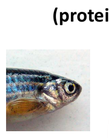
Coverage of Human genome (protein-coding genes) GENCODE24/Ensembl84

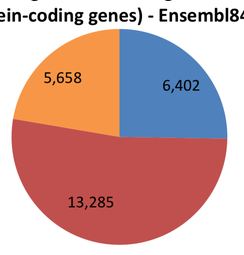


Coverage of Mouse genome (protein-coding genes) GENCODEM9/Ensembl84

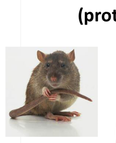


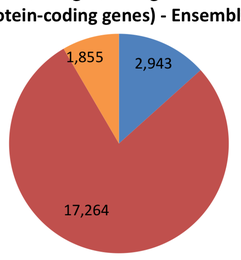
Coverage of Zebra-fish genome (protein-coding genes) - Ensembl84





Coverage of Rat genome (protein-coding genes) - Ensembl84





VALIDATION

APPRIS principal isoforms agrees with proteomics data

Here we compared the **main proteomics isoform** from multiple proteomics experiments (6) with (from left to right) the CCDS (4) variants from genes with a **unique CCDS variant**, **APPRIS principal isoforms**, the **longest annotated isoform**, and the dominant transcripts carried out using **RNAseq data** by Gonzalez-Porta et al. (4) on **cell lines and tissues**.

APPRIS principal isoforms, the main isoforms from proteomics experiments and the unique CCDS isoforms have an exceptionally **high level of agreement** (**99.5%** over those genes where each selects a reference isoform).

CONCLUSIONS

APPRIS principal isoforms have a wide range of uses and are **applicable in all fields of research**.

Determining a principal isoform is important for research groups **studying individual genes**, and the designation of a single variant as the principal isoform is a **critical first step for any genome analysis**, for example studies of cancer mutations would be able to use APPRIS data to determine whether **the mutations are in principal or alternative exons.**

We believe that the principal isoforms identified by APPRIS are a significant advance on the current practice of selecting the longest variants as the reference isoform.

The **APPRIS WebServer** allows for the **annotation of splice isoforms for individual genes**, and provides a range of **visual representations** and tools to allow researchers to identify the likely effect of splicing events.

The **APPRIS WebServices** have been implemented using REST architecture that permit users to **generate annotations** automatically in **high throughput mode**.

At present the APPRIS Database houses annotations for five Ensembl species (human, mouse, rat, pig, zebra fish, chimpanzee, fruitfly and C.elegans), the APPRIS WebServer allows users to check Ensembl annotations for six other species, dog, cat, cow, opossum, chicken and fugu.

REFERENCES

1. *Rodriguez, J. et al. (2013) Nucleic Acids Res., 41, D110-7.*
2. *Rodriguez, J. et al. (2015) Nucleic Acids Res., 43(W1):W455-9.*
3. *Gonzalez-Porta et al. Genome Biol. 2013,14:R70.*
4. *Pruitt, KD et al. (2009) Genome Res. 19(7):1316-23.*
5. *Ezkurdia et al. Mol Biol Evol. 2012. 29:2265-83.*
6. *Ezkurdia et al. J. Proteome Res. 2015. 14 (4), pp 1880–1887.*
7. *Harrow, J. et al. (2012) Genome Res. 22:1775-1789.*
8. *Lopez,G. et al. (2011) Nucleic Acids Res., 39(W1):W235-41.*
9. *Finn et al. (2014) Nucleic Acids Res., 42:D222-D230.*
10. *Pruitt KD et al. (2014) Nucleic Acids Res., 42:D756-63*
11. *The UniProt Consortium. (2015) Nucleic Acids Res. 43: D204-D212*

We would like to thank:

- Angel Carro, Iakes Ezkurdia and Paolo Maietta: CNIO, Spain.
- Adam Frankish, and Jennifer Harrow: Sanger, Cambridge.
- Amonida Zadissa, and Fergal Martin: Ensembl, Cambridge.
- Mark Diekhans: UCSC, California.

CONTACT: jmrodriguez@cnio.es

If you want this amazing POSTER :-)

here you are the QR code