APPRIS

ANNOTATION OF PRINCIPAL AND ALTERNATIVE SPLICE ISOFORMS[1]

http://appris.bioinfo.cnio.es

Jose Manuel Rodríguez*, Paolo Maietta, Iakes Ezkurdia, Alessandro Pietrelli, Jan-Jaap Wesselink, Gonzalo López, Alfonso Valencia, and Michael Tress.

Structural Biology and Biocomputing Programme, and *Spanish National Bioinformatics Institute (INB).
Spanish National Cancer Research Centre (CNIO). Madrid, Spain.

ABSTRACT

Alternative splicing generates different gene products. Studies have estimated that almost **100% of multi-exon human genes**[2]produce differently spliced mRNAs. It is important to designate one of the isoforms as the **main functional isoform** in order to predict the changes in function, structure or localisation brought about by alternative splicing[3].

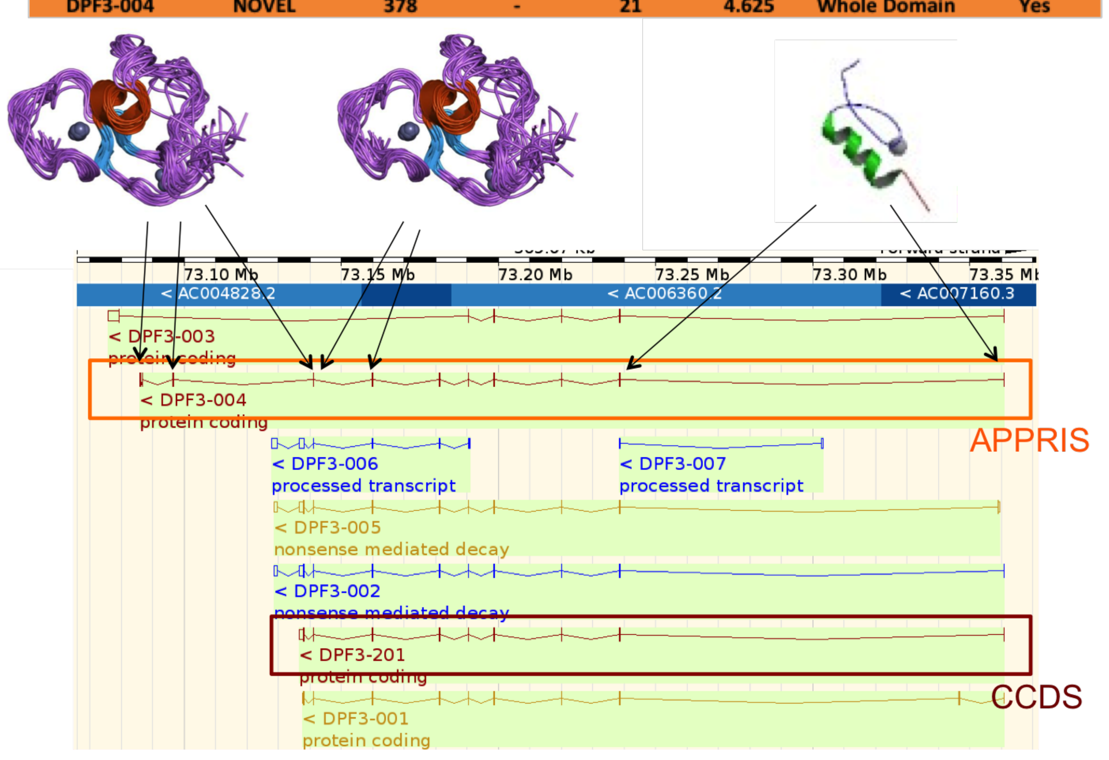
We have developed a pipeline that deploys a range of **computational methods** to annotate alternative splice isoforms by adding reliable **protein structural** and **functional** data and information from **cross-species conservation**. Based on these functional annotations, APPRIS also selects a single reference sequence for each gene, termed **the principal isoform**.

APPRIS has been developed within the **GENCODE consortium**[4] to annotate alternative human genes with reliable, biologically relevant data. APPRIS identifies a **principal isoform for 85% of the protein-coding genes** in GENCODE/Ensembl. Furthermore, APPRIS has been extended to **GENCODE Mouse consortium**, and it has been applied to other species such as **rat**, and **zebrafish**.

HOW PIPELINE WORKS with EXAMPLES

The variant selected by APPRIS (DPF3-004) clearly has all DPF3 domains (D4, zinc and double PFD fingers, family 3)

Transcript id	Status	Length (aa)	CCDS	firestar	Matador3D	SPADE	Principal
DPF3-005	KNOWN	412	-	10	2.125	No Domain	No
DPF3-002	KNOWN	**357**	**Yes**	10	2.125	No Domain	No
DPF3-003	NOVEL	195	-	0	0	No Domain	No
DPF3-201	KNOWN	**357**	**Yes**	10	2.125	No Domain	No
DPF3-001	NOVEL	367	-	10	2.125	No Domain	No
DPF3-004	**NOVEL**	**378**	**-**	**21**	**4.625**	**Whole Domain**	**Yes**



The principal isoform for DNAJC5G has 16 fewer residues than the CCDS variant

The principal isoform is highlighted.
APPRIS chooses DNAJC5G-004 isoform.

Transcript id	Name	Class	Status	Length (bp)	Length (aa)	Codons not found	CCDS	Annotated Isoform
ENST00000296097	DNAJC5G-001	protein_coding	KNOWN	2008	189	-	CCDS1744.1	✗
ENST00000402462	DNAJC5G-002	protein_coding	KNOWN	1904	189	-	CCDS1744.1	✗
ENST00000404433	DNAJC5G-004	protein_coding	NOVEL	1647	173	-	-	✓
ENST00000406962	DNAJC5G-003	protein_coding	NOVEL	1562	104	-	-	✗
ENST00000420191	DNAJC5G-007	protein_coding	NOVEL	593	62	stop	-	✗

The variant selected by APPRIS (DNAJC5G-004) has a conserved Pfam domain.

In contrast, the longer sequences would have broken Pfam domains and 3D structure.

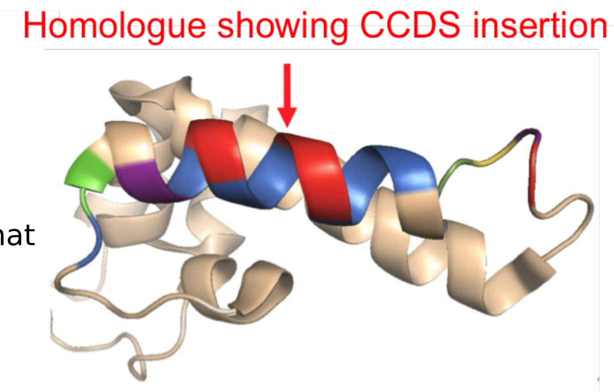
SPADE maps Pfam functional domains, and Matador3D maps 3D structure to the isoforms.

Transcript id	Status	Length (aa)	CCDS	Matador3D	SPADE	THUMP	Principal
DNAJC5G-001	KNOWN	189	Yes	1.75	Damage	0	No
DNAJC5G-002	KNOWN	189	Yes	1.75	Damage	0	No
DNAJC5G-004	**NOVEL**	**173**	**-**	**1.75**	**Whole**	**1**	**Yes**
DNAJC5G-003	NOVEL	104	-	0.75	Damage	1	No
DNAJC5G-007	NOVEL	62	-	0.75	Damage	0	No

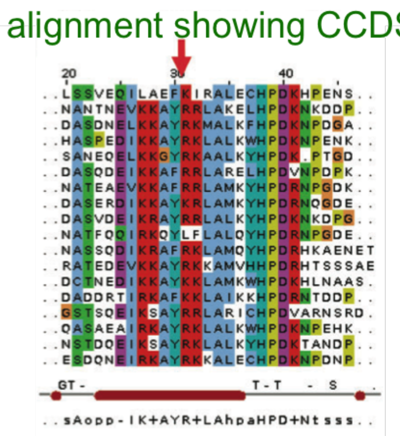
The 3D structure of mouse DNAJ subfamily C2 member 5 (PDB:2CTW), to which DNAJC5G-004 has 56% identity with no gaps

The large red arrows shows that the 16 extra residues present in the larger isoforms would have to be inserted into an important helix.

Homologue showing CCDS insertion



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 CCDS variants would need to be inserted into a critical region of the functional domain of DNAJC5G.

METHODS

firestar[5]

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

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.

CRASH and THUMP

Conservative predictions

Signal sequences and trans-membrane helices are also unlikely to have arisen by chance.

We have included conservative predictors of **signal peptide**, **mitochondrial signals** (**CRASH**), and **trans-membrane helices** (**THUMP**).

Matador3D

Protein structural information

Variants with large **inserts or deletions relative to their crystal structures** are also not likely to be the principal isoform.

Since protein structure is much **more conserved than sequence** this applies to all proteins that can be mapped reliably to **PDB structures**.

SPADE

Conservation of protein functional domains

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

Presence of protein domain is analysed with **Pfamscan**[7].

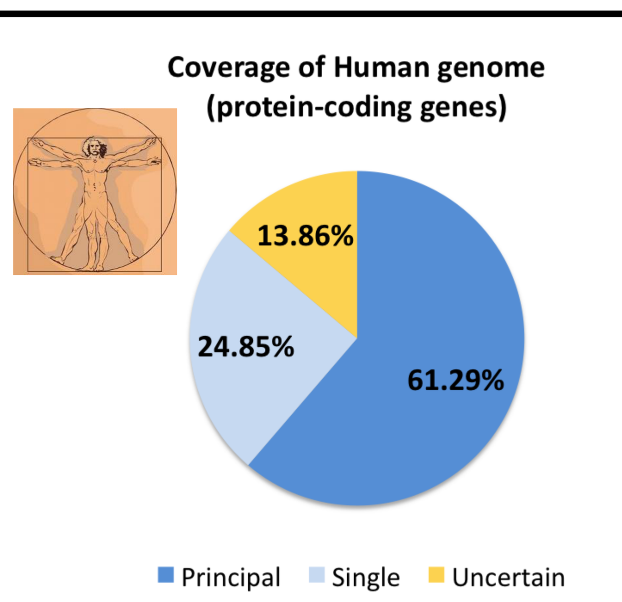
INERTIA

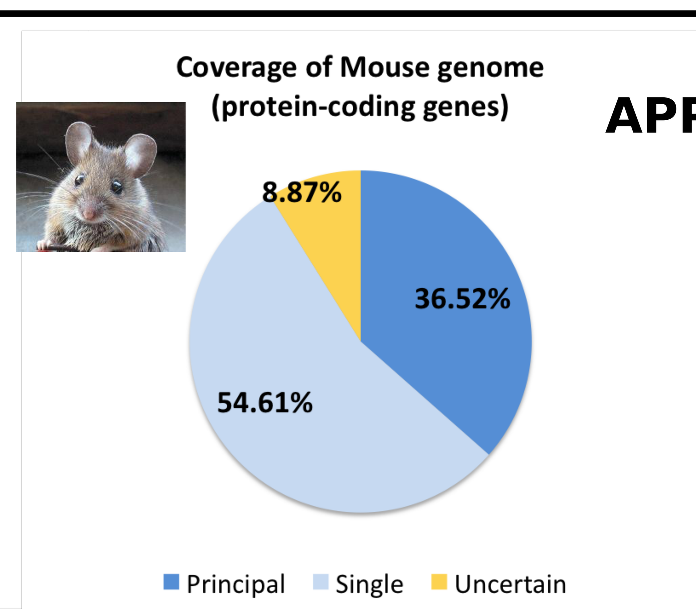
Non-neutral evolution of exons

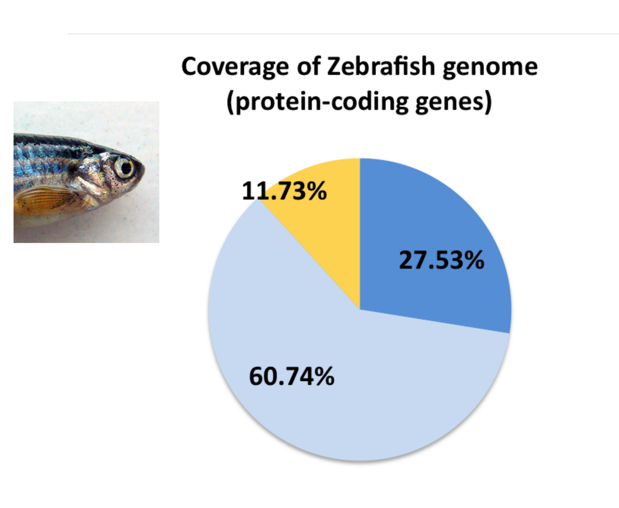
The method predicts exons with **non-neutral evolutionary rates** using **SLR**[6]. The principal isoform is not likely to contain exons that are evolving abnormally quickly or under **unusual selective pressures**.

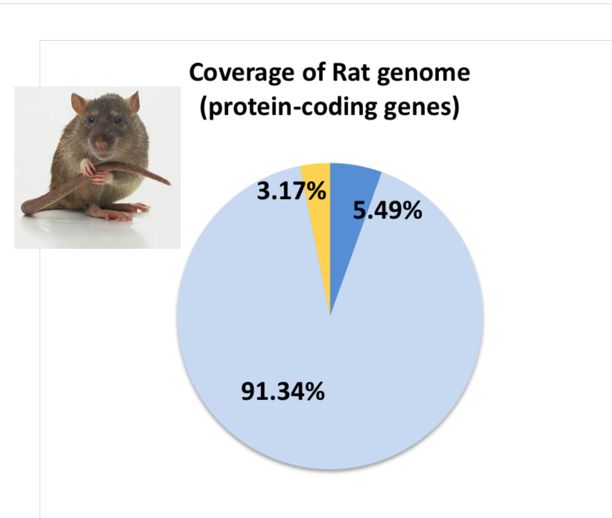
Transcripts are aligned against 46 vertabrate species using **PRANK**, **KALIGN**, and **MAF** alignments from **UCSC**.

RESULTS

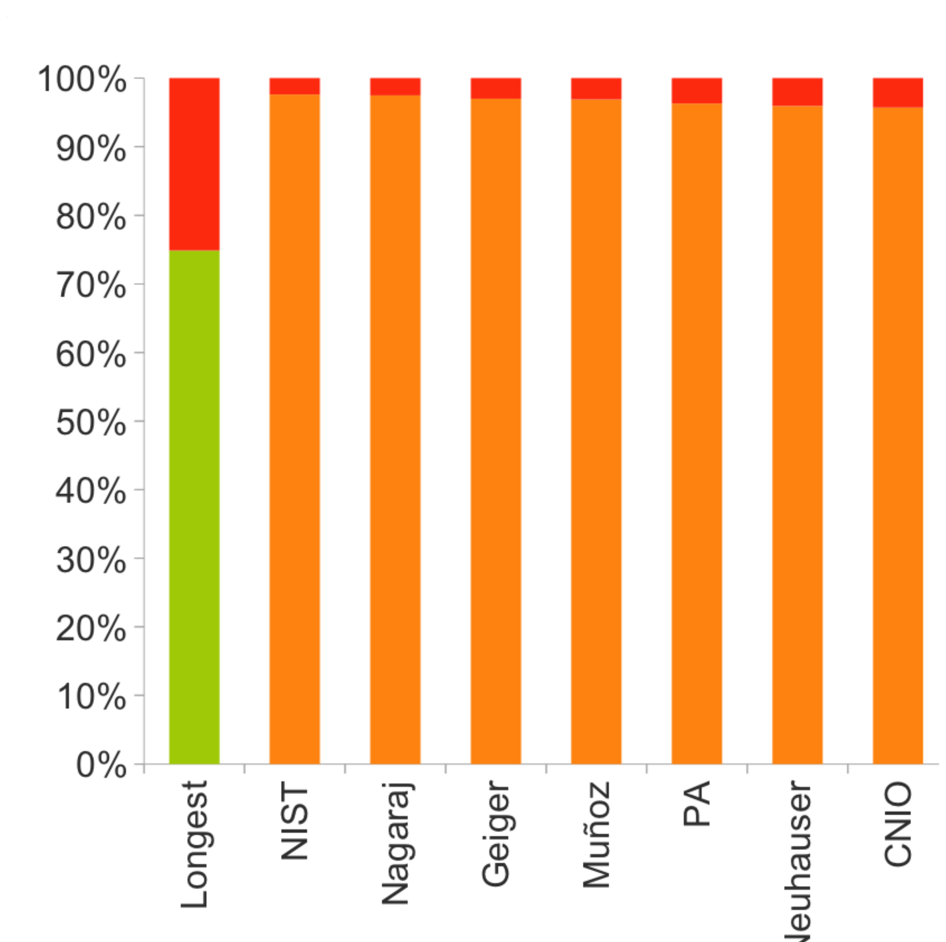








APPRIS principal isoforms - proteomics primary isoform agreement is **close to 100%** in all seven analysis



APPRIS vs. CCDS[8]

We analysed the results of APPRIS methods over the subset of genes where the CCDS project annotates a single variante.

For of the methods, firestar, Matador3D, CORSAIR, and SPADE almost never tag the CCDS annotated variant as alternative. While INERTIA, THUMP, and CRASH are slightly less reliable with a 93-98% agreement.

firestar	Matador3D	CORSAIR	SPADE	INERTIA	THUMP	CRASH
1.05%	0.79%	1.71%	0.88%	6.63%	2.18%	1.07%

For the genes that have multiple isoforms and a single CCDS variant, APPRIS is in agreement with the CCDS variant 96% of the time.

We also compared the selection of a principal isoform by APPRIS and the selection by choosing the longest isoform. This compares to an agreement of just 79.2%

Single CCDS	Longest Seq.
96%	79.2%

REFERENCES

1. *Rodriguez, J. et al. (2013) Nucleic Acids Res., 41, D110-7.*
2. *Wang ET, et al. (2008) Nature. Nov 27;456(7221):470-6.*
3. *Tress,M.L. et al. (2007) Proc Natl Acad Sci USA, 104:5495-5500.*
4. *Harrow, J. et al. (2012) Genome Res. 22:1775-1789.*
5. *Lopez,G. et al. (2007) Nucleic Acids Res., 35, W573-W577.*
6. *Massingham, T. et al (2005) Genetics 169: 1853-1762.*
7. *Finn et al. (2008) Nucleic Acids Res., 36, D281-D288.*
8. *Pruitt, KD et al. (2009) Genome Res. 19(7):1316-23.*

We would like to thank:

- Adam Frankish, Felix Kokocinski, Tim Hubbard and Jennifer Harrow, The HAVANA group, Sanger Centre, Cambridge.
- Tim Massingham, EBI, Cambridge.
- Mike Lin, MIT, Boston.
- Eduardo Andrés, Ángel Carro, CNIO, Spain.

If you want this amazing POSTER :-) here you are the QR code

CONTACT: jmrodriguez@cnio.es