APPRIS

SELECTION OF PRINCIPAL SPLICE ISOFORMS

http://appris.bioinfo.cnio.es

Jose Manuel Rodríguez[1], Paolo Maietta[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.

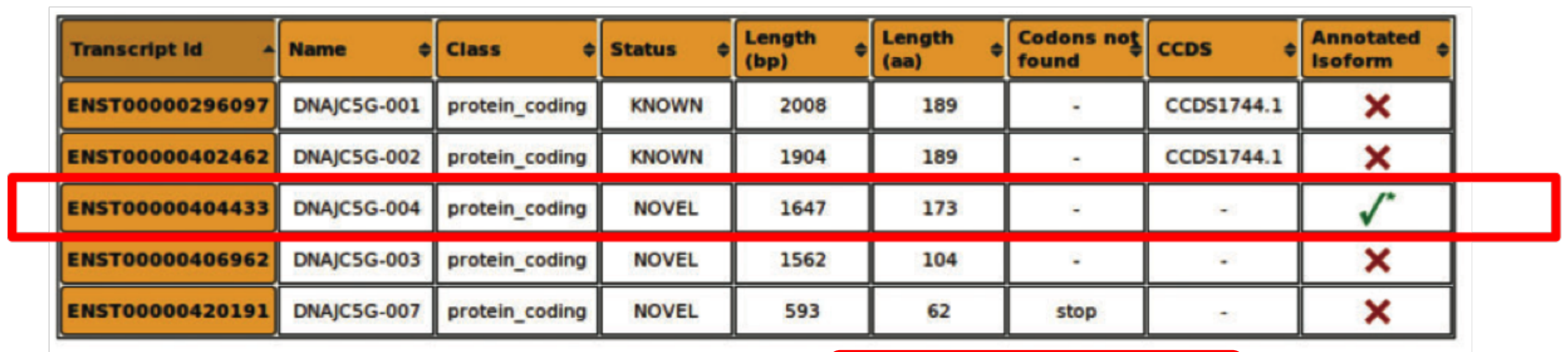
ABSTRACT

The cellular role of alternative protein isoforms is a topic of growing interest, both in normal cells and in cancer research (see the CLL example 1,2). We have developed the APPRIS database (3) to annotate splice variants with information relating to **protein structure**, **function** and **cross-species conservation**. APPRIS currently has annotations for 20,738 human genes and 95,309 transcripts.

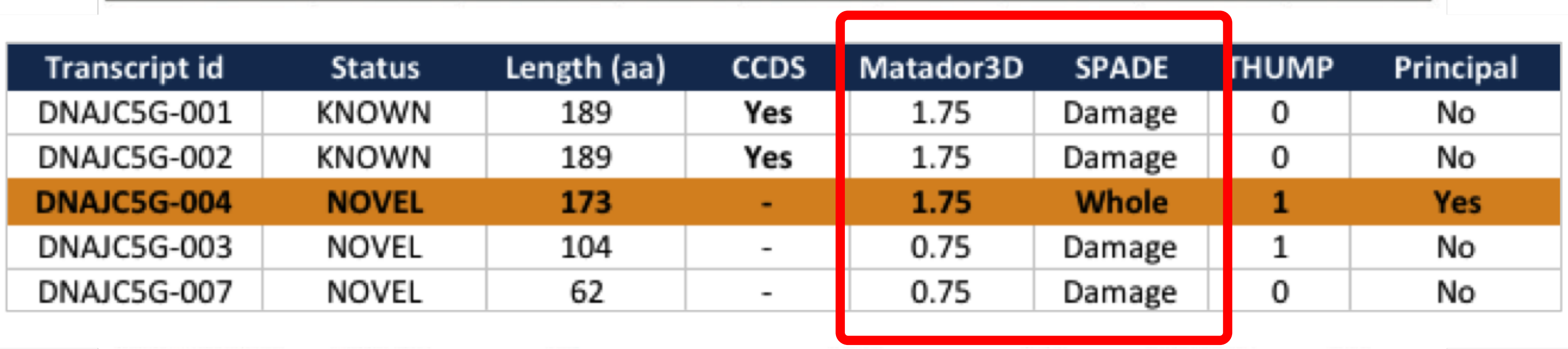
APPRIS makes use of the conservation of protein features to identify a **single dominant** (4) **isoform for each gene**. These principal isoforms are confirmed by orthogonal theoretical analyses (5) and by the results of multiple large-scale mass spectrometry experiments and databases (6,7).

The APPRIS database is stable and is implemented as part of the **GENCODE/Ensembl human genome annotation** (8), while the set of constitutive exons provided by APPRIS is also in use in our in-house cancer genome analysis pipeline (1,9).

HOW PIPELINE WORKS with an EXAMPLE



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	-	✗



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

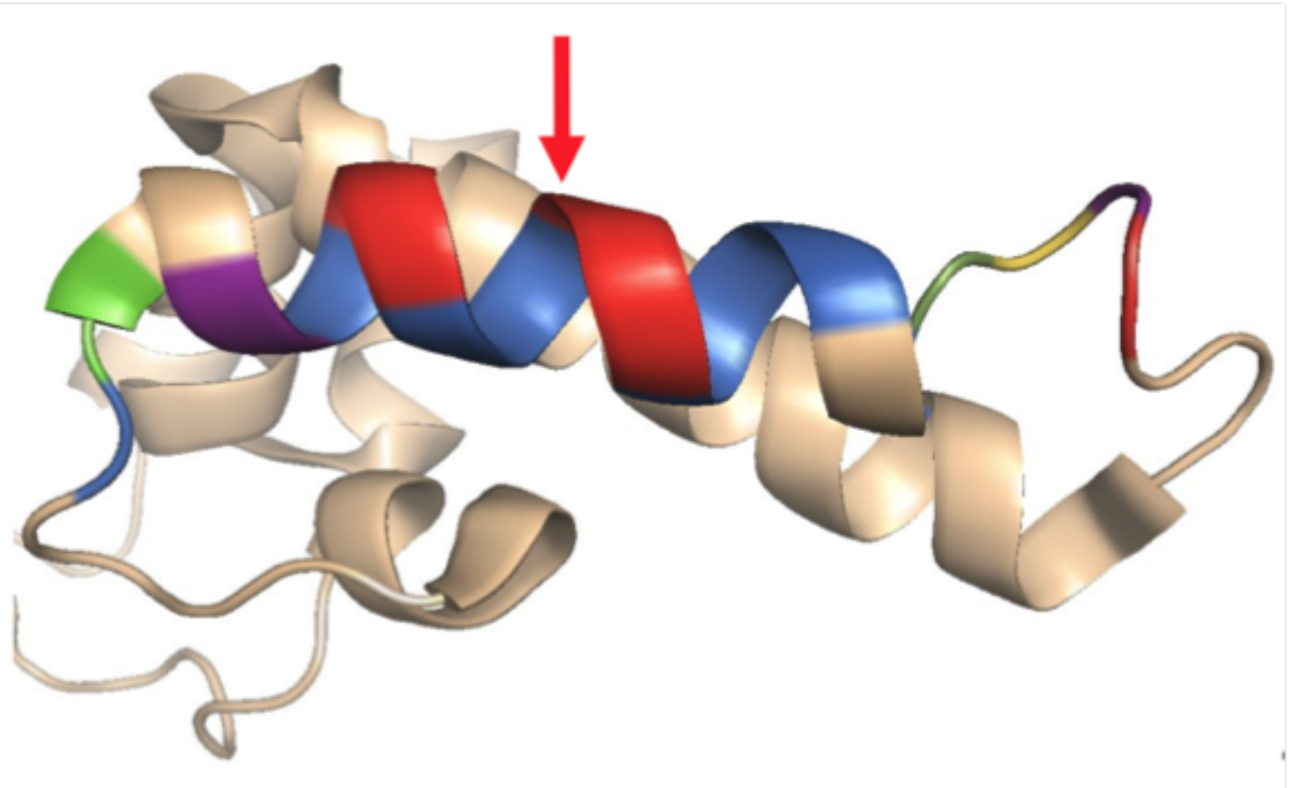
Snapshot of the APPRIS web page, showing the five protein-coding transcripts annotated by GENCODE/Ensembl and the selection of the principal isoform by APPRIS (green tick).

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 isoforms, which has an inserted exon that would compromise Pfam domains and 3D structure**.

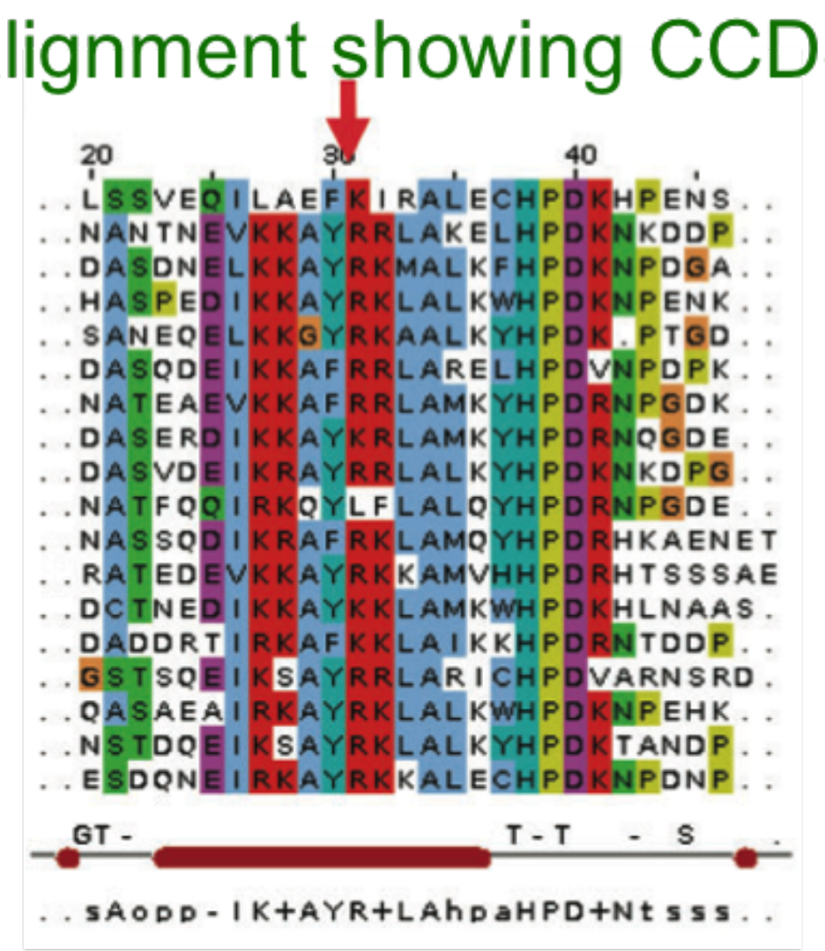
Homologue showing CCDS insertion



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 present 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 need to be inserted into a critical region of the functional domain of DNAJC5G.

METHODS

firestar(10)

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.

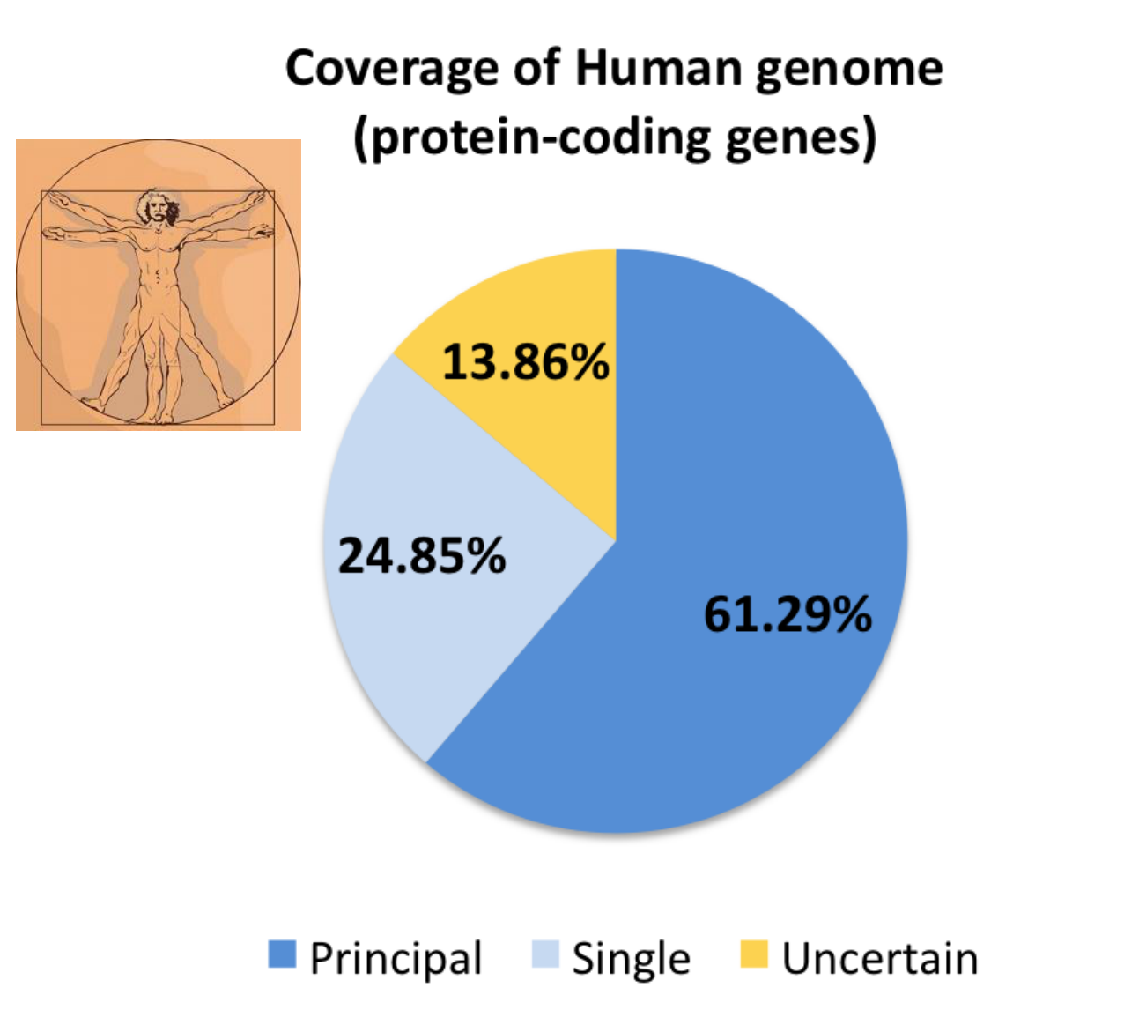
SPADE

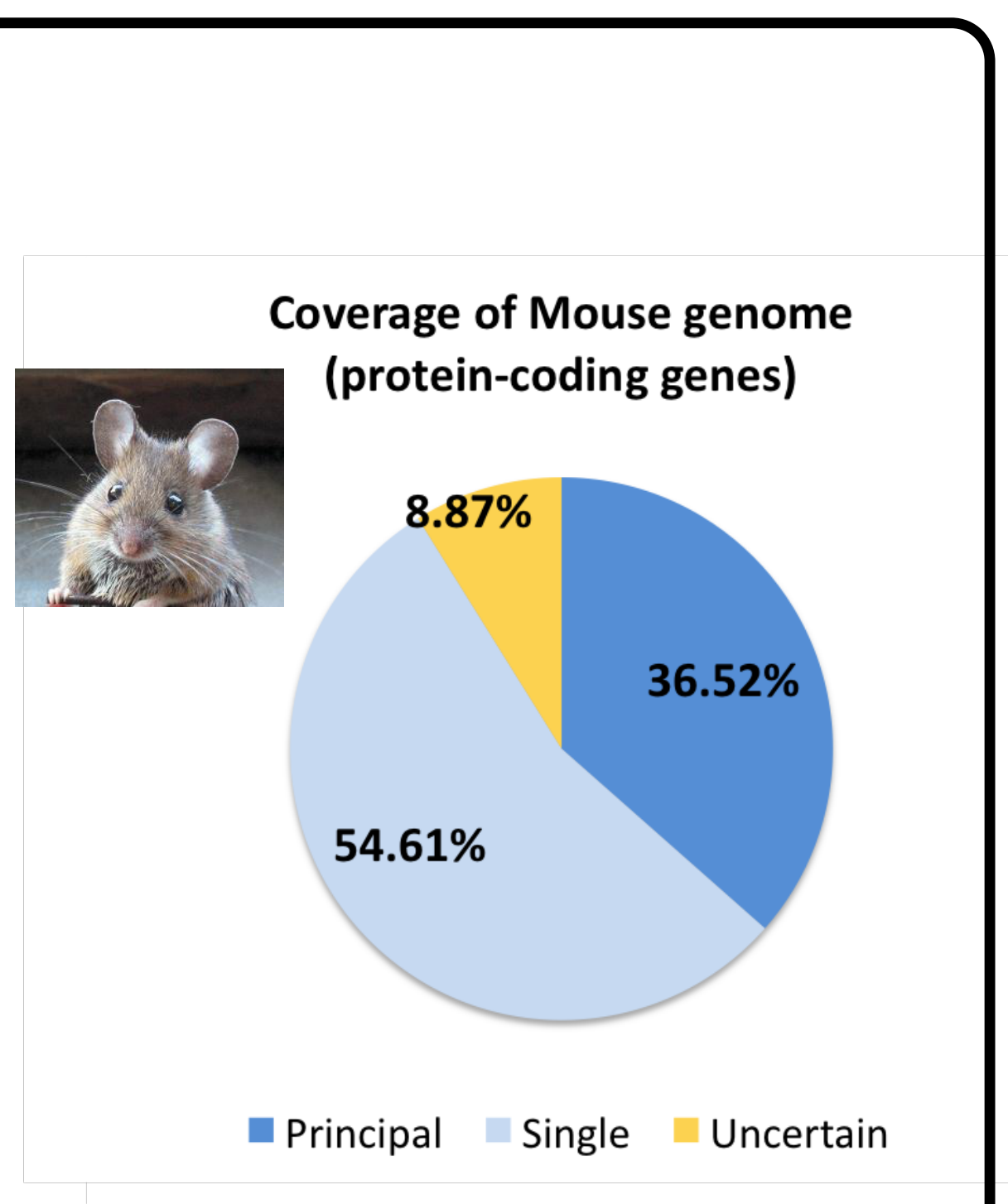
Conservation of protein functional domains

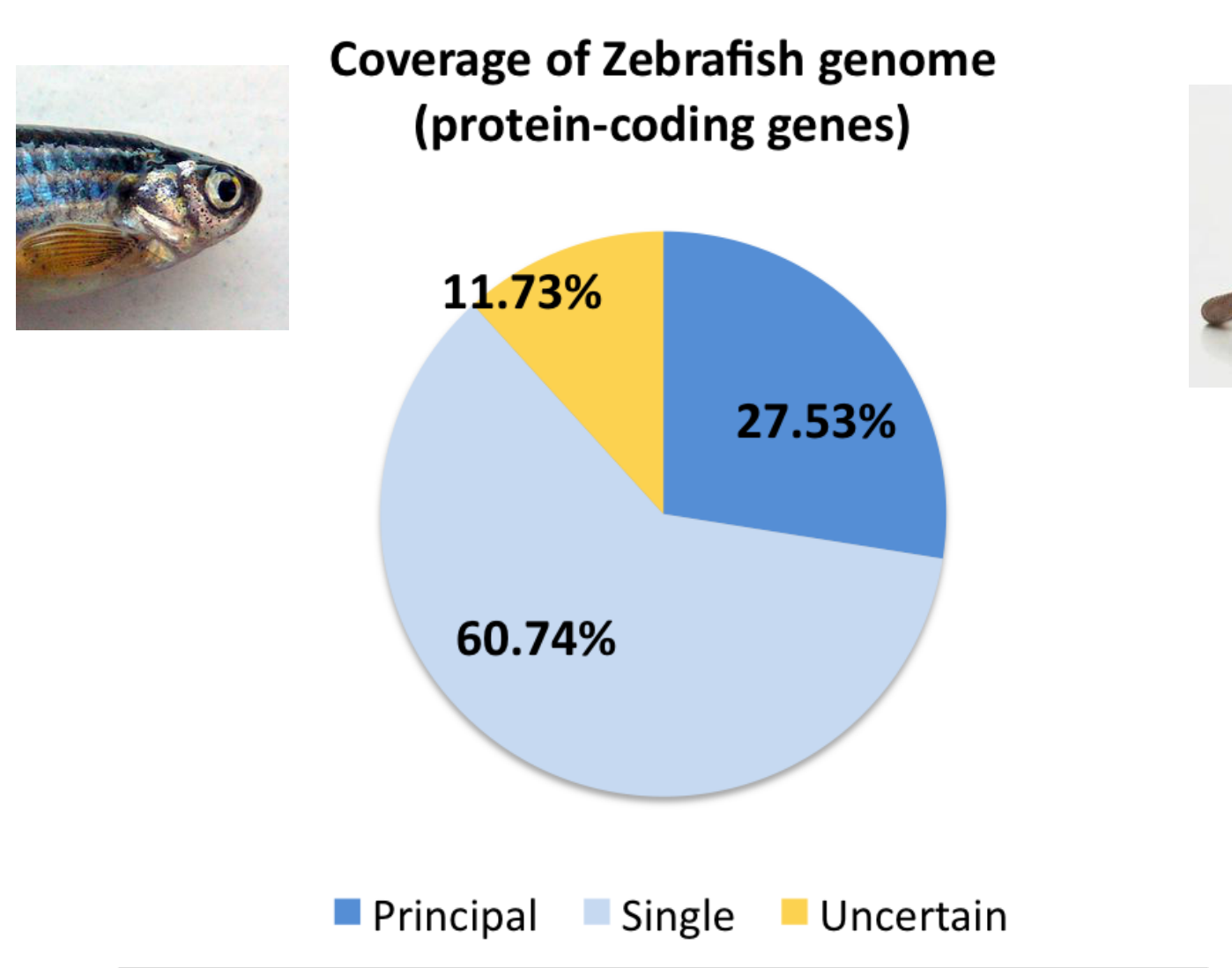
Identifying the functional domains present in a variant can **provide insights into the function**.

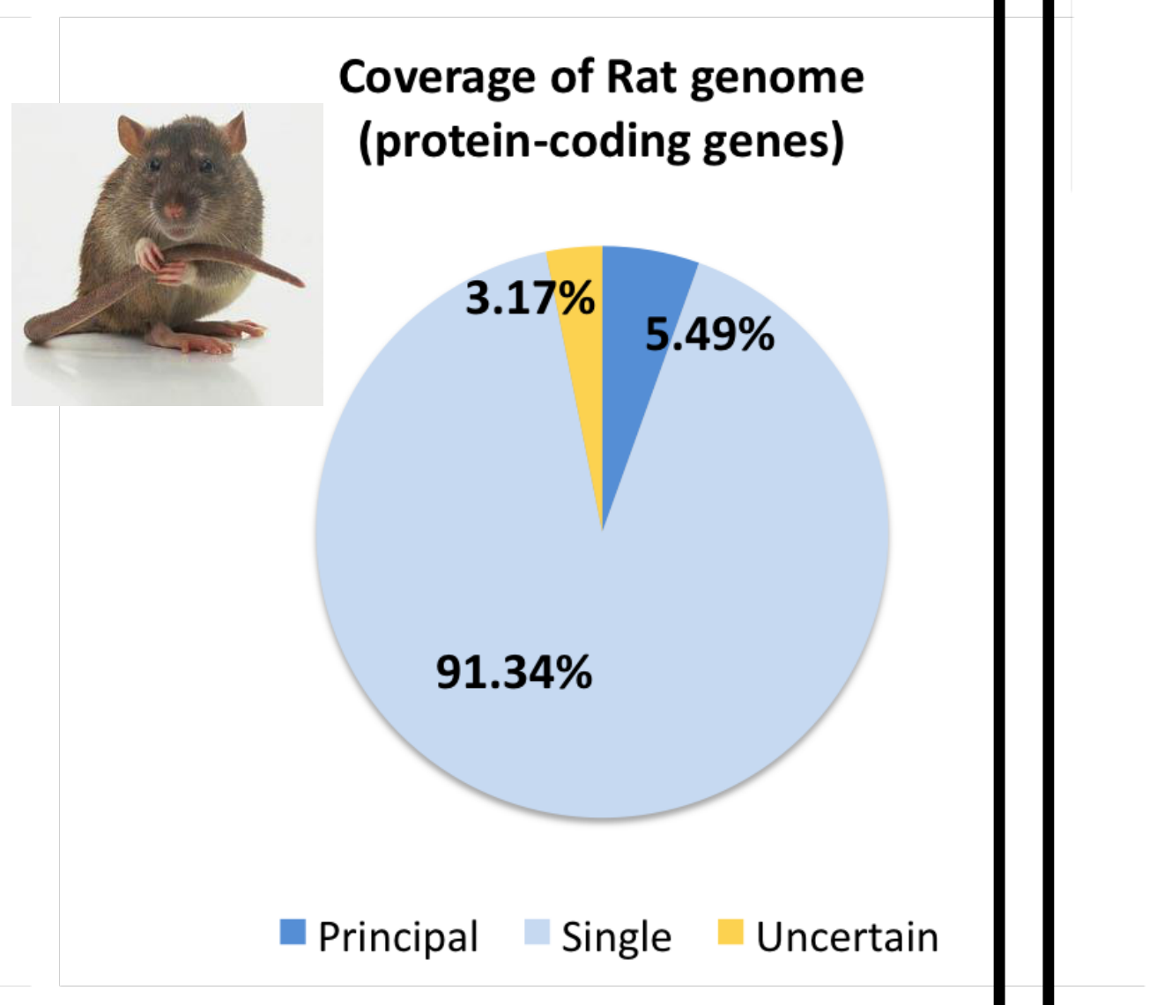
The presence of protein domains is analysed with **Pfamscan** (11).

GENOME COVERAGE

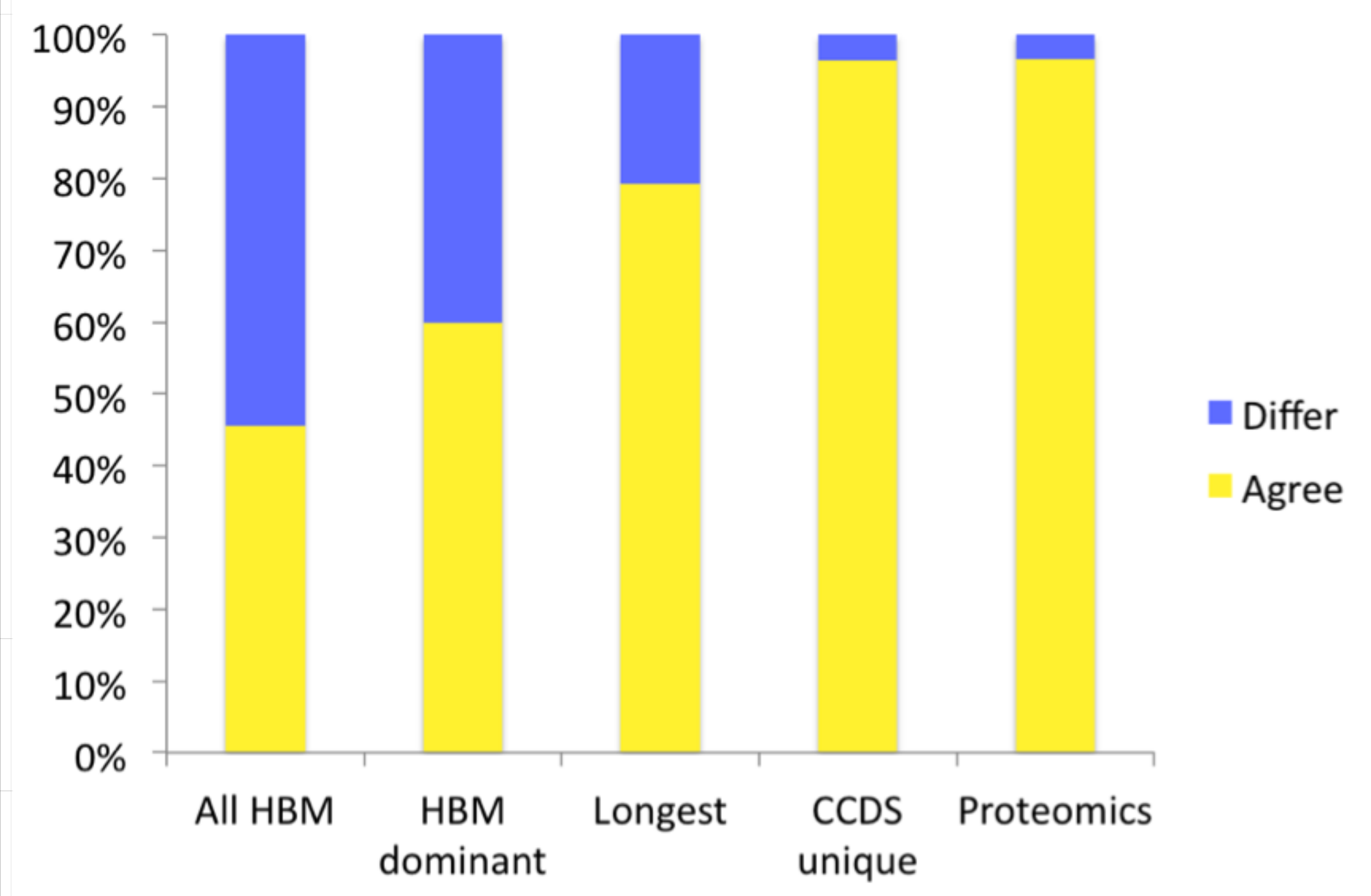








VALIDATION



Here we compared APPRIS principal variants with (from right to left) the main isoform identified in the proteomics experiments, with the CCDS (5) variants in those genes that have a unique CCDS variant and with the longest annotated isoform.

We also show the comparison with the dominant transcripts carried out using RNAseq data by Gonzalez-Porta et al. (4).

APPRIS principal isoforms, the main isoforms from proteomics experiments and the unique CCDS isoforms have an exceptionally **high level of agreement**.

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, since researchers need to be able to work with the isoform that is most likely to have **major functional activity**.

Likewise 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 potential for the use of APPRIS data in research is huge.

In the context of the **ICGC PAN CANCER** effort the information provided by APPRIS can be important for the interpretation of point **mutations in correct splice variants**, the **identification of principal isoforms** and the annotation of splice variants and **constitutive exons**.

REFERENCES

1. Quesada et al. Nature Genetics 2011, 44:47-52.
2. Ferreira et al. Genome Res. 2014 24:212-26.
3. Rodriguez, J. et al. (2013) Nucleic Acids Res., 41, D110-7.
4. Gonzalez-Porta et al. Genome Biol. 2013,14:R70.
5. Pruitt, KD et al. (2009) Genome Res. 19(7):1316-23.
6. Ezkurdia et al. Mol Biol Evol. 2012. 29:2265-83.
7. Farrah, T. et al. J. Proteome Res. 2013. 12(1):162-171.
8. Harrow, J. et al. (2012) Genome Res. 22:1775-1789.
9. Balbas-Martinez. Nature Genetics 2013, 45:1464-9.
10. Lopez,G. et al. (2007) Nucleic Acids Res., 35, W573-W577.
11. Finn et al. (2008) Nucleic Acids Res., 36, D281-D288.
12. Massingham, T. et al (2005) Genetics 169: 1853-1762.

We would like to thank:

- Gonzalo Lopez, Iakes Ezkurdia, Jan-Jaap Wesselink.
- 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