APPRIS

http://appris.bioinfo.cnio.es

Jose Manuel Rodríguez*, Paolo Maietta, Iakes Ezkurdia, Gonzalo López, Jan-Jaap Wesselink, Alessandro Pietrelli, 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. Recent studies have estimated that almost **100% of multi-exon human genes**[1]produce differently spliced mRNAs. It is important to designate one of the isoforms as the **"principal"** functional isoform in order to predict the changes in function, structure or localisation brought about by **Alternative Splicing**[2].

We have developed a pipeline to annotate principal functional variants works by a process of elimination. **APPRIS** deploys a range of computational methods including the ***conservation* of exonic structure**, the ***conservation* of protein structure** and **function** and a measure of **non-neutral evolution of exons**. The server is being used in the context of part of the scale up of the ***ENCODE***[2] project to annotate 100% of the human genome (**20,700 protein-coding genes** and **84,408 distinct alternative transcripts**).

METHODS

APPRIS combines **protein structural** and **functional information**, and **cross species conservation** to make automatic annotations of main functional isoforms.

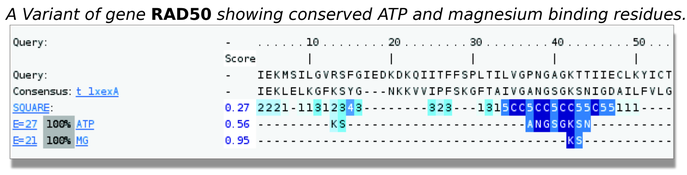
firestar[4]

Functionally Important Residues

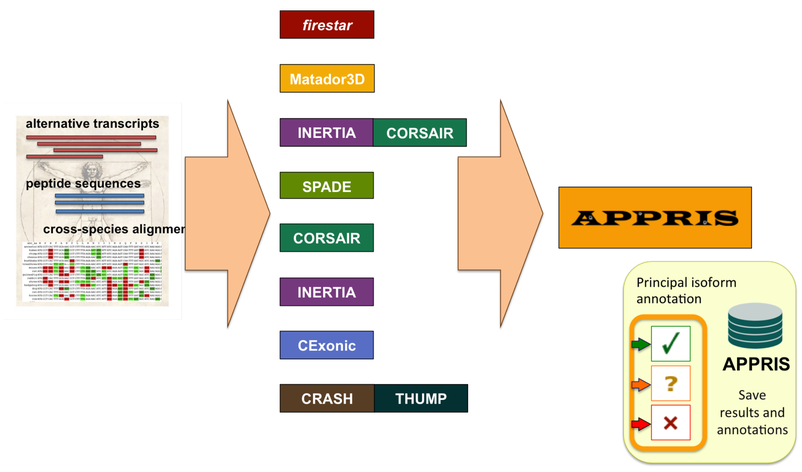
Functional residues are highly conserved, even across large evolutionary distances. Since these residues are **unlikely to have arisen by chance** we can use this to help determine the principal isoform.

Variants that have "***lost***" conserved functional residues are **not likely to be** as potential principal isoforms.

A Variant of gene **RAD50** *showing conserved ATP and magnesium binding residues.*



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



INERTIA

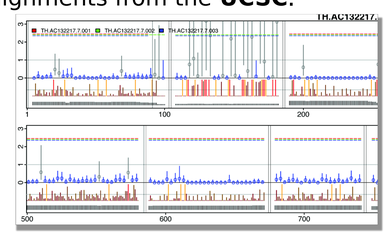
Non-neutral evolution of exons

The method predicts exons with **non-neutral evolutionary rates** using **SLR**[5]. 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 the **UCSC**.

The SLR output for three variants of the TH locus (tracks shown in red, green and blue).

The colours show different modes of selection. Abnormally fast sites are coloured orange or red.

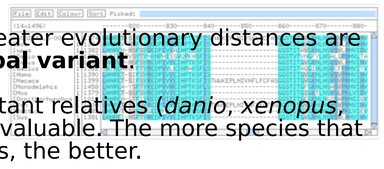


CORSAIR

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.



Matador3D

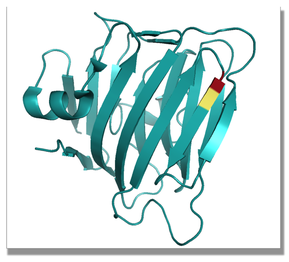
Protein structural information

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.

Those variants that **introduce gaps** are not likely to be as potential principal isoforms.

The sequence of variant 001 of neurexin 2 mapped onto the structure of a neurexin 1 domain. Variant 001 of neurexin would have a large insertion between the red and yellow residues.



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**).

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**[6].

CExonic

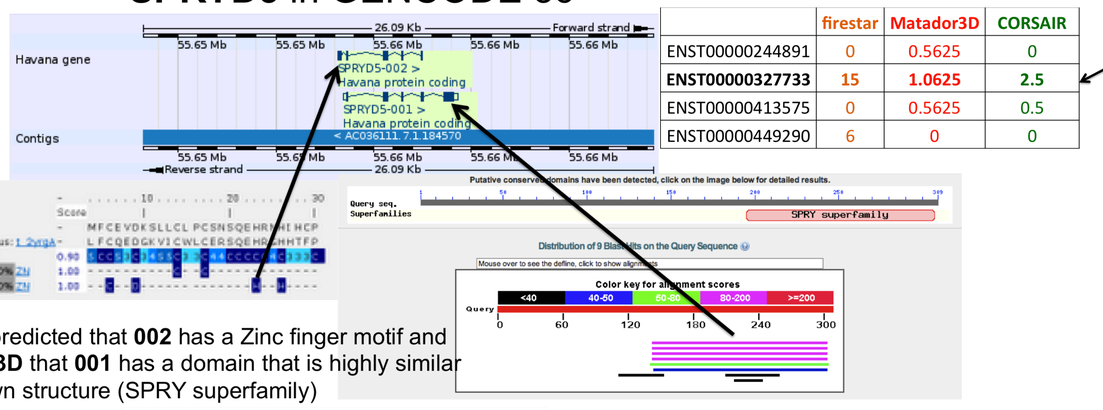
Conservation of exonic structure

CExonic evaluates the **conservation of exonic structure** between orthologous splice isoforms of two species.

http://cexonic.bioinfo.cnio.es/

RESULTS & SYSTEM CONTEXT

SPRYD5 in GENCODE 3c



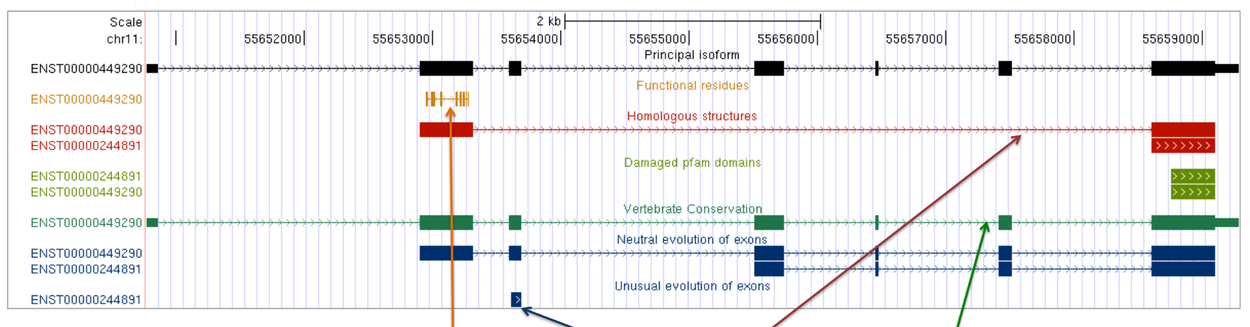
	firestar	Matador3D	CORSAIR
ENST00000244891	0	0.5625	0
ENST00000327733	**15**	**1.0625**	**2.5**
ENST00000413575	0	0.5625	0.5
ENST00000449290	6	0	0

New isoform in Gencode 3c added after APPRIS annotation. It was renamed in **Gencode 7** (**Ensembl 62**) as **ENST00000449290**

firestar predicted that **002** has a Zinc finger motif and **Matador3D** that **001** has a domain that is highly similar to a known structure (SPRY superfamily)

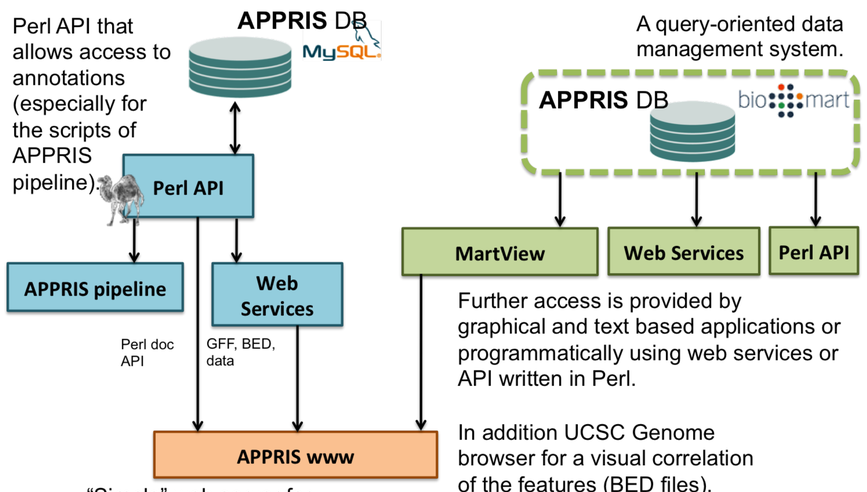
SPRYD5 in GENCODE 7

Transcript Id	Name	Class	Status	Length (bp)	Length (aa)	Codons not found	CCDS	Annotated Isoform
ENST00000244891	SPRYD5-001	protein_coding	PUTATIVE	1329	309	-	-	✗
ENST00000449290	SPRYD5-004	protein_coding	KNOWN	1629	452	-	-	✓**



Transcript Id	No. Functional Residues	Score of Homologous Structure	Neutral Evolution of Exons	Conservation score (vertebrates)
ENST00000244891	0	0.5625	X	0
ENST00000449290	16	0.8925	√	2.5

Perl API that allows access to annotations (especially for the scripts of APPRIS pipeline).



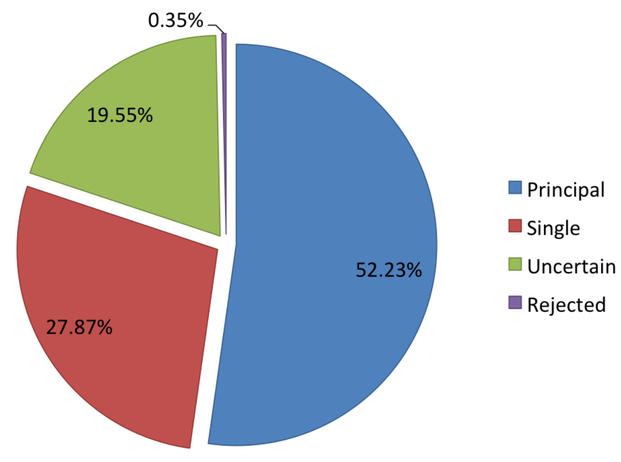
Further access is provided by graphical and text based applications or programmatically using web services or API written in Perl.

"Simple" web server for determining principal splice isoforms.

In addition UCSC Genome browser for a visual correlation of the features (BED files).

UCSC Genome Bioinformatics

Coverage for 20,700 genes in GENCODE 7



REFERENCES

1. *Wang ET, et al. (2008) Nature. Nov 27;456(7221):470-6.*
2. *Tress,M.L. et al. (2007) Proc Natl Acad Sci USA, 104:5495-5500.*
3. *The ENCODE Project Consortium. (2007) Nature, 447, 799-816.*
4. *Lopez,G. et al. (2007) Nucleic Acids Res., 35, W573-W577.*
5. *Massingham, T. et al (2005) Genetics 169: 1853-1762.*
6. *Finn et al. (2008) Nucleic Acids Res., 36, D281-D288.*

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.

POSTER: http://appris.bioinfo.cnio.es/download/docs/APPRIS-Poster_JBI_2012.pdf

CONTACT: jmrodriguez@cnio.es