APPRIS 2017: Principal isoforms for multiple gene sets (Supplementary Data)

Jose Manuel Rodriguez[1,*], Juan Rodriguez-Rivas[2], Tomás Di Domenico[2] , Jesús Vázquez[3,4], Alfonso Valencia[5,6], and Michael L. Tress[2,*]

1. Spanish National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), Madrid, 28029, Spain.
2. Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, 28029, Spain.
3. Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28029 Madrid, Spain.
4. CIBER de Enfermedades Cardiovasculares (CIBERCV), 28029 Madrid, Spain.
5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona E-08010, Spain.
6. Life Sciences Department, Barcelona Supercomputing Centre (BSC-CNS), Barcelona E-08034, Spain.

* To whom correspondence should be addressed. Tel: (+34) 91 732 80 00; Fax: (+34) 91 224 69 76; Email: mtress@cnio.es.
Correspondence may also be addressed to Jose Manuel Rodriguez. Tel: (+34) 914531200 Fax: (+34) 914531265; Email: jmrodriguezc@cnic.es

Current addresses:
Jose Manuel Rodriguez. Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28029 Madrid, Spain.
Juan Rodriguez-Rivas. Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain.

DESCRIPTION OF APPRIS TAGS

Principal Isoform labels

APPRIS selects a single CDS variant for each gene as the 'PRINCIPAL' isoform based on the range of protein features. Principal isoforms are tagged with the numbers 1 to 5, with 1 being the most reliable. The definition of the flags are as follows:

- **PRINCIPAL:1**

  Transcript(s) expected to code for the main functional isoform based solely on the core modules in the APPRIS database. The APPRIS core modules map protein structural and functional information and cross-species conservation to the annotated variants.

- **PRINCIPAL:2**

  Where the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding genes), the database chooses two or more of the CDS variants as "candidates" to be the principal variant.

  If one of these candidates has a distinct CCDS (1) identifier it is selected as the principal variant for that gene. A CCDS identifier shows that there is consensus between RefSeq and GENCODE/Ensembl for that variant, guaranteeing that the variant has cDNA support.

- **PRINCIPAL:3**

  Where the APPRIS core modules are unable to choose a clear principal variant and there more than one of the variants have distinct CCDS identifiers, APPRIS selects the variant with lowest CCDS identifier as the principal variant. The lower the CCDS identifier, the earlier it was annotated.

  Consensus CDS annotated earlier are likely to have more cDNA evidence. Consecutive CCDS identifiers are not included in this flag, since they will have been annotated in the same release of CCDS. These are distinguished with the next flag.

  In addition, there is more than one variant with a distinct (but consecutive) CCDS identifiers, APPRIS choose the variant when all splice junctions are supported by at least one non-suspect mRNA. This information is reported by the method Transcript Support Level (TSL), which is a method to highlight the well-supported and poorly-supported transcript models for users. The method relies on the primary data that can support full-length transcript structure: mRNA and EST alignments supplied by UCSC and Ensembl.

- **PRINCIPAL:4**

  Where the APPRIS core modules are unable to choose a clear principal CDS and there is more than one variant with a distinct (but consecutive) CCDS identifiers and all the splice junctions are not well-supported, APPRIS selects the longest CCDS isoform as the principal variant.

- **PRINCIPAL:5**

  Where the APPRIS core modules are unable to choose a clear principal variant and none of the candidate variants are annotated by CCDS, APPRIS selects the longest of the candidate isoforms as the principal variant.

Alternative Isoform labels

For genes in which the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding genes) the "candidate" variants not chosen as principal are labeled in the following way:

- **ALTERNATIVE:1**

  Candidate transcript(s) models that are conserved in at least three tested non-primate species.

- **ALTERNATIVE:2**

  Candidate transcript(s) models that appear to be conserved in fewer than three tested non-primate species.

Non-candidate transcripts are not flagged and are considered as "MINOR" transcripts.

REFINEMENTS TO CORE METHODS

In order to confirm that APPRIS is selecting the main isoform we validated the APPRIS identified principal isoforms, and those of the individual methods, using the annotations from the CCDS project (1). The CCDS project aims to identify a core set of consistently annotated, high quality protein coding variants for the human genome. CCDS variants are annotated based on agreement between the three main public annotation resources GENCODE/Ensembl (2,3), NCBI and (4) UCSC (5). We can analyse the performance of the APPRIS Database over the subset of genes where the CCDS project annotates a single sequence-unique variant.

The reliability of each APPRIS module is continually revised using the GENCODE human reference gene set. Since the first published version of APPRIS there have been an increase in the agremeent between the unique CCDS variants and the resuts from the indiviudual methods (and the APPRIS principal isoforms). In figure 1 we show the comparison between the CCDS analysis at the time of the first publication (GENCODE7) and using GENCODE19 (the last stable gene set for the hg19 assembly).

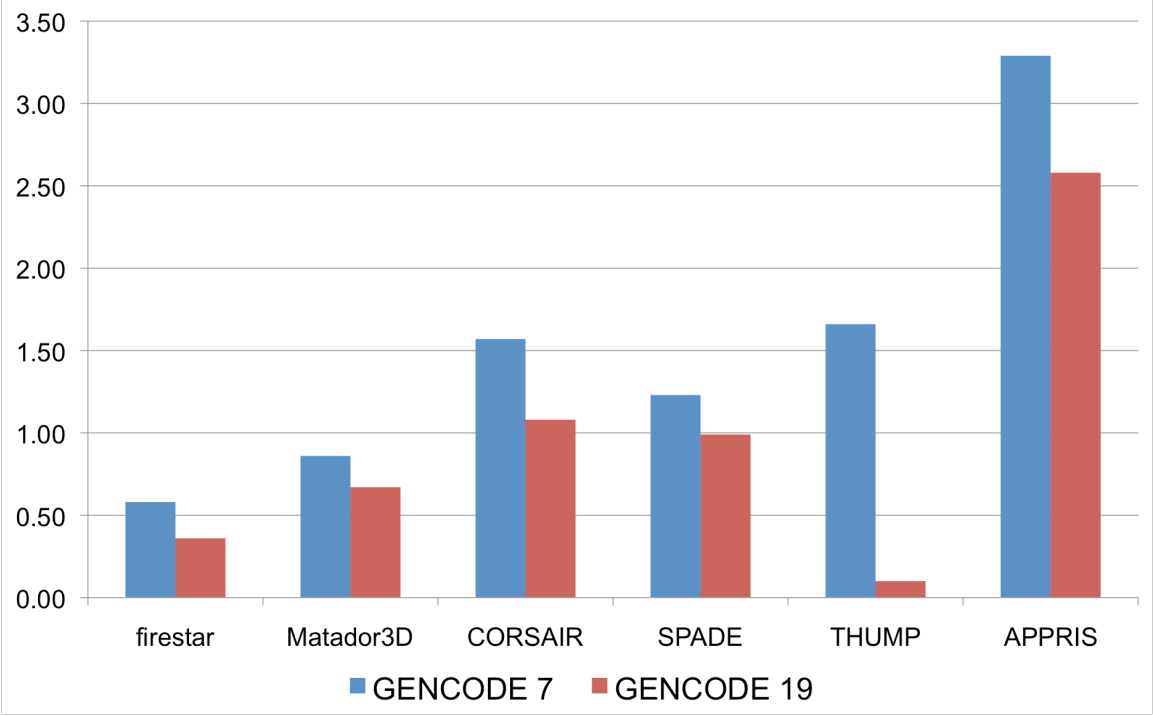

Figure 1. *Comparison between unique CCDS isoforms for each gene and those selected by the APPRIS methods from the human reference sets GENCODE 7 and GENCODE 19 (both in the hg19 assembly). The bars show the percentage of genes in which the unique CCDS isoform is not chosen as the principal isoform by APPRIS and the individual core methods. The APPRIS GENCODE 19 annotations agree with the unique CCDS variants more often.*

GENCODE/Ensembl, RefSeqGene, and UniProtKB INTERSECTION

We have created new gene sets (Intersections) based on the merge of Ensembl, RefSeqGene and UniProtKB isoforms for each gene. For the human genome we established a common gene set with the GENCODE (release 24), RefSeqGene (release 107), and UniProtKB (version 2016_06) reference sets (see Table 1). The initial cross-reference was generated with the data-mining tool, BioMart (6), and from there we re-annotated the cross-database relationships manually. For other species we generated the common gene sets with the BioMart tool solely.

Species	No. GENCODE/Ensembl Genes (release)	No. RefSeqGene Genes (release)	No. UniProtKB Genes (release)	No. Intersection Genes (release)
Human	20,250 (GENCODE_24/Ensembl_84)	20,066 (107)	21,608 (2016_06)	22,207 (a1)
Mouse	22,538 (GENCODE_M12/Ensembl_87)	21,879 (106)	22,852 (2016_10)	25,196 (a1)
Rat	22,041 (Ensembl_87)	22,640 (106)	21,837 (2016_10)	25,799 (a1)
Zebrafish	25,418 (Ensembl_87)	25,652 (105)	24,278 (2016_10)	28,279 (a1)
Pig	19,465 (Ensembl_88)	21,787 (105)	22,009 (2017_02)	14,746 (a1)
Chimp	18,282 (Ensembl_88)	21,139 (104)	18,878 (2017_02)	21,837 (a1)

Table 1. *Number of genes and releases of the reference genomes that were used to generate the common gene sets for the local species in the APPRIS Database. The common gene sets were established by cross-referencing GENCODE/Ensembl, RefSeqGene, and UniProtKB, using the data-mining tool, BioMart. For human, the cross-reference was also curated manually.*

APPRIS selects a single CDS variant for each gene as the “PRINCIPAL” isoform based on the annotated protein features. Principal isoforms are tagged with the numbers 1 to 5, with 1 being the most reliable. APPRIS determines a most reliable isoform for 75%-95% of protein-coding genes annotated depending on the gene set and the species (see Figure 2).

Where the APPRIS core modules are unable to choose a clear principal variant, the database chooses the variant based on the CCDS identifier and when all splice junctions are supported by at least one non-suspect mRNA (TSL). CCDS variants are annotated only for the human and mouse genomes, and TSL only for human. Where CCDS and TSL evidence is not decisive or available, APPRIS selects the longest of the candidate isoforms.

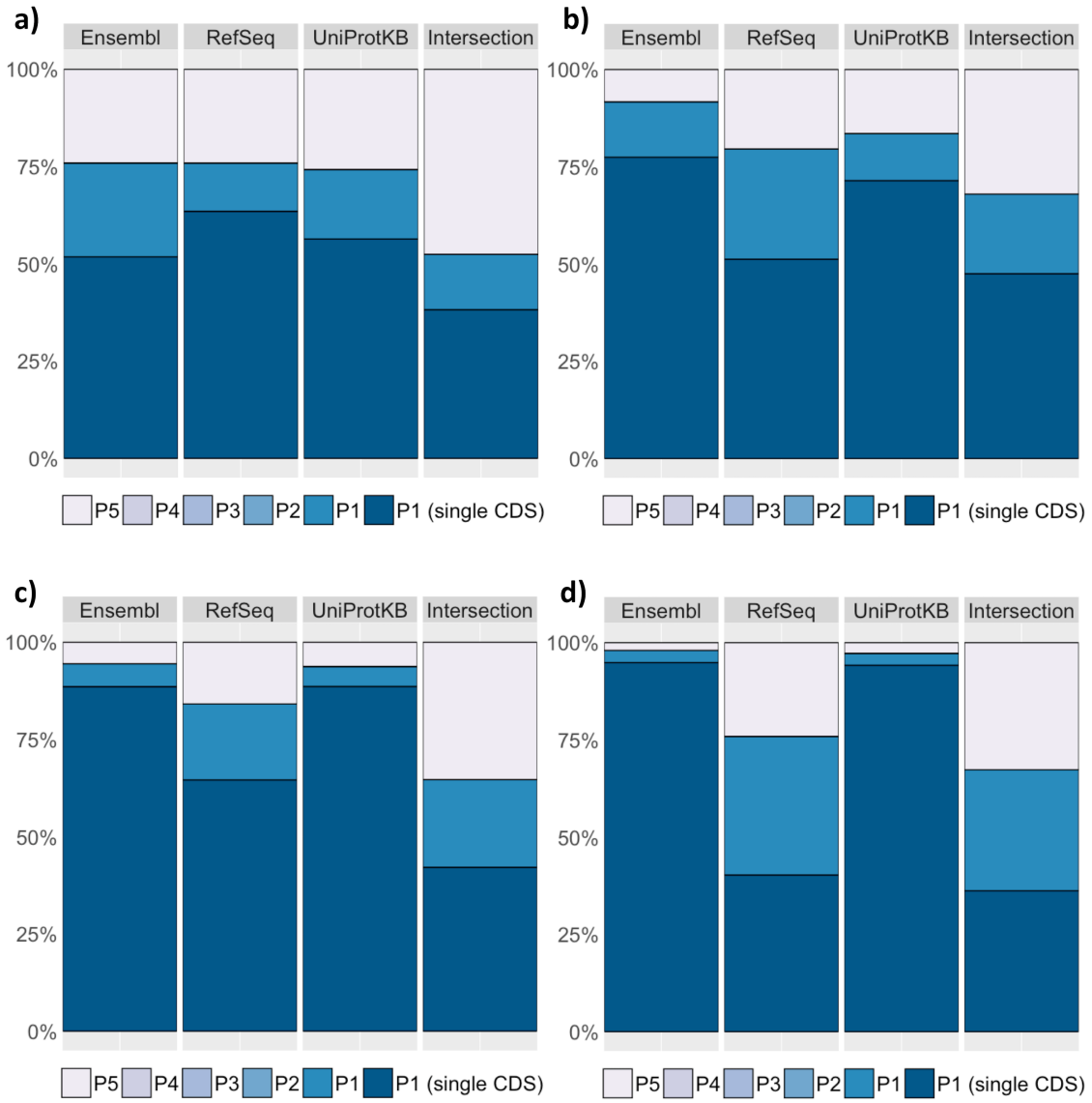

Figure 2. *Bar-plots with the percentage of genes identifiers with the final annotations of APPRIS for the vertebrate species zebra-fish (a), rat (b), pig (c), and chimpanzee (d). APPRIS identifies a principal isoform for each gene that are tagged with the numbers 1 to 5, with 1 being the most reliable, discarding the genes with a unique isoform. The APPRIS Database annotates the protein-coding genes in all public sets Ensembl, RefSeqGene, and UniProtKB. In addition, we established a common gene set (Intersection) with the Ensembl, RefSeqGene, and UniProtKB reference sets.*

The merged Intersection gene set allows us to identify principal isoforms missing in the individual gene sets. For example the principal isoform from the merged set for *GRIFIN* is annotated in GENCODE (ENST00000614228) and UniProtKB (A4D1Z8), but not in RefSeqGene (See Figure 2 in the main paper). The principal isoform has annotation evidence from cross-species alignments and the C-terminal extension in the GENCODE/Ensembl/UniProtKB principal isoform is established in mammals (see Figure 3).

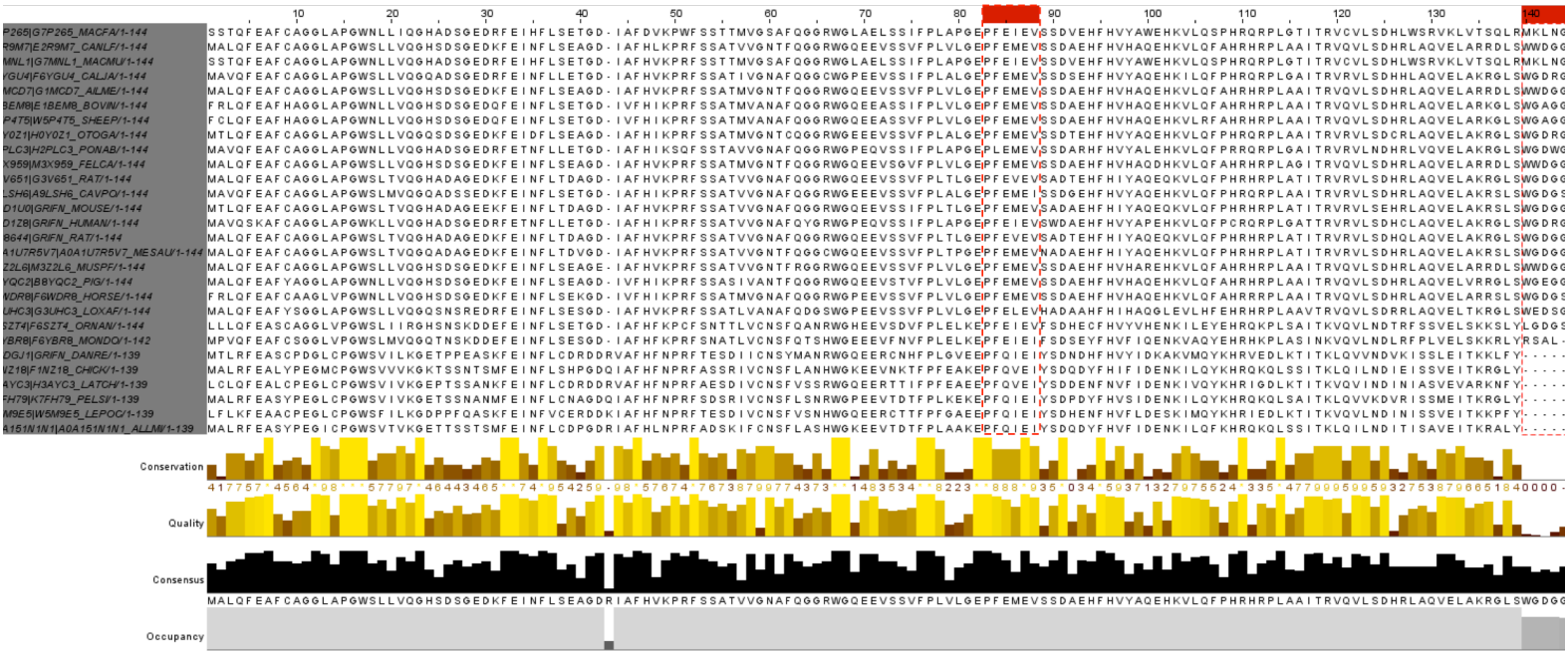

Figure 3. *Cross-Reference alignment of GRIFIN (ENST00000614228) variant. Clearly, we see an extension in the C-terminal established in mammals. In the region where the 8 extra residues would be to inserted for the alternative variants, we see evidences for cross-species conservation.*

REFERENCES

1. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**(7):1316-23, 10.1101/gr.080531.108, Epub 2009 Jun 4, Erratum in: *Genome Res.*, 2009 Aug;19(8):1506.

2. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M,Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**(9):1760-74, 10.1101/gr.135350.111.

3. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel JH, White S, Zadissa A, Flicek P, Searle SM. (2016) The Ensembl gene annotation system. *Database (Oxford)*, pii: baw093, 10.1093/database/baw093.

4. NCBI Resource Coordinators. (2017) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **45**(D1):D12-D17, 10.1093/nar/gkw1071.

5. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, Haeussler M, Heitner S, Hinrichs AS, Karolchik D, Lee BT, Lee CM, Nejad P, Raney BJ, Rosenbloom KR, Speir ML, Villarreal C, Vivian J, Zweig AS, Haussler D, Kuhn RM, Kent WJ. (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**(D1):D626-D634, 10.1093/nar/gkw1134.

6. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, Bardou P, Beck T, Blake A, Bonierbale M, Brookes AJ, Bucci G, Buetti I, Burge S, Cabau C, Carlson JW, Chelala C, Chrysostomou C, Cittaro D, Collin O, Cordova R, Cutts RJ, Dassi E, Di Genova A, Djari A, Esposito A, Estrella H, Eyras E, Fernandez-Banet J, Forbes S, Free RC, Fujisawa T, Gadaleta E, Garcia-Manteiga JM, Goodstein D, Gray K, Guerra-Assunção JA, Haggarty B, Han DJ, Han BW, Harris T, Harshbarger J, Hastings RK, Hayes RD, Hoede C,

Hu S, Hu ZL, Hutchins L, Kan Z, Kawaji H, Keliet A, Kerhornou A, Kim S, Kinsella R, Klopp C, Kong L, Lawson D, Lazarevic D, Lee JH, Letellier T, Li CY, Lio P, Liu CJ, Luo J, Maass A, Mariette J, Maurel T, Merella S, Mohamed AM, Moreews F, Nabihoudine I, Ndegwa N, Noirot C, Perez-Llamas C, Primig M, Quattrone A, Quesneville H, Rambaldi D, Reecy J, Riba M, Rosanoff S, Saddiq AA, Salas E, Sallou O, Shepherd R, Simon R, Sperling L, Spooner W, Staines DM, Steinbach D, Stone K, Stupka E, Teague JW, Dayem Ullah AZ, Wang J, Ware D, Wong-Erasmus M, Youens-Clark K, Zadissa A, Zhang SJ, Kasprzyk A. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**(W1):W589-98, 10.1093/nar/gkv350.