{APPRIS}

IMPROVING THE SELECTION OF PRINCIPAL ISOFORMS

http://appris-tools.org

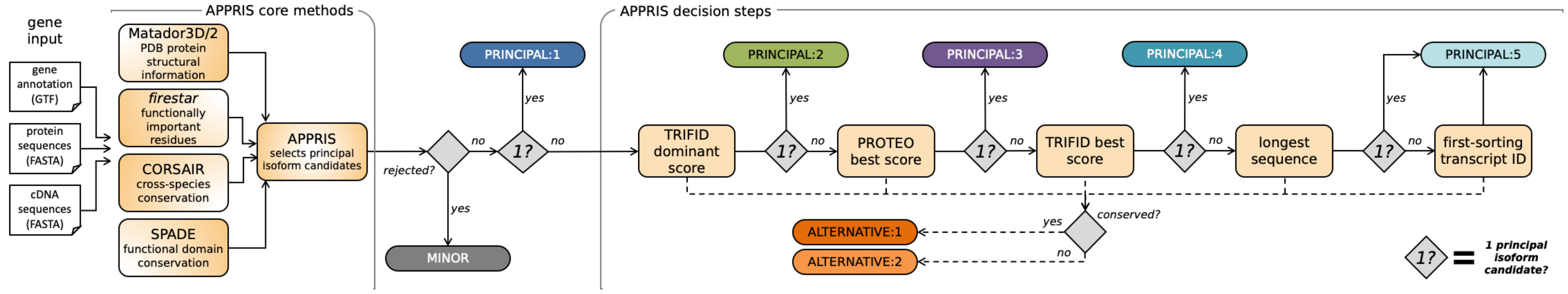
Thomas Walsh[1], Fernando Pozo Ocampo[1], José Manuel Rodriguez[2], and Michael Tress[1].
[1]Spanish National Cancer Research Centre (CNIO) and [2]Centro Nacional de Investigaciones Cardiovasculares (CNIC). Madrid, Spain.

ABSTRACT

Evidence suggests that a single main splice isoform reflects the biological reality of most protein-coding genes [1,2]. APPRIS [3,4] selects a single representative protein isoform for each coding gene based on cross-species conservation and the preservation of protein structural and functional features. The APPRIS principal isoform agrees with experimental protein evidence and expert manual curators over 99.5% of measurable coding genes [5] and the exons that produce APPRIS principal isoforms are under selective pressure, unlike the vast majority of alternative isoforms [3].

Improvements to the annotation system mean that APPRIS core methods are able to predict a principal for more than 75% of human genes, and the new TRIFID algorithm produces a score for the likely biological relevance of both principal and alternative isoforms.

METHODS



APPRIS core methods

The APPRIS score is a calculated from the weighted scores of 4 component methods.

- **Matador3D** scores each variant on its alignment to PBD [6] protein structures.
- **Firestar** [7] uses fireDB [8] to predict functionally important amino-acid residues. See also: http://firedb.bioinfo.cnio.es/index.php
- **CORSAIR** uses BLAST [9] to score variants on their conservation over evolutionary time.
- **SPADE** uses Pfamscan [10] to identify the set of functional domains in each variant.

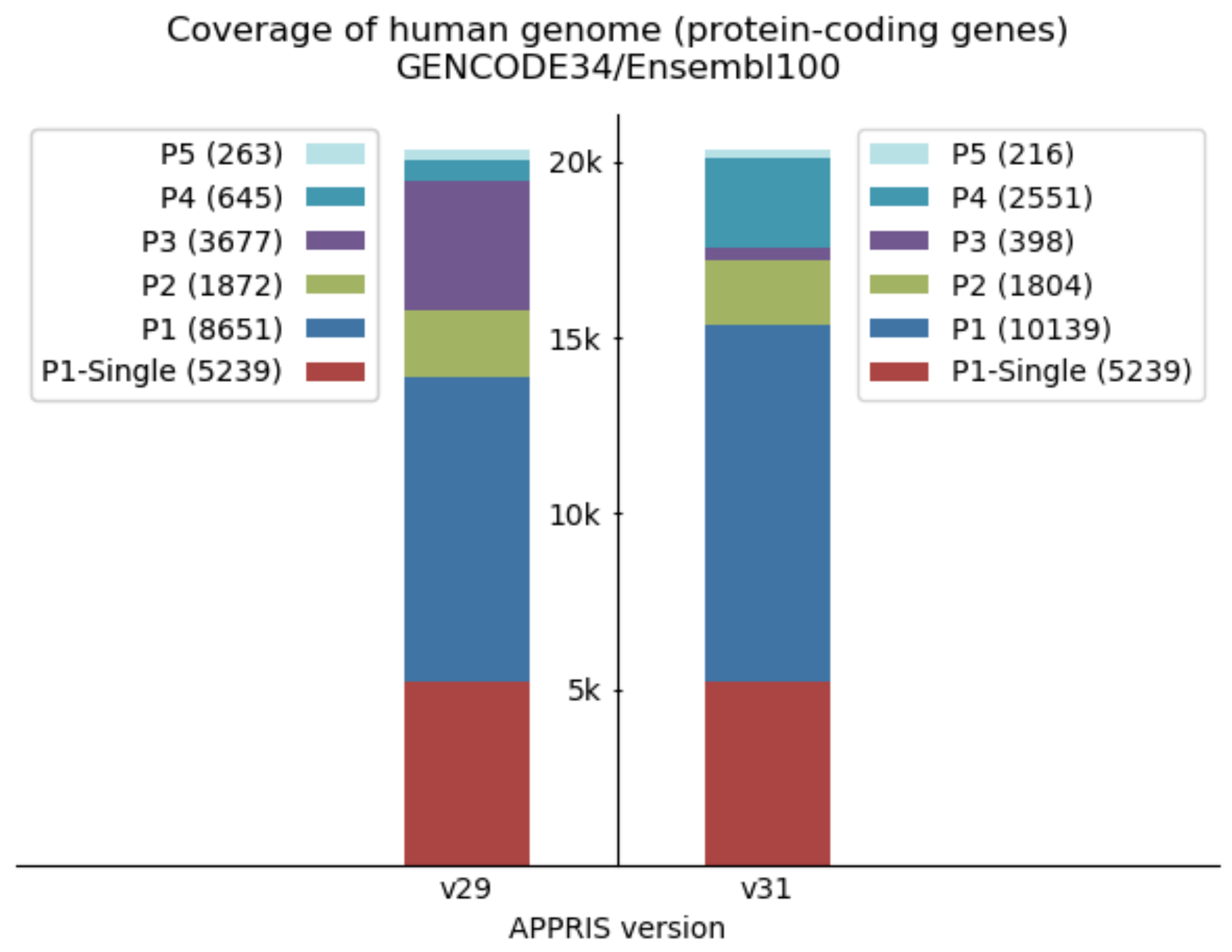
APPRIS decision steps

If APPRIS core methods cannot select a principal isoform, principal 'candidates' pass through a series of filters to select a single principal variant. There are currently two versions of this filtering process:

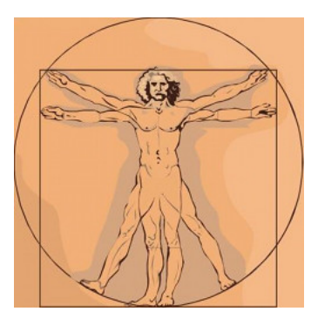
- **TRIFID transcript selection:** as outlined above, this incorporates predictions of functional importance by the TRIFID package — a machine learning classifier of isoform functionality. For more information about TRIFID, consult its GitLab repository at https://gitlab.com/bu_cnio/trifid — or visit the ISMB 2020 poster 'TRIFID: determining functional isoforms' by Fernando Pozo Ocampo.
- **classic transcript selection [4]:** used if TRIFID predictions are not available for a given species or dataset.

As TRIFID predictions become available for more datasets, the TRIFID transcript selection process will be used more frequently, with the ultimate aim of incorporating it as the default decision process.

GENOME COVERAGE



Organisms housed in APPRIS database


human

mouse

zebrafish

rat

pig

chimp

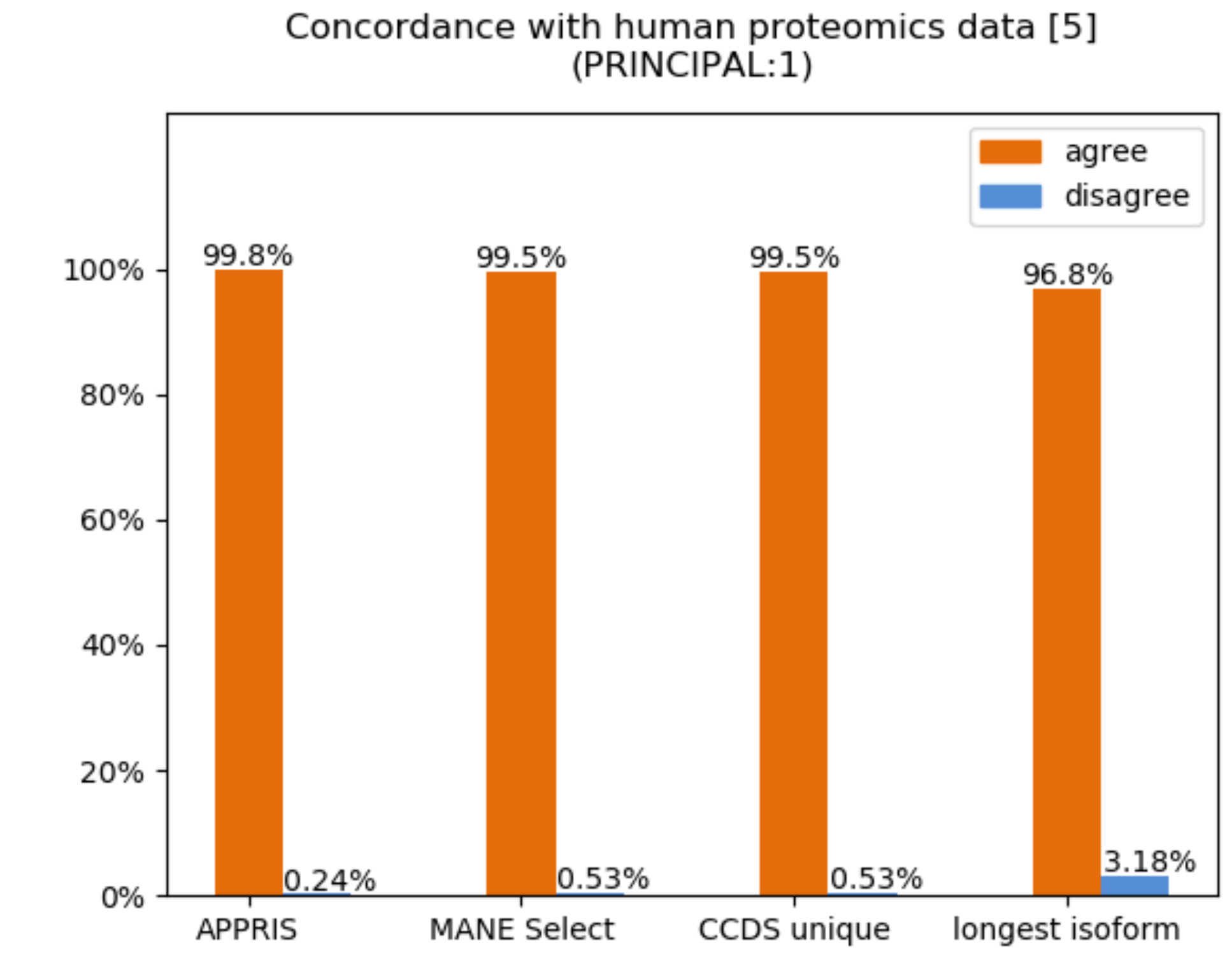
chicken

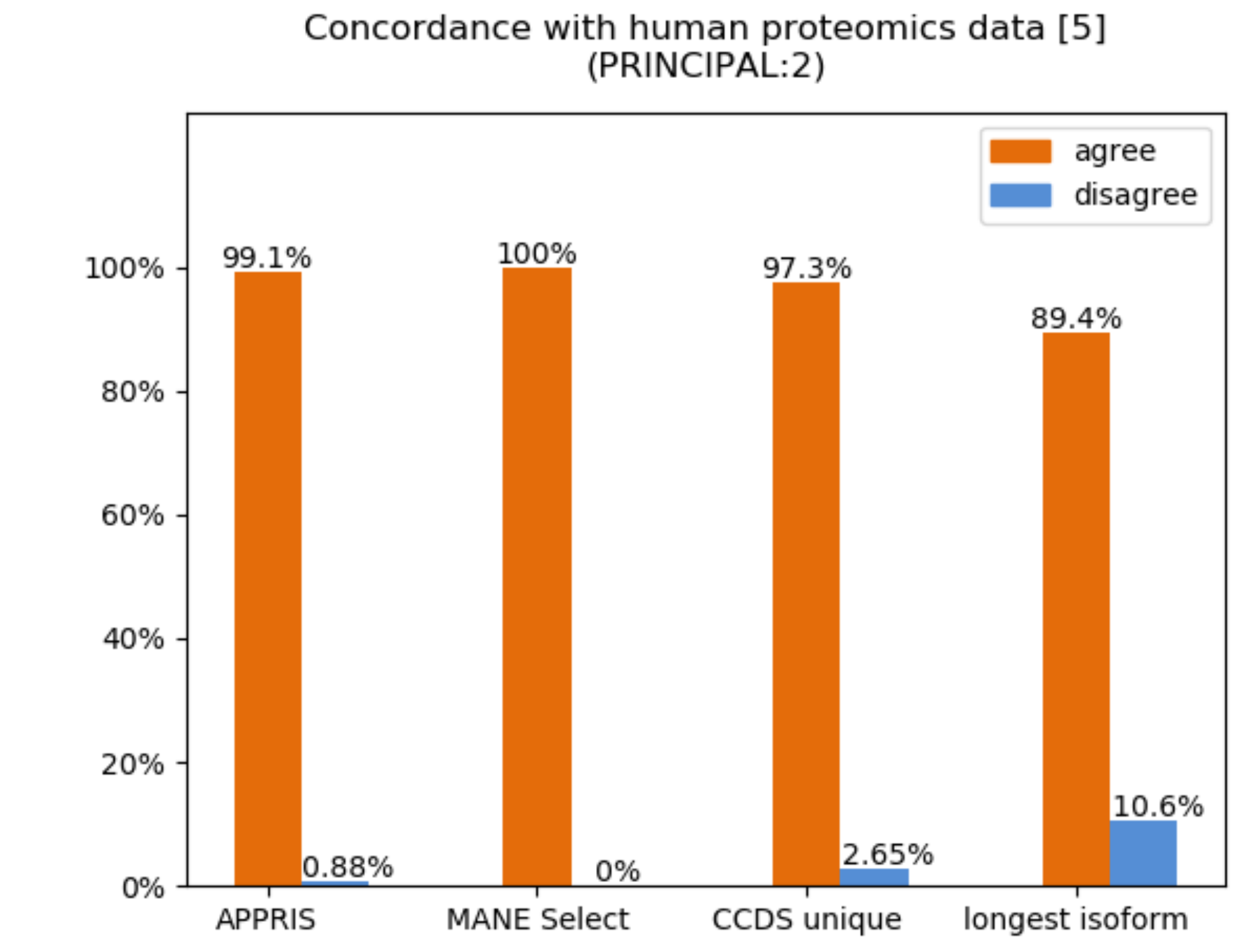
cow

fruitfly

C. elegans

VALIDATION





CONCLUSION

APPRIS principal isoforms are broadly applicable and are potentially useful in diverse areas of scientific research. Identifying the most functionally relevant or the most representative isoform is an essential first step in any genomic analysis, whether that be, for example, to ascertain if a CRISPR perturbation occurs in a principal or alternative transcript, or to choose a representative isoform for phylogenetic analysis.

Following recent updates to the APPRIS core methods and incorporation of TRIFID in the APPRIS decision steps, APPRIS predicts a PRINCIPAL:1 transcript for over 75% of genes, and identifies a highly reliable PRINCIPAL:2 or PRINCIPAL:3 isoform in a further 11% of genes using TRIFID and PROTEO, respectively.

In the human genome, APPRIS principal isoforms have high concordance with proteomics data [5], attaining a level of agreement comparable with MANE Select [11] and transcripts with a unique CCDS [12], while choosing a principal isoform for all protein-coding genes.

The APPRIS database currently houses annotations for 10 Ensembl species, of which 6 also have APPRIS annotations for RefSeq assemblies. We are open to requests for additional species.

REFERENCES

1. Tress ML et al. (2017) Trends Biochem Sci. 42(6):408-410.
2. Gonzàlez-Porta M et al. (2013) Genome Biol. 14(7):R70.
3. Rodriguez JM et al. (2013) Nucleic Acids Res. 41(D1):D110-D117.
4. Rodriguez JM et al. (2018) Nucleic Acids Res. 46(D1):D213-D217.
5. Ezkurdia I et al. (2015) J Proteome Res. 14(4):1880-1887.
6. Berman J et al. (2000) Nucleic Acids Res. 28: 235-242.
7. Lopez G et al. (2011) Nucleic Acids Res. 39(W1):W235-W241.
8. Maietta P et al. (2014) Nucleic Acids Res. 42(D1):D267-D272.
9. Altschul SF et al. (1997) Nucleic Acids Res. 25(17):3389-3402.
10. Finn RD et al. (2016) Nucleic Acids Res. 44(D1):D279-D285.
11. MANE Project www.ncbi.nlm.nih.gov/refseq/MANE [Accessed: 2020-06-30]
12. Pruitt KD et al. (2009) Genome Res. 19(7):1316-1323.

ACKNOWLEDGEMENTS

We would like to thank: Daniel Cerdán Vélez, Tomás Di Domenico, Laura Martínez-Gómez, and all in the CNIO Bioinformatics Unit.

We would also like to acknowledge the funding of the US National Institutes of Health grant 2 U41 HG007234.