Fully Automated and Accurate Annotation of Eukaryotic Genomes with BRAKER2

Katharina J. Hoff[1,2*], Tomáš Brůna[3*], Alexandre Lomsadze[3*], Mario Stanke[1,2] and Mark Borodovsky[3]

1) Institute for Mathematics and Computer Science, University of Greifswald, Greifswald, GERMANY
2) Center for Functional Genomics of Microbes, University of Greifswald, Greifswald, GERMANY
3) Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, U.S.A.

Contact: katharina.hoff@uni-greifswald.de, *) Authors contributed equally

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

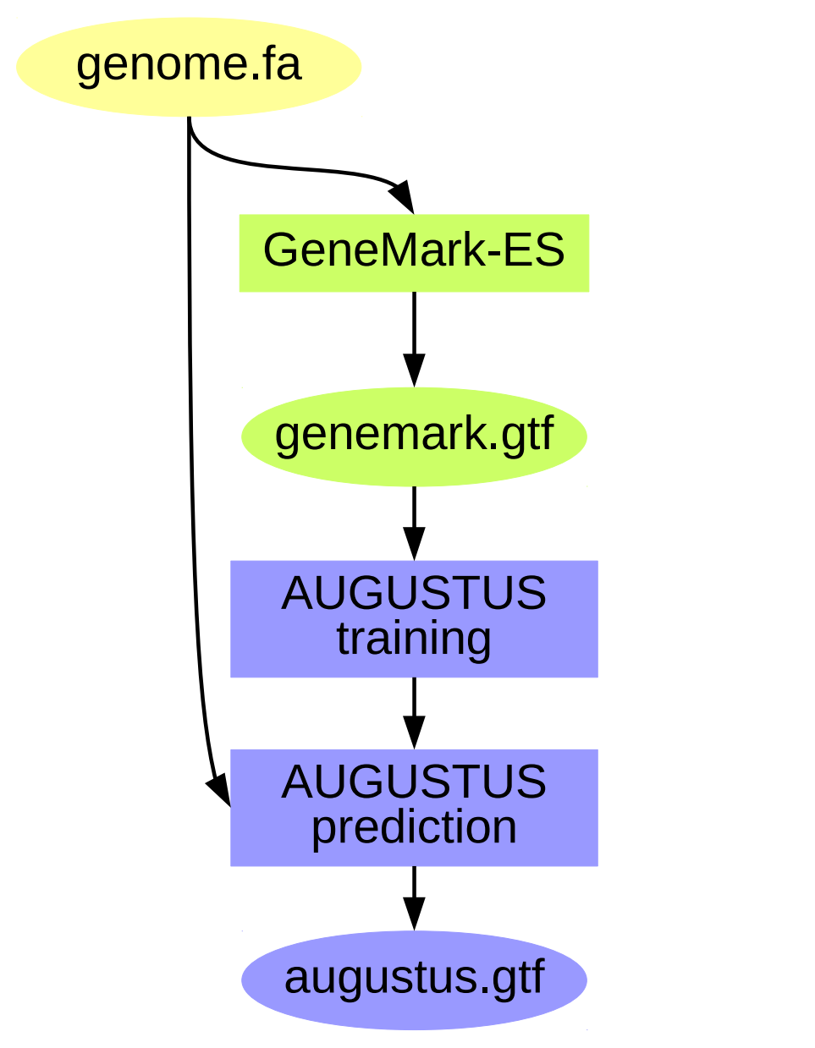
Abstract

While the number of sequenced genomes is ever growing, a vast majority of already available eukaryotic genomes may not be utilized to its full potential since it is lacking a high quality annotation of protein coding genes. Automation of the process of eukaryotic genome annotation is a challenging task due to diversity of input data situations.

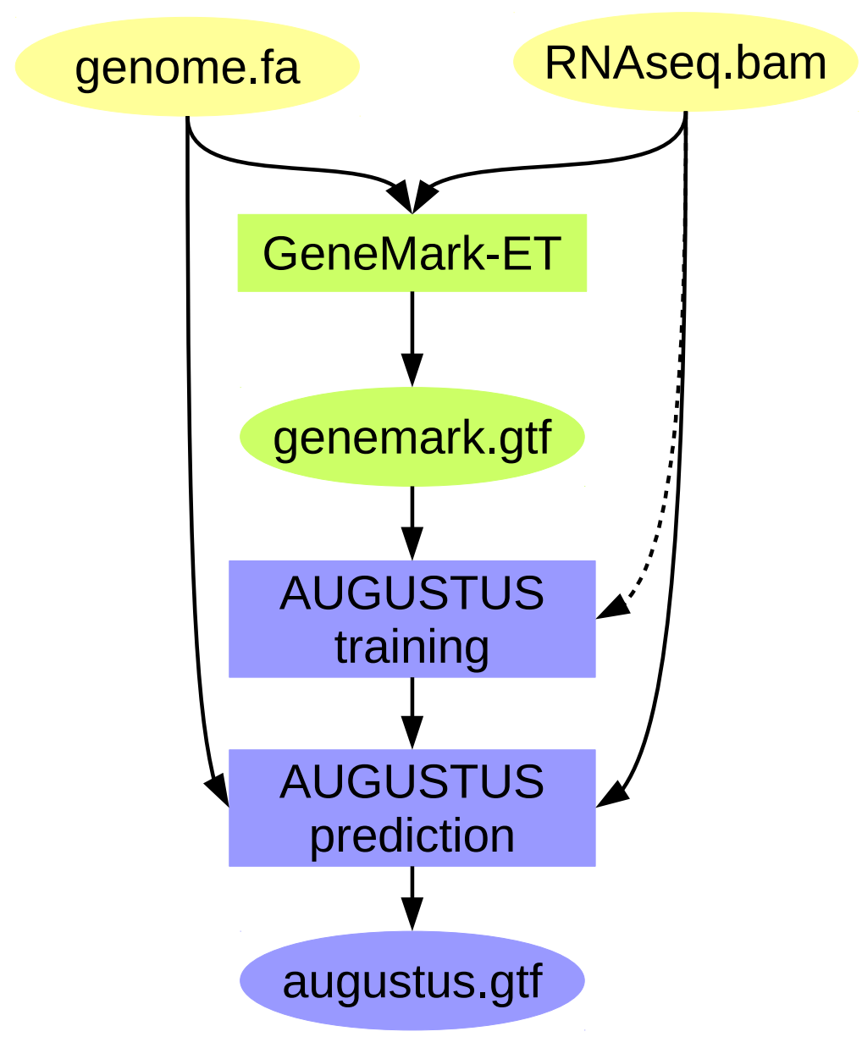
BRAKER2 [1] is an automated pipeline for annotation of protein coding genes in eukaryotic genomes. Common external data scenarios supported by BRAKER2 include the availability of i/ alignments of RNA-Seq short reads to the target genome, ii/ alignments of proteins of possibly distantly related species to the target genome or even iii/ absence of the evidence data. In all cases, BRAKER2 runs a self-training GeneMark-ET/-EP/-ES [2,3,4] depending on the external data situation, trains AUGUSTUS [5] on the genome annotation produced by GeneMark-ET/-EP/-ES and predicts genes (including alternative isoforms) with AUGUSTUS. Available extrinsic evidence is used by both tools. To use cross-species proteins, BRAKER2 automatically calls a novel ProtHint pipeline introduced in GeneMark-EP for generating protein evidence for gene prediction with GeneMark-EP and AUGUSTUS. ProtHint enables users to map proteomes of a large number of species to the target genome. Recent improvements in genome annotation accuracy with protein evidence reached in GeneMark-EP lead to an increase in genome annotation accuracy by BRAKER2.

The BRAKER2 project locations are **https://github.com/Gaius-Augustus/BRAKER** and **https://github.com/gatech-genemark/BRAKER2**.

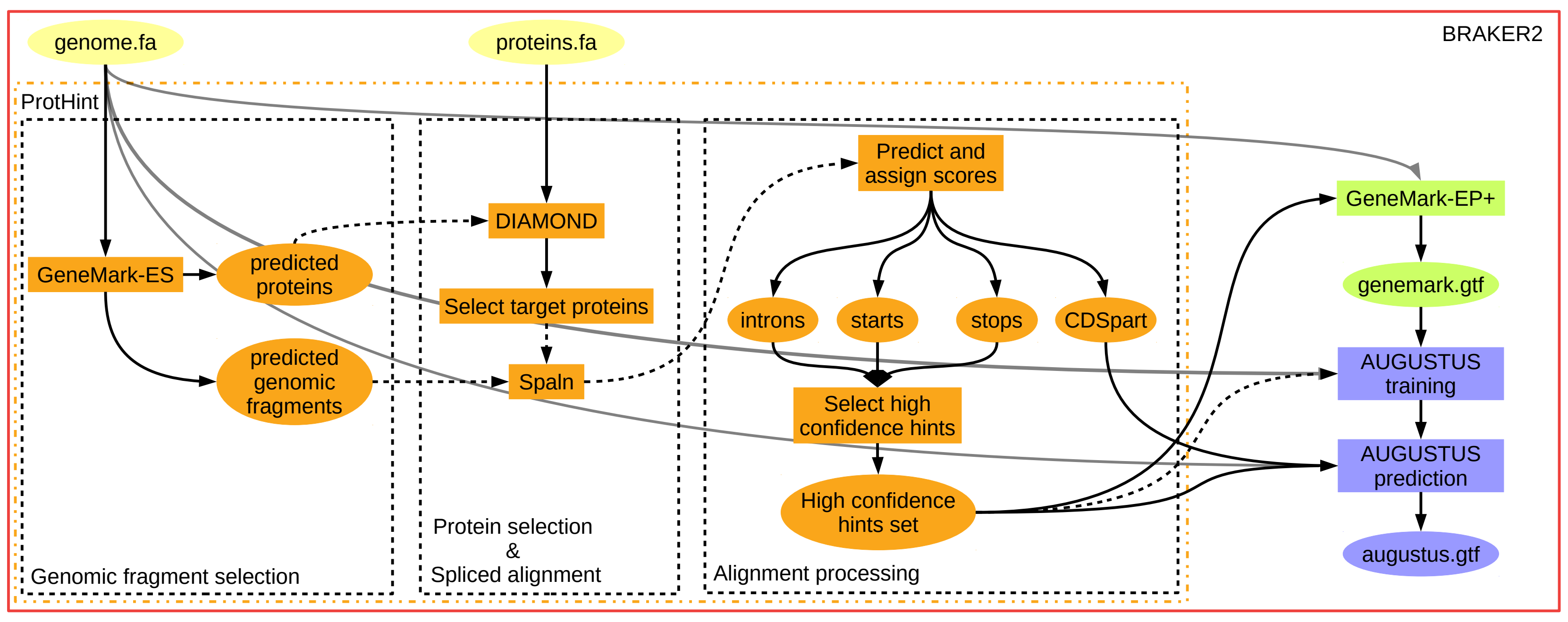
BRAKER-ES: Genome only



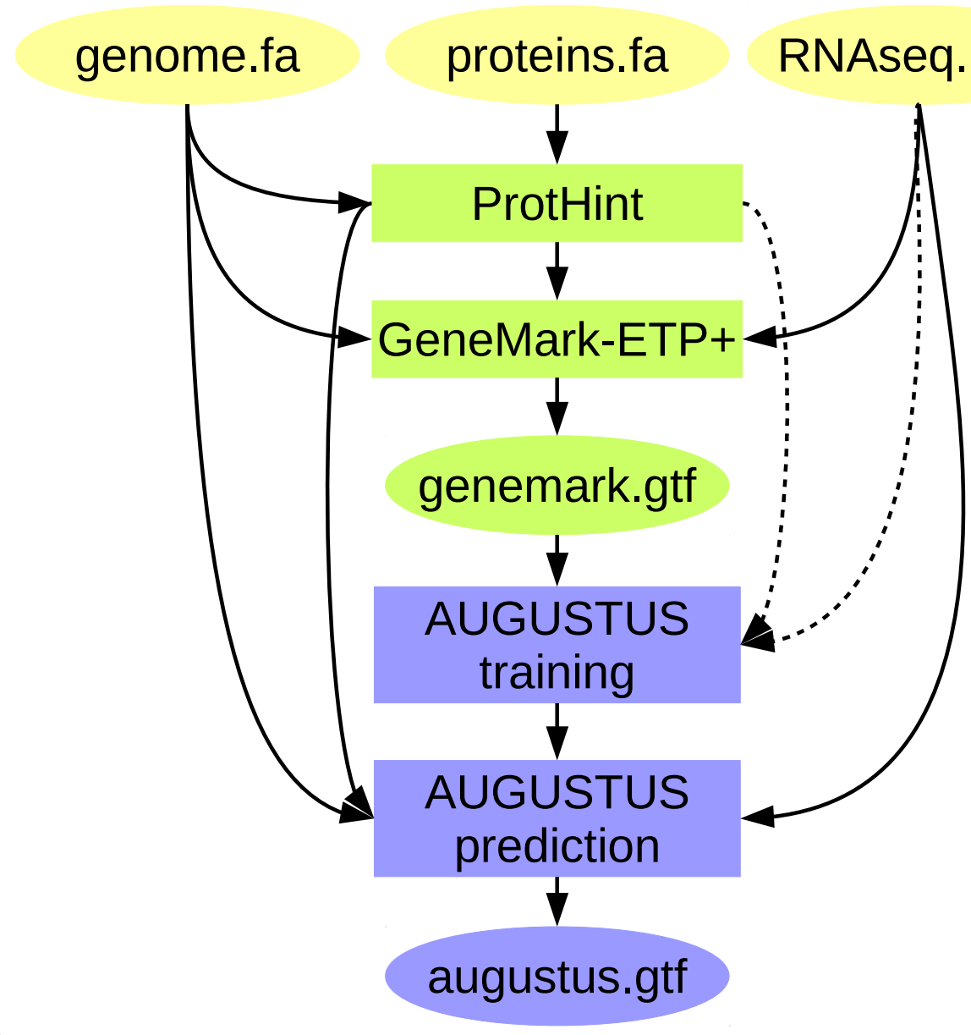
BRAKER-ET: Genome & RNA-seq



BRAKER-EP: Genome & proteins



BRAKER-ETP: Genome, RNA-seq & proteins*



*) Future version is shown. Current version uses GeneMark-ET instead of GeneMark-ETP+. Improvements in accuracy are expected with GeneMark-ETP+.

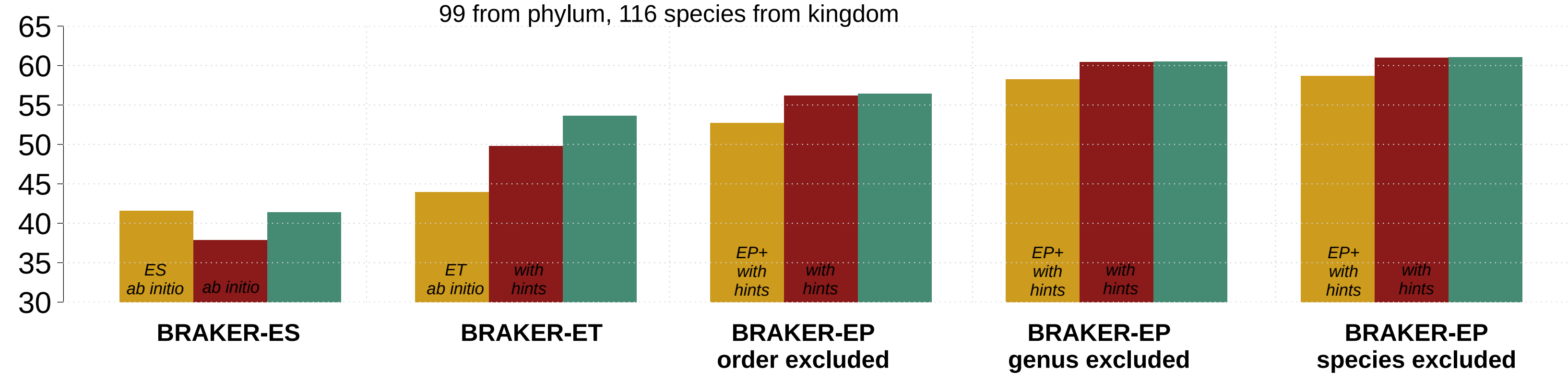
Transcript Prediction Accuracy with RNA-Seq or Proteins

$$\text{Transcript prediction accuracy } F1 = \frac{2 * Sensitivity * Specificity}{Sensitivity + Specificity}$$

🟧 GeneMark
🟥 AUGUSTUS
set union → 🟩 Union

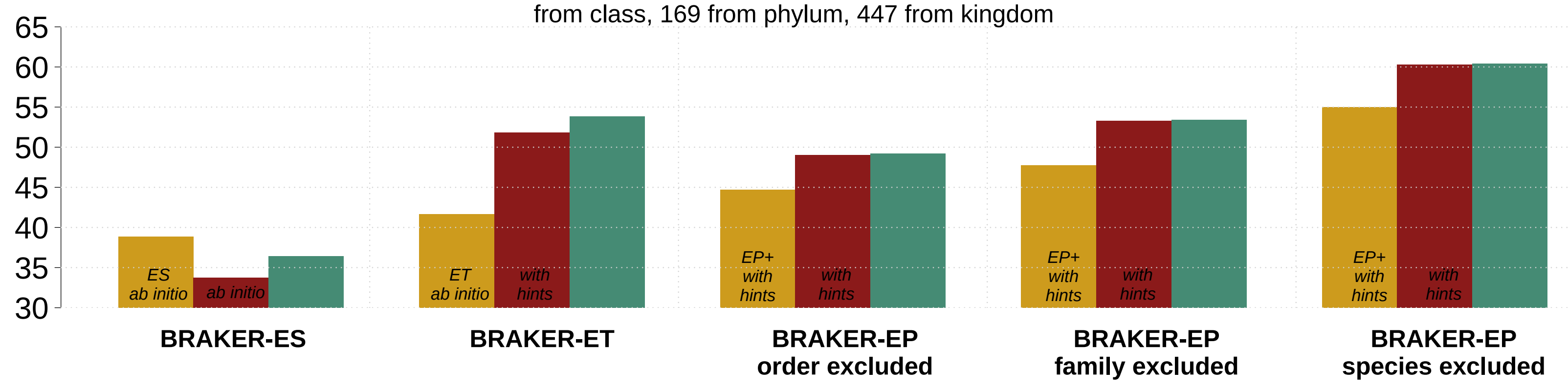
Arabidopsis thaliana

- Thale cress, genome size 151 MB, RNAseq from Varus [6] with Hisat2 [7]
- OrthoDB 10 [8] Plantae : 1 species from the same genus, 7 species from family, 9 from order, 99 from phylum, 116 species from kingdom



Drosophila melanogaster

- Fruit fly, genome size 130 MB, RNAseq from Varus with Hisat2
- OrthoDB 10 Arthropoda: 19 species from the same genus, 19 from family, 55 from order, 147 from class, 169 from phylum, 447 from kingdom



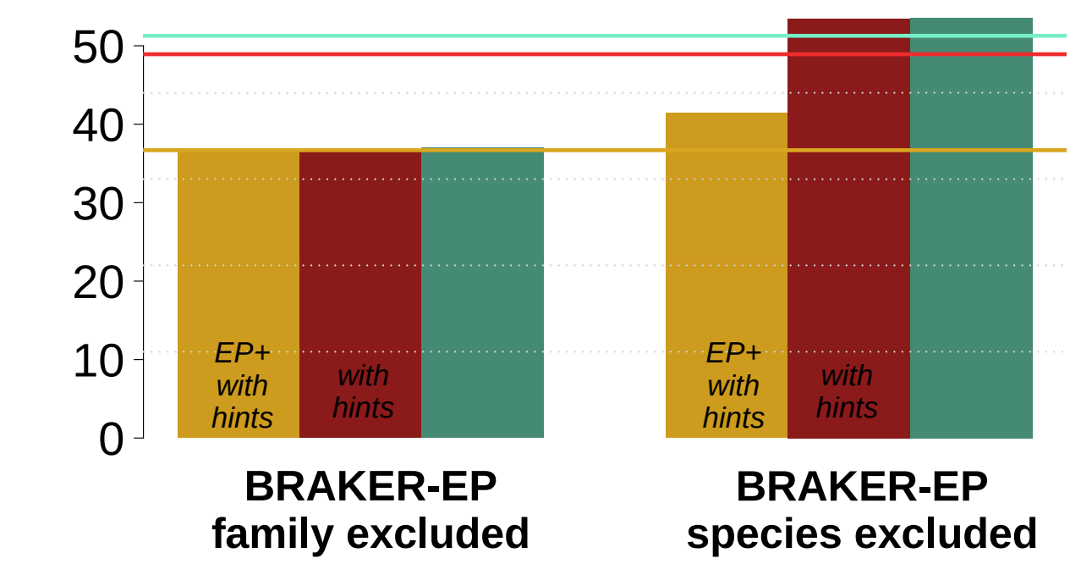
Runtime

BRAKER-EP incl. ProtHint on 8 CPUs / 8 GB RAM:

- *D. melanogaster* <9 h
- *S. lycopersicum* <20 h

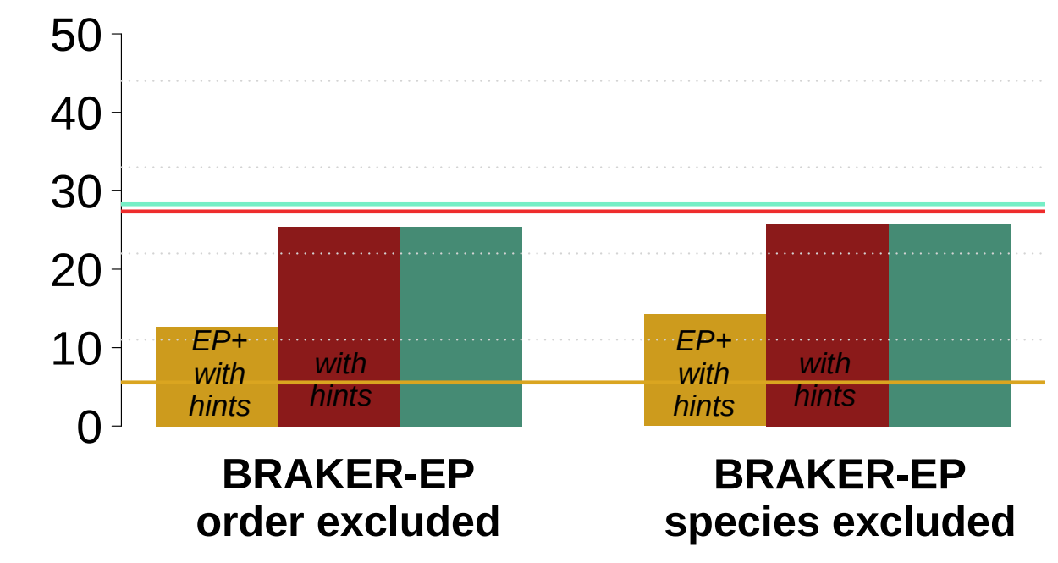
Caenorhabditis elegans

- Roundworm, genome size 84 MB, RNAseq from Varus with Hisat2
- OrthoDB 10 Metazoa : 2 species from the same genus, 2 from family, 4 from order, 5 from class, 6 from phylum, 447 from kingdom



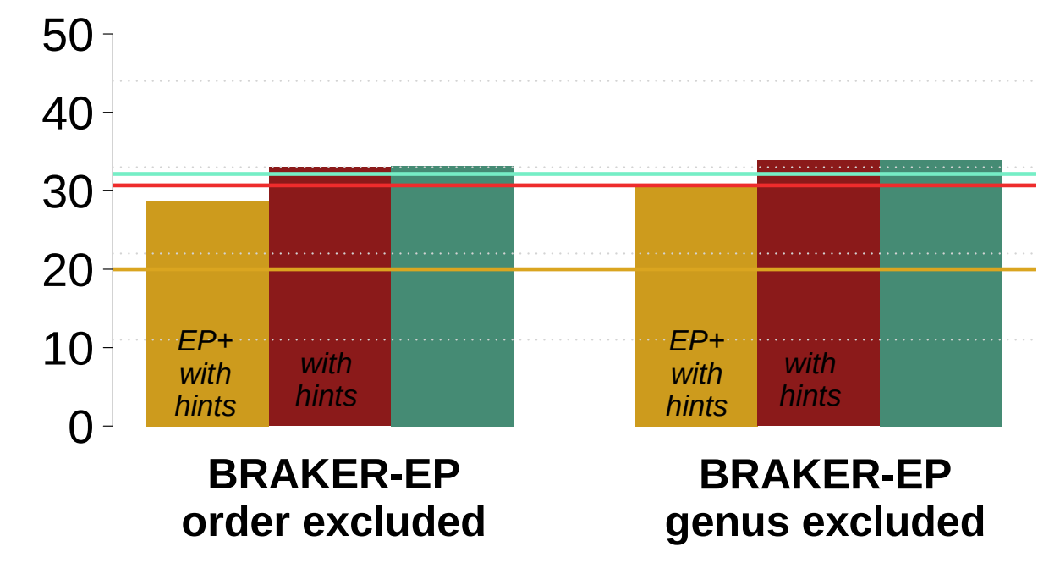
Danio rerio

- Zebrafish, genome size 1345 MB, RNAseq from Varus with Hisat2
- OrthoDB 10 Chordata : 4 species from the same family, 4 from order, 49 from class, 345 from phylum, 447 from kingdom



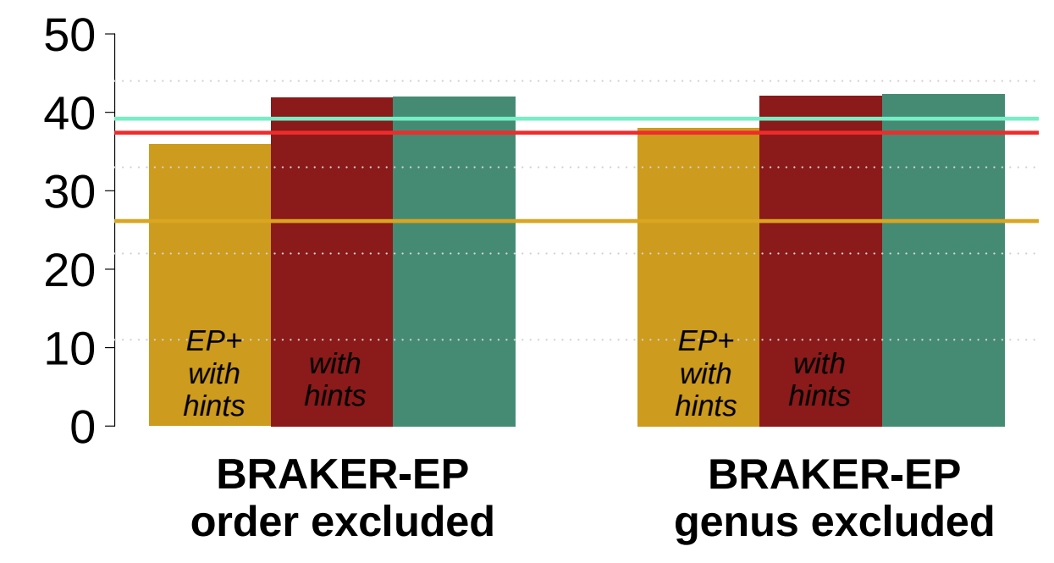
Solanum lycopersicum

- Tomato, genome size 1390 MB, RNAseq from Varus with Hisat2
- OrthoDB 10 Plantae : 1 species from the same genus, 9 from family, 10 from order, 99 from phylum, 116 from kingdom



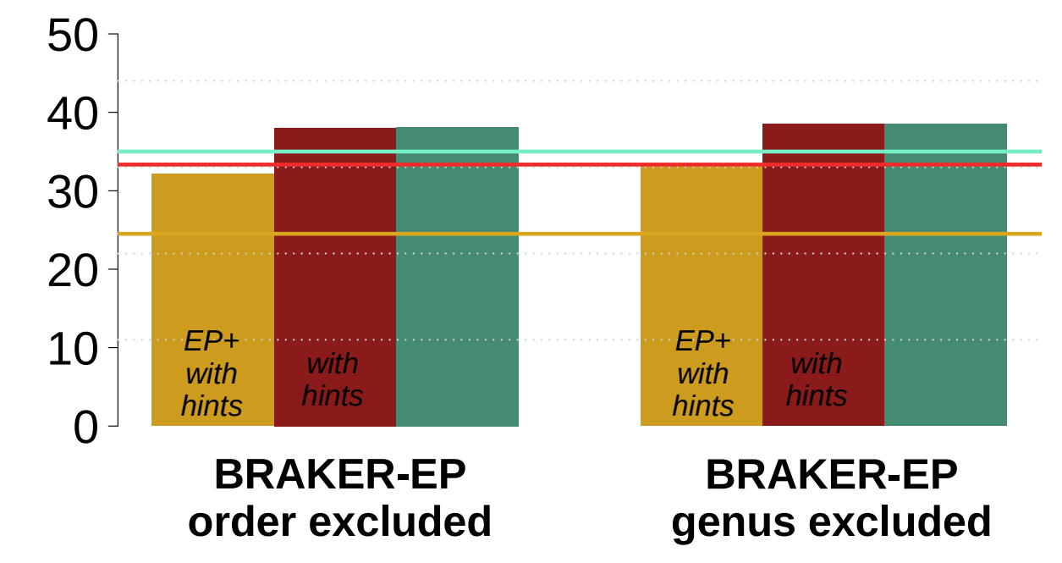
Medicago truncatula

- Barrel medic, genome size 807 MB, RNAseq from Varus with Hisat2
- OrthoDB 10 Plantae : 9 species from the same family, 9 from order, 99 from phylum, 116 from kingdom



Fragaria vesca

- Wild strawberry, genome size 198 MB, RNAseq from Varus with Hisat2
- OrthoDB 10 Plantae : 4 species from family, 9 from order, 99 from phylum, 116 from kingdom

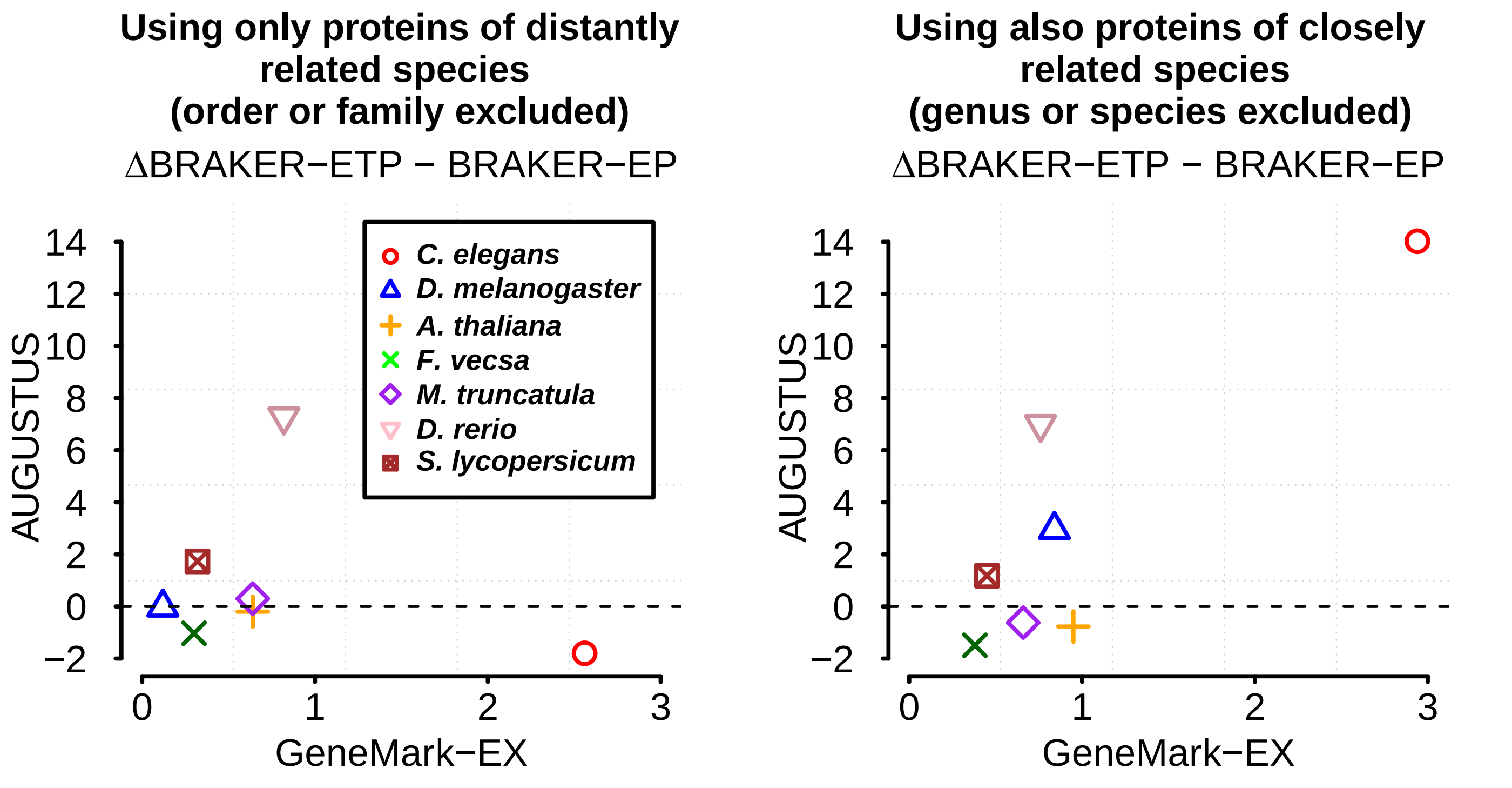


Lines show accuracy of BRAKER-ET:

- GeneMark-ET (ab initio)
- AUGUSTUS (with hints)
- Union

Using RNA-Seq and Proteins in Current BRAKER2

Difference in Transcript Prediction Accuracy F1



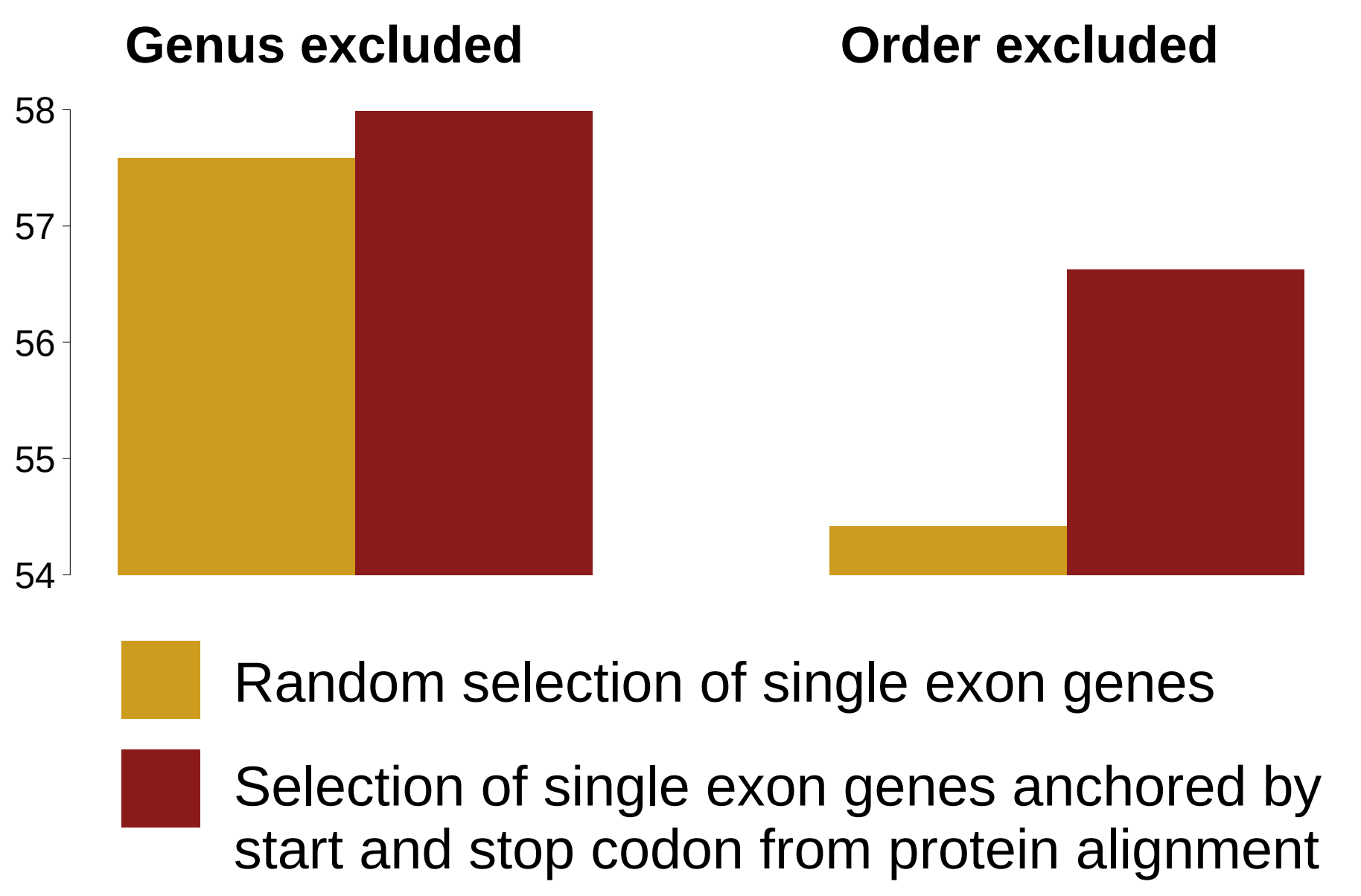
Species with long introns, such as *Danio rerio* or *Solanum lycopersicum*, typically benefit from combining RNA-Seq and protein evidence in the current version of BRAKER-ETP. Improvements for other species are to be expected, soon.

Training Gene Selection for AUGUSTUS

Anchored Single Exon Genes

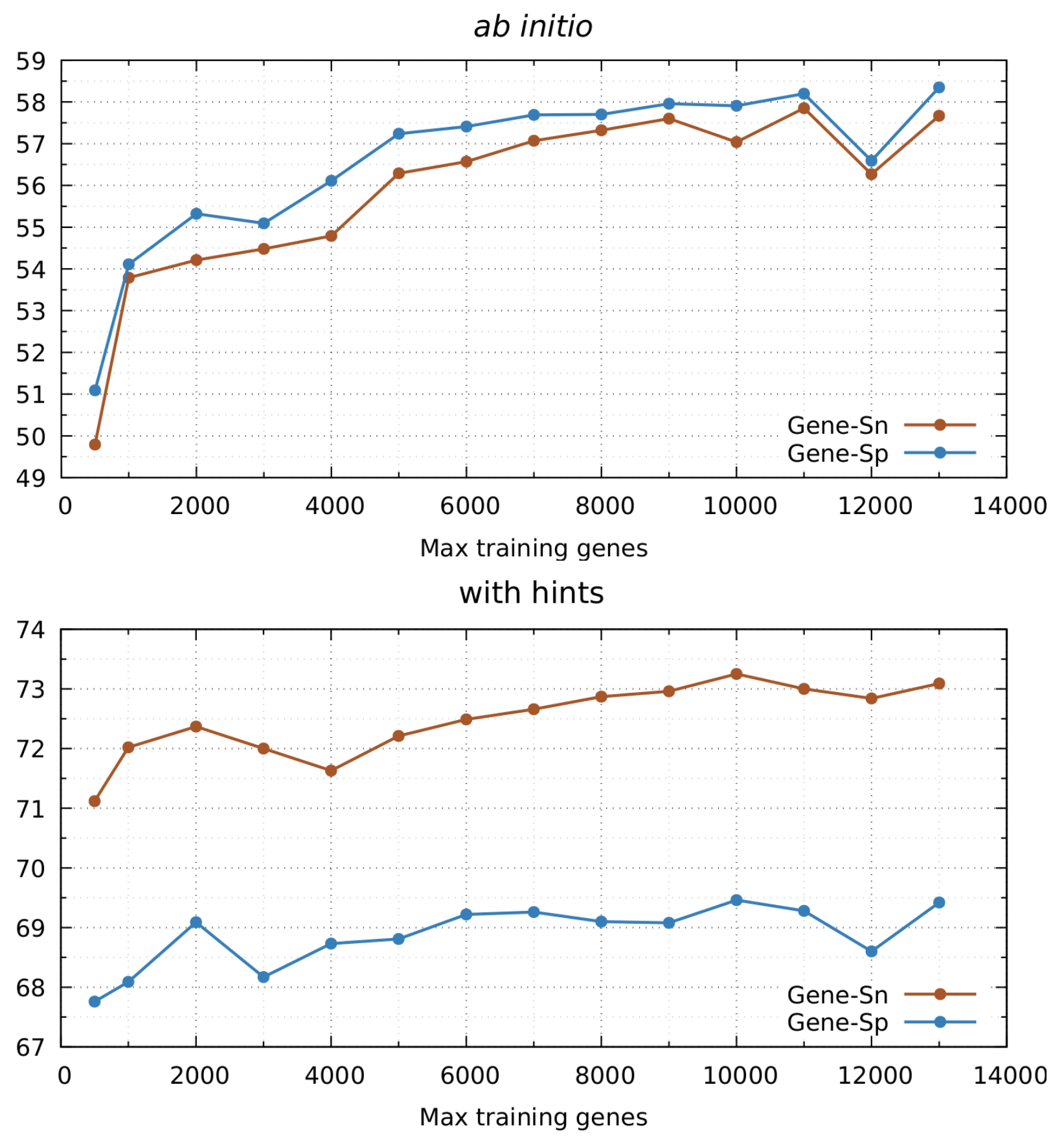
Protein evidence allows filtering single exon genes predicted by GeneMark-EP+ for those that are anchored by a start- and stop-codon from evidence prior training AUGUSTUS.

AUGUSTUS ab initio gene prediction accuracy F1 in BRAKER-EP based on different versions of selection of single-exon genes for training



Number of Genes

Influence of number of training genes for AUGUSTUS in BRAKER-EP



BRAKER2 by default uses a maximum of 8000 genes for training AUGUSTUS.

References

[1] Hoff KJ et al. (2019) “Whole-Genome Annotation with BRAKER” In Gene Prediction, pp. 65-95. Humana, New York, NY.
[2] Lomsadze A et al (2014) "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm." Nucleic Acids Research 42(15):e119-e119.
[3] Bruna T et al. (2020) “GeneMark-EP and-EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins.” bioRxiv https://doi.org/10.1101/2019.12.31.891218.
[4] Ter-Hovhannisyan V et al. (2008) “Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training.” Genome Research, 18(12):1979-90.
[5] Stanke M et al. (2008) "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics 24(5):637-644.
[6] Stanke M et al. (2019) “VARUS: sampling complementary RNA reads from the sequence read archive” BMC Bioinformatics 20:558.
[7] Daehwan K et al. (2015) “HISAT: a fast spliced aligner with low memory requirements.” Nature methods 12(4): 357.
[8] Kriventseva EV et al. (2018) “OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research: doi.org/10.1093/nar/gky1053s.

Funding

This research is supported by US National Institutes of Health grant GM128145.