Proteins Only: How Accurately Can We Annotate Large Genomes?

Tomáš Brůna[1], Heng Li[2,3], Joseph Guhlin[4], Daniel Honsel[5], Steffen Herbold[6], Mario Stanke[7], Natalia Nenasheva[7], Matthis Ebel[7], Lars Gabriel[7], Katharina J. Hoff[7]

[1]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, USA; [2]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, USA; [3]Department of Biomedical Informatics, Harvard Medical School, Boston, USA; [4]Genomics Aotearoa and Laboratory for Evolution and Development, Department of Biochemistry, University of Otago, Dunedin, New Zealand; [5]Institute of Computer Science, University of Göttingen, Göttingen, Germany; [6]Faculty for Computer Science and Mathematics, Universityof Passau, Passau, Germany; [7]Institute of Mathematics and Computer Science, and Center for Functional Genomics of Microbes, University of Greifswald, Greifswald, Germany.

Contacts: katharina.hoff@uni-greifswald.de, tbruna@lbl.gov

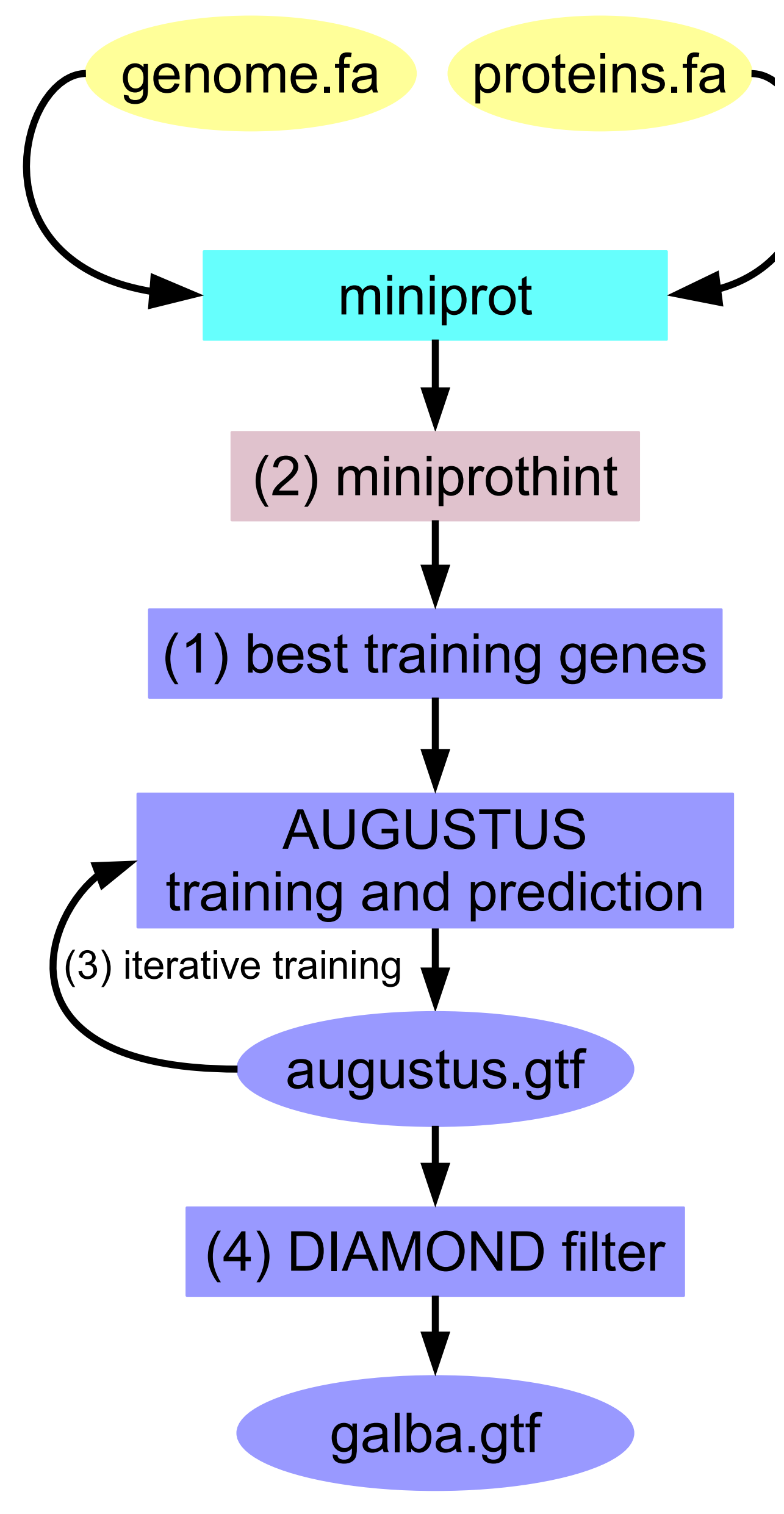
UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

JGI JOINT GENOME INSTITUTE

Abstract

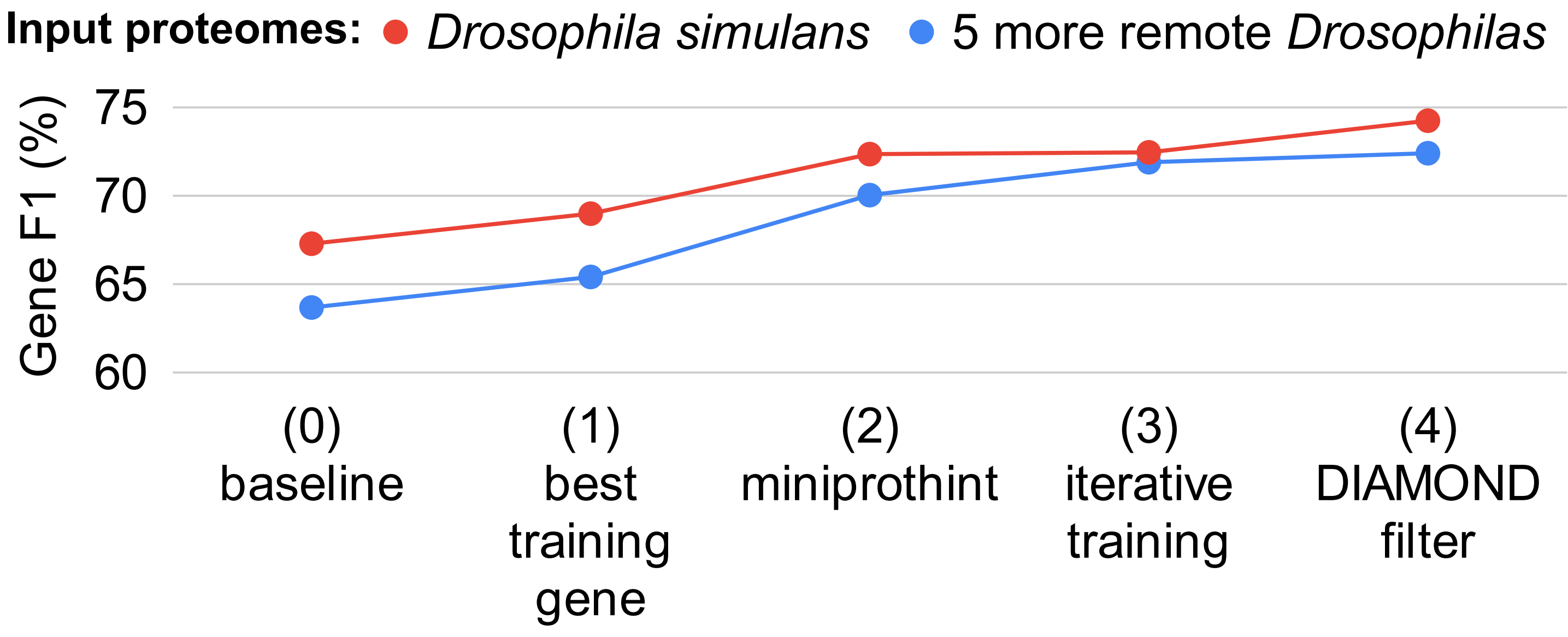
With the swift progress of the Earth Biogenome Project, the scientific community is gaining access to an unprecedented number of eukaryotic genomes. Despite this abundance, a significant portion of genomes remains unannotated for protein-coding genes, and a lack of transcriptome data in some cases further complicates the situation. Addressing this critical gap, we introduce GALBA [1], a fully automated pipeline for the structural annotation of protein coding genes, demonstrating robust performance, especially in annotating large vertebrate genomes. GALBA leverages the capabilities of miniprot, a fast protein-to-genome aligner, in synergy with AUGUSTUS [2], ensuring high prediction accuracy. Emphasizing accessibility and user-friendliness, GALBA is fully open source and readily available as a docker image, ensuring easy execution in high-performance computing environments via Singularity. Get GALBA at `https://github.com/Gaius-Augustus/GALBA`.

GALBA's workflow

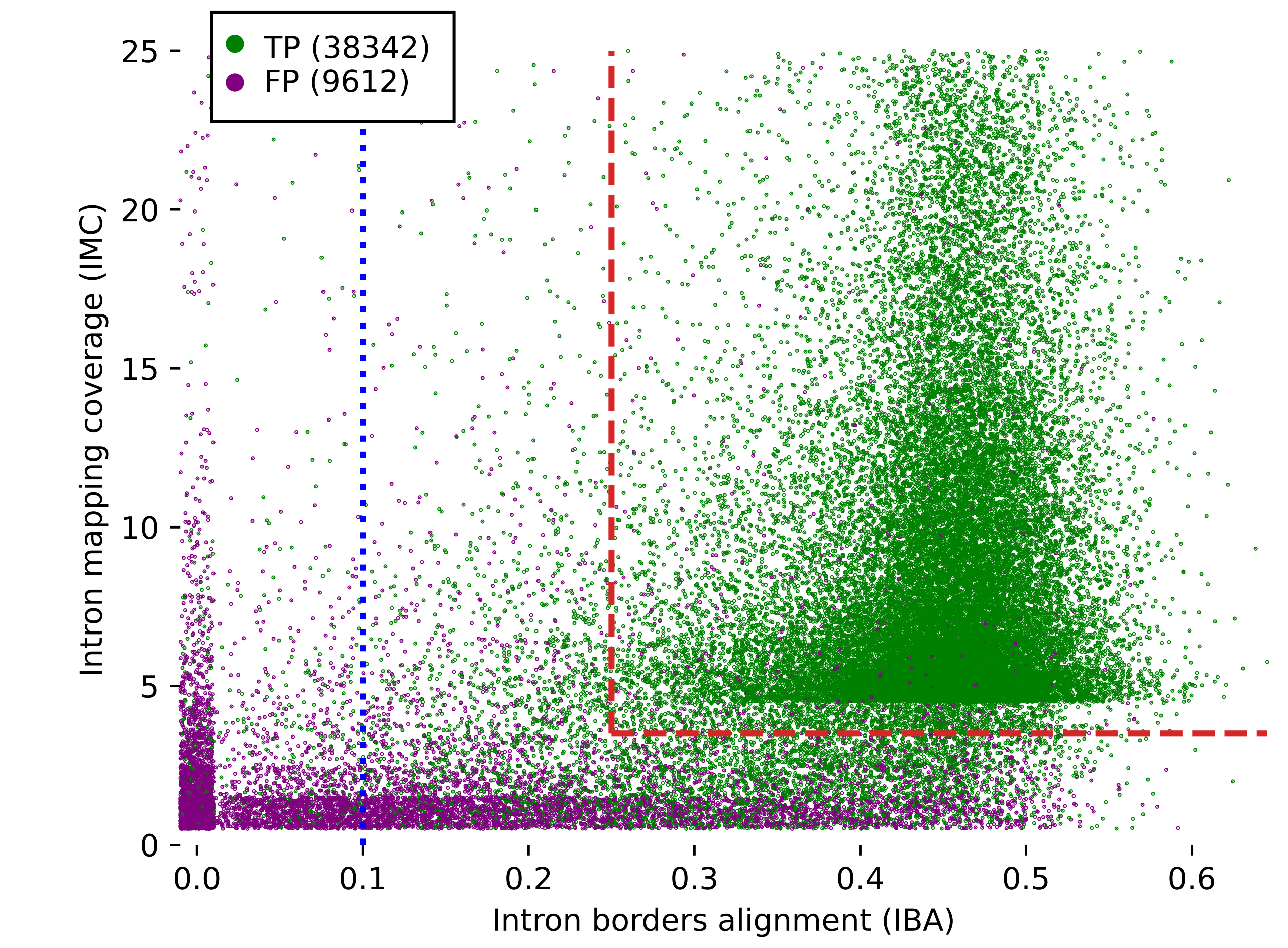


The GALBA pipeline. Initially, miniprot performs rapid spliced alignment of input proteins to the target genome. Subsequently, miniprothint scores and filters the alignments. Best-scoring alignments are employed to train AUGUSTUS, which generates an initial set of predicted genes. The initial set is then utilized to refine the AUGUSTUS model. The refined model, in turn, predicts the final set of genes, which undergoes filtering based on DIAMOND-derived homology [3, 4] to proteins in the input protein set.

The impact of major components on prediction accuracy in *D. melanogaster* is shown below.

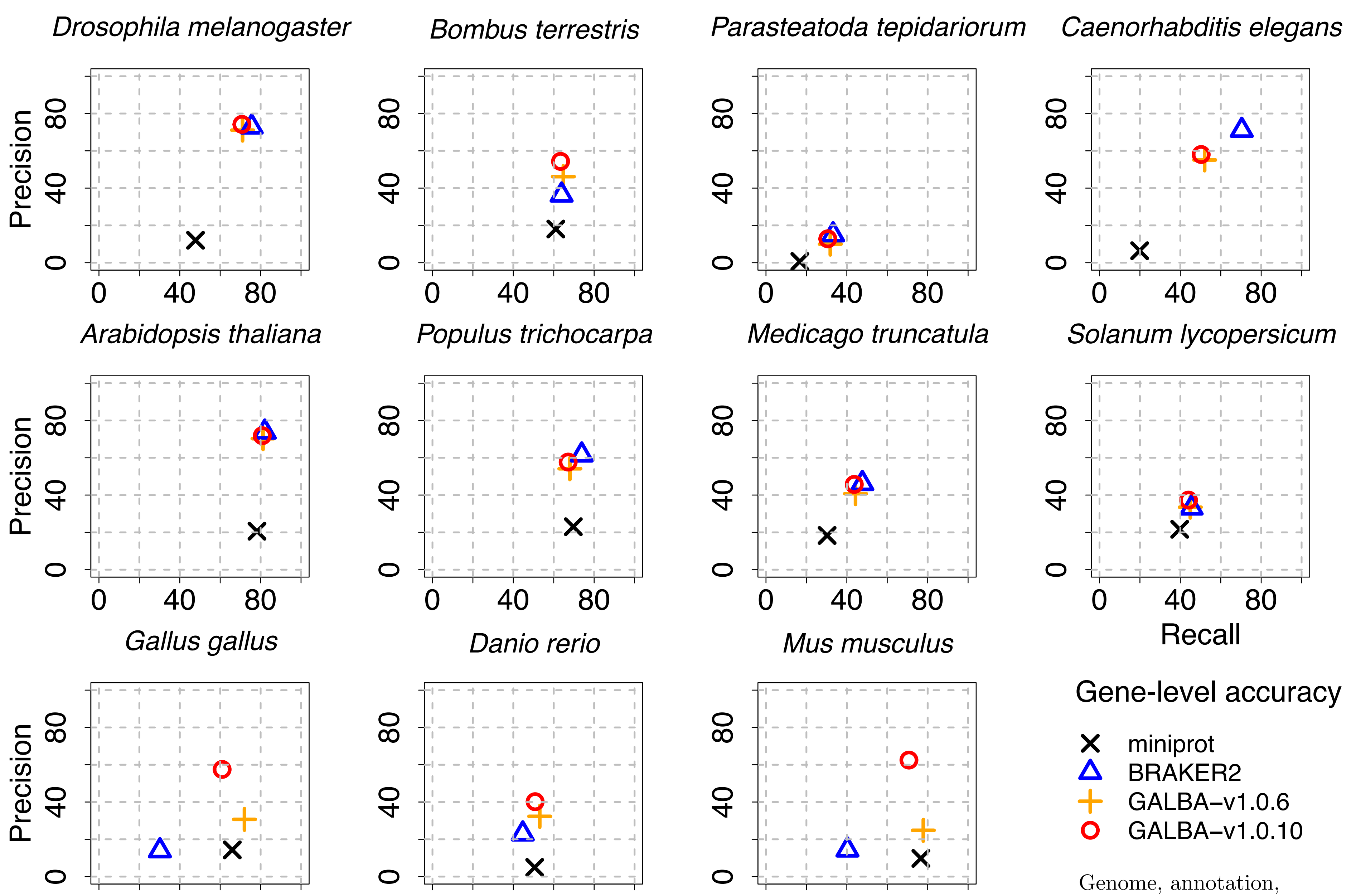


Miniprothint filtering

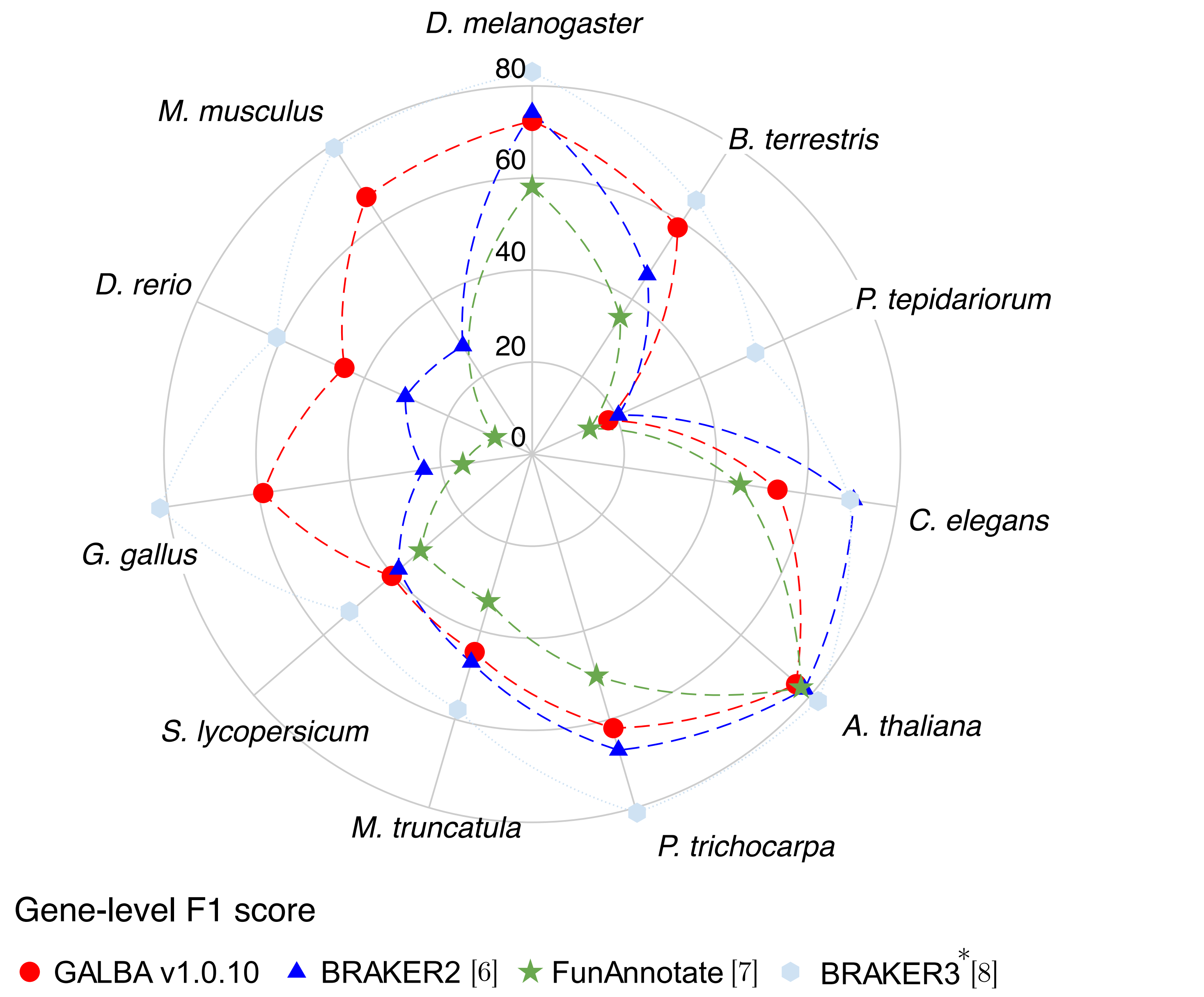


Miniprothint scores [5] and filters miniprot alignments. Above, miniprot-predicted introns are characterized by miniprothint-derived scores, facilitating a reliable distinction between true (green) and false (purple) positive predictions, indicated by the red dashed lines. Data is from proteins of 5 *Drosophilas* aligned to *D. melanogaster* [1].

Results



Genome, annotation, and protein sources are described in [1].



*Results of BRAKER3 with RNA-Seq and OrthoDB order-excluded inputs are from [8].

Recall: % of correctly found genes in the reference annotation
Precision: % of correctly found genes in the predicted set
F1 score: harmonic mean of Precision and Recall

- GALBA excels over other homology-based gene predictors in large genomes.
- BRAKER3 is the preferred choice when a transcriptome is available.

Summary

- GALBA is a fully automated pipeline for protein coding structural gene annotation in eukaryotes.
- Protein sequences of ≥ 1 related species serve as evidence.
- GALBA achieves good prediction accuracy across a diverse set of eukaryotic genomes.
- In larger, more complex genomes, GALBA significantly outperforms other homology-based gene predictors.
- GALBA is freely available and easy to execute.

References, funding & acknowledgements

[1] Tomáš Brůna et al. "Galba: genome annotation with miniprot and AUGUSTUS". In: *BMC bioinformatics* 24.1 (2023), p. 327.
[2] Mario Stanke et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding". In: *Bioinformatics* 24.5 (2008), pp. 637–644.
[3] Benjamin Buchfink et al. "Fast and sensitive protein alignment using DIAMOND". In: *Nature methods* 12.1 (2015), pp. 59–60.
[4] Ethan R Tolman et al. "Newly Sequenced Genomes Reveal Patterns of Gene Family Expansion in select Dragonflies (Odonata: Anisoptera)". In: *bioRxiv* (2023), pp. 2023–12.
[5] Tomáš Brůna et al. "GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins". In: *NAR genomics and bioinformatics* 2.2 (2020), lqaa026.
[6] Tomáš Brůna et al. "BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database". In: *NAR genomics and bioinformatics* 3.1 (2021), lqaa108.
[7] *funannotate*. `https://github.com/nextgenusfs/funannotate`.
[8] Lars Gabriel et al. "BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA". In: *bioRxiv* (2023).

Funding: German Research Foundation grant 277249973 to K.J.H.; Project Data Competency granted to K.J.H. and M.S. by the government of Mecklenburg-Vorpommern; US National Institute of Health grant R01HG010040 to H.L.; German Research Foundation grant 391397397 to S.H. and M.S.

Acknowledgements: We thank Ethan Tolman & Paul Frandsen for the idea of DIAMOND homology filtering.