Fully Automated and Accurate Annotation of Plant and Animal Genomes with **BRAKER2**

Plant and Animal Genomes XXVIII, January 12th 2020

Katharina J. Hoff,
Tomáš Brůna,
Alexandre Lomsadze,
Mario Stanke,
Mark Borodovsky

Poster PE0208

Presenting author e-mail: katharina.hoff@uni-greifswald.de

Contents

Structural Genome Annotation Problem

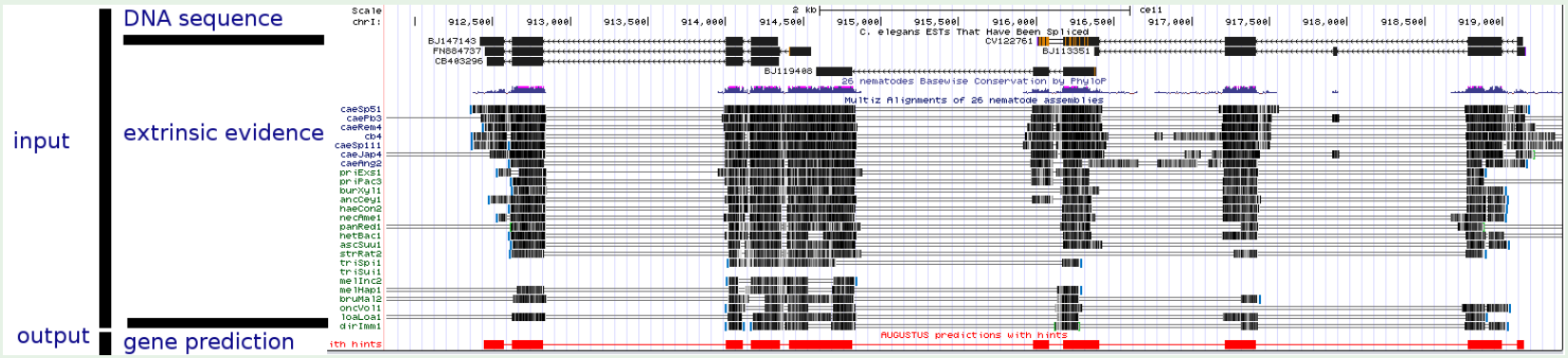
Input

- genome assembly
- extrinsic evidence, e.g. from RNA-Seq, **protein database**

Output

- protein-coding genes: exon-intron structures (`.gff`)

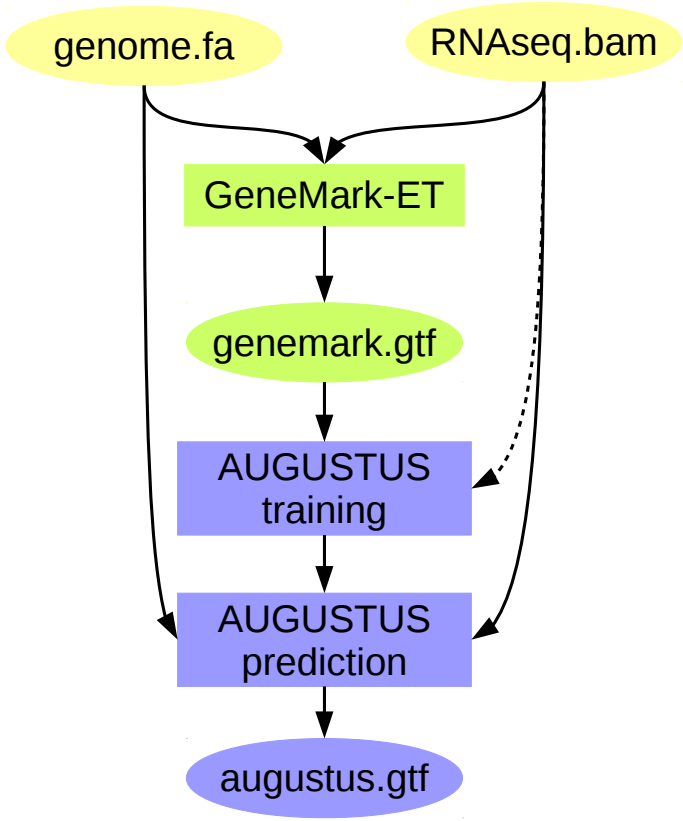
Example (from Chr I in *C. elegans*)



BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

Katharina J. Hoff ✉, Simone Lange, Alexandre Lomsadze, Mark Borodovsky ✉, Mario Stanke

Bioinformatics, Volume 32, Issue 5, 1 March 2016, Pages 767–769,
https://doi.org/10.1093/bioinformatics/btv661



- spliced alignments of RNA-Seq are used by GeneMark-ET and AUGUSTUS
- >4000 downloads since 2015
- 321 citations (Google Scholar)

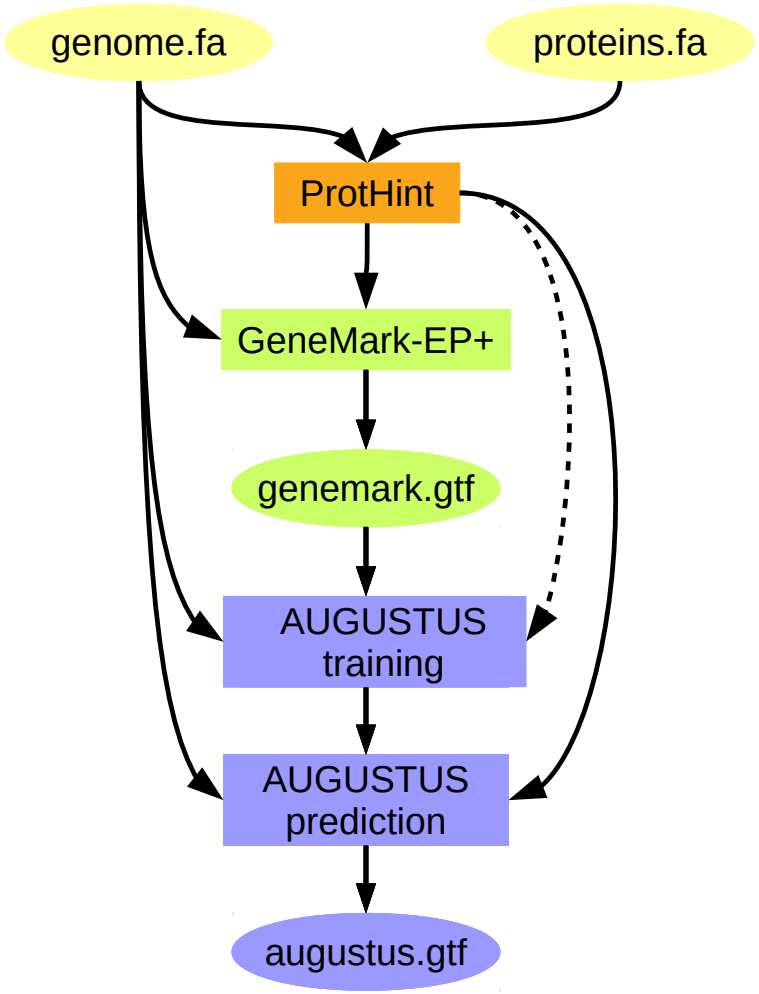
Whole-Genome Annotation with BRAKER

Katharina J. Hoff, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke

in Kollmar M. (eds) Gene Prediction. Methods in Molecular Biology, vol 1962. Humana, New York, NY, 2019

Katharina J. Hoff, Tomáš Brůna, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky

BRAKER2: GeneMark-EP+ and AUGUSTUS



Protein to Genome Alignment

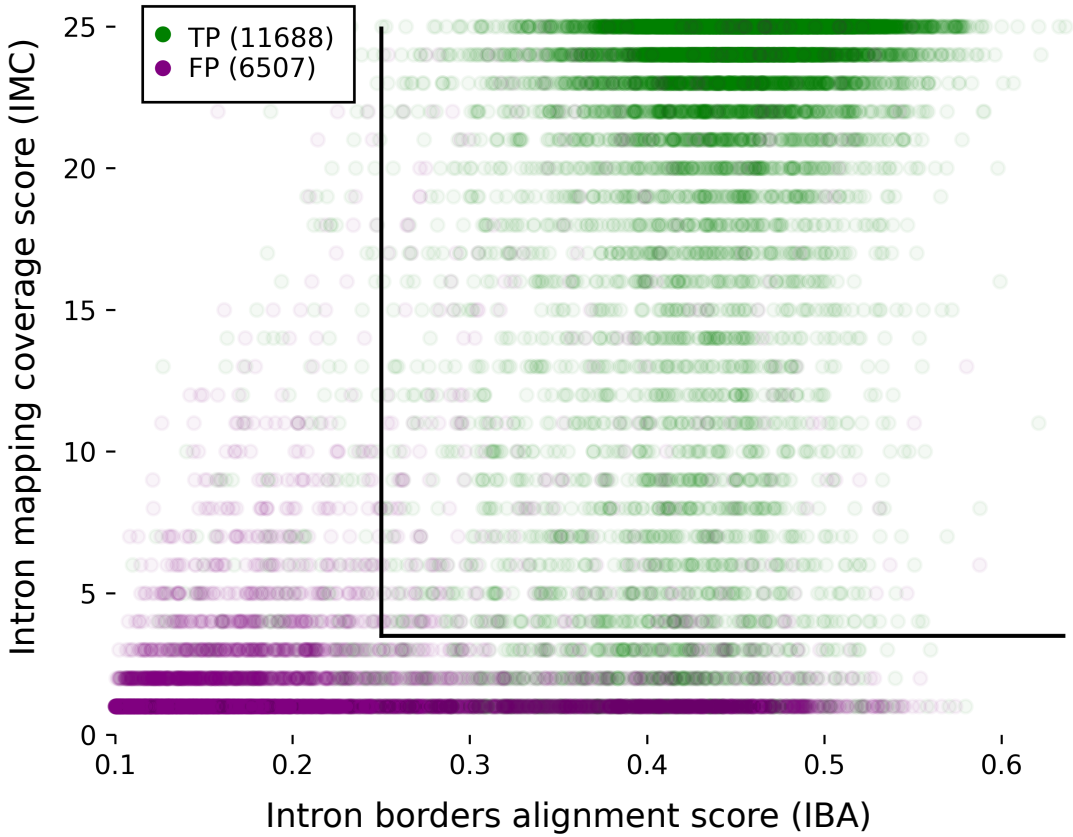
- end-to-end spliced alignment is difficult if sequence similarity is low
- ProtHint: fast large-scale spliced alignment
- both *large numbers* and *high qualities* of alignments may corroborate a gene feature

GeneMark-EP/EP+:

- Posters PO1003 & PE1004
- Talk by A. Lomsadze Jan 14 12:10 Golden West

Protein to Genome Alignment

ProtHint Maps and Scores Introns (and starts/stops)



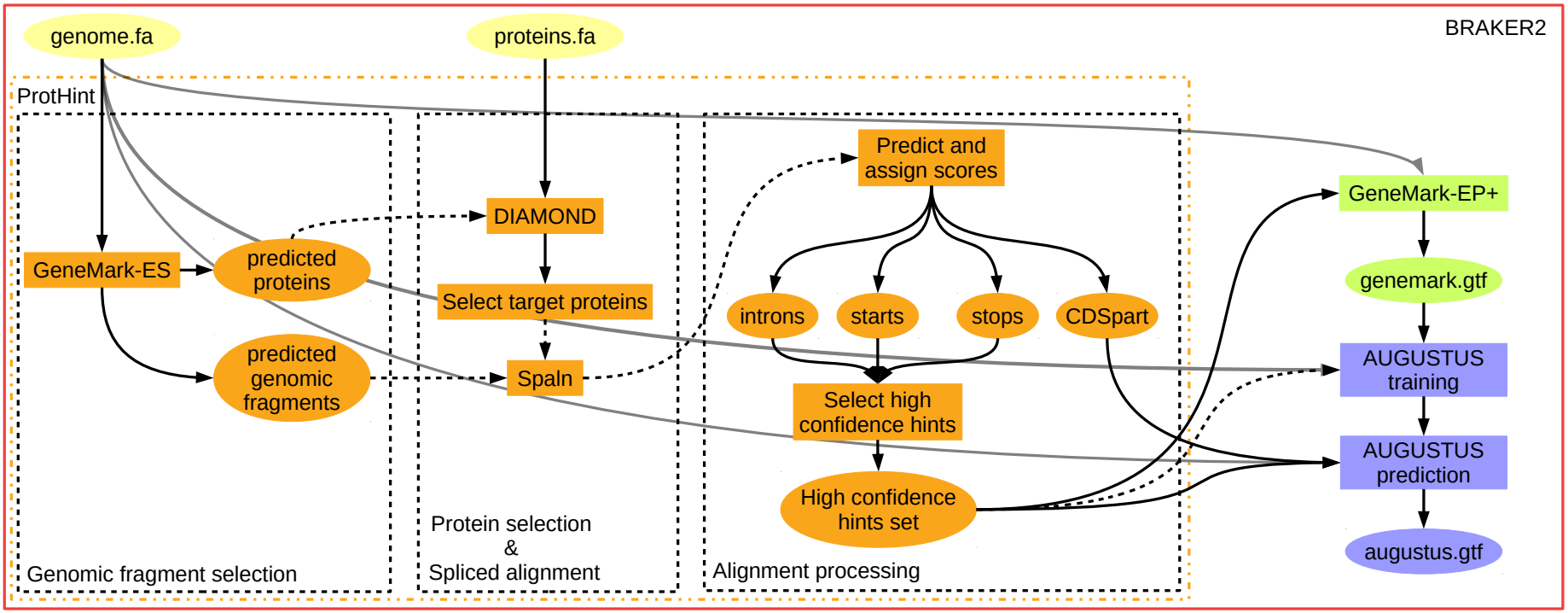
IMC: Intron Mapping Coverage
IBA: Intron Borders Alignment Score

GeneMark-EP and -EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins

Tomas Bruna[1,†], Alexandre Lomsadze[2†,] and Mark Borodovsky[1,2,3,*]

BRAKER2 with Proteins

Fast Genome Annotation with Large Protein Database

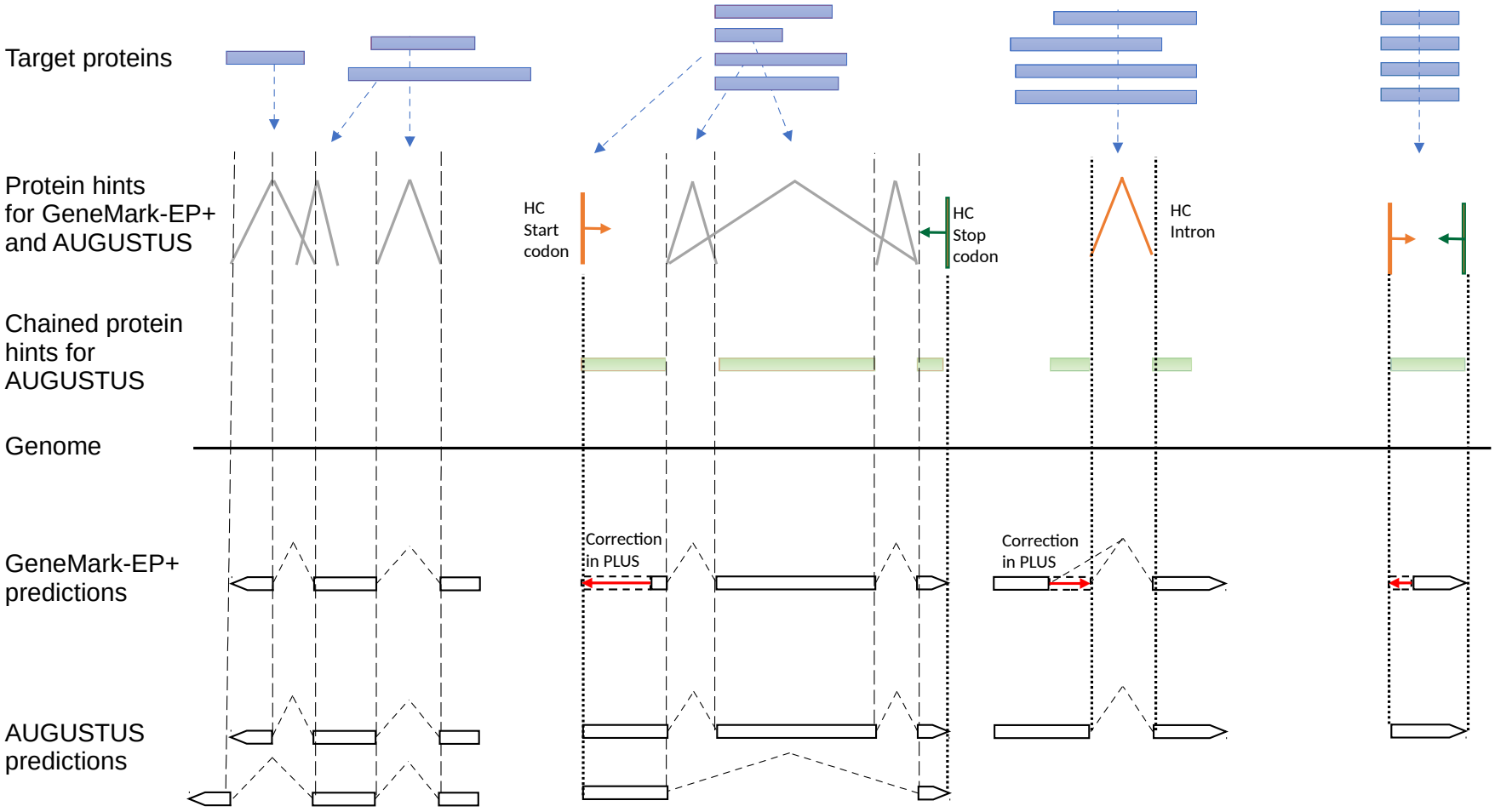


	Genome Size	# Proteins	Hours
D. melanogaster	130 MB	2,314K	< 9
S. lycopersicum	1390 MB	3,456K	< 20

- 8 CPUs
- 8 GB RAM

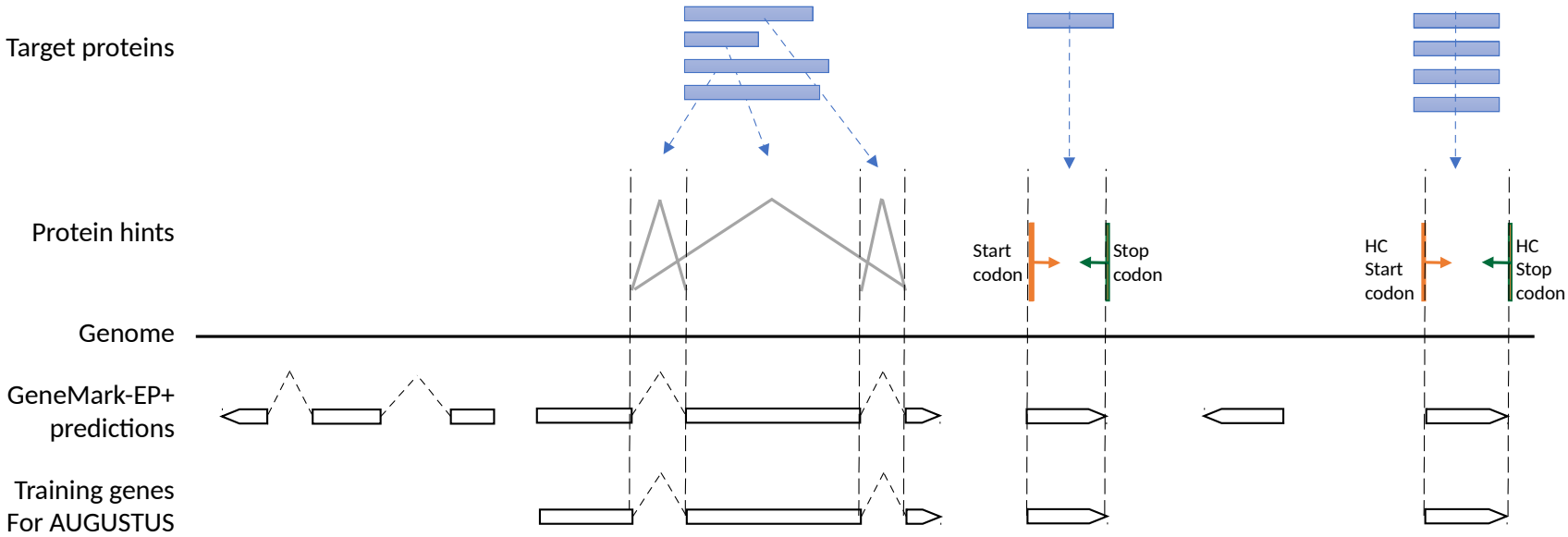
BRAKER2 with Proteins

Evidence Usage by GeneMark-EP+ & AUGUSTUS During Prediction



BRAKER2 with Proteins

Training Gene Selection for AUGUSTUS

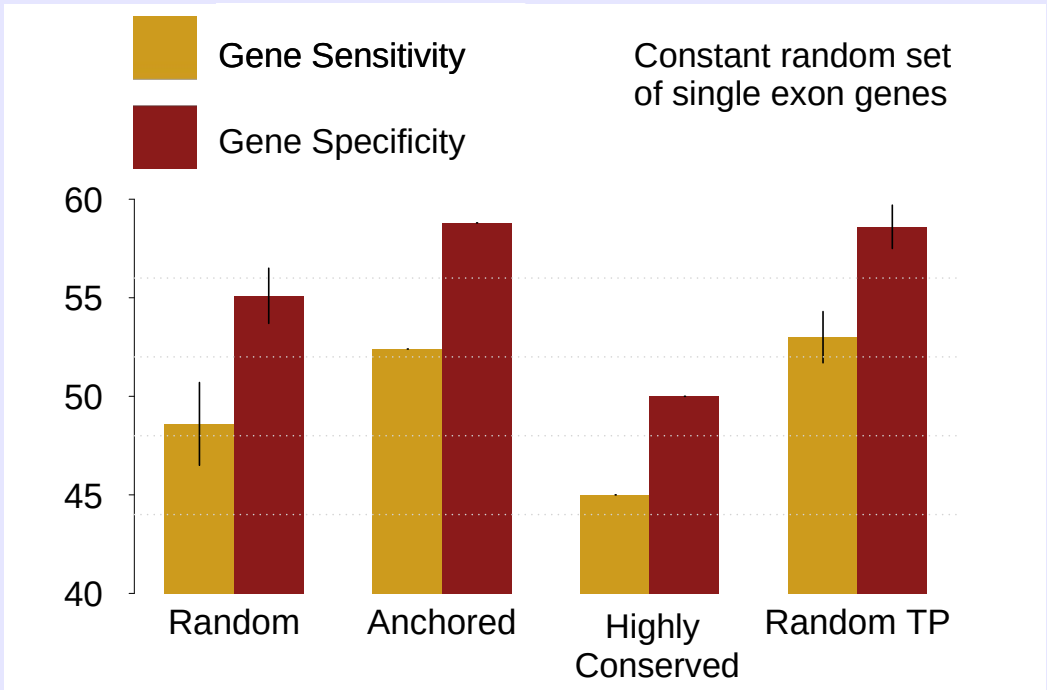


Katharina J. Hoff, Tomáš Brůna, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky

Evidence Based Multi Exon Training Gene Selection*

Effect on AUGUSTUS *ab initio* Gene Prediction Accuracy in BRAKER2

Drosophila melanogaster & OrthoDB Arthropoda (Phylum excl.)



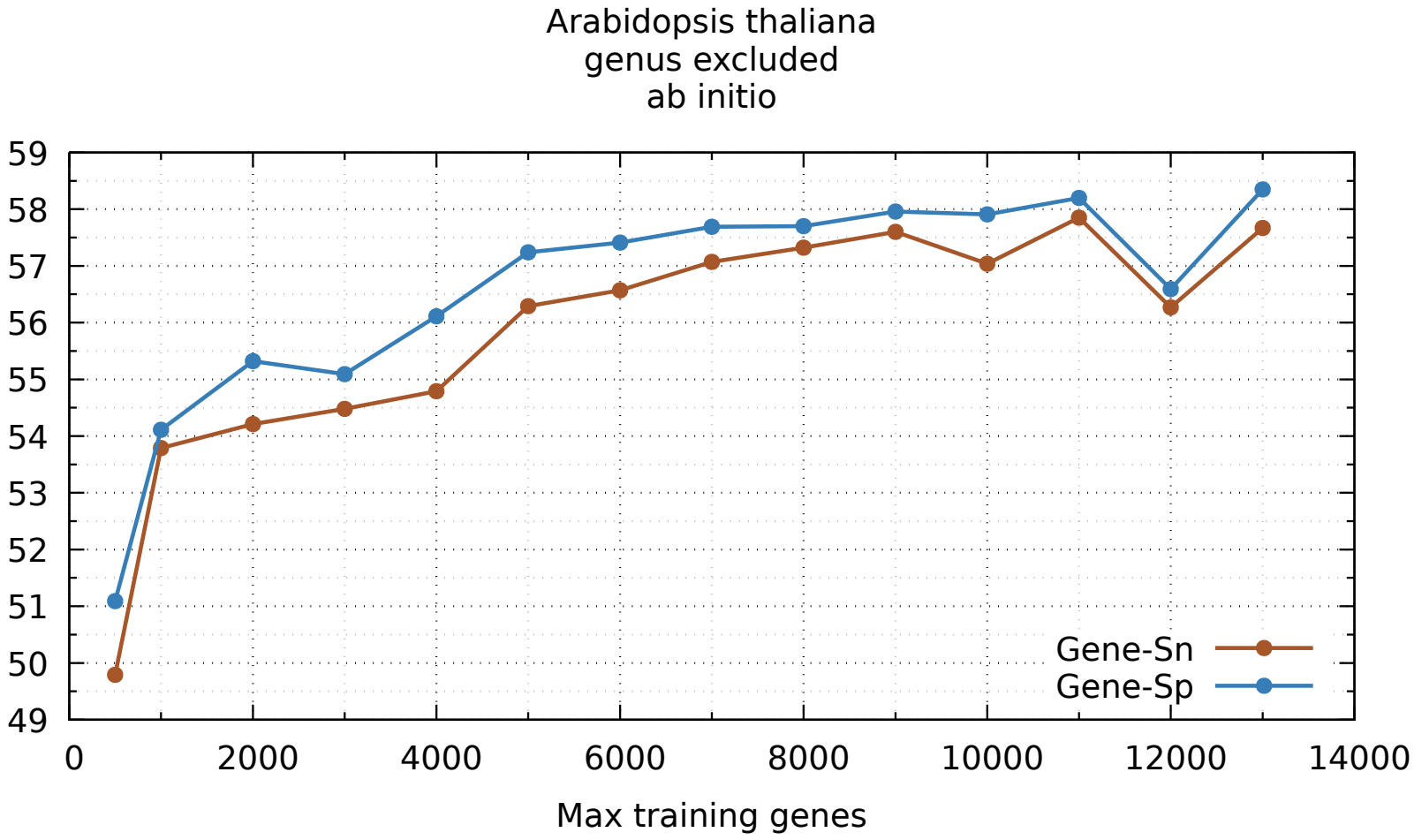
- Anchored genes and random TP $\rightarrow$ comparable accuracy
- Highly conserved genes are worse than random selection

$\Rightarrow$ BRAKER2 trains AUGUSTUS with anchored multi-exon genes

*) Single exon genes are anchored by starts and stops

Number of Training Genes

Effect on AUGUSTUS



⇒ AUGUSTUS benefits from high number of training genes

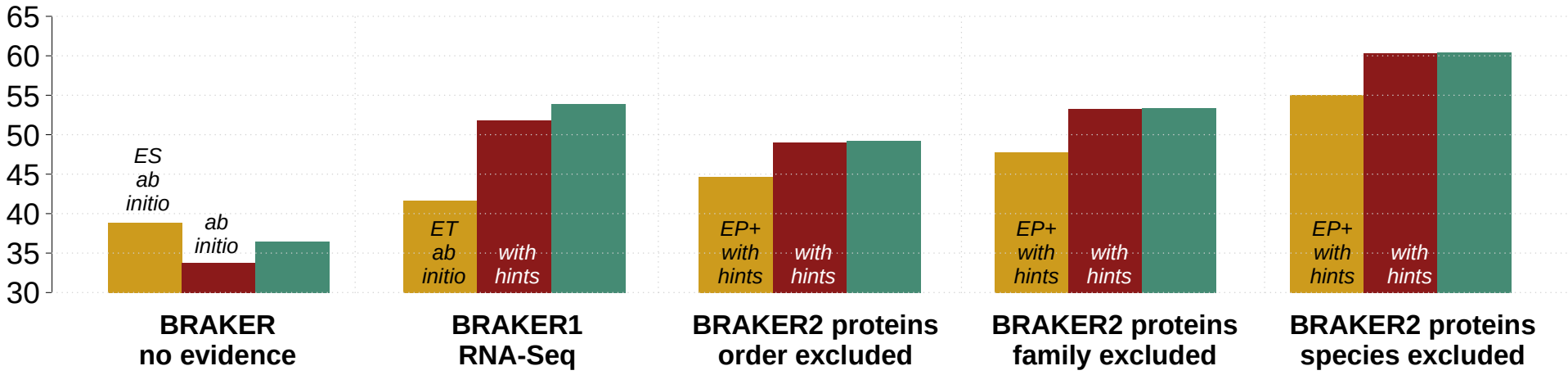
⇒ BRAKER uses up to 8000 training genes → speed

Transcript Prediction Accuracy F1

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

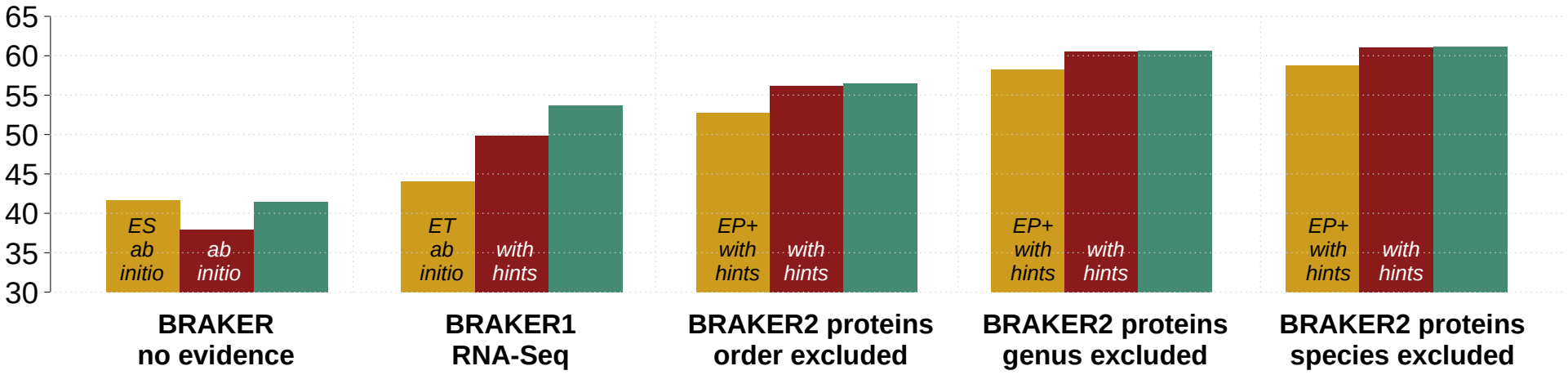
Drosophila melanogaster

- Fruit fly, genome size **130 MB**, RNAseq from Varus with Hisat2
- OrthoDB 10 **Arthropoda**



Arabidopsis thaliana

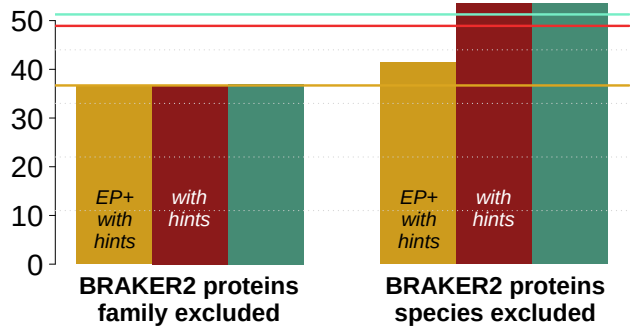
- Thale cress, genome size **151 MB**, RNAseq from Varus with Hisat2
- OrthoDB 10 **Plantae**



Transcript Prediction Accuracy F1

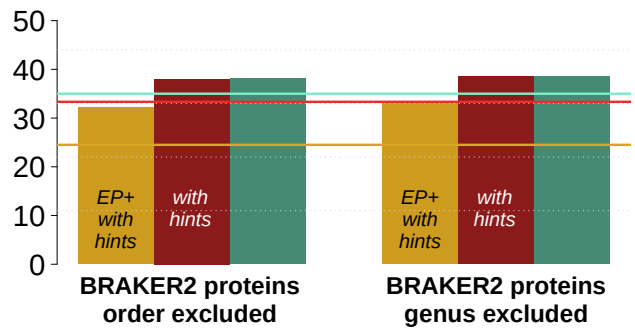
Caenorhabditis elegans

- Roundworm, genome size **84 MB**
- OrthoDB **Metazoa**



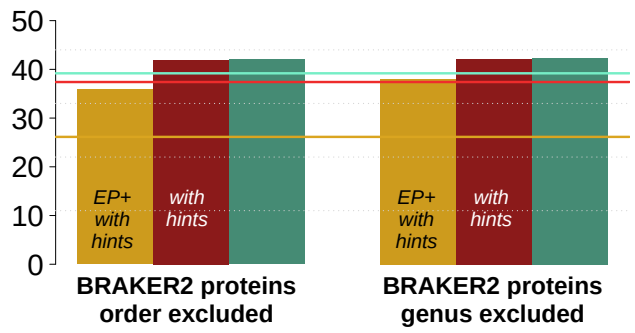
Fragaria vesca

- Wild strawberry, genome size **198 MB,**
- OrthoDB **Plantae**



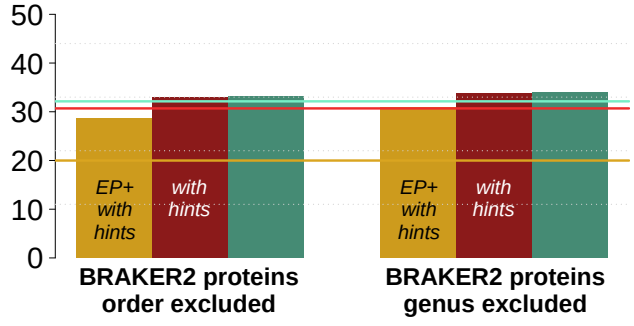
Medicago truncatula

- Barrel medic, genome size **807 MB,**
- OrthoDB **Plantae**



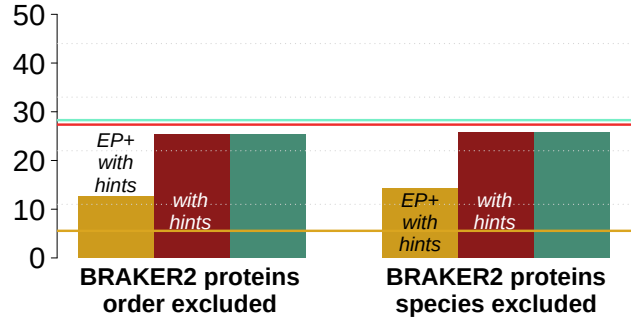
Solanum lycopersicum

- Tomato, genome size **1390 MB**
- OrthoDB **Plantae**



Danio rerio

- Zebrafish, genome size **1345 MB**
- OrthoDB **Chordata**:



Proteins from OrthoDB v10

Lines show accuracy of BRAKER1 with RNA-Seq from VARUS with Hisat2

GeneMark
AUGUSTUS
Union

Summary on BRAKER

- fully automatic pipeline
- state-of-the art accuracy
- fast
- easy to use
- BRAKER1: RNA-Seq spliced alignment information
- BRAKER2: + large scale protein sequence similarity

Ongoing Development

- BRAKER3 with RNA-Seq & protein
- UTR training & prediction
- protein family and intron profile integration (AUGUSTUS-PPX)

References

- Hoff KJ et al. (2016) “BRAKER1: unsupervised RNAseq-based genome annotation with GeneMark-ET and AUGUSTUS.”
- Hoff KJ et al. (2019) “Whole-genome annotation with BRAKER.“
- Lomsadze A et al. (2014) “Integration of mapped RNAseq reads into automatic training of eukaryotic gene finding algorithm.”
- Stanke M et al. (2008) “Using native and syntenically mapped cDNA alignments to improve de novo gene finding.”
- Brunå T et al. (2020) “GeneMark-EP and -EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins.”
- Lomsadze A et al. (2005) “Gene identification in novel eukaryotic genomes by self-training algorithm.”
- Buchfink B et al. (2015) “Fast and sensitive protein alignment using DIAMOND.”
- Iwata H & Gotoh O (2012) “Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features.”

BRAKER2 is Available for Download at

- `https://github.com/Gaius-Augustus/BRAKER`
- `https://github.com/gatech-genemark/BRAKER2`

Co-Authors

Alexandre Lomsazdze
Tomáš Brůna
Mario Stanke
Mark Borodovsky

Funding

This research is supported by US National Institutes of Health grant GM128145 to Mark Borodovsky and Mario Stanke.

Acknowledgements

Simone Lange
Anica Hoppe
Jens Keilwagen
Maria Hartmann
Ingo Bulla
Timon Kapischke
Holger Irrgang
Felix Becker
Matthis Ebel

BRAKER2 is Available for Download at

- `https://github.com/Gaius-Augustus/BRAKER`
- `https://github.com/gatech-genemark/BRAKER2`

Thank you for your attention!

BRAKER2 is Available for Download at

- `https://github.com/Gaius-Augustus/BRAKER`
- `https://github.com/gatech-genemark/BRAKER2`