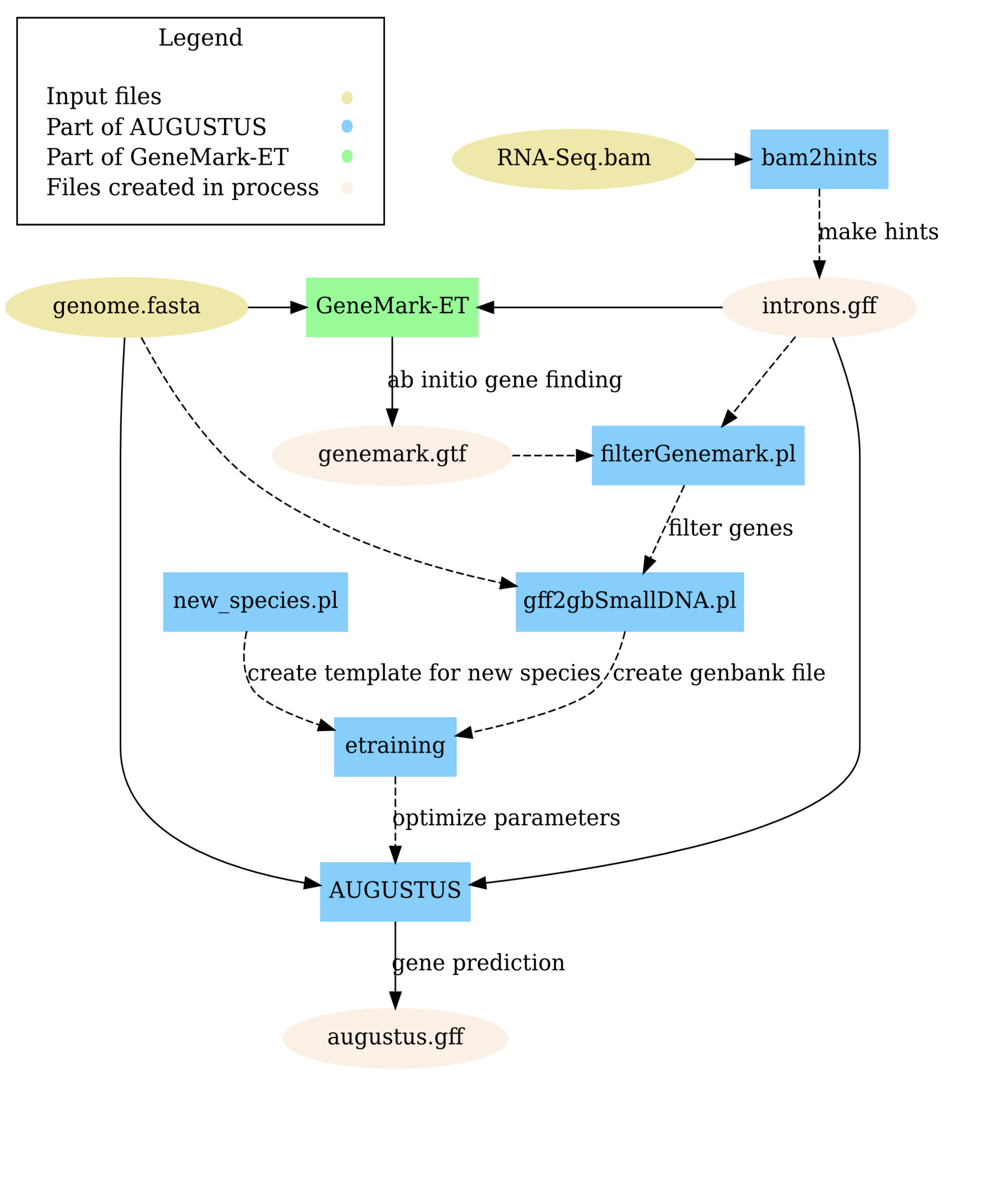
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

Simone Lange[1], Katharina Hoff[1], Alexandre Lomsadze[2], Mark Borodovsky[2], Mario Stanke[1]
1) University of Greifswald, GERMANY. 2) Georgia Institute of Technology, U.S.A. Contact: katharina.hoff@uni-greifswald.de

Abstract

Many genome sequencing projects are accompanied by transcriptome sequencing. The resulting RNA-Seq data is often assembled to aid structural genome annotation. However, as the RGASP [1] competition has shown the RNA-Seq assemblies contain errors and, as a result, the training of RNA-Seq-based gene finders can be involved and the prediction of protein-coding genes is still error prone. Therefore, there is a clear need for new easily applicable and accurate methods. Recently developed GeneMark-ET [2] is a gene prediction tool that incorporates unassembled RNA-Seq reads into unsupervised training and subsequently generates gene predictions as an *ab initio* gene prediction tool. AUGUSTUS [3] is a gene finder that usually requires supervised training; according to the RGASP results AUGUSTUS was one of the most accurate gene finders that uses RNA-Seq read information as extrinsic evidence in the prediction step. We saw a good potential in bypassing the RNA-Seq assembly step and developing a new method that would use mapped to genome RNA-Seq reads both in unsupervised automatic training and in gene prediction.
Here, we present **BRAKER1**, a pipeline for unsupervised RNA-Seq-based genome annotation that combines the advantages of GeneMark-ET and AUGUSTUS. BRAKER1 requires an RNA-Seq read alignment file (in bam format) and a genome file as input. First, GeneMark-ET performs iterative training and generates initial gene structures. Second, AUGUSTUS uses predicted genes for training and then integrates RNA-Seq read information as extrinsic evidence into final gene predictions. In our experiments we observed that BRAKER1 was more accurate than MAKER2 when it is using assembled RNA-Seq as sole source of extrinsic evidence. BRAKER1 does not require pre-trained parameters or a separate manually curated training step.
BRAKER1 is available for download at **http://bioinf.uni-greifswald.de/augustus/downloads/index.php** and **http://exon.gatech.edu/**.

BRAKER1 Pipeline



Running BRAKER1

```
braker.pl [OPTIONS]--genome=genome.fa --bam=rnaseq.bam
```

Test Data

C. elegans: genome and reference annotation version WS240 (wormbase)
RGASP RNA-Seq library
D. melanogaster: genome and reference annotation version R5 (flybase)
RGASP RNA-Seq library
A. thaliana: genome and reference annotation version TAIR 10
SRR934391
S. pombe: genome and reference annotation version ASM294v2.23 (pombase)
SRR097898, SRR097899, SRR097900, SRR097902, SRR097903, SRR097905, SRR097906, SRR097907, SRR097908, SRR097909, SRR097912, SRR097915, SRR097917, SRR097921, SRR097922, SRR097925, SRR402833

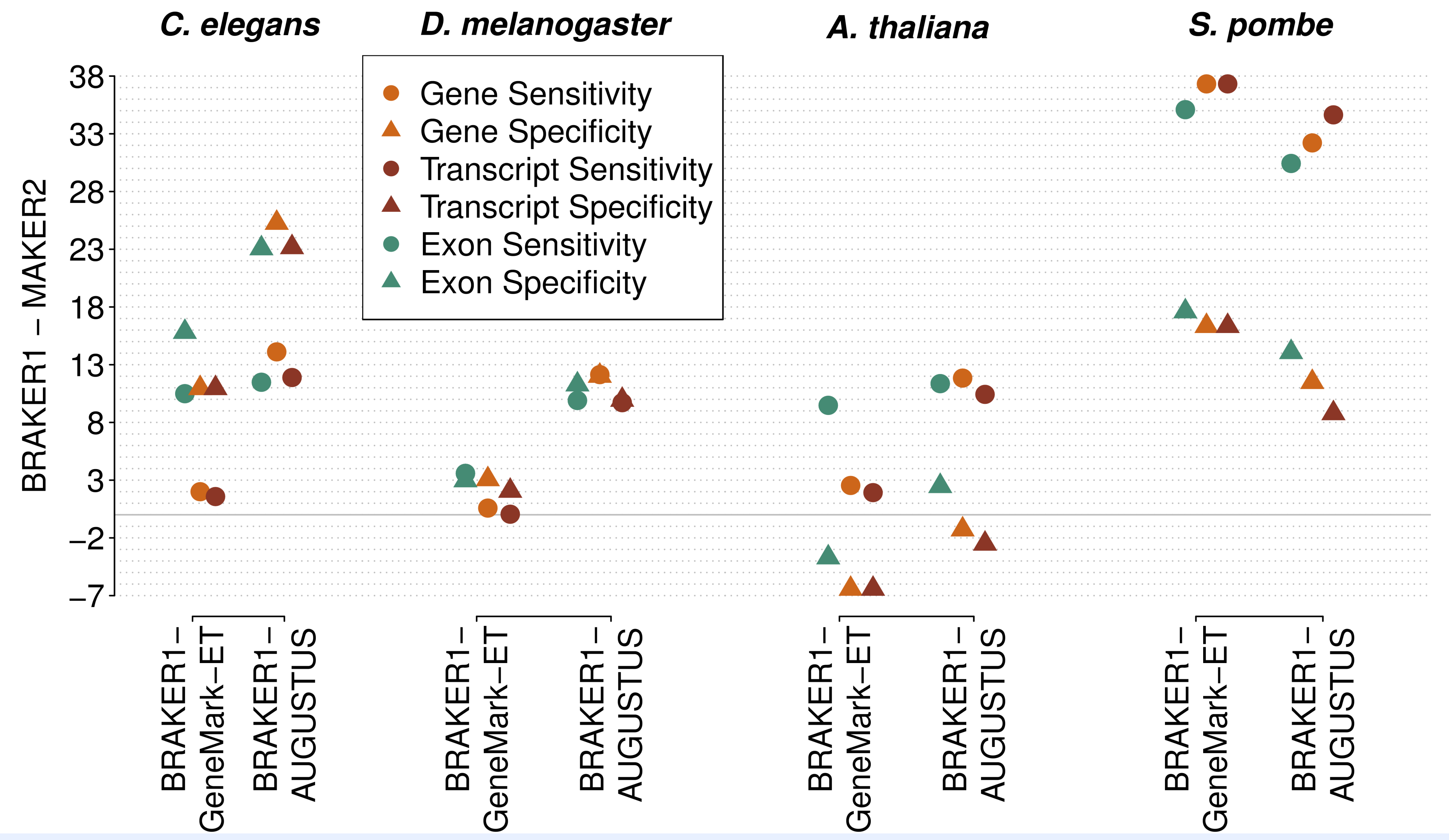
Accuracy in BRAKER1 GeneMark-ET and AUGUSTUS output

	C. elegans		*D. melanogaster*	
Level	BRAKER1-GeneMark-ET	BRAKER1-AUGUSTUS	BRAKER1-GeneMark-ET	BRAKER1-AUGUSTUS
Gene Sensitivity	43.0	55.1	58.5	70.2
Gene Specificity	41.7	56.1	49.9	59.0
Transcript Sensitivity	32.9	43.2	42.3	52.0
Transcript Specificity	41.7	54.0	49.9	57.8
Exon Sensitivity	79.9	80.9	68.5	75.1
Exon Specificity	78.2	85.4	57.9	66.2

	A. thaliana		*S. pombe*	
Level	BRAKER1-GeneMark-ET	BRAKER1-AUGUSTUS	BRAKER1-GeneMark-ET	BRAKER1-AUGUSTUS
Gene Sensitivity	53.9	63.2	80.0	77.3
Gene Specificity	46.1	51.3	84.9	81.2
Transcript Sensitivity	45.4	53.9	80.0	77.3
Transcript Specificity	46.1	50.0	84.9	77.4
Exon Sensitivity	81.1	83.0	85.2	84.2
Exon Specificity	72.4	78.5	89.0	82.6

Difference in accuracy parameters of BRAKER1 to MAKER2

"BRAKER1-GeneMark-ET" corresponds to genemark.gtf,
"BRAKER1-AUGUSTUS" corresponds to augustus.gff (top left illustration)



Gene finders AUGUSTUS, SNAP and GeneMark-ES were trained and MAKER2 was executed with RNA-Seq following the tutorial at
http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_GMOD_Online_Training_2014
We used no protein database, set the option keep_preds=1, included Cufflinks transcripts and Tophat2 read alignments, MAKER2 masked Repeats.

References

[1] T. Steijger, J.F. Abril, P.G. Engström, F. Kokocinski, The RGASP Consortium, T.J. Hubbard, R. Guigo, J. Harrow, P. Bertone (2013) "Assessment of transcript reconstruction methods for RNA-seq", *Nature Methods*, doi:10.1038/nmeth.271
[2] A. Lomsadze, P.D. Burns, M. Borodovsky (2014) "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm", *Nucleic Acids Research* doi: 10.1093/nar/gku557
[3] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler (2008) "Using native and syntenically mapped cDNA alignments to improve de novo gene finding", *Bioinformatics*, 24(5):637

Funding

This work is supported by the US National Institutes of Health grant HG000783.