The BRAKER3 Genome Annotation Pipeline

Lars Gabriel [1], Katharina J. Hoff [1,2], Tomáš Brůna [3], Alexandre Lomsadze [4], Mark Borodovsky [4,5], and Mario Stanke [1,2]

[1] Institute of Mathematics and Computer Science, University of Greifswald
[2] Center for Functional Genomics of Microbes, University of Greifswald
[3] US Department of Energy Joint Genome Institute, Berkeley
[4] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology
[5] School of Computational Science and Engineering, Georgia Institute of Technology

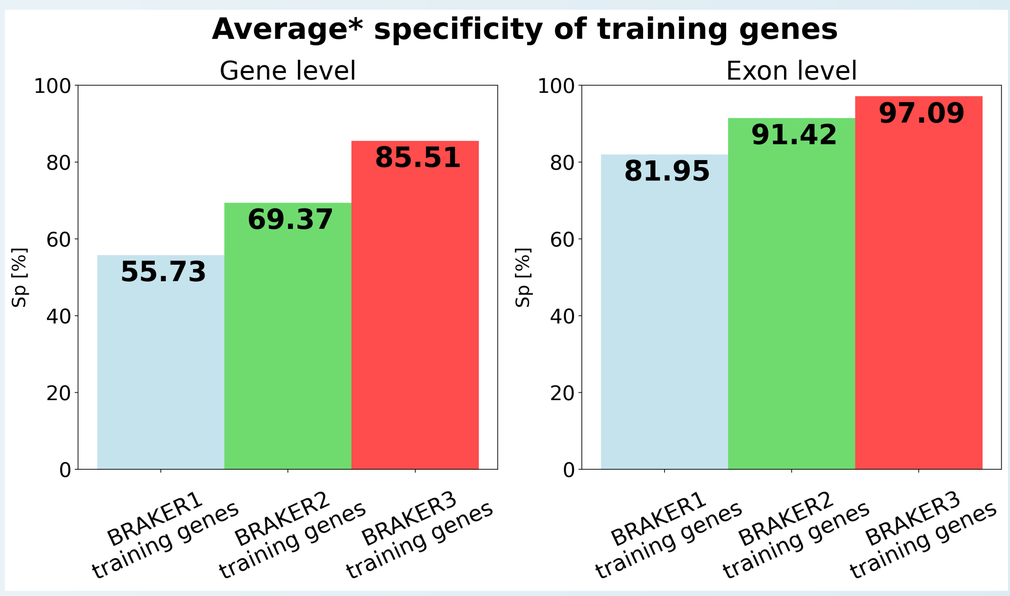
lars.gabriel@uni-greifswald.de
katharina.hoff@uni-greifswald.de
tbruna@lbl.gov
alexandre.lomsadze@bme.gatech.edu
borodovsky@gatech.edu
mario.stanke@uni-greifswald.de

Abstract

The increasing availability of databases that provide large amounts of extrinsic evidence in the form of protein sequences and RNA-Seq libraries provides powerful sources of information to improve methods for gene structure prediction of protein-coding genes. The BRAKER pipeline fully automates the annotation of novel eukaryotic genomes by utilizing the gene prediction tools GeneMark [1, 2] and AUGUSTUS [3] to provide an easy-to-use software tool. Previously published BRAKER pipelines offer a successful solution for genome annotation using either short read RNA-Seq (BRAKER1 [4]) or protein data (BRAKER2 [5]) alongside the intrinsic information of the nucleotide sequences.
We introduce BRAKER3, a novel pipeline in the BRAKER suite that enables the use of short read RNA-Seq together with a large protein database. It integrates the novel GeneMark-ETP tool and the BRAKER-related combiner software TSEBRA [6] into its annotation protocol. We showed on six species that BRAKER3 increases the annotation accuracy significantly compared to its predecessors, especially for large and complex genomes. In addition, we automated the genome annotation workflow further by adding more preprocessing steps for short read RNA-Seq and made BRAKER easier accessible by providing a Singularity container.

Training genes

A set of training genes is inferred directly from intrinsic and extrinsic evidence by GeneMark-ETP by assembling the short reads and subsequently predicting and filtering gene structures in these transcript sequences.



* Average over all six species listed in 'Results - Gene Prediction Accuracy'.

Usage

BRAKER3 can be run via a command line, e.g. with:

```
braker.pl --genome=genome.fa --prot_seq=proteins.fa \
      --rnaseq_sets_ids=SRA_ID1,SRA_ID2 \
      --rnaseq_sets_dirs=/path/to/RNASeq/
```

BRAKER3 uses aligned (`BAM`) or unaligned (`FASTQ`) RNA-Seq libraries named after their IDs* (here `SRA_ID1,SRA_ID2`) that are located at the specified directory (here `/path/to/RNASeq/`). Libraries for which local files are not provided are downloaded from SRA.

* The RNA-Seq IDs of local files do not have to match to libraries available at SRA.

Availability

GitHub:

```
https://github.com/Gaius-Augustus/BRAKER
```

Singularity:
Installing and Running BRAKER via Singularity*:

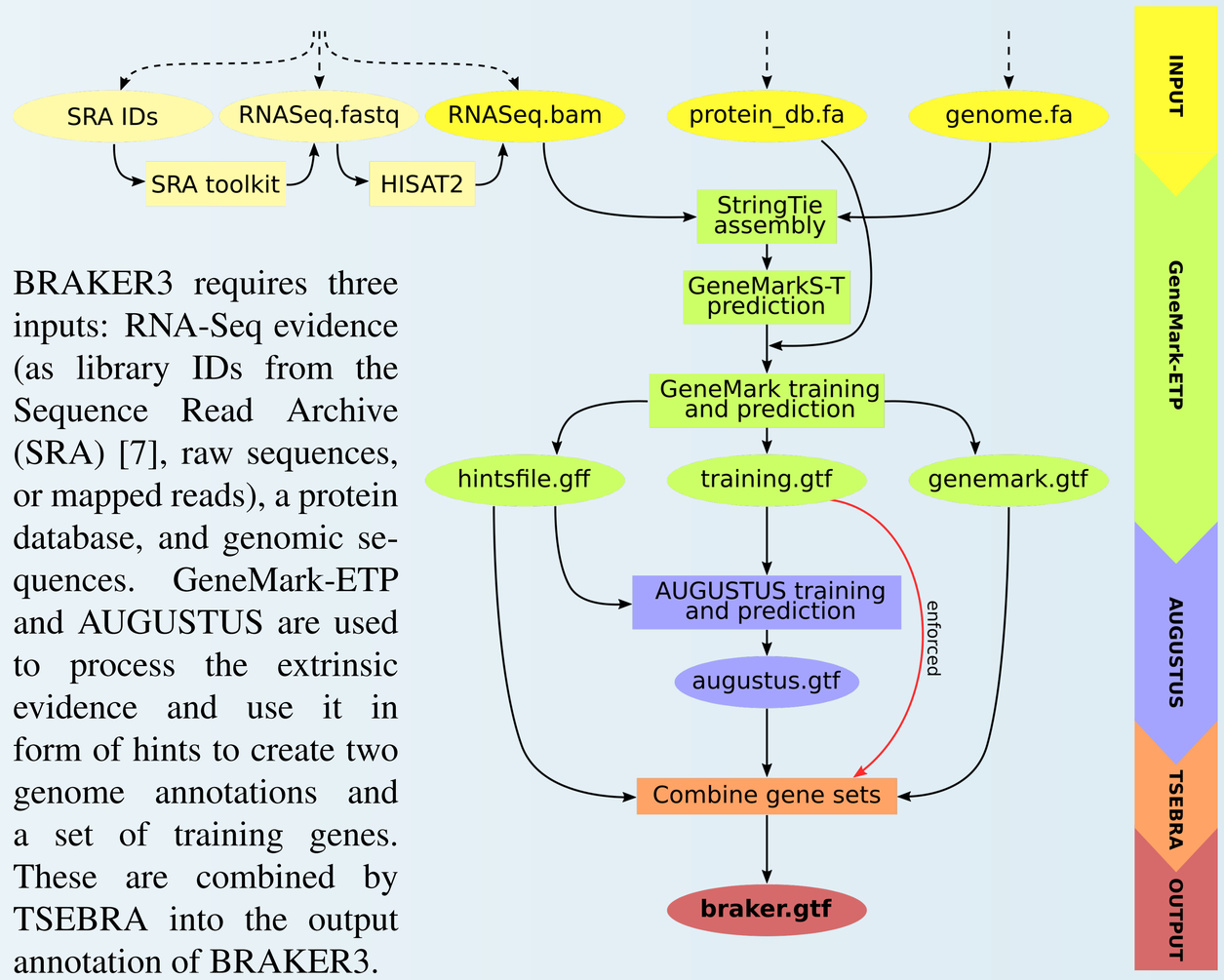
```
singularity build braker3.sif
       docker://teambraker/braker3:latest
singularity exec braker3.sif print_braker3_setup.py
singularity exec braker3.sif braker.pl [OPTIONS]
```

* At this point in time, GeneMark-ETP is not part of the container, yet. It needs to be installed following the printed setup instructions.

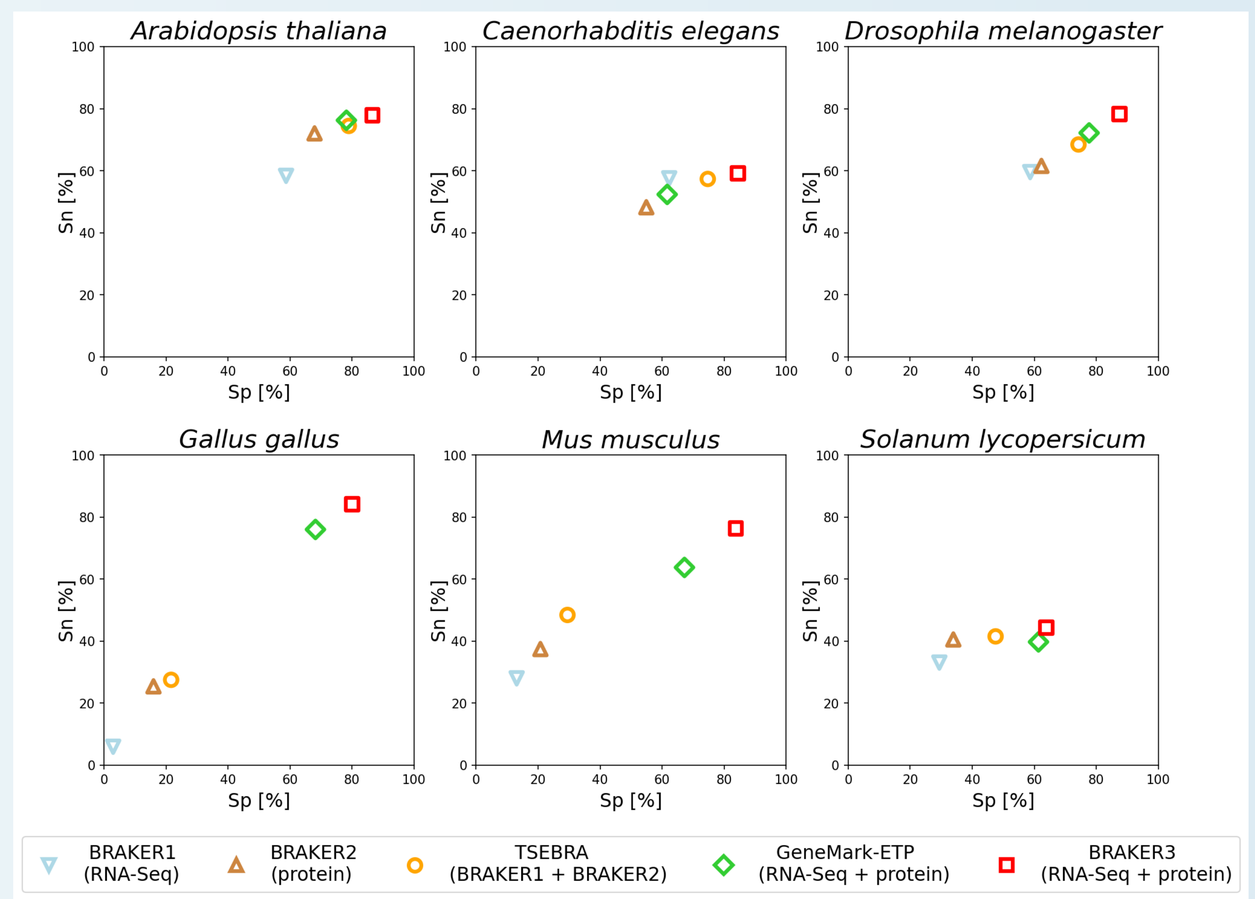
Funding: This research is supported by US National Institutes of Health grant GM128145

BRAKER3: Workflow

BRAKER3 requires three inputs: RNA-Seq evidence (as library IDs from the Sequence Read Archive (SRA) [7], raw sequences, or mapped reads), a protein database, and genomic sequences. GeneMark-ETP and AUGUSTUS are used to process the extrinsic evidence and use it in form of hints to create two genome annotations and a set of training genes. These are combined by TSEBRA into the output annotation of BRAKER3.



Results - Gene Prediction Accuracy



Gene level accuracy of genome annotations using short read RNA-Seq and large databases of protein sequences from which species of the same taxonomic order were removed.

Species	Size (Mb)	# Genes
Arabidopsis thaliana	119	27,444
Caenorhabditis elegans	100	20,172
Drosophila melanogaster	137	13,928
Gallus gallus	1,040	17,279
Mus musculus	2,650	22,378
Solanum lycopersicum	772	33,562

References

[1] Alexandre Lomsadze, Paul D Burns, and Mark Borodovsky. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm". In: *Nucleic Acids Research* 42.15 (2014), e119–e119.

[2] Tomáš Brůna, Alexandre Lomsadze, and Mark Borodovsky. "GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins". In: *NAR Genomics and Bioinformatics* 2.2 (2020), lqaa026.

[3] 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.

[4] Katharina J Hoff et al. "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS". In: *Bioinformatics* 32.5 (2016), pp. 767–769.

[5] 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.

[6] Lars Gabriel et al. "TSEBRA: transcript selector for BRAKER". In: *BMC Bioinformatics* 22.1 (2021), pp. 1–12.

[7] Rasko Leinonen et al. "The sequence read archive". In: *Nucleic Acids Research* 39.suppl_1 (2010), pp. D19–D21.