The BRAKER3 Genome Annotation Pipeline

Plant and Animal Genomes XXX
January 15th 2023

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

Presenting author e-mail: lars.gabriel@uni-greifswald.de

BRAKER - Genome Annotation

Task

- Find locations of protein-coding genes
- Predict their gene structure

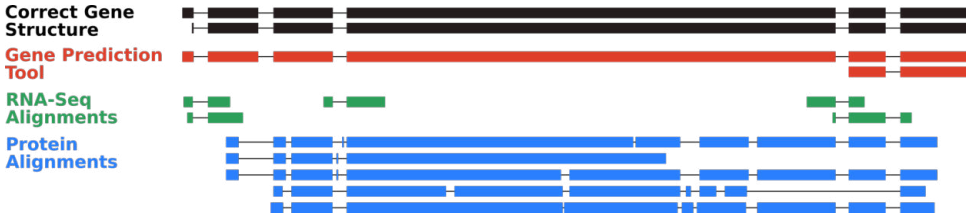
Evidence

Intrinsic: Nucleotide sequence:

- Predict gene structures *ab initio* using statistical models

Extrinsic: RNA-Seq reads, homologous proteins:

- E.g. infer exon borders from spliced alignments



BRAKER - Overview

BRAKER

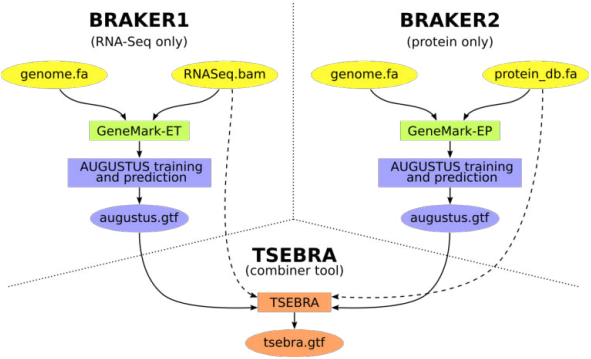
- Automated annotation of novel eukaryotic genomes
- Integrates extrinsic evidence, i.e. short read RNA-Seq, proteins

Gene Prediction Tools

- **GeneMark**: suite of self-training tools
- **AUGUSTUS**: highly accurate tool that requires a training gene set

TSEBRA

- Combiner tool for BRAKER predictions



BRAKER1: Hoff et al. 2016. *Bioinformatics*. 32(5):767–9.
BRAKER2: Brůna, Hoff et al. 2021. *NAR Genomics and Bioinform*. 3(1):lqaa108.
TSEBRA: Gabriel et al. 2021. *BMC Bioinformatics*. 22: 566.

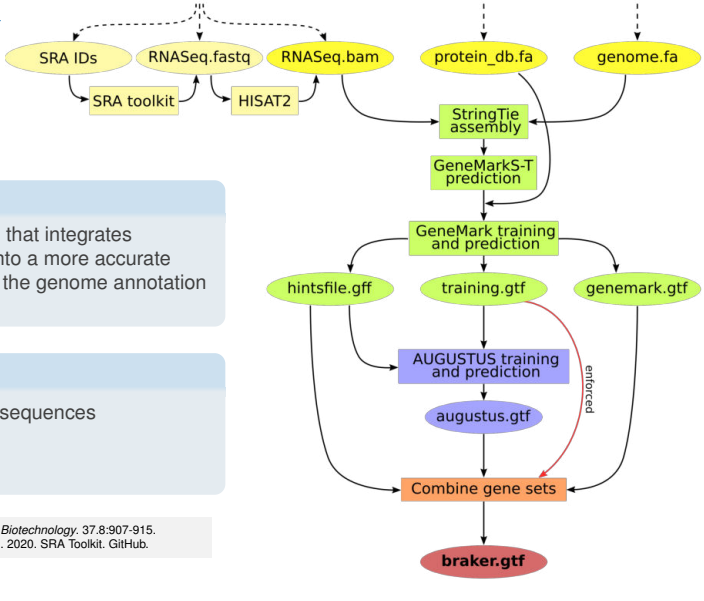
BRAKER3 - Workflow

Task

Develop a BRAKER pipeline that integrates RNA-Seq and protein data into a more accurate prediction, while automating the genome annotation process even further.

Input

- Softmasked genomic sequences
- Protein database
- Short read RNA-Seq



HISAT2: Kim, Daehwan, et al. 2019. *Nature Biotechnology*. 37.8:907-915.
SRA toolkit: SRA Toolkit Development Team. 2020. SRA Toolkit. GitHub.

BRAKER3 - Experiments

Experiments

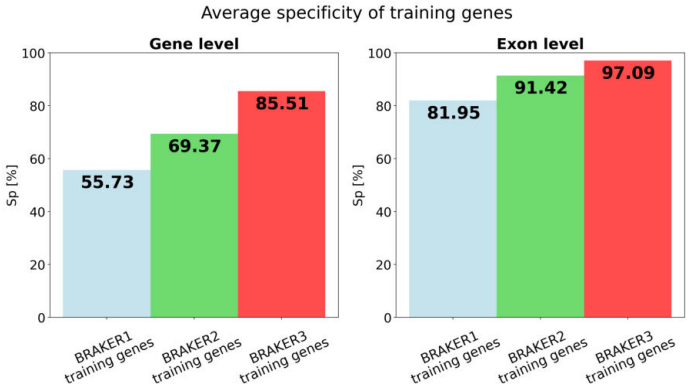
Accuracy assessment using genome-wide predictions of 6 species:

Species	Genome Size (Mb)	# Genes in Annotation
Arabidopsis thaliana (thale cress)	119	27,444
Caenorhabditis elegans (nematode)	100	20,172
Drosophila melanogaster (fruit fly)	137	13,928
Gallus gallus (chicken)	1,040	17,279
Mus musculus (mouse)	2,650	22,378
Solanum lycopersicum (tomato)	772	33,562

Accuracy metrics

Specificity [Sp]: Percentage of correctly found genes/transcripts/exons in the **predicted gene set**.
Sensitivity [Sn]: Percentage of correctly found genes/transcripts/exons in the **reference annotation**.

BRAKER3 - Training Genes

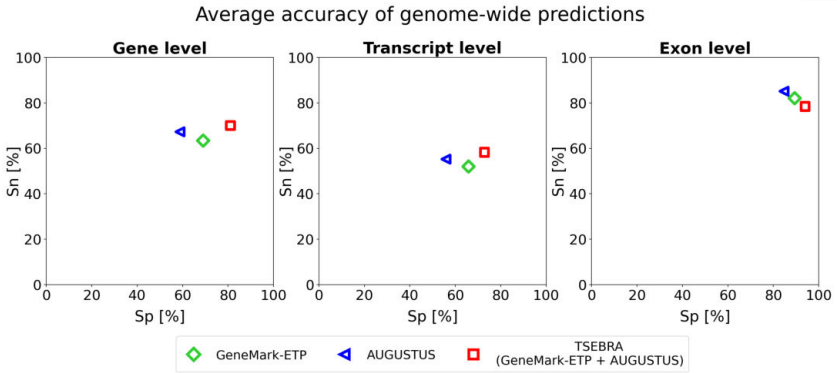


High-Confidents (HC) training genes

- Inferred from assembled short read RNA-Seq
- Used to train AUGUSTUS and GeneMark gene models

Species: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Gallus gallus*, *Mus musculus*, and *Solanum lycopersicum*

BRAKER3 - TSEBRA

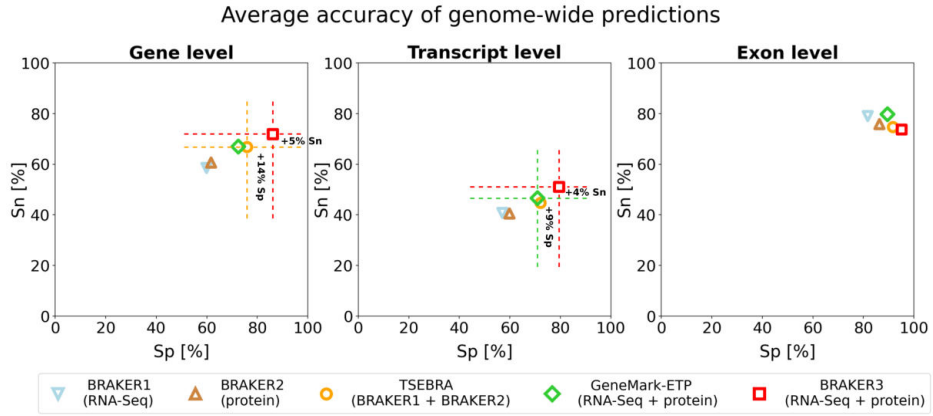


TSEBRA: Transcript selector for BRAKER

- Combines GeneMark-ETP and AUGUSTUS gene sets
- Uses the extrinsic evidence to compare and filter transcripts
- Enforces training genes

Species: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Gallus gallus*, *Mus musculus*, and *Solanum lycopersicum*

BRAKER3 Accuracy

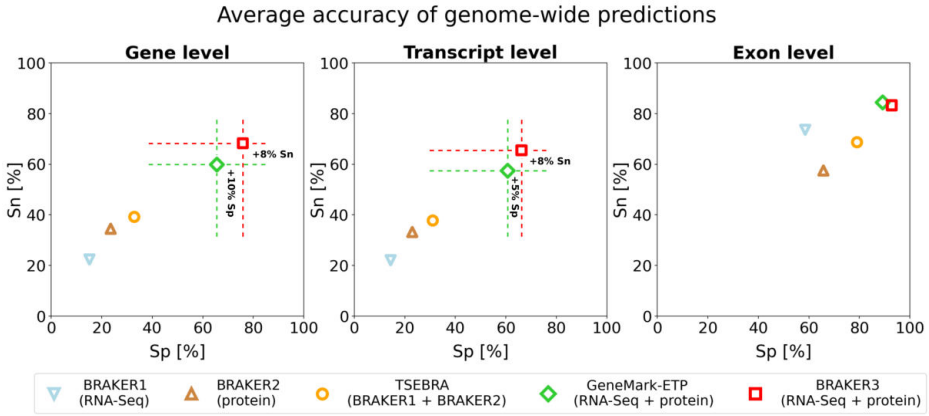


Species: *Arabidopsis thaliana*
Caenorhabditis elegans
Drosophila melanogaster

Extrinsic evidence:

- Paired RNA-Seq short reads
- OrthoDB protein database **(order excluded)**

BRAKER3 Accuracy



Species: *Gallus gallus*
Mus musculus
Solanum lycopersicum

Extrinsic evidence:

- Paired RNA-Seq short reads
- OrthoDB protein database **(order excluded)**

BRAKER3 - Usage and Runtime

Usage

Command line:

```
braker.pl --genome=genome.fa --prot_seq=protein_db.fa \
          --rnaseq_sets_ids=RNA_ID1,RNA_ID2 \
          --rnaseq_sets_dirs=/path/to/RNASeq/
```

Runtime

- Average runtime for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Gallus gallus*, *Mus musculus*, *Solanum lycopersicum*.
- Runtime on a 48 core cluster:

	BRAKER1	**BRAKER2**	**GM-ETP+**	**BRAKER3**
Runtime [h]	06:26	09:01	06:03	17:55

BRAKER3 - Availability

BRAKER3

GitHub:

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

Singularity:

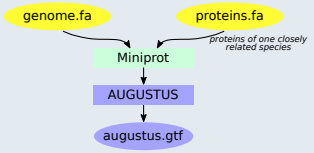
```
singularity build braker3.sif docker://teambraker/braker3:latest
singularity exec braker3.sif braker.pl [OPTIONS]
```

GeneMark-ETP

```
https://topaz.gatech.edu/GeneMark/etp.for_braker.tar.gz
```

Other Projects: GALBA

Workflow

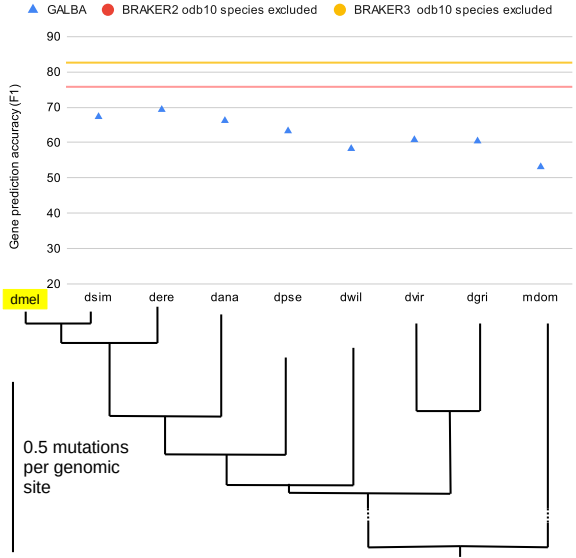


Availability

GitHub:
`https://github.com/Gaius-Augustus/GALBA`
Singularity:
`docker://katharinahoff/galba-notebook:latest`

Miniprot: Li, Heng. 2022. *arXiv preprint.* arXiv:2210.08052.

GALBA with miniprot in Drosophila melanogaster



Summary

BRAKER - Fully automated genome annotation pipeline

Depending on available extrinsic evidence, executes one of three pipelines:

	short read RNA-Seq	protein database
BRAKER1	X	
BRAKER2		X
BRAKER3	X	X

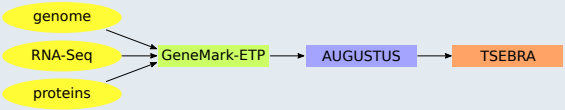
Availability

- **GitHub:** `https://github.com/Gaius-Augustus/BRAKER/`
- **Singularity**: `docker://teambraker/braker3:latest`

Poster numbers: **PE0922**, PE0162

BRAKER3

- Much higher accuracy than previous BRAKER predictions, especially for large and complex genomes.
- Adds automated download and alignment of RNA-Seq libraries.



Acknowledgements

Funding

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

Co-Authors

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

Availability

- BRAKER: `https://github.com/Gaius-Augustus/BRAKER/`
- GeneMark-ETP: `https://topaz.gatech.edu/GeneMark/etp.for_braker.tar.gz`
- AUGUSTUS: `https://github.com/Gaius-Augustus/Augustus`
- TSEBRA: `https://github.com/Gaius-Augustus/TSEBRA`
- GALBA: `https://github.com/Gaius-Augustus/Galba`

References

Lomsadze et al. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm." Nucleic acids research 42.15 (2014): e119-e119.

Brůna et al. "GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins." NAR genomics and bioinformatics 2.2 (2020): lqaa026.

Stanke et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics 24.5 (2008): 637-644.

Hoff et al. "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS." Bioinformatics 32.5 (2016): 767-769.

Brůna et al. "BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database." NAR genomics and bioinformatics 3.1 (2021): lqaa108.

Gabriel et al. "TSEBRA: Transcript Selector for BRAKER." BMC Bioinformatics 22: 566 (2021).

Tang et al. "Identification of protein coding regions in RNA transcripts." Nucleic acids research 43.12 (2015): e78-e78.

Wöhner et al. "The draft chromosome-level genome assembly of tetraploid ground cherry (Prunus fruticosa Pall.) from long reads." Genomics 113.6 (2021):4173-4183.

Kim, Daehwan, et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype." Nature Biotechnology 37.8 (2019): 907-915.

SRA Toolkit Development Team. "SRA Toolkit". (2020): `https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit`.

Li, Heng. "Protein-to-genome alignment with miniprot." arXiv preprint arXiv:2210.08052 (2022).