TSEBRA

Integrating various extrinsic evidence sources for more accurate genome annotation with BRAKER

Plant and Animal Genomes XXIX
January 9th 2022

Lars Gabriel
Katharina J. Hoff
Hannah Thierfeldt
Tomáš Brůna
Mark Borodovsky
Mario Stanke

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

Overview

1. BRAKER: Automated Whole-Genome Annotation

2. TSEBRA: Transcript Selector for BRAKER

3. Long-Read Integration Using GeneMarkS-T and TSEBRA

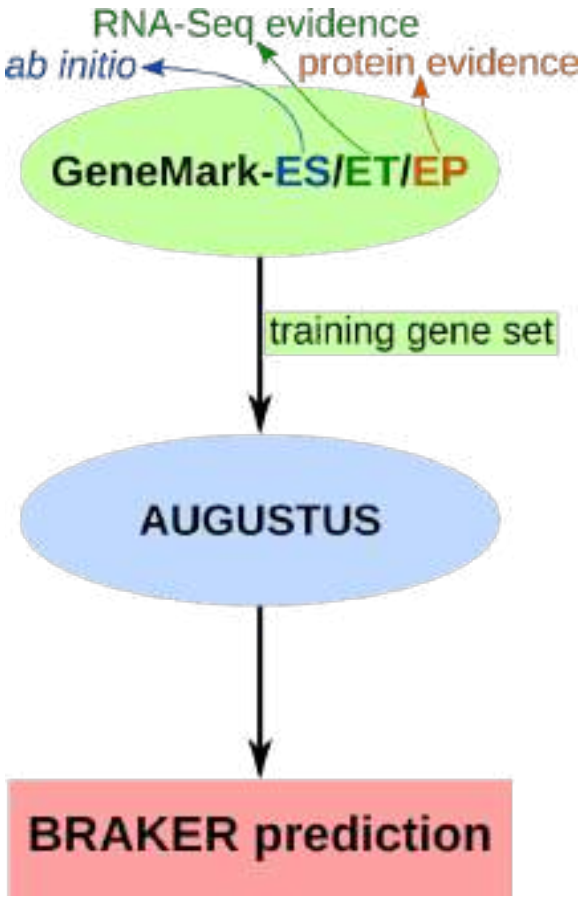
BRAKER - Overview

BRAKER

- Automated genome annotation of novel eukaryotic genomes
- Integrates extrinsic evidence from short-read RNA-Seq or homologous proteins

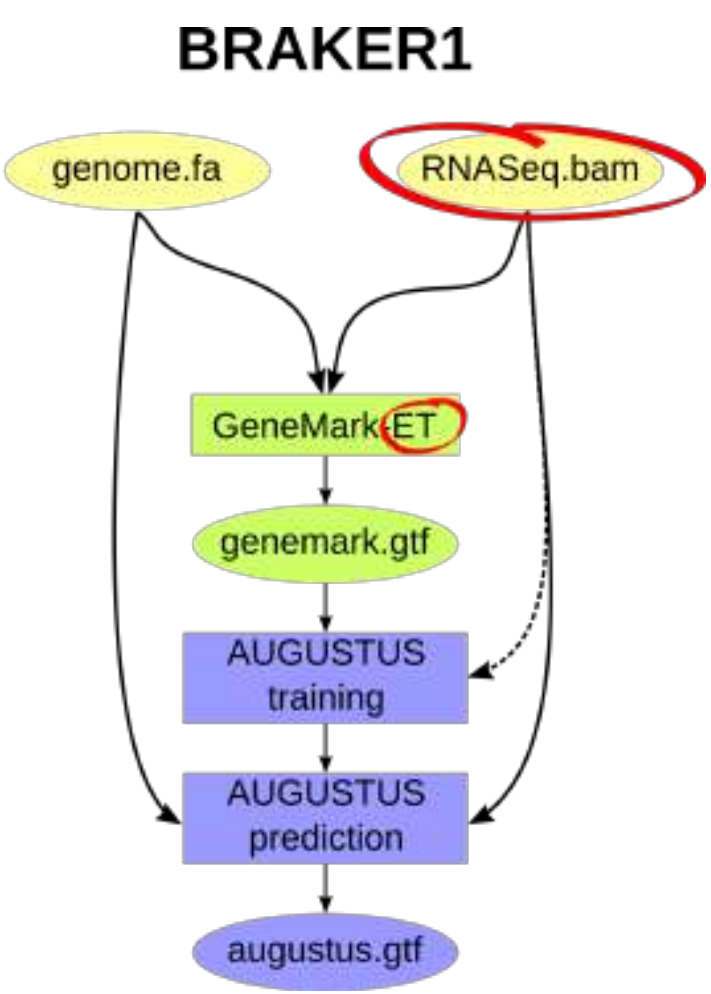
Gene Prediction Tools

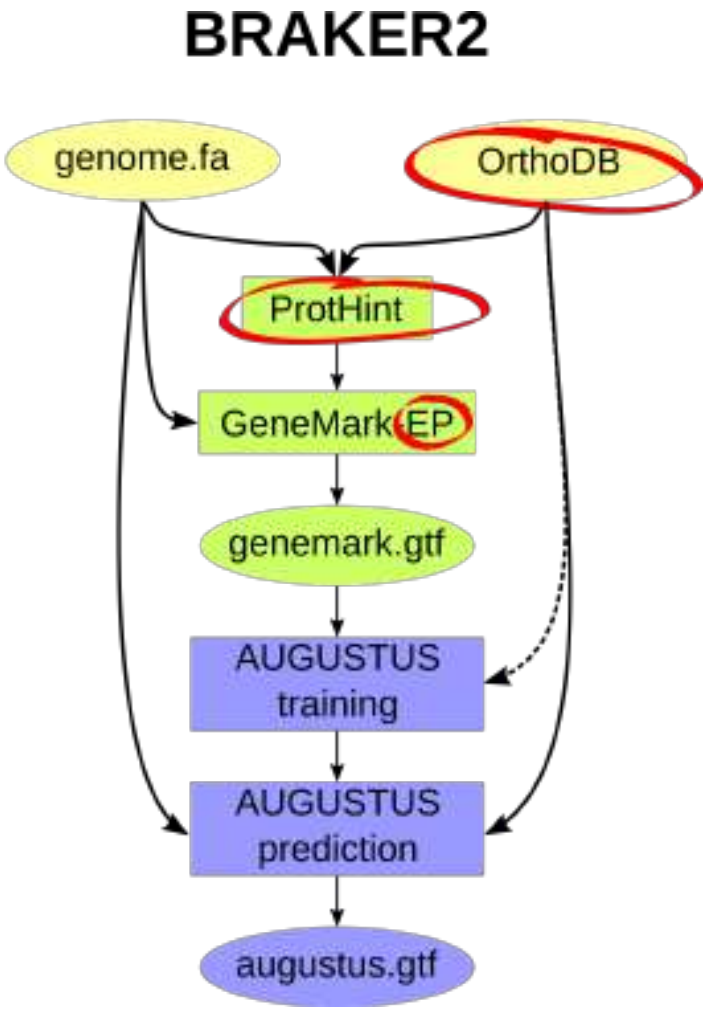
- **GeneMark**: suite of self-training methods
- **AUGUSTUS**: highly accurate method that requires a training gene set



Hoff et al. 2016. *Bioinformatics*. 32(5):767–9.
Brůna, Hoff et al. 2021. *NAR Genomics and Bioinform.* 3(1):lqaa108.

BRAKER - Genome Annotation Protocols

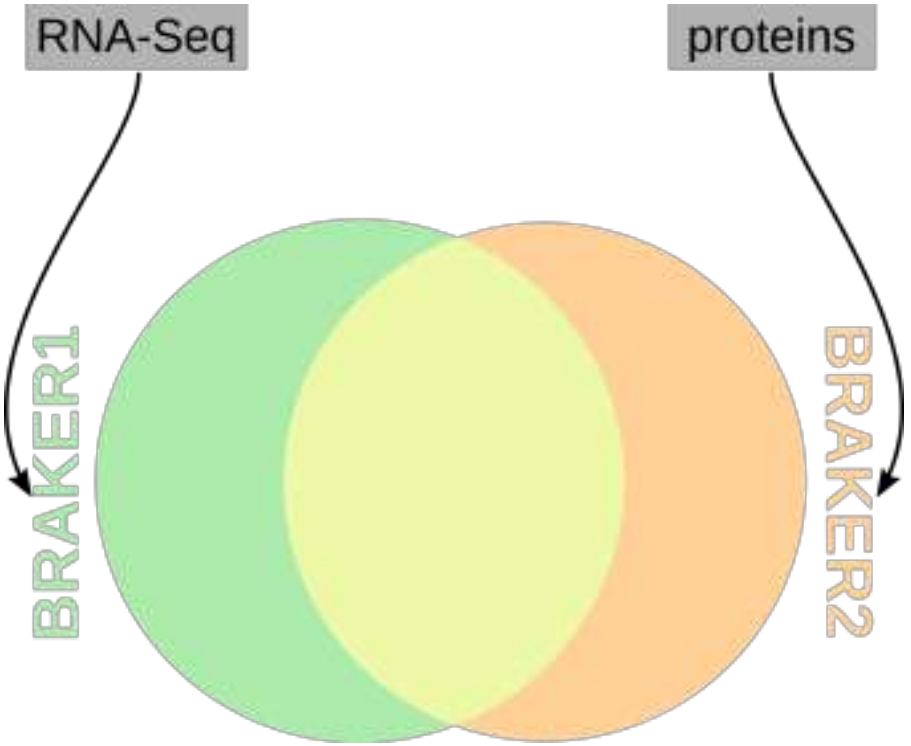




TSEBRA - Combiner for BRAKER

Goal

- BRAKER annotation with RNA-Seq and protein evidence
- Achieve higher accuracy than BRAKER1 and BRAKER2

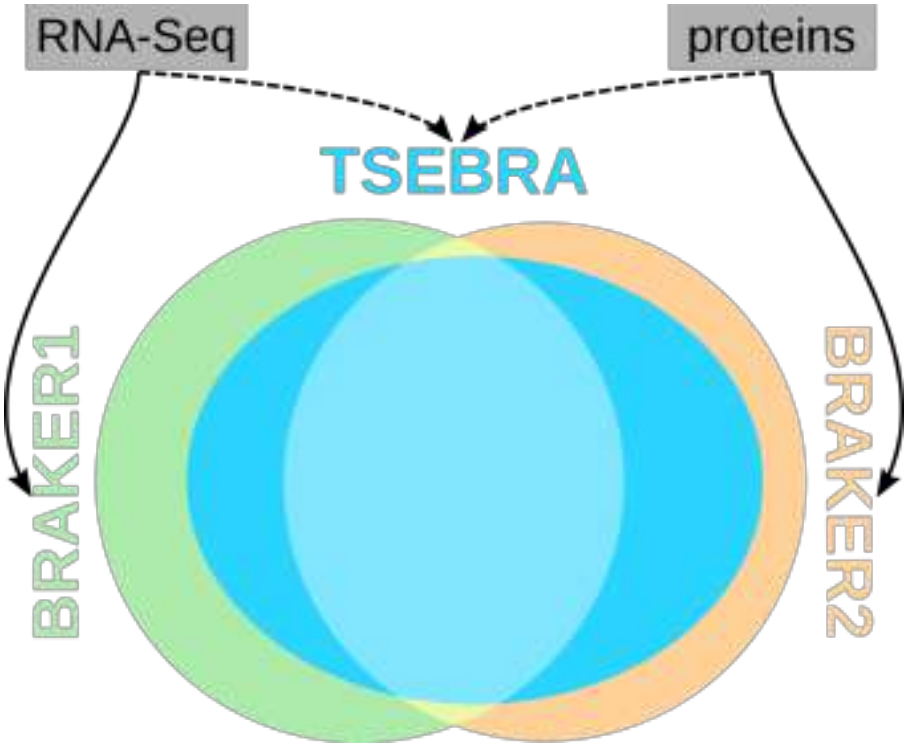


Transcript set comparison for different gene predictions.

TSEBRA - Combiner for BRAKER

Goal

- BRAKER annotation with RNA-Seq and protein evidence
- Achieve higher accuracy than BRAKER1 and BRAKER2

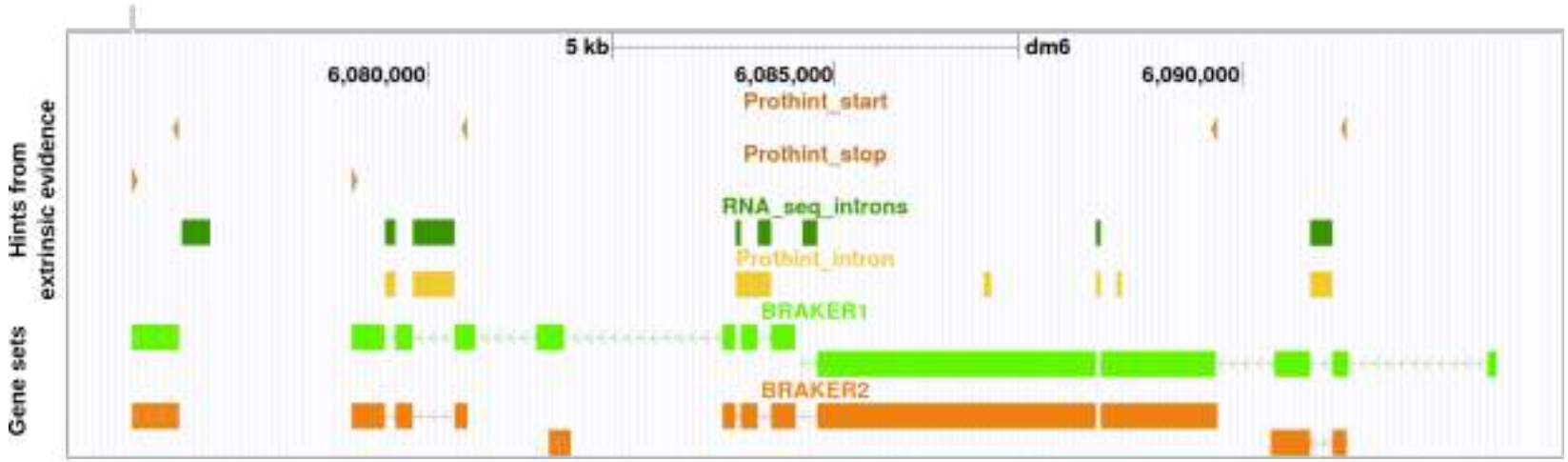


Transcript set comparison for different gene predictions.

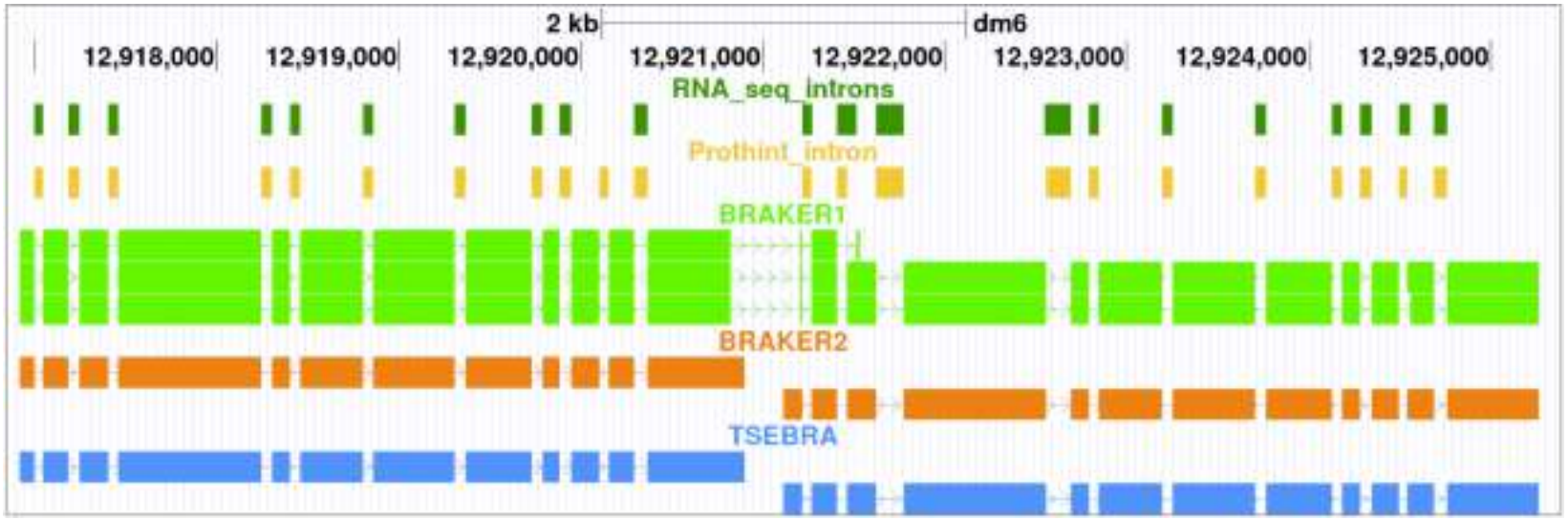
TSEBRA - Workflow

Input

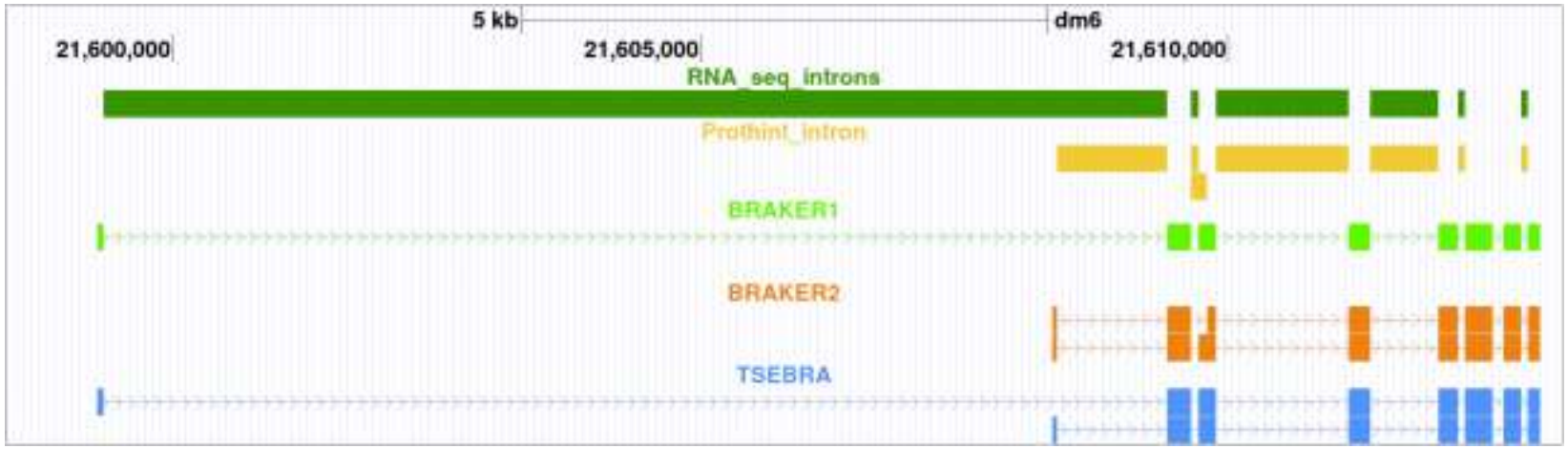
- Gene sets
- Hints from extrinsic evidence
- Configuration



TSEBRA - Workflow

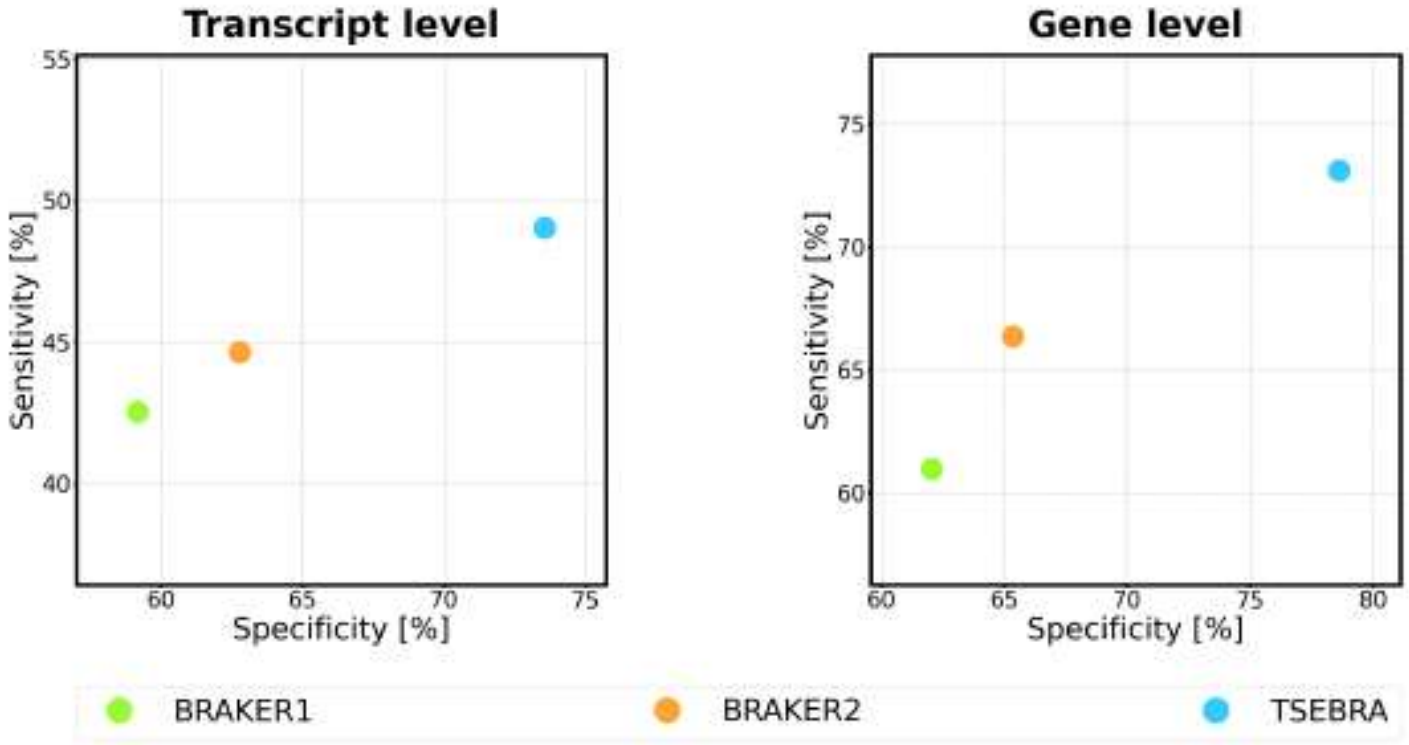


TSEBRA - Workflow



TSEBRA - Prediction Accuracy

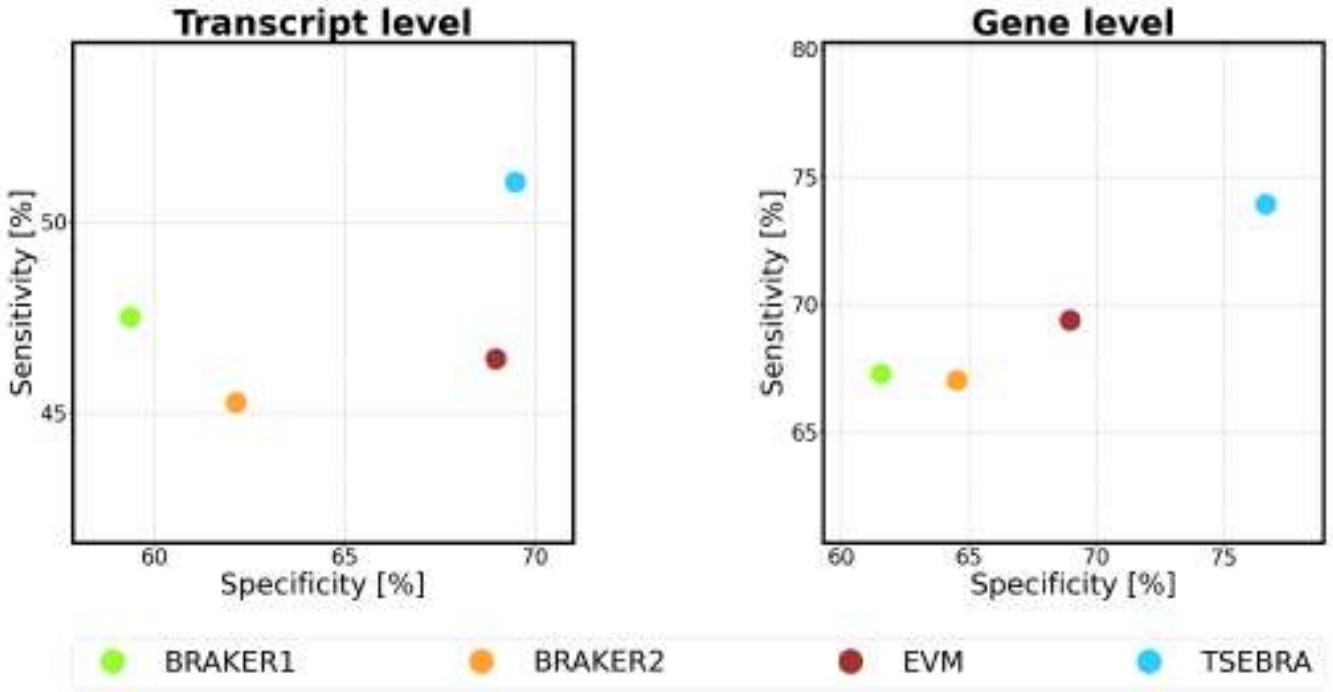
Average prediction accuracy for *A. thal.*, *C. ele.* and *D. mel.*



Gabriel et al. 2021. *BMC Bioinformatics* 22, 566.

TSEBRA - Comparison to EVidenceModeler

Average prediction accuracy for *A. thal.*, *C. ele.* and *D. mel.*

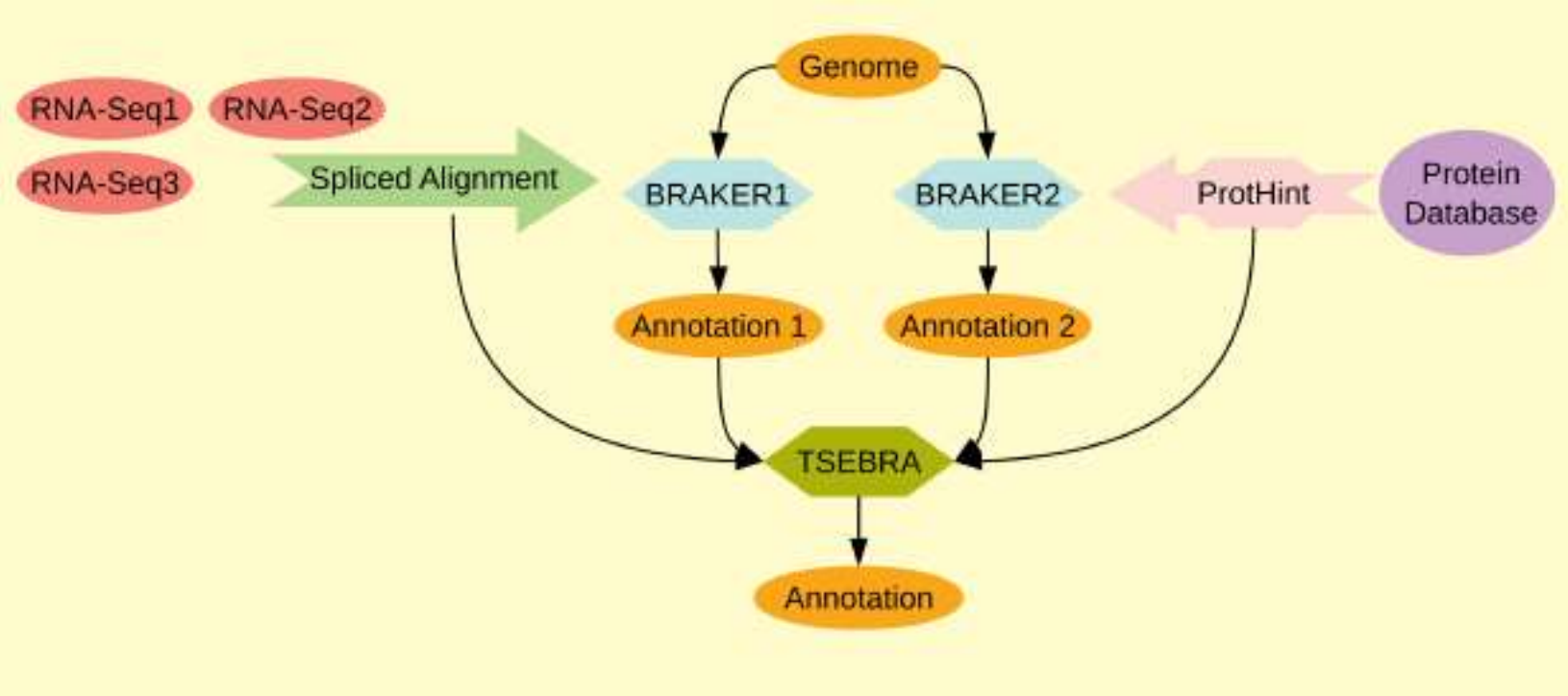


Average runtime

EVM : 35.28 min
TSEBRA : 0.37 min
* executed in parallel on 28 multi-core processor

Haas et al. 2008. *Genome biology* 9.1: 1-22.
Gabriel et al. 2021. *BMC Bioinformatics* 22, 566.

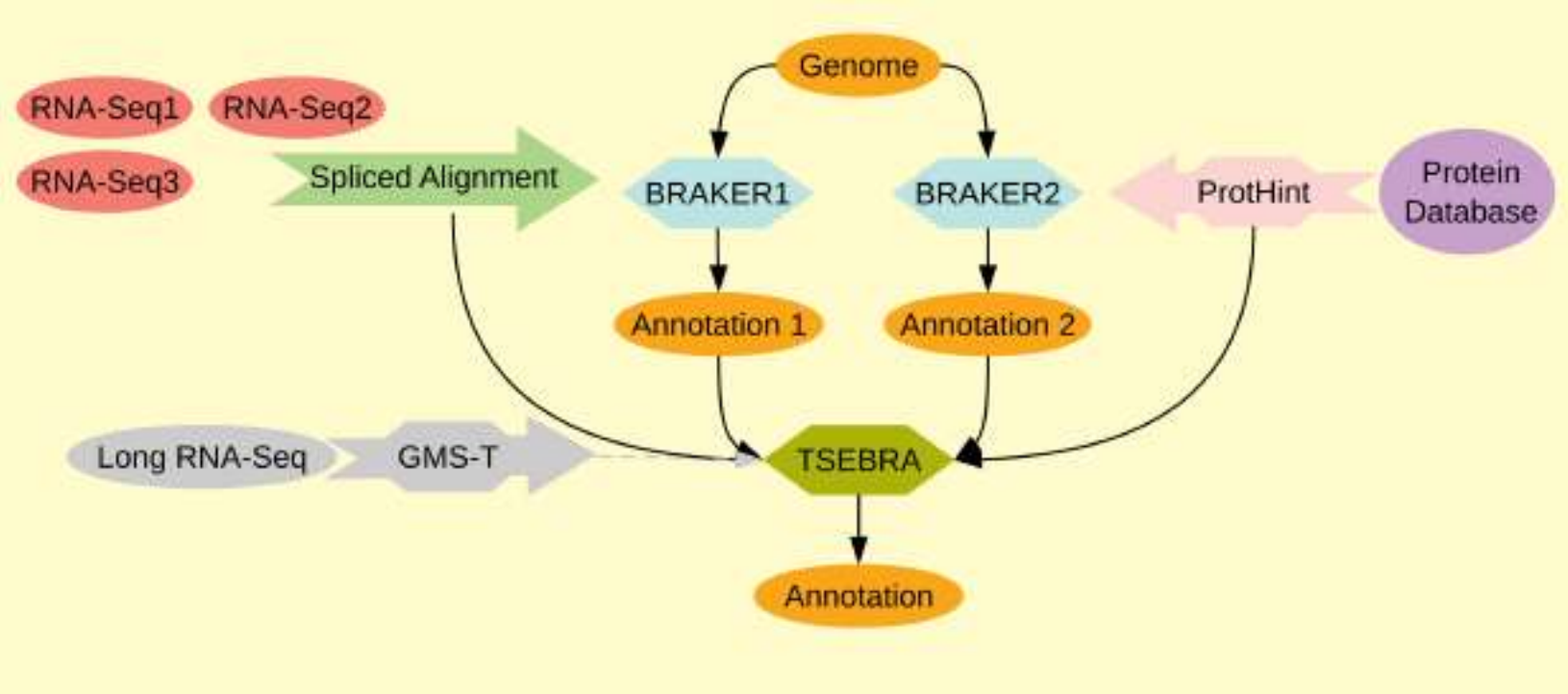
TSEBRA - Summary



Availability

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

Long-Read Integration



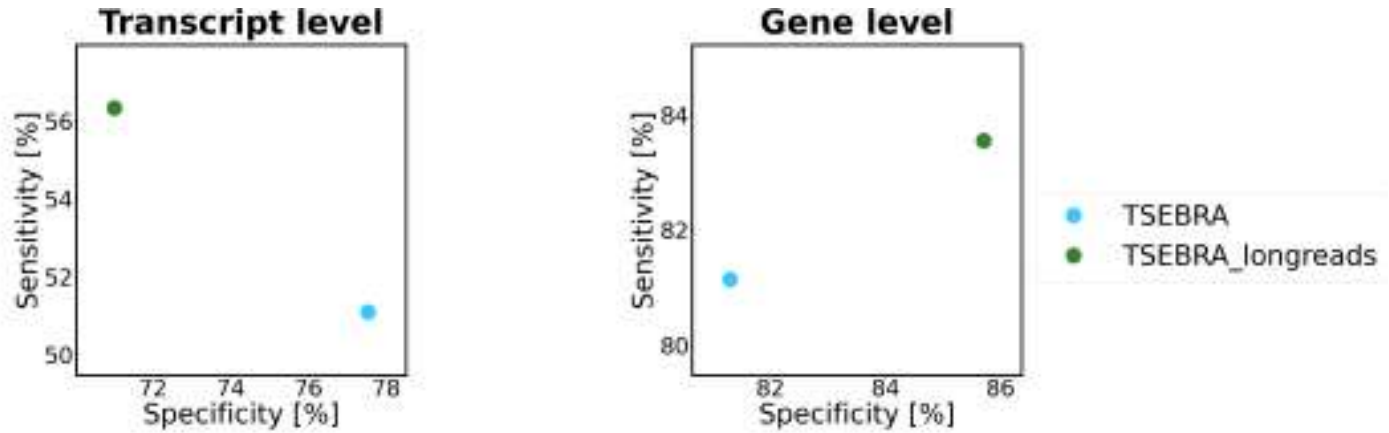
GeneMarkS-T

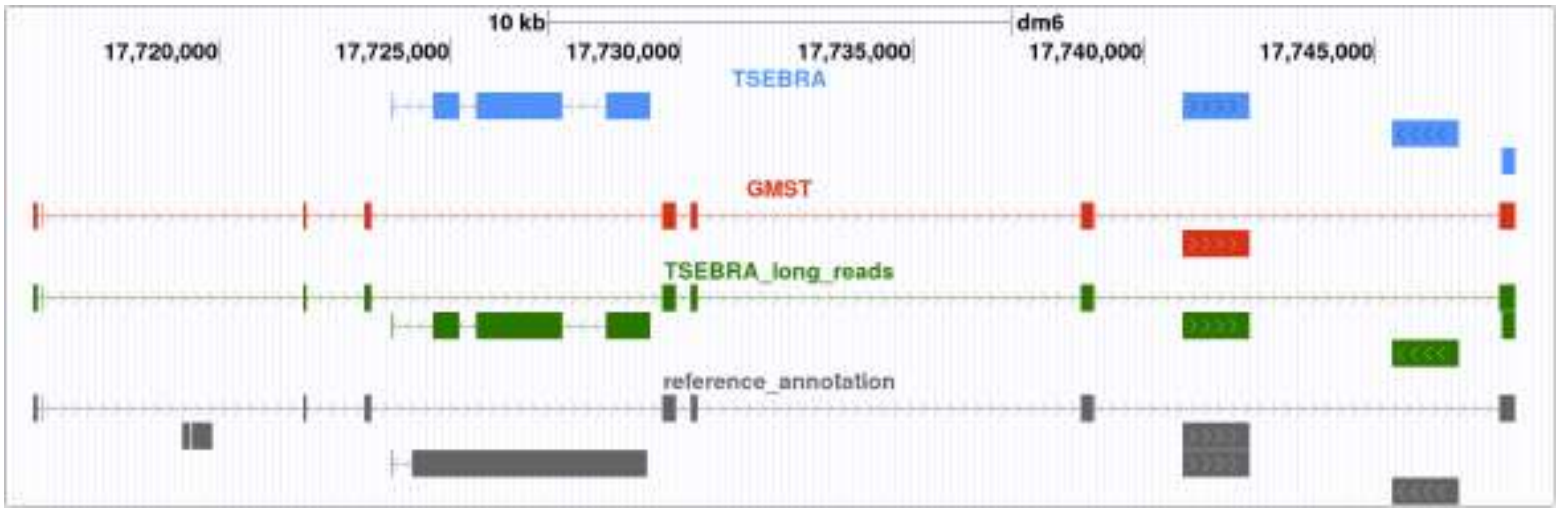
- Identifies protein-coding regions in assembled transcripts

Tang et al. 2015. *Nucleic acids research* 43.12:e78-e78.

Preliminary Results: Model Species

Species: ***Drosophila melanogaster*** (fruit fly)



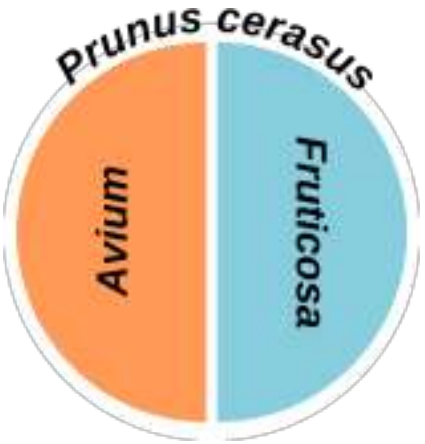


Preliminary Results: Novel Genome

Prunus cerasus (sour cherry)

- 2 subgenomes (*P. avium*, *P. fruticosa*)
- Extrinsic evidence:
  short+long read RNA-Seq data
  protein database (OrthoDB + *Prunus fruticosa*)

Morphology of *P. cerasus* L.. (a) mature tree habitus, (b) leaves, (c) inflorescence, (d) fruits.

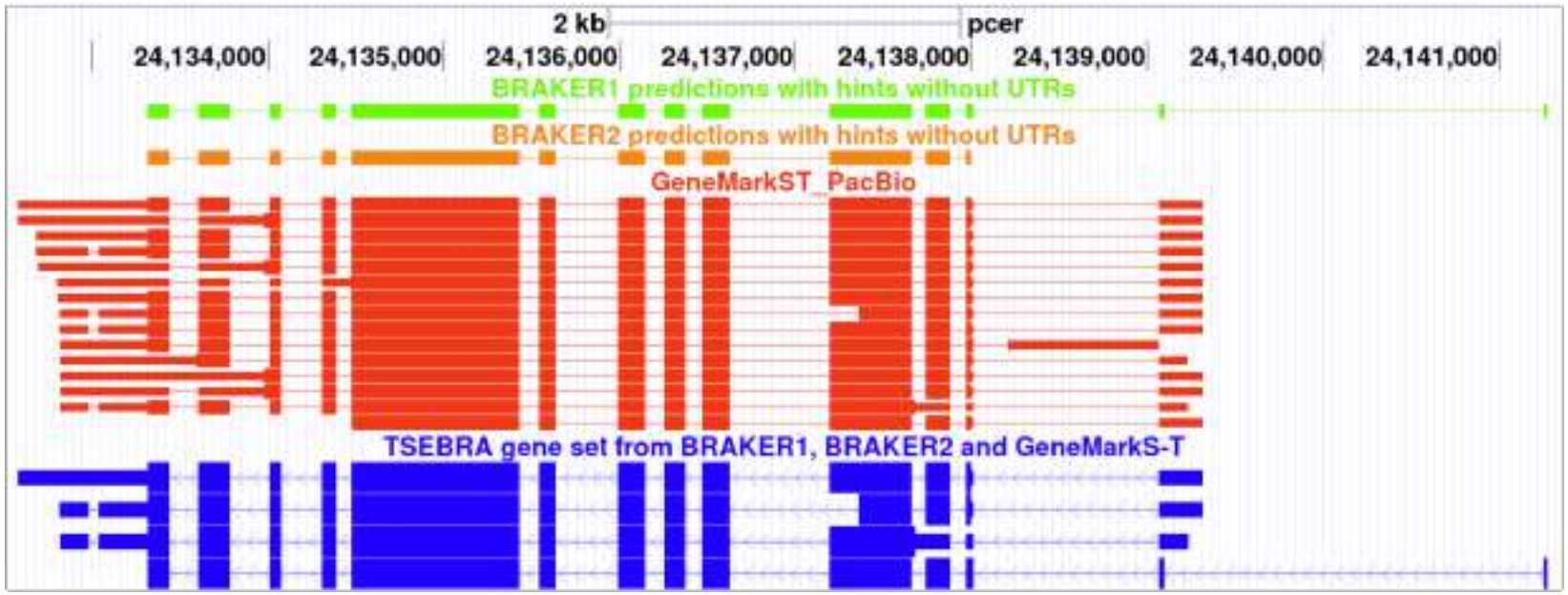


Images used with kind permission of Thomas Wöhner

Preliminary Results: Novel Genome

Prunus cerasus (sour cherry)

- BUSCO: 99.3% completeness
- Annotation: 63,132 genes; 110,062 transcripts
  49,807 transcripts from PacBio long-reads



Hoff et al. 2019. *Genomics, proteomics & bioinformatics* 17.5: 546-549.

Summary

BRAKER Annotation with Extrinsic Evidence

- BRAKER1: short-read RNA-Seq
- BRAKER2: homologous proteins
- GMS-T Protocol: long-read RNA-Seq
- TSEBRA: short(+long)-read RNA-Seq and homologous proteins

Availability

- `https://github.com/Gaius-Augustus/TSEBRA`
- `https://github.com/Gaius-Augustus/BRAKER`
  - `http://exon.gatech.edu/GeneMark/license_download.cgi`
  - `https://github.com/Gaius-Augustus/Augustus`
  - `https://github.com/gatech-genemark/ProtHint`

Acknowledgements

Acknowledgements

Thomas Wöhner
Alexandre Lomsadze

Funding

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

Co-Authors

Katharina J. Hoff
Hannah Thierfeldt
Tomáš Brůna
Mark Borodovsky
Mario Stanke

Availability

- `https://github.com/Gaius-Augustus/TSEBRA`
- `https://github.com/Gaius-Augustus/BRAKER`
  - `http://exon.gatech.edu/GeneMark/license_download.cgi`
  - `https://github.com/Gaius-Augustus/Augustus`
  - `https://github.com/gatech-genemark/ProtHint`

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.

Evgenia et al. ”OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.” Nucleic Acids Research 47.D1 (2019): D807–D811.

Haas, Brian J., et al. ”Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.” Genome biology 9.1 (2008): 1-22.

Hoff et al. ”MakeHub: fully automated generation of UCSC genome browser assembly hubs.” Genomics, proteomics & bioinformatics 17.5 (2019): 546-549.