Enhancing BRAKER3 for Eukaryotic Genome Annotation: Improved Transcript Selection with TSEBRA and a Step Towards Isoseq Integration

Plant and Animal Genome 31

Neng Huang,
Tomáš Brůna,
Katharina J. Hoff

Poster PO0719

Contact: katharina.hoff@uni-greifswald.de
Twitter: @katharina_hoff

Contents

1. **Gene Prediction**
2. **BRAKER**
3. **BUSCO Drop**
   - TSEBRA
   - compleasm
   - Solution
4. **BRAKER3 & IsoSeq**
   - StringTie Option
   - Data
   - Results
5. **Availability**

Structural Genome Annotation Problem

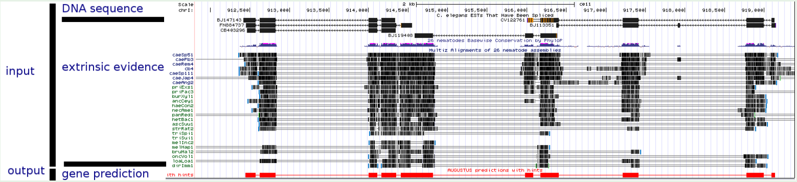
Input

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

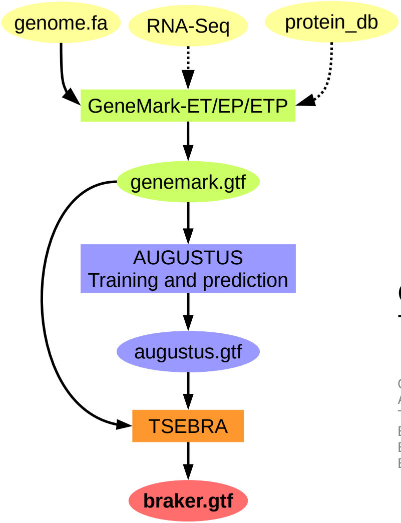
Output

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

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


DNA sequence
input
extrinsic evidence
output
gene prediction
C. elegans ESTs That Have Been Spliced
26 nematodes Basewise Conservation by PhyloP
Multiz Alignments of 26 nematode assemblies
AUGUSTUS predictions with hints

BRAKER: Using RNA-Seq and/or Protein Evidence with GeneMark-ET/EP/ETP, AUGUSTUS and TSEBRA



genome.fa

RNA-Seq

protein_db

GeneMark-ET/EP/ETP

genemark.gtf

AUGUSTUS
Training and prediction

augustus.gtf

TSEBRA

braker.gtf

- BRAKER1: RNA-Seq 1316 citations
- BRAKER2: Proteins DB 728 citations
- BRAKER3: RNA-seq & Protein DB, 15 citations
- 5.3k docker pulls (since 2023)

GeneMark-ETP: **PO0709**
Talk Tuesday, 10:50, Pacific A

GeneMark-ETP: Bruna *et al.* (2023)
AUGUSTUS: Stanke *et al.* (2008)
TSEBRA: Gabriel *et al.* (2021)
BRAKER1: Hoff *et al.* (2016, 2019)
BRAKER2: Bruna *et al.* (2021)
BRAKER3: Gabriel *et al.* (2032)

Measuring Accuracy of Genome Annotation

Developer Approach

Experiments

Accuracy assessment using genome-wide predictions:

Species	Genome Size (Mb)	# Genes in Annotation
Arabidopsis thaliana (thale cress)	119	27,444
Bombus terrestris (bumble bee)	249	10,581
Caenorhabditis elegans (nematode)	100	20,172
Danio rerio (zebrafish)	1,345	25,611
Drosophila melanogaster (fruit fly)	137	13,928
Gallus gallus (chicken)	1,040	17,279
Medicago truncatula (barrelclover)	420	44,464
Mus musculus (mouse)	2,650	22,378
Parasteatoda tepidariorium (house spider)	1,445	18,602
Populus trichocarpa (poppy)	389	34,488
Solanum lycopersicum (tomato)	772	33,562

Accuracy metrics

Precision: Percentage of correctly found genes/transcripts/exons in the **predicted gene set**.

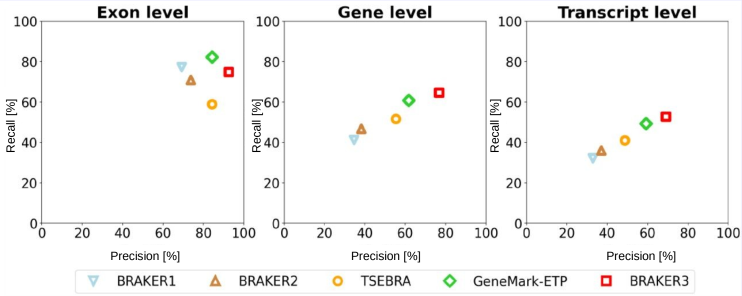
Recall: Percentage of correctly found genes/transcripts/exons in the **reference annotation**.

F1-Score: $\frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$

Measuring Accuracy of Genome Annotation

Developer Approach

Gabriel *et al.* (2023), adapted from Figure 2,
https://doi.org/10.1101/2023.06.10.544449

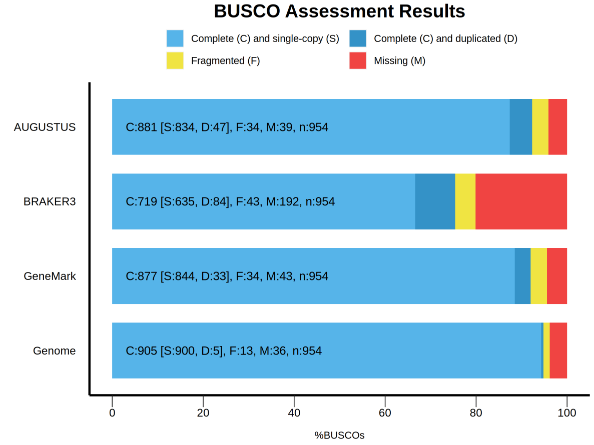


Average accuracy for the genomes of 11 different species with short-read RNA-Seq libraries and protein databases (order excluded)

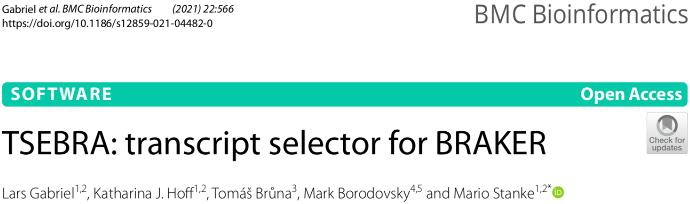
Measuring Accuracy of Genome Annotation

User Approach

Example: *Pycnopodia helianthoides*, Schiebelhut *et al.* (2023)



BUSCO: Waterhouse *et al.* (2018)



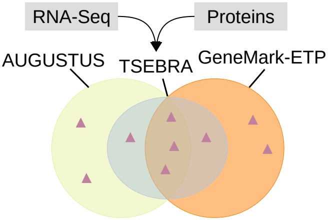
Gabriel *et al. BMC Bioinformatics* (2021) 22:566
https://doi.org/10.1186/s12859-021-04482-0

BMC Bioinformatics

SOFTWARE Open Access

TSEBRA: transcript selector for BRAKER

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

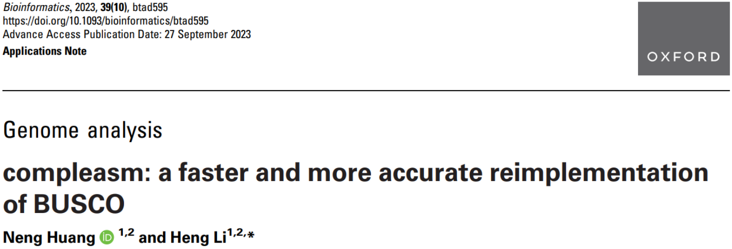


▲ Transcript that matches BUSCO

- combine several gene sets
- increase accuracy
- 68 citations (Google Scholar)
- **may discard BUSCOs**

BUSCO: **B**enchmarking **U**niversal **S**ingle **C**opy **O**rthologs

compleasm



Bioinformatics, 2023, **39(10)**, btad595
https://doi.org/10.1093/bioinformatics/btad595
Advance Access Publication Date: 27 September 2023
Applications Note

OXFORD

Genome analysis

compleasm: a faster and more accurate reimplementation of BUSCO

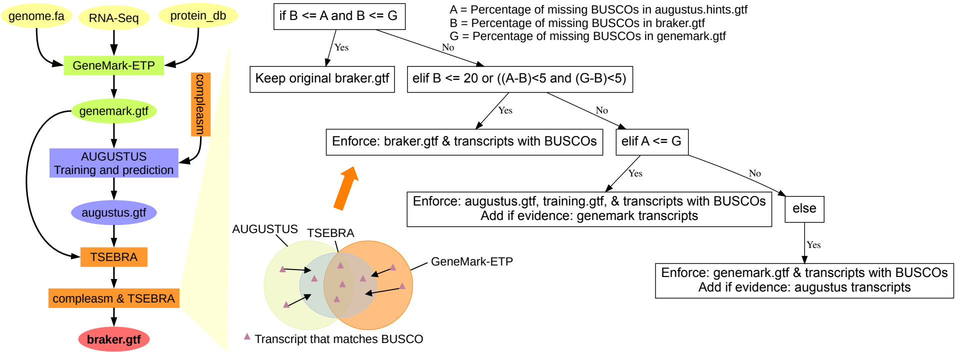
Neng Huang [1,2] and Heng Li [1,2,*]

- originally developed for BUSCO detection in genomes
- recently extended to BUSCO detection in proteins

⇒ This can solve our BRAKER-BUSCO problem

Improving BRAKER with Compleasm

Scenario 1: Good Evidence



genome.fa
RNA-Seq
protein_db
GeneMark-ETP
genemark.gtf
compleasm
AUGUSTUS
Training and prediction
augustus.gtf
TSEBRA
compleasm & TSEBRA
braker.gtf

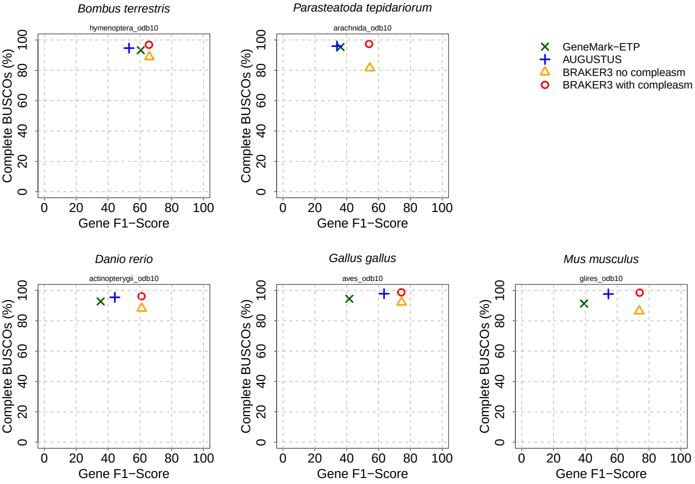
A = Percentage of missing BUSCOs in augustus.hints.gtf
B = Percentage of missing BUSCOs in braker.gtf
G = Percentage of missing BUSCOs in genemark.gtf

- if B <= A and B <= G
  - Yes: Keep original braker.gtf
  - No: elif B <= 20 or ((A-B)<5 and (G-B)<5)
    - Yes: Enforce: braker.gtf & transcripts with BUSCOs
    - No: elif A <= G
      - Yes: Enforce: augustus.gtf, training.gtf, & transcripts with BUSCOs
Add if evidence: genemark transcripts
      - No: else
        - Yes: Enforce: genemark.gtf & transcripts with BUSCOs
Add if evidence: augustus transcripts

▲ Transcript that matches BUSCO

Improving BRAKER with Compleasm

Scenario 1: Good Evidence

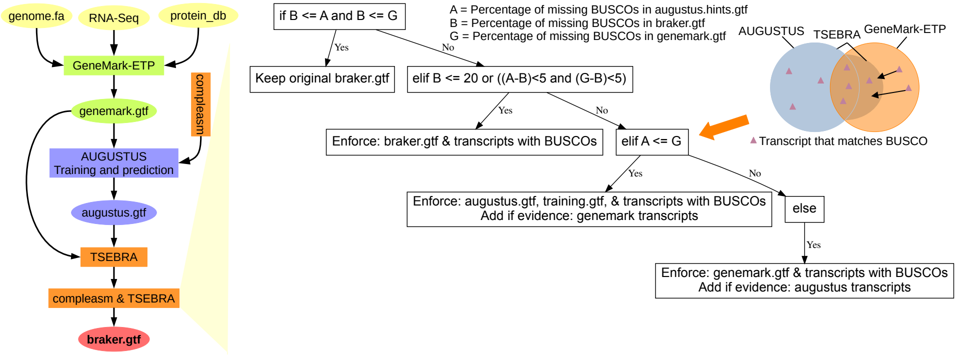


Input data see Gabriel *et al.* (2023)

`https://doi.org/10.1101/2023.06.10.544449`

Improving BRAKER with Compleasm

Scenario 2: Poor Evidence



genome.fa

RNA-Seq

protein_db

GeneMark-ETP

genemark.gtf

compleasm

AUGUSTUS
Training and prediction

augustus.gtf

TSEBRA

compleasm & TSEBRA

braker.gtf

if B <= A and B <= G

Yes: Keep original braker.gtf

No: elif B <= 20 or ((A-B)<5 and (G-B)<5)

Yes: Enforce: braker.gtf & transcripts with BUSCOs

No: elif A <= G

Yes: Enforce: augustus.gtf, training.gtf, & transcripts with BUSCOs
Add if evidence: genemark transcripts

No: else

Yes: Enforce: genemark.gtf & transcripts with BUSCOs
Add if evidence: augustus transcripts

A = Percentage of missing BUSCOs in augustus.hints.gtf
B = Percentage of missing BUSCOs in braker.gtf
G = Percentage of missing BUSCOs in genemark.gtf

AUGUSTUS

TSEBRA

GeneMark-ETP

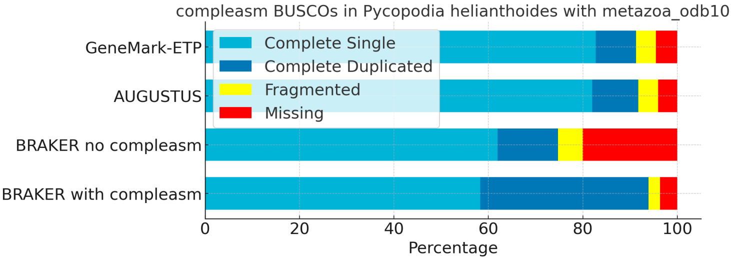
▲ Transcript that matches BUSCO

Improving BRAKER with Compleasm

Scenario 2: Poor Evidence

Input data see Schiebelhut *et al.* (2023)

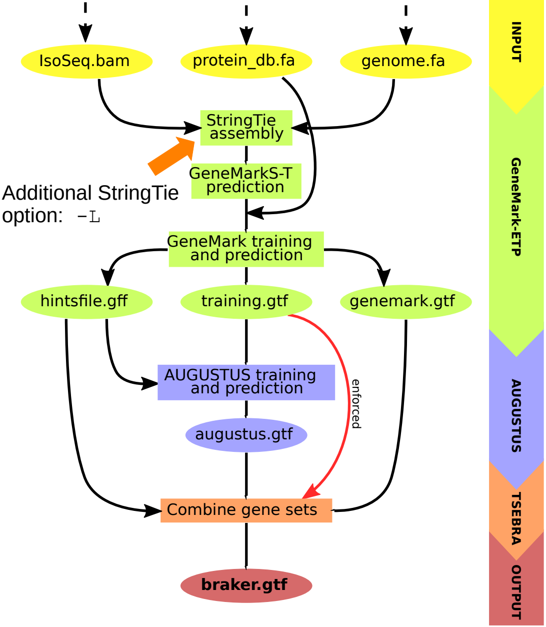
https://doi.org/10.1093/jhered/esad054



	AUGUSTUS	BRAKER3 no compleasm	BRAKER3 with compleasm
#Genes	24,184	15,598	25,601
#Transcripts	26,581	16,473	30,626
Single:Mult ratio	0.29	0.2	0.32

Related seastar *Asterias rubens* has 19,938 genes

BRAKER3 with IsoSeq Data



Additional StringTie option: `-L`

StringTie: Pertea *et al.* (2015); GeneMarkS-T: Tang *et al.* (2015)
Figure adapted from Figure 1 in Gabriel *et al.* (2023)

GeneMark-ETP: **PO0709**

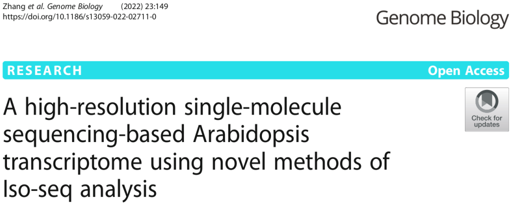
Tue 10:50 Pacific A

Data for Experiment: *Arabidopsis thaliana*

Goal

Comparison of BRAKER3 with short reads (SR) & proteins against BRAKER3 with long reads (LR) & proteins

Genome, reference annotation, protein database (order excluded), & short read RNA-Seq data, see Gabriel *et al.* (2023), IsoSeq reads from



Zhang *et al. Genome Biology* (2022) 23:149
https://doi.org/10.1186/s13059-022-02711-0

Genome Biology

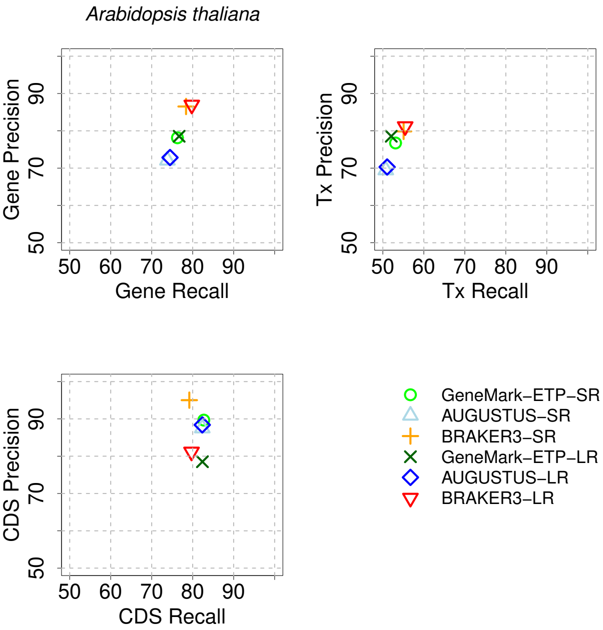
RESEARCH Open Access

A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis

Check for updates

- 11 GB
- 7,604,981 reads
- 96% alignment rate to genome
- spliced alignment IsoSeq to genome with minimap2

Accuracy Results



Availability

GitHub

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

Docker/Singularity

```
singularity build braker.sif \
    docker://teambraker/braker:latest

singularity exec braker.sif braker.pl [OPTIONS]
```

→ Running BRAKER3 with IsoSeq instructions at PO0719

Licenses

- BRAKER: Artistic License
- most dependencies: open source software licenses
- GeneMark-ETP: CC BY-**NC**-SA

Summary

- BRAKER is a highly accurate & fully automatic pipeline
- new: maximizing BUSCOs with compleasm & TSEBRA
- new: IsoSeq as input

BRAKER is Available for Download at

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

Acknowledgements

Neng Huang

Tomas Bruna

Simone Lange
Hannah Thierfeldt
Anica Hoppe

Lars Gabriel

Mark Borodovsky

Matthis Ebel

Alexandre Lomsadze

Mario Stanke

Thank you for your attention!

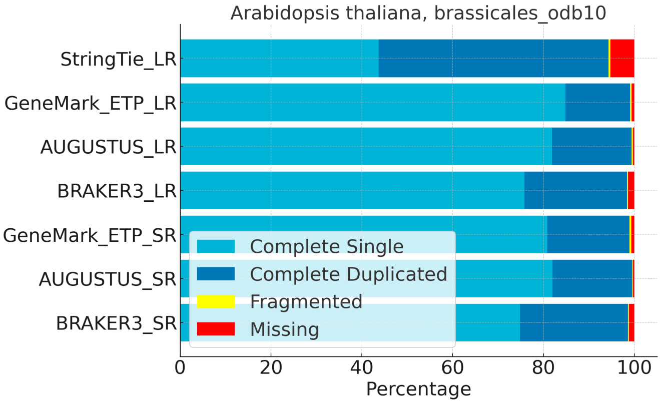
Talk on genome annotation pipeline GALBA on Tue 10:30 Pacific A, **PO0711**

References

- Gabriel *et al.* (2021) “TSEBRA: transcript selector for BRAKER”
- Bruna *et al.* (2021) “BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database”
- Hoff *et al.* (2016) “BRAKER1: unsupervised RNAseq-based genome annotation with GeneMark-ET and AUGUSTUS.”
- Hoff *et al.* (2019) “Whole-genome annotation with BRAKER.“
- Stanke *et al.* (2008) “Using native and syntenically mapped cDNA alignments to improve de novo gene finding.”
- Bruna *et al.* (2023) “GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data”
- Gabriel *et al.* (2023) “BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA”
- Schiebelhut *et al.* (2023) “A reference genome for ecological restoration of the sunflower sea star, Pycnopodia helianthoides”
- Waterhouse *et al.* (2018) “BUSCO applications from quality assessments to gene prediction and phylogenomics”
- Huang & Li (2023) “compleasm: a faster and more accurate reimplementation of BUSCO”
- Pertea *et al.* (2015) “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”
- Tang *et al.* (2015) “Identification of protein coding regions in RNA transcripts”
- Zhang *et al.* (2022) “A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis”.
- Li (2018) “Minimap2: pairwise alignment for nucleotide sequences”

Appendix Slides

BUSCO Scores



StringTie assessment with BUSCO, protein assessments with compleasm