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

Neng Huang[1,2], Tomáš Brůna[3], and Katharina J. Hoff[4,5]

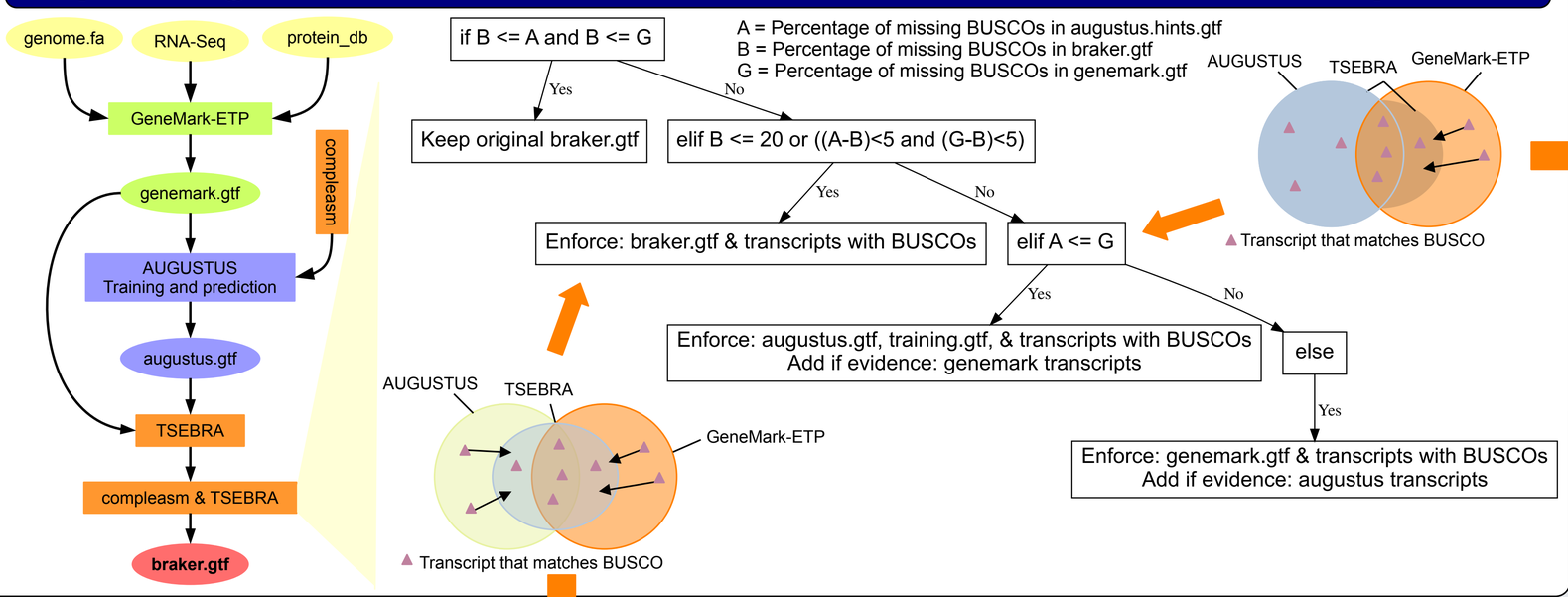
1) Department of Data Sciences, Dana-Farber Cancer Institute, Boston, USA
2) Department of Biomedical Informatics, Harvard Medical School, Boston, USA
3) U.S. Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, USA
4) Institute for Mathematics and Computer Science, University of Greifswald, Greifswald, Germany
5) Center for Functional Genomics of Microbes, University of Greifswald, Greifswald, Germany

Contact: katharina.hoff@uni-greifswald.de

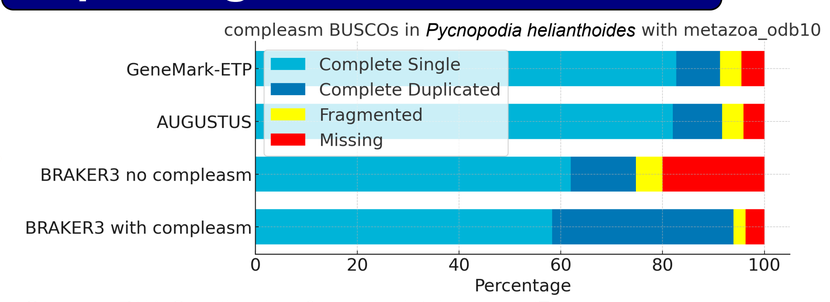
Abstract

BRAKER3 [1] is a cutting-edge, fully automated pipeline for structural annotation of eukaryotic protein-coding genes, leveraging evidence from short read RNA-Seq data and an extensive protein database. This pipeline integrates a number of bioinformatics tools, most importantly GeneMark-ETP [2], AUGUSTUS [3], and TSEBRA [4], the transcript selector for BRAKER. We previously demonstrated BRAKER3's high accuracy in various test genomes, provided there is ample extrinsic evidence. However, challenges arise in less ideal evidence scenarios, where TSEBRA's reliance on extrinsic evidence might lead to excessive transcript rejection. To mitigate this, we introduced a TSEBRA extension employing the rapid marker gene assessment tool compleasm [5]. Compleasm, first developed for assessing the presence of marker genes in genomic sequences, was here equipped with a mode for finding marker genes in protein sequences. The TSEBRA extension evaluates the the number of missing marker genes in gene sets pre- and post-TSEBRA processing. A significant reduction in BUSCO [6] presence triggers a re-run of TSEBRA, prioritizing transcripts from the gene set with the fewest missing BUSCOs. While this approach trades off precision, it substantially enhances the presence of marker genes in the final gene set. Additionally, responding to user requests, we made initial steps towards isoseq data handling by GeneMark-ETP, allowing BRAKER3 to process this data type. Initial findings suggest that given appropriate quantity, quality, and coverage by isoseq reads, accuracy can be comparable to short read application scenarios.
The updated BRAKER3 pipeline has been containerized with Docker and is easily executable with Singularity. BRAKER3 is available at **https://github.com/Gaius-Augustus/BRAKER.**

Improving BUSCO Scores of BRAKER3 Gene Sets with Compleasm



Improving with Poor Evidence



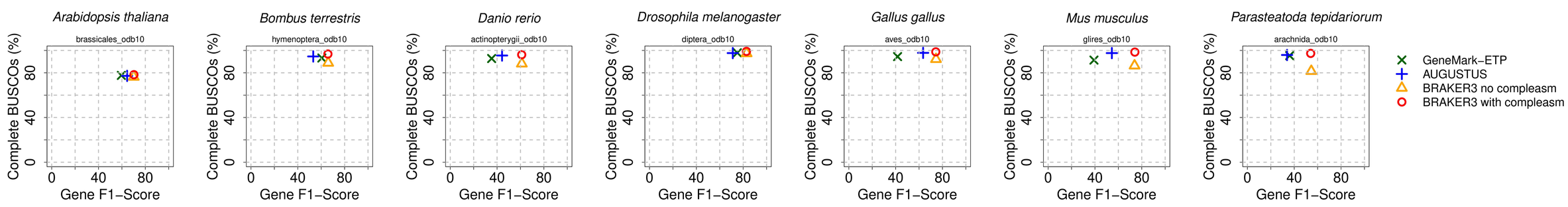
- Genome, RNA-Seq libraries, & protein database see [7].
- Originally published the AUGUSTUS gene set because TSEBRA drops too many transcripts in BRAKER3 due to a lack of evidence support
- compleasm and TSEBRA restore BUSCO completeness in BRAKER3 gene set

\	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

⚠ This approach may lead to an overprediction of genes. GNOMON (usually less sensitive) annotated 19,938 genes in related seastar *Asterias rubens.*

Improving Compleasm BUSCO Completeness with Good Evidence

Genomes, RNA-Seq libraries, & order excluded protein databases see [1].



Recall: Percentage of correctly found features in the reference annotation

Precision: Percentage of correctly found features in the predicted feature set

$$\text{F1-score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

Running BRAKER3 with Compleasm in Singularity

Building the Singularity image:

```
singularity build braker3.sif docker://teambraker/braker3:latest
```

Calling BRAKER3 with compleasm support:

```
singularity exec -B ${PWD}:${PWD} braker3.sif braker.pl --genome=genome.fa --prot_seq=protein_db.fa --rnaseq_sets_ids=RNA_ID1,RNA_ID2 --rnaseq_sets_dirs=/RNASeq/dir/ \
   --busco_lineage=lineage_odb10 --threads=48
```

BRAKER3 with IsoSeq Reads

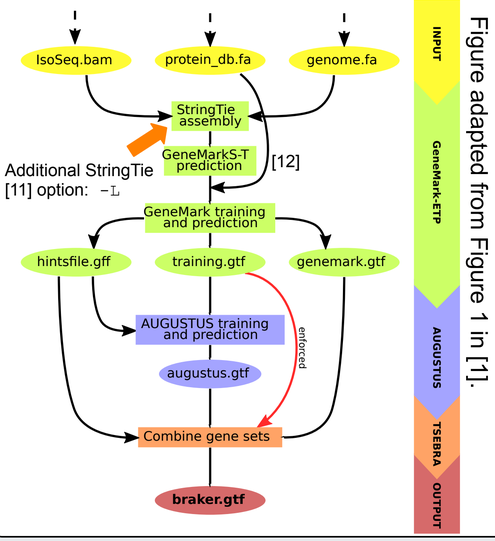
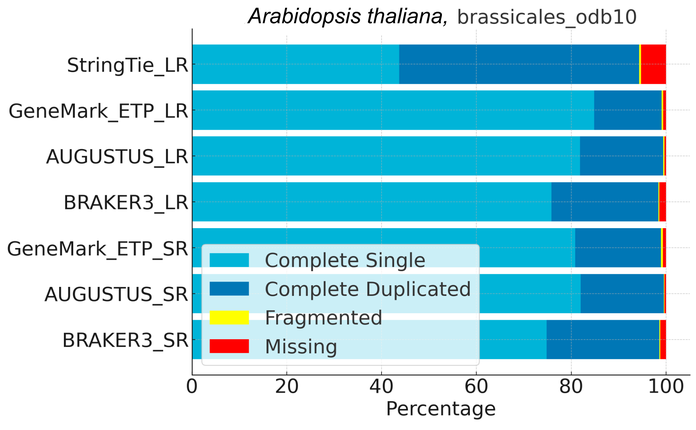


Figure adapted from Figure 1 in [1].

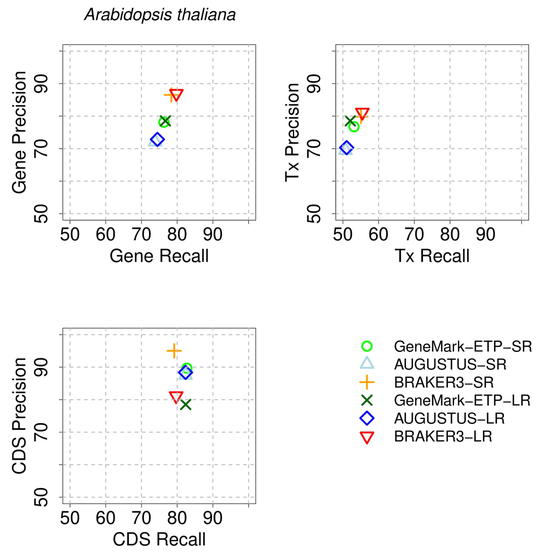
Data for Experiments:
Genome, reference annotation, protein database (order excluded), & short read RNA-Seq data, see [1], IsoSeq reads from [10]. IsoSeq reads were aligned to the genome with minimap2 [8] prior running BRAKER3.

Properties of IsoSeq data:
- 11 GB
- 7,604,981 reads
- 96% alignment rate to genome

Conclusion:
On gene and transcript level, the usage of high quality long read transcriptome data in combination with a large OrthoDB partition yields slightly better results than the usage of short read RNA-Seq data in combination with the same OrthoDB partition in BRAKER3. This does not hold for CDS level precision.



BUSCO detection in protein sets was performed with compleasm. BUSCO detection in the StringTie transcriptome was performed with BUSCO v5.4.2.



Running BRAKER3 with IsoSeq Reads in Singularity

Prepraring BAM file:

```
minimap2 -t48 -ax splice:hq -uf genome.fa isoseq.fa > isoseq.sam
Samtools view -bS --threads 48 isoseq.sam -o isoseq.bam                    #[9]
```

Building the Singularity image:

```
singularity build braker3_lr.sif docker://teambraker/braker3:isoseq
```

Calling BRAKER3 with a BAM file of spliced-aligned IsoSeq Reads:

```
singularity exec -B ${PWD}:${PWD} braker3_lr.sif braker.pl --genome=genome.fa --prot_seq=protein_db.fa --bam=isoseq.bam --threads=48
```

⚠ This is an experimental container, do not mix short reads and long reads! Do not input fastq or SRA IDs!

⚠ Long reads do not always yield results as accurate as short reads because coverage tends to be worse!

Author Contributions
N.H. implemented the protein mode of compleasm; T.B. designed the long reads experiment, performed quality control on IsoSeq raw data and participated in design of the long reads variant of BRAKER3; K.J.H. devised and implemented the compleasm extension of BRAKER3, implemented the StringTie option in GeneMark-ETP, built the containers, performed the experiments.

Acknowledgements
We thank Stefan Prost for testing the long reads container.

References

[1] Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2023). BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv.
[2] Bruna, T., Lomsadze, A., & Borodovsky, M. (2023). GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data. bioRxiv.
[3] Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), 637-644.
[4] Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: transcript selector for BRAKER. BMC Bioinformatics, 22(1), 1-12.
[5] Huang, N., & Li, H. (2023). compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics, 39(10), btad595.
[6] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212.
[7] Schiebelhut, L. M., DeBiasse, M. B., Gabriel, L., Hoff, K. J., & Dawson, M. N. (2023). A reference genome for ecological restoration of the sunflower sea star, *Pycnopodia helianthoides*. Journal of Heredity, esad054.
[8] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100.
[9] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.
[10] Zhang, R., Kuo, R., Coulter, M., Calixto, C. P., Entizne, J. C., Guo, W., ... & Brown, J. W. (2022). A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. Genome biology, 23(1), 149.
[11] Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome biology, 20(1), 1-13.
[12] Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic acids research, 43(12), e78-e78.