Proteins Only: How Accurately Can We Annotate Large Genomes?

Plant and Animal Genome 31

Tomáš Brůna,
Heng Li,
Joseph Guhlin,
Daniel Honsel,
Steffen Herbold,
Mario Stanke,
Natalia Nenasheva,
Matthis Ebel,
Lars Gabriel,
Katharina J. Hoff

Contact: katharina.hoff@uni-greifswald.de, **Poster PO0711**
Twitter: @katharina_hoff

Contents

1. Gene Prediction
2. GALBA
3. The Idea
4. Accuracy Metrics
5. Development Steps
6. Accuracy Results
   - Effect of Mutations
   - Annotated Reference Species
7. Availability

Structural Genome Annotation Problem

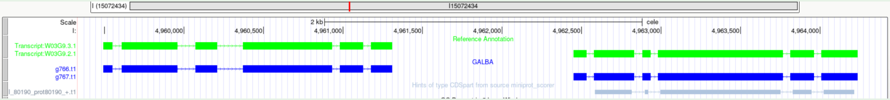
Input

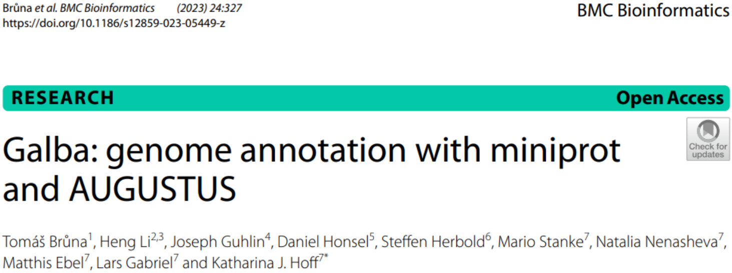
- genome assembly
- extrinsic evidence, e.g. **protein sequences of related species**

Output

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

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





Brůna *et al. BMC Bioinformatics* (2023) 24:327
https://doi.org/10.1186/s12859-023-05449-z

BMC Bioinformatics

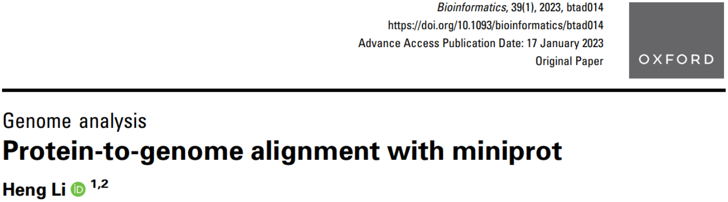
RESEARCH **Open Access**

Galba: genome annotation with miniprot and AUGUSTUS

Tomáš Brůna[1], Heng Li[2,3], Joseph Guhlin[4], Daniel Honsel[5], Steffen Herbold[6], Mario Stanke[7], Natalia Nenasheva[7], Matthis Ebel[7], Lars Gabriel[7] and Katharina J. Hoff[7*]

- 752 docker pulls
- 4 citations (Google Scholar Jan 2nd 2024)

Miniprot



Bioinformatics, 39(1), 2023, btad014
https://doi.org/10.1093/bioinformatics/btad014
Advance Access Publication Date: 17 January 2023
Original Paper

OXFORD

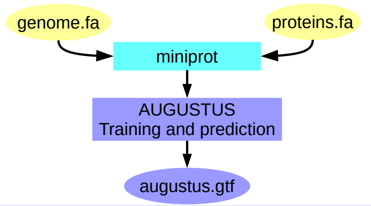
Genome analysis

Protein-to-genome alignment with miniprot

Heng Li [1,2]

“*Miniprot is a fast protein-to-genome aligner comparable to existing tools in accuracy. Its primary use case is to assist gene annotation.*”

GALBA



Do we need another pipeline?

- ∼1000 vertebrate genomes: no RNA-Seq
- BRAKER2 less accurate in large genomes
- Free Open Source Software

Measuring Accuracy of Genome Annotation

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

Protein sequence donor list at `https://doi.org/10.1186/s12859-023-05449-z`

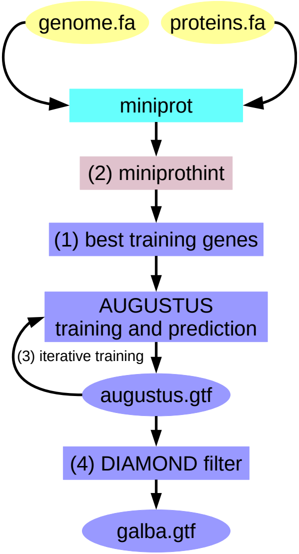
Accuracy metrics

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

GALBA: Using Proteins of Related Species



genome.fa

proteins.fa

miniprot

(2) miniprothint

(1) best training genes

AUGUSTUS training and prediction

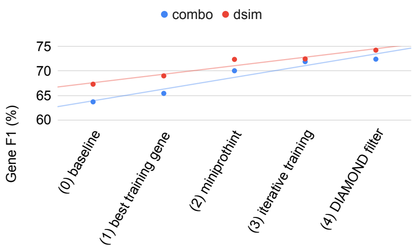
(3) iterative training

augustus.gtf

(4) DIAMOND filter

galba.gtf

Development steps in D. melanogaster



Donor proteins from

dsim	*D. simulans*
combo	*D. ananassae*, *D. pseudoobscura*, *D. willistoni*, *D. virilis*, *D. grimshawi*

Idea for DIAMOND filter from Tolman *et al.* (2023)

DIAMOND: Buchfink *et al.* (2015)

Accuracy of GALBA with Different Protein Donors

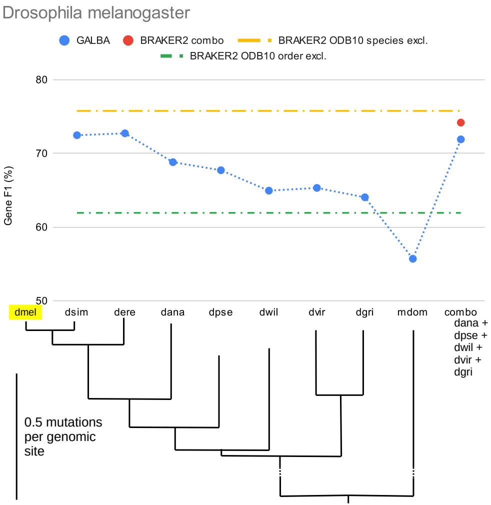
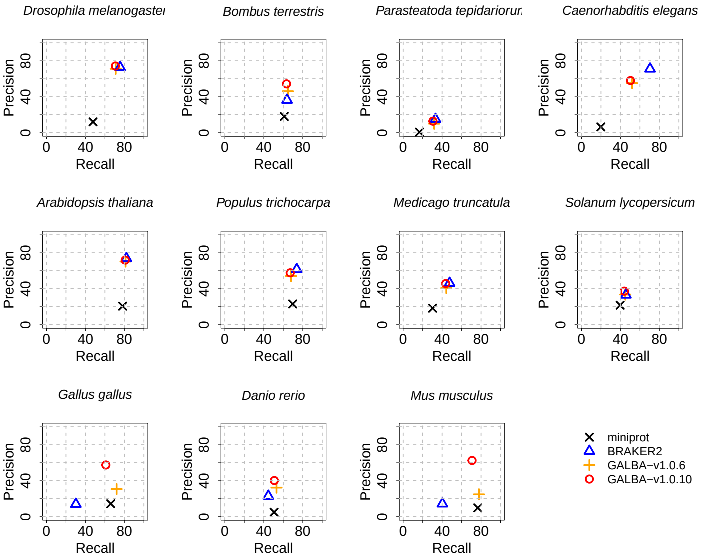


Image: Brůna *et al.* `https://doi.org/10.1186/s12859-023-05449-z`, Fig. 2

BRAKER2: Bruna *et al.* (2021); OrthoDB: Kuznetsov *et al.* (2023)

Accuracy in Reference Species



Drosophila melanogastei

Bombus terrestris

Parasteatoda tepidariorur

Caenorhabditis elegans

Arabidopsis thaliana

Populus trichocarpa

Medicago truncatula

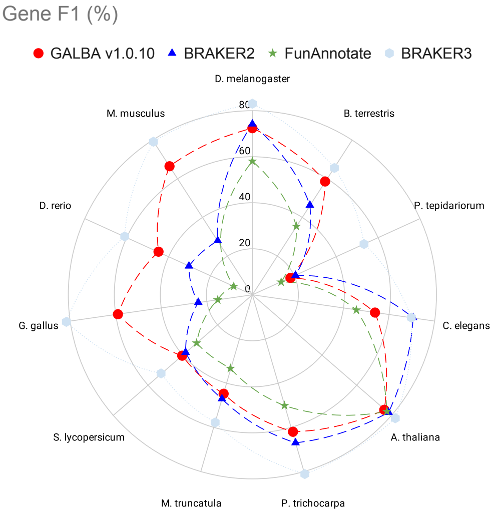
Solanum lycopersicum

Gallus gallus

Danio rerio

Mus musculus

Proteins Only (GALBA, BRAKER2, FunAnnotate) vs. BRAKER3 with RNA-Seq & Proteins



→ Use BRAKER3 with RNA-Seq if available!

Results of BRAKER3 with RNA-Seq and OrthoDB order excluded from Gabriel *et al.* (2023) `https://doi.org/10.1101/2023.06.10.544449`

Availability

GitHub

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

Docker/Singularity

```
singularity build galba.sif \
   docker://katharinahoff/galba:latest

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

Licenses

- GALBA: Artistic License
- all dependencies have Open Source Licenses

Summary

- GALBA is a fully automated pipeline for protein coding gene annotation in eukaryotes
- protein sequences of $n \geq 1$ related species serve as evidence
- GALBA has good accuracy in large vertebrate genomes
- precision improvement slightly decreases recall
- RNA-Seq & proteins are superior to proteins only
- GALBA is freely available and easy to execute

GALBA Contributors

Tomáš Brůna

Heng Li

Joseph Guhlin

Lars Gabriel

Natalia Nenasheva

Ethan Tolman

Paul Frandsen

Matthis Ebel

Mario Stanke

Katharina Hoff

Also: Daniel Honsel, & Steffen Herboldt

Funding

- German Research Foundation grant 277249973 to K.J.H.
- Project Data Competency granted to K.J.H. and M.S. by the government of Mecklenburg-Vorpommern
- US National Institute of Health grant R01HG010040 to H.L.
- German Research Foundation grant 391397397 to S.H. and M.S.

Thank you for your attention!

References

- Bruna *et al.* (2023) “Galba: genome annotation with miniprot and AUGUSTUS”
- Bruna *et al.* (2020) “GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins”
- Bruna *et al.* (2021) “BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database”
- Li (2023) “Protein-to-genome alignment with miniprot.”
- Tolman *et al.* (2023) “Newly Sequenced Genomes Reveal Patterns of Gene Family Expansion in select Dragonflies (Odonata: Anisoptera)“
- Stanke *et al.* (2008) “Using native and syntenically mapped cDNA alignments to improve de novo gene finding.”
- Buchfink *et al.* (2015) “Fast and sensitive protein alignment using DIAMOND.”
- Kuznetsov *et al.* (2023) “OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity.”
- FunAnnotate: `https://github.com/nextgenusfs/funannotate`