Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

2023 Workshop on Genomics, Cesky Krumlov

Katharina J. Hoff & Natalia Nenasheva

Contact: katharina.hoff@uni-greifswald.de

Katharina J. Hoff

Senior Scientist in Bioinformatics at University of Greifswald

Short CV

2022	Habilitation (Greifswald)
2009	Ph.D. Molecular Biology (Göttingen)
2005	B.Sc. Plant Biotechnology (Hanover, stays abroad: Budapest & Alnarp)

Research

- eukaryotic genome annotation, metagenomics
- best known for: **BRAKER** & other **Gaius-Augustus** software
- 27 peer-reviewed research articles with currently 4,759 citations
- ∼1 Mio € grants → **open PhD student position!**

Teaching

- currently 1(+2) PhD students, 1 MSc student, 1 BSc student
- applied bioinformatics, programming, statistics, & data science

Natalia Nenasheva

PhD Student in Bioinformatics at University of Greifswald

Short CV

2022 - now	Institute for Mathematics and Computer Science, University of Greifswald
2021 - 2022	Bioinformatician at the Genotek, Moscow, Russia
2020 - 2022	M.Sc. in Applied Mathematics and Physics (the Moscow Institute of Physics and Technology, Department of Biological and Medical Physics)
2019 - 2020	Junior bioinformatician at the Boston Gene, Moscow, Russia
2016 - 2020	B.Sc. in Applied Mathematics and Physics (the Moscow Institute of Physics and Technology, Department of Biological and Medical Physics)

Research

- gene prediction in eukaryotic metagenomes
- affiliated with POMPU

Teaching

- Python, Bash, and other data science tools

We are Looking for a PhD Student!

Amazing Project

- develop novel methods for pathogen screening in dairy farm environments
- NGS, AI, & applied bioinformatics
- interdisciplinary team

Position

- full-time research assistant
- 3 years, starting on July 1st 2023

Application Deadline

- send cover letter, CV, copies of degree certificates to katharina.hoff@uni-greifswald.de until June 3rd 2023

Contents

Eukaryotic Gene

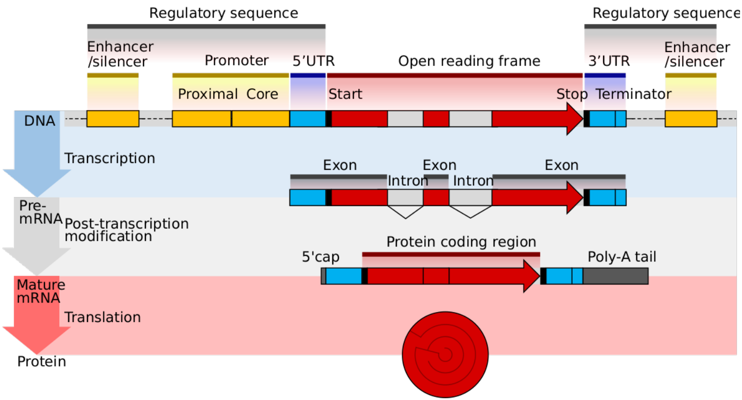
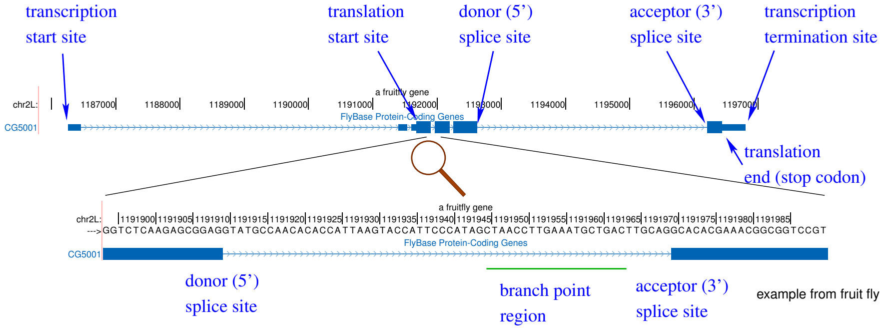


Image: Wikipedia, CC BY 4.0

Information for Genome Annotaiton

- genome sequence: **mathematical model**
- observed mRNA expression (parts)
- known proteins of relatives

Signals



transcription start site

translation start site

donor (5’) splice site

acceptor (3’) splice site

transcription termination site

translation end (stop codon)

donor (5’) splice site

branch point region

acceptor (3’) splice site

example from fruit fly

Sequence Content

Besides the signals, position-unspecific frequencies of nucleotide patterns (e.g. k-mer frequencies) can be used to guess biological classification of longer sequence intervals.

⇒ species specific parameters ⇒ need training ⇒ need training examples

Structural Genome Annotation Problem

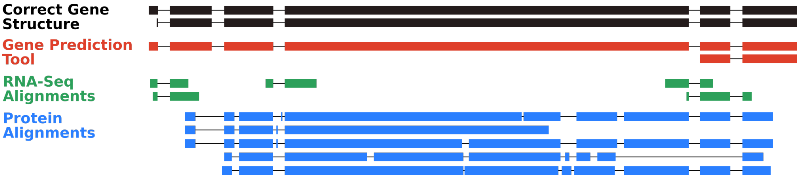
Input

- genome assembly
- extrinsic evidence, e.g. from RNA-Seq, protein sequences

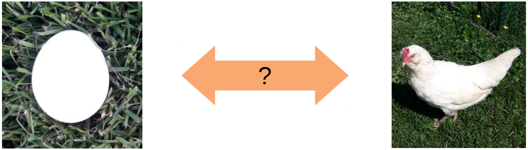
Output

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

Task: find and predict gene structures of protein-coding genes



Gene Finders Need Training Examples



Genes
(Parameter training)

Genes
(Prediction)

Gene Finders Need Training Examples

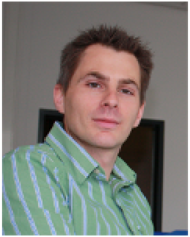
Image: credits to DALL-E2, modified by human

The BRAKER Team

University of Greifswald & Georgia Tech University

Lars Gabriel

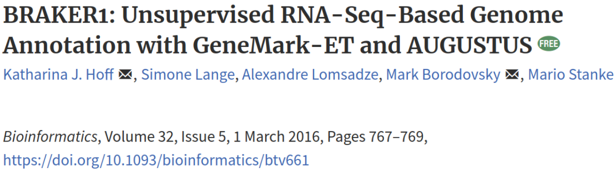
Alexandre Lomsadze, Katharina Hoff, Tomáš Brůna



Mario Stanke

Mark Borodovsky

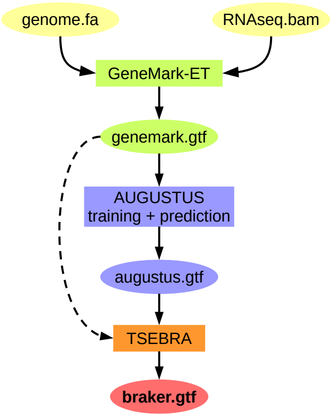
Also: Simone Lange, Matthis Ebel, Hannah Thierfeldt, Anica Hoppe



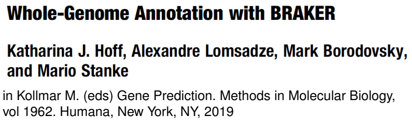
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS FREE

Katharina J. Hoff ✉, Simone Lange, Alexandre Lomsadze, Mark Borodovsky ✉, Mario Stanke

Bioinformatics, Volume 32, Issue 5, 1 March 2016, Pages 767–769, https://doi.org/10.1093/bioinformatics/btv661



- spliced alignments of RNA-Seq are used by GeneMark-ET and AUGUSTUS
- 1,157 citations (Google Scholar)



Whole-Genome Annotation with BRAKER

Katharina J. Hoff, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke

in Kollmar M. (eds) Gene Prediction. Methods in Molecular Biology, vol 1962. Humana, New York, NY, 2019

GeneMark-ET uses RNA-Seq for Training

Anchors from RNA-Seq for training

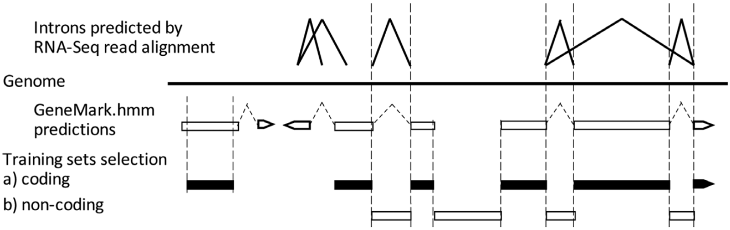
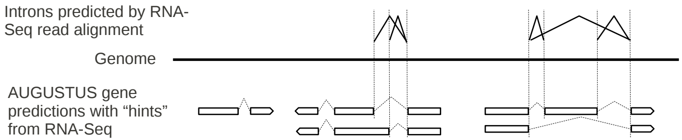


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- employs unsupervised training
- includes in training introns and exons anchored by mapped RNA-Seq reads
- does not require RNA-Seq reads assembly
- does not use RNA-Seq information in the *prediction* step

AUGUSTUS uses RNA-Seq for Prediction



Introns predicted by RNA-Seq read alignment

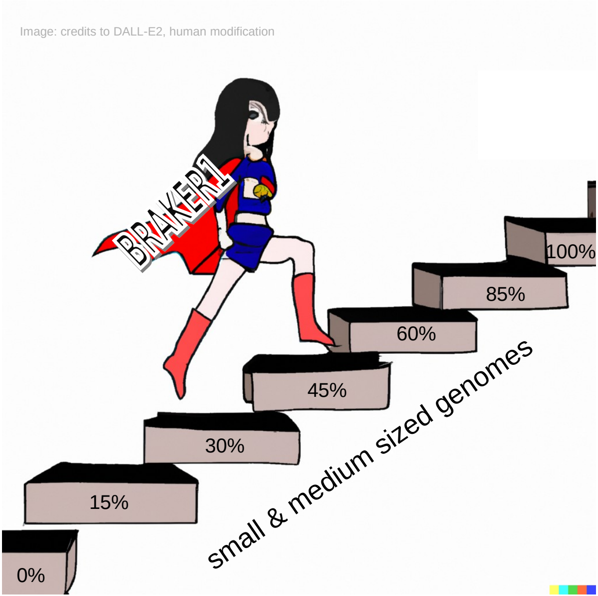
Genome

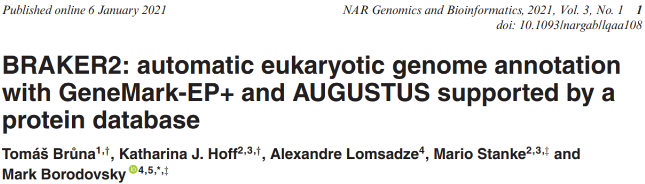
AUGUSTUS gene predictions with “hints” from RNA-Seq

- requires “prior data” for training
- uses intron information from RNA-seq for *prediction*
- no RNA-Seq assembly required

BRAKER1 Gene F1 Accuracy

Image: credits to DALL-E2, human modification

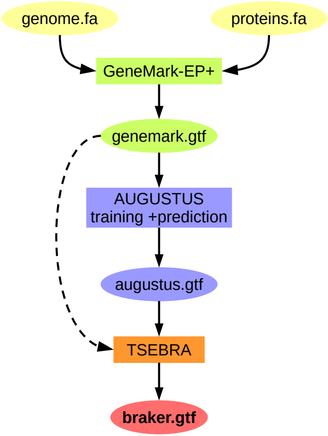




Published online 6 January 2021 *NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1* **1**
doi: 10.1093/nargab/lqaa108

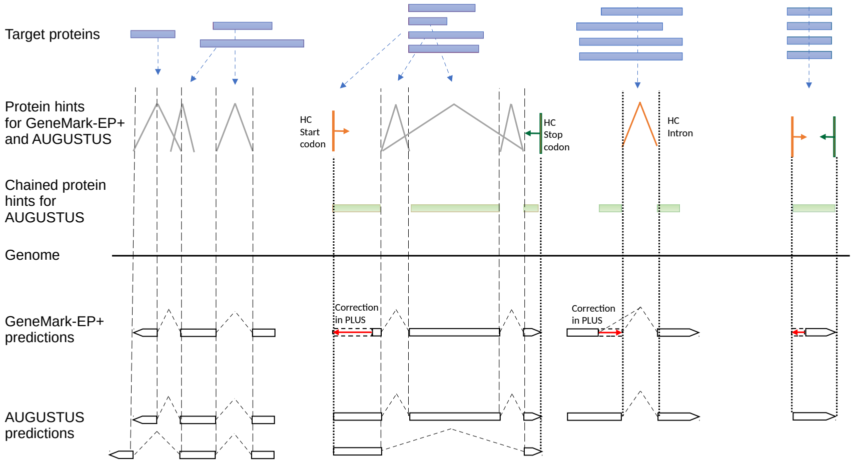
BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database

Tomáš Brůna[1,†], Katharina J. Hoff[2,3,†], Alexandre Lomsadze[4], Mario Stanke[2,3,‡] and Mark Borodovsky[4,5,*,‡]



- spliced alignments of a large number of proteins
- 428 citations (Google Scholar)

Evidence Usage by GeneMark-EP+ & AUGUSTUS During Prediction



Target proteins

Protein hints for GeneMark-EP+ and AUGUSTUS

Chained protein hints for AUGUSTUS

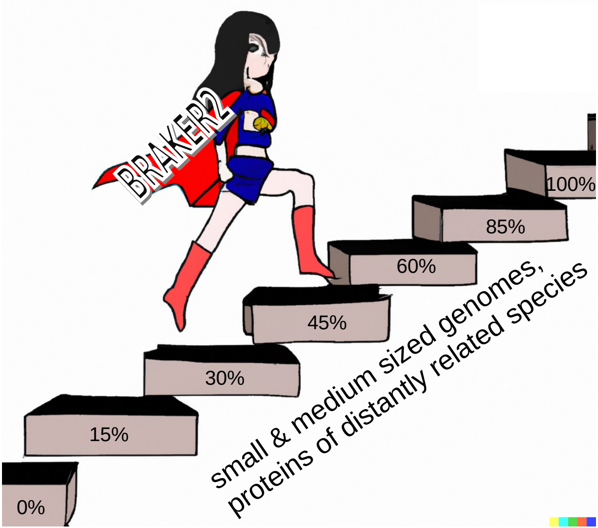
Genome

GeneMark-EP+ predictions

AUGUSTUS predictions

BRAKER2 Gene F1 Accuracy

Image: credits to DALL-E2, human modification



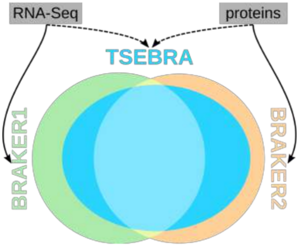
Gabriel *et al. BMC Bioinformatics* (2021) 22:566
https://doi.org/10.1186/s12859-021-04482-0

BMC Bioinformatics

SOFTWARE **Open Access**

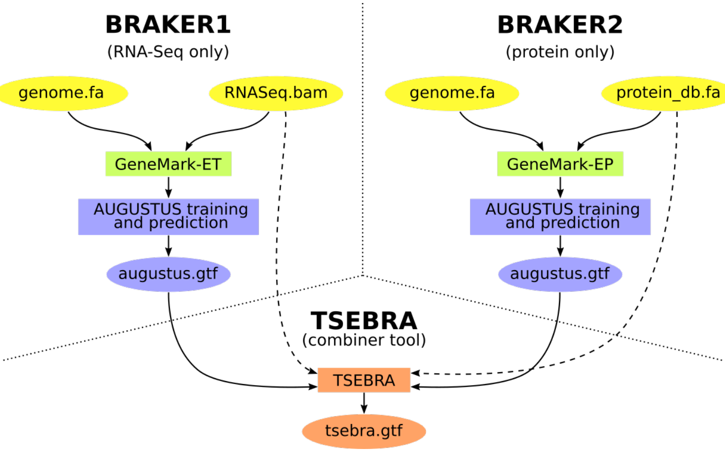
TSEBRA: transcript selector for BRAKER

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



- run both BRAKER1 and BRAKER2
- combine & increase accuracy
- 34 citations (Google Scholar)

BRAKER1 + BRAKER2 → TSEBRA



TSEBRA: BRAKER1 + BRAKER2 Gene F1 Accuracy

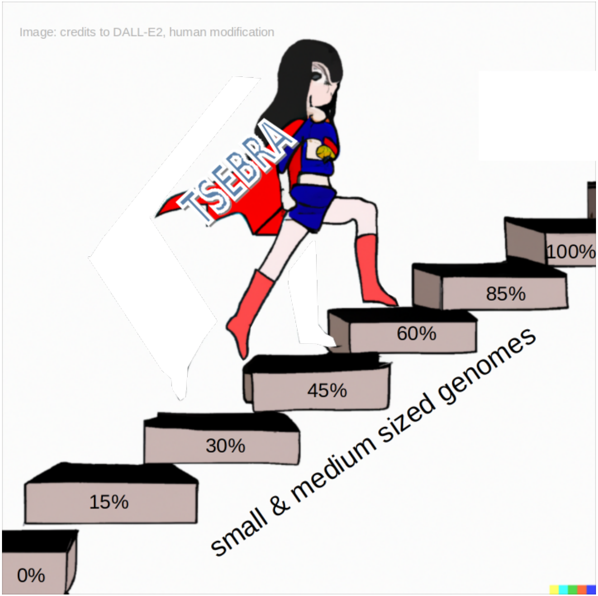


Image: credits to DALL-E2, human modification

BRAKER3 Gene F1 Accuracy - Climbing the Top

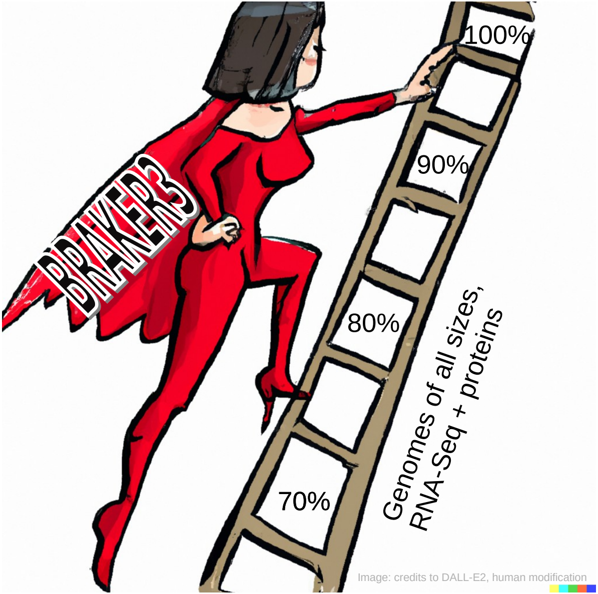
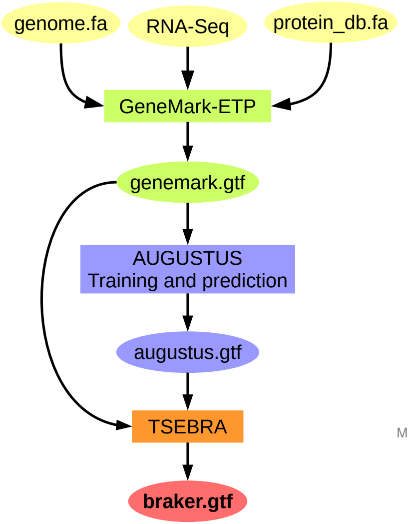


Image: credits to DALL-E2, human modification

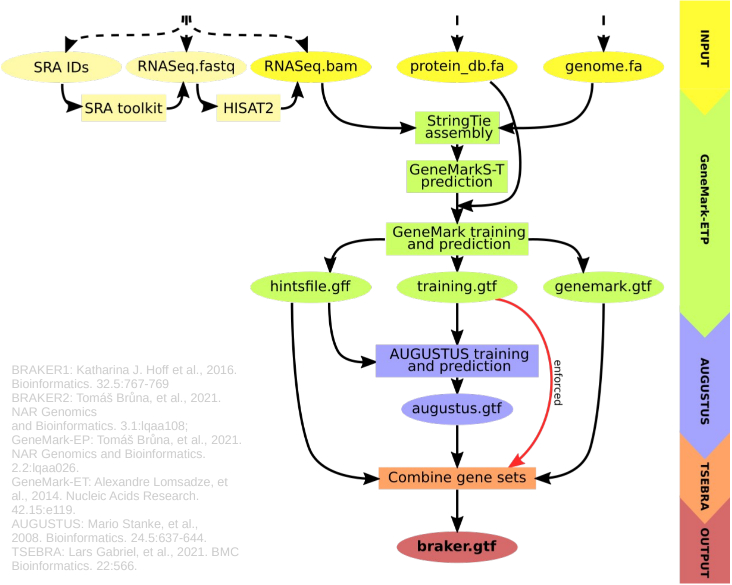
BRAKER3: Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA



- spliced aligned and **assembled** RNA-Seq
- large protein database
- combines GeneMark-ETP and AUGUSTUS gene sets with TSEBRA

Manuscript by Lars Gabriel et. al in preparation

BRAKER3: Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA



BRAKER1: Katharina J. Hoff et al., 2016. Bioinformatics. 32.5:767-769
BRAKER2: Tomáš Brůna, et al., 2021. NAR Genomics and Bioinformatics. 3.1:lqaa108;
GeneMark-EP: Tomáš Brůna, et al., 2021. NAR Genomics and Bioinformatics. 2.2:lqaa026.
GeneMark-ET: Alexandre Lomsadze, et al., 2014. Nucleic Acids Research. 42.15:e119.
AUGUSTUS: Mario Stanke, et al., 2008. Bioinformatics. 24.5:637-644.
TSEBRA: Lars Gabriel, et al., 2021. BMC Bioinformatics. 22:566.

BRAKER3 Experiments

Experiments

Accuracy assessment using genome-wide predictions in 6 species:

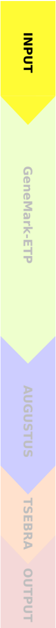
Species	Genome Size (Mb)	# Genes in Annotation
Arabidopsis thaliana (thale cress)	119	27,444
Caenorhabditis elegans (nematode)	100	20,172
Drosophila melanogaster (fruit fly)	137	13,928
Gallus gallus (chicken)	1,040	17,279
Mus musculus (mouse)	2,650	22,378
Solanum lycopersicum (tomato)	772	33,562

Accuracy metrics

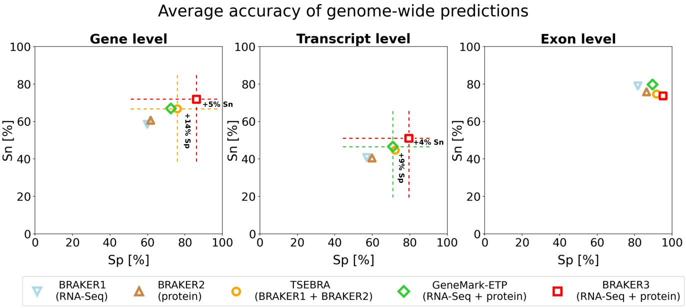
Specificity [Sp]: Percentage of correctly found genes/transcripts/exons in the **predicted gene set**.

Sensitivity [Sn]: Percentage of correctly found genes/transcripts/exons in the **reference annotation**.

Harmonic Mean [F1]: $$\frac{2 \cdot \text{Sensitivity} \cdot \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$



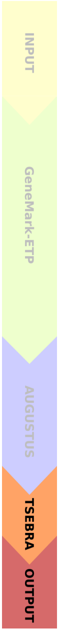
BRAKER3 Accuracy in Small Genomes



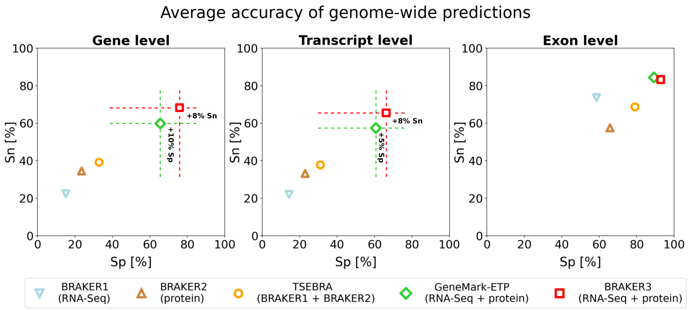
Species: *D. melanogaster*
A. thaliana
C. elegans

Extrinsic evidence:

- paired RNA-Seq short reads
- OrthoDB v.10 clade partitions **(order excluded)**



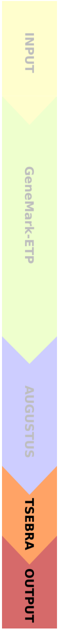
BRAKER3 Accuracy in Larger Genomes



Species: *M. musculus*
G. gallus
S. lycopersicum

Extrinsic evidence:

- paired RNA-Seq short reads
- OrthoDB v.10 clade partitions **(order excluded)**



Usage & Runtime

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 for *A. thaliana*, *C. elegans*, *D. melanogaster*, *G. gallus*, *M. musculus*, *S. lycopersicum*.
- with 48 threads:

	BRAKER1	BRAKER2	GM-ETP	BRAKER3
Runtime (h)	06:26	09:01	06:03	17:55

Availability

GitHub

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

Docker/Singularity

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

singularity exec braker3.sif braker.pl [OPTIONS]
```

Licenses

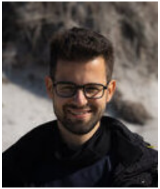
- BRAKER: Artistic License
- GeneMark-ETP: License for GeneMark family software `http://topaz.gatech.edu/genemark/license_download.cgi`

The GALBA Team

DOE Joint Genome Institute & Harvard Medical School & University of Greifswald

Tomáš Brůna

Heng Li



Lars Gabriel

Natalia Nenasheva

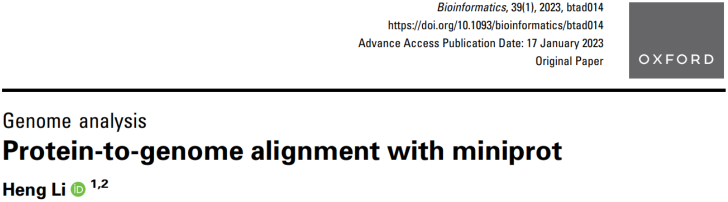
Matthis Ebel

Mario Stanke

Katharina Hoff

Also: Daniel Honsel, & Steffen Herboldt

Miniprot



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

OXFORD

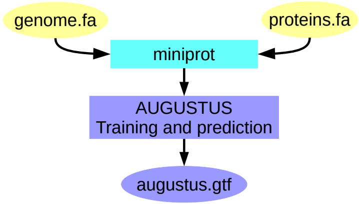
Genome analysis

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

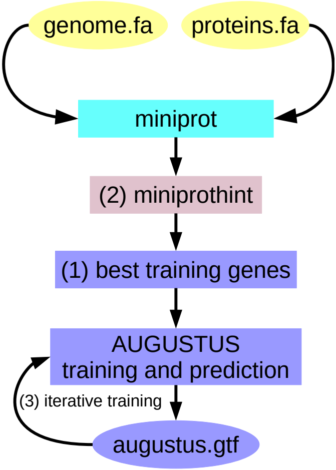


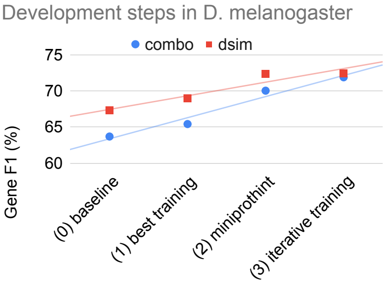
genome.fa
proteins.fa
miniprot
AUGUSTUS
Training and prediction
augustus.gtf

Do we need another pipeline?

- no RNA-Seq, large genome
- proteins of few or 1 reference species
- fully open source

GALBA: Proteins of Closely Related Species

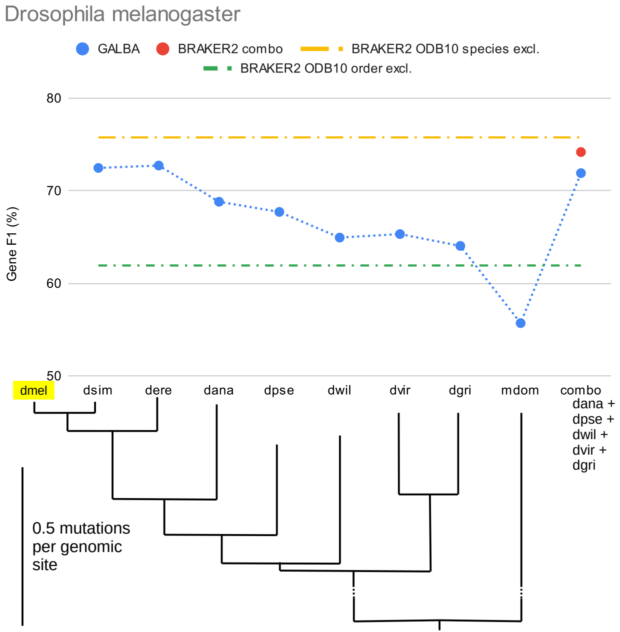

genome.fa
proteins.fa
miniprot
(2) miniprothint
(1) best training genes
AUGUSTUS training and prediction
(3) iterative training
augustus.gtf



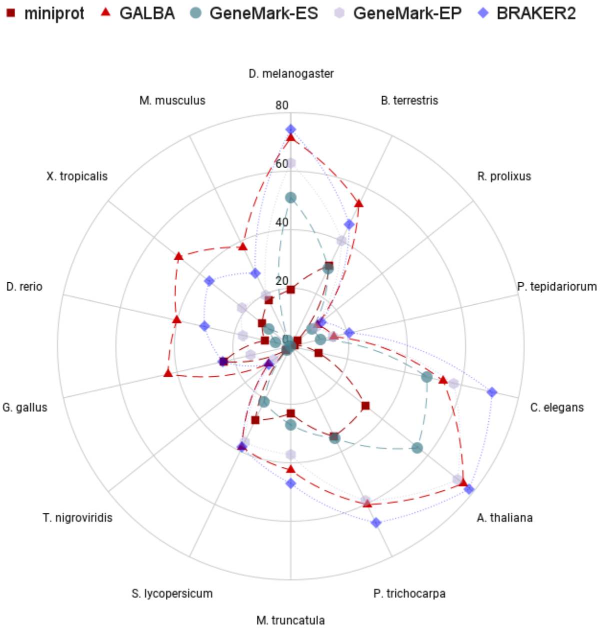
Donor proteins from

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

Accuracy of GALBA with Different Protein Donors



Gene F1 of GALBA in 14 Species



GALBA: Gene F1 Accuracy

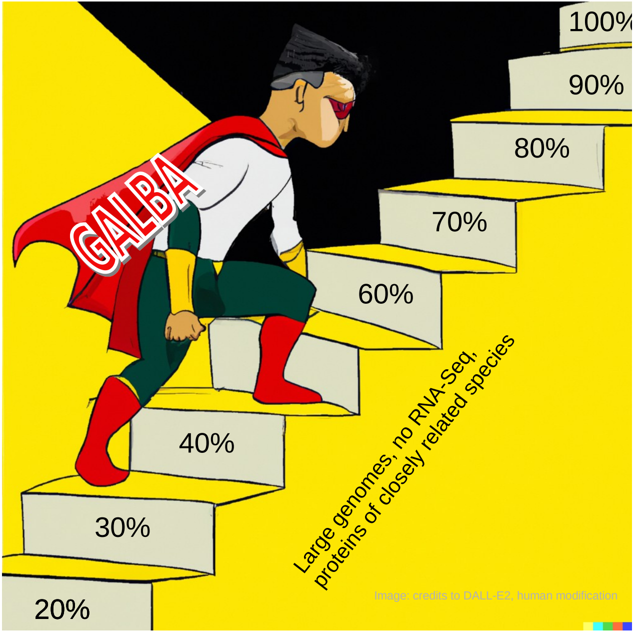


Image: credits to DALL-E2, human modification

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

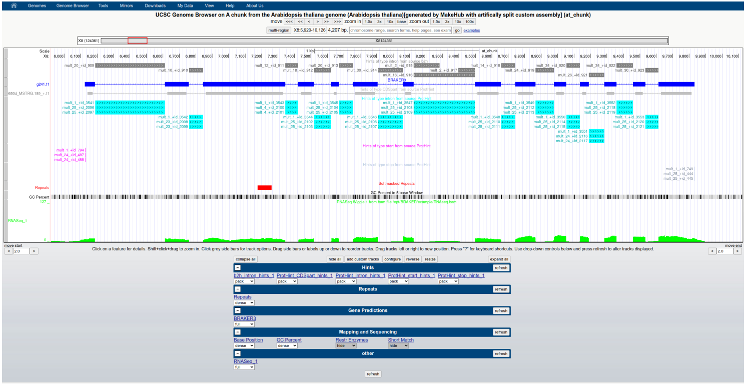
- GALBA: Artistic License
- miniprot: MIT License

Did We Do a Good Job?

Genome Browsers

Visualize your Annotation in Context with Evidence

- UCSC Genome Browser, MakeHub
- JBrowse
- ...



UCSC Genome Browser on A chunk from the Arabidopsis thaliana genome (Arabidopsis thaliana)[generated by MakeHub with artifically split custom assembly] (at_chunk)

X8:5,920-10,126 4,207 bp.

Click on a feature for details. Shift+click+drag to zoom in. Click grey side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. Use drop-down controls below and press refresh to alter tracks displayed.

Hints

b2h_intron_hints_1: pack
ProtHint_CDSpart_hints_1: pack
ProtHint_intron_hints_1: pack
ProtHint_start_hints_1: pack
ProtHint_stop_hints_1: pack

Repeats

Repeats: dense

Gene Predictions

BRAKER3: full

Mapping and Sequencing

Base Position: dense
GC Percent: dense
Restr Enzymes: hide
Short Match: hide

other

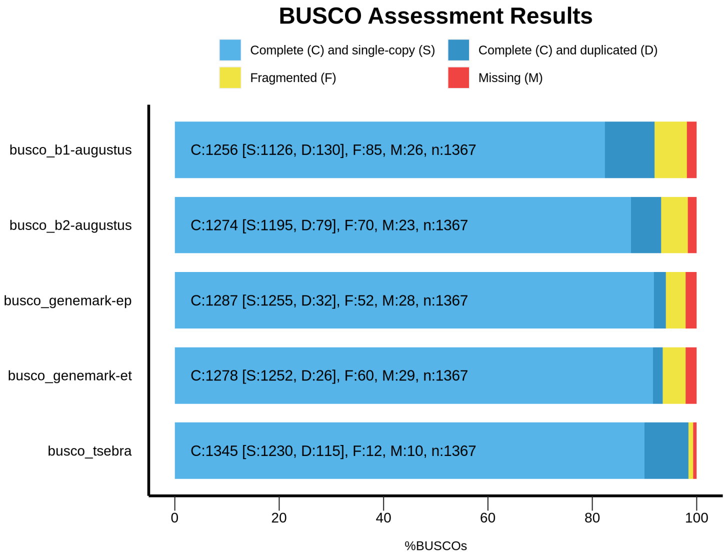
RNASeq_1: full

Describe Your Annotation

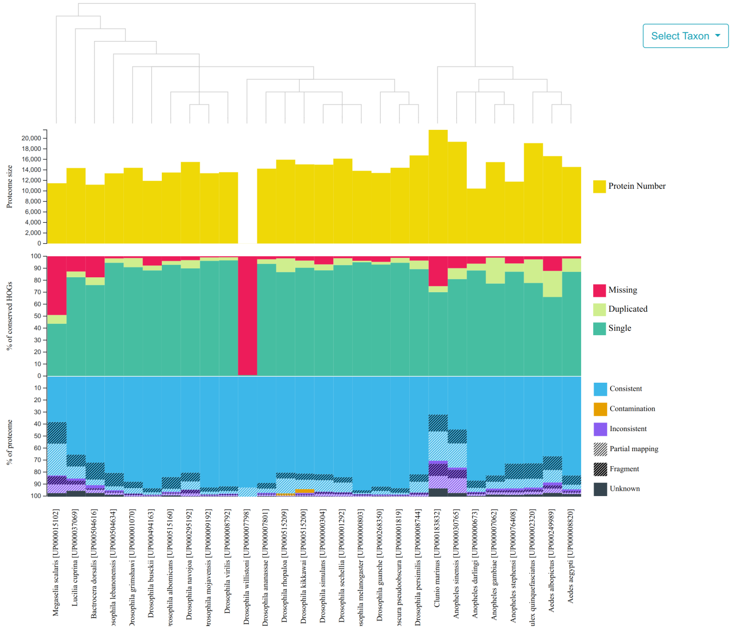
- number of genes
- number of transcripts
- ratio of mono-exonic to multi-exonic genes
- median number of exons per transcript
- maximal number of exons per transcript
- median transcript length
- ...

If possible, compare to annotated close relatives.
Consider effect of individual annotation pipelines.

BUSCO: Sensitivity in Clade-Specific Conserved Genes



OMArk: Sensitivity, Contaminations, & More



Hands on Lab Session

Organization of Session

1. start JupyterLab with Singularity
2. familiarize yourself with JupyterNotebooks (modify & run cells)
3. we walk you through GenomeAnnotation.ipynb
4. Annotate_Babesia_duncani.ipynb is for advanced learners, organized as group work

Getting Started

Go to `https://github.com/KatharinaHoff/GenomeAnnotation_Workshop2023` and follow instructions

Funding

- Research on BRAKER is supported by US National Institutes of Health grant GM128145 to M.B. and M.S.
- Research on GALBA is supported by German Research Foundation grant 277249973 to K.J.H., and by Project Data Competency granted to K.J.H. and M.S. by the government of Mecklenburg-Vorpommern, and by US National Institute of Health grant R01HG010040 to H.L., and by German Research Foundation grant 391397397 to S.H. and M.S.