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

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session

2023 Workshop on Genomics, Cesky Krumlov

Katharina J. Hoff & Natalia Nenasheva

Contact: katharina.hoff@uni-greifswald.de
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
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA
Katharina J. Hoff & Natalia Nenasheva

Contents

1 Who we are
2 Gene Prediction
3 BRAKER
   BRAKER1: RNA-Seq
   BRAKER2: Proteins
   TSEBRA
   BRAKER3: RNA-Seq + Proteins
   Accuracy Results
   Availability
4 GALBA
   GALBA: Proteins
   Accuracy Results
   Availability
5 Annotation Quality
   Genome Browsers
   Descriptive Statistics
   BUSCO
   OMArk
6 Hands on Lab Session
Information for Genome Annotation

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

Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Gene Finders Need Training Examples

Image: credits to DALL-E2, modified by human
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session

The BRAKER Team
University of Greifswald & Georgia Tech University

Lars Gabriel
Alexandre Lomsadze, Katharina Hoff, Tomáš Brůna

Mario Stanke

Mark Borodovsky

Also: Simone Lange, Matthias Ebel, Hannah Thierfeldt, Anica Hoppe
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

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

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)

**Who we are**
Gene Prediction

**BRAKER**
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins

**Accuracy Results**

**Availability**

**GALBA**
GALBA: Proteins

**Annotation Quality**
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

**Hands on Lab Session**

1.11

- 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

GeneMark-ET uses RNA-Seq for **Training**

**Anchors from RNA-Seq for training**

- 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**

- requires “prior data” for training
- uses intron information from RNA-seq for *prediction*
- no RNA-Seq assembly required
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

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

Tomáš Brůna¹, Katharina J. Hoff²,³, Alexandre Lomsadze⁴, Mario Stanke²,³,⁴ and Mark Borodovsky⁴,⁵,⁶,†

- 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

small & medium sized genomes, proteins of distantly related species
TSEBRA: transcript selector for BRAKER

Lars Gabriel\textsuperscript{1,2}, Katharina J. Hoff\textsuperscript{1,2}, Tomáš Brůna\textsuperscript{3}, Mark Borodovsky\textsuperscript{4,5} and Mario Stanke\textsuperscript{1,2}\textsuperscript{*}

- run both BRAKER1 and BRAKER2
- combine & increase accuracy
- 34 citations (Google Scholar)
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins

TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session

BRAKER1 + BRAKER2 → TSEBRA

BRAKER1
(RNA-Seq only)

BRAKER2
(protein only)

TSEBRA
(combiner tool)

genome.fa
RNASEq.bam

GeneMark-ET

AUGUSTUS training and prediction

augustus.gtf

genome.fa
protein_db.fa

GeneMark-EP

AUGUSTUS training and prediction

augustus.gtf

tsebra.gtf
TSEBRA: BRAKER1 + BRAKER2 Gene F1 Accuracy

Image: credits to DALL-E2, human modification

small & medium sized genomes
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction
BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability
GALBA
GALBA: Proteins
Accuracy Results
Availability
Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk
Hands on Lab Session

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

**Who we are**
Gene Prediction

**BRAKER**
- BRAKER1: RNA-Seq
- BRAKER2: Proteins
- TSEBRA

**BRAKER3: RNA-Seq + Proteins**

**Accuracy Results**
**Availability**

**GALBA**
- GALBA: Proteins
- Accuracy Results
- Availability

**Annotation Quality**
- Genome Browsers
- Descriptive Statistics
- BUSCO
- OMArk

**Hands on Lab Session**

Manuscript by Lars Gabriel et. al in preparation
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins

Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session

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. 3.1:1qaa108;
### BRAKER3 Experiments

#### Experiments

Accuracy assessment using genome-wide predictions in 6 species:

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

#### 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:

<table>
<thead>
<tr>
<th>Runtime (h)</th>
<th>BRAKER1</th>
<th>BRAKER2</th>
<th>GM-ETP</th>
<th>BRAKER3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>average</strong></td>
<td>06:26</td>
<td>09:01</td>
<td>06:03</td>
<td>17:55</td>
</tr>
</tbody>
</table>
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
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins

Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session

The GALBA Team
DOE Joint Genome Institute & Harvard Medical School & University of Greifswald

Tomáš Brůna
Heng Li
Lars Gabriel

Natalia Nenasheva
Matthias Ebel
Mario Stanke

Also: Daniel Honsel, & Steffen Herboldt

Also: Daniel Honsel, & Steffen Herboldt
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?

- no RNA-Seq, large genome
- proteins of few or 1 reference species
- fully open source
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session

GALBA: Proteins of Closely Related Species

development steps in D. melanogaster

Donor proteins from

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

Gene F1 (%)
60  65  70  75
(0) baseline  (1) best training  (2) miniprothint  (3) iterative training
combo  dsim
Accuracy of GALBA with Different Protein Donors

Drosophila melanogaster

Accuracy of GALBA with Different Protein Donors

- GALBA
- BRAKER2 combo
- BRAKER2 ODB10 species excl.
- BRAKER2 ODB10 order excl.

0.5 mutations per genomic site
Gene F1 of GALBA in 14 Species

![Gene F1 of GALBA in 14 Species diagram](image-url)
GALBA: Gene F1 Accuracy

Large genomes, no RNA-Seq, proteins of closely related species

20%
30%
40%
60%
70%
80%
90%
100%

GALBA: Gene F1 Accuracy

Who we are
Gene Prediction
BRAKER
- BRAKER1: RNA-Seq
- BRAKER2: Proteins
TSEBRA
- BRAKER3: RNA-Seq + Proteins
Accuracy Results
Availability

GALBA
- GALBA: Proteins
Accuracy Results
Availability

Annotation Quality
- Genome Browsers
- Descriptive Statistics
- BUSCO
- OMArk

Hands on Lab Session
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
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Did We Do a Good Job?
Automated and Accurate Genome Annotation with BRAKER, GALBA, & TSEBRA

Katharina J. Hoff & Natalia Nenasheva

Who we are
Gene Prediction

BRAKER
BRAKER1: RNA-Seq
BRAKER2: Proteins
TSEBRA
BRAKER3: RNA-Seq + Proteins

Accuracy Results
Availability

GALBA
GALBA: Proteins

Accuracy Results
Availability

Annotation Quality

Genome Browsers
Descriptive Statistics
BUSCO
OMArk

Hands on Lab Session
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

BUSCO Assessment Results

- **busco_b1-augustus**
  - C:1256 [S:1126, D:130], F:85, M:26, n:1367

- **busco_b2-augustus**
  - C:1274 [S:1195, D:79], F:70, M:23, n:1367

- **busco_genemark-ep**
  - C:1287 [S:1255, D:32], F:52, M:28, n:1367

- **busco_genemark-et**
  - C:1278 [S:1252, D:26], F:60, M:29, n:1367

- **busco_tsebra**
  - C:1345 [S:1230, D:115], F:12, M:10, n:1367
OMArk: Sensitivity, Contaminations, & More

OMArk: Sensitivity, Contaminations, & More

OMArk: Sensitivity, Contaminations, & More

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
• 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.