

Introduction

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

# Structural Genome Annotation with **BRAKER3**

## BGA23 Workshop

Katharina J. Hoff

Contact: katharina.hoff@uni-greifswald.de

## Introduction

### 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

# Eukaryotic Gene

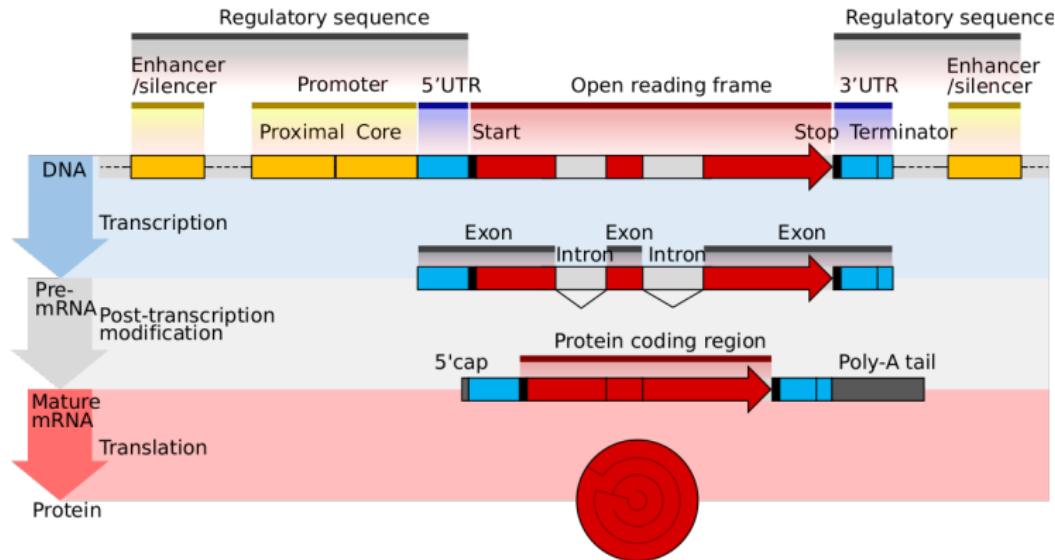
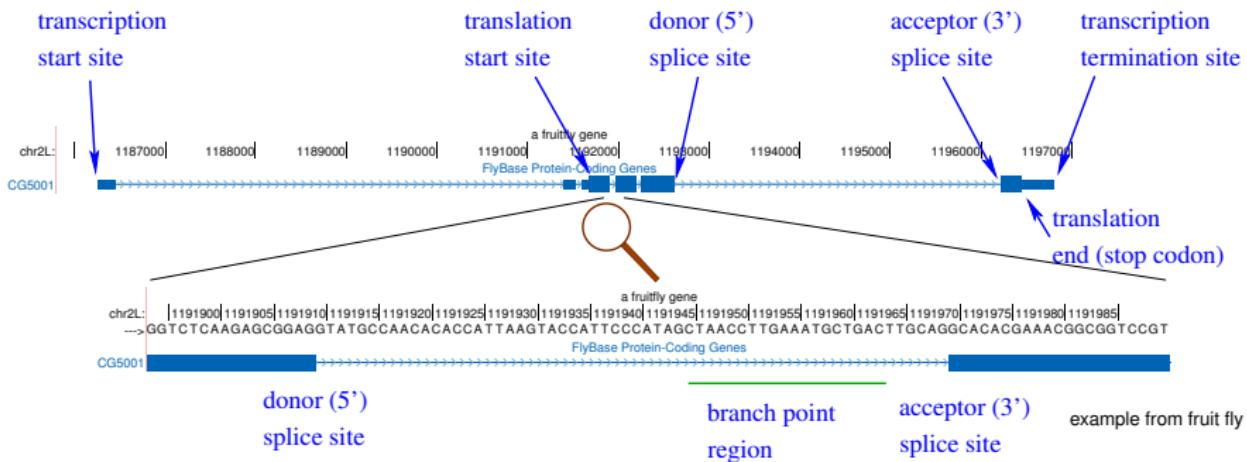


Image: Wikipedia, CC BY 4.0

## Information for Genome Annotation

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

# Signals



## 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

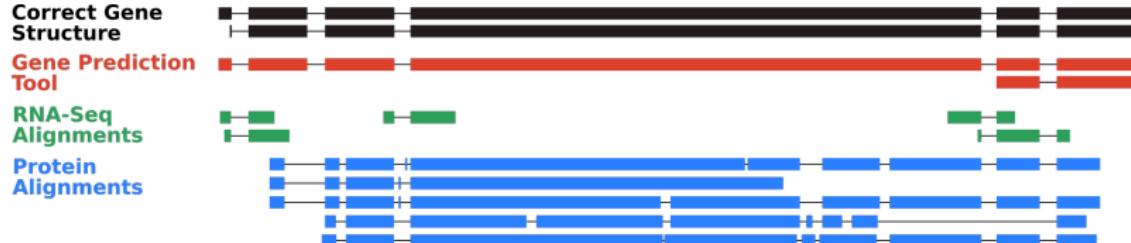
## Input

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

## Output

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

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



## Introduction

### BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

### GALBA

GALBA: Proteins

Accuracy Results

Availability

### Annotation Quality

Genome Browsers

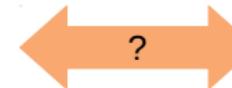
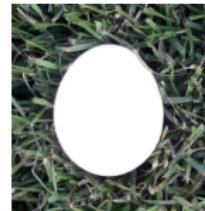
Descriptive Statistics

BUSCO

OMArk

### Hands on Lab Session

# Gene Finders Need Training Examples



Genes  
(Parameter training)

Genes  
(Prediction)

## Introduction

### 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



Image: credits to DALL-E2, modified by human



Introduction

**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



Introduction

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

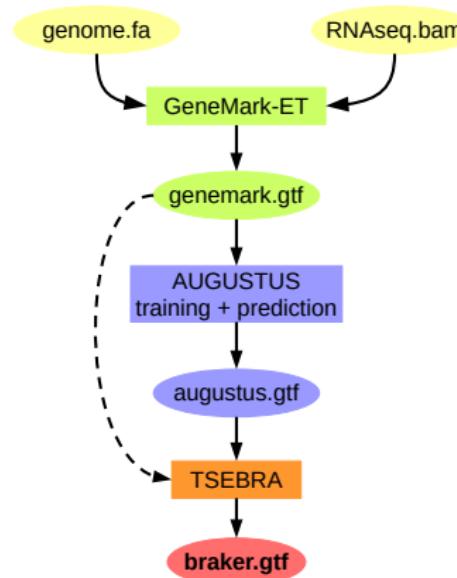
Hands on Lab Session

# BRAKER1: 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
- 1,218 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

Introduction

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

# GeneMark-ET uses RNA-Seq for Training

## Anchors from RNA-Seq for training

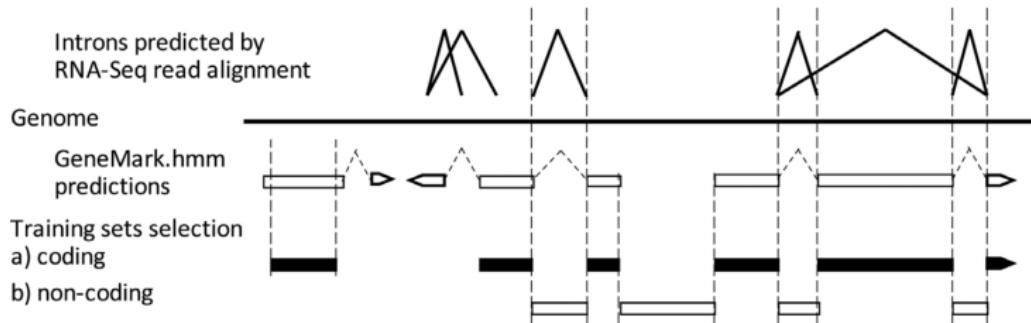


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one ‘anchored splice site’ as well as long exons predicted *ab initio* (>800 nt).

- 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

Introduction

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

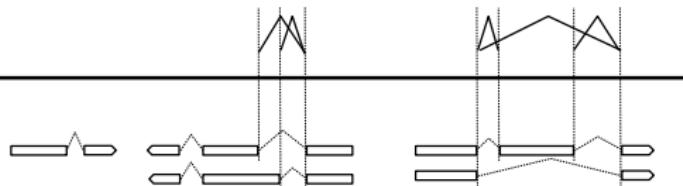
OMArk

Hands on Lab Session

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

Image: credits to DALL-E2, human modification

## Introduction

### BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

### GALBA

GALBA: Proteins

Accuracy Results

Availability

### Annotation Quality

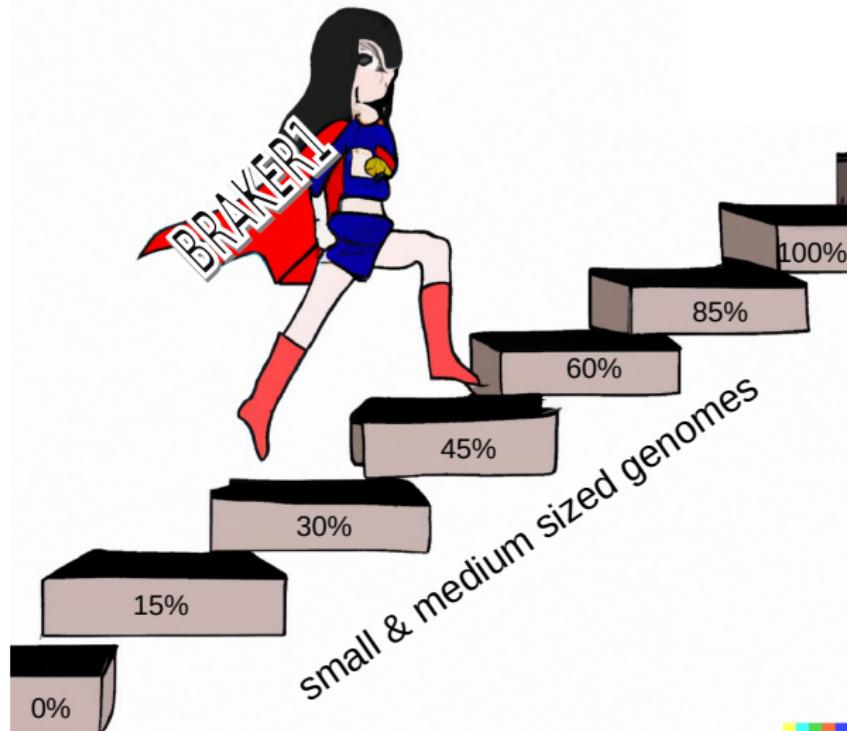
Genome Browsers

Descriptive Statistics

BUSCO

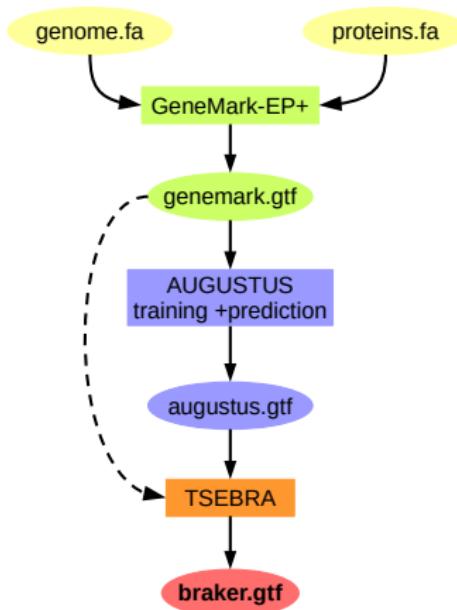
OMArk

### Hands on Lab Session



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

Tomáš Brůna<sup>1,†</sup>, Katharina J. Hoff<sup>2,3,†</sup>, Alexandre Lomsadze<sup>4</sup>, Mario Stanke<sup>2,3,‡</sup> and Mark Borodovsky<sup>2,4,5,\*;‡</sup>



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

# Evidence Usage by GeneMark-EP+ & AUGUSTUS During Prediction

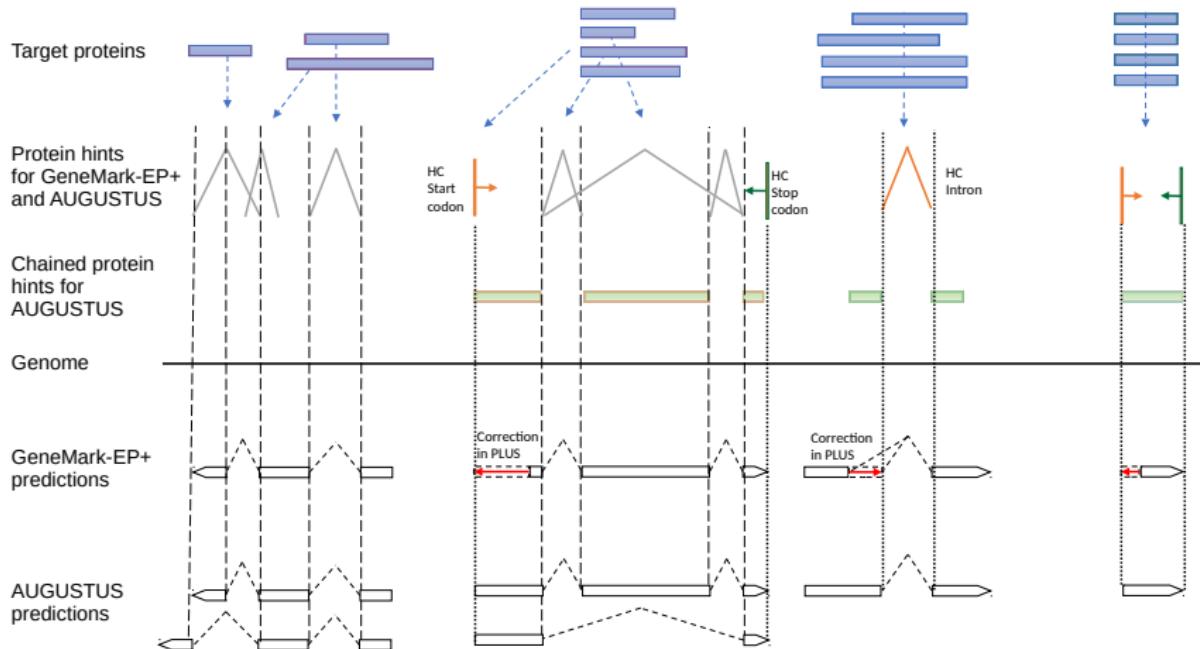
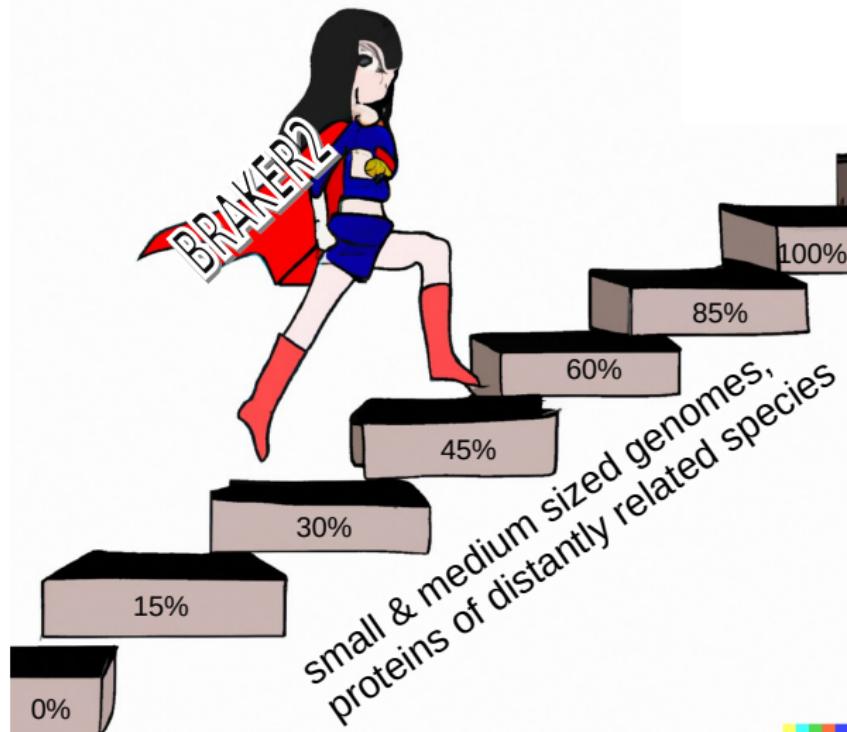


Image: credits to DALL-E2, human modification



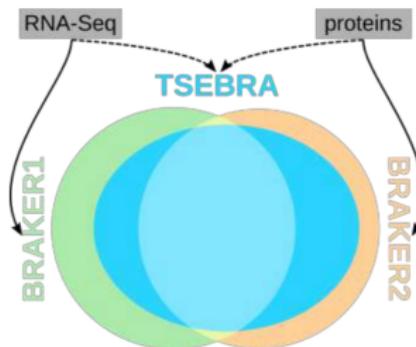
**SOFTWARE**

Open Access



# TSEBRA: transcript selector for BRAKER

Lars Gabriel<sup>1,2</sup>, Katharina J. Hoff<sup>1,2</sup>, Tomáš Brůna<sup>3</sup>, Mark Borodovsky<sup>4,5</sup> and Mario Stanke<sup>1,2\*</sup>



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

Introduction

BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

GALBA

GALBA: Proteins

Accuracy Results

Availability

Annotation Quality

Genome Browsers

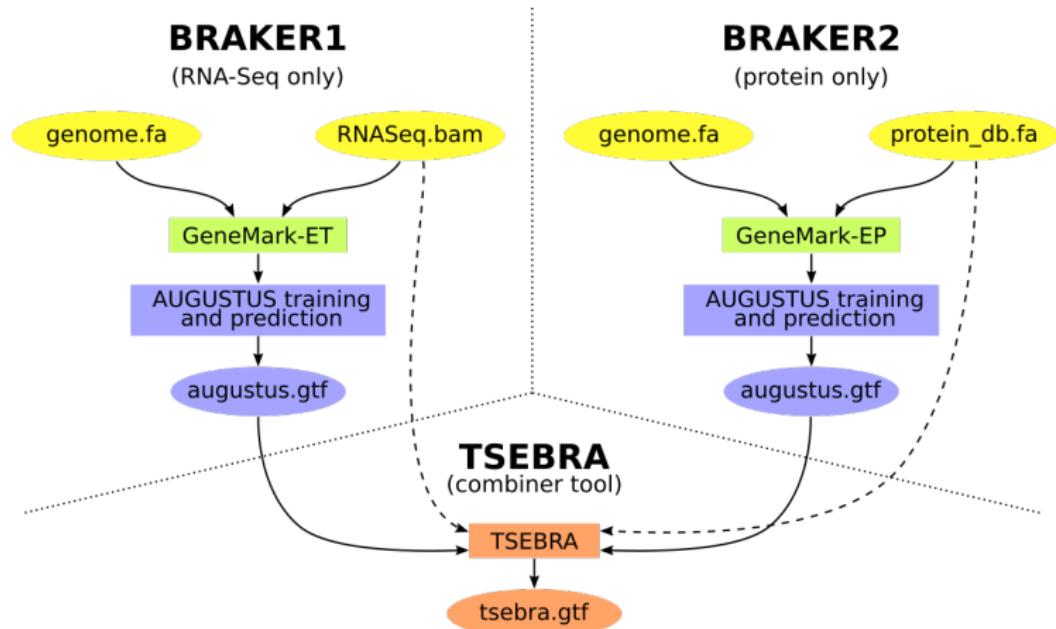
Descriptive Statistics

BUSCO

OMArk

Hands on Lab Session

# BRAKER1 + BRAKER2 → TSEBRA



Introduction

**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

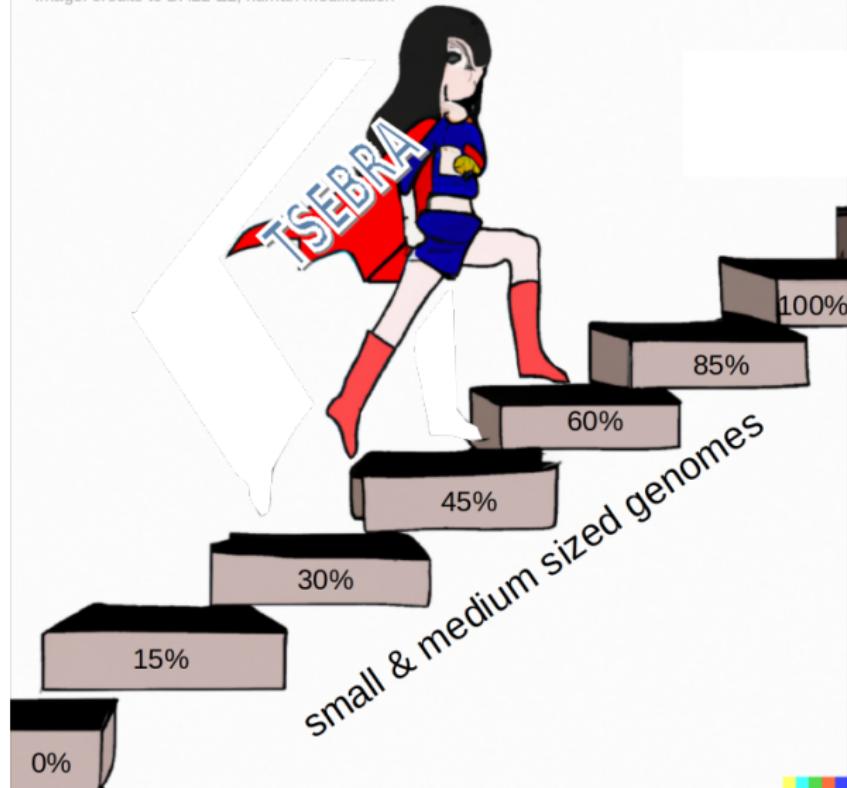
BUSCO

OMArk

Hands on Lab Session

# TSEBRA: BRAKER1 + BRAKER2 Gene F1 Accuracy

Image: credits to DALL-E2, human modification



Introduction

**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



Image: credits to DALL-E2, human modification



Introduction

BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

GALBA

GALBA: Proteins

Accuracy Results

Availability

Annotation Quality

Genome Browsers

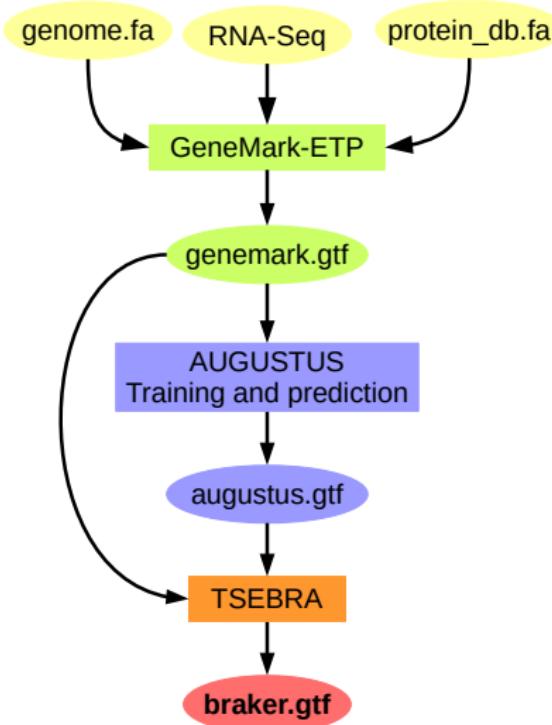
Descriptive Statistics

BUSCO

OMArk

Hands on Lab Session

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



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

Gabriel et al. (2023) [https://www.biorxiv.org/  
content/10.1101/2023.06.10.544449v1](https://www.biorxiv.org/content/10.1101/2023.06.10.544449v1)

## Introduction

### BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

### GALBA

GALBA: Proteins

Accuracy Results

Availability

### Annotation Quality

Genome Browsers

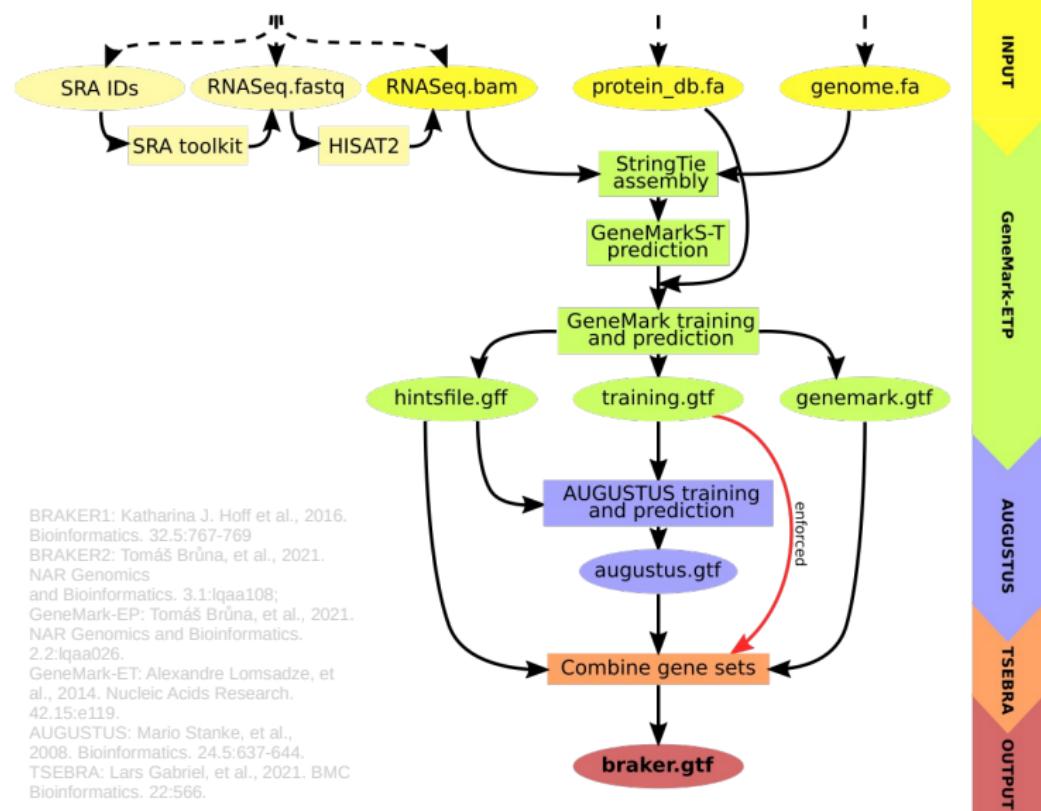
Descriptive Statistics

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: Tomás Brúna, et al., 2021.  
NAR Genomics  
and Bioinformatics. 3:1:lqaai08;  
GeneMark-EP: Tomás Brúna, 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.



Introduction

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 Experiments

## Experiments

Accuracy assessment using genome-wide predictions in 6 species:

Species	Genome Size (Mb)	# Genes in Annotation
<i>Arabidopsis thaliana</i> (thale cress)	119	27,444
<i>Caenorhabditis elegans</i> (nematode)	100	20,172
<i>Drosophila melanogaster</i> (fruit fly)	137	13,928
<i>Gallus gallus</i> (chicken)	1,040	17,279
<i>Mus musculus</i> (mouse)	2,650	22,378
<i>Solanum lycopersicum</i> (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]:** 
$$\frac{2 \cdot \text{Sensitivity} \cdot \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$

INPUT

GeneMark-ETP

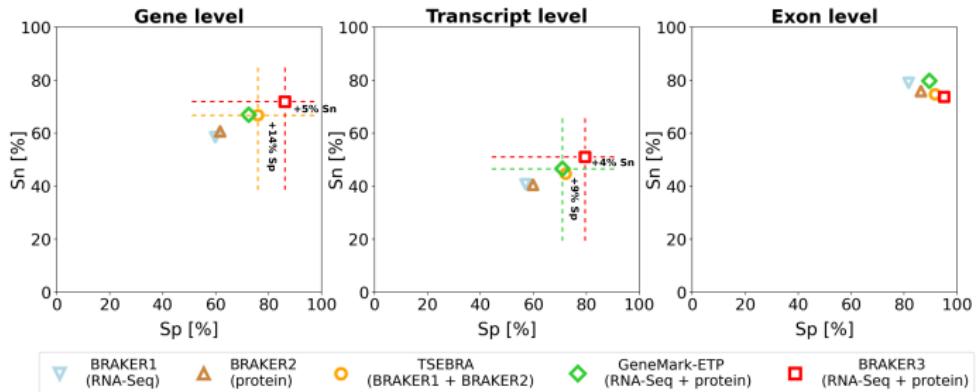
AUGUSTUS

TSEBRA

OUTPUT

# BRAKER3 Accuracy in Small Genomes

Average accuracy of genome-wide predictions



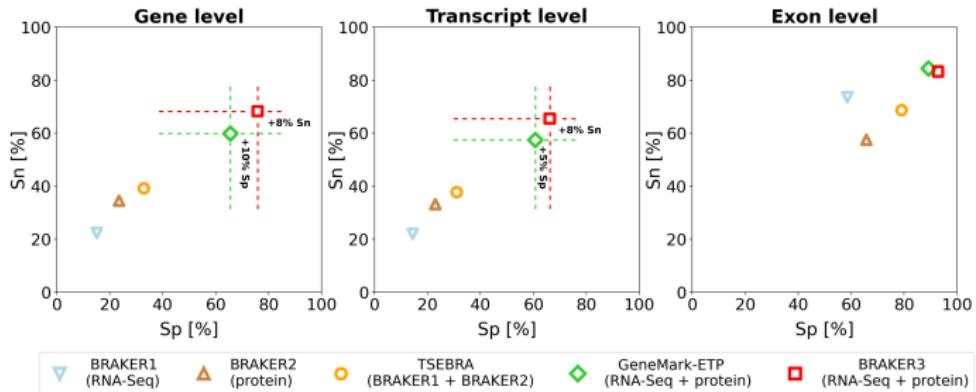
Species: *D. melanogaster*  
*A. thaliana*  
*C. elegans*

Extrinsic evidence:

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

# BRAKER3 Accuracy in Larger Genomes

Average accuracy of genome-wide predictions



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

Extrinsic evidence:

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

Introduction

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

# 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

Introduction

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/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](http://topaz.gatech.edu/genemark/license_download.cgi)

Introduction

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

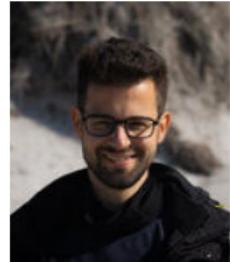
DOE Joint Genome Institute & Harvard Medical School  
& University of Greifswald



Tomáš Brůna



Heng Li



Lars Gabriel



Natalia Nenasheva



Matthis Ebel



Mario Stanke



Katharina Hoff

Also: Daniel Honsel, & Steffen Herboldt

## Introduction

### 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

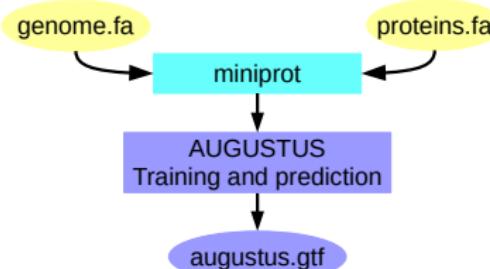
## Genome analysis

# 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

Introduction

BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

GALBA

GALBA: Proteins

Accuracy Results

Availability

Annotation Quality

Genome Browsers

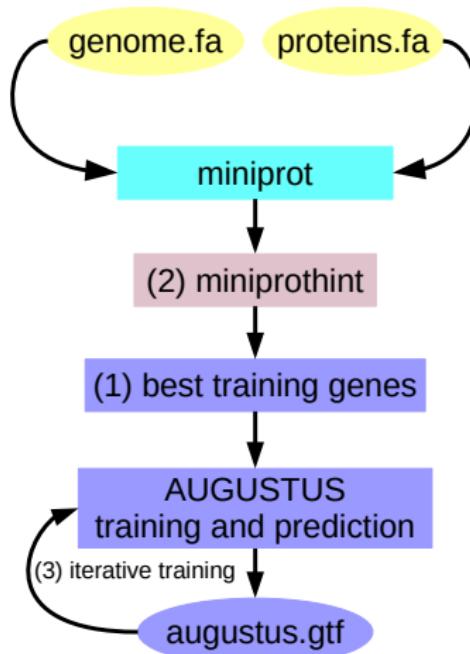
Descriptive Statistics

BUSCO

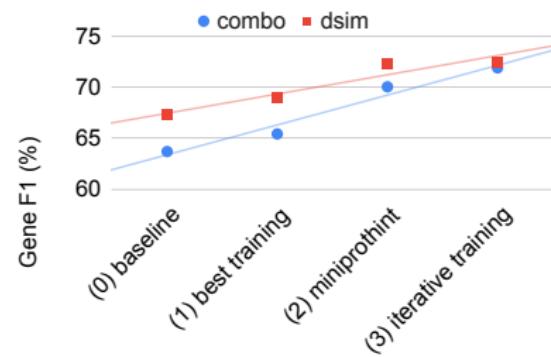
OMArk

Hands on Lab Session

# GALBA: Proteins of Closely Related Species



Development steps in *D. melanogaster*



Donor proteins from

`dsm` *D. simulans*

`combo` *D. ananassae*,  
*D. pseudoobscura*,  
*D. willistoni*,  
*D. virilis*,  
*D. grimshawi*

## Introduction

### 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

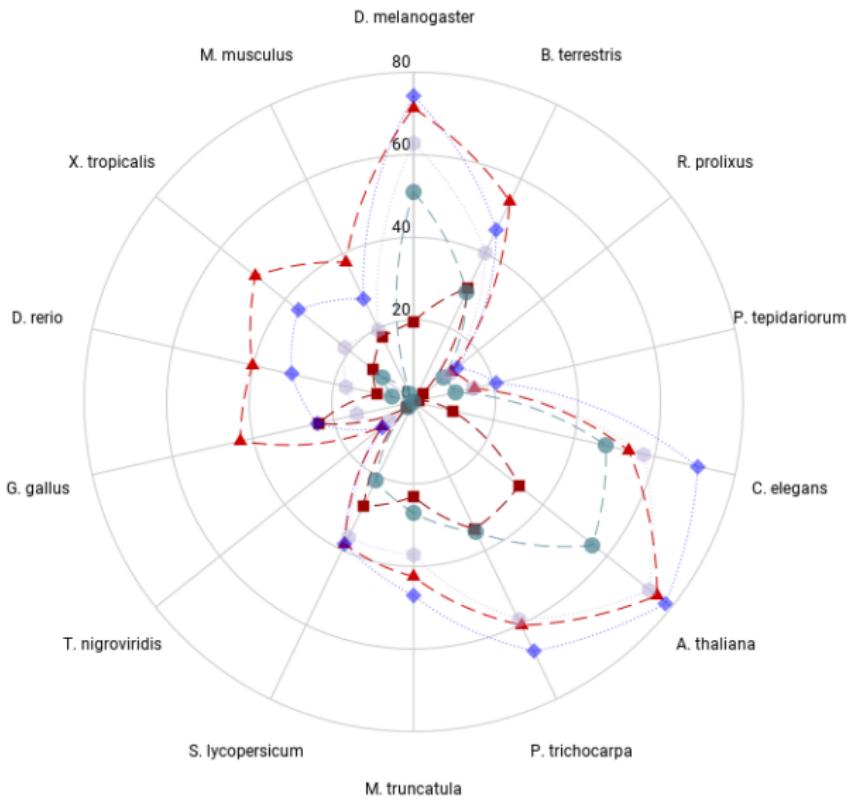
BUSCO

OMark

### Hands on Lab Session

# Gene F1 of GALBA in 14 Species

- miniprot
- ▲ GALBA
- GeneMark-ES
- GeneMark-EP
- ◆ BRAKER2



Introduction

**BRAKER**

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

**GALBA**

GALBA: Proteins

Accuracy Results

Availability

Annotation Quality

Genome Browsers

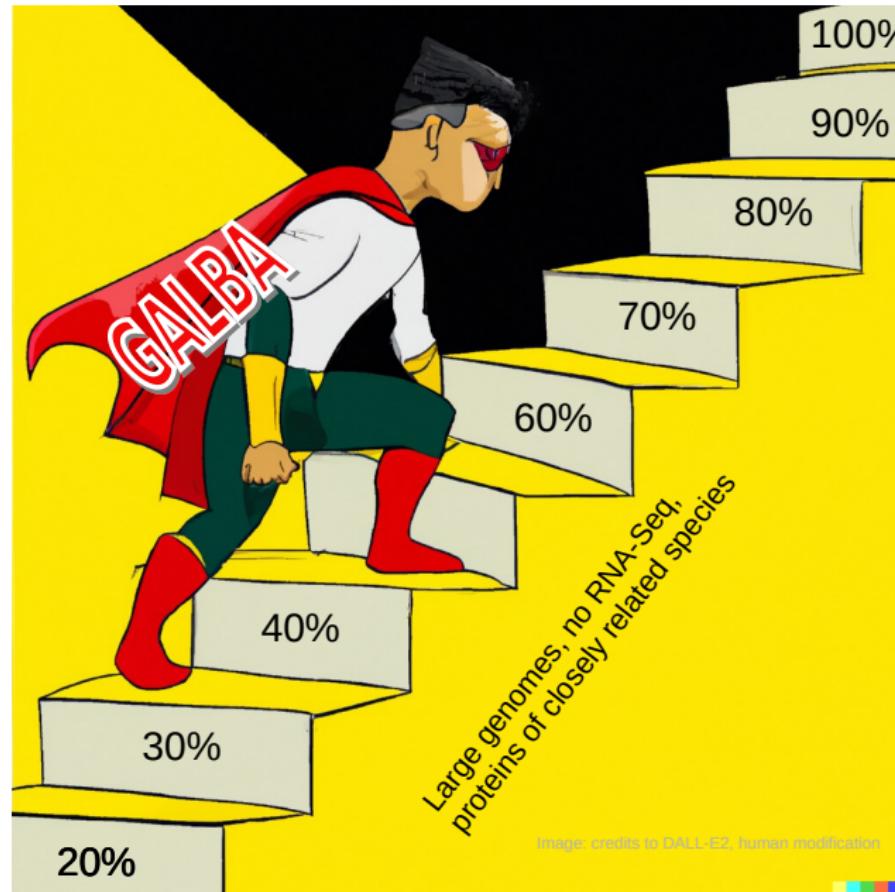
Descriptive Statistics

BUSCO

OMArk

Hands on Lab Session

# GALBA: Gene F1 Accuracy



Introduction

**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

# Did We Do a Good Job?



Introduction

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

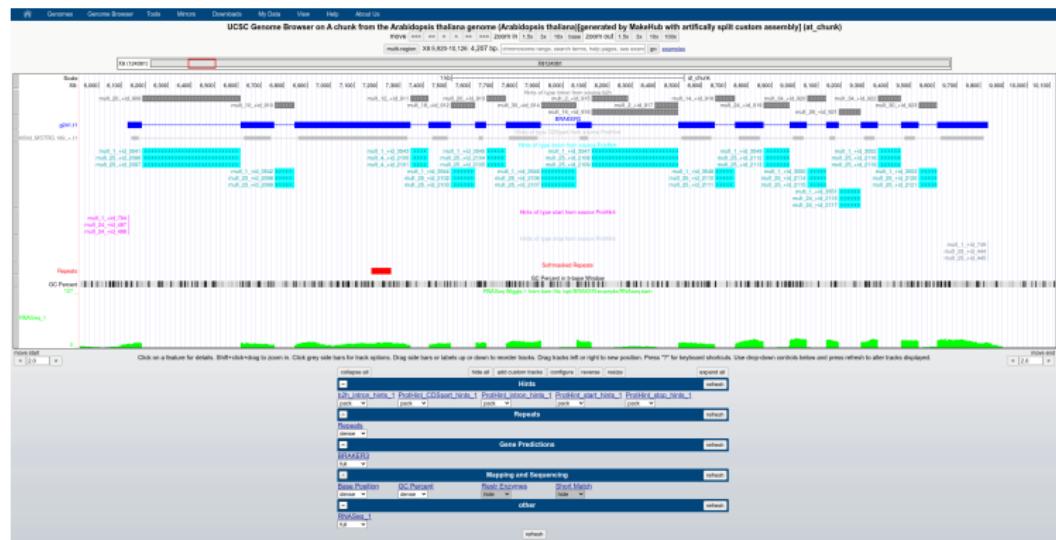
OMArk

Hands on Lab Session

# Genome Browsers

## Visualize your Annotation in Context with Evidence

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



Introduction

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

- 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.

## Introduction

### 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

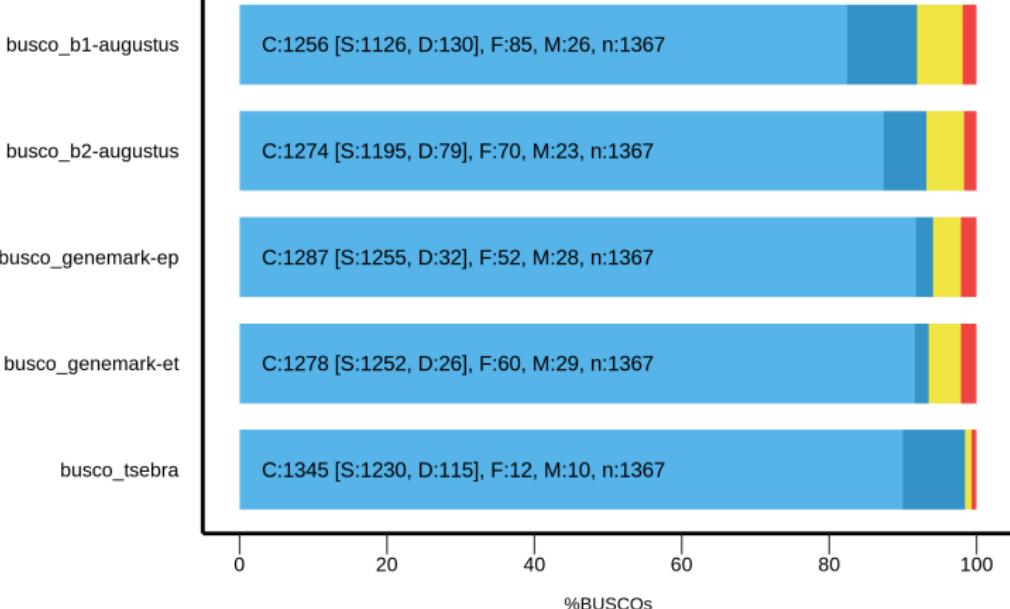
### BUSCO

OMArk

### Hands on Lab Session

# BUSCO: Sensitivity in Clade-Specific Conserved Genes

## BUSCO Assessment Results



# OMArk: Sensitivity, Contaminations, & More

Katharina J. Hoff



Select Taxon ▾

## Introduction

### BRAKER

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

BRAKER3: RNA-Seq +  
Proteins

Accuracy Results

Availability

### GALBA

GALBA: Proteins

Accuracy Results

Availability

### Annotation Quality

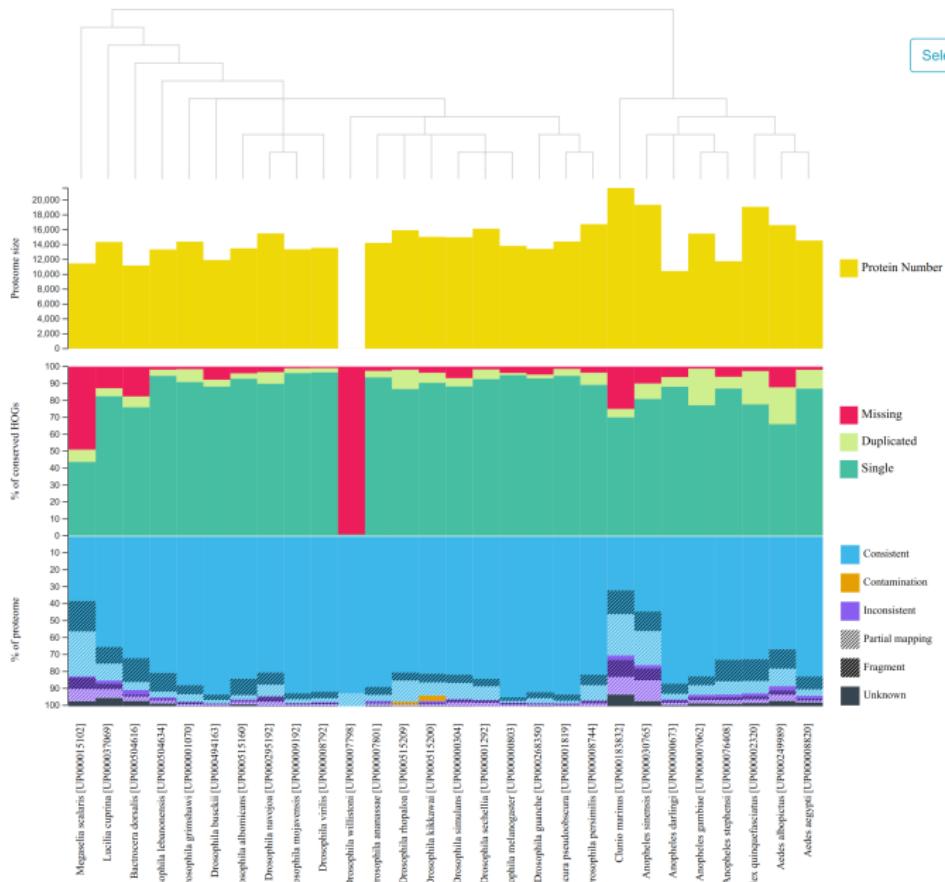
Genome Browsers

Descriptive Statistics

BUSCO

OMArk

### Hands on Lab Session



Introduction

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

# Hands on Lab Session

## Organization of Session

- ① start GitPod with the B3 repository
- ② open the file tutorial.md
- ③ we walk you through that file
- ④ copy & paste the BRAKER3 command into the terminal and execute
- ⑤ (runtime is 30 minutes with 8 threads)
- ⑥ we will inspect a pre-computed results folder
- ⑦ time for your questions and discussion

## Getting Started

Go to <https://gitpod.io/#https://github.com/BGAcademy23/B3>



Introduction

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

# Funding

## BRAKER

US National Institutes of Health grant GM128145 to M.B. and M.S.

## GALBA

- 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.