

Annotation of Protein Coding Genes

January 9th 2024

Katharina J. Hoff

Twitter: @katharina_hoff

Bluesky: @katharinahoff.bsky.social

Mastodon: @KatharinaHoff@fosstodon.org

E-Mail: katharina.hoff@uni-greifswald.de

Katharina J. Hoff

Group Leader in Applied Bioinformatics at University of Greifswald

2022 Habilitation (Greifswald)

2009 Ph.D. Molecular Biology (Göttingen)

2005 B.Sc. Plant Biotechnology (Hanover, stays abroad: Budapest & Alnarp)

- eukaryotic genome annotation, metagenomics
- best known for: **BRAKER** & other **Gaius-Augustus** software
- 31 peer-reviewed research articles with currently 5,535 citations

- currently 4 PhD students, 2 MSc students, 1 BSc student
- applied bioinformatics, programming, statistics, & data science

... I love to sail, have a dog, a cat, and a 7-years old daughter...

- understand what genome annotation in eukaryotes is
- know the basics of a Hidden Markov Model
- have a vague idea of INSDC annotation pipelines
- roughly understand methods sections on genome annotation
- know what's happening in MAKER, BRAKER, and GALBA
- have an idea of quality control methods

Materials at [https:](https://github.com/KatharinaHoff/GenomeAnnotation_Workshop)

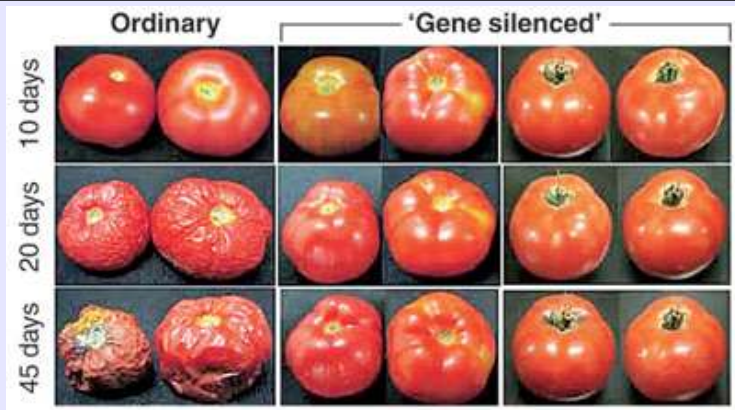
[//github.com/KatharinaHoff/GenomeAnnotation_Workshop](https://github.com/KatharinaHoff/GenomeAnnotation_Workshop)

Where are the protein coding genes?

Genomic sequence: chicken

```
cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtagactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcagggtgtaggcagagtccttttttctagtcattggctggcaaacagtgaggatctggggatggg
acaaaaggcagctaggaagattgccatgtagctctgctgctaagtgtagagctctagtagatattcagtaa
cattcaagttcctatttttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagctg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttcccc
atctgtcatatgaaggagtgcgatgtgttctgagactgaatccagttccaatcttctagattttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggttctggtagctccacagggtgctcaatgaagatgcaaaattagaagtaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgaggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccttgggaggcaagggtgaggg
ctggatttttaaaagggggcctgttttggggaggggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaagggggttcttggggaggaagttattcagaggacaggggaagcaggggaaggcag
acaggtcccatgagatattggaccaatttccttaaacatgctagaaaaacatgtggaaaagtcactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttccctaatctgggcaacc
cctggggccacagctaaatccagtgagtggaagttacagggagtgctgcttccagtgctgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagagggaagcaagtgctgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagtcctc
```

Examples for the importance of genome annotation



Sheeny et al. (1988) Proc. Natl. Acad. Sci. USA 85:8805-8809; Image: adapted from

<http://luisbarbosa2.blogspot.com/2013/06/flavr-savr-tomato.html>, Original: Asia Datta, Subhra Chakraborty, National Institute of Plant Genome Research, New Delhi

Examples for the importance of genome annotation

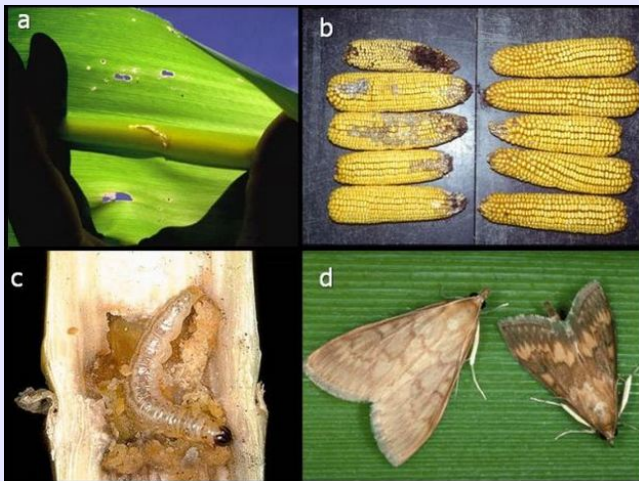


Image: Hellmich & Hellmich (2012) Nature Education Knowledge 3(10):4

http://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/46977030/1_2.jpg

Examples for the importance of genome annotation

Number of authors on genome papers more recently:

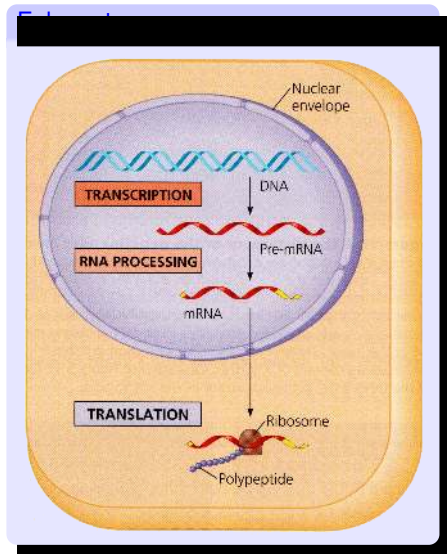
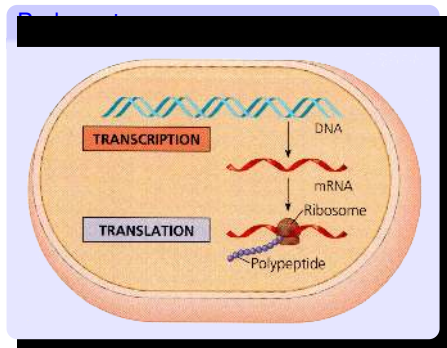
- **Wheat:** Gao et al. (2018) Gene 642, 284-292, **9 authors**
- **Goat:** Bickhart et al. (2017) Nature Genetics, 49(4), 643, **30 authors**
- **Wheat pathogenic fungus:** Plissonneau et al. (2016) MBio 7.5, e01231-16, **3 authors**
- **Quinoa:** Yasui et al. (2016) Dna Research, 23(6), 535-546, **16 authors**

In the past:

- Mosquito: Nene et. al (2007) Science doi: 10.1126/science.1138878 **95 authors**
- Human: International Human Genome Sequencing Consortium (2001) Nature, 409(6822), 860 **248 authors**

How does a cell recognize protein-coding genes?

Transcription & Translation



Images: Campbell et al., Biology, San Francisco, 2008, p. 329, Fig. 17.3

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

- every protein coding gene has an ORF
- not every ORF is a protein coding gene

How does a cell recognize protein-coding genes?

Eukaryotes: Splicing of introns

The Genome Annotation Problem

Genomic Sequence: chicken

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtagactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtcctttttctagtcattggctggcaaacagtgaggatctggggatggg
acaaaaggcagctaggaagattgccatgtagctctgctgctaagtgtagagctctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagactg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttcccc
atctgtcatatgaaggagtgcgatgtgttctgagactgaatccagttccaatcttctagattttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggtctggtagctccacagggtgctcaatgaagatgcaaaattagaagcaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgaggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccttgggaggcaagggtgaggg
ctggatttttaaaagggggcctgttttggggagggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggttcttggggaggaagttattcagaggacaggggaagcaggggaaggcag
acaggtcccatgagatattggaccaattccttaaacatgctagaaaaacatgtggaaaagtcactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttctaatctgggcaacc
cctggggccacagctaaatccagttagtggaagttacagggagctgtgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagagggaagcaagctgtgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagctcttc

The Genome Annotation Problem

Genomic sequence: chicken (1 gene: macrophage inflammatory protein-1 b)

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcaactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtcctttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagctctgctgctaagtgtagagctctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagctg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtcctctctgtgccttggtttccc
atctgtcatatgaaggagtgcatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggtctggtagctccacagggtgctcaatgaagatgcaaaattagaagtaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccctgggaggcaagggtgaggg
ctggatttttaaaagggggcctgttttggggagggggtgatgagcgtggggaggcagctctcagggtg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggttctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatattggaccaattccttaaacatgctagaaaaacatgtgaaaagtactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttccctaatctgggcaacc
cctggggccacagctaaatccagtgagtggaagttacaggagtgctgcttccagtgctgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagagggaagcaagtgctgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagtcttc

What aids in the identification of genes in genomes?

Evidence data from transcription

What aids in the identification of genes in genomes?

Evidence data from translation

What aids in the identification of genes in genomes?

Mathematical models

What aids in the identification of genes in genomes?

Mathematical models

Basis of highly accurate gene prediction tools

Hidden Markov Model

- There are only 2 nucleotides: A, B
- There are only 2 sequence states: intergenic (I), coding sequence (K)

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

- 1 The current value of the hidden state depends exclusively on the state of its predecessor.

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

- 1 The current value of the hidden state depends exclusively on the state of its predecessor.
- 2 The current value of the visible observation depends exclusively on the value of the current, hidden state.

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Basis of highly accurate gene prediction tools

Hidden Markov Model

Idea:

- 1 Generate all possible state sequences
- 2 Calculate the probability for each state sequence
- 3 Choose the state sequence with the highest probability

⇒ too expensive!

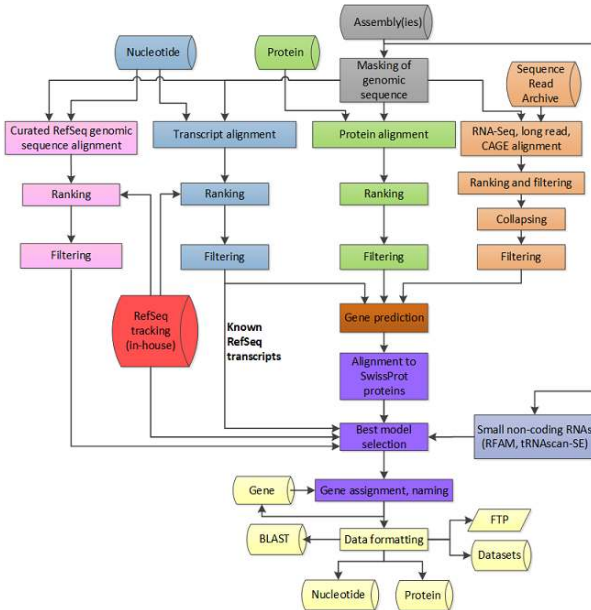
Basis of highly accurate gene prediction tools

Hidden Markov Model

Figure 1. The effect of the concentration of the monomer on the polymerization of *N*-vinylcarbazole initiated by *N*-vinylcarbazole. The polymerization was carried out at 60 °C for 24 h in the presence of 0.05 mol/L of *N*-vinylcarbazole. The concentration of the monomer was 0.05 mol/L (○), 0.1 mol/L (□), 0.2 mol/L (△), 0.3 mol/L (◇), 0.4 mol/L (×), 0.5 mol/L (●), 0.6 mol/L (◇), 0.7 mol/L (×), 0.8 mol/L (◇), 0.9 mol/L (×), 1.0 mol/L (◇).

Hidden Markov Model for gene identification in practice

Annotation at the NCBI: GNOMON



- Applied to a large number of genomes
- Can only be installed and used by the NCBI
- Formal description may be outdated
- Not benchmarked against other pipelines

Image: https://www.ncbi.nlm.nih.gov/core/assets/genome/images/Pipeline_sm_

- Katharina's best guess about what is happening under the hood

EBI: Ensembl annotation system

- **Ensembl vertebrate pipeline:** https://rapid.ensembl.org/info/genome/genebuild/full_genebuild.html
- **Ensembl non-vertebrate pipeline:** <https://rapid.ensembl.org/info/genome/genebuild/anno.html>
- **BRAKER2 in Ensembl:** <https://rapid.ensembl.org/info/genome/genebuild/braker.html>

- **Ensembl core species:**
<https://www.ensembl.org/info/about/species.html>
- **Everything from June 2022:**
<https://rapid.ensembl.org/info/about/species.html>

- Can (probably) only be installed and executed by EBI
- Not publicly benchmarked against other pipelines

Read your methods snippet

Structural annotation of protein coding genes only!

- 1 Go to `https://www.menti.com/7zgomub8sx`
- 2 Enter the names of tools involved in structural annotation of protein coding genes

Categorize tool names

Go to

<https://padlet.com/katharinahoff1/tools-for-structural-annotation-of-protein-coding-genes-c8desilwro11qp0h>

and sort the tools names from your methods snippet into categories

MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes

Brandi L. Cantarel,¹ Ian Korf,² Sofia M.C. Robb,³ Genis Parra,² Eric Ross,⁴ Barry Moore,¹ Carson Holt,¹ Alejandro Sánchez Alvarado,^{3,1} and Mark Yandell^{1,5}

¹Evans Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA; ²Department of Molecular and Cellular Biology and Genome Center, UC Davis, Davis, California 95616, USA; ³Department of Neurobiology & Anatomy, University of Utah School of Medicine, Salt Lake City, Utah 84132, USA; ⁴Howard Hughes Medical Institute, University of Utah School of Medicine, Salt Lake City, Utah 84132, USA

Cell and Tissue Res. 350:127-137
<https://doi.org/10.1007/s001070120001>



SOFTWARE

Open Access

MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects

MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations¹[W][OPEN]

Michael S. Campbell, MeiYee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, Rujia Achawanantakun, Dian Jiao, Carolyn J. Lawrence, Doreen Ware, Shio-Han Shiu, Kevin L. Childs, Yanni Sun, Ning Jiang, and Mark Yandell*

- first highly popular community annotation pipeline
- free for academic purposes
- good tutorials
- very runtime consuming
- efficiently parallelized
- accuracy today: low (see BRAKER publications)

MAKER Workflow

MAKER: important facts for users

- + MAKER is highly flexible:
 - ▶ use one or all gene finders
 - ▶ use proteins and/or transcriptome evidence
- MAKER does not automatically train any gene finder
- MAKER was designed to execute pre-trained gene finders
- + Tutorial(s) provide suggestions on how to train gene finders in a multi-step process via MAKER
- MAKER does generate repeat libraries but performs repeat masking
- + MAKER annotates also tRNAs and snoRNAs
- + MAKER tutorials are very helpful, easy to run
- Authors of BRAKER have repeatedly optimized MAKER protocols and assessed upper boundary limits of accuracy → below expectations!
- Download Software from MAKER website

Improving the field of fully automated genome annotation

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, Neng Huang

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
- 1,312 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

Annotation of RNA-Seq reads

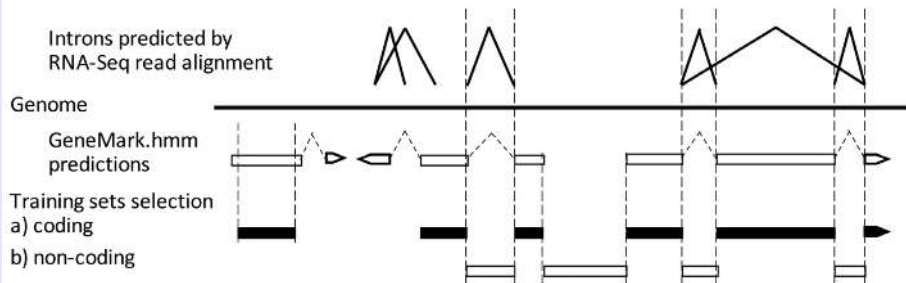


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- employs unsupervised training
- training includes 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

- requires “prior data” for training
- uses intron information from RNA-seq for *prediction*
- no RNA-Seq assembly required
- optional input: BUSCO lineage (compleasm)

Measuring accuracy of genome annotation

Accuracy assessment after applying tool to genome with reference annotation:

Species	Genome Size (Mb)	# Genes in Annotation
<i>Arabidopsis thaliana</i> (thale cress)	119	27,444
<i>Bombus terrestris</i> (bumble bee)	249	10,581
<i>Caenorhabditis elegans</i> (nematode)	100	20,172
<i>Danio rerio</i> (zebrafish)	1,345	25,611
<i>Drosophila melanogaster</i> (fruit fly)	137	13,928
<i>Gallus gallus</i> (chicken)	1,040	17,279
<i>Medicago truncatula</i> (barrelclover)	420	44,464
<i>Mus musculus</i> (mouse)	2,650	22,378
<i>Parasteatoda tepidariorum</i> (house spider)	1,445	18,602
<i>Populus trichocarpa</i> (poppy)	389	34,488
<i>Solanum lycopersicum</i> (tomato)	772	33,562

Precision = Specificity: Percentage of correctly found genes/transcripts/exons in the **predicted gene set**.

Recall = Sensitivity: Percentage of correctly found genes/transcripts/exons in the **reference annotation**.

F1-Score:
$$\frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

BRAKER1 gene F1 accuracy

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

Tomáš Brůna^{1,†}, Katharina J. Hoff^{2,3,†}, Alexandre Lomsadze⁴, Mario Stanke^{2,3,‡} and Mark Borodovsky^{4,5,*}

- spliced alignments of a large number of proteins (e.g. OrthoDB partition)
- optional input: BUSCO lineage (compleasm)
- 719 citations (Google Scholar)

Evidence usage by GeneMark-EP+ & AUGUSTUS during prediction

BRAKER2 gene F1 accuracy

BRAKER3 gene F1 accuracy - climbing the top

BRAKER3: using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA

- Gabriel *et al.* (2023), bioarxiv
- 15 citations (Google Scholar)
- spliced aligned and **assembled** RNA-Seq
- large protein database
- optional input: BUSCO lineage (compleasm)
- combines GeneMark-ETP and AUGUSTUS gene sets with TSEBRA

BRAKER3: using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA

SOFTWARE

Open Access

TSEBRA: transcript selector for BRAKER



Lars Gabriel^{1,2}, Katharina J. Hoff^{1,2}, Tomáš Brůna³, Mark Borodovsky^{4,5} and Mario Stanke^{1,2*} 

- combine several gene sets
- increases accuracy
- originally for combining BRAKER1 & BRAKER2 gene sets
- today combines GeneMark-ETP & AUGUSTUS in BRAKER
- 67 citations (Google Scholar)
- **may discard BUSCOs**

Genome analysis

compleasm: a faster and more accurate reimplementa- tion of BUSCO

Neng Huang ^{1,2} and Heng Li^{1,2,*}

- originally developed for BUSCO detection in genomes
- recently extended to BUSCO detection in proteins

⇒ This can solve our BRAKER-BUSCO problem

Improving BRAKER with Compleasm

Scenario 1: Good Evidence

Improving BRAKER with Compleasm

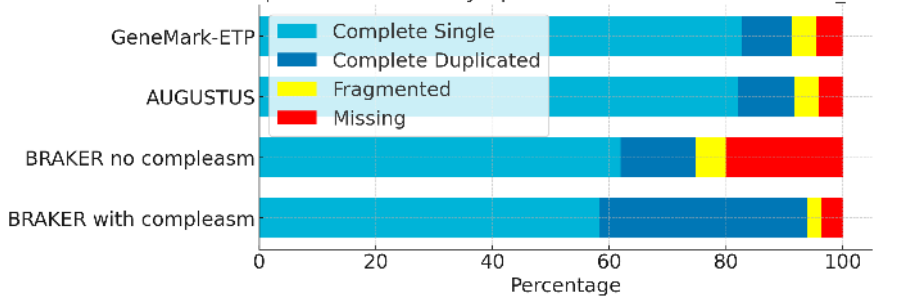
Scenario 1: Poor Evidence

Improving BRAKER with Compleasm

Scenario 1: Poor Evidence

Input data see Schiebelhut *et al.* (2023) <https://doi.org/10.1093/jhered/esad054>

compleasm BUSCOs in Pycopodia helianthoides with metazoa odb10



	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

Related seastar *Asterias rubens* has 19,938 genes

Accuracy of genome annotation approaches by BRAKER team

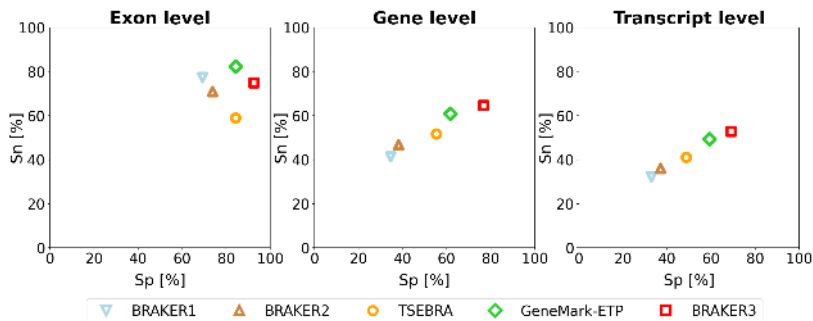


Figure 2: Average specificity and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in Supplemental Table S1). Inputs were the genomic sequences, short-read RNA-Seq libraries and protein databases (*order excluded*).

Image: Gabriel *et al.* (2023), biorxiv

Accuracy of genome annotation approaches including competitive tools

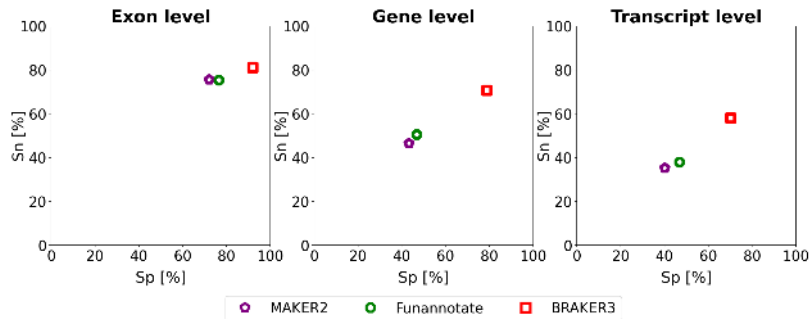


Figure 5: Average specificity and sensitivity of gene predictions made by MAKER2, Funannotate, and BRAKER3 for a subset of 8 species (excluding the mouse, spider and fish genome). Inputs were the genomic sequences, short-read RNA-Seq libraries, and protein databases (*close relatives included*). The accuracy of MAKER2 reported here can be regarded as an upper limit of what can be expected when annotating a previously unannotated genome (see Experiments section).

Image: Gabriel *et al.* (2023), biorxiv

GitHub

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

Docker/Singularity

```
singularity build braker.sif \  
    docker://teambraaker/braker:latest  
  
singularity exec braker.sif braker.pl [OPTIONS]
```

License

- BRAKER: Artistic License
- most components under open source software licenses
- GeneMark-ETP: CC BY-NC

GALBA Contributors



Tomáš Brůna



Heng Li



Joseph Guhlin



Lars Gabriel



Natalia Nenasheva



Ethan Tolman



Paul Frandsen



Matthis Ebel



Mario Stanke



Katharina Hoff

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.”

Summary

Protein-to-genome alignment with miniprot

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

GALBA: using proteins of related species

dsim	<i>D. simulans</i>
combo	<i>D. ananassae</i> , <i>D. pseudoobscura</i> , <i>D. willistoni</i> , <i>D. virilis</i> , <i>D. grimshawi</i>

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

<https://doi.org/10.1101/2023.12.11.569651>

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

Important: **If you have RNA-Seq, use it! It's better!**

GitHub

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

Docker (Singularity)

```
singularity build galba.sif \  
    docker://katharinahoff/galba:latest  
  
singularity exec galba.sif galba.pl [OPTIONS]
```

License

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

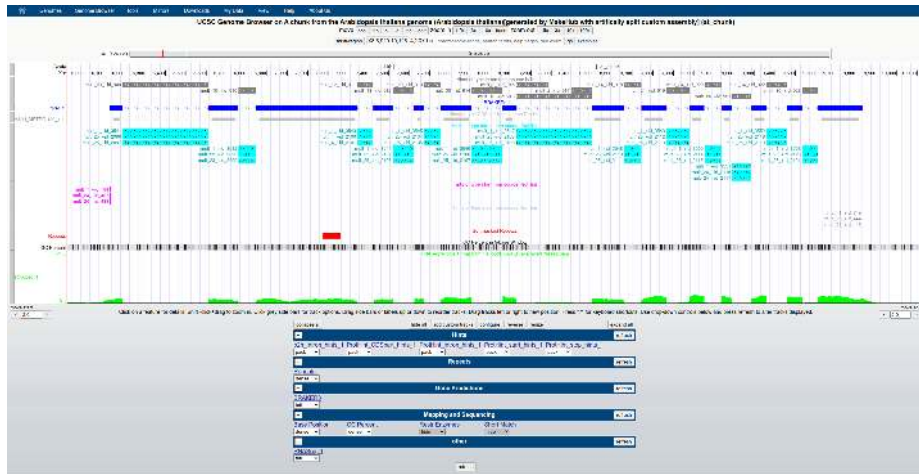
Did We Do a Good Job?



Genome Browsers

Visualize your Annotation in Context with Evidence

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



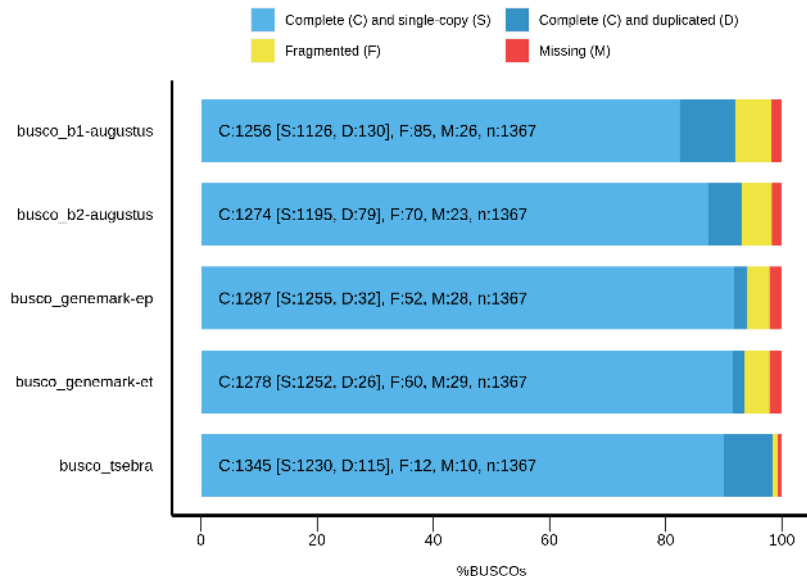
Describe Your Annotation

- 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



Revisiting annotation methods sections

- 1 Read your methods snippet, again
- 2 Use our categorized tool name board if you are still unsure what a tool does
- 3 Ask if you remain unsure what a method is good for
- 4 Fill the poll at <https://www.menti.com/alsied94gmi6>

Most important stuff on genome annotation

- structural genome annotation in eukaryotes is hard
- Hidden Markov Models are essential
- evidence helps a lot
- majority of genomes is annotated by large centers
- popular community annotation pipelines:
 - 1 MAKER
 - 2 BRAKER
 - 3 (GALBA may become popular)
- accuracy matters
- "looking nice" is not always "correct"
- BUSCO completeness is a widely used sensitivity measure
- OMArk might be more appropriate