

Annotation of Protein Coding Genes

January 8th 2025

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

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

- eukaryotic genome annotation, metagenomics
- best known for: **BRAKER** & other **Gaius-Augustus** software
- 37 peer-reviewed research articles with currently 7,186 citations

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

... I love to sail, have a dog, a cat, and an 8-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 BRAKER & GALBA
- be aware of the rapid advances with Deep Learning
- have an idea of quality control methods

Materials at

https:

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

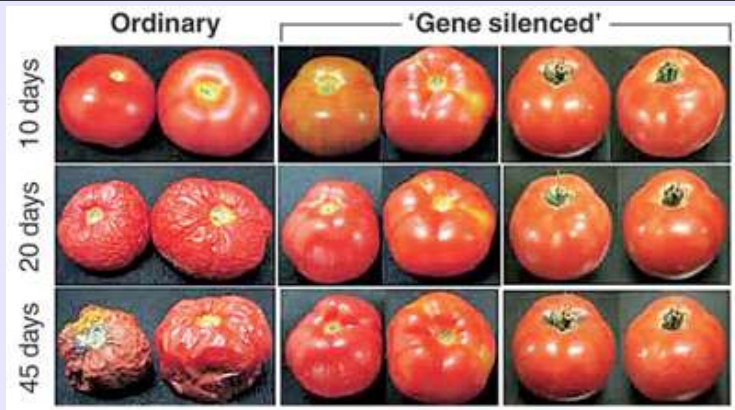
(Some images have been removed on Github because I do not have permission to share them.)

Where are the protein coding genes?

Genomic sequence: chicken

```
cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtagactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgaggatctggggatggg
acaaaaggcagctaggaagattgccatgtagctctgctgctaagtgtagagctctagtagatattcagtaa
cattcaagttcctatttttcttaagaattagcaaccagcagaggaaaacgatgggctggaagttagactg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttcccc
atctgtcatatgaaggagtgcgatgtgttctgagactgaatccagttccaatcttctagattttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggttctggtagctccacagggtgctcaatgaagatgcaaaattagaagtaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgaggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggttagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccttgggaggcaagggtgaggg
ctggatttttaaaagggggcctgttttggggaggggggtgatgagcgtggggaggcagctctcagggtg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaagggggttcttggggaggaagttattcagaggacaggggaagcaggggaaggcag
acaggtcccatgagatattggaccaatttccttaaacatgctagaaaaacatgtggaaaagtactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttcctaattctgggcaacc
cctggggccacagctaaatccagttagtggaagttacagggagctctgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagaggcaagcaagtctgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagtcttc
```

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

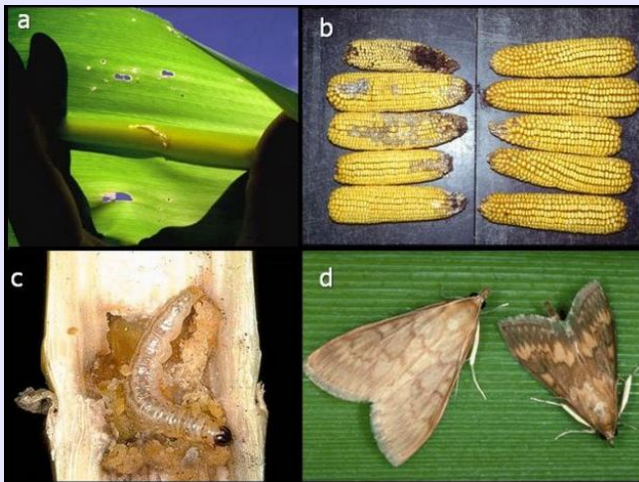


Image: Hellmich & Hellmich (2012) Nature Education Knowledge 3(10):4

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

It does not take a village to publish a genome!

- In the past:
 - ▶ Human: International Human Genome Sequencing Consortium (2001), Nature 409(6822), 860 **248 authors**
 - ▶ Mosquito: Nene et. al (2007) **95 authors**

It does not take a village to publish a genome!

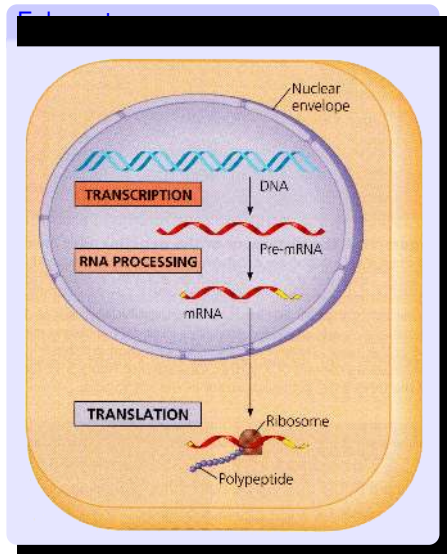
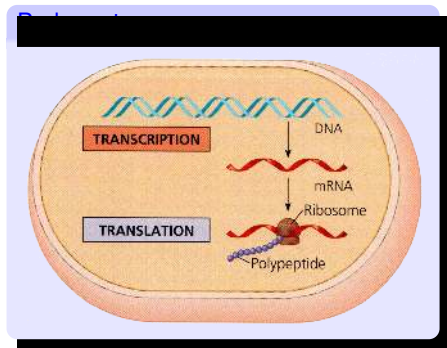
- In the past:
 - ▶ Human: International Human Genome Sequencing Consortium (2001), Nature 409(6822), 860 **248 authors**
 - ▶ Mosquito: Nene et. al (2007) **95 authors**
- More recently:
 - ▶ 4 *Botrytis cinerea*: Adhikari et al. (2025), **5 authors**
 - ▶ European harvest mouse: O'Brien & Colom (2024), **2 authors**
 - ▶ Great wood-rush: Goodwin et al. (2024), **4 authors**

It does not take a village to publish a genome!

- In the past:
 - ▶ Human: International Human Genome Sequencing Consortium (2001), Nature 409(6822), 860 **248 authors**
 - ▶ Mosquito: Nene et. al (2007) **95 authors**
- More recently:
 - ▶ 4 *Botrytis cinerea*: Adhikari et al. (2025), **5 authors**
 - ▶ European harvest mouse: O'Brien & Colom (2024), **2 authors**
 - ▶ Great wood-rush: Goodwin et al. (2024), **4 authors**
- **You can do it!**

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
ggaccataatttgggggttctggtagctccacagggtgctcaatgaagatgcaaaattagaagcaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgaggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgcctgggaggcaagggtgaggg
ctggatttttaaggggggcctgttttggggagggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaagggggttctggggaggaagttattcagaggacaggggaagcaggggaaggcag
acaggtcccatgagatattggaccaattccttaaacatgctagaaaaacatgtggaaaagtcactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttctaatctgggcaacc
cctggggccacagctaaatccagtgagtggaagttacagggagctgtgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagagggaagcaagctgtgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagctctt
```

The Genome Annotation Problem

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

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcaactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtcctttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaagtgtagagctctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagctg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttccc
atctgtcatatgaaggagtgcatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggtctggtagctccacagggtgctcaatgaagatgcaaaattagaagtaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccctgggaggcaaggggtgaggg
ctggatttttaaaagggggcctgttttggggaggggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgtgaaggcggctgagt
ggcagccgagacagaaggggttctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaacatgctagaaaaacatgtggaaaagtcactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttccctaatctgggcaacc
cctggggccacagctaaatccagtgagtggaaattacagggagctctgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaacaaaagagggaagcaagctctgcgctgacccagtgagtctgggtccaggagtagctgtatgac
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 Fe^{2+} on the adsorption of Pb^{2+} by the Fe^{2+} -modified bentonite. The concentration of the Pb^{2+} was 100 mg/L, the pH was 5.0, and the adsorption time was 24 h.

Hidden Markov Model for gene identification in practice

EBI: Ensembl annotation system

Ensembl annotation pipelines

Ensembl annotation pipeline for non-vertebrates

- **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:** <https://beta.ensembl.org/>

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

- Internal: The NCBI Eukaryotic Genome Annotation Pipeline (EGAP)
- (Internal: RefSeq curation)
- You can run it: **EGAPx**

Annotation with EGAPx (NCBI)

Annotation with EGAPx (NCBI)

- Containerized with Docker/Singularity
- Documentation: <https://github.com/ncbi/egapx>
- Currently supported clades (protein sets):
 - ▶ Chordata
 - ▶ Insecta
 - ▶ Arthropoda
 - ▶ Monocots
 - ▶ Eudicots
- Easy to use
- Benchmarking possible: good accuracy!

Read your methods snippet

Focus on structural annotation of protein coding genes only!

- 1 We move to Wooclap
- 2 Enter the names of tools involved in structural annotation of protein coding genes

Categorize tool names

Go to

<https://shorturl.at/uA0Tg>

and sort the tools names from your methods snippet into categories



Categorize tool names

Bioinformatics Tools for Genome Annotation

Sort tools into categories

Protein-to-Genome Aligners

Diamond

Not a spliced aligner, a fast-trapper. Does not directly return evidence for gene finding but can be used to select candidates for accurate spliced alignment to save runtime. Also used for removing redundancy in training gene sets (in protein-protein mode)

TBLASTN

Not a spliced aligner, a fast-trapper. Does not directly return evidence for gene finding. Not much used in the field anymore because DIAMOND is faster

GenomeThreader

Not used much anymore because pinpoint is faster and more robust towards increasing distance between donor and recipient

miniprot

very fast

Spaln

Spaln2 is also very fast

(GeMoMa)

Does not require protein sequence as input but the genomes and annotations of closely related species. Returns accurate gene-models if mapping is successful

Exonerate

very slow

(EviAnn)

see note on GeMoMa, not related tool but works in a similar way in terms of possible inputs

Transcriptome Processing Tools

HISAT2

Fast and efficient short reads spliced aligner

STAR

Remember to run 2-pass mapping

Minimap

For long reads

StringTie

State of the art and fast transcriptome assembler (genome-guided assembly). Can use short and/or long reads

Trinity

For de novo transcriptome assembly. This is often not very helpful for structural genome annotation and needs a bit of runtime...

PASA

Old tool but still pretty good, includes Transdecoder for finding ORFs in assembled transcripts, can perform UTR annotation

cDNA cupcake

Outdated

cd-hit

Used to cluster transcripts, not directly helpful for genome annotation but sometimes during data preparation

ab initio Gene Finders

AUGUSTUS

Older but still very accurate gene finder. Can run ab initio or with evidence. Was in the past also used in MAKER, is at the core of BRAKER and Gaija

GeneMark

Suite of self-training gene finding tools. GeneMarkES-2: finding genes in transcripts. GeneMarkES: ab initio, genome sequence input only

Helixer

deep learning gene finder, quite changer in the field, good BUSCO accuracy, poor gene structure accuracy, great web service, several stable models available

Tiberius

deep learning gene finder, building on results of Helixer team. So far only trained for mammals (poor parameters for diatoms also exist)

FgeneSH

Older gene finder that can run ab initio or with evidence. Was often used in MAKER in the past

GlimmerHMM

Older gene finder, was an early community project. Was in the past often used in MAKER

SNAP

Older gene finder that can run ab initio or with evidence. Was often used in MAKER in the past. Very easy to train and use

Gene Finders that Use Evidence

PASA

Can be used to generate training genes for gene finders from transcriptome evidence. Not much used anymore for this

GeMoMa

Only evidence based gene models. Requires as input genomes and annotations of related species. Can add transcriptome data

EviAnn

Similar to GeMoMa

AUGUSTUS

Can use lots of evidence. There is also a comparative version (AUGUSTUS-CGP) that annotates genes consistently in a multi-species genome alignment

GeneMark

GeneMarkES-2: genome + mass mapping input. GeneMarkES: genome + protein db input. GeneMarkES-TP: genome+mass mapping+protein db input

SNAP

FGenesH

Gene Set Combiners

TSEBRA

Custom tailored to AUGUSTUS/GeneMark output and evidence, strong focus on splice site support

EVM

Widely used, kind of maximizes coverage information for building an optimal gene set

(MAKER)

contains a combiner part to build a consensus gene set

Categorize tool names

Gene Set Combiners	Complex Annotation Pipelines	Functional Genome Annotation Tools	Repeat Masking Tools	Others
TSEBRA Custom tailored to AUGUSTUS/Genewise output and evidence, strong focus on splice site support	BRAKER BRAKER2: RNA-Seq + protein database BRAKER3: protein database BRAKER3: RNA-Seq + protein mapping file, better to use BRAKER3	InterProScan	RepeatMasker	BUSCO Marker gene detection in genomes, proteomes and transcriptomes. Widely used.
EVM Widely used, kind of maximizes coverage information for building an optimal gene set	EASEL Nextflow pipeline, includes several transcriptome assemblies, protein mappers, and gene finders	Blastp	RepeatModeler2	OMark Marker gene detection in proteomes. Handles alternative splicing/isoforms well. Larger number of marker genes than in BUSCO.
(MAKER) combines a combiner part to build a consensus gene set	TOGA Interesting if you have multi-genome alignment. Integrates AUGUSTUS-CGP	EggNOG-mapper	Deep TE	AGAT Useful for many gff handling tools
	MAKER2 The first highly popular community annotation pipeline. Now outdated because accuracy is not that good. Does not train gene finders automatically.	CD-SEARCH	LTR FINDER	OrthoDB Database
	PGAP (prokaryotes)	InterproScan	GMATA	Gffcompare Useful for handling gff files
	Prokka (prokaryotes)	FANTASIA Very fast, only GO term assignment	RepeatScout	
			Tandem Repeat Finder	
			EDTA pipeline	
			Red	

The aera of genome annotation super heroes

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
- 1,677 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

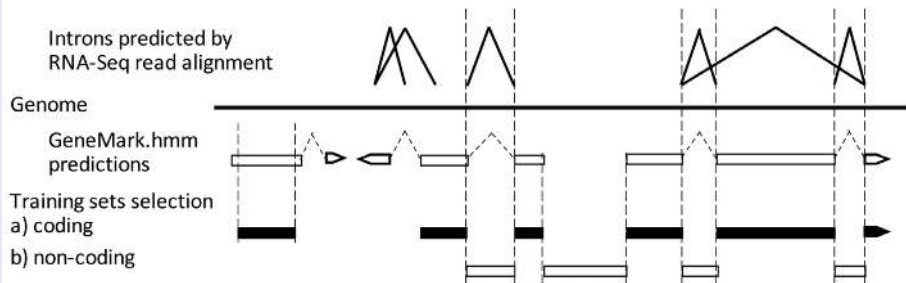


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- 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

Use only if not enough RNA-Seq for BRAKER3!

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^{2,3,1} and Mark Borodovsky^{4,5,*}

- spliced alignments of a large number of proteins (e.g. OrthoDB partition)
- optional input: BUSCO lineage (compleasm)
- 1,269 citations (Google Scholar)

BRAKER2 gene F1 accuracy

Use only if you have no RNA-Seq data on genomes <1 Gbp

BRAKER3 gene F1 accuracy - climbing the top

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

- Gabriel *et al.* (2024)
- 137 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

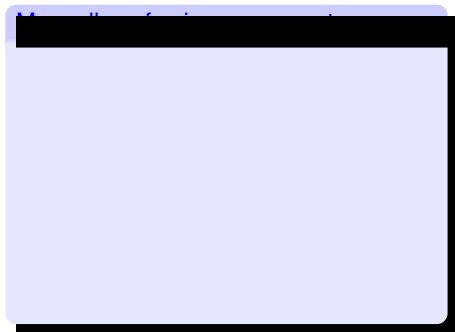
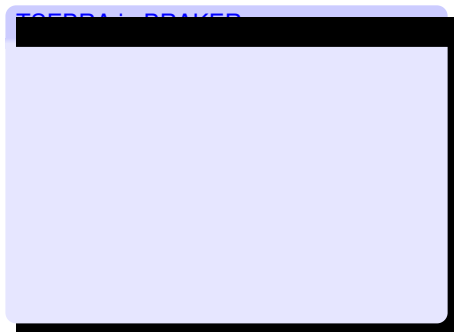
Open Access

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*} 

- **combines** several gene sets according to evidence
- 154 citations (Google Scholar)



Can be used to combine BRAKER1 and BRAKER2 output if BRAKER3 fails.

Accuracy of genome annotation approaches by BRAKER team

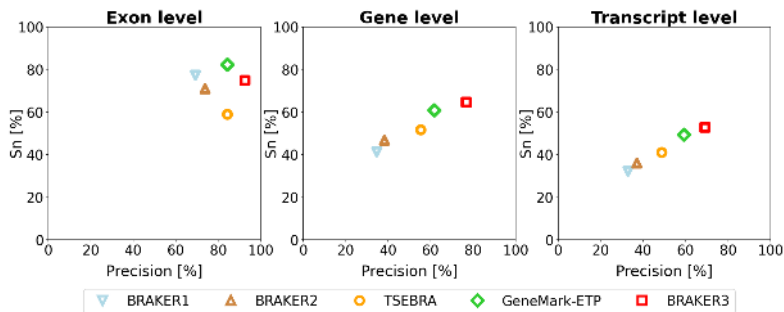


Figure 2. Average precision and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in [Supplemental Table S1](#)). Inputs were the genomic sequences, short-read RNA-seq libraries, and protein databases (*order excluded*).

Image: Gabriel *et al.* (2024), Genome Research

GitHub

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

Docker/Singularity

```
singularity build braker.sif \  
    docker://teambraker/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

RESEARCH

Open Access

Galba: genome annotation with miniprot and AUGUSTUS



- 28 citations (Google Scholar)
- for genomes >1Gbp
- proteins of close relatives

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

If you have RNA-Seq, use BRAKER3!

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

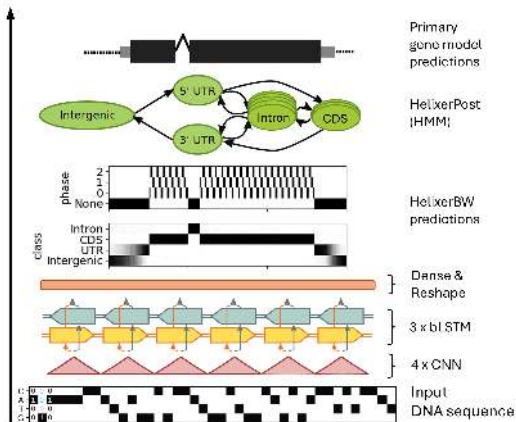
- GALBA: Artistic License
- all dependencies have Open Source Licenses

Helixer: bringing deep learning into genome annotation

Image: ChatGPT by OpenAI, manual editing

HELIXER—*de novo* PREDICTION OF PRIMARY EUKARYOTIC GENE MODELS COMBINING DEEP LEARNING AND A HIDDEN MARKOV MODEL

Felix Holst^{1†}, Anthony Bolger^{2†}, Christopher Günther¹, Janina Maß³,
Sebastian Triesch^{1,4}, Felicitas Kindel¹, Niklas Kiel^{1,4}, Nima Saadat^{3,4}, Oliver Ebenhöf^{3,4},
Björn Usadel^{2,4,5}, Rainer Schwacke², Marie Bolger², Andreas P.M. Weber^{1,4}, Alisandra K. Denton^{1,4}



- 27 citations (Google Scholar)
- cross-species gene finder
- *ab initio* prediction
- Pre-trained models for:
 - ▶ fungi
 - ▶ land plant
 - ▶ vertebrate
 - ▶ invertebrate
- accuracy (BUSCO): good
- web service

Availability: <https://github.com/weberlab-hhu/Helixer>

Image of Helixer: <https://github.com/weberlab-hhu/Helixer/blob/main/img/network.png>

Tiberius: improved genome annotation with deep learning

Image: ChatGPT by OpenAI, manual editing



Lars Gabriel, Felix Becker, Katharina Hoff, Mario Stanke

Tiberius: end-to-end deep learning with an HMM for gene prediction

Lars Gabriel ^{1,2,*}, Felix Becker ^{1,2}, Katharina J. Hoff^{1,2}, Mario Stanke ^{1,2,*}

- builds on findings by Helixer team
- cross-species gene finder
- faster
- higher accuracy
- *ab initio* prediction
- Pre-trained model(s) for:
 - ▶ mammals
 - ▶ (diatoms)
- container for A100 GPU

Availability: <https://github.com/Gaius-Augustus/Tiberius>

Accuracy of state of the art gene finders

No alternative splicing isoforms

Image: Lars Gabriel, PAG presentation 2025

Current shortcomings of deep learning gene finders

- no evidence integration
- no alternative splicing isoform prediction
- require expensive GPU for feasible runtime
- limited to specific clades

→ BRAKER3, Galba & EGAPx currently remain important

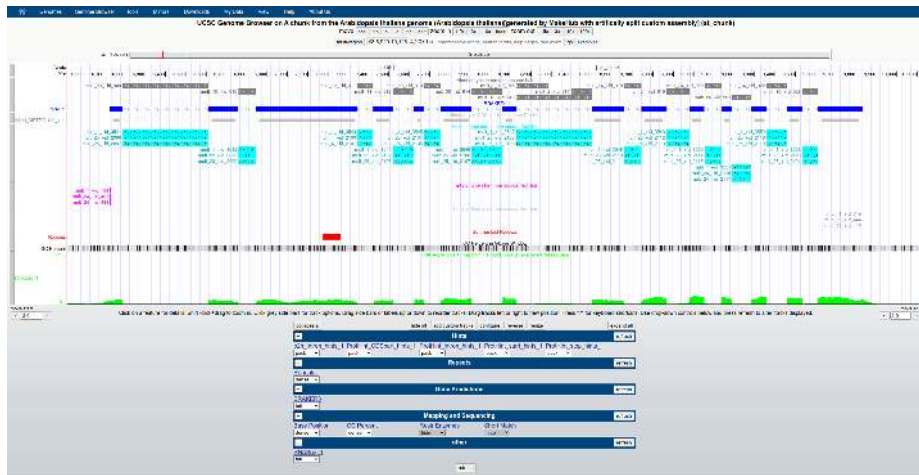
Did We Do a Good Job?



Genome Browsers

Visualize your Annotation in Context with Evidence

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



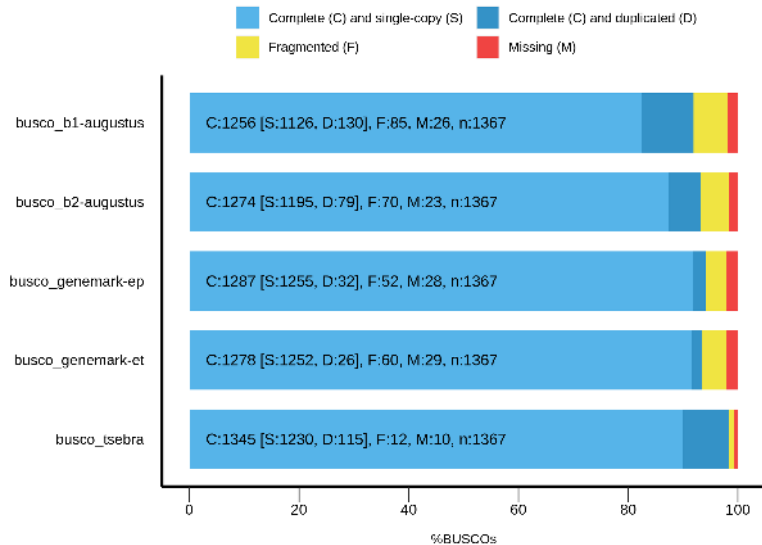
Describe Your Annotation

- 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



Beware! BUSCO completeness does not warrant correct gene structures!

Revisiting annotation methods sections

- 1 Read your methods snippet, again
- 2 Use our categorized tool name board at <https://shorturl.at/uA0Tg> 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 on Wooclap

Revisiting annotation methods sections

Revisiting Annotation Methods (1 = disagree strongly, 5 = completely agree)



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 BRAKER
 - 2 GALBA
 - 3 (EGAPx may become popular)
- deep learning is changing the field
 - 1 Helixer (careful with accuracy)
 - 2 Tiberius (only for two clades)
- "looking nice" is not always "correct"
- BUSCO completeness is widely used
- OMArk might be more appropriate
- high marker gene detection rate $\neq$ high accuracy

