

Annotation of Protein Coding Genes

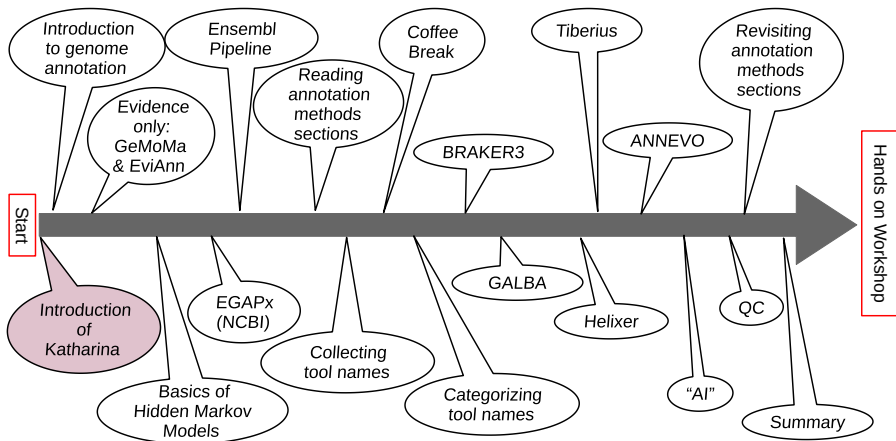
January 16th 2026

Katharina J. 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

Short CV

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

2009 Ph.D. Molecular Biology (Göttingen)

2022 Habilitation (Greifswald)

2025 **DFG** Heisenberg grant → professor in 2026?

Research

- eukaryotic genome annotation, metagenomics
- best known for: **BRAKER** & other **Gaius-Augustus** software
- 41 peer-reviewed research articles with currently 9,357 citations
- ~1.9 Mio € grants

Teaching

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

... I love to sail, have 2 dogs, a cat, and an 9-years old daughter...

After this lecture, you will...

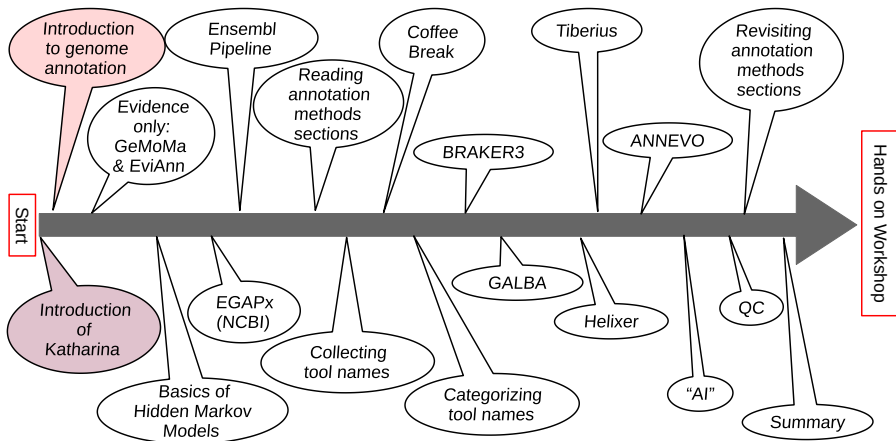
- understand what genome annotation in eukaryotes is
- know the basics of a Hidden Markov Model
- have a vague idea of INSDC annotation pipelines
- roughly understand methods sections on genome annotation
- know what's happening in BRAKER & GALBA
- be aware of the rapid advances with Deep Learning
- 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
cattcaagttcctatttttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagctg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttcccc
atctgtcatatgaaggagtgcgatgtgttctgagactgaatccagttccaatcttctagattttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggttctggtagctccacagggtgctcaatgaagatgcaaaattagaagtcaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccttgggaggcaagggtgaggg
ctggatttttaaaagggggcctgttttggggaggggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgtgaaggcggctgagt
ggcagccgagacagaaggggttcttggggaggaagttattcagaggacaggggaagcaggggaaggcag
acaggtcccatgagatattggaccaattccttaaacatgctagaaaaacatgtgaaaagtactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttctaatctgggcaacc
cctggggccacagctaaatccagtgagtggaagttacagggagctgtgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttgttctacagattc
caaaccaaaagagggaagcaagctgtgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagctctt
```

Examples for the importance of genome annotation

Silencing polygalacturonase activity in tomato



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

Bacillus thuringiensis toxin against European corn borer

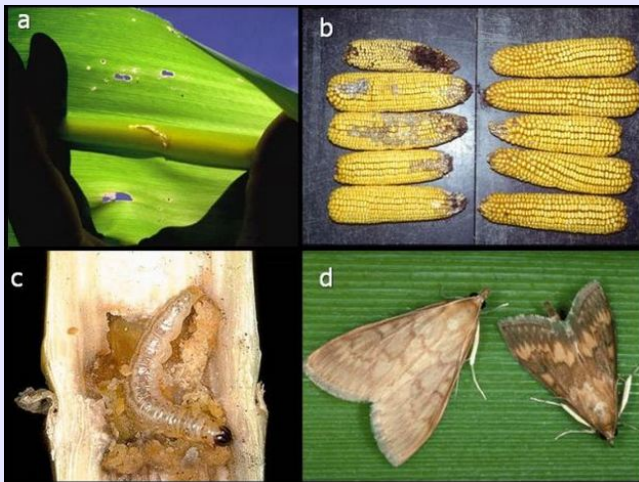


Image: Hellmich & Hellmich (2012) Nature Education Knowledge 3(10):4

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**

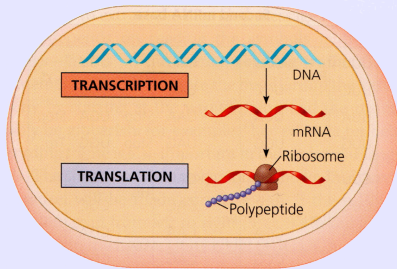
It does not take a village to publish a genome!

- 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!**

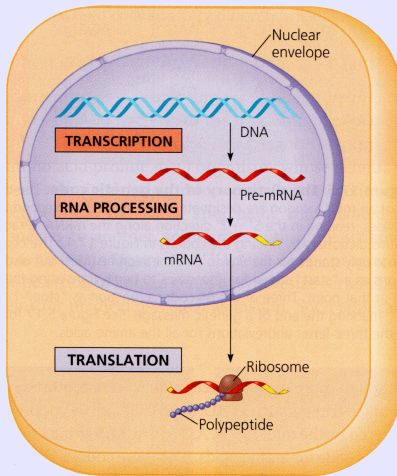
How does a cell recognize protein-coding genes?

Transcription & Translation

Prokaryotes



Eukaryotes



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

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

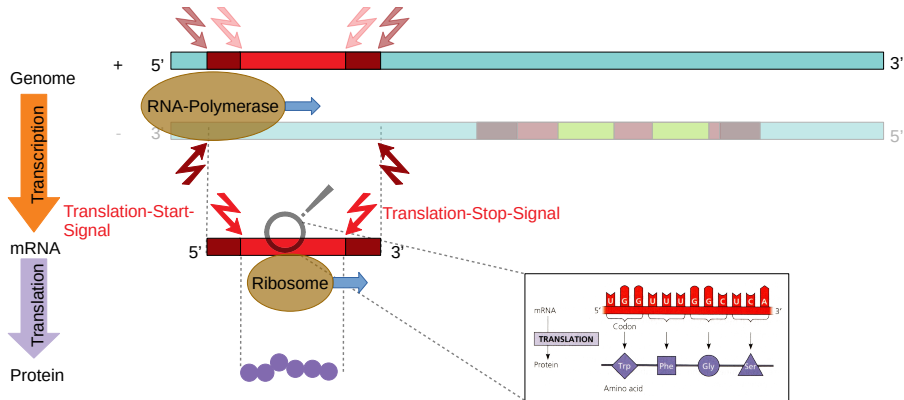


Image: Campbell et al., Biology, San Francisco, 2008, p. 329, Fig. 17.4

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

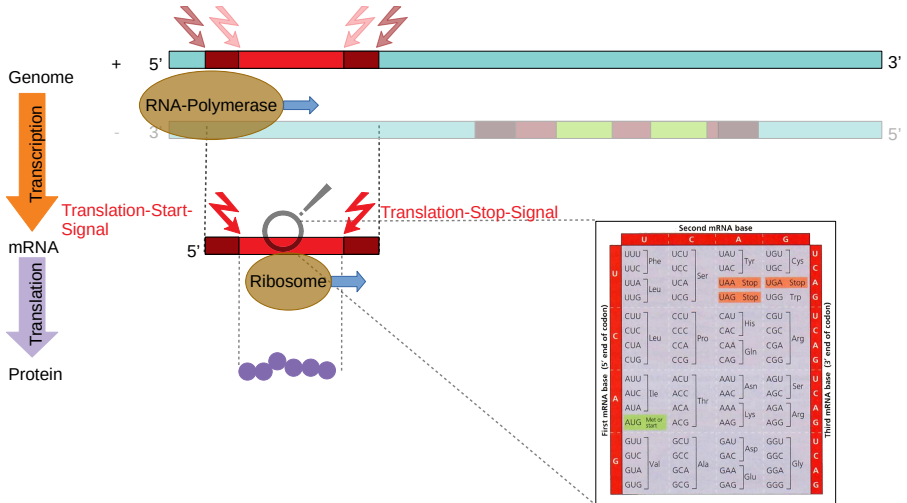
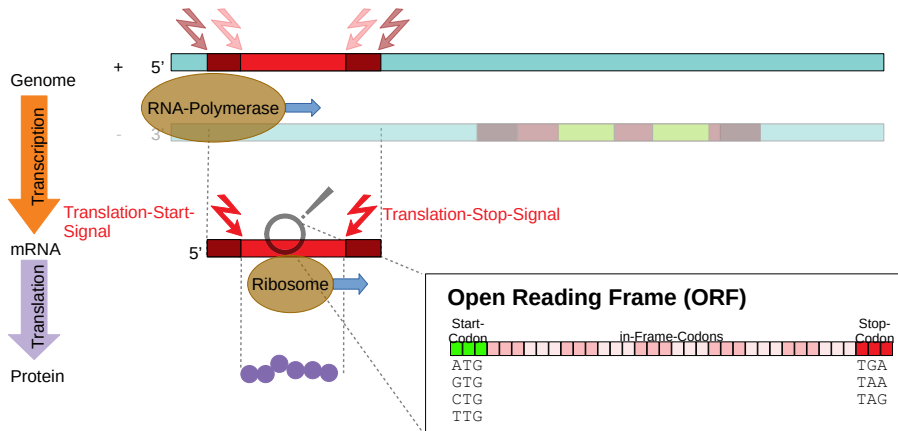


Image: Campbell et al., Biology, San Francisco, 2008, p. 339, Fig. 17.5

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

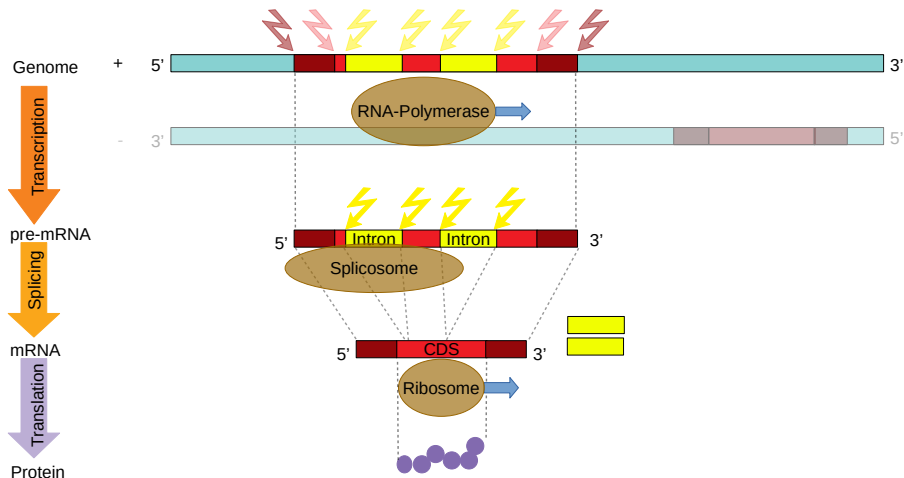
Prokaryotes & Eukaryotes*



- 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
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagctg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttcccc
atctgtcatatgaaggagtgcgatgtgttctgagactgaatccagttccaatcttctagattttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggtctggtagctccacagggtgctcaatgaagatgcaaaattagaagtcaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgcctgggaggcaagggtgaggg
ctggatttttaagggggcctgttttggggagggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgtgaaggcggctgagt
ggcagccgagacagaagggggttctggggaggaagttattcagaggacaggaagcaggggaaggcag
acaggtcccatgagatattggaccaattccttaaacatgctagaaaaacatgtgaaaagtactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggtcctaattctgggcaacc
cctggggccacagctaaatccagtgagtggaagttacagggagctgtgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagagggaagcaagctgtgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagtcttc
```

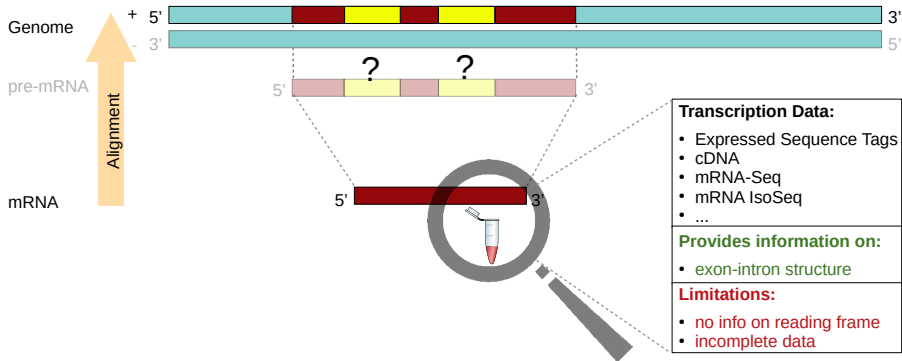
The Genome Annotation Problem

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

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcaactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcagggtgtaggcagagtcctttttctagtcattggctggcaaacagtgaggatctggggatggg
acaaaaggcagctaggaagattgccatgtagctctgctgctaagtgtagagctctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtgcagctg
ttgaattggctctgcctttaattatttgttcaagcaagccctgtccctctctgtgccttggtttccc
atctgtcatatgaaggagtgcatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttgggggtctggtagctccacagggtgctcaatgaagatgcaaaattagaagtcaaaat
aacagctcccatgggcagtggtgatctcacctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccctgggaggcaagggtgaggg
ctggatttttaaaagggggcctgttttggggagggggtgatgagcgtggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggttctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaacatgctagaaaaacatgtgaaaagtactacca
ggctggcaggggaatggggcaatctattcatactgattgcaatgccactgggttctaatctgggcaacc
cctggggccacagctaaatccagtgagtggaagttacagggagctctgcttccagtgtgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaaagccagatgacctcaaaggtcccatgggattctaattctgtctgctccttggttctacagattc
caaaccaaaagagggaagcaagctctgcgctgacccagtgagtcctgggtccaggagtagctgtatgac
ctggaactgaactgagctgctcagagacaggaagtcttc

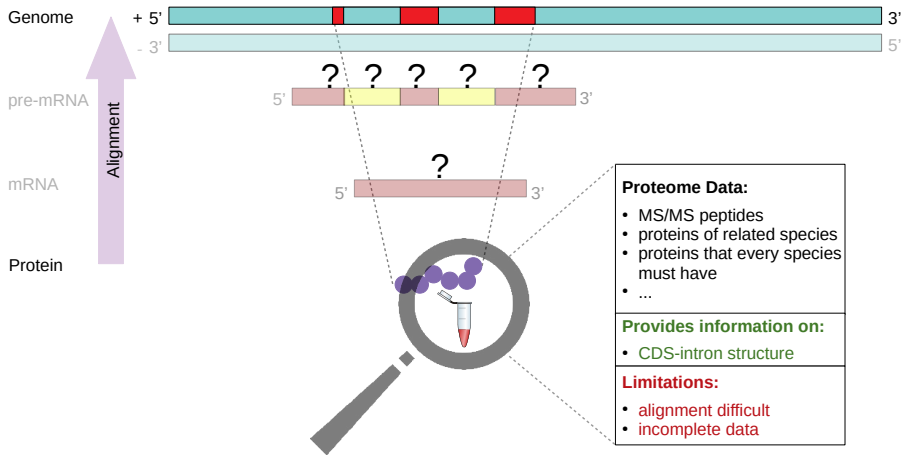
What aids in the identification of genes in genomes?

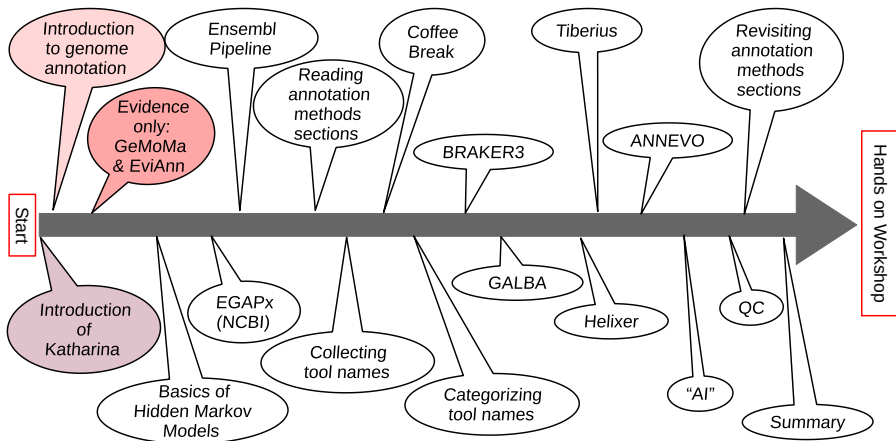
Evidence data from transcription



What aids in the identification of genes in genomes?

Evidence data from translation





Genome Annotation Tools that Rely on Evidence Data

GeMoMa

- Java tool
- required inputs:
 - ▶ genome file
 - ▶ genome files of close relative species
 - ▶ annotation files from (several) close relative species
 - ▶ optional: transcriptome data

Keilwagen et al. (2016); Keilwagen et al. (2018)

EviAnn Input Data

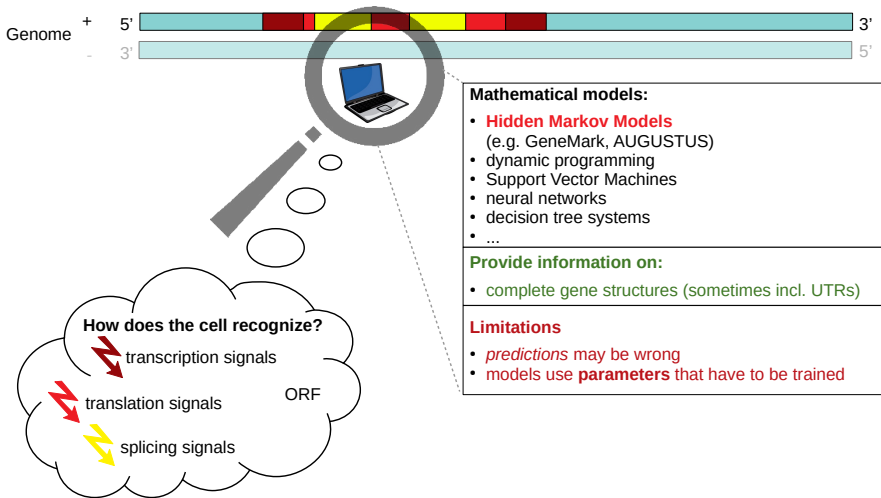
- Perl pipeline
- required inputs:
 - ▶ genome file
 - ▶ protein sequences from (several) close relative species
 - ▶ transcriptome data

Zimin *et al.* (2025)

- + high accuracy if annotations of very close relatives are available
- cannot predict genes that have no evidence

What aids in the identification of genes in genomes?

Mathematical models



What aids in the identification of genes in genomes?

Mathematical models



Mathematical models:

- **Hidden Markov Models**
(e.g. GeneMark, AUGUSTUS)
- dynamic programming
- Support Vector Machines
- neural networks
- decision tree systems
- ...

Provide information on:

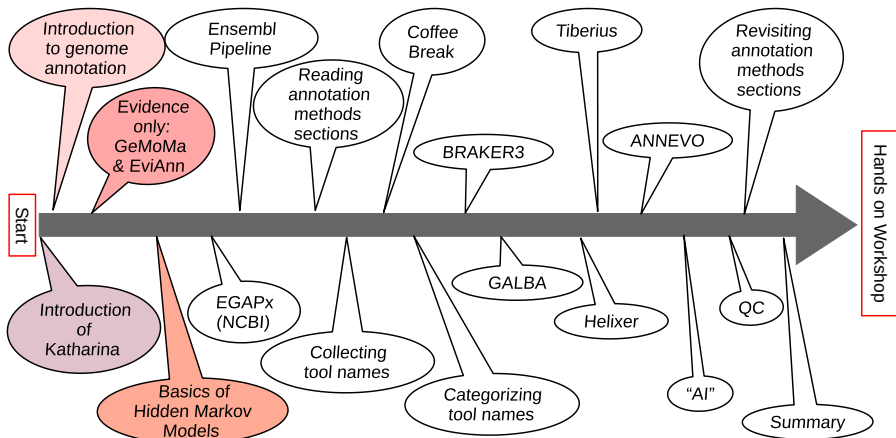
- complete gene structures (sometimes incl. UTRs)

Limitations

- *predictions* may be wrong
- models use **parameters** that have to be trained

A **Hidden Markov Model** can read the genome sequence from left to right and, through knowledge of signals for transcription and translation, assign a probable state to each nucleotide (e.g., intergenic region or CDS).





Basis of highly accurate gene prediction tools

Hidden Markov Model

Simplifications

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

Input: “Genome sequence”

e.g. AABBBAB

Goal: “Most likely path through hidden states”

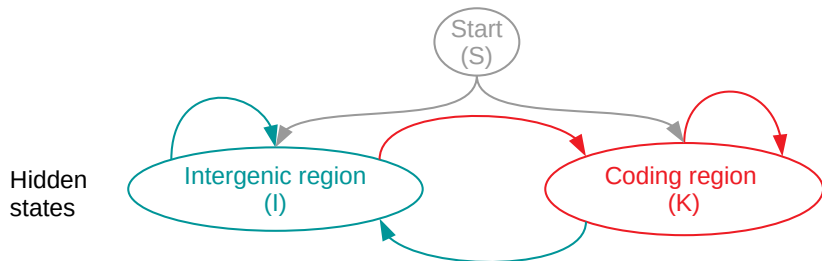
e.g. AABBBAA

or IIKKIKI

$P(\text{path}) = 0.3\%$

Basis of highly accurate gene prediction tools

Hidden Markov Model

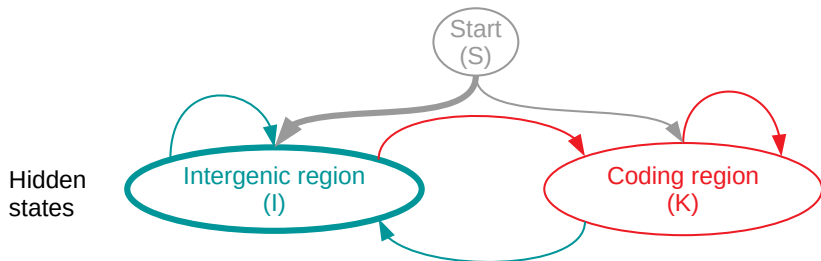


A possible 'state path' for the genome sequence:

AABBBA

Basis of highly accurate gene prediction tools

Hidden Markov Model



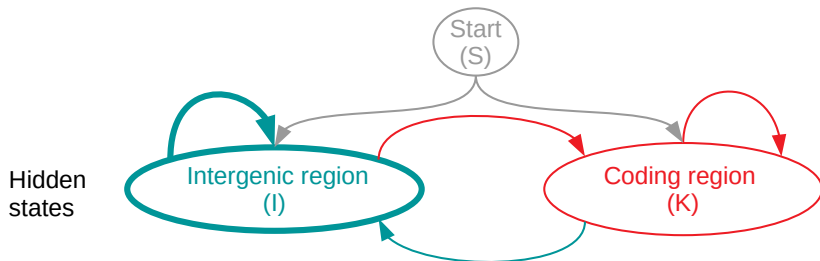
A possible 'state path' for the genome sequence:

AABBBAA

I

Basis of highly accurate gene prediction tools

Hidden Markov Model



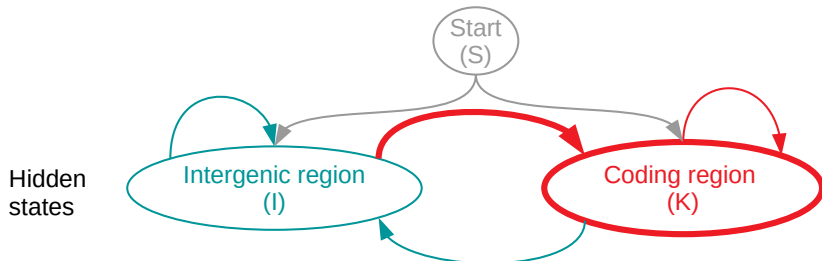
A possible 'state path' for the genome sequence:

AABBBAA

II

Basis of highly accurate gene prediction tools

Hidden Markov Model



A possible 'state path' for the genome sequence:

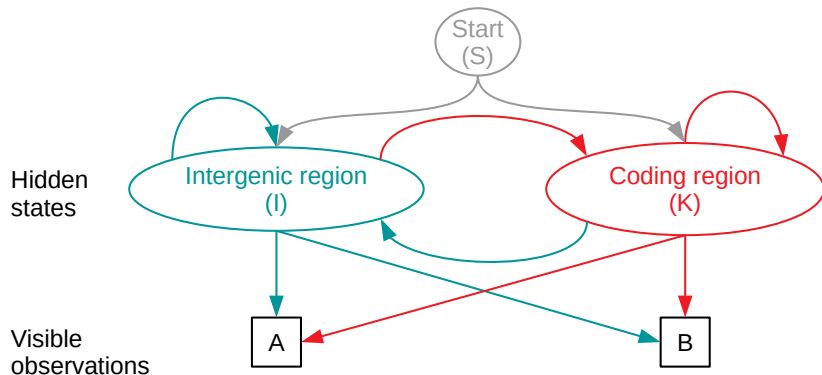
AABBBAA
IIK...

Model properties

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

Basis of highly accurate gene prediction tools

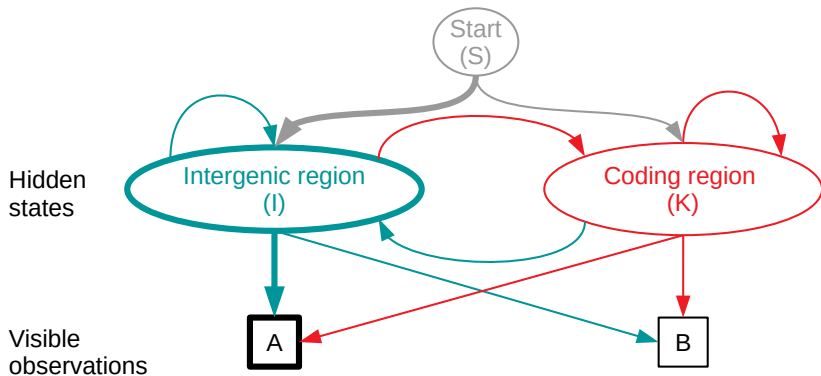
Hidden Markov Model



A possible 'state path' for the genome sequence:

Basis of highly accurate gene prediction tools

Hidden Markov Model

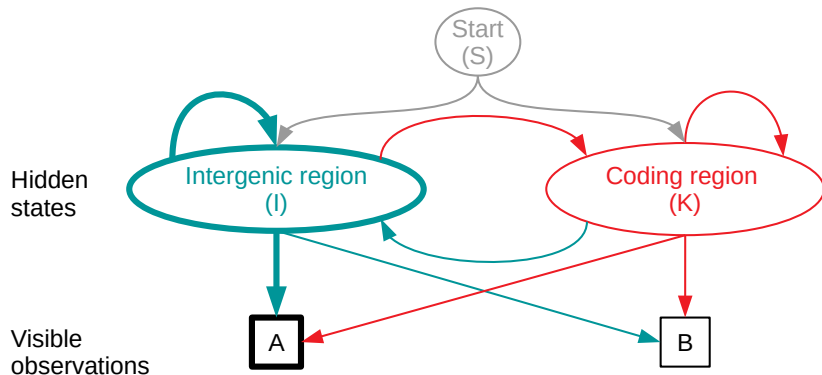


A possible 'state path' for the genome sequence:

A
I

Basis of highly accurate gene prediction tools

Hidden Markov Model

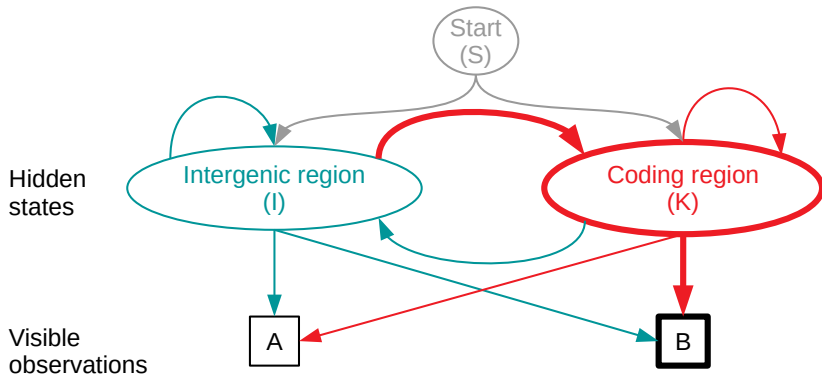


A possible 'state path' for the genome sequence:

AA
II

Basis of highly accurate gene prediction tools

Hidden Markov Model



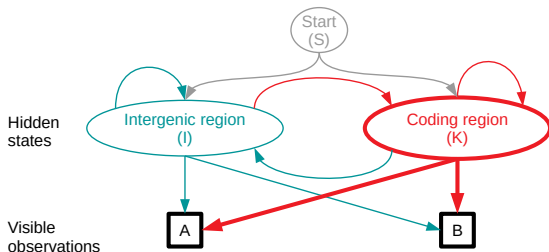
A possible 'state path' for the genome sequence:

AAB...

IIK...

Basis of highly accurate gene prediction tools

Hidden Markov Model

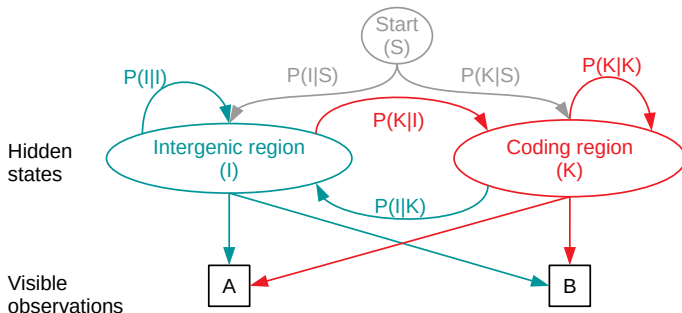


Model properties

- 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



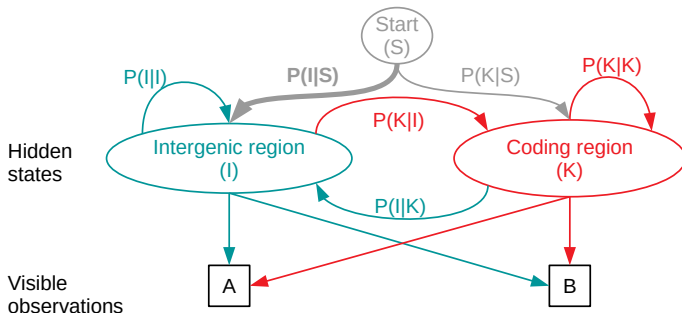
How likely are the state transitions?

Use data with known state transitions for learning!



Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

AABABA

IKKIII

Start probability

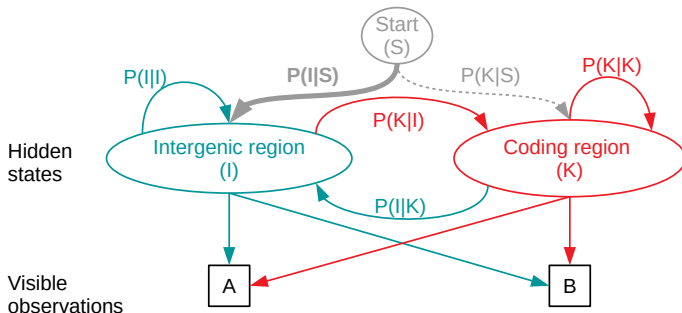
$P(I|S) = ?$

Use data with
known state
transitions for
learning!



Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

AABABA

IKKIII

+

Start probability

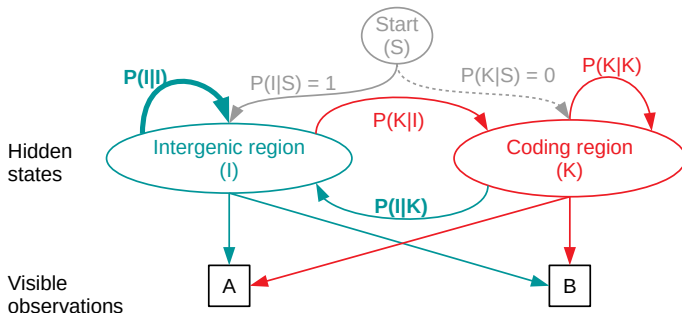
$$P(I|S) = 1$$

Use data with
known state
transitions for
learning!



Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

AABABA

IKKIII

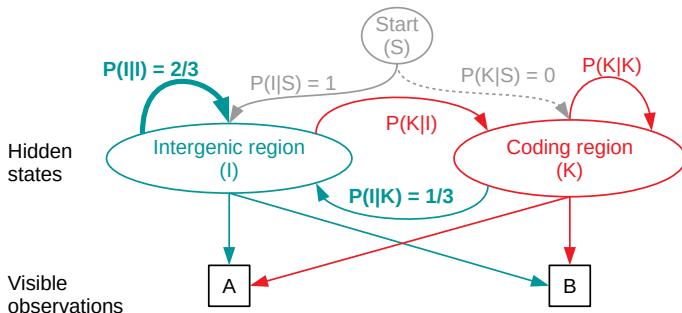
$P(I|I) = ?$

Use data with
known state
transitions for
learning!



Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

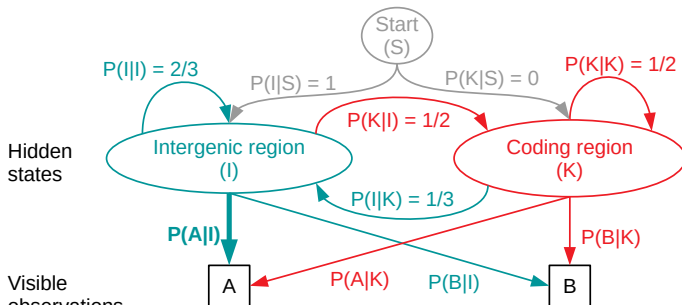
AABABA
IKKIII
-++

$$P(I|I) = 2/3$$

$$P(I|K) = 1 - P(I|I) = 1/3$$

Basis of highly accurate gene prediction tools

Hidden Markov Model



Visible
observations
(emissions)

How likely are the observations?

AABABA

I K K I I I

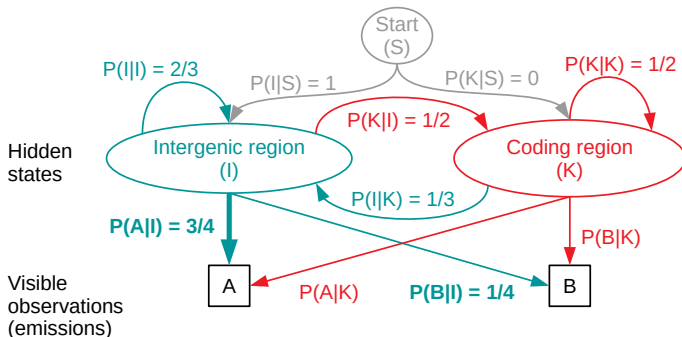
$P(A|I) = ?$

Use data with
known
"emissions" for
learning!



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely are the observations?

AABABA

I K K I I I

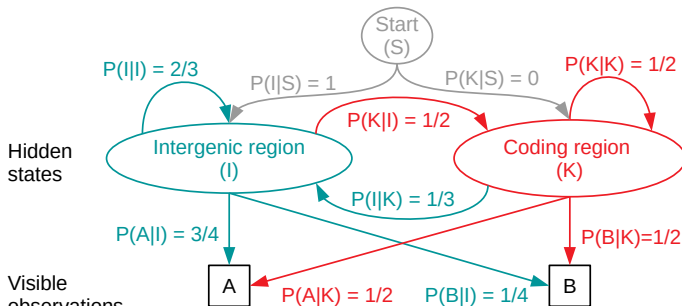
+ - - +

$$P(A|I) = \frac{3}{4}$$

$$P(B|I) = 1 - P(A|I) = 1 - \frac{3}{4} = \frac{1}{4}$$

Basis of highly accurate gene prediction tools

Hidden Markov Model



Visible observations (emissions)
Training data:

AABABA

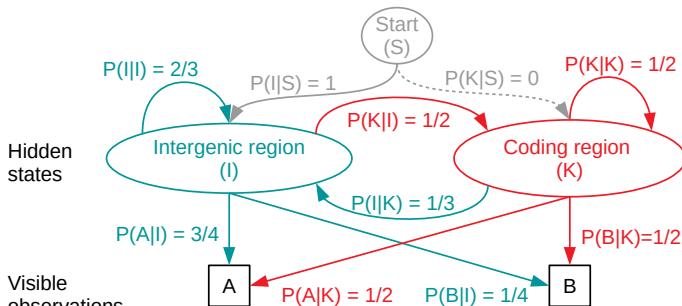
I K K I I I

In practice, more training data and training algorithm!



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

Path = AAB
I K K

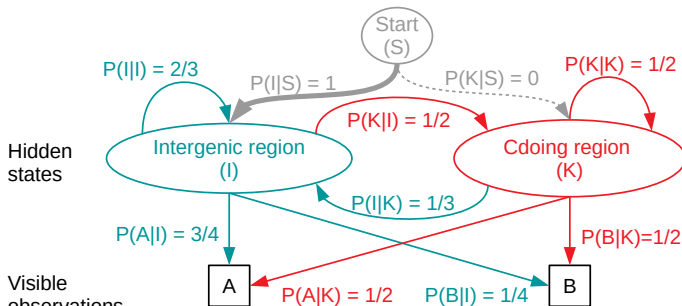
$P(\text{Path}) = ?$

Multiply the probabilities along the state-emission path!



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

Path = AAB

IKK

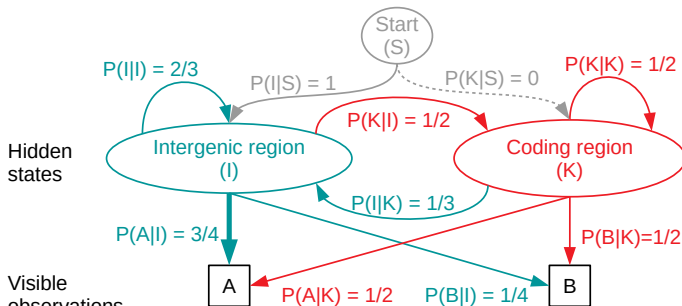
$P(\text{Path}) = P(I|S)$

Multiply the probabilities along the state-emission path!



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

Path = **AAB**

IKK

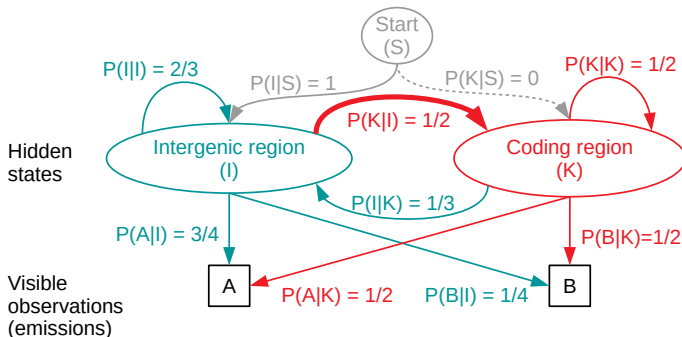
$$P(\text{Path}) = P(I|S) * P(A|I)$$

Multiply the probabilities along the state-emission path!



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

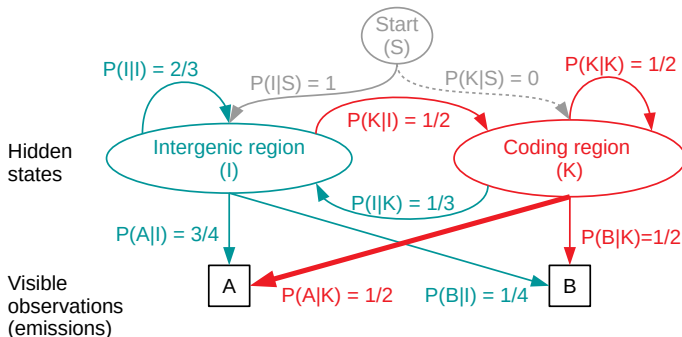
Path = AAB

IKK

$$P(\text{Path}) = P(I|S) * P(A|I) * P(K|I)$$

Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

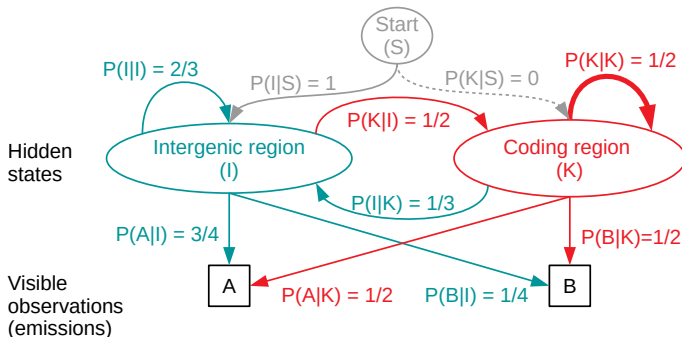
Path = AAB

I K K

$$P(\text{Path}) = P(I|S) * P(A|I) * P(K|I) * P(A|K)$$

Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

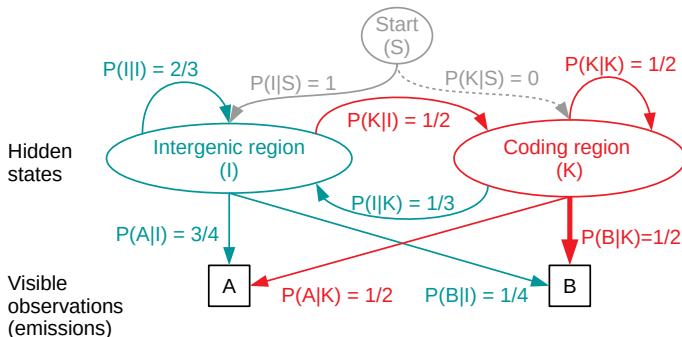
Path = AAB

IKK

$$P(\text{Path}) = P(I|S) * P(A|I) * P(K|I) * P(A|K) * P(K|K)$$

Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

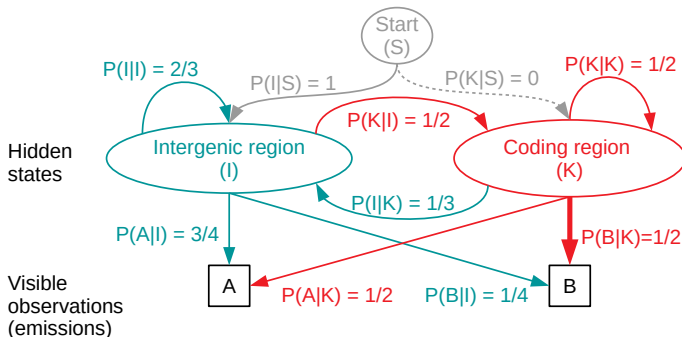
Path = AAB

I K K

$$P(\text{Path}) = P(I|S) * P(A|I) * P(K|I) * P(A|K) * P(K|K) * P(B|K)$$

Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

Path = AAB

I K K

$$P(\text{Path}) = P(I|S) * P(A|I) * P(K|I) * P(A|K) * P(K|K) * P(B|K)$$

$$= 1 * \frac{3}{4} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2}$$

$$= \frac{3}{64}$$

Basis of highly accurate gene prediction tools

Hidden Markov Model

Find the most probable state sequence for a given sequence

Input: “genome sequence”

AABBBABA

Problem: “too many possible state sequences”

IIIKKKKKK
KKIKKIIK
IIKIIKIK
IKKIKIIK
KIKIKKKIK
KKKIKIKK
...

Idea:

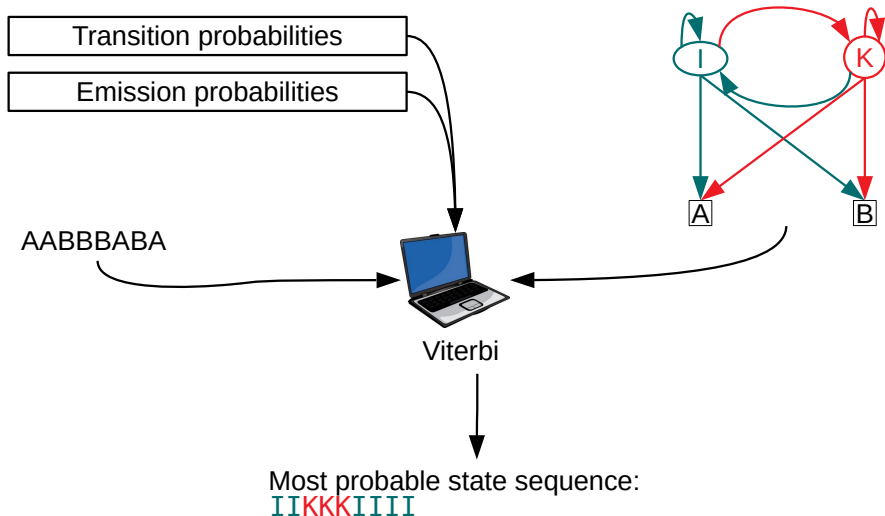
- 1 Generate all possible state sequences
- 2 Calculate the probability for each state sequence
- 3 Choose the state sequence with the highest probability

⇒ too expensive!

Basis of highly accurate gene prediction tools

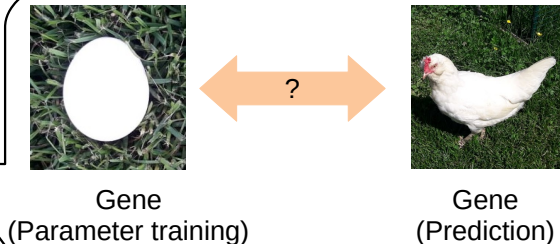
Hidden Markov Model

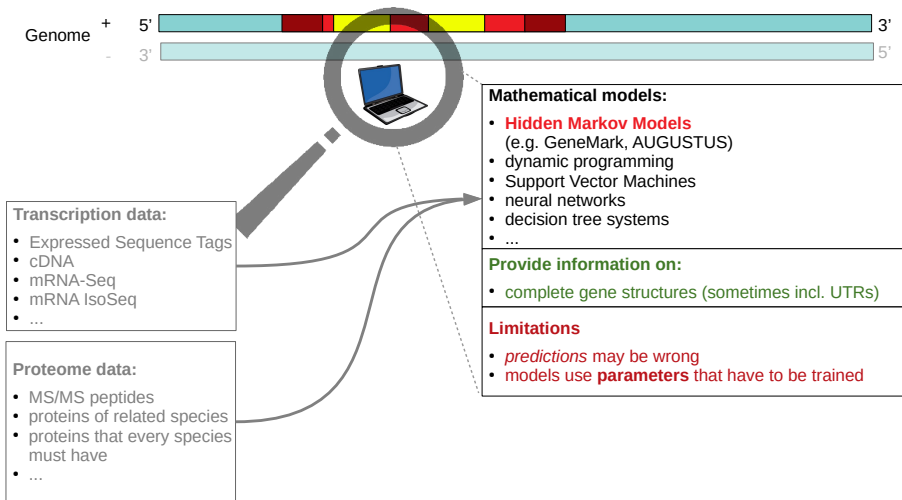
Find the most probable state sequence for a sequence: Viterbi Algorithm.



Hidden Markov Model prior Deep Learning in Practice

- 4096 observed nucleotide hexamers
- Many more hidden states
(e.g. 3'-UTR, 5'-UTR, Intron, ...)





Many genomes are annotated by database providers

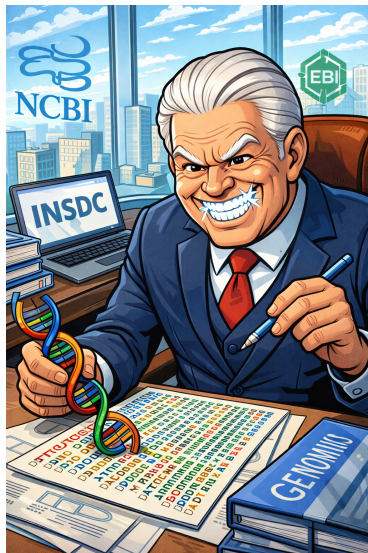
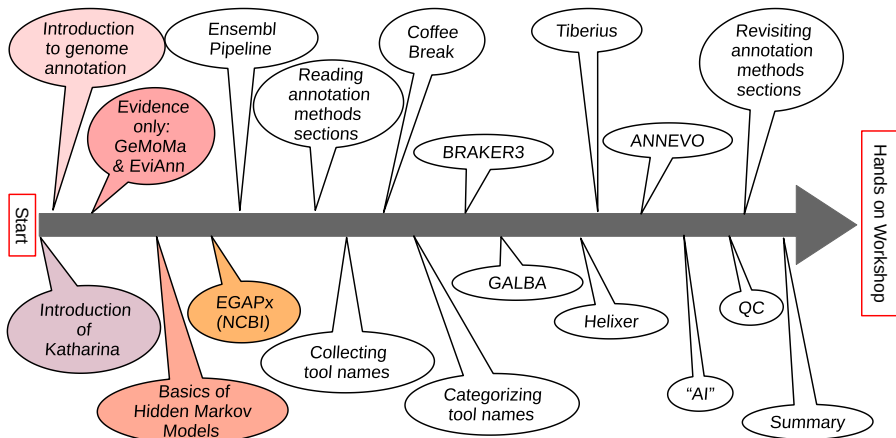


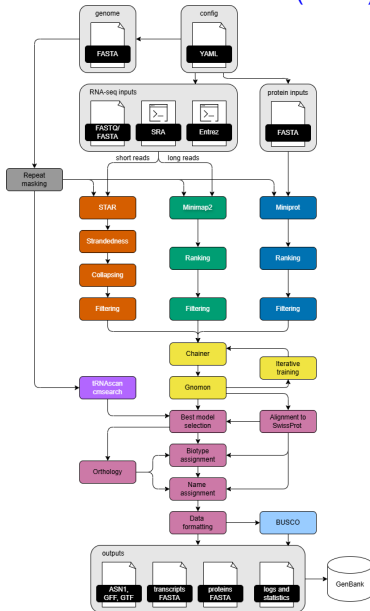
Image: ChatGPT



Different Annotation Scenarios

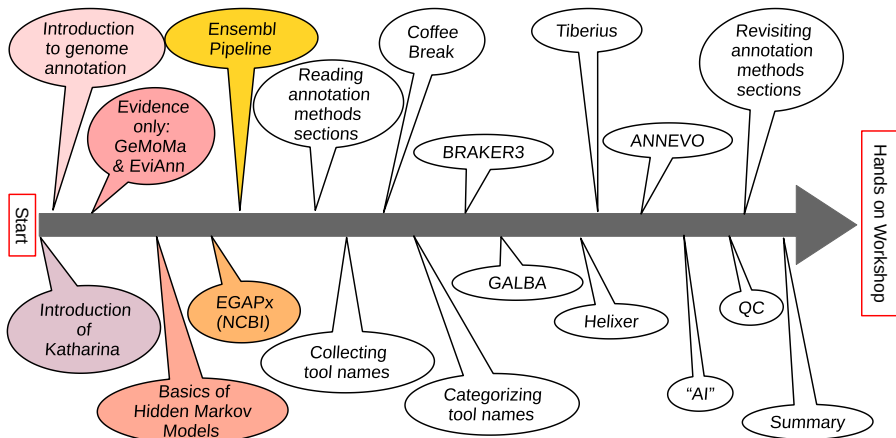
- Internal: The NCBI Eukaryotic Genome Annotation Pipeline (EGAP)
- (Internal: RefSeq curation)
- You can run it: **EGAPx**

Annotation with EGAPx (NCBI)



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

Image: <https://github.com/ncbi/egapx>



EBI: Ensembl annotation system

Ensembl annotation pipelines

Ensembl annotation pipeline for non-vertebrates

Documentation

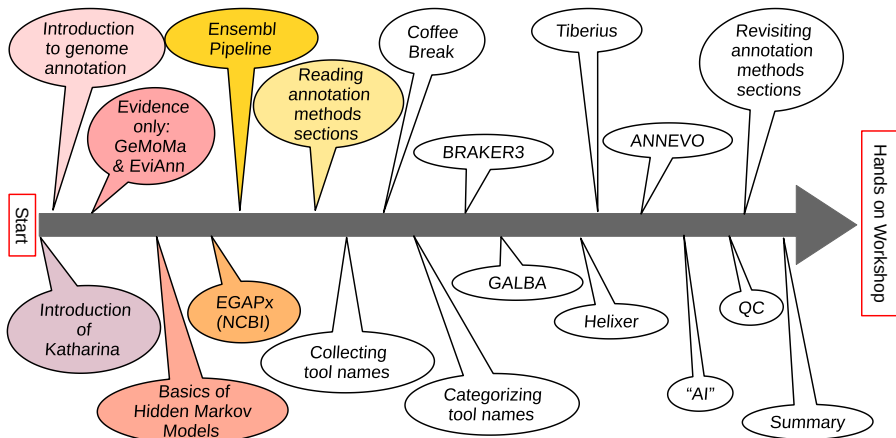
- **Ensembl vertebrate pipeline:** <https://beta.ensembl.org/help/articles/vertebrate-genome-annotation>
- **Ensembl non-vertebrate pipeline:** <https://beta.ensembl.org/help/articles/non-vertebrate-genome-annotation>
- **BRAKER2 in Ensembl:** <https://beta.ensembl.org/help/articles/braker-2-genome-annotation>

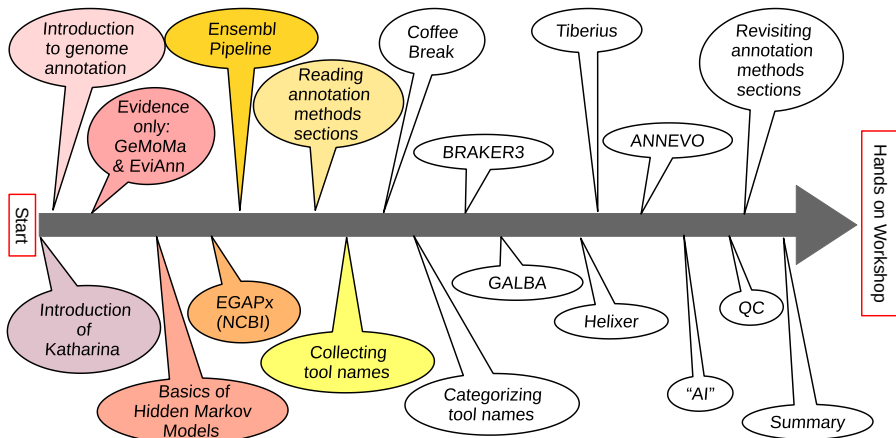
Where to find annotations

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

Notes by Katharina

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

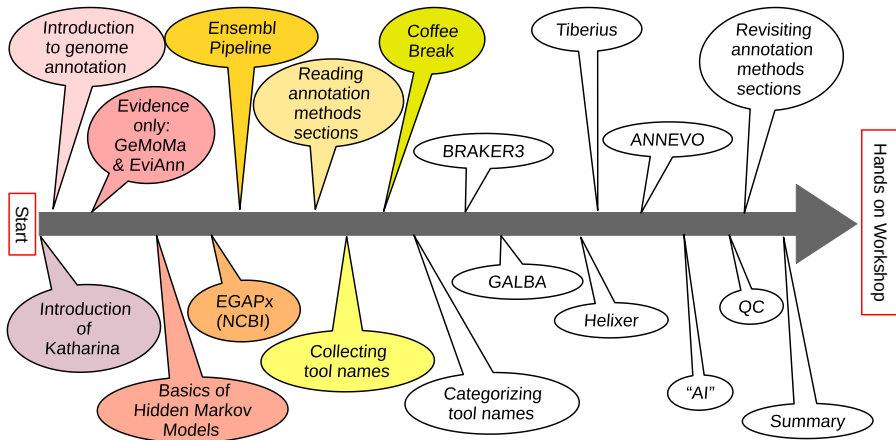


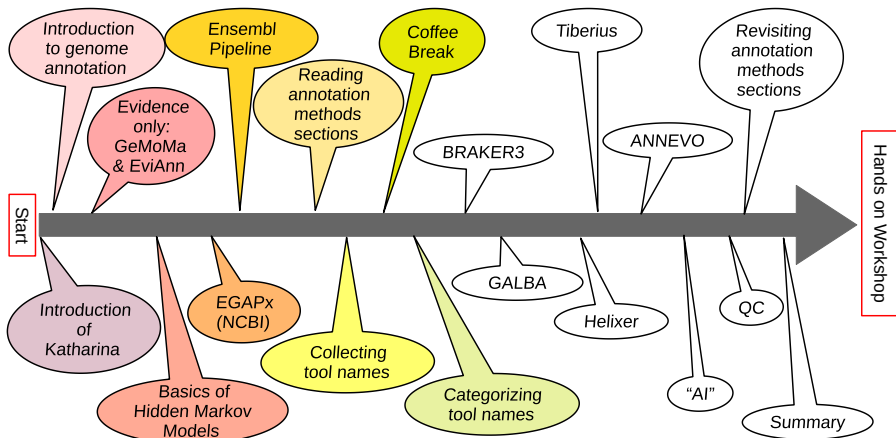


Read your methods snippet

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





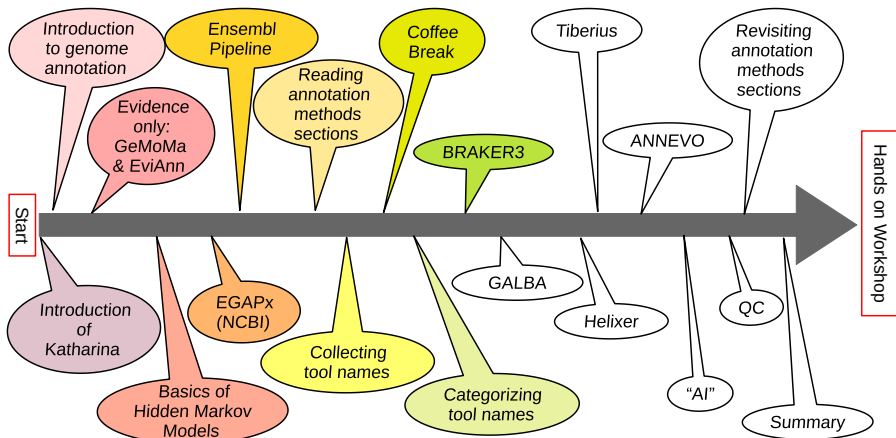
Categorize tool names

Go to

<https://shorturl.at/uA0Tg>

and sort the tools names from your methods snippet into categories





Genome annotation super heroes enable the community



Image: ChatGPT

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

Historical context on BRAKER

Once upon a time...

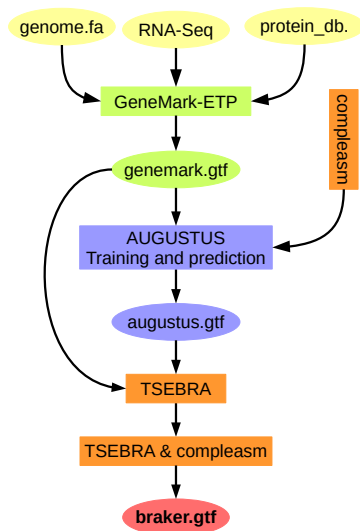
- AUGUSTUS requires supervised training & predicts alternative isoforms
- GeneMark performs unsupervised training
- MAKER integrated both tools for inference, not for training
- BRAKER1: GeneMark-ET + AUGUSTUS + RNA-Seq
- BRAKER2: GeneMark-EP + AUGUSTUS + proteins
- BRAKER3: GeneMark-ETP + AUGUSTUS + RNA-Seq + proteins

... but it is all one pipeline: `braker.pl`

Reasons for making BRAKER

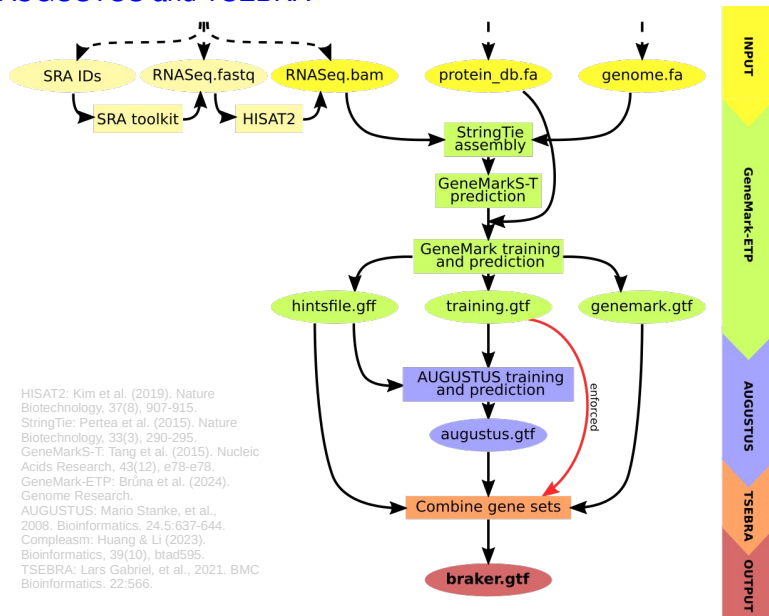
- increasing accuracy
- providing automation
- restoring citations

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



- Gabriel *et al.* (2024)
- 4,371 citations (all BRAKER publications, 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



HISAT2: Kim et al. (2019). Nature Biotechnology, 37(8), 907-915.
StringTie: Pertea et al. (2015). Nature Biotechnology, 33(3), 290-295.
GeneMarkS-T: Tang et al. (2015). Nucleic Acids Research, 43(12), e78-e78.
GeneMark-ETP: Bruna et al. (2024). Genome Research.
AUGUSTUS: Mario Stanke, et al., 2008. Bioinformatics, 24.5:637-644.
Compleasm: Huang & Li (2023). Bioinformatics, 39(10), btad595.
TSEBRA: Lars Gabriel, et al., 2021. BMC Bioinformatics, 22:566.

SOFTWARE

Open Access

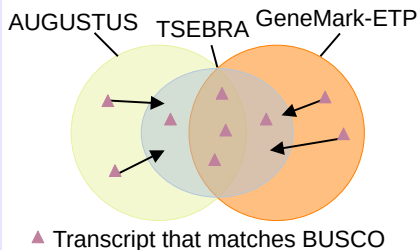
TSEBRA: transcript selector for BRAKER



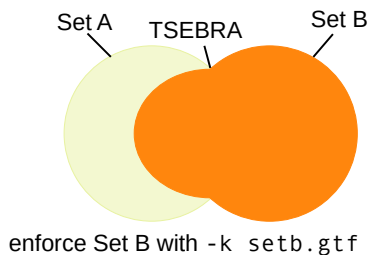
Lars Gabriel^{1,2}, Katharina J. Hoff^{1,2}, Tomáš Brůna³, Mark Borodovsky^{4,5} and Mario Stanke^{1,2*}

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

TSEBRA in BRAKER



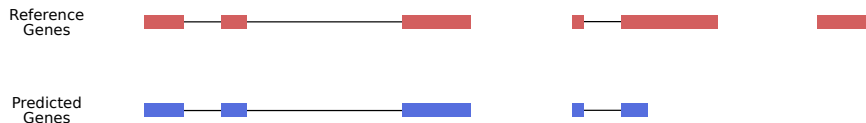
Manually enforcing a gene set



Is also used in Tiberius evidence processing pipeline.

Measuring accuracy of genome annotation

Compare prediction against a high quality reference.

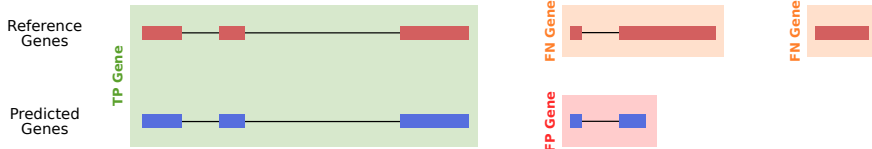


We only look at CDS exons, ignoring UTRs.

Measuring accuracy of genome annotation

Gene Level Accuracy

TP = True Positive
FP = False Positive
FN = False Negative



$$\text{Sensitivity (Sn)} = \frac{\#TP \text{ Genes}}{\#TP \text{ Genes} + \#FN \text{ Genes}}$$

$$\text{Precision (Pr)} = \frac{\#TP \text{ Genes}}{\#TP \text{ Genes} + \#FP \text{ Genes}}$$

$$\text{F1-Score} = \frac{2 \cdot Sn \cdot Pr}{Sn + Pr}$$

We only look at CDS exons, ignoring UTRs.

Accuracy of genome annotation approaches by BRAKER team

	Tools	Evidence
BRAKER1	GeneMark-ET + AUGUSTUS	RNA-Seq
BRAKER2	GeneMark-EP + AUGUSTUS	proteins
BRAKER3	GeneMark-ETP + AUGUSTUS	RNA-Seq + proteins
TSEBRA	BRAKER1 + BRAKER2	RNA-Seq + proteins
GeneMark-ETP		RNA-Seq + proteins

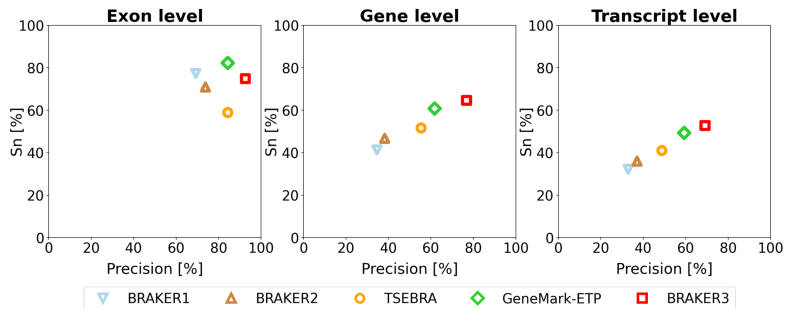


Figure 2. Average precision and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in [Supplemental Table S1](#)). Inputs were the genomic sequences, short-read RNA-seq libraries, and protein databases (*order excluded*).

Availability

GitHub

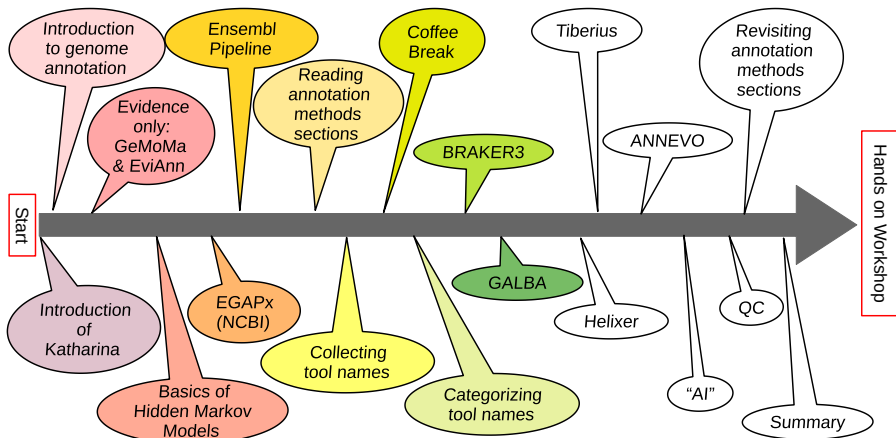
<https://github.com/Gaius-Augustus/BRAKER>

Docker/Singularity

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

Licenses

- 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



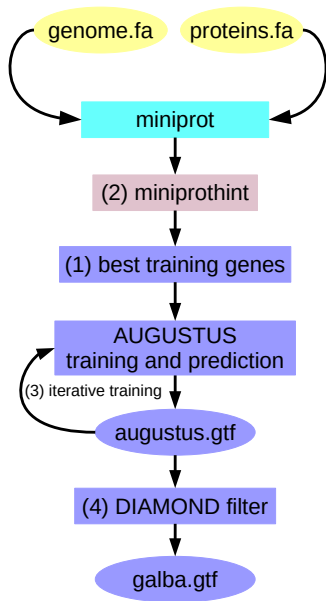
Matthis Ebel



Mario Stanke



Katharina Hoff



RESEARCH

Open Access

Galba: genome annotation with miniprot and AUGUSTUS

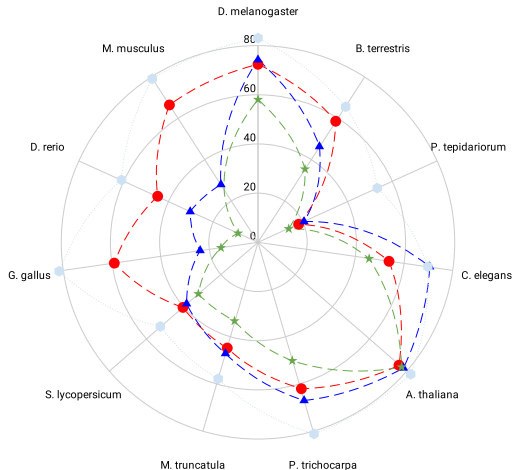


- 84 citations (Google Scholar)
- for genomes >1Gbp
- proteins of close relatives

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

Gene F1 (%)

● GALBA v1.0.10 ▲ BRAKER2 ★ FunAnnotate ● BRAKER3



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
- all dependencies have Open Source Licenses

Interlude: EASEL can annotate genomes with very large introns



Image: ChatGPT by OpenAI

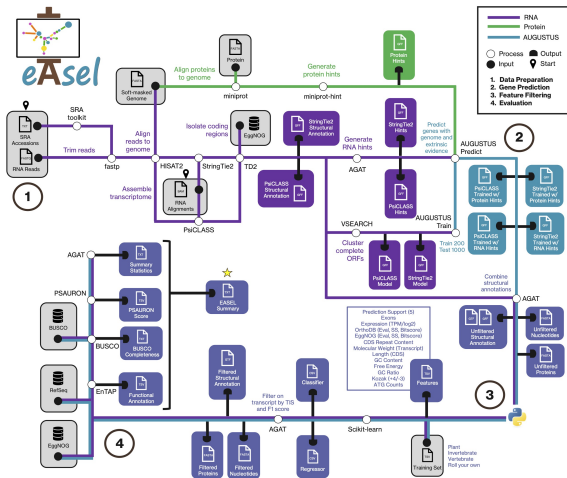
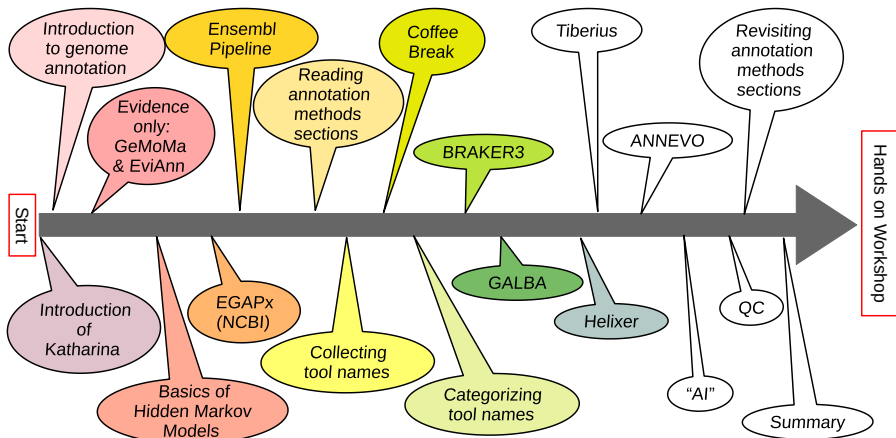


Image: courtesy of Cynthia Webster



Helixer: bringing deep learning into genome annotation



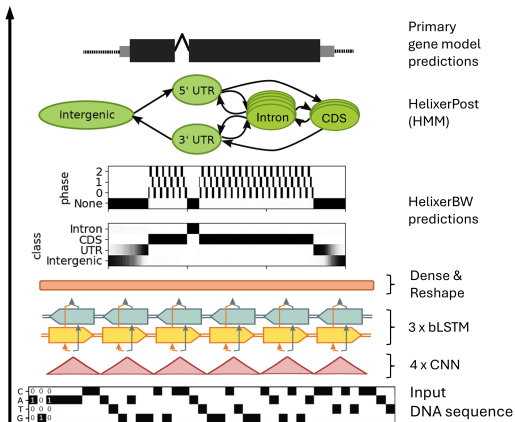
Image: ChatGPT by OpenAI, manual editing

Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model

Felix Holst, Anthony M. Bolger, Felicitas Kindel, Christopher Günther, Janina Maß, Sebastian Triesch,

Niklas Kiel, Nima Saadat, Oliver Ebenhöf, Björn Usadel, Rainer Schwacke, Andreas P. M. Weber, Marie E.

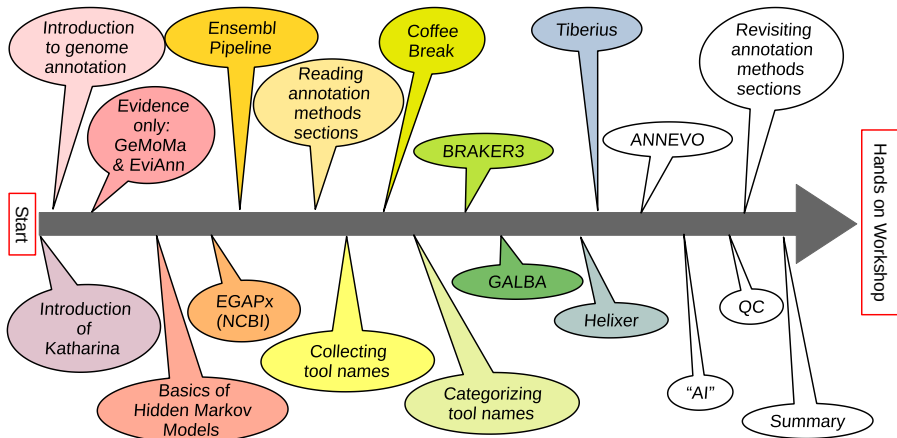
[Bolger](#) & Alisandra K. Denton



- 206 citations (all pubs, 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

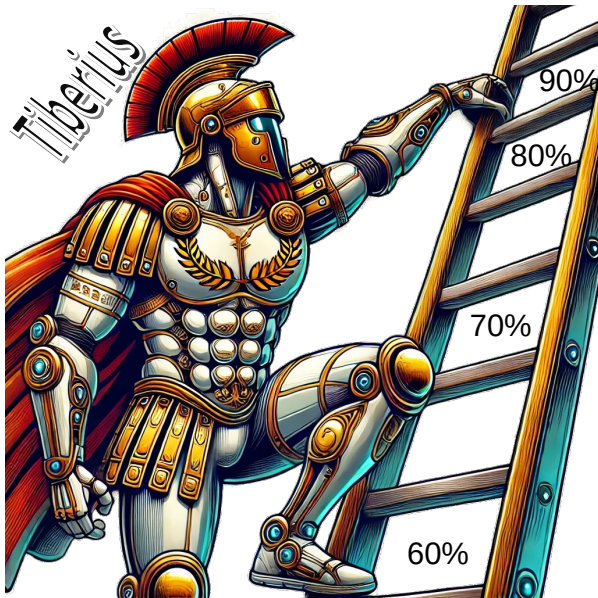


Image: ChatGPT by OpenAI, manual editing



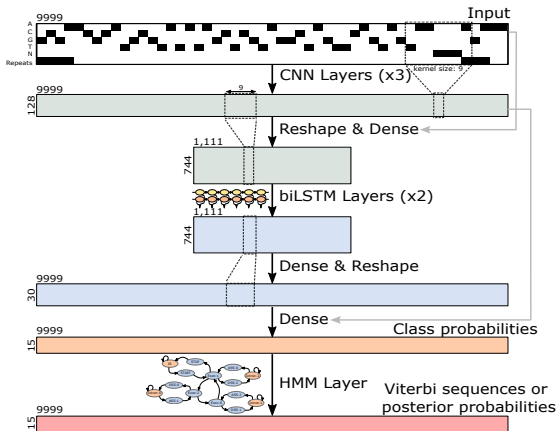
Lars Gabriel, Felix Becker, Katharina Hoff, Mario Stanke

Contributing to clade parameters:

Tomás Bruna, Asees Kaur, Anish Krishnan, Felix Ortmann, Asaf Salamov,
Igor Grigoriev, Samuel Talbot, Christopher Wheat, Richard Krieg

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 than Helixer
- higher accuracy
- (*ab initio* prediction)
- container for A100 GPU

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

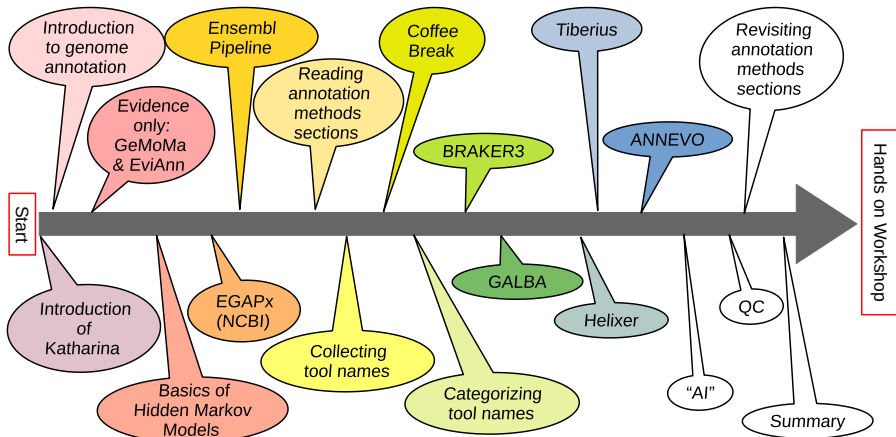
Training Tiberius for clades

Selected examples

Clade	# Train Species	Test species
Mammals	32	<i>Bos taurus</i> , <i>Delphinapterus leucas</i> , <i>Homo sapiens</i>
Eudicots	25	<i>Arabidopsis thaliana</i> , <i>Eschscholzia californica</i> , <i>Mimulus guttatus</i>
Monocots	20	<i>Brachypodium stacei</i> , <i>Freycinetia multiflora</i> , <i>Sorghum bicolor</i> , <i>Urochloa brizantha</i>
Insecta	100	<i>Bombyx mori</i> , <i>Colias croceus</i> , <i>Danaus plexippus</i>
Diatoms	10	<i>Phaeodactylum tricornutum</i> , <i>Thalassiosira pseudonana</i>
Vertebrates	66	<i>Archocentrus centrarchus</i> , <i>Gallus gallus</i> , <i>Homo sapiens</i> , <i>Zootoca vivipara</i>

Supervised Training Procedure

- Requires GPU and annotated genomes
- Training time can take up to 15 days
- ... more parameter sets are available



ANNEVO: HiC inspired fast genome annotation

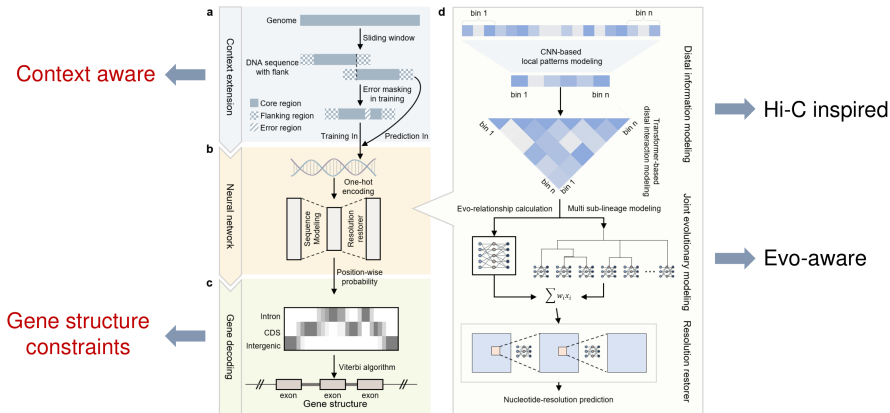


Image: ChatGPT by OpenAI

Highly accurate ab initio gene annotation with ANNEVO

Kai Ye, Pengyu Zhang, Tun Xu, Songbo Wang, Xiaofei Yang, Peisen Sun, Peng Jia, Bo Wang, Stephen Bush, Zemin Ning.

Under review at Nature Methods

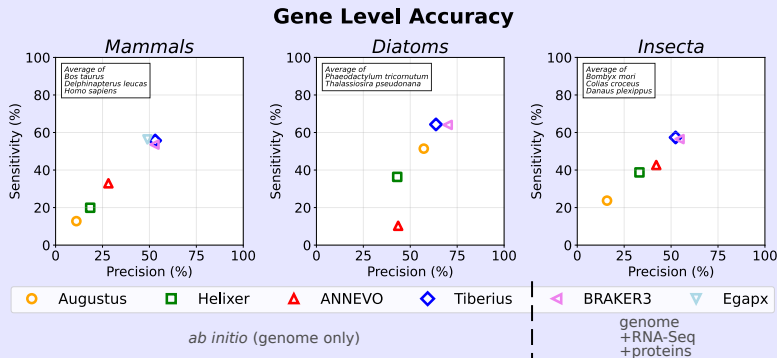


No RNA, No protein, No repeat masking

Image: courtesy of Kai Ye

Benchmark: Deep Learning *ab initio* vs. traditional pipelines with evidence

Average accuracy in Tiberius' test species



Non-dedicated model used for ANNEVO (invertebrates) & Helixer (landplants) on diatoms

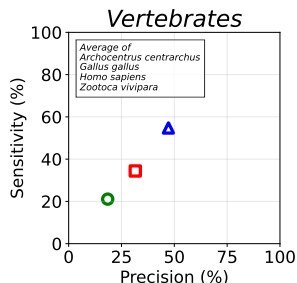
Take home message

- *ab initio* accuracy of Tiberius comparable to BRAKER3 & EGAPx
- much faster, no extrinsic data, only one isoform per gene

Benchmark: *ab initio* gene prediction tools

Average accuracy in Tiberius' test species

Gene Level Accuracy



○ Helixer □ ANNEVO ▲ Tiberius

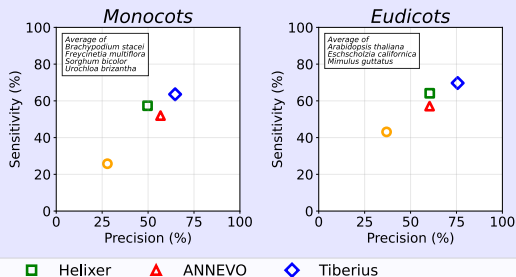
Test species include

- fish
- bird
- mammal
- reptile

Benchmark: *ab initio* gene prediction tools

Average accuracy in Tiberius' test species

Gene Level Accuracy



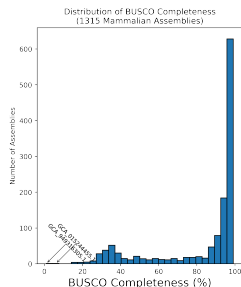
Tiberius enables large scale *ab initio* annotation of genomes

Mammalian Genomes

- 1314 assemblies from 788 species
- 344 primate and 970 assemblies not from primates
- Hiram Clawson (UCSC) has added 701 annotation sets to GenArk browser
(<https://hgdownload.soe.ucsc.edu/hubs/>)

Download:

<https://bioinf.uni-greifswald.de/bioinf/tiberius/genes/tib-tbl.html>



Tiberius Gene Predictions on Mammals

Below tables contains the predictions of the gene finder Tiberius on 344 primate assemblies and 970 mammalian assemblies that are not from primates.
[Tiberius v1.1.3](#) (GitHub commit d0c17c5) was run with this command line:

```
python3 tiberius.py --genome $(ACC).fa --out $(ACC).gtf
```

where ACC is the accession. The genomes were downloaded from Genbank in December 2024 and January 2025. [Mario Stanke](#), Greifswald, Germany, February 2, 2025

#	Accession	Species	Clade	Tiberius				Reference	Alternate species name	Genome Size
				Number of Genes	GTF	AA	codingseq			
1	GCA_001443585.1	Acinonyx jubatus	mammals	17566	GTF	AA	codingseq		cheetah	2372536860
2	GCF_027475565.1	Acinonyx jubatus	mammals	19294	GTF	AA	codingseq	reference genome	cheetah	2377417351
3	GCA_004027535.1	Acomys cahirinus	mammals	14398	GTF	AA	codingseq		Egyptian spiny mouse	2306070819

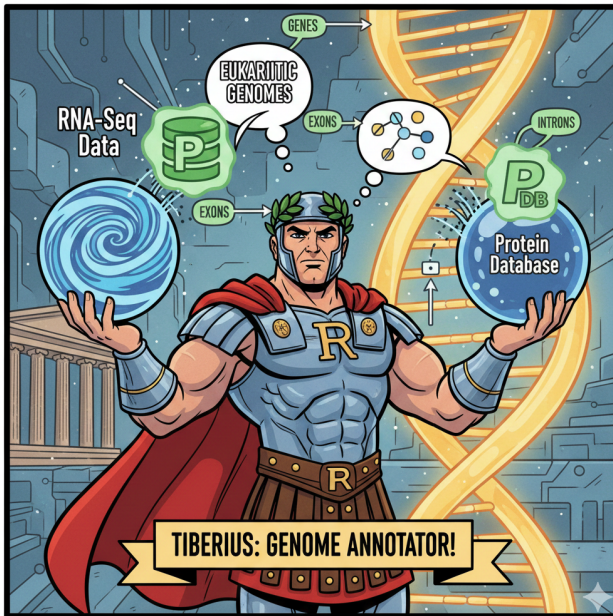
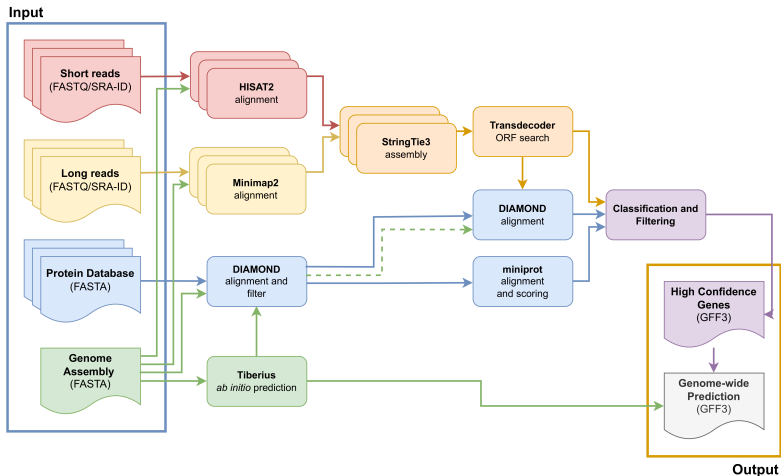
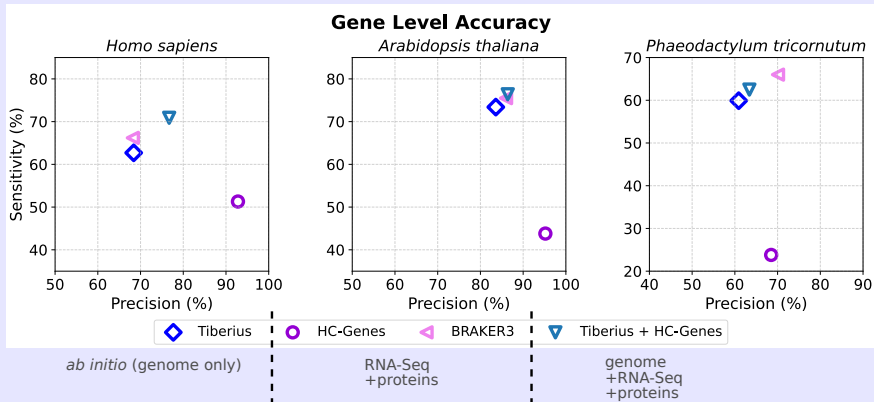


Image: Gemini with NanoBanana

Tiberius evidence processing pipeline

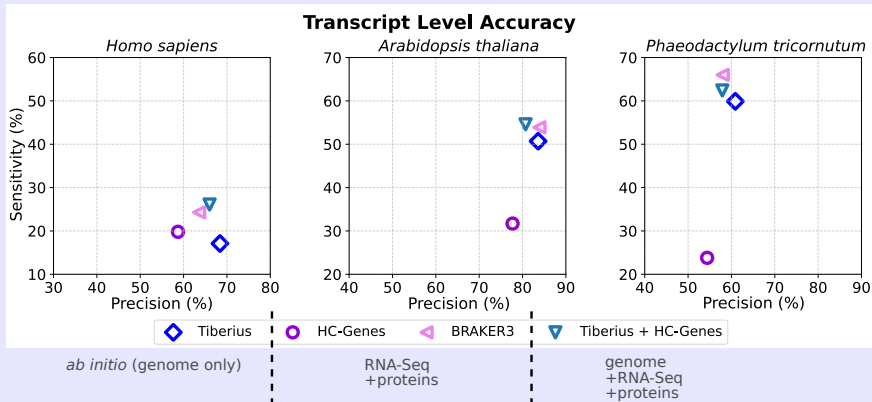


Benchmark: Tiberius with extrinsic evidence



Proteins: OrthoDB excluding proteins from species of the same order.

Benchmark: Tiberius with extrinsic evidence



Proteins: OrthoDB excluding proteins from species of the same order.

Runtime comparison

Species	Runtime (min:sec)				
	ANNEVO	GPU Tiberius*	Helixer	CPU BRAKER3	HC-Genes*
<i>Arabidopsis thaliana</i>	3:49	6:45	13:12	242:02	202:28
<i>Homo sapiens</i>	87:09	116:02	567:36	2663:15	233:43
<i>Phaeodactylum tricornutum</i>	0:46	3:29	8:24	307:20	209:40

Conditions

- Runtime for whole-genome annotation
- GPU-based methods: 1x A100
- CPU-based methods: 100 threads
- * Can be further parallelized with Nextflow on an HPC

Take home message

- using evidence increases accuracy of Tiberius
- runtime shorter than BRAKER3 but substantial
- very new, we can't cover it in the hands-on session this year

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

Tiberius *ab initio*

```
tiberius.py --genome genome.fa --out tiberius.gtf \  
            --model_cfg eudicots.cfg
```

Tiberius with evidence

```
tiberius.py --params_yaml parameter.yaml \  
            --nf_config nextflow.config
```

Tiberius via singularity

```
singularity build tiberius.sif \  
    docker://larsgabriel23/tiberius:latest  
singularity run --nv tiberius.sif \  
    tiberius.py --genome genome.fa \  
    --out tiberius.gtf \  
    --model_cfg eudicots.cfg
```

Tiberius is under MIT License

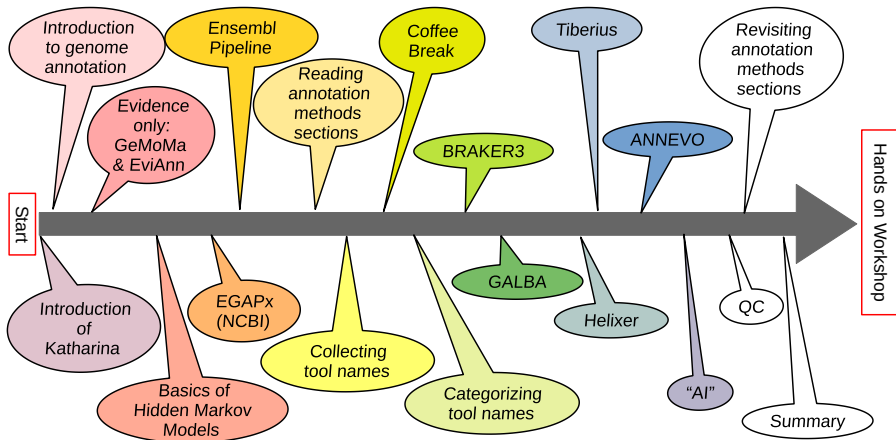


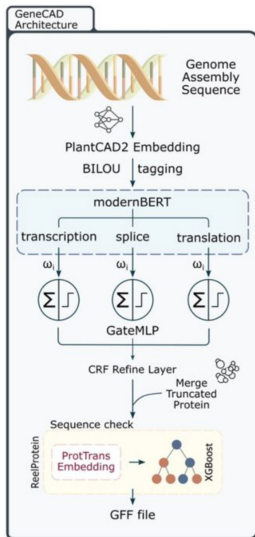


Image: Gemini with NanoBanana

GeneCAD: Plant Genome Annotation with a DNA Foundation Model

✉ Zong-Yan Liu, ✉ Ana Berthel, ✉ Eric Czech, ✉ Michelle Stitzer, ✉ Sheng-Kai Hsu, Matt Pennell,
✉ Edward S. Buckler, ✉ Jingjing Zhai

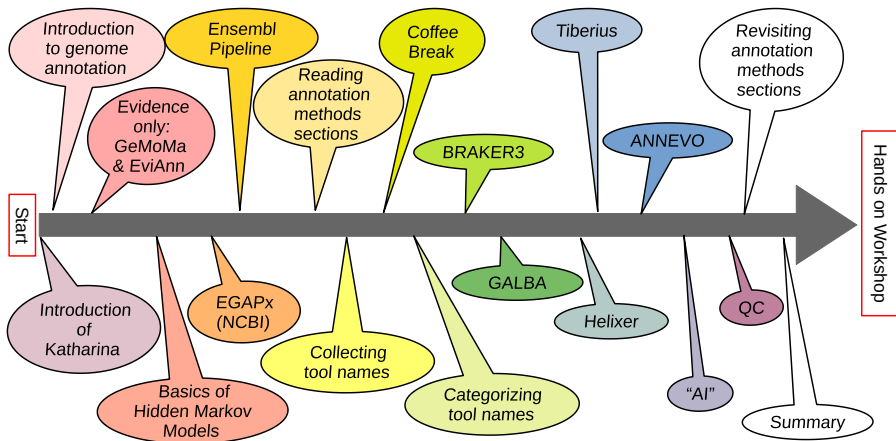
doi: <https://doi.org/10.1101/2025.10.31.685877>



- for plants only
- CDS locus accuracy unclear
- demonstrates potential

Vipsania: A New Genome Foundation Model for Gene Finding





Did we do a good job?



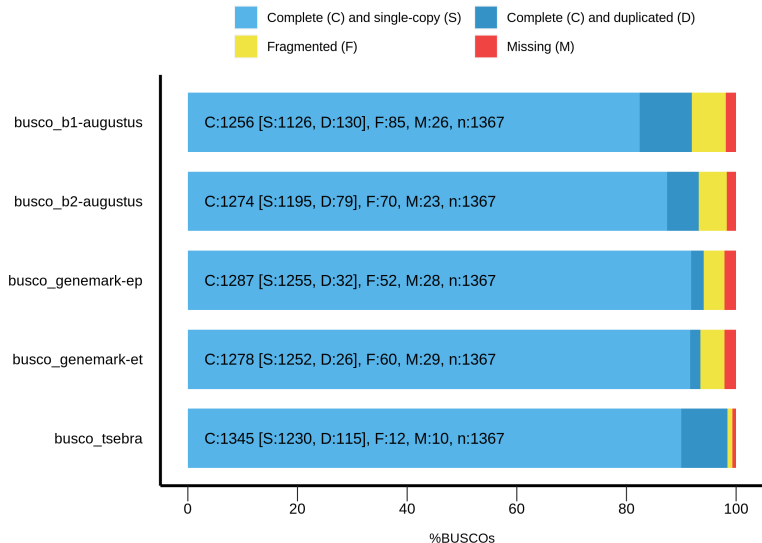
Describe your annotation

- 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.

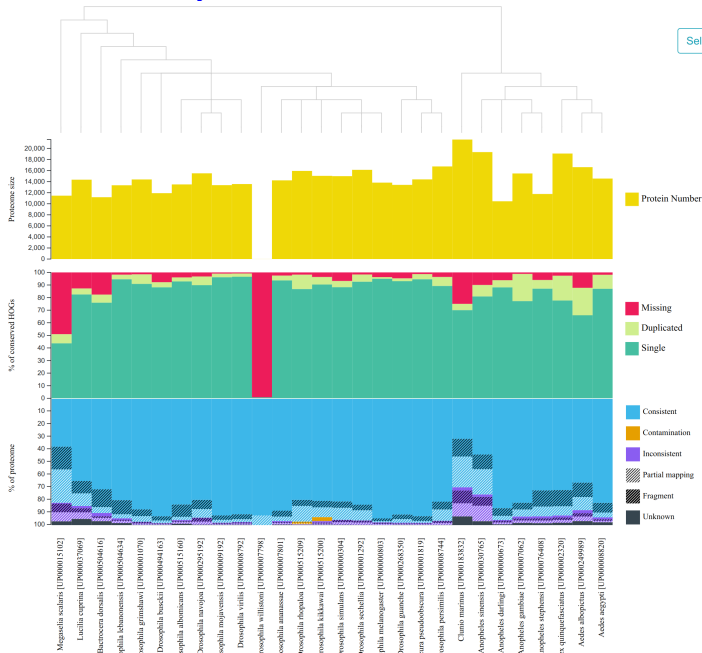
BUSCO: Sensitivity in clade-specific conserved genes

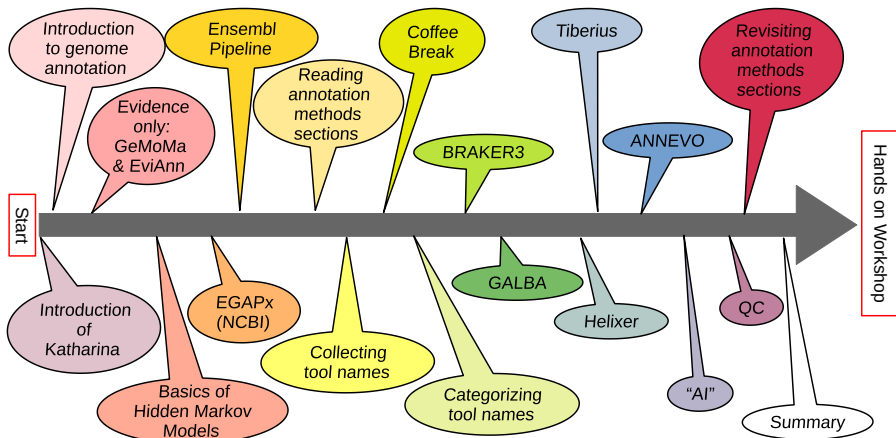
BUSCO Assessment Results



Beware! BUSCO completeness does not warrant correct gene structures!

OMark: sensitivity, contaminations, & more

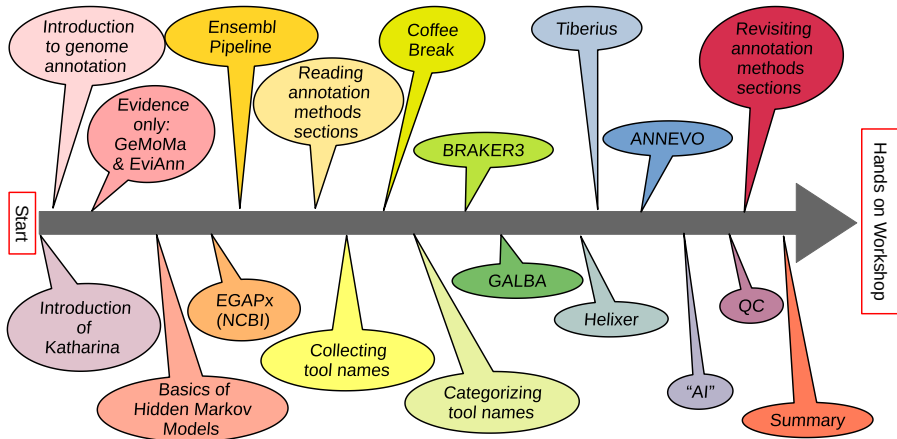




Revisiting annotation methods sections

Your tasks

- 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



Most important stuff on genome annotation

- structural genome annotation in eukaryotes is hard
- Hidden Markov Models (& CRFs) remain essential
- evidence improves accuracy
- popular community annotation pipelines:
 - ① BRAKER
 - ② GALBA
 - ③ EGAPx
- deep learning is changing the field
 - ① Tiberius most accurate
 - ② Tiberius can become even more accurate with evidence
- "looking nice" is not always "correct"
- BUSCO & OMArk
- high marker gene detection rate $\neq$ high accuracy

Navigating the maize of tools

What would Katharina try?

