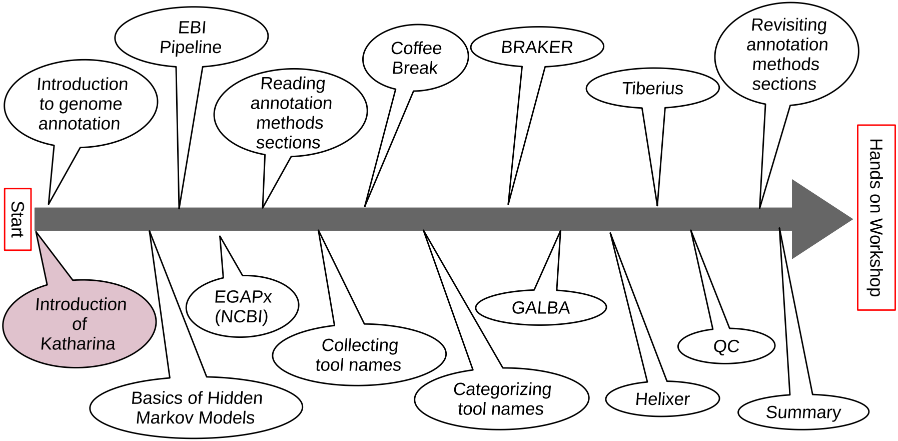
Annotation of Protein Coding Genes

January 8th 2025

Katharina J. Hoff

Twitter: @katharina_hoff
Bluesky: @katharinahoff.bsky.social
Mastodon: @KatharinaHoff@fosstodon.org

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


Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
EGAPx (NCBI)
Reading annotation methods sections
Collecting tool names
Coffee Break
Categorizing tool names
BRAKER
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

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)

Research

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

Teaching

- 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...

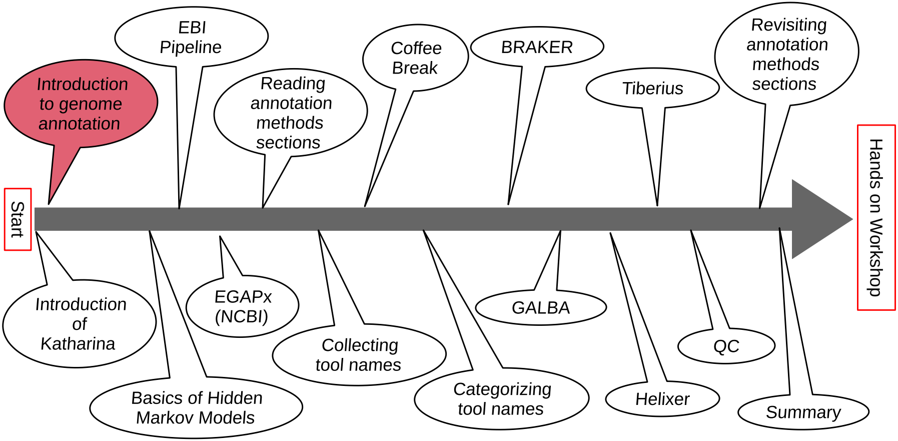
After this lecture, you will...

- 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

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


Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
EGAPx (NCBI)
Reading annotation methods sections
Collecting tool names
Coffee Break
Categorizing tool names
BRAKER
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

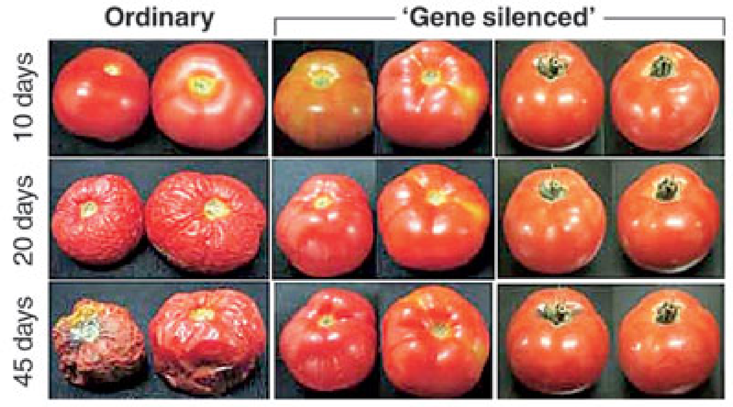
Where are the protein coding genes?

Genomic sequence: chicken

```
cctcacctctgagaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctactttgcagctgctatt
tcgagtcaaggtgtaggcagagtcctttttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagtagatattcagtaa
cattcaagttcctatttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtcagactg
ttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctctctgtgccttggtttcccc
atctgtcatatgaagggagtgcgatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttggggttctggtagctccacagggatgctcaatgaagatgcaaaattagaagtcaaaat
aaacagctcccatgggcagtgttgatctcaccctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaacccctggctgccctgggaggcaagggtgaggg
ctggatttttaaaggggggcctgttttggggagggggtgatgagcgctggggaggcagctctcaggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggggttcctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaaccatgctagaaaaacatgtggaaaagtcactacca
ggctggcagggaatggggcaatctattcatactgattgcaatgcccactggttcctaatctgggcaacc
cctggggcccacagctaaatccagtgagtggaagttacaggagtctgcttccagtgctgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaagccagatgacctcaaaggtcccatgggattctaatctgtctgctccttgttctacagattc
caaaccaaaagaggcaagcaagtctgcgctgaccccagtgagtcctgggtccaggagtacgtgtatgac
ctggaactgaactgagctgctcagagacaggaagtcttc
```

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

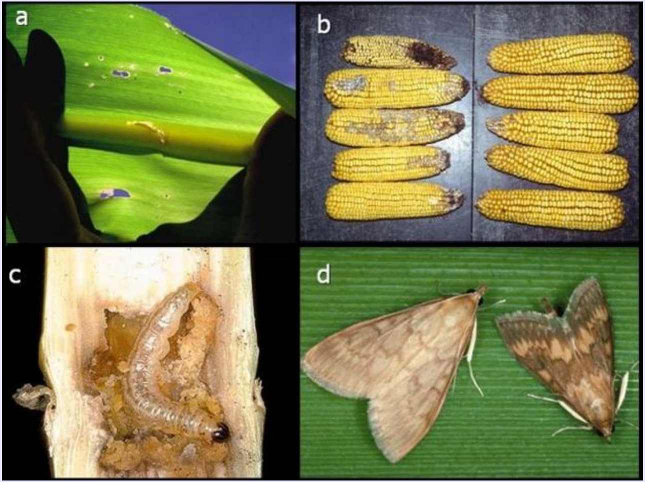


Image: Hellmich & Hellmich (2012) Nature Education Knowledge 3(10):4

`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**

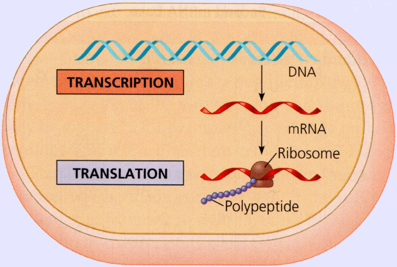
It does not take a village to publish a genome!

- 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?

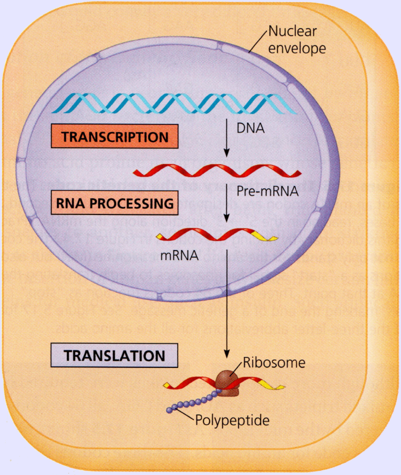
Transcription & Translation

Prokaryotes



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

Eukaryotes



How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

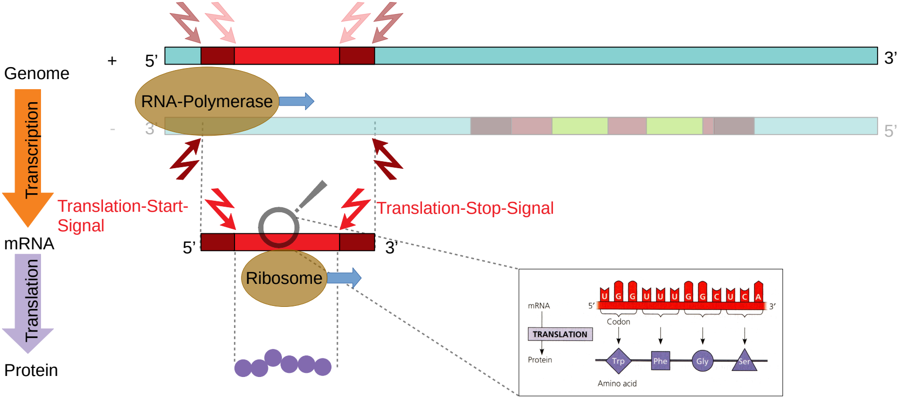
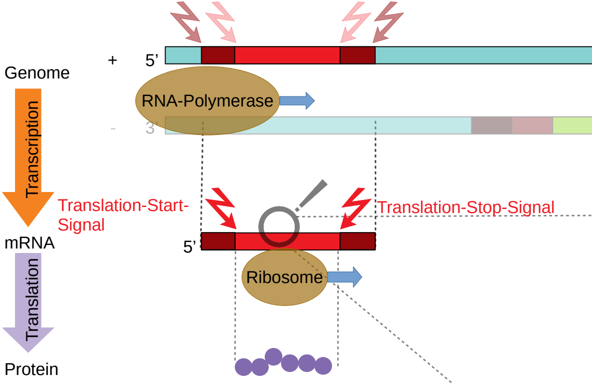


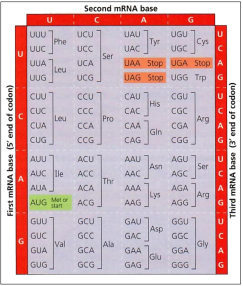
Image: Campbell et al., Biology, San Francisco, 2008, p. 329, FIg. 17.4

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*





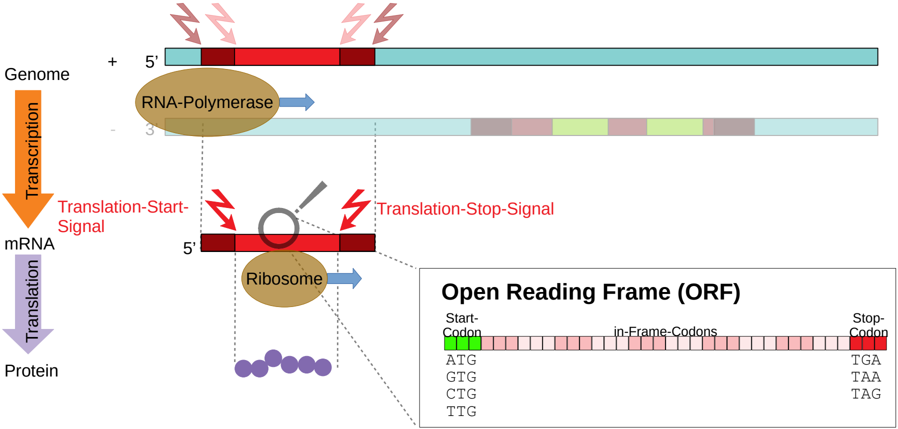
		Second mRNA base				
First mRNA base (5' end of codon)		U	C	A	G	Third mRNA base (3' end of codon)
U		UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
		UUC	UCC	UAC	UGC	C
		UUA Leu	UCA	UAA Stop	UGA Stop	A
		UUG	UCG	UAG Stop	UGG Trp	G
C		CUU Leu	CCU Pro	CAU His	CGU Arg	U
		CUC	CCC	CAC	CGC	C
		CUA	CCA	CAA Gln	CGA	A
		CUG	CCG	CAG	CGG	G
A		AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
		AUC	ACC	AAC	AGC	C
		AUA	ACA	AAA Lys	AGA Arg	A
		AUG Met or start	ACG	AAG	AGG	G
G		GUU Val	GCU Ala	GAU Asp	GGU Gly	U
		GUC	GCC	GAC	GGC	C
		GUA	GCA	GAA Glu	GGA	A
		GUG	GCG	GAG	GGG	G

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?

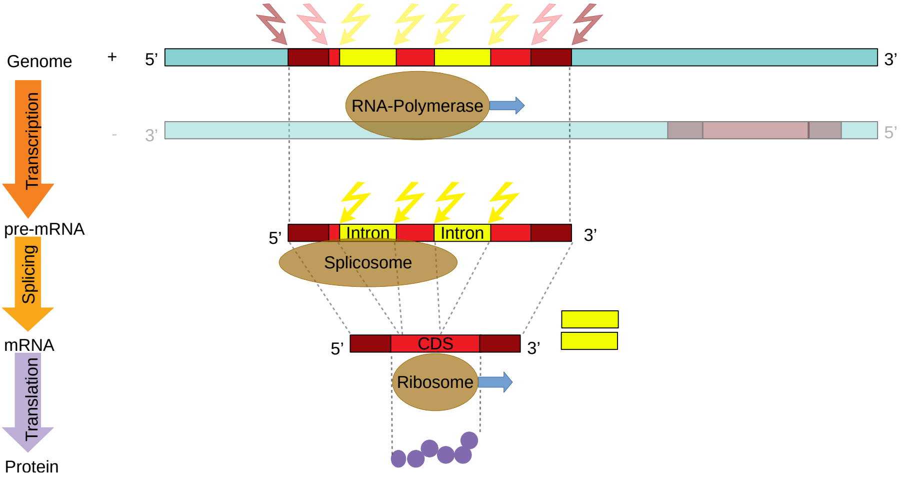
Prokaryotes & Eukaryotes*



- 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

```
cctcacctctgagaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtcctttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagtagatattcagtaa
cattcaagttcctatttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtcagactg
ttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctctctgtgccttggtttcccc
atctgtcatatgaagggagtgcgatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttggggttctggtagctccacagggatgctcaatgaagatgcaaattagaagtcaaaat
aaacagctcccatgggcagtgttgatctcaccctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaacccctggctgccctgggaggcaagggtgaggg
ctggattttaaagggggcctgttttggggagggggtgatgagcgctggggaggcagctctcaggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggggttcctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaaccatgctagaaaaacatgtggaaaagtcactacca
ggctggcagggaatggggcaatctattcatactgattgcaatgcccactggttcctaatctgggcaacc
cctggggcccacagctaaatccagtgagtggaagttacagggagtctgcttccagtgctgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaagccagatgacctcaaaggtcccatgggattctaatctgtctgctccttgttctacagattc
caaaccaaaagaggcaagcaagtctgcgctgaccccagtgagtcctgggtccaggagtacgtgtatgac

ctggaactgaactgagctgctcagagacaggaagtcttc
```

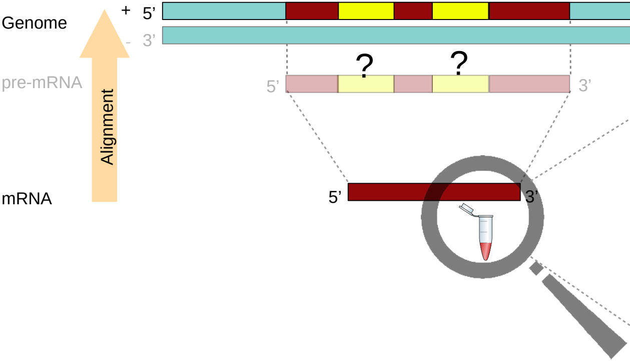
The Genome Annotation Problem

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

```
cctcacctctgagaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctactttgcagctgctatt
tcgagtcaaggtgtaggcagagtccttttttctagtcatggctggcaaacagtgggatctgggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagtagatattcagtaa
cattcaagttcctatttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtcagactg
ttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctctctgtgccttggtttcccc
atctgtcatatgaagggagtgcgatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttggggttctggtagctccacagggatgctcaatgaagatgcaaattagaagtcaaaat
aaacagctcccatgggcagtgttgatctcaccctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaacccctggctgccctgggaggcaagggtgaggg
ctggattttaaagggggcctgttttgggagggggtgatgagcgctggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggggttcctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaaccatgctagaaaaacatgtggaaaagtcactacca
ggctggcagggaatggggcaatctattcatactgattgcaatgcccactggttcctaatctgggcaacc
cctggggcccacagctaaatccagtgagtggaagttacaggagtctgcttccagtgctgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaagccagatgacctcaaaggtcccatgggattctaatctgtctgctccttgttctacagattc
caaaccaaaagaggcaagcaagtctgcgctgaccccagtgagtcctgggtccaggagtacgtgtatgac
ctggaactgaactgagctgctcagagacaggaagtcttc
```

What aids in the identification of genes in genomes?

Evidence data from transcription



Transcription Data:

- Expressed Sequence Tags
- cDNA
- mRNA-Seq
- mRNA IsoSeq
- ...

Provides information on:

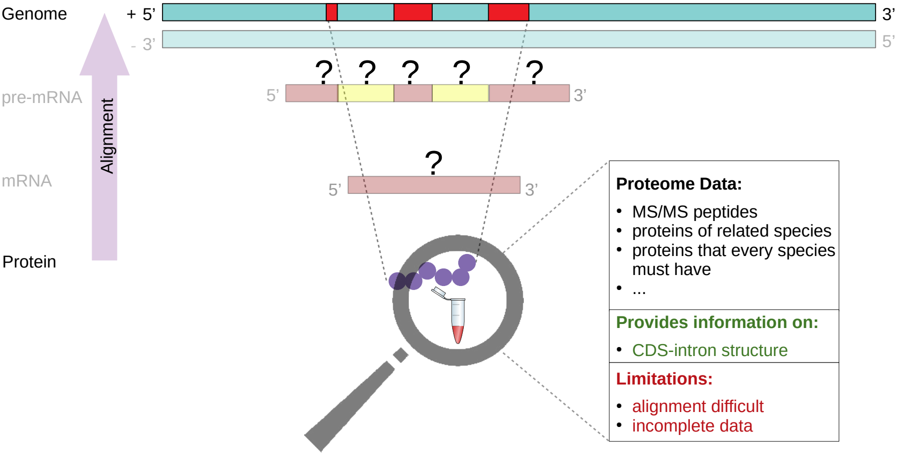
- exon-intron structure

Limitations:

- no info on reading frame
- incomplete data

What aids in the identification of genes in genomes?

Evidence data from translation



Proteome Data:

- MS/MS peptides
- proteins of related species
- proteins that every species must have
- ...

Provides information on:

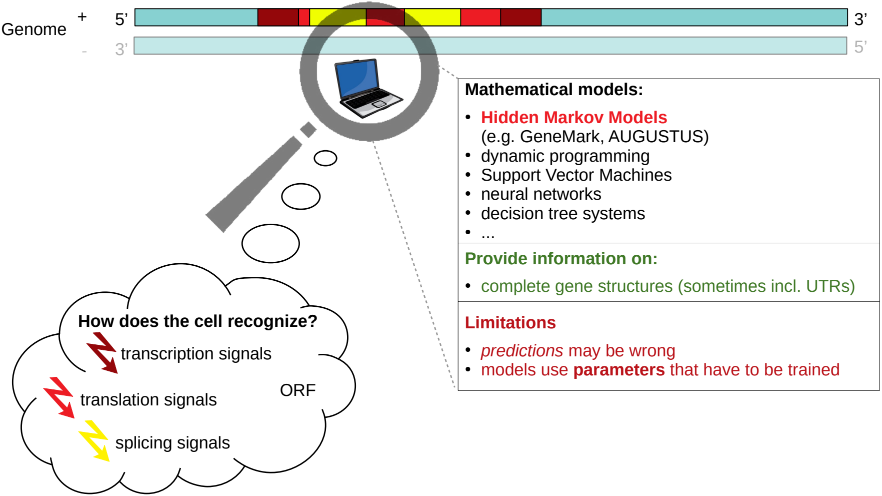
- CDS-intron structure

Limitations:

- alignment difficult
- incomplete data

What aids in the identification of genes in genomes?

Mathematical models



How does the cell recognize?

- transcription signals
- translation signals
- splicing signals

Mathematical models:

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

Provide information on:

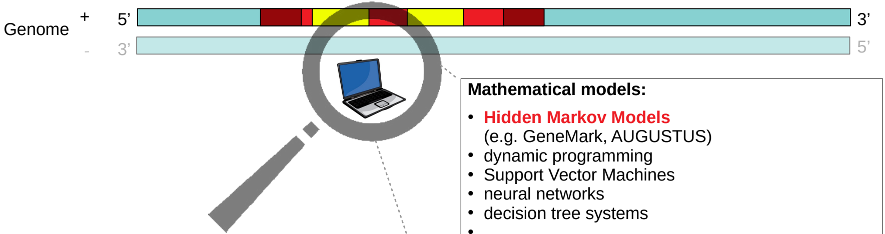
- complete gene structures (sometimes incl. UTRs)

Limitations

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

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
- ...

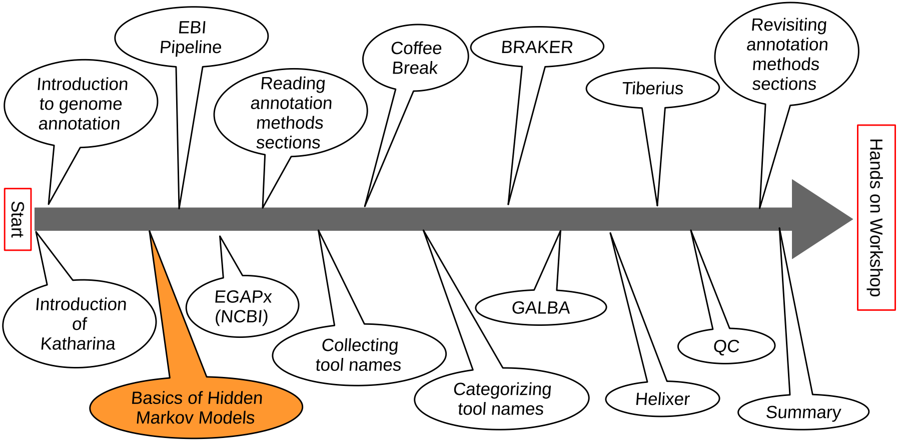
Provide information on:

- 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).


Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
BRAKER
Categorizing tool names
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

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

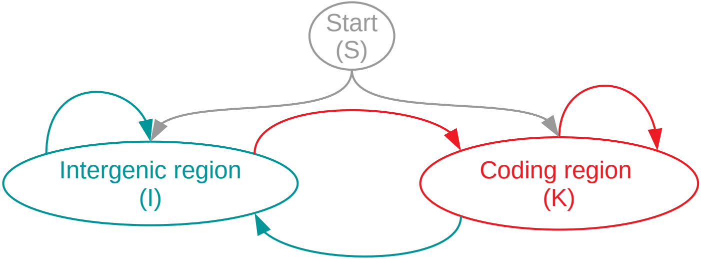
Goal: “Most likely path through hidden states”

```
e.g. AABBBAA
or   IIKKIKI    P(path) = 0.3%
```

Basis of highly accurate gene prediction tools

Hidden Markov Model

Hidden states

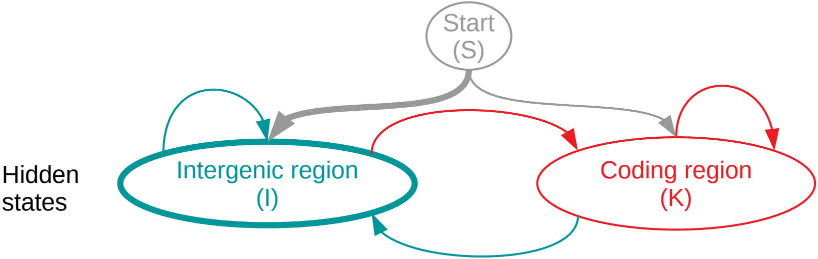


A possible 'state path' for the genome sequence:

AABBBAA

Basis of highly accurate gene prediction tools

Hidden Markov Model

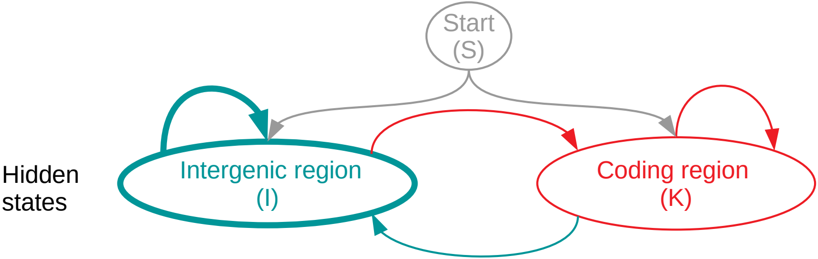


A possible 'state path' for the genome sequence:

```
AABBBAA
I
```

Basis of highly accurate gene prediction tools

Hidden Markov Model

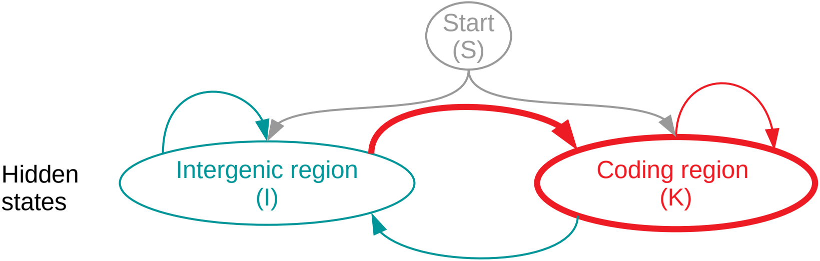


A possible 'state path' for the genome sequence:

```
AABBBAA
II
```

Basis of highly accurate gene prediction tools

Hidden Markov Model



A possible 'state path' for the genome sequence:

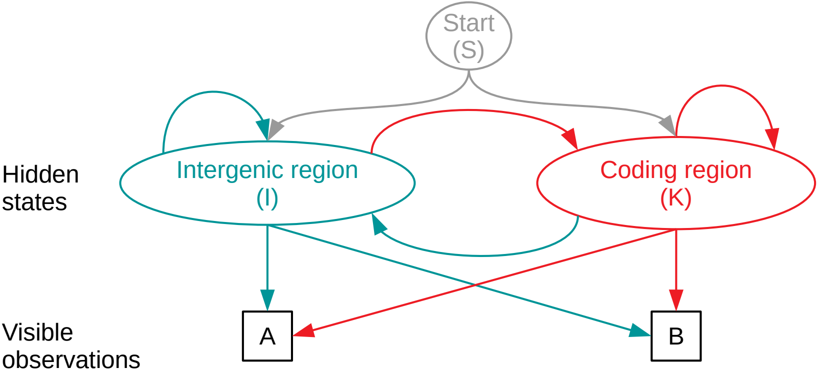
```
AABBBAA
IIK...
```

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

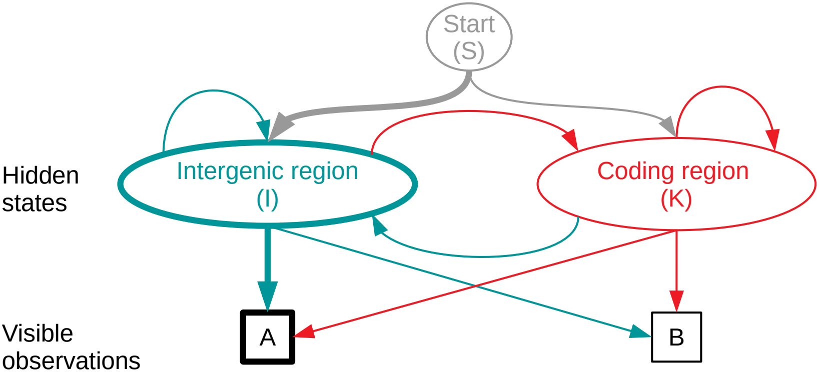
Hidden Markov Model



A possible ‘state path’ for the genome sequence:

Basis of highly accurate gene prediction tools

Hidden Markov Model

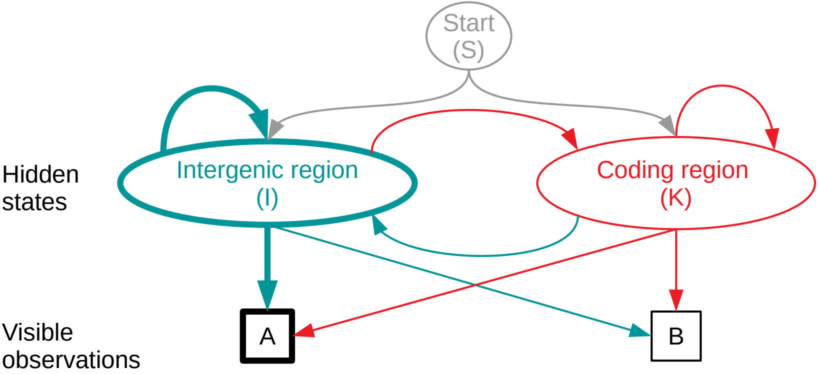


A possible ‘state path’ for the genome sequence:

```
A
I
```

Basis of highly accurate gene prediction tools

Hidden Markov Model

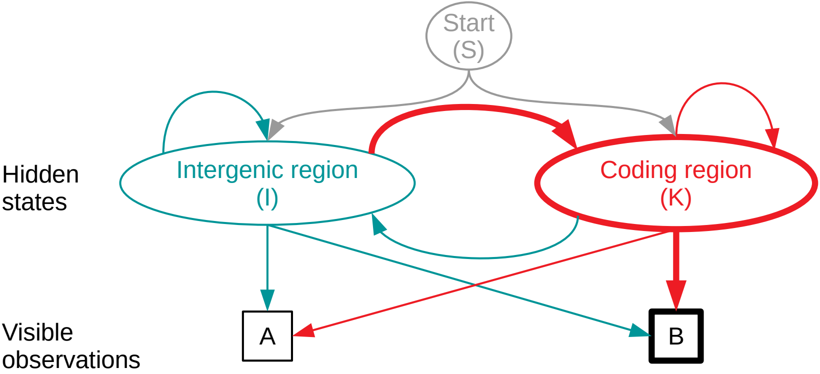


A possible 'state path' for the genome sequence:

```
AA
II
```

Basis of highly accurate gene prediction tools

Hidden Markov Model

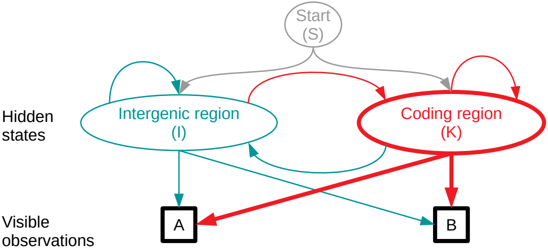


A possible ‘state path’ for the genome sequence:

```
AAB...
IIK...
```

Basis of highly accurate gene prediction tools

Hidden Markov Model

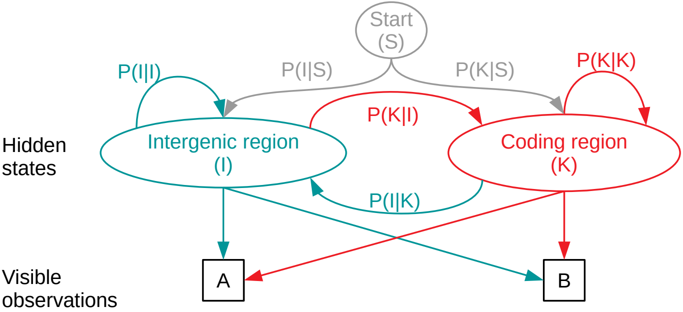


Model properties

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

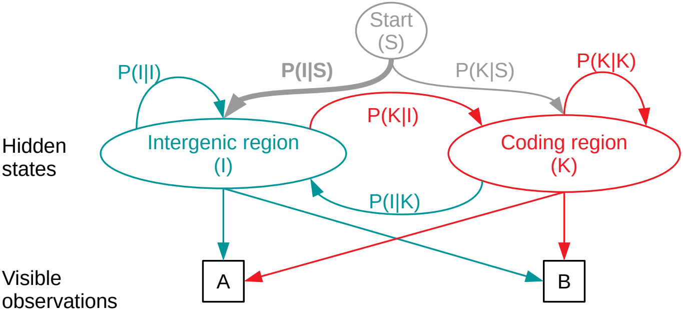


How likely are the state transitions?

Use data with known state transitions for learning!

Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

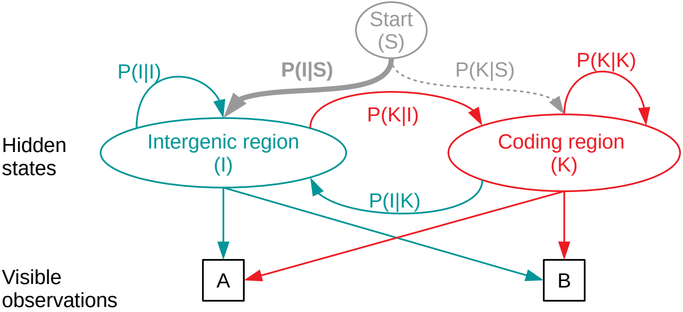
AABABA
IKKIII

Start probability
P(I|S) = ?

Use data with known state transitions for learning!

Basis of highly accurate gene prediction tools

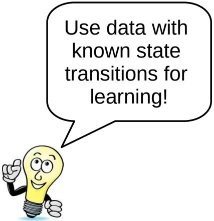
Hidden Markov Model



Training data:

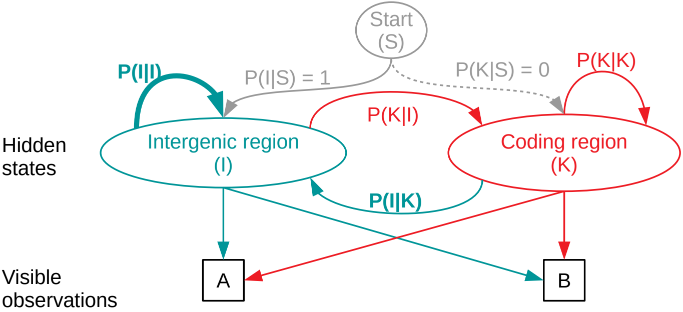
AABABA
IKKIII
+

Start probability
P(I|S) = 1



Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

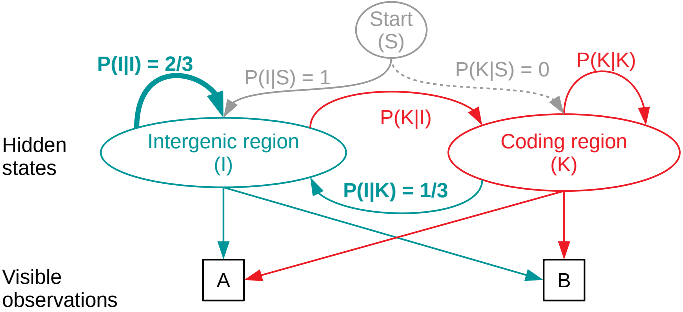
AABABA
IKKIII

P(I|I) = ?

Use data with known state transitions for learning!

Basis of highly accurate gene prediction tools

Hidden Markov Model



Training data:

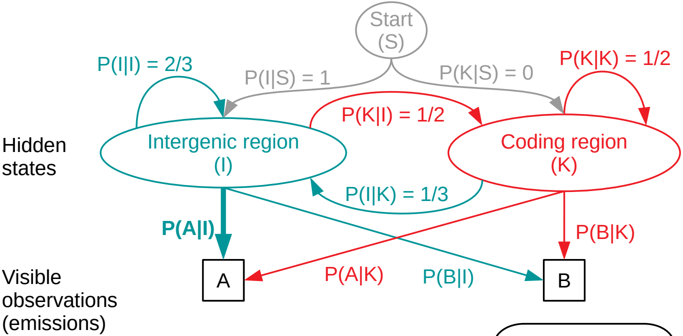
```
AABABA
IKKIII
   -++
```

P(I|I) = 2/3

P(I|K) = 1 – P(I|I) = 1/3

Basis of highly accurate gene prediction tools

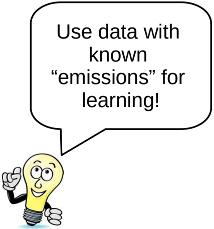
Hidden Markov Model



How likely are the observations?

AABABA
IKKIII

P(A|I) = ?

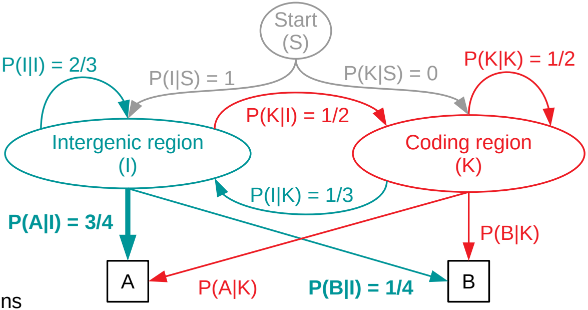


Basis of highly accurate gene prediction tools

Hidden Markov Model

Hidden states

Visible observations (emissions)



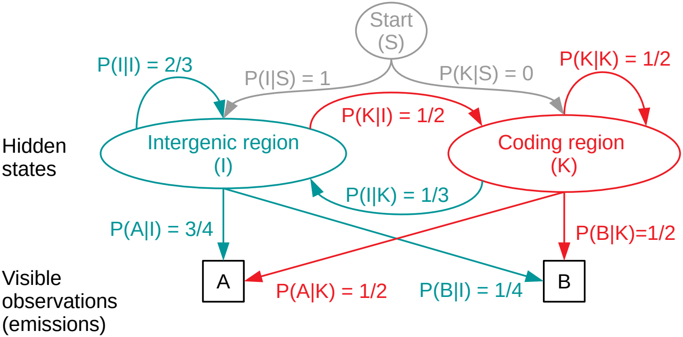
How likely are the observations?

```
AABABA
IKKIII
+  +-+
```

P(A|I) = ¾   P(B|I) = 1 – P(A|I) = 1 – ¾ = ¼

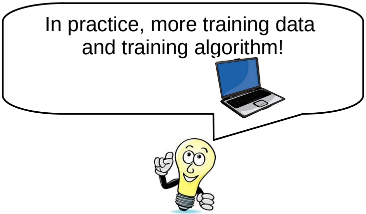
Basis of highly accurate gene prediction tools

Hidden Markov Model



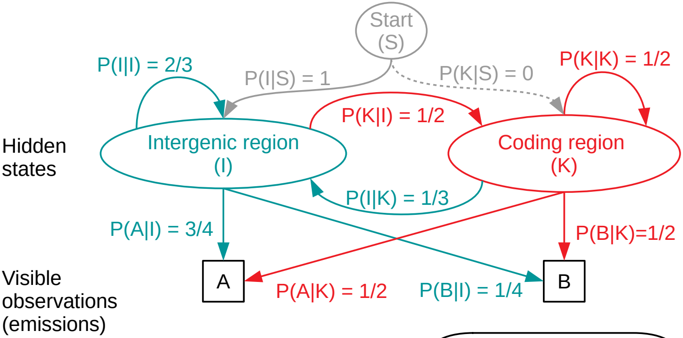
Training data:

AABABA
IKKIII



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

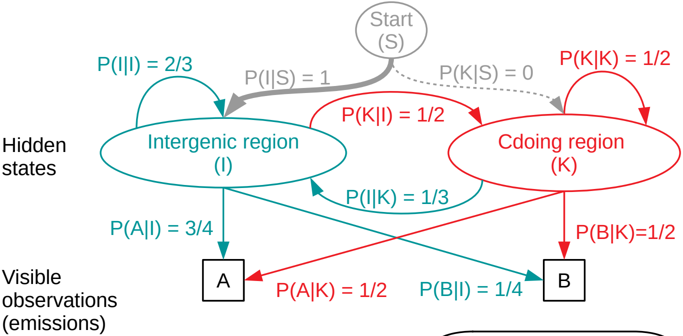
```
Path = AAB
       IKK

P(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?

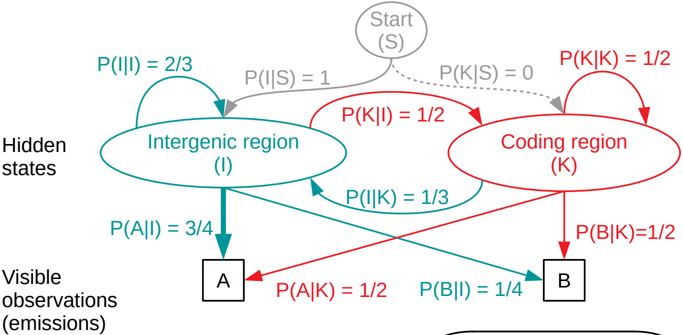
```
Path = AAB
       IKK

P(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?

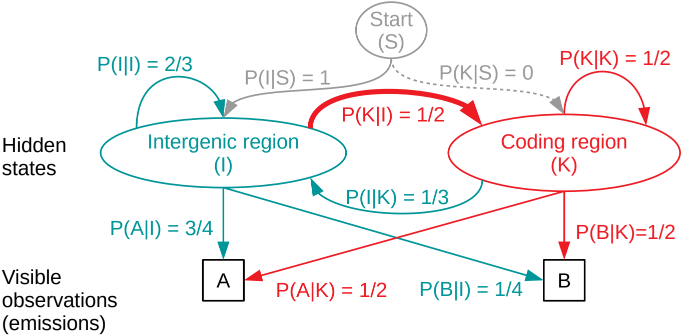
```
Path = AAB
       IKK

P(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



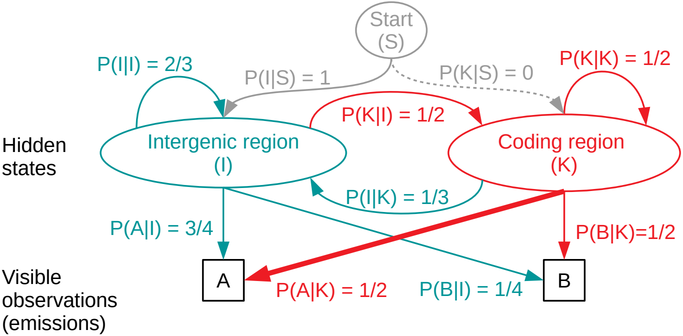
How likely is a given state-emission path?

```
Path = AAB
       IKK

P(Path) = P(I|S)*P(A|I)*P(K|I)
```

Basis of highly accurate gene prediction tools

Hidden Markov Model



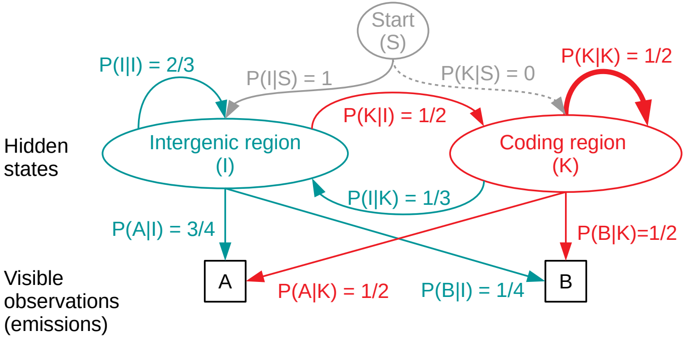
How likely is a given state-emission path?

```
Path = AAB
       IKK

P(Path) = P(I|S)*P(A|I)*P(K|I)*P(A|K)
```

Basis of highly accurate gene prediction tools

Hidden Markov Model



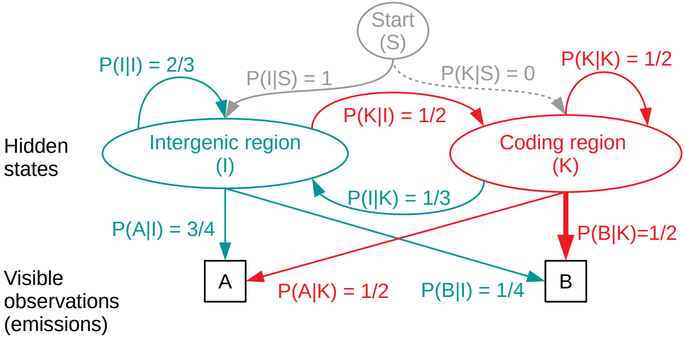
How likely is a given state-emission path?

```
Path = AAB
       IKK

P(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



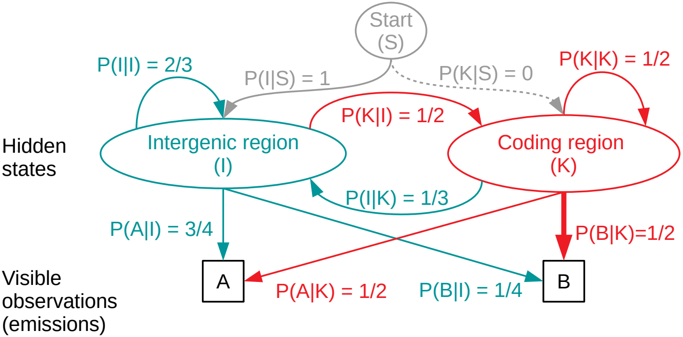
How likely is a given state-emission path?

```
Path = AAB
       IKK
```

```
P(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
       IKK
```

```
P(Path) = P(I|S)*P(A|I)*P(K|I)*P(A|K)*P(K|K)*P(B|K)

        = 1 * ¾ * ½ * ½ * ½ * ½

        = 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”

```
IIIKKKKK
KKIKKIIIK
IIKIIIKIK
IKKIKIIIK
KIKIKKKIK
KKKIKIKKK
```

...

Idea:

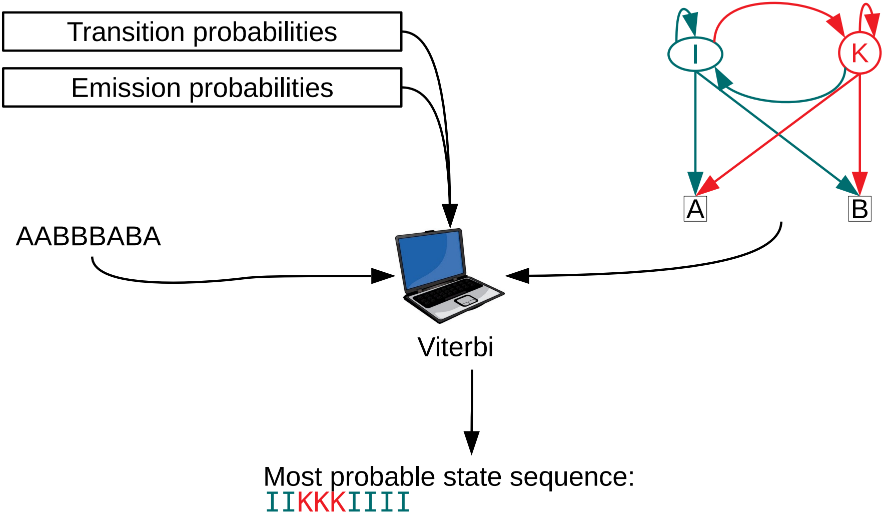
1. Generate all possible state sequences
2. Calculate the probability for each state sequence
3. Choose the state sequence with the highest probability

⇒ too expensive!

Basis of highly accurate gene prediction tools

Hidden Markov Model

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



Transition probabilities

Emission probabilities

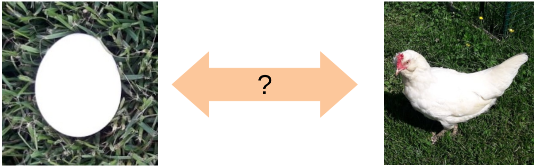
AABBBABA

Viterbi

Most probable state sequence:
IIKKIIII

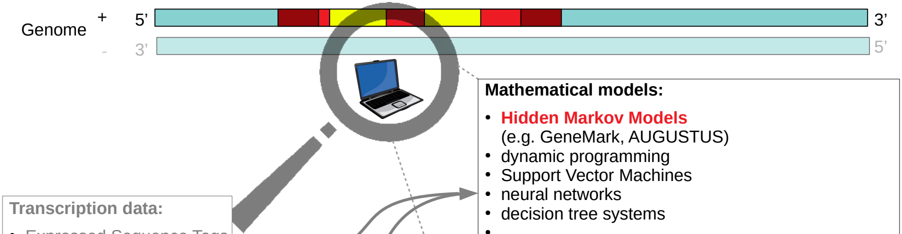
Hidden Markov Model for gene identification in practice

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



Gene
(Parameter training)

Gene
(Prediction)



Transcription data:

- Expressed Sequence Tags
- cDNA
- mRNA-Seq
- mRNA IsoSeq
- ...

Proteome data:

- MS/MS peptides
- proteins of related species
- proteins that every species must have
- ...

Mathematical models:

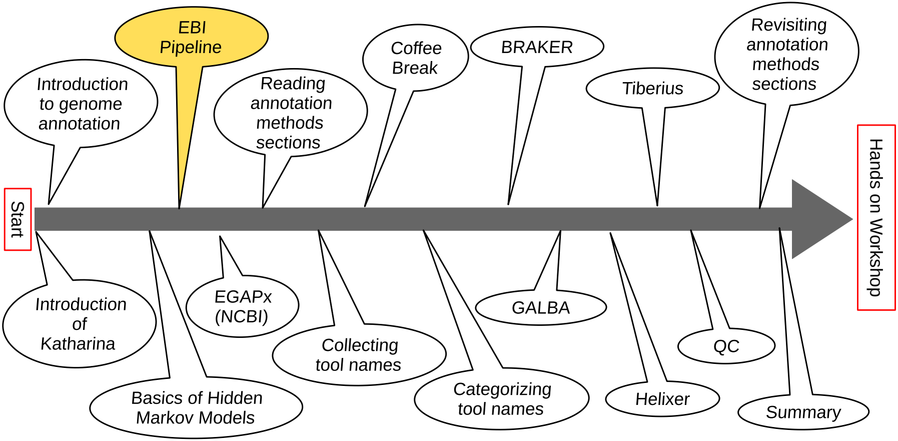
- **Hidden Markov Models** (e.g. GeneMark, AUGUSTUS)
- dynamic programming
- Support Vector Machines
- neural networks
- decision tree systems
- ...

Provide information on:

- complete gene structures (sometimes incl. UTRs)

Limitations

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



Start

Introduction of Katharina

Introduction to genome annotation

Basics of Hidden Markov Models

EBI Pipeline

EGAPx (NCBI)

Reading annotation methods sections

Collecting tool names

Coffee Break

Categorizing tool names

BRAKER

GALBA

Helixer

Tiberius

QC

Revisiting annotation methods sections

Summary

Hands on Workshop

EBI: Ensembl annotation system

Ensembl annotation pipelines

Ensembl annotation pipeline for non-vertebrates

Documentation

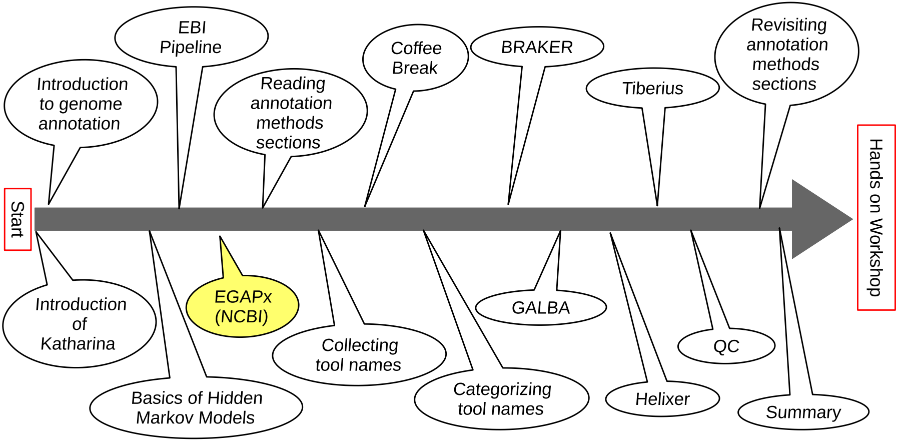
- Ensembl vertebrate pipeline: `https://rapid.ensembl.org/info/genome/genebuild/full_genebuild.html`
- Ensembl non-vertebrate pipeline: `https://rapid.ensembl.org/info/genome/genebuild/anno.html`
- BRAKER2 in Ensembl: `https://rapid.ensembl.org/info/genome/genebuild/braker.html`

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


Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
BRAKER
Categorizing tool names
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

Genome Annotation at the NCBI

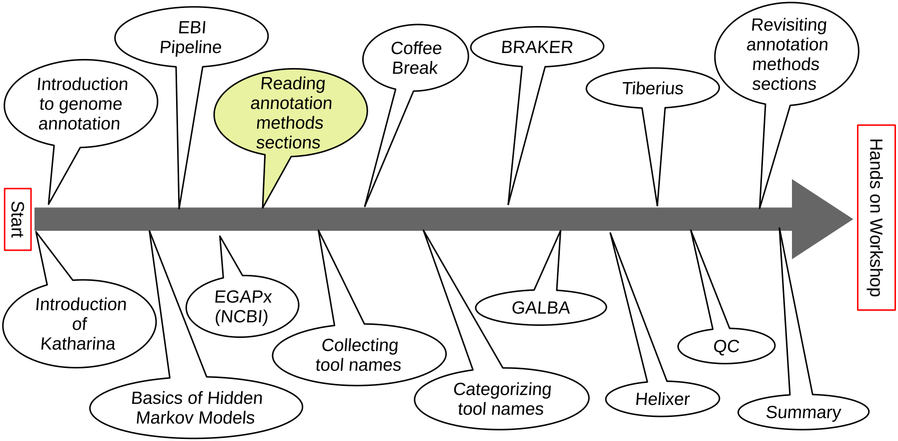
Different Annotation Scenarios

- 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!



Start

Introduction of Katharina

Introduction to genome annotation

Basics of Hidden Markov Models

EBI Pipeline

EGAPx (NCBI)

Reading annotation methods sections

Collecting tool names

Coffee Break

Categorizing tool names

BRAKER

GALBA

Helixer

Tiberius

QC

Revisiting annotation methods sections

Summary

Hands on Workshop

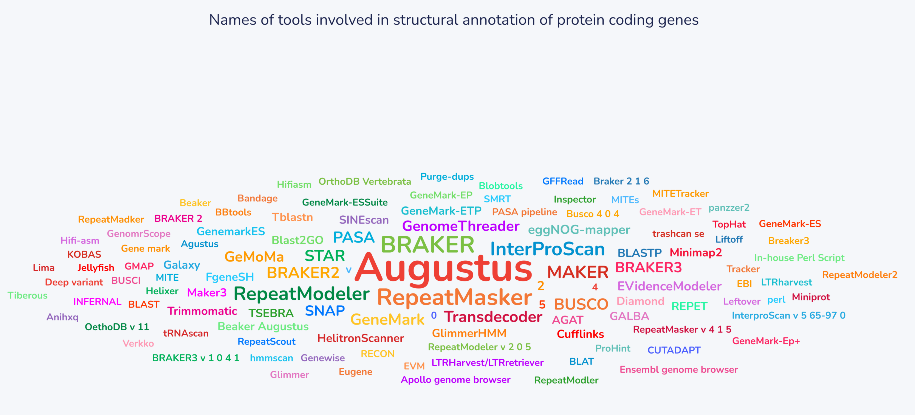
Read your methods snippet

Focus on structural annotation of protein coding genes only!

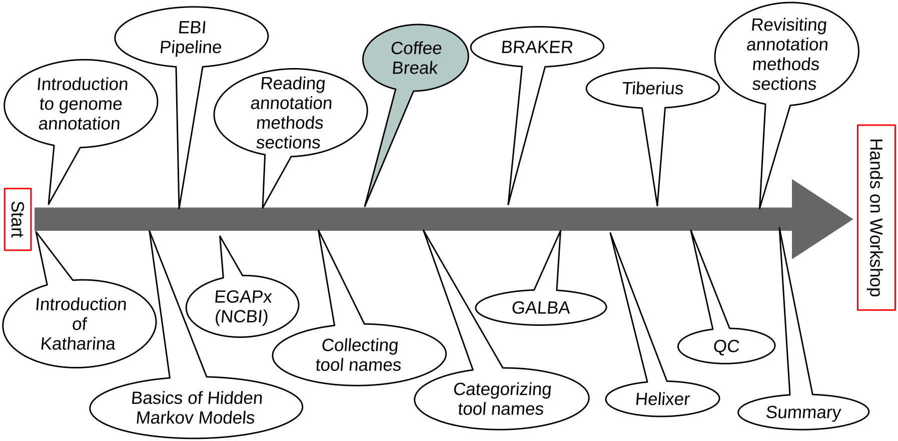
1. We move to Wooclap
2. Enter the names of tools involved in structural annotation of protein coding genes

Read your methods snippet

Focus on structural annotation of protein coding genes only!



Names of tools involved in structural annotation of protein coding genes



Start

Introduction of Katharina

Introduction to genome annotation

Basics of Hidden Markov Models

EBI Pipeline

EGAPx (NCBI)

Reading annotation methods sections

Collecting tool names

Coffee Break

Categorizing tool names

BRAKER

GALBA

Helixer

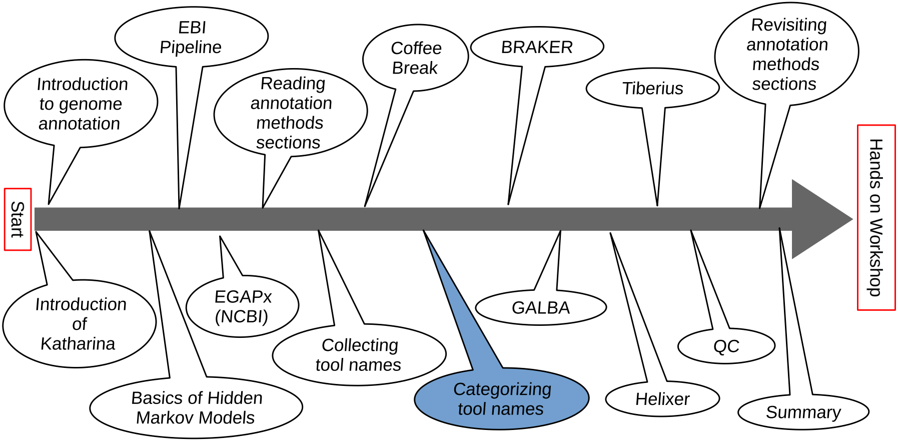
Tiberius

QC

Revisiting annotation methods sections

Summary

Hands on Workshop


Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
Categorizing tool names
BRAKER
GALBA
Tiberius
Helixer
Revisiting annotation methods sections
QC
Summary
Hands on Workshop

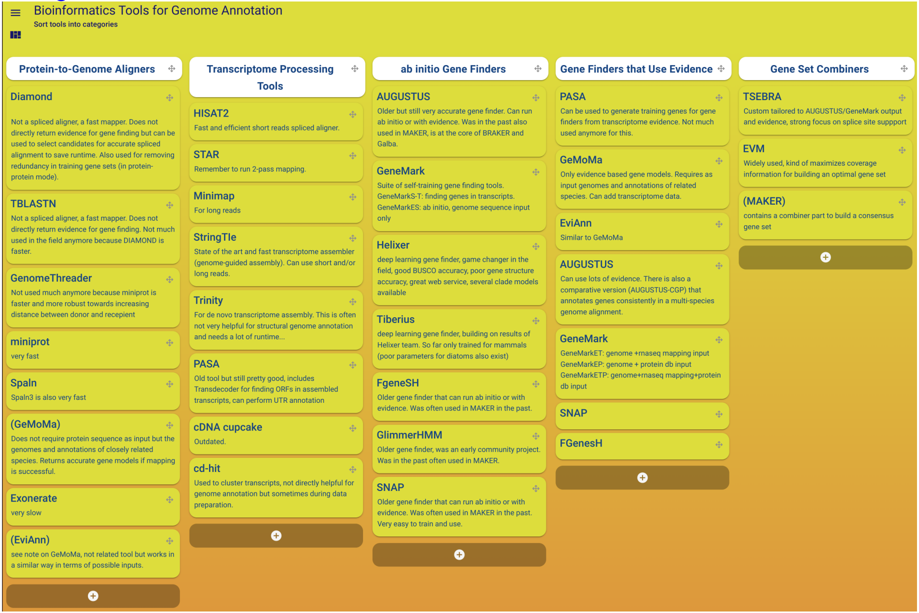
Categorize tool names

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 mapper. 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 mapper. 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 miniprot is faster and more robust towards increasing distance between donor and recepient

miniprot

very fast

Spaln

Spaln3 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 lot 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 Galba.

GeneMark

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

Helixer

deep learning gene finder, game changer in the field, good BUSCO accuracy, poor gene structure accuracy, great web service, several clade 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

GeneMarkET: genome +rnaseq mapping input
GeneMarkEP: genome + protein db input
GeneMarkETP: genome+rnaseq mapping+protein db input

SNAP

FGenesH

Gene Set Combiners

TSEBRA

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

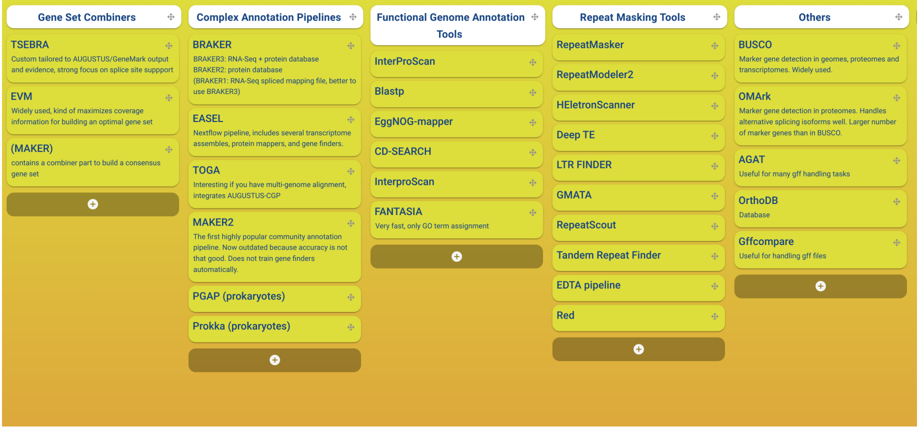
EVM

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

- **TSEBRA**
  Custom tailored to AUGUSTUS/GeneMark output and evidence, strong focus on splice site suppport
- **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

Complex Annotation Pipelines

- **BRAKER**
  BRAKER3: RNA-Seq + protein database
  BRAKER2: protein database
  (BRAKER1: RNA-Seq spliced mapping file, better to use BRAKER3)
- **EASEL**
  Nextflow pipeline, includes several transcriptome assembles, protein mappers, and gene finders.
- **TOGA**
  Interesting if you have multi-genome alignment, integrates AUGUSTUS-CGP
- **MAKER2**
  The first highly popular community annotation pipeline. Now outdated because accuracy is not that good. Does not train gene finders automatically.
- **PGAP (prokaryotes)**
- **Prokka (prokaryotes)**

Functional Genome Annotation Tools

- **InterProScan**
- **Blastp**
- **EggNOG-mapper**
- **CD-SEARCH**
- **InterproScan**
- **FANTASIA**
  Very fast, only GO term assignment

Repeat Masking Tools

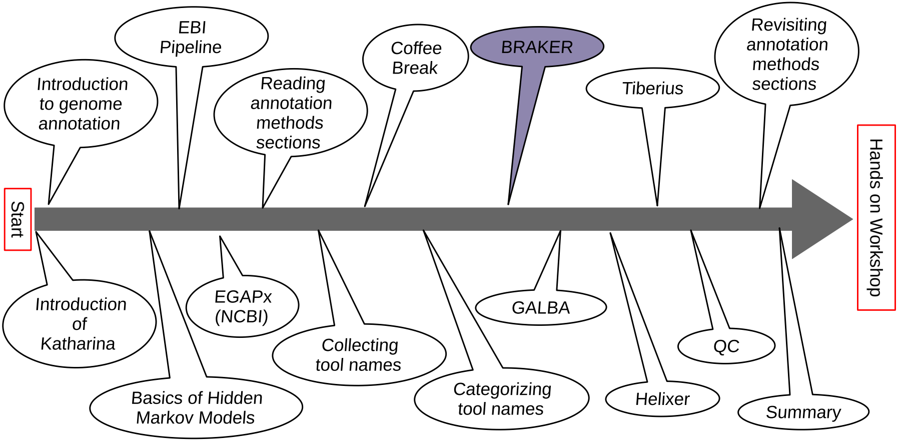
- **RepeatMasker**
- **RepeatModeler2**
- **HEletronScanner**
- **Deep TE**
- **LTR FINDER**
- **GMATA**
- **RepeatScout**
- **Tandem Repeat Finder**
- **EDTA pipeline**
- **Red**

Others

- **BUSCO**
  Marker gene detection in geomes, proteomes and transcriptomes. Widely used.
- **OMArk**
  Marker gene detection in proteomes. Handles alternative splicing isoforms well. Larger number of marker genes than in BUSCO.
- **AGAT**
  Useful for many gff handling tasks
- **OrthoDB**
  Database
- **Gffcompare**
  Useful for handling gff files

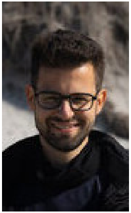
The aera of genome annotation super heroes

Image: credits to DALL-E2, modified by human


Start
Introduction to genome annotation
EBI Pipeline
Reading annotation methods sections
Coffee Break
BRAKER
Tiberius
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
Basics of Hidden Markov Models
EGAPx (NCBI)
Collecting tool names
Categorizing tool names
GALBA
Helixer
QC
Summary

The BRAKER Team

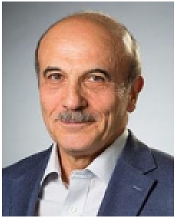
University of Greifswald & Georgia Tech University



Lars Gabriel

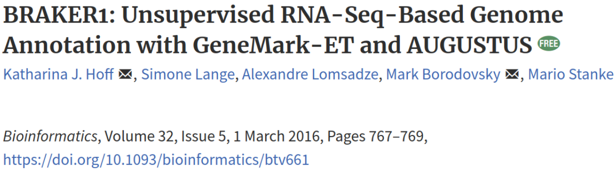
Alexandre Lomsadze, Katharina Hoff, Tomáš Brůna

Mario Stanke



Mark Borodovsky

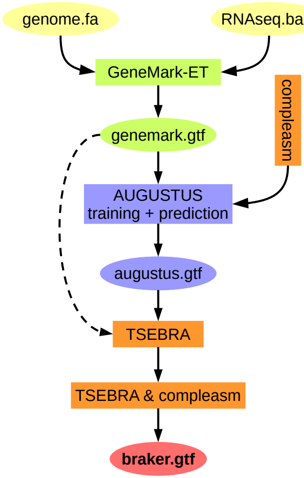
Also: Simone Lange, Matthis Ebel, Hannah Thierfeldt, Anica Hoppe, Neng Huang



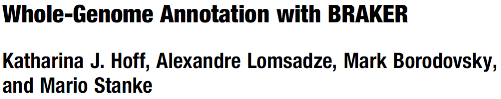
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

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

Anchors from RNA-Seq for training

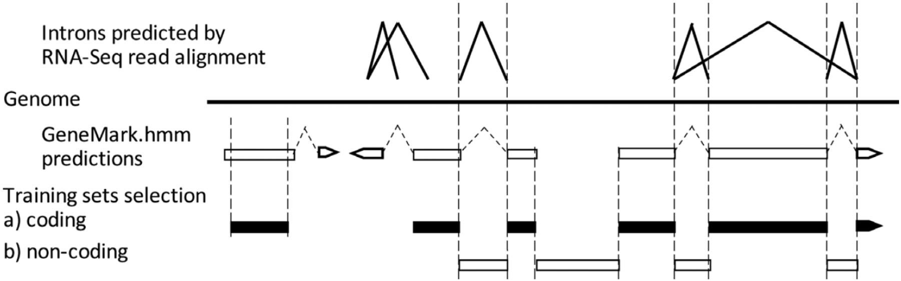
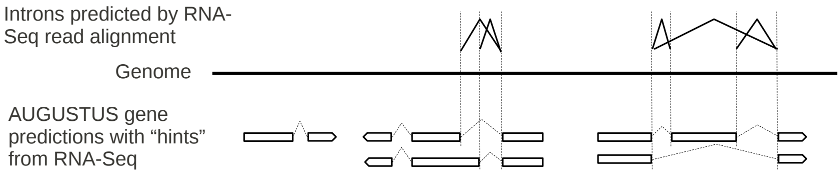


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- employs unsupervised training
- 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

AUGUSTUS uses RNA-Seq for **Prediction**



Introns predicted by RNA-Seq read alignment

Genome

AUGUSTUS gene predictions with “hints” from RNA-Seq

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

Measuring accuracy of genome annotation

Experiments

Accuracy assessment after applying tool to genome with reference annotation:

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

Accuracy metrics

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

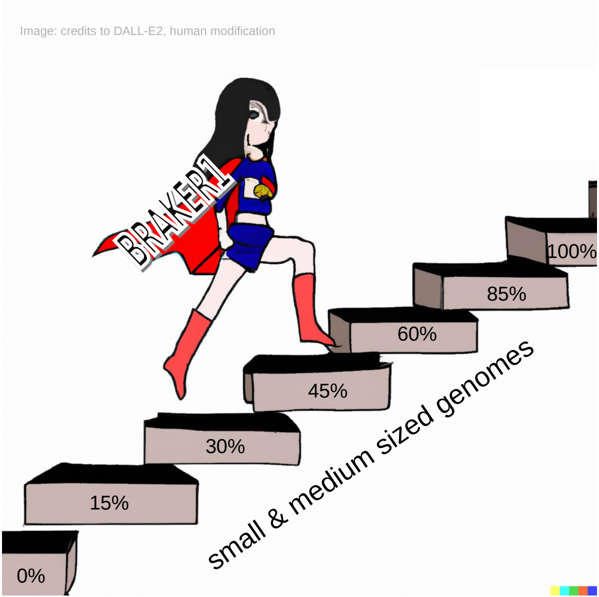
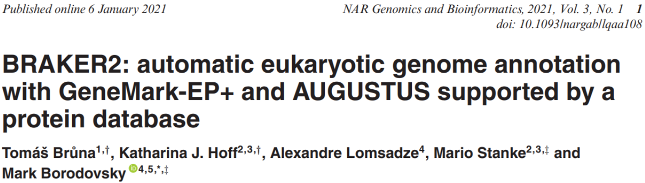


Image: credits to DALL-E2, human modification

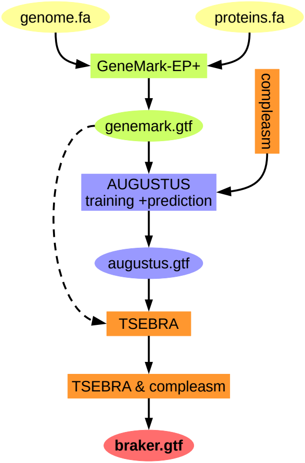
Use only if not enough RNA-Seq for BRAKER3!



Published online 6 January 2021 *NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1* **1**
doi: 10.1093/nargab/lqaa108

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

Tomáš Brůna[1,†], Katharina J. Hoff[2,3,†], Alexandre Lomsadze[4], 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)
- 1,269 citations (Google Scholar)

BRAKER2 gene F1 accuracy

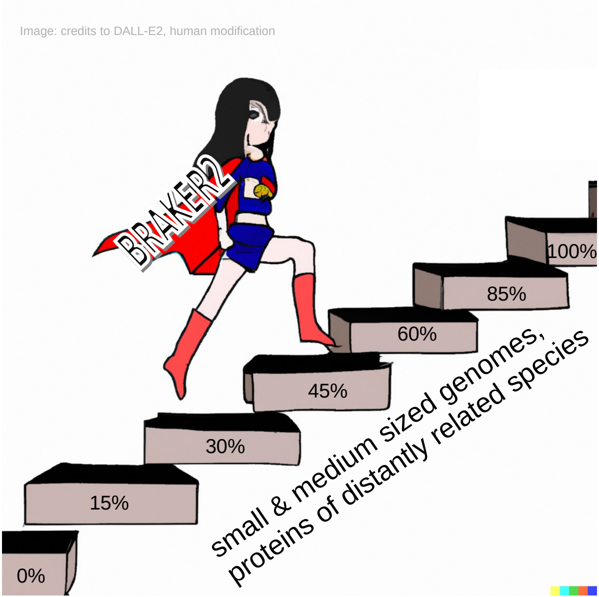


Image: credits to DALL-E2, human modification
BRAKER2
100%
85%
60%
45%
30%
15%
0%
small & medium sized genomes,
proteins of distantly related species

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

BRAKER3 gene F1 accuracy - climbing the top

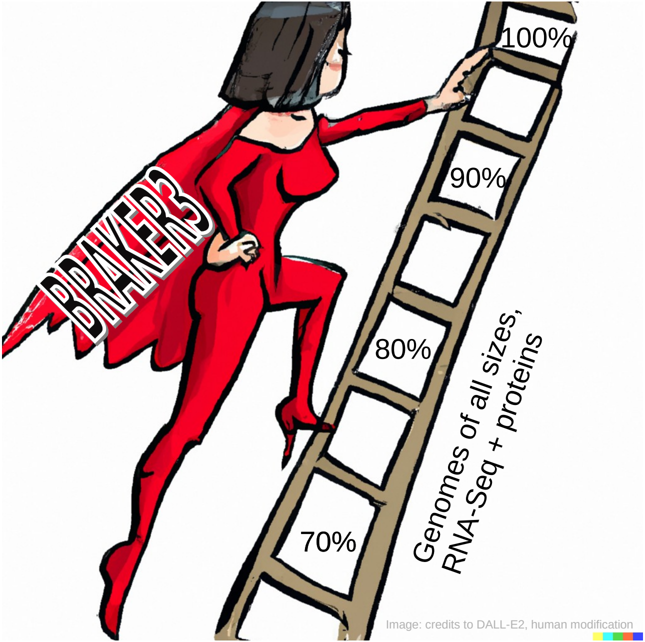
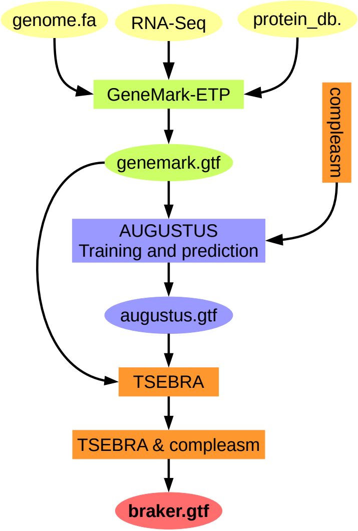


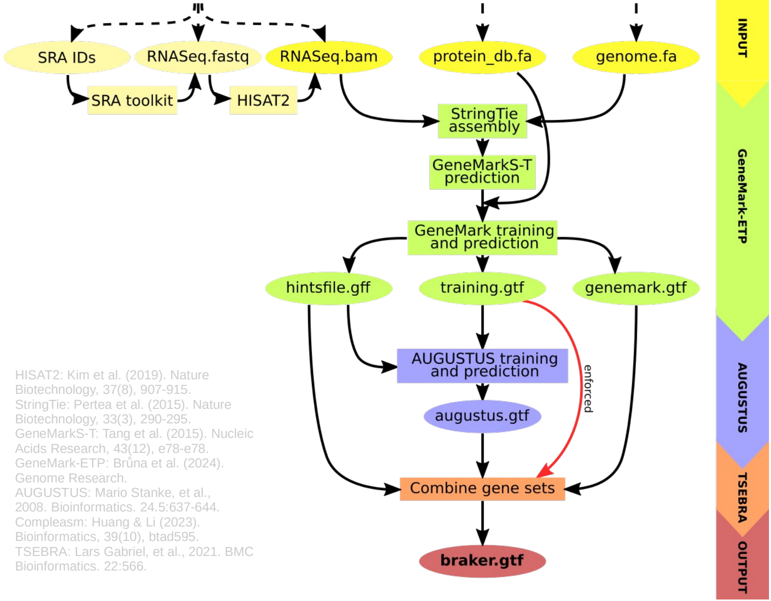
Image: credits to DALL-E2, human modification

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



- 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



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: Brůna 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.

Gabriel *et al. BMC Bioinformatics* (2021) 22:566
https://doi.org/10.1186/s12859-021-04482-0

BMC Bioinformatics

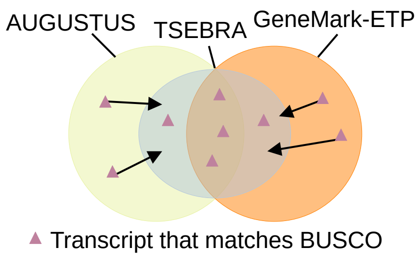
SOFTWARE Open Access

TSEBRA: transcript selector for BRAKER

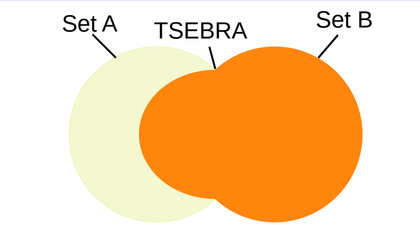
Lars Gabriel[1,2], Katharina J. Hoff[1,2], Tomáš Brůna[3], Mark Borodovsky[4,5] and Mario Stanke[1,2*]

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

TSEBRA in BRAKER



Manually enforcing a gene set



enforce Set B with `-k setb.gtf`

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

Accuracy of genome annotation approaches by BRAKER team

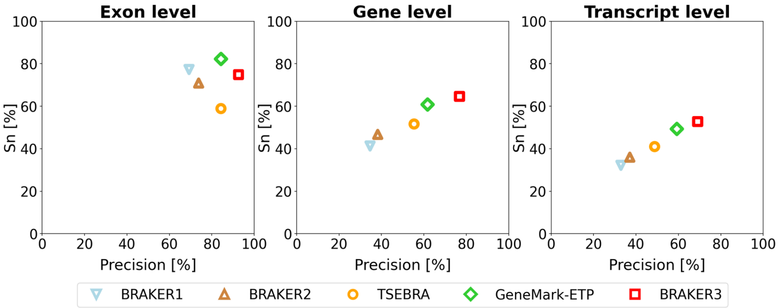


Figure 2. Average precision and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in Supplemental Table S1). Inputs were the genomic sequences, short-read RNA-seq libraries, and protein databases (*order excluded*).

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

Availability

GitHub

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

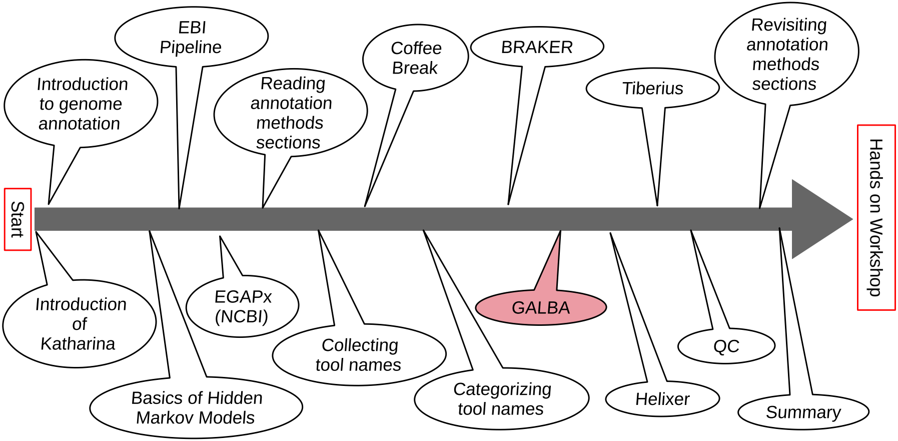
Docker/Singularity

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

singularity exec braker.sif braker.pl [OPTIONS]
```

Licenses

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

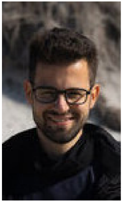

Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
EGAPx (NCBI)
Reading annotation methods sections
Collecting tool names
Coffee Break
Categorizing tool names
BRAKER
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

GALBA Contributors

Tomáš Brůna

Heng Li

Joseph Guhlin

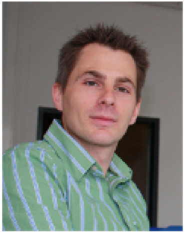

Lars Gabriel

Natalia Nenasheva

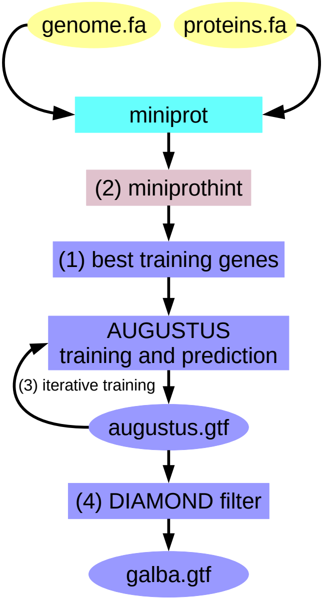
Ethan Tolman

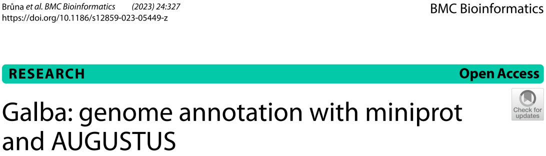
Paul Frandsen

Matthis Ebel


Mario Stanke

Katharina Hoff




Brůna *et al. BMC Bioinformatics* (2023) 24:327
https://doi.org/10.1186/s12859-023-05449-z

BMC Bioinformatics

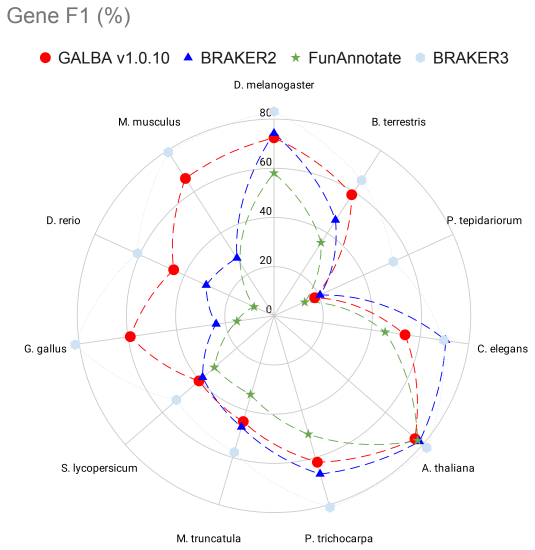
RESEARCH Open Access

Galba: genome annotation with miniprot and AUGUSTUS

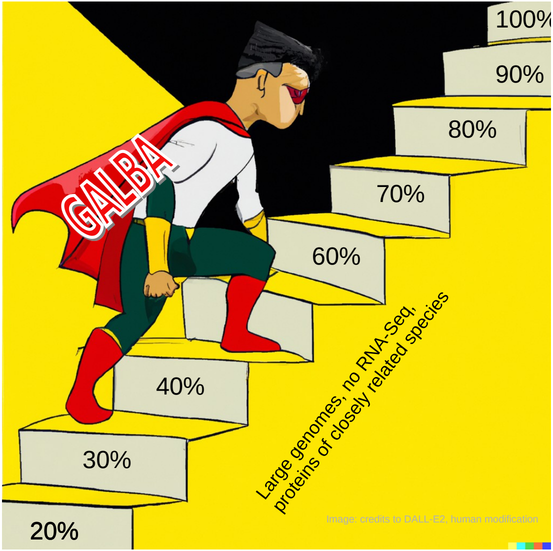
Check for updates

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

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



GALBA: Gene F1 Accuracy



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

Image: credits to DALL-E2, human modification

If you have RNA-Seq, use BRAKER3!

Availability

GitHub

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

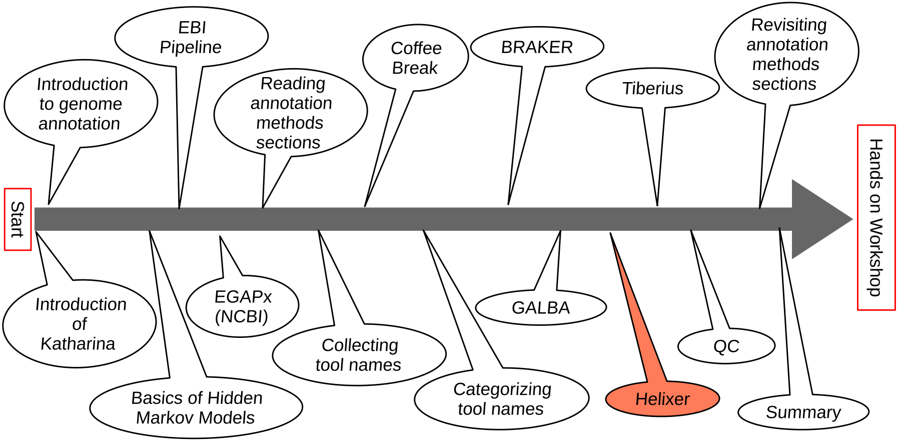
Docker/Singularity

```
singularity build galba.sif \
        docker://katharinahoff/galba:latest

singularity exec galba.sif galba.pl [OPTIONS]
```

Licenses

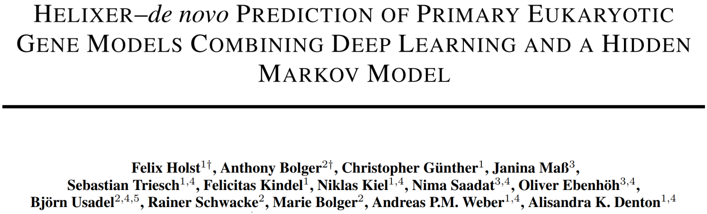
- GALBA: Artistic License
- all dependencies have Open Source Licenses



Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
Categorizing tool names
BRAKER
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

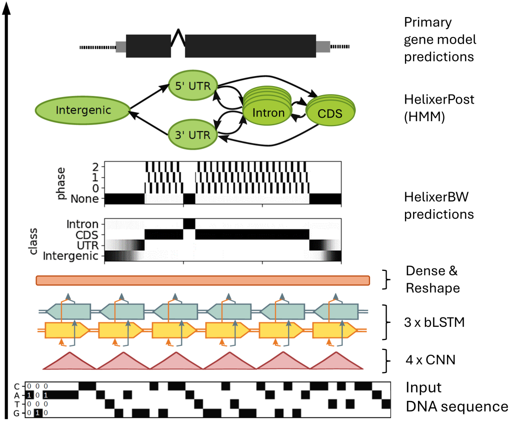
Helixer: bringing deep learning into genome annotation

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 Günther[1], Janina Maß[3],
Sebastian Triesch[1,4], Felicitas Kindel[1], Niklas Kiel[1,4], Nima Saadat[3,4], Oliver Ebenhöh[3,4],
Björn Usadel[2,4,5], Rainer Schwacke[2], Marie Bolger[2], Andreas P.M. Weber[1,4], Alisandra K. Denton[1,4]



Primary gene model predictions

HelixerPost (HMM)

HelixerBW predictions

Dense & Reshape

3 x bLSTM

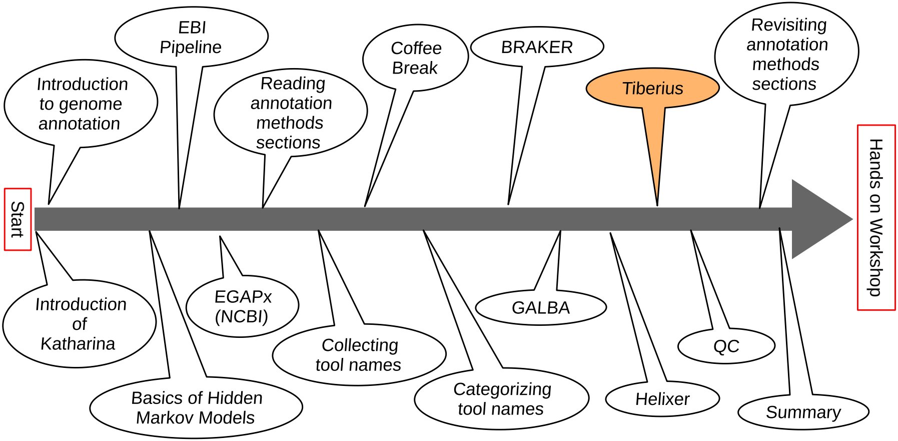
4 x CNN

Input DNA sequence

- 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`



Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
BRAKER
Categorizing tool names
GALBA
Tiberius
Helixer
Revisiting annotation methods sections
QC
Summary
Hands on Workshop

Tiberius: improved genome annotation with deep learning

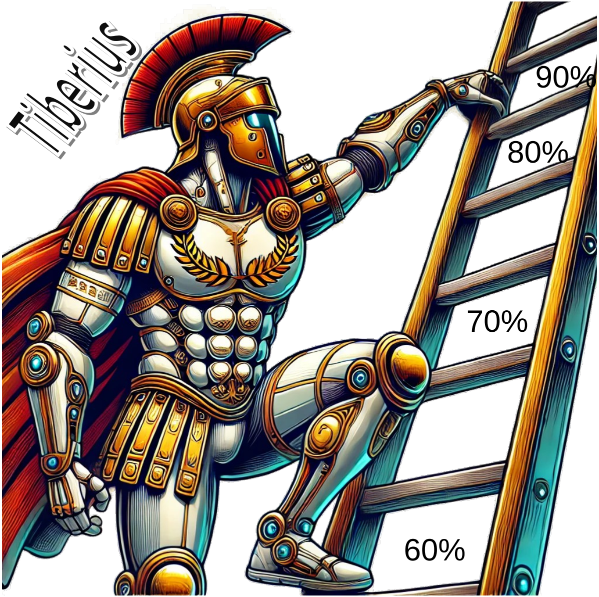


Image: ChatGPT by OpenAI, manual editing

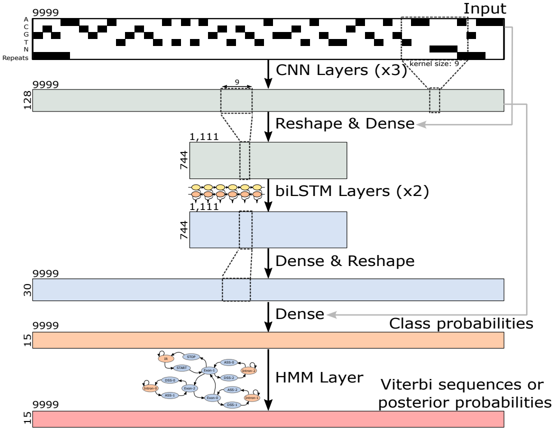
The Tiberius Team

University of Greifswald

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,*]



Input

CNN Layers (x3)

Reshape & Dense

biLSTM Layers (x2)

Dense & Reshape

Dense

Class probabilities

HMM Layer

Viterbi sequences or posterior probabilities

- 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

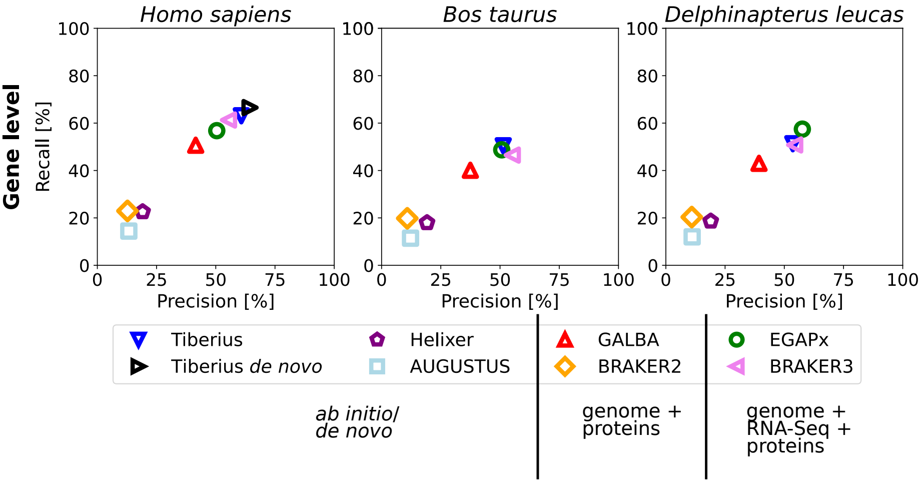
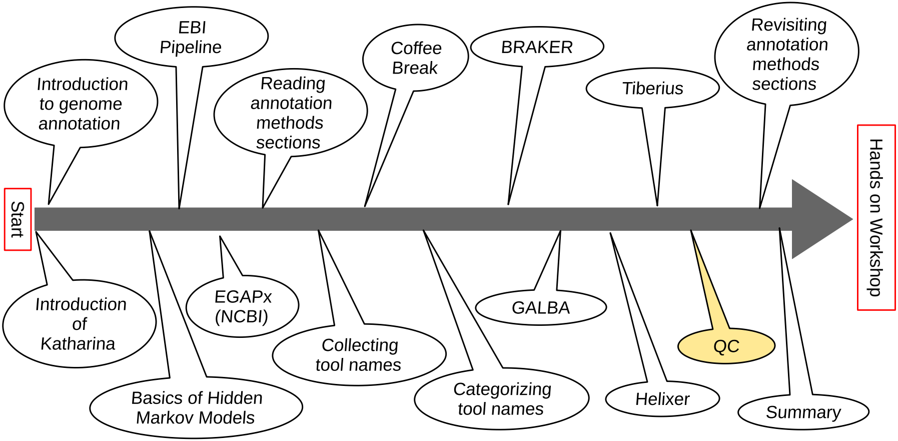


Image: Lars Gabriel, PAG presentation 2025

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



Start

Introduction of Katharina

Introduction to genome annotation

Basics of Hidden Markov Models

EBI Pipeline

EGAPx (NCBI)

Reading annotation methods sections

Collecting tool names

Coffee Break

Categorizing tool names

BRAKER

GALBA

Helixer

Tiberius

QC

Revisiting annotation methods sections

Summary

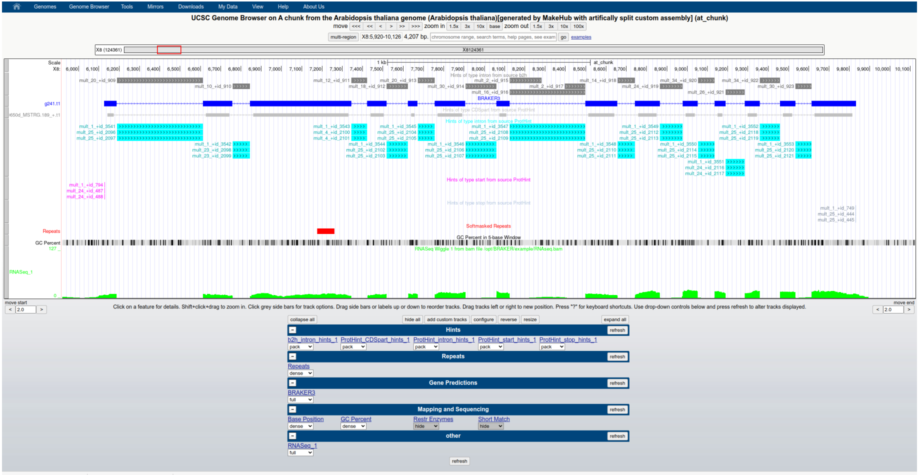
Hands on Workshop

Did We Do a Good Job?

Genome Browsers

Visualize your Annotation in Context with Evidence

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



Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us

UCSC Genome Browser on A chunk from the Arabidopsis thaliana genome (Arabidopsis thaliana)[generated by MakeHub with artifically split custom assembly] (at_chunk)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
go
examples

X8:5,920-10,126 4,207 bp. chromosome range, search terms, help pages, see exam

X8 (124361)
X8124361
Scale
X8:
1 kb
at_chunk
Hints of type intron from source b2h
BRAKER3
g241.t1
Hints of type CDSpart from source ProtHint
Hints of type intron from source ProtHint
Hints of type start from source ProtHint
Hints of type stop from source ProtHint
Repeats
Softmasked Repeats
GC Percent
GC Percent in 5-base Window
127
RNASeq Wiggle 1 from bam file /opt/BRAKER/example/RNAseq.bam
RNASeq_1
0
move start
move end

Click on a feature for details. Shift+click+drag to zoom in. Click grey side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. Use drop-down controls below and press refresh to alter tracks displayed.

Hints

b2h_intron_hints_1: pack
ProtHint_CDSpart_hints_1: pack
ProtHint_intron_hints_1: pack
ProtHint_start_hints_1: pack
ProtHint_stop_hints_1: pack

Repeats

Repeats: dense

Gene Predictions

BRAKER3: full

Mapping and Sequencing

Base Position: dense
GC Percent: dense
Restr Enzymes: hide
Short Match: hide

other

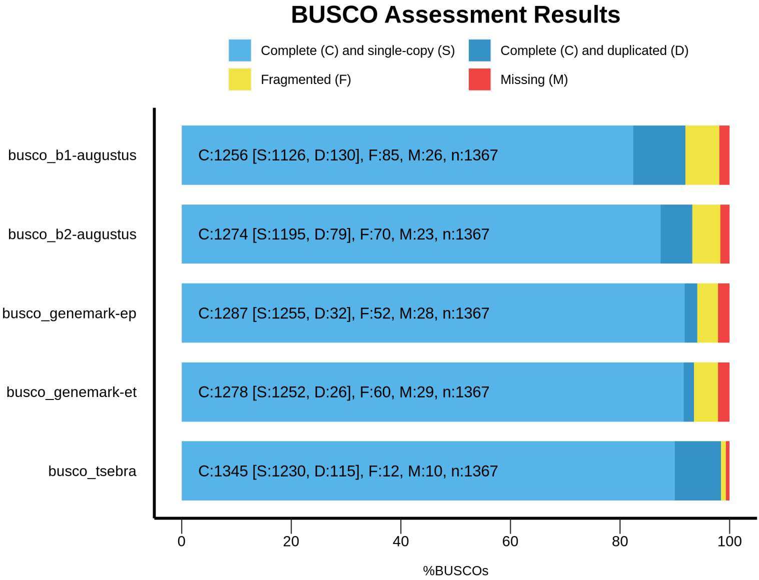
RNASeq_1: full

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
- ...

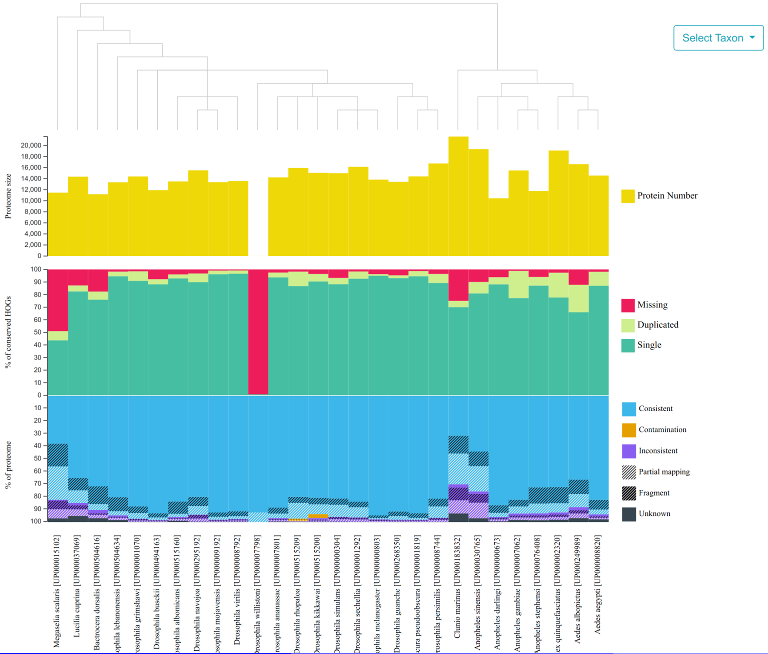
If possible, compare to annotated close relatives.
Consider effect of individual annotation pipelines.

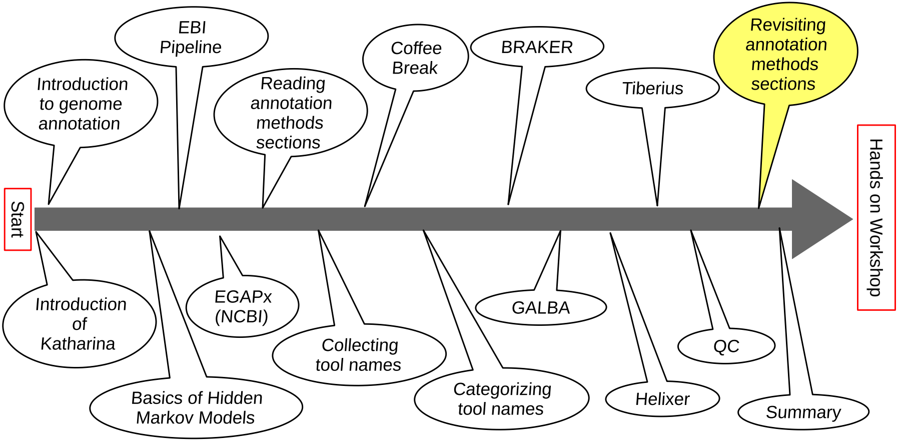
BUSCO: Sensitivity in Clade-Specific Conserved Genes



Beware! BUSCO completeness does not warrant correct gene structures!

OMArk: Sensitivity, Contaminations, & More



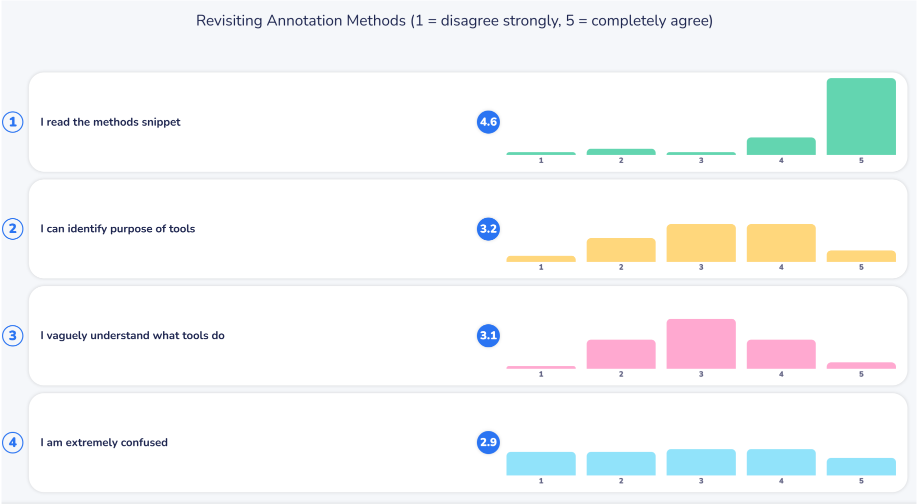

Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
BRAKER
Categorizing tool names
GALBA
Tiberius
Helixer
QC
Revisiting annotation methods sections
Summary
Hands on Workshop

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

Revisiting annotation methods sections



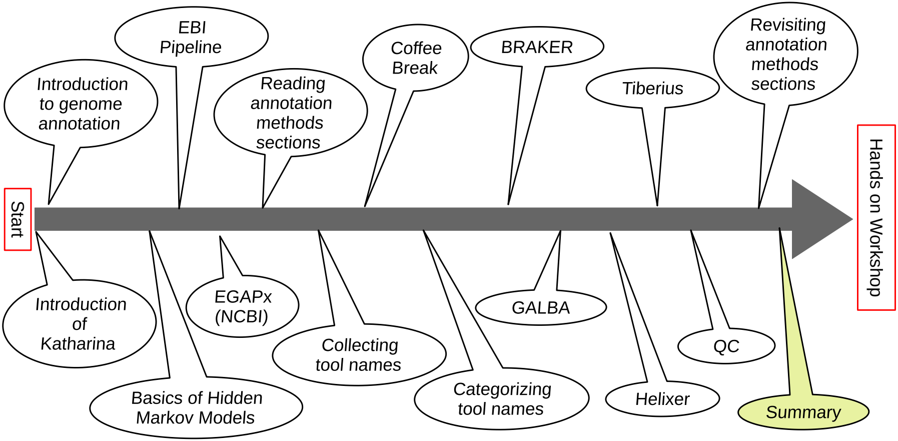
Revisiting Annotation Methods (1 = disagree strongly, 5 = completely agree)

1 I read the methods snippet 4.6

2 I can identify purpose of tools 3.2

3 I vaguely understand what tools do 3.1

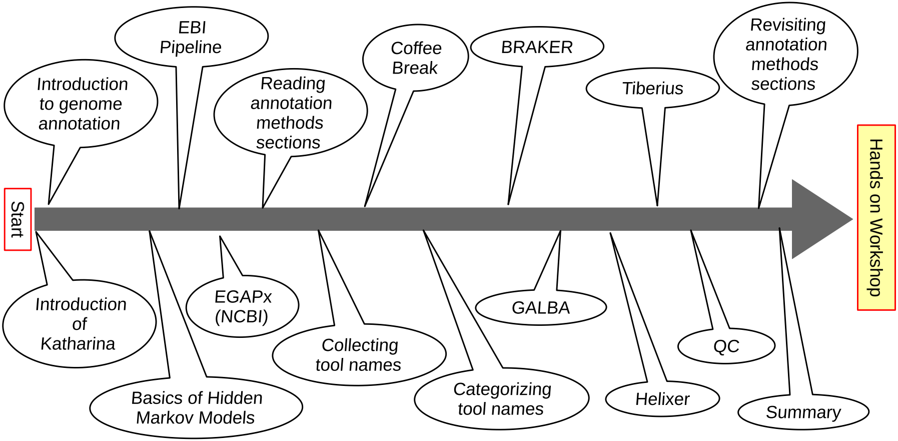
4 I am extremely confused 2.9



Start
Introduction to genome annotation
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Reading annotation methods sections
EGAPx (NCBI)
Coffee Break
Collecting tool names
BRAKER
Categorizing tool names
GALBA
Tiberius
Helixer
Revisiting annotation methods sections
QC
Summary
Hands on Workshop

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



Start

Introduction of Katharina

Introduction to genome annotation

Basics of Hidden Markov Models

EBI Pipeline

EGAPx (NCBI)

Reading annotation methods sections

Collecting tool names

Coffee Break

Categorizing tool names

BRAKER

GALBA

Helixer

Tiberius

QC

Revisiting annotation methods sections

Summary

Hands on Workshop