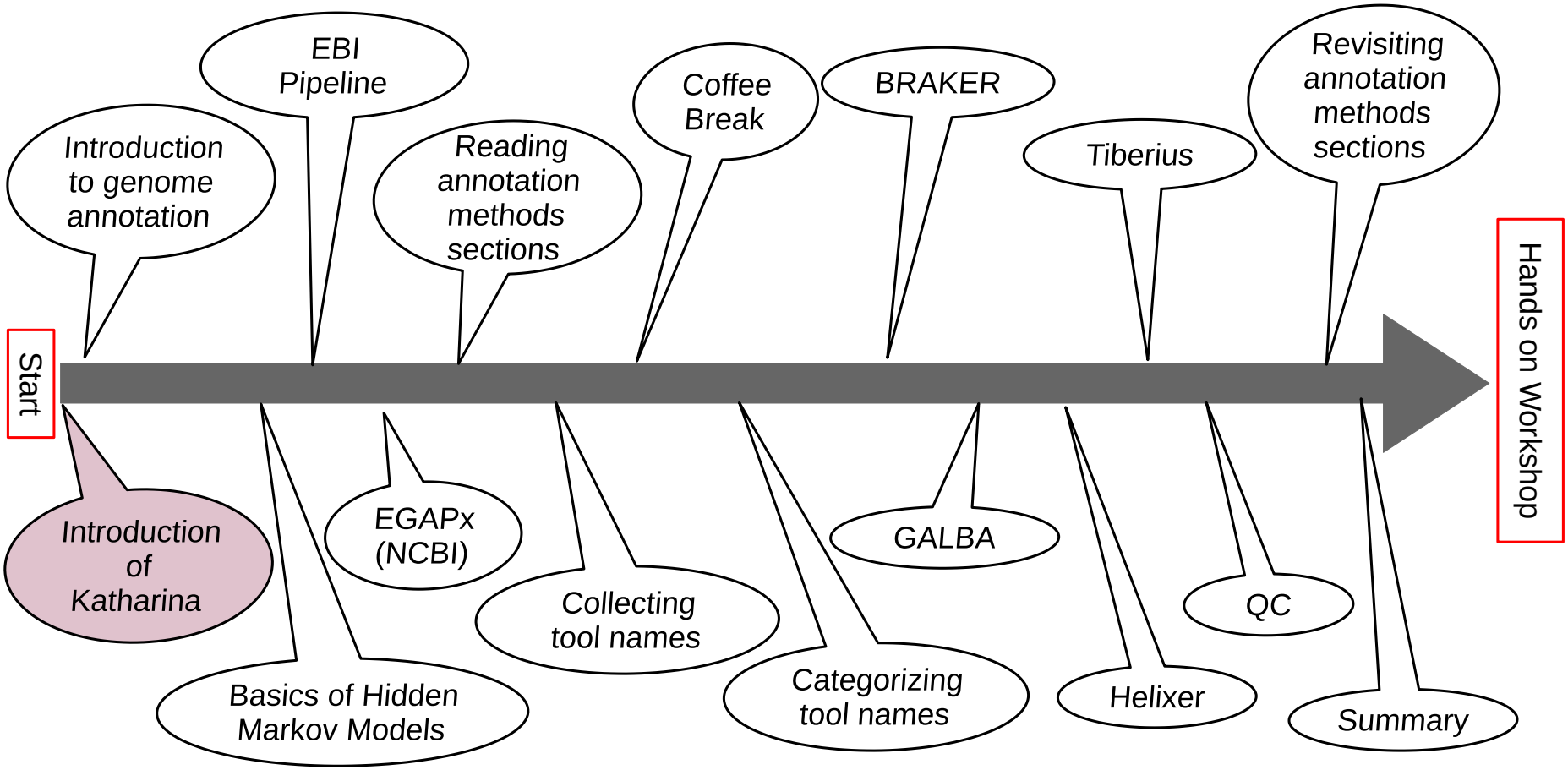
Annotation of Protein Coding Genes

May 6th 2025

Katharina J. Hoff

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

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


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

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
- 38 peer-reviewed research articles with currently 7,902 citations

Teaching

- currently 1 postdoc, 4 (+4) PhD students, 1 MSc student, 3 BSc students
- applied bioinformatics, programming, statistics, & data science

... I love to sail, have a dog, a cat, and an 8-years old daughter...

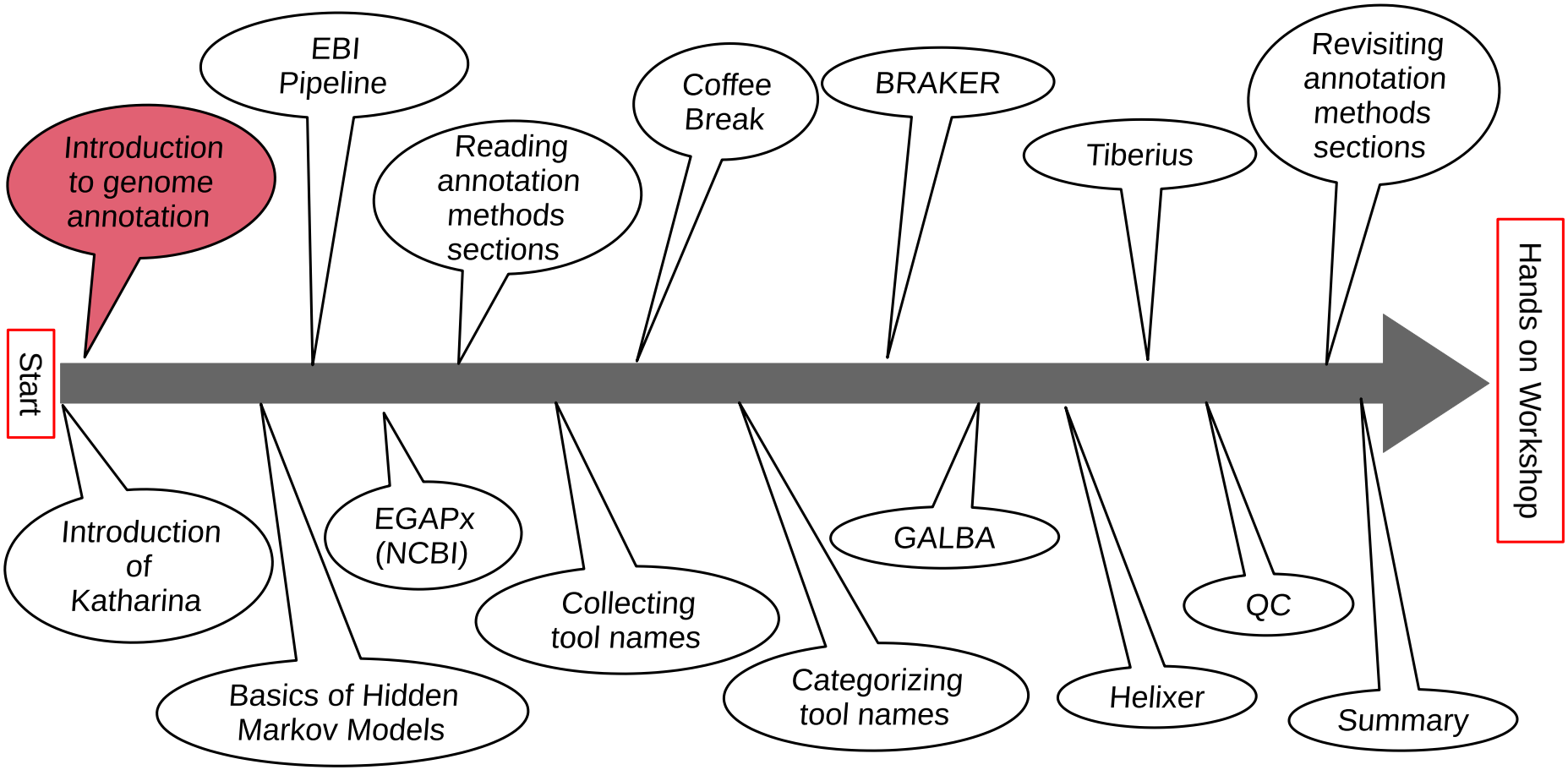
After this lecture, you will...

- 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
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

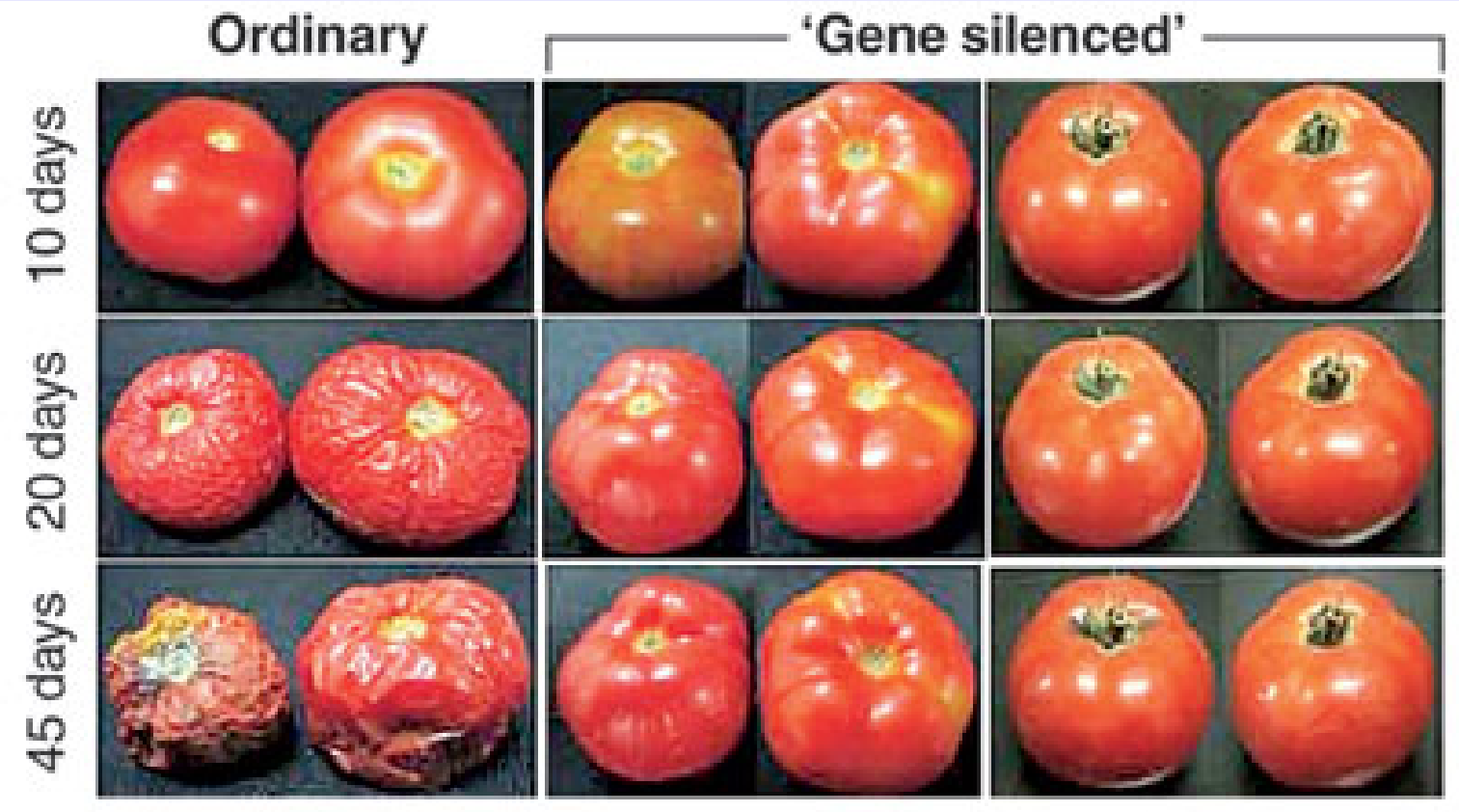
Where are the protein coding genes?

Genomic sequence: chicken

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtccttttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtcagactg
ttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctctctgtgccttggtttcccc
atctgtcatatgaagggagtgcgatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttggggttctggtagctccacagggatgctcaatgaagatgcaaaattagaagtcaaaat
aaacagctcccatgggcagtgttgatctcaccctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaacccctggctgccctgggaggcaagggtgaggg
ctggatttttaaaggggcctgttttggggaggggtgatgagcgctggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggttcctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaaccatgctagaaaaacatgtggaaaagtcactacca
ggctggcagggaatggggcaatctattcatactgattgcaatgcccactggttcctaatctgggcaacc
cctggggcccacagctaaatccagtgagtggaagttacagggagtctgcttccagtgctgctcgaggaa
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

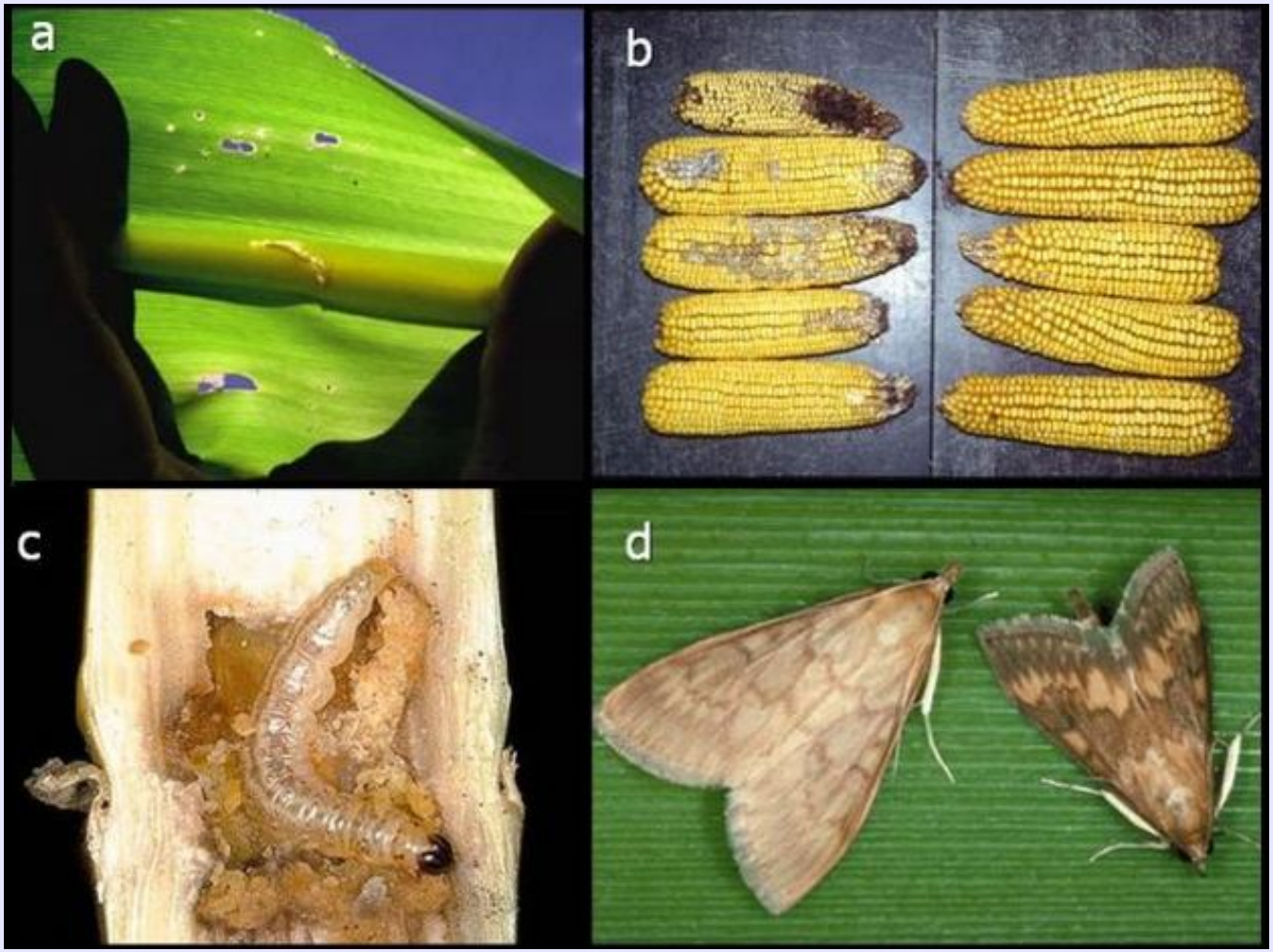


Image: Hellmich & Hellmich (2012) Nature Education Knowledge 3(10):4

`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**

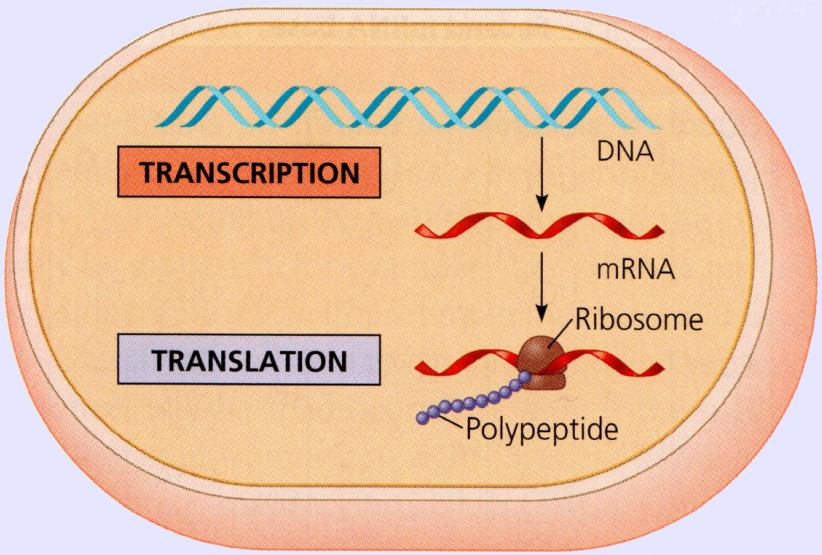
It does not take a village to publish a genome!

- 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!**

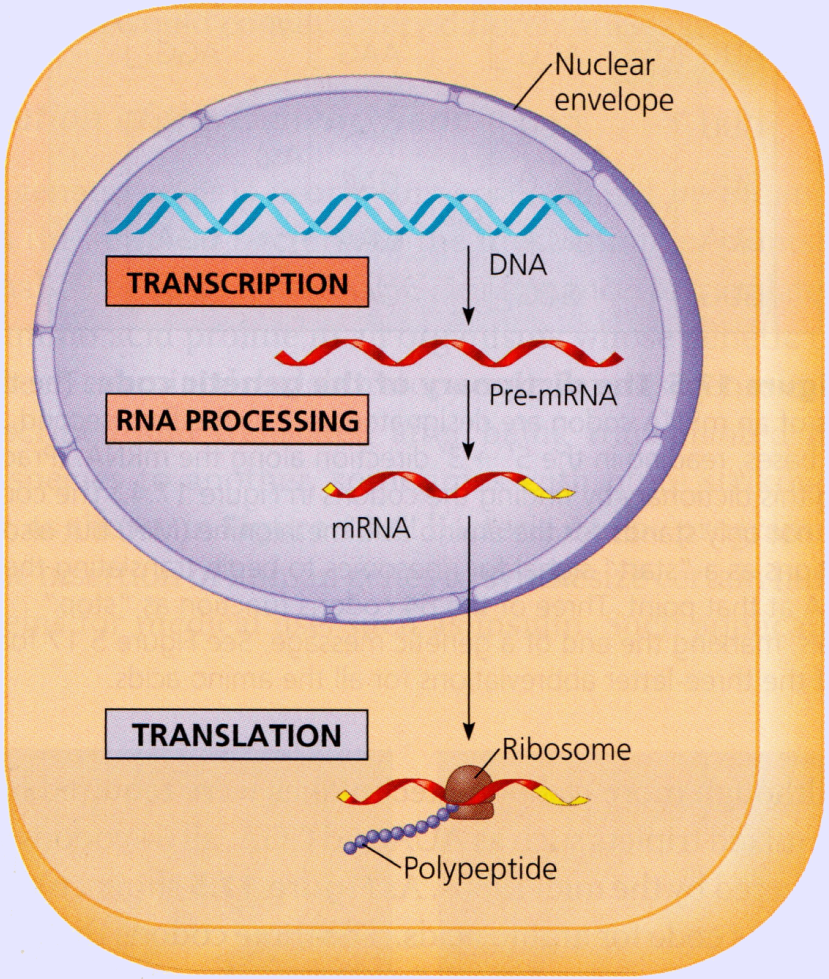
How does a cell recognize protein-coding genes?

Transcription & Translation

Prokaryotes



Eukaryotes



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

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

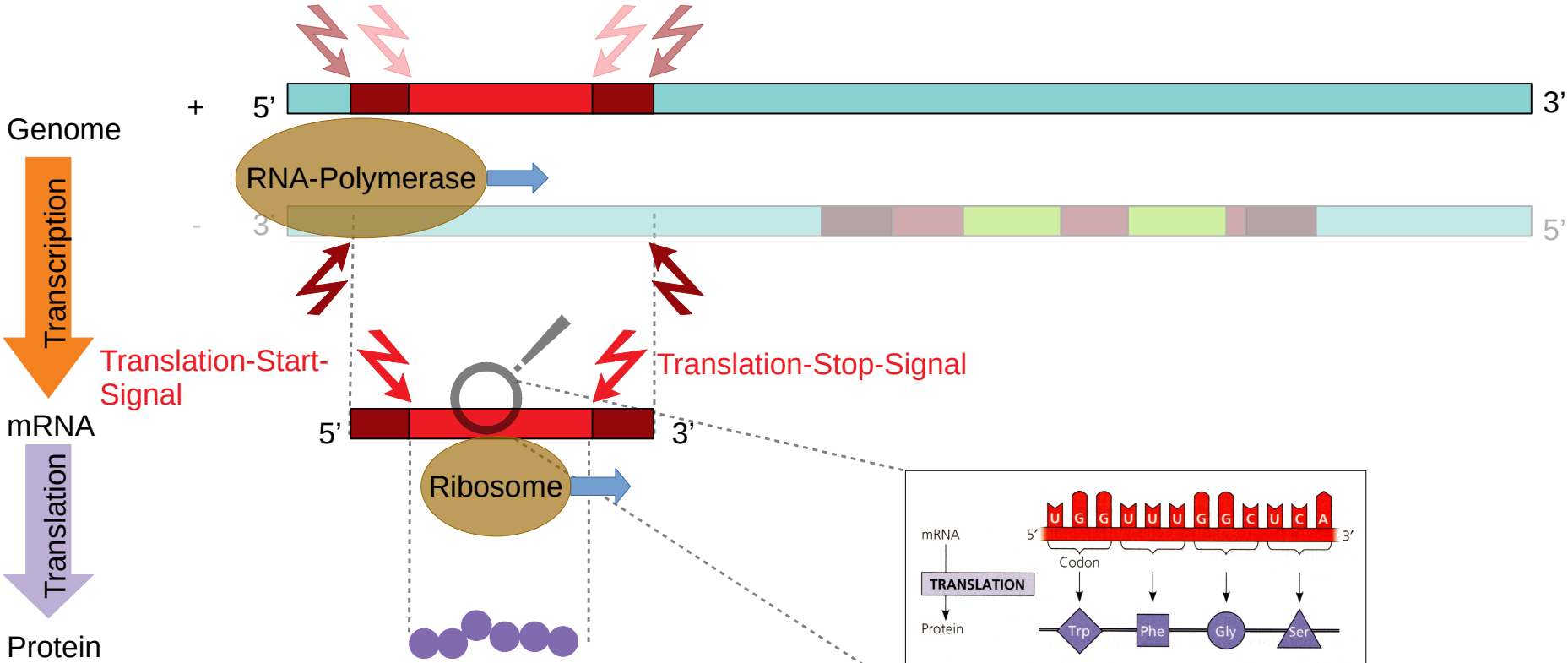


Image: Campbell et al., Biology, San Francisco, 2008, p. 329, FIg. 17.4

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

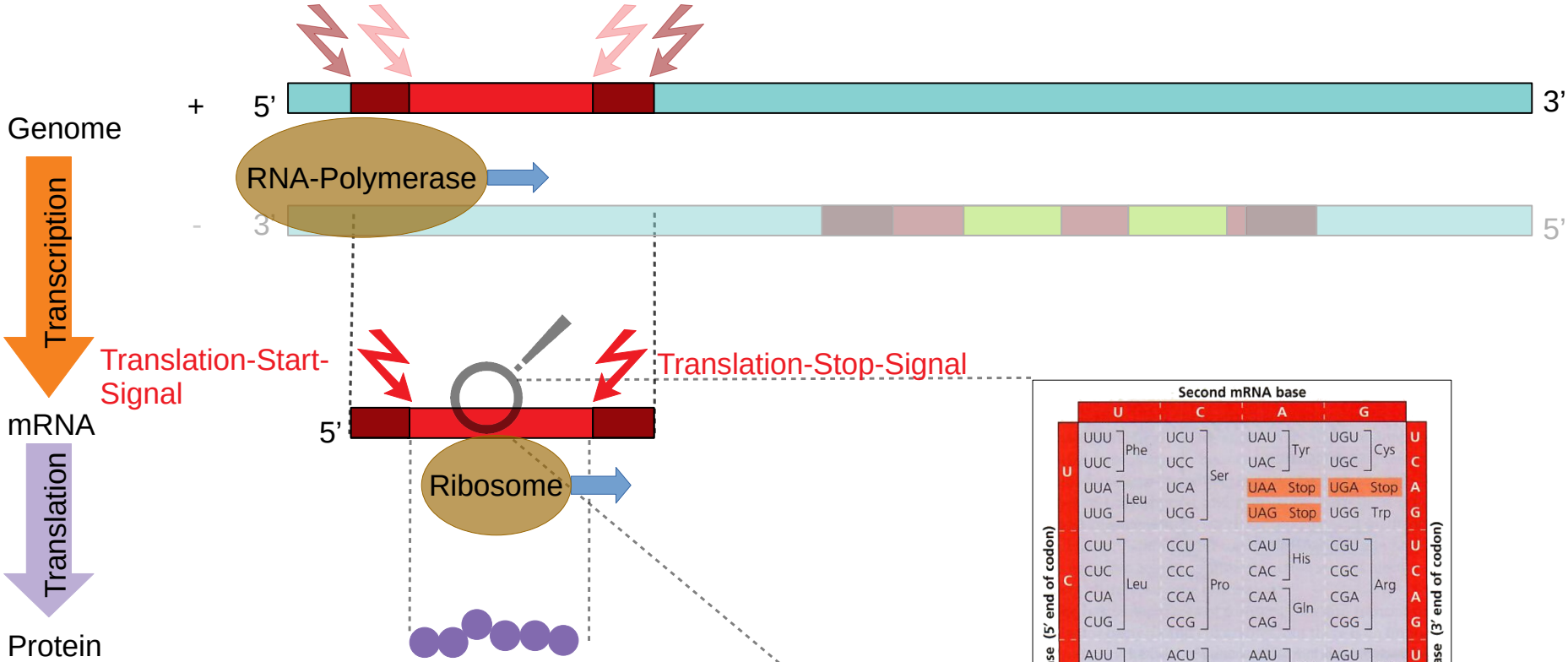
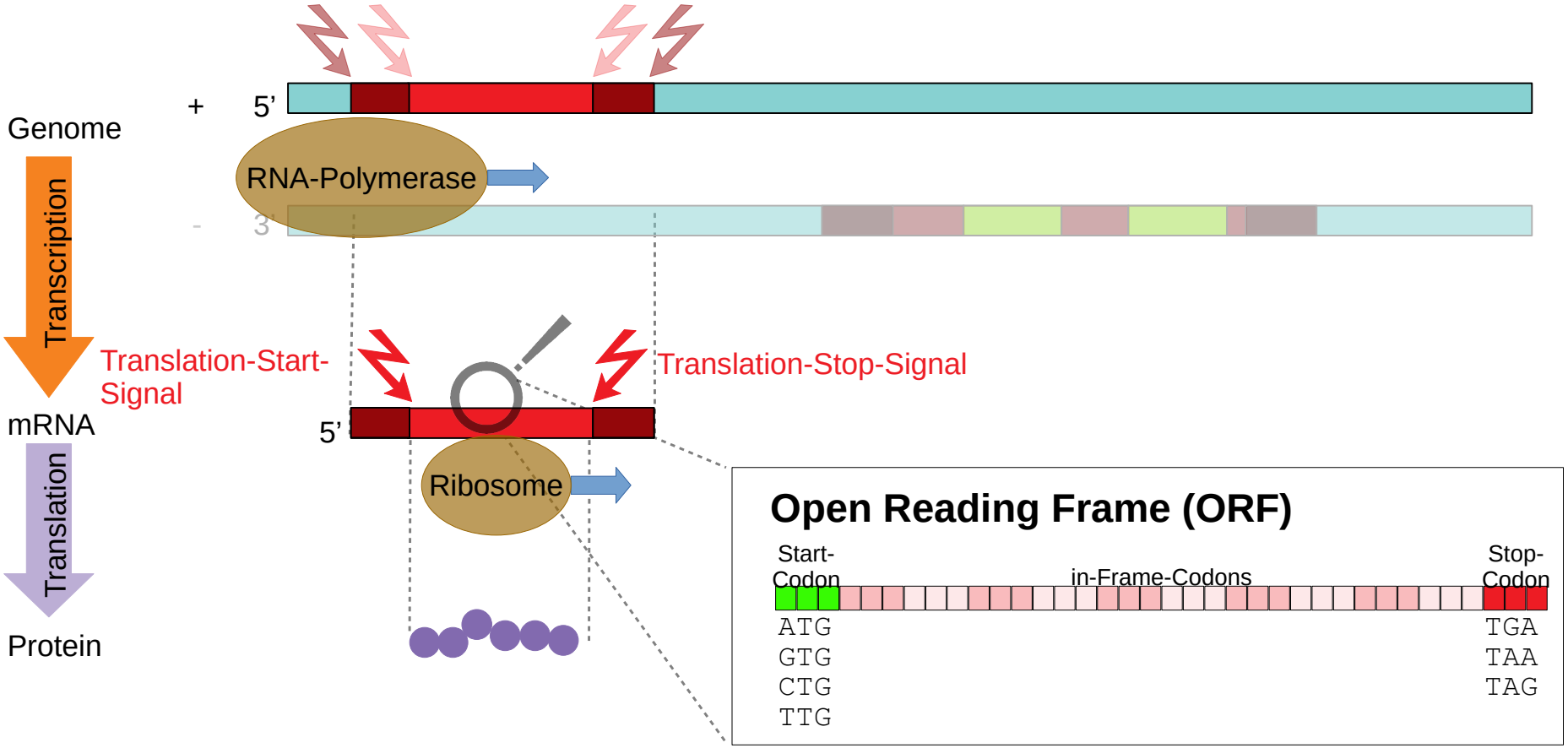


Image: Campbell et al., Biology, San Francisco, 2008, p. 339, Fig. 17.5

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

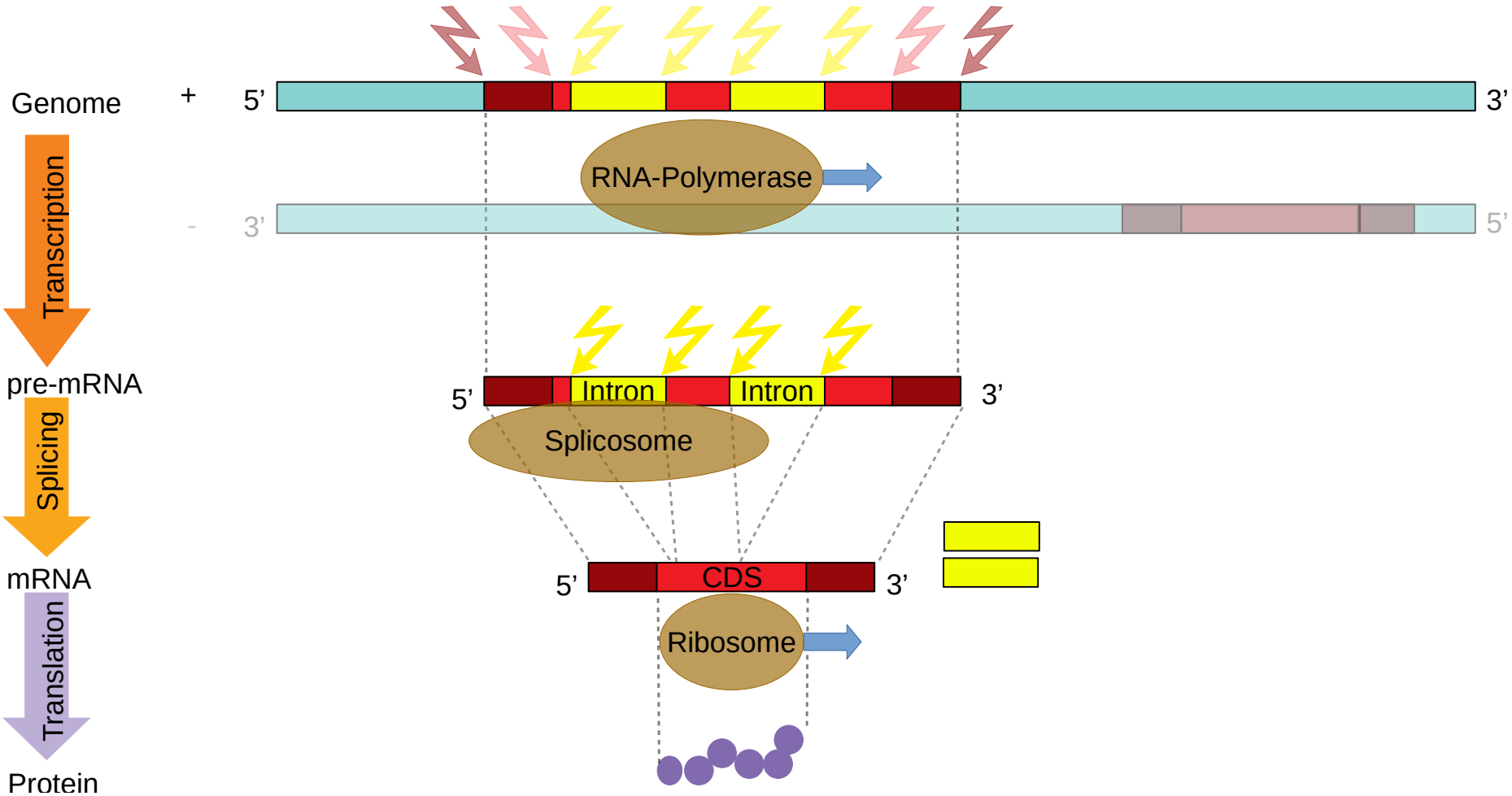
Prokaryotes & Eukaryotes*



- 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

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtccttttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtcagactg
ttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctctctgtgccttggtttcccc
atctgtcatatgaagggagtgcgatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttggggttctggtagctccacagggatgctcaatgaagatgcaaaattagaagtcaaaat
aaacagctcccatgggcagtgttgatctcaccctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaacccctggctgccctgggaggcaagggtgaggg
ctggatttttaaaggggcctgttttggggaggggtgatgagcgctggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggttcctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaaccatgctagaaaaacatgtggaaaagtcactacca
ggctggcagggaatggggcaatctattcatactgattgcaatgcccactggttcctaatctgggcaacc
cctggggcccacagctaaatccagtgagtggaagttacagggagtctgcttccagtgctgctcgaggaa
ggatcccatccaccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaa
gggataaagccagatgacctcaaaggtcccatgggattctaatctgtctgctccttgttctacagattc
caaaccaaaagaggcaagcaagtctgcgctgaccccagtgagtcctgggtccaggagtacgtgtatgac

ctggaactgaactgagctgctcagagacaggaagtcttc

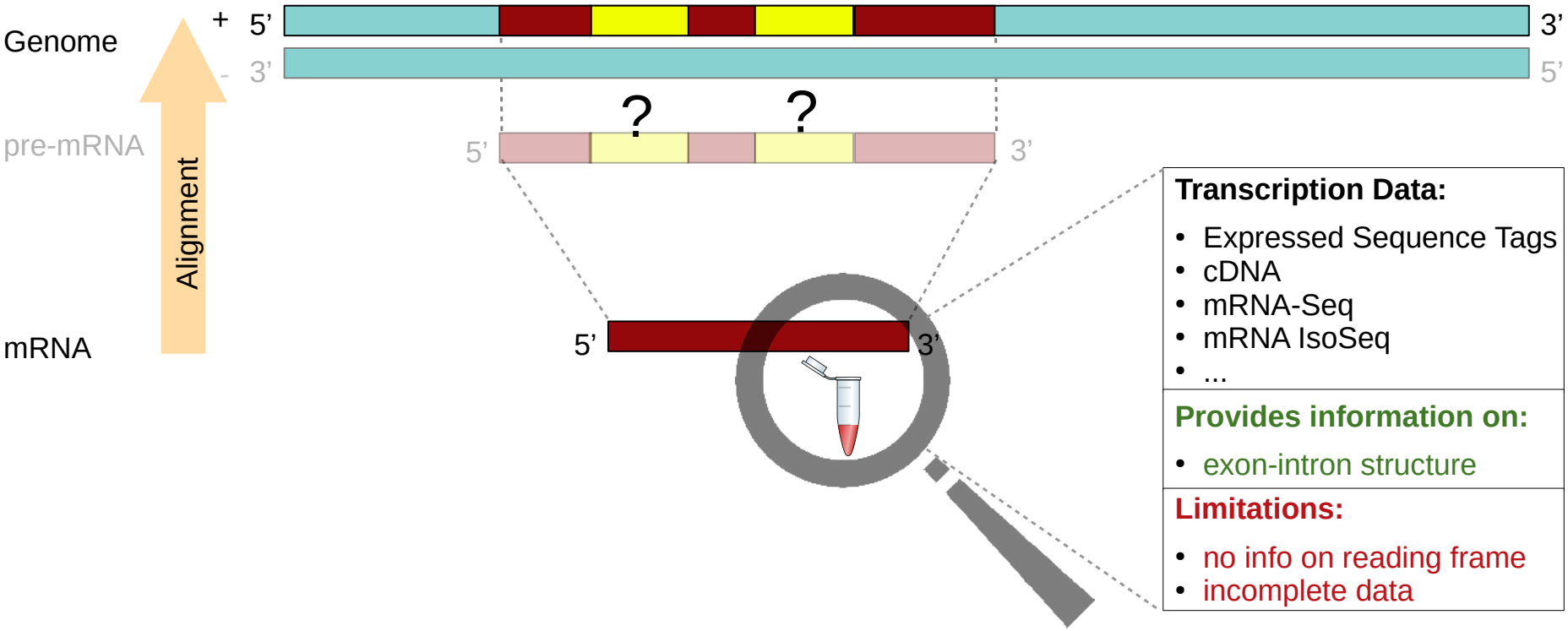
The Genome Annotation Problem

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

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctcctc
gtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagctgctatt
tcgagtcaaggtgtaggcagagtccttttttctagtcatggctggcaaacagtgggatctggggatggg
acaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagtagatattcagtaa
cattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatgggctggaagtcagactg
ttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctctctgtgccttggtttcccc
atctgtcatatgaagggagtgcgatgtgttctgagactgaatccagttccaatcttctagatttctttc
tcgttcttctctgaagatccactattcagaataagactcctgctcatgttaggtgggaatggatacaag
ggaccatatttggggttctggtagctccacagggatgctcaatgaagatgcaaaattagaagtcaaaat
aaacagctcccatgggcagtgttgatctcaccctggcctttcctttcagtgggctcagaccctcccacc
gcctgctgcttttcttacaccgcgaggaagcttcctcgcaactttgtggtagattactatgagaccagc
agcctctgctcccagccagctgtggtgtgagtatcaaccctggctgccctgggaggcaagggtgaggg
ctggatttttaaaggggcctgttttggggaggggtgatgagcgctggggaggcagctctcagggctg
aagccttccctgacagcagtgaggtcacaggtcatgaactcacttttcaagtgctgaaggcggctgagt
ggcagccgagacagaaggggttcctggggaggaagttattcagaggacagggaagcaggggaaggcag
acaggtcccatgagatatggaccaattccttaaaccatgctagaaaaacatgtggaaaagtcactacca
ggctggcagggaatggggcaatctattcatactgattgcaatgcccactggttcctaatctgggcaacc
cctggggcccacagctaaatccagtgagtggaagttacagggagtctgcttccagtgctgctcgaggaa
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:

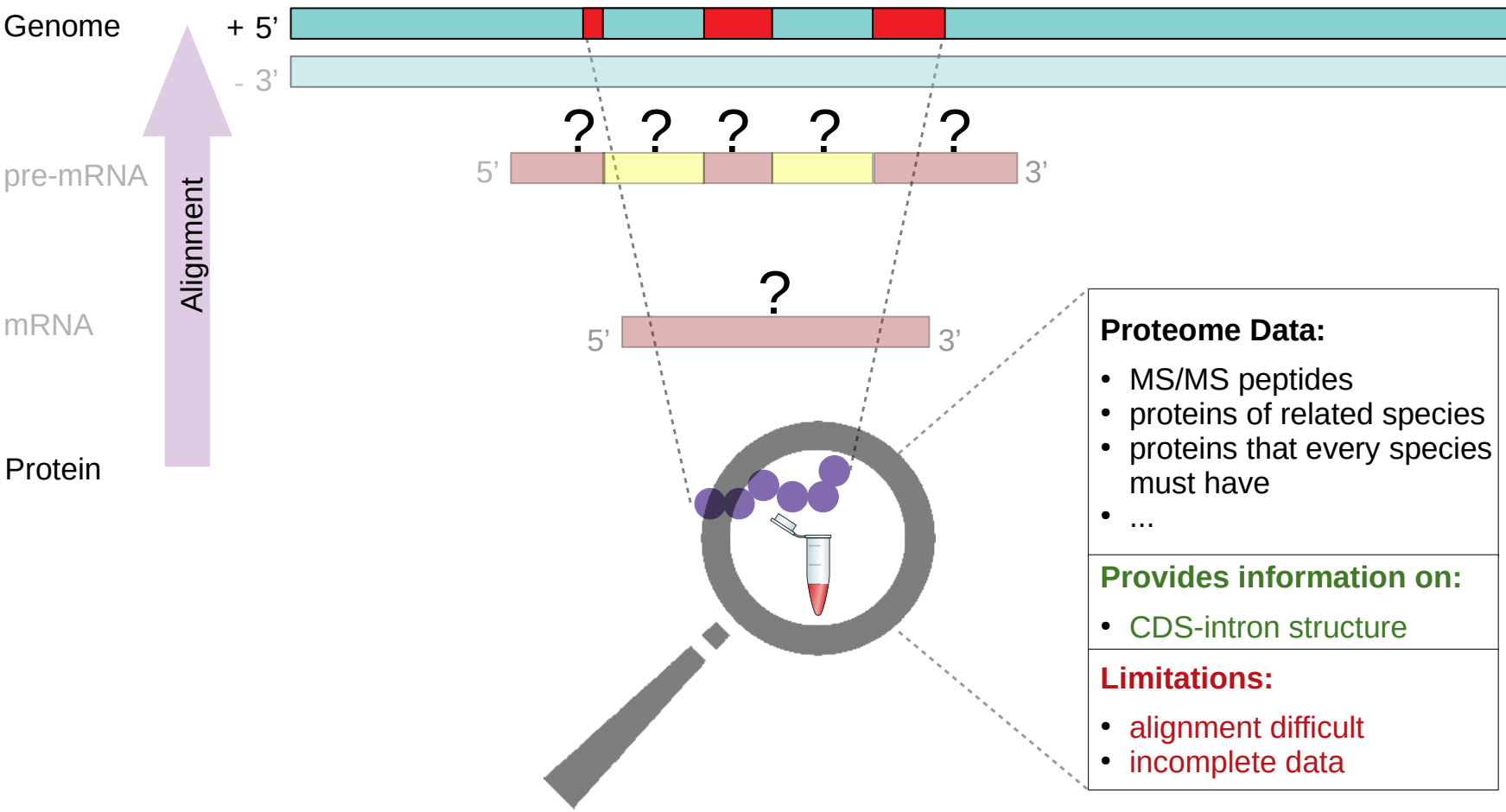
- exon-intron structure

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:

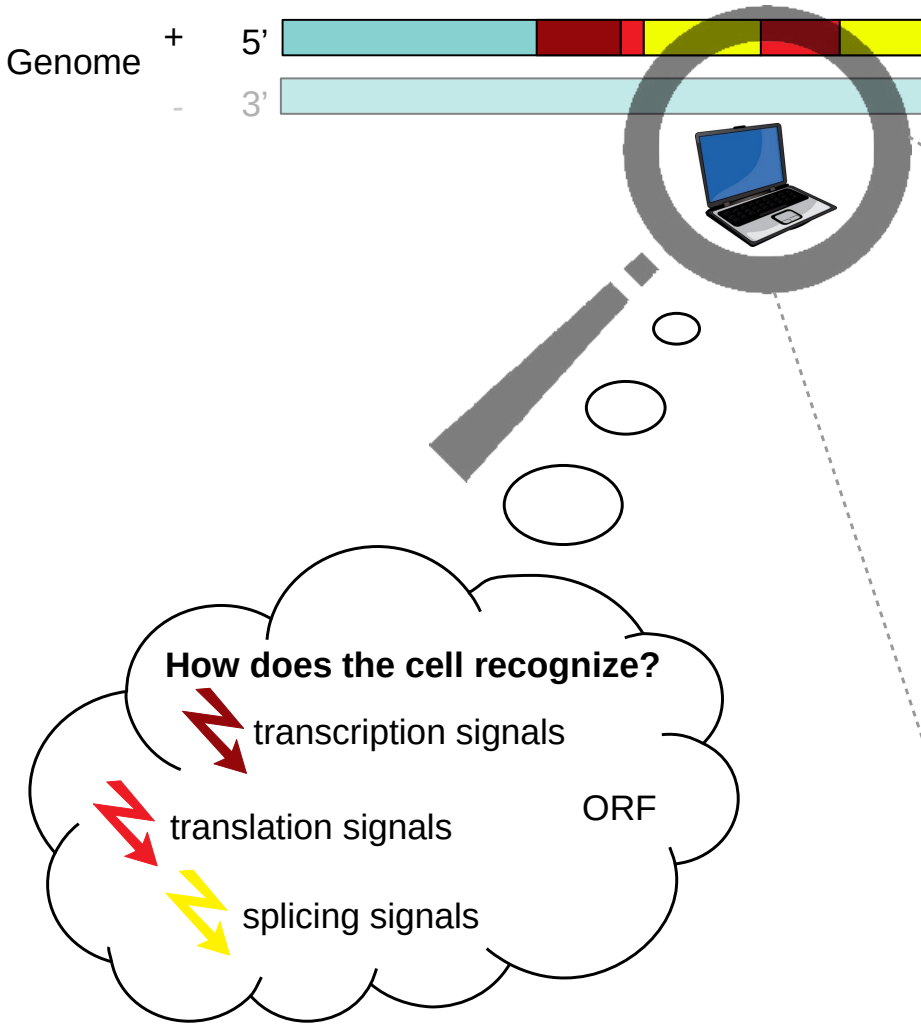
- CDS-intron structure

Limitations:

- alignment difficult
- incomplete data

What aids in the identification of genes in genomes?

Mathematical models



Mathematical models:

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

Provide information on:

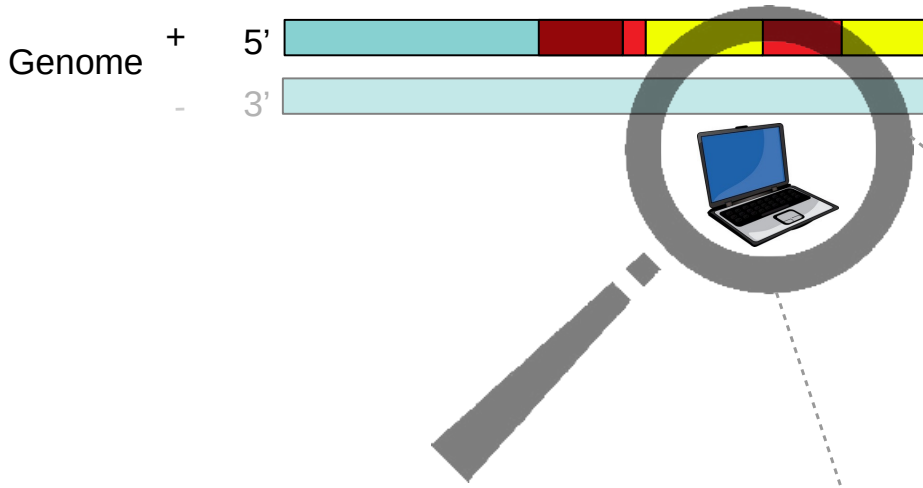
- complete gene structures (sometimes incl. UTRs)

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
- ...

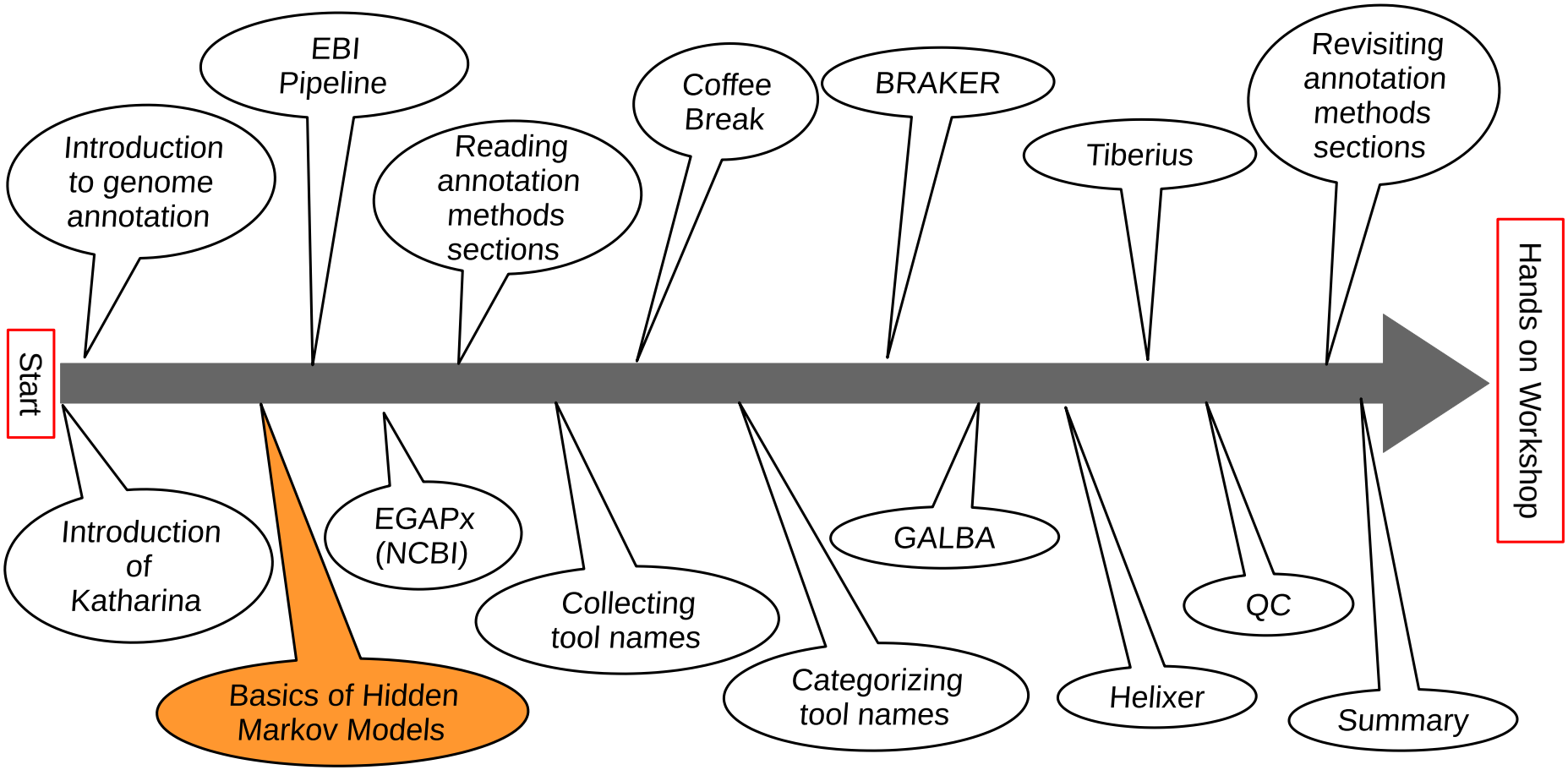
Provide information on:

- 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
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

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

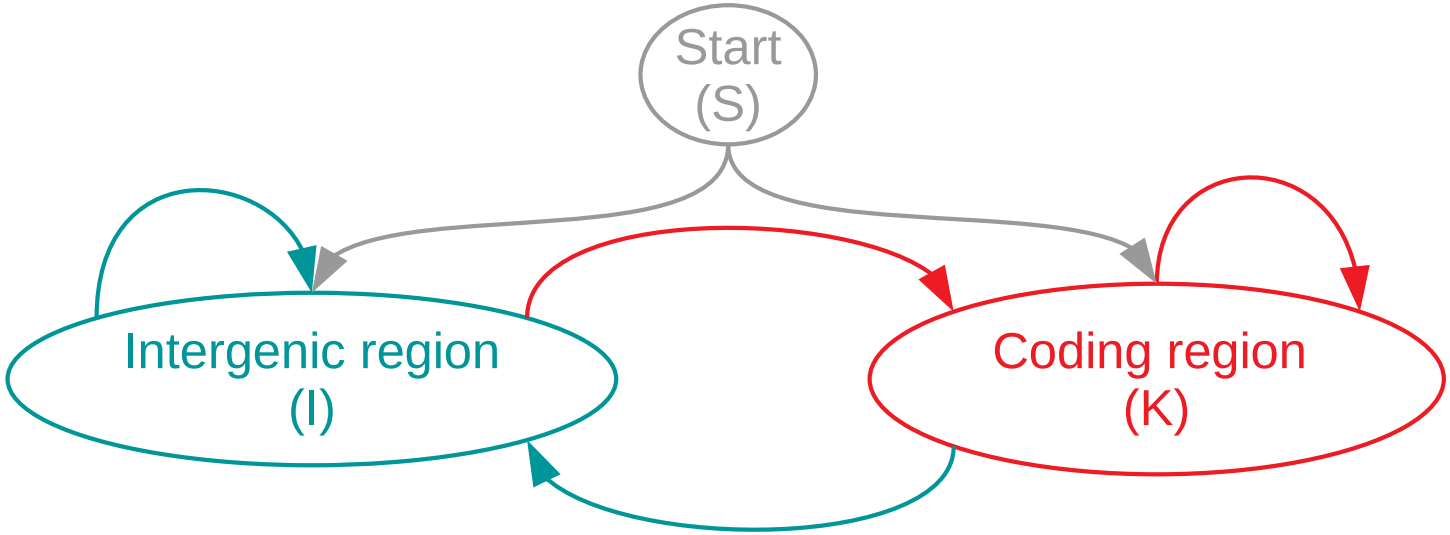
Goal: “Most likely path through hidden states”

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

Basis of highly accurate gene prediction tools

Hidden Markov Model

Hidden states

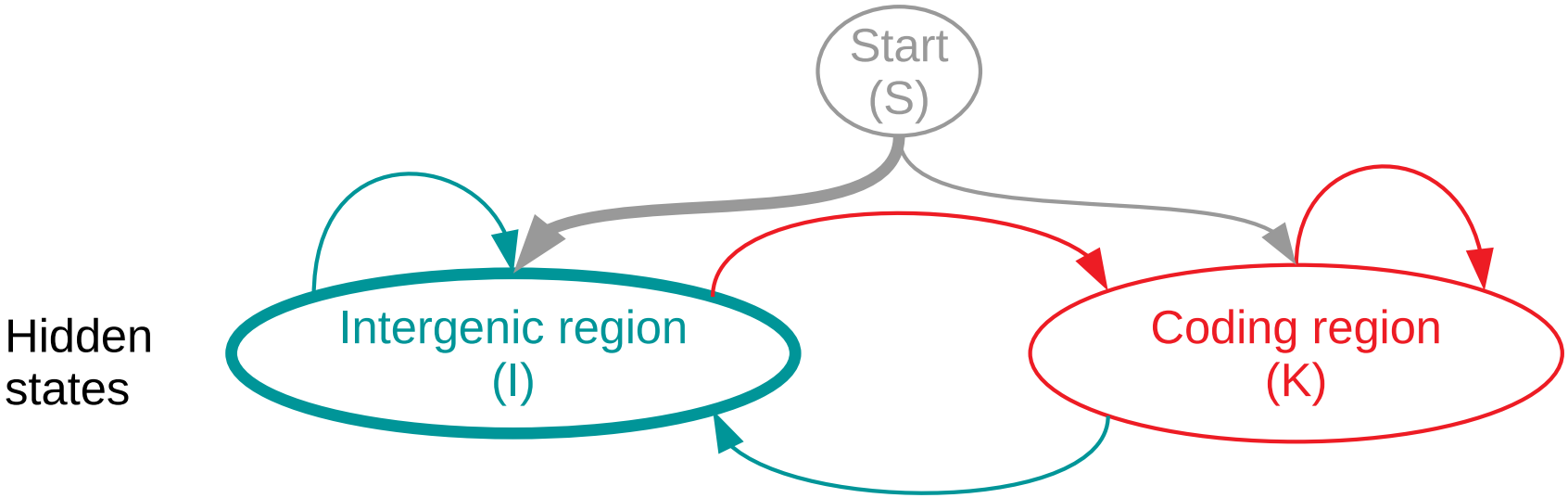


A possible 'state path' for the genome sequence:

AABBBAA

Basis of highly accurate gene prediction tools

Hidden Markov Model



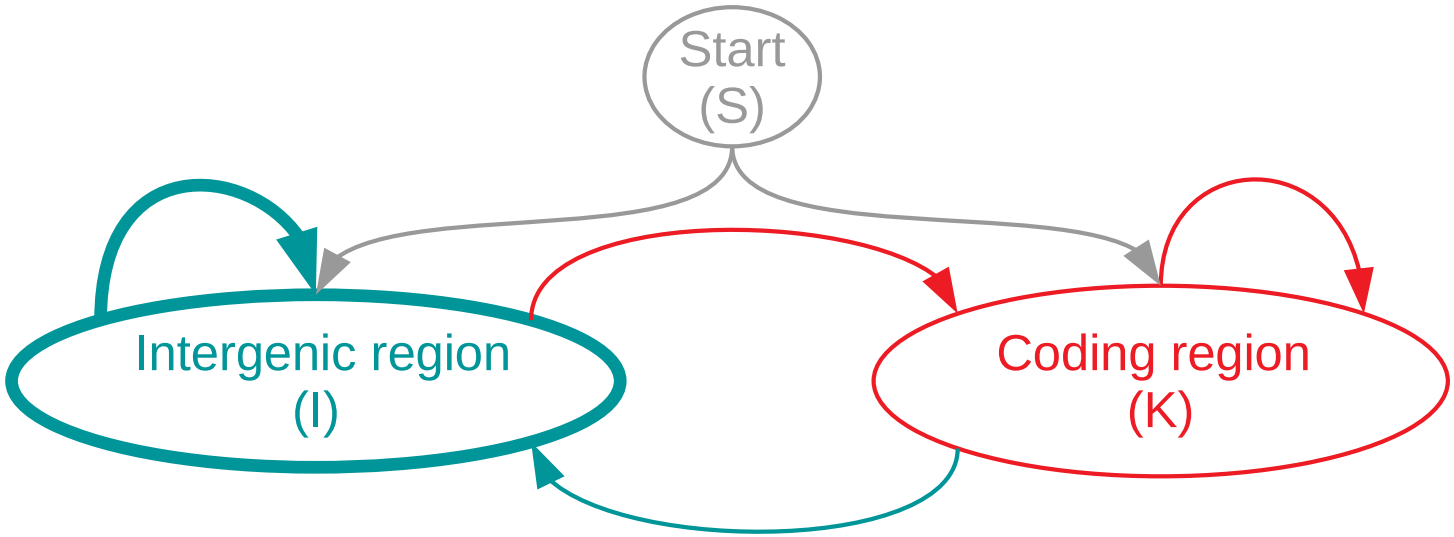
A possible 'state path' for the genome sequence:

AABBBAA
I

Basis of highly accurate gene prediction tools

Hidden Markov Model

Hidden states

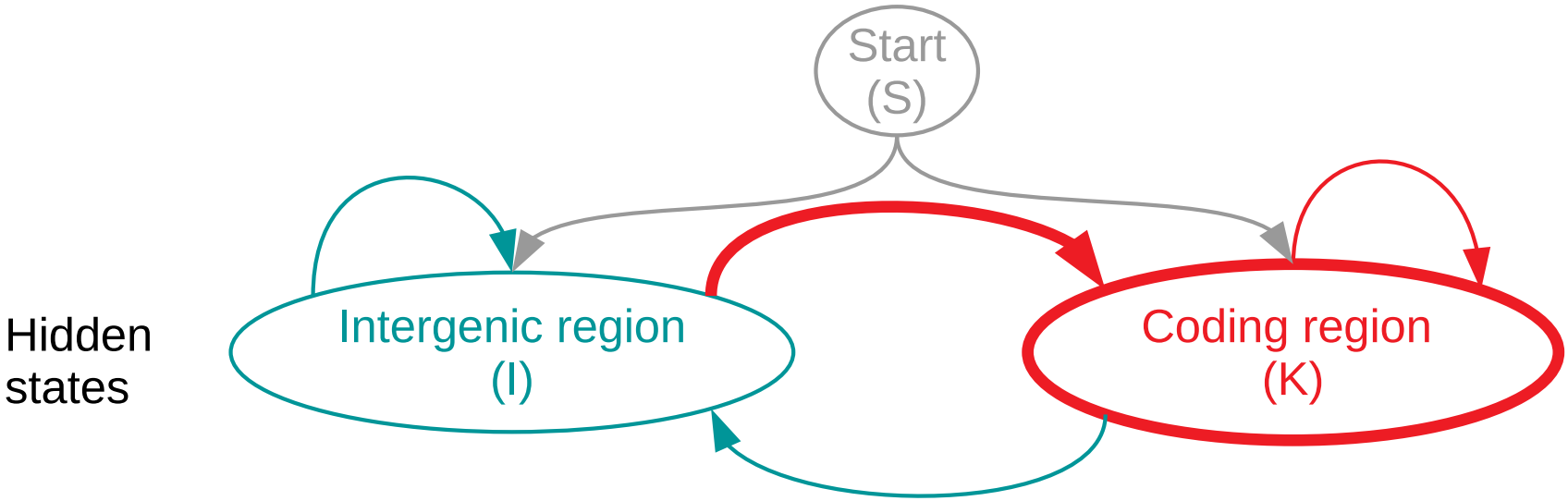


A possible 'state path' for the genome sequence:

AABBBAA
II

Basis of highly accurate gene prediction tools

Hidden Markov Model



A possible 'state path' for the genome sequence:

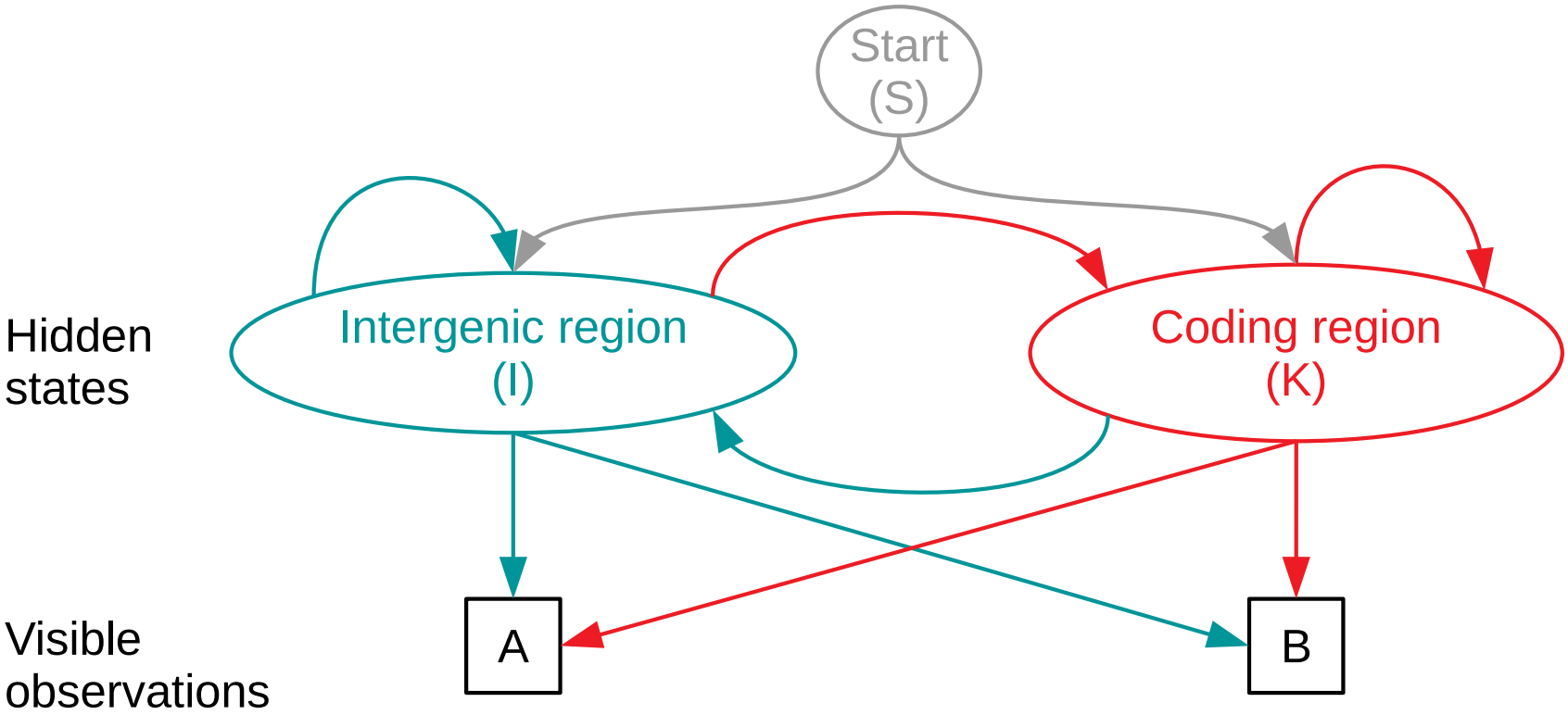
```
AABBBAA
IIK...
```

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

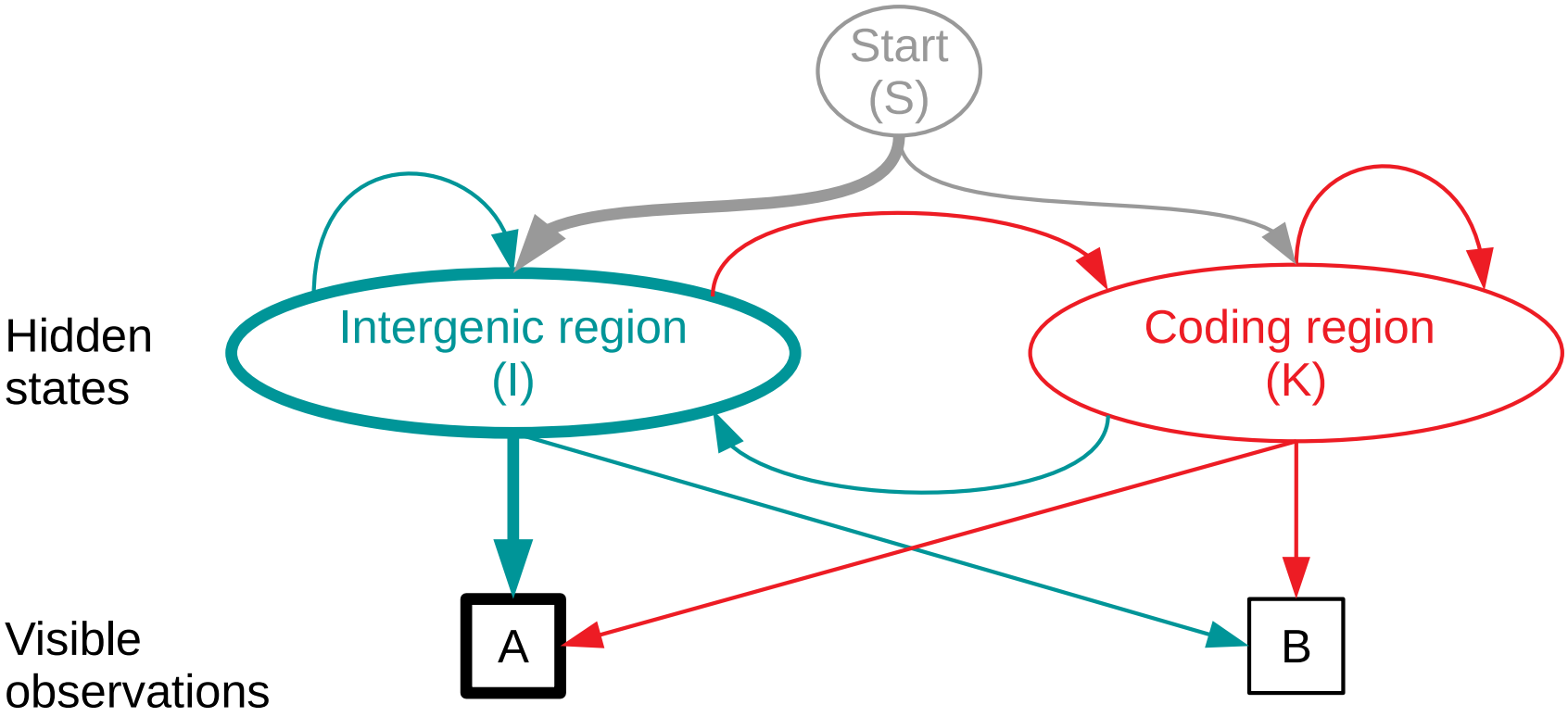
Hidden Markov Model



A possible ‘state path’ for the genome sequence:

Basis of highly accurate gene prediction tools

Hidden Markov Model

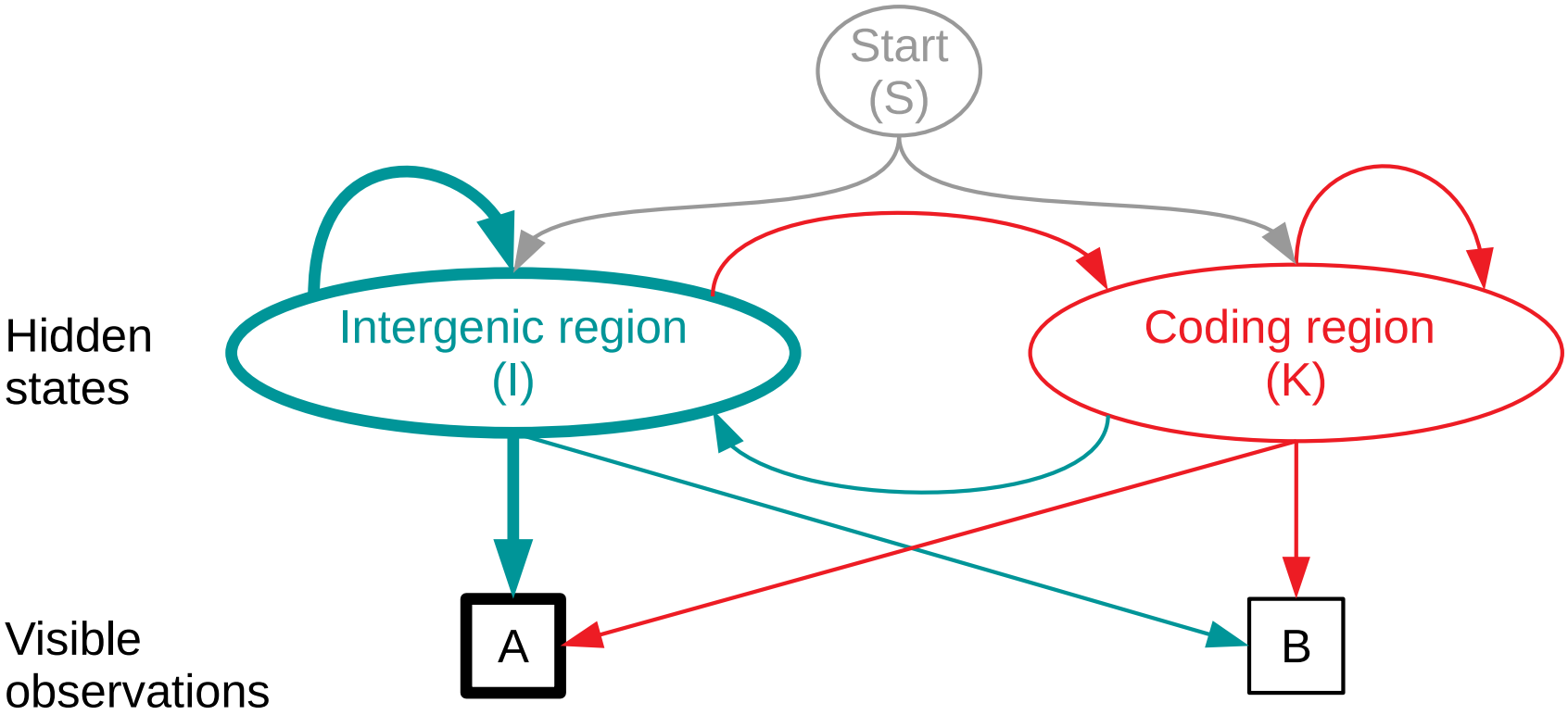


A possible ‘state path’ for the genome sequence:

A
I

Basis of highly accurate gene prediction tools

Hidden Markov Model

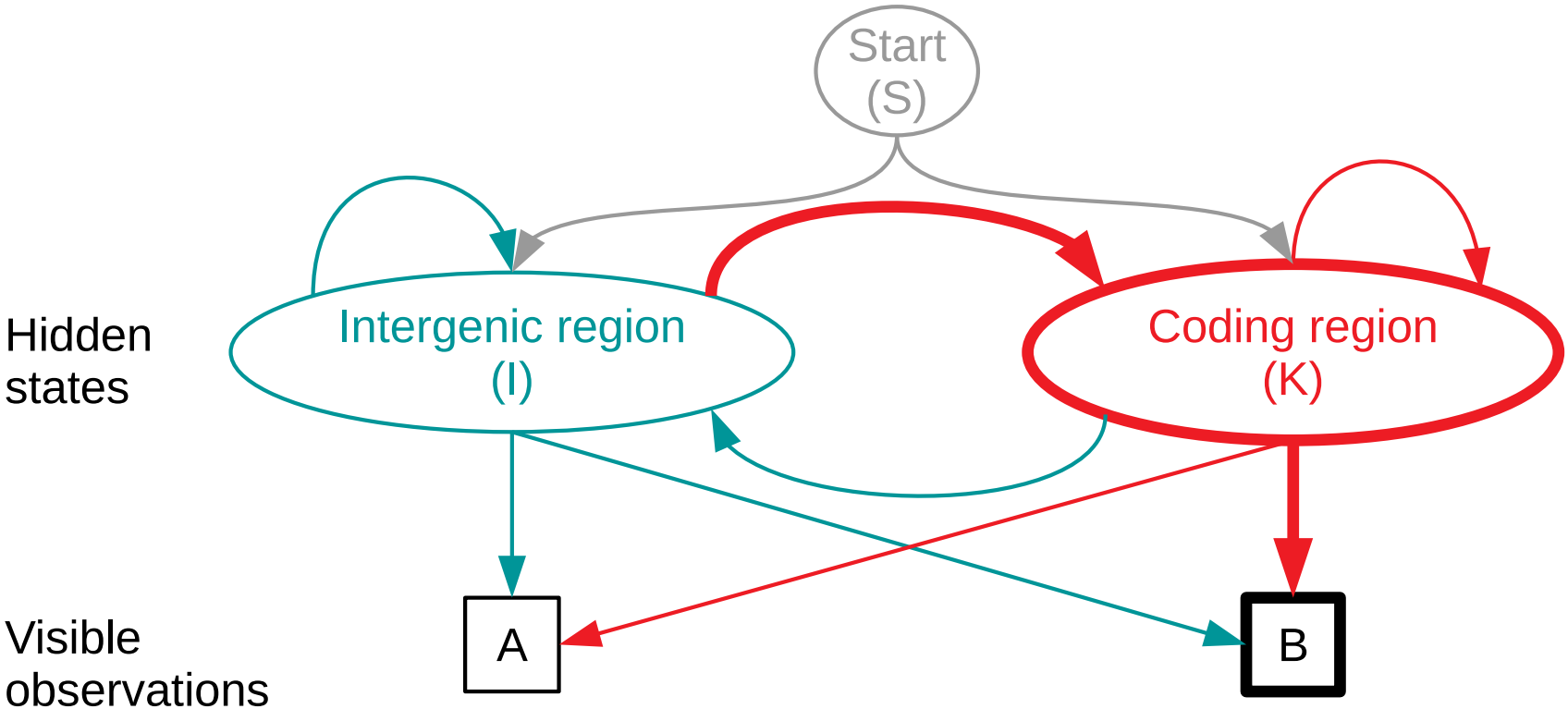


A possible ‘state path’ for the genome sequence:

AA
II

Basis of highly accurate gene prediction tools

Hidden Markov Model

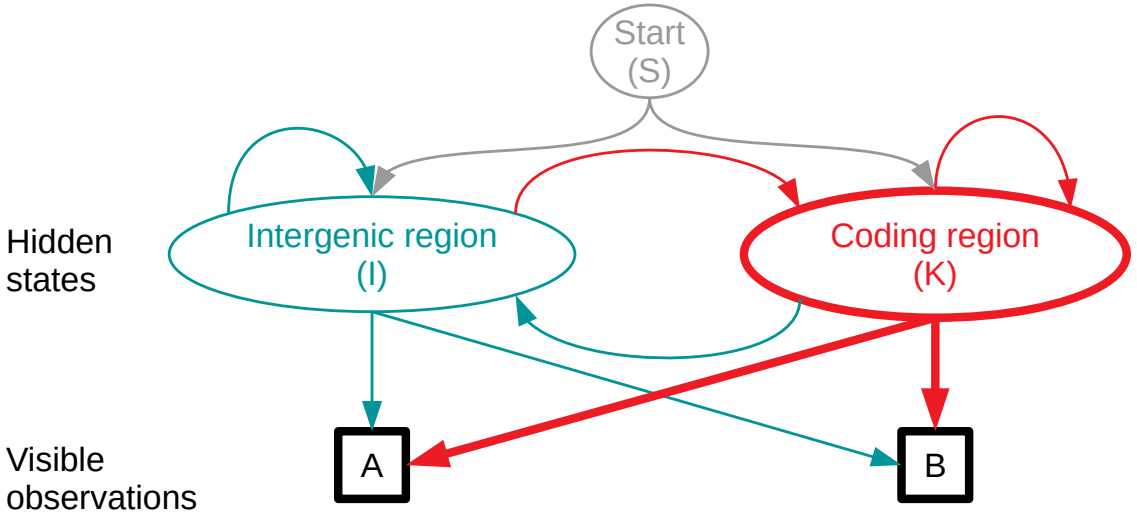


A possible ‘state path’ for the genome sequence:

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

Basis of highly accurate gene prediction tools

Hidden Markov Model

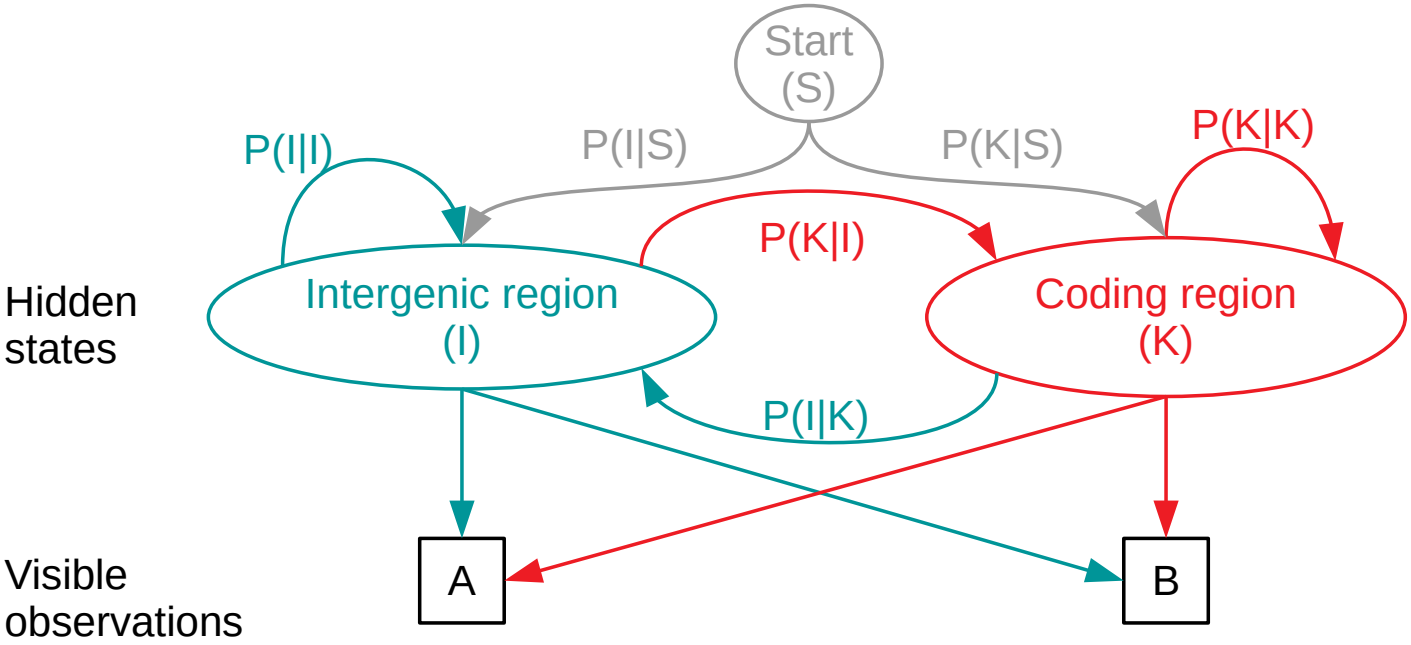


Model properties

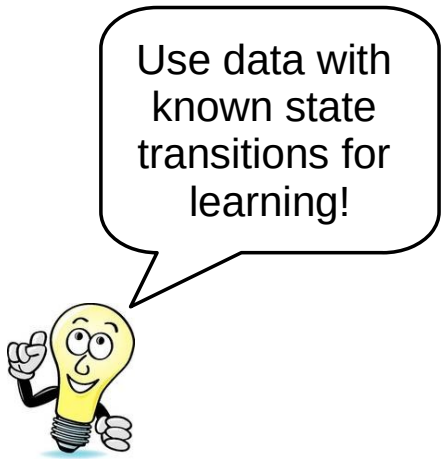
1. The current value of the hidden state depends exclusively on the state of its predecessor.
2. The current value of the visible observation depends exclusively on the value of the current, hidden state.

Basis of highly accurate gene prediction tools

Hidden Markov Model

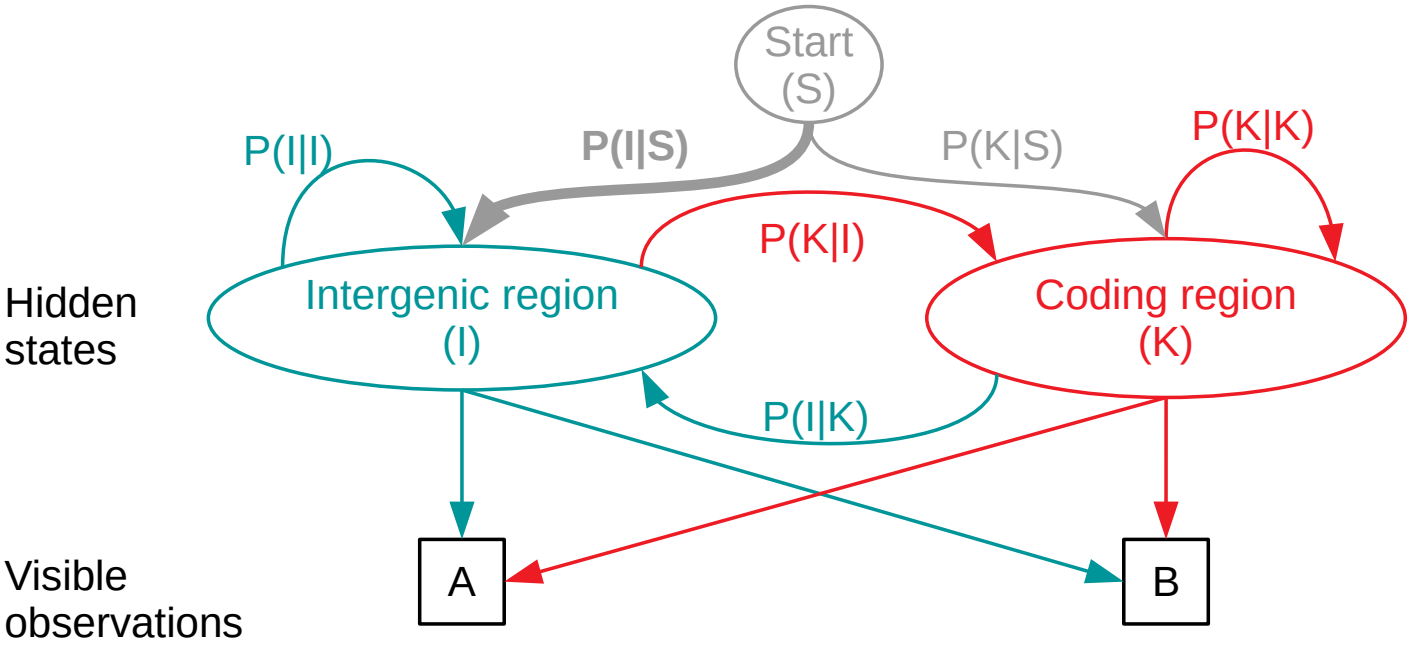


How likely are the state transitions?



Basis of highly accurate gene prediction tools

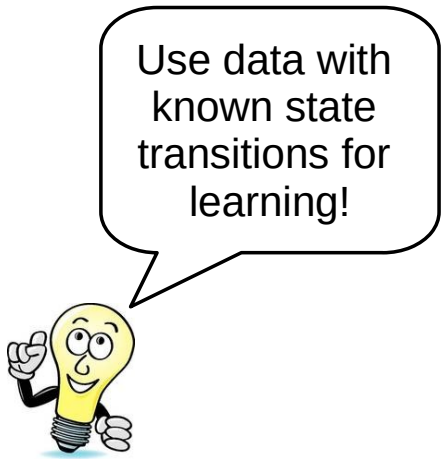
Hidden Markov Model



Training data:

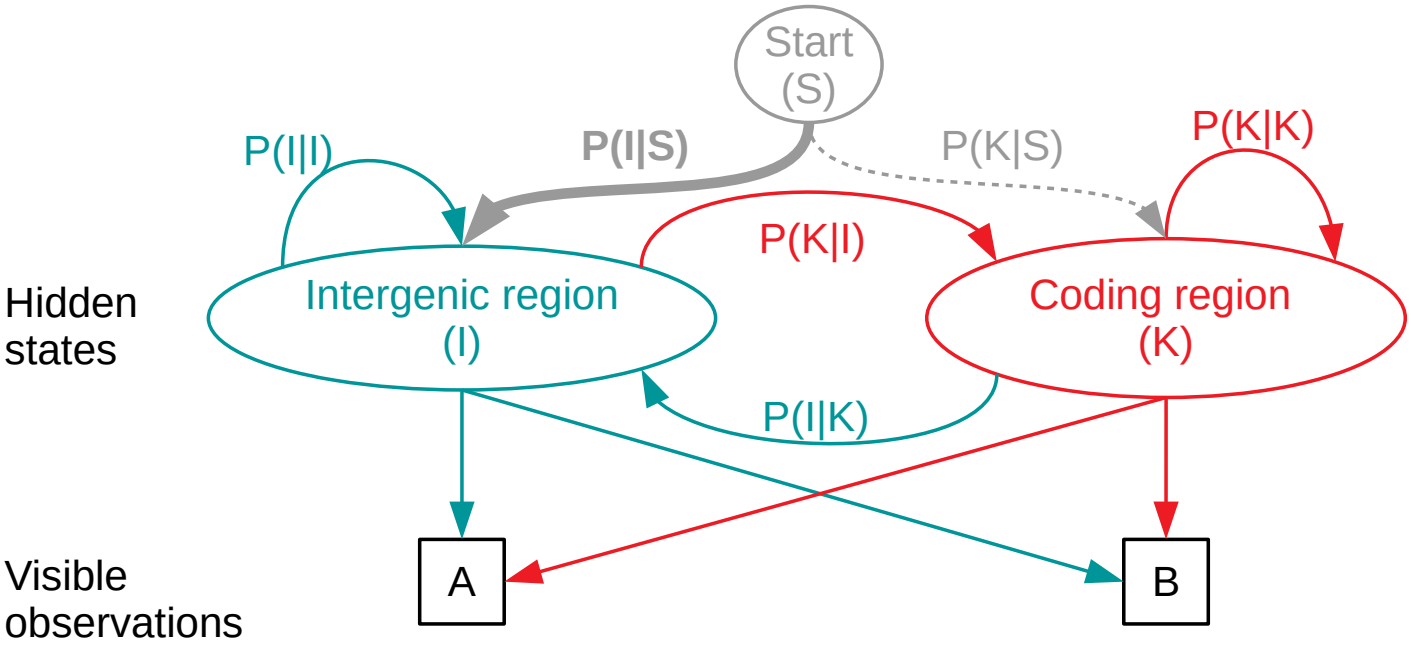
AABABA
IKKIII

Start probability
P(I|S) = ?



Basis of highly accurate gene prediction tools

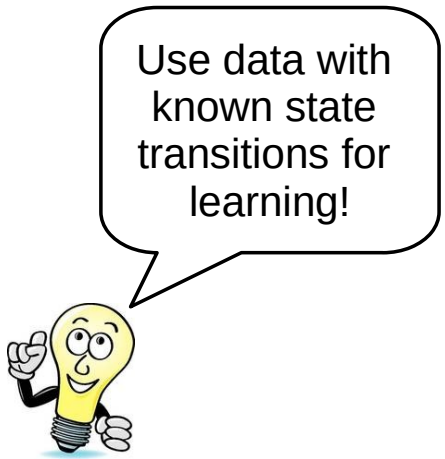
Hidden Markov Model



Training data:

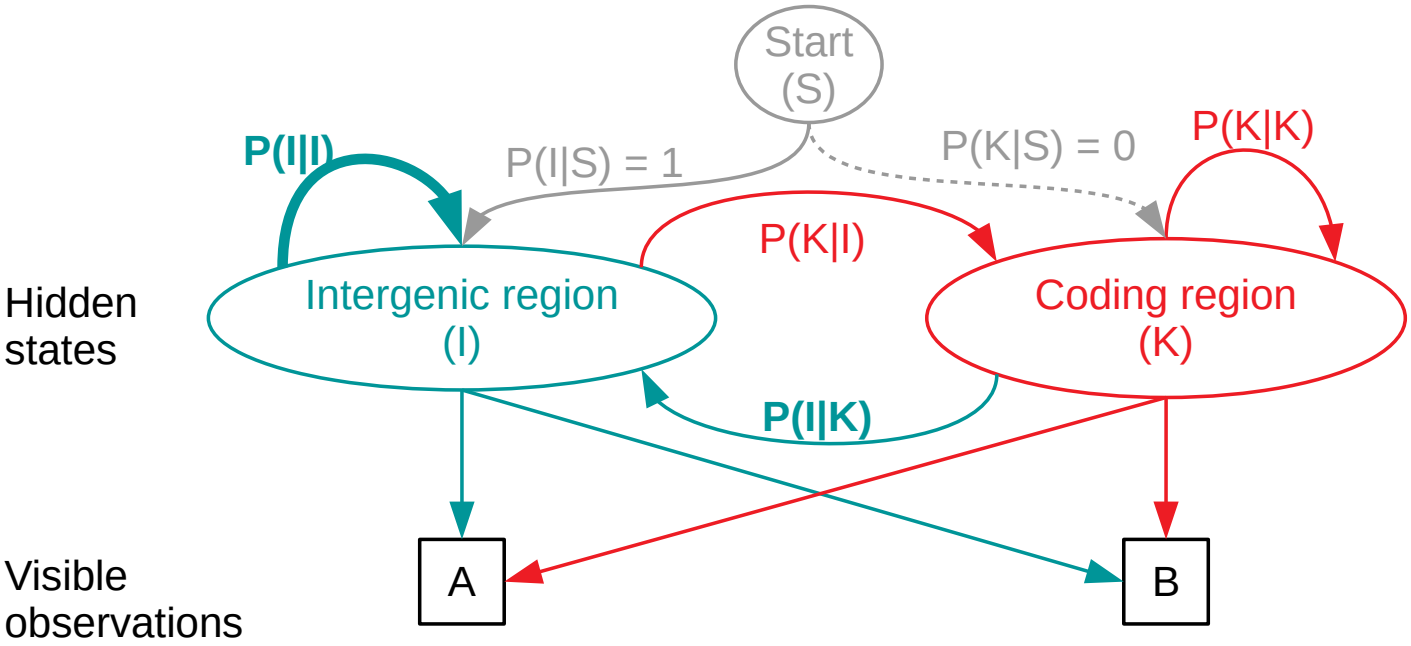
AABABA
IKKIII
+

Start probability
P(I|S) = 1



Basis of highly accurate gene prediction tools

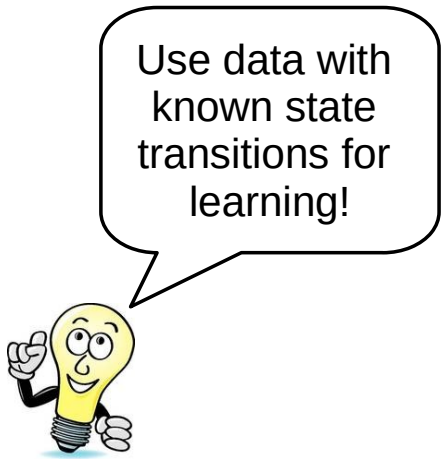
Hidden Markov Model



Training data:

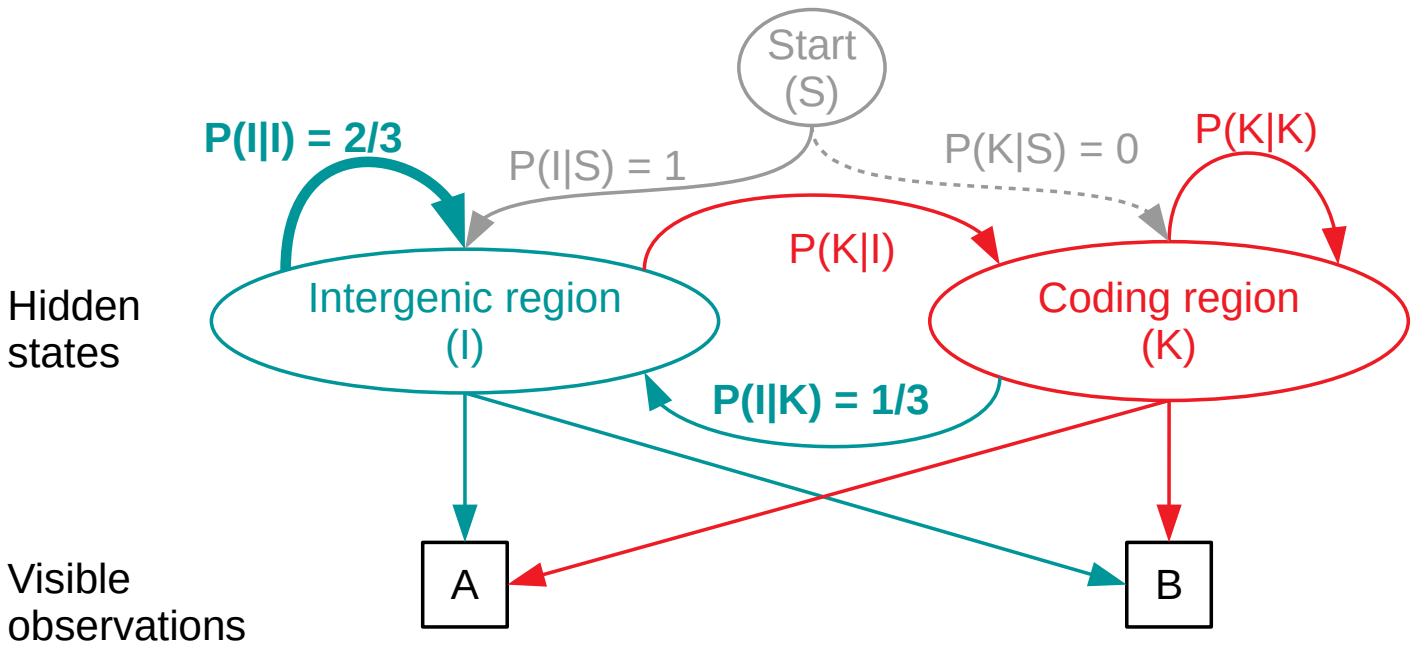
AABABA
IKKIII

P(I|I) = ?



Basis of highly accurate gene prediction tools

Hidden Markov Model



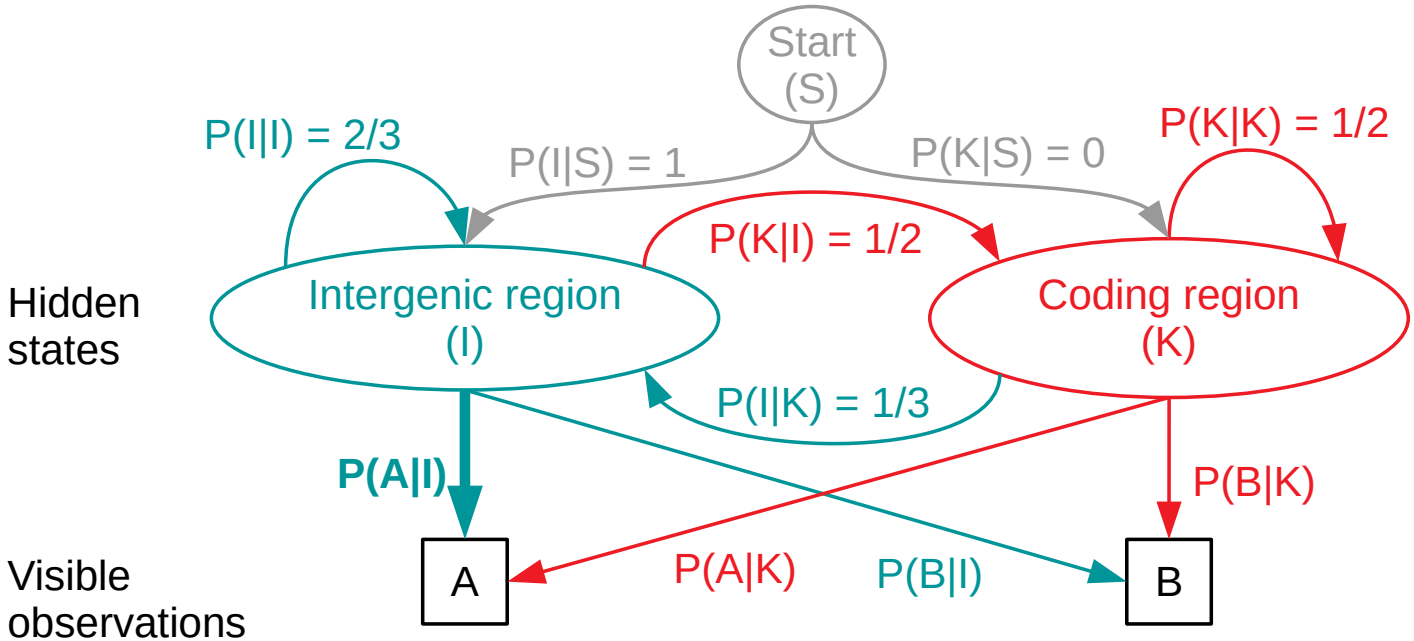
Training data:

AABABA
IKKIII
-++

P(I|I) = 2/3 P(I|K) = 1 – P(I|I) = 1/3

Basis of highly accurate gene prediction tools

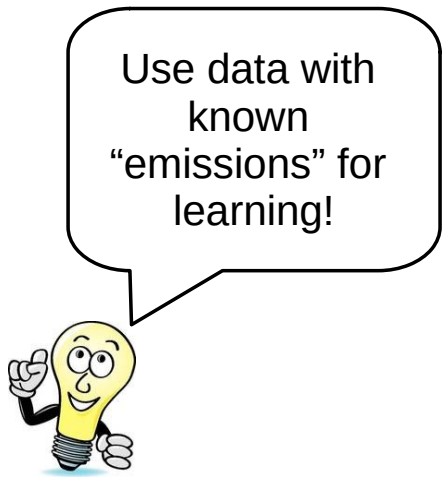
Hidden Markov Model



How likely are the observations?

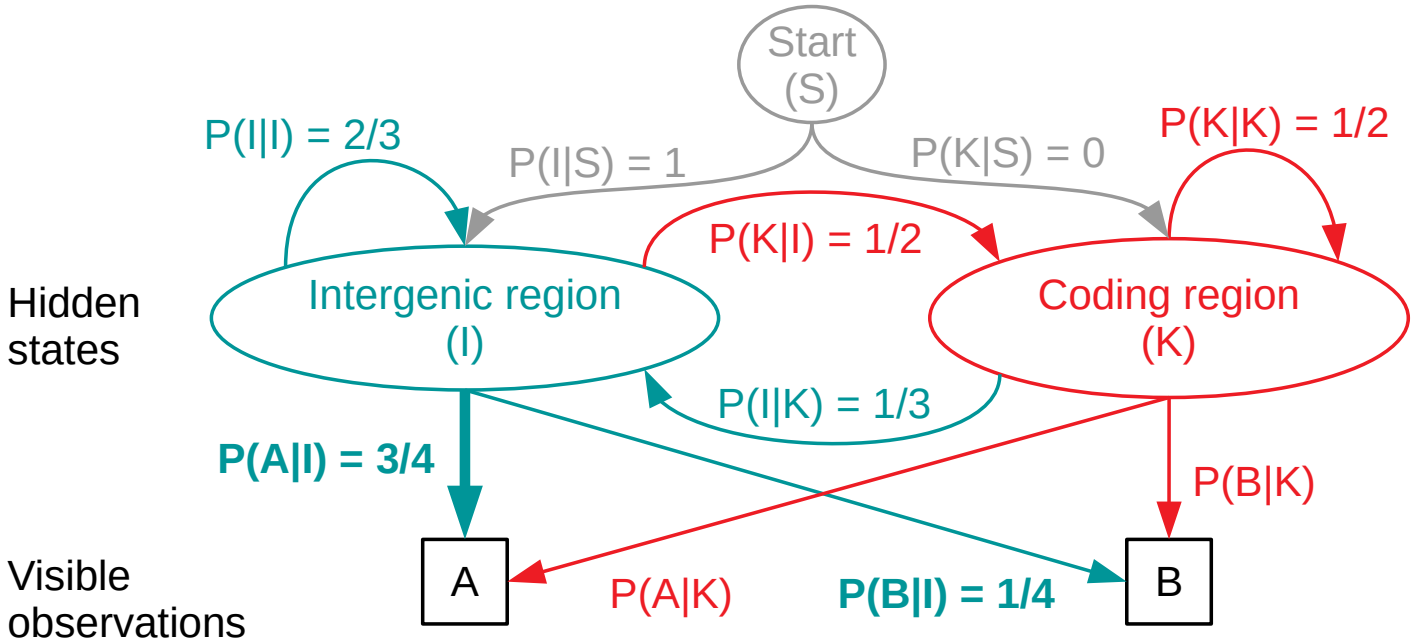
AABABA
IKKIII

P(A|I) = ?



Basis of highly accurate gene prediction tools

Hidden Markov Model



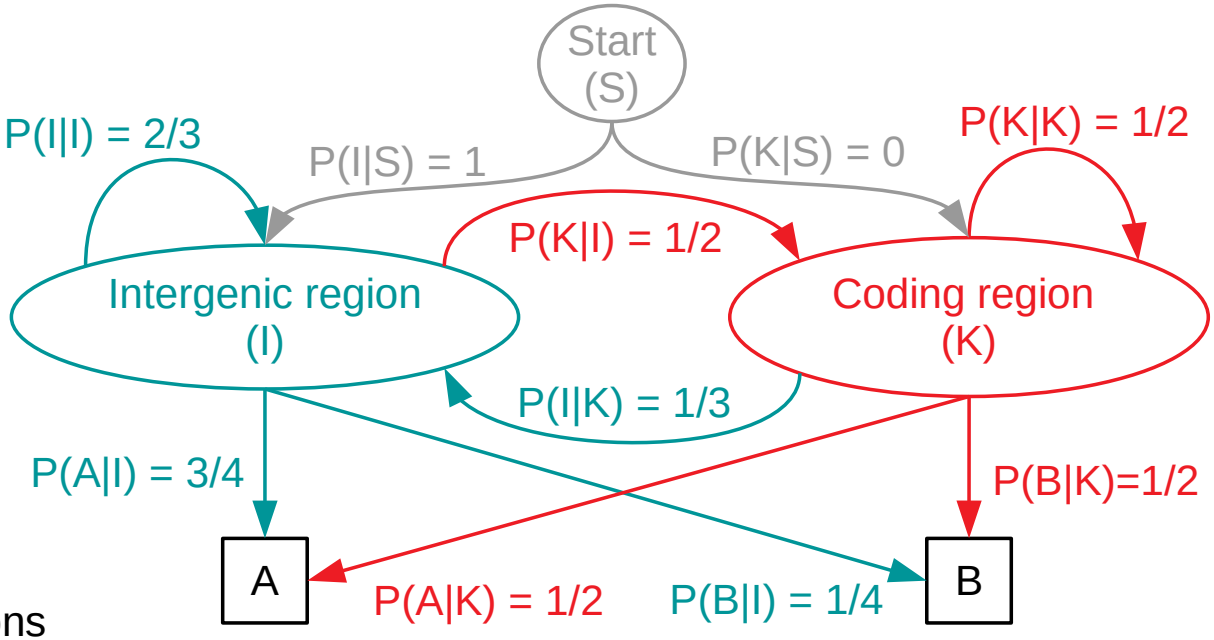
How likely are the observations?

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

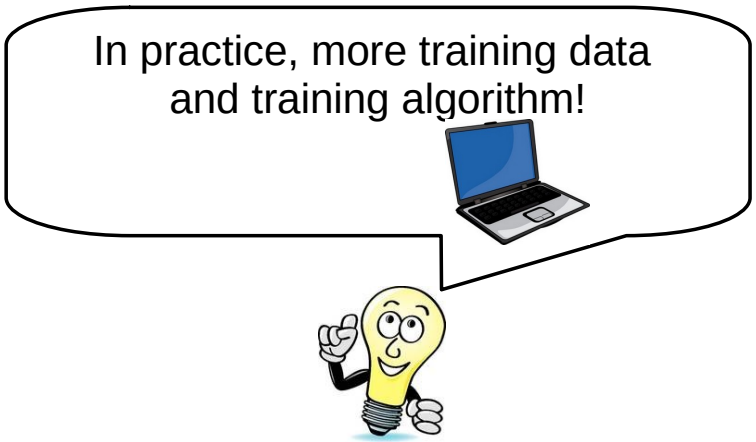


Hidden states

Visible observations (emissions)

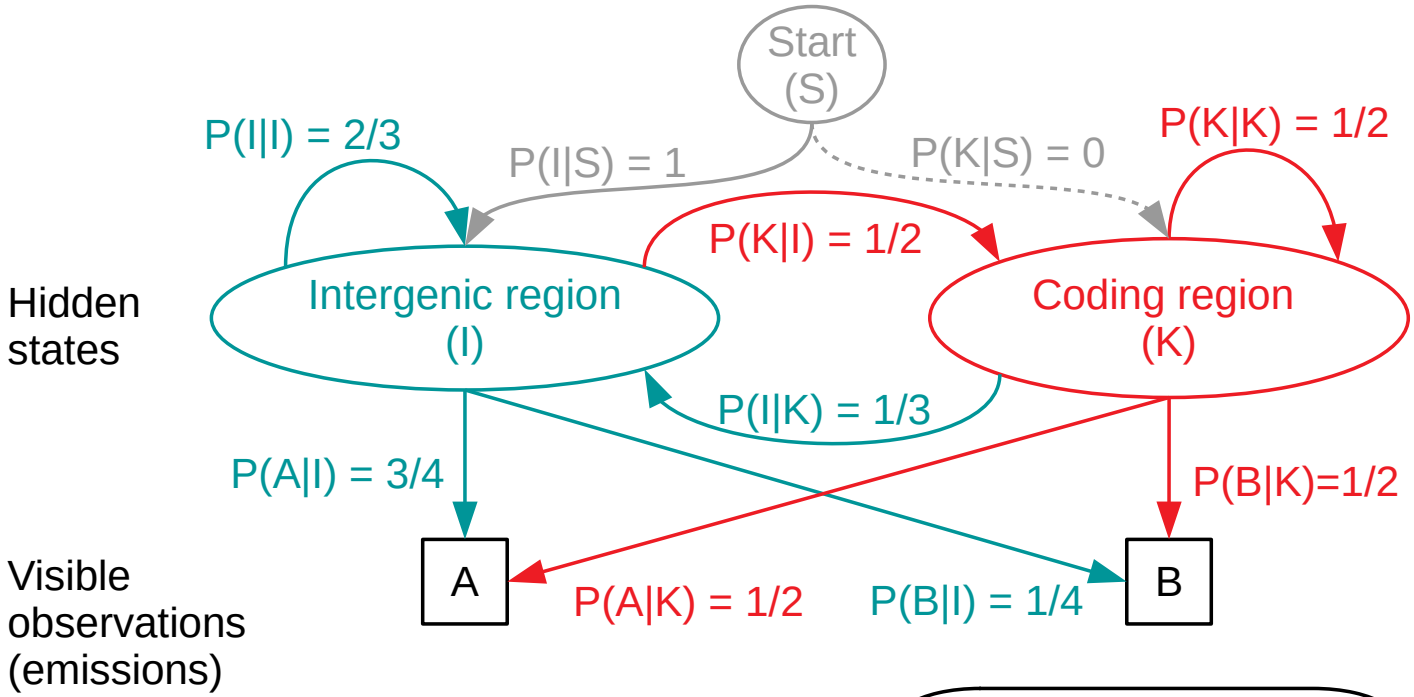
Training data:

AABABA
IKKIII



Basis of highly accurate gene prediction tools

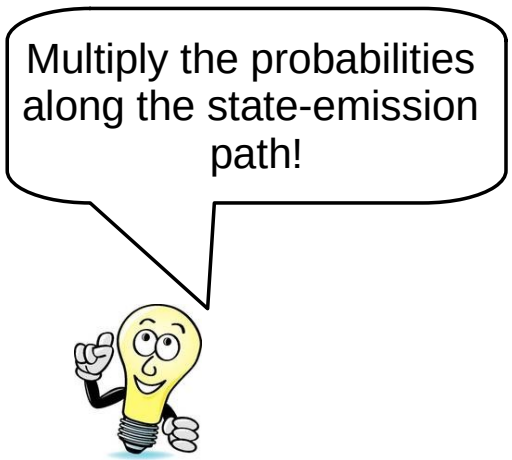
Hidden Markov Model



How likely is a given state-emission path?

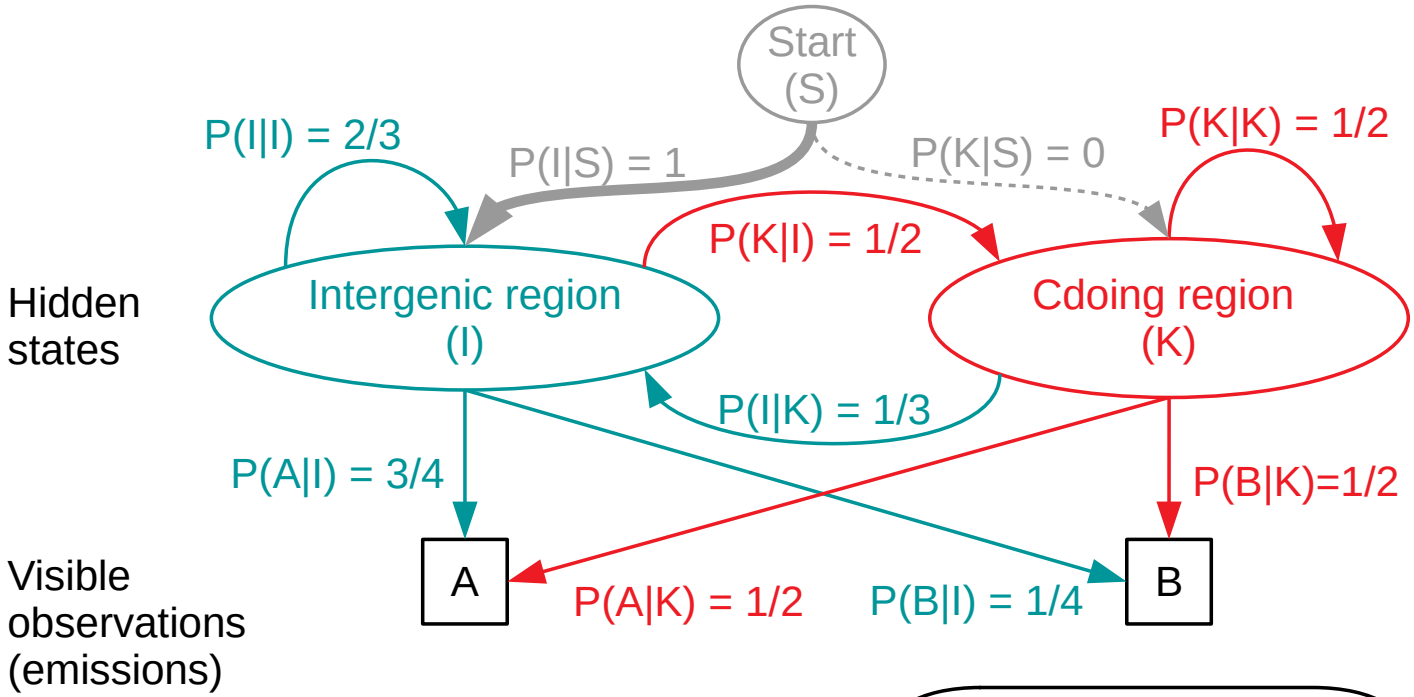
```
Path = AAB
       IKK

P(Path) = ?
```



Basis of highly accurate gene prediction tools

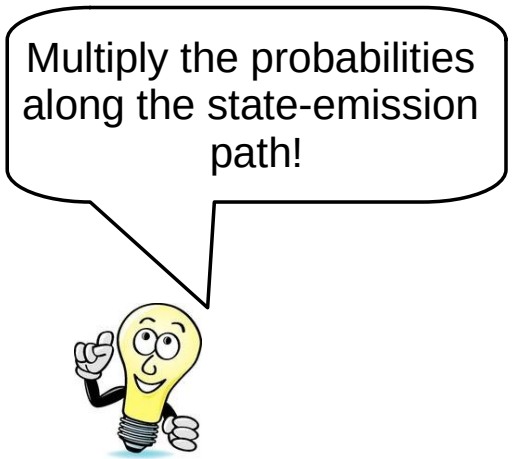
Hidden Markov Model



How likely is a given state-emission path?

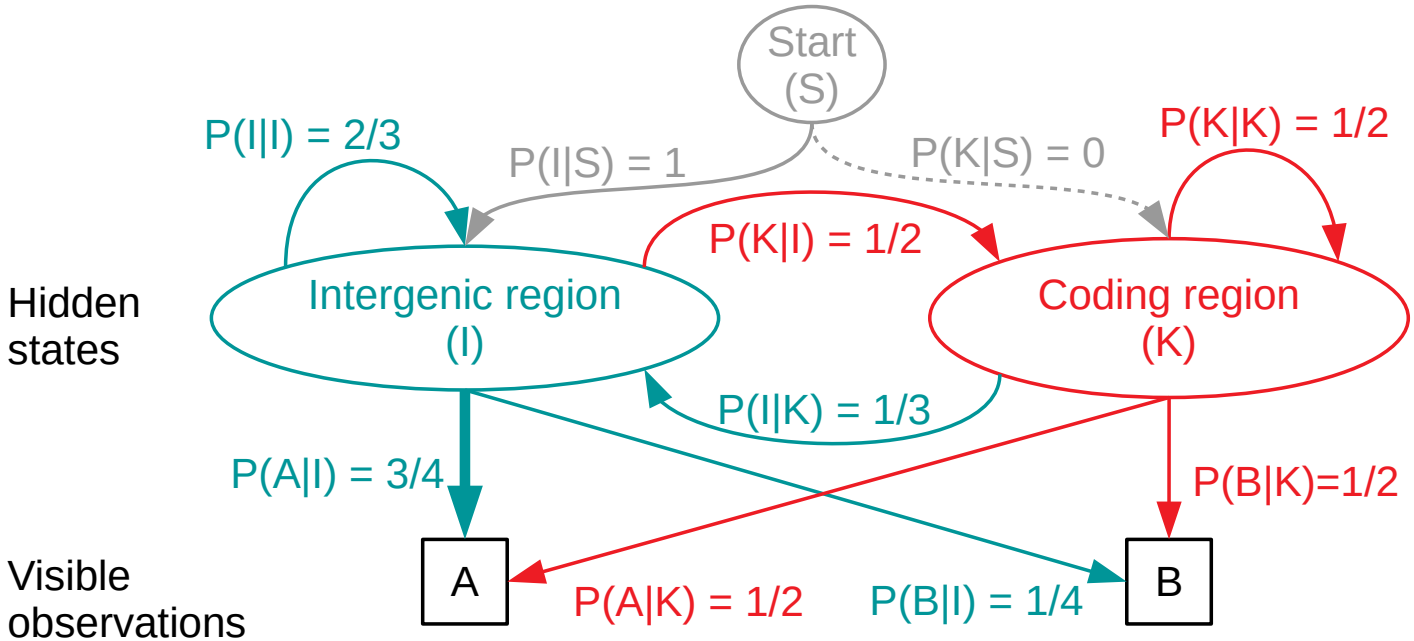
```
Path = AAB
       IKK

P(Path) = P(I|S)
```



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

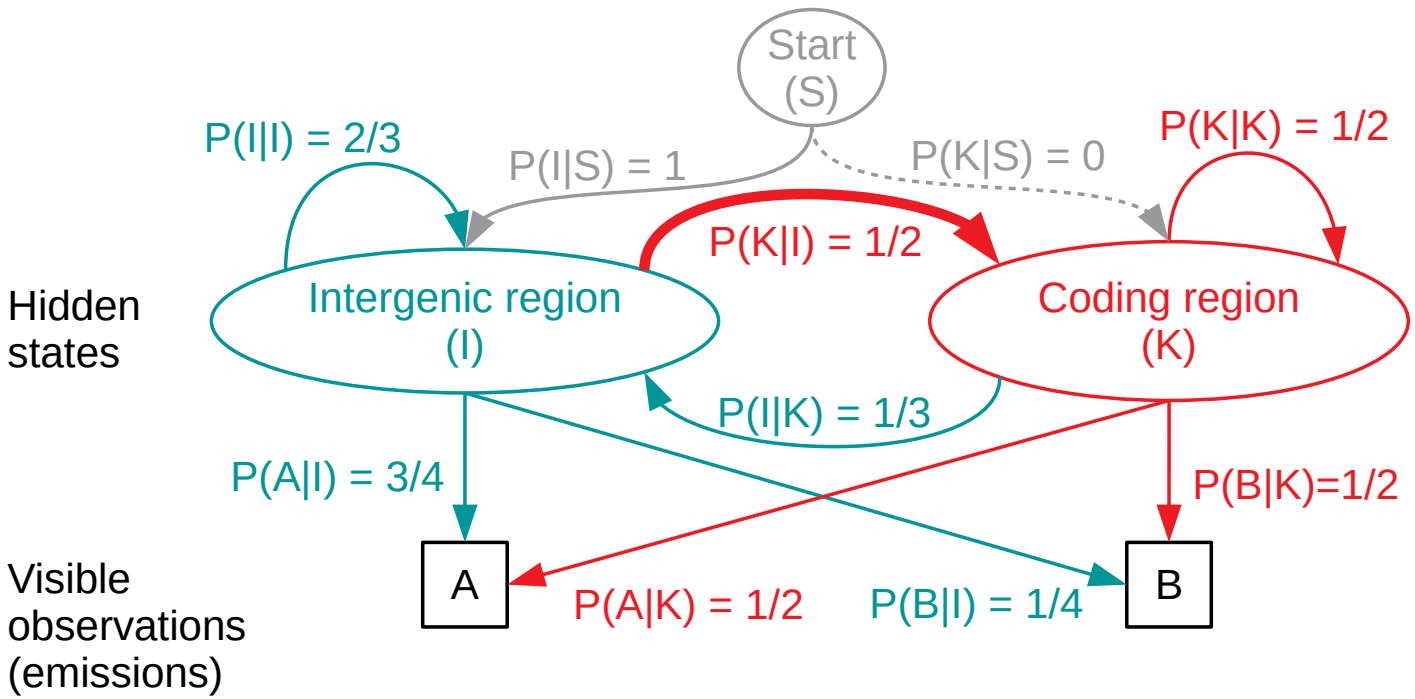
```
Path = AAB
       IKK

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



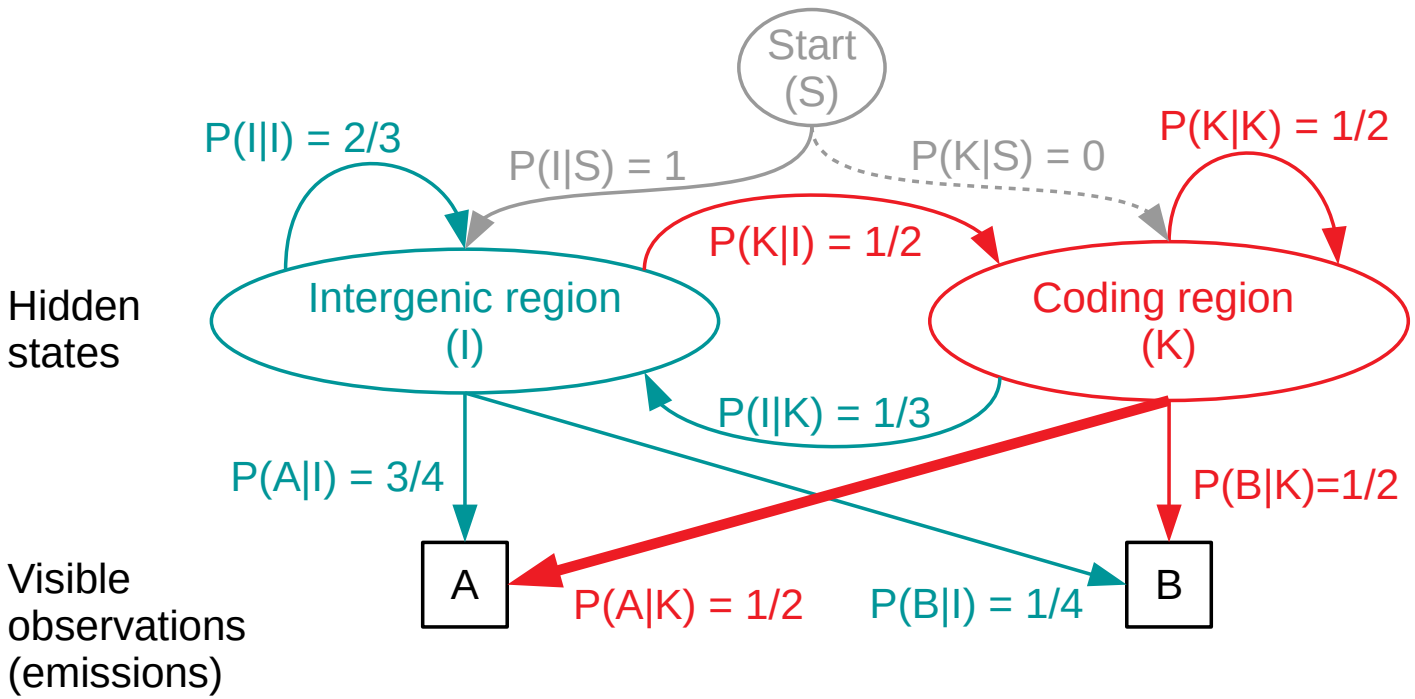
How likely is a given state-emission path?

```
Path = AAB
       IKK

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

Basis of highly accurate gene prediction tools

Hidden Markov Model



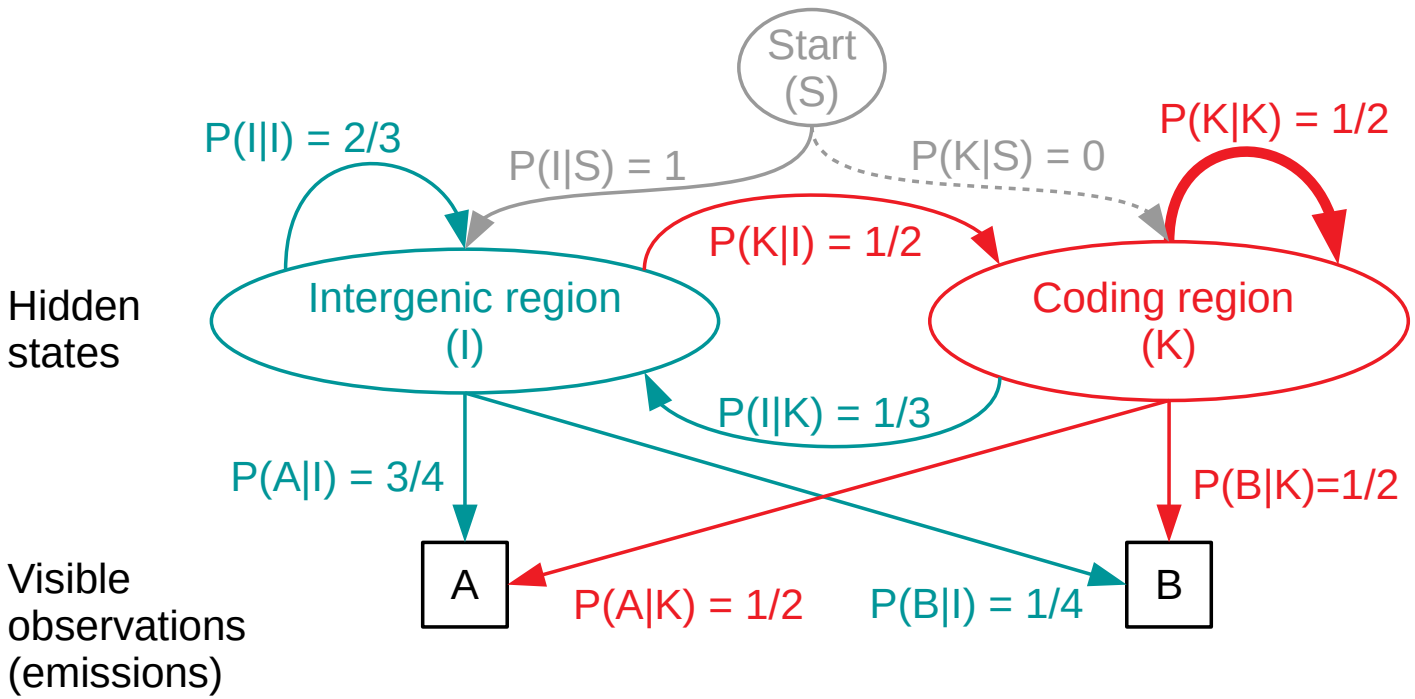
How likely is a given state-emission path?

```
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



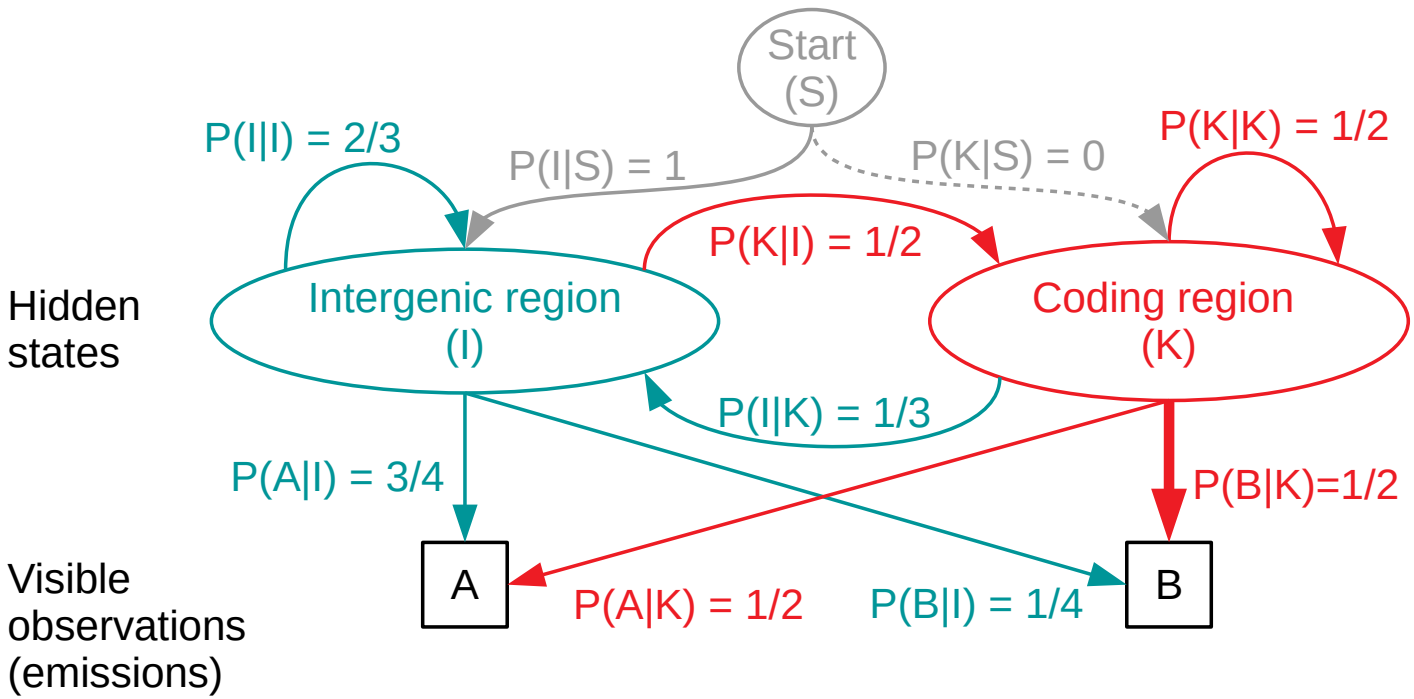
How likely is a given state-emission path?

```
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



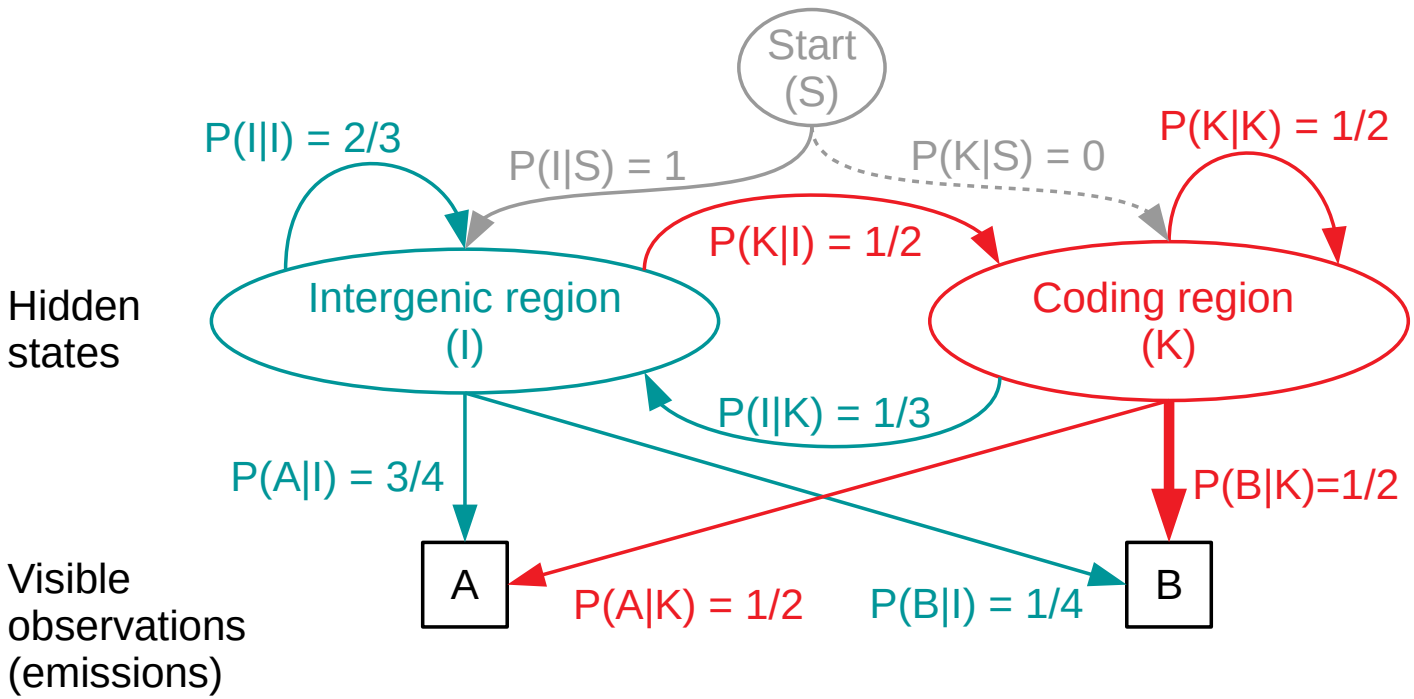
How likely is a given state-emission path?

```
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
```

Find the most probable state sequence for a given sequence

Input: “genome sequence”

AABBBABA

Problem: “too many possible state sequences”

IIIKKKKKK
KKIKKIIIK
IIKIIIKIK
IKKIKIIIK
KIKIKKKIK
KKKIKIKKK
...

Idea:

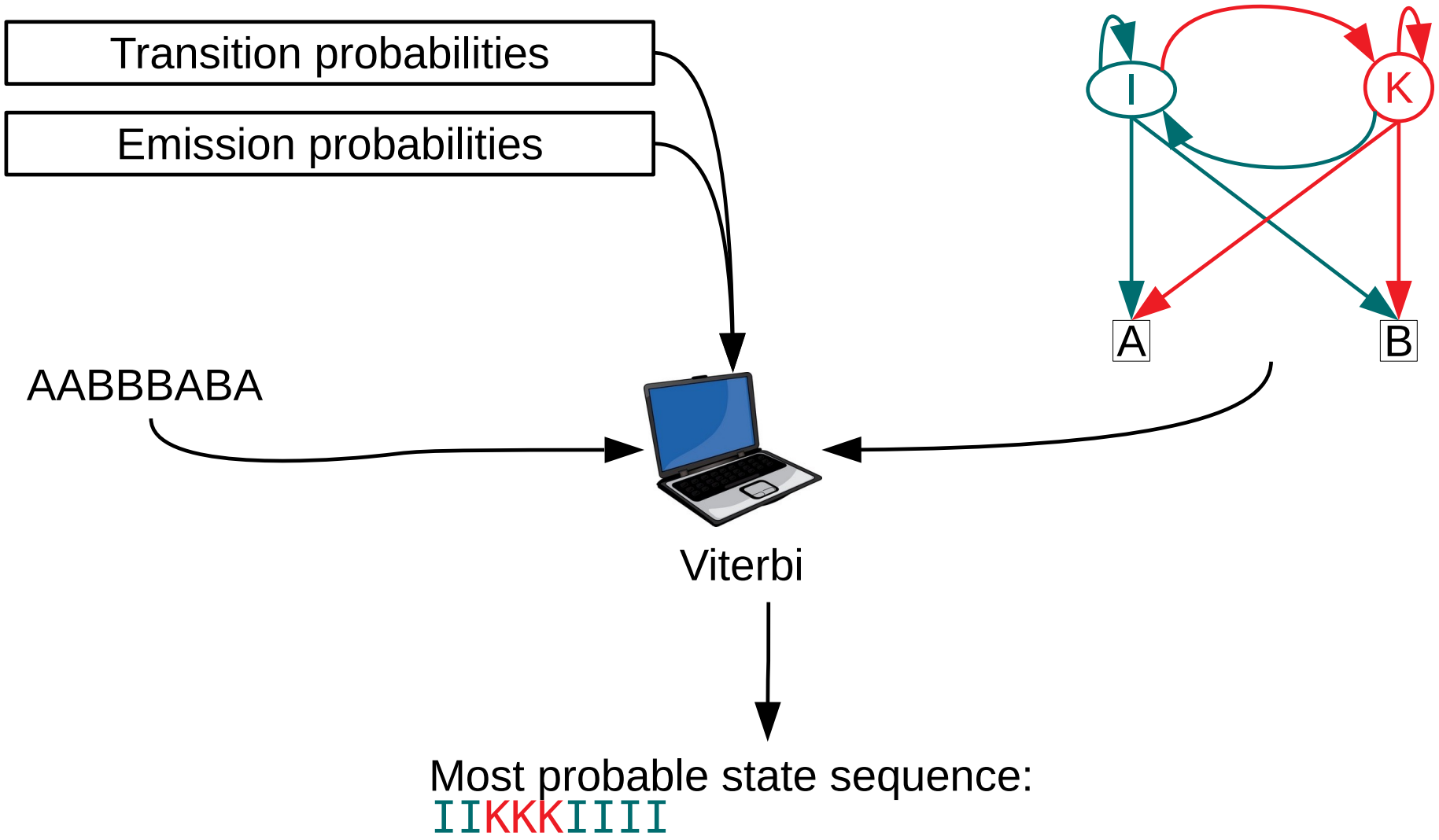
1. Generate all possible state sequences
2. Calculate the probability for each state sequence
3. Choose the state sequence with the highest probability

$\Rightarrow$ too expensive!

Basis of highly accurate gene prediction tools

Hidden Markov Model

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



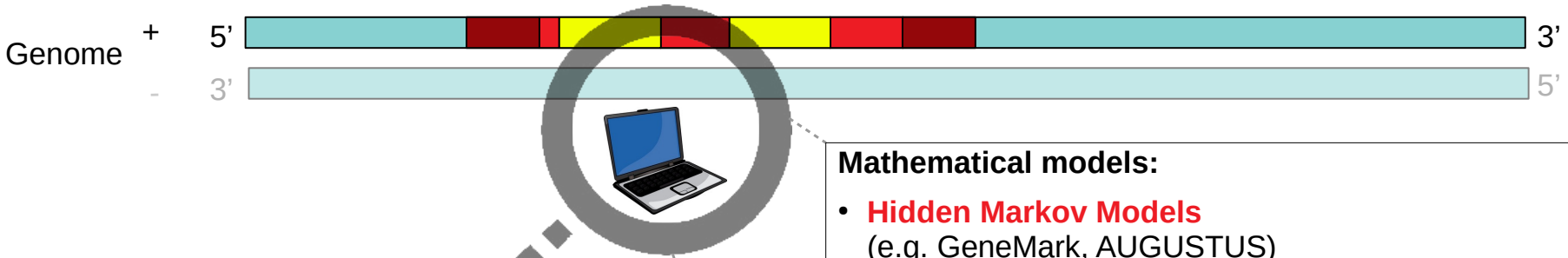
Hidden Markov Model for gene identification in practice

- 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:

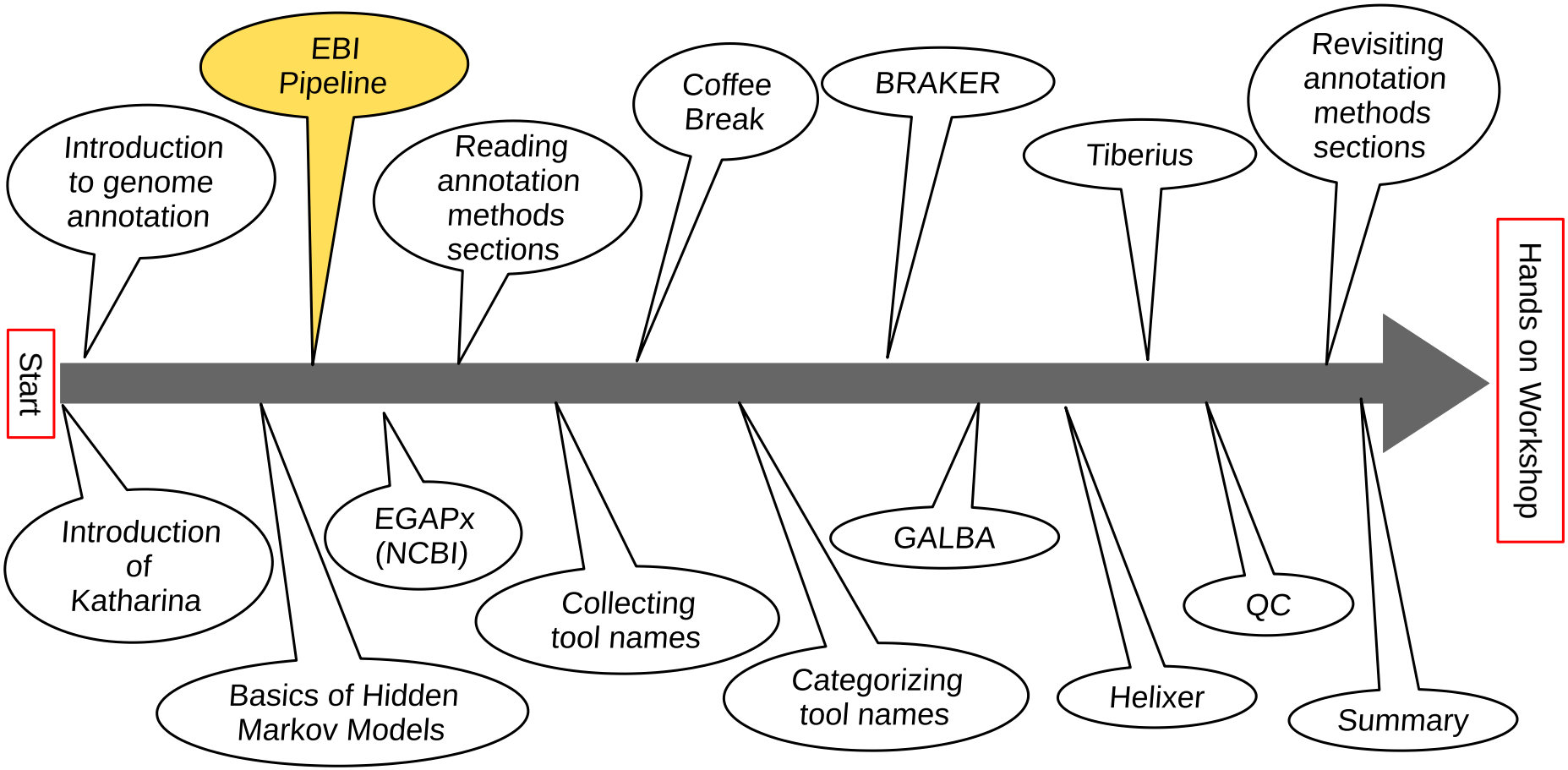
- **Hidden Markov Models** (e.g. GeneMark, AUGUSTUS)
- dynamic programming
- Support Vector Machines
- neural networks
- decision tree systems
- ...

Provide information on:

- complete gene structures (sometimes incl. UTRs)

Limitations

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


Start
Introduction 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

EBI: Ensembl annotation system

Ensembl annotation pipelines

Ensembl annotation pipeline for non-vertebrates

Documentation

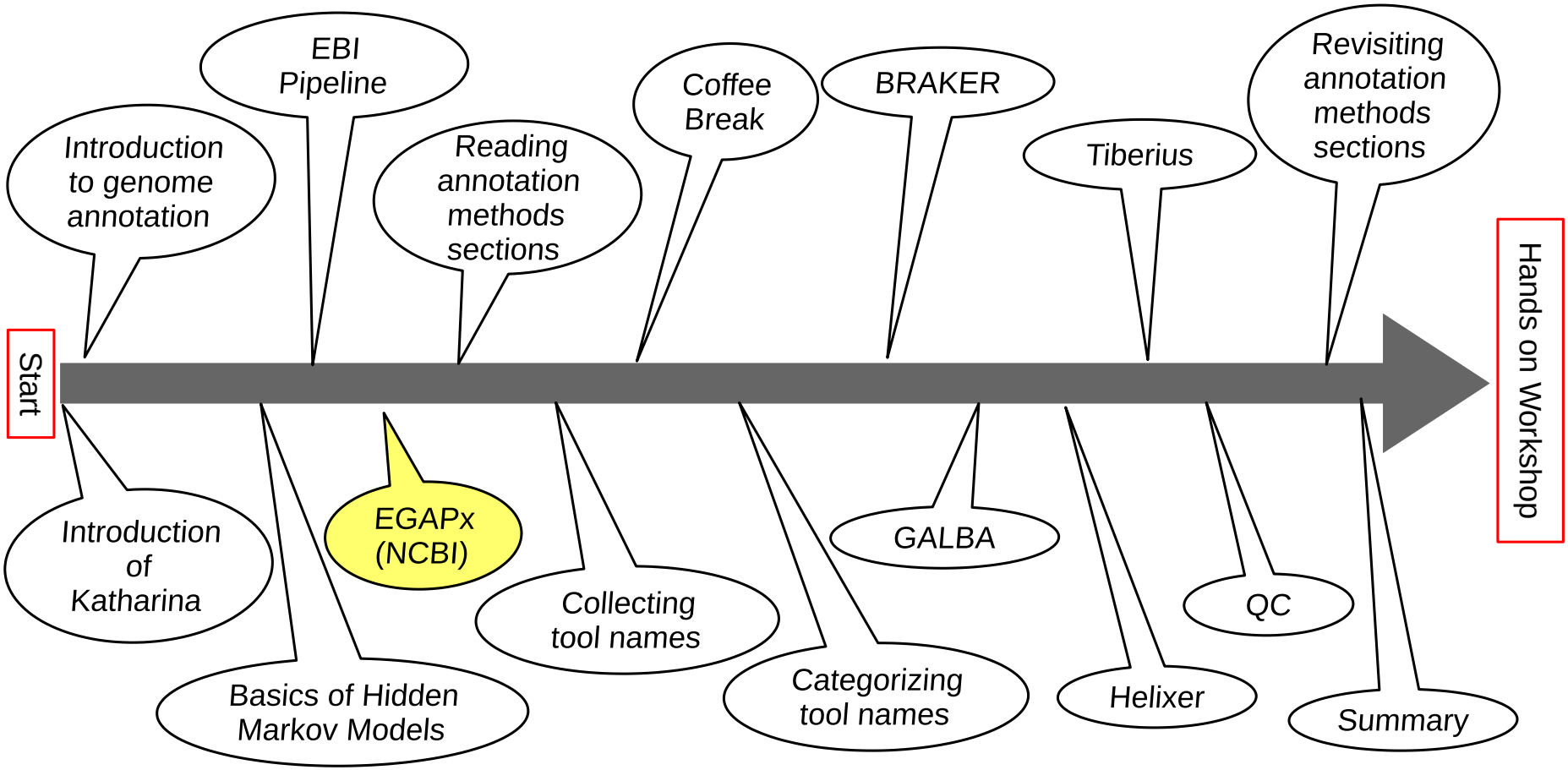
- Ensembl vertebrate pipeline: `https://rapid.ensembl.org/info/genome/genebuild/full_genebuild.html`
- Ensembl non-vertebrate pipeline: `https://rapid.ensembl.org/info/genome/genebuild/anno.html`
- BRAKER2 in Ensembl: `https://rapid.ensembl.org/info/genome/genebuild/braker.html`

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
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

Genome Annotation at the NCBI

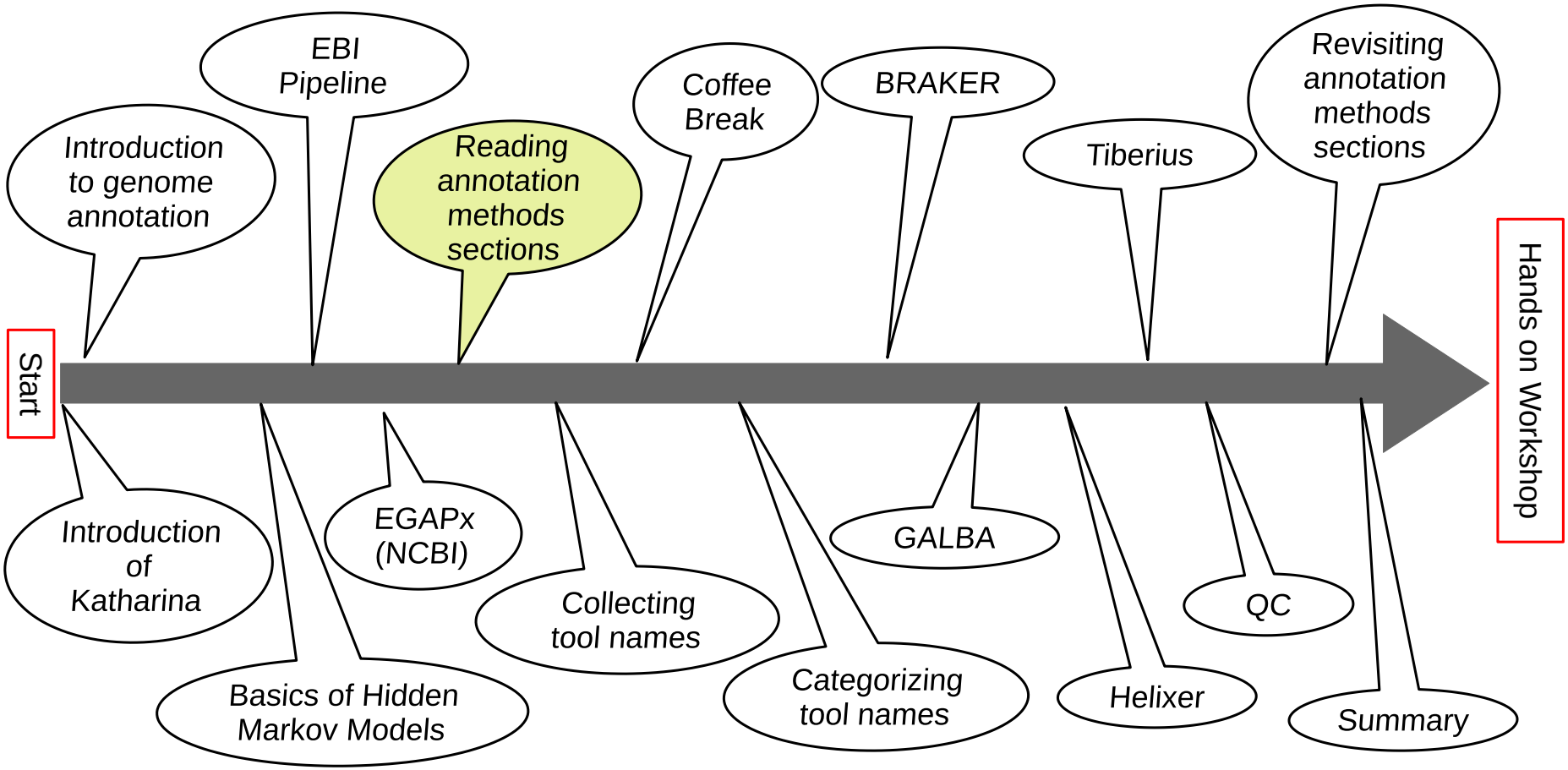
Different Annotation Scenarios

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

Annotation with EGAPx (NCBI)

Annotation with EGAPx (NCBI)

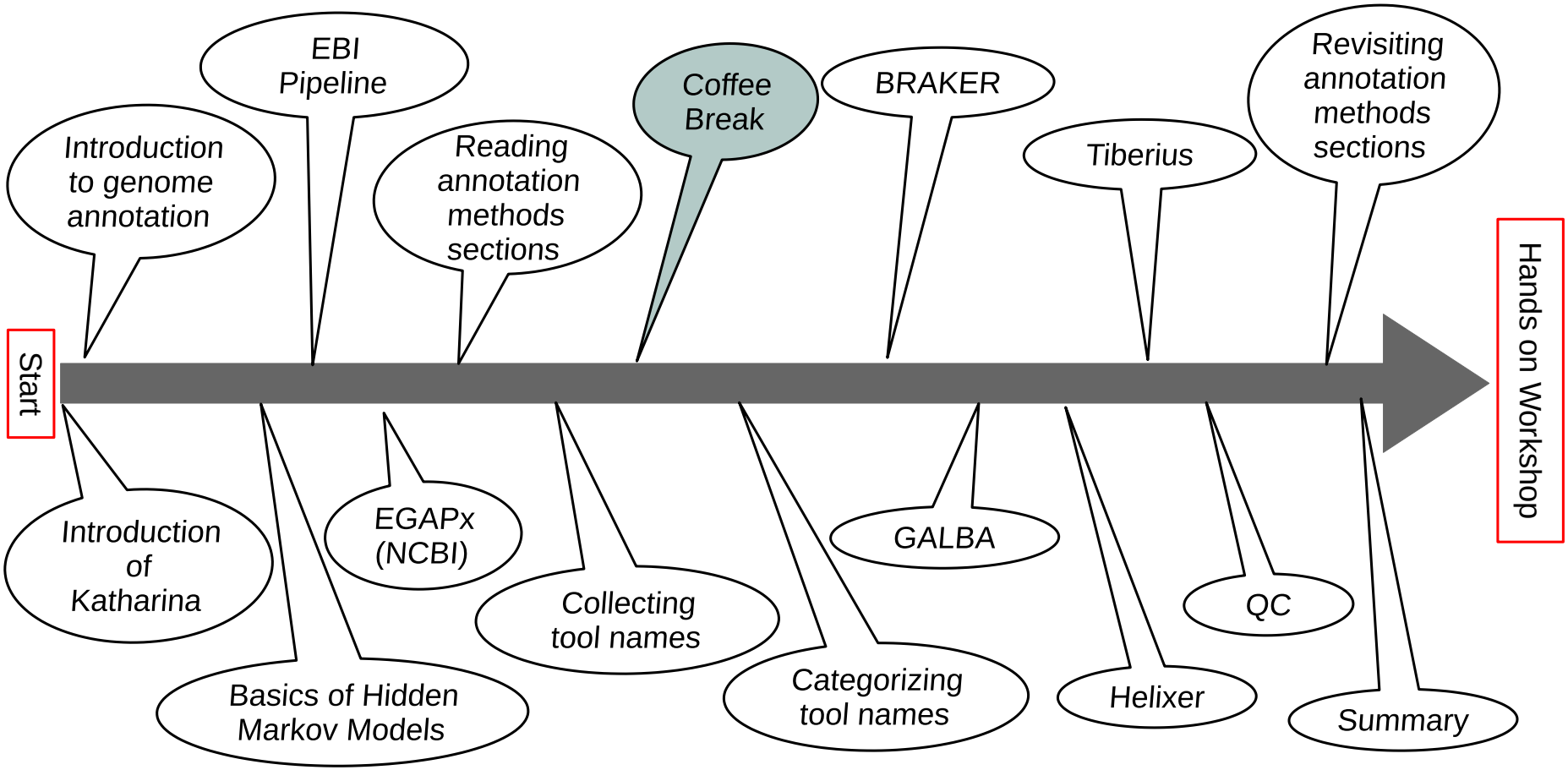
- Containerized with Docker/Singularity
- Documentation: `https://github.com/ncbi/egapx`
- Currently supported clades (protein sets):
  - Chordata
  - Insecta
  - Arthropoda
  - Monocots
  - Eudicots
- Easy to use
- Benchmarking possible: good accuracy!

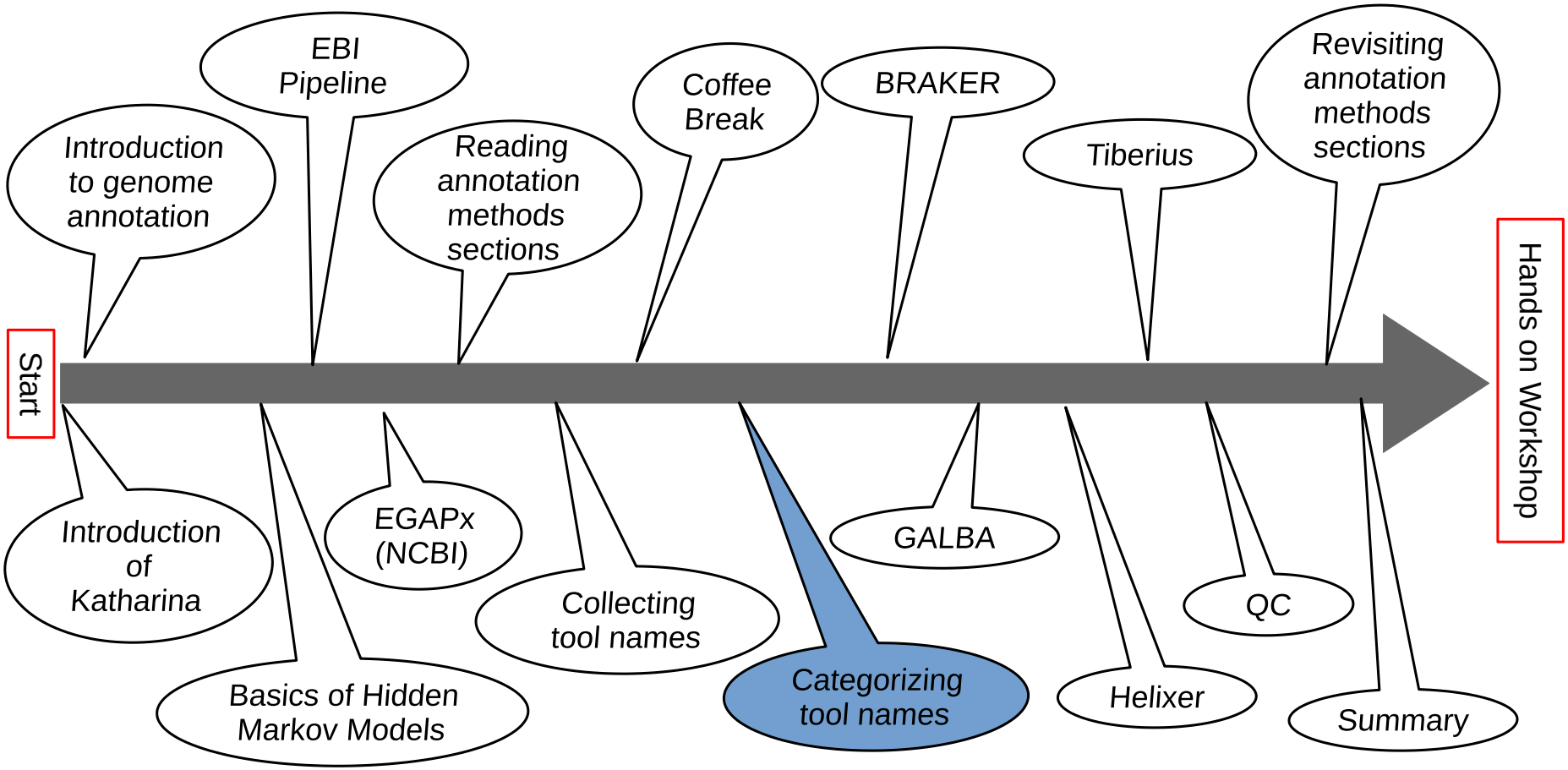

Start
Introduction to genome annotation
EBI Pipeline
Reading annotation methods sections
Coffee Break
BRAKER
Tiberius
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EGAPx (NCBI)
Collecting tool names
Basics of Hidden Markov Models
Categorizing tool names
GALBA
Helixer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

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


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


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

Categorize tool names

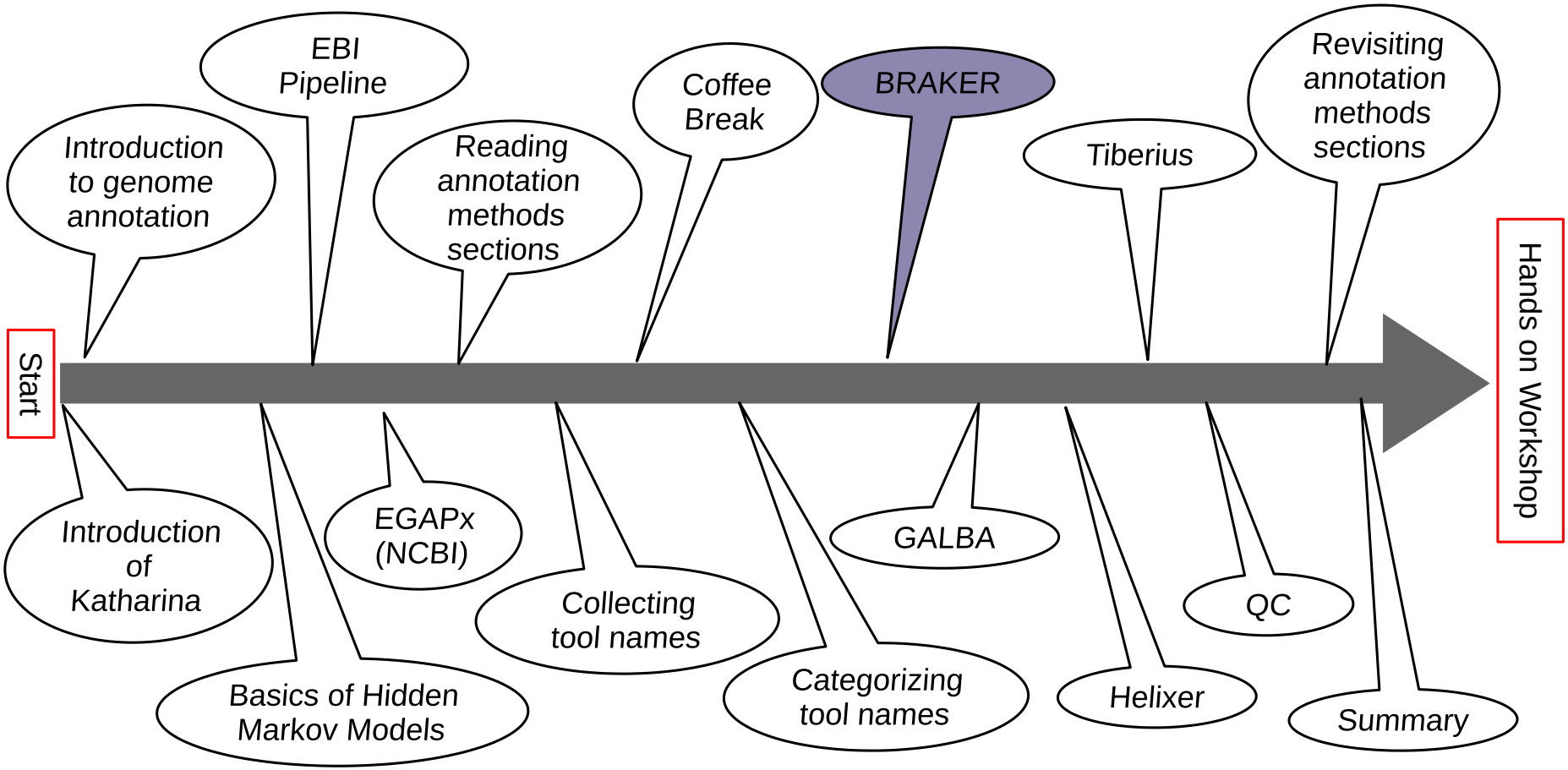
Go to

`https://shorturl.at/uA0Tg`

and sort the tools names from your methods snippet into categories

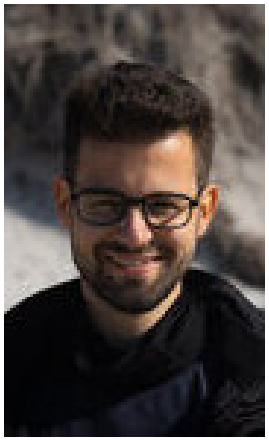
The aera of genome annotation super heroes

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

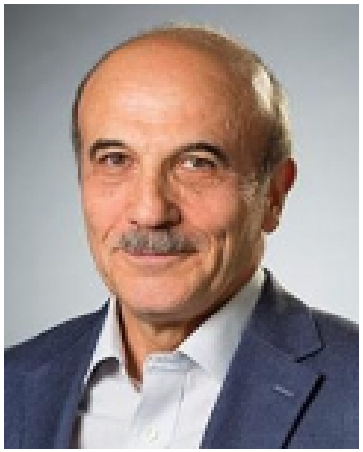
University of Greifswald & Georgia Tech University



Lars Gabriel

Alexandre Lomsadze, Katharina Hoff, Tomáš Brůna

Mario Stanke



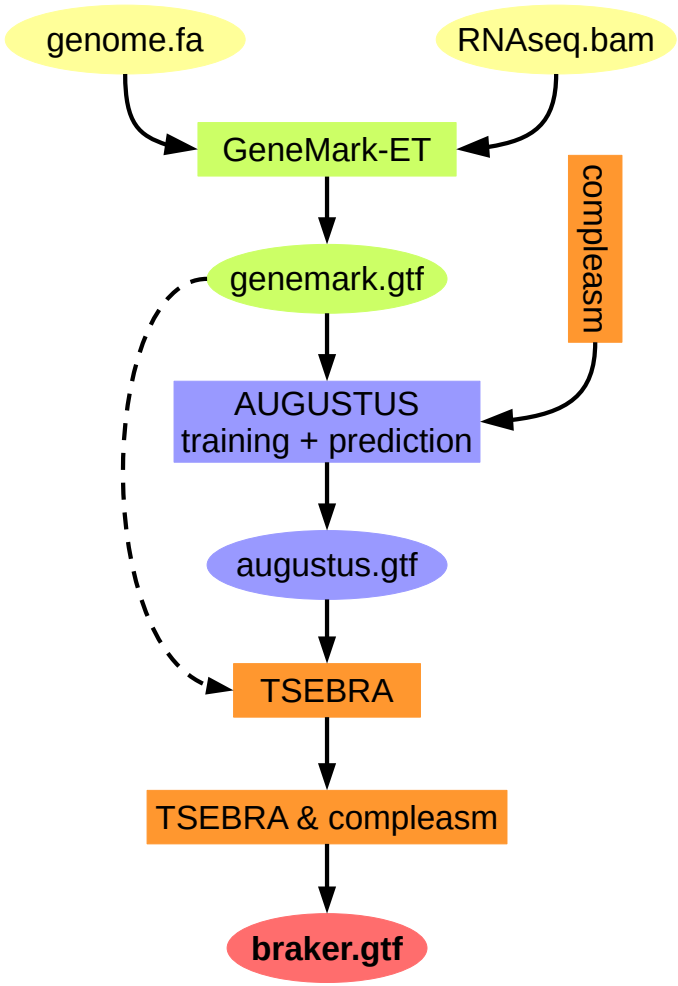
Mark Borodovsky

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
- 2,174 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

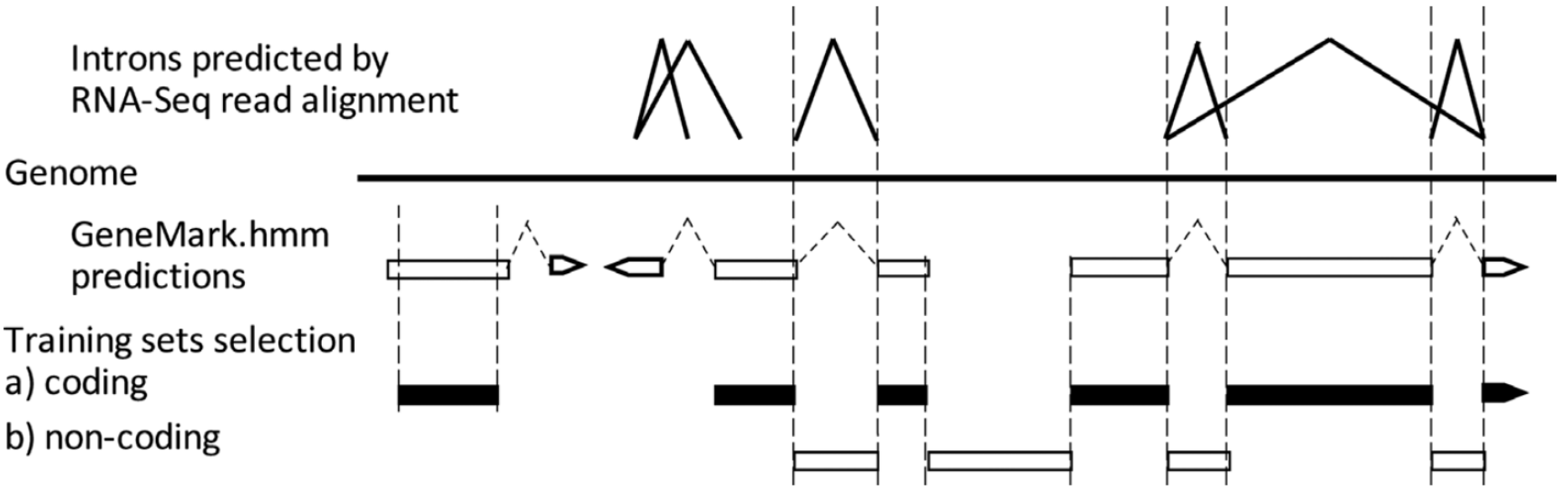
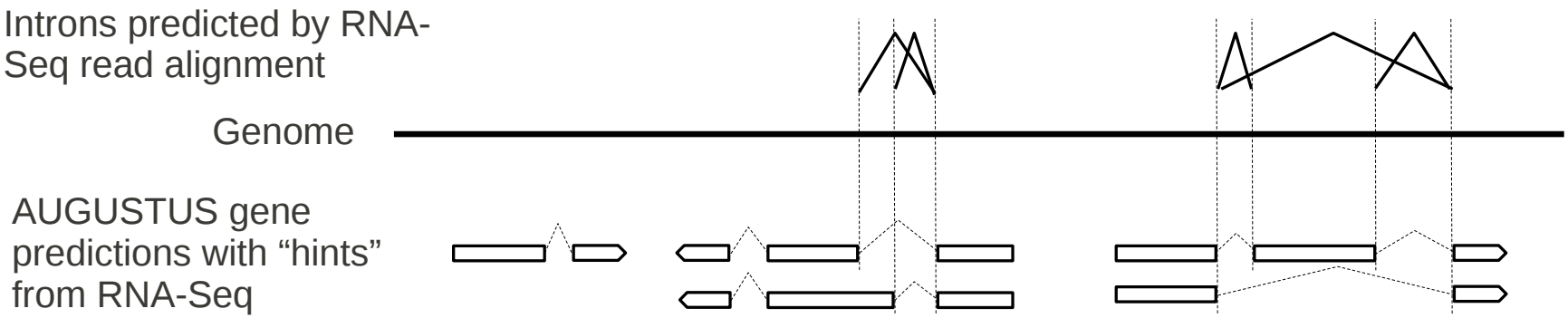


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- 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**



- 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

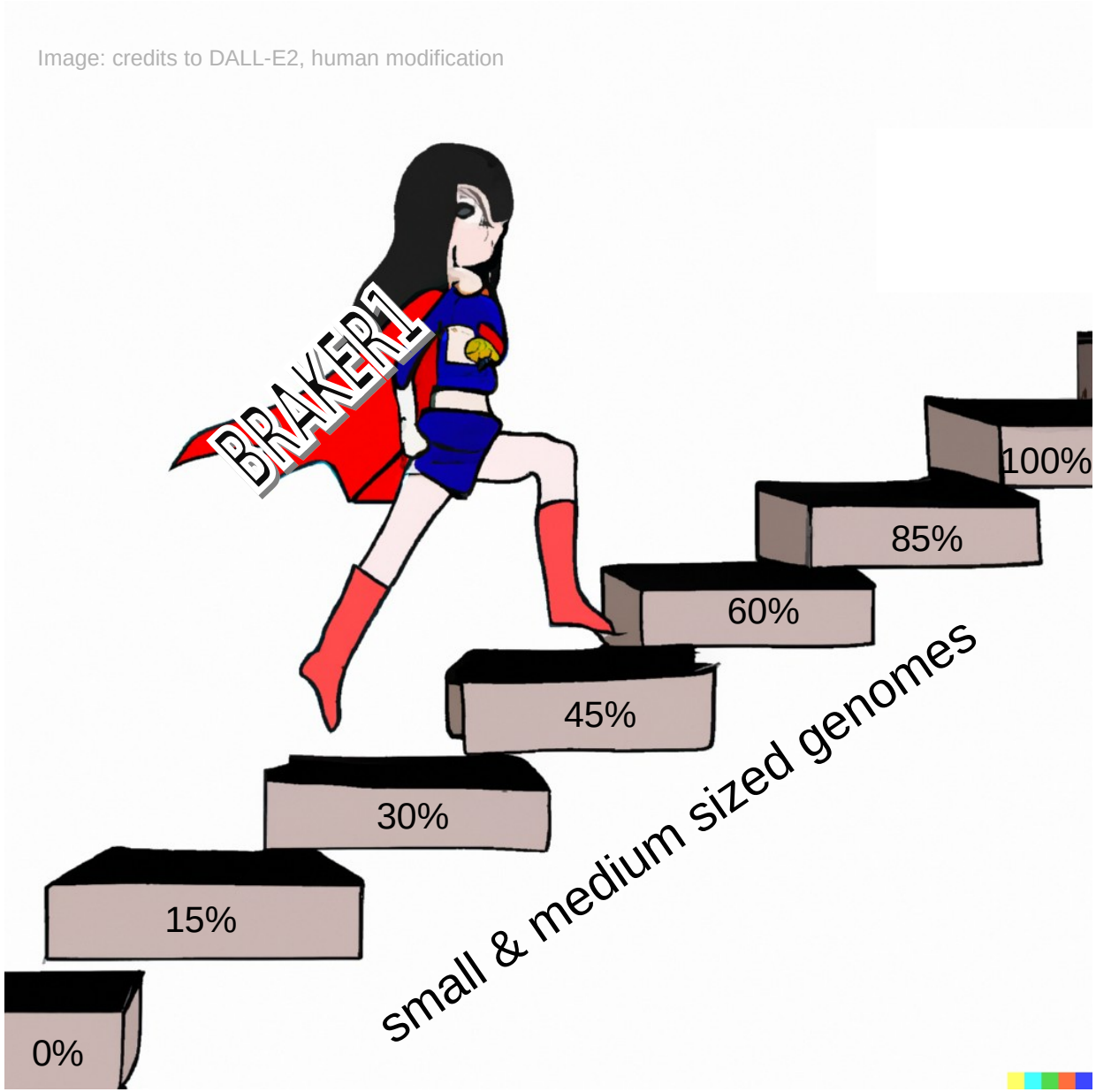
Accuracy metrics

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

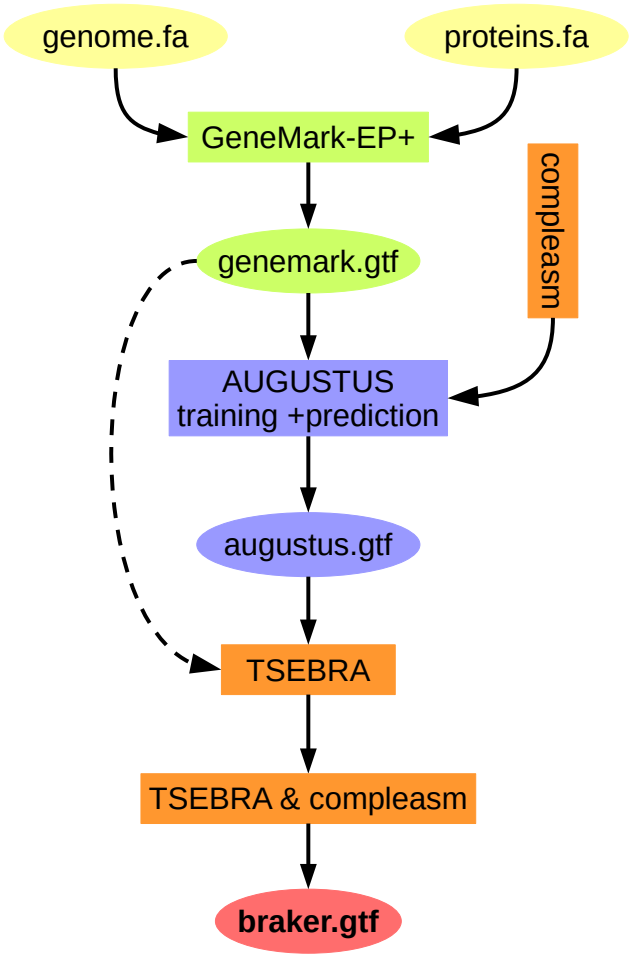


Use only if not enough RNA-Seq for BRAKER3!

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

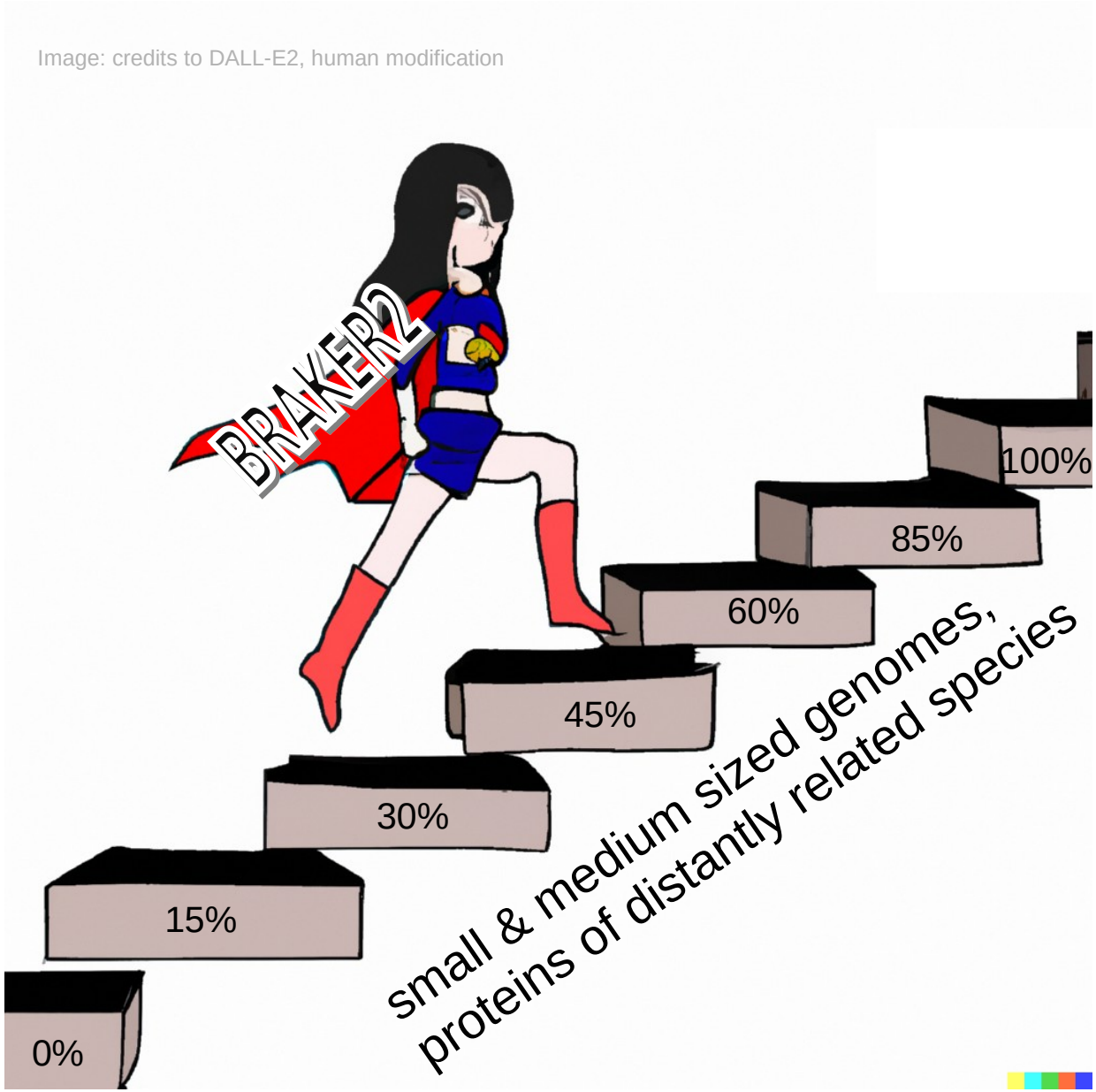
BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database

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)
- 1,480 citations (Google Scholar)

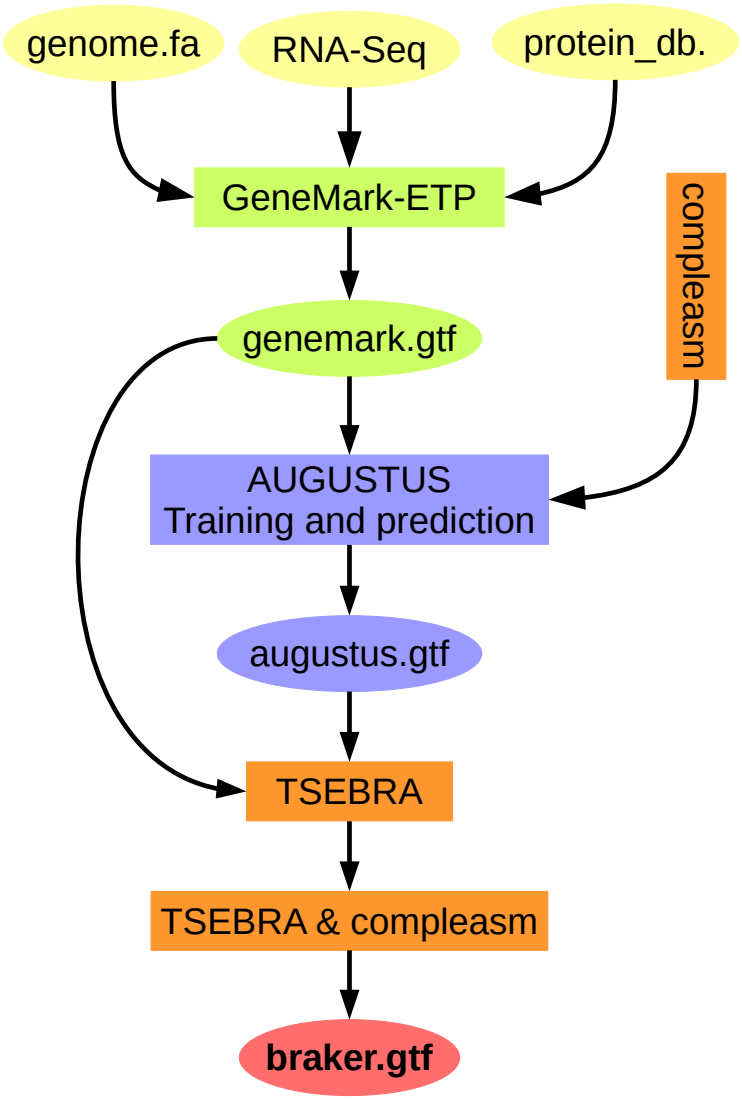
BRAKER2 gene F1 accuracy



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

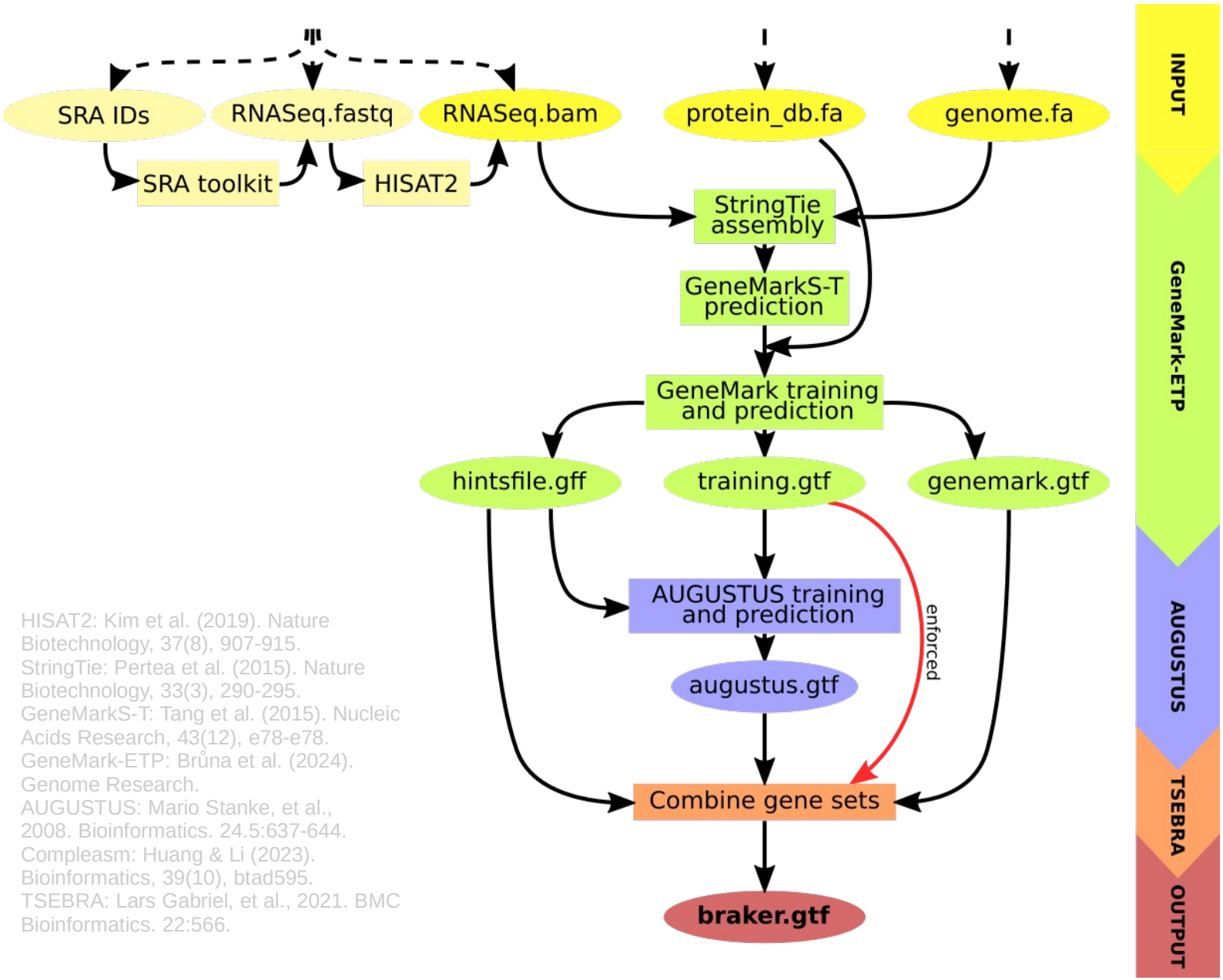
BRAKER3 gene F1 accuracy - climbing the top

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



- Gabriel *et al.* (2024)
- 249 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

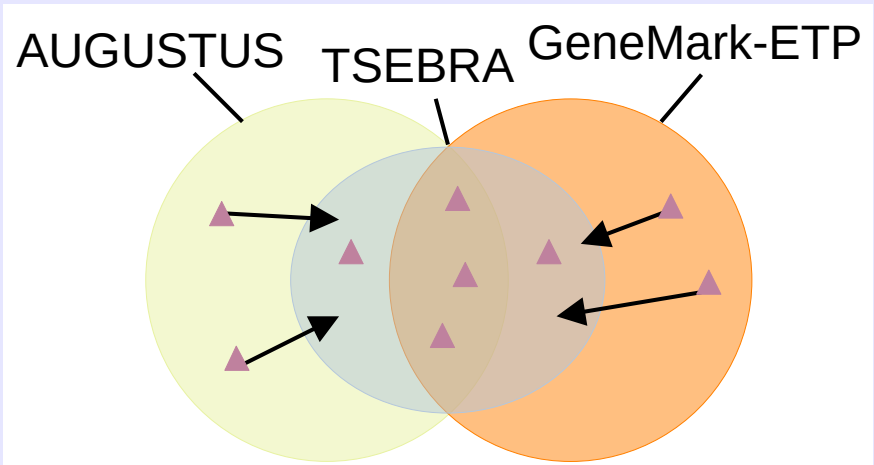
SOFTWARE Open Access

TSEBRA: transcript selector for BRAKER

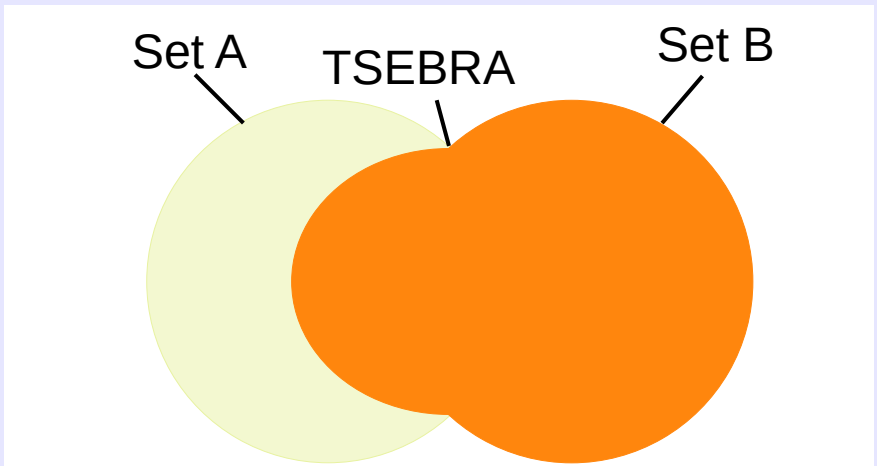
Lars Gabriel[1,2], Katharina J. Hoff[1,2], Tomáš Brůna[3], Mark Borodovsky[4,5] and Mario Stanke[1,2*]

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

TSEBRA in BRAKER



Manually enforcing a gene set



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

Accuracy of genome annotation approaches by BRAKER team

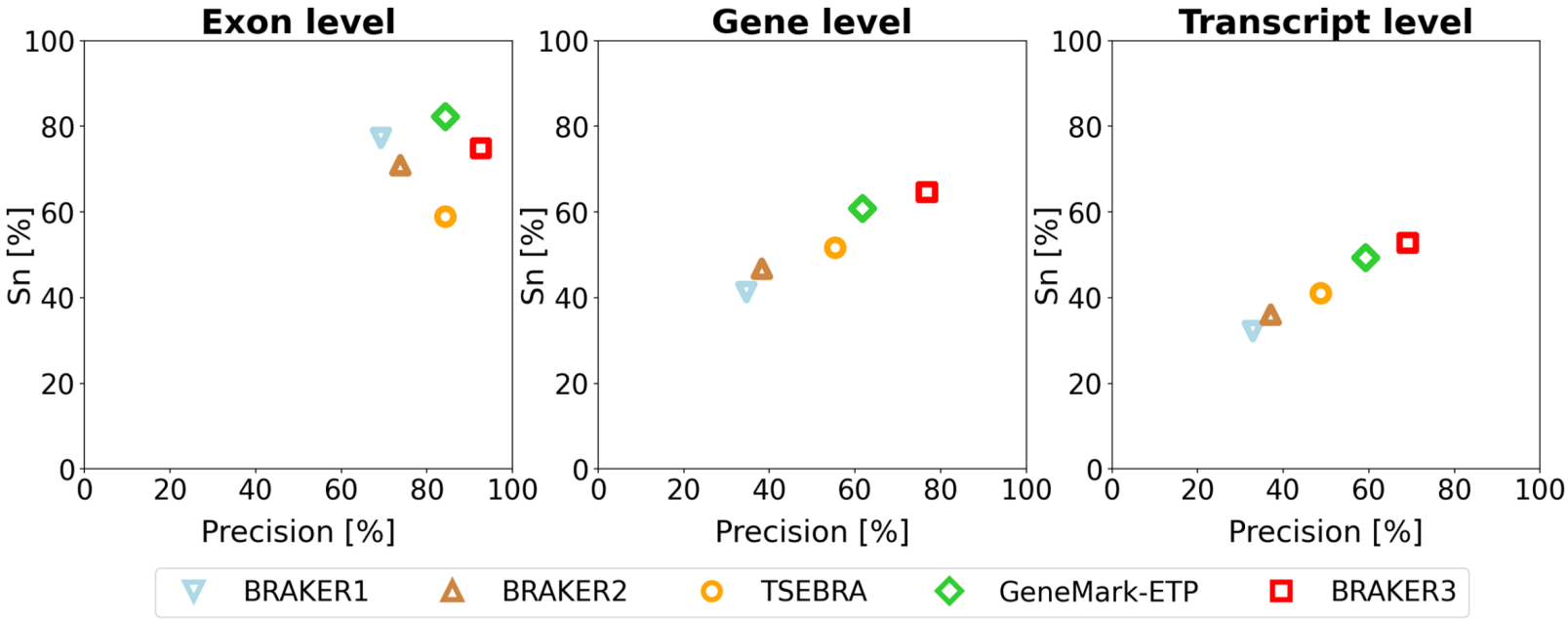


Figure 2. Average precision and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in Supplemental Table S1). Inputs were the genomic sequences, short-read RNA-seq libraries, and protein databases (*order excluded*).

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

Availability

GitHub

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

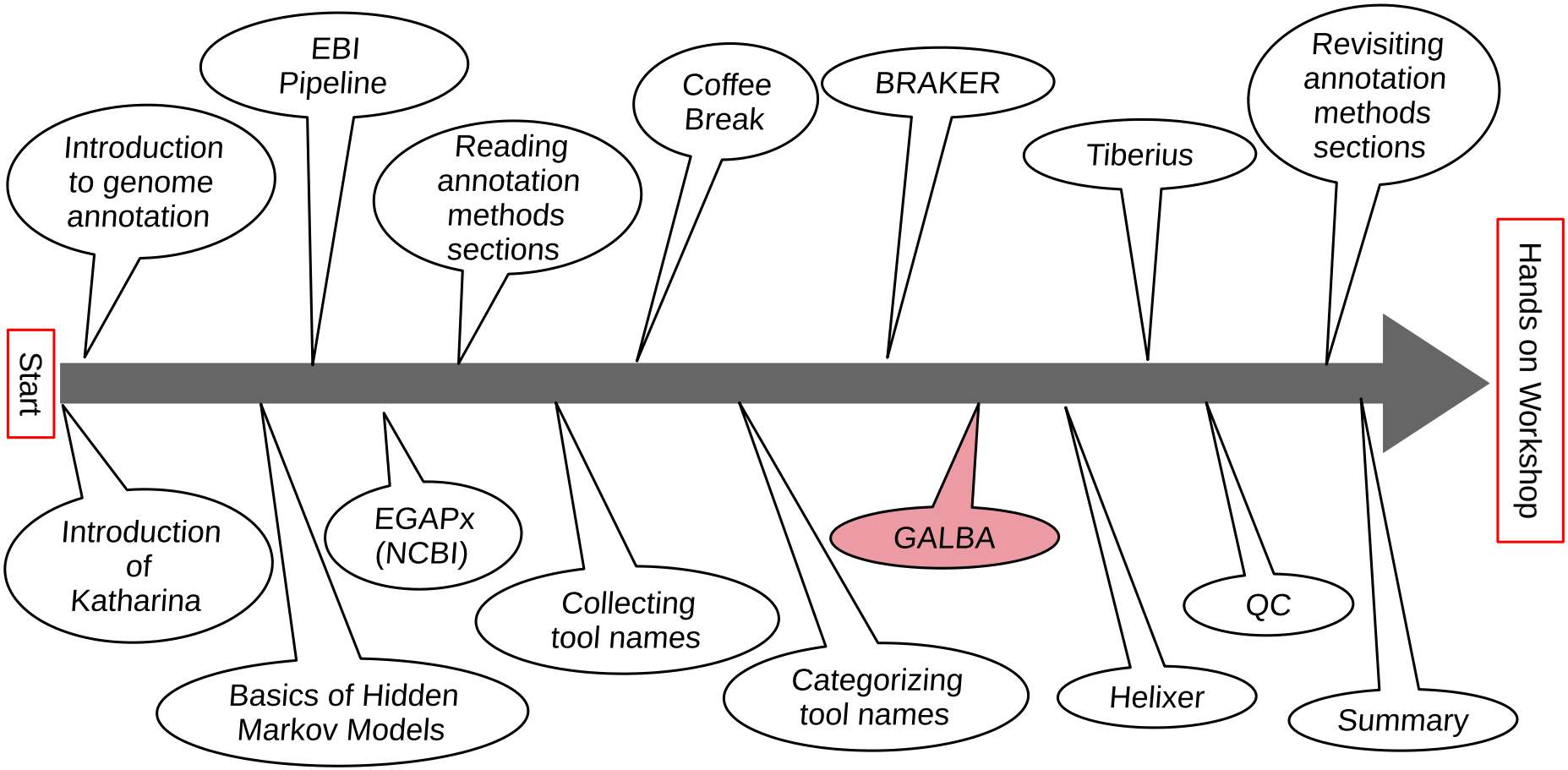
Docker/Singularity

```
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

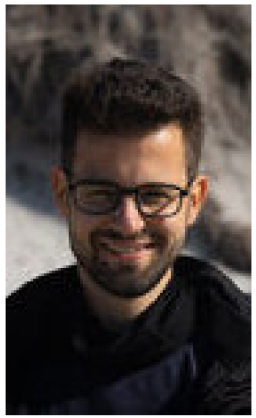

Start
Introduction to genome annotation
EBI Pipeline
Reading annotation methods sections
Coffee Break
BRAKER
Tiberius
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EGAPx (NCBI)
Collecting tool names
GALBA
QC
Basics of Hidden Markov Models
Categorizing tool names
Helixer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

GALBA Contributors

Tomáš Brůna

Heng Li

Joseph Guhlin



Lars Gabriel

Natalia Nenasheva

Ethan Tolman

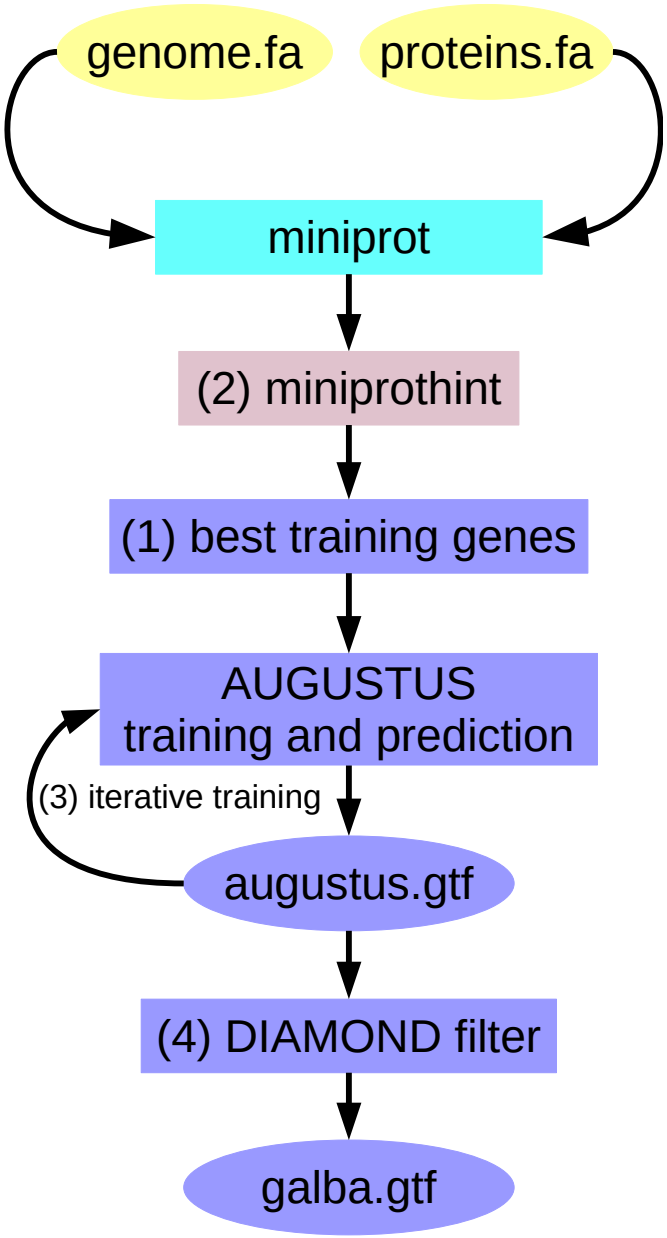
Paul Frandsen

Matthis Ebel

Mario Stanke

Katharina Hoff

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

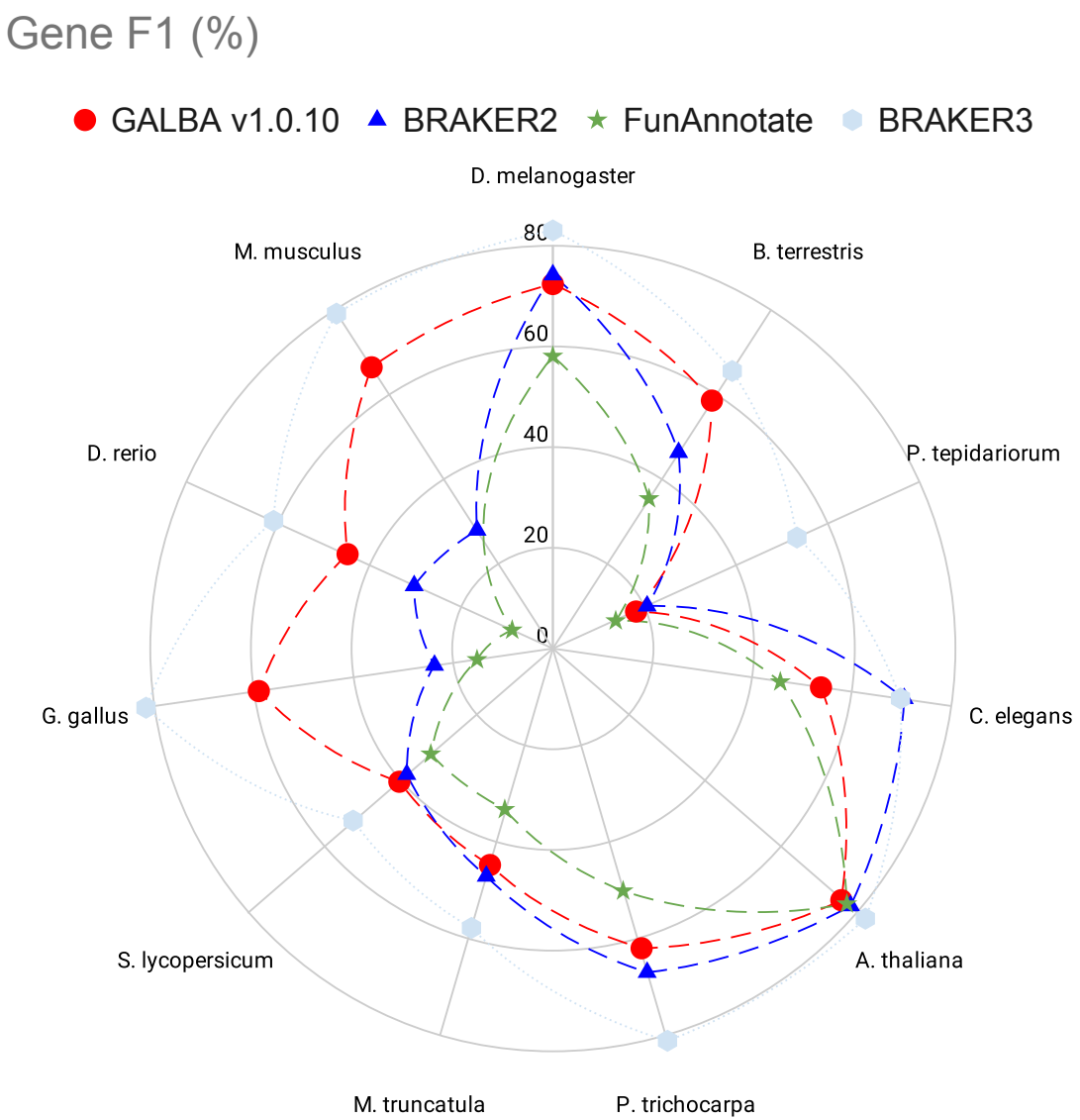


RESEARCH Open Access

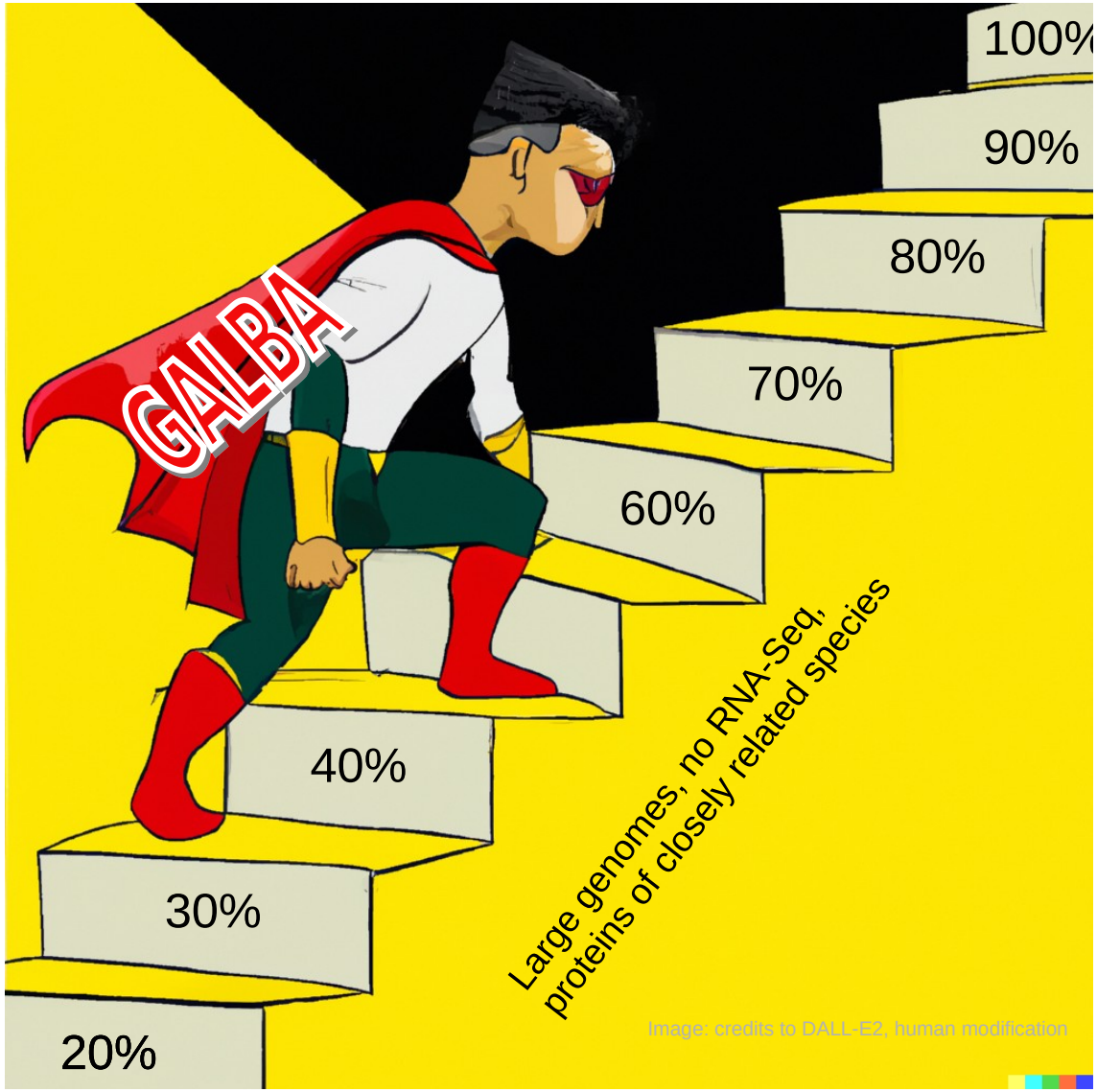
Galba: genome annotation with miniprot and AUGUSTUS

- 42 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



If you have RNA-Seq, use BRAKER3!

Availability

GitHub

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

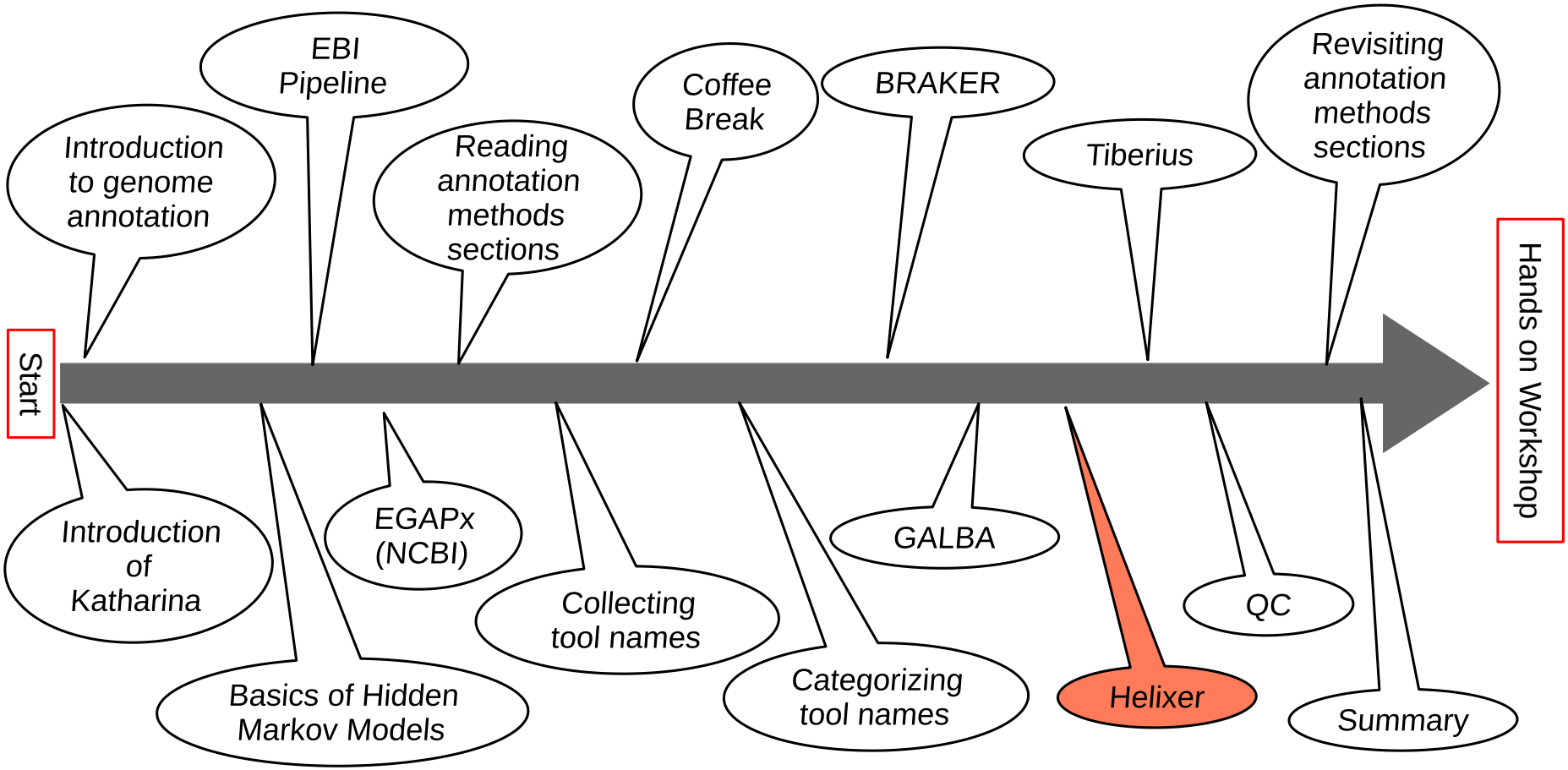
Docker/Singularity

```
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

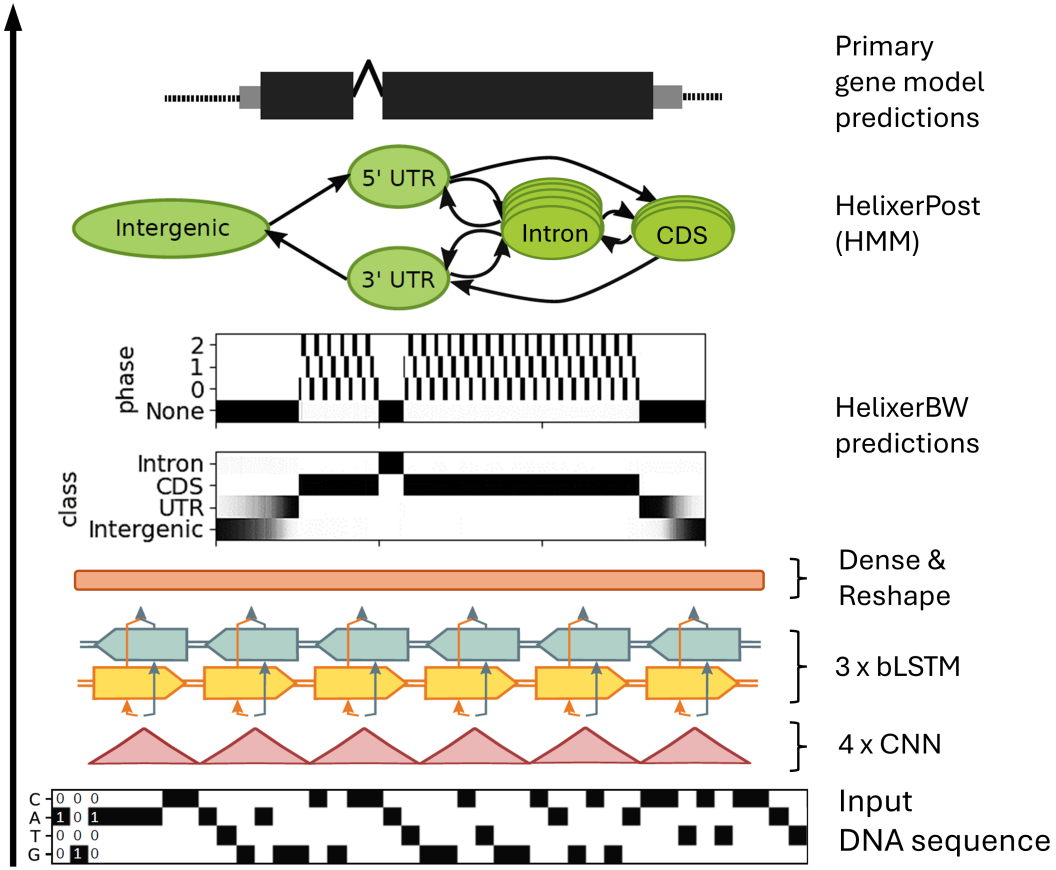

Start
Introduction to genome annotation
EBI Pipeline
Reading annotation methods sections
Coffee Break
BRAKER
Tiberius
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EGAPx (NCBI)
Collecting tool names
Basics of Hidden Markov Models
Categorizing tool names
GALBA
Helixer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

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

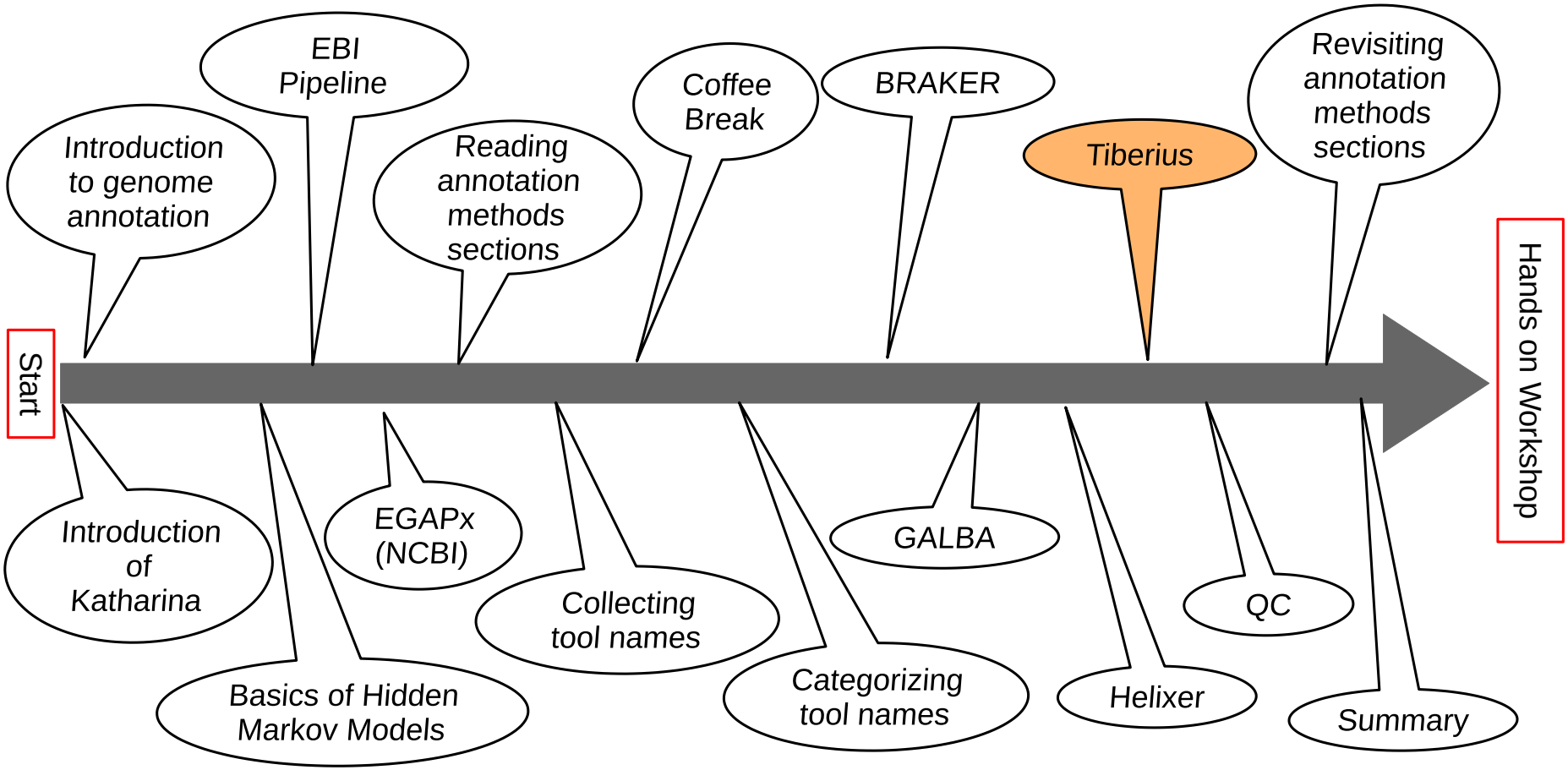
Felix Holst[1†], Anthony Bolger[2†], Christopher Günther[1], Janina Maß[3], Sebastian Triesch[1,4], Felicitas Kindel[1], Niklas Kiel[1,4], Nima Saadat[3,4], Oliver Ebenhöh[3,4], Björn Usadel[2,4,5], Rainer Schwacke[2], Marie Bolger[2], Andreas P.M. Weber[1,4], Alisandra K. Denton[1,4]



- 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
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

Tiberius: improved genome annotation with deep learning

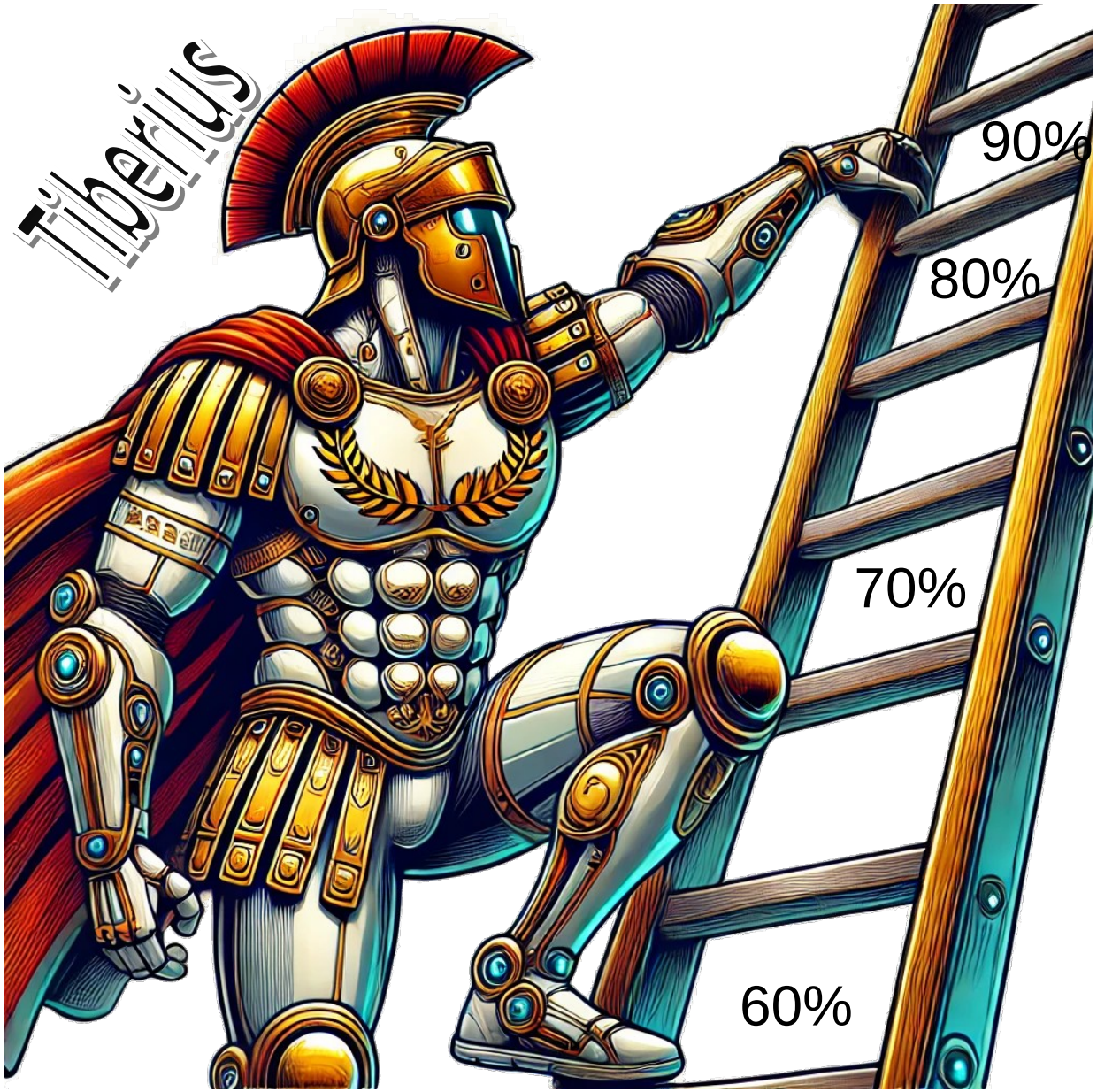


Image: ChatGPT by OpenAI, manual editing

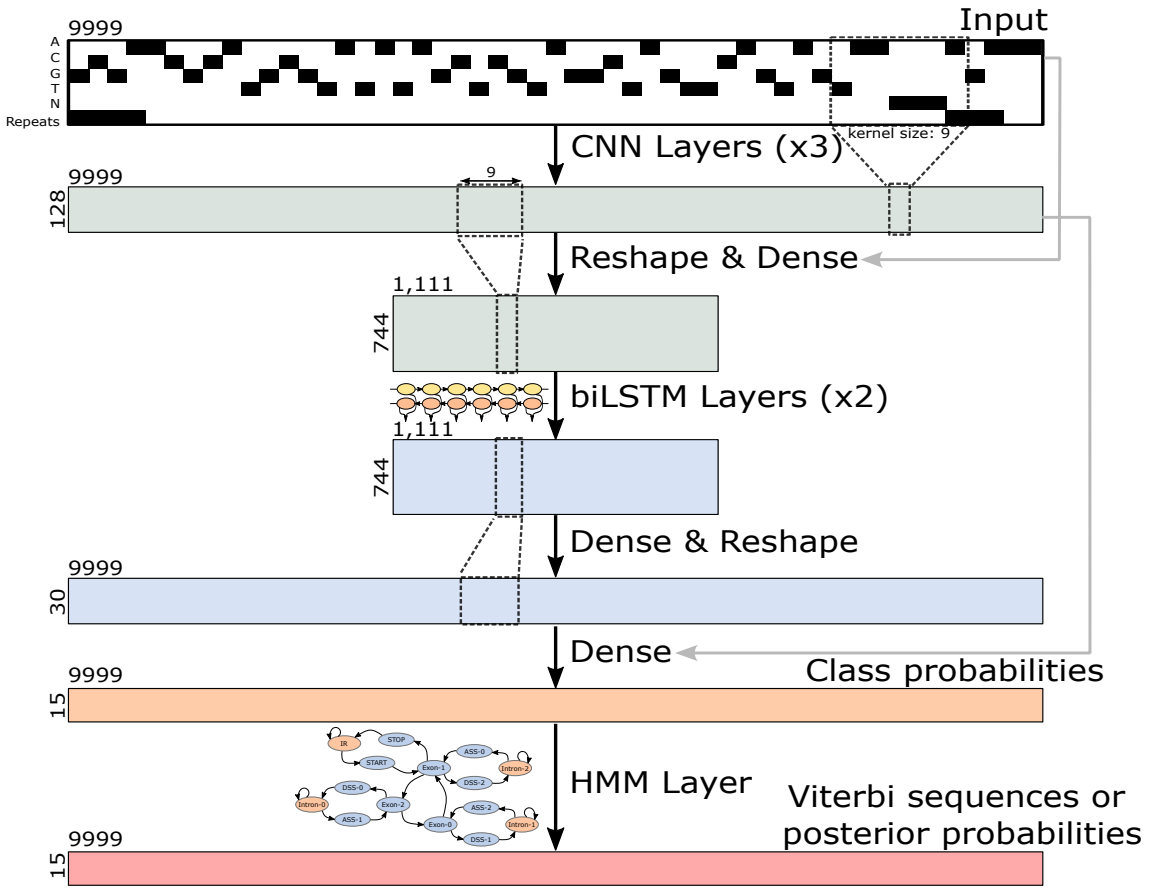
The Tiberius Team

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,*]



- 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

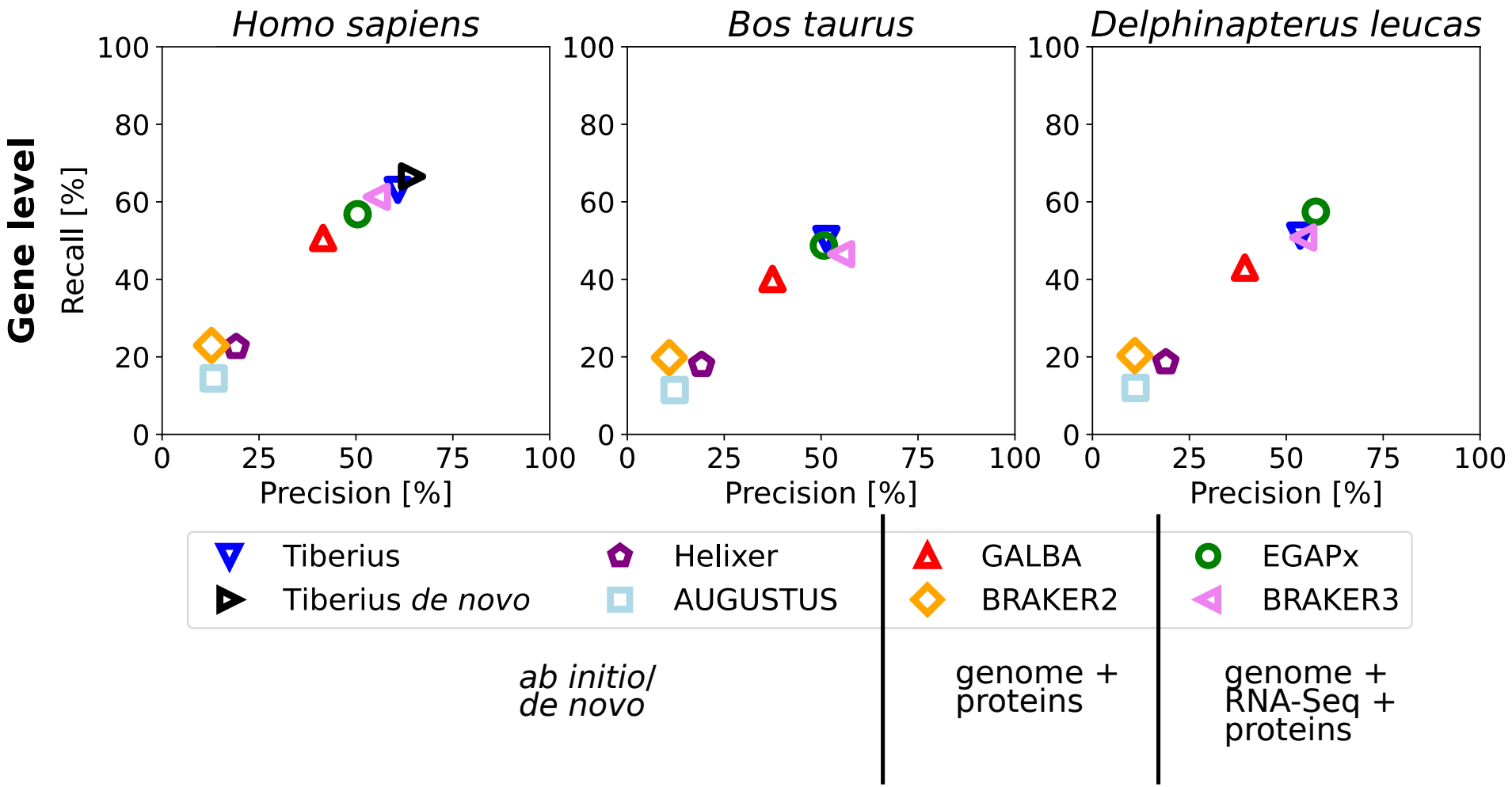
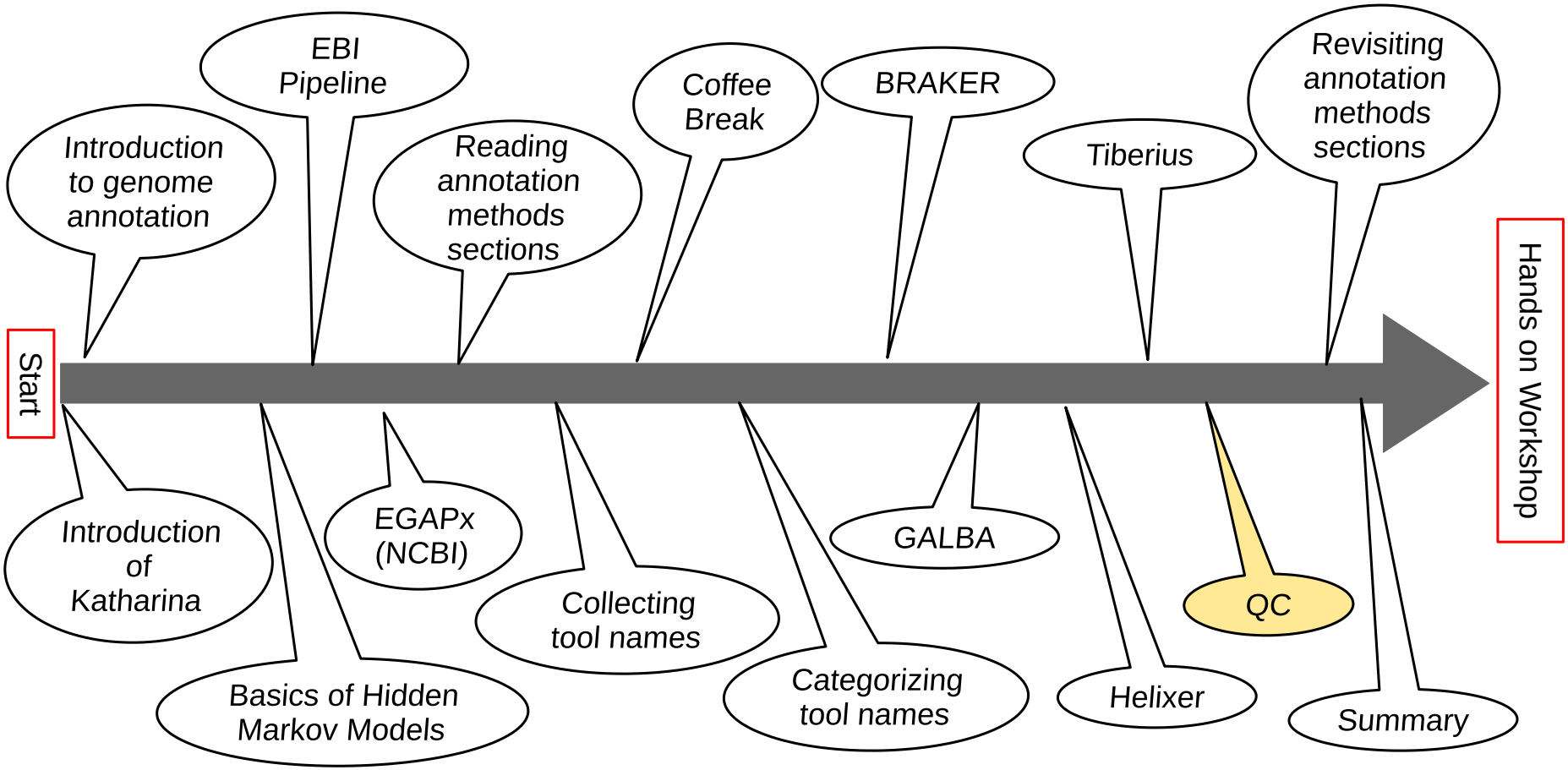


Image: Lars Gabriel, PAG presentation 2025

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

$\rightarrow$ BRAKER3, Galba & EGAPx currently remain important

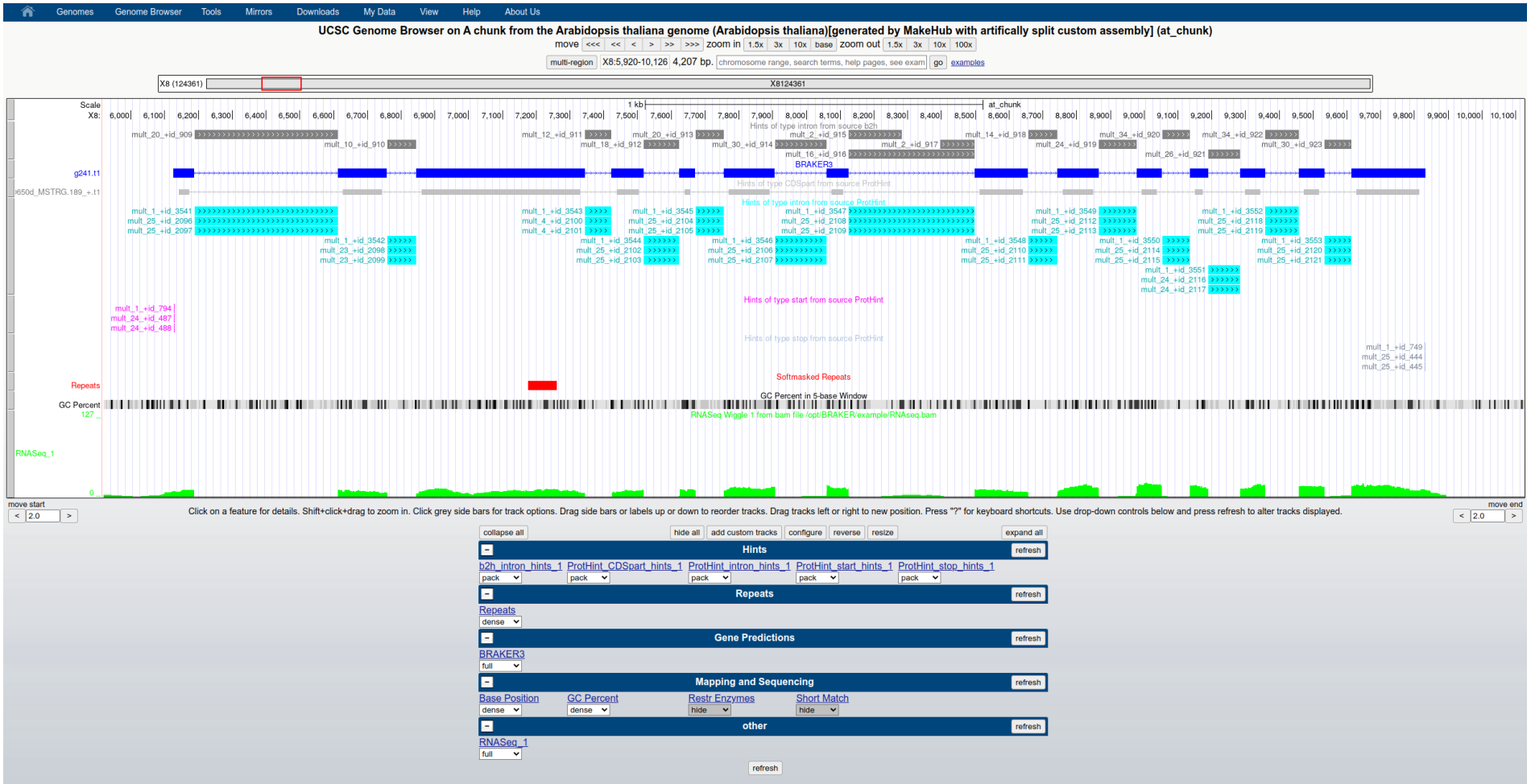

Start
Introduction to genome annotation
EBI Pipeline
Reading annotation methods sections
Coffee Break
BRAKER
Tiberius
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
Basics of Hidden Markov Models
EGAPx (NCBI)
Collecting tool names
Categorizing tool names
GALBA
Helixer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

Did We Do a Good Job?

Genome Browsers

Visualize your Annotation in Context with Evidence

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

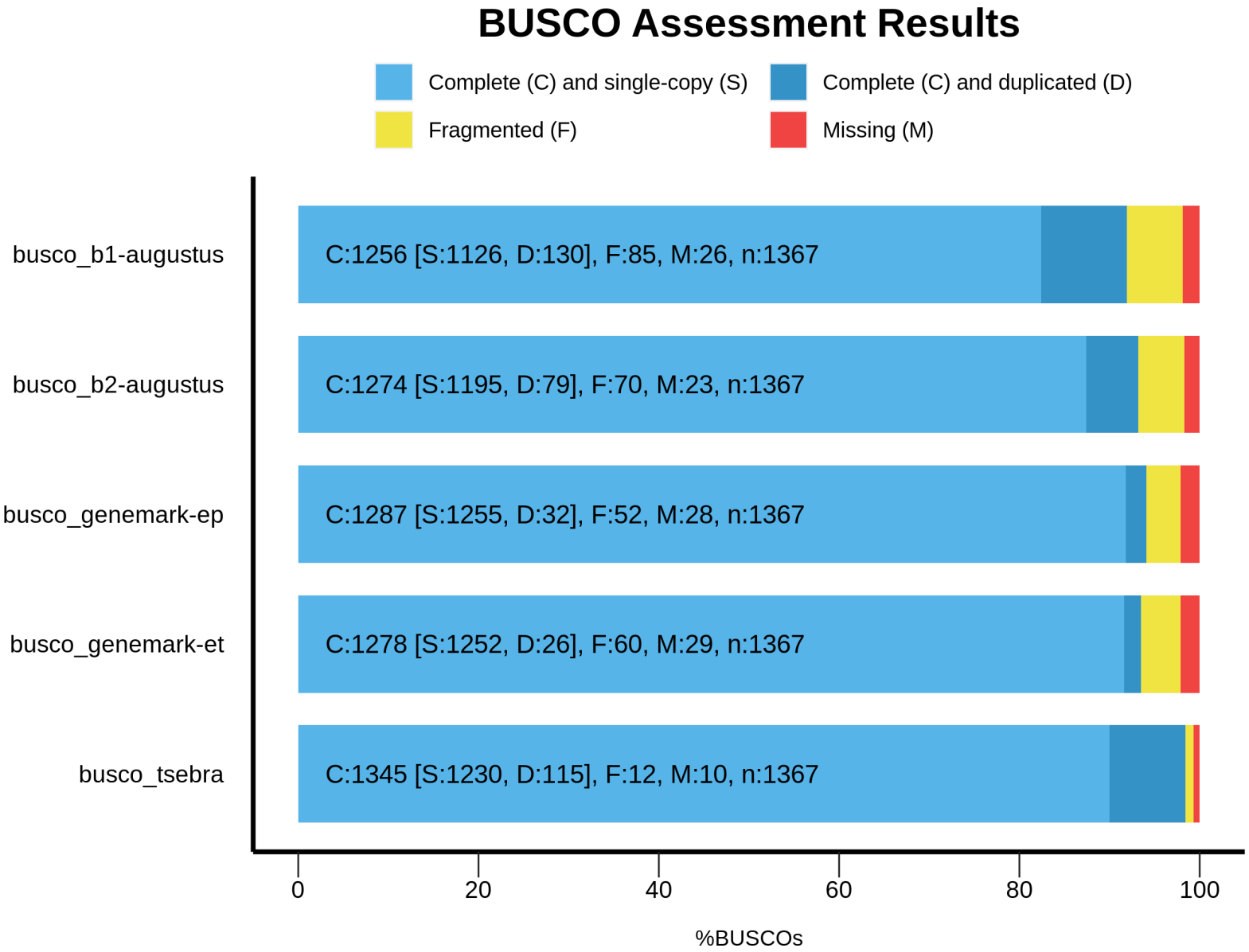


Describe Your Annotation

- number of genes
- number of transcripts
- ratio of mono-exonic to multi-exonic genes
- median number of exons per transcript
- maximal number of exons per transcript
- median transcript length
- ...

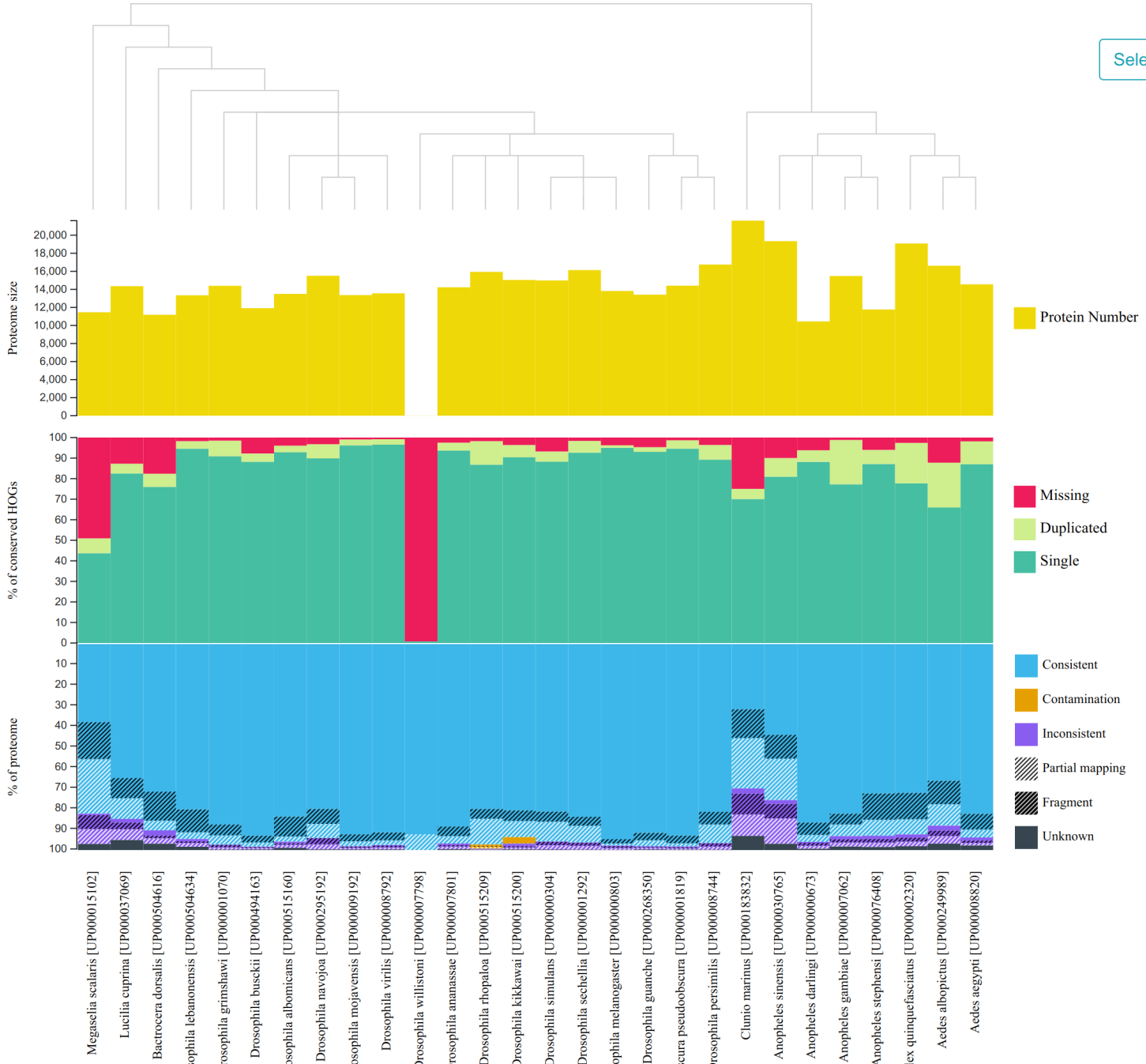
If possible, compare to annotated close relatives.
Consider effect of individual annotation pipelines.

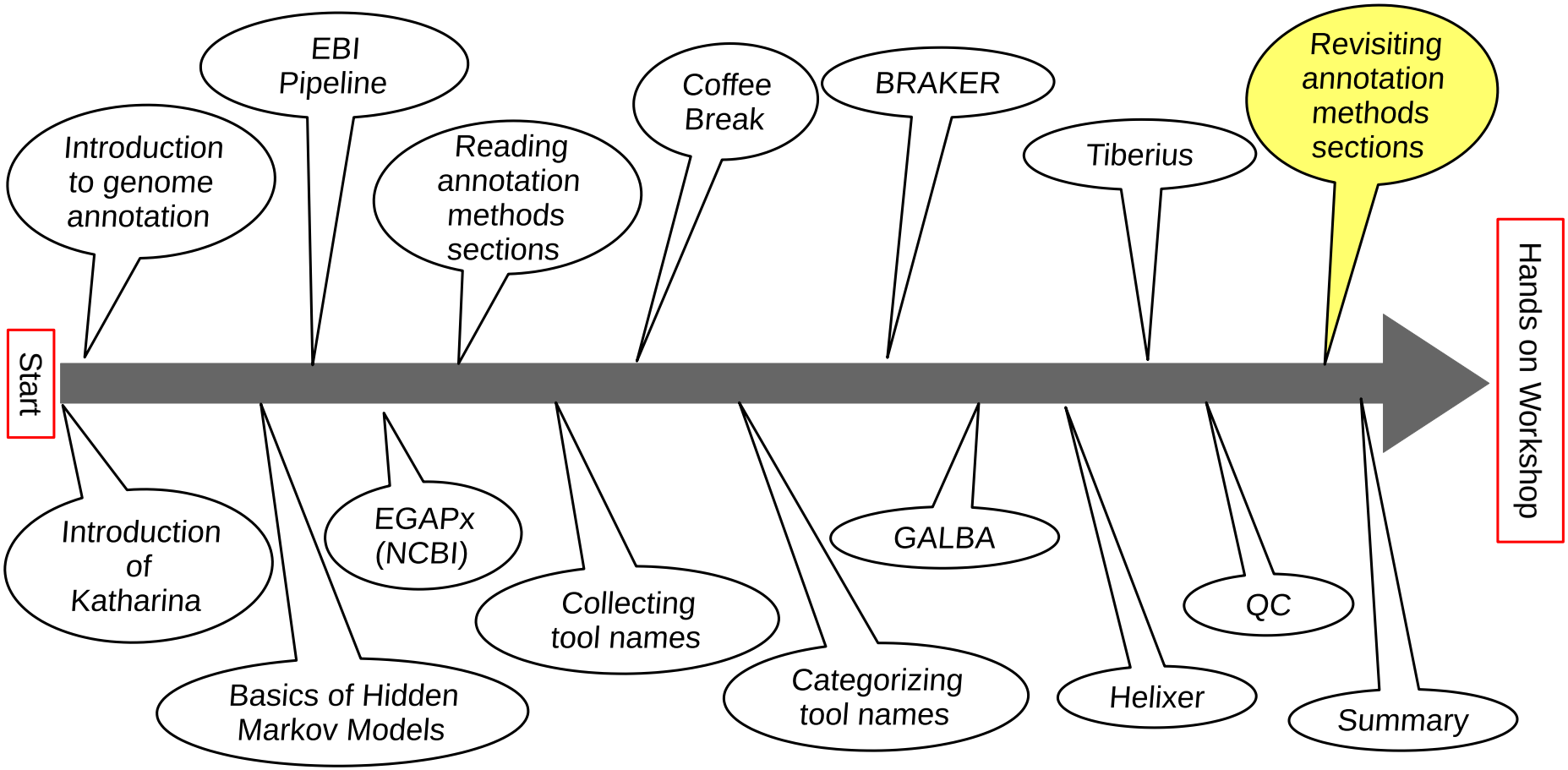
BUSCO: Sensitivity in Clade-Specific Conserved Genes



Beware! BUSCO completeness does not warrant correct gene structures!

OMArk: Sensitivity, Contaminations, & More

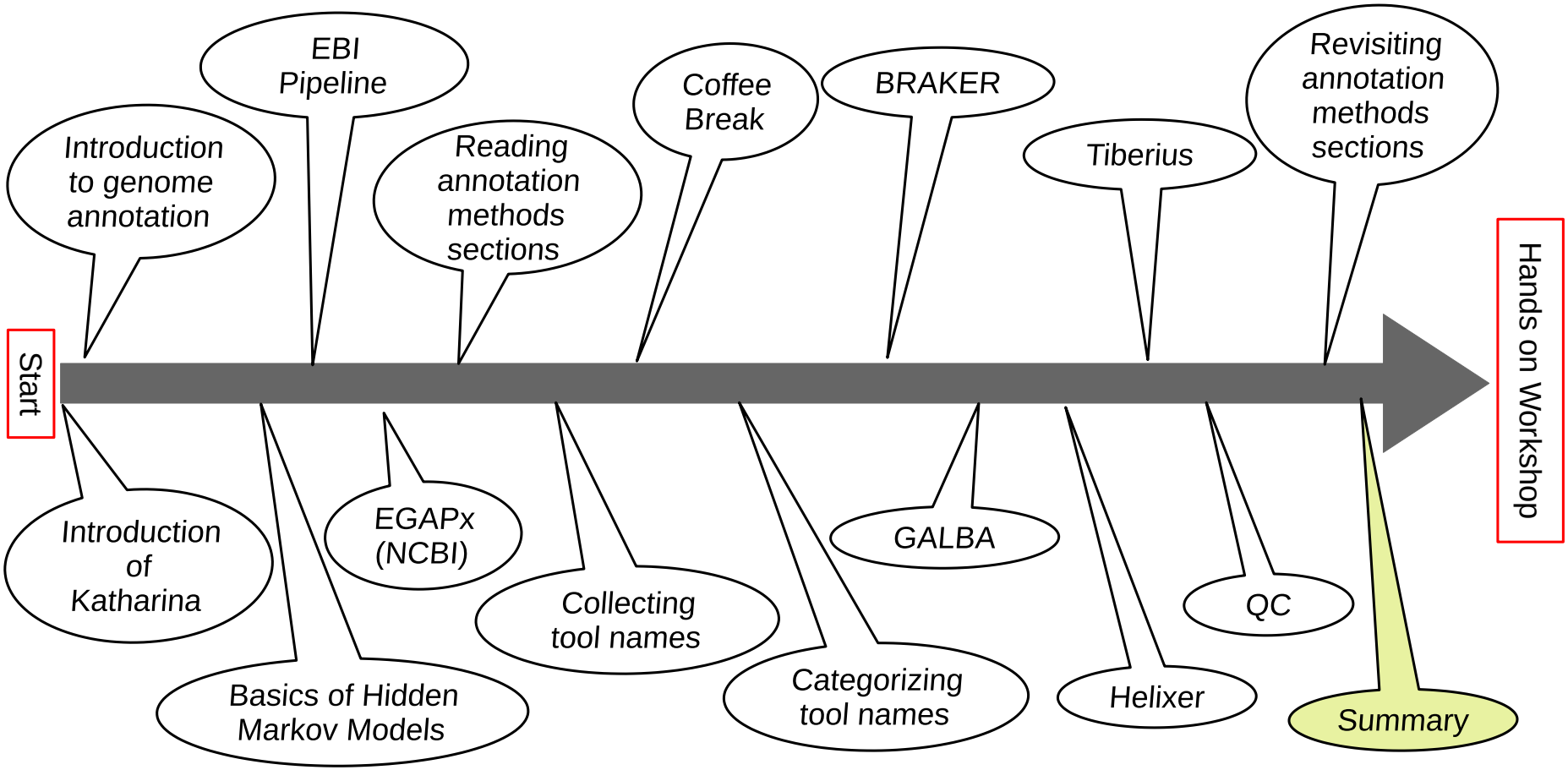



Start
Introduction to genome annotation
EBI Pipeline
Reading annotation methods sections
Coffee Break
BRAKER
Tiberius
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
Basics of Hidden Markov Models
EGAPx (NCBI)
Collecting tool names
Categorizing tool names
GALBA
Helixer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

Revisiting annotation methods sections

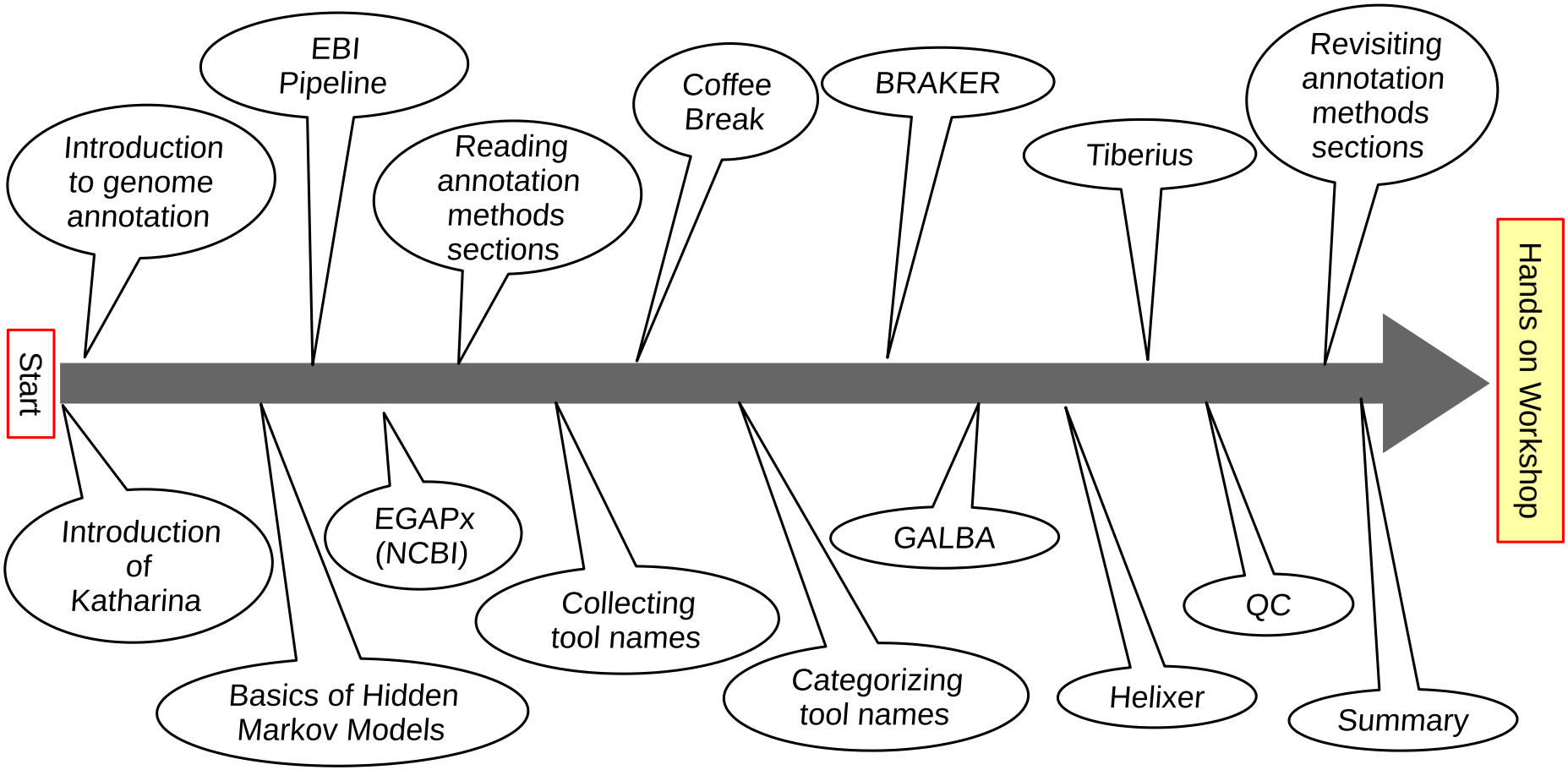
Your tasks

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


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

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