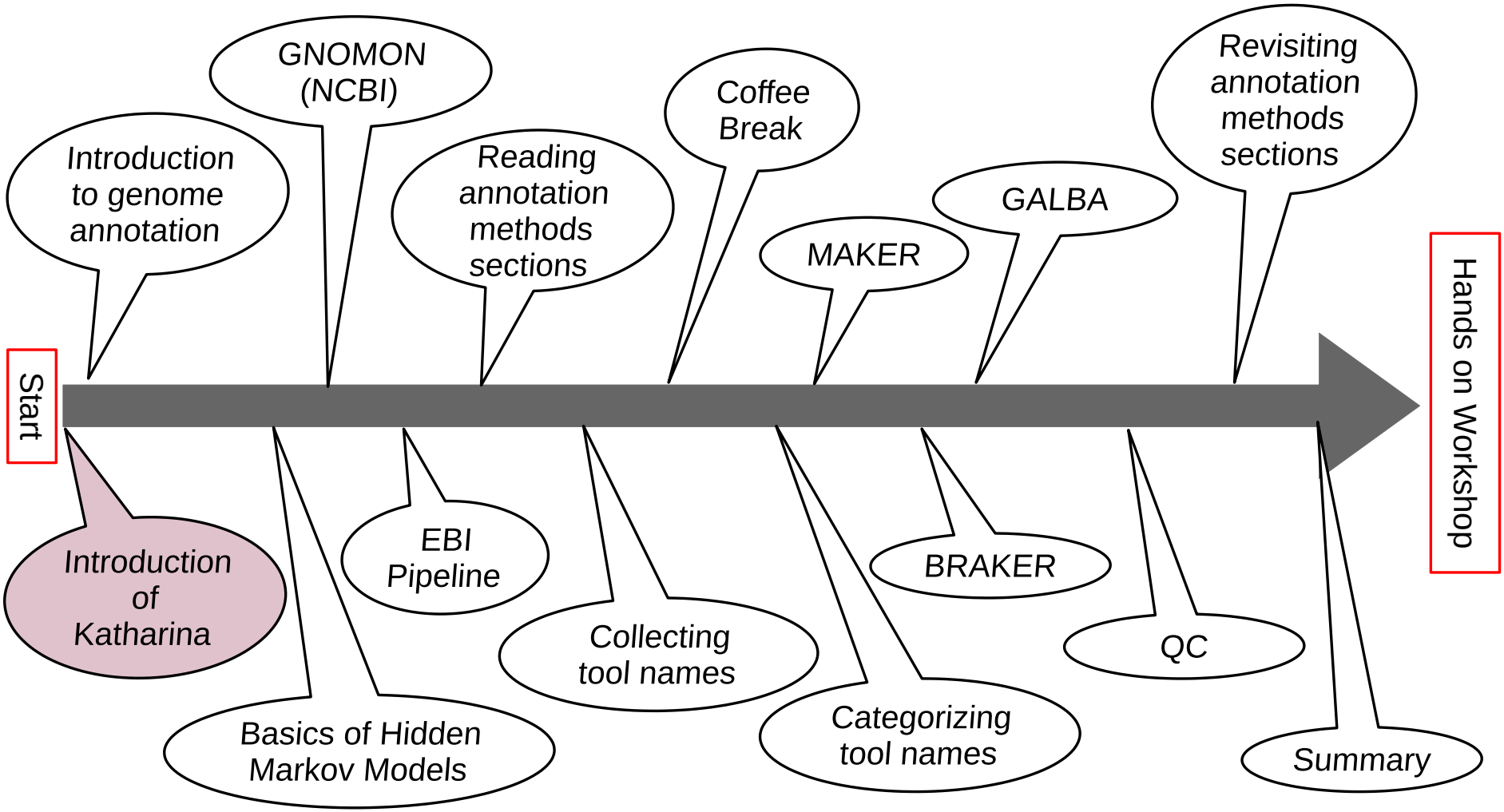
Annotation of Protein Coding Genes

January 9th 2024

Katharina J. Hoff

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

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


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

Katharina J. Hoff
Group Leader in Applied Bioinformatics at University of Greifswald

Short CV

2022 Habilitation (Greifswald)

2009 Ph.D. Molecular Biology (Göttingen)

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

Research

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

Teaching

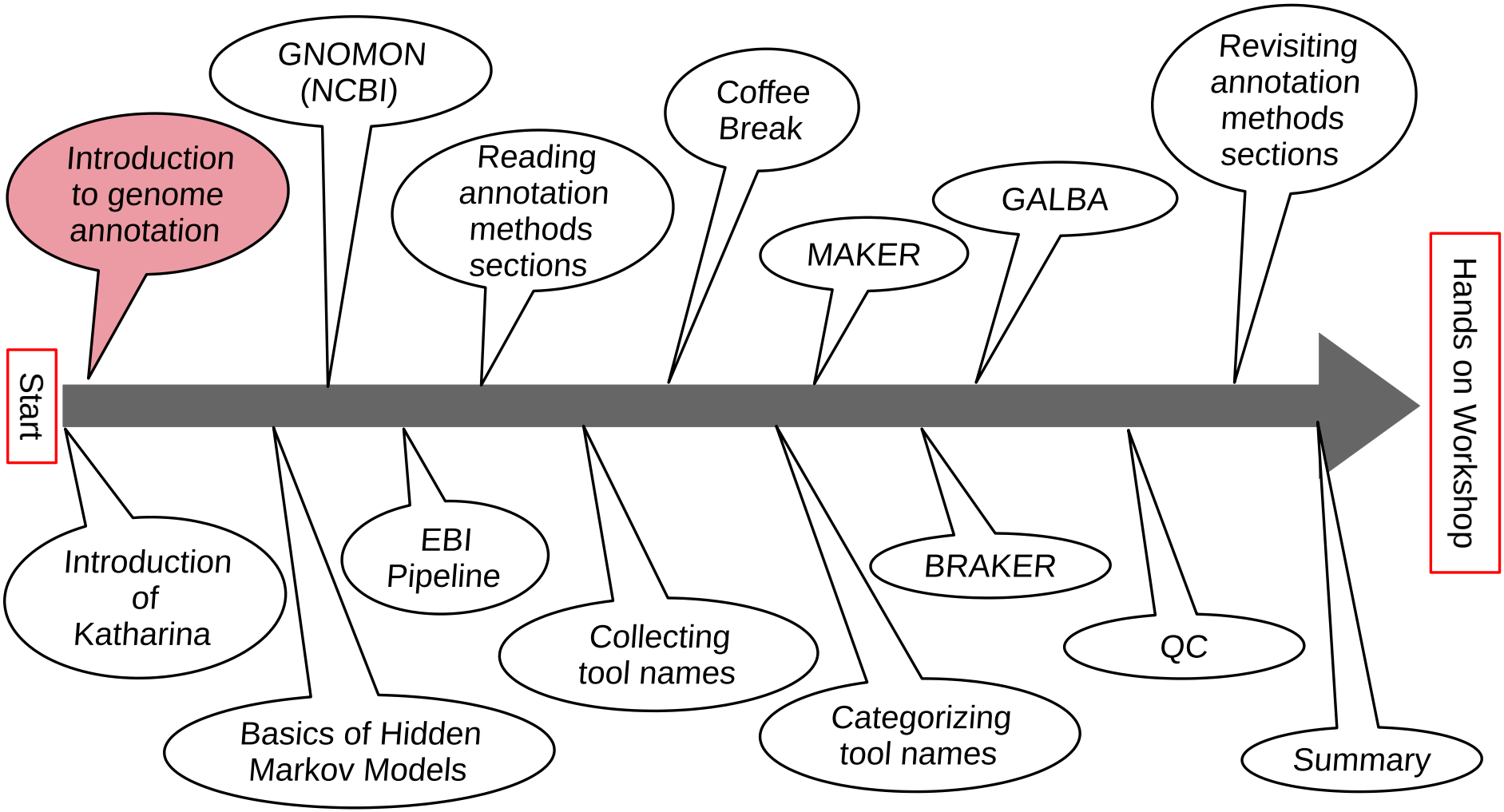
- currently 4 PhD students, 2 MSc students, 1 BSc student
- applied bioinformatics, programming, statistics, & data science

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

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 MAKER, BRAKER, and GALBA
- have an idea of quality control methods

Materials at `https://github.com/KatharinaHoff/GenomeAnnotation_Workshop`


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

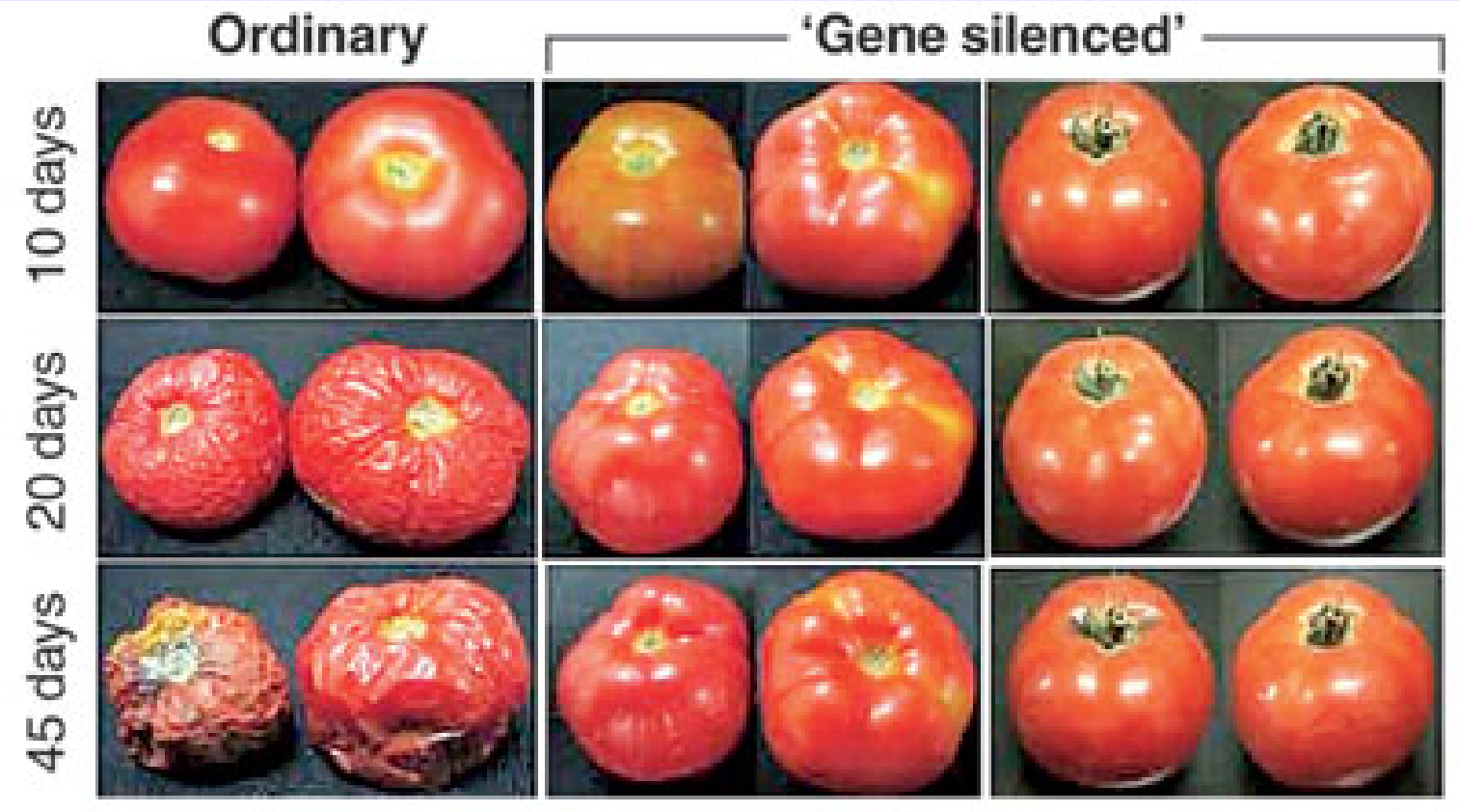
Where are the protein coding genes?

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

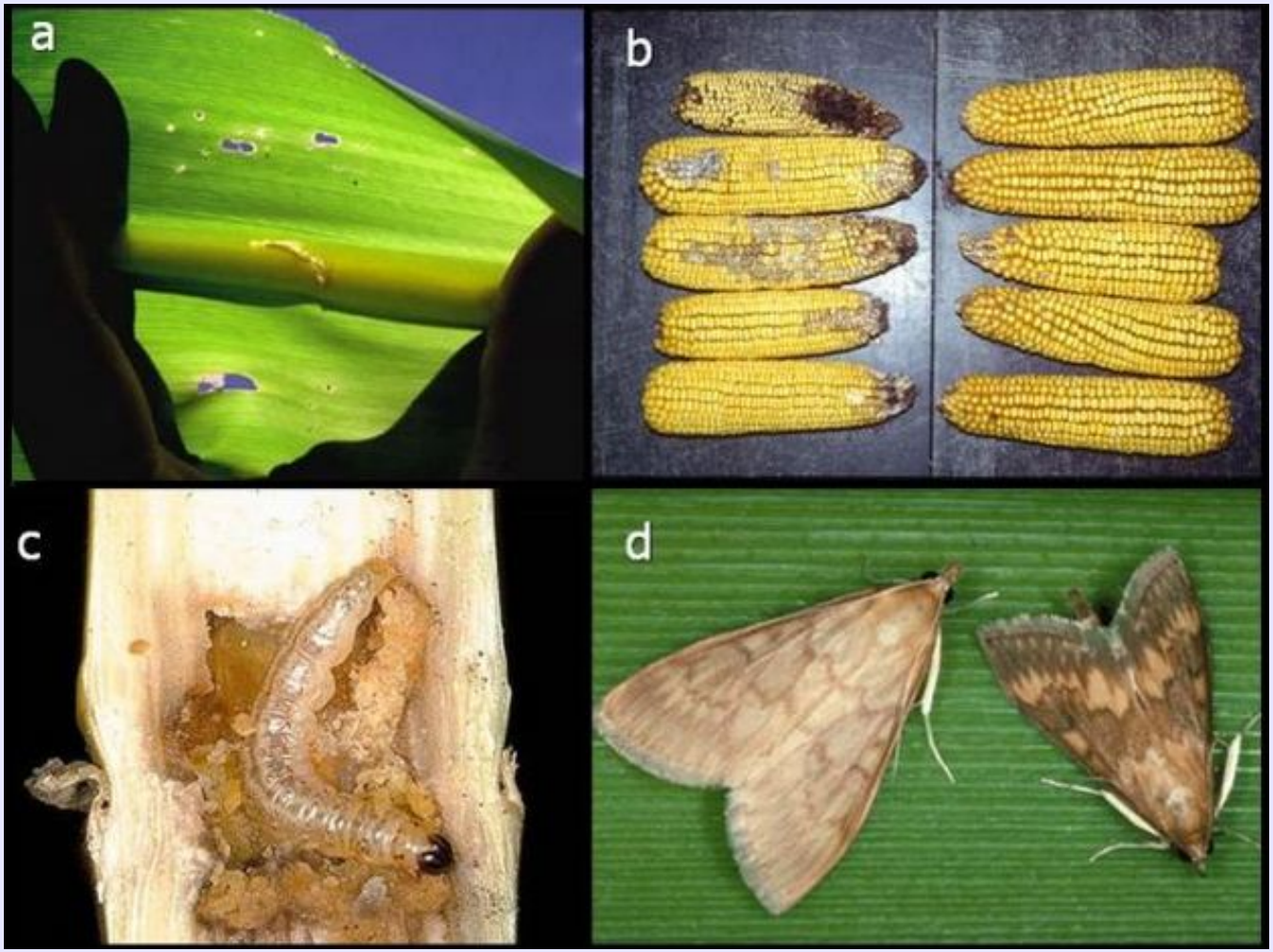


Image: Hellmich & Hellmich (2012) Nature Education Knowledge 3(10):4

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

Examples for the importance of genome annotation

Number of authors on genome papers more recently:

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

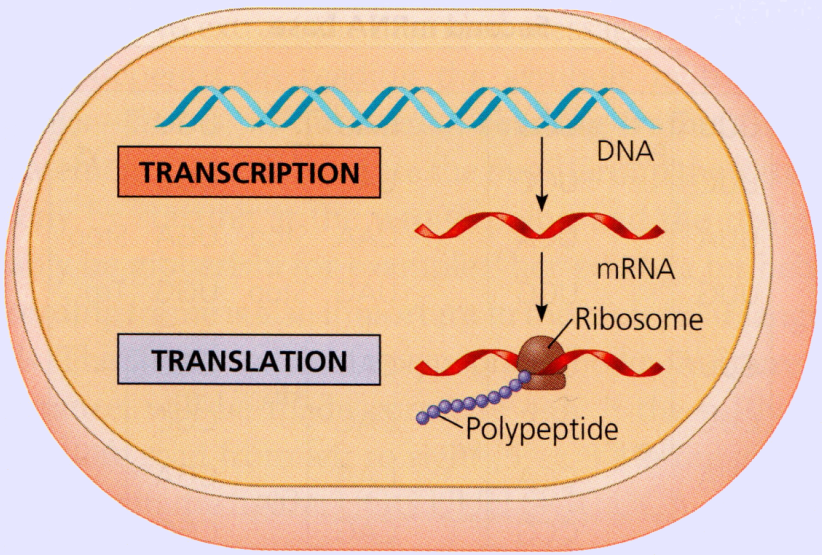
In the past:

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

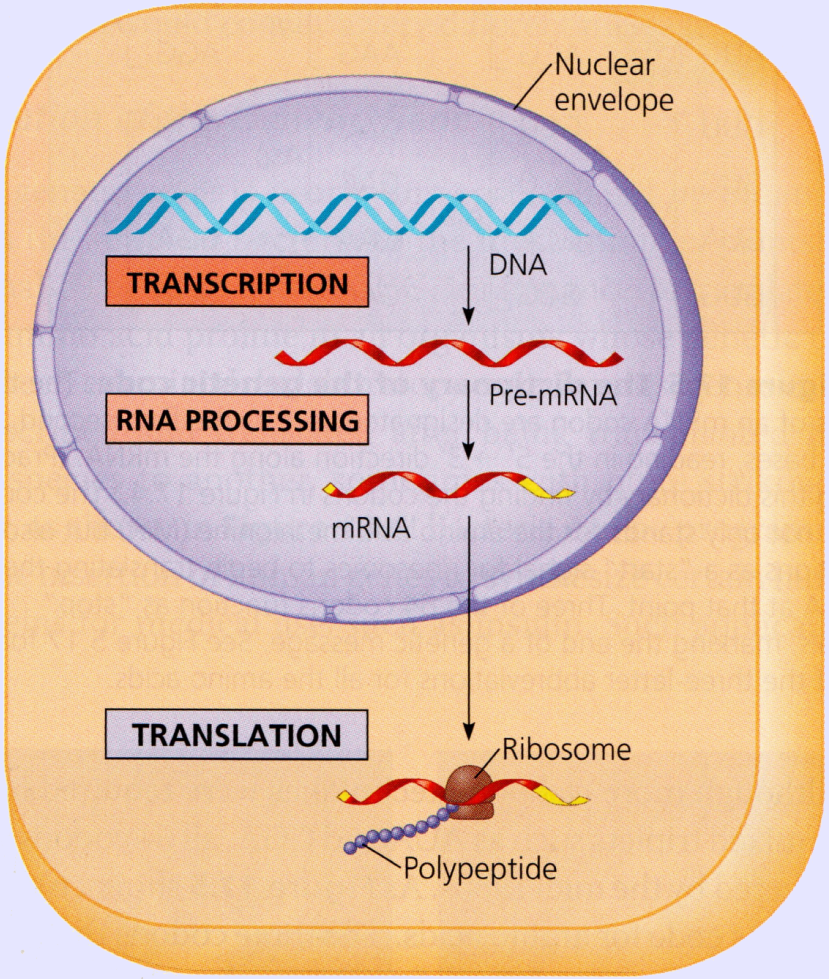
How does a cell recognize protein-coding genes?

Transcription & Translation

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*

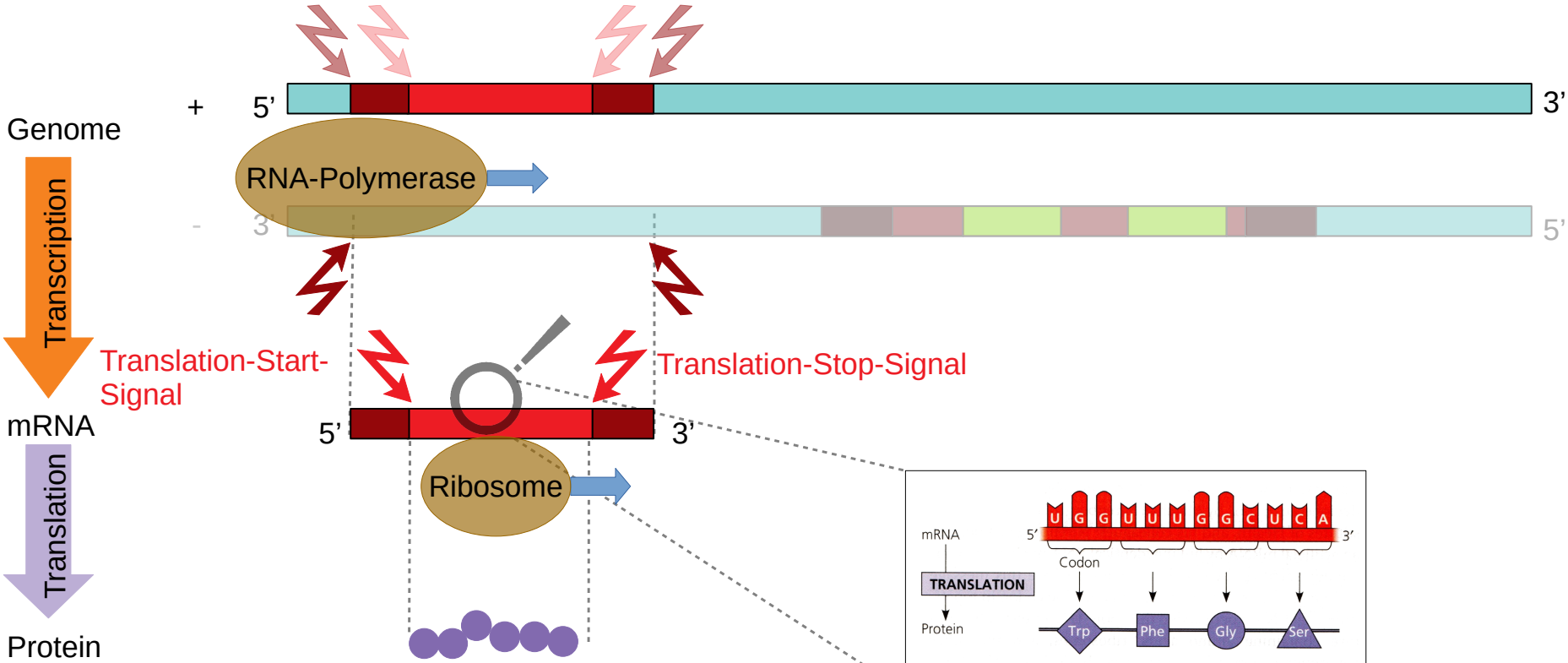


Image: Campbell et al., Biology, San Francisco, 2008, p. 329, FIg. 17.4

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

Prokaryotes & Eukaryotes*

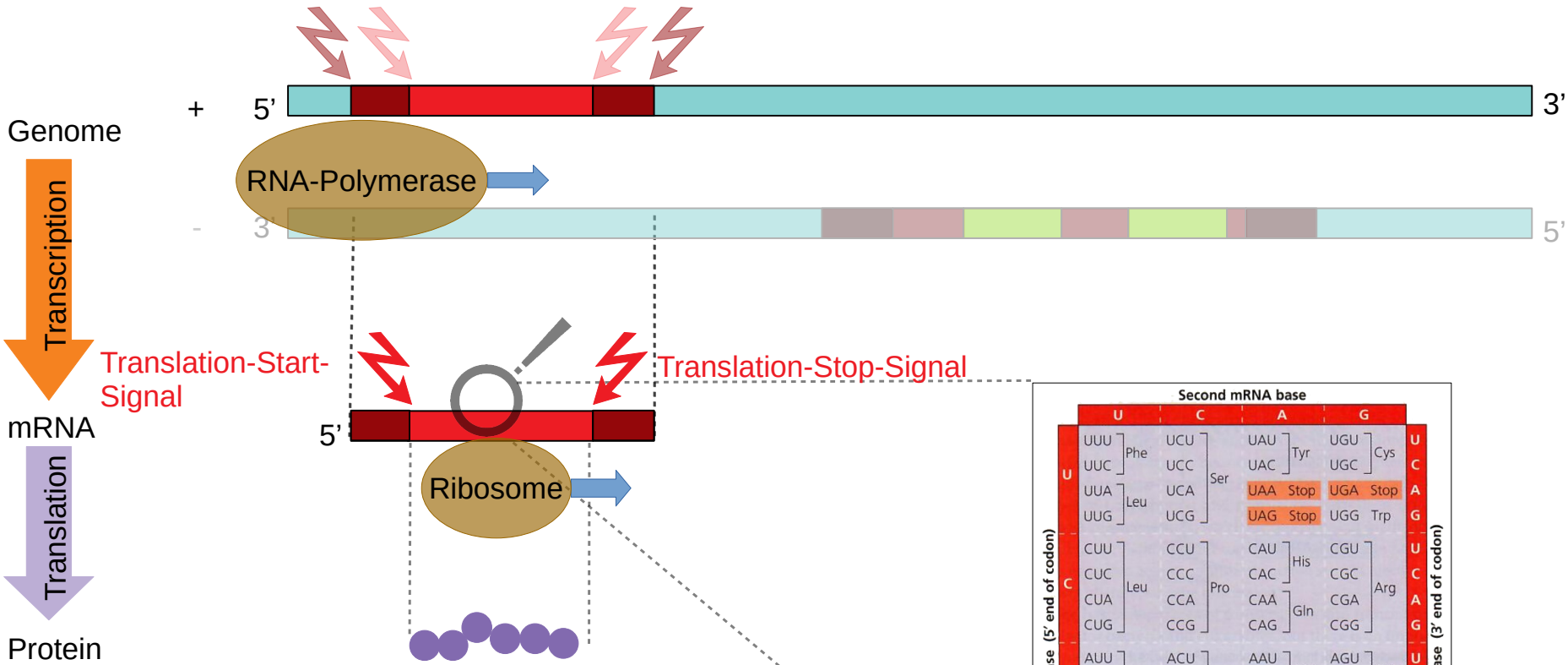
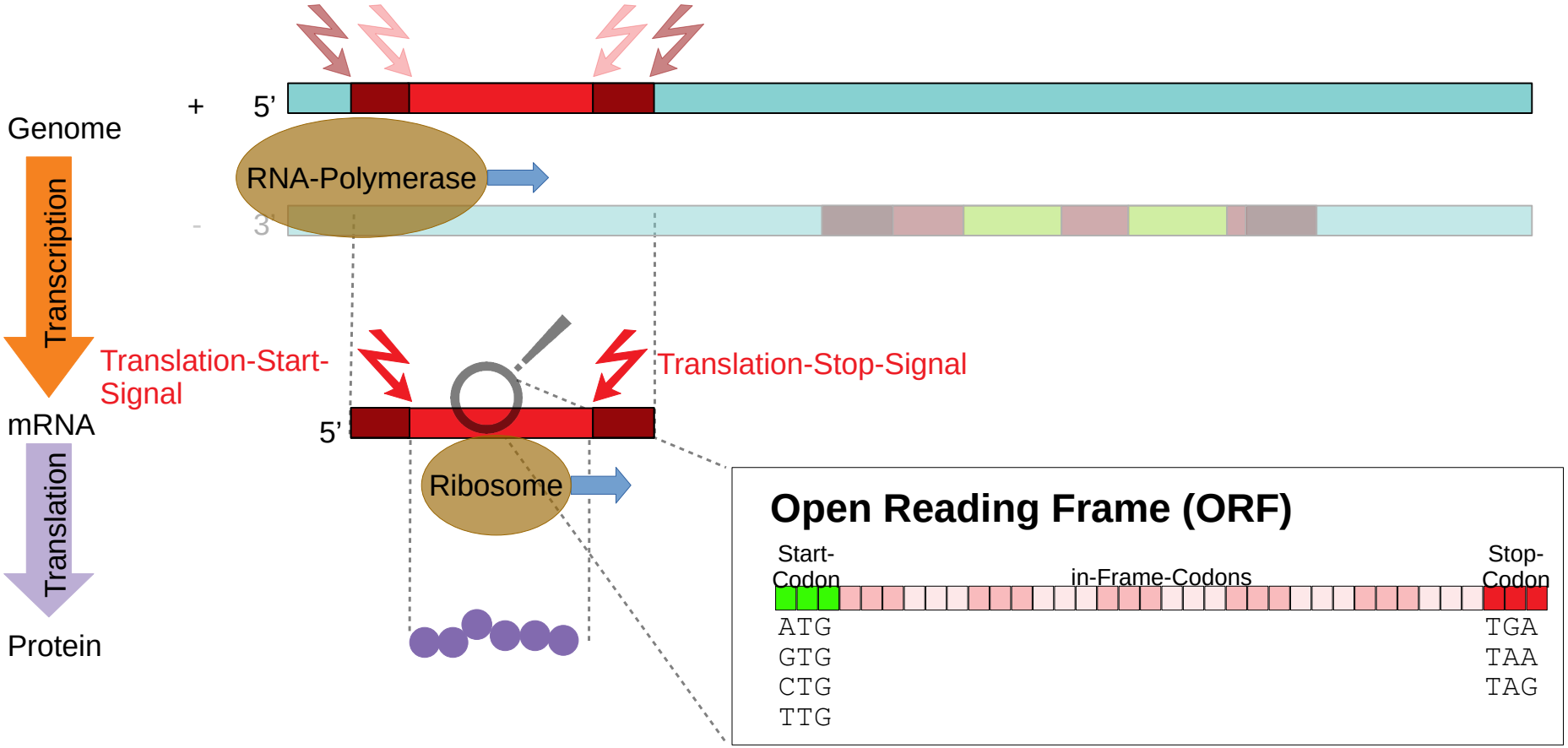


Image: Campbell et al., Biology, San Francisco, 2008, p. 339, Fig. 17.5

*) only some of the genes in eukaryotes

How does a cell recognize protein-coding genes?

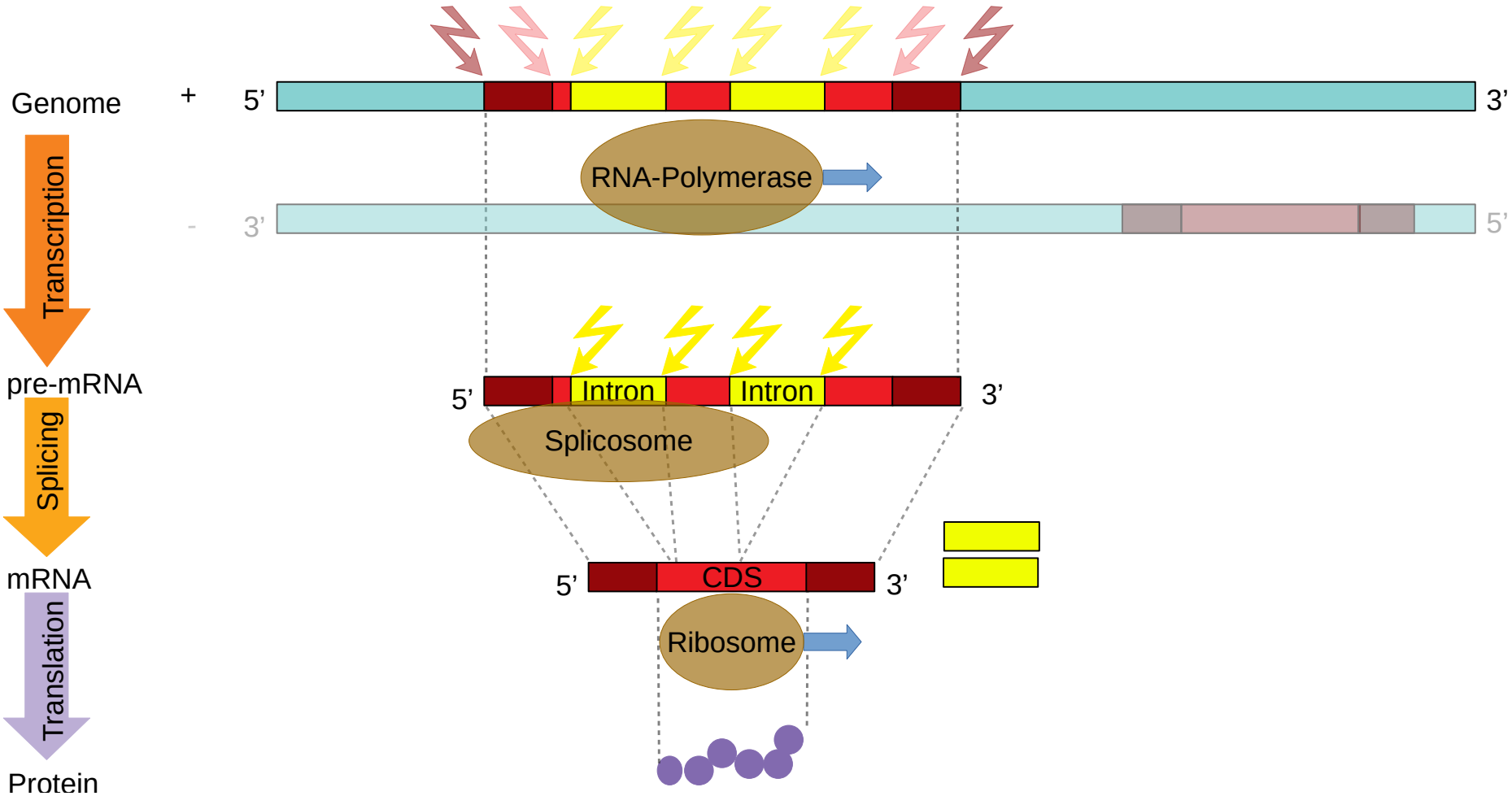
Prokaryotes & Eukaryotes*



- 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

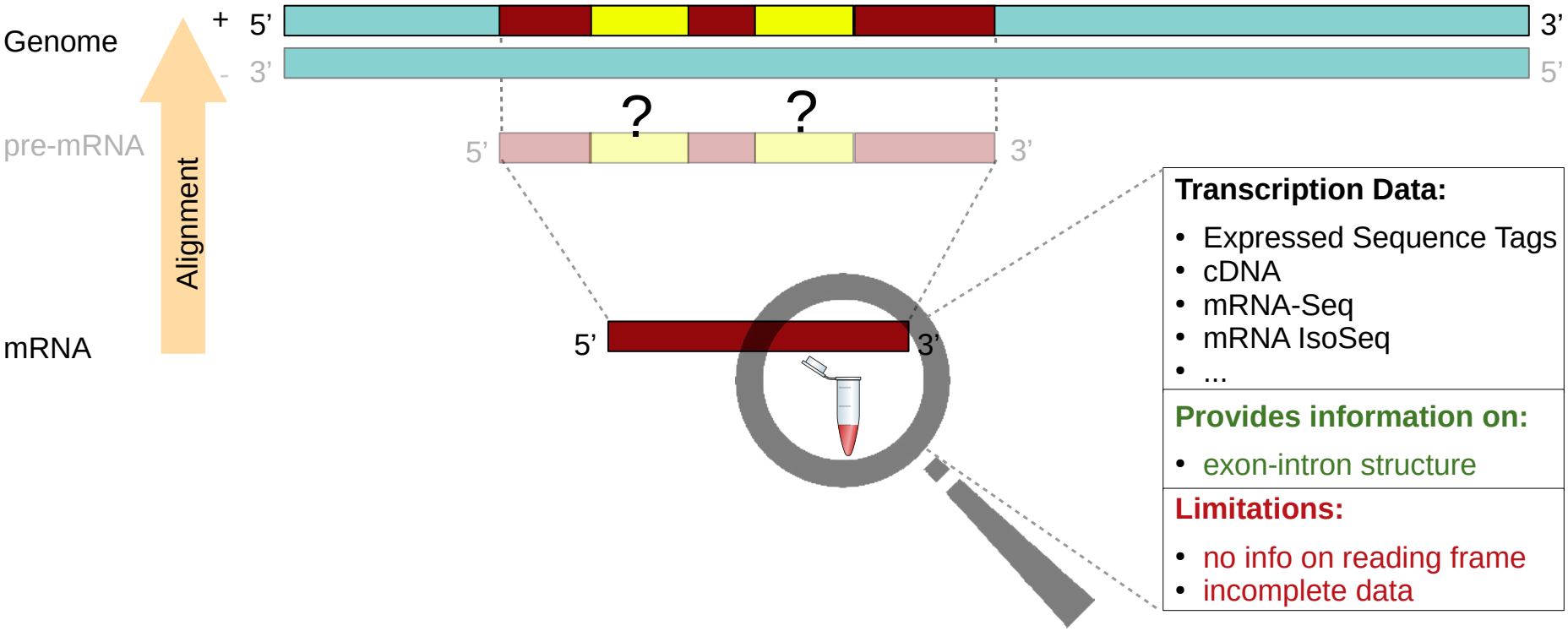
The Genome Annotation Problem

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:

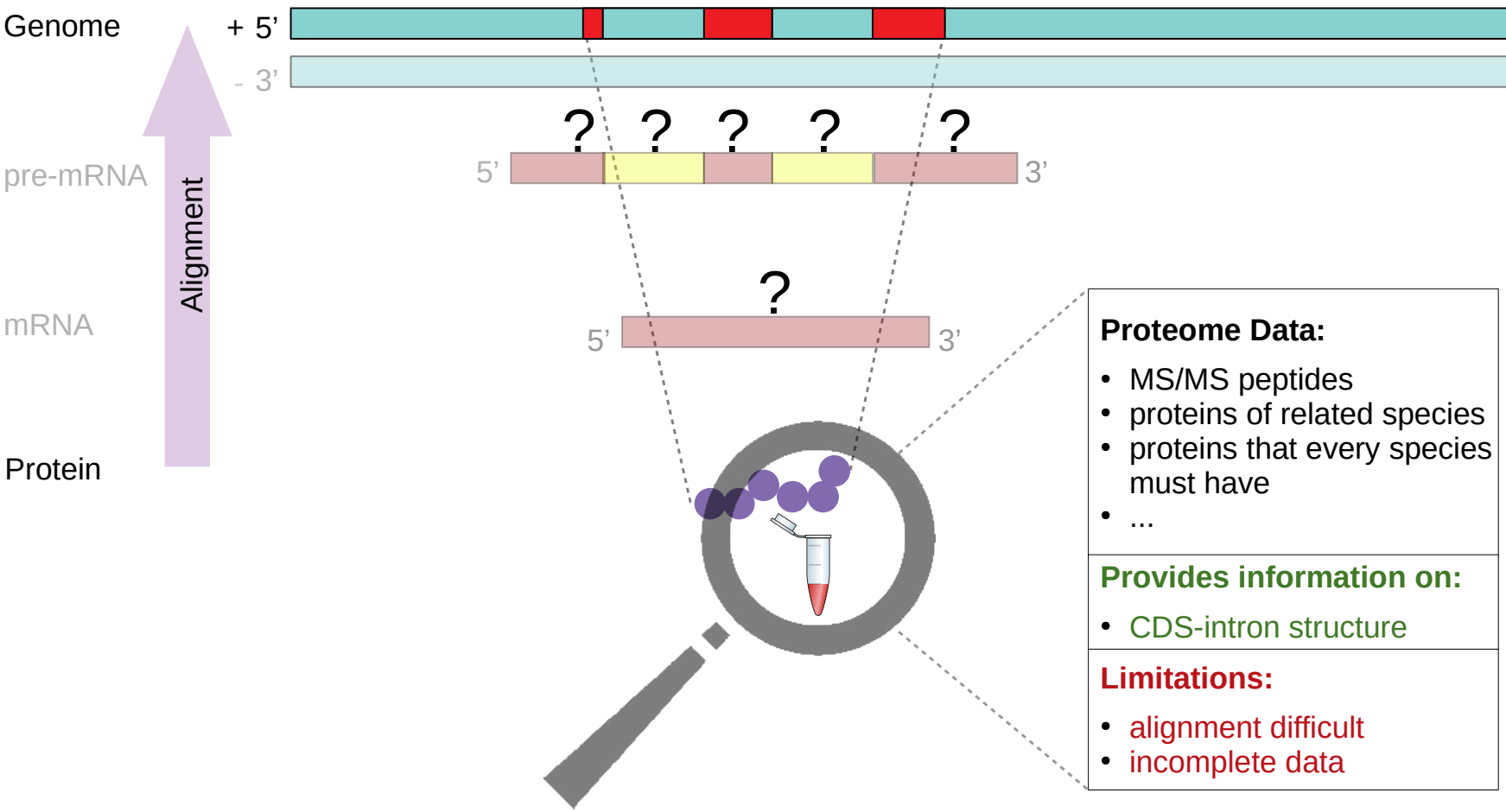
- exon-intron structure

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:

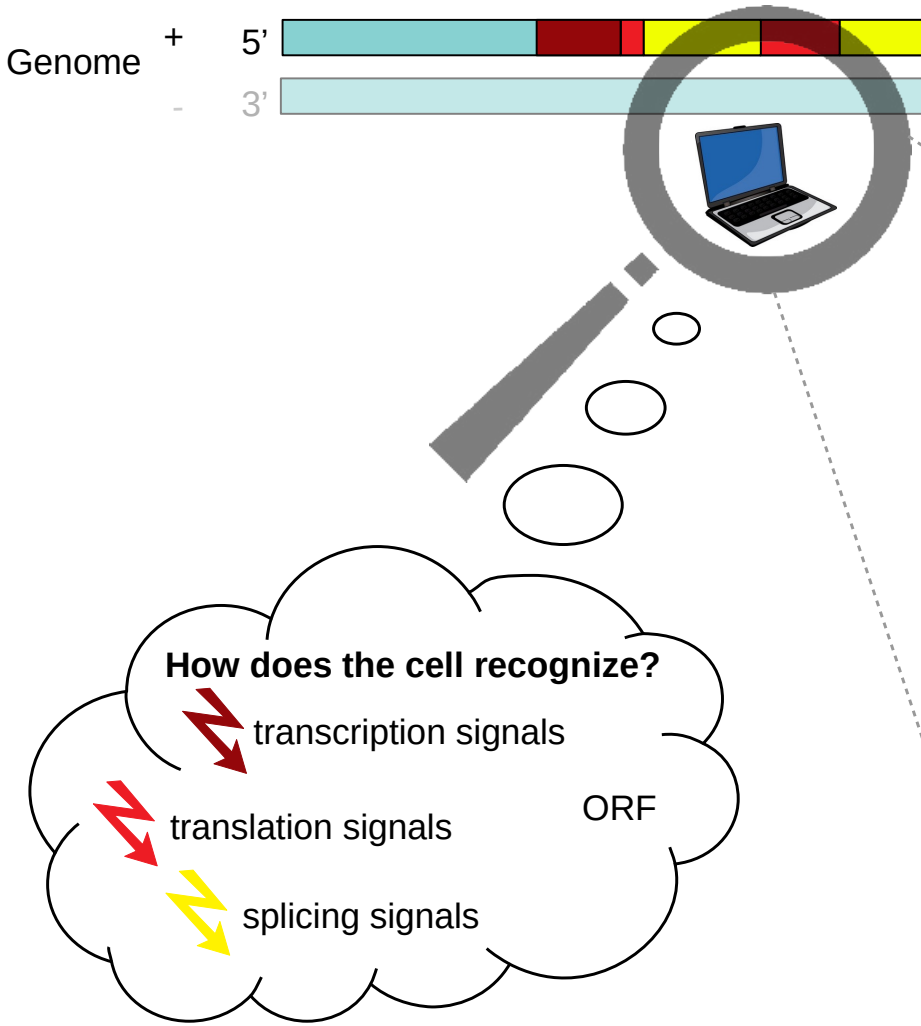
- CDS-intron structure

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:

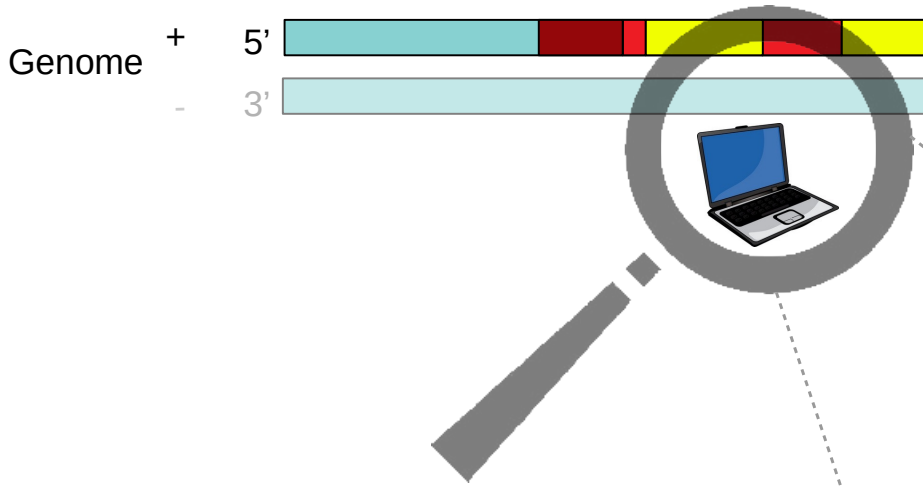
- complete gene structures (sometimes incl. UTRs)

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
- ...

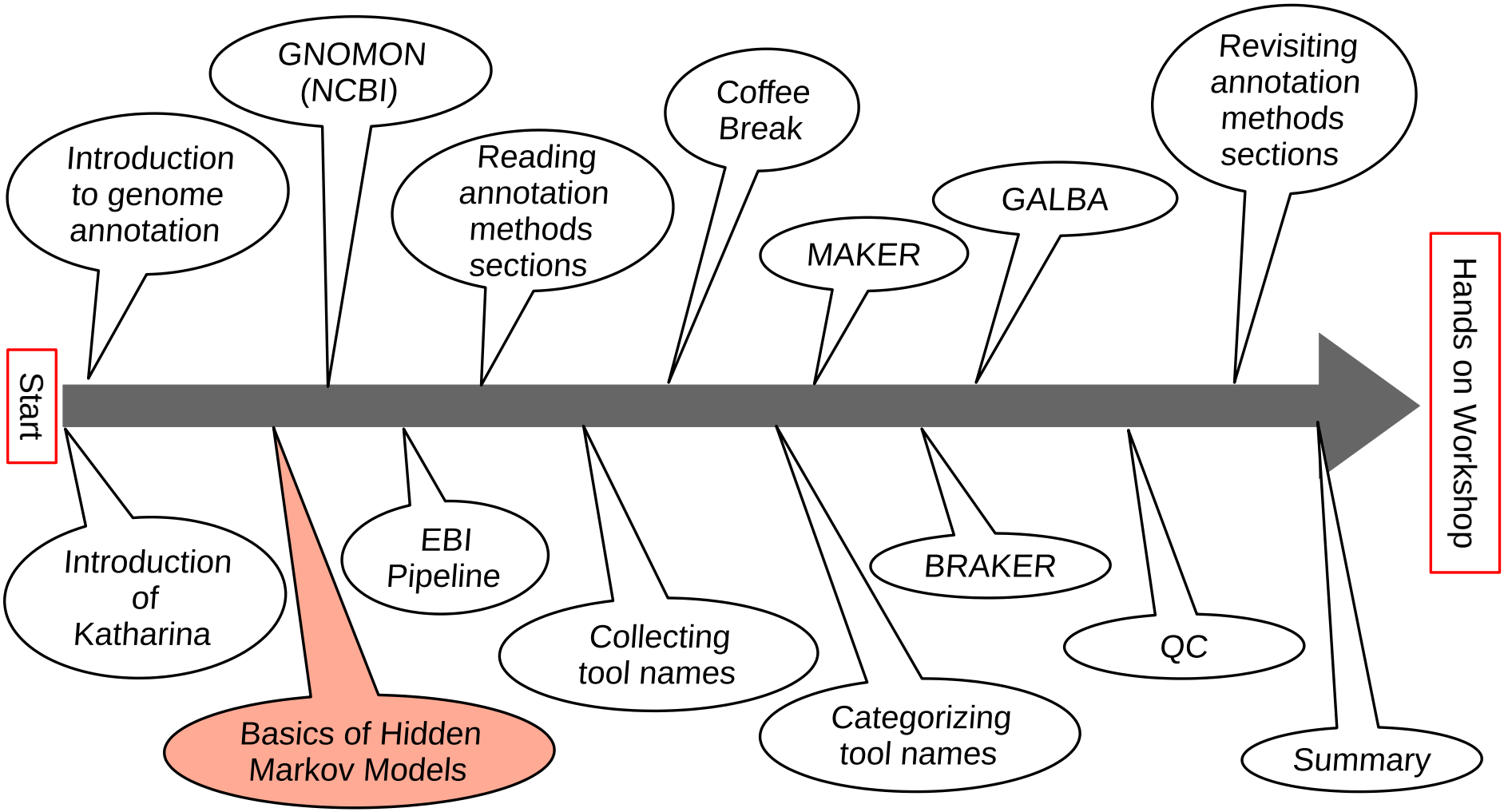
Provide information on:

- 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
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
Basics of Hidden Markov Models
EBI Pipeline
Collecting tool names
Categorizing tool names
BRAKER
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

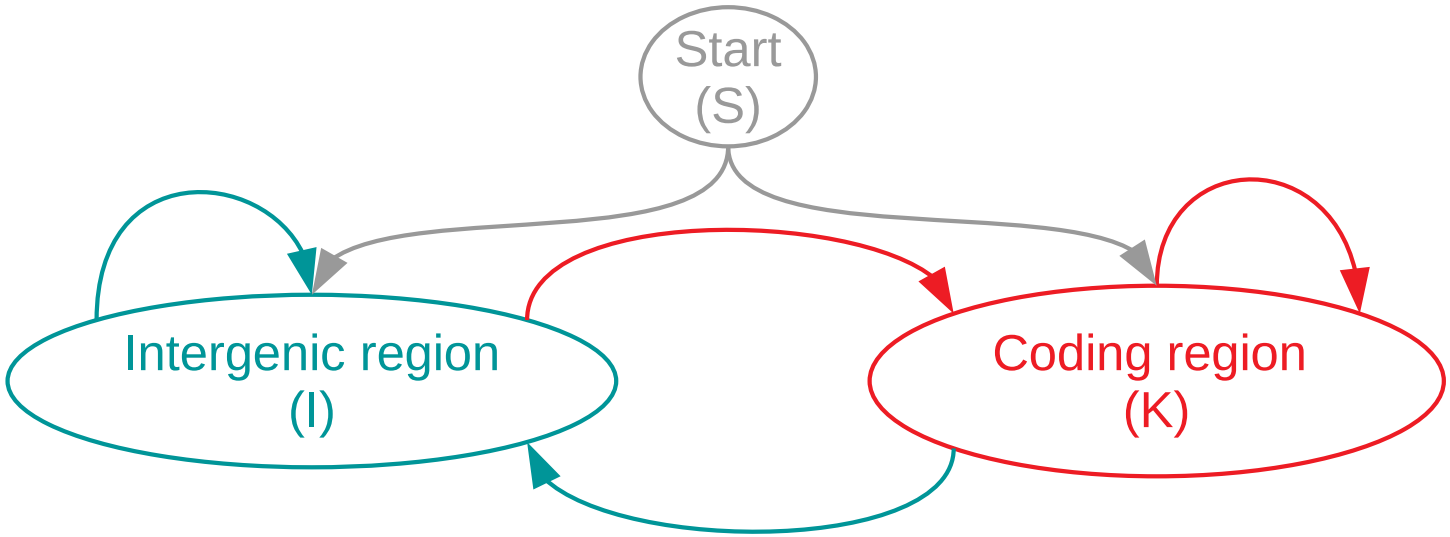
Goal: “Most likely path through hidden states”

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

Basis of highly accurate gene prediction tools

Hidden Markov Model

Hidden states



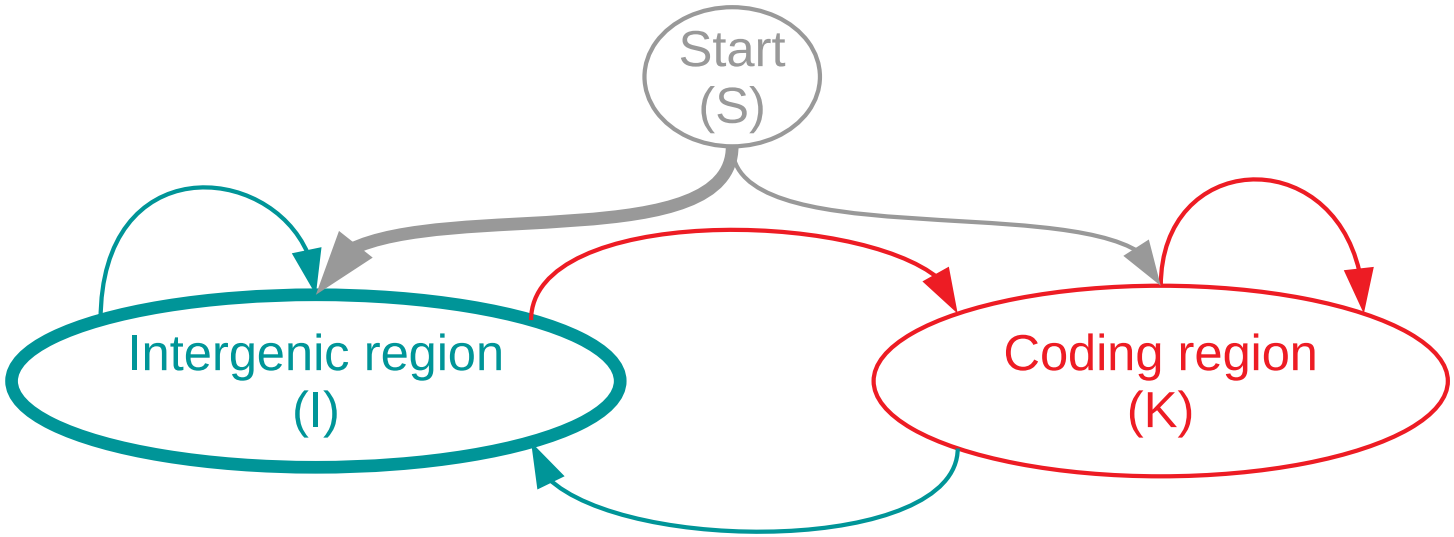
A possible 'state path' for the genome sequence:

AABBBAA

Basis of highly accurate gene prediction tools

Hidden Markov Model

Hidden states

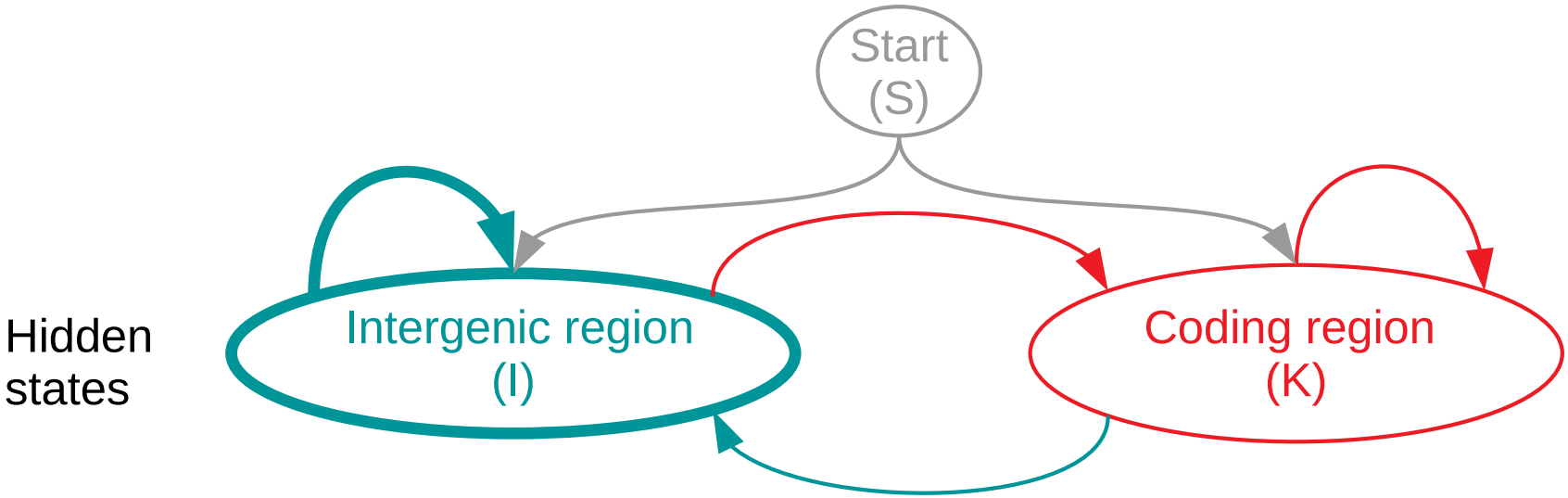


A possible 'state path' for the genome sequence:

AABBBAA
I

Basis of highly accurate gene prediction tools

Hidden Markov Model

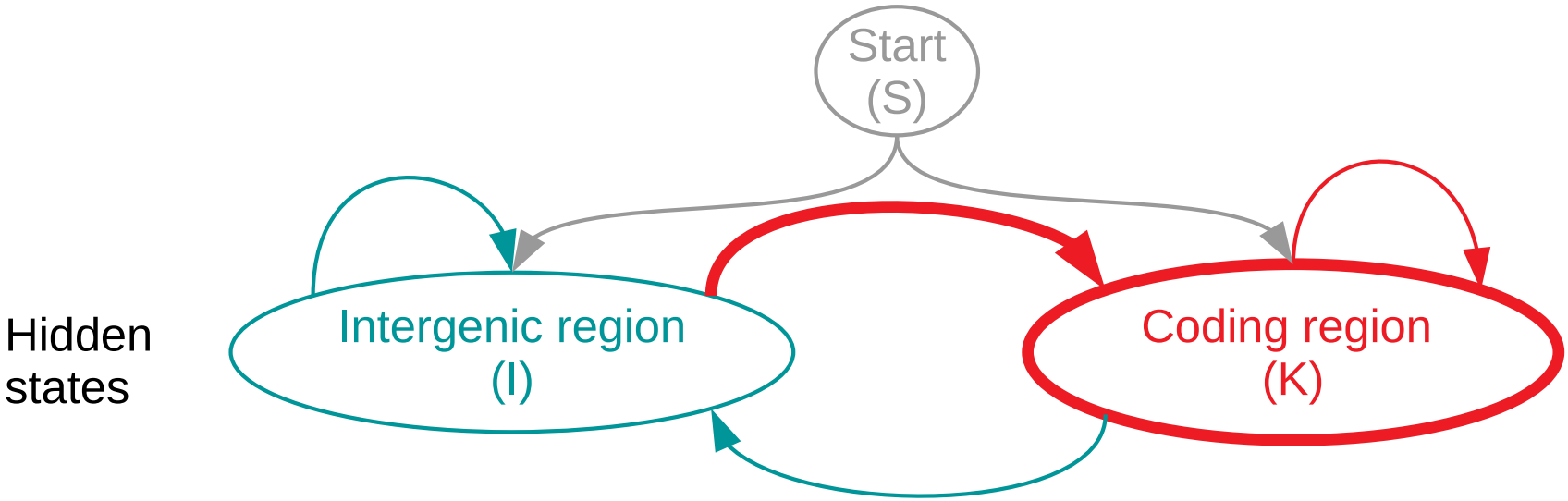


A possible 'state path' for the genome sequence:

AABBBAA
II

Basis of highly accurate gene prediction tools

Hidden Markov Model



A possible 'state path' for the genome sequence:

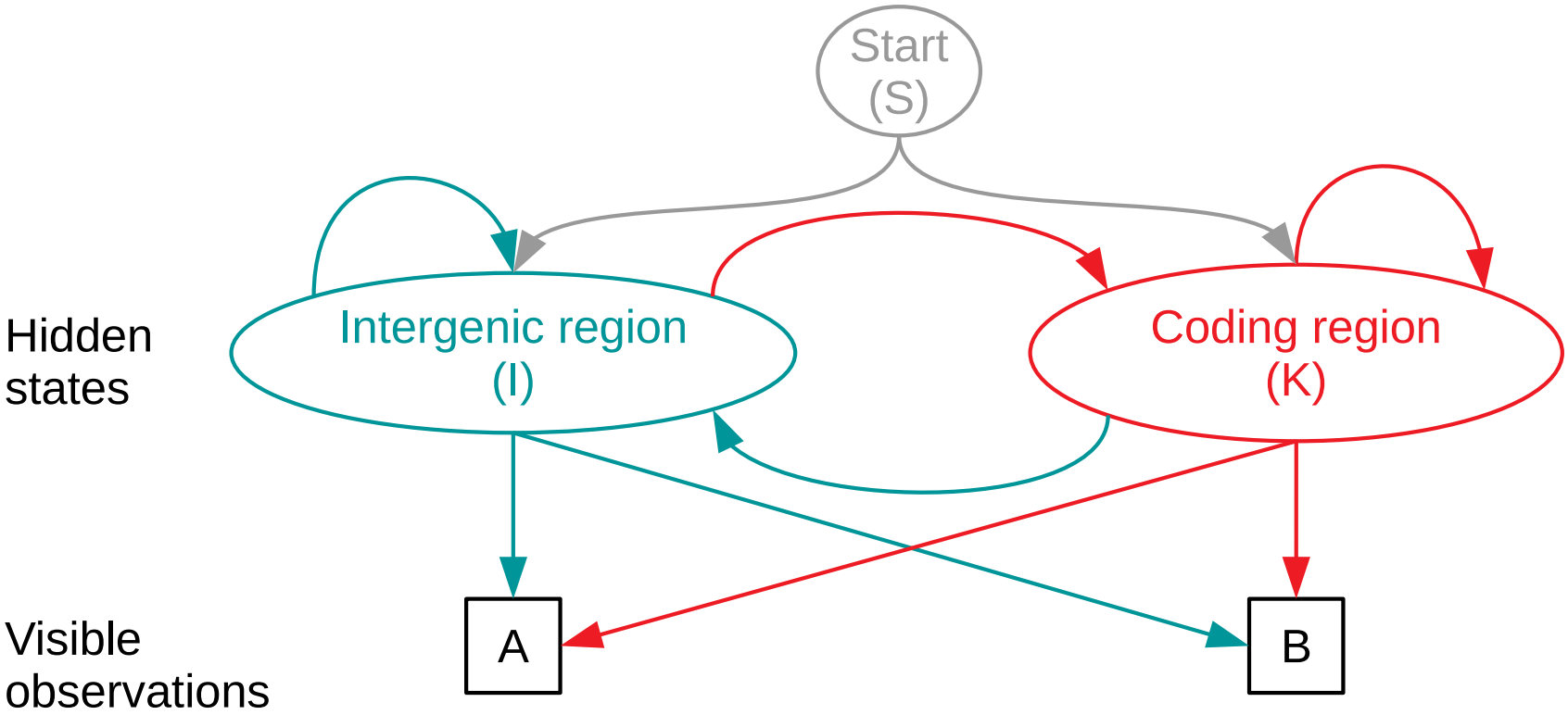
```
AABBBAA
IIK...
```

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

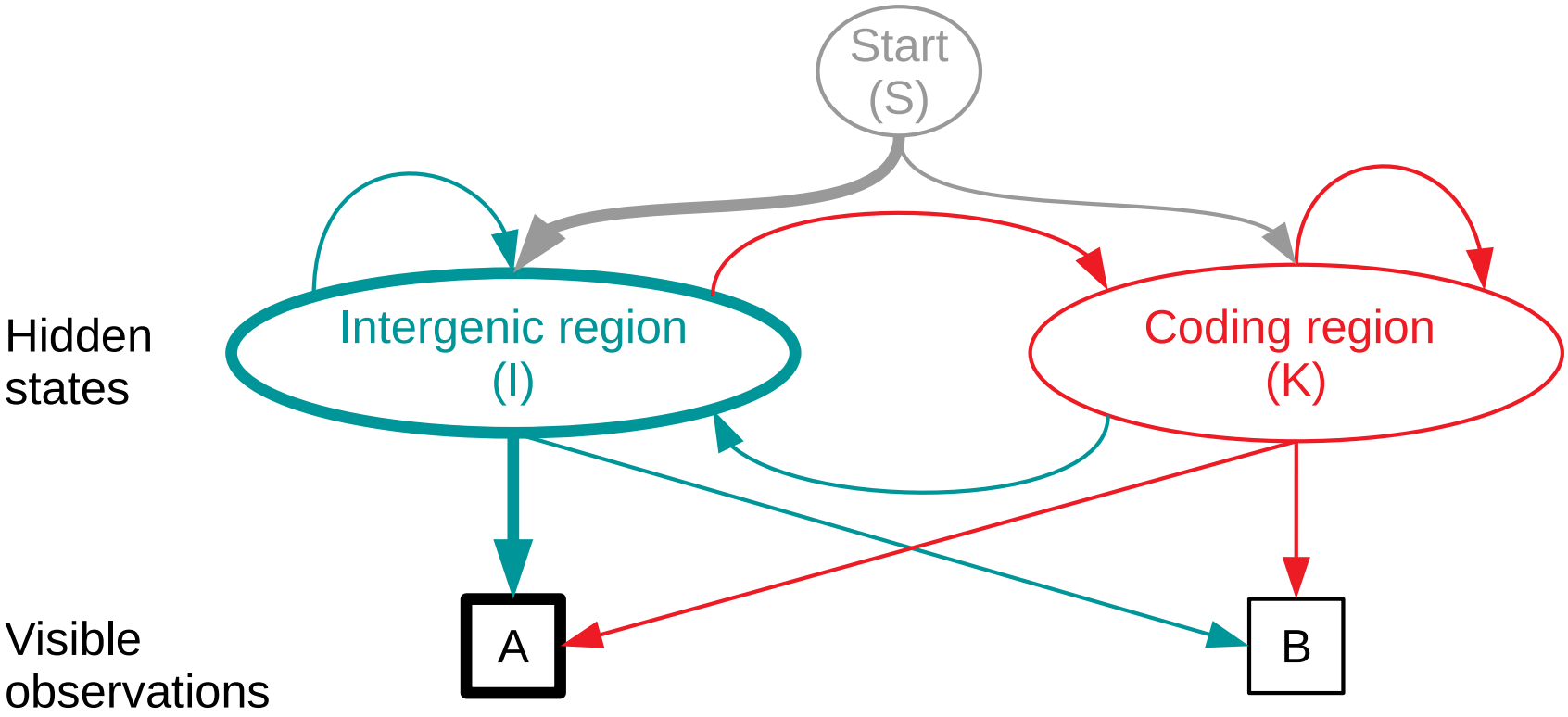
Hidden Markov Model



A possible ‘state path’ for the genome sequence:

Basis of highly accurate gene prediction tools

Hidden Markov Model

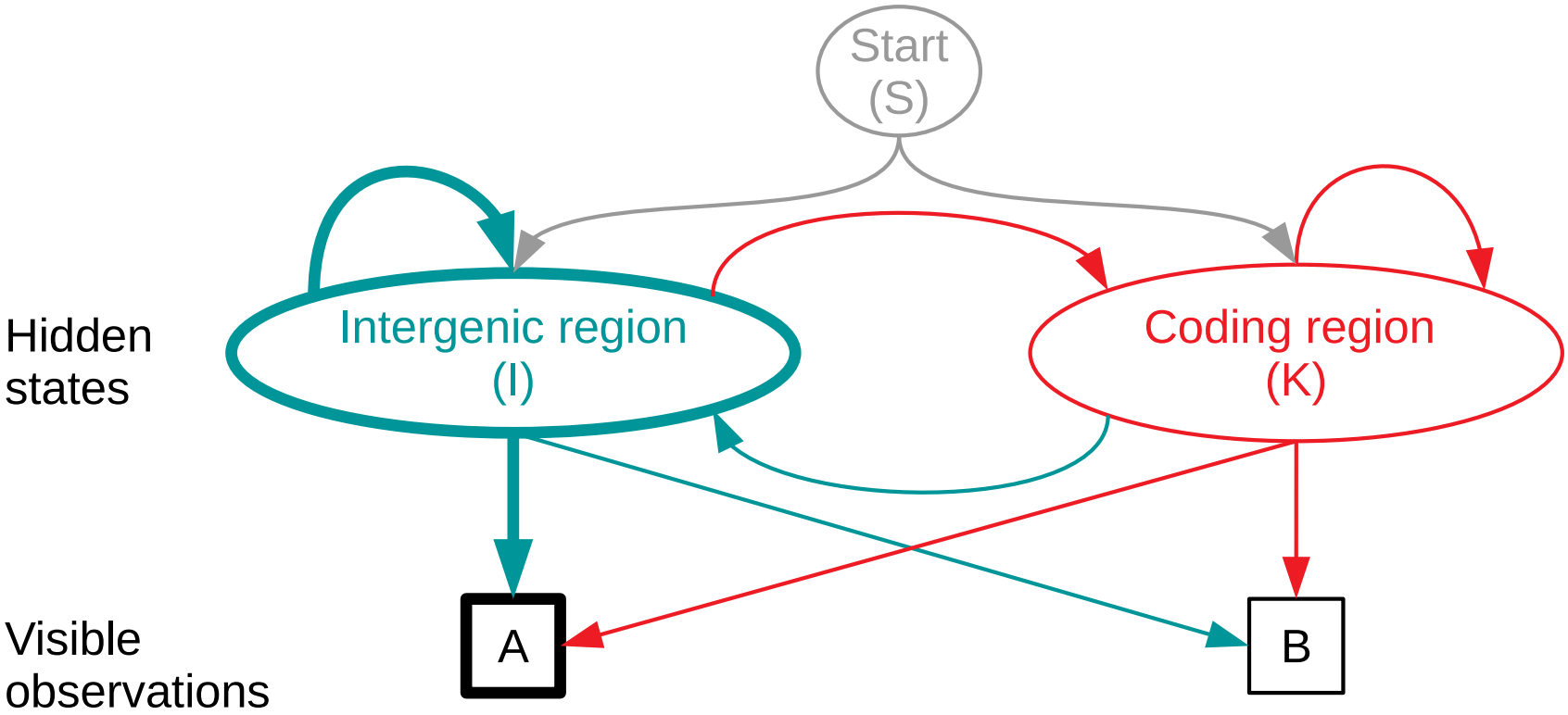


A possible ‘state path’ for the genome sequence:

A
I

Basis of highly accurate gene prediction tools

Hidden Markov Model

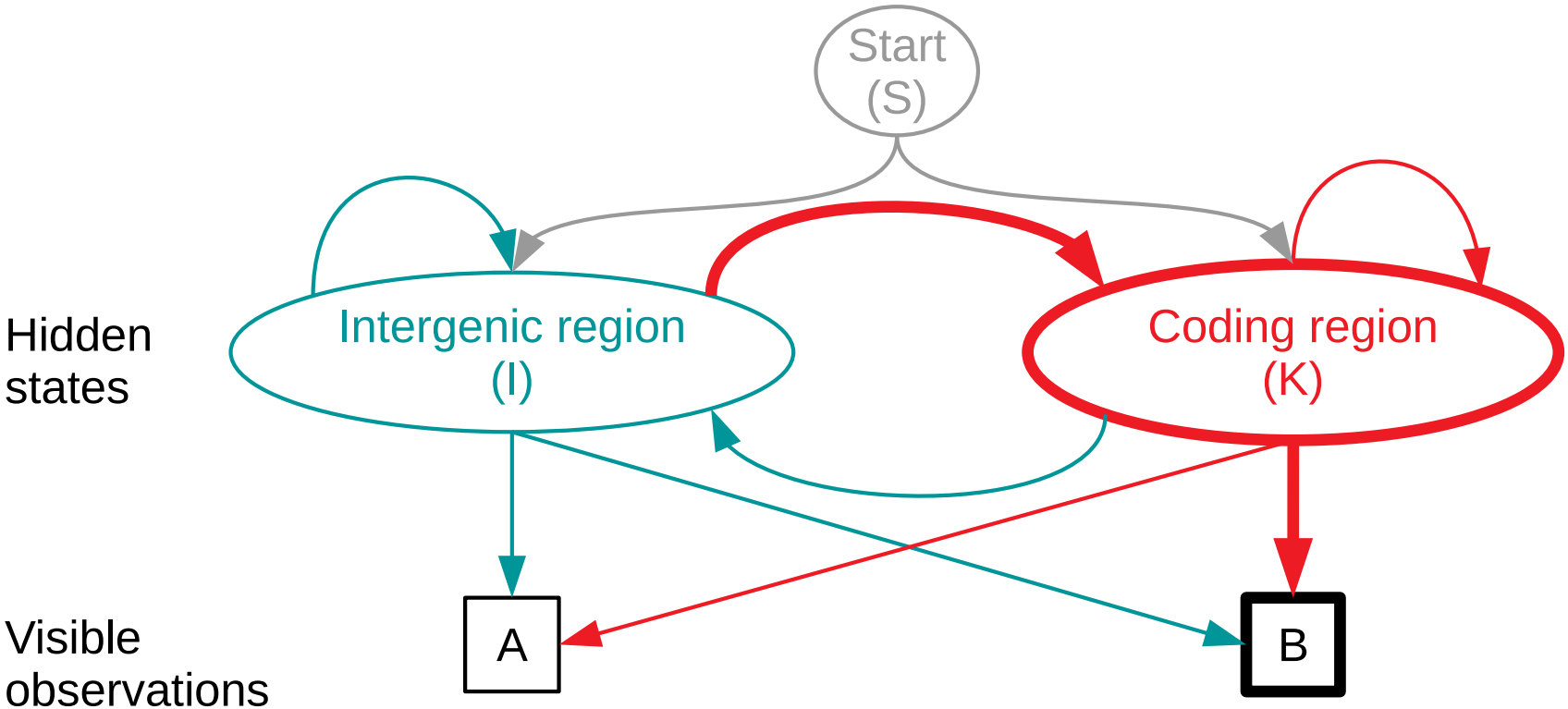


A possible 'state path' for the genome sequence:

AA
II

Basis of highly accurate gene prediction tools

Hidden Markov Model

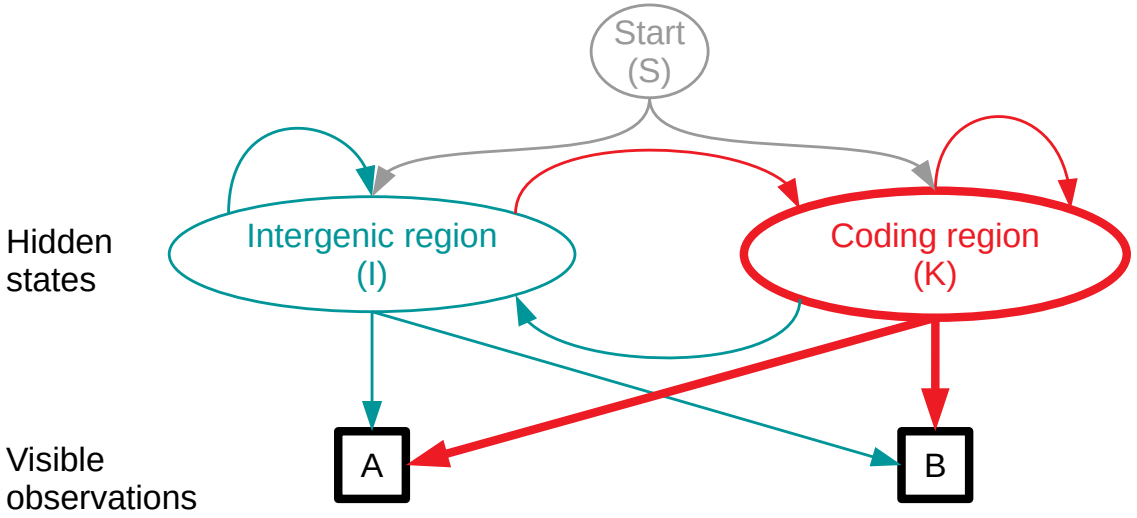


A possible 'state path' for the genome sequence:

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

Basis of highly accurate gene prediction tools

Hidden Markov Model

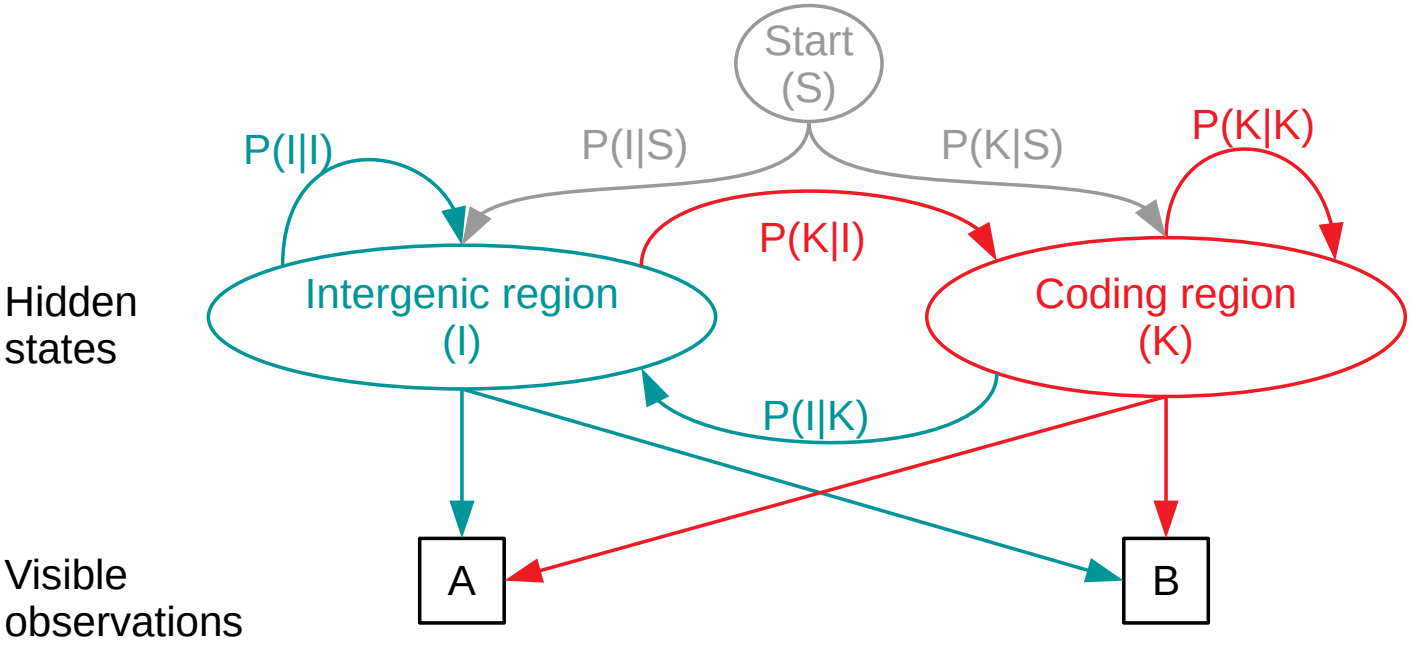


Model properties

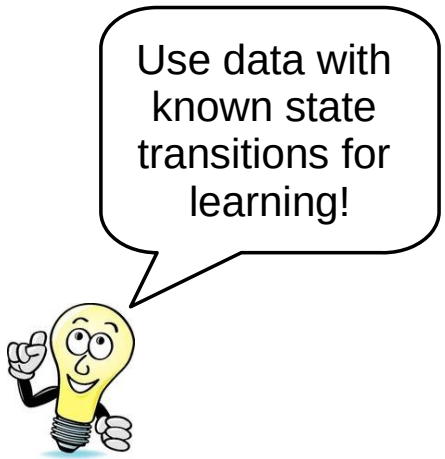
1. The current value of the hidden state depends exclusively on the state of its predecessor.
2. The current value of the visible observation depends exclusively on the value of the current, hidden state.

Basis of highly accurate gene prediction tools

Hidden Markov Model

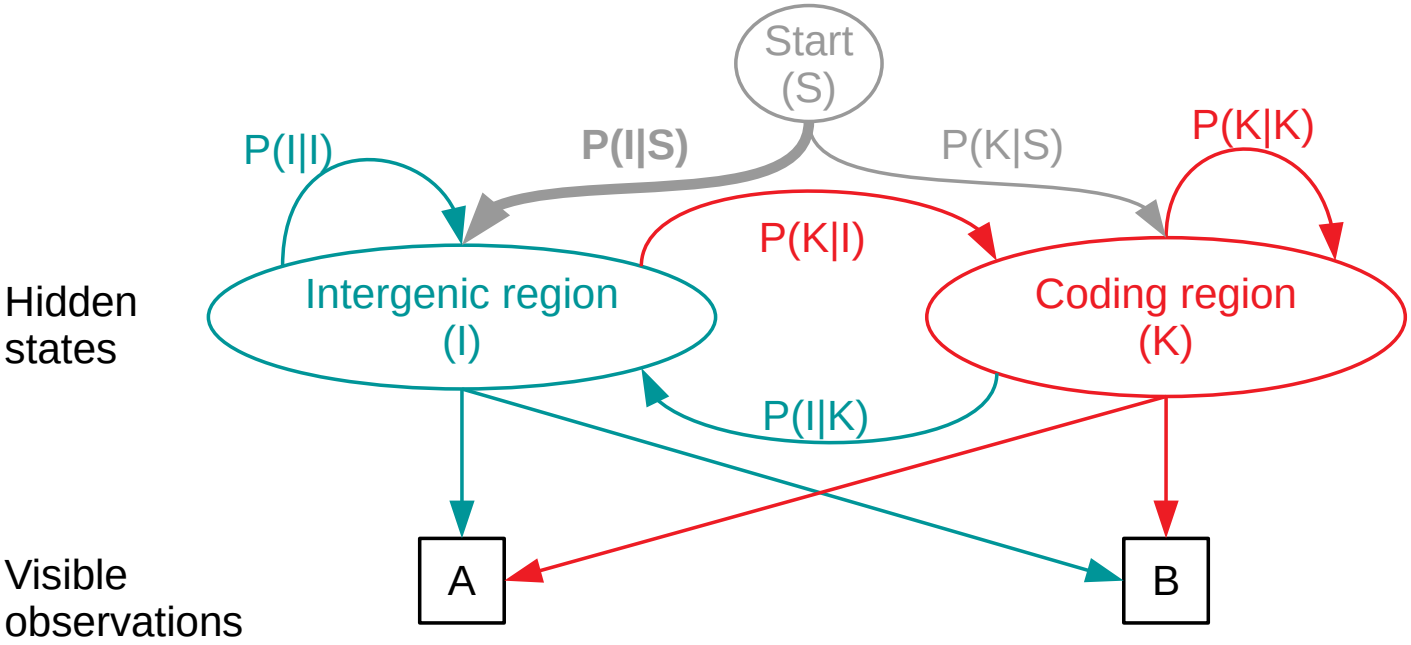


How likely are the state transitions?



Basis of highly accurate gene prediction tools

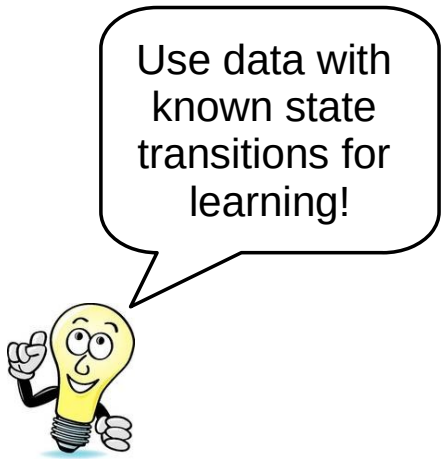
Hidden Markov Model



Training data:

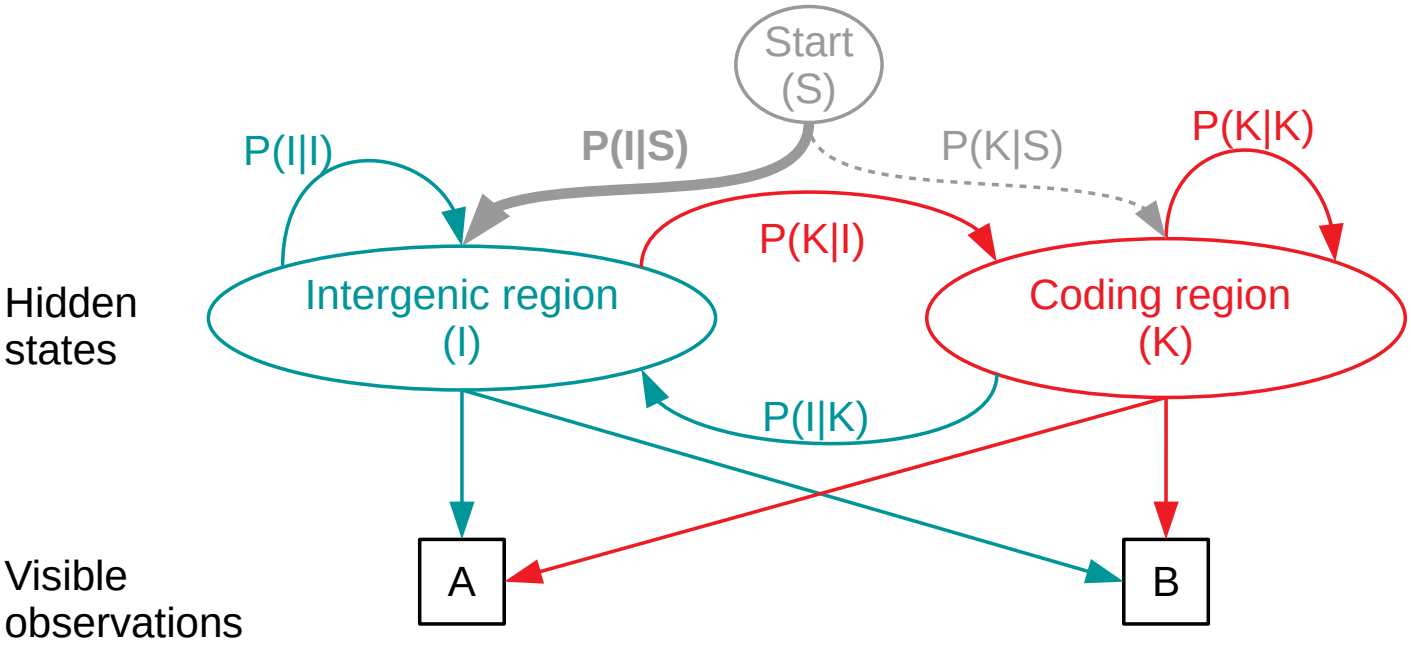
AABABA
IKKIII

Start probability
P(I|S) = ?



Basis of highly accurate gene prediction tools

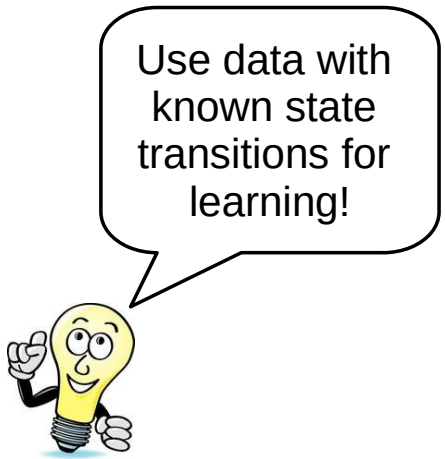
Hidden Markov Model



Training data:

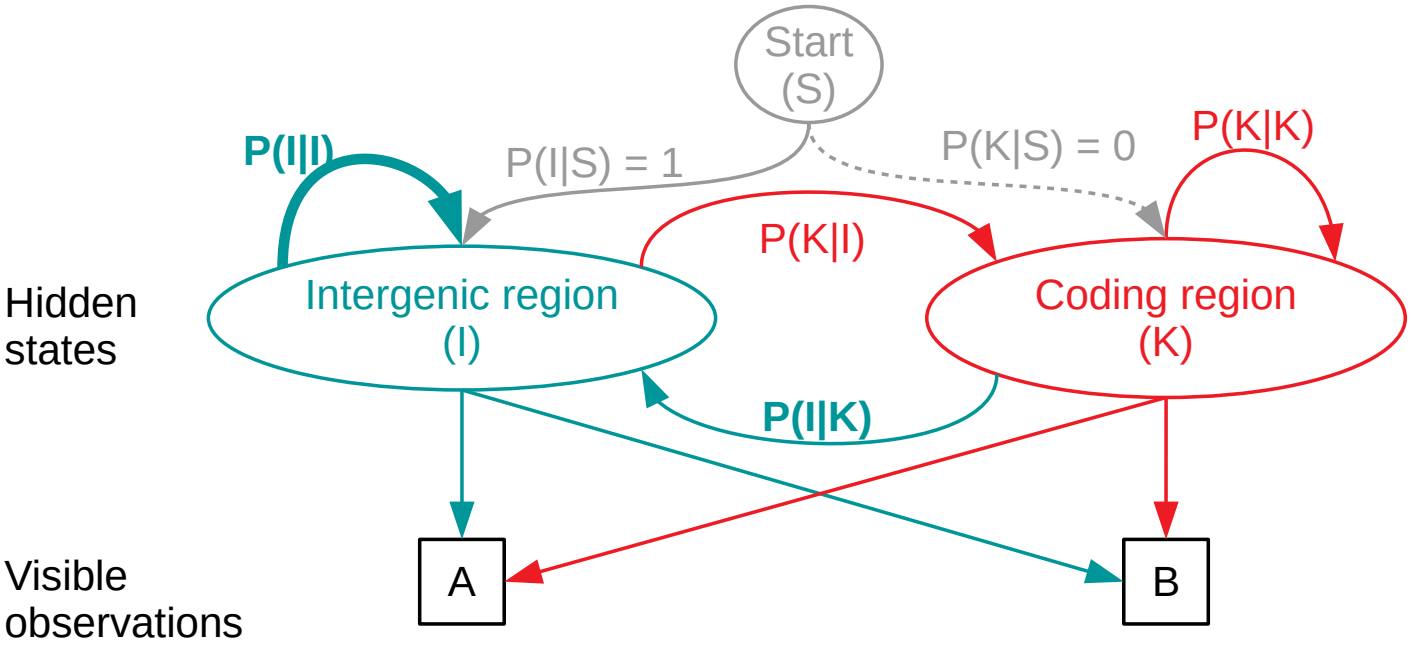
AABABA
IKKIII
+

Start probability
P(I|S) = 1



Basis of highly accurate gene prediction tools

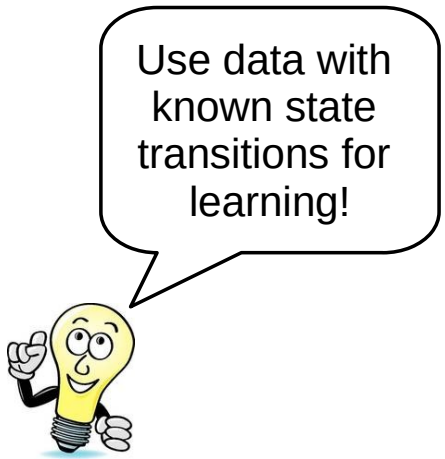
Hidden Markov Model



Training data:

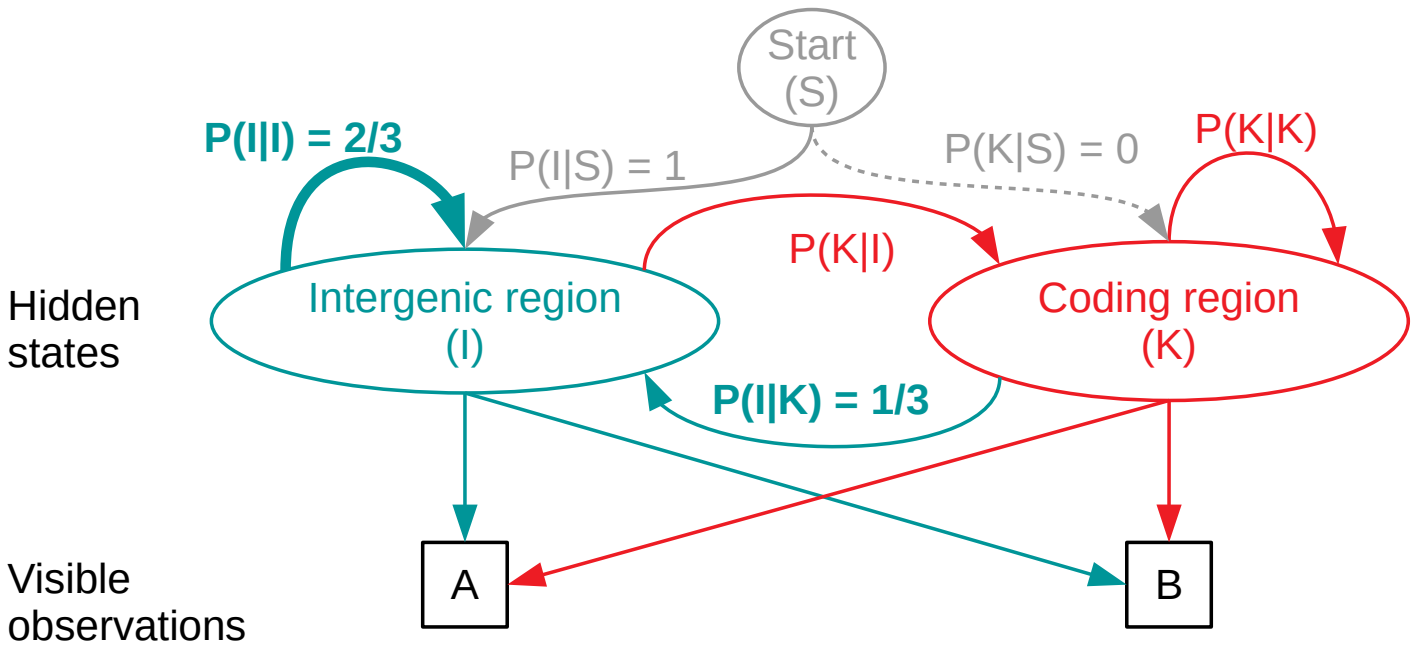
AABABA
IKKIII

P(I|I) = ?



Basis of highly accurate gene prediction tools

Hidden Markov Model



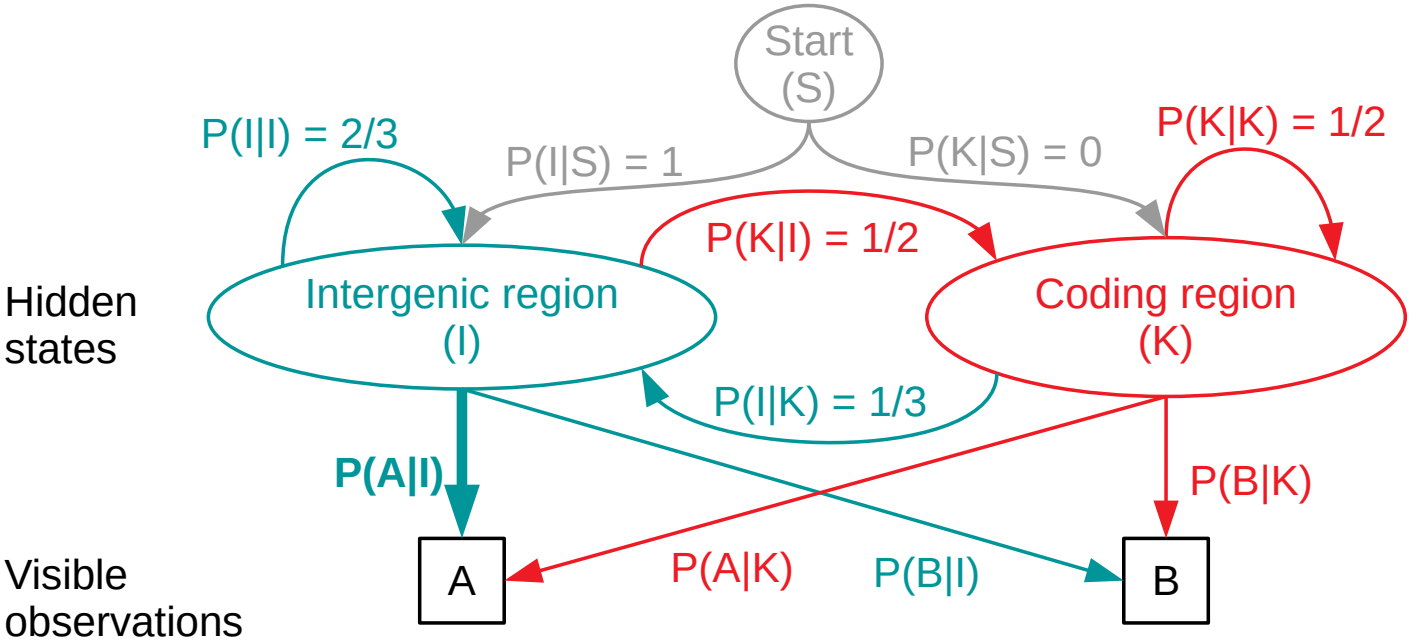
Training data:

AABABA
IKKIII
-++

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

Basis of highly accurate gene prediction tools

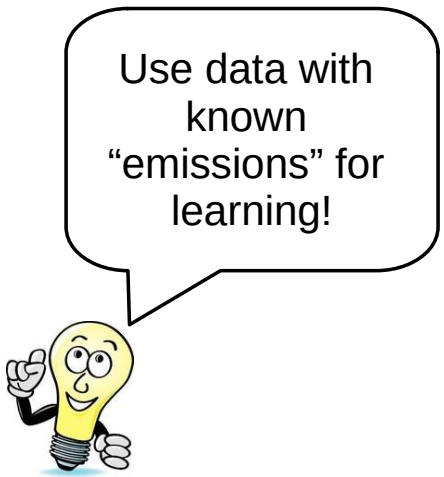
Hidden Markov Model



How likely are the observations?

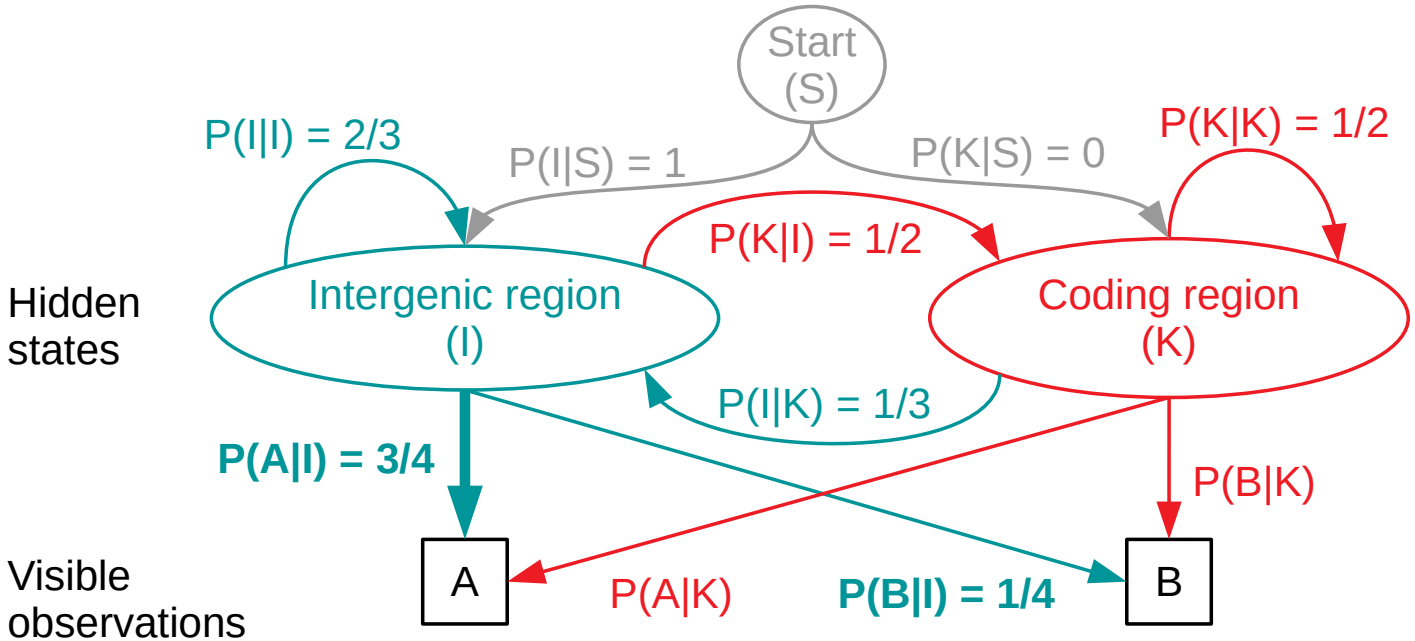
AABABA
IKKIII

P(A|I) = ?



Basis of highly accurate gene prediction tools

Hidden Markov Model



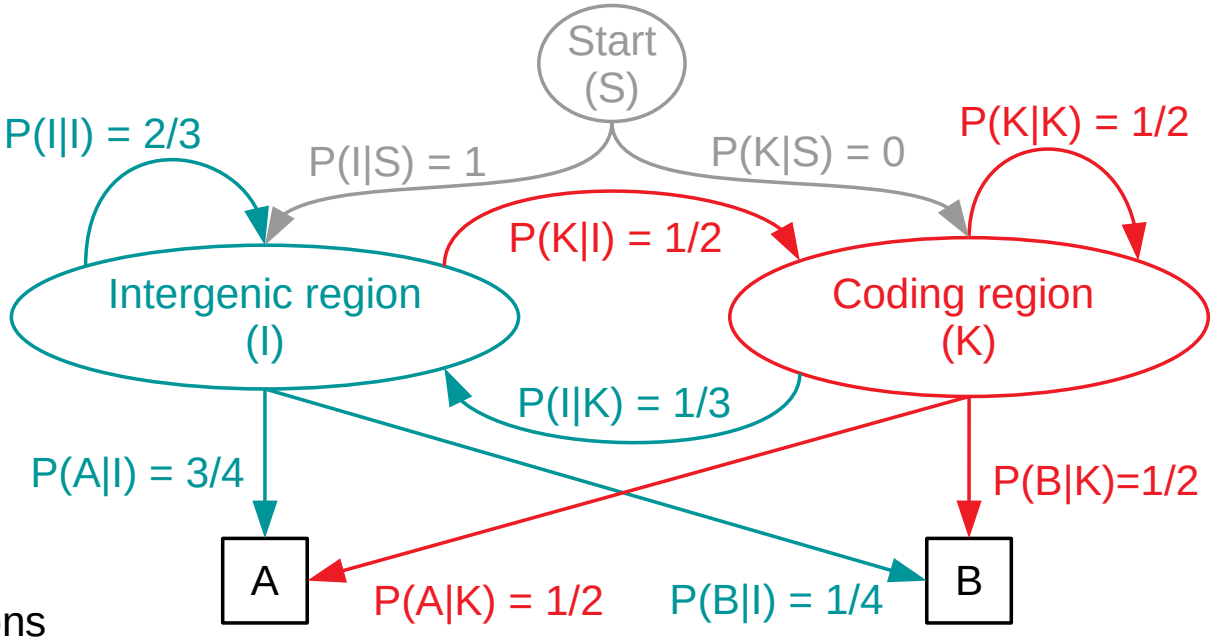
How likely are the observations?

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

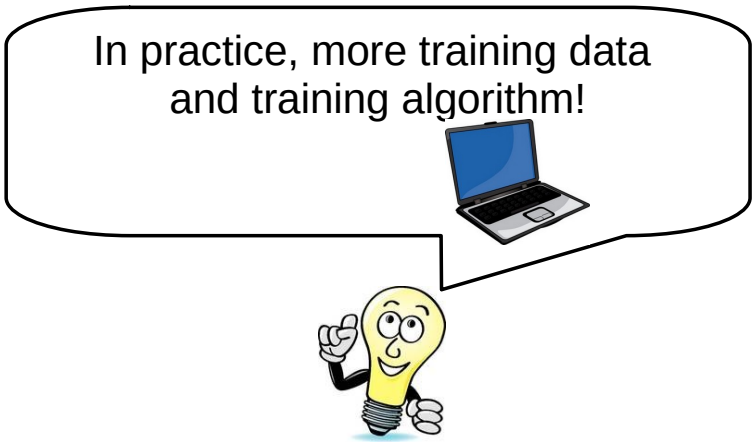


Hidden states

Visible observations (emissions)

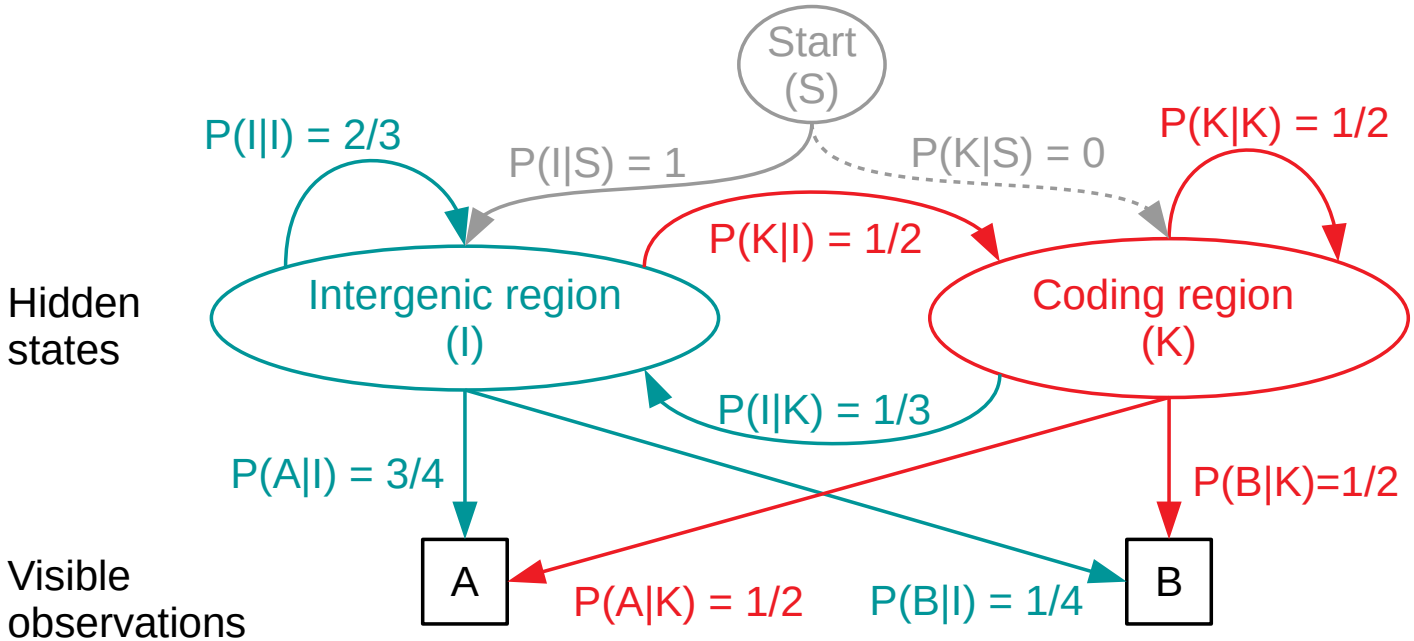
Training data:

AABABA
IKKIII



Basis of highly accurate gene prediction tools

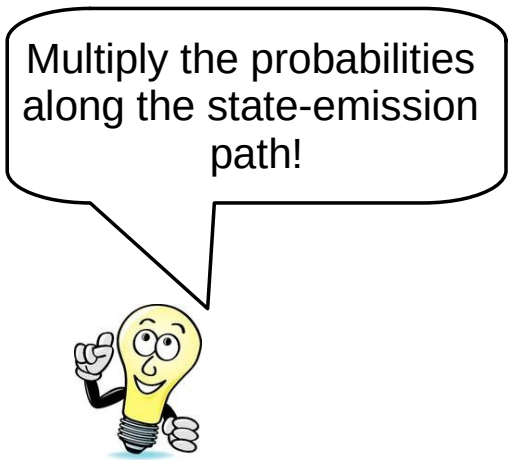
Hidden Markov Model



How likely is a given state-emission path?

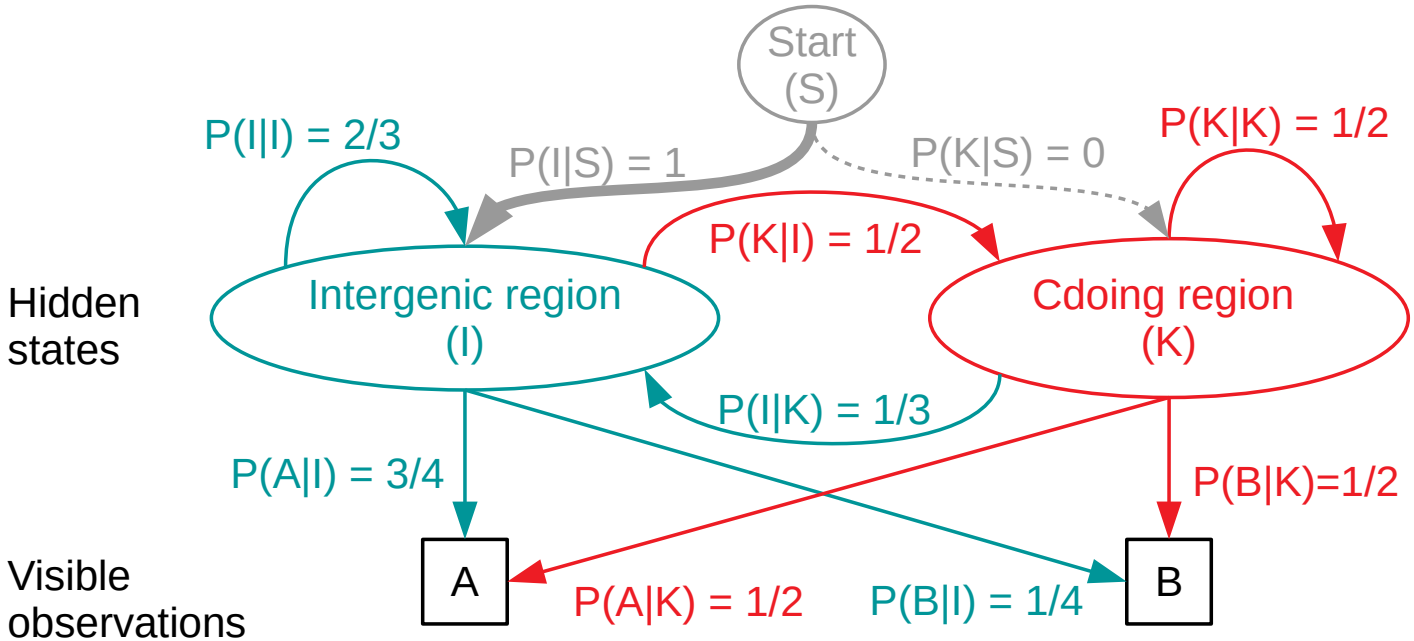
```
Path = AAB
       IKK

P(Path) = ?
```



Basis of highly accurate gene prediction tools

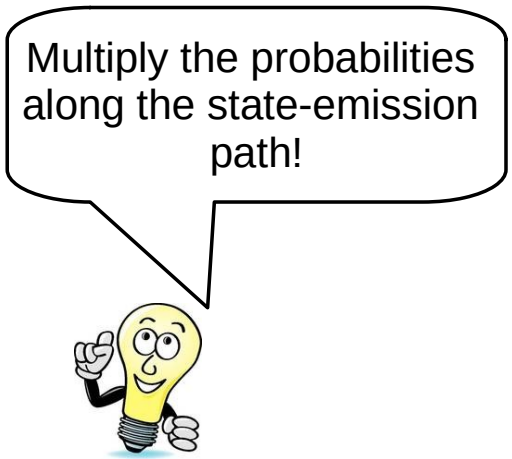
Hidden Markov Model



How likely is a given state-emission path?

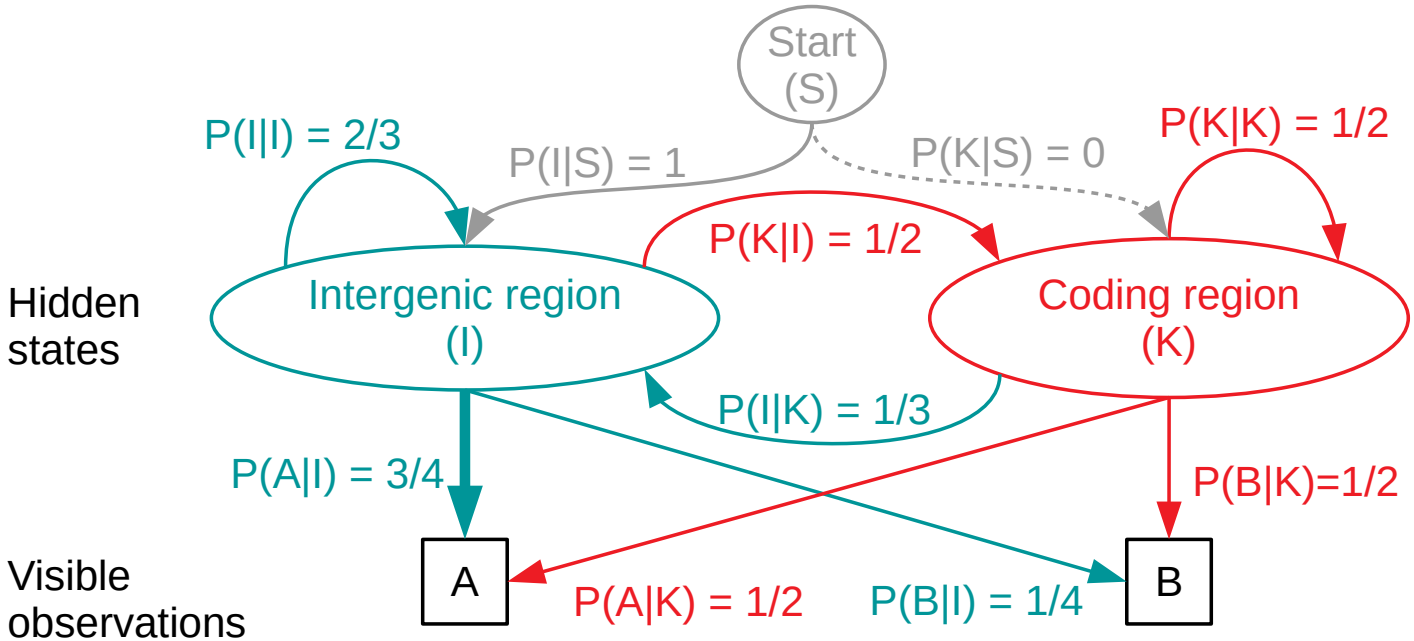
```
Path = AAB
       IKK

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



Basis of highly accurate gene prediction tools

Hidden Markov Model



How likely is a given state-emission path?

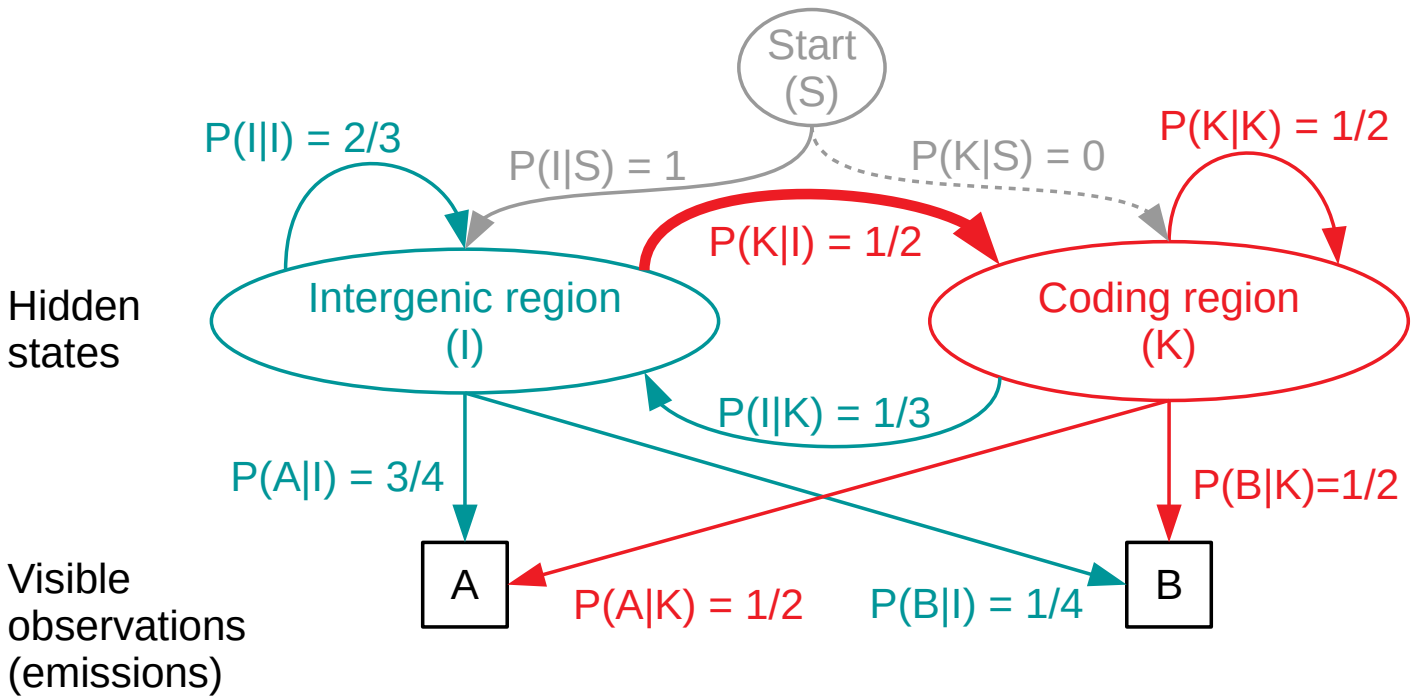
```
Path = AAB
       IKK

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



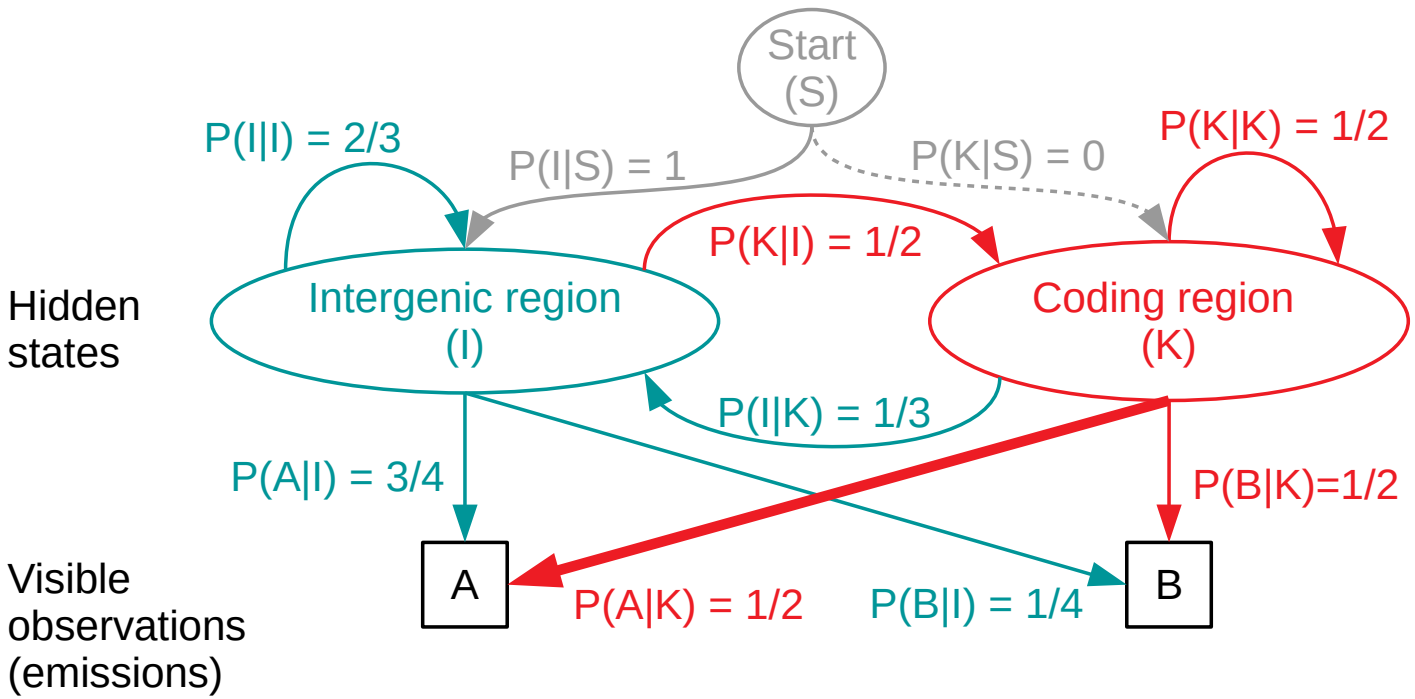
How likely is a given state-emission path?

```
Path = AAB
       IKK

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

Basis of highly accurate gene prediction tools

Hidden Markov Model



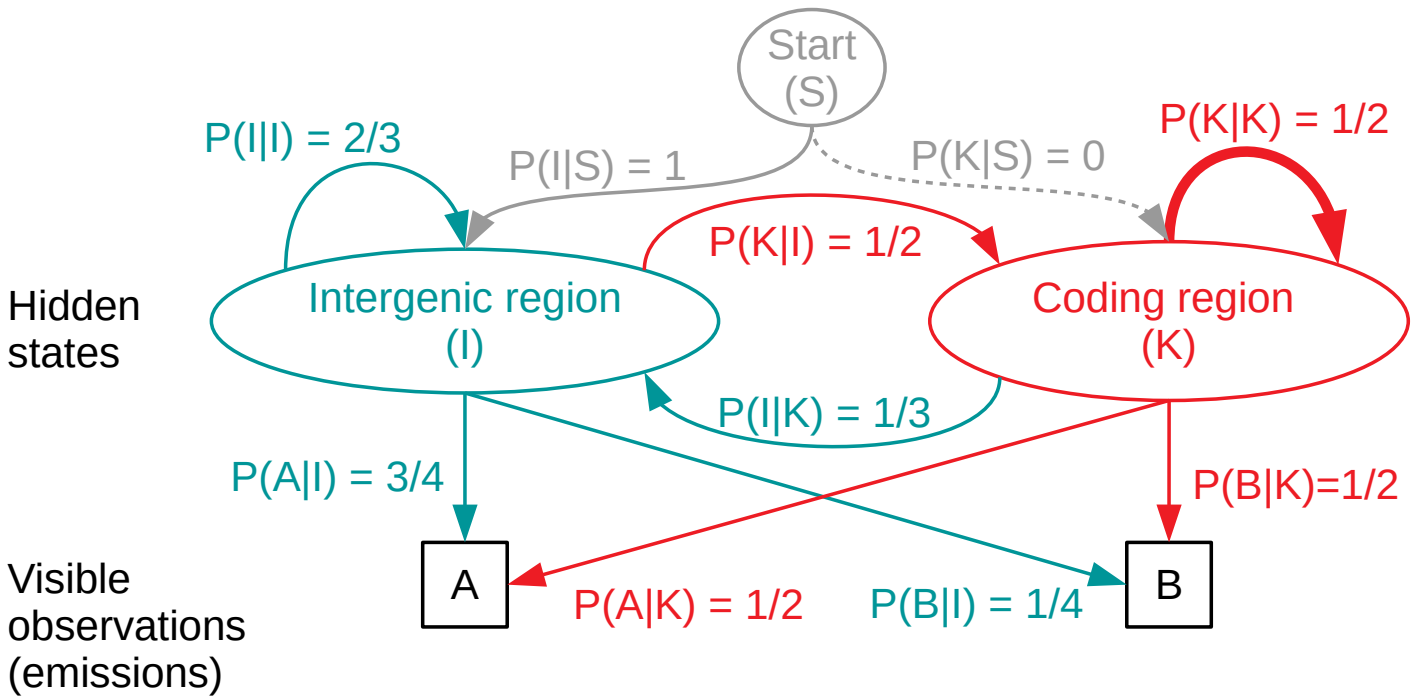
How likely is a given state-emission path?

```
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



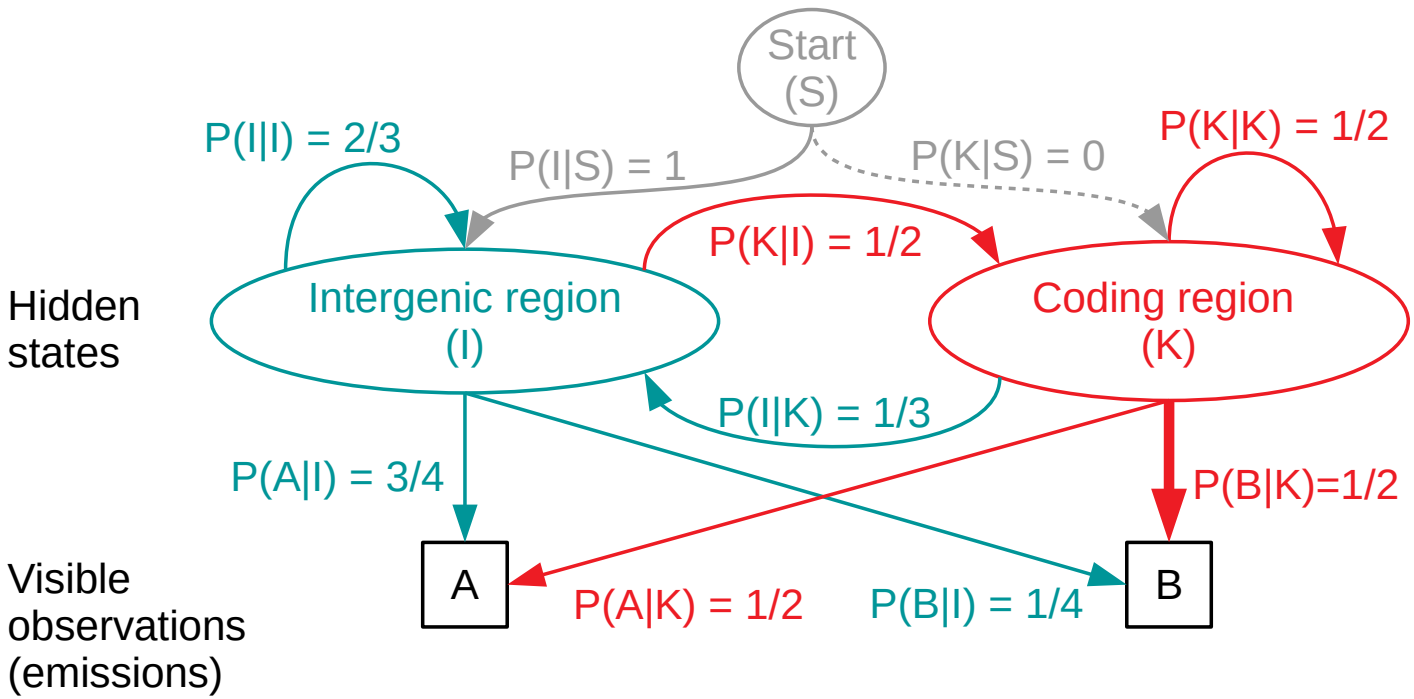
How likely is a given state-emission path?

```
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



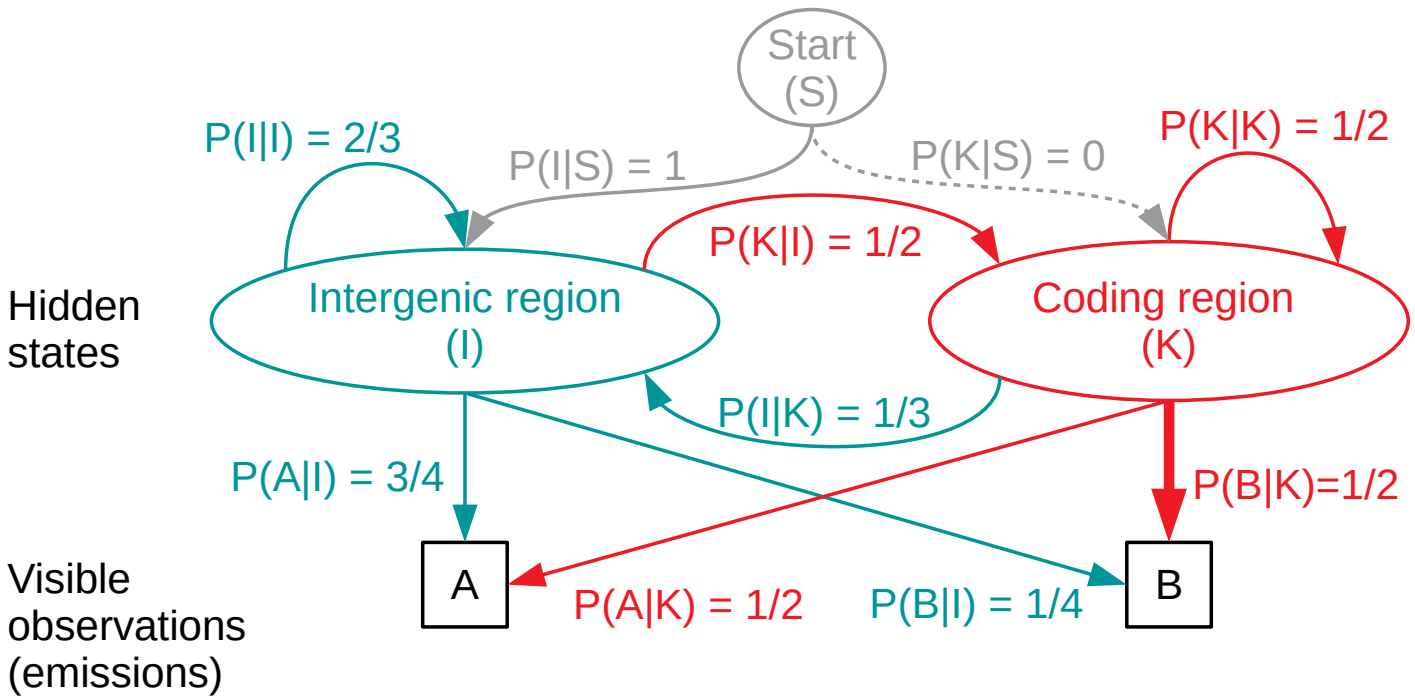
How likely is a given state-emission path?

```
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:

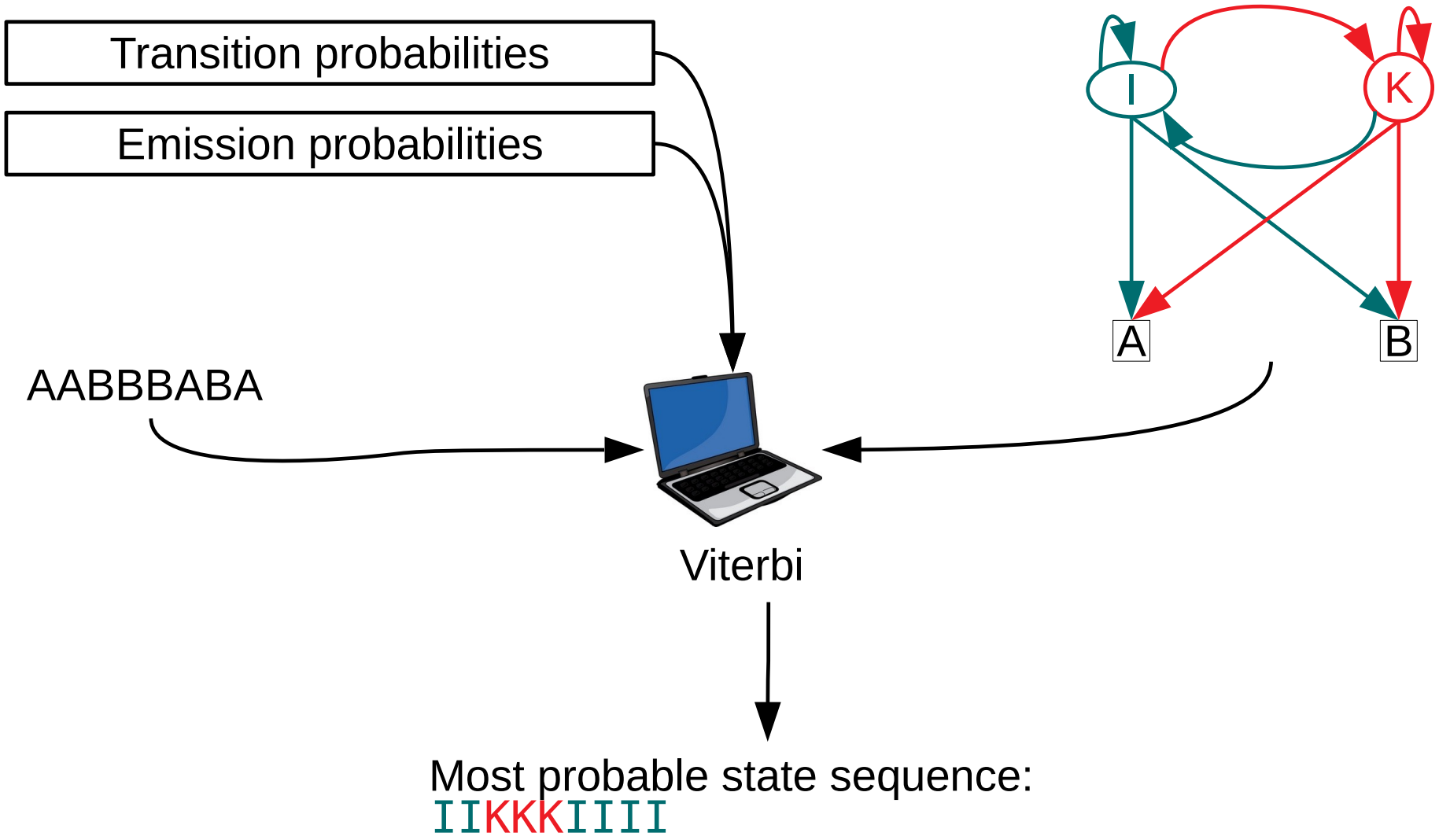
1. Generate all possible state sequences
2. Calculate the probability for each state sequence
3. Choose the state sequence with the highest probability

$\Rightarrow$ too expensive!

Basis of highly accurate gene prediction tools

Hidden Markov Model

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



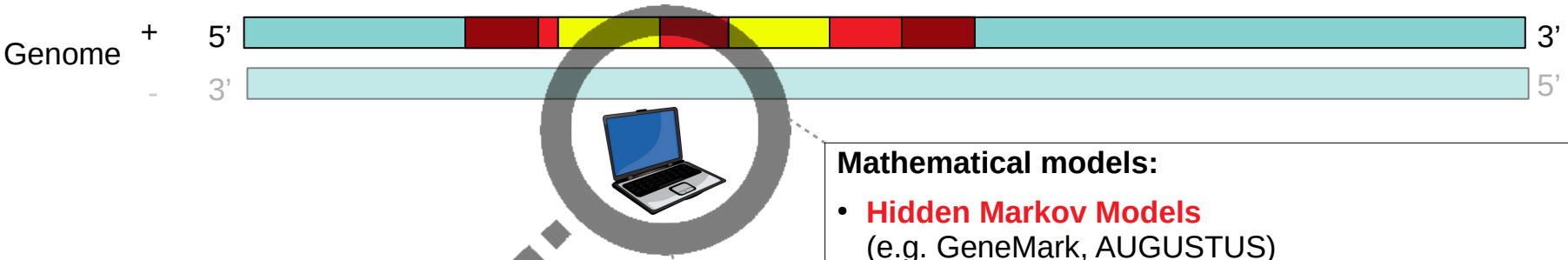
Hidden Markov Model for gene identification in practice

- 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:

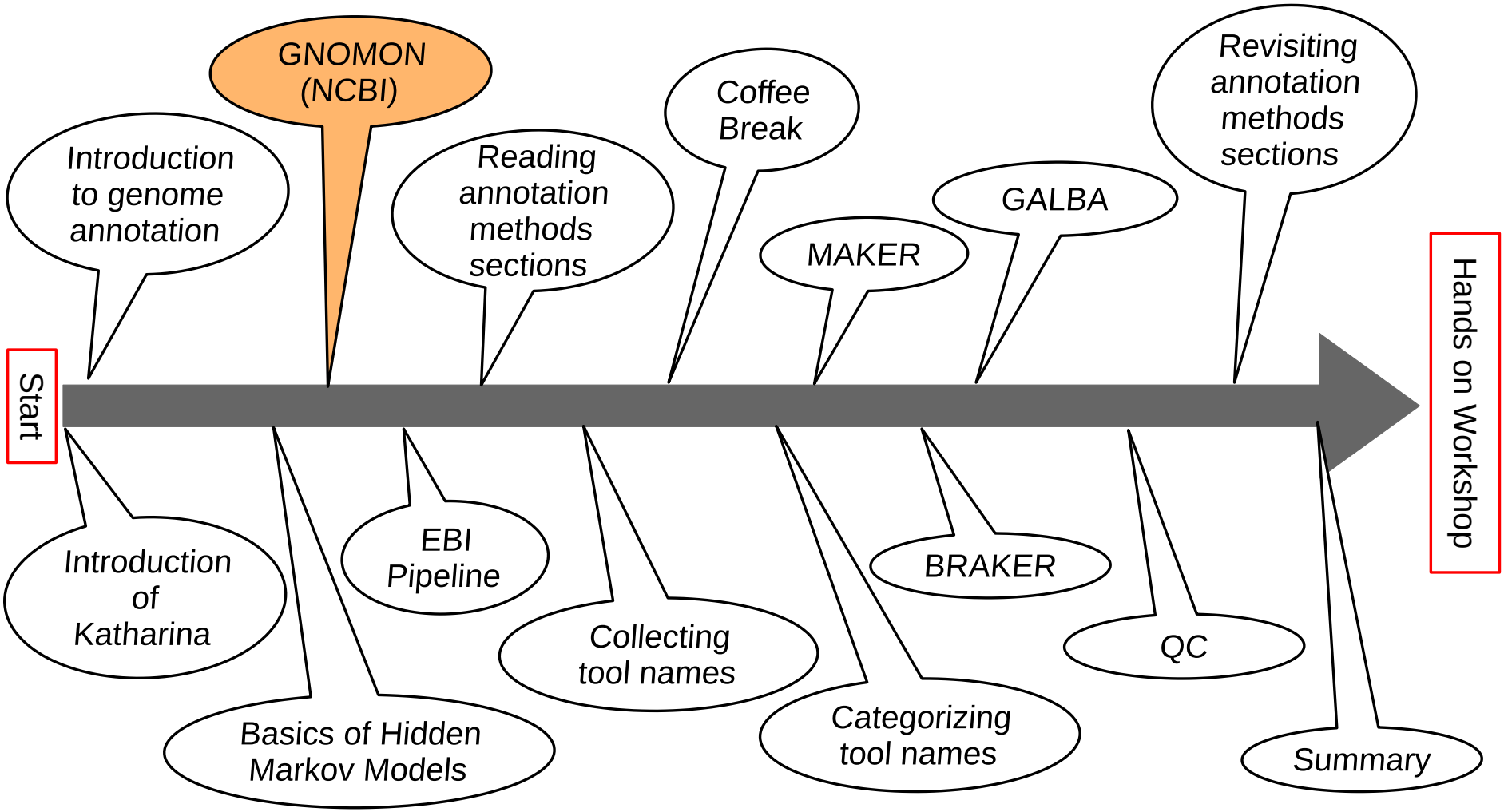
- **Hidden Markov Models** (e.g. GeneMark, AUGUSTUS)
- dynamic programming
- Support Vector Machines
- neural networks
- decision tree systems
- ...

Provide information on:

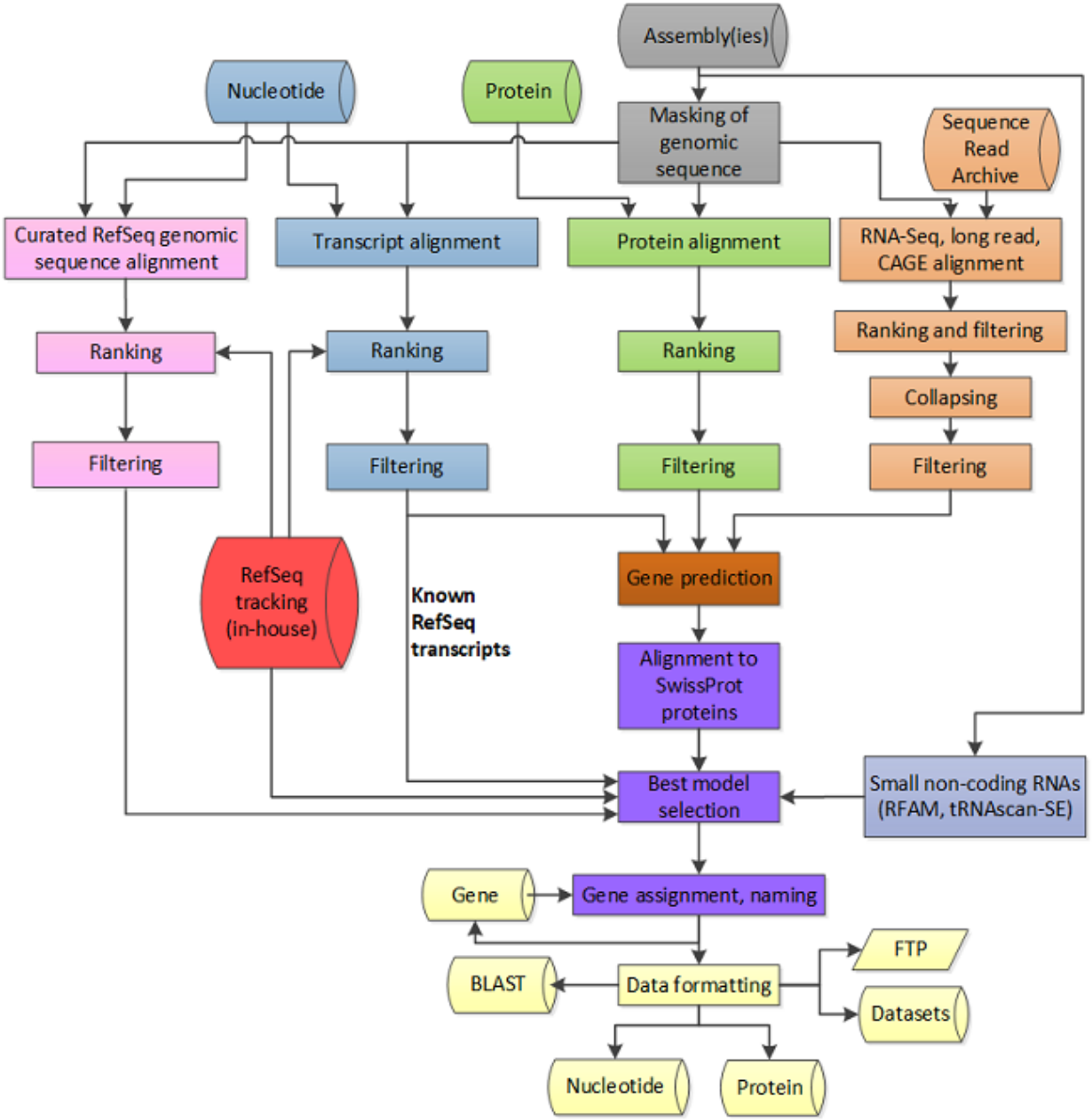
- complete gene structures (sometimes incl. UTRs)

Limitations

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


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

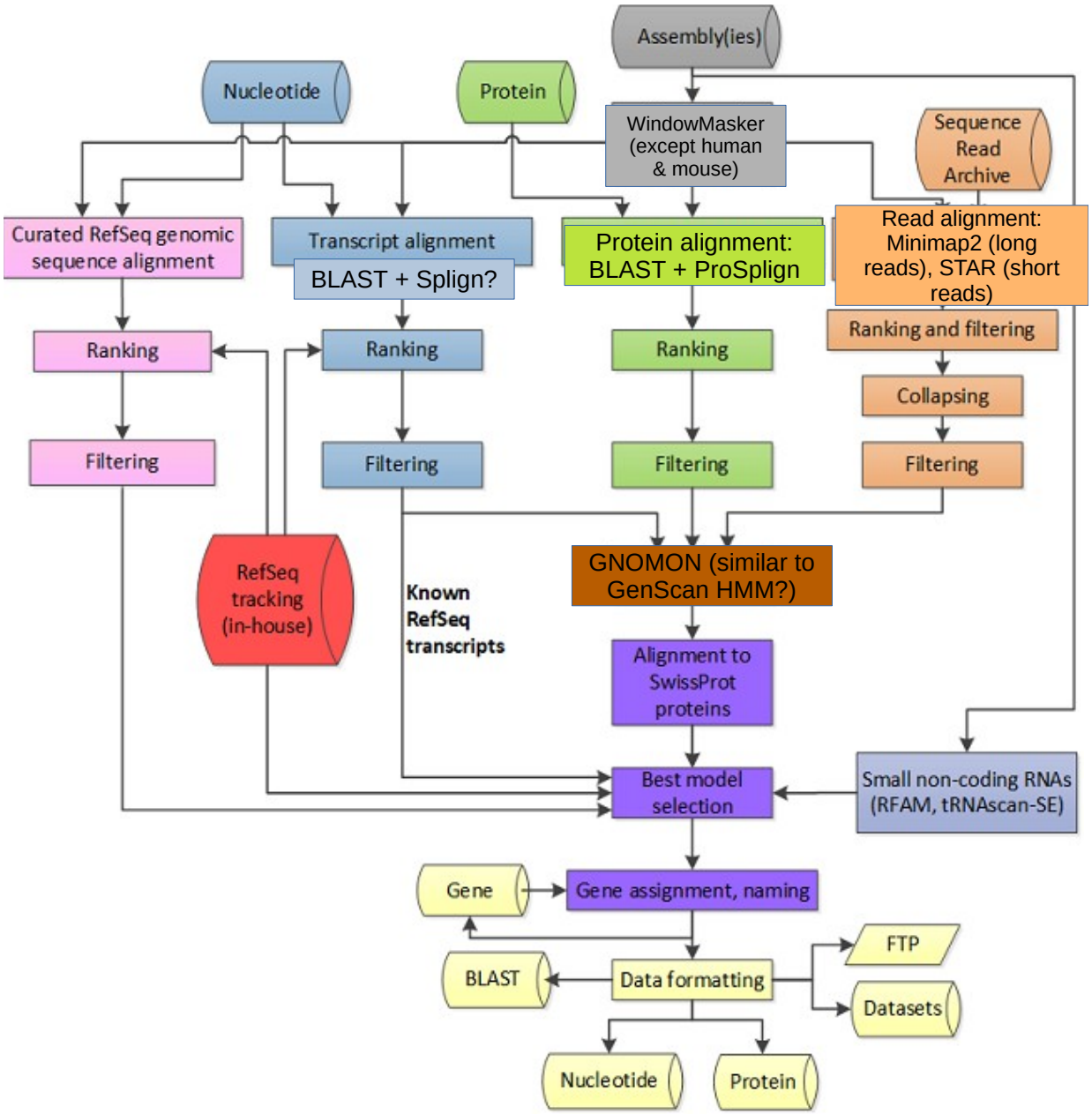
Annotation at the NCBI: GNOMON



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

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

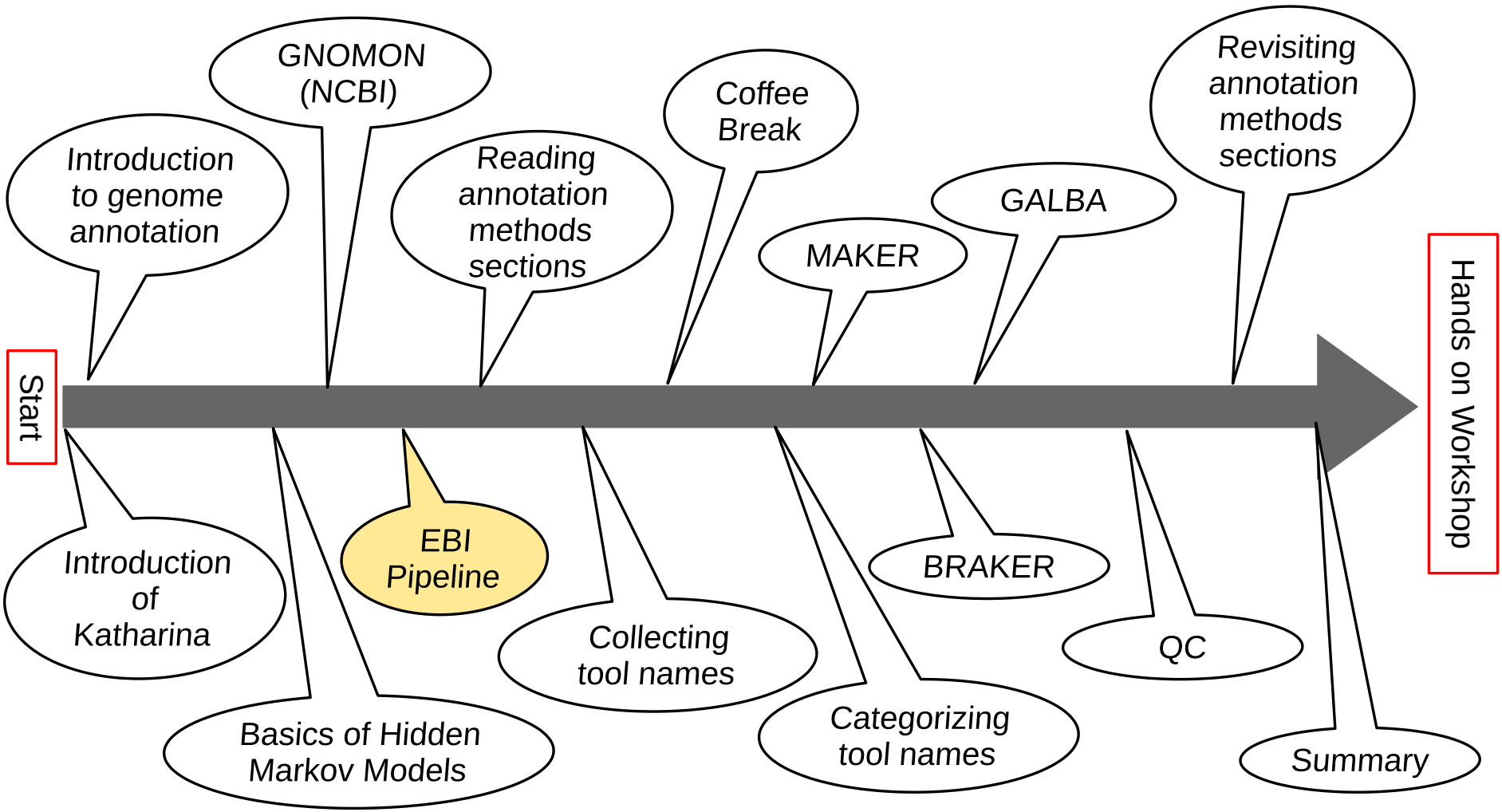
Annotation at the NCBI: GNOMON



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

Image adapted from: `https://www.ncbi.nlm.nih.gov/core/assets/genome/images/`

EBI: Ensembl annotation system



Documentation

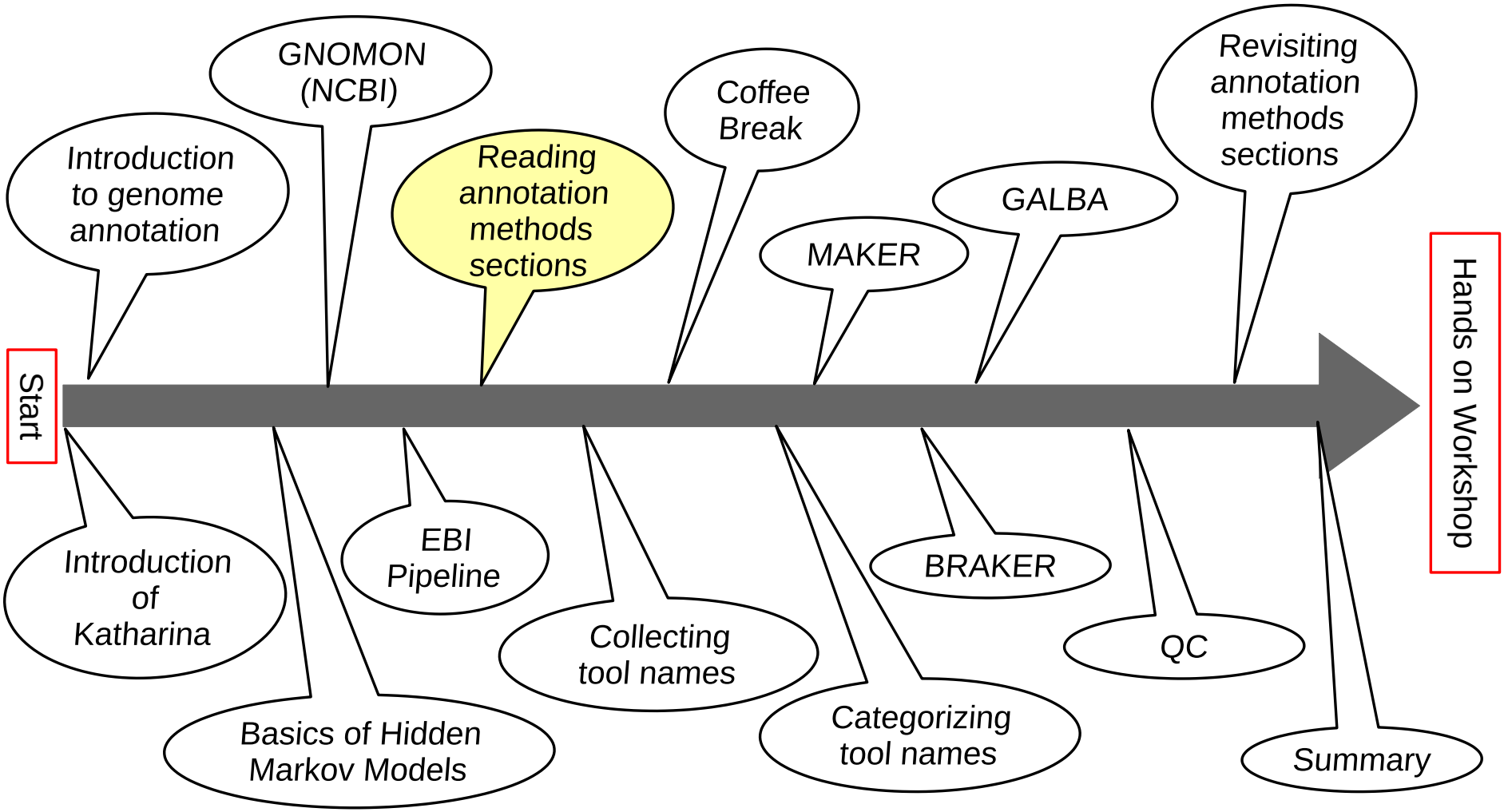
- Ensembl vertebrate pipeline: `https://rapid.ensembl.org/info/genome/genebuild/full_genebuild.html`
- Ensembl non-vertebrate pipeline: `https://rapid.ensembl.org/info/genome/genebuild/anno.html`
  BRAKER2 in Ensembl: `https://rapid.ensembl.org/info/genome/genebuild/braker.html`

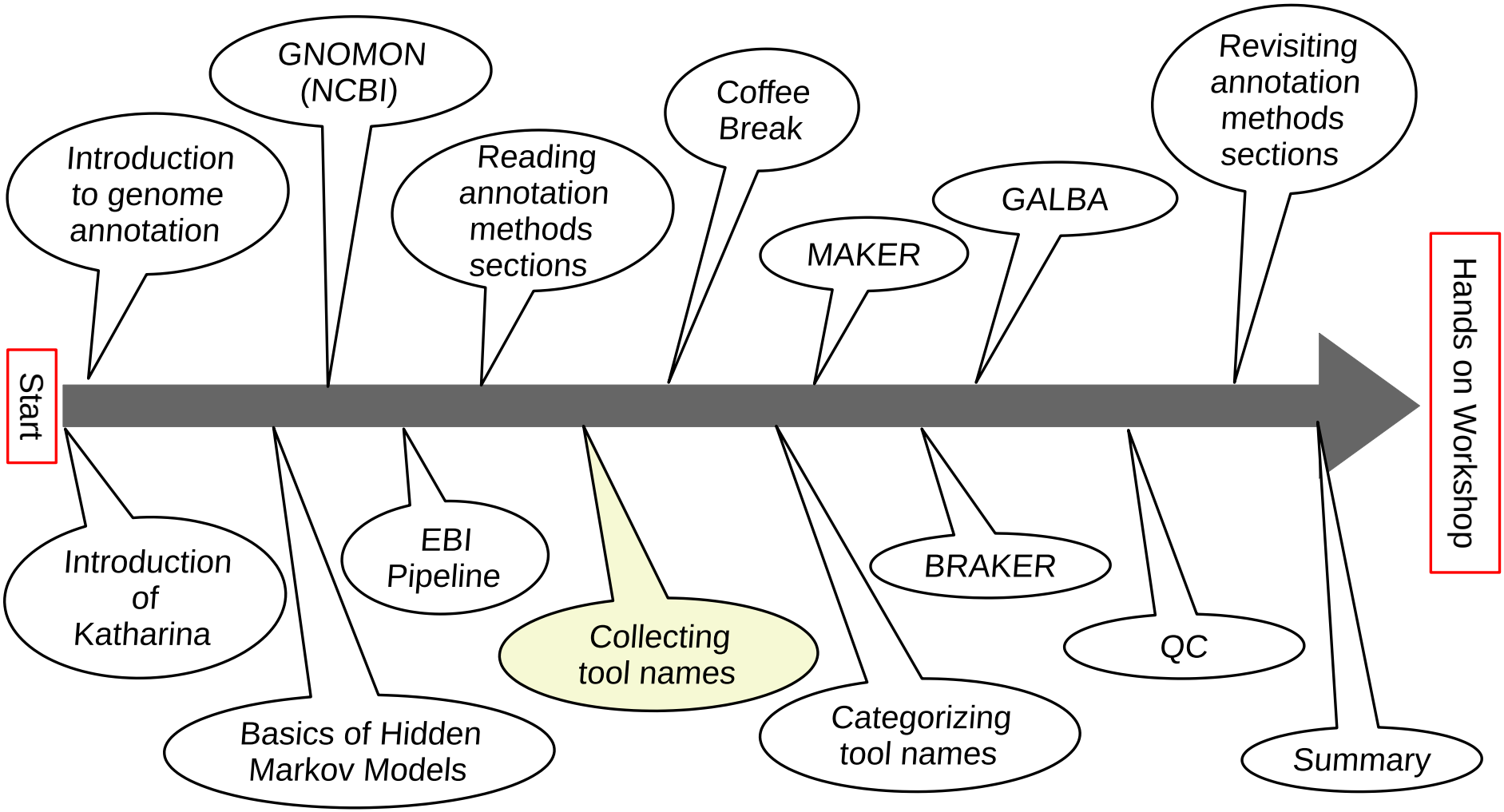
Where to find annotations

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

Notes by Katharina

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

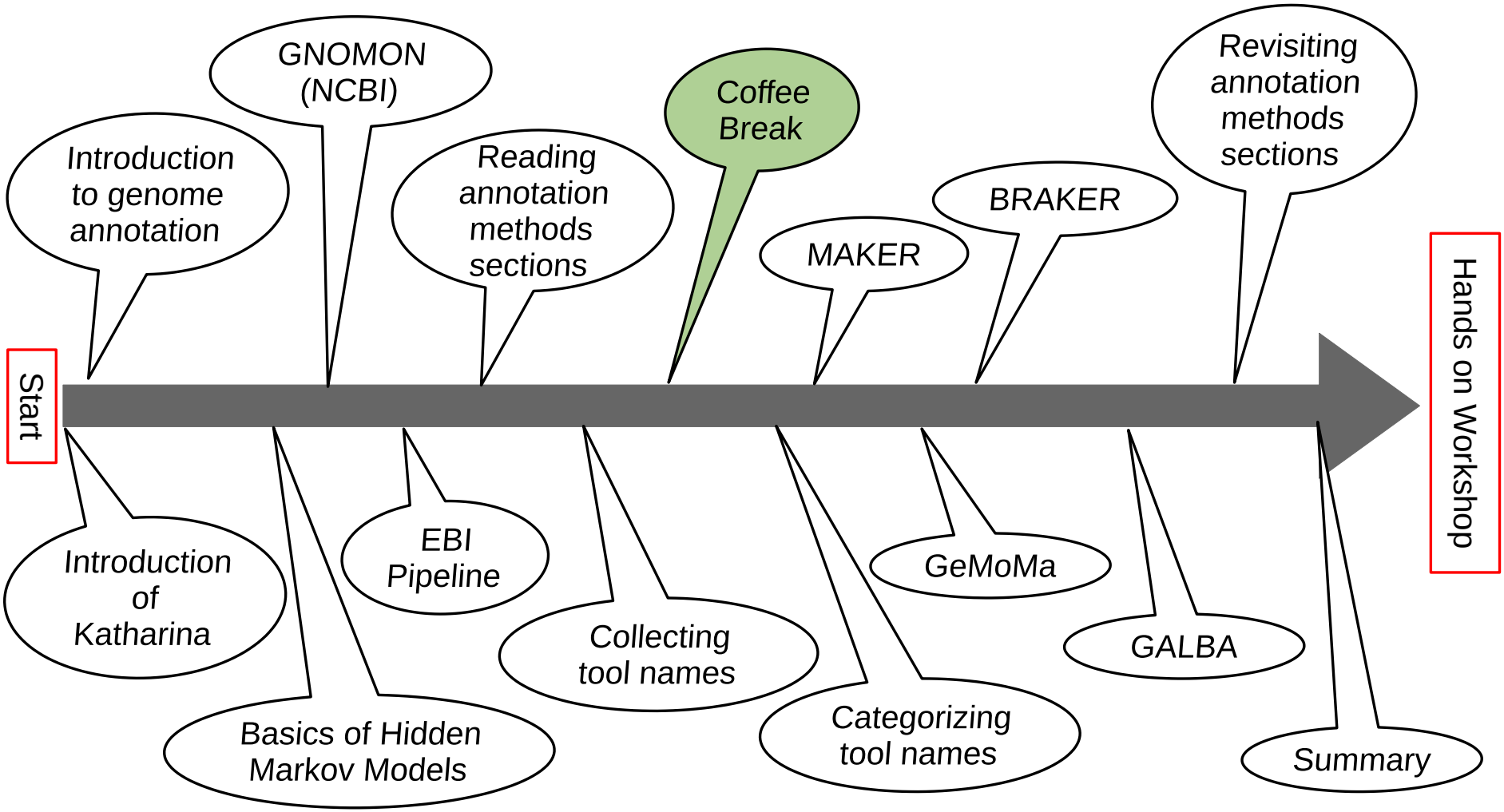

Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
QC
Summary

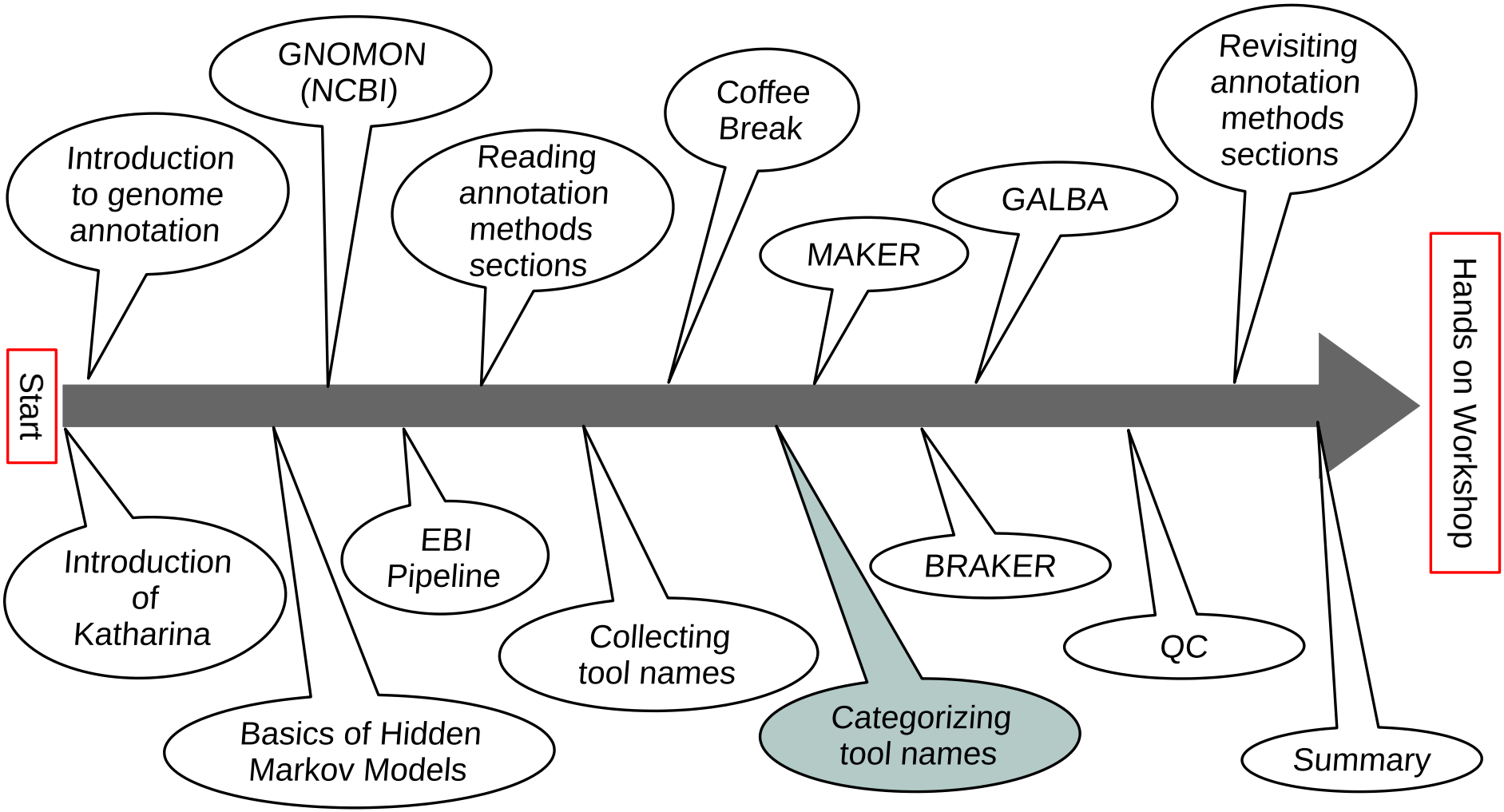

Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
QC
Summary

Read your methods snippet

Structural annotation of protein coding genes only!

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


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

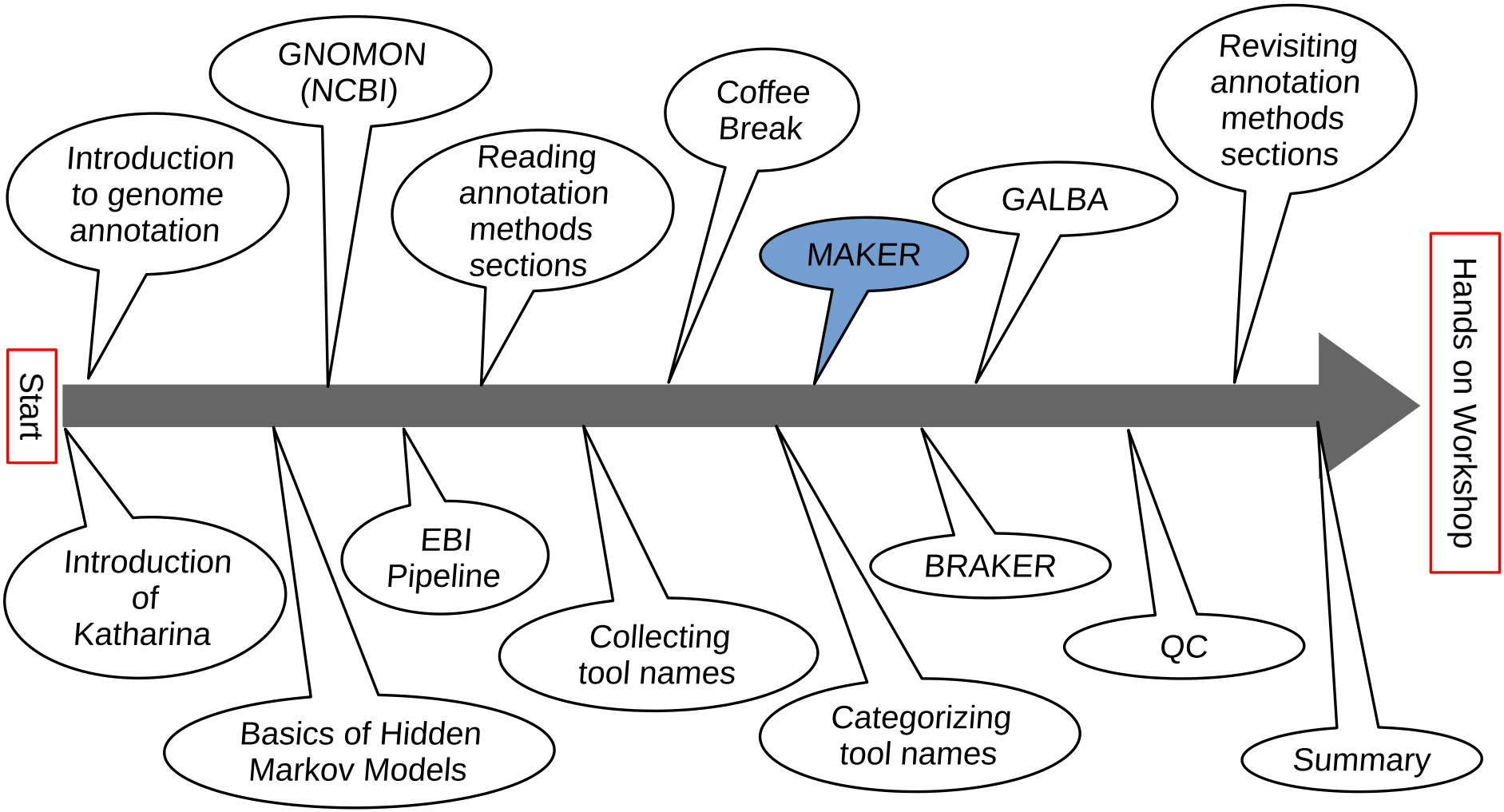

Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
QC
Summary

Categorize tool names

Go to

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

and sort the tools names from your methods snippet into categories


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

MAKER

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

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

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

Holt and Yandell *BMC Bioinformatics* 2011, **12**:491
http://www.biomedcentral.com/1471-2105/12/491

BMC Bioinformatics

SOFTWARE Open Access

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

MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations[1[W][OPEN]]

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

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

MAKER Workflow

MPI-enabled to allow parallel operation on large compute clusters

Ab initio prediction

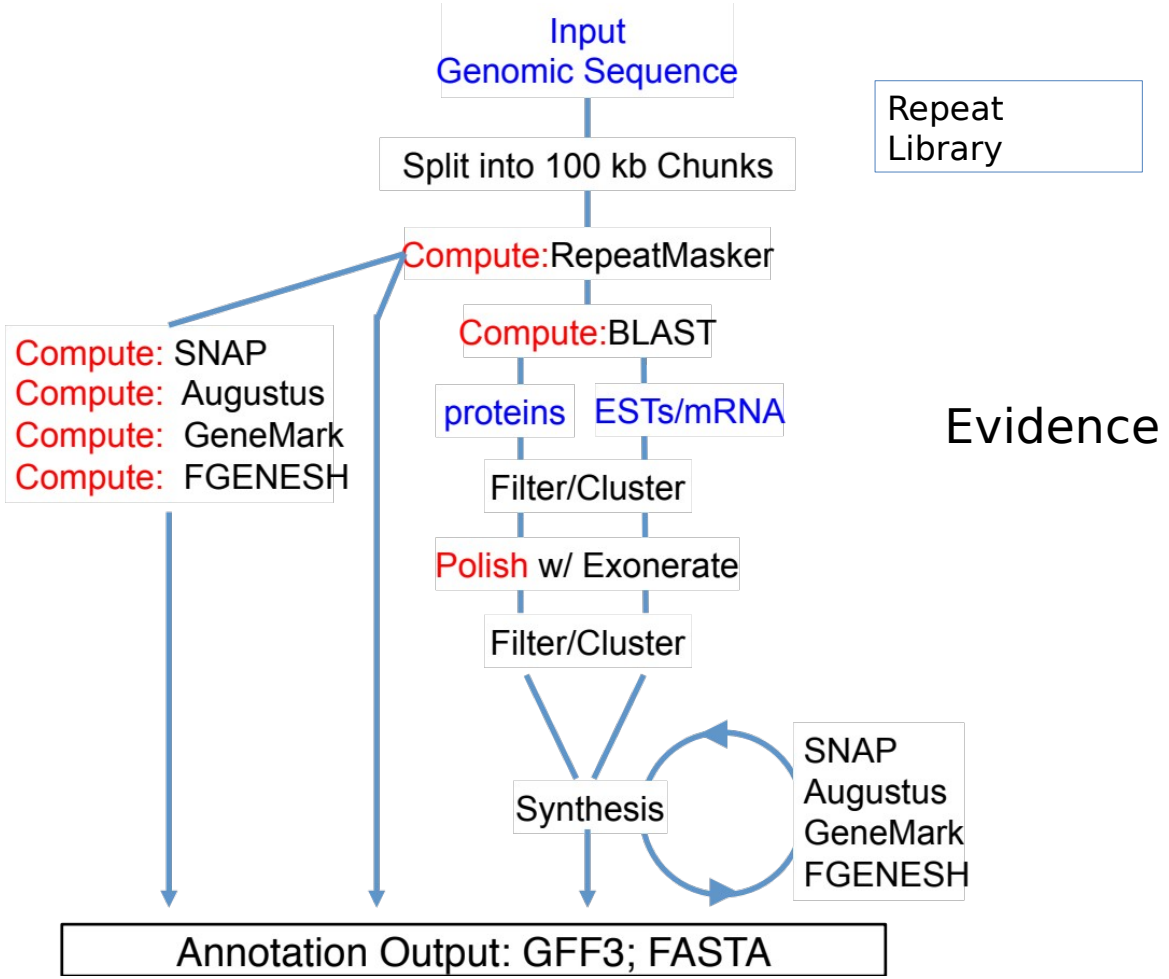


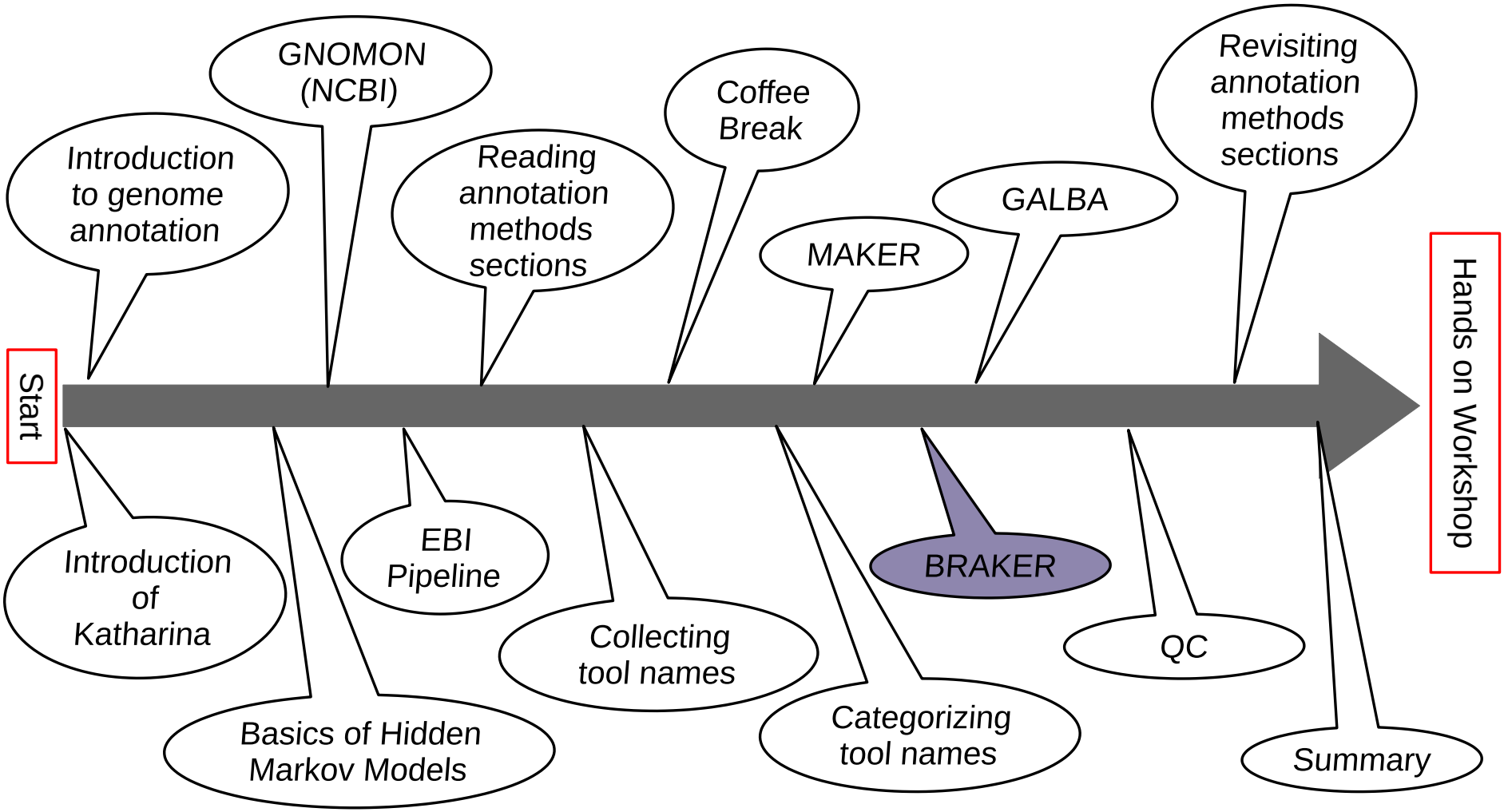
Image from slide 8 Cyverse Workshop on Genome Annotation w/MAKER hosted at https://player.slideplayer.com/download/91/14950916/UYt-fGaBni4baa3z4hU5gg/1704645110/14950916.ppt
authors unclear, Mark Yandell probably contributed the figure?

MAKER: important facts for users

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

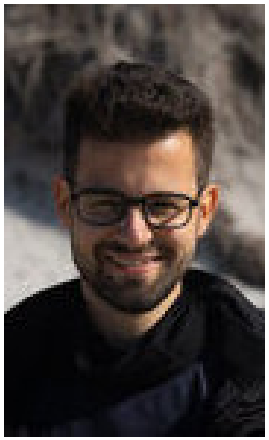
Improving the field of fully automated genome annotation

Image: credits to DALL-E2, modified by human


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

The BRAKER Team

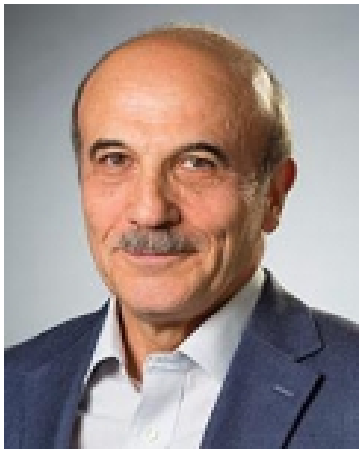
University of Greifswald & Georgia Tech University



Lars Gabriel

Alexandre Lomsadze, Katharina Hoff, Tomáš Brůna

Mario Stanke



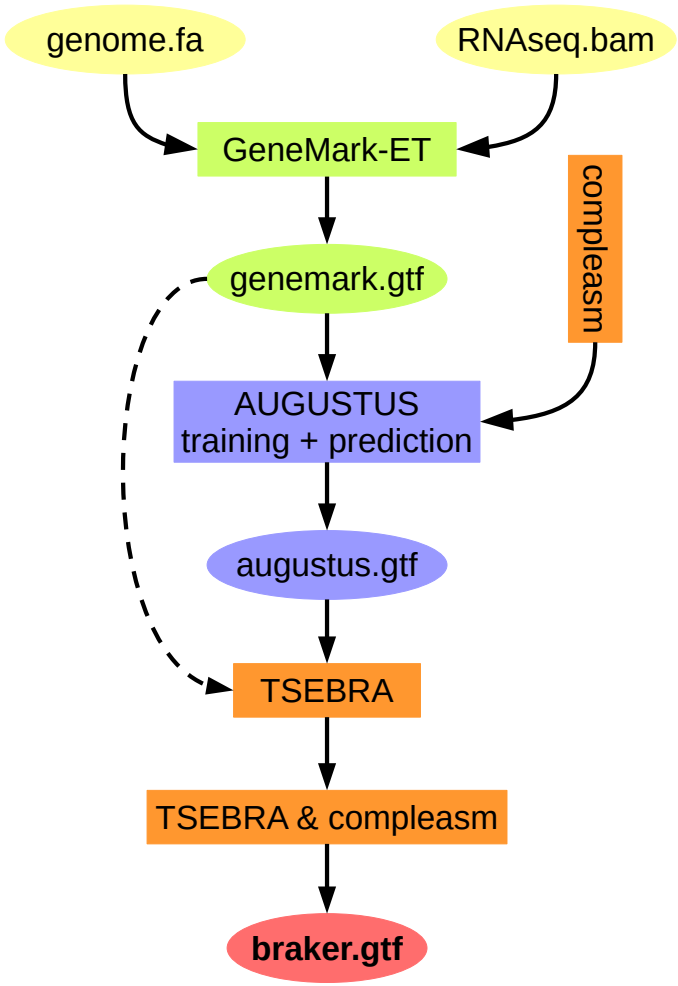
Mark Borodovsky

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
- 1,312 citations (Google Scholar)

Whole-Genome Annotation with BRAKER

Katharina J. Hoff, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke

in Kollmar M. (eds) Gene Prediction. Methods in Molecular Biology, vol 1962. Humana, New York, NY, 2019

GeneMark-ET uses RNA-Seq for Training

Anchors from RNA-Seq for training

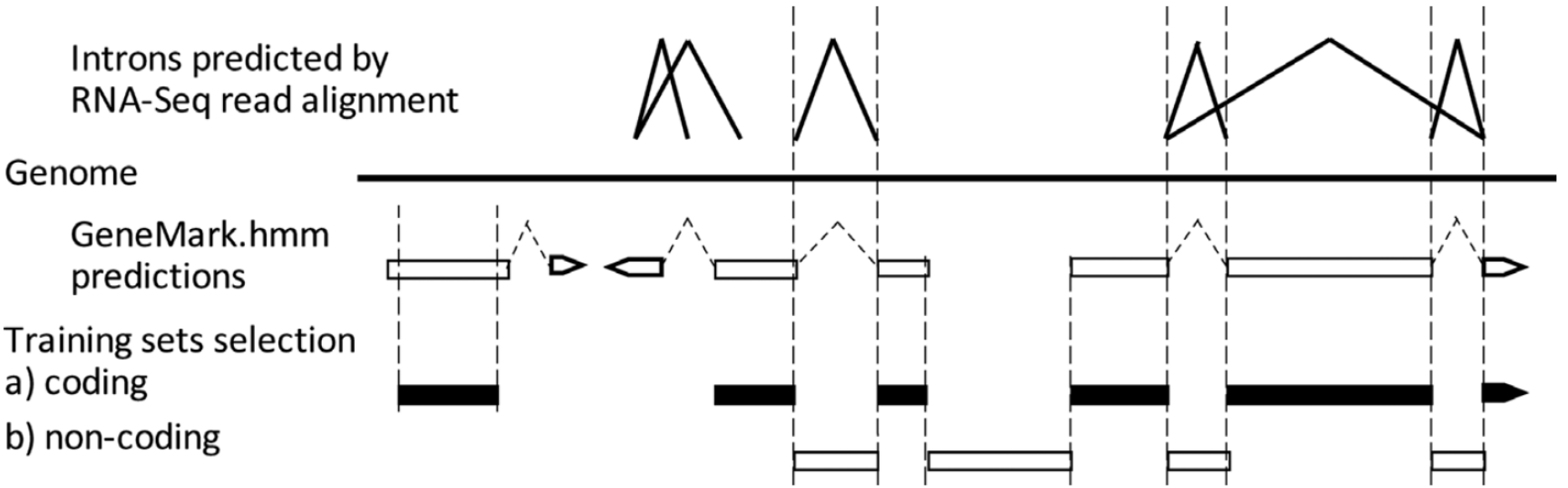
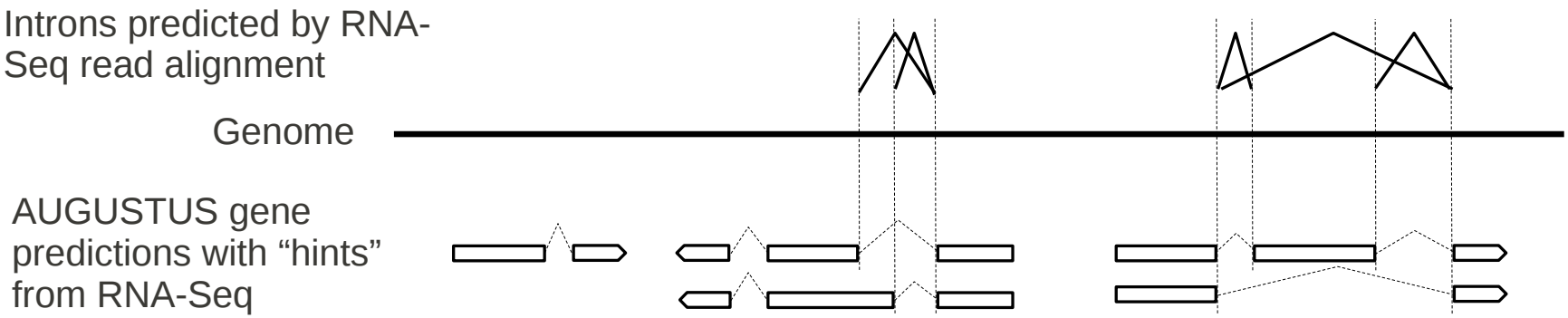


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- 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

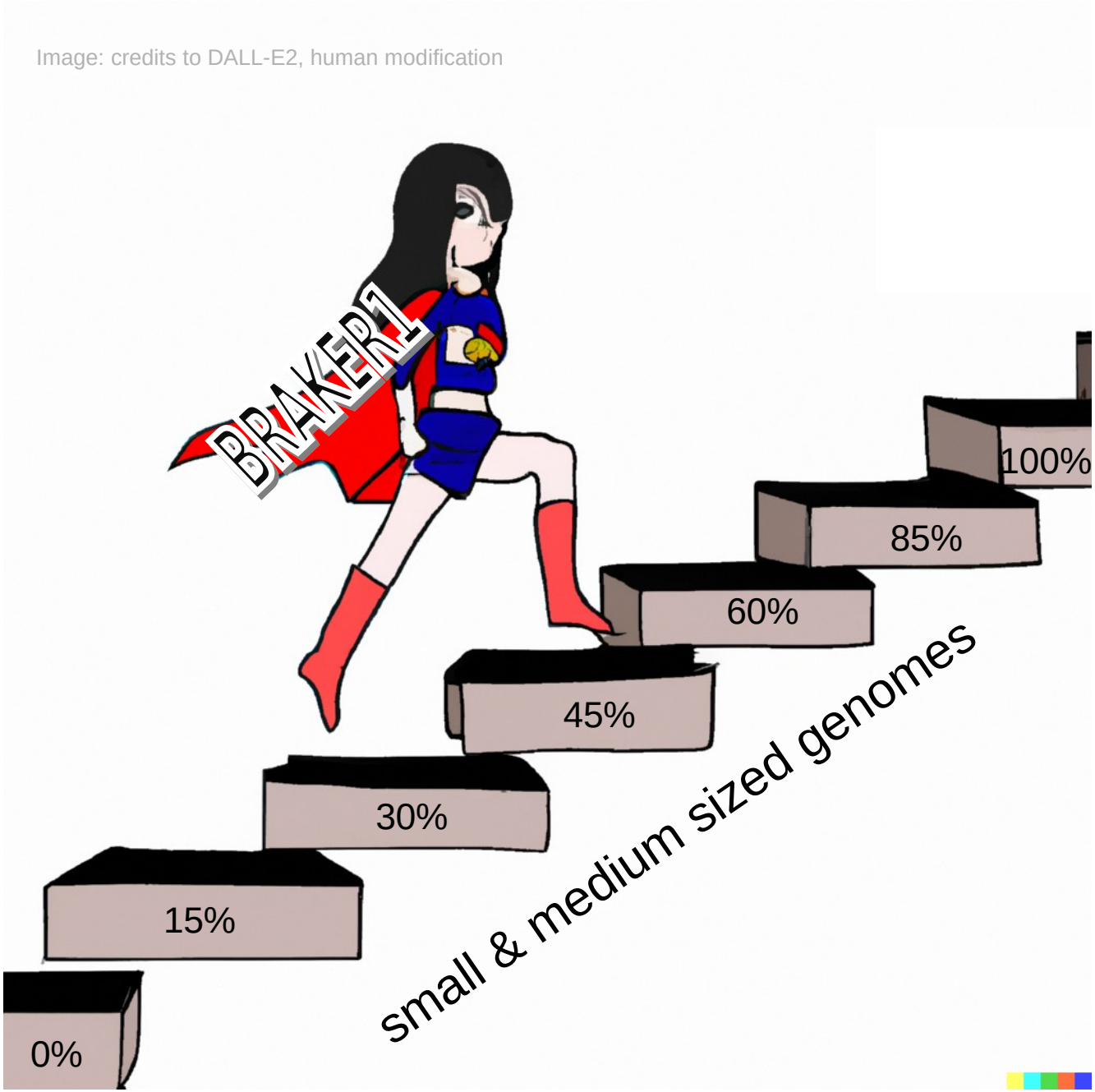
Accuracy metrics

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

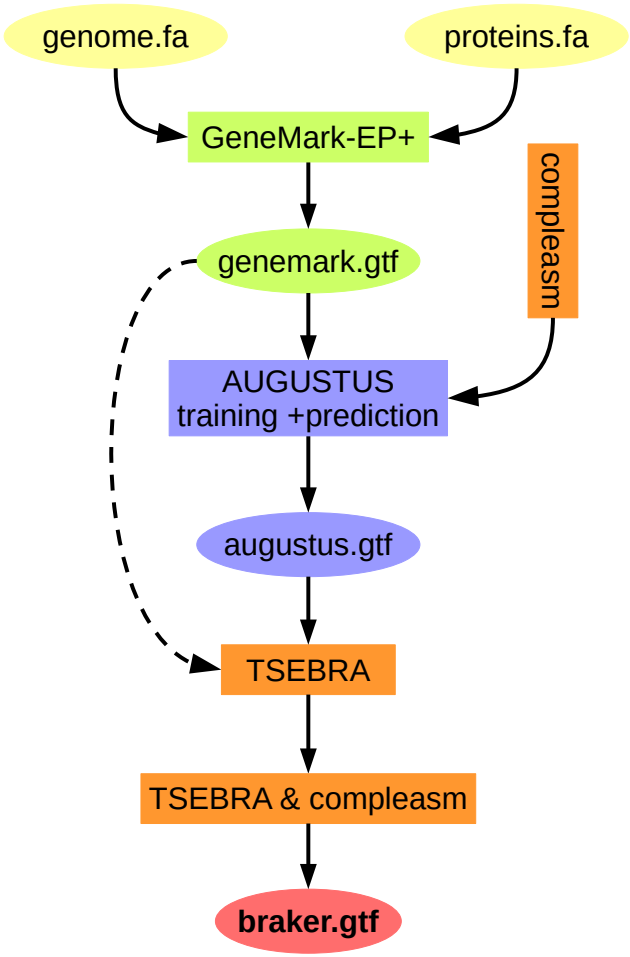
BRAKER1 gene F1 accuracy



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

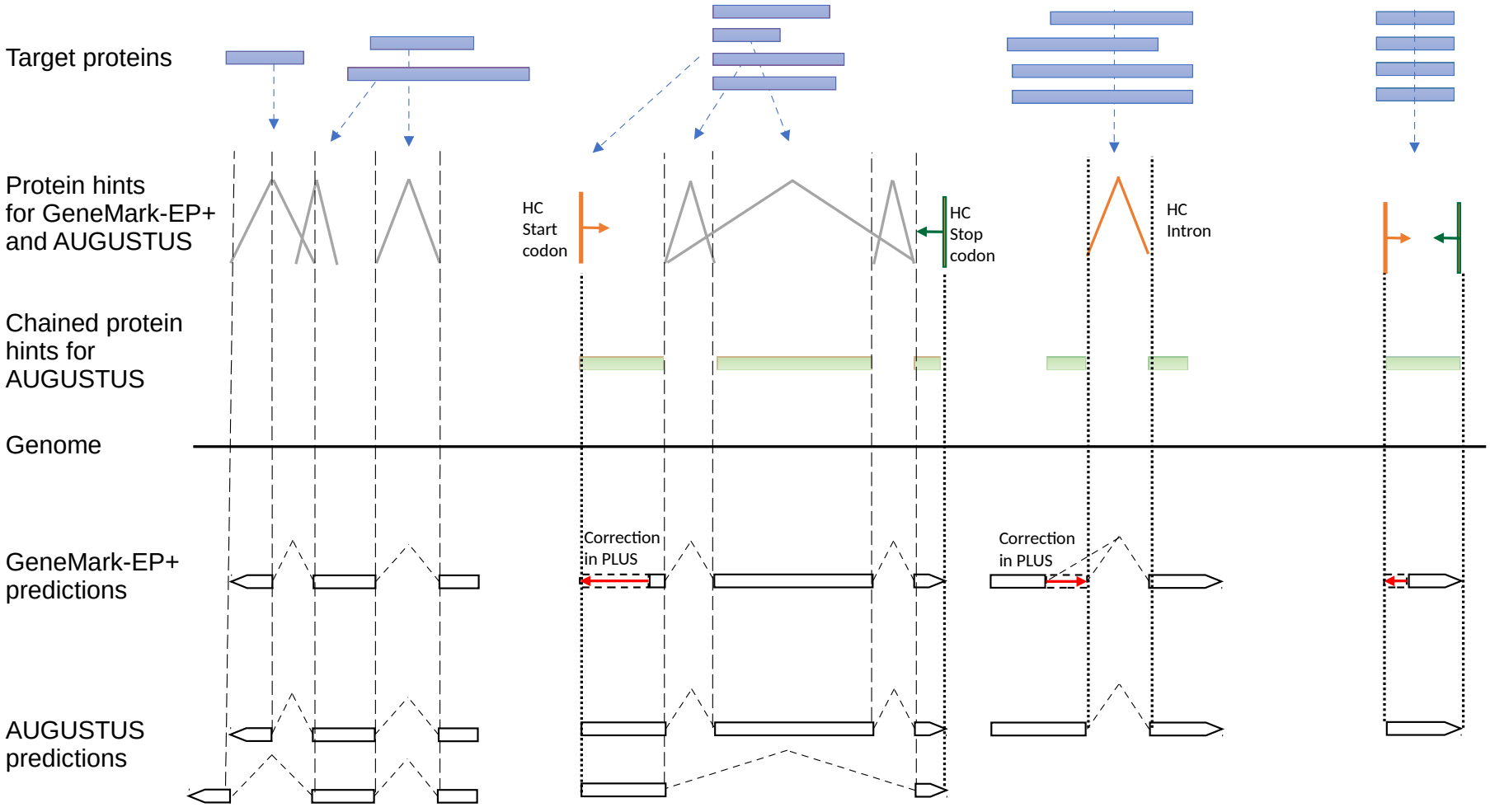
BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database

Tomáš Brůna[1,†], Katharina J. Hoff[2,3,†], Alexandre Lomsadze[4], Mario Stanke[2,3,‡] and Mark Borodovsky[4,5,*,‡]

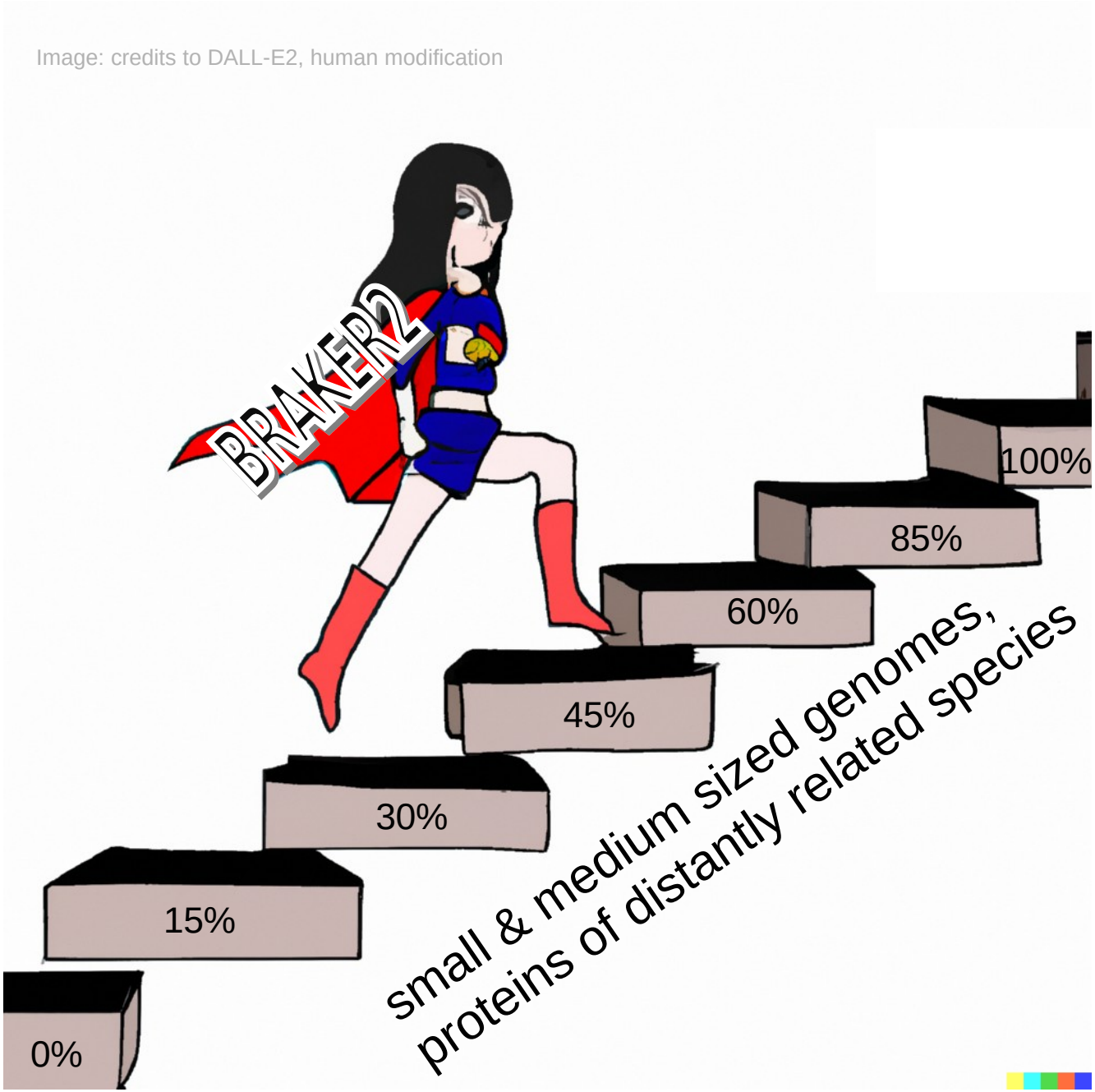


- spliced alignments of a large number of proteins (e.g. OrthoDB partition)
- optional input: BUSCO lineage (compleasm)
- 719 citations (Google Scholar)

Evidence usage by GeneMark-EP+ & AUGUSTUS during prediction

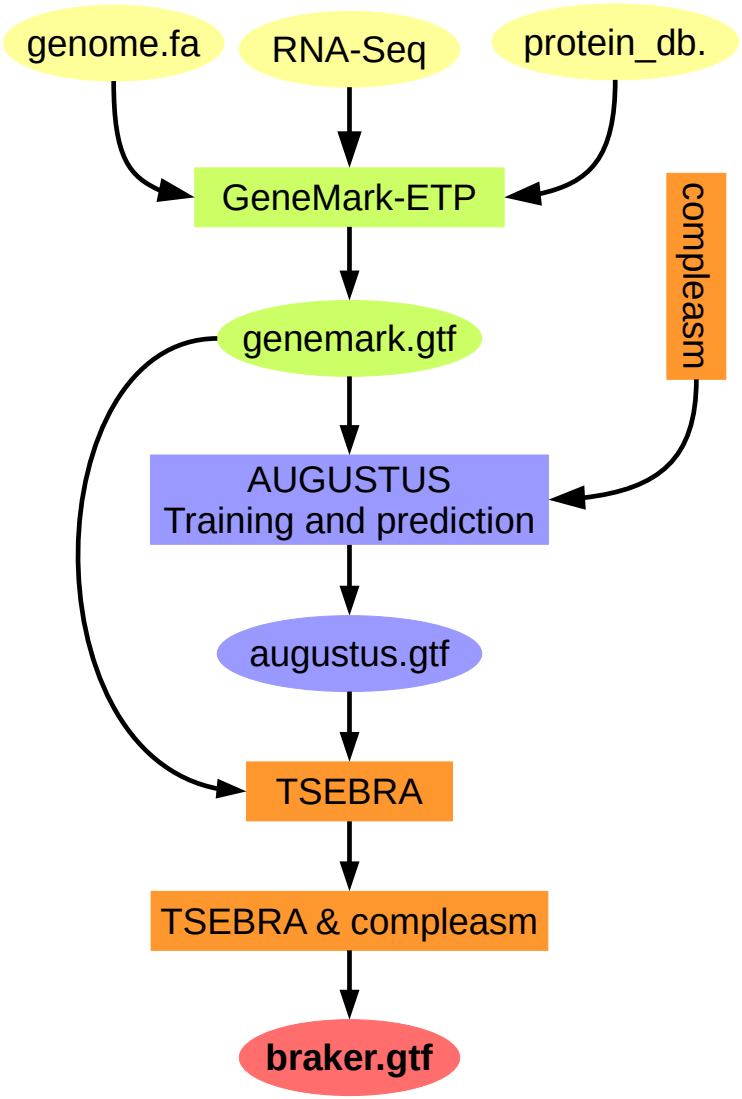


BRAKER2 gene F1 accuracy



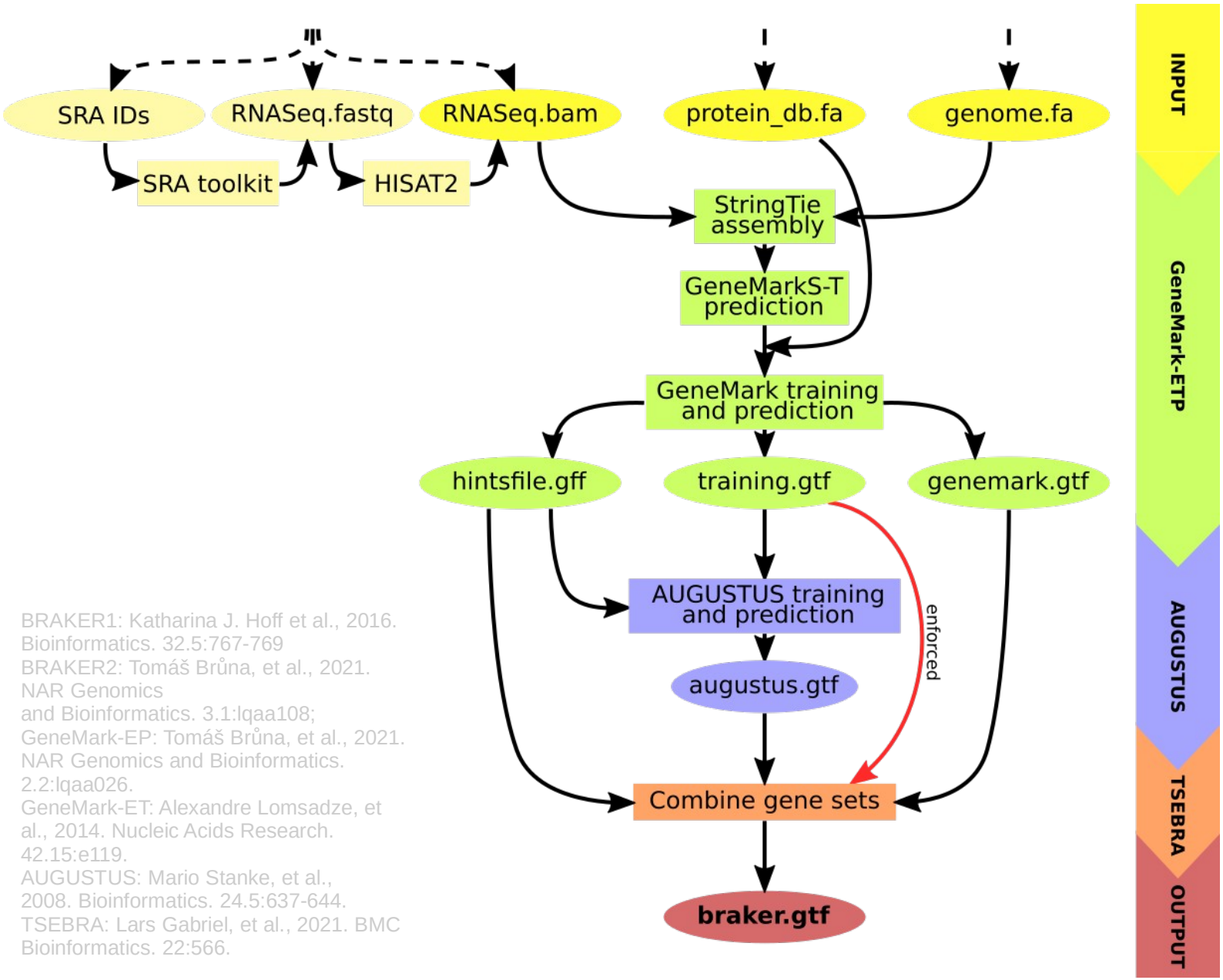
BRAKER3 gene F1 accuracy - climbing the top

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



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

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



BRAKER1: Katharina J. Hoff et al., 2016. Bioinformatics. 32.5:767-769
BRAKER2: Tomáš Brůna, et al., 2021. NAR Genomics and Bioinformatics. 3.1:lqaa108;
GeneMark-EP: Tomáš Brůna, et al., 2021. NAR Genomics and Bioinformatics. 2.2:lqaa026.
GeneMark-ET: Alexandre Lomsadze, et al., 2014. Nucleic Acids Research. 42.15:e119.
AUGUSTUS: Mario Stanke, et al., 2008. Bioinformatics. 24.5:637-644.
TSEBRA: Lars Gabriel, et al., 2021. BMC Bioinformatics. 22:566.

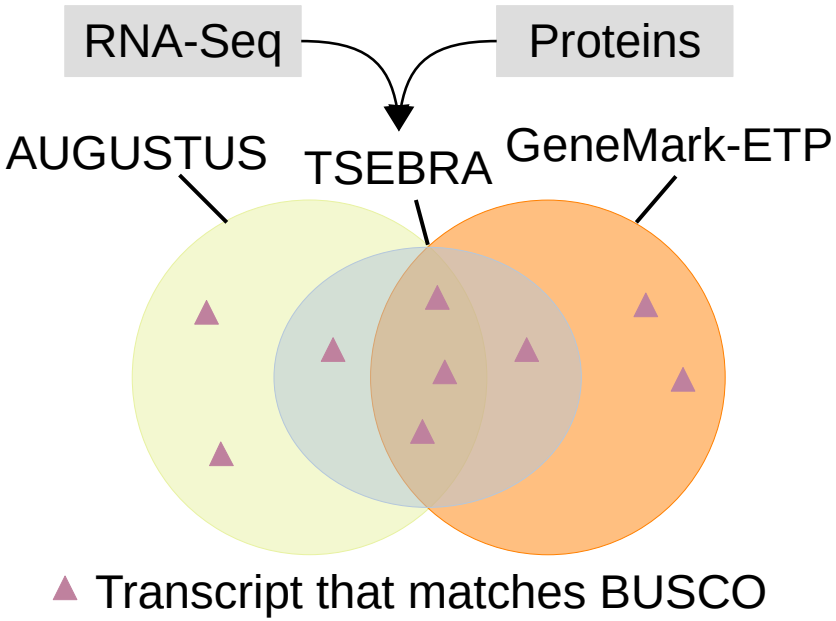
Gabriel *et al. BMC Bioinformatics* (2021) 22:566
https://doi.org/10.1186/s12859-021-04482-0

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*]



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

compleasm

Bioinformatics, 2023, **39(10)**, btad595
https://doi.org/10.1093/bioinformatics/btad595
Advance Access Publication Date: 27 September 2023
Applications Note

OXFORD

Genome analysis

compleasm: a faster and more accurate reimplementation of BUSCO

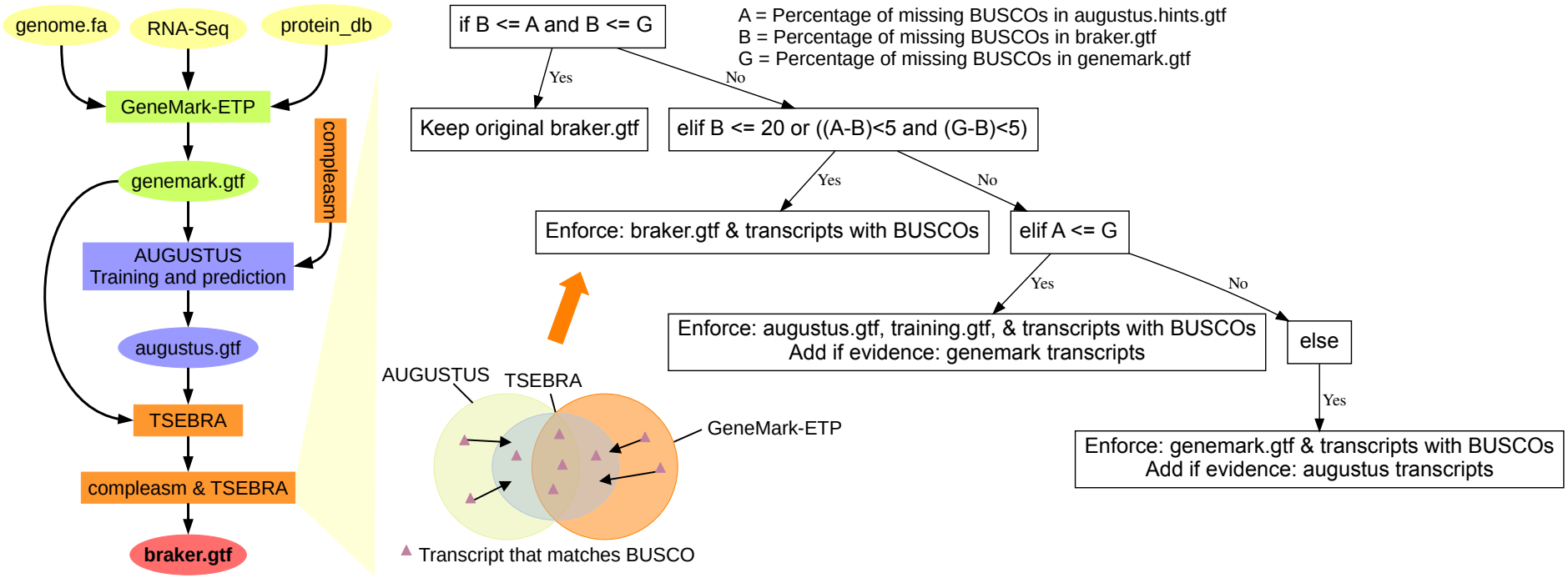
Neng Huang [1,2] and Heng Li[1,2,*]

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

$\Rightarrow$ This can solve our BRAKER-BUSCO problem

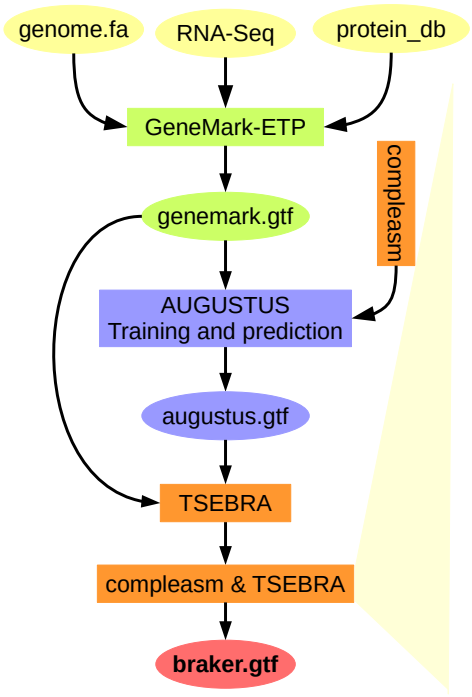
Improving BRAKER with Compleasm

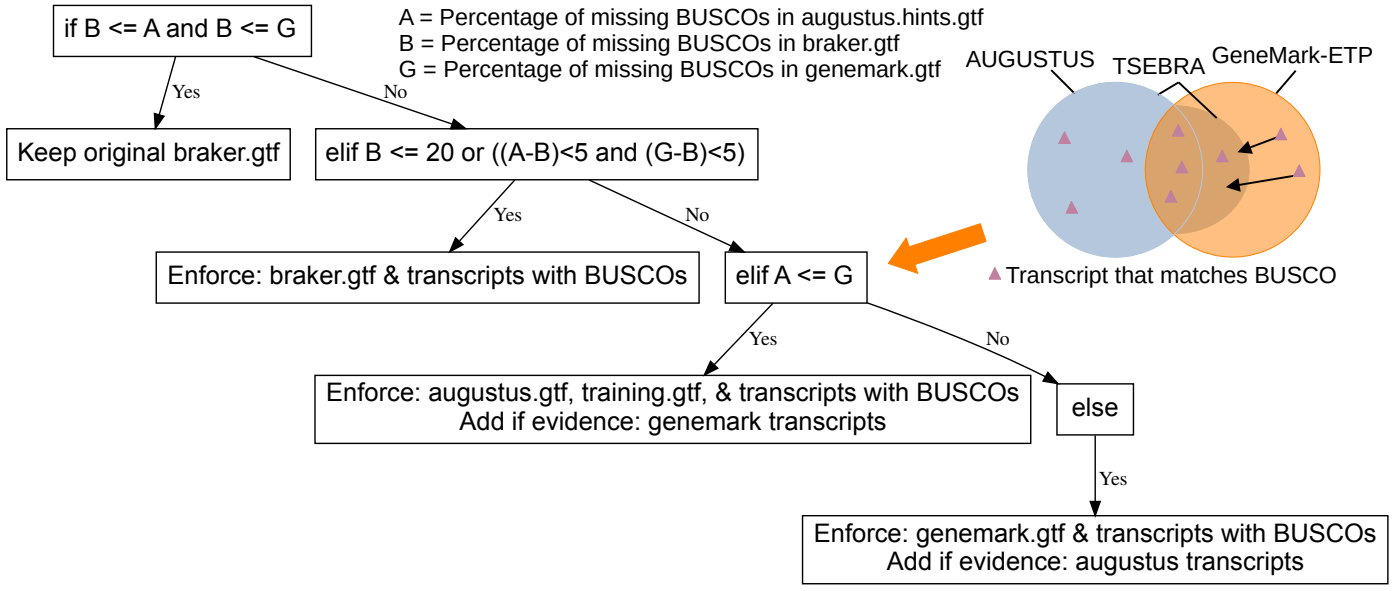
Scenario 1: Good Evidence



Improving BRAKER with Compleasm

Scenario 1: Poor Evidence

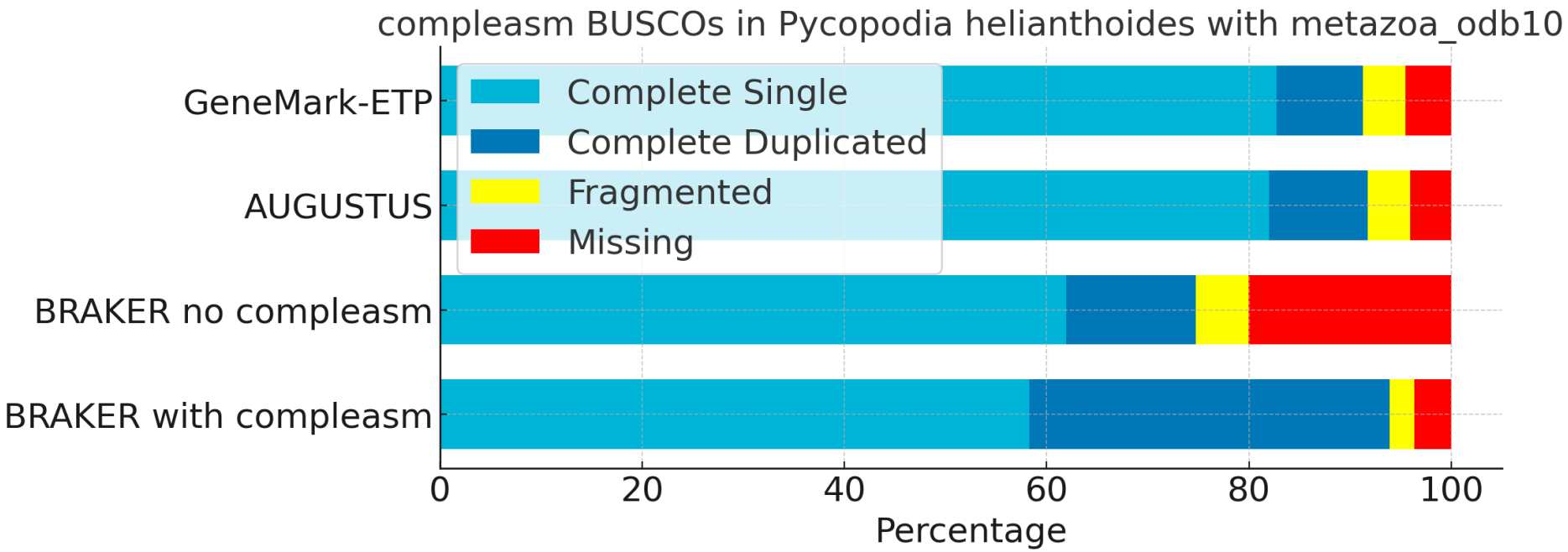




Improving BRAKER with Compleasm

Scenario 1: Poor Evidence

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



	AUGUSTUS	BRAKER3 no compleasm	BRAKER3 with compleasm
#Genes	24,184	15,598	25,601
#Transcripts	26,581	16,473	30,626
Single:Mult ratio	0.29	0.2	0.32

Related seastar *Asterias rubens* has 19,938 genes

Accuracy of genome annotation approaches by BRAKER team

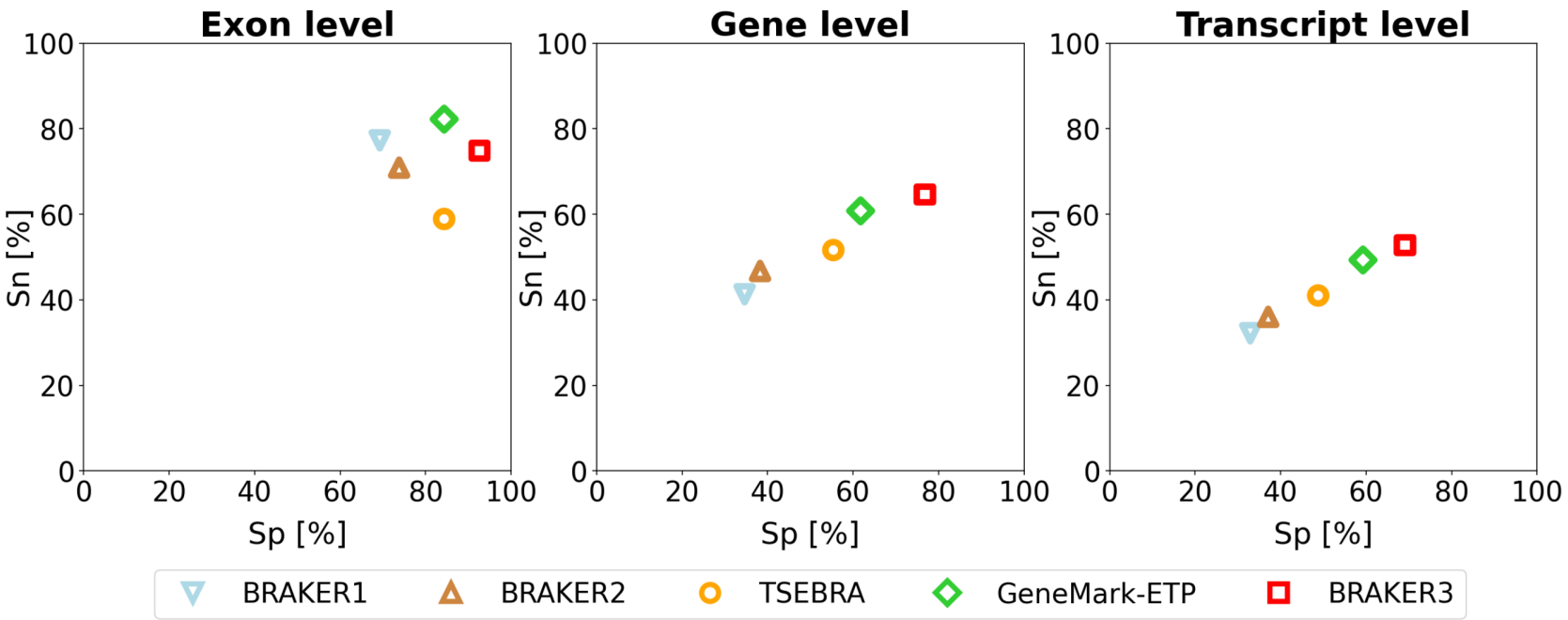


Figure 2: Average specificity and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in Supplemental Table S1). Inputs were the genomic sequences, short-read RNA-Seq libraries and protein databases (*order excluded*).

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

Accuracy of genome annotation approaches including competitive tools

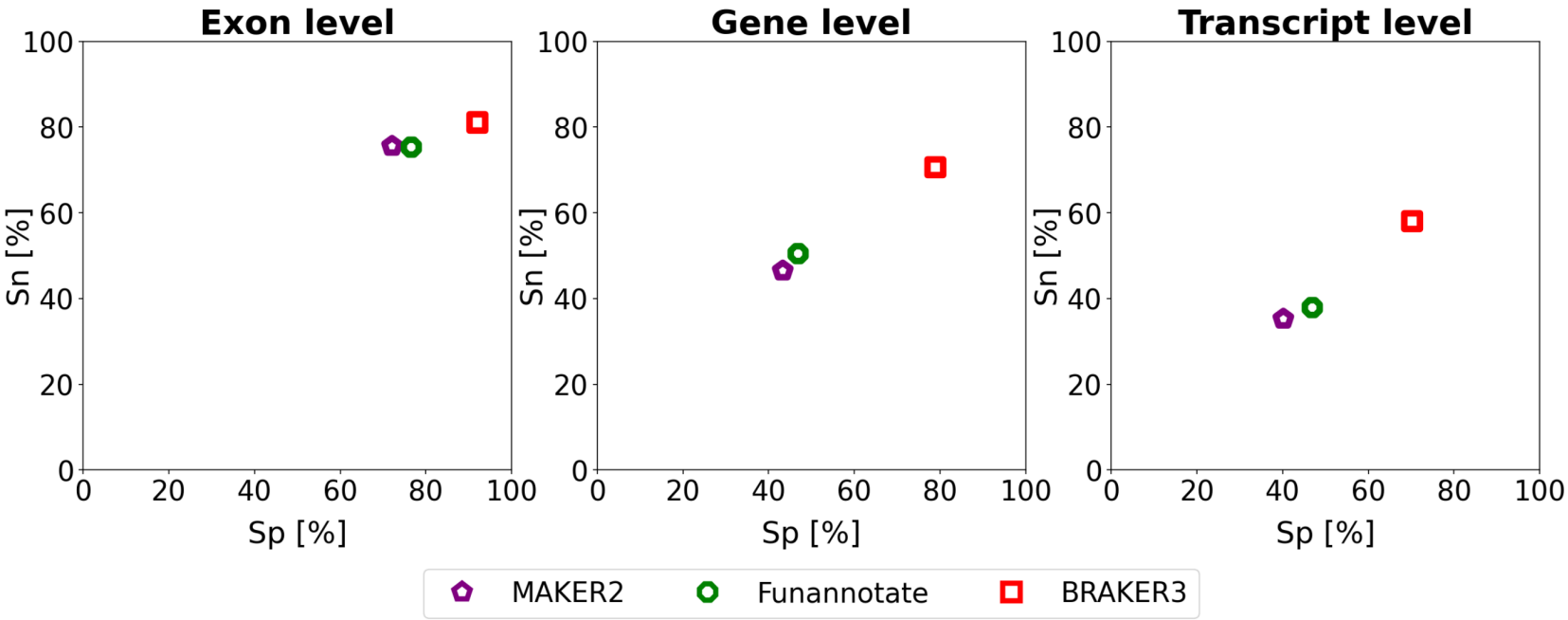


Figure 5: Average specificity and sensitivity of gene predictions made by MAKER2, Funannotate, and BRAKER3 for a subset of 8 species (excluding the mouse, spider and fish genome). Inputs were the genomic sequences, short-read RNA-Seq libraries, and protein databases (*close relatives included*). The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

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

Availability

GitHub

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

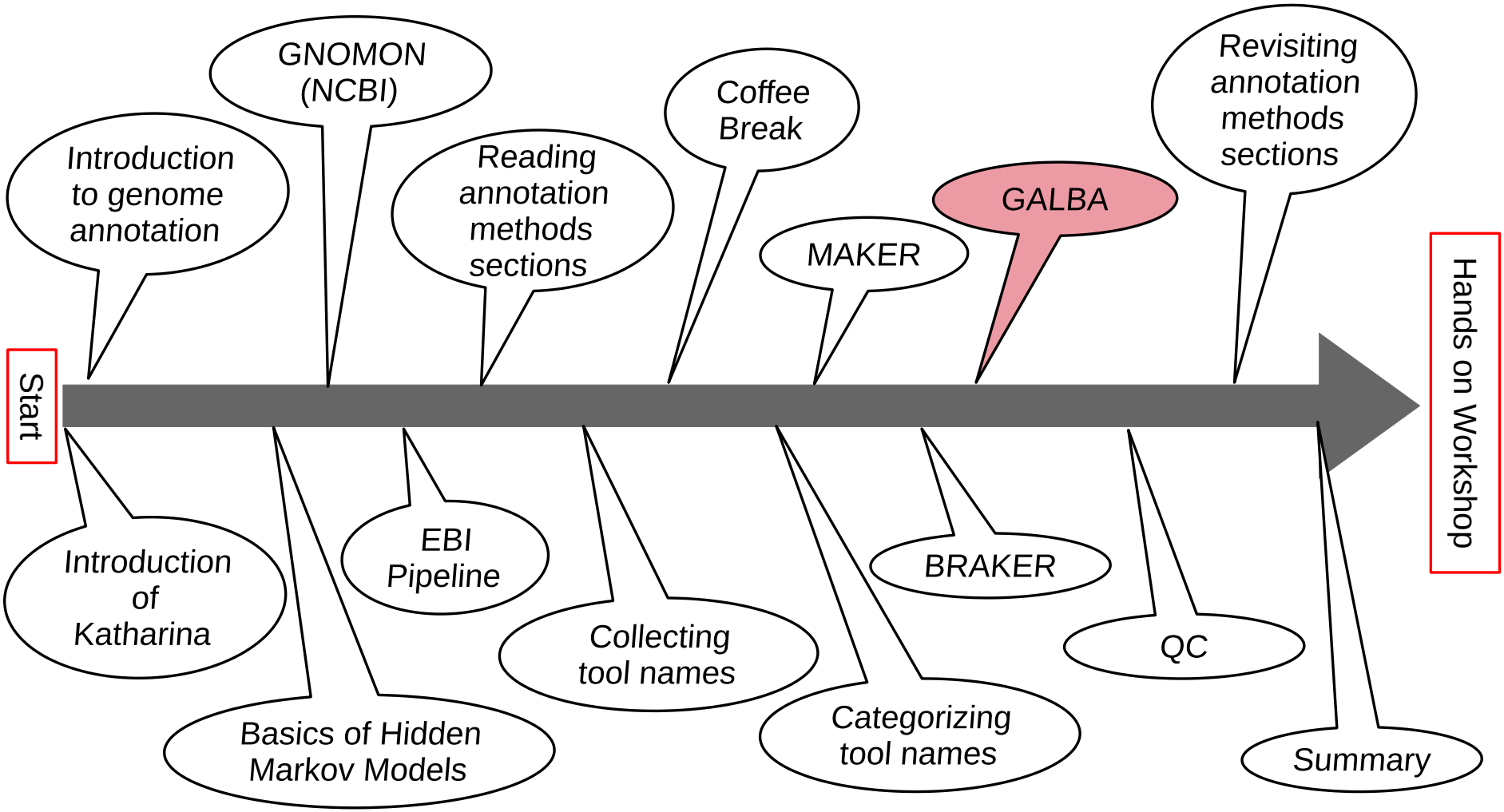
Docker/Singularity

```
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

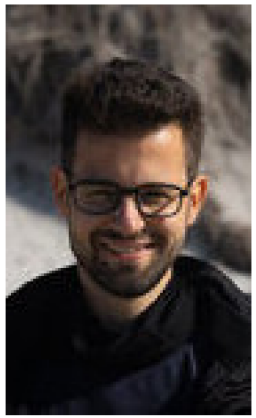

Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
QC
Summary

GALBA Contributors


Tomáš Brůna


Heng Li


Joseph Guhlin


Lars Gabriel


Natalia Nenasheva


Ethan Tolman


Paul Frandsen


Matthis Ebel


Mario Stanke


Katharina Hoff

Miniprot

Bioinformatics, 39(1), 2023, btad014
https://doi.org/10.1093/bioinformatics/btad014
Advance Access Publication Date: 17 January 2023
Original Paper

OXFORD

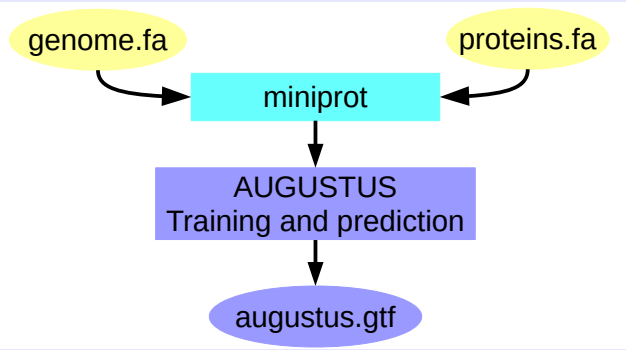
Genome analysis

Protein-to-genome alignment with miniprot

Heng Li [1,2]

"*Miniprot is a fast protein-to-genome aligner comparable to existing tools in accuracy. Its primary use case is to assist gene annotation.*"

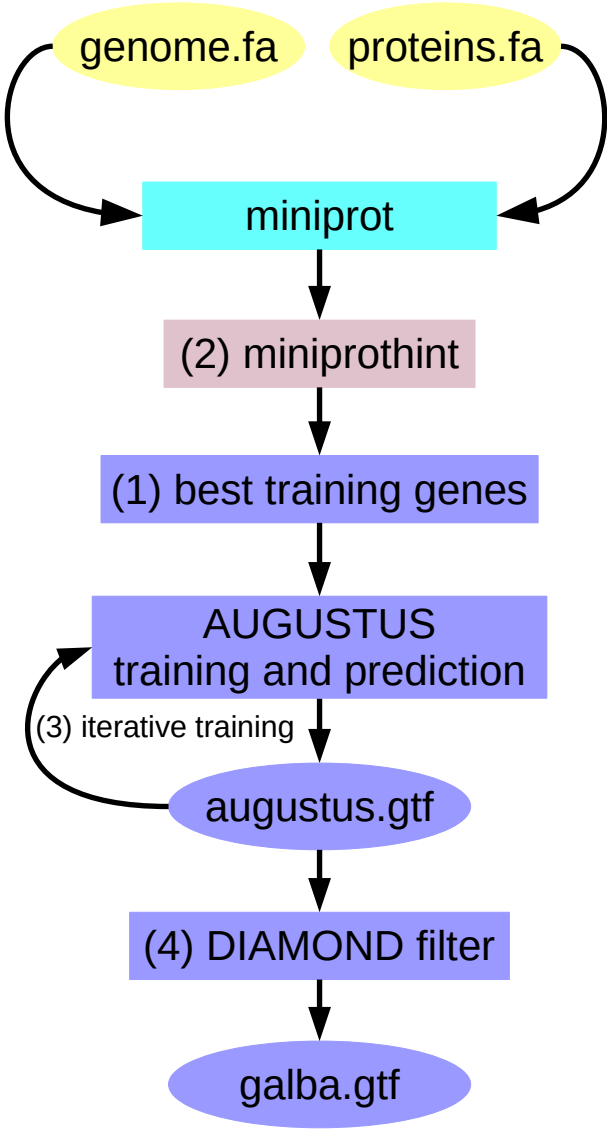
GALBA

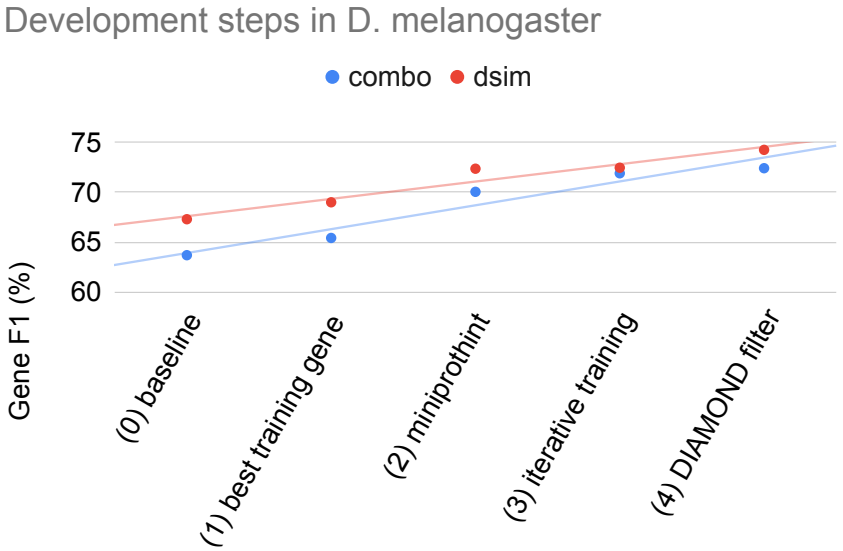


Do we need another pipeline?

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

GALBA: using proteins of related species





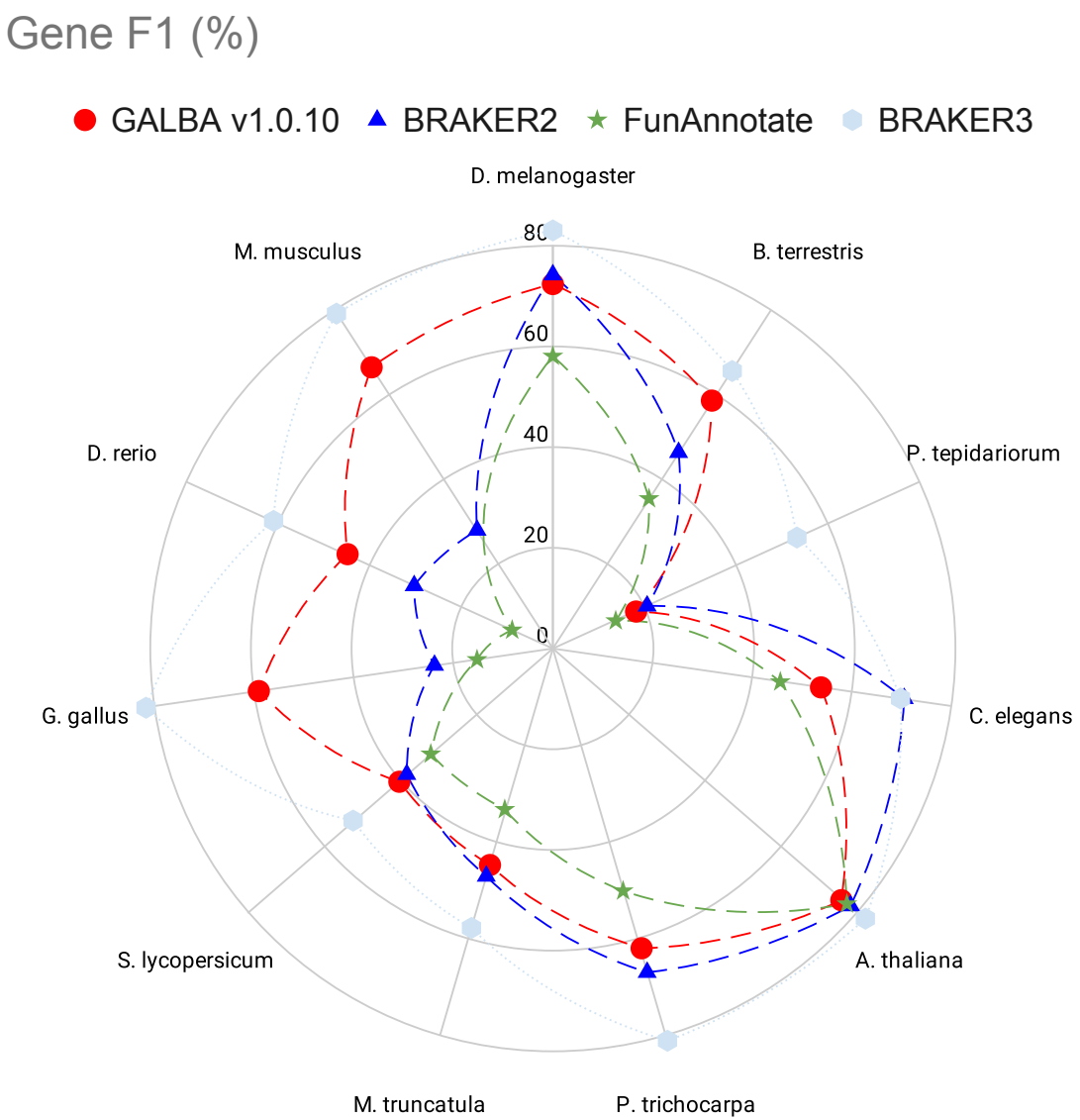
Donor proteins from

dsim	*D. simulans*
combo	*D. ananassae*,
	D. pseudoobscura,
	D. willistoni,
	D. virilis,
	D. grimshawi

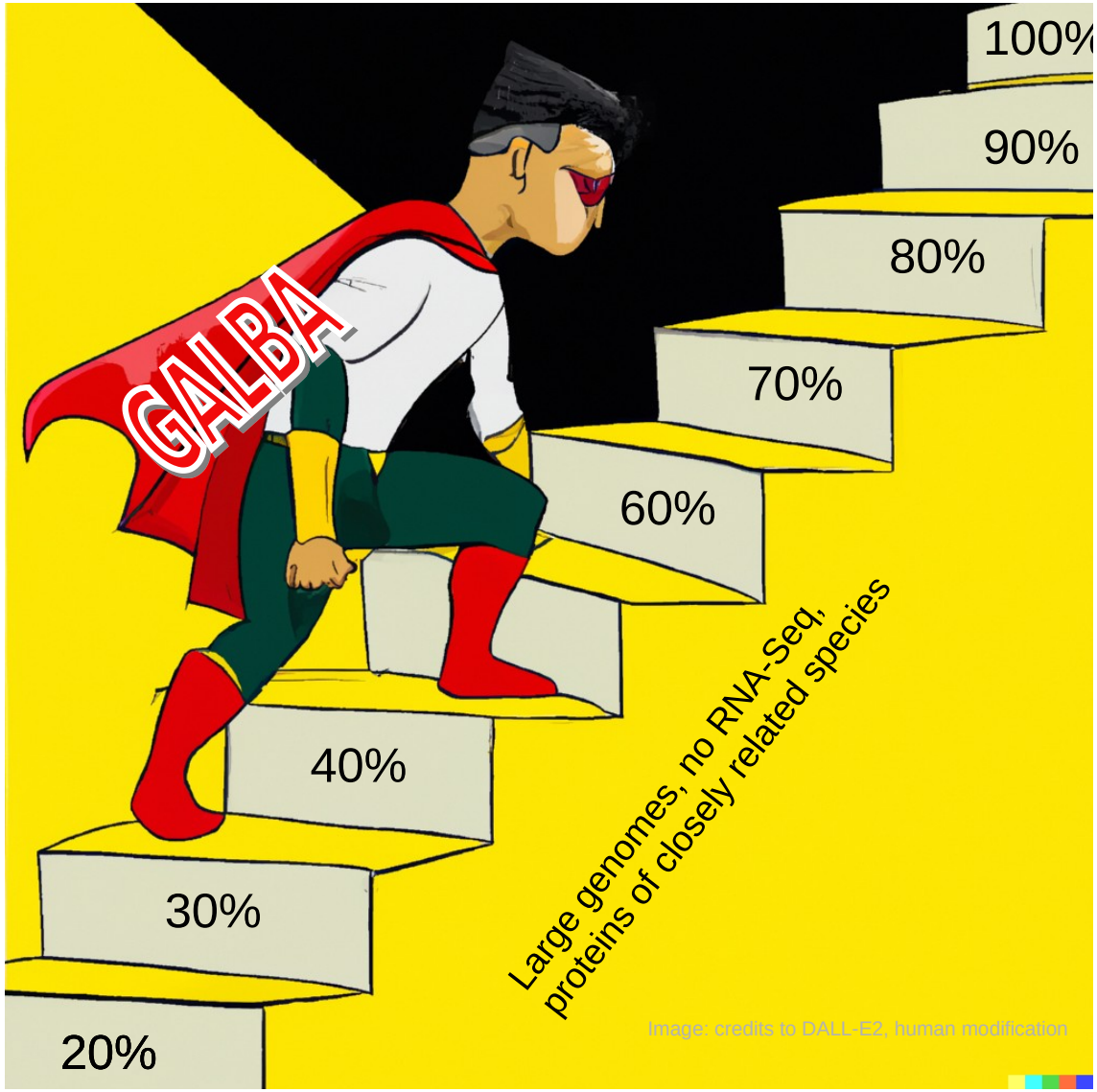
Idea for DIAMOND filter from Tolman *et al.* (2023)

`https://doi.org/10.1101/2023.12.11.569651`

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



GALBA: Gene F1 Accuracy



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

Availability

GitHub

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

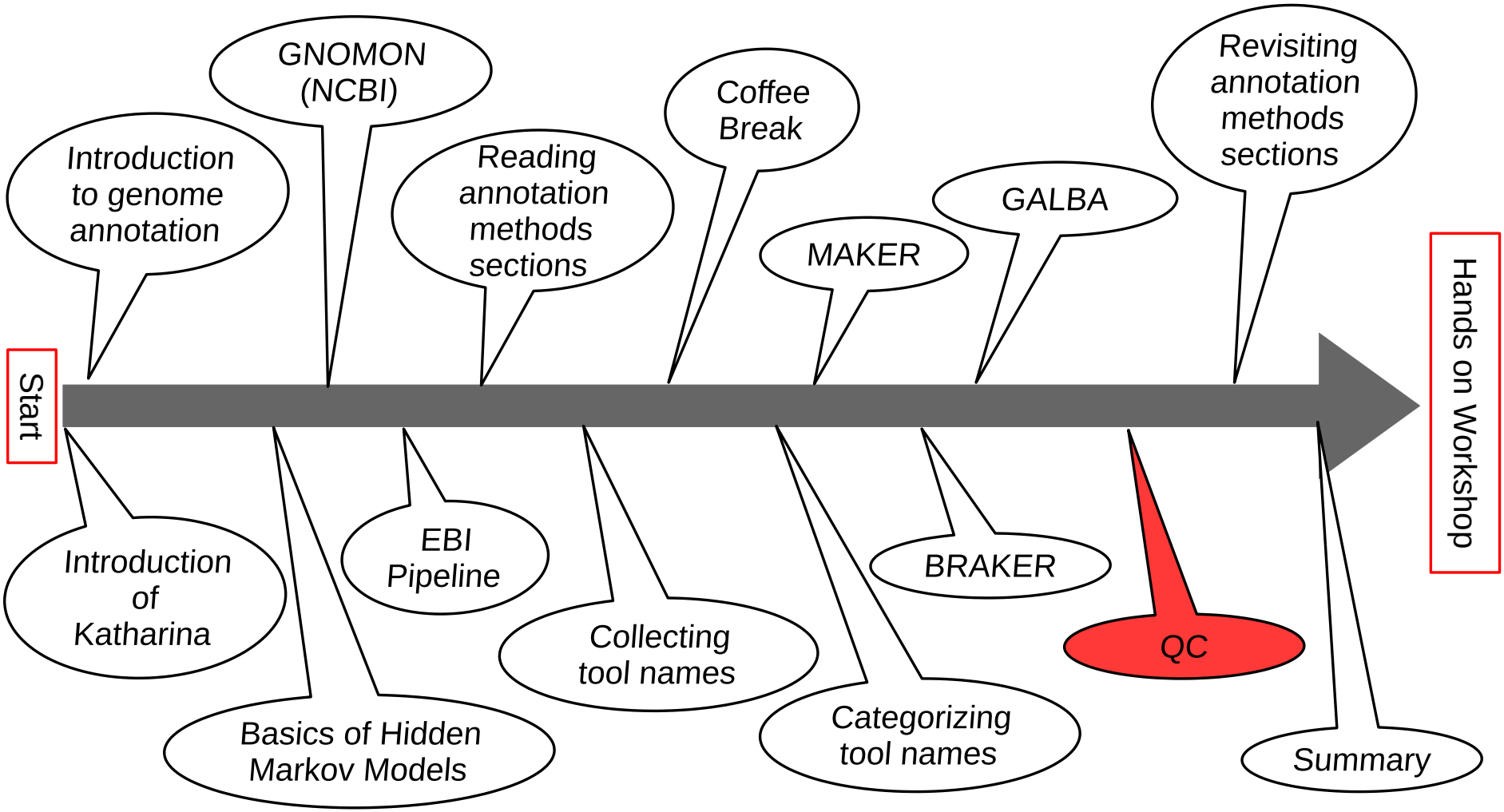
Docker/Singularity

```
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

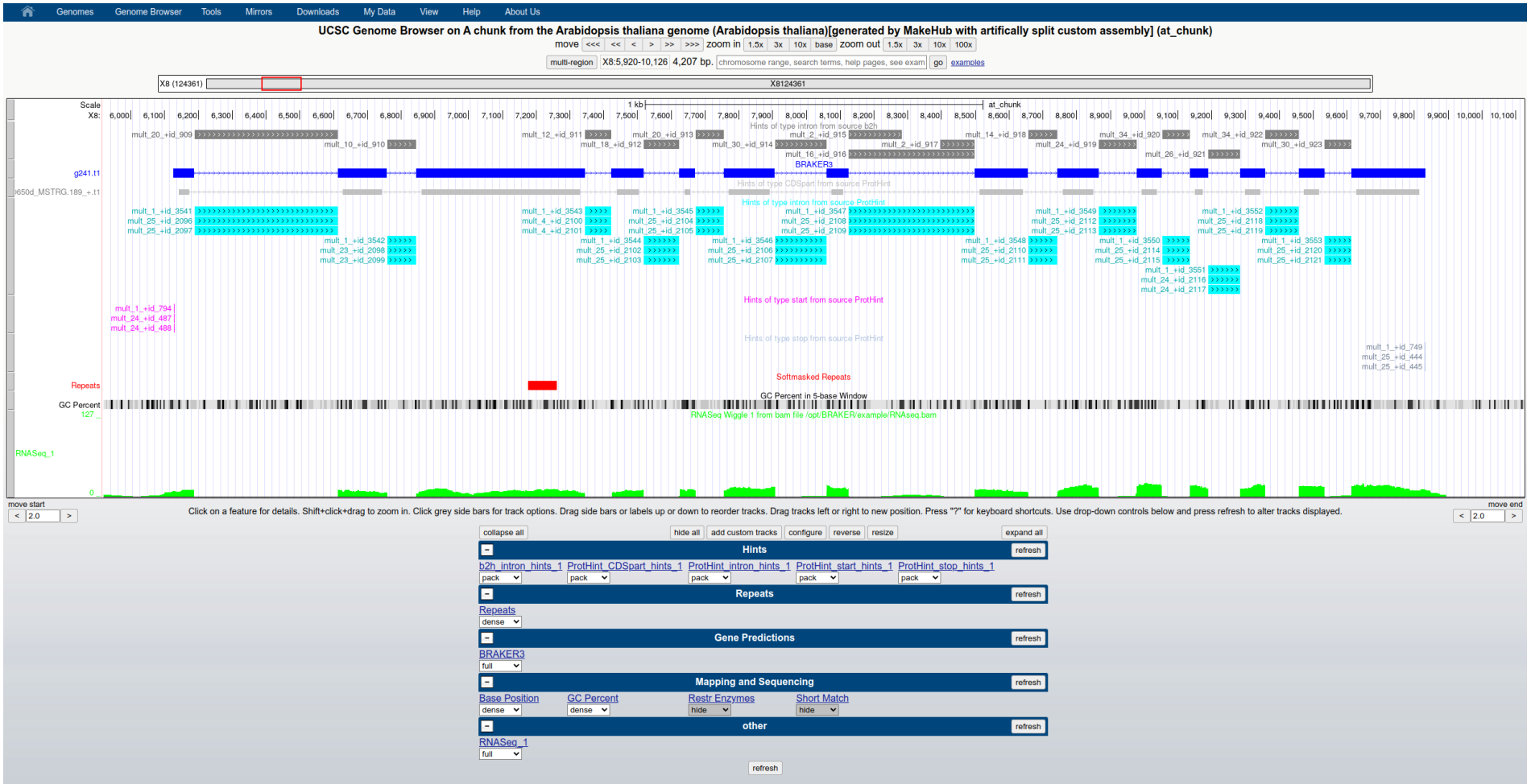

Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
QC
Summary

Did We Do a Good Job?

Genome Browsers

Visualize your Annotation in Context with Evidence

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

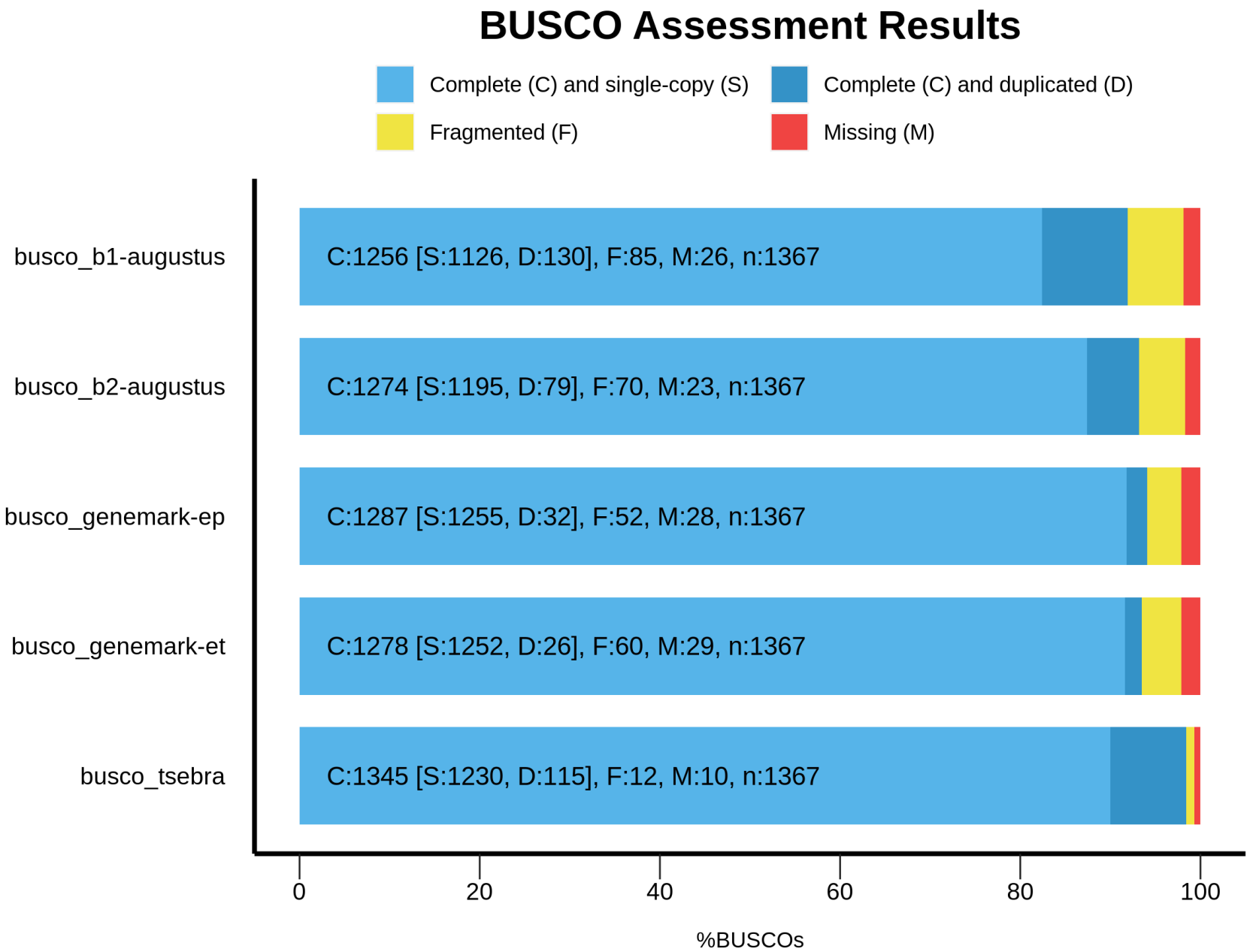


Describe Your Annotation

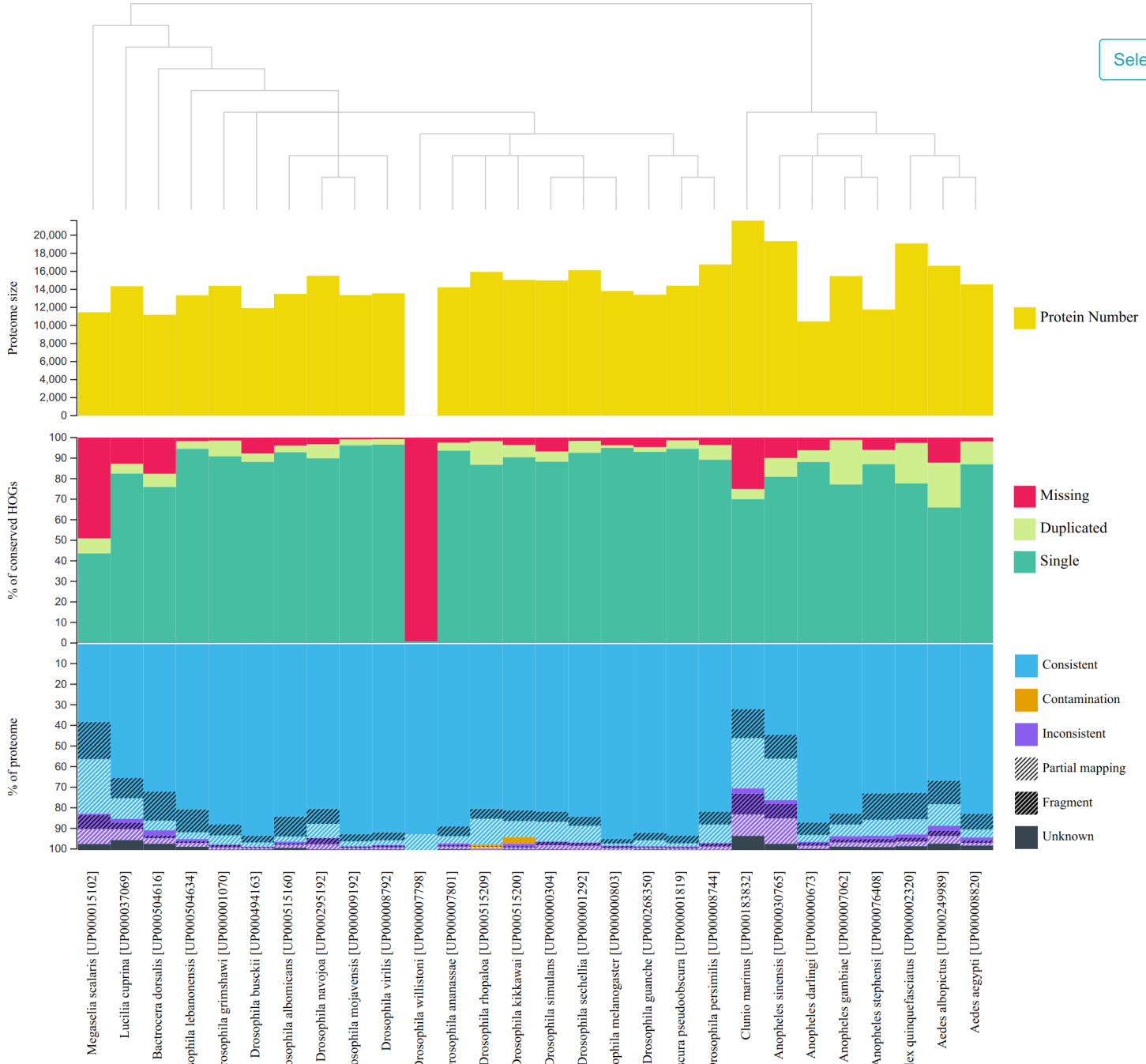
- number of genes
- number of transcripts
- ratio of mono-exonic to multi-exonic genes
- median number of exons per transcript
- maximal number of exons per transcript
- median transcript length
- ...

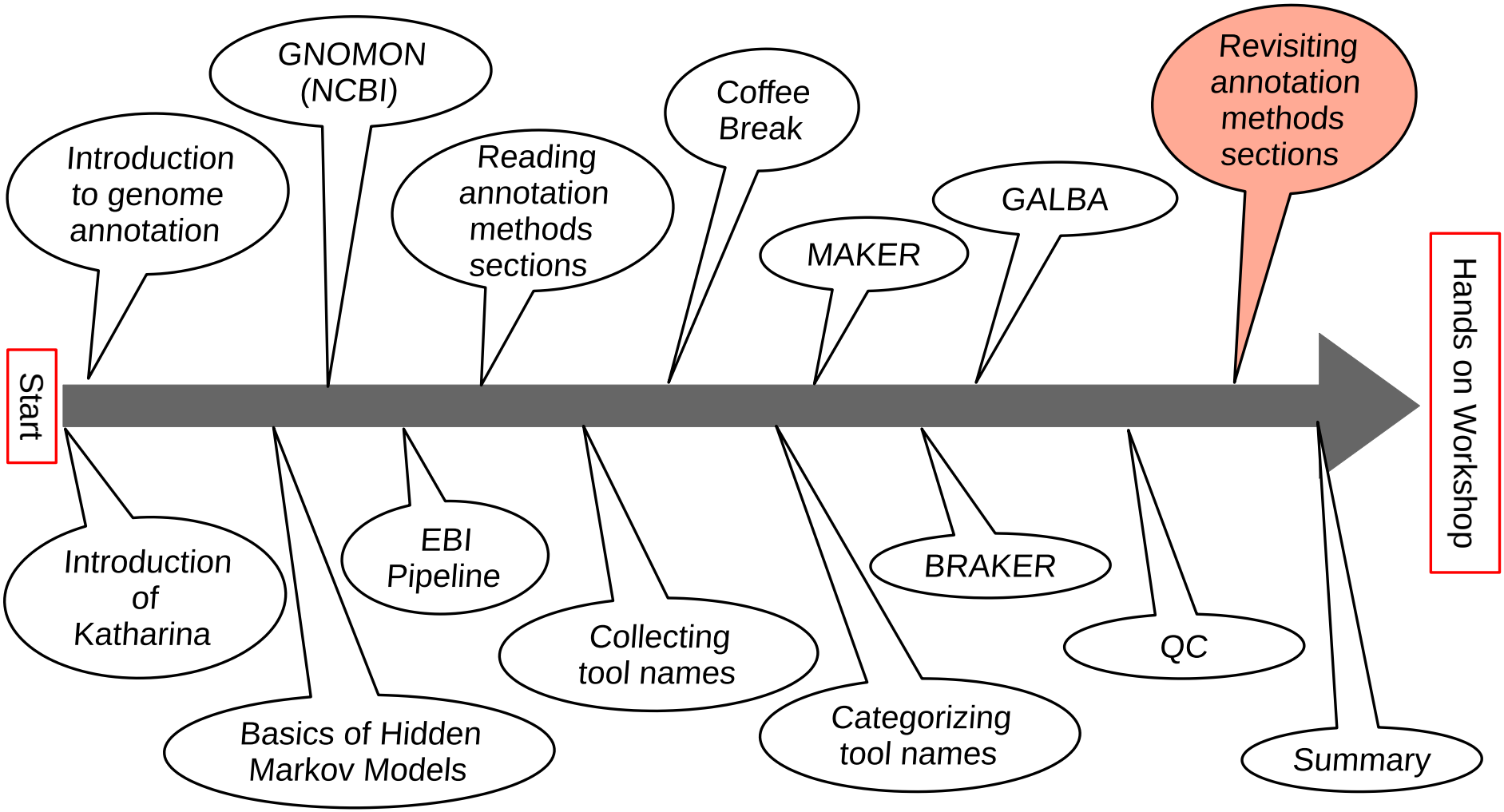
If possible, compare to annotated close relatives.
Consider effect of individual annotation pipelines.

BUSCO: Sensitivity in Clade-Specific Conserved Genes



OMArk: Sensitivity, Contaminations, & More

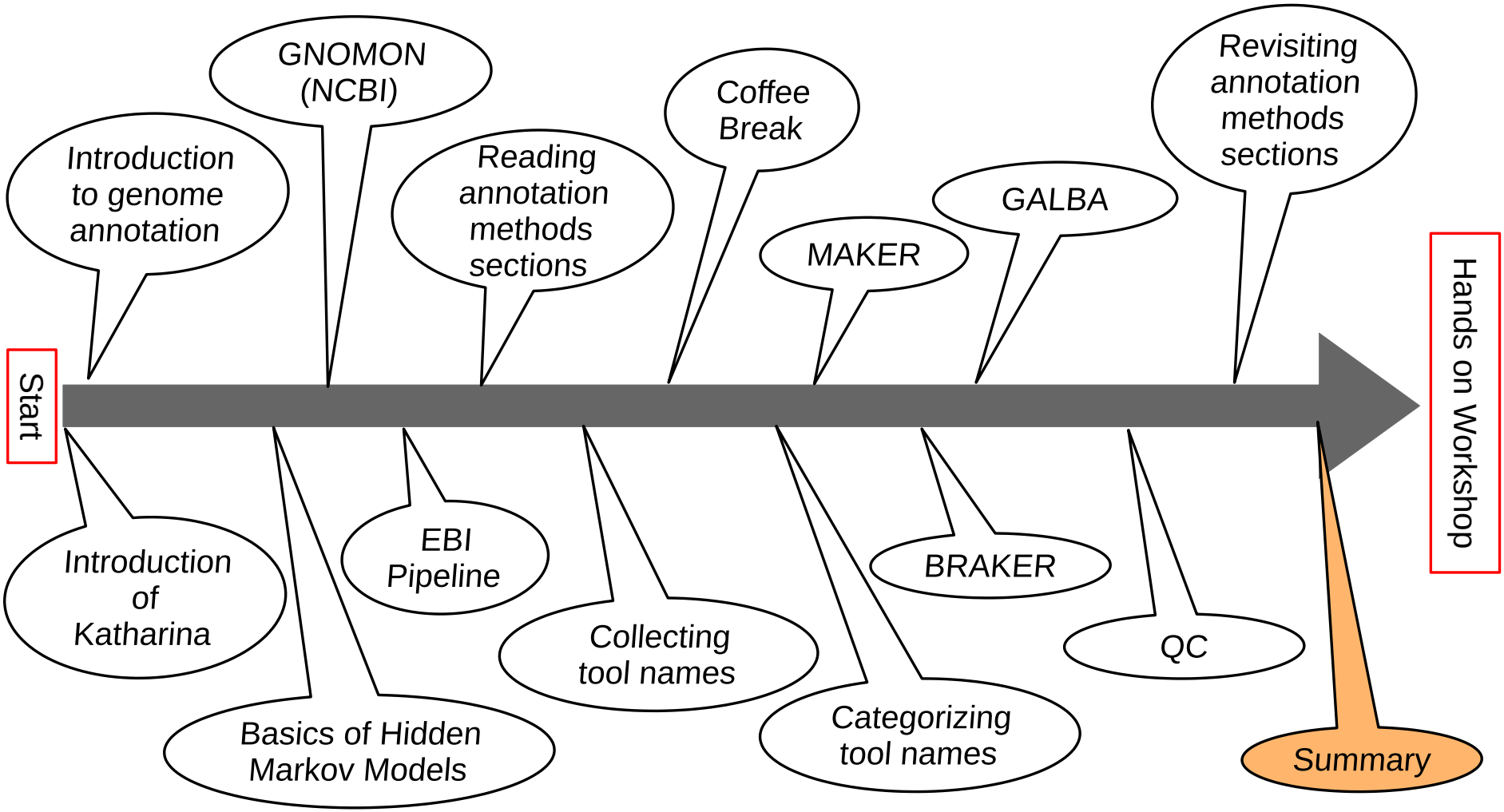



Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
QC
Summary

Revisiting annotation methods sections

Your tasks

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


Start
Introduction to genome annotation
GNOMON (NCBI)
Reading annotation methods sections
Coffee Break
MAKER
GALBA
Revisiting annotation methods sections
Hands on Workshop
Introduction of Katharina
EBI Pipeline
Basics of Hidden Markov Models
Collecting tool names
Categorizing tool names
BRAKER
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. MAKER
  2. BRAKER
  3. (GALBA may become popular)
- accuracy matters
- "looking nice" is not always "correct"
- BUSCO completeness is a widely used sensitivity measure
- OMArk might be more appropriate