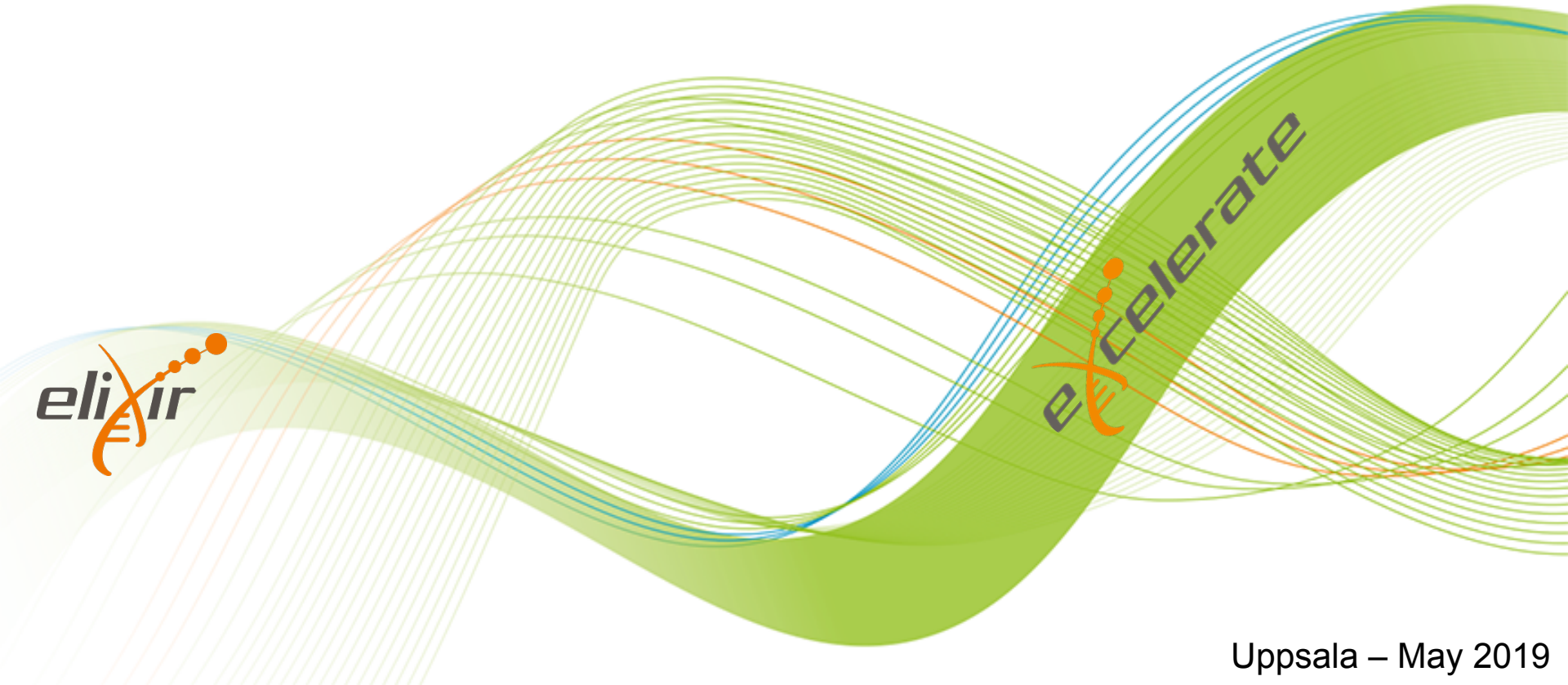
Jacques Dainat PhD

Methods in genome annotation



Uppsala – May 2019

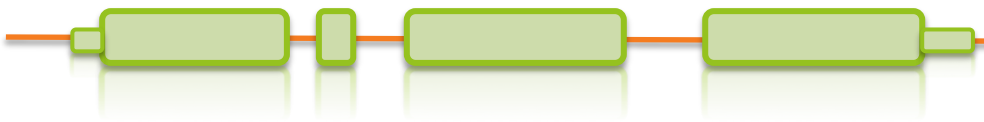
This lecture will focus on eukaryotes

1. The different annotation approaches (coding genes)
   1.1 Introduction
   1.2 Ab initio
   1.3. Hybrid
   1.4 Chooser, combiner
   1.5 Pipeline

2. Annotation of other genome features

3. Assessing an annotation

4. Closing remarks



1. The different annotation approaches

1.1 Introduction

❑ Annotation tools

First annotation tools — **>100 annotation tools***

90's — Nowadays

GFFs

Structural annotation + functional annotation + …

*estimation

https://github.com/NBISweden/GAAS/blob/master/annotation/CheatSheet/annotation_tools.md

The different approaches

- Similarity-based methods :

  These use similarity to annotated sequences like proteins, cDNAs, or ESTs

- *Ab initio* prediction :

  Likelihood based methods

- Hybrid approaches :

  Ab initio tools with the ability to integrate external evidence/hints

- Comparative (homology) based gene finders :

  These align genomic sequences from different species and use the alignments to guide the gene predictions

- Chooser, combiner approaches :

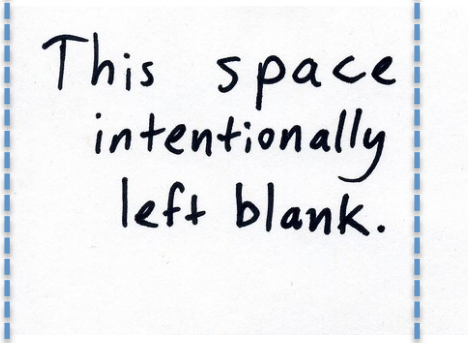
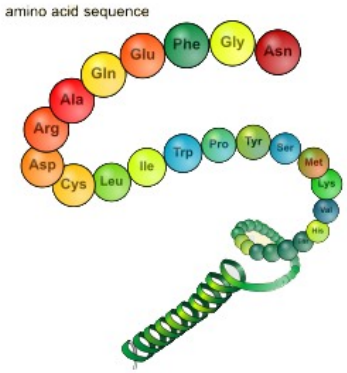
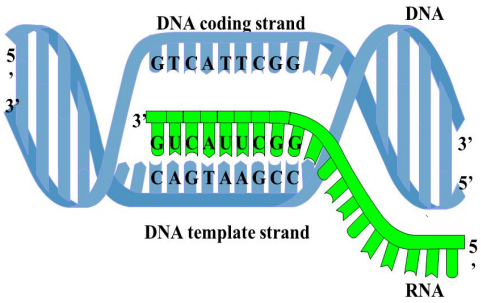
  These combine gene predictions of other gene finders

- Pipelines :

  These combine multiple approaches

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Types data used vs methods

Annotation approach	∅ 	Proteins Known amino acid sequences from other organisms 	Transcripts Assembled from RNA-seq or downloaded ESTs 
Similarity		X	X
Pure ab initio	X		
Hybrid		X	X
Comparative	X	X	X
Chooser/combiner	X	X	X
Pipeline	X	X	X

The different approaches

- Similarity-based methods :

  These use similarity to annotated sequences like proteins, cDNAs, or ESTs

- *Ab initio* prediction :

  Likelihood based methods

- Hybrid approaches :

  Ab initio tools with the ability to integrate external evidence/hints

- Comparative (homology) based gene finders :

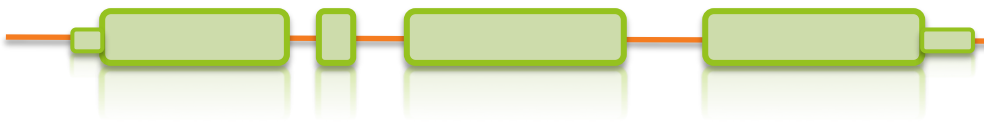
  These align genomic sequences from different species and use the alignments to guide the gene predictions

- Chooser, combiner approaches :

  These combine gene predictions of other gene finders

- Pipelines :

  These combine multiple approaches



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

1. The different annotation approaches

1.2 *Ab-initio* annotation tools
“intrinsic approach”

Ab initio method

- Uses likelihoods to find the most likely gene models
- Easy to use!
- augustus --species=chicken contig.fa > augustus_chicken.gff

Ab initio method

method based on **gene content :**
(statistical properties of protein-coding sequence)

- codon usage
- hexamer usage
- GC content
- compositional bias between codon positions
- nucleotide periodicity
- exon/intron size
- …

and on **signal detection**:

- Promoter
- ORF
- Start codon
- Splice site (Donor and acceptor)
- Stop codon
- Poly(A) tail
- CpG islands
- …

=> *Ab initio* tools will combine this information through different Probabilistic models: HMM, GHMM, WAM, etc.

These models need to be created if not already existing for your organism => **training!**

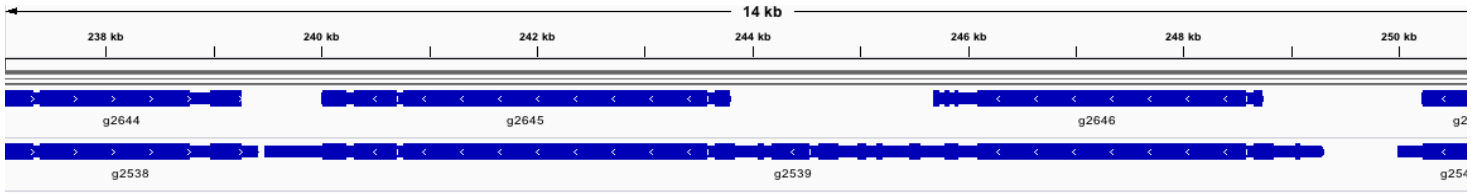
Training *ab-initio* gene-finders

- Some gene-finders train themselves, others need a separate training procedure
- Around 500 already known genes are usually needed to train the gene-finder

  => These ”known” genes can be inferred from aligned transcripts or proteins
- The quality of the gene-finder results hugely relies on the quality of the training!

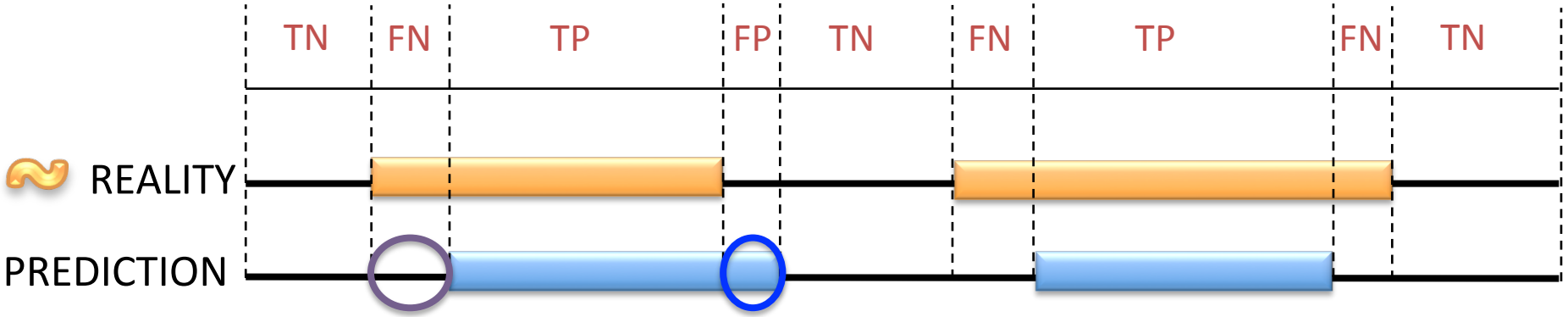
A fungal genome

Fungi
Plants



Ab initio method

Assess the quality of the *ab-initio* model/training:



Sensitivity is the proportion of true predictions compared to the total number of correct genes (including missed predictions)

$$Sn = \frac{TP}{TP + FN}$$

Specificity is the proportion of true predictions among all predicted genes (including incorrectly predicted ones)

$$Sp = \frac{TP}{TP + FP}$$

Ab Initio methods can approach 100% sensitivity, however as the sensitivity increases, accuracy suffers as a result of increased false positives.

Ab initio method

```
*******      Evaluation of gene prediction     *******

-----------------------------------------\
                | sensitivity | specificity |
----------------|-------------|-------------|
nucleotide level |       0.987 |       0.896 |
-----------------------------------------/

-------------------------------------------------------------------------------------------------\
           |  #pred  |  #anno  |      |    FP = false pos.  |    FN = false neg.  |             |             |
           | total/  | total/  |  TP  |---------------------|---------------------| sensitivity | specificity |
           | unique  | unique  |      | part | ovlp | wrng | part | ovlp | wrng |             |             |
-----------|---------|---------|------|------|------|------|------|------|------|-------------|-------------|
           |         |         |      |                 85 |                 45 |             |             |
exon level |     512 |     472 |  427 | ---------------- | ---------------- |       0.905 |       0.834 |
           |     512 |     472 |      |   29 |    2 |   54 |   30 |    1 |   14 |             |             |
-------------------------------------------------------------------------------------------------/

-----------------------------------------------------------------------\
transcript | #pred | #anno |   TP |   FP |   FN | sensitivity | specificity |
-----------|-------|-------|------|------|-------|-------------|-------------|
gene level |   105 |   100 |   67 |   38 |    33 |        0.67 |       0.638 |
-----------------------------------------------------------------------/
```

Ab initio method

Popular tools:

- **SNAP** Works ok, easy to train, not as good as others especially on longer intron genomes.
- **Augustus** Works great, hard to train (but getting better).
- **GeneMark-ES** Self training, no hints, buggy, not good for fragmented genomes or long introns (Best suited for Fungi).
- **FGENESH** Works great, costs money even for training.

Supported by MAKER

http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial

- **GlimmerHMM** (Eukaryote)
- **GenScan**
- **Gnomon** (NCBI)

Ab initio method

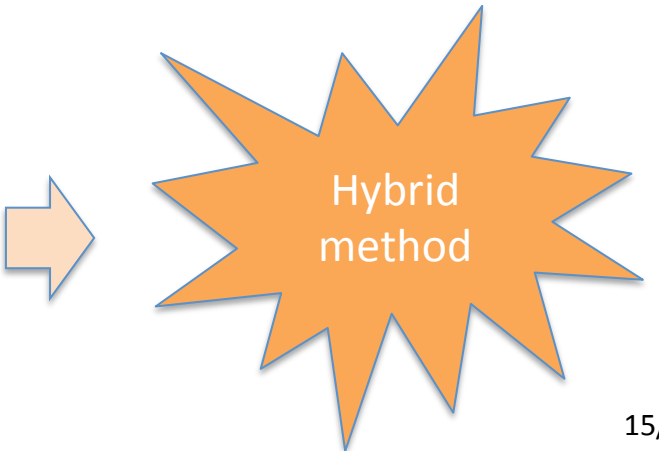
Strengths :

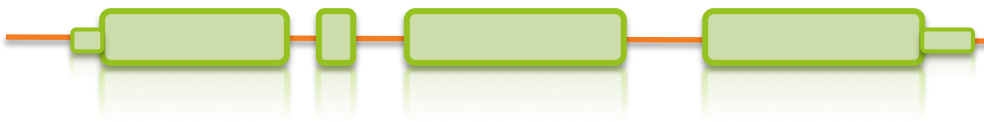
- Fast and easy means to identify genes
- Annotate unknown genes
- “Exhaustive” annotation
- Need no external evidence

Limits :

- No UTR*
- No alternatively spliced transcripts*
- Over prediction (exons or genes)
- **Training** needed to perform well in *terra incognita*’

- Split single gene into multiple predictions
- Fused with neighboring genes
- Less accurate than homology based method:
  - Exon boundaries
  - Splicing sites





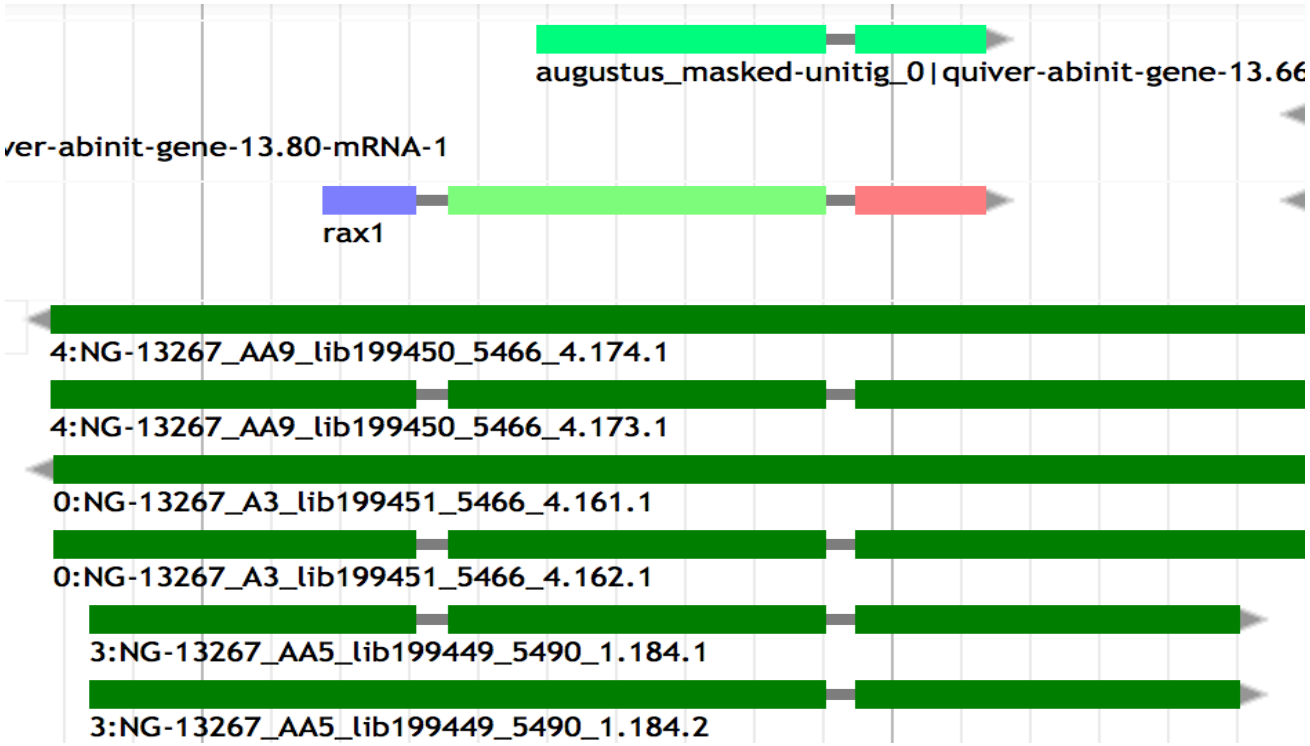
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

1. The different annotation approaches

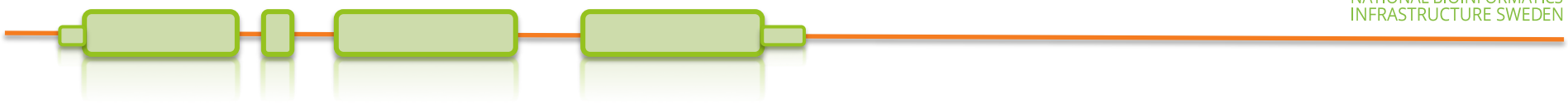
1.3 Hybrid approaches

Hybrid method

Hybrid (*evidence-drivable gene predictors*) approaches incorporate hints in the form of EST or protein alignments to increase the accuracy of the gene prediction.



Hybrid method



Hybrid (*evidence-drivable gene predictors*) approaches incorporate hints in the form of EST alignments or protein profiles to increase the accuracy of the gene prediction.

GenomeScan	Blast hit used as extra guide
Augustus	16 types of hints accepted (gff): start, stop, tss, tts, ass, dss, exonpart, exon, intronpart, intron, CDSpart, CDS, UTRpart, UTR, irpart, nonexonpart.
GeneMark-ET	EST-based evidence hints
GeneMark-EP	Protein-based evidence hints
SNAP	Accepts EST and protein-based evidence hints.
Gnomon	Uses EST and protein alignments to guide gene prediction and **add UTRs**
FGENESH+	Best suited for plant
EuGene*	Any kind of evidence hints. Hard to configure (best suited for plant)

Self training !

Hybrid method

Strength : High accuracy

Limits :

- Extra computation to generate alignments

- heterogeneous sequence quality :

Incomplete,
Error during transcriptome assembly
Contamination
Sequence missing
Orientation error

The BRAKER1 gene finding pipeline:

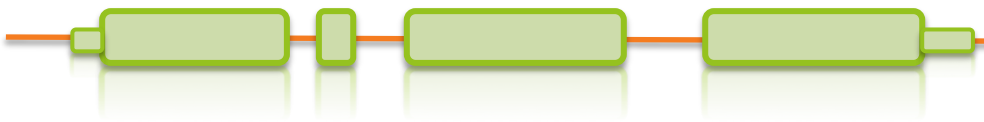
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

Katharina J. Hoff et *al.*

Bioinformatics (2016) 32 (5): 767-769. doi: 10.1093/bioinformatics/btv661

- BRAKER1 was more accurate than MAKER2 when it is using RNA-Seq as sole source for training and prediction.
- BRAKER1 does not require pre-trained parameters or a separate expert-prepared training step.

BRAKER2 since 2019 (Incorporate Protein Homology Information)



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

1. The different annotation approaches

1.4 Chooser / combiner

Overview

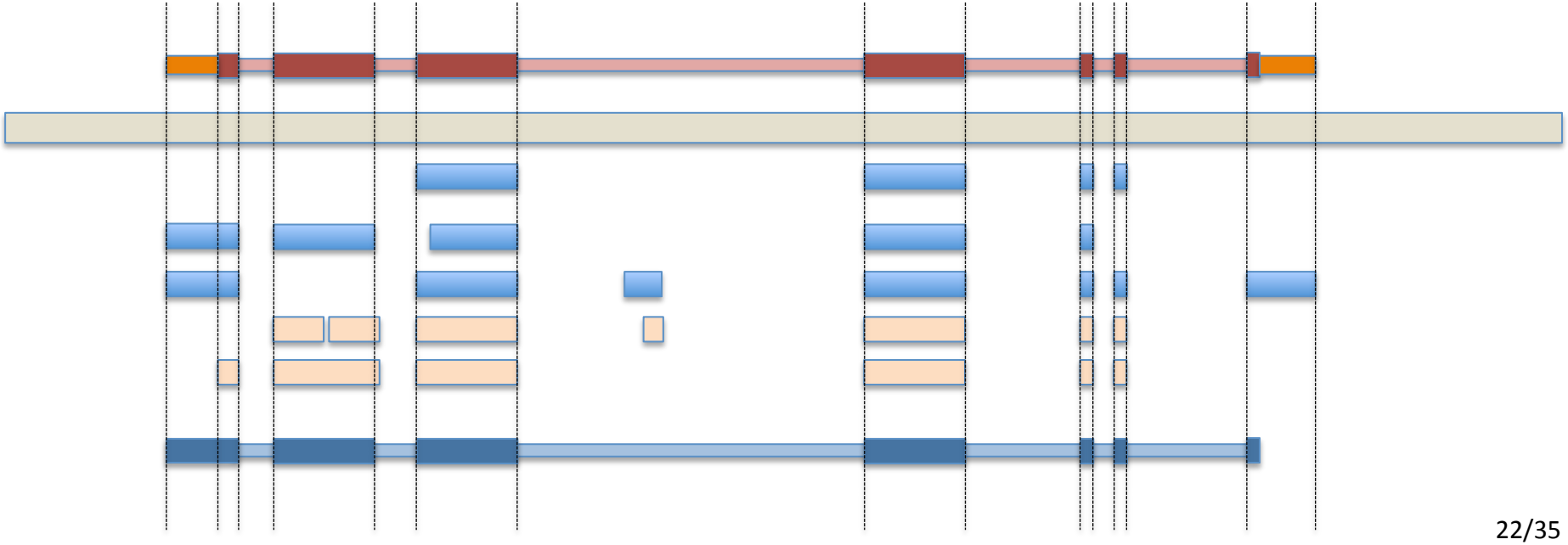
combining different lines of evidence into gene models

Evidence: ESTs / Transcripts

Proteins

Ab-initio prediction

Combining

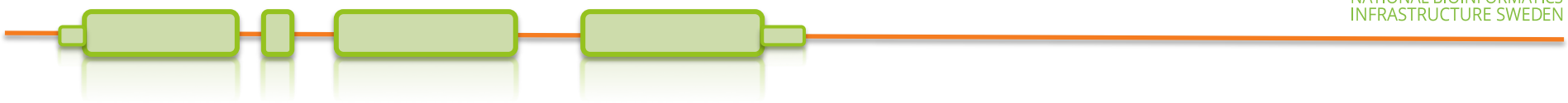


Chooser / combiner

Use battery of gene finders and evidence (EST, RNAseq, protein) alignments and:

Tool	Consensus based chooser	Evidence based chooser	weight of different sources	Comment
A) Choose the prediction whose best matches the evidence				
MAKER*		X		
PASA*		X		
B) Choose the prediction whose structure best represents the consensus				
JIGSAW	X			
C) Choose the best possible set of exons and combine them in a gene model				
EVM Evidencemodeler	X	X	X	User can set the expected evidence error rate manually or/and learn from a training set
Evigan	X		X	Unsupervised learning method
Ipred		X		Does not require any a priori knowledge Can also combine only evidences to create a gene model

Strength => They improve on the underlying gene prediction models



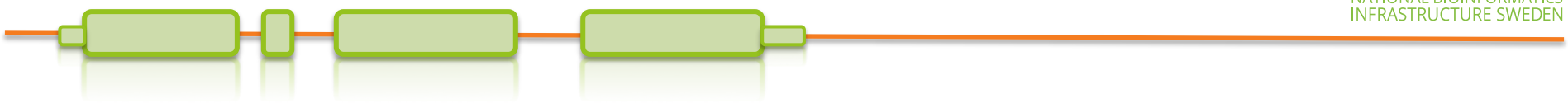
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

1. The different annotation approaches

1.5 Gene annotation pipelines
(The ultimate step)

Align evidence, add UTRs and more

Annotation pipeline



PASA Produces evidence-driven consensus gene models

- \- minimalist pipeline ()
- \+ good for detecting isoforms
- \+ biologically relevant predictions

=> using *Ab initio* tools and combined with **EVM** it does a pretty good job !

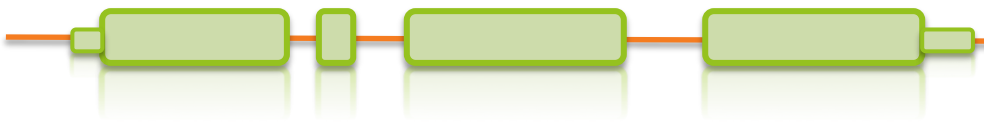
- \- PASA + Ab initio + EVM not automatized

NCBI pipeline Evidence + *ab initio* (Gnomon), repeat masking, gene naming, data formatting, miRNAs, tRNAs

Ensembl Evidence based only (comparative + homology) …

MAKER2 Evidence based and/or *ab initio* …

…

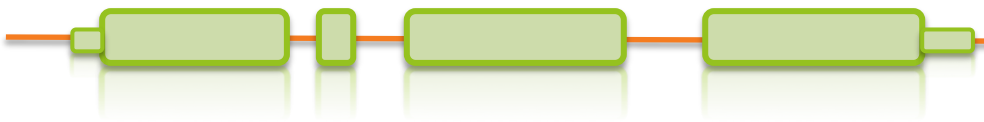


NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

2. Annotation of other genome features

Other genome features

Feature type	DB associated	Tool example	approach
ncRNA	Rfam	infernal	HMM + CM
tRNA	Sprinzl database	tRNAscan-SE	CM + WMA
snoRNA		snoscan	HMM + SCFG
miRNA	miRBase	Splign miR-PREFeR (for plant)	sequence alignment Based on expression patterns
Repeats	Repbase, Dfam	repeatMasker	HMM, blast
Pseudogenes		pseudopipe	homology-based (blast)
…			

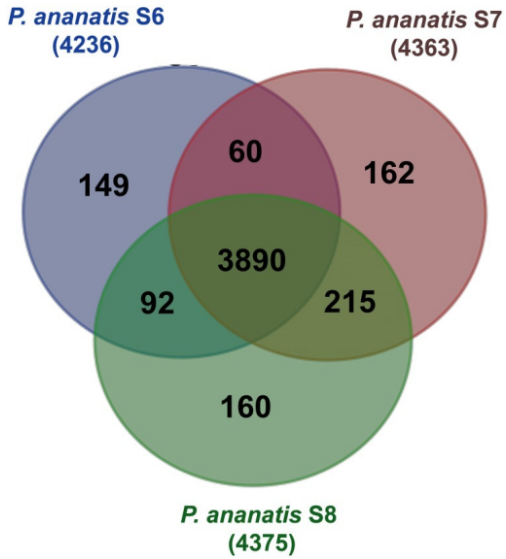


NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

3. Assessing an annotation

Assessing an annotation

- Simple statistics (number genes / number exon per gene)
- BUSCO (and compare against assembly result)
- Protein/transcript evidence (AED score in MAKER)
- Comparative genomics (OrthoMCL)
- Domain / Function attached
- Visualization

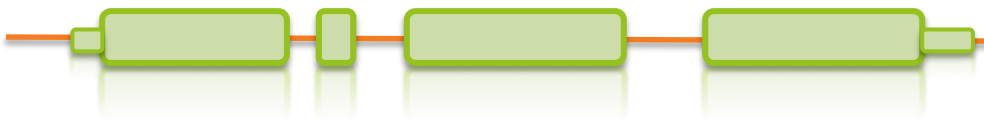


NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Selection of most common visualization or/and Manual curation tools

Name	Standalone	Web tool	Manual curation	year	comment
Artemis	X		X	2000	Can save annotation in EMBL format
IGV	X			2011	Popular
Savant	X			2010	Sequence Annotation, Visualization and ANalysis Tool. enable Plug-ins
Tablet	X		X	2013	
IGB	X			2008	enable Plug-ins. Can load local and remote data (dropbox, UCSC genome, etc)
Jbrowse		X		2010	GMOD (successor of Gbrowse)
Web Apollo		X	X	2013	Active community (gmod). Based on Jbrowse. Real-time collaboration
UCSC		X		2000	A large amount of locally stored data must be uploaded to servers across the internet
Ensembl genome browsers		X		2002	A large amount of locally stored data must be uploaded to servers across the internet

FOR AN EXHAUSTIVE LIST: https://en.wikipedia.org/wiki/Genome_browser



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

4. *To resume / Closing remarks*

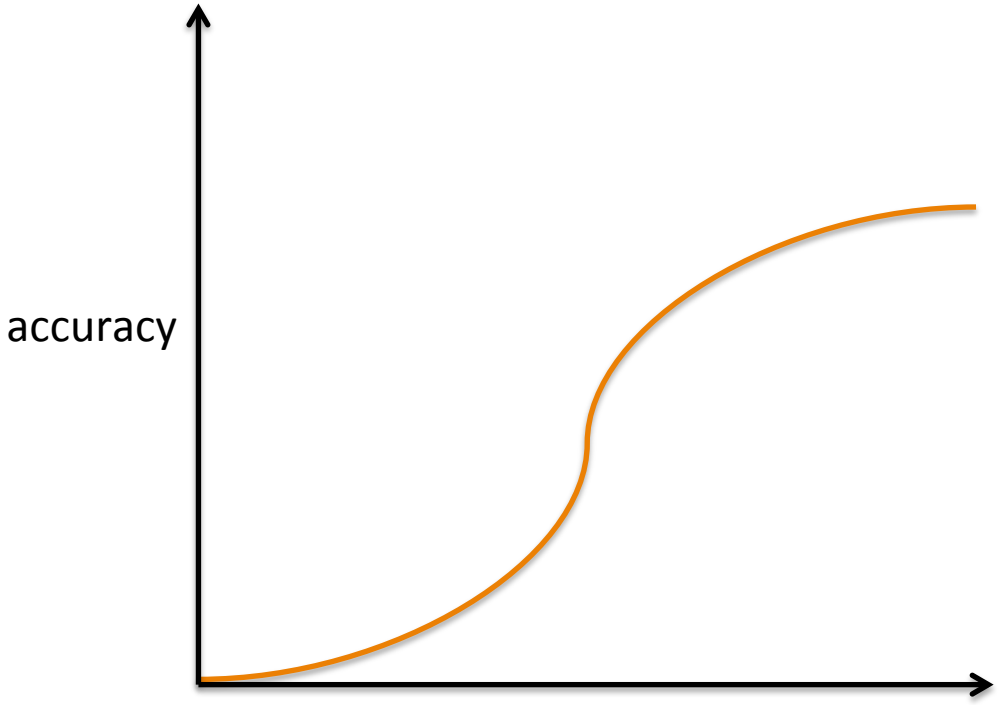
Closing remarks

- >100 annotation tools − as many methods
  (https://github.com/NBISweden/GAAS/blob/master/annotation/CheatSheet/annotation_tools.md)
- 6 main class of approaches (Similarity-based, *ab initio*, hybrid, comparative, combiner, pipeline)

How to choose Method:

- Scientific question behind (need of a conservative annotation vs exhaustive)
- Species dependent (plant / Fungi / eukaryotes)
- phylogenetic relationship of the investigated genome to other annotated genomes (Terra incognita, close, already annotated).
- Data available (hmm profile, RNAseq, etc…)
- Depending on computing resources (*ab initio* ~ hours < VS > pipeline ~ weeks)

Effort versus accuracy



Closing remarks

- Several *ab-initio* tools together give better result that one alone
  (they complement each other)

- Pipelines give good results
  MAKER2 the most flexible, adjustable

- Most methods only build gene models, no **functional inference**

- No annotation method is perfect, they do mistakes !!

- Annotation requires **manual curation**

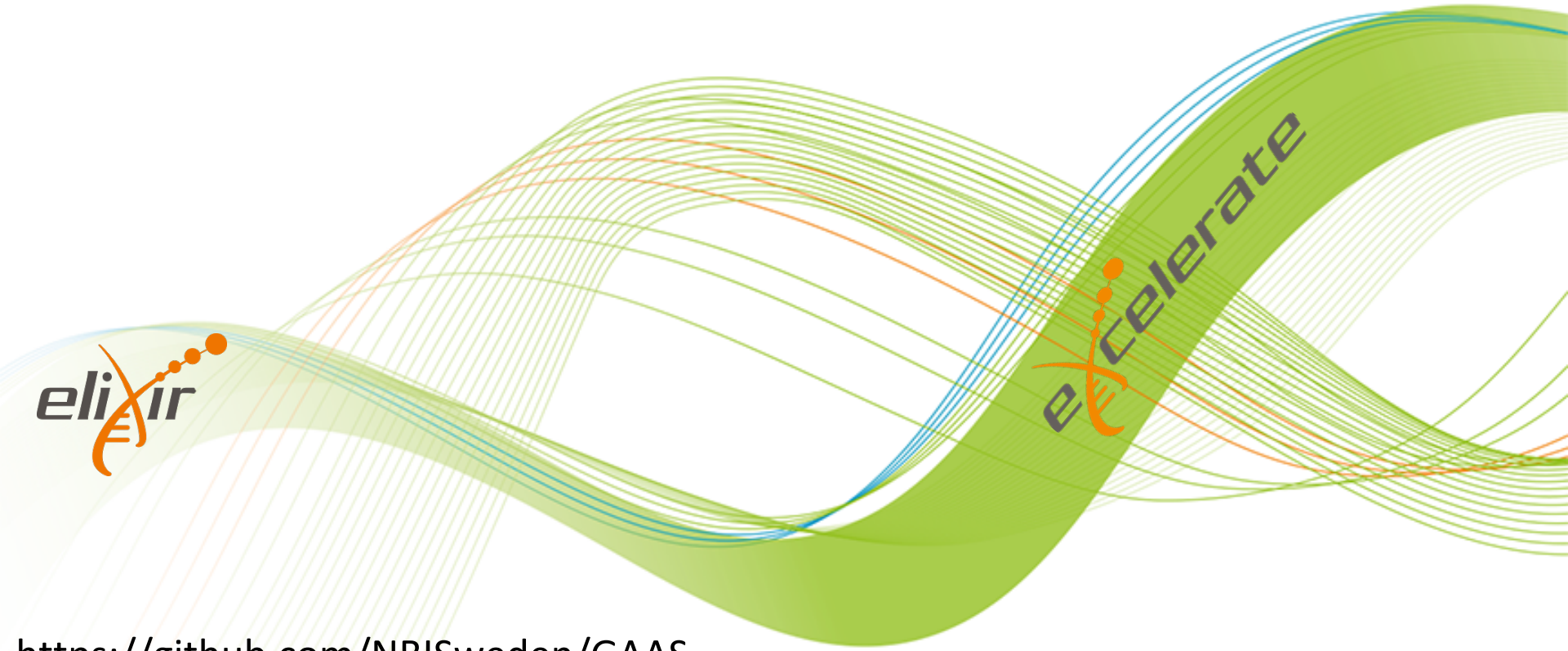
- As for assembly, an annotation is never finished, it can always be improved
  => e.g. Human (to know how to stop)

- Submit your annotation in public archive

SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

THE END



https://github.com/NBISweden/GAAS