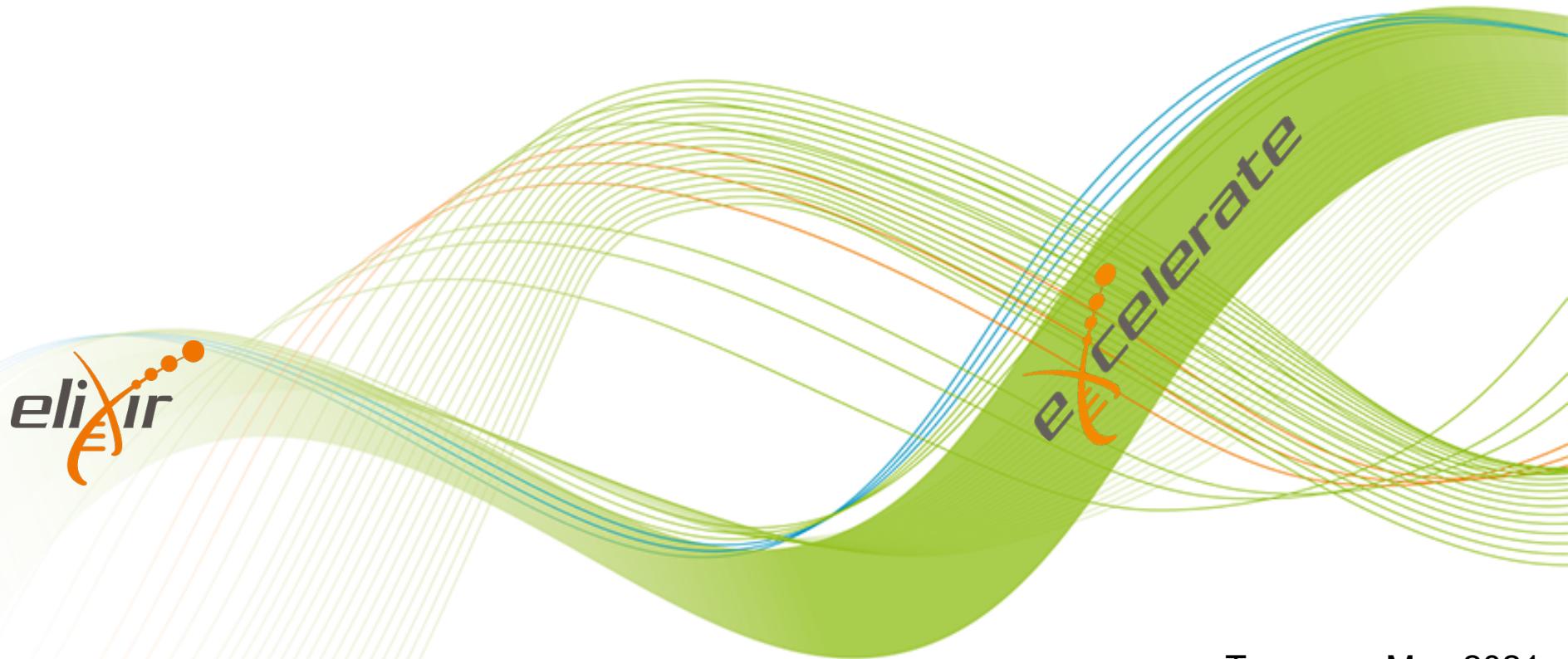




Lucile Soler, PhD
Jacques Dainat, PhD

Methods in genome annotation



- 
1. Introduction
 2. The different annotation methods (coding genes)
 - 2.1 Extrinsic / similarity-based
 - 2.2 Intrinsic / *ab initio*
 - 2.3 Hybrid : *Ab initio* evidence-driven
 - 2.4 comparative
 - 2.5 Combiner / chooser
 - 2.6 Pipeline
 3. Annotation of other genome features
 4. Annotation assessment
 5. To resume / Closing remarks



1. Introduction



What is annotation?

Structural annotation:



Find out where the regions of interest (usually genes) are in the sequence data and what they look like.

=> **Gene prediction / Gene Finding**

functional annotation:

Find out what the regions do.
What do they code for?

*It is the **annotation** that bridges the gap from the sequence to the biology of the organism*

Introduction: Formats



From a genome...

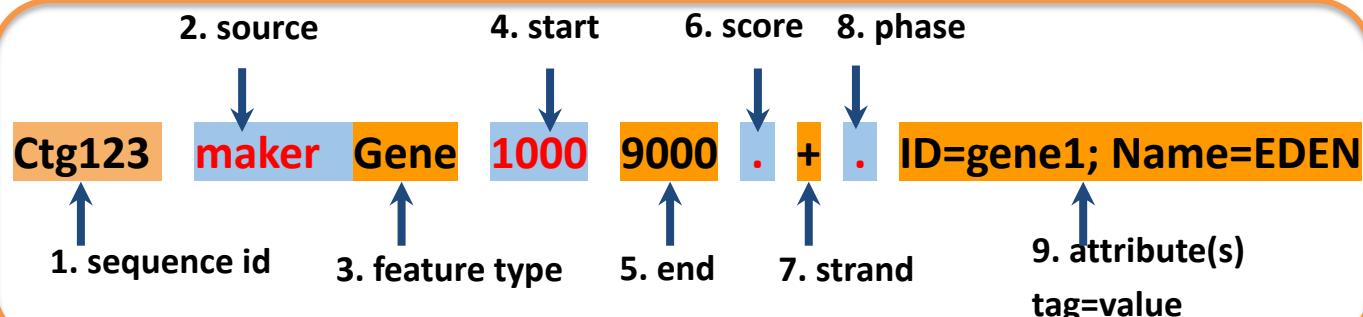
FASTA

...to an annotated gene

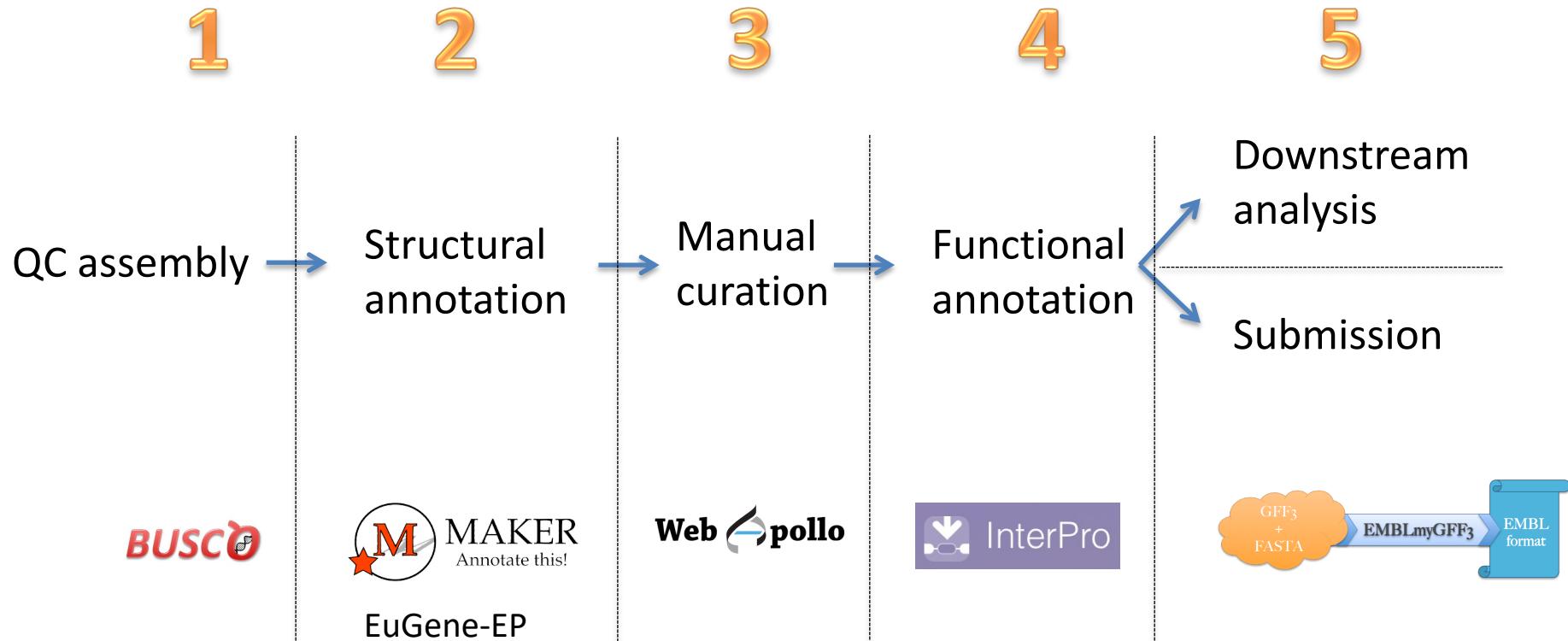
GTF/GFF



- 9 columns
 - 1 feature = 1 line



The main steps in genome annotation





Experimental (ESTs, cDNAs, RNA-seq)

Isolate and clone cognate transcripts (as cDNA), sequence them and compare cDNA with genomic DNA

=> It's the ONLY secure method but:

- Cloning is time consuming.
- Lowly expressed genes are difficult to detect.
- ...

Predictive

- Intrinsic / *ab initio*
- Extrinsic
- Hybrid



Intrinsic / *Ab-initio*

Extrinsic

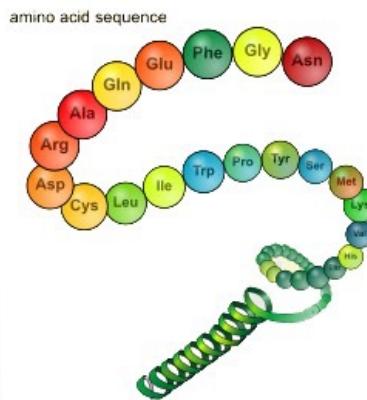
∅

- Use of information/features from the sequence itself

This space intentionally left blank.

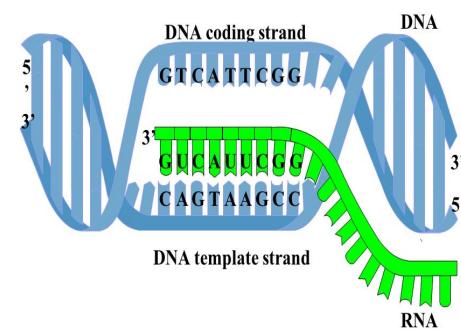
Proteins

- Known amino acid sequences from other organisms



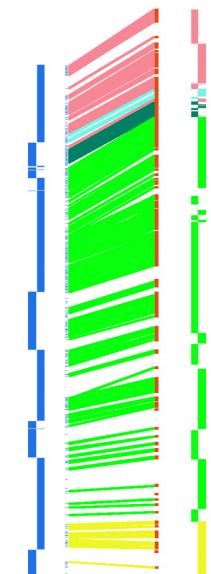
Transcripts

- Assembled from RNA-seq or downloaded ESTs



Genomes

- Close relative genomes





2. The different methods/approaches

The different approaches



- **Similarity-based methods :**
These use similarity to annotated sequences like proteins, cDNAs, or ESTs
- ***Ab initio* prediction :**
Likelihood based methods
- **Hybrid (*Ab initio* evidence-driven) 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



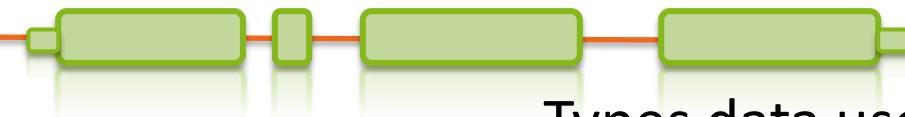
2.1. Extrinsic approaches

Similarity-based methods

These use similarity to annotate sequences like

- Proteins
- Transcripts
- ESTs

The different approaches



Types data used vs methods

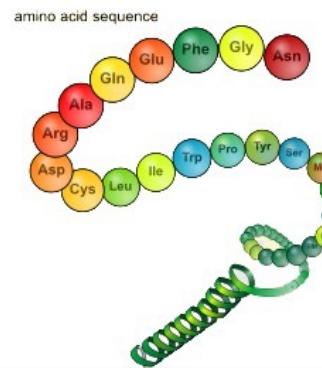
Annotation approach

∅

This space
intentionally
left blank.

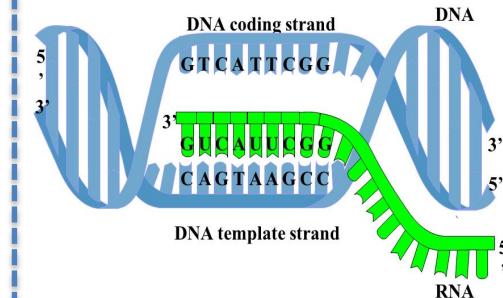
Proteins

Known amino acid sequences
from other organisms



Transcripts

Assembled from RNA-seq or
downloaded ESTs



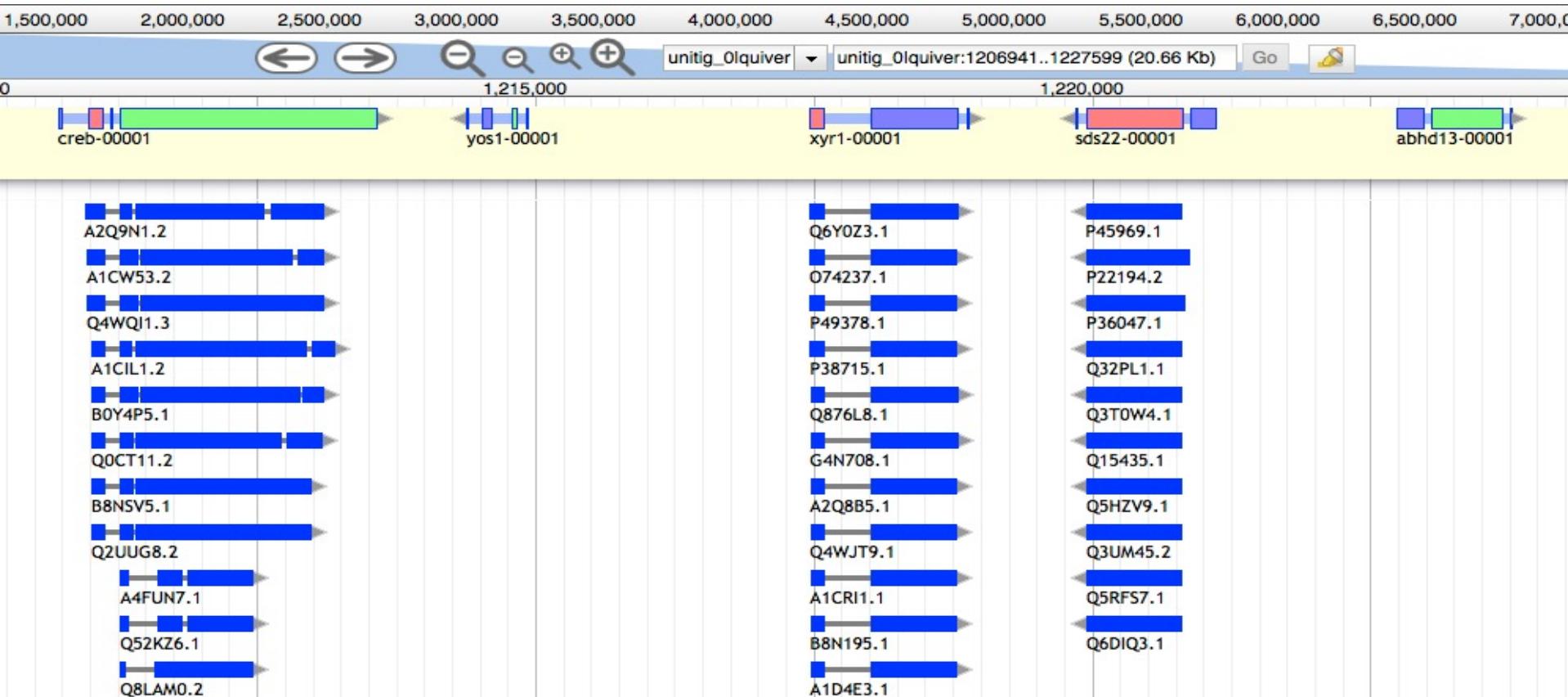
Similarity		X	X
Pure ab initio	X		
Hybrid	X	X	X
Comparative	X	X	X
Chooser/combiner	X	X	X
Pipeline	X	X	X

Similarity-based method: Protein data



Protein sequences are aligned to the genome

- Conserved in sequence => conserved annotation with little noise



Similarity-based method: Protein data

Limits :

- Related to pre-existing data
- Proteins from model organisms often used => bias?
- Proteins can be incomplete
- Protein can be wrong (PE)
- No UTR

```
>sp|Q9NSK7|CS012_HUMAN Protein C19orf12 OS=Homo sapiens OX=9606 GN=C19orf12 PE=1 SV=3
MERLKSHKPATMTIMVEDIMKLLCSLSGERKMKAAVKHSGKGALVTGAMAFVGGLVGGPP
GLAVGGAVVGGLLGAWMTSGQFKPVPQILMELPPAEQQRLFNEAAAIIRHLEWTDAVQLTA
LVMGSEALQQQLLAMLVNYVTKELRAEIQYDD
```

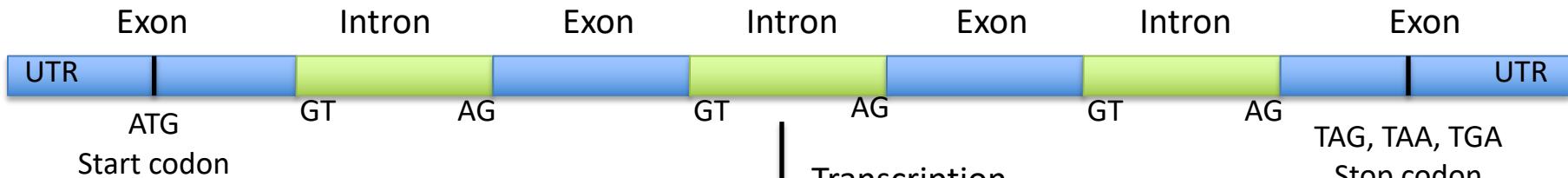
```
>sp|Q2V2T9|SCL16_ARATH Putative scarecrow-like protein 16 OS=Arabidopsis thaliana OX=3702 GN=SCL16 PE=5 SV=1
MQIPTLIDSMANKLHKPPPLKLTVIASDAEFHPPPLGISYEELGSKLVNFATTRNVA
MEFRIISSSSYSDGLSSLIEQLRIDPFVFNEALVNCHMMHLHYIPDEILTSNLRSVFLKEL
RDLNPTIVTLIDEDSDFTSTNFISRLRSLYNMWIPYDTAEMFLTRGSEQRQWYEADISW
KIDNVVAKEGAERVERLEPKSR
```

Similarity-based method: RNA-seq data



Types of data used: RNA-seq

DNA



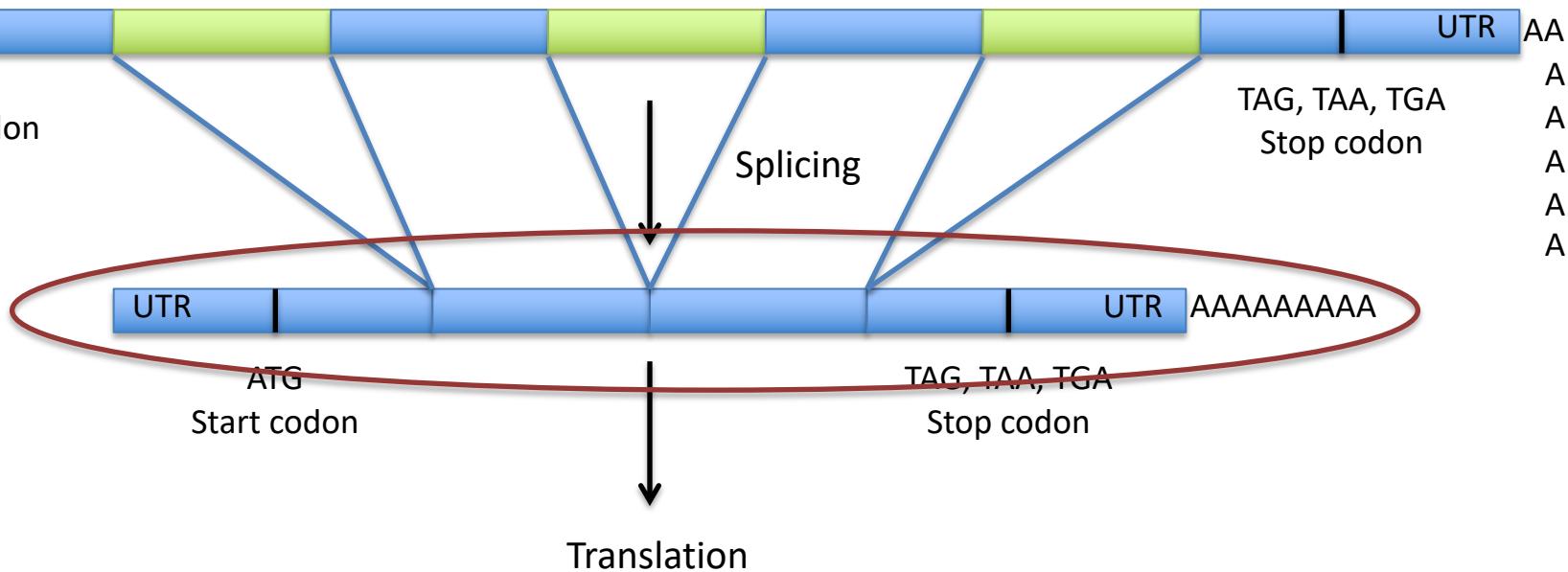
Transcription

Pre-mRNA



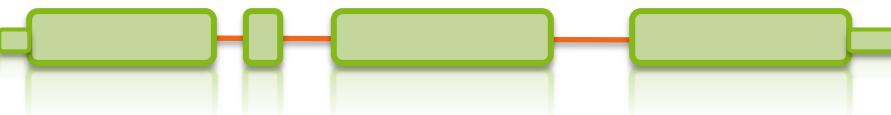
Splicing

mRNA



Translation

Similarity-based method: RNA-seq data



- From the same organism as the genomic data => unbiased
- Sample different tissues or life stages if possible
- short-reads:
 - need to be assembled first
 - Genome guided assembly
⇒ Stringtie: mapped reads -> transcripts
 - *De novo*
=> Trinity: assembles transcripts without a genome
- Long-reads: IsoSeq ...





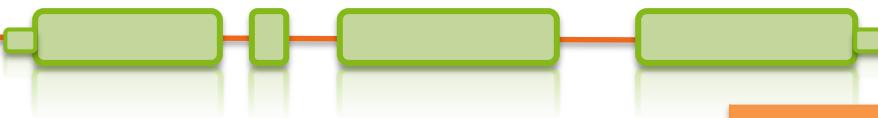
Limits :

/!\ Can be very noisy (tissue/species dependent), can include pre-mRNA
=> Avoid gonads and brain ; muscle is good

Hard to catch low expressed / peculiar expressed (stage of life, condition, etc...) / isoforms

- short-reads:
 - Transcriptome assembly errors
- Long-reads:
 - error rate / frameshift / indels

Similarity-based method: Protein / transcripts



- Rough approximation (fast)

DNA	AA
Blastn	Pmatch
Vsearch	tblastn
NSimScan	PSimScan

- Splice-site aware alignment (slow – moderately slow)

		Human				
		Mouse		Chicken		
CDS	EXALIN	3h	5min	41.3s	2h	1min
	Exonerate			41.1s		9min
	GeneSequer	7h	14min	48.2s	3h	2min
	GMAP			49.4s	1min	37.9s
	PairagonX	274h	1min	16.0s	500h	57min
	sim4cc			16.8s		33.9s
	Spaln2TBZX		6min	55.2s		19.3s
	SplignX		14min	2.0s		44.3s
	XAT		2min	2.8s		6min
protein	Exonerate	12h	36min	10.3s	7h	33min
	genBlastG		3min	30.3s		17.0s
	GeneSequer	10h	10min	24.0s	6h	20min
	GeneWise	69h	17min	36.1s	47h	36min
	ProSplign		2h	24.9s	1h	17min
	Spaln2TBZ		4min	32.0s		39.1s

DNA	AA
Exonerate	Exonerate
Gmap	Genewise
GenomeThreader	GenomeThreader



2.2 Intrinsic / *ab initio*

Using a probabilistic models to predict features

The different approaches

Types data used vs methods

Annotation approach	∅	Proteins	Transcripts
Similarity		X	X
Pure ab initio	X		
Hybrid	X	X	X
Comparative	X	X	X
Chooser/combiner	X	X	X
Pipeline	X	X	X



How it works?

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
- ...

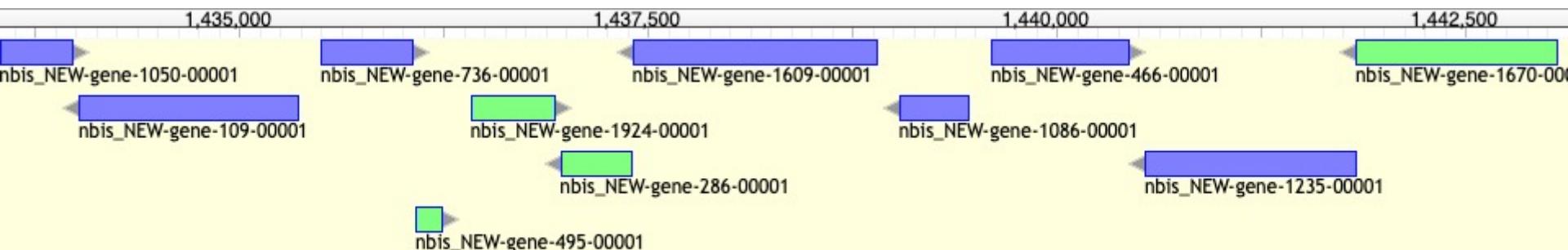
=> *Ab initio* tools will combine this information through different Probabilistic models: HMM, GHMM, WAM, etc.

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

Intrinsic / *ab initio*

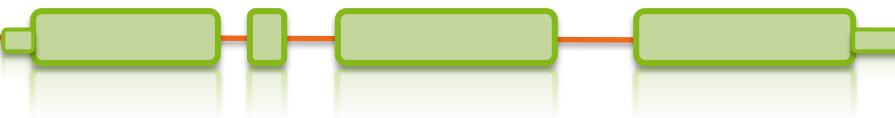


- Bacterial gene prediction
 - Short intergenic regions
 - Uninterrupted ORFs (No intron)
 - Very conserved signals



⇒ easy problem: Accuracy > 90%

Intrinsic / *ab initio*

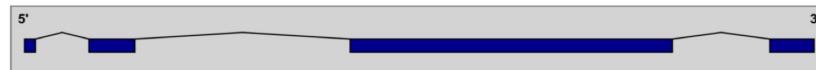


- Eukaryotic gene prediction
 - Presence of intron => structures Differ
 - Isoforms
 - New features (e.g. lncRNAs)

Ostreococcus



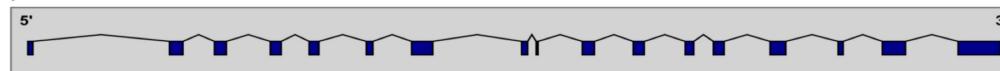
Populus



Spruce



Ectocarpus

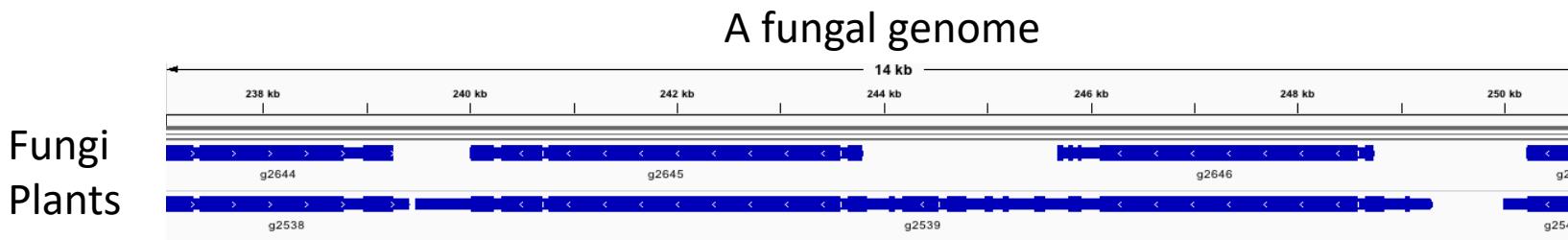


⇒ Hard problem: Accuracy < 70%
⇒ **Requires training**



Training *ab-initio* gene-finders

- *Ab-initio* tools need a probabilistic model (also called profile) => Training
 - Few self-trained tools, most need a separate training procedure
 - The quality of the gene-finder results, hugely, relies on the quality of the training!
 - Needs training for every genome (= different training sets)



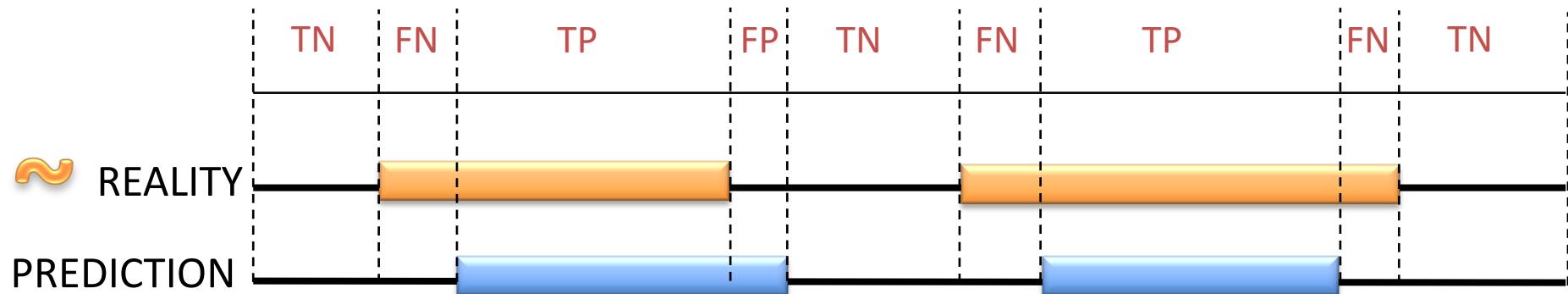
Training:

- sets of high quality genes (>500)
=> These "known" genes are usually inferred from aligned transcripts or proteins

Intrinsic / *ab initio*



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 calculated as the number of correct negative predictions divided by the total number of negatives.

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

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

Intrinsic / *ab initio*



Evaluation of gene prediction										
sensitivity specificity										
nucleotide level	#pred total/ unique	#anno total/ unique	TP	FP = false pos. part ovlp wrng	FN = false neg. part ovlp wrng	sensitivity	specificity	\		
-----/										
exon level	512	472	427	85	45	0.905	0.834	\		
	512	472	427	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	\		
-----/										

0.67

Intrinsic / *ab initio*



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



Supported
by MAKER

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

Intrinsic / *ab initio*



Strengths :

- Fast and easy
- Annotate unknown genes
- Sensitivity ok
- Need no external evidence

Limits :

- No UTR
- No alternatively spliced transcripts (augustus does)
- Bad specificity (Over prediction of exons or/and genes)
- **Training needed** (Need external evidence)

Common errors in annotation:

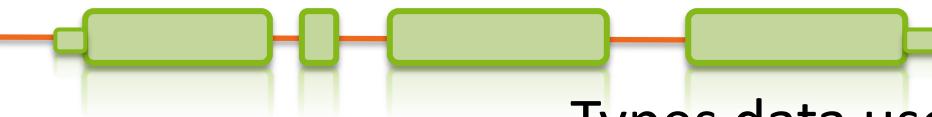
- Split single gene into multiple predictions
- Fused with neighboring genes
- Less accurate than homology based method:
 - Exon boundaries
 - Splicing sites



2.3 Hybrid approaches (*Ab initio* evidence-driven)

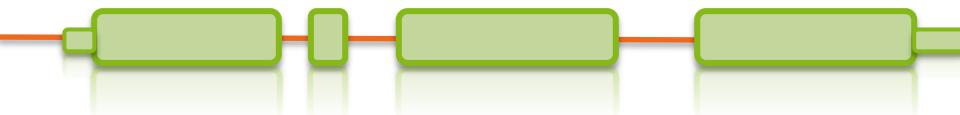
Ab initio tools with the ability to integrate external evidence/hints

The different approaches



Types data used vs methods

Annotation approach	∅	Proteins	Transcripts
Similarity		X	X
Pure ab initio	X		
Hybrid (<i>Ab initio</i> evidence-driven)		X	X
Comparative	X	X	X
Chooser/combiner	X	X	X



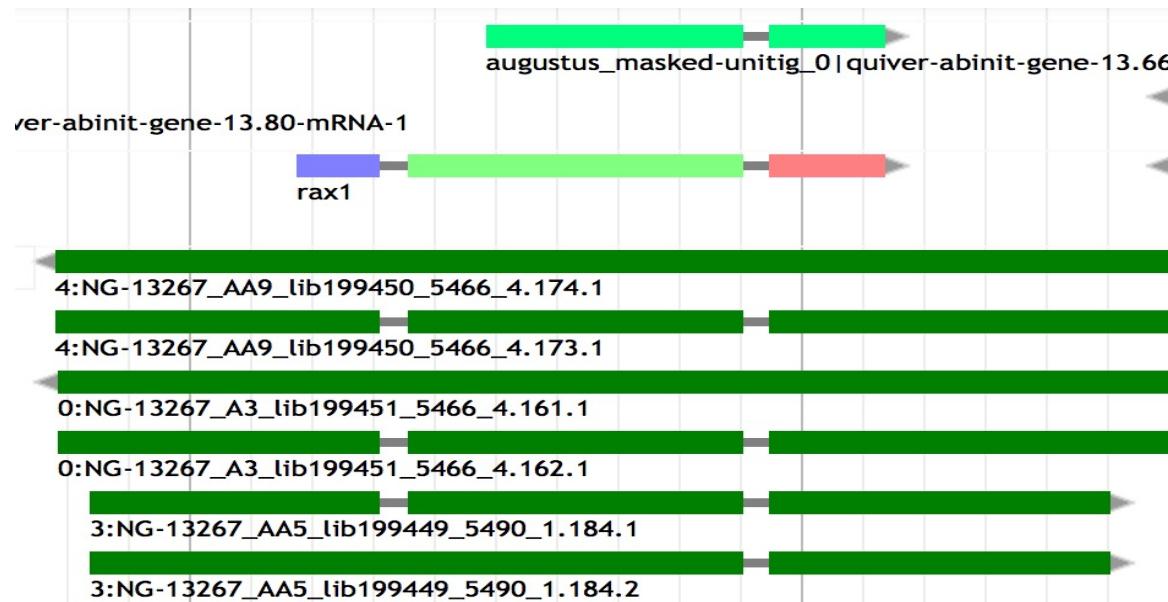
Evidence-drivable gene predictors approaches incorporate hints/extrinsic information in the form of alignments (transcript, protein, whole genome) to increase the accuracy of the gene prediction.

- Can predict locus without extrinsic information
- Improve prediction when extrinsic information available



Strength :

- Lot of data available
- Protein well conserved
- High accuracy



Limits :

- Extra computation to generate alignments
- heterogeneous sequence quality :
 - Incomplete
 - Error during transcriptome assembly
 - Contamination
 - Sequence missing
 - Orientation error



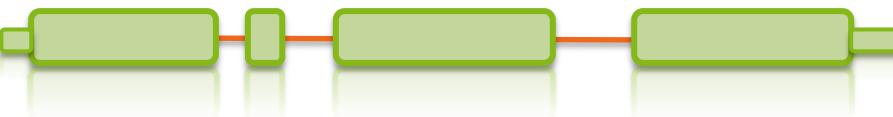
Tools

- **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 !

* Can be seen as combiner



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.

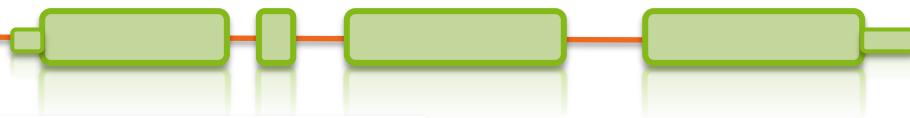
BRAKER2 since 2019 (Incorporate Protein Homology Information)



2.4 Comparative based method

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

The different approaches



Intrinsic / *Ab-initio*

Extrinsic

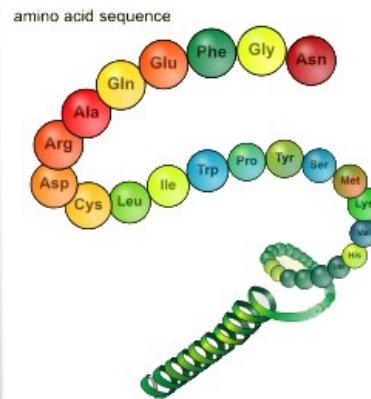
∅

- Use of information/features from the sequence itself

This space intentionally left blank.

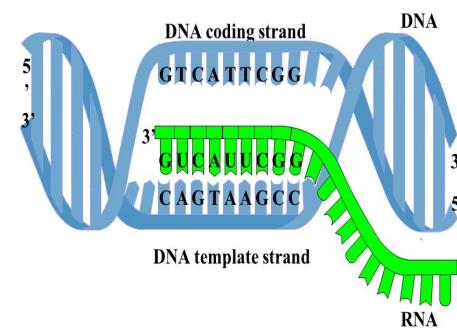
Proteins

- Known amino acid sequences from other organisms



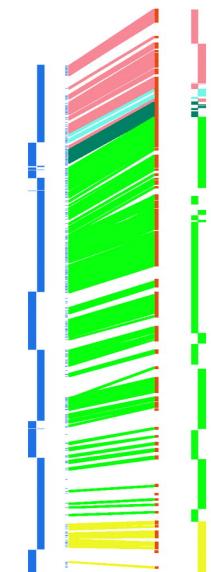
Transcripts

- Assembled from RNA-seq or downloaded ESTs

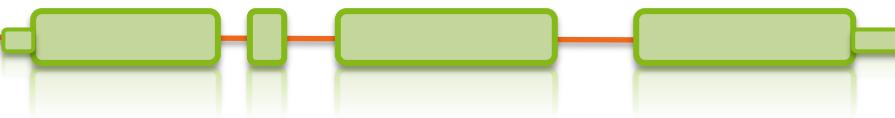


Genomes

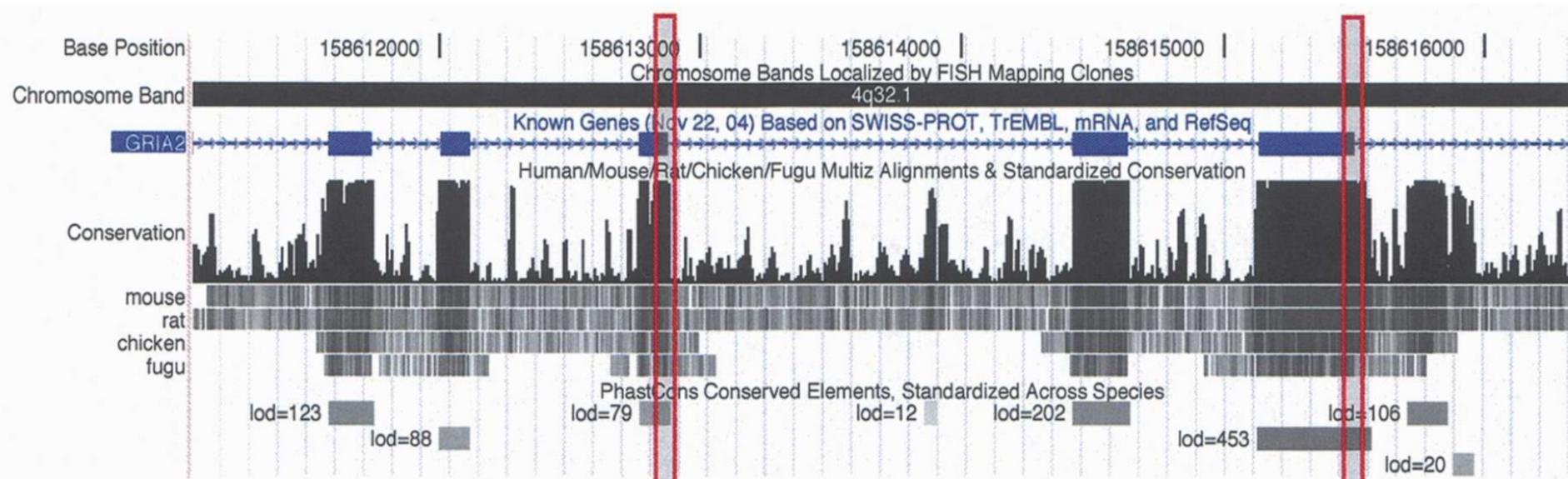
- Close relative genomes



Comparative-based method



The main assumption of these methods is that the functional parts of an eukaryotic genomic sequence, the exons, tend to be more conserved than the non-functional ones, the introns.



These align genomic sequences from different species and use the alignments to guide the ab-initio gene predictions.

- Limits :**
- Whole genome alignment is time/memory consuming
 - Need relatively close related genome (<50 My)

Comparative-based method: Tools



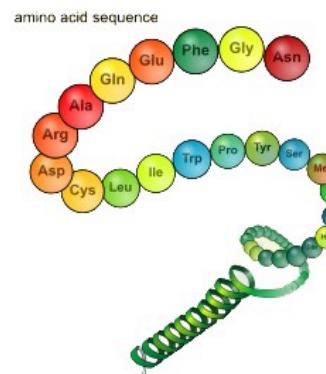
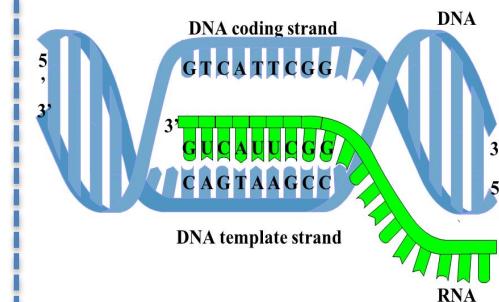
- **Dual genome, de novo gene structure prediction:**
 - **RosettaCM** (Pioneer – 2000)
 - **SGP-2** (2001) – considers only the conservation in protein-coding regions
 - **TWINSCAN** (2001) - included models of conservation in splice sites and start and stop codons
 - **SLAM** (2003)
 - **TWAIN** (2005)
- More than 2 genomic sequences:
 - **NSCAN*** (2006)
 - **Conrad ***(CRF, 2007)
 - **CONTRAST*** (CRF, 2008) -> 58% accuracy
 - **GSA-MPSA**
 - **Augustus-CGP***



2.5 Combiner / Chooser

- *Combining heterogeneous data into gene models*
- *Selection of gene models*

Types data used vs methods

Annotation approach	∅	Proteins	Transcripts
	This space intentionally left blank.	 <p>Known amino acid sequences from other organisms</p>	 <p>Assembled from RNA-seq or downloaded ESTs</p>
Similarity		X	X
Pure ab initio	X		
Hybrid	X	X	X
Comparative	X	X	X
Chooser/combiner	X	X	X
Pipeline	X	X	X

Chooser / combiner

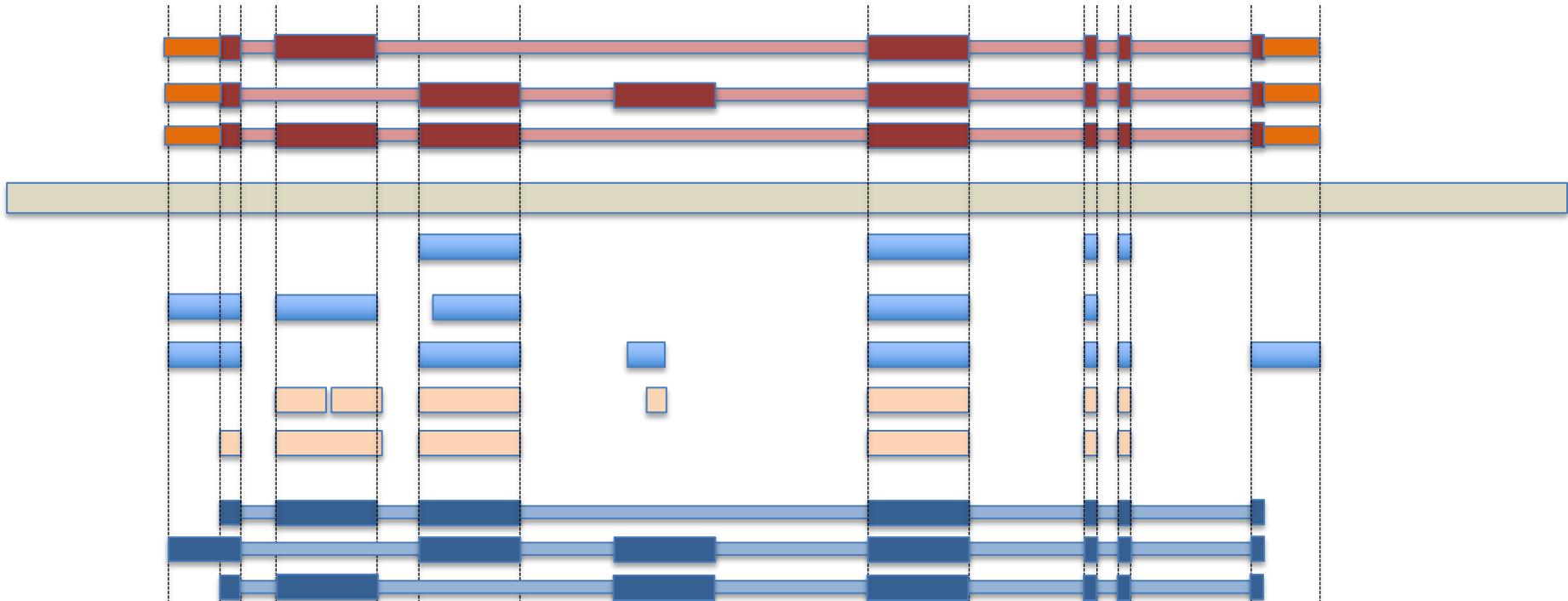
Combiner concept: combining different lines of evidence into gene models

Evidence: ESTs / Transcripts Proteins

Gene prediction (*ab-initio* or evidence-based)

Gene models

=> add untranslated regions (UTR)



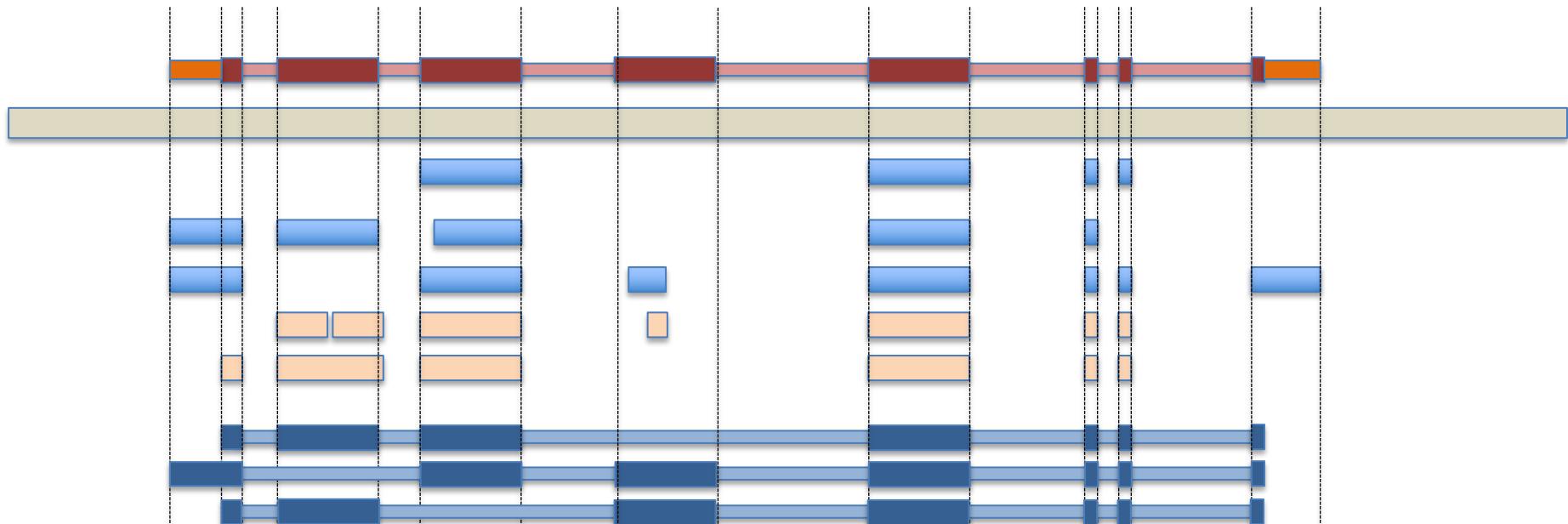
Chooser / combiner

Combiner concept: combining different lines of evidence into gene models

Evidence: ESTs / Transcripts / Proteins

Gene prediction (*ab-initio* or evidence-based)

=> Select the best possible set of exons and combine them in a consensus gene model



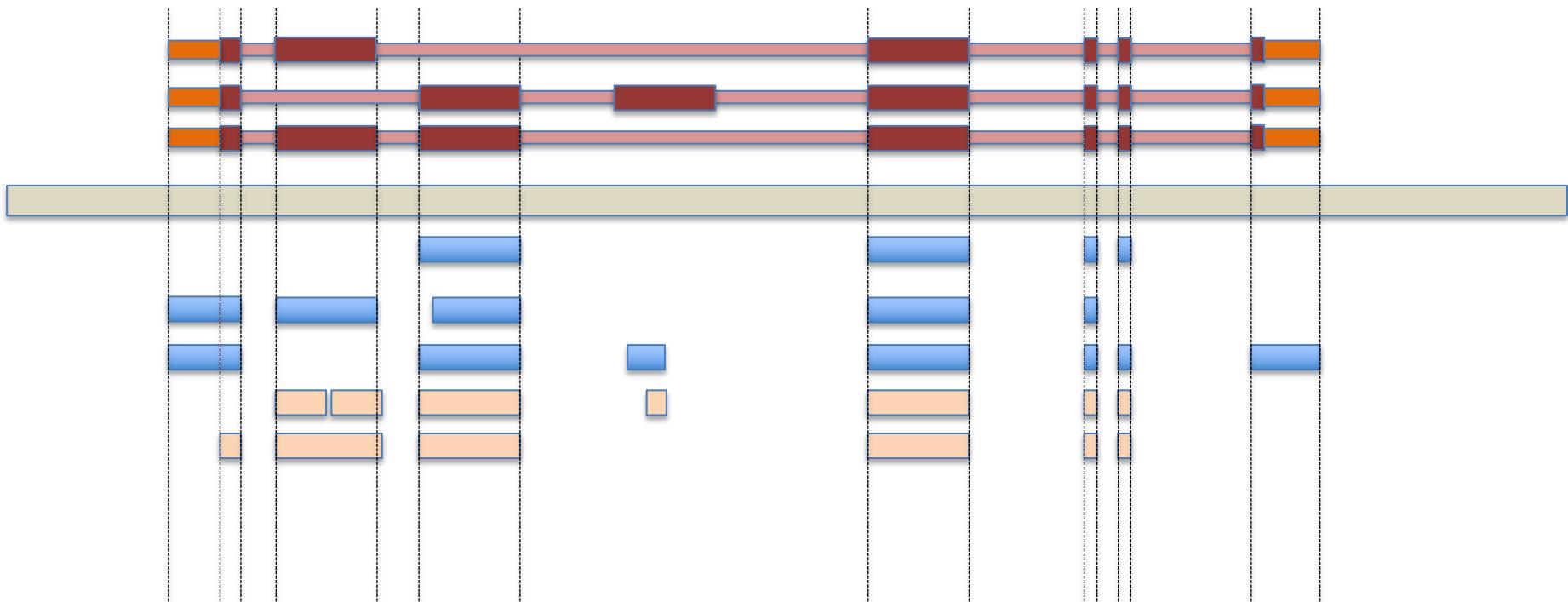
Chooser / combiner

Chooser concept: Select best gene models

Evidence: ESTs / Transcripts / Proteins

Gene models

=> Choose the prediction whose best matches the evidence

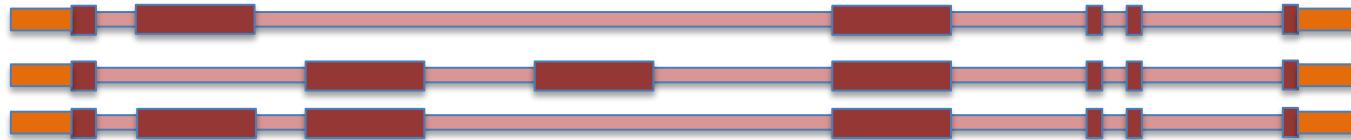


Chooser / combiner

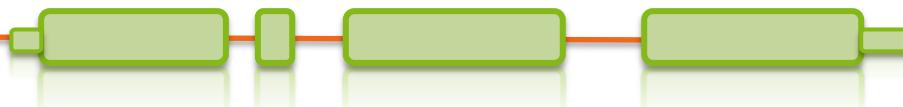
Chooser concept: Select best gene models

Gene models

=> Choose the prediction whose structure best represents the consensus



Chooser / combiner

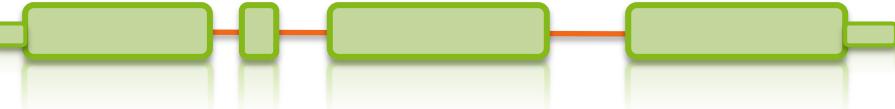


Use gene finders and evidence (EST, RNAseq, protein) alignments and:

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 Evidencemodeleur	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

* Are pipelines that contain chooser



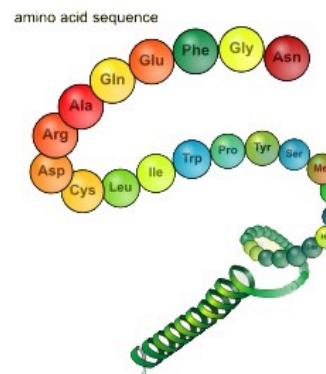
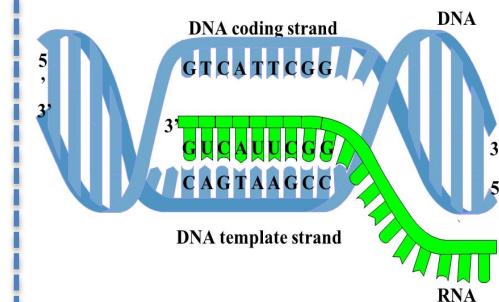
2.6 Pipelines

(The ultimate step)

- *Align evidence*
- *add annotation (UTR, score, gene name)*
- *Annotation of other features (Repeat, tRNA, etc)*
- *And more...*

The different approaches

Types data used vs methods

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

Annotation pipeline



* Evidence-based only

** May use *ab initio*

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, miRNAs, tRNAs, ...

Ensembl** Evidence based only (comparative + homology) ...

Comparative Annotation Toolkit (CAT) *ab-initio* (Augustus) evidence driven + comparative

MAKER2

Evidence based and/or *ab initio* ...



3. Annotation of other genome features

Other genome features



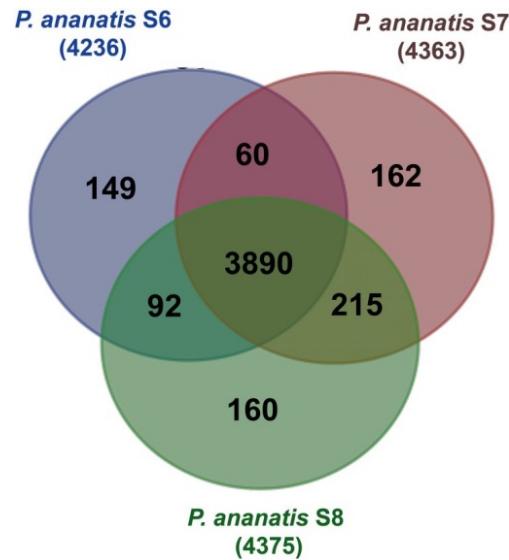
Feature type	DB associated	Tool example	approach
ncRNA	Rfam	infernal	HMM + CM
tRNA	Sprinl database	tRNAscan-SE	CM + WMA
snoRNA		snoscan	HMM + SCFG
miRNA	miRBase	Splign miR-PREFeR (for plant)	sequence alignment Based on expression patterns
Repeats	Repbase, Dfam	repeatMasker	HMM, blast
Pseudogenes		pseudopipe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



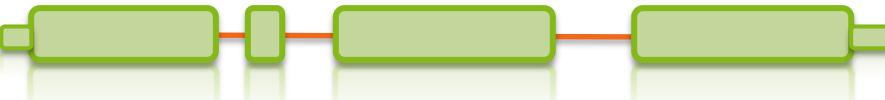
4. Assessing an annotation



- 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



Assessing an annotation



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



5. To resume / Closing remarks

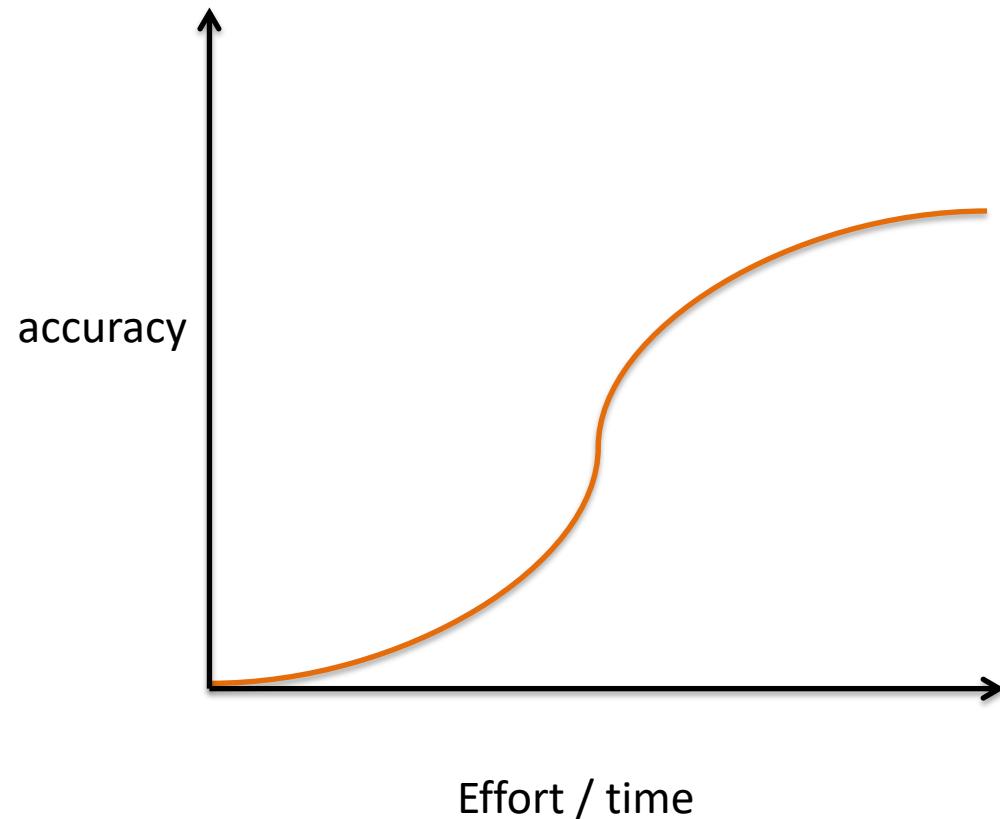


- >100 annotation tools – as many methods
(https://github.com/NBISweden/GAAS/blob/master/annotation/knowledge/annotation_tools_genome.md)
- GTF/GFF format
- 6 main class of approaches (Similarity-based, *ab initio*, hybrid, comparative, combiner, pipeline)

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)

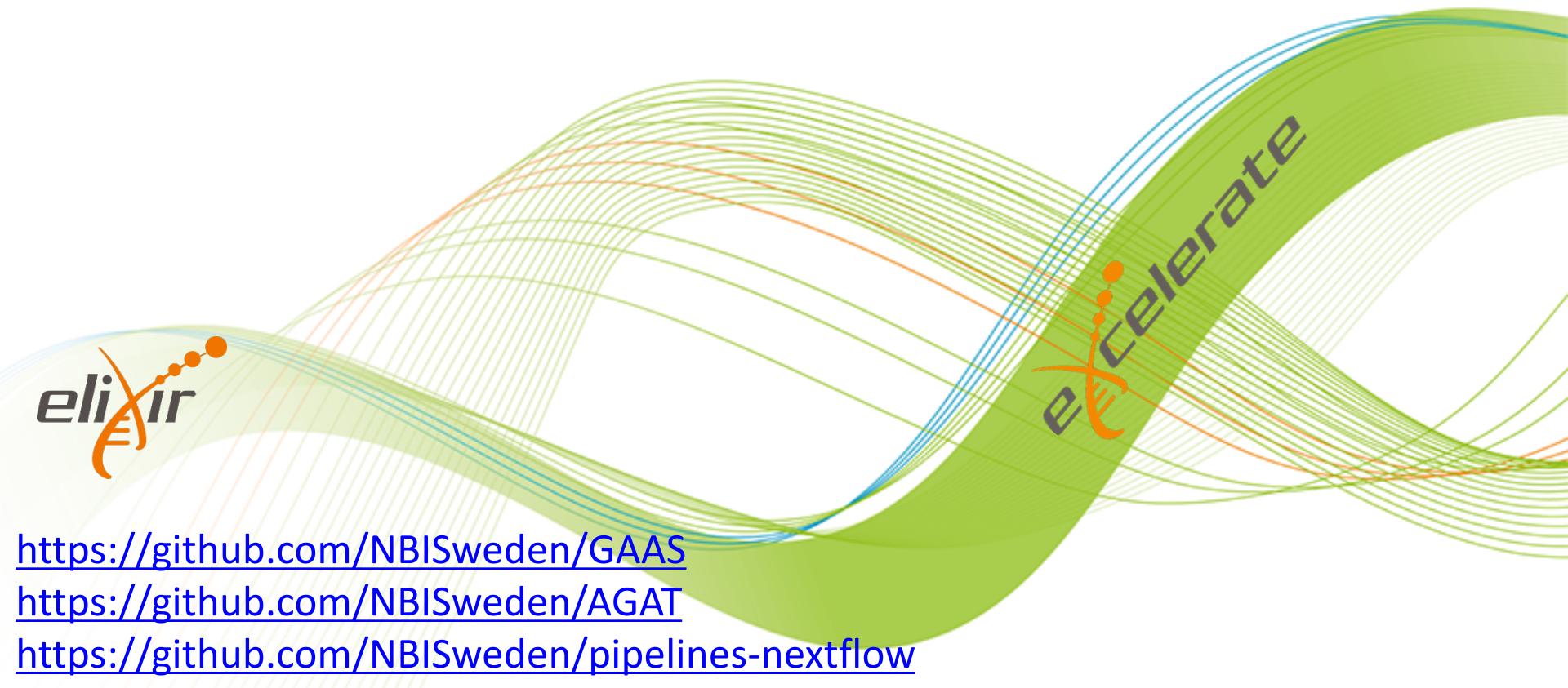
Effort versus accuracy



- Several *ab-initio* tools together give better result than 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 make mistakes !!
- Annotation requires **manual curation**
- As for assembly, an annotation is never finished, it can always be improved
=> e.g. Human (to know when to stop)
- Submit your annotation in public archive



THE END



<https://github.com/NBISweden/GAAS>

<https://github.com/NBISweden/AGAT>

<https://github.com/NBISweden/pipelines-nextflow>