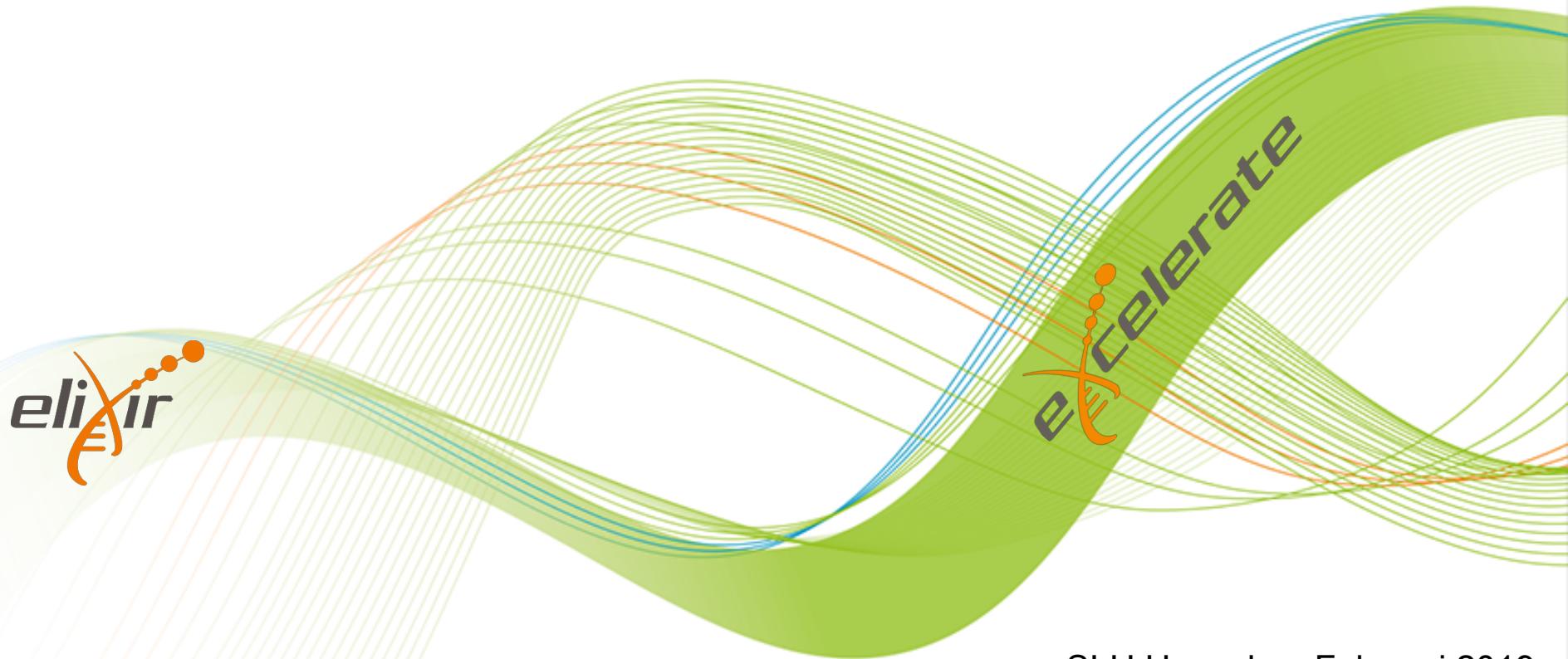




Jacques Dainat PhD

## Methods in genome annotation



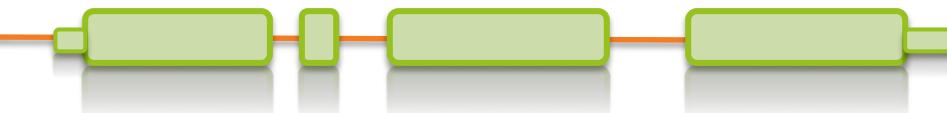


This lecture will focus on eukaryotes

1. Introduction to annotation
2. The different annotation approaches
3. Assessing an annotation
4. Closing remarks



## 1. Introduction to annotation

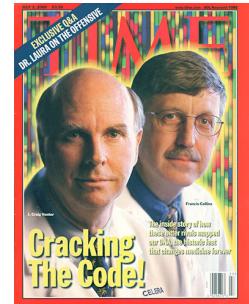


... prices go down

## Human genome sequencing:

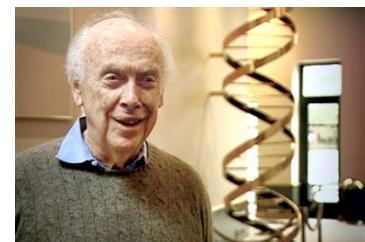
2004: Genome of Craig Wenter costs 70 mln \$

- Sanger's sequencing



2007: Genome of James Watson costs 2 mln \$

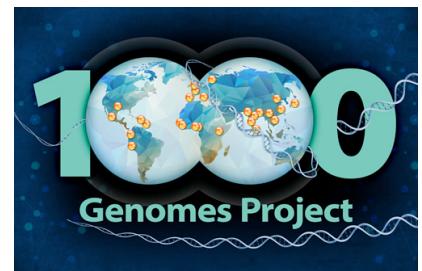
- 454 pyrosequencing

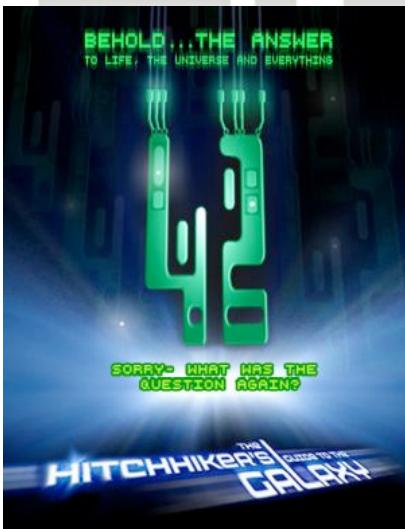


2014: Ultimate goal: 1000 \$ / individual

2016: Illumina Xten: Almost there! (1200 \$)

2017: NovaSeq: "Hold my beer..." (100 \$)





... scientific value diminishes

Science 5 September 1997:  
Vol. 277 no. 5331 pp. 1453-1462  
DOI: 10.1126/science.277.5331.1453

IF 31.6

< Prev | Table of Contents | Next >

ARTICLES

## The Complete Genome Sequence of *Escherichia coli* K-12

Frederick R. Blattner\*, Guy Plunkett III\*, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau and Ying Shao

Journal of Biotechnology  
Article in Press, Corrected Proof - Note to users

IF 2.9

doi:10.1016/j.biote.2010.12.018 | How to Cite or Link Using DOI

Permissions & Reprints

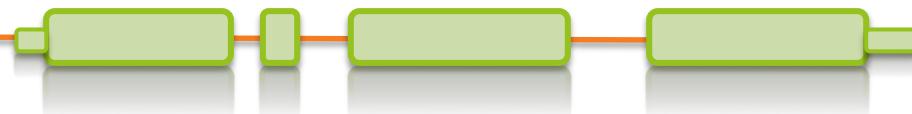


## The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome

Susanne Schneiker-Bekel<sup>a</sup>, Daniel Wibberg<sup>a</sup>, Thomas Bekel<sup>b</sup>, Jochen Blom<sup>b</sup>, Burkhard Linke<sup>b</sup>, Heiko Neuweger<sup>b</sup>, Michael Stiens<sup>a, c</sup>, Frank-Jörg Vorhölter<sup>a</sup>, Stefan Weidner<sup>a</sup>, Alexander Goesmann<sup>b</sup>, Alfred Pühler<sup>a</sup> and Andreas Schlüter<sup>a</sup>

# Let's get philosophical





# What is annotation ?

### Structural annotation:

**vs**

### functional annotation:

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

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 to annotation



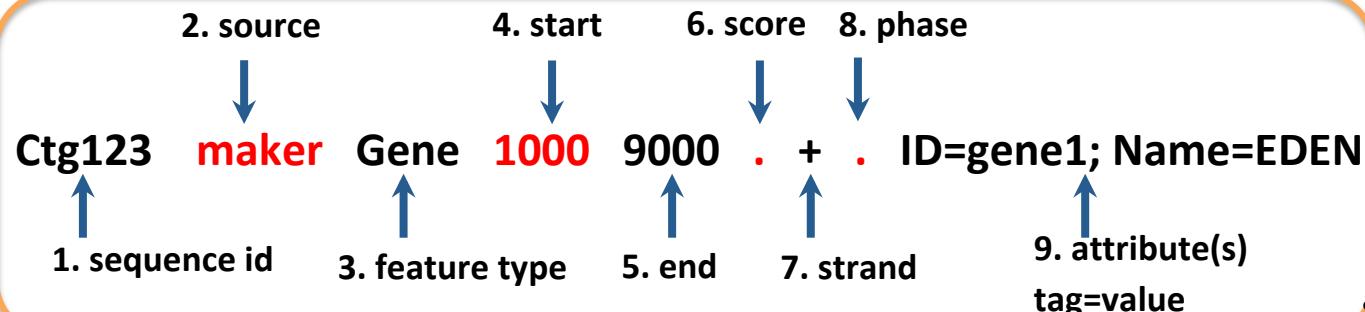
## From a genome...

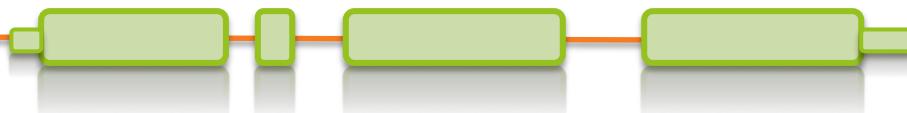
## FASTA

...to an annotated gene  
**GFF**



- 9 columns
  - 1 feature = 1 line





## One gene in GFF3 format:

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . Gene 1000 9000 . + . ID=gene1;Name=EDEN
ctg123 . mRNA 1050 9000 . + . ID=mRNA1;Parent=gene1
ctg123 . exon 1050 1500 . + . ID=exon1;Parent=mRNA1
ctg123 . exon 7000 9000 . + . ID=exon2;Parent=mRNA1
ctg123 . CDS 1201 1500 . + 0 ID=cds1;Parent=mRNA1;Name=edenprotein.1
ctg123 . CDS 7000 7600 . + 0 ID=cds1;Parent=mRNA1;Name=edenprotein.1
```

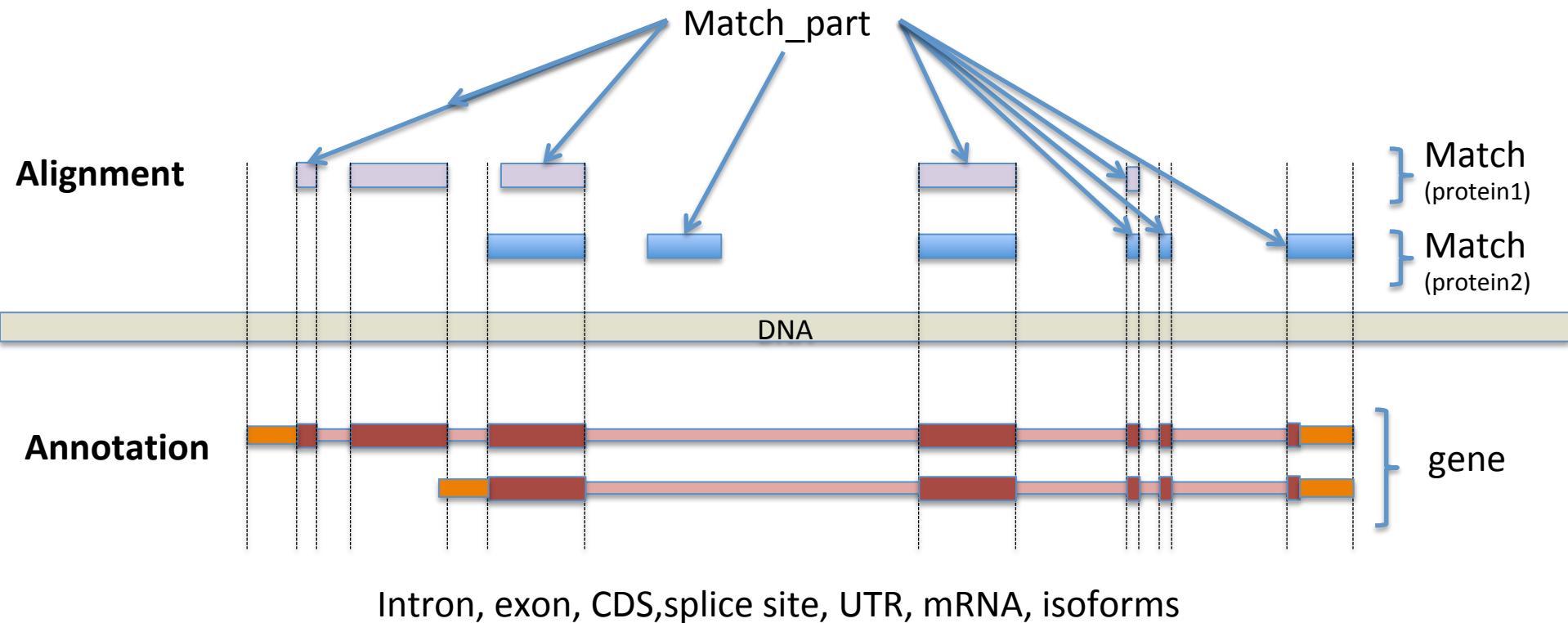
/!\ different version 1, 2, 2.5, 3

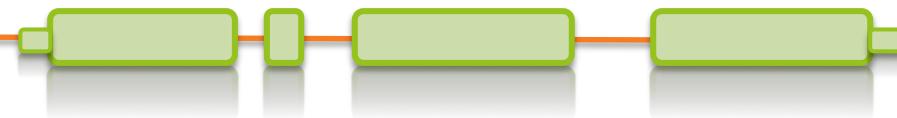
GTF = GFF version 2

## Introduction to annotation

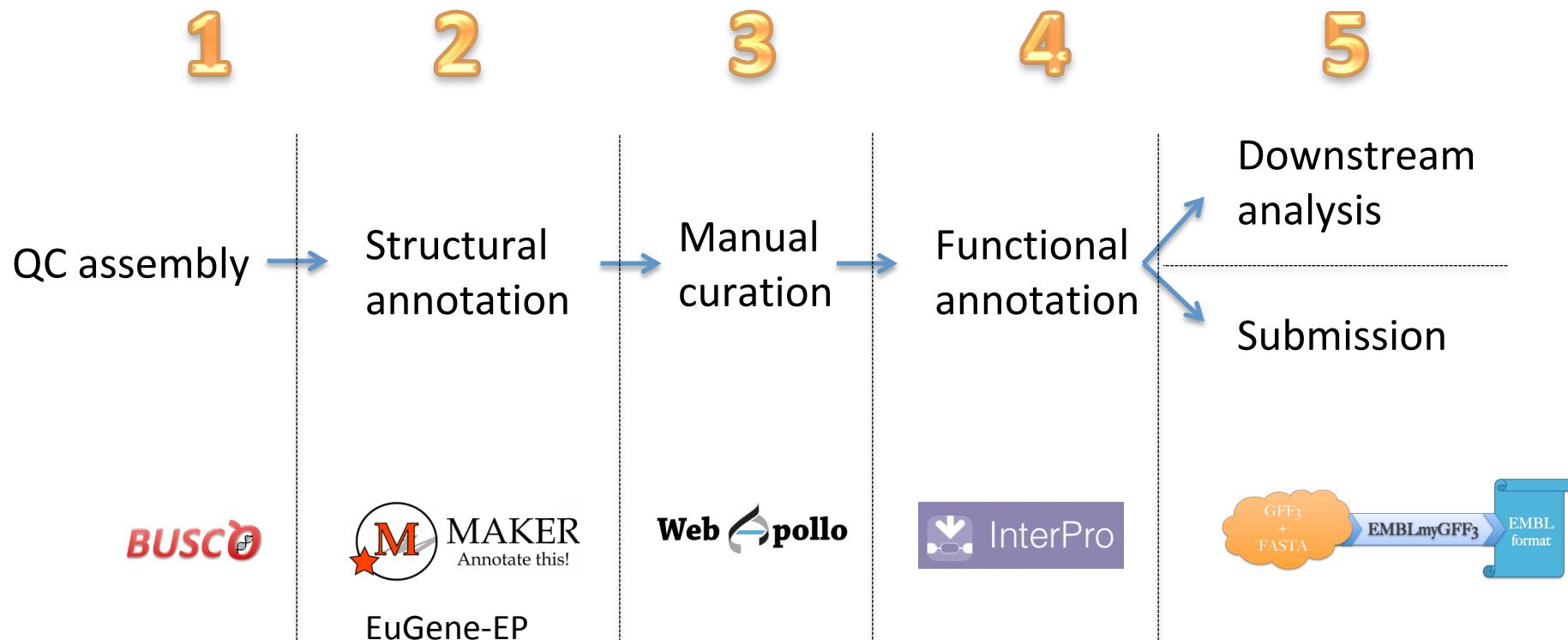


/!\ different type of gff: **annotation / alignment / other**





## The main steps in genome annotation





### Before annotation – check assembly quality

- The quality of the assembly will heavily influence the quality of the annotation
  - ❑ SNP-errors can change start/stop-codons
  - ❑ Indels can cause frame-shifts
  - ❑ High fragmentation could break loci
  - ❑ missing loci cannot be annotated

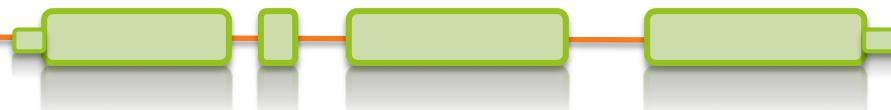
=> Annotation tools have difficulties to deal with those problems



### Assembly check and preparation

- Fragmentation (N50, number of sequences, how many small contigs)
- Sanity of the fasta file (Ns, IUPAC, lowercase nucleotides)
- Completeness / duplication / fragmentation
- Presence of Organelles
- Other (GC content, how distant from other species)





BUSCO output

```
# BUSCO version is: 3.0.2
# The lineage dataset is: fungi_odb9 (Creation date: 2016-02-13,
number of species: 85, number of BUSCOs: 290)
#
# Summarized benchmarking in BUSCO notation for file genome.fa
# BUSCO was run in mode: genome
```

C: 98.6% [S: 97.9%, D: 0.7%], F: 0.0%, M: 1.4%, n: 290

286 Complete BUSCOs (C)

284 Complete and single-copy BUSCOs (S)

2 Complete and duplicated BUSCOs (D)

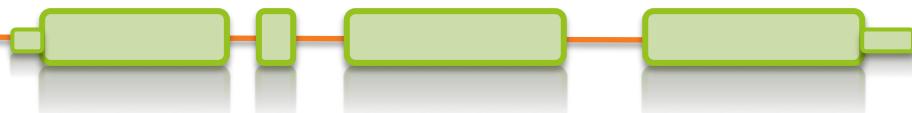
0 Fragmented BUSCOs (F)

4 Missing BUSCOs (M)

290 Total BUSCO groups searched

### Repeat Masking

- Repeatmodeler to find new repeats  
<http://www.repeatmasker.org/RepeatModeler/>
- Repeatmasker to mask known repeats  
<http://www.repeatmasker.org>
  - + Save time
  - + Increase quality of the annotation



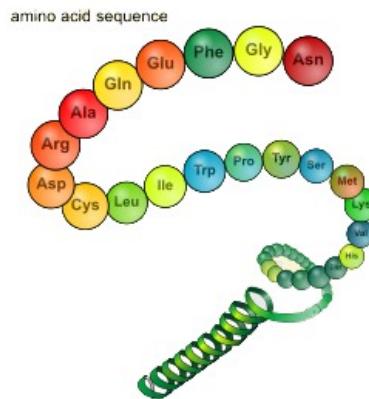
## Types of external data used

∅

### Proteins

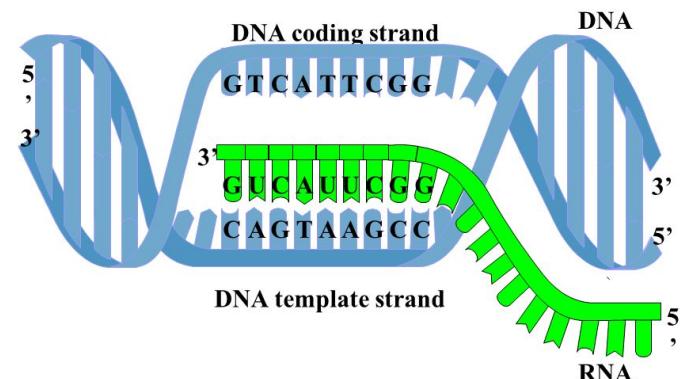
- Known amino acid sequences from other organisms

This space intentionally left blank.



### Transcripts

- Assembled from RNA-seq or downloaded ESTs





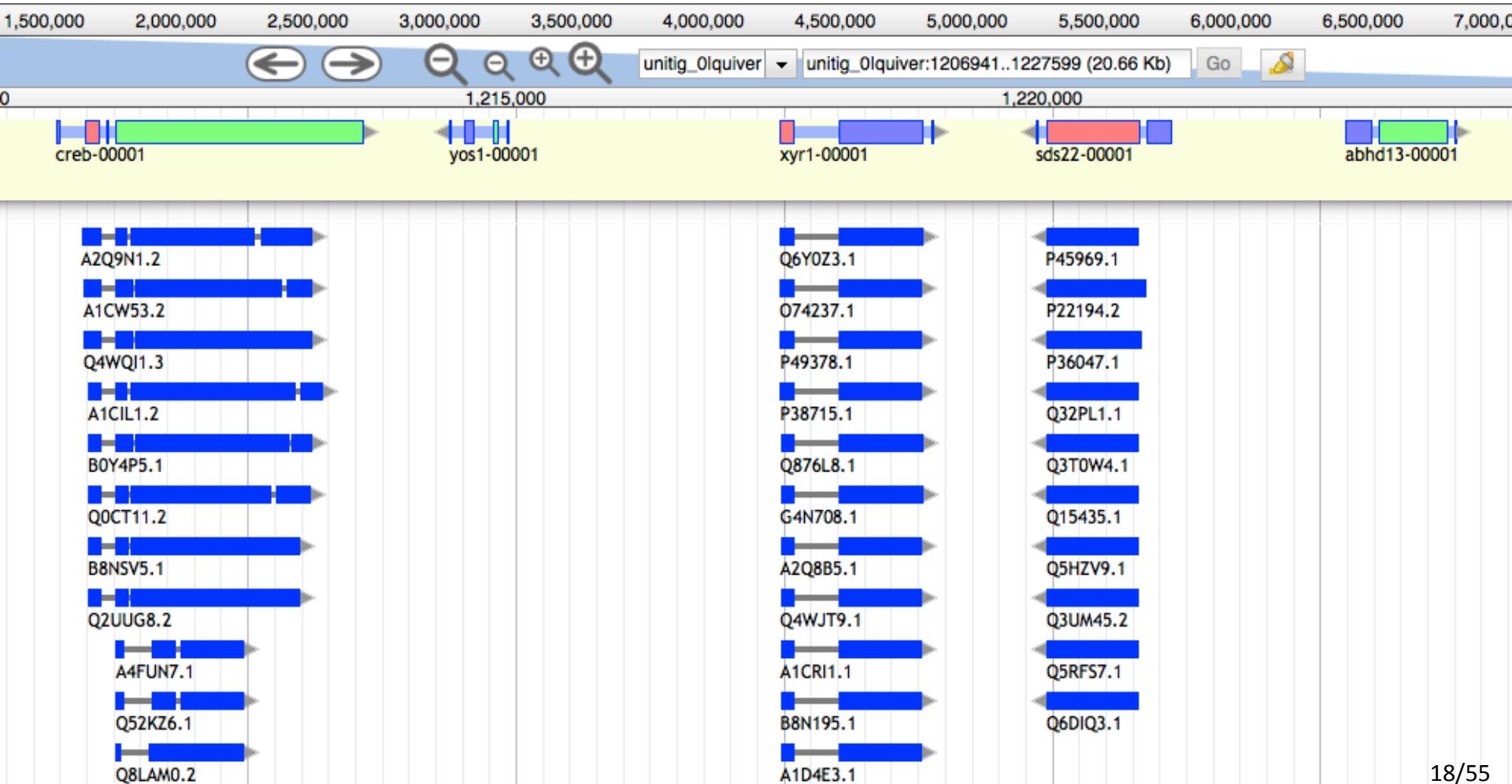
### Types of data used: Proteins

- Conserved in sequence => conserved annotation with little noise
- Proteins from model organisms often used => bias?
- Proteins can be incomplete => problems as many annotation procedures are heavily dependent on protein alignments

```
>ENSTGUP00000017616 pep:novel chromosome:taeGut3.2.4:8_random:2849599:2959678:-1 gene:ENSTGUG00000017338 transcript:ENSTGUT00000018018 gene:  
RSPNATEYNWHLRYPKIPERLNPPAAGPALSTAEGWMILPWNGNGQHPLLARAPGKGRER  
DGKELIKKPPTFKFTFLKKKKKKKKTFK  
>ENSTGUP00000017615 pep:novel chromosome:taeGut3.2.4:23_random:205321:209117:1 gene:ENSTGUG00000017337 transcript:ENSTGUT00000018017 gene:  
PDLRELVLMFEHLHRVRNGGFRNSEVKWPDRSPPYHSFTPQAQKSFSLAGCSEGESTKMG  
IKERMRLSSSQRQGSRGRQQHLGPPLHRSPSPEDVAEATSPTKVQKWSFNDRTRFRASL  
RLKPRIPAEGDCPPEDSGEERSSPCDLTfedIMPavKTLIRAVRLKFVAKRKFKETLR  
PYDVKDVIQEYQSAGHLDMLGRIKSLQTRVEQIVGRDRALPADKKVREKGEKPALEAELVD  
ELSMMMGRVVKVERQVQSIIEHKLDLLGLYSRCLRGSA NSLVLA A VRVPPGEPDVTSDYQ  
SPVEHEDISTSAQSLISRLASTNMD
```



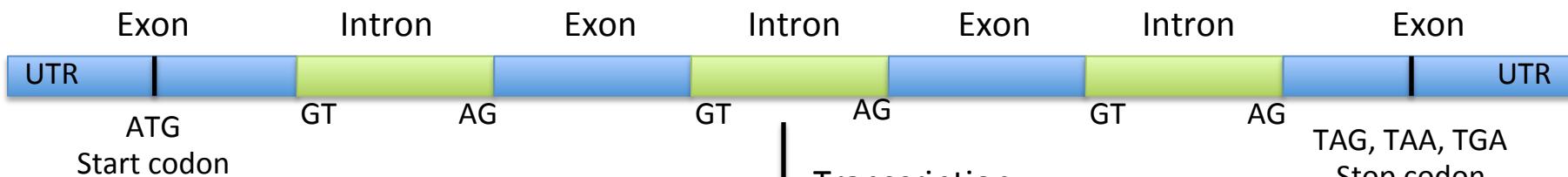
## Protein sequences are aligned to the genome





## Types of data used: RNA-seq

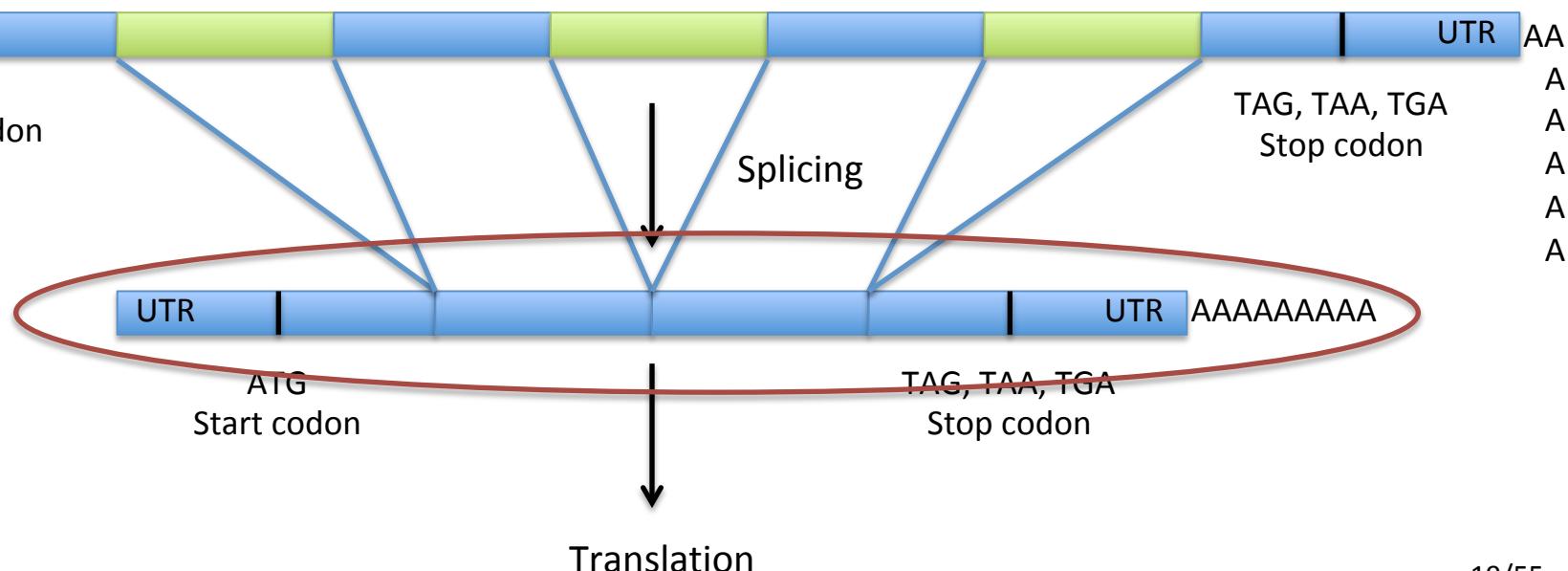
### DNA



### Pre-mRNA



### mRNA



Translation

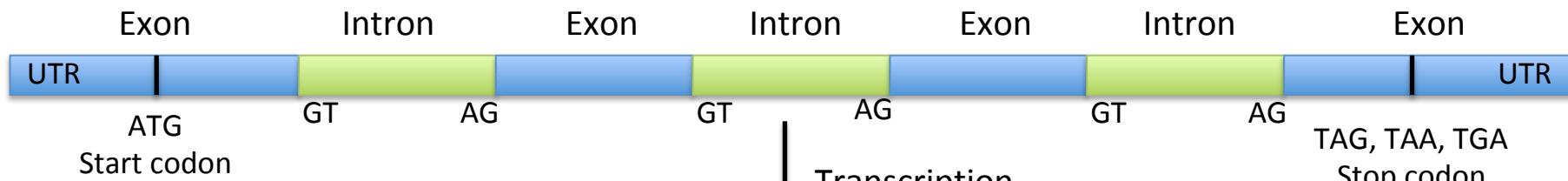


### Types of data used: RNA-seq

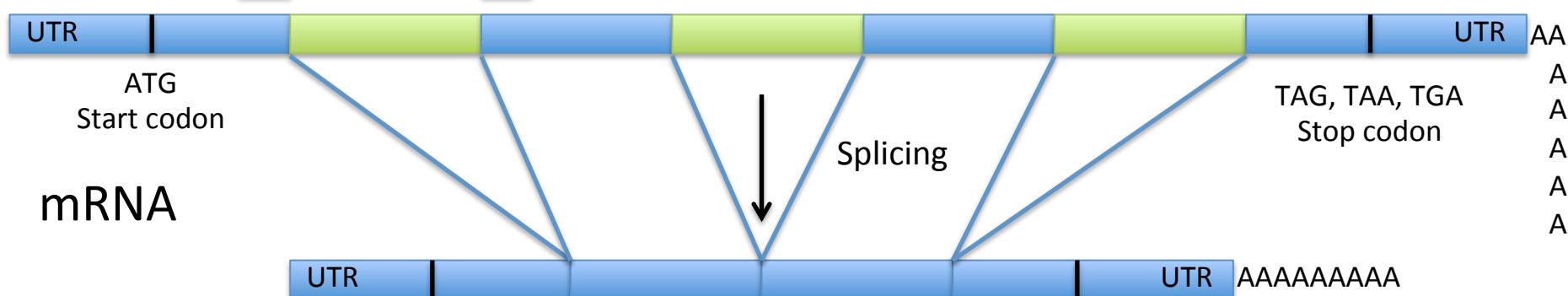
- Should always be included in an annotation project
- From the same organism as the genomic data => unbiased
- /!\ Can be very noisy (tissue/species dependent), can include pre-mRNA
- Sample different tissues or life stages if possible
- Avoid gonads; muscle or liver is good

## RNA-seq - Spliced reads

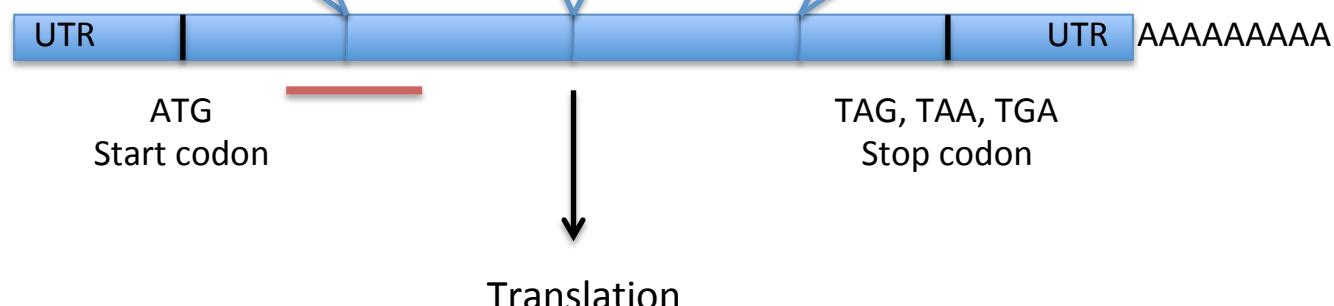
DNA



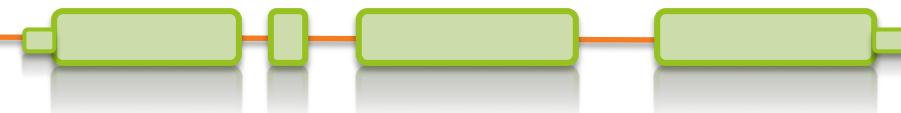
Pre-mRNA



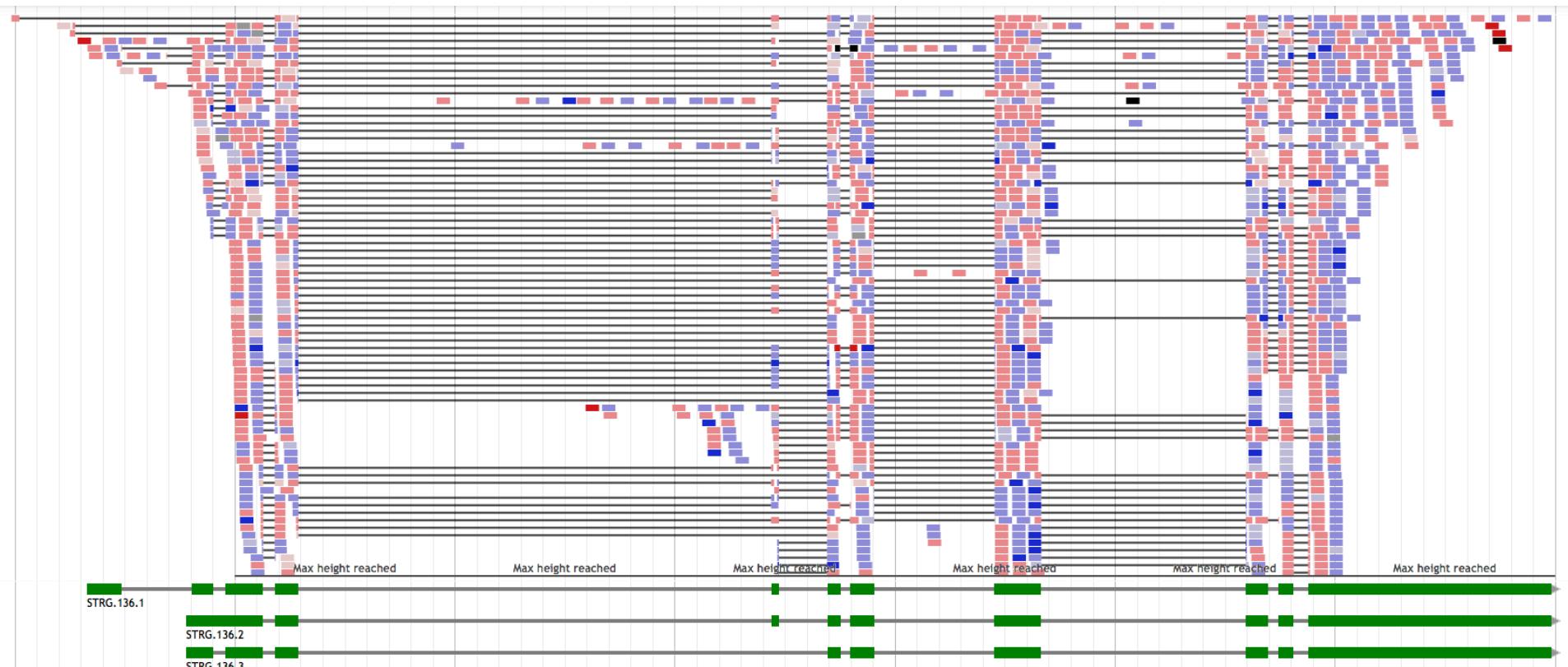
mRNA



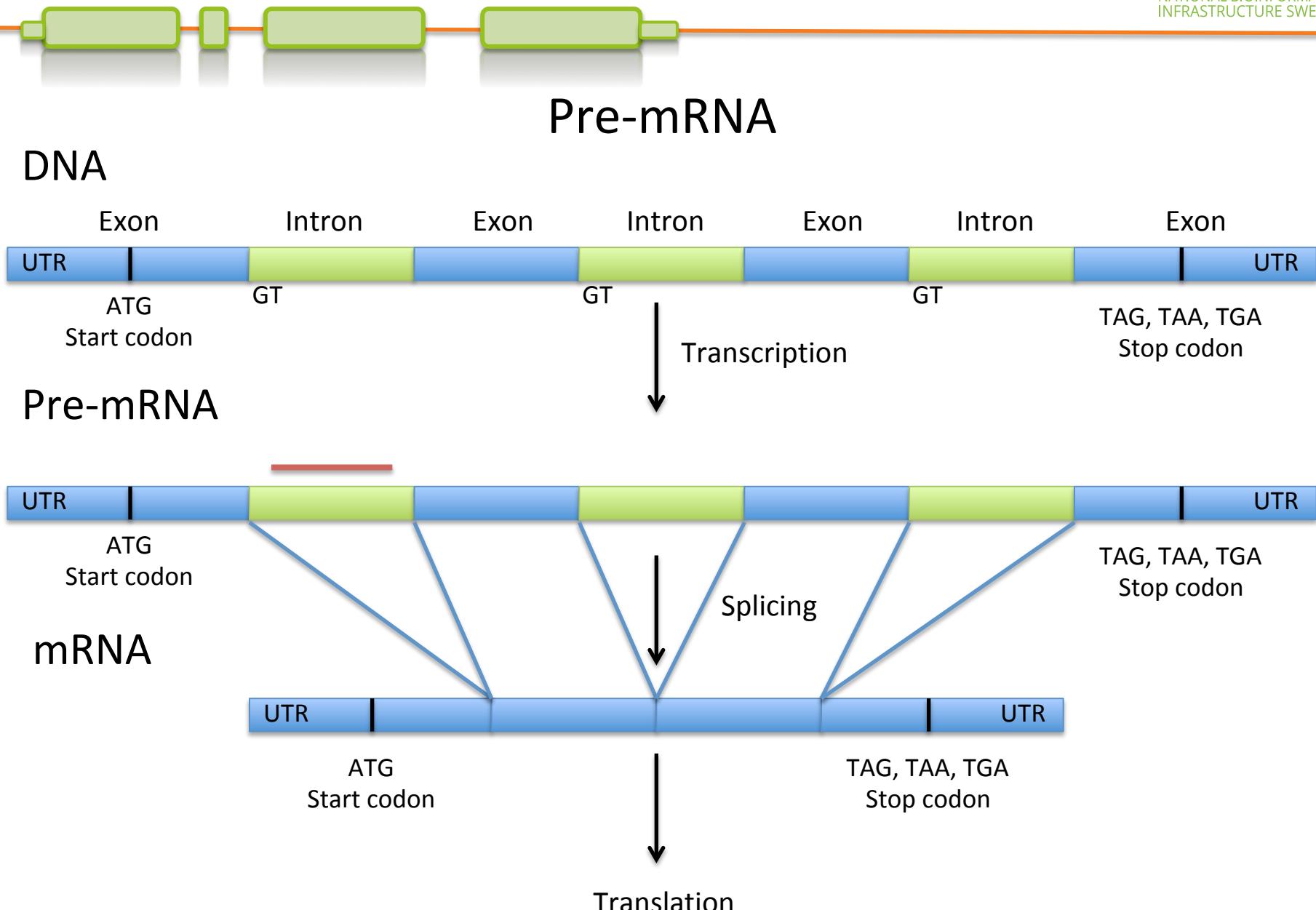
Translation

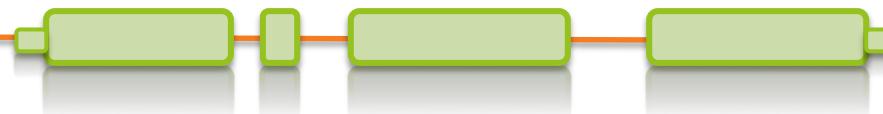


## RNA-seq - Spliced reads

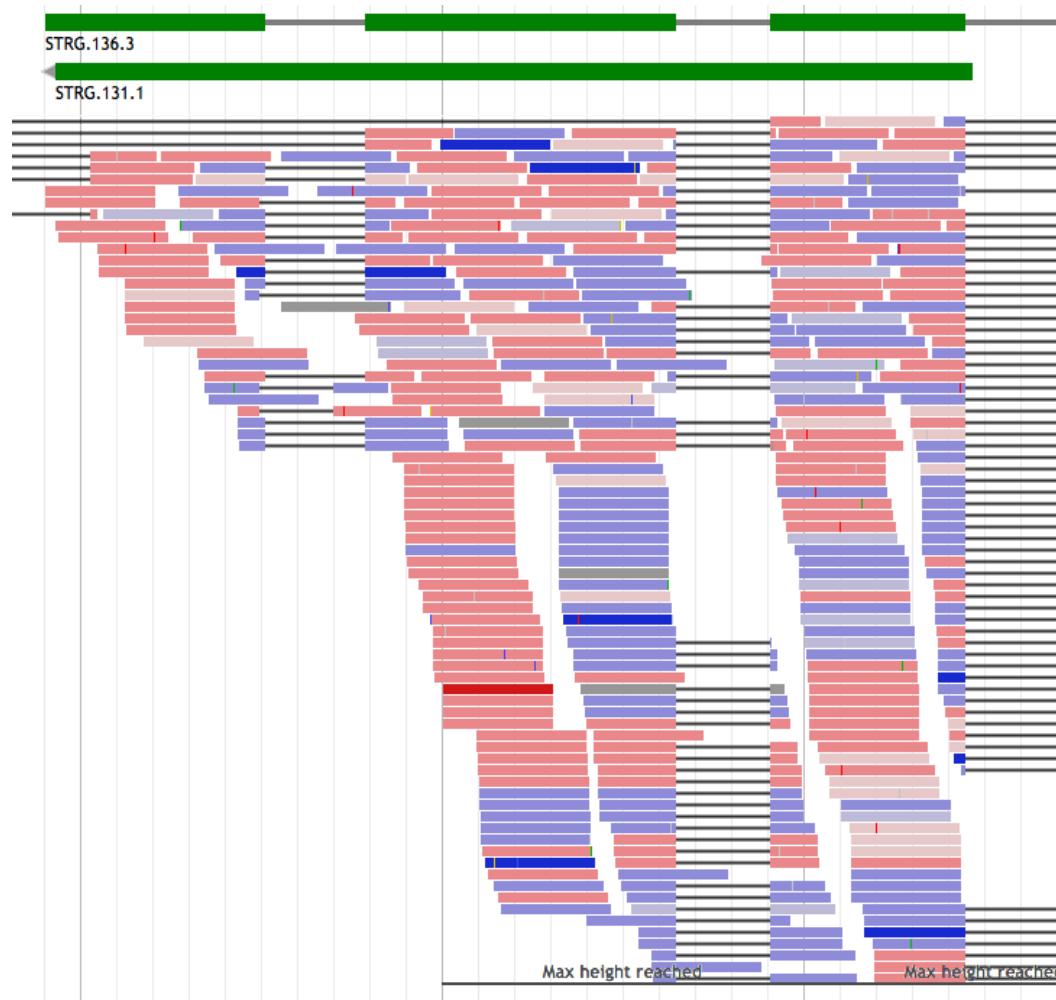


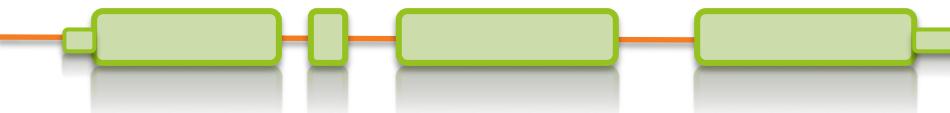
# Introduction to annotation





## RNA-seq – pre-mRNA noise





# Types of data used: RNA-seq

RNA-seq (short-reads) need to be assembled first

- Genome guided assembly  
=> Cufflinks/Stringtie/...: mapped reads -> transcripts
  - *De novo*  
=> Trinity: assembles transcripts without a genome





## 2. The different annotation approaches

# The different approaches



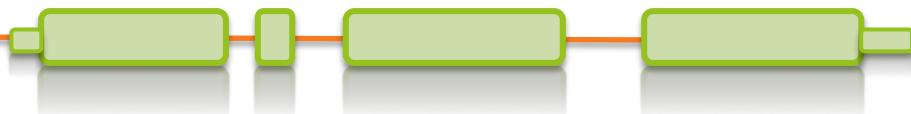
- **Similarity-based methods :**  
These use similarity to annotated sequences like proteins, cDNAs, or ESTs
- ***Ab initio* prediction :**  
Likelihood based methods
- **Hybrid approaches :**  
*Ab initio* tools with the ability to integrate external evidence/hints
- **Comparative (homology) based gene 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. The different annotation approaches

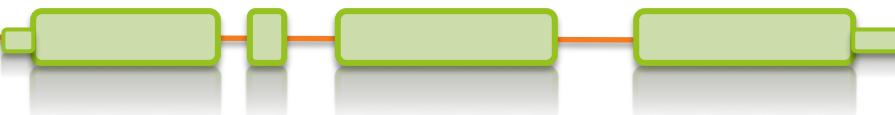
### 2.1 *Ab-initio* annotation tools “intrinsic approach”

## *Ab initio* method



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





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

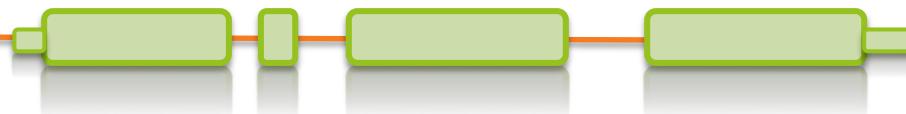
- codon usage
- hexamer usage
- GC content
- compositional bias between codon positions
- nucleotide periodicity
- exon/intron size
- ...

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.

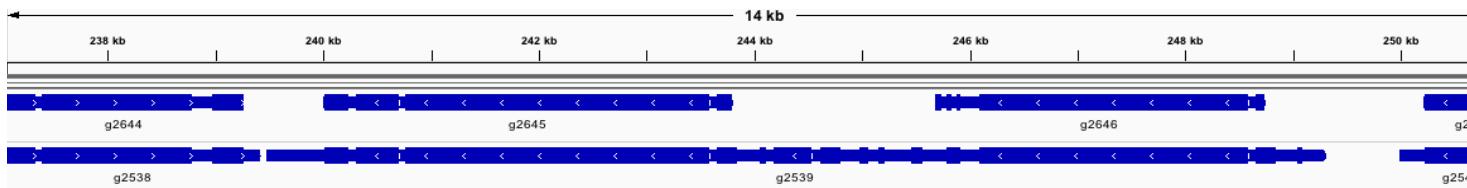
These models need to be created if not already existing for your organism => **training!**



## Training *ab-initio* gene-finders

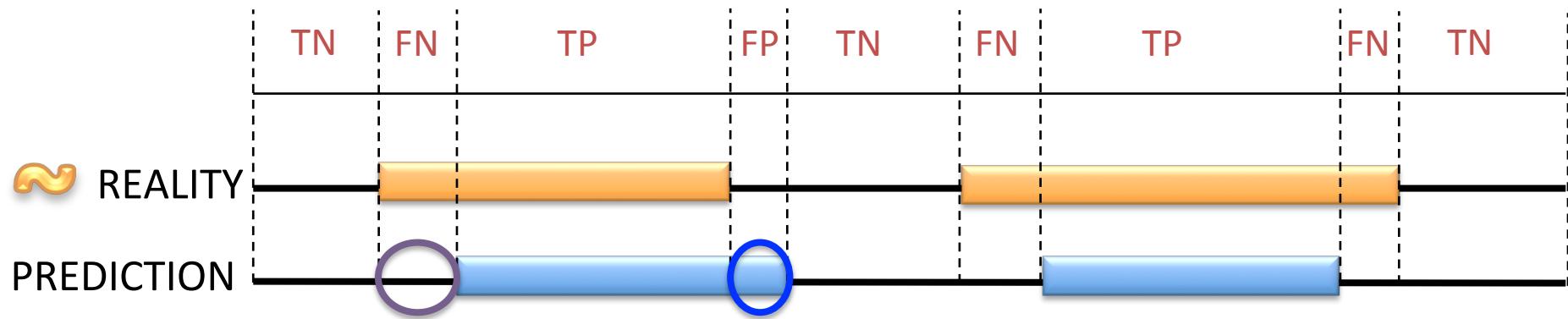
- Some gene-finders train themselves, others need a separate training procedure
- Around 500 already known genes are usually needed to train the gene-finder
  - => These "known" genes can be inferred from aligned transcripts or proteins
- The quality of the gene-finder results hugely relies on the quality of the training!

## A fungal genome





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)

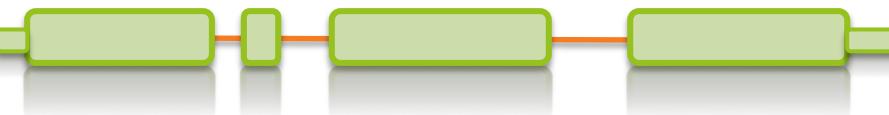
$$Sn = \frac{TP}{TP + FN}$$

**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   TP   FP = false pos.   FN = false neg.   sensitivity   specificity									
	total/ unique	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	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		



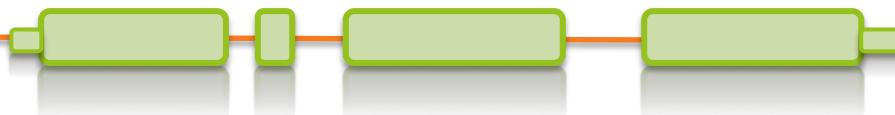
### Popular tools:

- **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](http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial)



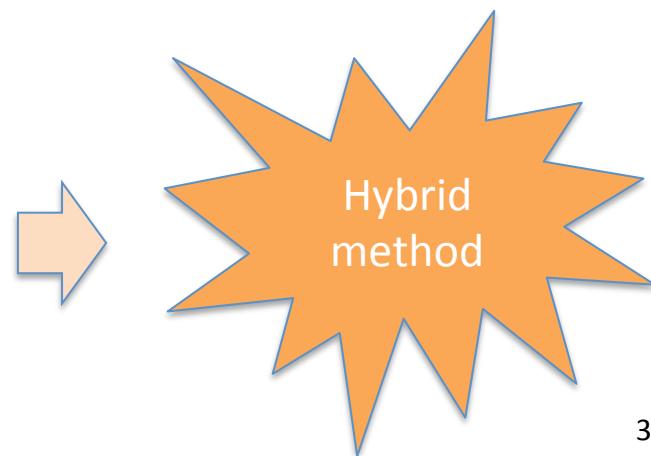
### Strengths :

- 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

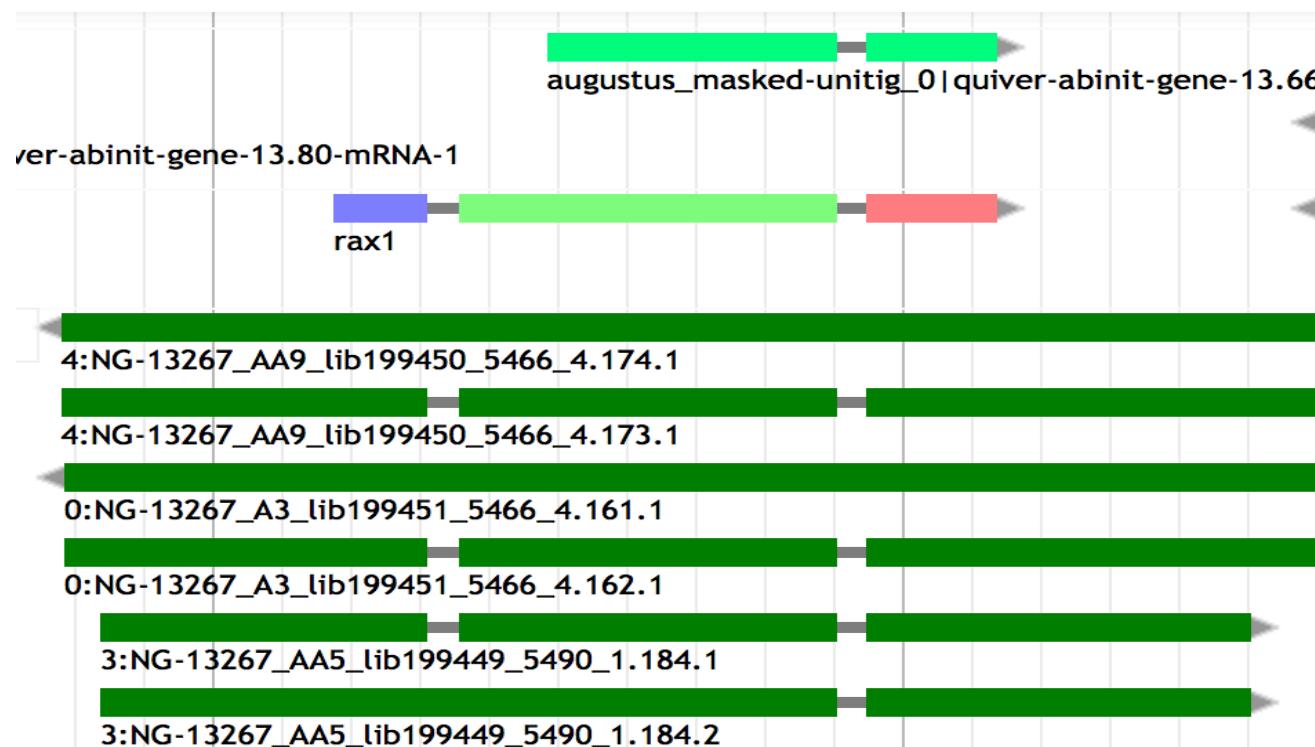




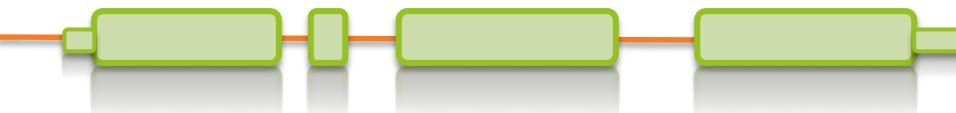
## 2. The different annotation approaches

### 2.2 Hybrid approaches

**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

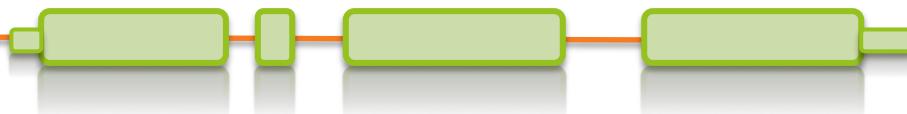
} Self training !

**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

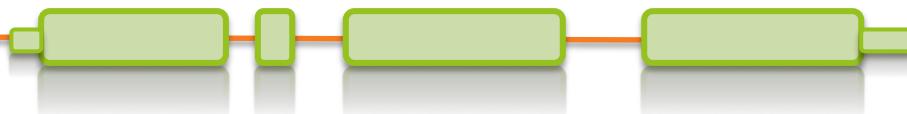
**EuGene\*** Any kind of evidence hints. Hard to configure (best suited for plant)



**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.



## 2. The different annotation approaches

### 2.3 Chooser / combiner

## Overview

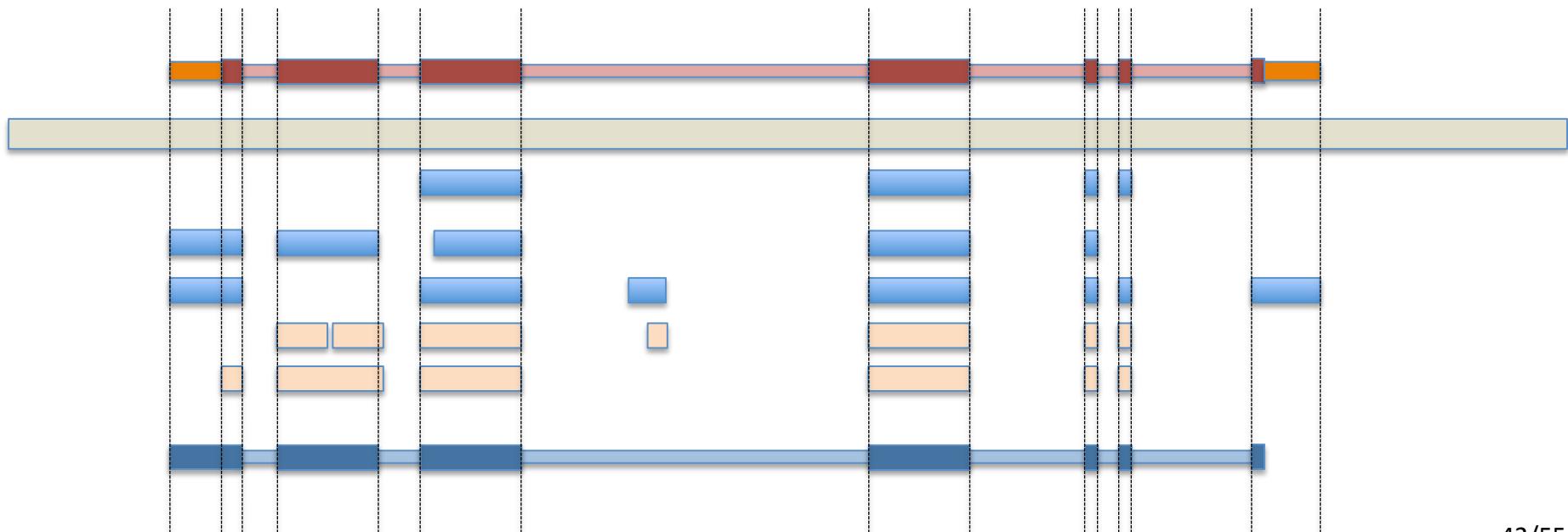
combining different lines of evidence into gene models

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 Evidencemodeles	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



## 2. The different annotation approaches

### 2.4 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* ...

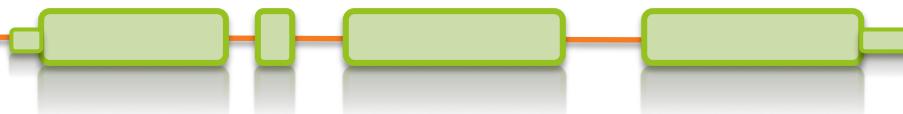
...



## 2. The different annotation approaches

### 2.5 Annotation of other genome features

## Other genome features



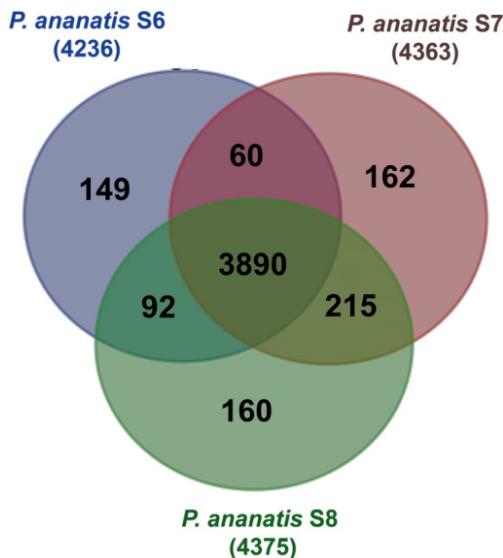
Feature type	DB associated	Tool example	approach
ncRNA	Rfam	infernal	HMM + CM
tRNA	Sprinl database	tRNAscan-SE	CM + WMA
snoRNA		snoscan	HMM + SCFG
miRNA	miRBase	Splign miR-PREFeR (for plant)	sequence alignment Based on expression patterns
Repeats	Repbase, Dfam	repeatMasker	HMM, blast
Pseudogenes		pseudopipe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



### 3. 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



4. *To resume / Closing remarks*

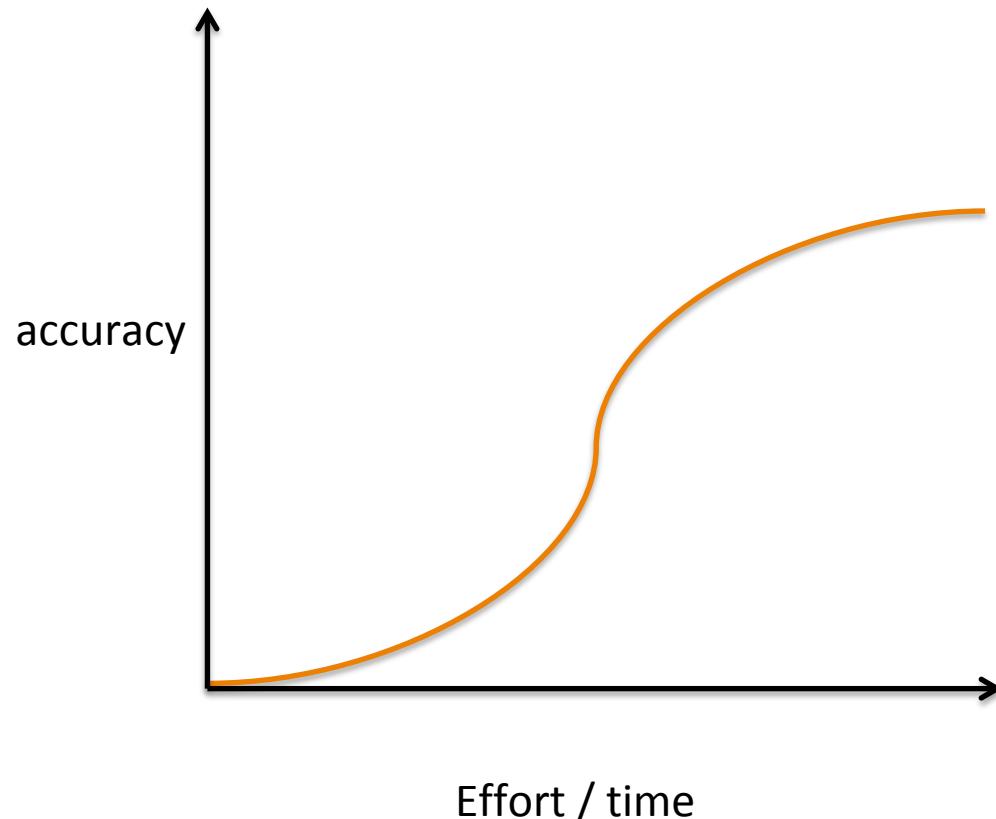


- >100 annotation tools – as many methods  
([https://github.com/NBISweden/GAAS/blob/master/annotation/CheatSheet/annotation\\_tools.md](https://github.com/NBISweden/GAAS/blob/master/annotation/CheatSheet/annotation_tools.md))
- 6 main class of approaches (Similarity-based, *ab initio*, hybrid, comparative,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



- 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 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



***THE END***

