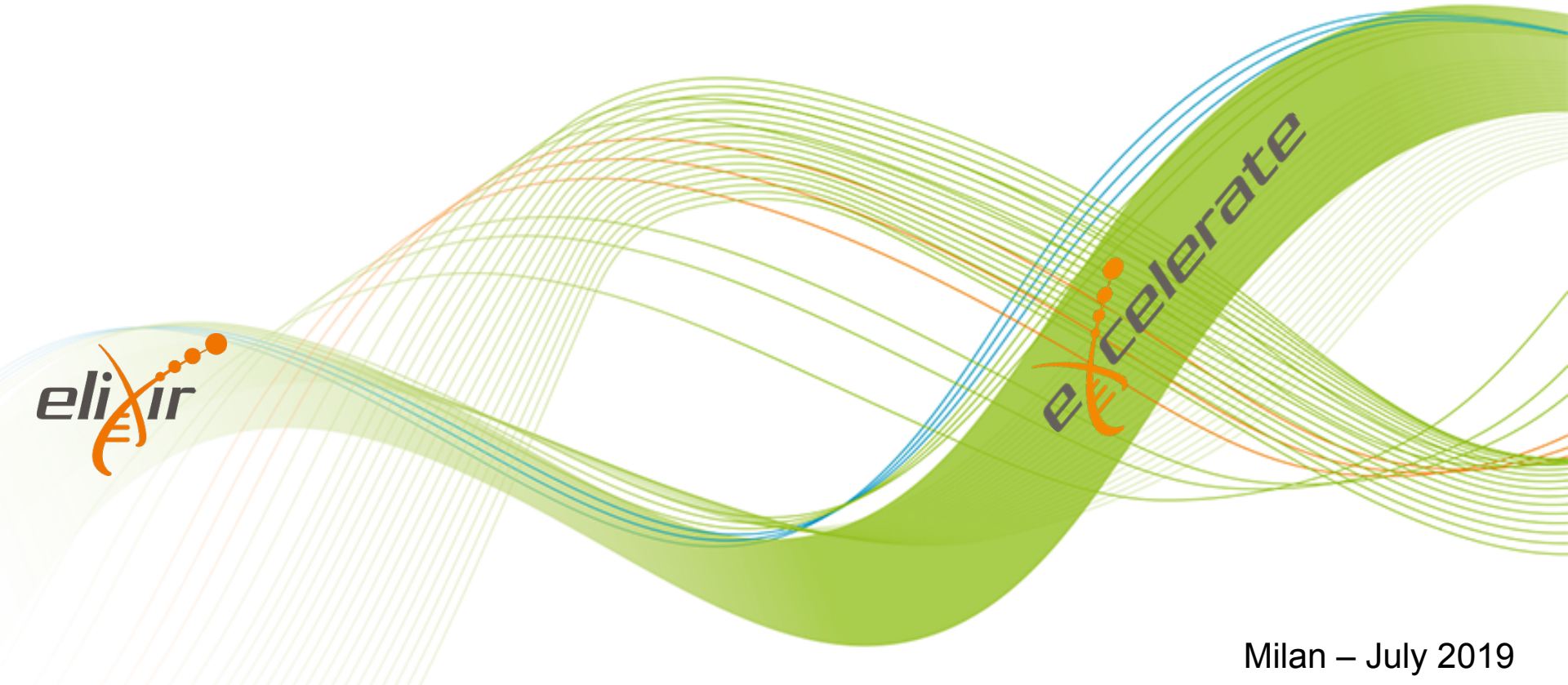
Jacques Dainat PhD

Methods in genome annotation



Milan – July 2019

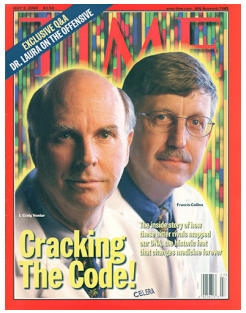
Contents

1. Introduction

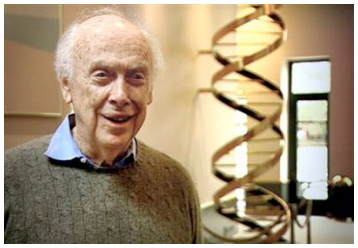
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

… prices go down

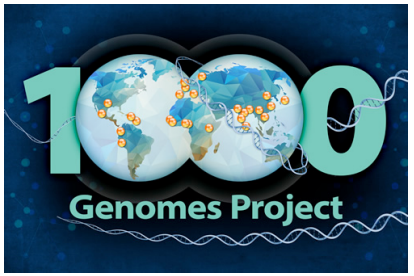
- Human genome sequencing:
  - 2004: Genome of Craig Wenter costs 70 mln $
    - Sanger's sequencing

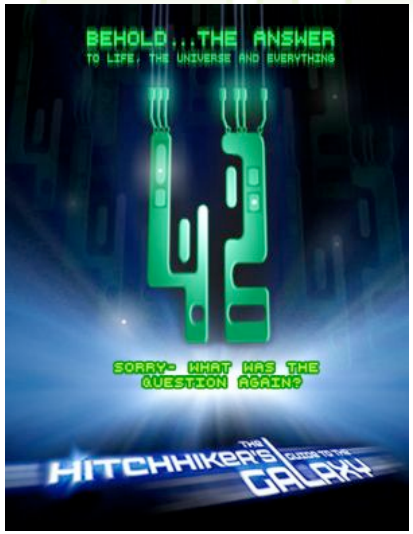


  - 2007: Genome of James Watson costs 2 mln $
    - 454 pyrosequencing



  - 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

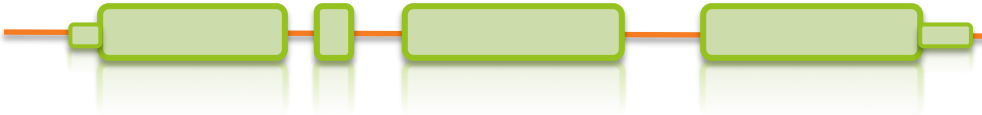
IF 2.9

doi:10.1016/j.jbiotec.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[a], Daniel Wibberg[a], Thomas Bekel[b], Jochen Blom[b], Burkhard Linke[b], Heiko Neuweger[b], Michael Stiens[a, c], Frank-Jörg Vorhölter[a], Stefan Weidner[a], Alexander Goesmann[b], Alfred Pühler[a] and Andreas Schlüter[a],

Introduction: Let's take a step back



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

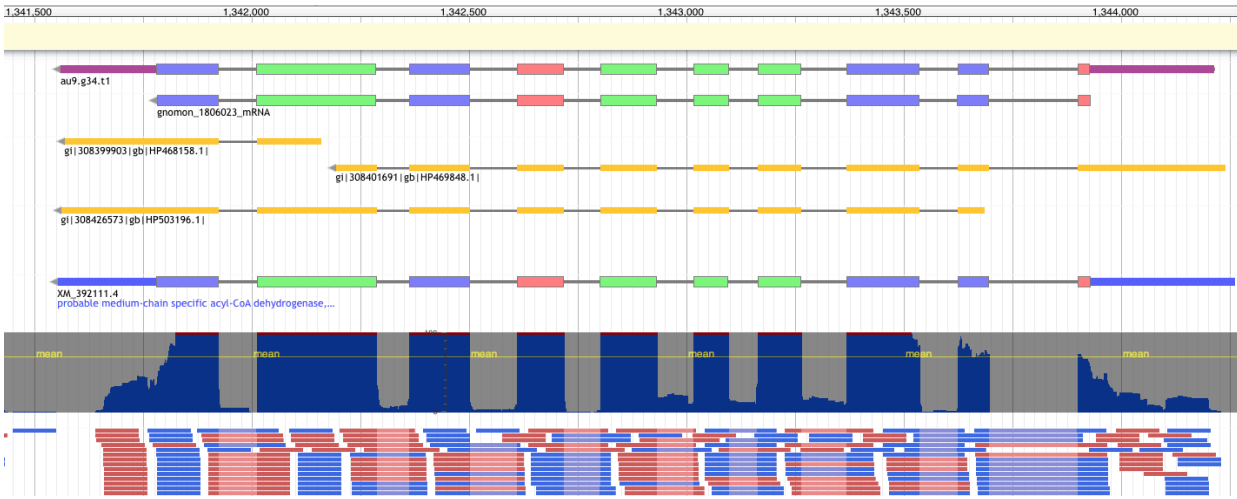
From a genome…

FASTA

```
>scaffold_26
AGTCACACACCCTTCAGCTTACACCCTGACTGCAGCCCTTACTCAAAACA
TTCCAGCCAGGAAGATGCTCCGACACAGCTTCTGGATGCCGCTCCTCGAC
GTCGAACGGCCCGCGCCGGGAAAATCGGCAGCGTCGGTGACCGCGGAGAT
CCGAAGCCGCCTCGGGGACCTGCGAGACAACGGGAGGCGGTCAACGAGAC
GCCGAGGGCTGGGAGTTATTCCCACACCGGGCCCGTAAGTTTTCTACCCA
AAAACCCATAGAAAAGAGATGAACCACTAAGTTTGATAACTCTTCTACTT
AACCGTGACCCTACGTGCCGGGGCAGGGCAGCTCTGACCCTAAGCGGCAC
ACGAACAAGGTGGTGCGCCCAATATAAACAAAGATGATGCAAGGGCTTGA
AATAAATCTCCGGAAGATTAATTCTCGAGCCCGACACGCTTTGAGGCAGC
GGAACCTACAGAACCACCGCAGTCACGTGAGAAGAGTCTAATACTCTCCA
AAGAGAAGTCCAAGGGAATGGAACGTGAAAAGAAGGTGCTTATCAAAAGC
GAGAAGGAAGATGGATGAGAACATCTTGTGTACTTCTTCTGGTCTCAAAA
AGCAAAAATGTAAAGATGCCAGACTAAGCCCGATCTGAGAAAGTACGCGA
GCAGAGACCCCCGCTGCCGATGTGGCCCAGAACGATGCCGATAAAGCACC
GAGACATAACAAAGCCCTGTGACACACAAGACGATGGACACAAACTACAT
AACACAGACACAAACTAAATGACACAGAGAGAAGTTGAAACTTCTGGGGA
AGTAAACATTTCTGAAACATCTACCAACAATCCGTCCATATATATTTCCA
TTCCACGGGACTCTTGGTTTGATATATGCGTGTTAACAGTAATCCCCGCT
GTAGCAATCACCACTATGCATAATTCATTAATTCTTTGGAGTTGCTGAGT
ATCATCTTATCAGTCTTATTTTTTTCCTTGGCTCTGGTTTCGGGCTTTTT
TTTTTTCTTCTGATAAGATTTTCCAGGAATGTGAAGACCCCCTGCATCCT
TCCCAAACTGACCACCCAAACTACAGACATTCTATAGCATTACATTACAC
AACCTAGGCAAAGTTTTTCTAACATTAAGGAACATGAAAAAAGCCAACAT
CACAATATATTCATAACAATTATGGAACATGCGAAAAGCCAATACCACAG
TACATTTATAACAATACCTCCCTTTTCCTTTCTTTAGAGATCATATGGCT
TGACCGCCGCCTCCTCGCCCGCCACCGCTGAGTACTGCCGTGCCGGAGTC
ACGGAGCCAGTCCCCCGCGGCCCCACCGCCTCCTCGCCCGCCGCCACGGA
GATCGGCTGCGCCACTCCCGAGCTCGGCCGTGCCATCGCCGCCCCCGCCG
GGGTCCCCCGGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

…to an annotated gene

GTF/GFF

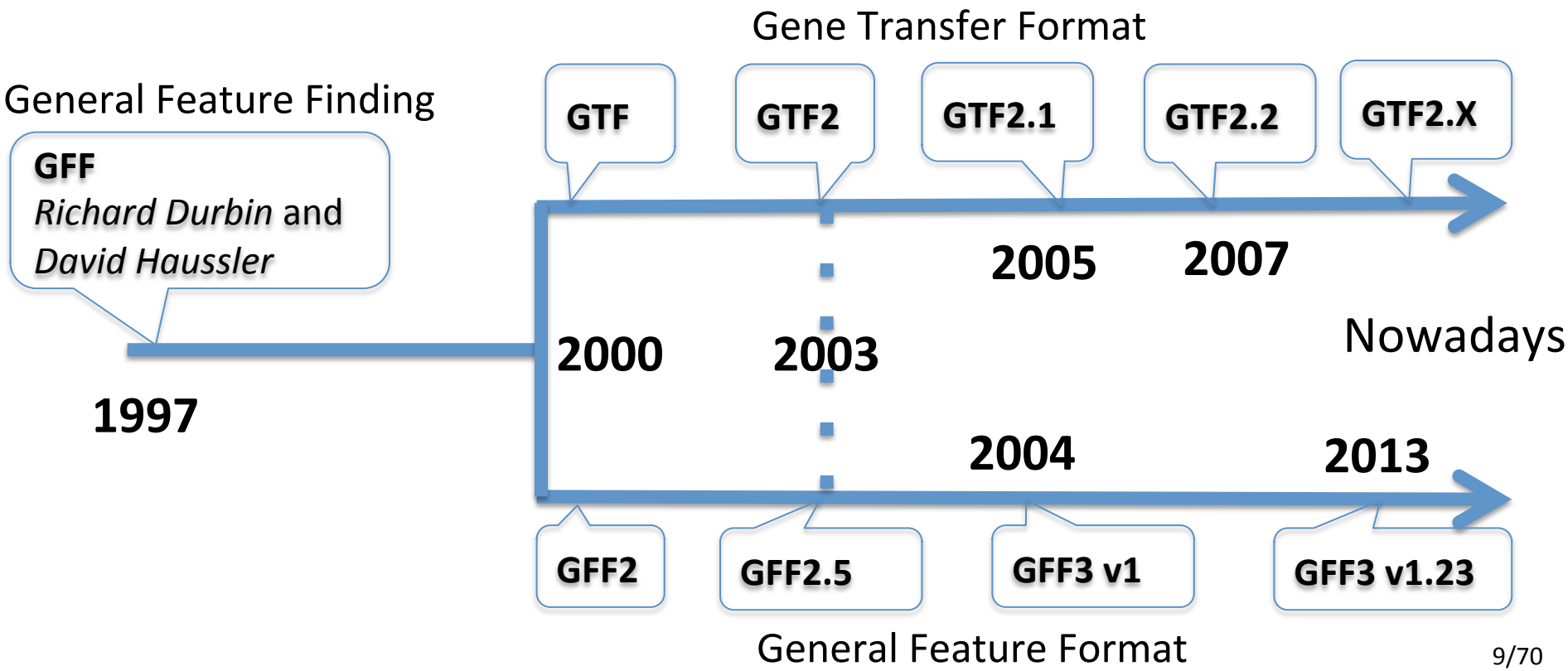


- 9 columns
- 1 feature = 1 line

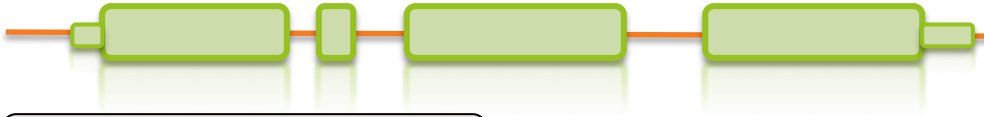
1. sequence id	2. source	3. feature type	4. start	5. end	6. score	7. strand	8. phase	9. attribute(s) tag=value
Ctg123	maker	Gene	1000	9000	.	+	.	ID=gene1; Name=EDEN

❑ GFF / GTF formats

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



Introduction: Formats: GTF2.X



- 9 columns
- 1 feature = 1 line

```
#!genome-build GRCz11
#!genome-date 2017-05
```

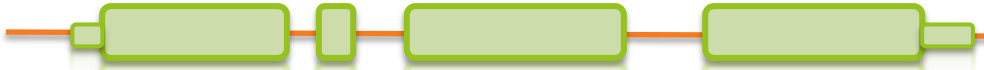
Header

Ctg123	.	Gene	1000	9000	.	+	.	gene_id gene1; name EDEN;
ctg123	.	Transcript	1050	9000	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	Transcript	1050	9000	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	exon	1300	1500	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	1050	1500	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
tg123	.	exon	1050	1500	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	exon	3000	3902	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	5000	5500	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	5000	5500	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	exon	7000	9000	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	7000	9000	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	CDS	1201	1500	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	CDS	3000	3902	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	CDS	5000	5500	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	CDS	7000	7600	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
Ctg123	.	CDS	1201	1500	.	+	0	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	CDS	5000	5500	.	+	0	gene_id gene1; transcript_id=t2; name EDEN;
Ctg123	.	CDS	7000	7600	.	+	0	gene_id gene1; transcript_id=t2; name EDEN;

1) sequence id
2) source
3) feature type (9 possibilities)
4) start
5) end
6) score
7) strand
8) phase
9) attributes ***tag value;***

! Features grouped by a **common attribute** (gene_id / transcript_id)

Introduction: Formats: GFF3



```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
```

← Header

- 9 columns
- 1 feature = 1 line

Ctg123	.	Gene	1000	9000	.	+	.	ID=gene1;Name=EDEN
ctg123	.	mRNA	1050	9000	.	+	.	ID=mRNA1;Parent=gene1;Name=EDEN.1
ctg123	.	mRNA	1050	9000	.	+	.	ID=mRNA2;Parent=gene1;Name=EDEN.2
ctg123	.	exon	1300	1500	.	+	.	ID=exon1;Parent=mRNA3
ctg123	.	exon	1050	1500	.	+	.	ID=exon2;Parent=mRNA1,mRNA2
ctg123	.	exon	3000	3902	.	+	.	ID=exon3;Parent=mRNA1
ctg123	.	exon	5000	5500	.	+	.	ID=exon4;Parent=mRNA1,mRNA2
ctg123	.	exon	7000	9000	.	+	.	ID=exon5;Parent=mRNA1,mRNA2
ctg123	.	CDS	1201	1500	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
ctg123	.	CDS	3000	3902	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
ctg123	.	CDS	5000	5500	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
ctg123	.	CDS	7000	7600	.	+	0	ID=cds1;Parent=mRNA1;Name=eden1
Ctg123	.	CDS	1201	1500	.	+	0	ID=cds2;Parent=mRNA2;Name=eden2
ctg123	.	CDS	5000	5500	.	+	0	ID=cds2;Parent=mRNA2;Name=eden2
Ctg123	.	CDS	7000	7600	.	+	0	ID=cds2;Parent=mRNA2;Name=eden2

1) sequence id

2) source

3) feature type
(SO term = 2278 possibilities)

4) start

5) end

6) score

7) strand

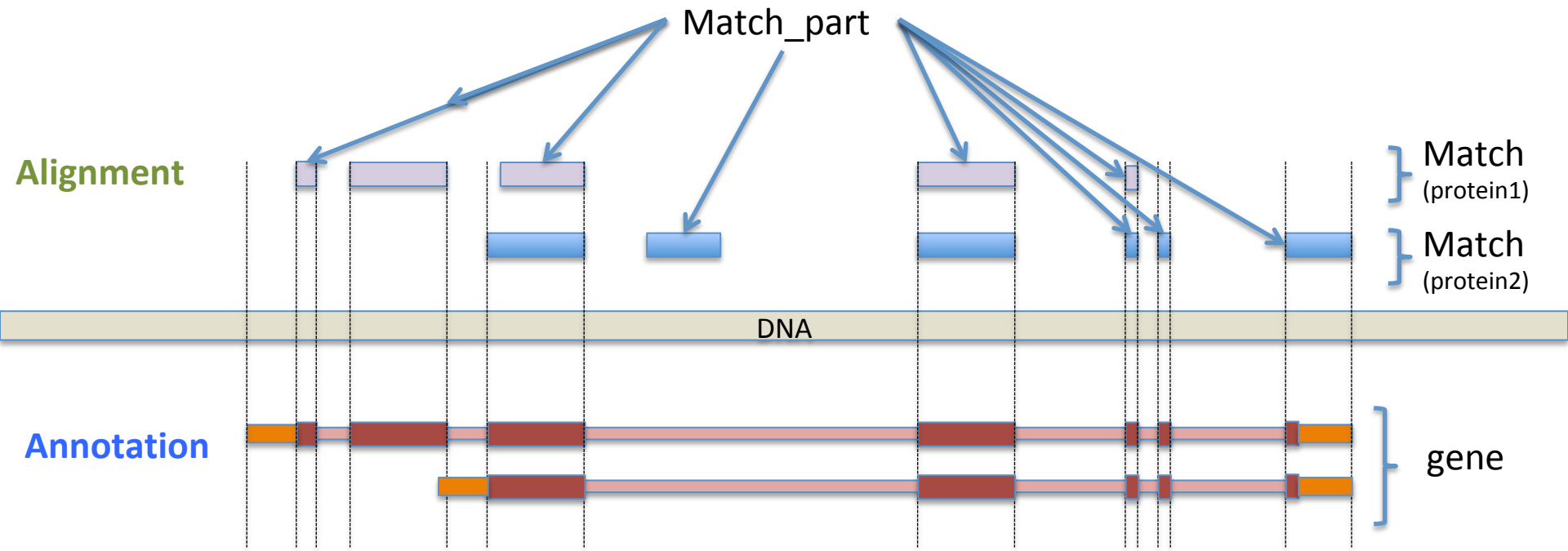
8) phase

9) attributes
tag=value

! Features are grouped by **parent** relationship

Introduction: Formats: GFF3

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



Intron, exon, CDS,splice site, UTR, mRNA, isoforms

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

The main steps in genome annotation

1 2 3 4 5

QC assembly → Structural annotation → Manual curation → Functional annotation → Downstream analysis / Submission

BUSCO

MAKER Annotate this!

EuGene-EP

Web Apollo

InterPro

GFF3 + FASTA → EMBLmyGFF3 → EMBL format

Introduction: Before annotation

- Get the best assembly! 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

- Freeze the assembly!

  => Updating assembly ~ annotation from scratch

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Always check :

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

Introduction: BUSCO

BUSCO used on assembly and annotation

Example of 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 annotation 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
```

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**

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

First of all: 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 gene coding annotation

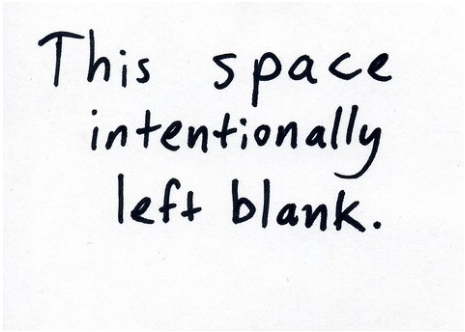
Introduction: Predictive methods

Intrinsic / *Ab-initio*

Extrinsic

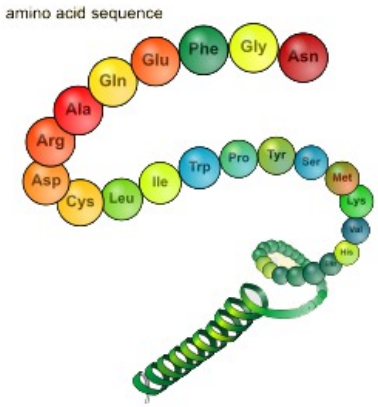
Ø

- Use of information/ features from the sequence itself



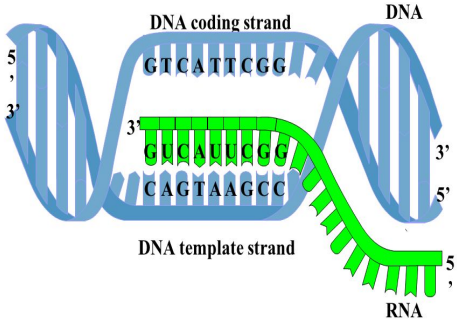
Proteins

- Known amino acid sequences from other organisms



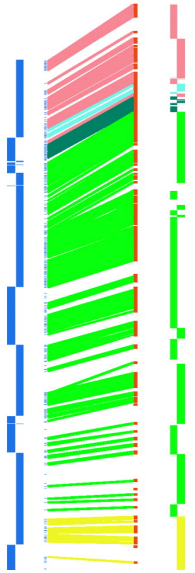
Transcripts

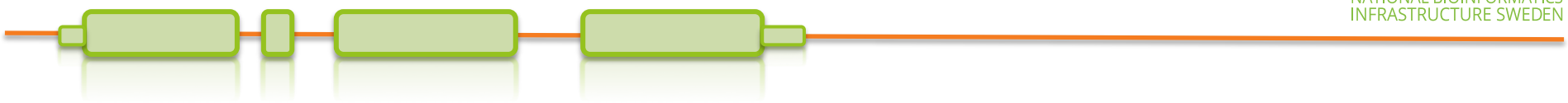
- Assembled from RNA-seq or downloaded ESTs



Genomes

- Close relative genomes

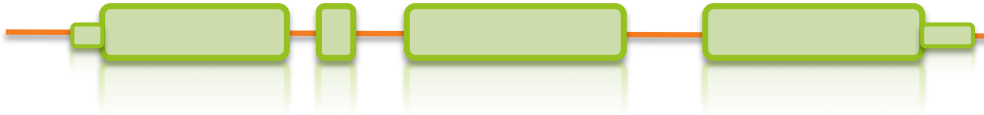




2.1 Intrinsic / *ab initio*

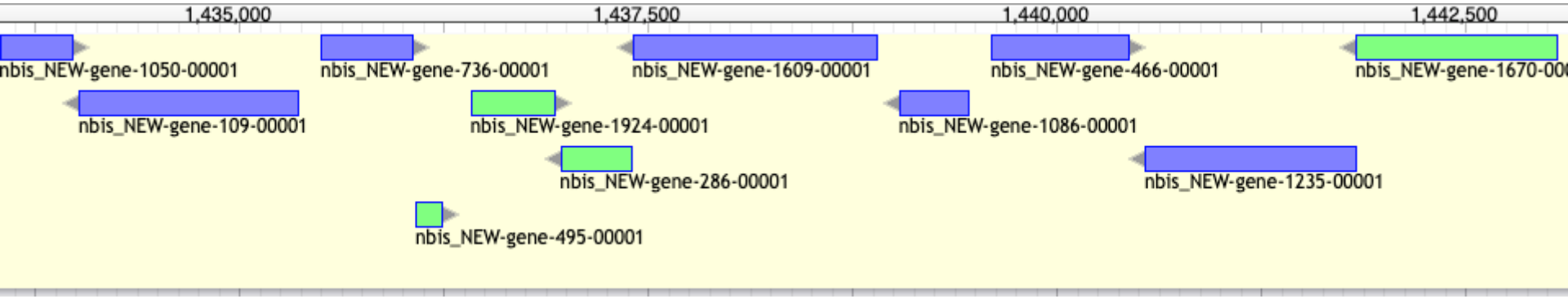
Using a probabilistic models to predict features

Intrinsic / *ab initio*



- Bacterial gene prediction

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



⇒ easy problem: Accuracy > 90%

Intrinsic / *ab initio*

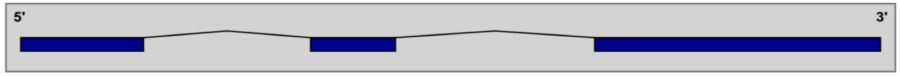
- Eukaryotic gene prediction

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

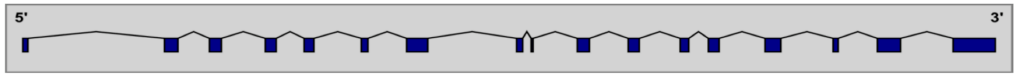
Ostreococcus

Populus

Spruce



Ectocarpus



⇒ Hard problem: Accuracy < 70%

⇒ **Requires training**

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

How it works?

Methods are 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
- …

signal detection

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

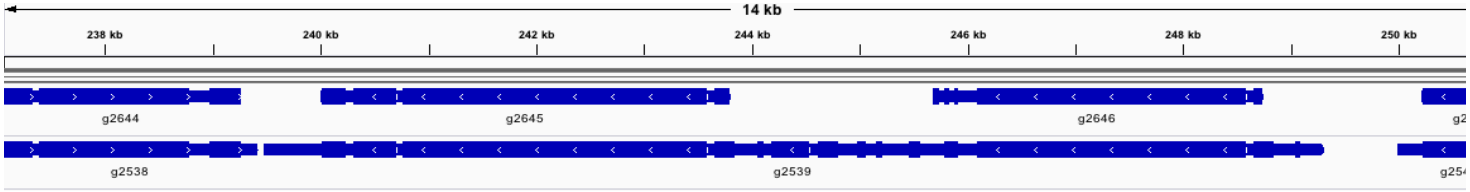
=> *Ab initio* tools will combine this information through different Probabilistic models: HMM, GHMM, WAM, etc.

Training *ab-initio* gene-finders

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

A fungal genome

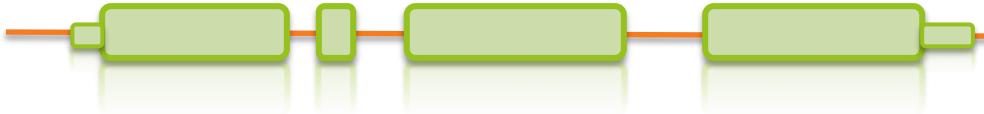
Fungi
Plants



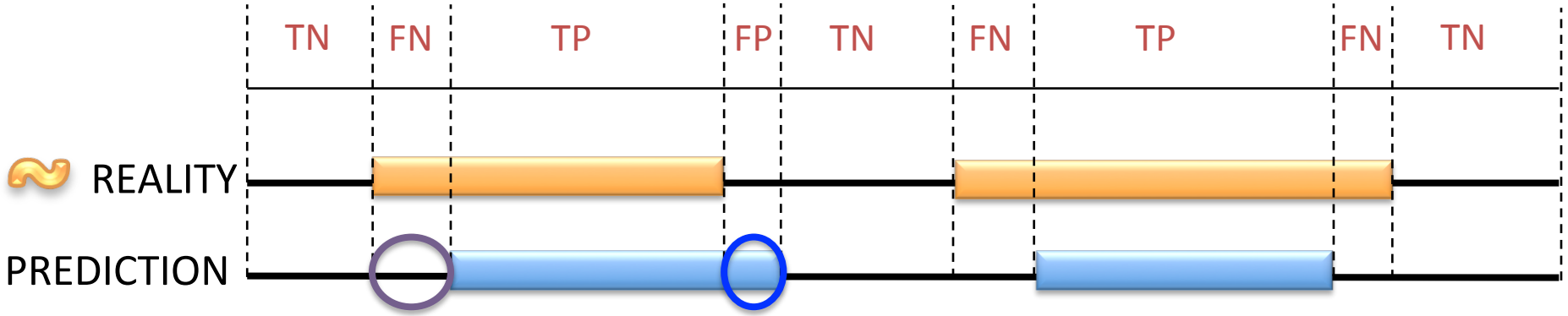
<u>Training:</u>

- sets of high quality genes (>500)

  => These ”known” genes are usually inferred from aligned transcripts or proteins



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)

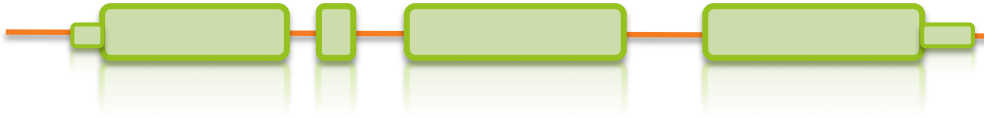
$$Sn = \frac{TP}{TP + FN}$$

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.

Intrinsic / *ab initio*



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

```
*******     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 |
-------------------------------------------------------------------------/
```

Intrinsic / *ab initio*

Popular tools:

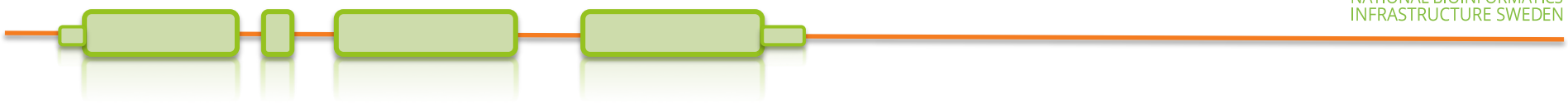
- **SNAP** Works ok, easy to train, not as good as others especially on longer intron genomes.
- **Augustus** Works great, hard to train (but getting better).
- **GeneMark-ES** Self training, no hints, buggy, not good for fragmented genomes or long introns (Best suited for Fungi).
- **FGENESH** Works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

Supported by MAKER

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

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

Intrinsic / *ab initio*



Strengths :

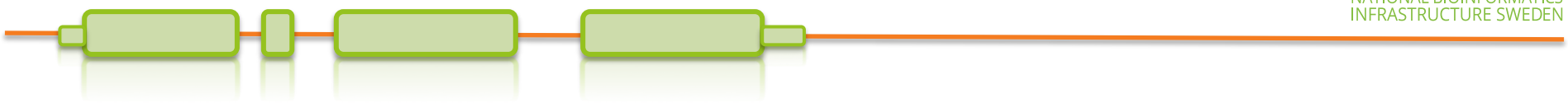
- Fast and easy
- Annotate unknown genes
- Sensitivity ok
- Need no external evidence**

Limits :

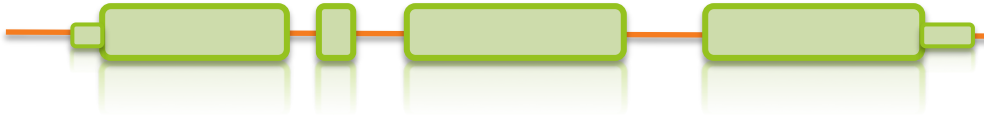
- No UTR*
- No alternatively spliced transcripts*
- Bad specificity (Over prediction of exons or/and genes)
- **Training** needed (Need external evidence**)

Common errors in annotation:

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



Exercise



2.2. Extrinsic approaches

Similarity-based methods

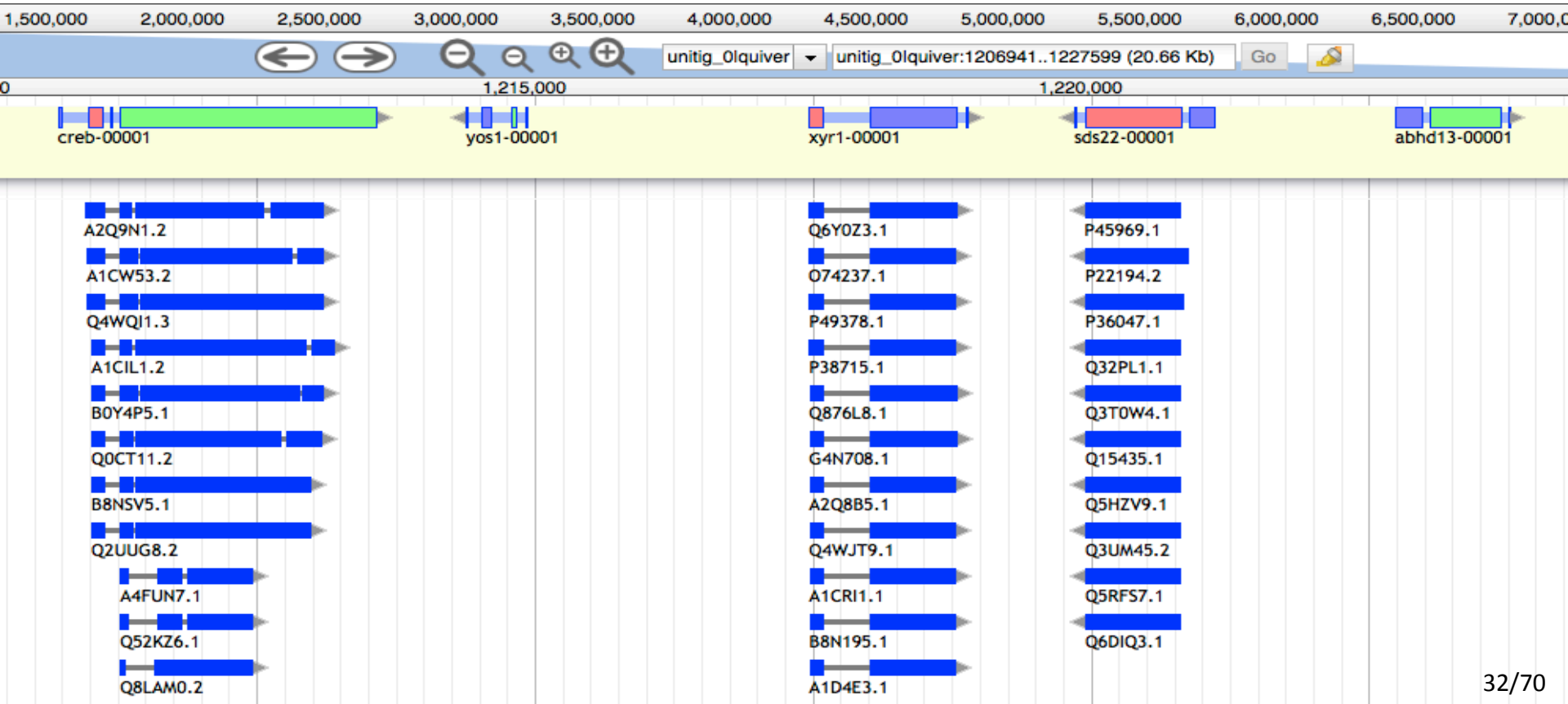
These use similarity to annotated sequences like

- Proteins
- Transcripts
- ESTs

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

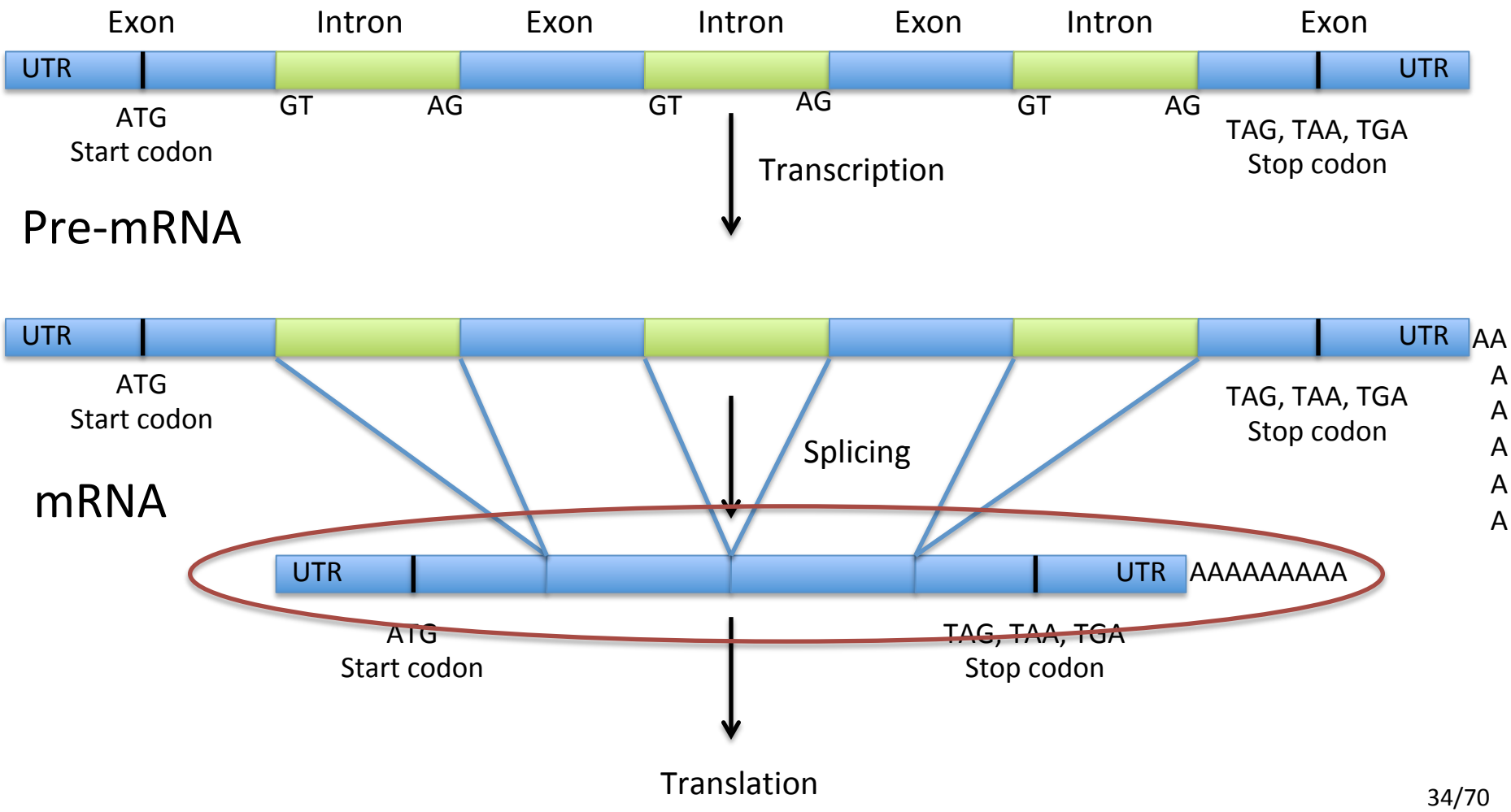
Limits :

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

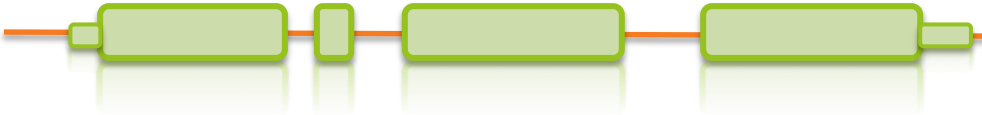
>sp|Q9NSK7|CS012_HUMAN Protein C19orf12 OS=Homo sapiens OX=9606 GN=C19orf12 **PE=1** SV=3
MERLKSHKPATMTIMVEDIMKLLCSLSGERKMKAAVKHSGKGALVTGAMAFVGGLVGGPP
GLAVGGAVGGLLGAWMTSGQFKPVPQILMELPPAEQQRLFNEAAAIIRHLEWTDAVQLTA
LVMGSEALQQQLLAMLVNYVTKELRAEIQYDD

>sp|Q2V2T9|SCL16_ARATH Putative scarecrow-like protein 16 OS=Arabidopsis thaliana OX=3702 GN=SCL16 **PE=5** SV=1
MQIPTLIDSMANKLHKKPPPLLKLTVIASDAEFHPPPLLGISYEELGSKLVNFATTRNVA
MEFRIISSSYSDGLSSLIEQLRIDPFVFNEALVVNCHMMLHYIPDEILTSNLRSVFLKEL
RDLNPTIVTLIDEDSDFTSTNFISRLRSLYNYMWIPYDTAEMFLTRGSEQRQWYEADISW
KIDNVVAKEGAERVERLEPKSR

Types of data used: RNA-seq



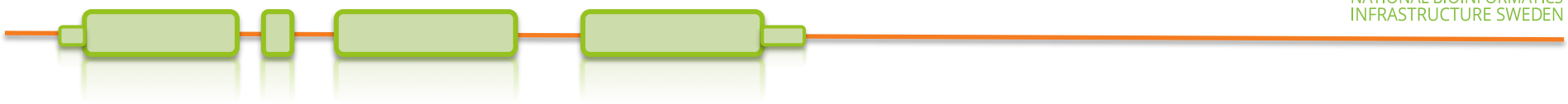
Similarity-based method: RNA-seq data



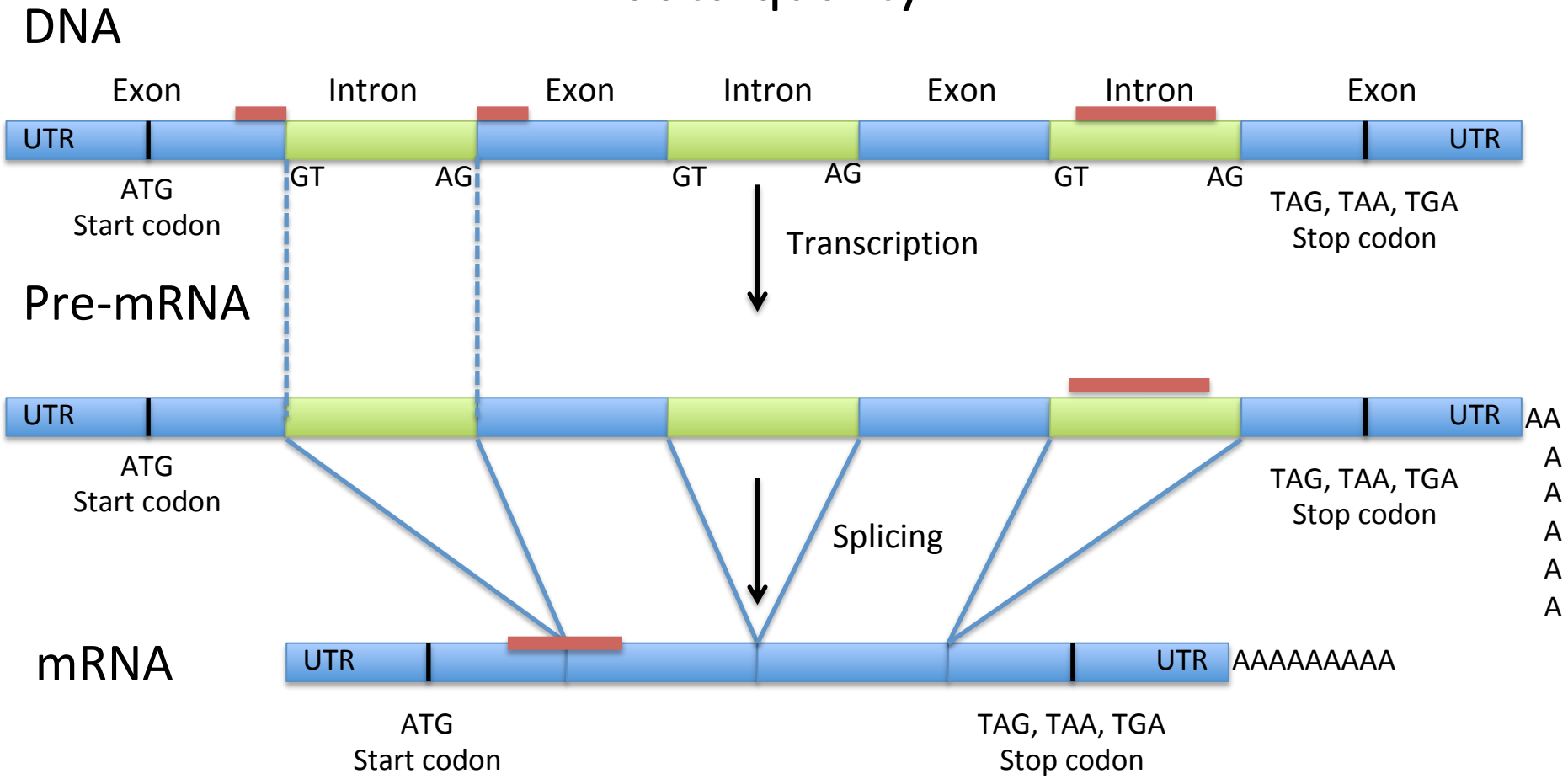
- Should always be included in an annotation project
- From the same organism as the genomic data => unbiased
- Sample different tissues or life stages if possible
- /!\ Can be very noisy (tissue/species dependent), can include pre-mRNA

  => Avoid gonads; muscle or liver is good

Similarity-based method: RNA-seq data

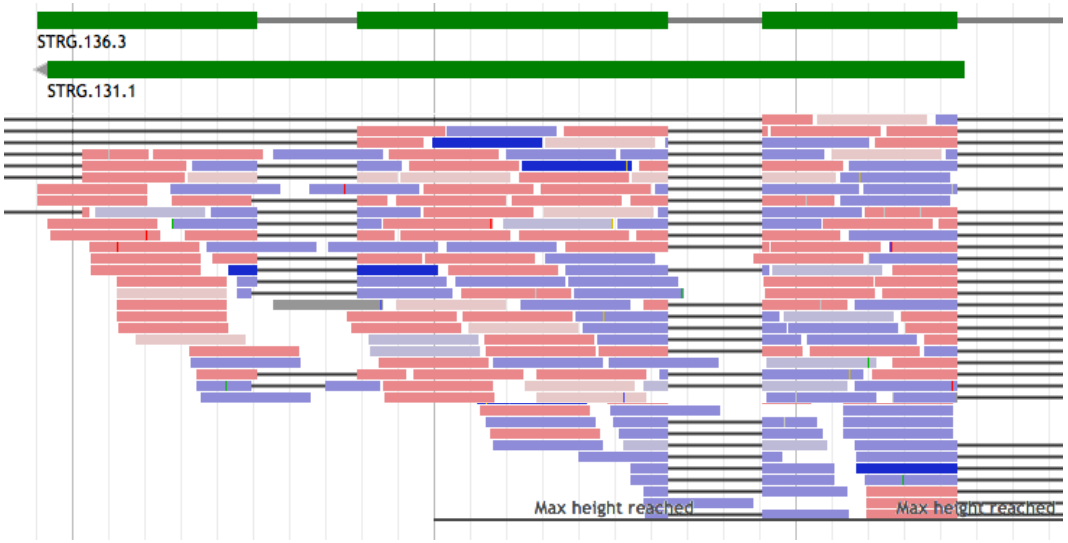


data quality



RNA-seq - Spliced reads

RNA-seq − pre-mRNA noise



Limits :

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

- short-reads:
  - need to be assembled first
    - Genome guided assembly
      ⇒ Stringtie: mapped reads -> transcripts
    - *De novo*
      => Trinity: assembles transcripts without a genome
  - Transcriptome assembly errors

- Long-reads:
  - error rate / frameshift / indels

Similarity-based method: Protein / transcripts

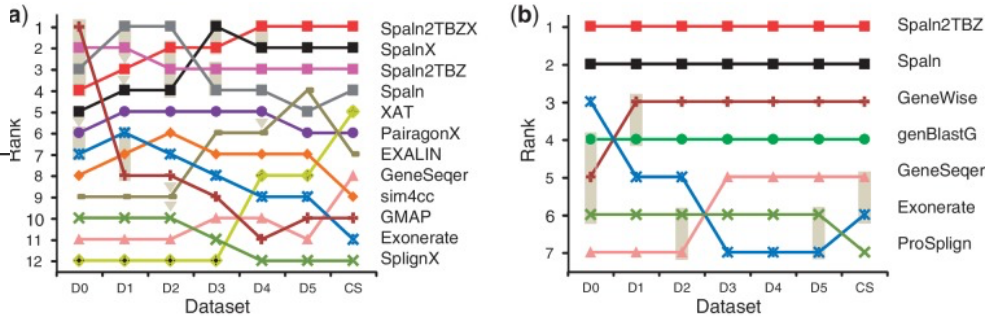
- Rough approximation (fast)

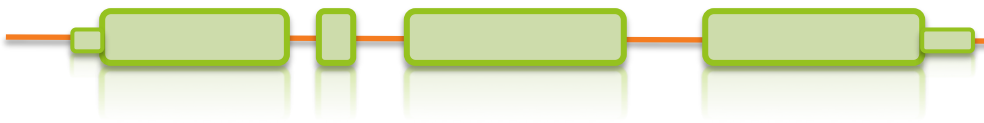
DNA	AA
Blastn	Pmatch
Vsearch	tblastn
NSimScan	PSimScan

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

	Method	Human					
		Mouse			Chicken		
CDS	EXALIN	3h	5min	41.3s	2h	1min	41.1s
	Exonerate		9min	30.1s		3min	31.7s
	GeneSeqer	7h	14min	48.2s	3h	2min	49.4s
	GMAP		1min	43.5s		1min	37.9s
	PairagonX	274h	1min	16.0s	500h	57min	16.8s
	sim4cc			33.9s			19.3s
	Spaln2TBZX		6min	55.2s		9min	44.3s
	SplignX		14min	2.0s		6min	24.6s
	XAT		2min	2.8s		1min	17.9s
protein	Exonerate	12h	36min	10.3s	7h	33min	17.0s
	genBlastG		3min	30.3s		2min	16.6s
	GeneSeqer	10h	10min	24.0s	6h	20min	40.0s
	GeneWise	69h	17min	36.1s	47h	36min	6.6s
	ProSplign	2h	18min	24.9s	1h	17min	39.1s
	Spaln2TBZ		4min	32.0s		4min	41.8s

DNA	AA
Exonerate	Exonerate
Gmap	Genewise
GenomeThreader	GenomeThreader

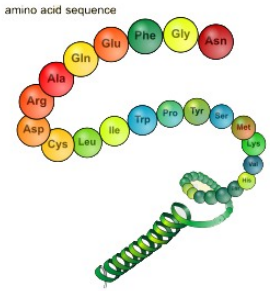
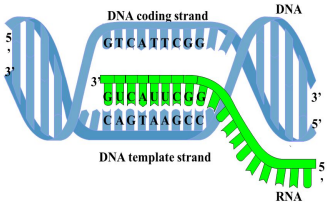
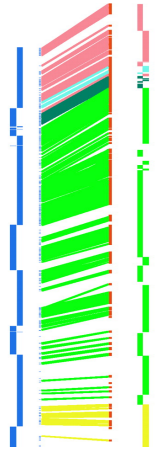


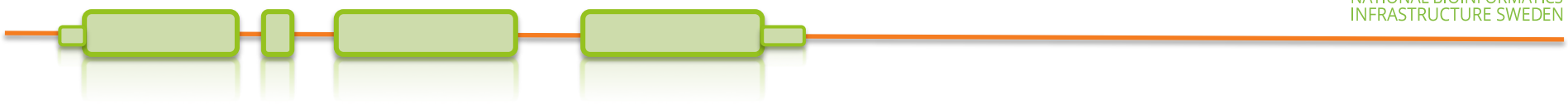


NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Exercise

Types data used vs methods

Annotation approach	∅	Proteins	Transcripts	Genomes
	This space intentionally left blank.			
Pure *ab initio*	X			
Similarity		X	X	
Ab initio evidence-driven (hybrid)	X	X	X	X

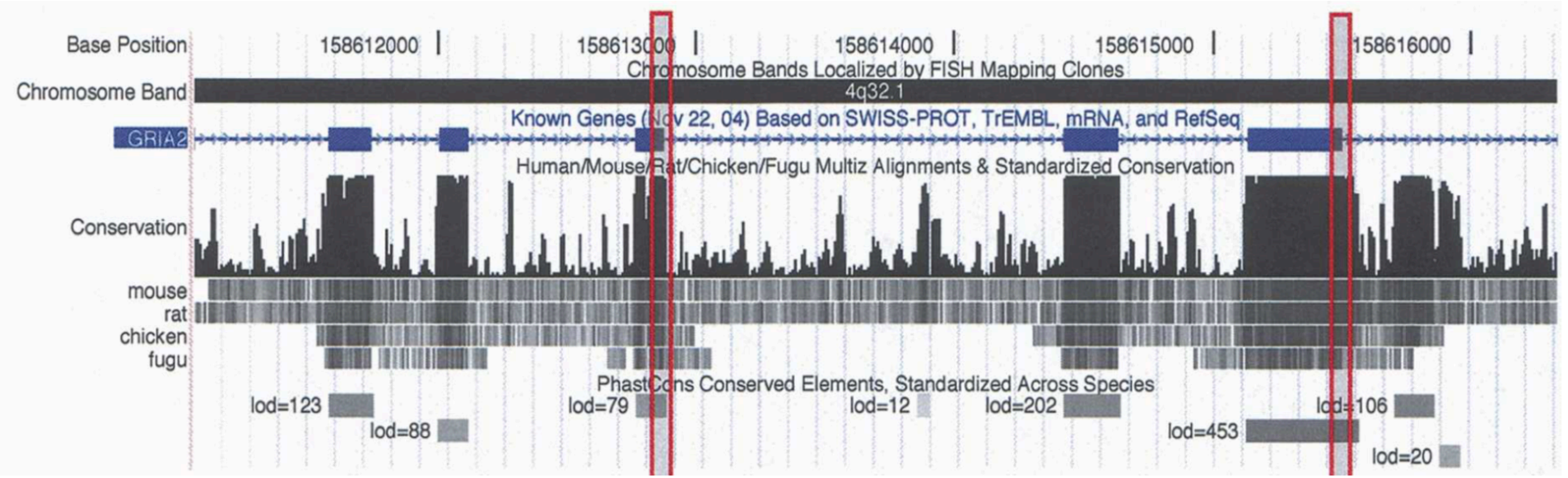


Evidence-drivable gene predictors approaches incorporate hints/extrinsic information in the form of alignments (transcript, protein, whole genome) to increase the accuracy of the gene prediction.

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

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

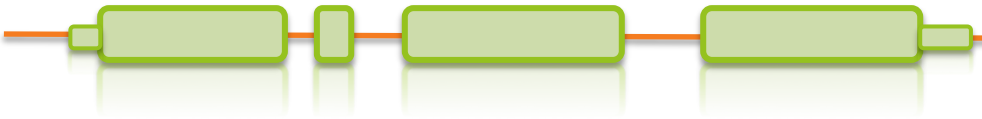
- Need relatively close related genome (<50 My)

Comparative-based method: Tools

- **Dual genome**, de novo gene structure prediction:
  - **Rosetta** (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
  - **Augustus-CPG***

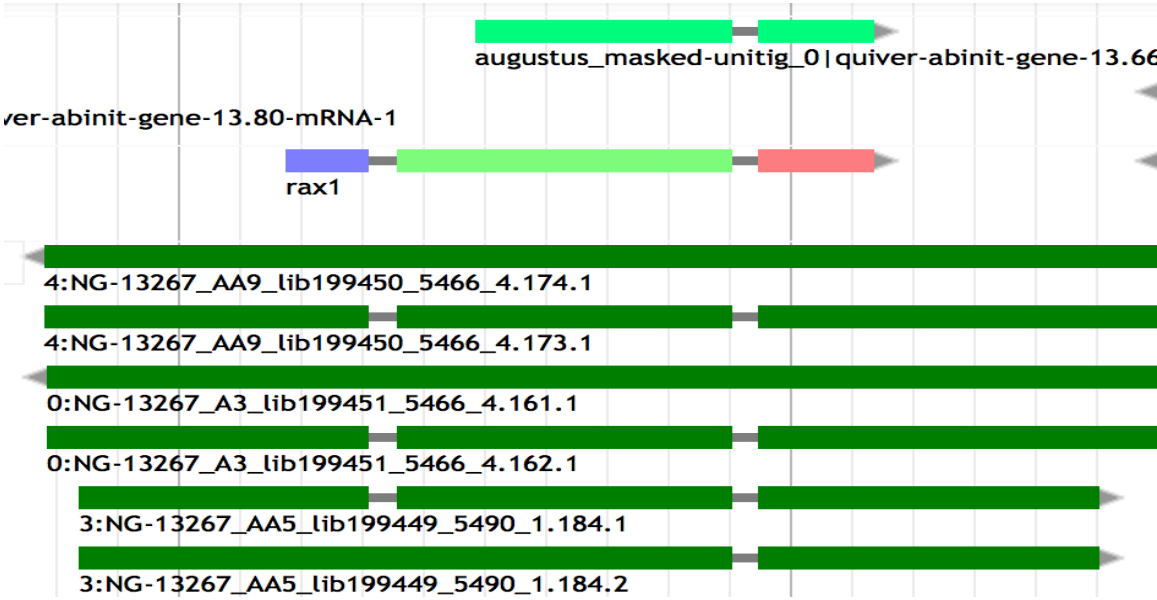
*Can use similarity-based information

Ab initio evidence-based



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 } 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)

The BRAKER1 gene finding pipeline:

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

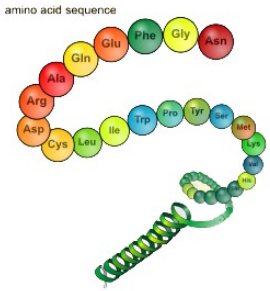
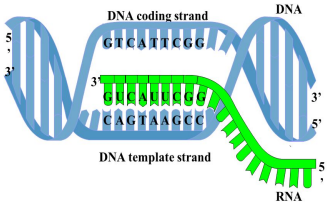
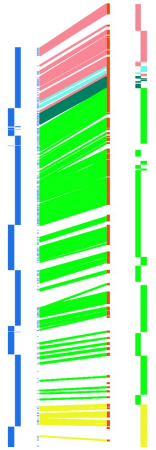
Katharina J. Hoff et *al.*

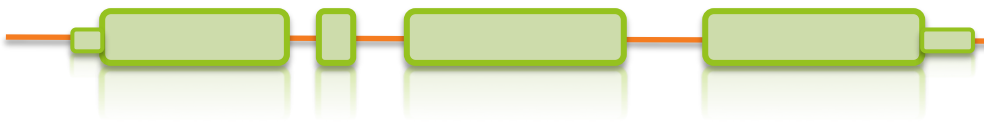
Bioinformatics (2016) 32 (5): 767-769. doi: 10.1093/bioinformatics/btv661

- 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)

Types data used vs methods

Annotation approach	∅	Proteins	Transcripts	Genomes
	This space intentionally left blank.			
Pure *ab initio*	X			
Similarity		X	X	
Ab initio evidence-driven (hybrid)	X	X	X	X
Chooser/combiner	X	X	X	X
Pipeline	X	X	X	X



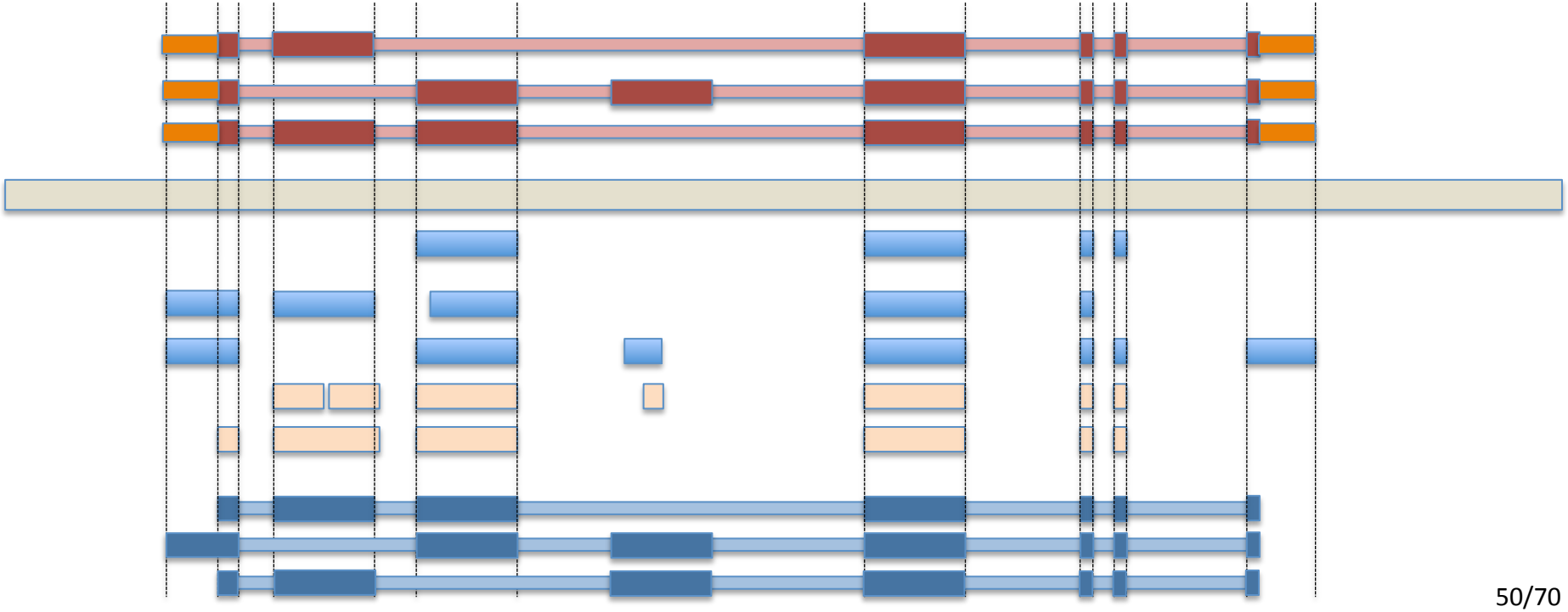
2.4. Combiner / Chooser

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

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)

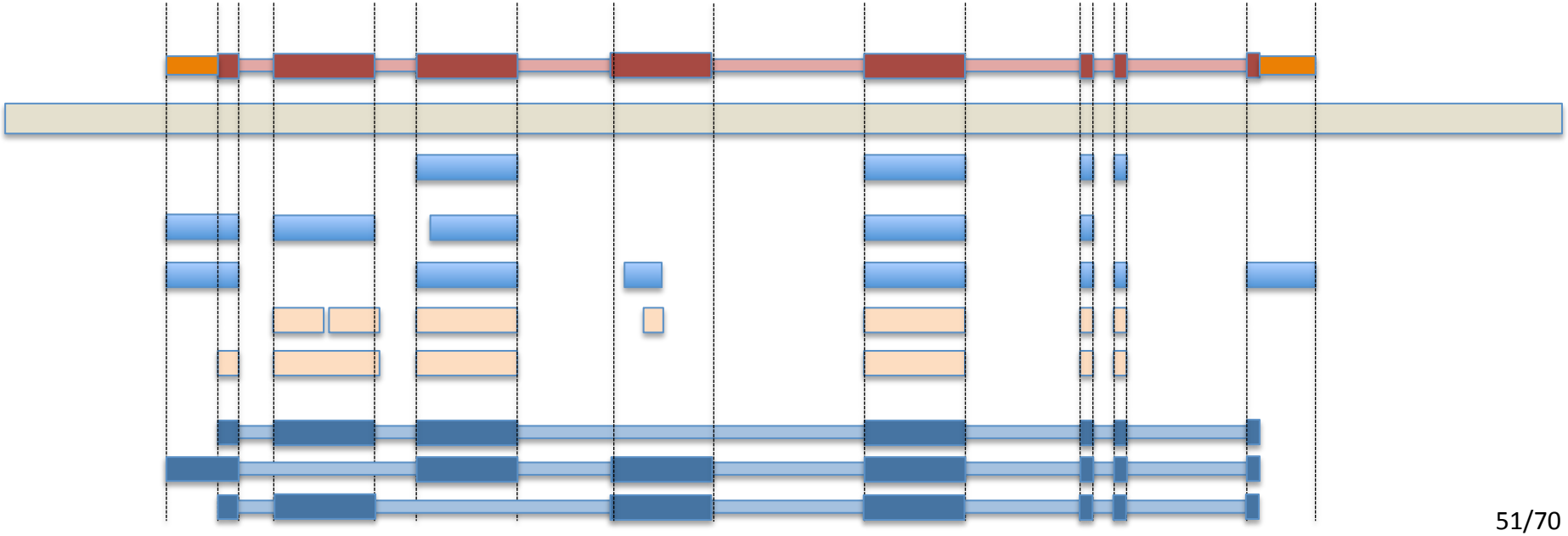


Combiner concept: combining different lines of evidence into gene models

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



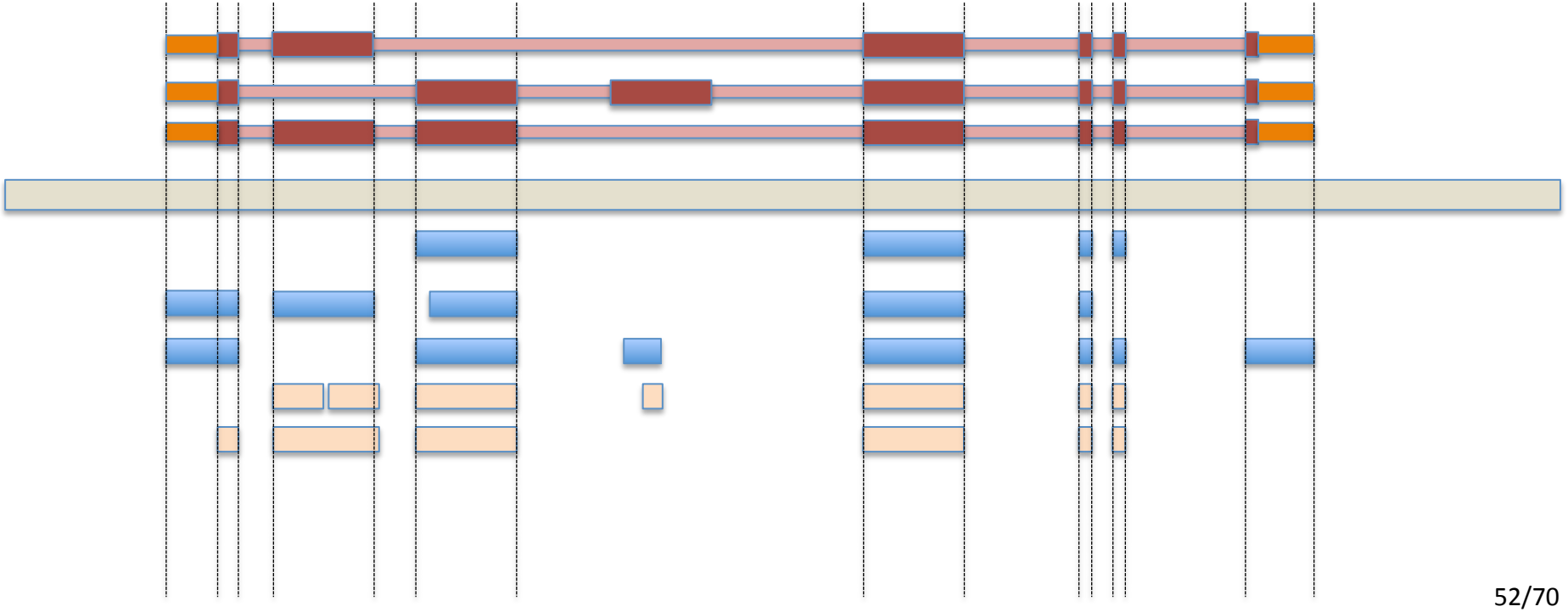
Chooser / combiner

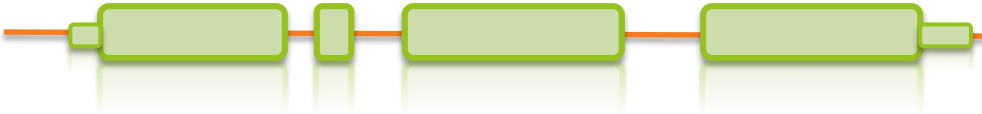
Chooser concept: Select best gene models

Evidence: ESTs / Transcripts / Proteins

Gene models

=> Choose the prediction whose best matches the evidence

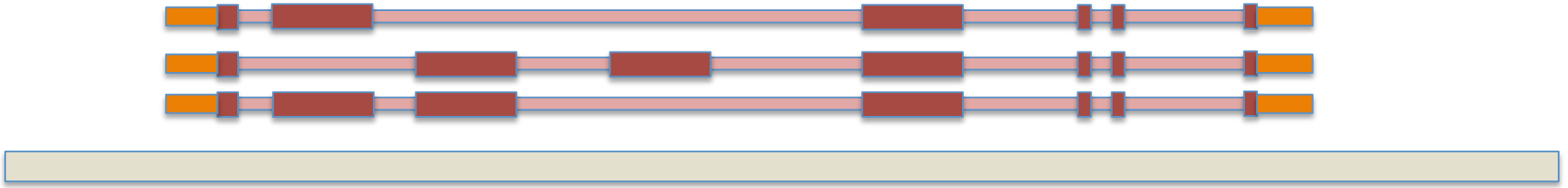




Chooser concept: Select best gene models

Gene models

=> Choose the prediction whose structure best represents the consensus

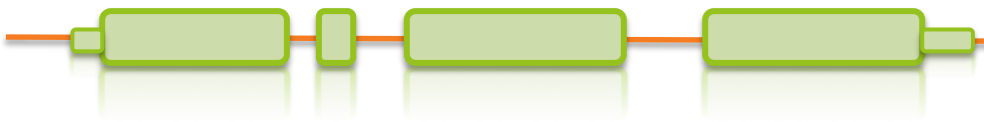


Chooser / combiner: resume

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 best possible set of exons and combine them in a new consensus gene model				
JIGSAW	X			
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

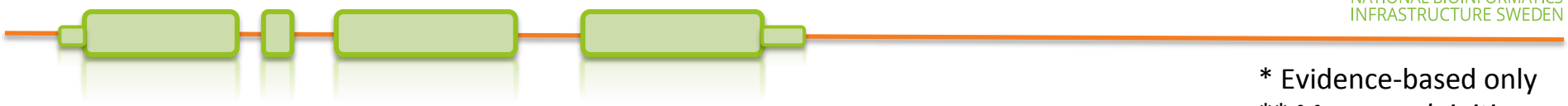


Pipelines

(The ultimate step)

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

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

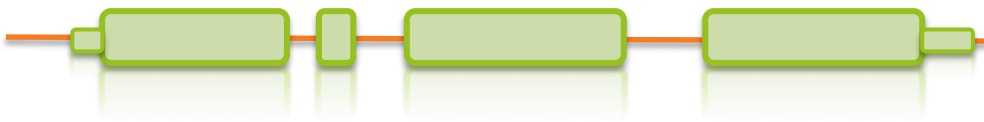
NCBI pipeline Evidence + *ab initio* (Gnomon), repeat masking, gene naming, miRNAs, tRNAs, …

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

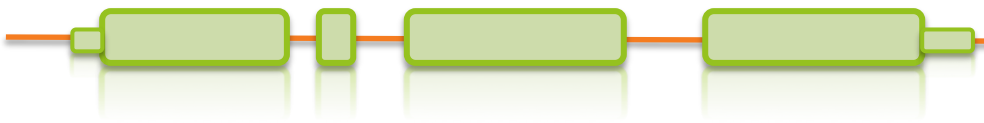
Comparative Annotation Toolkit (CAT) ab-initio (Augustus) evidence driven + comparative

MAKER2 Evidence based and/or *ab initio …*

…



MAKER lecture

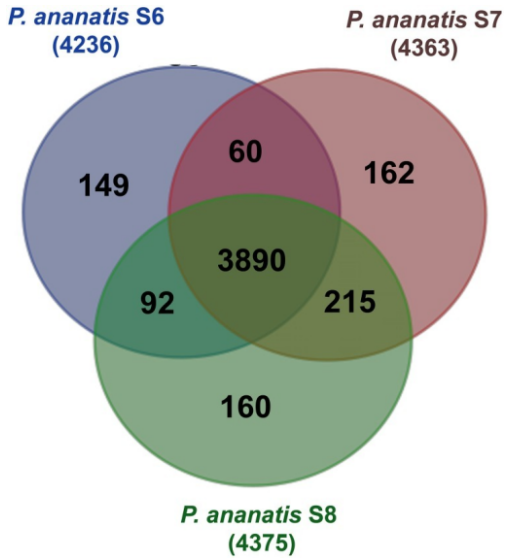


NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Exercise

3. Annotation assessment

- 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



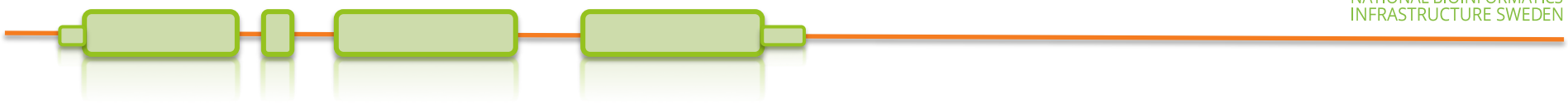
Annotation assessment: Visualization

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



Exercice

4. *To resume / Closing remarks*

To resume: 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 :

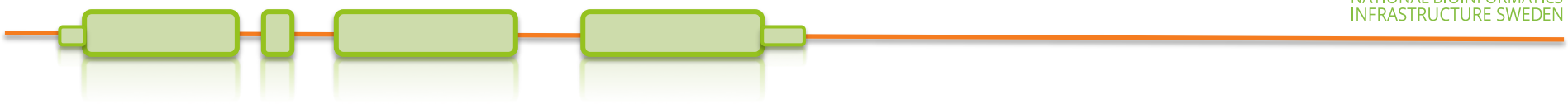
  These combine multiple approaches

To resume: Other genome features

Coding genes are nice, but what about the other 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)
…			

To resume: Tools and formats



❑ Annotation tools

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

90's — Nowadays

❑ Formats

GTFs / GFFs

Structural annotation + functional annotation + …

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

To resume: Tools and methods

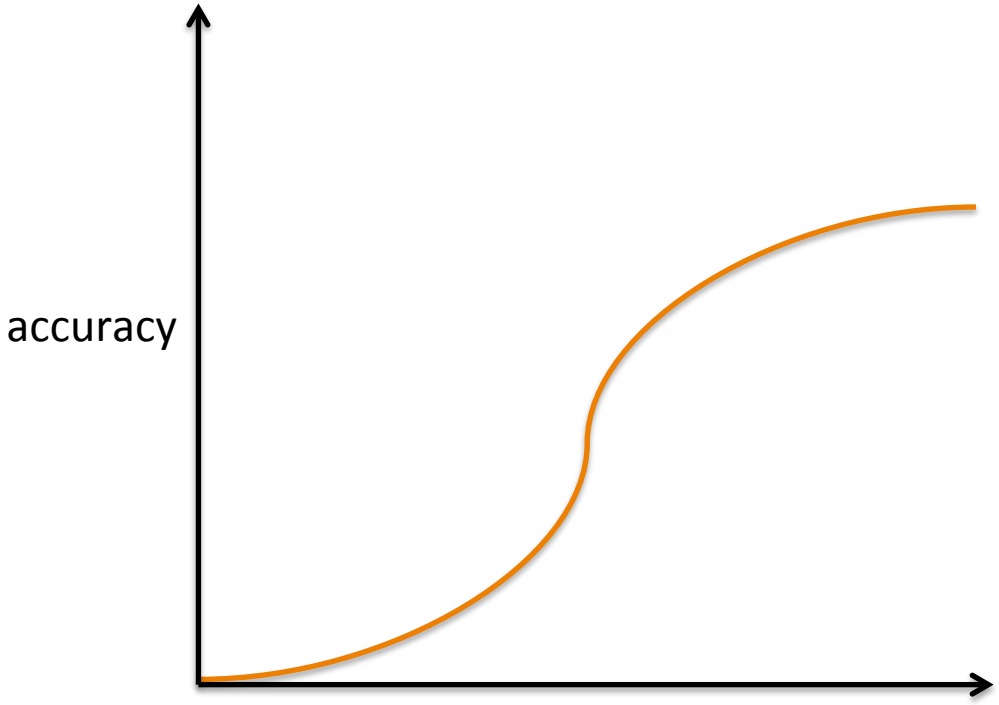
- >100 annotation tools − as many methods
  (https://github.com/NBISweden/GAAS/blob/master/annotation/CheatSheet/annotation_tools.md)
- 6 main class of approaches (Similarity-based, *ab initio*, hybrid, comparative, combiner, pipeline)

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)

To resume: Tools and methods

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/MAKER3 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** to be close to perfect

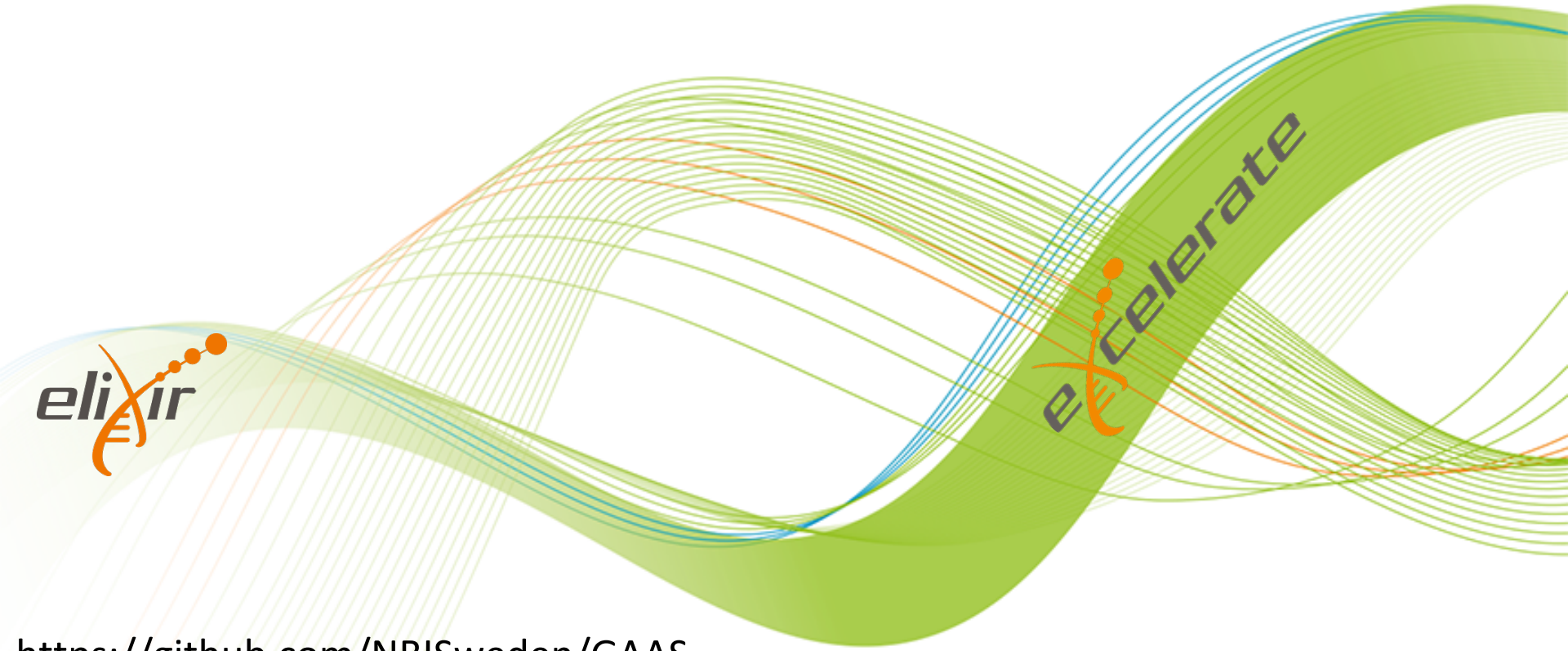
- As for assembly, an annotation is never finished, it can always be improved (e.g. Human) => **to know how to stop**

- Submit your annotation in public archive

SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

THE END



https://github.com/NBISweden/GAAS