

Gene annotation

Adapted from the courses of the Bonsai team,

CRISTAL UMR 9189

Sylvain.legrand@univ-lille.fr

Introduction

Challenge

- The constant **decrease** in **sequencing costs** makes it increasingly easy to obtain the sequence of the genome of a species
- However, in many respects, **genome annotation** has become more difficult!
 - The NGS **short reads** (Illumina) do not allow to obtain the **quality of assembly** of the first genomes (Drosophila, human, Arabidopsis...) obtained using Sanger technology
 - Genome sequencing projects with **unusual characteristics** and without prior data
 - Genome sequencing projects are now done "in house", by biologists who sometimes have **little bioinformatics skills**

Challenge

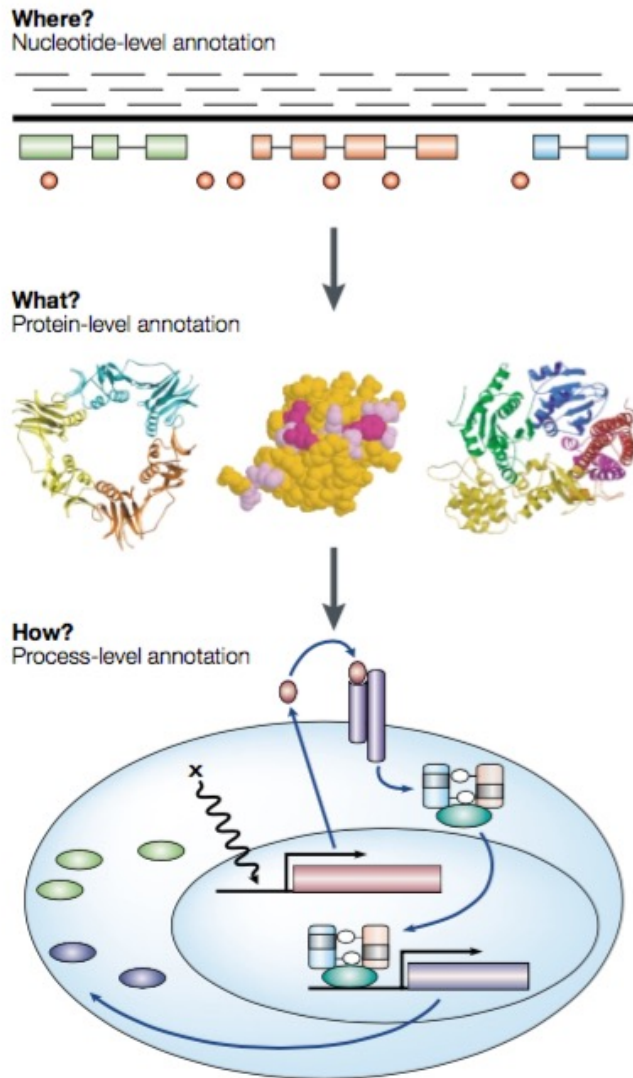


Figure 1 | **The three layers of genome annotation: where, what and how?**

Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001 Jul;2(7):493-503.

- **Objective of the lesson: gene annotation**
 - background
 - main methods
 - application to bacteria
 - application to eukaryotes
- **Going further: protein annotation**
 - context
 - function prediction
 - prediction of cellular localisation
 - study of 2D and 3D structures

Gene annotation

Quality of assembly

- The first step is to **validate the assembly**
- Observe the **metrics** (N50, L50..)

	<i>A. halleri</i> <i>halleri</i>	<i>A. halleri</i> <i>gemmifera</i>	<i>A. lyrata</i>	<i>A. thaliana</i>
<u>Nb scaffolds</u>	3 152	2 239	695	7
Total length	174 Mb	196 Mb	207 Mb	120 Mb
<u>Genome cov.</u>	68.3 %	76.9 %	89.9 %	88.9 %
<u>Longest scaff.</u>	1.5 Mb	4.3 Mb	33.1 Mb	30.4 Mb
N50	279 389	712 249	24 464 547	23 459 830
L50	177	71	4	3

Box 1 | Common statistics for describing genome assemblies

Genome assemblies are composed of scaffolds and contigs. Contigs are contiguous consensus sequences that are derived from collections of overlapping reads. Scaffolds are ordered and orientated sets of contigs that are linked to one another by mate pairs of sequencing reads.

Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012 Apr 18;13(5):329-42.

Scaffold and contig N50s

By far the most widely used statistics for describing the quality of a genome assembly are its scaffold and contig N50s. A contig N50 is calculated by first ordering every contig by length from longest to shortest. Next, starting from the longest contig, the lengths of each contig are summed, until this running sum equals one-half of the total length of all contigs in the assembly. The contig N50 of the assembly is the length of the shortest contig in this list. The scaffold N50 is calculated in the same fashion but uses scaffolds rather than contigs. The longer the scaffold N50 is, the better the assembly is. However, it is important to keep in mind that a poor assembly that has forced unrelated reads and contigs into scaffolds can have an erroneously large N50.

Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012 Apr 18;13(5):329-42.

Quality of assembly

- BUSCO <http://busco.ezlab.org/>
- Search for **universal single copy genes** in the assembly

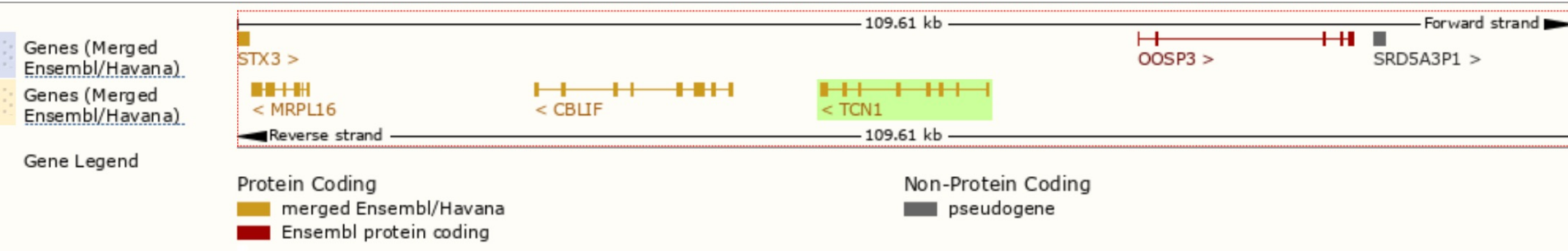
	<i>A. halleri</i> <i>halleri</i>	<i>A. halleri</i> <i>gemmaifera</i>	<i>A. lyrata</i>	<i>A. thaliana</i>
Complete universal single-copy <u>orthologs</u>	95.3%	97.6%	98.5%	98.2%
Fragmented universal single-copy <u>orthologs</u>	1.5%	0.3%	0.3%	0.5%
Missing universal single-copy <u>orthologs</u>	3.2%	2.1%	1.2%	1.3%



Assessing genome assembly and annotation completeness with **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Identification of repeated sequences

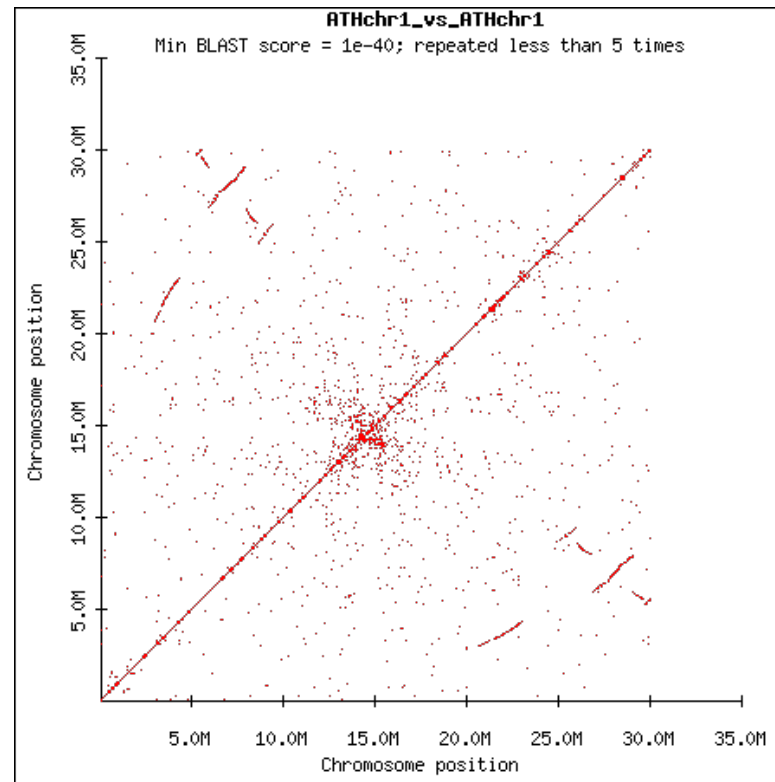
- **Eukaryotic** genomes can be **rich in repeated sequences**: 47% of the human genome, and only 1-2% of the genome is coding!
- At first sight, the **human genome** seems to be **a model of inefficiency**: genes spaced by large regions (10-100 kb), introns
- **In yeast**: 60% of the genome encodes the 6000 proteins. The 35,000 human genes are encoded by a genome 300 x larger



Screenshot from Ensembl.org Human genome 11:59,804,864-59,914,472

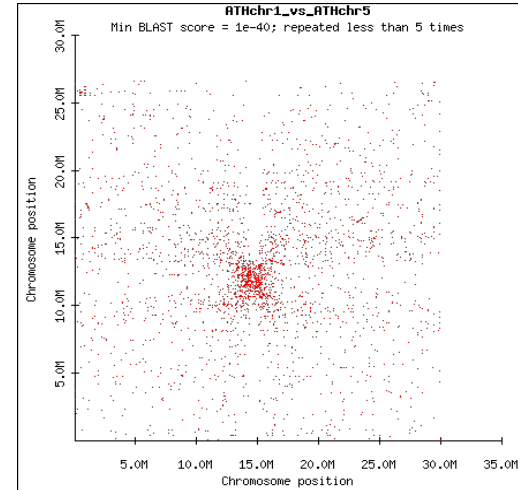
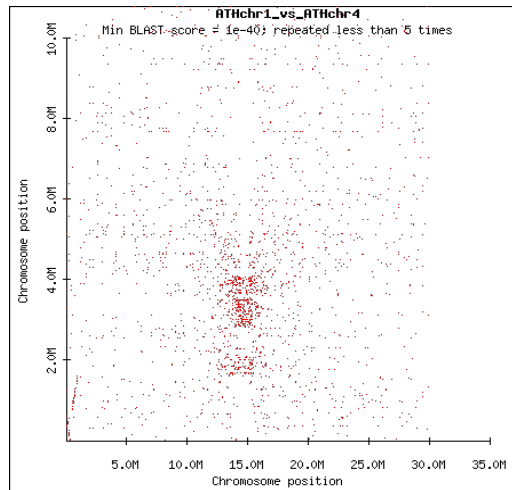
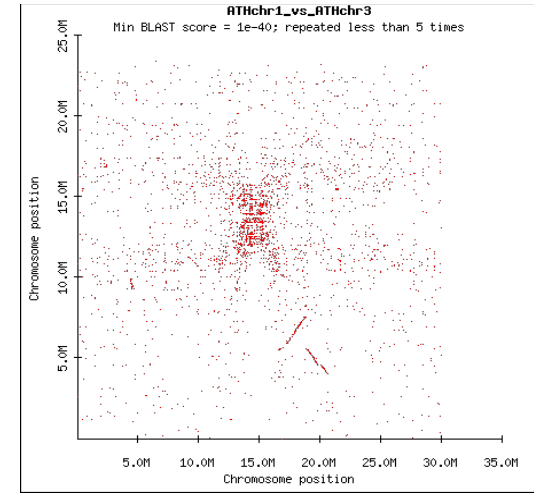
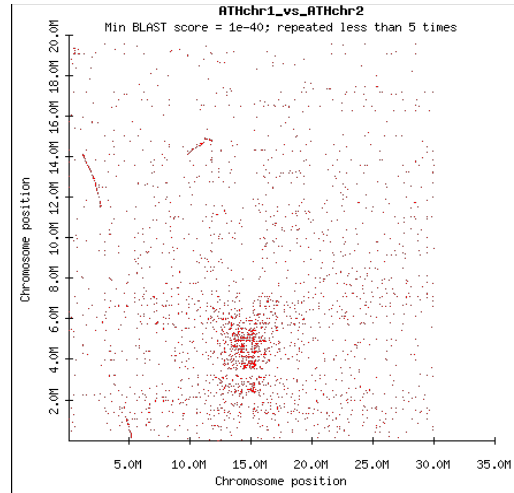
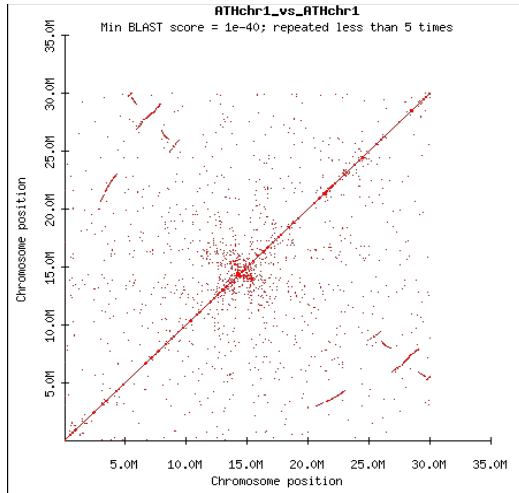
Identification of repeated sequences

- **Dotplot** Chr1 *Arabidopsis thaliana* aligned against itself
- There are **many repeats**
- There are more repeats at the **centromere** in *At*



http://biolinx.bios.niu.edu/t80maj1/rice/arab_mega_dotplots.htm

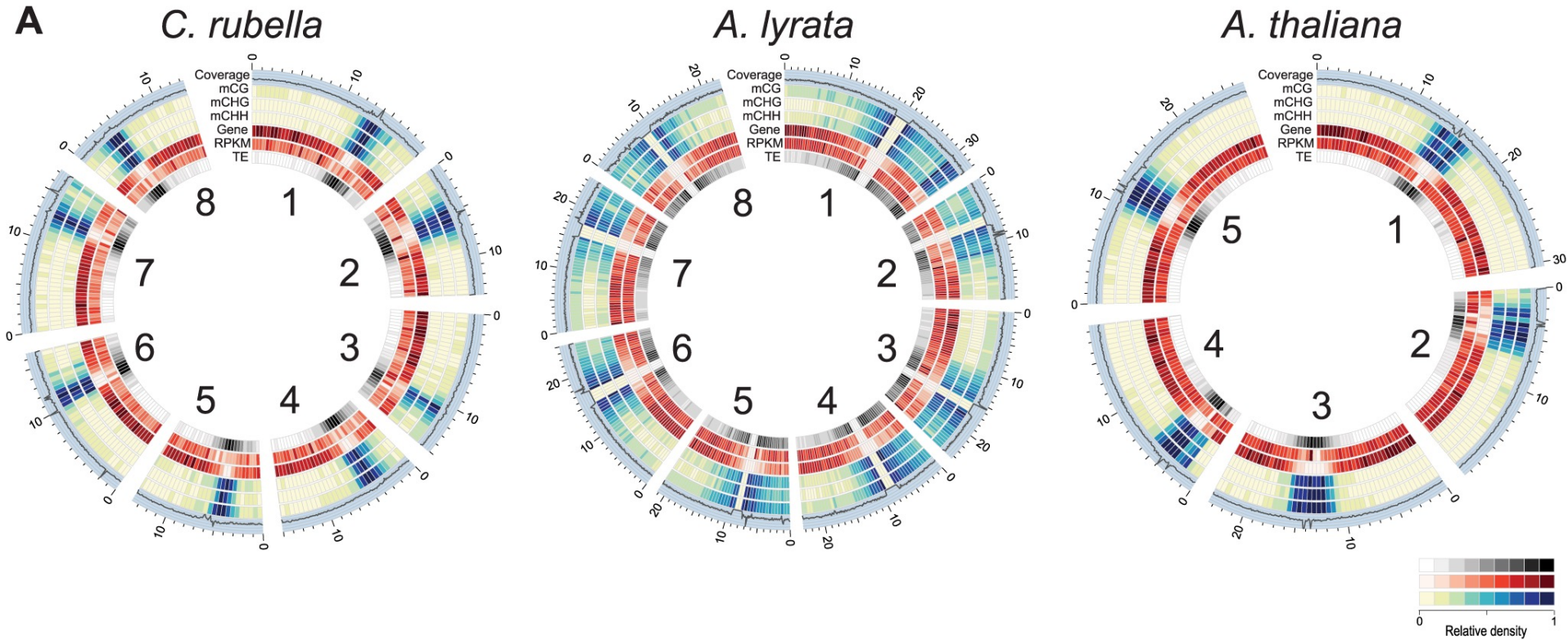
Identification of repeated sequences



http://biolinx.bios.niu.edu/t80maj1/rice/arab_mega_dotplots.htm

Identification of repeated sequences

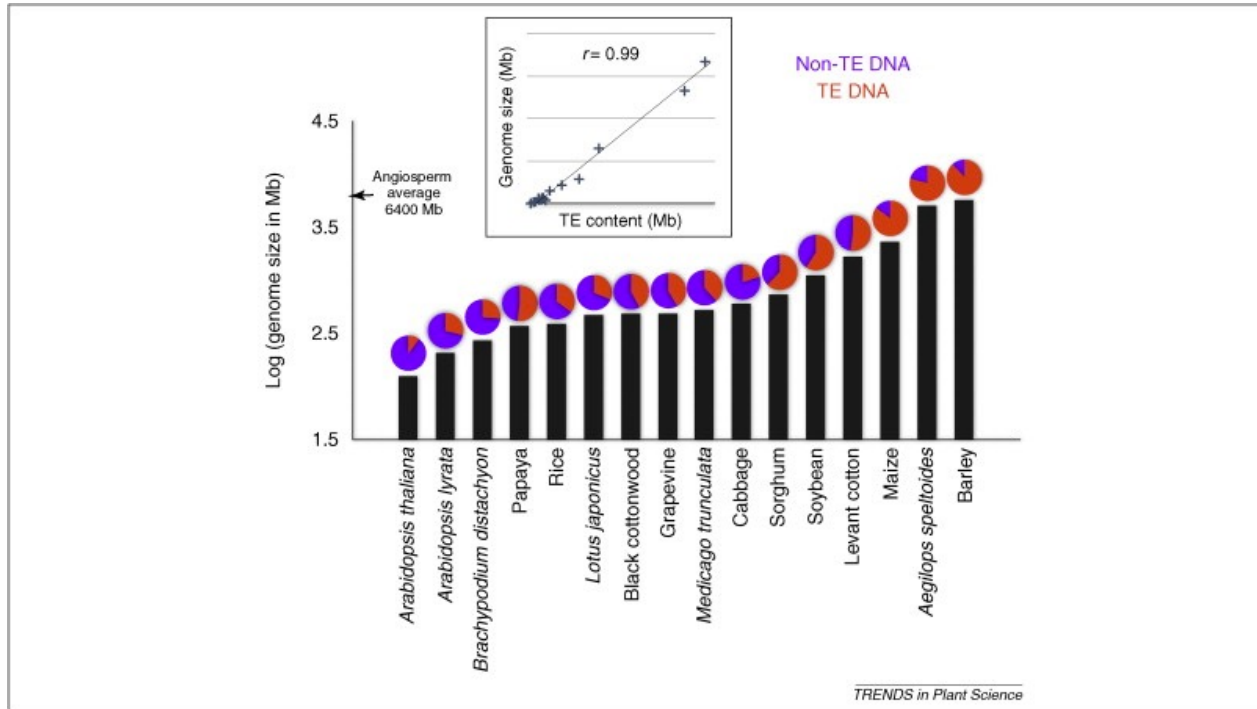
- Transposable elements can have **different locations** in the genome



Seymour et al., Plos Genet 2014

Identification of repeated sequences

- **Genome size is correlated** with **transposable element content**



Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 2010 Aug;15(8):471-8.

Identification of repeated sequences

- **Repeated sequences interfere** with **genome assembly** and **gene prediction**: ORFs of transposable elements are identified as genes of the host organism. They can also produce errors in the annotation of neighbouring genes
- **Identification** of repeated sequences and their **masking** are usually **the first steps** in annotating a (eukaryotic) genome
- **Masking**: replace these regions with "N" or lower case letters (softmasking)

Identification of repeated sequences

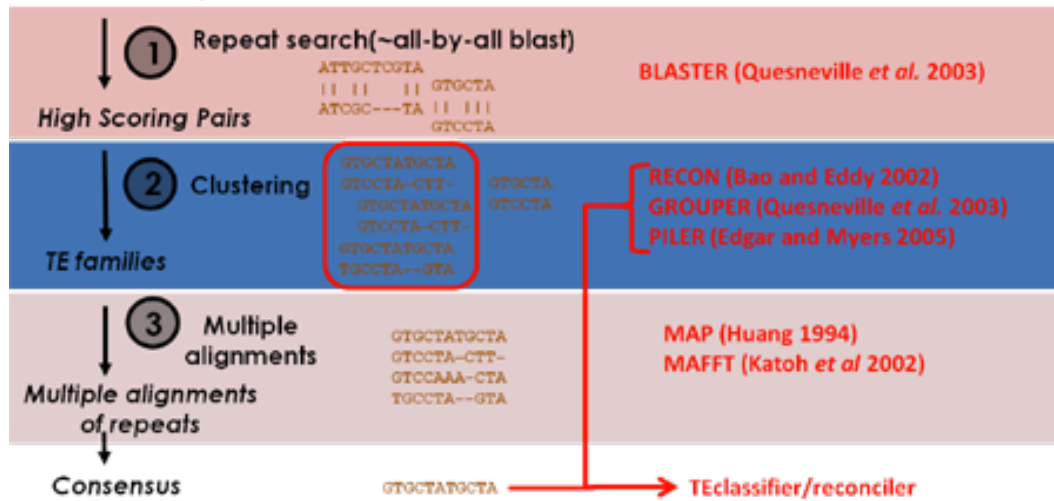
- Two types of analysis: **homology-based** or ***de novo***
- As transposable elements are **poorly conserved** between species, *de novo* analysis has the advantage of being able to identify specific families of elements
- Once a database of transposable elements has been obtained, the elements can be **identified** using tools such as RepeatMasker, Crossmatch... It is also possible to combine different tools
- In addition to transposable elements, the identified repeats can also include regions of **low complexity** and **repeated genes**: histones, tubulins...

Identification of repeated sequences

• TE identification (de novo)

Genomic sequence ...TATGTGCTATTACTATTAGATTACCATGCGT...

Pipeline TEdenovo
(Flutre et al. In prep.)

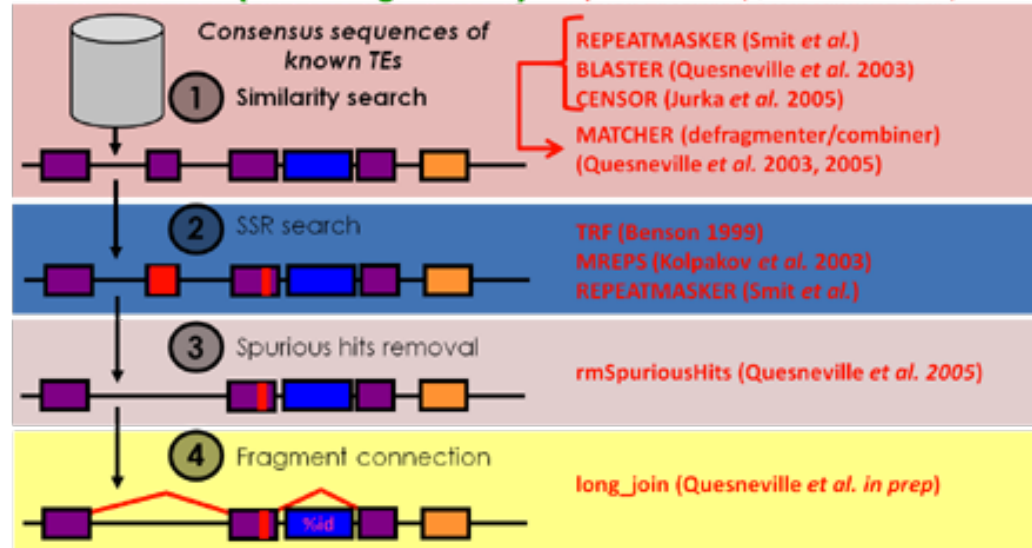


The REPET package (URGI)

Flutre T. et al, 2011

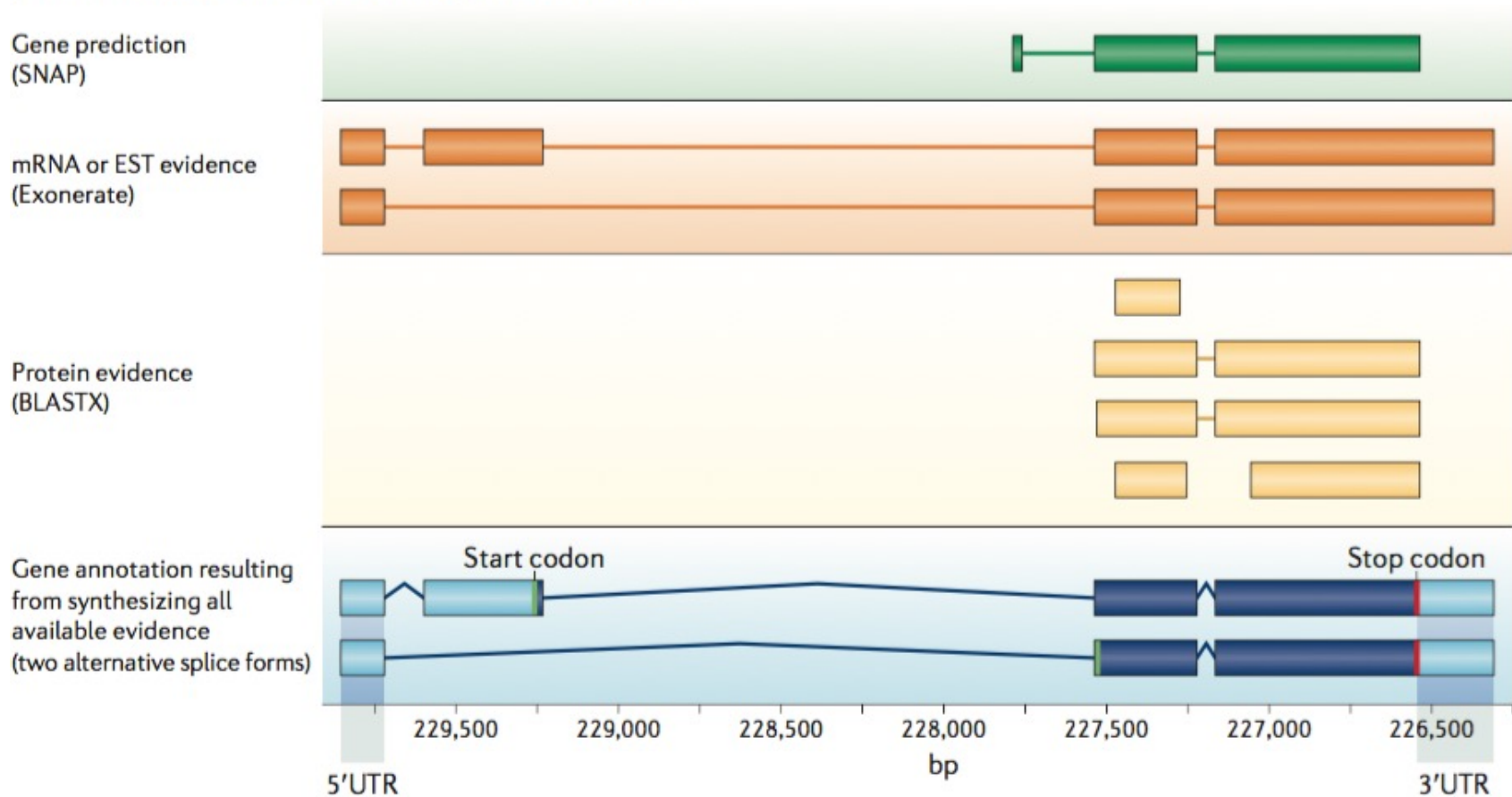
• TE annotation (knowledge based)

Pipeline TEannot (Quesneville et al. 2005)



Gene annotation

Box 2 | Gene prediction versus gene annotation

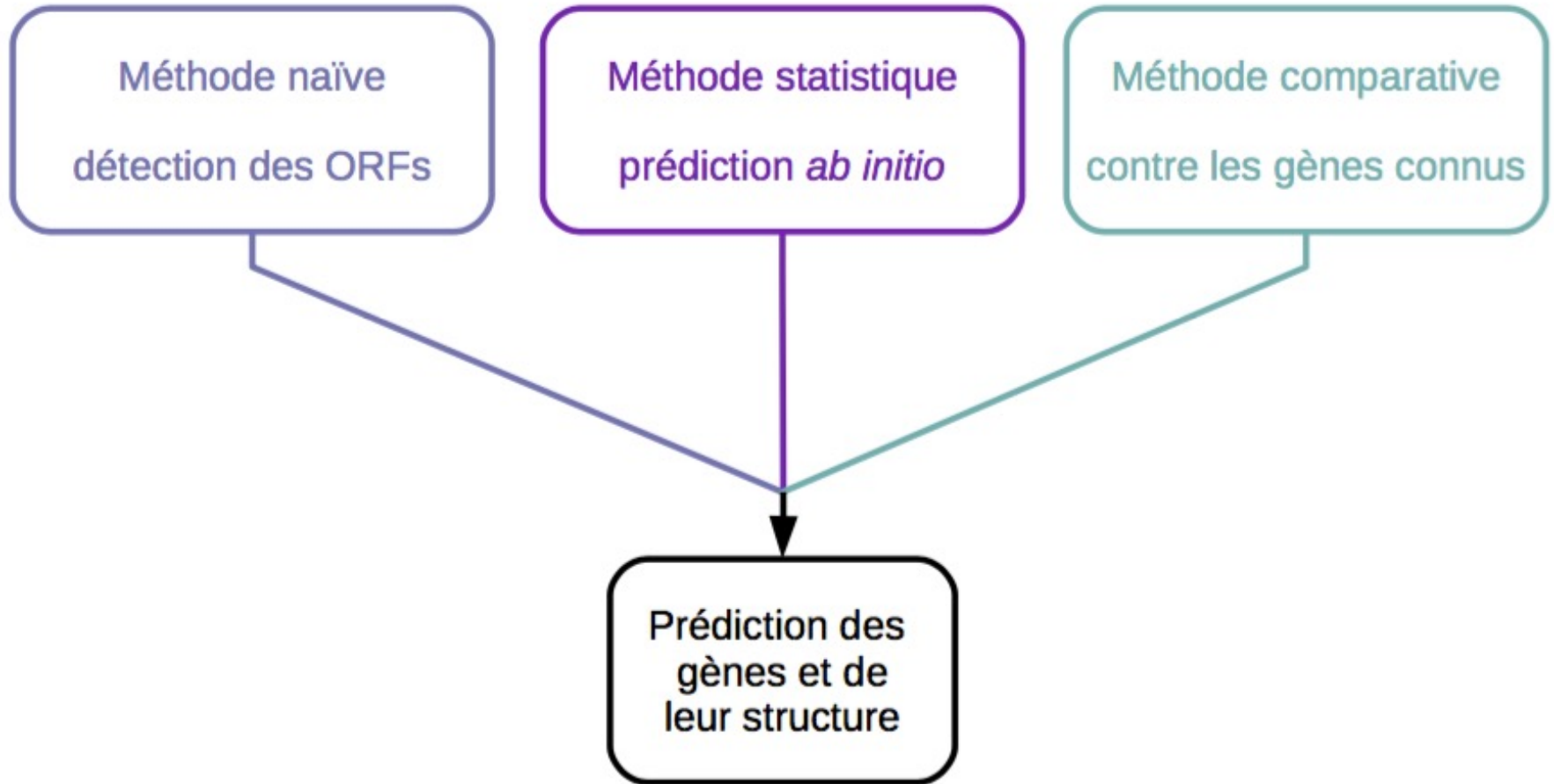


Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012 Apr 18;13(5):329-42.

Gene annotation

- **Starting point:** raw nucleic acid sequences
- **Output**
 - Start and end positions of genes
 - Transcription, splicing and translation signals
 - Idea of the function of proteins encoded by genes
- **Limits**
 - Some genes are not predicted (false negatives)
 - Some predicted genes are not true genes (false positives)
 - Precise gene boundaries are sometimes wrong (wrong initiation codon choice...)

Main methods



Naive approach

- **Principle:** search for coding sequence signals
 - Start with an initiation codon ATG + other
 - End with a termination codon TAA, TAG or TGA
 - Have a size multiple of 3 (for genes without introns)
- **Implementation:** detect open reading frames
 - ORFs = Open Reading Frames
 - Frames that may contain a gene
 - >50 nt between an initiation codon and a termination codon
 - Blind translation in the 6 reading frames (3 frames per DNA strand)

Naive approach

- The **6 reading frames** of a nucleic sequence



Naive approach

- **Advantages**

- *ab initio* method: without prior knowledge
- Reduces the amount of data to be analysed for sequence comparison

- **Limits**

- Not all ORFs are genes
- Sensitive to sequencing errors
- Not very useful for eukaryotic genes (presence of introns)

Statistical approach

- **Principle:** discriminate between coding and non-coding sequences
 - Based on code usage bias
- **Implementation:**
 - Learning the code usage for a given organism from reliable coding sequences
 - Calculation of the probability that a portion of a sequence is coding
 - Analysis of transcription and translation signals to determine gene boundaries

Bias in the use of the genetic code

- 1 amino acid is encoded by N codons → synonymous codons
- Non-uniform distribution of codons used

aa	codons	% par aa	Nb
A	GCA	0,65	11
	GCC	0	0
	GCG	0	0
	GCT	0,35	6
F	TTC	0,21	7
	TTT	0,79	27
G	GGA	0,50	11
	GGC	0	0
	GGG	0,05	1
	GGT	0,45	10

Exemple : gène *cytB* de *P. falciparum*
 G+C = 27.59 % du génome

Statistical approach

- **Advantages**

- *ab initio* approach: without prior annotation of genes
- More reliable criteria than the naive approach

- **Limits**

- Need for a training dataset: confirmed coding sequences
- Does not detect small genes/exons (below detection threshold)
- CDS identification only, no identification of UTRs
- No identification of alternative splicing

- Some tools such as TwinScan, FGENESH, Augustus, Gnomon, GAZE and SNAP, can use evidence (mRNA, proteins) to improve evidence-driven predictions (compared to *ab initio*)

Comparative approach

- **Principle:** locate annotations from databases on the genome sequence
 - Alignments with known proteins → location of CDS, including introns
 - Alignments with mRNAs (ESTs, cDNAs, etc.) → location of CDSs + UTRs, including introns
- **Implementation**
 - Sequence comparison against libraries of mRNA or protein using Blast
 - Alignment of matched mRNAs or proteins using a specialised software

- **Comparison of the DNA sequence to nucleic acid databases** using Blastn (or equivalent)
 - (- Detection of contaminating sequences (vectors...)
Specialised blast: VecScreen)
 - Detection of mRNAs potentially derived from the DNA sequence → comparison with mRNAs obtained from the same species or from closely related species
- **Alignment** between the **DNA** sequence and the **matched mRNAs** using **specialised software**
 - Fine determination of 5' and 3' UTRs and exons
 - Software: EST2genome, Splign

Use of RNA-seq data

- This is the data that has the greatest potential to improve annotation
- Allows a better delimitation of exons, splice sites, and alternative splicing events
- But large amount of data, complex because often short Illumina reads
- 2 ways to use the reads
 - Genome-independent *de novo* assembly of reads (ABYSS, SOAPdenovo, Trinity). The resulting transcripts are then aligned to the genome in the same way as seen for mRNA
 - Directly aligned to the genome (TopHat, GSNAP, Scripture), then the alignments are assembled using Cufflink

Comparison to proteins

- **Comparison** of the translated DNA sequence (in the 6 frames) **to protein databases** using BLASTX
 - Detection of proteins potentially encoded by the sequence
- **Alignment** of **matched proteins** using specialised software
 - Determination of initiation codon and intron/exon junctions
 - Software: GeneWise

Comparative approach

- **Advantages**

- Validates potential genes by comparison with experimental data (mRNA, proteins)
- Provides clues to protein function

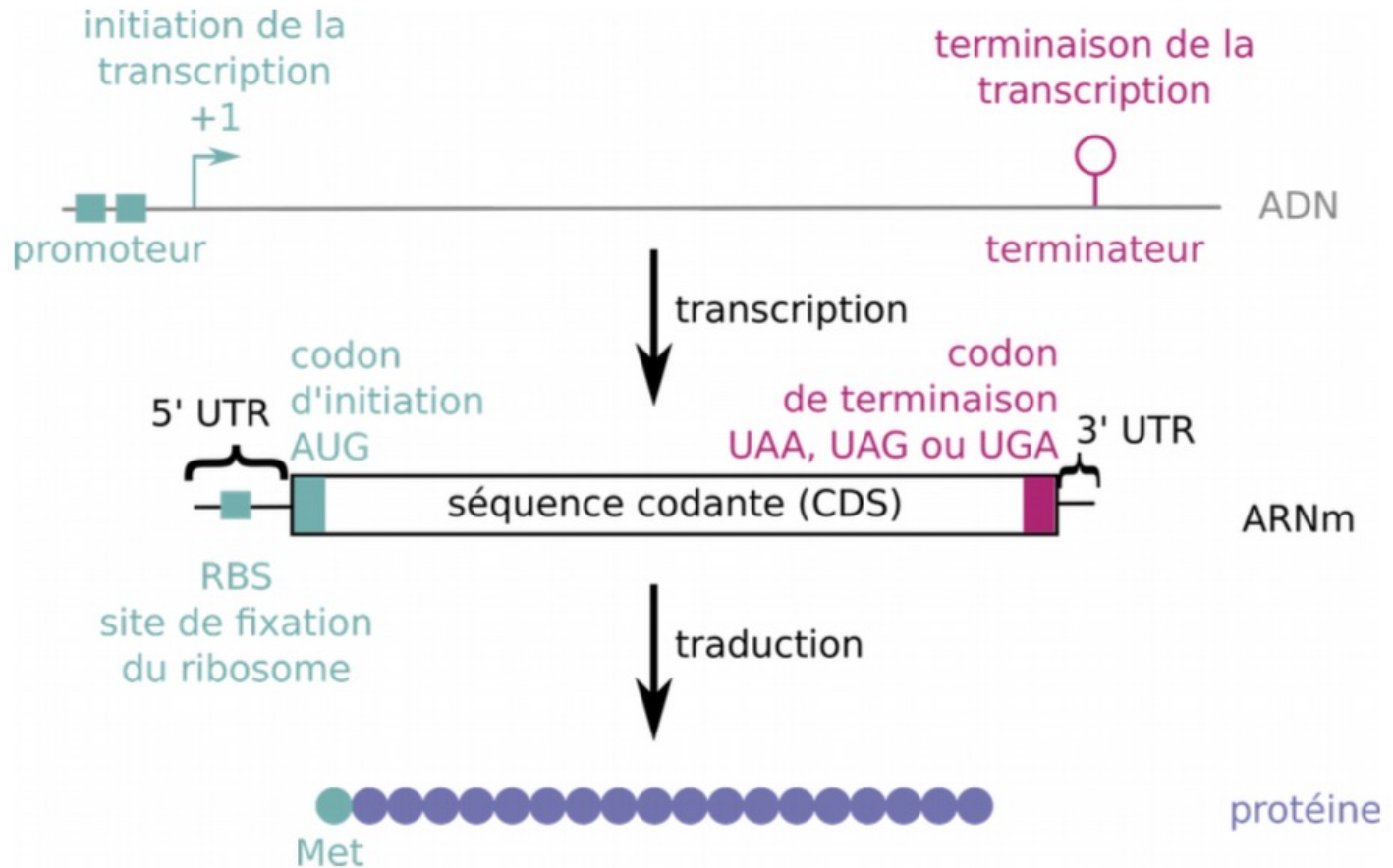
- **Limits**

- Requires *a priori* knowledge
- Does not find orphan sequences
- Difficult with isolated genomes from a taxonomic point of view
- Propagates errors in libraries

Structure of prokaryotic genes

- Over **80% of the genome is coding**
 - Short intergenic sequences
 - On average: one gene per 1,000 nucleotides (kb)
- **Simple gene structure**
 - Short transcribed but untranslated regions (3' and 5' UTR)
 - No intron (with some exceptions)

Structure of prokaryotic genes



Structure of prokaryotic genes

- Here is an extract from the **genome sequence** of *Pseudoalteromonas* sp.

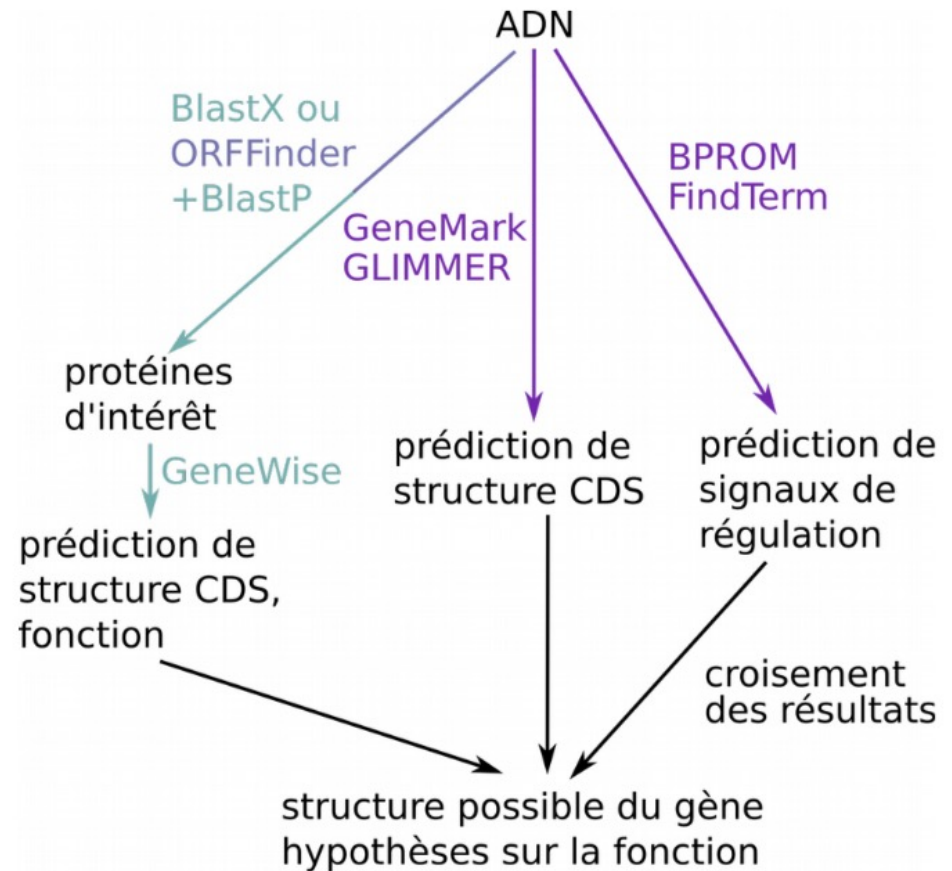
>AB057417

```
aacgaaaagattaaaaatcatttttctcttggaatcttactctacccccatta  
atgaatgcaaattagaaaagcttttttctgtactgttcagaaactgttaggagaactaaa  
aaacatgaacattcgtcctttacaagatcgcgtaatcgttaaacgtctagaagaagaac  
aaaatctgctggcgggtattgtattaactggctctgcagctgaaaaatcaactcgcggaga  
agtagtagccgtaggtaatggtcgtatcttagataacggtgacgttagagctttagaagt  
aaaagccggtgacactgtgttatttggctcatatggttgagaaaactgaaaagatcgaagg  
tcaagagtacctgatcatgcgtgaagacaacattttaggcattgtaggcctaagcctactt  
ttcgtttaacacacatttaagaatttagagg
```

Proposed workflow

▪ Analyse in 4 steps

1. ORF identification
 - ORFFinder
2. ORF validation
 - SmartBlast (GeneWise if needed)
3. Statistical prediction of CDS
 - GeneMark, GLIMMER
4. Statistical prediction of regulatory signals
 - BPRM



ORFfinder

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for [Linux x64](#).

Examples (click to set values, then click Submit button) :

- NC_0111604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt



Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

From: To:

Choose Search Parameters

Minimal ORF length (nt):

Genetic code:

ORF start codon to use:

"ATG" only

"ATG" and alternative initiation codons

Any sense codon

Ignore nested ORFs:

Start Search / Clear

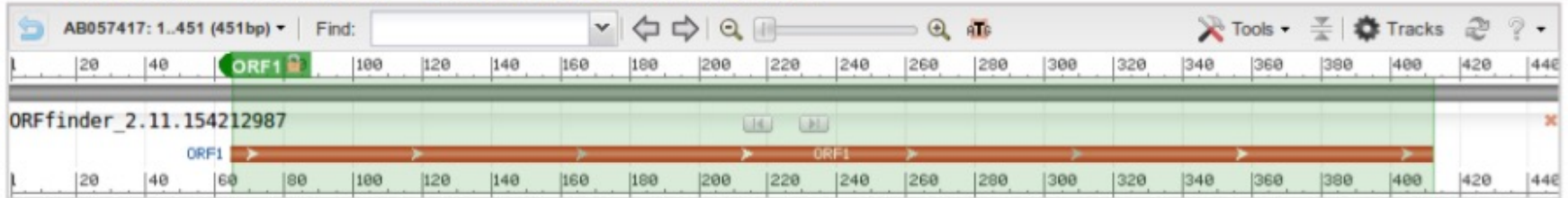
<https://www.ncbi.nlm.nih.gov/orffinder/>

ORFfinder results

Open Reading Frame Viewer

AB057417

ORFs found: 1 Genetic code: 1 Start codon: 'ATG' and alternative codons



brin +, positions : 65..412

ORF1 (115 aa)

Display ORF as...

Mark

Mark subset...

Marked: 0

Download marked set

Add six-frame translation track

as Protein FASTA

```
>lcl|ORF1
MQIRKAFFCTVQKLLGELKNMNIPLQDRVIVKRL EEETK
SAGGIVLTGSAAEKSTRGEVVAVGNRILDNGDVRAL EVK
AGD TVLFGSYVEKTEKIEGQEYLIMREDNILGIVG
```

SmartBLAST ORF1

BLAST ORF1

BLAST marked set

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF1	+	2	65	412	348 115

1 seule ORF prédite

=> sûrement 1 seul gène, voir aucun

possibilité de lancer BlastP

avec la protéine codée par l'ORF
contre des familles de protéines

(<https://ncbiinsights.ncbi.nlm.nih.gov/2015/07/29/smartblast/>)

SmartBlast

- **SmartBlast** compares the ORF against database (« landmark database ») consisting of the proteomes of 27 species spread over a large phylogeny. Also compares against the nr database https://blast.ncbi.nlm.nih.gov/smartblast/smartBlast.cgi?CMD=Web&PAGE_TYPE=BlastDocs#searchSets
- It returns the 5 best results obtained against the « landmark database »
- He then returns the results obtained against the « nr » database

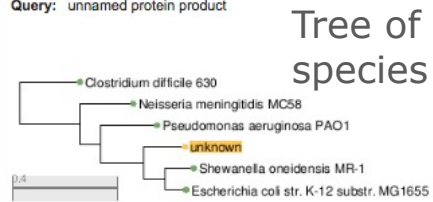
SmartBlast results

Summary

[Report description](#)

Query: unnamed protein product

Query length: 115 aa



Domaines
Fonction

DOMAIN: co-chaperonin GroES

chaperonin GroES

co-chaperonin GroES

co-chaperonin GroES

Your query: unnamed protein product

10 kDa chaperonin GroES

Cpn10 chaperonin GroES, small subunit of GroESL

Visualization of alignments



[See full multiple alignment](#)

Legend

[About the database](#)

Descriptions

Best hits

5 best results against « landmark » database

Select: [All](#) [None](#) Selected:0

[Alignments](#) [GenPept](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
	10 kDa chaperonin GroES [Shewanella oneidensis MR-1]	130	130	82%	1e-39	72%	NP_716336.1
	Cpn10 chaperonin GroES, small subunit of GroESL [Escherichia coli str. K-12 substr. MG1655]	127	127	81%	2e-38	72%	NP_418568.1
	co-chaperonin GroES [Pseudomonas aeruginosa PAO1]	119	119	81%	6e-35	62%	NP_253076.1
	co-chaperonin GroES [Neisseria meningitidis MC58]	111	111	81%	7e-32	59%	NP_274967.1
	chaperonin GroES [Clostridioides difficile 630]	86.3	86.3	81%	4e-22	46%	YP_001086663.1

Additional BLAST Hits

Results against « nr » database

Select: [All](#) [None](#) Selected:0

[Alignments](#) [GenPept](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
	MULTISPECIES: co-chaperone GroES [Pseudoalteromonas]	186	186	82%	3e-65	100%	WP_006791252.1
	MULTISPECIES: co-chaperone GroES [Pseudoalteromonas]	184	184	82%	8e-65	98%	WP_004587676.1
	co-chaperone GroES [Pseudoalteromonas sp. TMED43]	184	184	82%	2e-64	98%	OUX91642.1

SmartBlast results

- Best hit against « landmark » database

Alignments

GenPept ▼ Next ▲ Previous ▲ Descriptions

10 kDa chaperonin GroES [Shewanella oneidensis MR-1]
Sequence ID: [NP_716336.1](#)

Range 1: 1 to 96 [GenPept](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
130 bits(327)	1e-39()	Compositional matrix adjust.	69/96(72%)	79/96(82%)	1/96(1%)	

Query 21 MNIRPLQDRVIVKRLLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNGRILDNGDVRALEVK 80
MNIRPL DRVIVKRL E+ SAGGIVLTGSAAEKSTRGEV+AVGNGRIL+NG VR L+VK
Sbjct 1 MNIRPLHDRVIVKRLVEVETSAGGIVLTGSAAEKSTRGEVLAVGNGRILENGTVRPLDVK 60

Query 81 AGDTVLFG-SYVEKTEKIEGQEYLIMREDNILGIVG 115
GD V+F Y K EKI+GQE LI+ E +++ IVG
Sbjct 61 VGDVVIFNEGYGVKKEIDGQEVLLSEADLMAIVG 96

pas 100 % id

pas début ORF,
mais début prot

GenPept ▼ Next ▲ Previous ▲ Descriptions

Cpn10 chaperonin GroES, small subunit of GroESL [Escherichia coli str. K-12 substr. MG1655]
Sequence ID: [NP_418566.1](#)

Range 1: 1 to 95 [GenPept](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
127 bits(320)	2e-38()	Compositional matrix adjust.	68/95(72%)	78/95(82%)	1/95(1%)	

Query 21 MNIRPLQDRVIVKRLLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNGRILDNGDVRALEVK 80
MNIRPL DRVIVKR E ETKSAGGIVLTGSA KSTRGEV+AVGNGRIL+NG+V+ L+VK
Sbjct 1 MNIRPLHDRVIVKRKEVETKSAGGIVLTGSAAAKSTRGEVLAVGNGRILENGEVKPLDVK 60

Query 81 AGDTVLFGS-YVEKTEKIEGQEYLIMREDNILGIV 114
GD V+F Y K+EKI+ +E LIM E +IL IV
Sbjct 61 VGDIVIFNDGYGVKSEKIDNEEVLIMSESDILAIV 95

Related Information
[Gene](#) - associated gene details
[Identical Proteins](#) - Identical proteins to WP_011071021.1

Related Information
[Gene](#) - associated gene details
[Identical Proteins](#) - Identical proteins to WP_001026276.1

SmartBlast results

- Best hit against « nr » database

Additional BLAST Hits

Select: [All](#) [None](#) Selected: 0

Alignments [GenPept](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	MULTISPECIES: molecular chaperone GroES [Pseudoalteromonas]	186	186	82%	3e-65	100%	WP_006791252.1
<input type="checkbox"/>	MULTISPECIES: molecular chaperone GroES [Pseudoalteromonas]	184	184	82%	8e-65	98%	WP_004587676.1

mêmes espèces

GenPept

Next Previous Descriptions

MULTISPECIES: molecular chaperone GroES [Pseudoalteromonas]

Sequence ID: [WP_006791252.1](#) Length: 95 Number of Matches: 1

100 % id

Range 1: 1 to 95 [GenPept](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
186 bits(471)	3e-65()	Compositional matrix adjust	95/95(100%)	95/95(100%)	0/95(0%)	

Query 21 MNIRPLQDRVIVKRL EETKSAGGIVLTGSAAEKSTRGEVVAVGNRILNDGVRAL EVK 80
 Sbjct 1 MNIRPLQDRVIVKRL EETKSAGGIVLTGSAAEKSTRGEVVAVGNRILNDGVRAL EVK 60

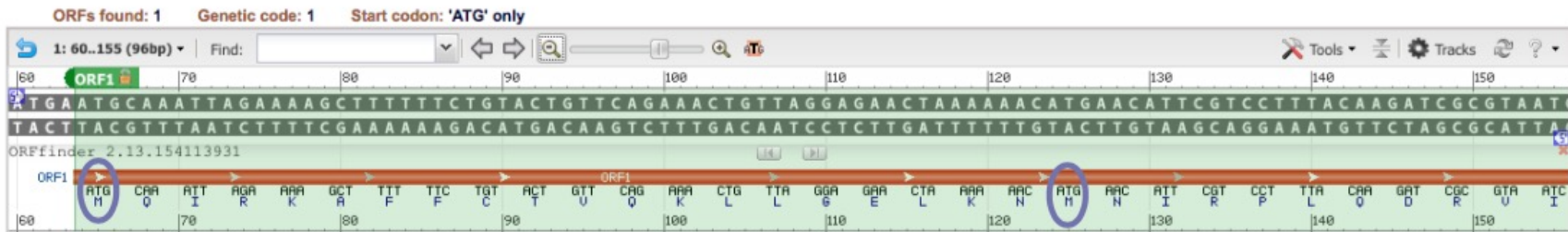
Query 81 AGD TVLFGSYVEKTEKIEGQEYLIMREDN ILGIVG 115
 Sbjct 61 AGD TVLFGSYVEKTEKIEGQEYLIMREDN ILGIVG 95

Related Information

[Identical Proteins](#) - Identical proteins to WP_006791252.1

pas début ORF, mais début prot

fin ORF et fin protéine



SmartBlast results

- **ORF : 65..412** on the strand + of the DNA sequence
 - Codes a 115 aa protein + stop codon
- Alignments provided by SmartBLAST
 - Query 21..115 : only a part of the ORF protein
 - So the **ORF is not fully coding**
 - The alignment starts at 21 => the CDS starts at $65+(21-1)*3 = 125$
 - End of the CDS at 412
 - Sbjct 1..95 : The protein from the database is complete
 - The predicted coding sequence is complete**
- The alignments obtained with different sequences are good
 - Prediction is **reliable**, no need for GeneWise

GeneMark

A family of gene prediction programs developed at
[Georgia Institute of Technology](http://www.gatech.edu), Atlanta, Georgia, USA.

Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes



Novel genomic sequences can be analyzed either by the self-training program **GeneMarkS** (sequences longer than 50 kb) or by **GeneMark.hmm with Heuristic models**. For many species pre-trained model parameters are ready and available through the **GeneMark.hmm** page. Metagenomic sequences can be analyzed by **MetaGeneMark**, the program optimized for speed.

Gene Prediction in Eukaryotes



Novel genomes can be analyzed by the program **GeneMark-ES** utilizing unsupervised training. Note that GeneMark-ES has a special mode for analyzing fungal genomes. Recently, we have developed a semi-supervised version of GeneMark-ES, called GeneMark-ET that uses RNA-Seq reads to improve training. For several species pre-trained model parameters are ready and available through the **GeneMark.hmm** page.

Gene Prediction in Transcripts



Sets of assembled eukaryotic transcripts can be analyzed by the modified **GeneMarkS** algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of **GeneMark.hmm with Heuristic models**. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for [download](#).

Gene Prediction in Viruses, Phages and Plasmids



Sequences of viruses, phages or plasmids can be analyzed either by the **GeneMark.hmm with Heuristic models** (if the sequence is shorter than 50 kb) or by the self-training program **GeneMarkS**.

<http://exon.gatech.edu/GeneMark/>

- **GeneMark.hmm with Heuristic models**
- Same result as ORFfinder; in contradiction with the start identified by SmartBlast

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Tue Jan  2 06:16:26 2018
Sequence file name: seq.fna
Model file name: GeneMark_hmm_heuristic.mod
RBS: false
Model information: Heuristic_model_for_genetic_code_11_and_GC_36
```

FASTA definition line: AB057417

Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	65	412	348	1

GeneMark results

- **GeneMark.hmm** model adapted for *Pseudoalteromonas sp.*
- Result **in agreement** with SmartBlast
- **Best results with a model** defined for the studied species

GeneMark.hmm PROKARYOTIC (Version 3.26)

Date: Tue Jan 2 06:25:38 2018

Sequence file name: seq.fna

Model file name: /home/genemark/parameters/prokaryotic/Pseudoalteromonas_atlantica_T6c/

RBS: true

Model information: Pseudoalteromonas_atlantica_T6c

FASTA definition line: AB057417

Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	125	412	288	1

BPRM

Used in more than [800 publications](#).

Reference: V. Solovyev, A Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78

BPRM - Prediction of bacterial promoters

BPRM is bacterial sigma70 promoter recognition program with about 80% accuracy and specificity. It is best used in regions immediately upstream from ORF start for improved gene and operon prediction in bacteria.

Paste nucleotide sequence here (plain or in fasta format):

```
>AB057417
aacgaaaagattaaaaattatcatttttctcttggatttttactctacccccatta
atgaatgcaaattagaaaagcttttctgtactgttcagaaactgttaggagaactaaa
```

Alternatively, load a local file with sequence:

Local file name:

Choisissez un fichier

Aucun fichier choisi

Process Reset

[\[Help\]](#)

[\[Example\]](#)

Return to page with other programs of group: [Operon and gene finding in bacteria](#)

<http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>

Bprom results

Threshold for promoters - 0.20

Number of predicted promoters - 2

Promoter Pos: 418 LDF- 4.85

-10 box at pos. 402 ttgtaggct Score 44

-35 box at pos. 381 gtgaag Score 21

Promoter Pos: 80 LDF- 2.31

-10 box at pos. 65 atgcaaatt Score 29

-35 box at pos. 43 tttact Score 42

proche des vraies positions
cf. page suivante

Oligonucleotides from known TF binding sites:

For promoter at 418:

fnr:	TCAAGAGT	at position	361	Score -	13
purR:	TTTTCGTT	at position	419	Score -	5
purR:	TTTCGTTT	at position	420	Score -	6
rpoD15:	TTAACACA	at position	426	Score -	12
crp:	ACACACAT	at position	429	Score -	12
glpR:	CACACATT	at position	430	Score -	6

For promoter at 80:

soxS:	TATCATTT	at position	20	Score -	9
fur:	ATCATTTT	at position	21	Score -	8

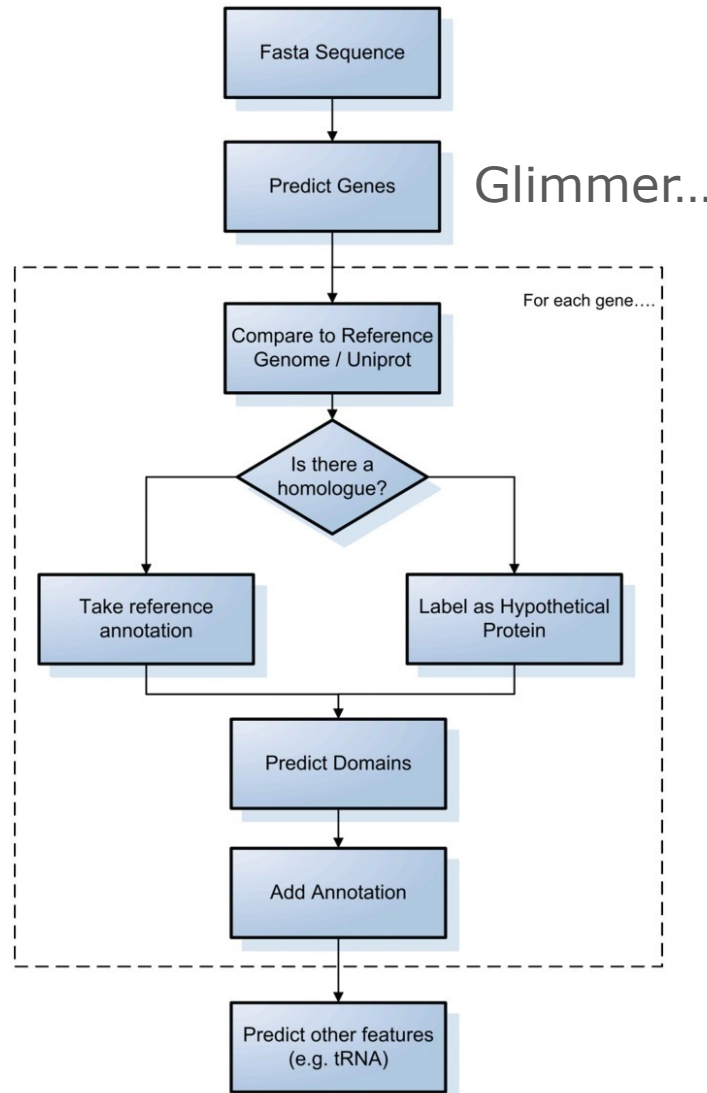
Summary of the analysis

- The methods lead to the **same conclusion**: the sequence contains a single CDS 125...412 (including the termination codon)
- Additional information given by SmartBLAST: the CDS encodes a **chaperone** of the type **Cpn10 / GroES**
- Glimmer (statistical prediction) finds no CDS

Prediction in bacteria: some pitfalls

- **Several initiation codons** (AUG) on the sequence: Which is the right one?
- Possibility of **alternative initiation codons** (GUG, UUG)
Confirmation by:
 - Presence of RBS (Ribosome Binding Site)
 - Comparative analysis with other species
 - Statistical prediction
- **Incomplete genes** (early stop codon, phase shift)
 - Real (corrected during translation, pseudogenes)
 - Sequencing errors
 - Detection by: BlastX reports inconsistencies (different frames); comparison + prediction
- **Overlapping genes**
 - Common in viruses, sometimes in bacteria (gene ends)

Alternative pipeline

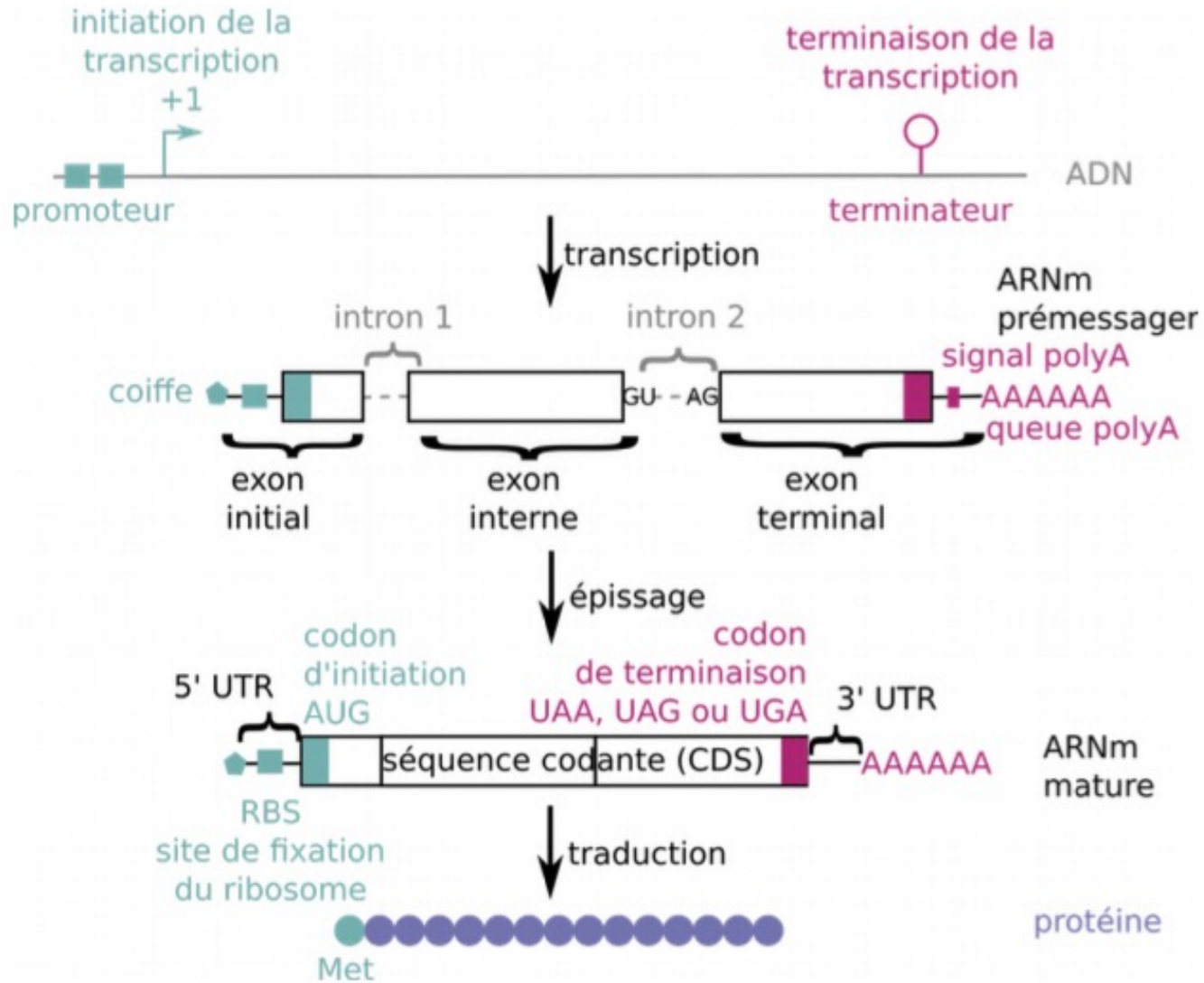


Richardson EJ, Watson M. The automatic annotation of bacterial genomes. *Brief Bioinform.* 2013 Jan;14(1):1-12.

Eukaryotic gene structure

- **Low proportion** of protein coding sequences in genomes
 - About 2% of the human genome
- Presence of a very large number of **repeated sequences**
 - ~ 50% of the human genome
- **Complex gene structure**
 - Long 3' and 5' untranslated regions (non-coding exons)
 - Presence of introns, alternative splicing

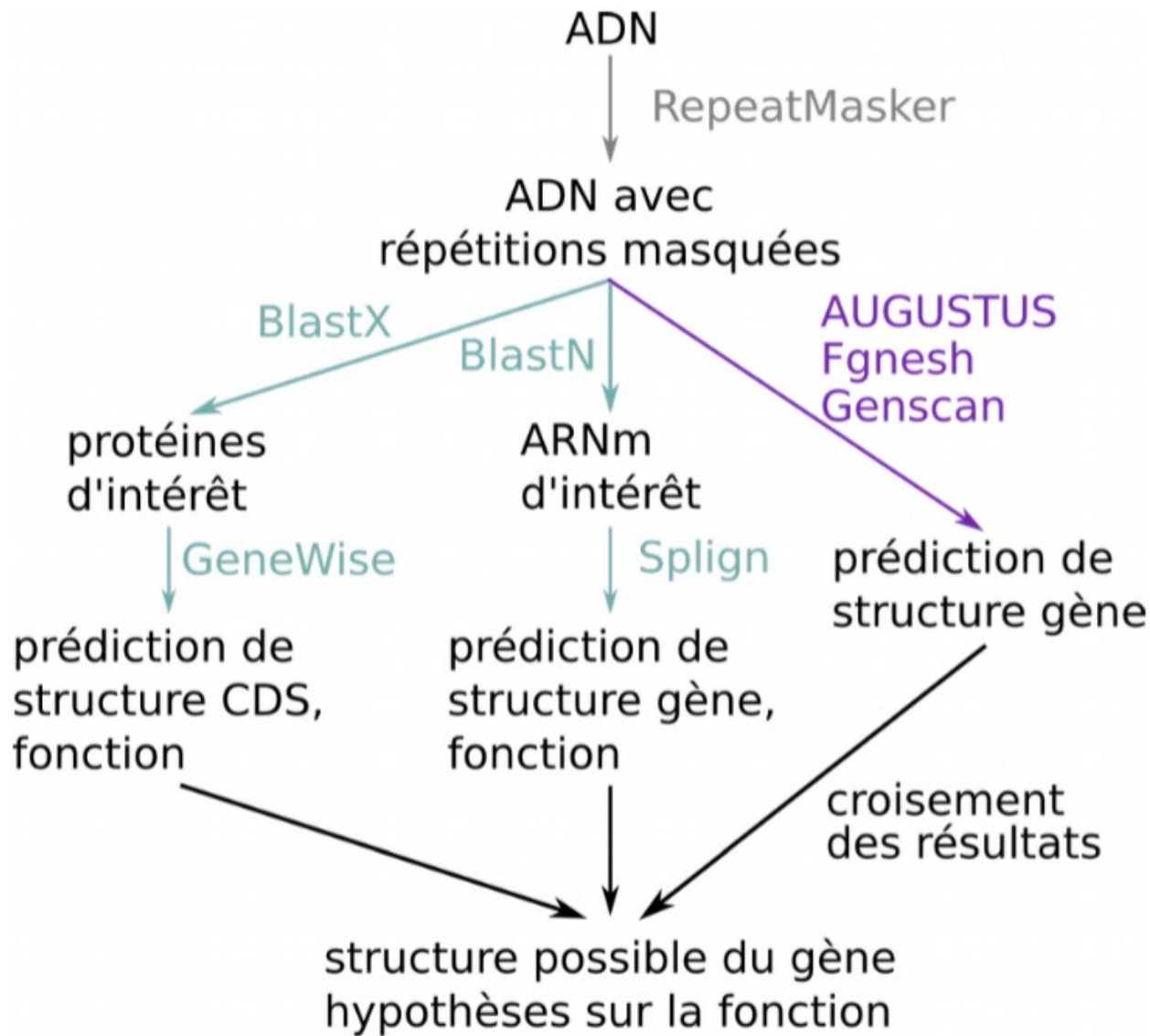
Eukaryotic gene structure



Intron consequences

- **Exon size not a multiple of 3**
 - Codons cut by an intron
 - Frame shift from one exon to another
 - No strand change
- Existence of **short exons** (~ 10 nt)
 - Above the resolution limits of software
- Existence of **very long introns** ($>$ exons)
 - Difficulty in locating exons
- **Alternative splicing**
 - Concerns $> 50\%$ of human genes

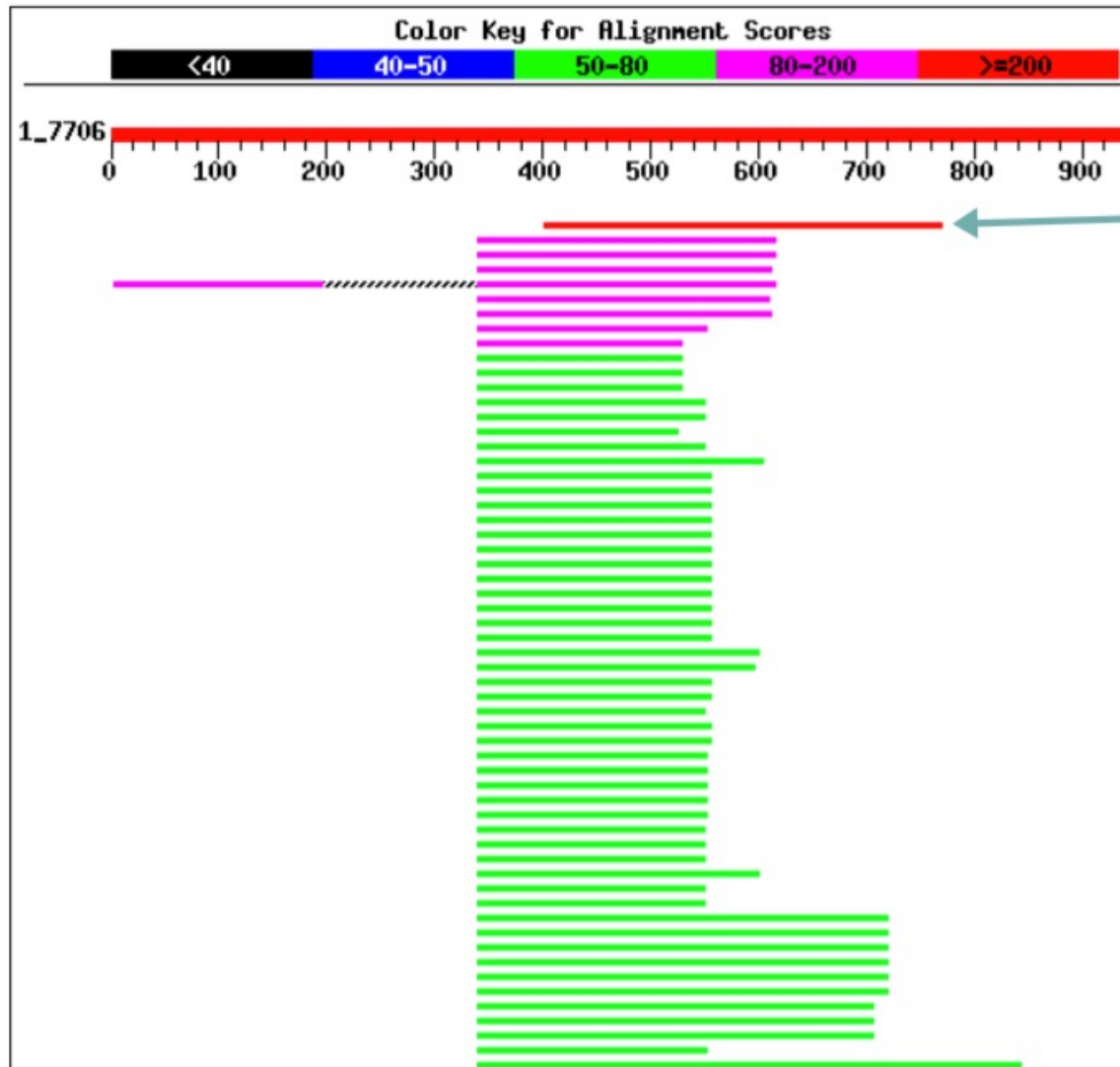
Suggested workflow



Example: study of an mRNA

- 905 bp **mRNA** from a human cell
- **Three steps** analysis
 1. Search for CDS in mRNA
BlastX + GeneWise
 2. Localization of the gene corresponding to the mRNA
Blast "Genomes" + Est2genomes or Splign
 3. Testing of statistical methods on the genome sequence
FGENESH, AUGUSTUS, GeneScan

Blastx results, graphic



bon score mais
seule protéine
s'alignant avec
cette région

Blastx results, alignments

annotation automatique
=> peu fiable

```
>gi|55641083|ref|XP_529628.1 PREDICTED: hypothetical protein XP_529628 [Pan troglodytes]
      Length = 155           Score = 227 bits (578), Expect = 4e-58
      Identities = 109/123 (88%), Positives = 110/123 (89%)   Frame = +1
Q 403 APGERRPGETERGSTQGDQAAHRGTEVLHVGAEQPRAPVLGAGRQHALAPRGGVQRPRIP 582
      +PGERRPGETERGSTQGDQAAH GTEVLHVGAEQPRAPVLGAGRQHALAPRGGVQRPRIP
S 33  SPGERRPGETERGSTQGDQAAHGTEVLHVGAEQPRAPVLGAGRQHALAPRGGVQRPRIP 92

Q 583 PTSCQLPALPALSFRCGESRASGGAHRLWQSCAHPAEAPVHLETRRQRPXXXXXXXXXXXX 762
      PTSCQLPALPALSFRCGESRASGGAHRLWQSCAHPAEAPVHLETRRQRP
S 93  PTSCQLPALPALSFRCGESRASGGAHRLWQSCAHPAEAPVHLETRRQRPGQGVNTGTVTT 152

Q 763 XRA 771
      RA
S 153 GRA 155
```

sp = SwissProt

=> fiable

```
>gi|32171340|sp|Q16528|B-ATF_HUMAN Gene info ATF-like basic leucine zipper transcriptional
factor B-ATF (SF-HT-activated gene-2) (SFA-2)
      Length = 125           Score = 185 bits (470), Expect = 1e-45
      Identities = 92/92 (100%), Positives = 92/92 (100%)   Frame = +2
Q 241 EKNRIAAQKSRQRQTQKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLC 525
      EKNRIAAQKSRQRQTQKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLC
S 34  EKNRIAAQKSRQRQTQKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLC 93

Q 521 SVLAASTPSPPEVVYSAHAFHQPHVSSPRFQP 616
      SVLAASTPSPPEVVYSAHAFHQPHVSSPRFQP
S 94  SVLAASTPSPPEVVYSAHAFHQPHVSSPRFQP 125
```

pas même phases
(les protéines suivantes
s'alignent aussi avec +2)

très bon alignement

Study of the alignment with the 2nd protein (1st is not relevant)

- **BATF_HUMAN**

Human protein, 100% identity => protein of interest

- Frame = +2: Coding sequence is on the + strand
- Query 341..616 / Sbjct 34..125
 - Need a specialised software to align this protein to the mRNA
- ATF-like basic leucine zipper transcriptional factor
 - May be a bZIP-type transcription factor

Pairwise Sequence Alignment

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

STEP 1 - Enter your sequences

Enter or paste your **protein** sequence in any supported format:

Or, upload a file: Aucun fichier choisi

AND

Enter or paste your **DNA** sequence in any supported format:

<https://www.ebi.ac.uk/Tools/psa/genewise/>

Wise results

```
BATF_HUMAN      1  MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ ← protéine d'intérêt
MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ
MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ ← prot codée par ARNm
ARNm_hsp        243 accatgaagtatactcccgcggttggaagcaagaacaggcaaccacac } codons en colonne
tcagcaggacgtggccccgaaaccaatggtaggaaagtccaaggagaca
gtccccctcccccttctcagcatttgaatggggatttccggcagggag

BATF_HUMAN      50  KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA
KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA
KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA
ARNm_hsp        390  aggacccgaggcgacaggccagaaccaggcattatgcaacgccttgcg
acactatagaataaaacctgaataatcaataatccttagaactgcttcc
gcccgcggcacgggagcgtacggcggcagaggccggggcccgcgcggggc

BATF_HUMAN      99  STPSPPEVVYSAHAFHQPHVSSPRFQP
STPSPPEVVYSAHAFHQPHVSSPRFQP
STPSPPEVVYSAHAFHQPHVSSPRFQP
ARNm_hsp        537  aactccgggtagcgtccccgatcctcc
gccccattagcactaacatgccgtac
cgcgccgggcccaccattcccgcgcg

FT              CDS      243..617
```

← début et fin de la CDS
sans le codon de terminaison

Wise results report

- Comparison with the protein of interest (BATF_HUMAN)
 - BlastX does not align the whole protein with the mRNA because the beginning of the protein contains an low complexity region that has been masked by BlastX
- GeneWise gives a CDS at position 243..617+3 on the mRNA
 - The protein is fully aligned with the mRNA

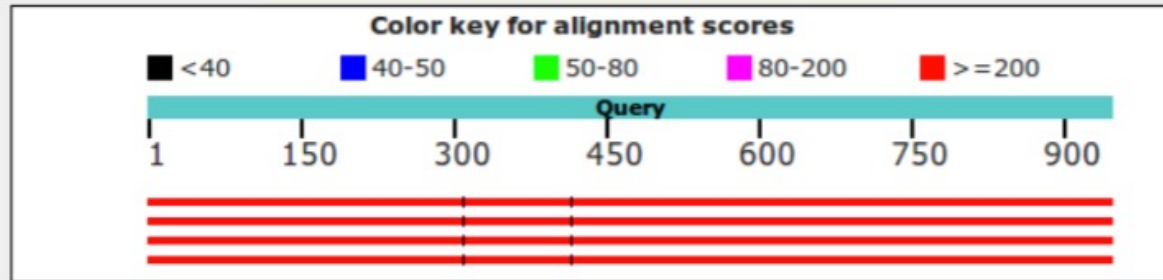


Blastn result against the genome

Graphic Summary

Distribution of the top 12 Blast Hits on 4 subject sequences

Mouse over to see the title, click to show alignments



Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly	977	1746	100%	0.0	100%	NC_000014.9
<input type="checkbox"/>	Homo sapiens chromosome 14 genomic scaffold, GRCh38.p7 Primary Assembly HSCHR1	977	1746	100%	0.0	100%	NT_026437.13
<input type="checkbox"/>	Homo sapiens chromosome 14, alternate assembly CHM1 1.1, whole genome shotgun se	977	1746	100%	0.0	100%	NC_018925.2
<input type="checkbox"/>	Homo sapiens chromosome 14 genomic scaffold, alternate assembly CHM1 1.1, whole ge	977	1746	100%	0.0	100%	NW_004929393.1

Blastn result against the genome, alignments

Download ▾ GenBank Graphics Sort by: Query start position ▾ Next Previous Descriptions

Homo sapiens chromosome 14 GRCh38.p7 Primary Assembly

Sequence ID: NC_000014.9 Length: 107043718 Number of Matches: 3

Range 1: 75522441 to 75522746 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
566 bits(306)	4e-158	306/306(100%)	0/306(0%)	Plus/Plus

Features: basic leucine zipper transcriptional factor ATF-like

Query 1 CAagagagagagagagCGTGCAAGCCCCAAAGCGAGCGACATGTCCCTTTGGGGAGCAGT 60
 Sbjct 75522441 CAAGAGAGAGAGAGAGAGCGTGCAAGCCCCAAAGCGAGCGACATGTCCCTTTGGGGAGCAGT 75522500

Query 301 AACAGG 306
 Sbjct 75522741 AACAGG 75522746

Range 2: 75525081 to 75525189 GenBank Graphics Next Match Previous Match First Match

Score	Expect	Identities	Gaps	Strand
202 bits(109)	1e-48	109/109(100%)	0/109(0%)	Plus/Plus

Features: basic leucine zipper transcriptional factor ATF-like

Query 303 CAGGACTCATCTGATGATGTGAGAAGAGTTCAGAGGAGGGAGAAAAATCGTATTGCCGCC 362
 Sbjct 75525081 CAGGACTCATCTGATGATGTGAGAAGAGTTCAGAGGAGGGAGAAAAATCGTATTGCCGCC 75525140

Query 363 CAGAAGAGCCGACAGAGGCAGACACAGAAGGCCGACACCCCTGCACCTGG 411
 Sbjct 75525141 CAGAAGAGCCGACAGAGGCAGACACAGAAGGCCGACACCCCTGCACCTGG 75525189

Range 3: 75546461 to 75546989 GenBank Graphics Next Match Previous Match First Match

Score	Expect	Identities	Gaps	Strand
977 bits(529)	0.0	529/529(100%)	0/529(0%)	Plus/Plus

Features: basic leucine zipper transcriptional factor ATF-like

Query 410 GGAGAGCGAAGACCTGGAGAAACAGAACCGCGGCTCTACGCAAGGAGATCAAGCAGCTCAC 469
 Sbjct 75546461 GGAGAGCGAAGACCTGGAGAAACAGAACCGCGGCTCTACGCAAGGAGATCAAGCAGCTCAC 75546520

Query 890 AGCAAGGCGGGCAGGGAAACGGTTATTTTTCTAAATAAATGCTTTAAAAG 938
 Sbjct 75546941 AGCAAGGCGGGCAGGGAAACGGTTATTTTTCTAAATAAATGCTTTAAAAG 75546989

Related Information

PubChem BioAssay - bioactivity screening
 Map Viewer - aligned genomic context

Gène sur chr14, brin +
 3 régions s'alignent => 3 exons ?
 Début..fin : 75522441..75546989
 Taille : 24550 nt
 => région à aligner avec l'ARNm :
 chr14+ 75522400..75547000

est2genome

Align EST sequences to genomic DNA sequence ([read the manual](#))

Unshaded fields are optional and can safely be ignored. ([hide optional fields](#))

Input section

Spliced EST nucleotide sequence(s). Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here:

3. To enter the sequence data manually, type here:

Unspliced genomic nucleotide sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here:

3. To enter the sequence data manually, type here:

<http://www.bioinformatics.nl/cgi-bin/emboss/est2genome>



EST2genome results

- Determination of **the position of exons** on **chr 14, region 75522400..75547000**

Exon	305	100.0	42	346	NC_000014	1	305	ARNm_hsp
+Intron	-20	0.0	347	2684	NC_000014			
Exon	105	100.0	2685	2789	NC_000014	306	410	ARNm_hsp
+Intron	-20	0.0	2790	24062	NC_000014			
Exon	528	100.0	24063	24590	NC_000014	411	938	ARNm_hsp
Span	898	100.0	42	24590	NC_000014	1	938	ARNm_hsp
Segment	305	100.0	42	346	NC_000014	1	305	ARNm_hsp
Segment	105	100.0	2685	2789	NC_000014	306	410	ARNm_hsp
Segment	528	100.0	24063	24590	NC_000014	411	938	ARNm_hsp

- So exon 1 start: $75522400+42-1$
- CDS location on chromosome 14 region:
`join(284..346,2685..2789,24063..24272)`

Splign results

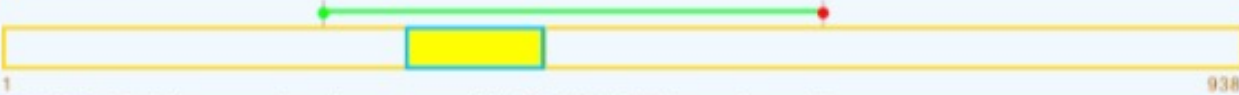
HOME SEARCH SITE MAP
Overview
Online
Download
Documentation
Contacts

#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	ARNm_hsp(+)	568815584(+)	75522441-75546989	100.00	100.00	100.00	0.00	0.00

[Graphics](#) | [Text](#)

Model 1	Coverage	100.00%	CDS	0.00%	Mismatches and indels	0
	Overall	100.00%	In-frame	0.00%	Exons (min/max/ave), bp	105 / 528 / 312
	Exon	100.00%	Primary transcript	938 bp	Introns (min/max/ave), bp	2339 / 21274 / 11806

ARNm_hsp (+)



1 568815584 (+) Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly 938

75522441 75546989

Segments Alignment

1 2 3		D S S D D V R R V Q R R E K N R I A A Q K S			
	306 GACTCATCTGATGATGTGAGAAGAGTTCAGAGGAGGGAGAAAATCGTATTGCCGCCCAGAAGAG		75525079	CCCAGGACTCATCTGATGATGTGAGAAGAGTTCAGAGGAGGGAGAAAATCGTATTGCCGCCCAGAAGAG
	371	R Q R Q T Q K A D T L H L		75525149	CCGACAGAGGCAGACACAGAAGGCCGACACCCTGCACCTG
					CCGACAGAGGCAGACACAGAAGGCCGACACCCTGCACCTGGTAAG

Fgenesh results

G	Str	Feature	Start	End	Score	ORF	Len
1	+	1 CDSf	284 -	346	9.43	284 - 346	63
1	+	2 CDSi	1573 -	1644	0.66	1573 - 1644	72
1	+	3 CDSi	2685 -	2789	18.51	2685 - 2789	105
1	+	4 CDSl	24063 -	24272	19.16	24063 - 24272	210
1	+	PolA	24574		1.12		

1 exon supplémentaire
épissage alternatif ?

3 exons identiques aux
prédictions par comparaison
de séquences

CDSf = CDS first (commence par un codon d'initiation)

CDSi = CDS internal (ni codon d'initiation, ni codon de terminaison)

CDSl = CDS last coding segment (se termine par un codon de terminaison)

PolA = signal pour la queue polyA

Augustus results

NC_000014	AUGUSTUS	gene	284	24272	0.89	+	.	Gene	g1
NC_000014	AUGUSTUS	mRNA	284	24272	0.89	+	.	mRNA	g1.t1
NC_000014	AUGUSTUS	start_codon	284	286	.	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	initial	284	346	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	internal	2685	2789	0.99	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	terminal	24063	24272	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	CDS	284	346	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	CDS	2685	2789	0.99	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	CDS	24063	24272	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	stop_codon	24270	24272	.	+	0	mRNA	g1.t1

- **3 exons:** identical to results obtained by sequence comparison approach
 - Only coding regions (the term mRNA is abused)

GenScan results

Gn.Ex	Type	S	Begin	..End	Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	284	346	63	1	0	83	80	45	0.914	4.35
1.02	Intr	+	2685	2789	105	2	0	114	119	114	0.996	17.51
1.03	Term	+	24063	24272	210	2	0	83	49	404	0.985	33.29
1.04	PlyA	+	24574	24579	6							1.05

- **3 exons:** identical to results obtained by sequence comparison approach
 - Only coding parts
- **PolyA tail** predicted at the same location as FGENESH

Summary of the analysis

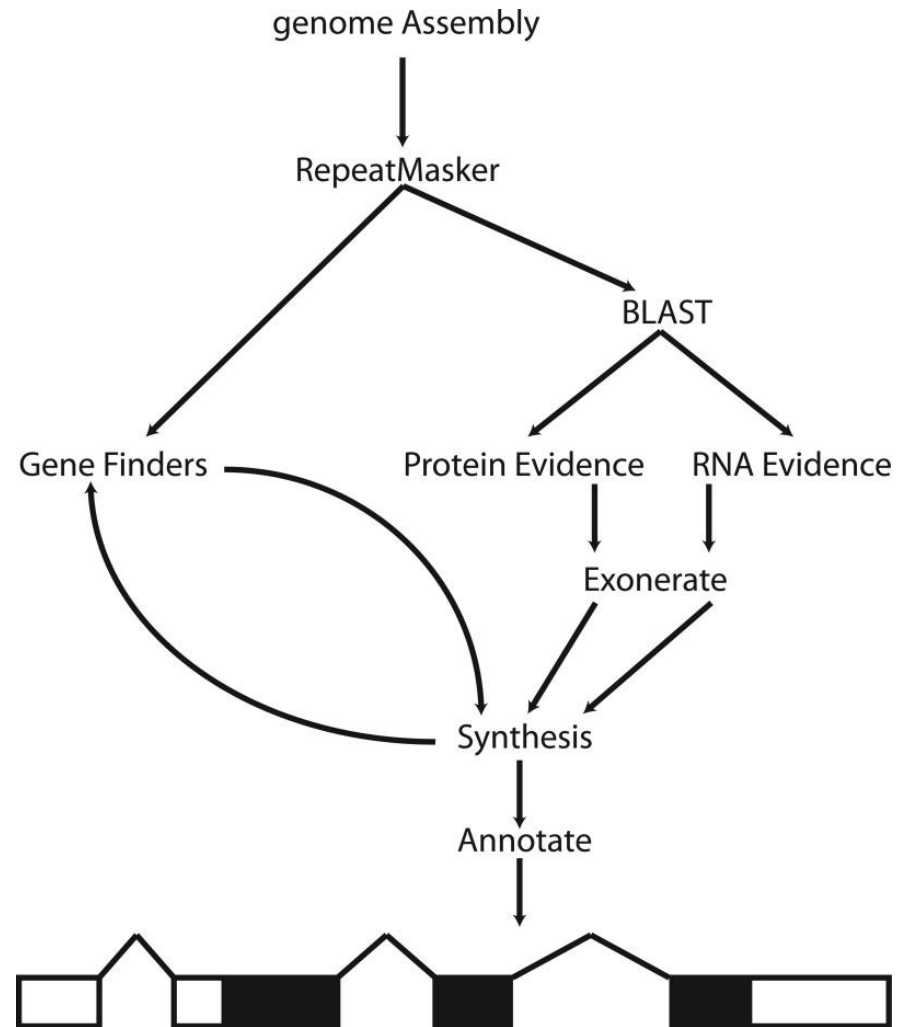
- All predictions agree on **3 exons**
- **One additional coding exon** predicted by FGENESH
Alternative splicing ?
- The encoded protein is probably a **B-zip transcription factor**



Automatic annotation of a whole genome

- Use an annotation pipeline: Maker, PASA, Gnomon ...
- Example: Maker

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 2008 Jan;18(1):188-96.

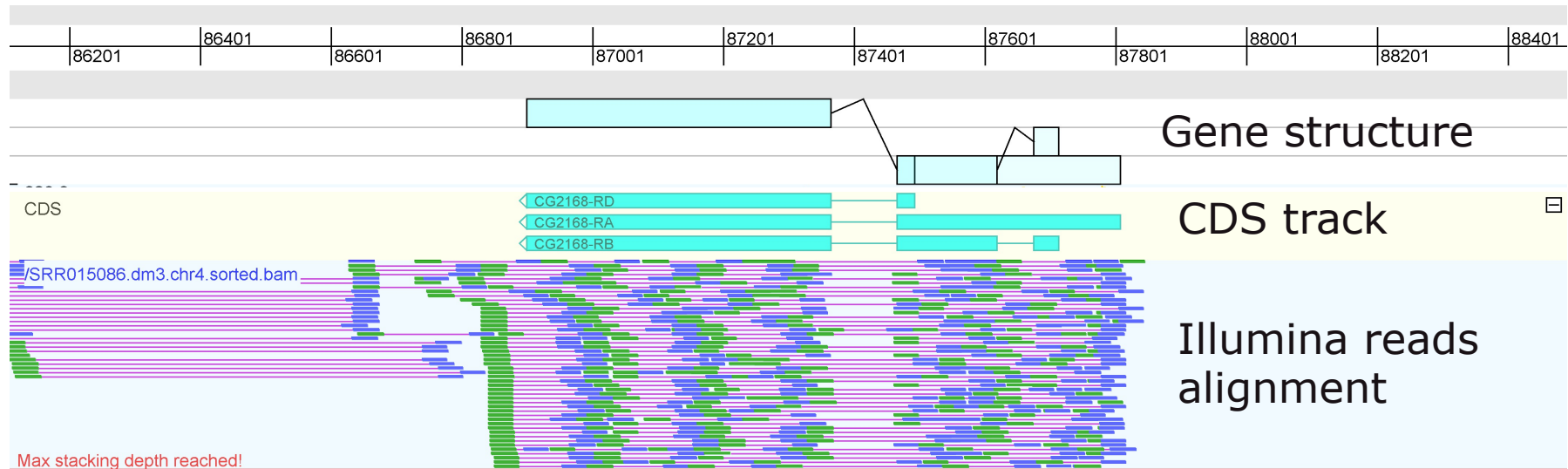


Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. *Curr Protoc Bioinformatics.* 2014 Dec 12;48:4.11.1-39.

Viewing annotation data

- Five commonly used formats for annotations: **GFF3**, GenBank, BED, GTF and EMBL
GFF3 → <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>
- Use of a genome visualisation tool (IGV, Jbrows, GenomeView...)

Sequence in bp



Exemple de visualisation par GenomeView

Viewing annotation data

Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000014.9

[GenBank](#) [Graphics](#)

>NC_000014.9:75522400-75547000 Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly

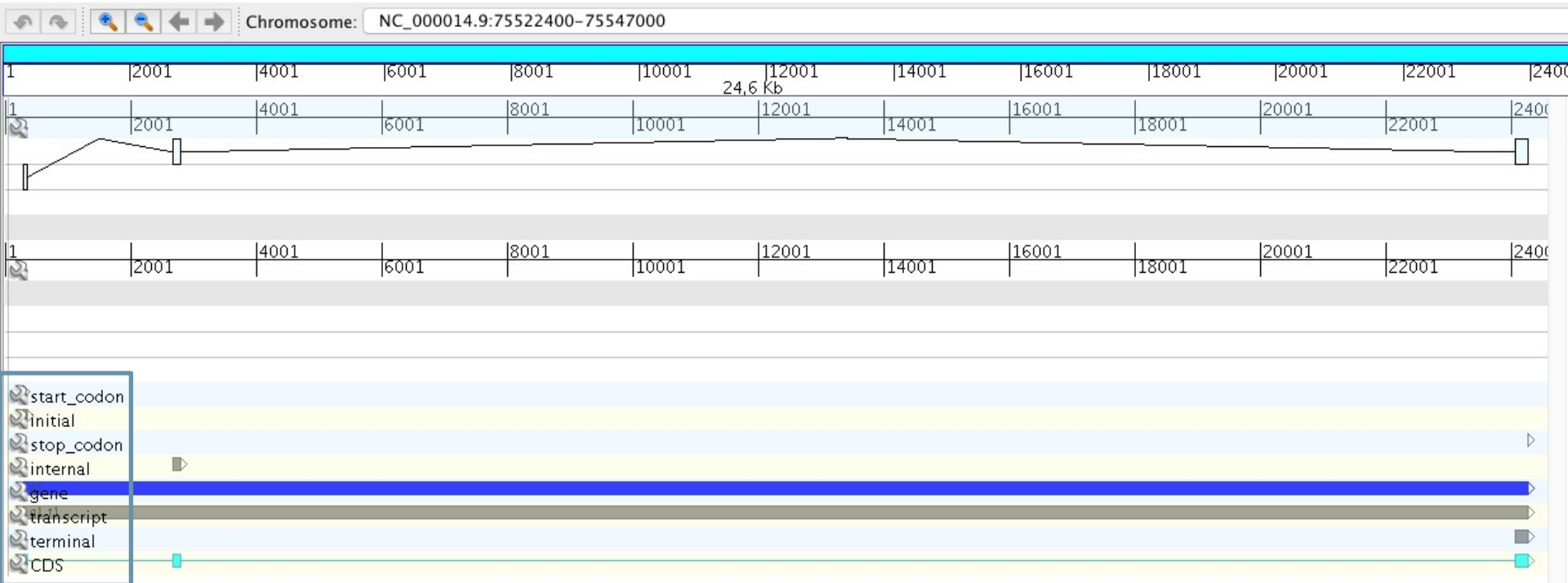
```
TTTCCGCCCATGTGACTTCCAGCGTGAGTTACCAGAAACCACAAGAGAGAGAGAGCGTGCAGCCCCA
AAGCGAGCGACATGTCCCTTTGGGGAGCAGTCCCTCTGCACCCCAGAGTGAGGAGGACGCAGGGGTGAGA
GGTGGCTACAGGGCAGGCAGAGGAGGCACCTGTAGGGGGTGGTGGGCTGGTGGCCCAGGAGAAGTCAGGA
AGGGAGCCCAGCTGGTGACAAGAGAGCCCAGAGGTGCCTGGGGCTGAGTGTGAGAGCCCAGGAAAGATTCA
GCCATGCCTCACAGCTCCGACAGCAGTACTCCAGCTTCAGCCGCTCTCCTCCCCCTGGCAAACAGGTAG
AGTCCTCCTTTTCTCTCTCTACCTTCTGATTCTCCTGGGGGATGGAAAAGAGAGCCAGGCTTCTCTGTC
CTGCCCAGGGAGCTGAGGATGGAGGAAGTGGCTCGTTGCACGGGCACTCTGTTAGACTTAGGACATGGAA
TTTGCTACTAAGCTGTGCATATTGGCAGAGATCCTCATCCTTCCACCCATTCTGCCAAAGCCCTTTTTC
TCTCCATTTTCCAAGGCTGCCTATCACCTCTGCCTCACTGGGGTTGCCACCCTAAAAAGCTTTCTAGGAA
CAAAGAGGAGGATGAACATCAAAGAATGCAGAGAAAAGAGTCTACTGTTCTCCAAGGCTGTAGAAAAGT
```

Sequence
in FASTA
forma

NC_000014	AUGUSTUS	gene	284	24272	0.89	+	.	Gene	g1
NC_000014	AUGUSTUS	mRNA	284	24272	0.89	+	.	mRNA	g1.t1
NC_000014	AUGUSTUS	start_codon	284	286	.	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	initial	284	346	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	internal	2685	2789	0.99	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	terminal	24063	24272	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	CDS	284	346	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	CDS	2685	2789	0.99	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	CDS	24063	24272	1	+	0	mRNA	g1.t1
NC_000014	AUGUSTUS	stop_codon	24270	24272	.	+	0	mRNA	g1.t1

Gene prediction by
AUGUSTUS in GFF3
format

Visualising annotation data in GenomeView



Different tracks

References

- Annotation of **prokaryotic** genomes
 - Richardson EJ, Watson M. The automatic annotation of bacterial genomes. *Brief Bioinform.* 2013 Jan;14(1):1-12.
- Annotation of **eukaryotic** genomes
 - Stein L. Genome annotation: from sequence to biology. *Nat Rev Genet.* 2001 Jul;2(7):493-503.
 - Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet.* 2012 Apr 18;13(5):329-42.
 - Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. *Nat Rev Genet.* 2016 Dec;17(12):758-772.

Sylvain Legrand
Maître de Conférences
UMR CNRS 8198 EVO-ECO-PALEO
Evolution, Ecologie et Paléontologie
Université de Lille – Faculté des Sciences et Technologies
Bât SN2, bureau 208 - 59655 Villeneuve d'Ascq

sylvain.legrand@univ-lille.fr | www.univ-lille.fr
Tél. +33 (0)3 20 43 40 16