Université de Lille

Prédiction de gènes

Contenus basés sur les cours des enseignants en bioinformatique de l'université de Lille

Sylvain.legrand@univ-lille.fr

Introduction

Problématique

- La **baisse** constante des **couts de séquençage** permet d'obtenir de plus en plus facilement le génome d'une espèce

- Cependant, à bien des égards, l'**annotation** de génomes est devenue **plus difficile** !

  - Les **short reads** des séquenceurs de nouvelle génération (Illumina) ne permet pas d'obtenir la qualité d'assemblage des premiers génomes (drosophile, homme, arabidopsis…)

  - Projets de séquençage de génomes avec des **caractéristiques inhabituelles** et sans données préalables

  - Les projets de séquençage de génomes se font maintenant « à la maison », par des biologistes qui possèdent parfois **peu de compétences en bioinformatique**

Problématique

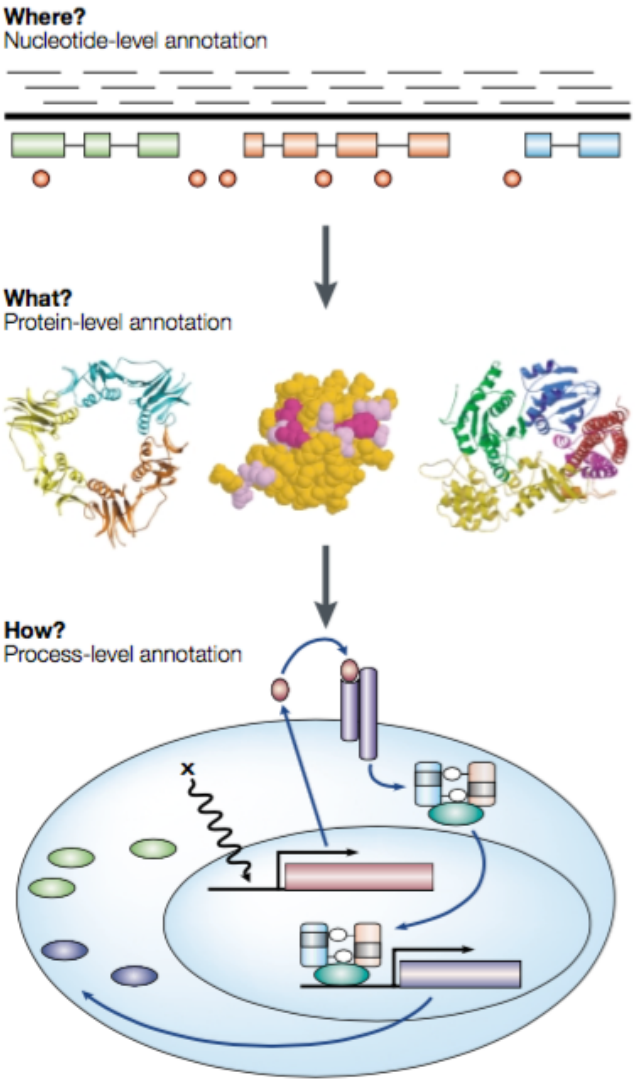


Figure 1 | **The three layers of genome annotation: where, what and how?**

- **Objectif du module : prédiction de gènes**
  - contexte
  - principales méthodes
  - application aux bactéries
  - application aux eucaryotes
- **Pour aller plus loin : annotation de protéines**
  - contexte
  - prédiction de la fonction
  - prédiction de la localisation cellulaire
  - étude des structures 2D et 3D

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

Prédiction de gènes

Qualité de l'assemblage

- La première étape consiste à **valider l'assemblage** obtenu

- Observer les **métriques** (N50, L50..)

	A. halleri halleri	*A. halleri gemmifera*	*A.lyrata*	*A. thaliana*
Nb scaffolds	3 152	2 239	695	7
Total length	174 Mb	196 Mb	207 Mb	120 Mb
Genome cov.	68.3 %	76.9 %	89.9 %	88.9 %
Longest scaff.	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.

Qualité de l'assemblage

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.

Qualité de l'assemblage

- BUSCO http://busco.ezlab.org/
- Rechercher dans l'assemblage des **gènes simples copies, universels**

	A. halleri halleri	*A. halleri gemmifera*	*A.lyrata*	*A. thaliana*
Complete universal single-copy orthologs	95.3%	97.6%	98.5%	98.2%
Fragmented universal single-copy orthologs	1.5%	0.3%	0.3%	0.5%
Missing universal single-copy orthologs	3.2%	2.1%	1.2%	1.3%

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

Identification des séquences répétées

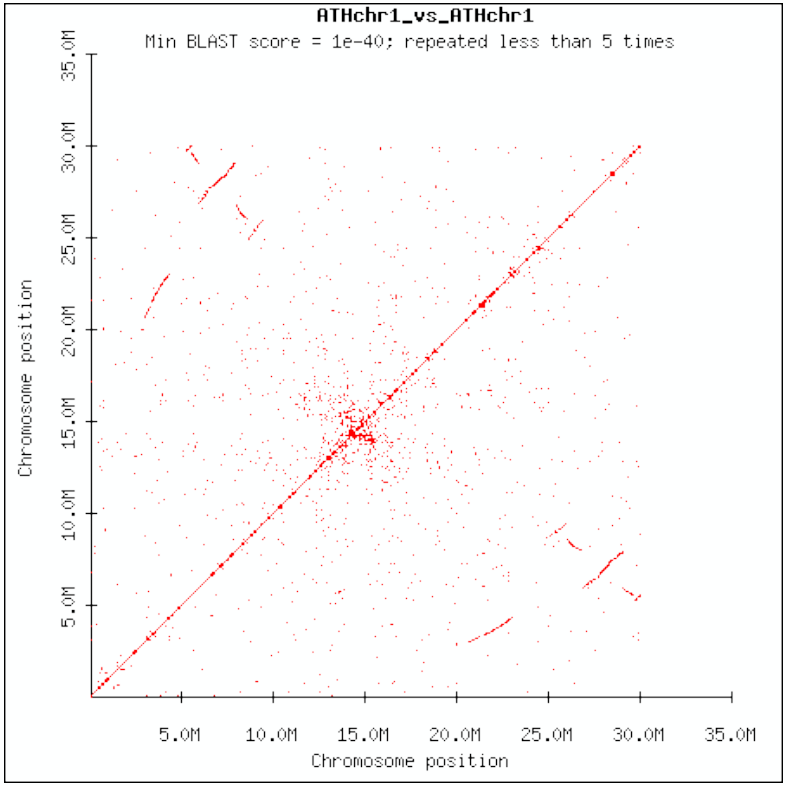
- Les génomes **eucaryotes** peuvent être **riches** en séquences répétées : 47% du génome humain, seul 1-2% du génome est codant !

- A première vue, le **génome humain** semble être un modèle d'**inefficacité** : gènes séparés par de larges régions (10-100 kb), introns

- Chez la **levure** : 60% du génome code les 6000 protéines. Les 35000 gènes humains sont codés par un génome 300 x plus grands

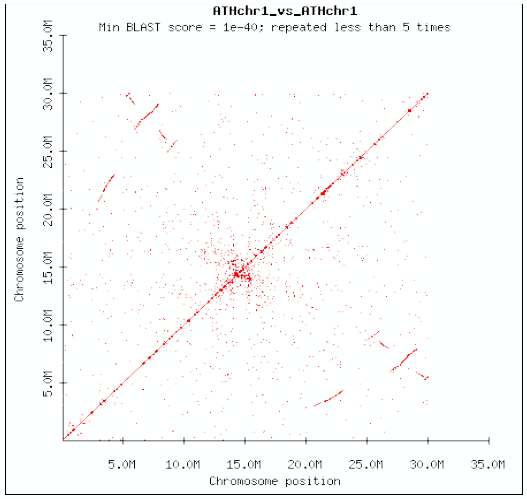
Identification des séquences répétées

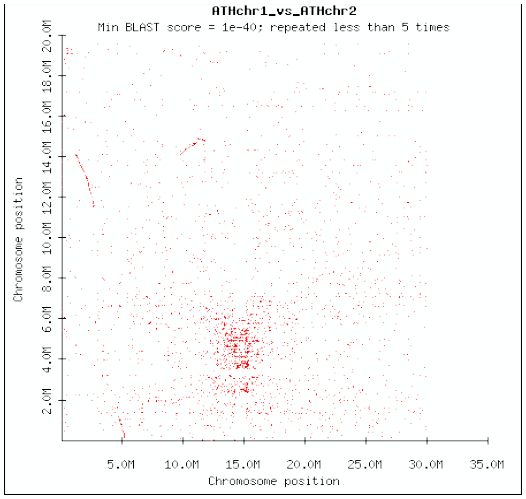
- **Dotplot** Chr1 *Arabidopsis thaliana* aligné contre lui-même
- On observe de **nombreuses séquences répétées**
- Il y a d'avantage de répétitions au niveau du **centromère** chez *At*

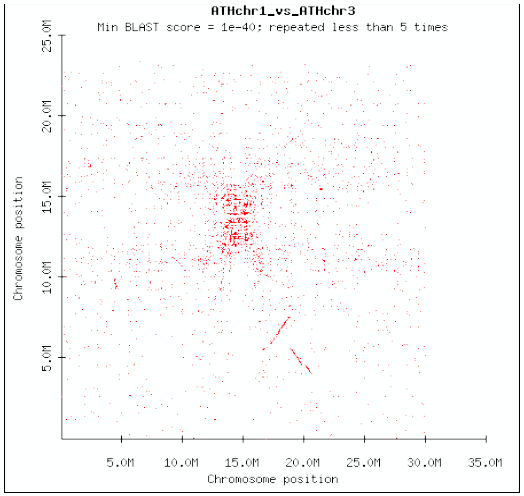


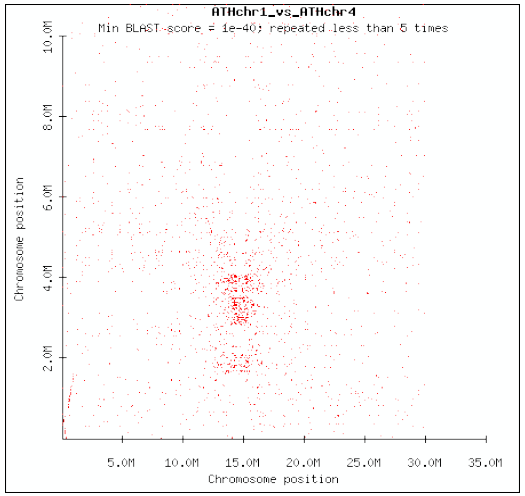
http://biolinx.bios.niu.edu/t80maj1/rice/arab_mega_dotplots.htm

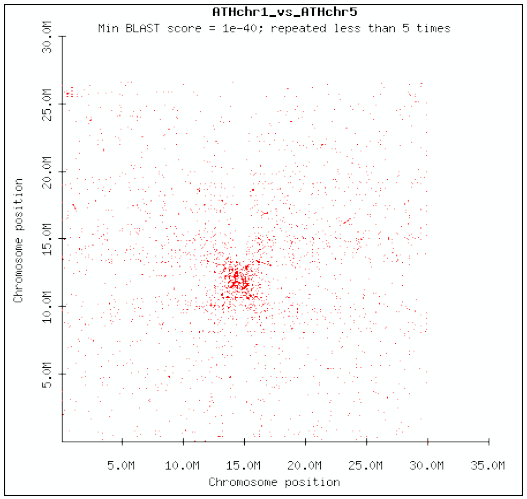
Identification des séquences répétées







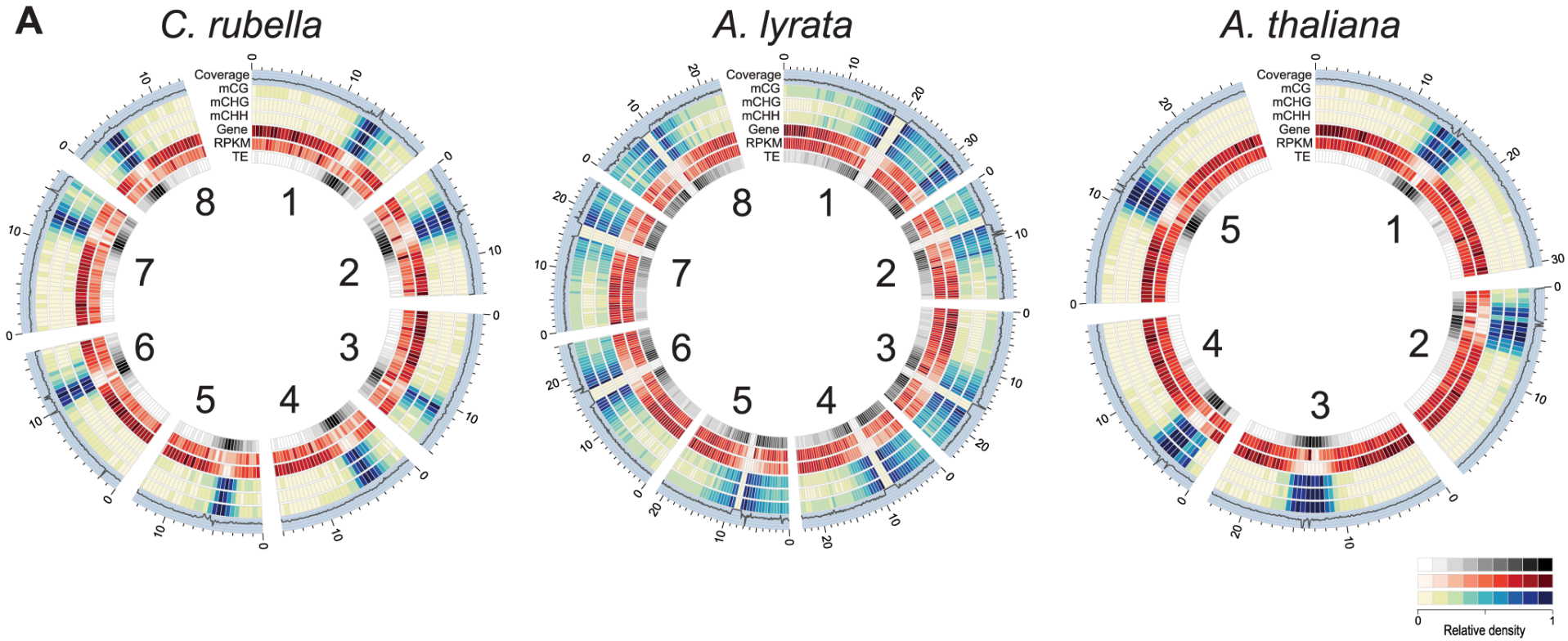




http://biolinx.bios.niu.edu/t80maj1/rice/arab_mega_dotplots.htm

Identification des séquences répétées

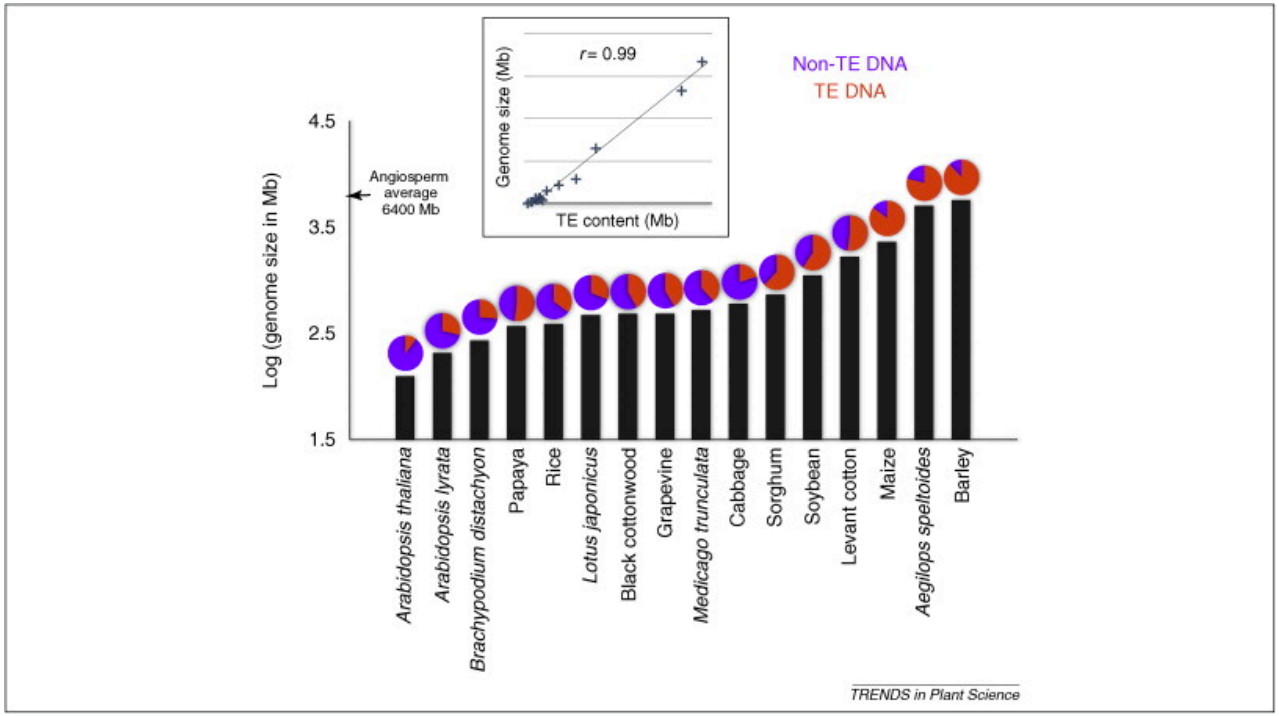
- Les éléments transposables peuvent avoir **différentes localisations** dans le génome



Seymour et al., Plos Genet 2014

Identification des séquences répétées

- La **taille** du génome est **corrélée** au contenu en **éléments transposables**



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 des séquences répétées

- **Les séquences répétées gênent l'assemblage** du génome **et la prédiction des gènes** : en effet les ORF des éléments transposables sont identifiés comme des gènes de l'organisme hôte. Peuvent aussi produire des erreurs dans l'annotation des gènes voisins

- L'identification des séquences répétées et leur masquage sont généralement les **premières étapes de l'annotation** d'un génome (eucaryote)

- **Masquage** : remplacer ces régions par des « N » ou par des lettres minuscules (softmasking)

Identification des séquences répétées

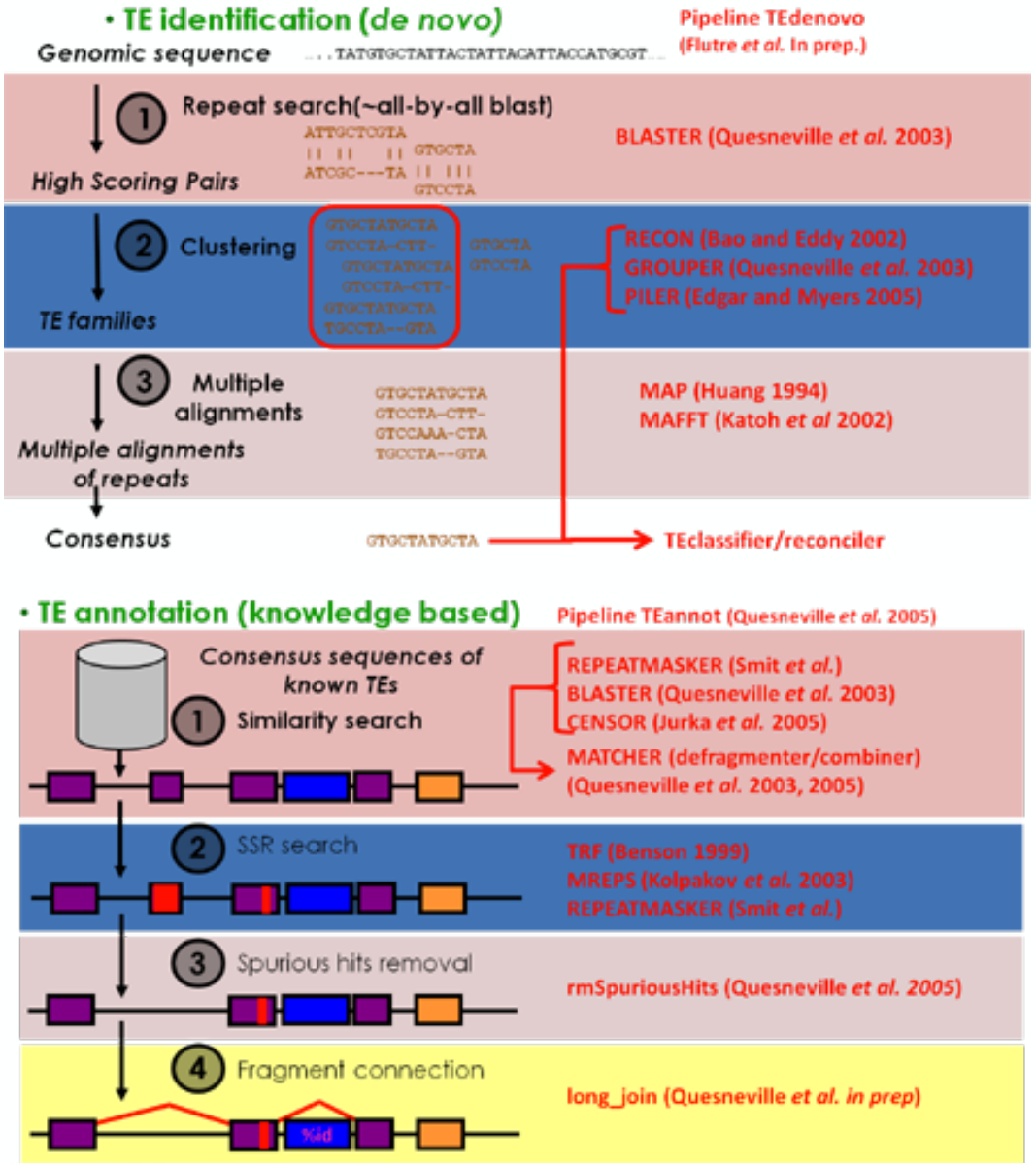
- **Deux types d'analyse** : basée sur **homologie** ou ***de novo***

Les éléments transposables étant peu conservés d'une espèce à l'autre, l'analyse *de novo* présente l'avantage de pouvoir identifier des familles spécifiques d'éléments

- Une fois une banque d'éléments transposable obtenue, les éléments peuvent être **identifiés** à l'aide d'outils tels que RepeatMasker, Crossmatch… Il est possible également de combiner différents outils

- En plus des éléments transposables, les répétitions identifiées peuvent englober également les **régions de faibles complexités** et les **gènes répétés** : histones, tubulines…

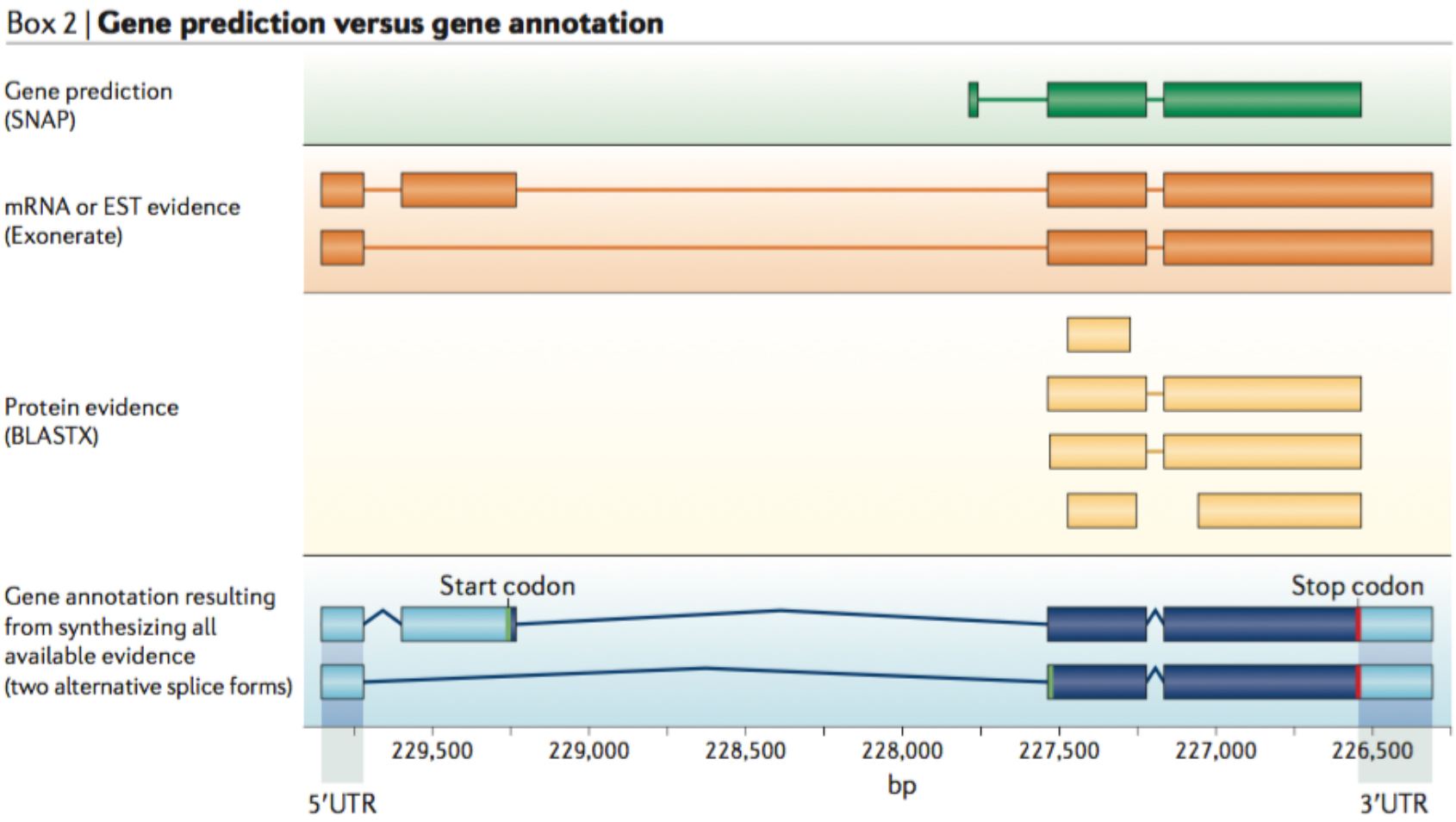
Identification des séquences répétées



The REPET package (URGI)

Flutre T. et al, 2011

Annotation des gènes



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

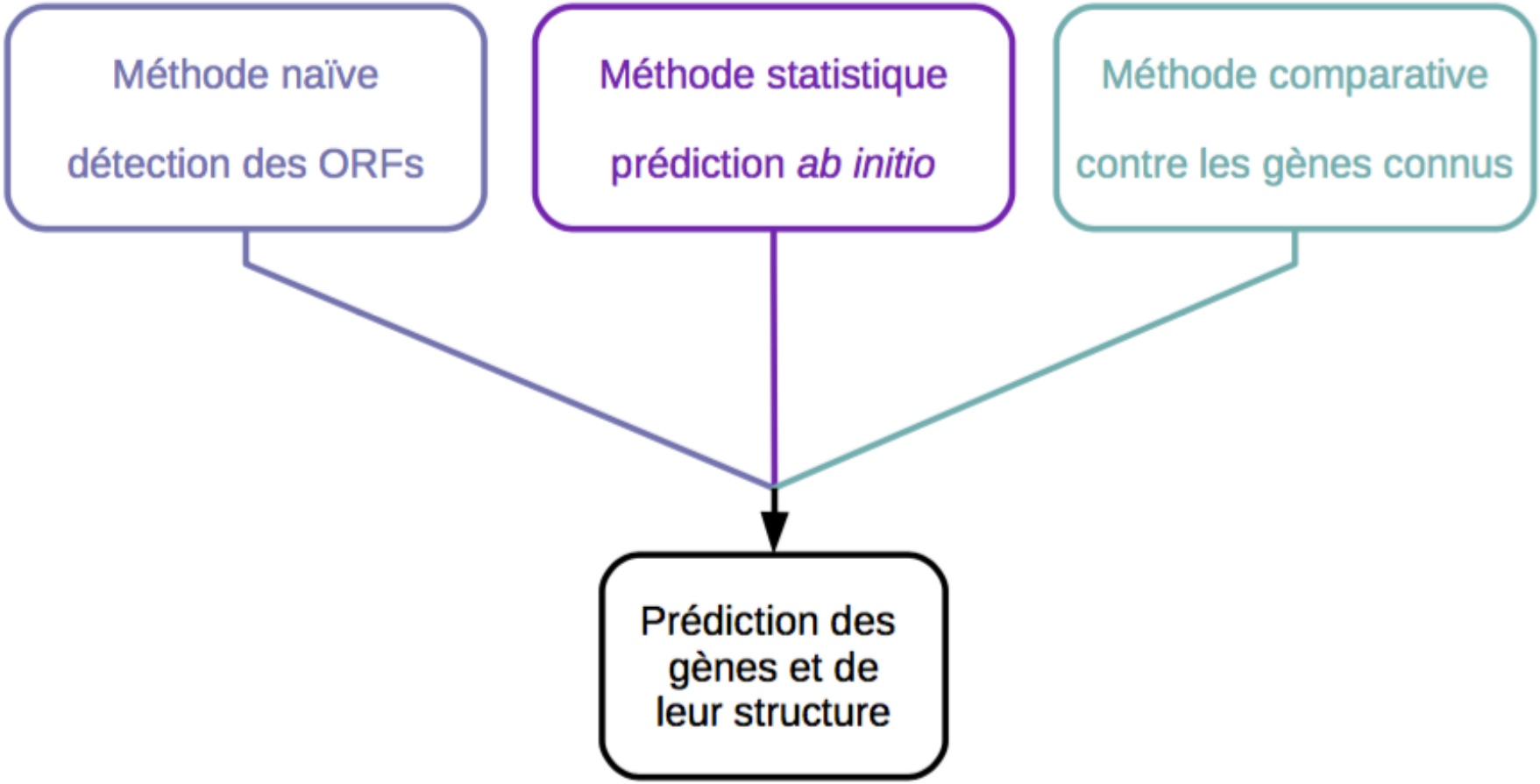
Annotation des gènes

- **Point de départ** : séquences nucléiques brutes

- **Obtenu à l'arrivée**
  - Positions de début et de fin des gènes
  - Signaux de transcription, épissage et traduction
  - Idée de la fonction des protéines codées par les gènes

- **Limites**
  - Certains gènes ne sont pas prédits (faux négatifs)
  - Certains gènes prédits ne sont pas de vrais gènes (faux positifs)
  - Les limites précises des gènes sont parfois erronées (choix du mauvais codon d'initiation…)

Principales méthodes



Méthode naïve

- **Principe** : recherche les signaux des séquences codantes
  - Débutent par un codon d'initiation ATG + autres
  - Se terminent par un codon de terminaison TAA, TAG ou TGA
  - Possèdent une taille multiple de 3 Cas des gènes sans intron

- **Mise en œuvre** : détecter les phases ouvertes de lecture
  - ORFs = Open Reading Frames
  - Phases (cadres) pouvant contenir un gène
  - >50 nt entre un codon d'initiation et un codon de terminaison
  - Traduction à l'aveugle dans les 6 phases de lecture → 3 phases par brin d'ADN

Méthode naïve

- Les **6 phases de lectures** d'une séquence nucléique

Méthode naïve

- **Avantages**
  - méthode *ab initio* : sans connaissances préalables
  - diminue la quantité de données à analyser pour la comparaison de séquences
- **Limites**
  - toutes les ORFs ne sont pas des gènes
  - sensible aux erreurs de séquençage
  - peu utile pour les gènes eucaryotes (présence d'introns)

- **Principe** : Discriminent les séquences codantes des non codantes
  - Se basent sur les biais d'usage du code

- **Mise en œuvre** :
  - Apprentissage de l'usage du code pour un organisme donné à partir de séquences codantes fiables
  - Calcul de la probabilité pour qu'une portion de séquence soit codante
  - Analyse des signaux de transcription et traduction pour déterminer les limites des gènes

Biais d'usage du code génétique

- 1 acide aminé est codé par N codons
  - codons synonymes
- Répartition non uniforme des codons utilisés

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

- **Avantages**
  - Méthode *ab initio* : sans annotation préalable de gènes de la même famille
  - Critères plus fiables que la méthode naïve

- **Limites**
  - Besoin d'un jeu de données d'apprentissage
    Séquences codantes confirmées
  - Ne détecte pas les petits gènes/exons
    En-dessous du seuil de détection
  - Identification CDS uniquement, pas UTRs
  - Pas d'identification d'épissage alternatif

- Certains outils tels que TwinScan, FGENESH, Augustus, Gnomon, GAZE et SNAP, peuvent utiliser des évidences (ARNm, protéines) pour améliorer les prédictions→ evidence-driven (en comparaison de *ab initio*)

Méthodes de comparaison

- **Principe** : cherche à localiser les annotations des banques sur la séquence à annoter
  - Alignements contre les protéines connues
  Localisation des CDS, y compris avec introns
  - Alignements contre les ARNm (EST, cDNA...)
  Localisation des CDS + UTR, y compris avec introns

- **Mise en œuvre**
  - Comparaison de la séquence contre les banques d'ARNm ou de protéines
  - Alignement des ARNm ou protéines extraites à l'aide de logiciels spécialisés

Comparaison aux ARNm

- **Comparaison** de la séquence nucléique **aux banques nucléiques** à l'aide de BLASTN (ou équivalent)
  - Détection de séquences contaminantes (vecteurs...)
    blast spécialisé : VecScreen
  - Détection d'ARNm potentiellement issus de la séquence
    Comparaison aux ARNm de l'organisme étudié ou d'organismes proches

- **Alignement des ARNm** identifiés à l'aide de **logiciels spécialisés**
  - Détermination fine des régions 5' et 3’ UTR et des exons
  - Logiciels : EST2genome, Splign

Utilisation des données de RNA-seq

- Ce sont les données qui ont le plus fort potentiel pour améliorer l'annotation

- Permettent de mieux délimiter les exons, les sites d'épissage, et les évènements d'épissage alternatif

- Mais grande quantité de donnée, complexe car souvent lectures courtes Illumina

- 2 façons de les utiliser les lectures
  - Assemblage *de novo* des lectures, indépendamment du génome, (ABySS, SOAPdenovo, Trinity). Les transcrits obtenus sont ensuite alignés sur le génome de la même façon que vu pour les ARNm
  - Directement alignées sur le génome (TopHat, GSNAP, Scripture), puis les alignements sont assemblés à l'aide de Cufflink

Comparaison aux protéines

- **Comparaison** de la **séquence nucléique traduite** dans les 6 phases aux **banques protéiques** à l'aide de BLASTX
  - Détection de protéines potentiellement codées par la séquence

- **Alignement des protéines identifiées** à l'aide de logiciels spécialisés
  - Détermination du codon d'initiation et des jonctions introns/exons
  - Logiciel : GeneWise

- **Avantages**
  - Valide des gènes potentiels par comparaison aux données expérimentales (ARNm séquencés, protéines étudiées)
  - Donne des indices sur la fonction de la protéine

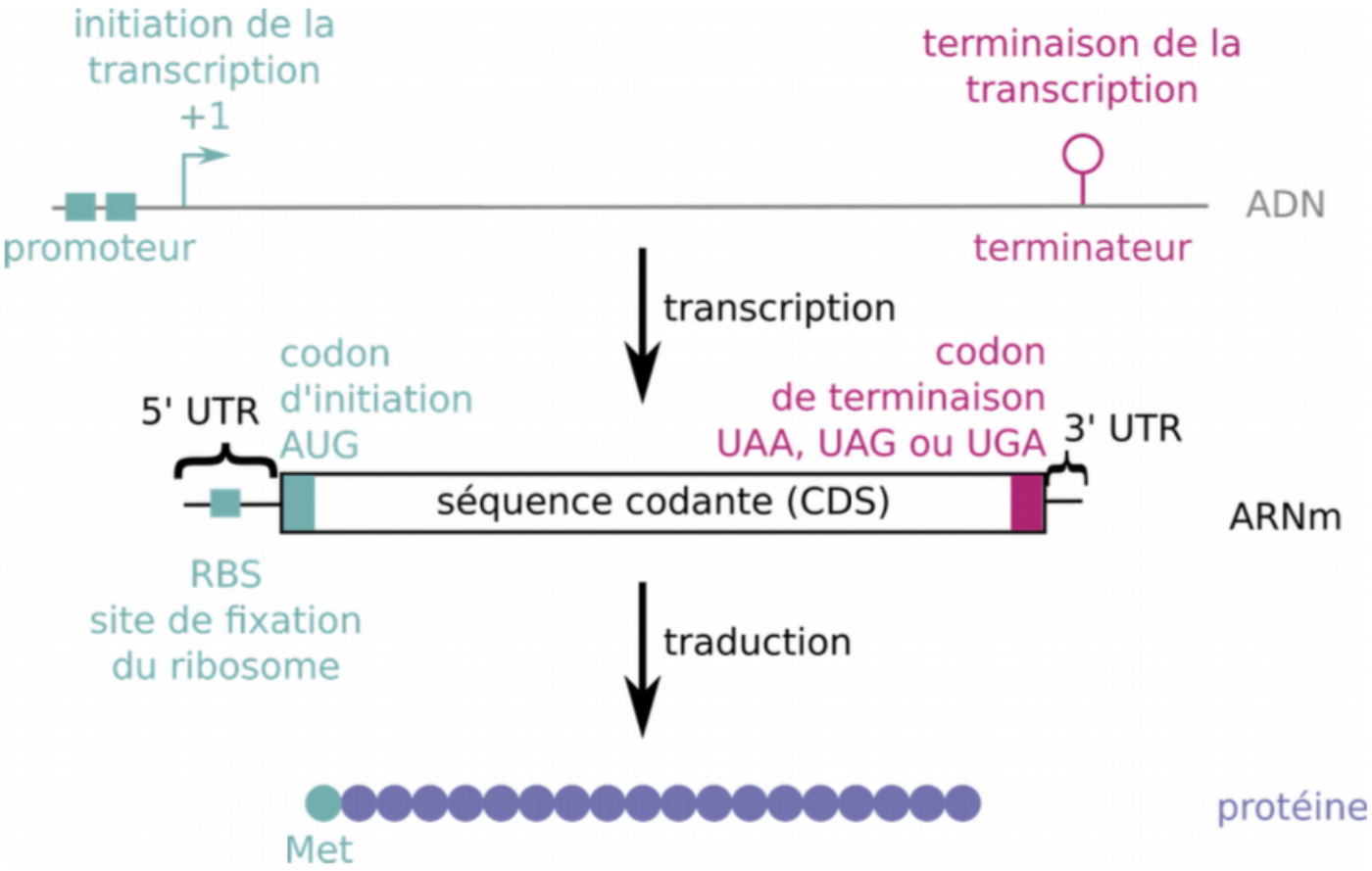
- **Limites**
  - Nécessite des connaissances a priori
  Ne trouve pas les séquences orphelines
  Difficile avec les génomes isolés d'un point de vue taxonomique
  - Propage les erreurs présentes dans les banques

Structure des gènes procaryotes

- Plus de **80% du génome est codant**
  - Séquences intergéniques courtes
  - En moyenne : un gène pour 1.000 nucléotides (kb)

- **Structure simple** des gènes
  - Régions transcrites mais non traduites (3' et 5' UTR) courtes
  - Pas d'intron (sauf exception)

Structure des gènes procaryotes



Structure des gènes procaryotes

- Voici un **extrait du génome** de la bactérie *Pseudoalteromonas sp.*

>AB057417
aacgaaaagattaaaaatttatcatttttctcttggaatttttactctacccccatta
atgaatgcaaattagaaaagctttttctgtactgttcagaaactgttaggagaactaaa
aaacatgaacattcgtcctttacaagatcgcgtaatcgttaaacgtctagaagaagaaac
aaaatctgctggcggtattgtattaactggctctgcagctgaaaaatcaactcgcggaga
agtagtagccgtaggtaatggtcgtattttagataacggtgacgttagagctttagaagt
aaaagccggtgacactgtgttatttggctcatatgttgagaaaactgaaaagatcgaagg
tcaagagtacctgatcatgcgtgaagacaacattttaggcattgtaggctaagcctactt
ttcgtttaacacacatttaagaatttagagg

Workflow proposé

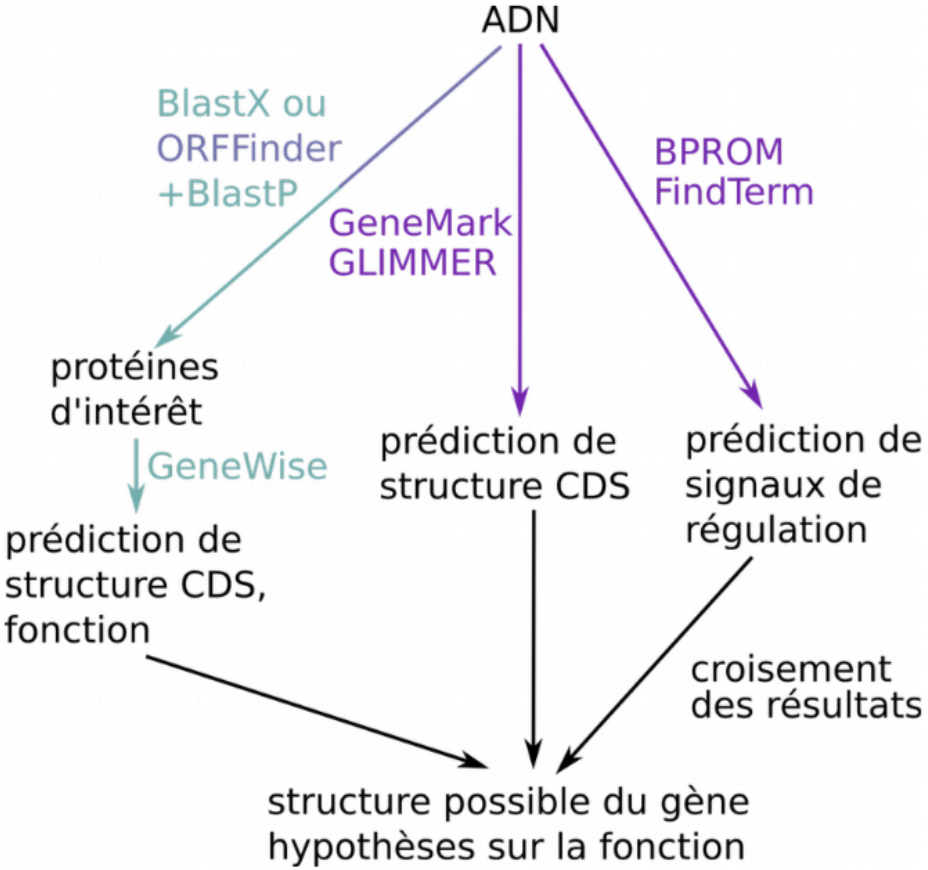
- **4 étapes** d'analyse :

1. Identification des ORF
   - ORFFinder

2. Validation des ORF
   - SmartBlast (GeneWise si besoin)

3. Prédiction statistique des CDS
   - GeneMark, GLIMMER

4. Prédiction statistique des signaux de régulation
   - BPROM



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_011604 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): 75

Genetic code: 1. Standard

ORF start codon to use:

- "ATG" only
- "ATG" and alternative initiation codons
- Any sense codon

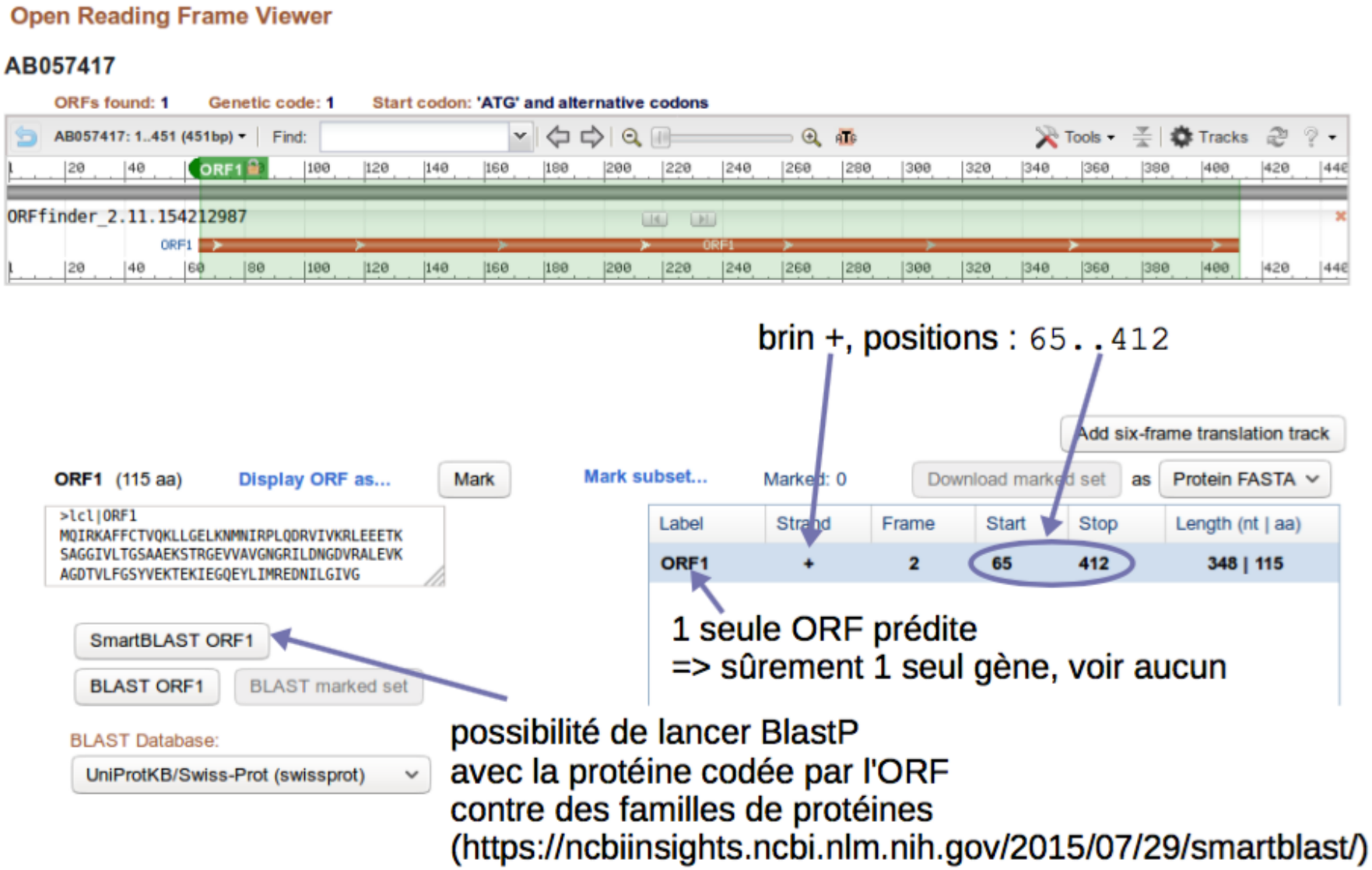
Ignore nested ORFs:

Start Search / Clear

Submit Clear

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

Résultats d'ORFfinder



SmartBlast

- **SmartBlast** compare l'ORF contre une banque (landmark database) constituée des protéomes de 27 espèces réparties sur une large phylogénie. Compare également contre la banque nr

https://blast.ncbi.nlm.nih.gov/smartblast/smartBlast.cgi?CMD=Web&PAGE_TYPE=BlastDocs#searchSets

- Il retourne les 5 meilleurs résultats obtenus contre la banque landmark

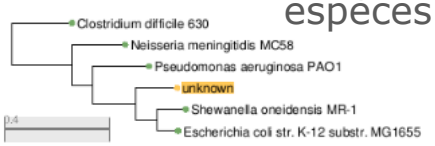
- Il retourne ensuite les résultats obtenus contre la banque nr

Résultats de SmartBlast

Summary — Report description

Query: unnamed protein product — **Query length:** 115 aa

Arbre des espèces

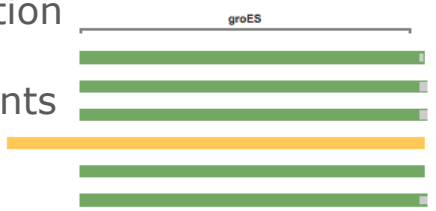


Domaines

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

Visualisation des alignements



About the database — See full multiple alignment — Legend

Descriptions

Best hits

5 meilleurs résultats contre banque landmark

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_418566.1
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

Résultats contre banque nr

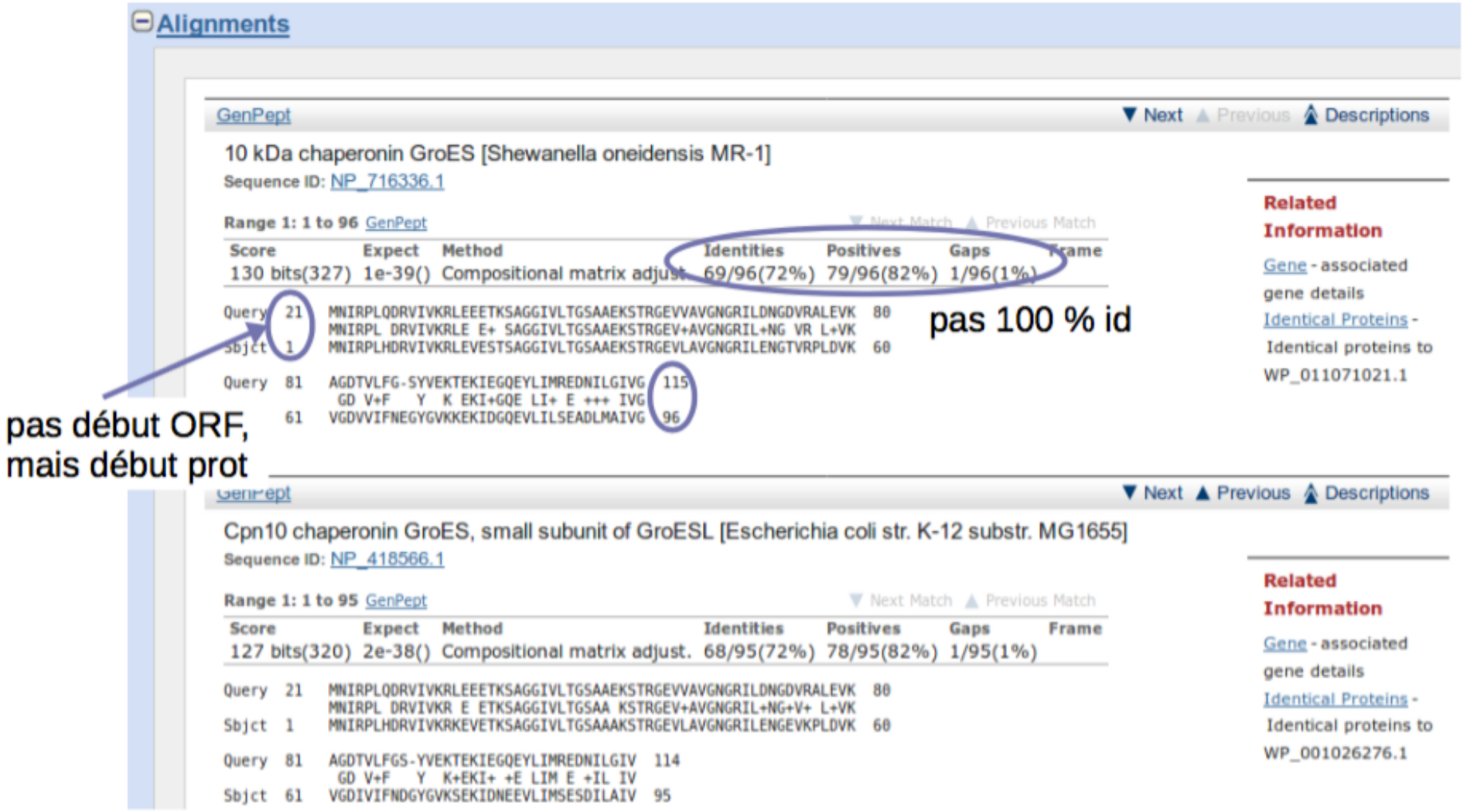
Select: All None Selected:0

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

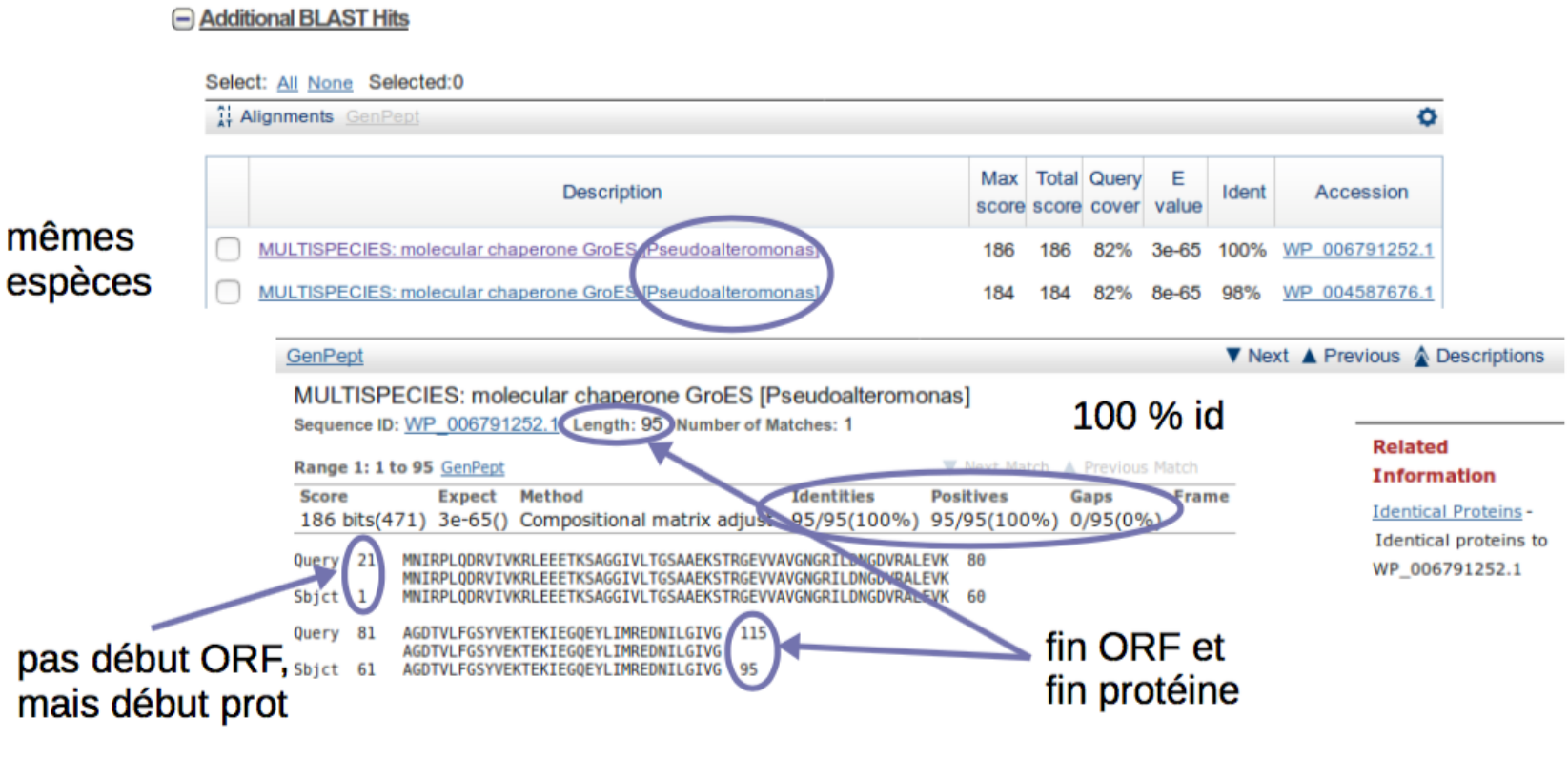
Résultats de SmartBlast

- Best hit obtenu contre banque landmark



Résultats de SmartBlast

- Best hit obtenu contre banque nr



Résultats de SmartBlast

- **ORF : 65..412** sur le brin + de la séquence ADN
  - Code une protéine de 115 aa + le codon de terminaison

- Alignements fournis par SmartBLAST
  - Query 21..115 : seulement une partie de la protéine de l'ORF
  Donc **l'ORF entière n'est pas codante**
  L'alignement débute en 21 => la CDS débute en 65+(21-1)*3 = 125
  Fin de la séquence codante en 412
  - Sbjct 1..95 : la protéine de la banque est entière
  La séquence codante prédite est complète

- Les alignements obtenus avec différentes séquences sont bons
  - La **prédiction est fiable**, pas besoin de GeneWise

GeneMark

A family of gene prediction programs developed at
Georgia Institute of Technology , 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/

Résultats de GeneMark

- **GeneMark.hmm with Heuristic models**
- Même résultat que ORFfinder, **en contradiction** avec début identifié par 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      Class
    #                                            Length
    1         +           65          412           348         1
```

Résultats de GeneMark

- **GeneMark.hmm** modèle adapté pour *Pseudoalteromonas sp.*
- Résultat **en accord** avec SmartBlast
- **Meilleur résultat avec un modèle** défini pour l'espèce étudiée

```
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      Class
    #                                             Length
    1         +          125          412           288         1
```

BPROM

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

BPROM - Prediction of bacterial promoters

BPROM 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
aacgaaaagattaaaaatttatcatttttttctcttggaatttttttactctacccccatta
atgaatgcaaattagaaaagctttttttctgtactgttcagaaactgttaggagaactaaa
```

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

Résultats de BPROM

```
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

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
```

proche des vraies positions
cf. page suivante

Bilan de l'analyse

- Les méthodes aboutissent à la **même conclusion** : la séquence contient **une seule CDS 125..412** (codon de terminaison compris)
- Information supplémentaire donnée par SmartBLAST : la CDS code une **chaperonne du type Cpn10 / GroES**
- Glimmer (prédiction statistique) ne trouve pas de CDS

Prédiction chez les bactéries : quelques pièges

- **Plusieurs Cinit** (AUG) sur la séquence : lequel prendre ?

- Possibilité de **Cinit alternatifs** (GUG, UUG)

  Confirmation par :
  - Présence de RBS (Ribosome Binding Site)
  - Comparaison (analyse comparative avec autres espèces)

  Prédiction statistique

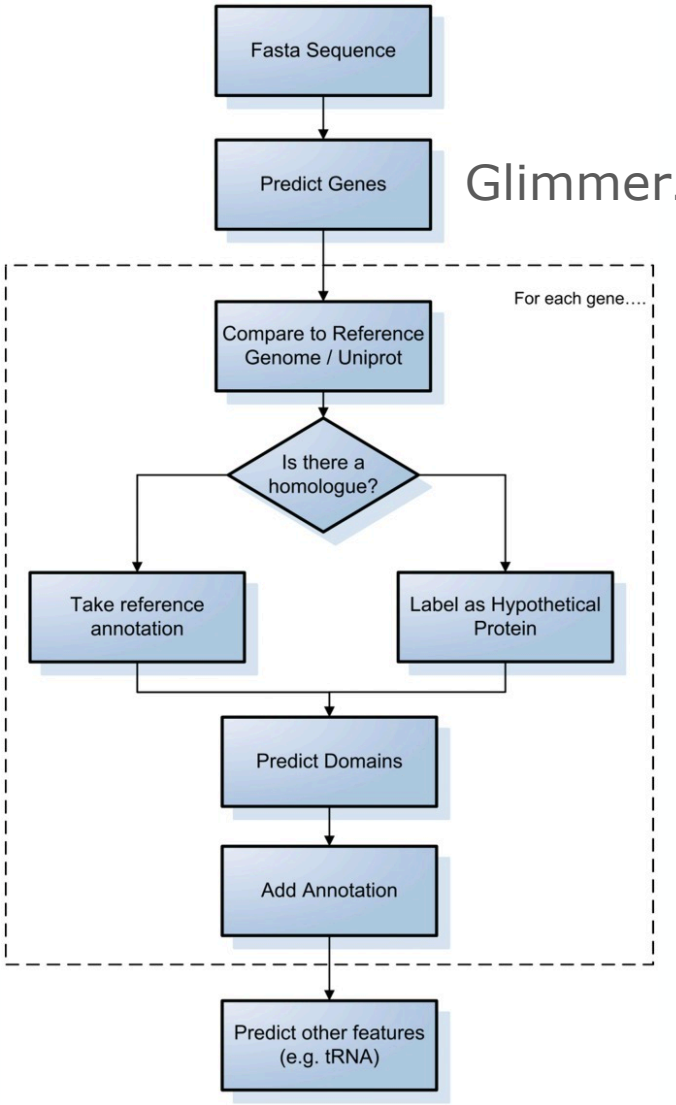
- **Gènes incomplets** (Cterm prématuré, décalage de phase)
  - Réel (corrigé lors de la traduction, pseudogènes)
  - Erreurs de séquençage
  - Détection par :

  BlastX signale des incohérences (phases différentes)

  Comparaison + Prédiction

- **Gènes chevauchants**
  - Fréquent chez les virus, quelquefois sur bacteries (fins de gènes)

Pipeline alternatif



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

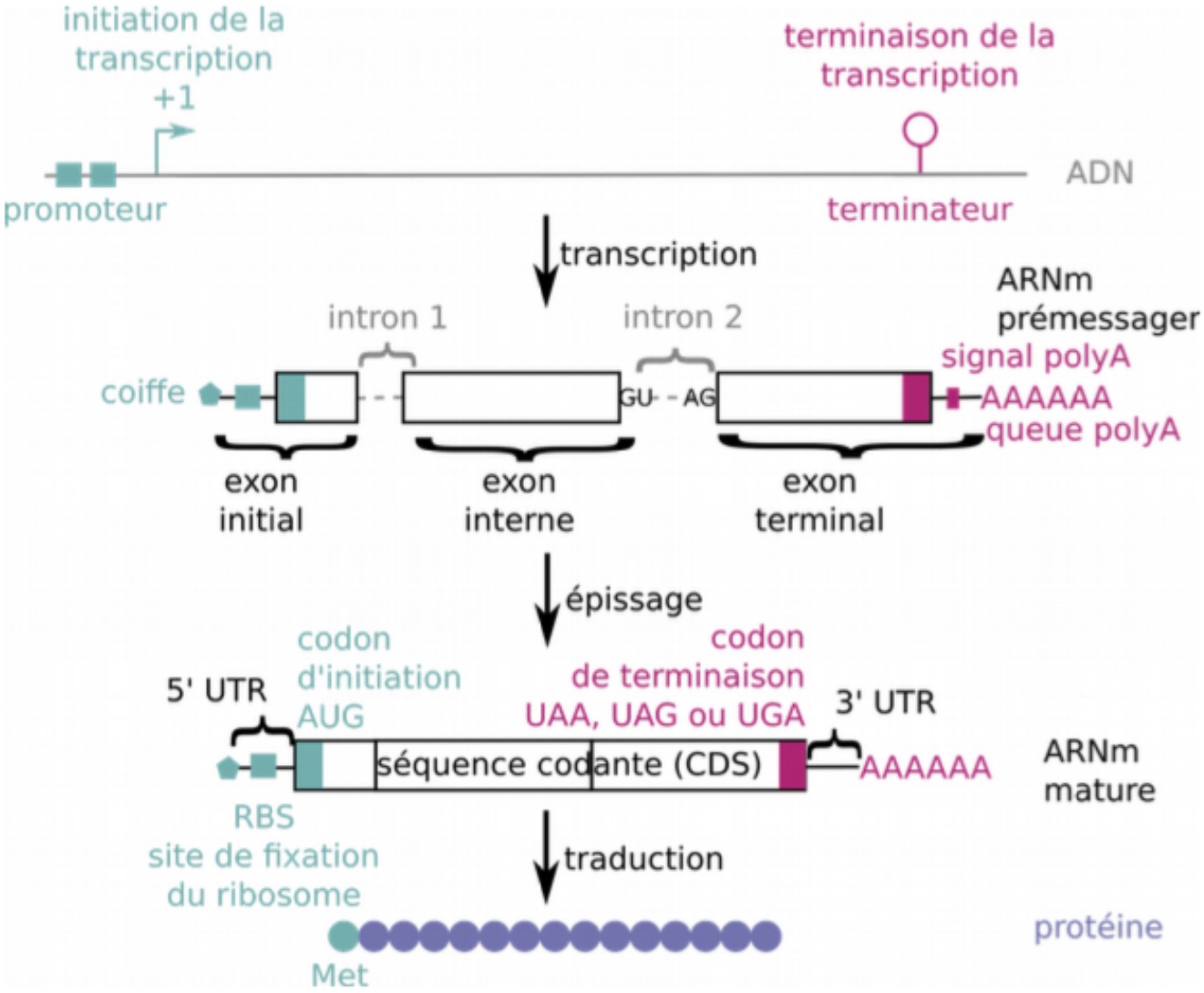
Structure des gènes eucaryotes

- **Faible pourcentage** de séquences **codant** pour des protéines
  Environ 2% du génome humain

- Présence de **séquences répétées** en très grande quantité
  ~ 50 % du génome humain

- **Structure complexe** des gènes
  - Longues régions 3' et 5' non traduites (exons non codants)
  - Présence d'introns, épissage alternatif

Structure des gènes eucaryotes



Conséquences des introns

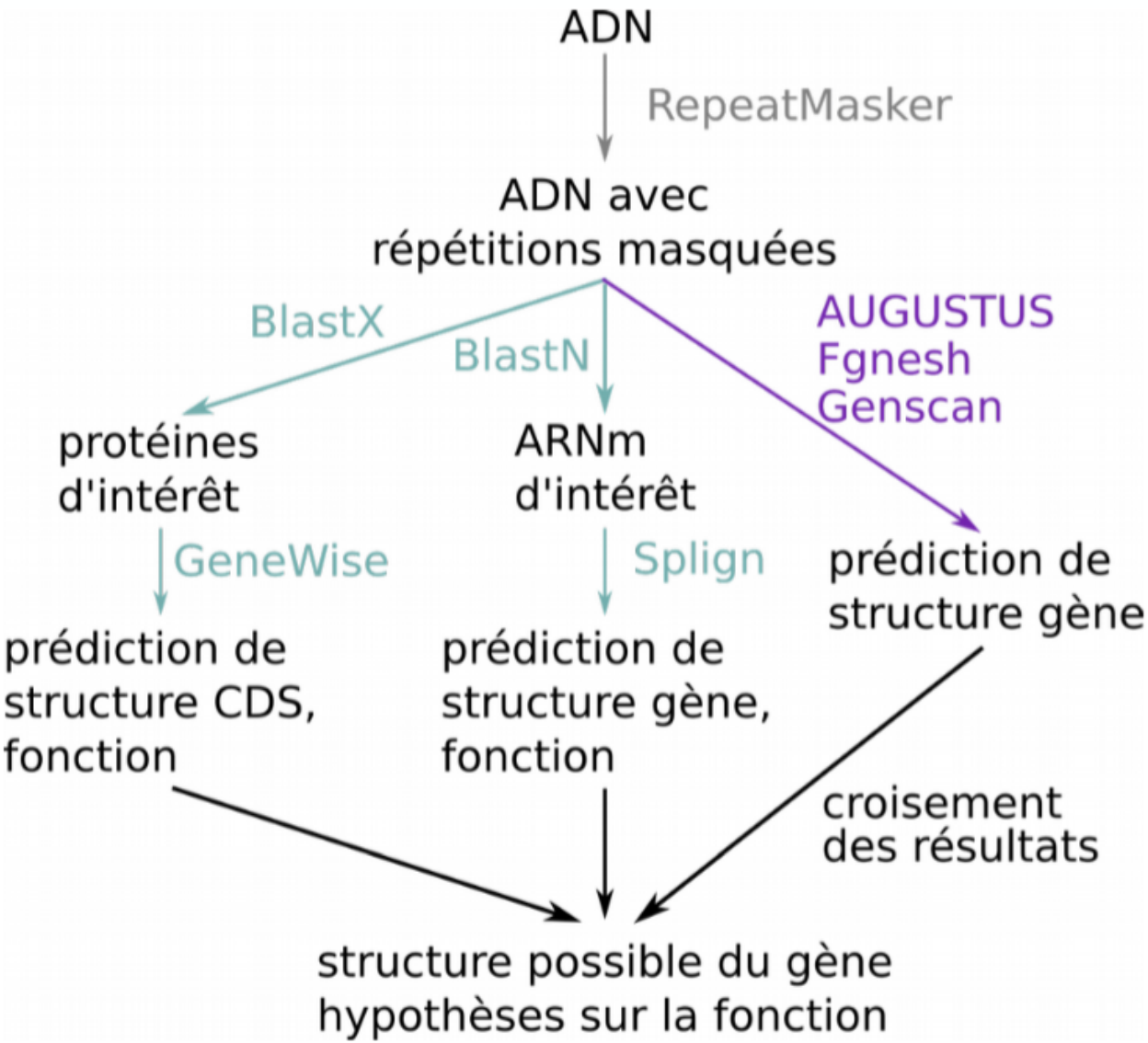
- Taille des **exons non multiple de 3**
  - Codons coupés par un intron
  - Changement de phase d'un exon à un autre
  - Pas de changement de brin

- Existence d'**exons courts** (~ 10 nt)
  En-dessus des limites de résolution des logiciels

- Existence d'**introns très longs** (> exons)
  Difficulté pour localiser les exons

- **Epissage alternatif**
  Concerne > 50 % des gènes humains

Workflow proposé



Exemple étude d'un ARNm

- **ARNm** de 905 bp, issu d'une cellule humaine

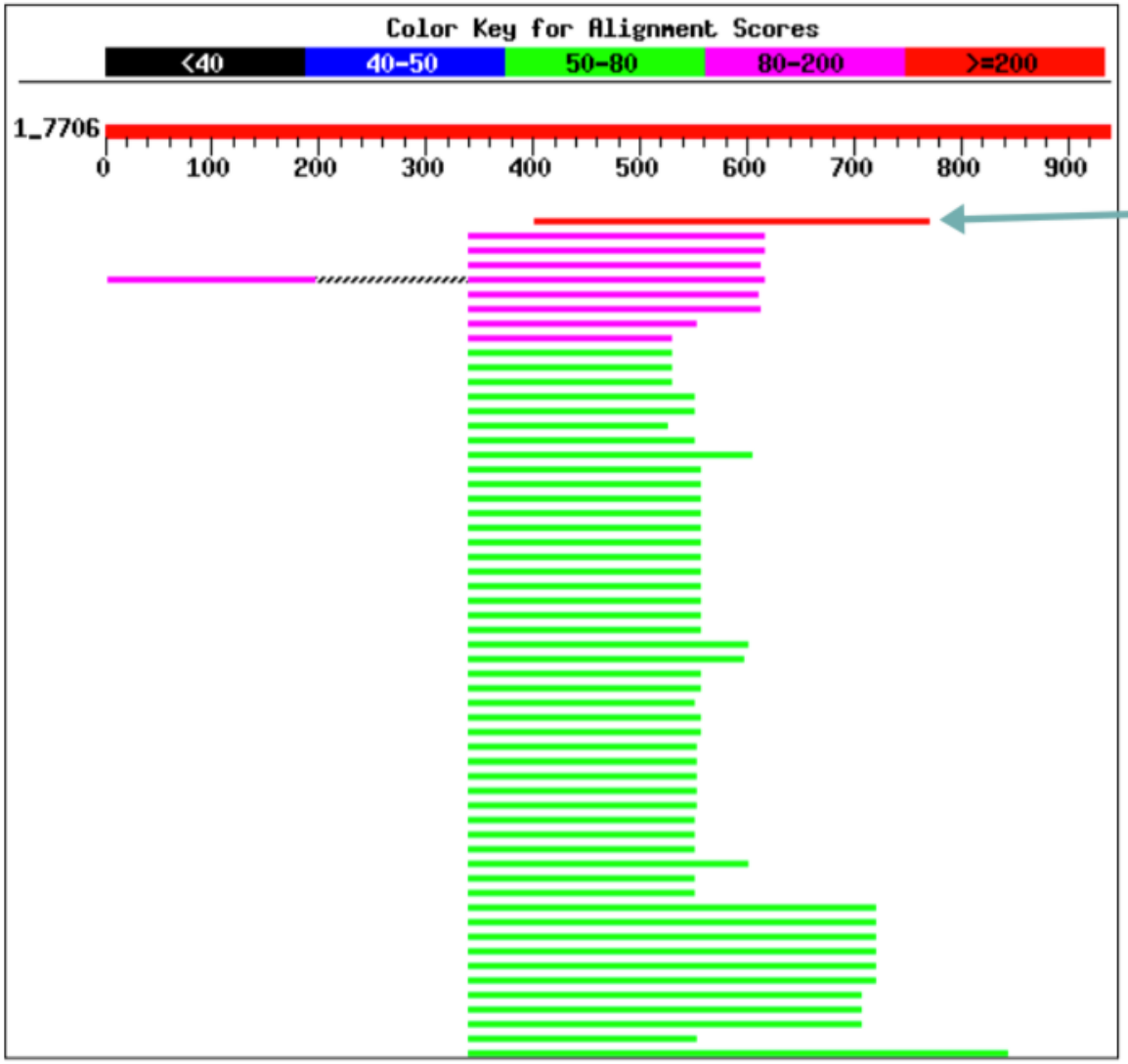
- **Trois étapes** d'analyse :

  1. Recherche de la CDS de l'ARNm
  BlastX + GeneWise

  2. Localisation du gène correspondant à l'ARNm
  Blast « Genomes » + Est2genomes ou Splign

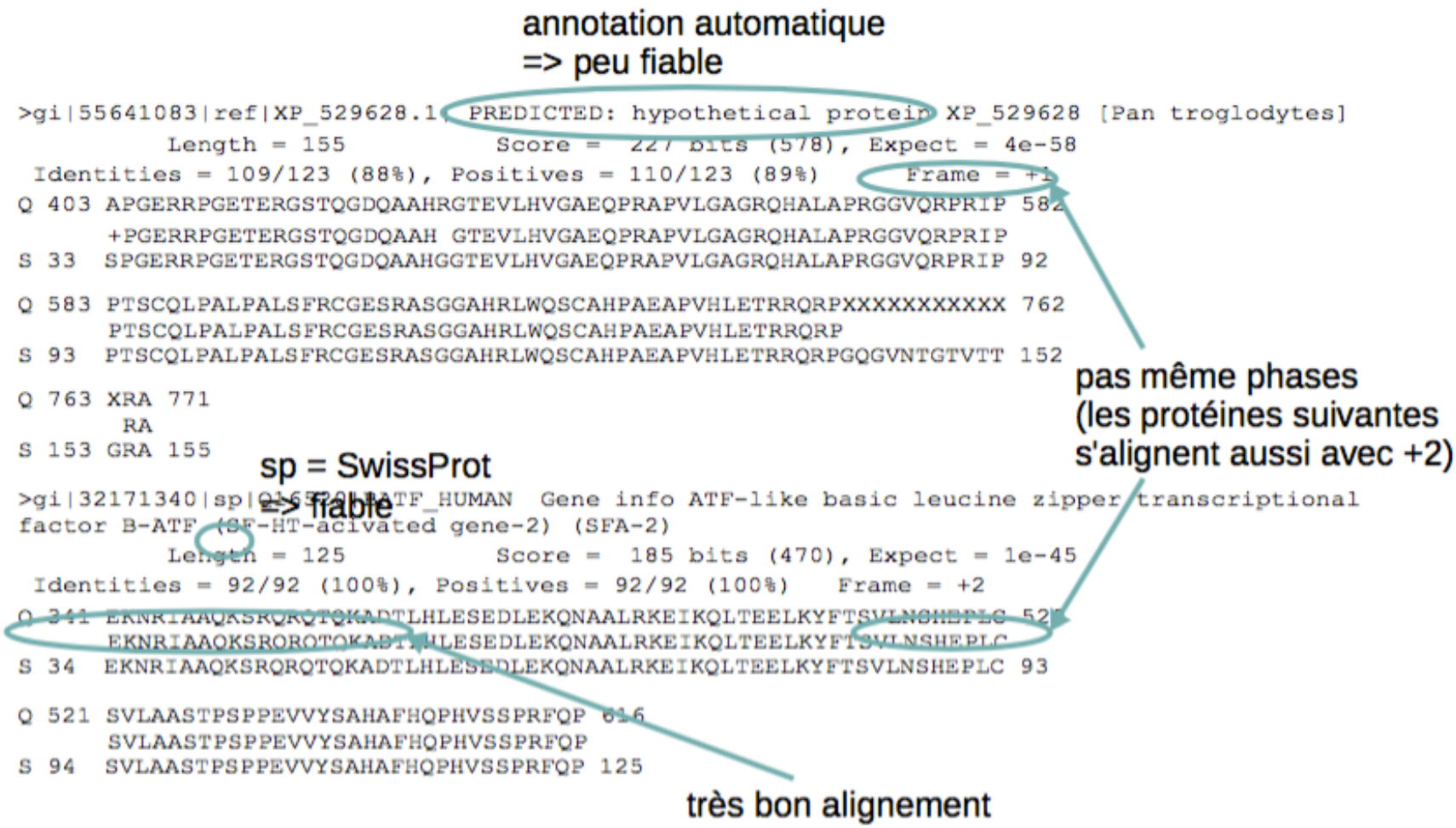
  3. Test des méthodes statistiques sur le génome
  FGENESH, AUGUSTUS, GeneScan

Résultat de Blastx, graphique



bon score mais seule protéine s'alignant avec cette région

Résultat de Blastx, alignements



Etude de l'alignement avec 2ème protéine (1ère non pertinente)

- BATF_HUMAN
  Protéine humaine, 100 % id => protéine d'intéret

- Frame = +2 :
  La séquence codante est sur le brin +

- Query 341..616 / Sbjct 34..125
  Il manque le début de la protéine de la banque
  Besoin d'un logiciel spécialisé pour aligner cette protéine à l'ARNm

- ATF-like basic leucine zipper transcriptional factor
  C'est peut-être un facteur de transcription du type bZIP

Wise

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: Choisissez un fichier Aucun fichier choisi

AND

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

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

Résultat de Wise

```
BATF_HUMAN          1 MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ
                      MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ
                      MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ
ARNm_hsp          243 accatgaagtatactcccgacgttgggaagcaagaacaggcaaccacac
                      tcagcaggacgtggccccgaaaccaatggtaggaaagtccaaggagaca
                      gtccccctcccccttctcagcatttgaatggggatttccggcagggag

BATF_HUMAN         50 KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA
                      KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA
                      KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA
ARNm_hsp          390 aggacccgaggcgacaggccagaaccaggcattatgcaacgccttgcgg
                      acactatagaataaaacctgaataatcaataatccttagaactgcttcc
                      gcccgcggcacggagcgtacggcggcagaggccggggcccgcgcgggcc

BATF_HUMAN         99 STPSPPEVVYSAHAFHQPHVSSPRFQP
                      STPSPPEVVYSAHAFHQPHVSSPRFQP
                      STPSPPEVVYSAHAFHQPHVSSPRFQP
ARNm_hsp          537 aactccgggtagcgtccccgatcctcc
                      gccccccattagcactaacatgccgtac
                      cgcgccgggccccaccattcccgccgc

FT          CDS     243..617
```

protéine d'intérêt

prot codée par ARNm

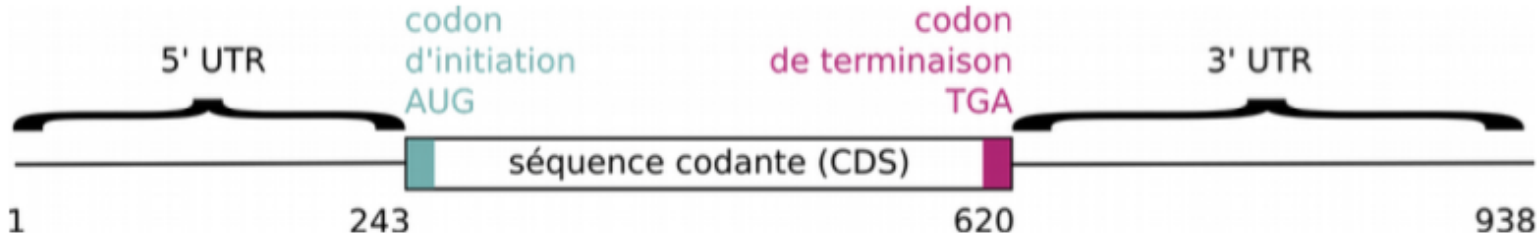
codons en colonne

début et fin de la CDS
sans le codon de terminaison

Résultat de Wise, bilan

- Comparaison avec la protéine d'intérêt (BATF_HUMAN)
  BlastX n'aligne pas la protéine entière avec l'ARNm car le début de la protéine contient une zone de faible complexité qui a été masquée par BlastX

- GeneWise donne une CDS en position 243..617+3 sur l'ARNm
  La protéine est alignée entièrement avec l'ARNm

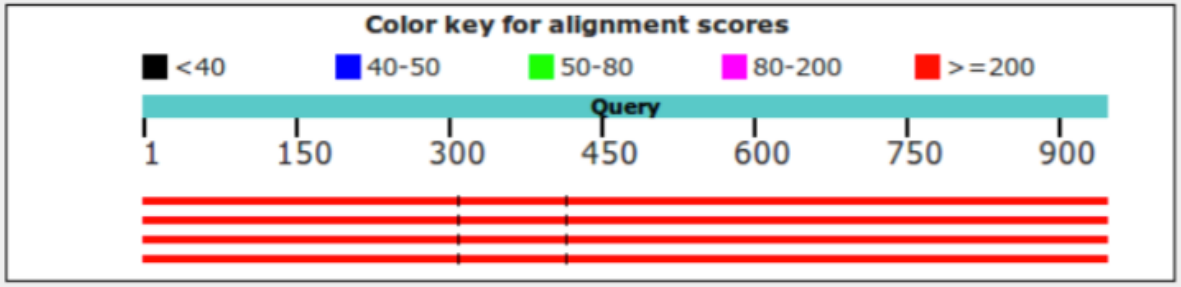


Résultat de Blastn contre le génome

Graphic Summary

Distribution of the top 12 Blast Hits on 4 subject sequences

Mouse over to see the title, click to show alignments



Descriptions

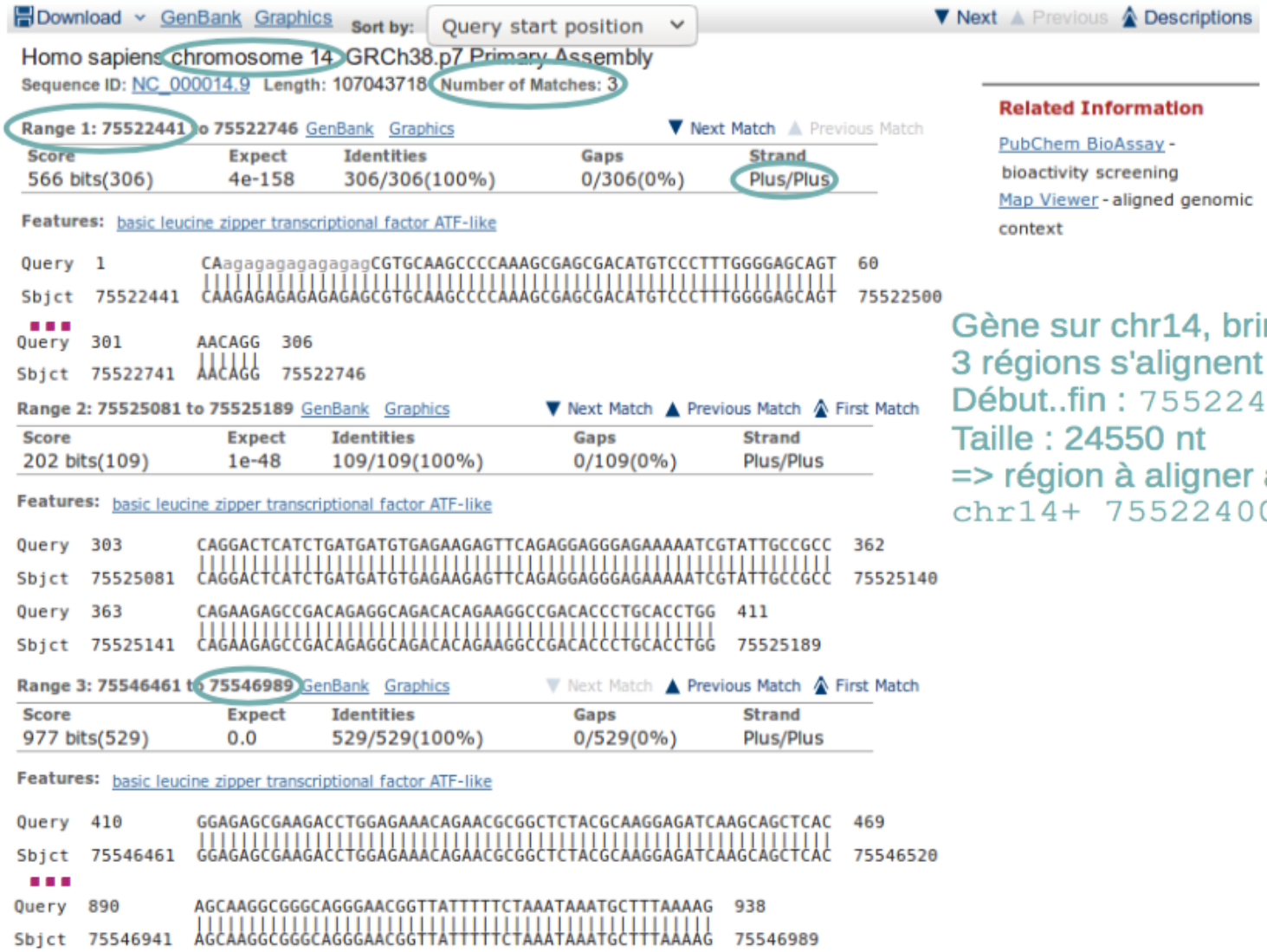
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

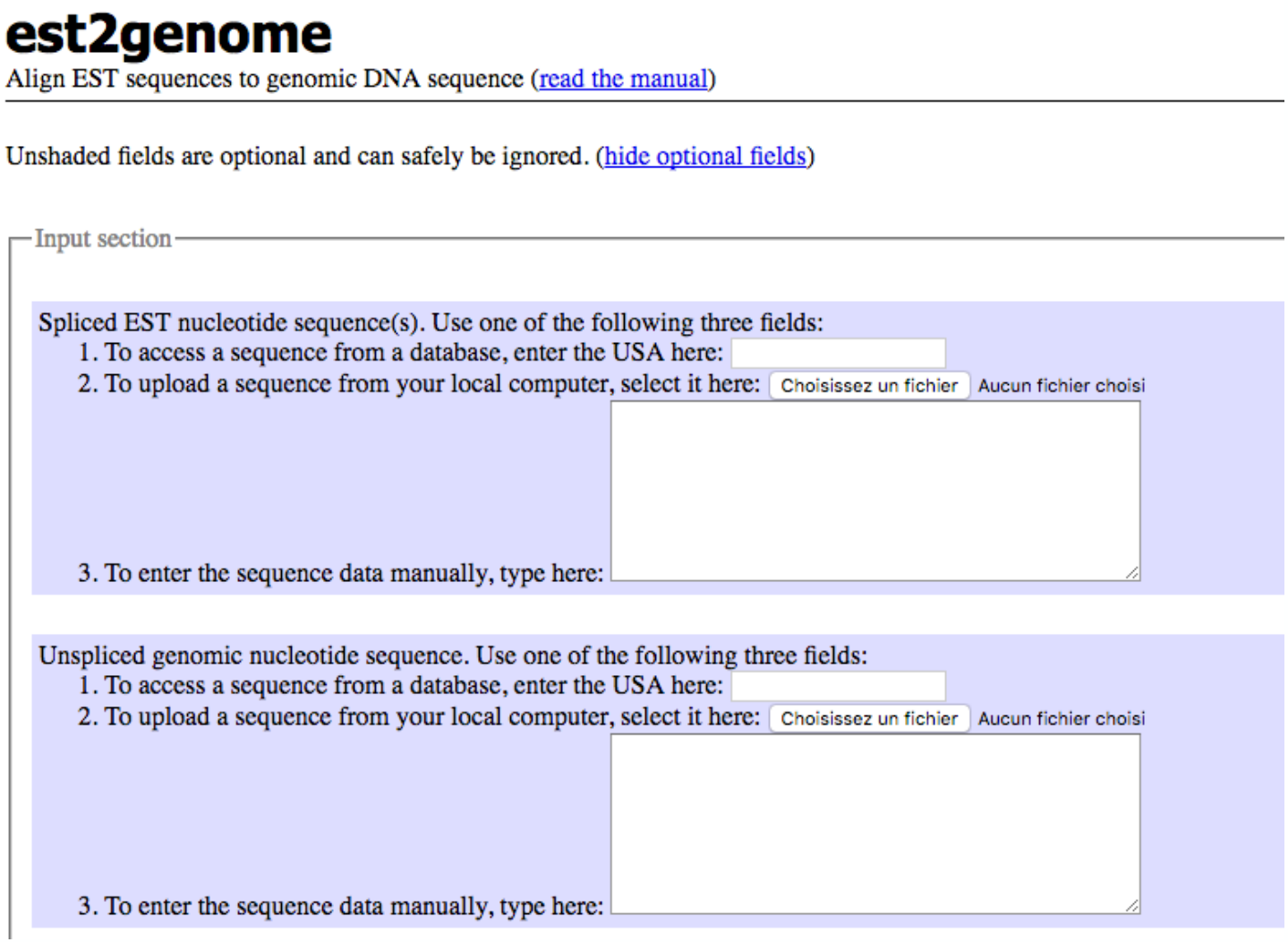
Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly	977	1746	100%	0.0	100%	NC_000014.9
Homo sapiens chromosome 14 genomic scaffold, GRCh38.p7 Primary Assembly HSCHR1	977	1746	100%	0.0	100%	NT_026437.13
Homo sapiens chromosome 14, alternate assembly CHM1_1.1, whole genome shotgun se	977	1746	100%	0.0	100%	NC_018925.2
Homo sapiens chromosome 14 genomic scaffold, alternate assembly CHM1_1.1, whole ge	977	1746	100%	0.0	100%	NW_004929393.1

Résultat de Blastn contre le génome, alignements



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



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

Résultat de EST2genome

- Détermination de la **position des exons sur** le **chr 14, région 75522400..75547000**

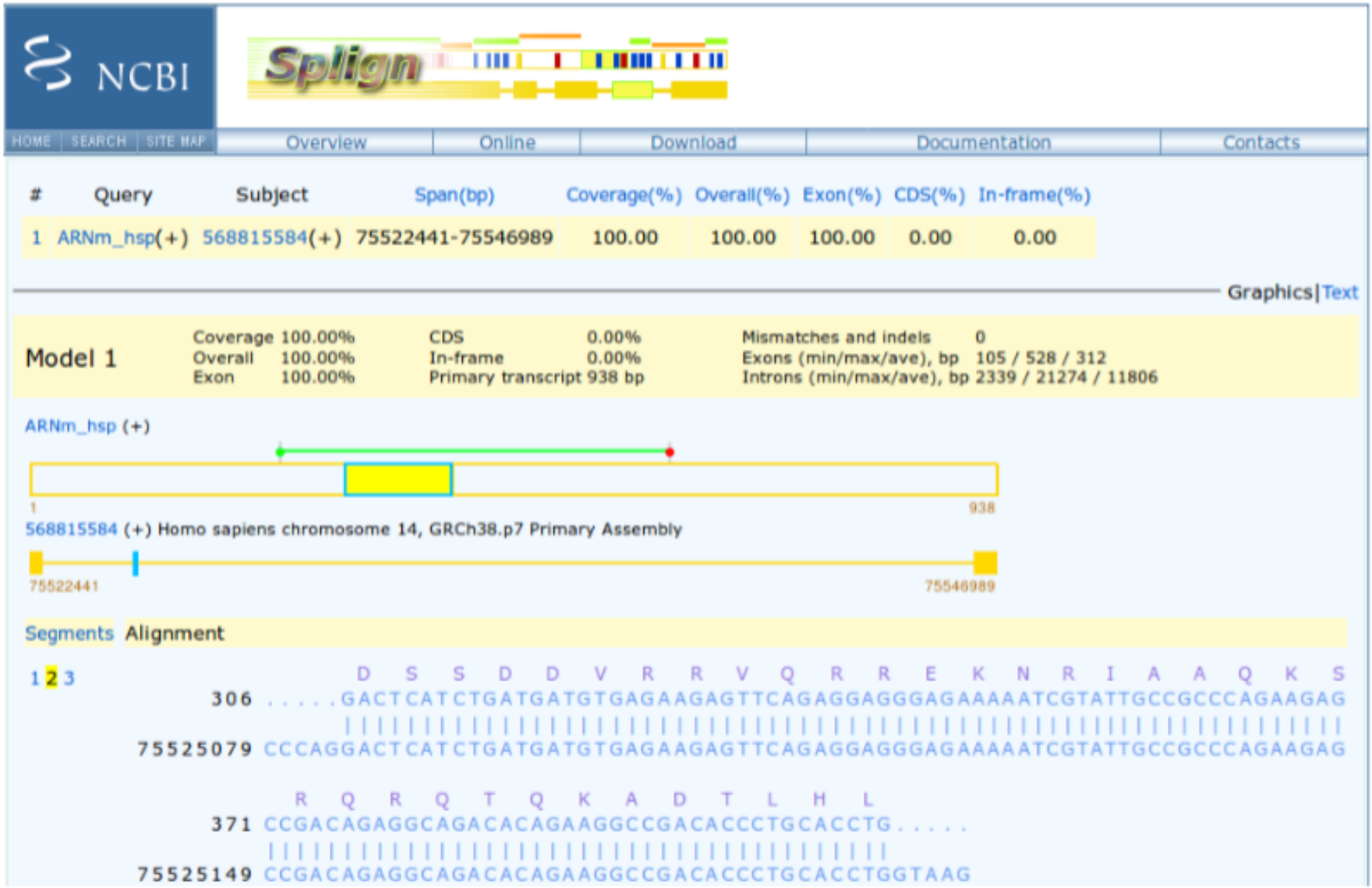
Exon	305	100.0	42	346	NC_000014	1	305	ARNm_hsp
+Intron	-20	0.0	347	2684	NC_000014			
Exon	105	100.0	2685	2789	NC_000014	306	410	ARNm_hsp
+Intron	-20	0.0	2790	24062	NC_000014			
Exon	528	100.0	24063	24590	NC_000014	411	938	ARNm_hsp
Span	898	100.0	42	24590	NC_000014	1	938	ARNm_hsp
Segment	305	100.0	42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

- Donc début exon 1 : 75522400+42-1

- Positions CDS sur extrait chromosome 14 :

join(284..346,2685..2789,24063..24272)

Résultat de Splign



Résultat de FGENESH

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 conding segment (se termine par un codon de terminaison)
PolA = signal pour la queue polyA

Résultat d'AUGUSTUS

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 identiques aux prédictions par comparaison de séquences

  Uniquement les régions codantes (le terme mRNA est abusif)

Résultat de GenScan

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 identiques aux prédictions par comparaison de séquences - uniquement les parties codantes
- Queue polyA prédite au même endroit que FGENESH

Bilan de l'analyse

- Toutes les prédictions concordent sur **3 exons**
- Un **exon codant supplémentaire** prédit par FGENESH
  épissage alternatif ?
- La protéine codée est sûrement **facteur de transcription B-zip**

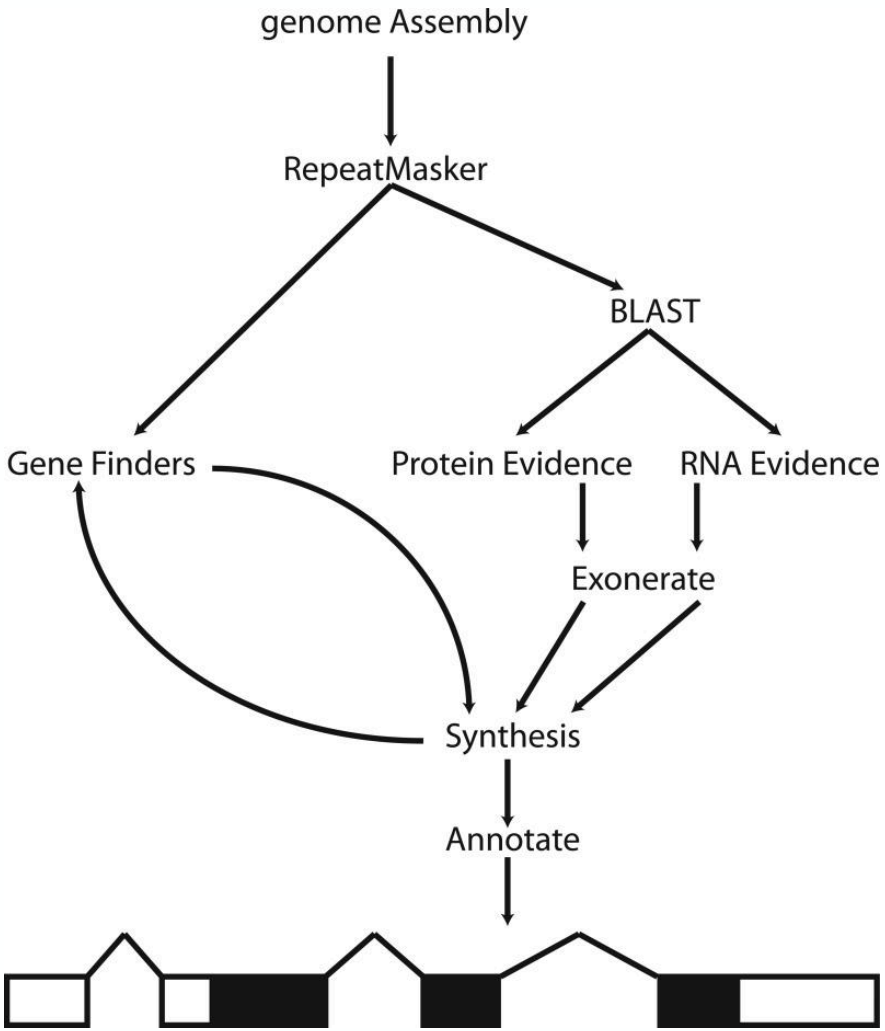
Annotation automatique d'un génome complet

- Utilisation d'un **pipeline d'annotation**
  Maker, PASA, Gnomon

...

- Exemple Maker (depuis 2008)

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.

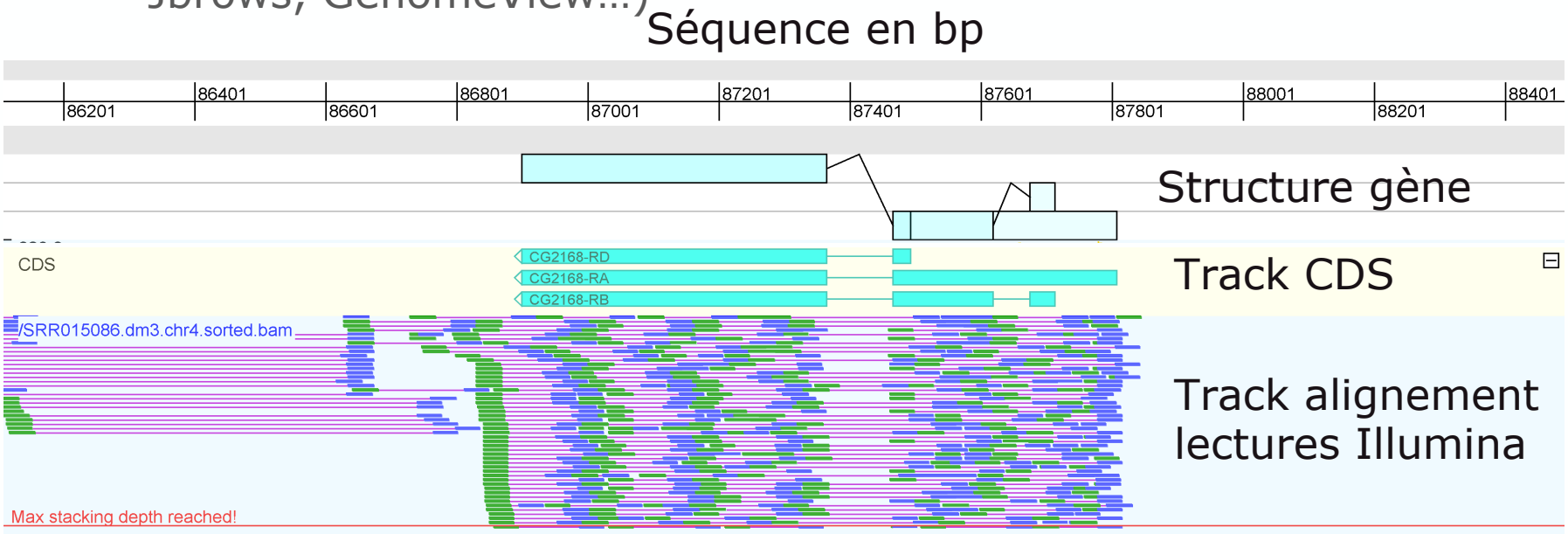


Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics. 2014 Dec 12;48:4.11.1-39.

Visualisation des données d'annotation

- Cinq formats utilisés couramment pour décrire les annotations sont les formats **GFF3**, GenBank, BED, GTF et EMBL; GFF3 → https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

- Utilisation d'un outil de visualisation de génome (IGV, Jbrows, GenomeView…)



Exemple de visualisation par GenomeView

Visualisation des données d'annotation

Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000014.9

GenBank Graphics

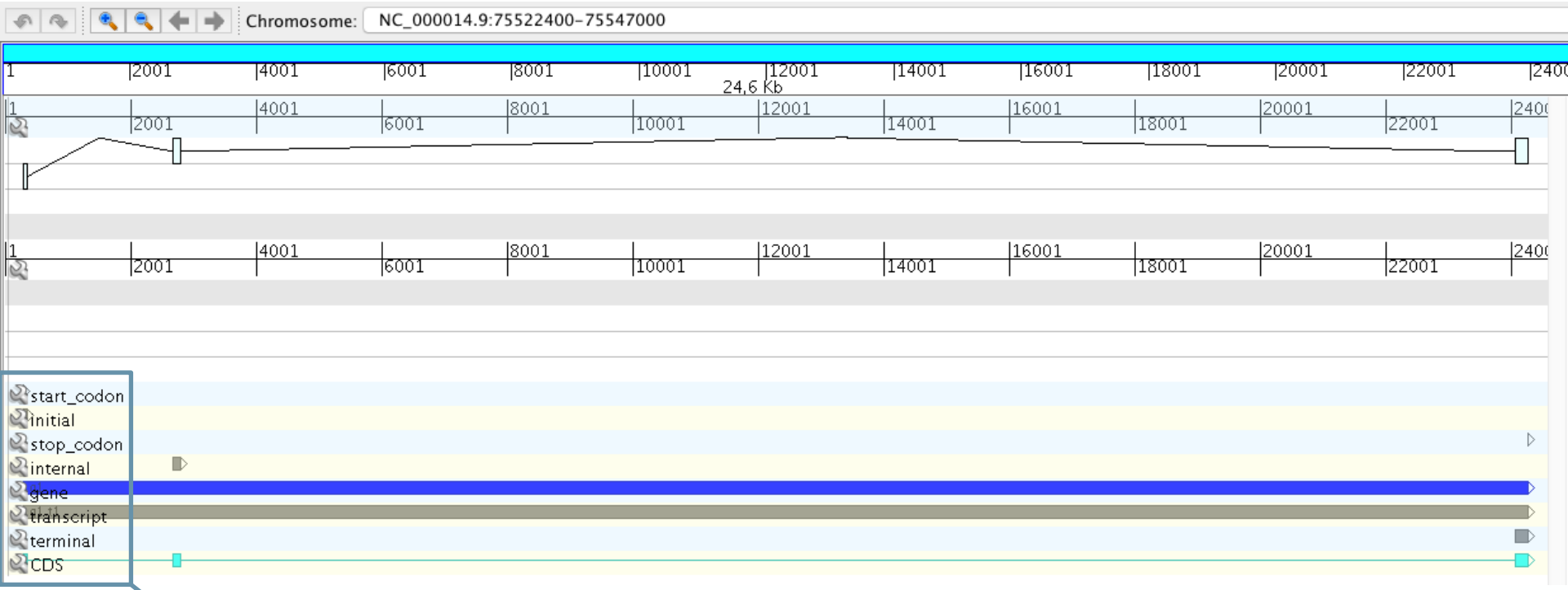
```
>NC_000014.9:75522400-75547000 Homo sapiens chromosome 14, GRCh38.p7 Primary
Assembly
TTTCCGCCCATGTGACTTCCAGCGTGAGTTACCAGAAACCACAAGAGAGAGAGAGAGCGTGCAAGCCCCA
AAGCGAGCGACATGTCCCTTTGGGGAGCAGTCCCTCTGCACCCCAGAGTGAGGAGGACGCAGGGGTCAGA
GGTGGCTACAGGGCAGGCAGAGGAGGCACCTGTAGGGGGTGGTGGGCTGGTGGCCCAGGAGAAGTCAGGA
AGGGAGCCCAGCTGGTGACAAGAGAGCCCAGAGGTGCCTGGGGCTGAGTGTGAGAGCCCGGAAGATTTCA
GCCATGCCTCACAGCTCCGACAGCAGTGACTCCAGCTTCAGCCGCTCTCCTCCCCCTGGCAAACAGGTAG
AGTCCTCCTTTTTCTCTCTCCTACCTTCTGATTCTCCTGGGGGATGGAAAGAGAGCCAGGCTTCCTTGTC
CTGCCCAGGGAGCTGAGGATGGAGGAAGTGGCTCGTTGCACGGGCACTCTGTTAGACTTAGGACATGGAA
TTTGCTACTAAGCTGTGCATATTGGCAGAGATCCTCATCCTTCCACCCATTCTGCCAAAGCCCCTTTTTC
TCTCCATTTTCCAAGGCTGCCTATCACCTCTGCCTCACTGGGGTTGCCACCCTAAAAAGCTTTCTAGGAA
GCAAAGAGGAGGATGAACATCAAAGAATGCAGAGAAAAGAGTCTACTGTTCTCCAAGGCTGTAGAAAAGT
```

Séquence au format FASTA

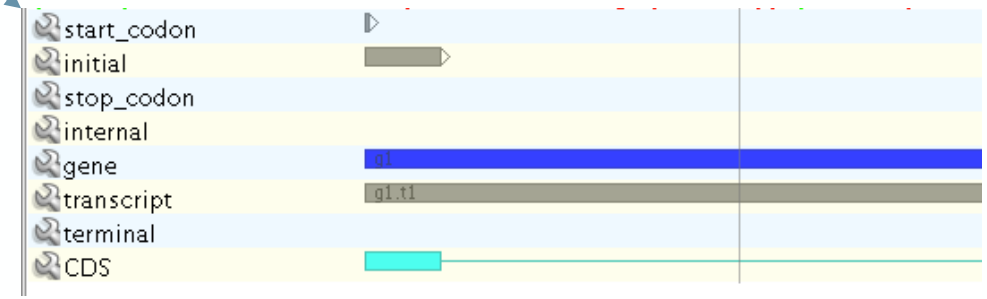
NC_000014	AUGUSTUS	gene	284	24272	0.89	+	.	Gene g1
NC_000014	AUGUSTUS	mRNA	284	24272	0.89	+	.	mRNA g1.t1
NC_000014	AUGUSTUS	start_codon	284	286	.	+	0	mRNA g1.t1
NC_000014	AUGUSTUS	initial	284	346	1	+	0	mRNA g1.t1
NC_000014	AUGUSTUS	internal	2685	2789	0.99	+	0	mRNA g1.t1
NC_000014	AUGUSTUS	terminal	24063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

Prédiction de gène par AUGUSTUS au Format GFF3

Visualisation des données d'annotation dans GenomeView



Les différentes tracks



Référence bibliographiques

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

- Annotation génomes **eucaryotes**
  - 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