

Introduction à la bioinformatique (UE SSV3U15)

Chapitre 3. Du gène au génome

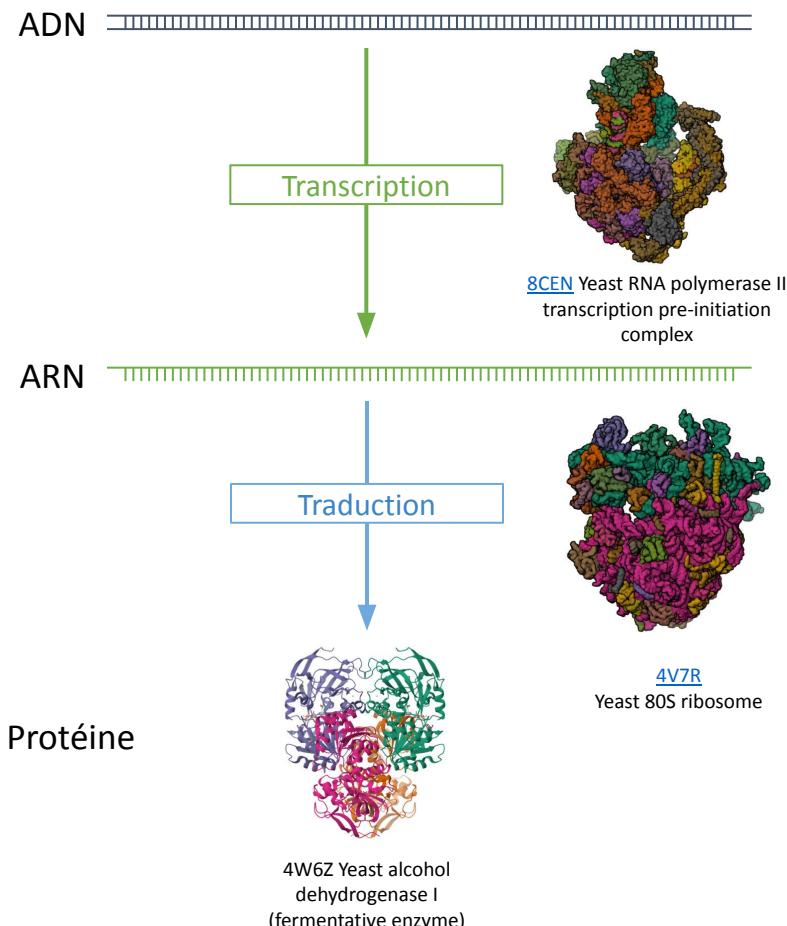
Jacques van Helden (Aix-Marseille Université)
ORCID [0000-0002-8799-8584](https://orcid.org/0000-0002-8799-8584)

1. Structure d'un gène
2. Disponibilité des génomes
3. Composition et organisation des génomes
4. Annotation des génomes : où sont les gènes ?
5. Annotation des génomes : que font les gènes ?
 - Assignation de fonction par similarité de séquences
 - Un élément structurant des génomes: la régulation
 - Génomique comparative
 - Coupable par association
 - La Gene Ontology – Définir et structurer les termes d'annotation des gènes et de leurs produits

Structure d'un gène

Le cas simple : l'ADN fait l'ARN fait la protéine

- Le modèle de base (et un peu trop simpliste) de l'expression des gènes repose sur une relation simple
 - Transcription** : synthèse d'une molécule d'ARN sur modèle, à partir d'une région de l'ADN
 - Traduction** : synthèse d'un polypeptide à partir de l'ARN messager (mRNA)

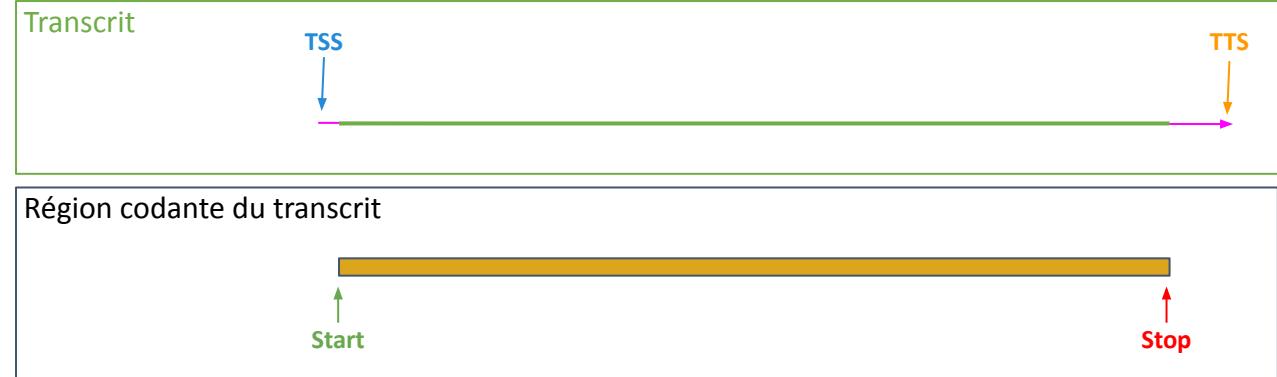
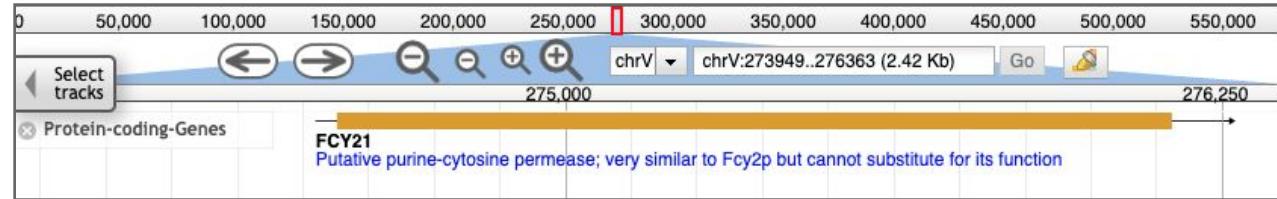


Un cas simple: le gène de levure FCY21

Le navigateur de génomes yeastgenomes.org permet de visualiser des régions génomiques et leurs annotations.

Exemple : le gène **FCY21** de la levure du boulanger (*Saccharomyces cerevisiae*) code pour une perméase “putative” des purines et de la cytosine.

- Le rectangle ocre indique la **région codante**, qui s'étend du **codon start** au **codon stop**.
- La ligne noire (interrompue par la boîte ocre) indique l'étendue du transcript, délimitée par le site d'initiation (*transcription start site, TSS*) et le site de terminaison (*transcription termination site, TTS*) de la transcription.
- La flèche indique le sens de la transcription.
- Les régions du transcript en amont et en aval correspondent aux régions non traduites (*untranslated regions, UTR*) en 5' et en 3'.

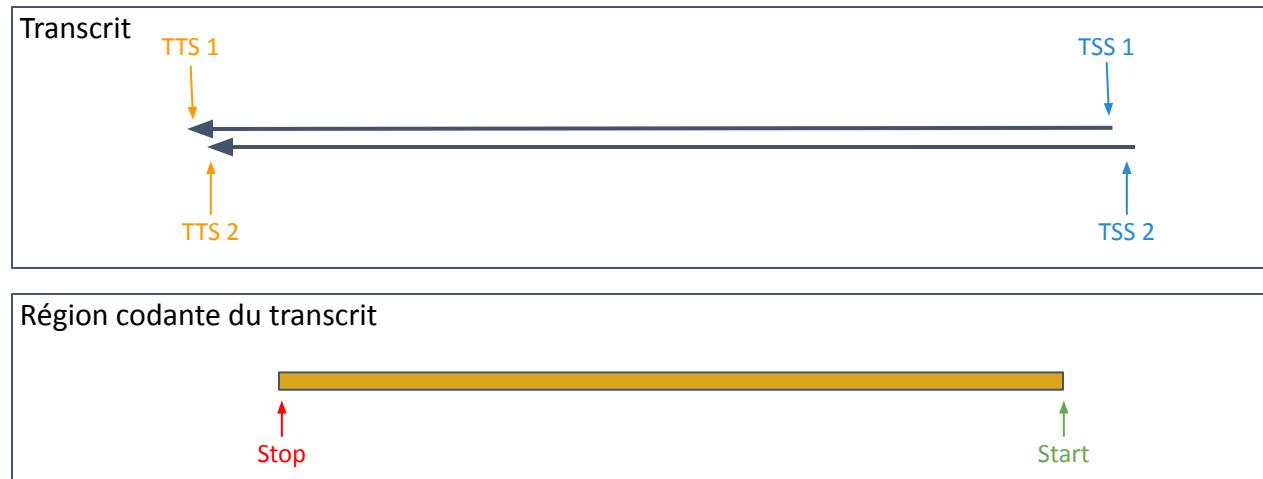
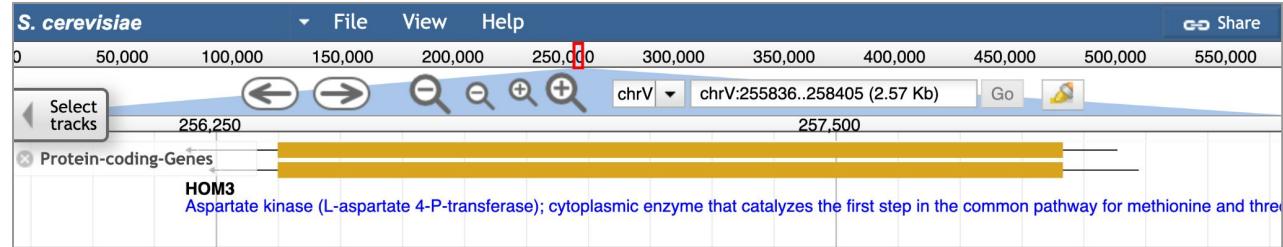


Transcrits alternatifs

Le navigateur de génomes yeastgenomes.org permet de visualiser des régions génomiques et leurs annotations (indication de tous les éléments qu'on y détecte).

Le gène **HOM3** code pour l'enzyme **aspartate kinase**, qui catalyse la première étape de la biosynthèse de l'homosérine.

- La ligne noire (partiellement marquée par la boîte ocre) indique l'étendue du transcript.
- La flèche indique le sens de la transcription.
- Pour ce gène, il existe deux transcrits alternatifs, qui diffèrent par le site d'initiation de la transcription ([Transcription Start Site, TSS](#)) et par le site de terminaison ([Transcription Termination Site, TTS](#))
- Le rectangle ocre indique la **région codante**, qui s'étend du **codon start** au **codon stop**.



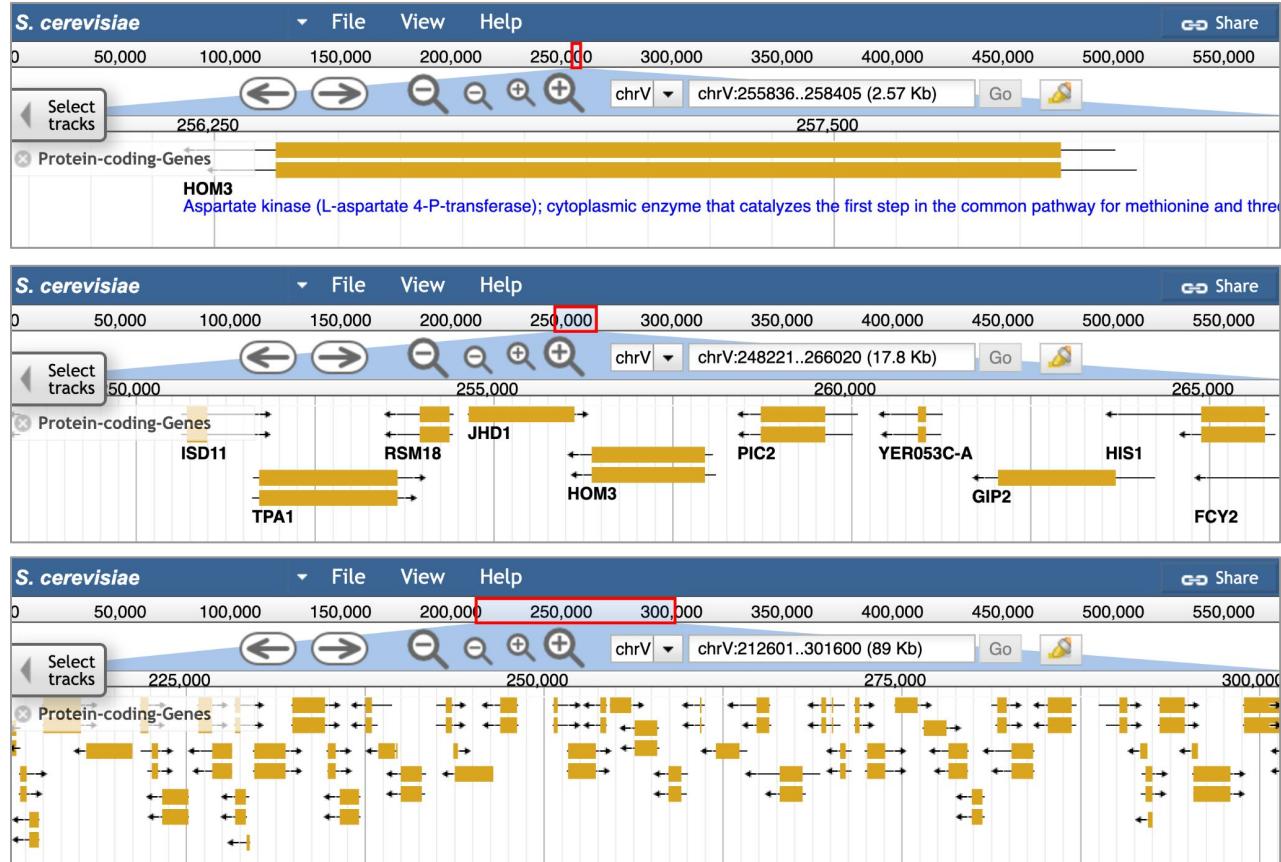
Disposition des gènes dans une région génomique de levure

En dézoomant, on peut observer la disposition des gènes dans la région génomique avoisinante.

Orientation : sur l'un ou l'autre brin, sans logique apparente.

Notation des brins

- + = D (direct) = W (Watson)
- - = R (réverse) = C (Crick)



<https://jbrowse.yeastgenome.org/?loc=chrV%3A255836..258405&tracks=Protein-Coding-Genes%2CNon-coding-RNA-Genes>

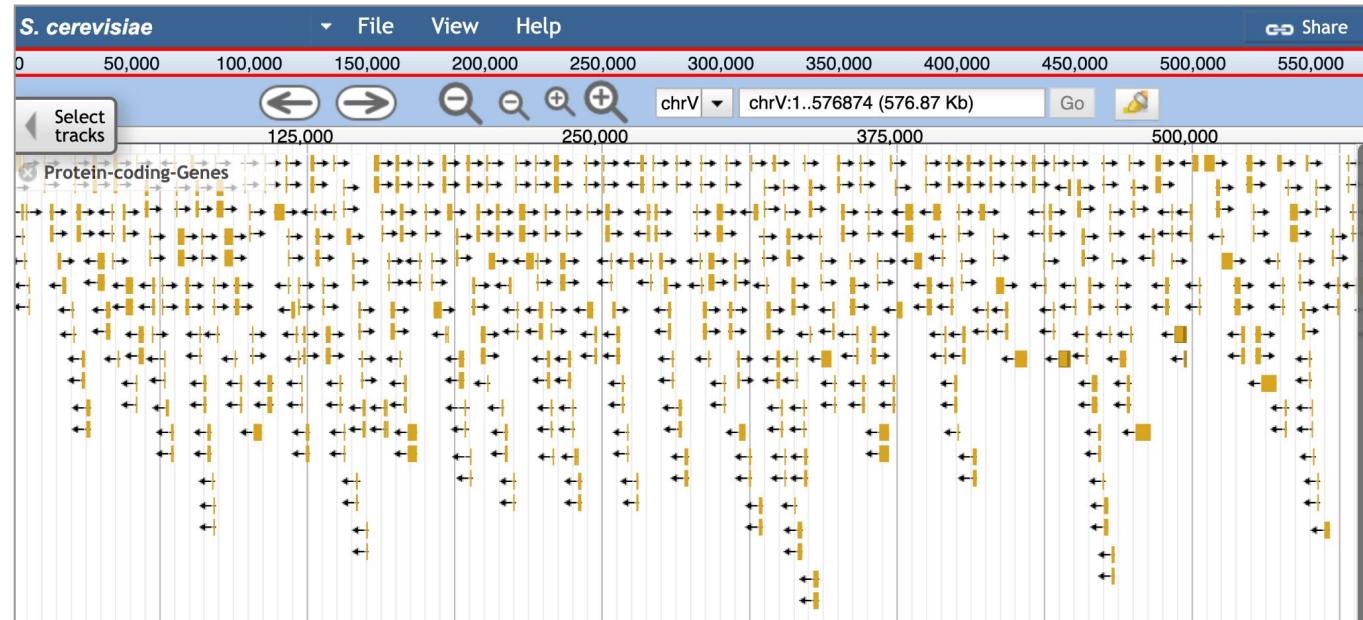
<https://jbrowse.yeastgenome.org/?loc=chrV%3A248221..266020&tracks=Protein-Coding-Genes%2CNon-coding-RNA-Genes>

<https://jbrowse.yeastgenome.org/?loc=chrV%3A212601..301600&tracks=Protein-Coding-Genes%2CNon-coding-RNA-Genes>

Disposition des gènes sur un chromosome de levure

On voit ici la disposition des gènes codants sur l'ensemble du cinquième chromosome (chrV) de levure.

- Longueur totale du chromosome : 576 874 bases.
- Nombre de gènes codants: 289
- Densité moyenne : 1 gène / 2kb



Les gènes non-codants

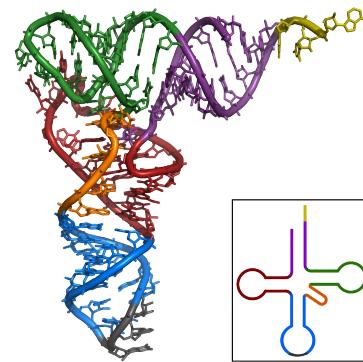
Eviter l'erreur fréquente qui consiste à ne prendre en considération que les gènes codants.

Les ARN ne font pas que servir de modèle à la synthèse des protéines.

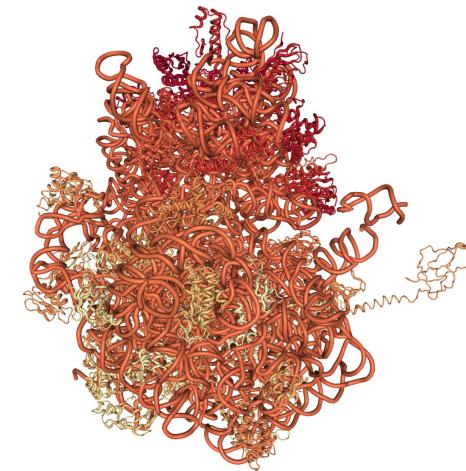
Il existe des gènes qui sont transcrits mais pas traduits.

- tRNA : ARN de transfert
- rRNA : ARN ribosomique
 - Le ribosome est un assemblage complexe d'ARN et de protéines
- lncRNA : long non-coding RNA (lncRNA)
- microRNA : petits ARN impliqués dans la régulation de l'expression des gènes

tRNA



Ribosome



PDB 4V6C. Crystal structure of the *E. coli* 70S ribosome in an intermediate state of ratcheting

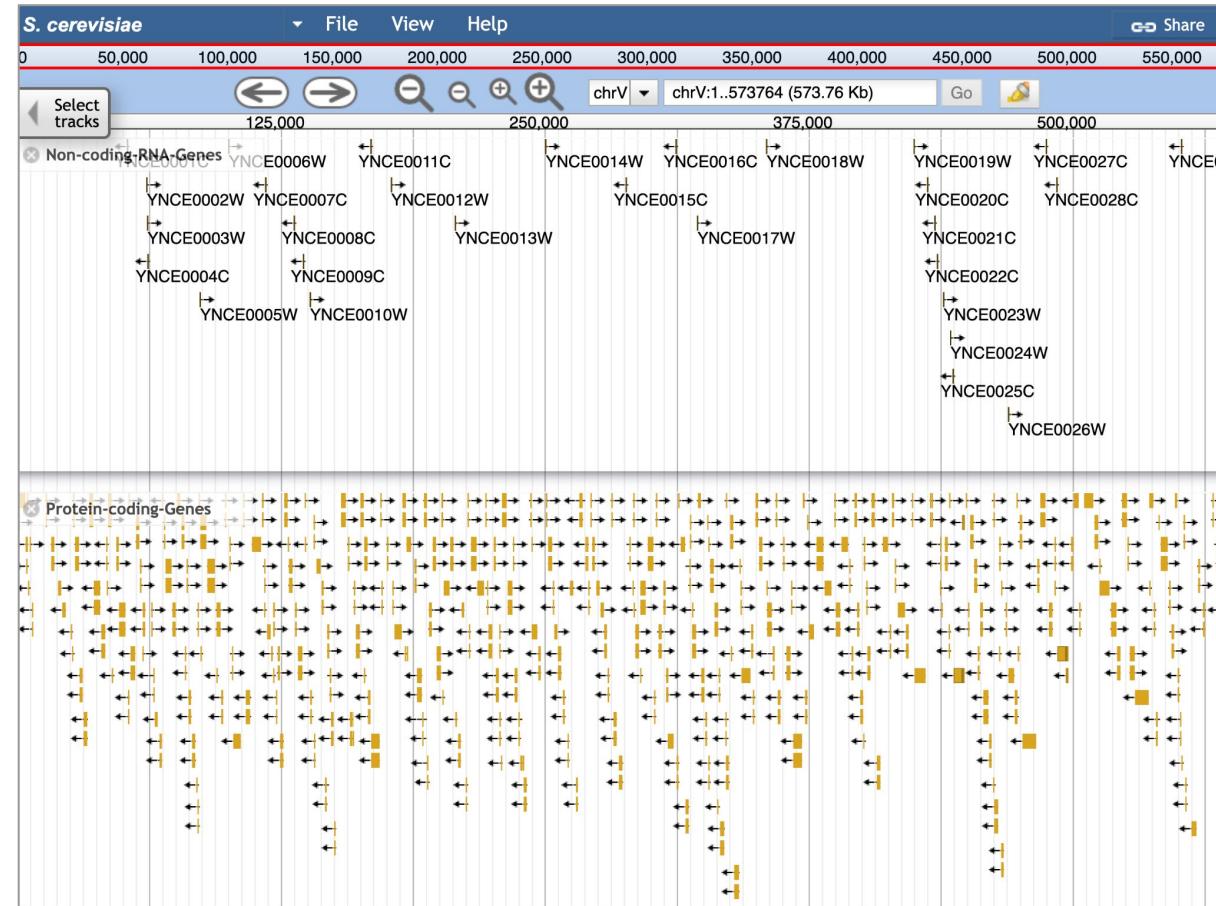
Les gènes non-codants

Le navigateur de génomes permet de sélectionner différentes **pistes d'annotation (annotation tracks)**.

Le chromosome V de la levure inclut 28 gènes non-codants (haut de la figure).

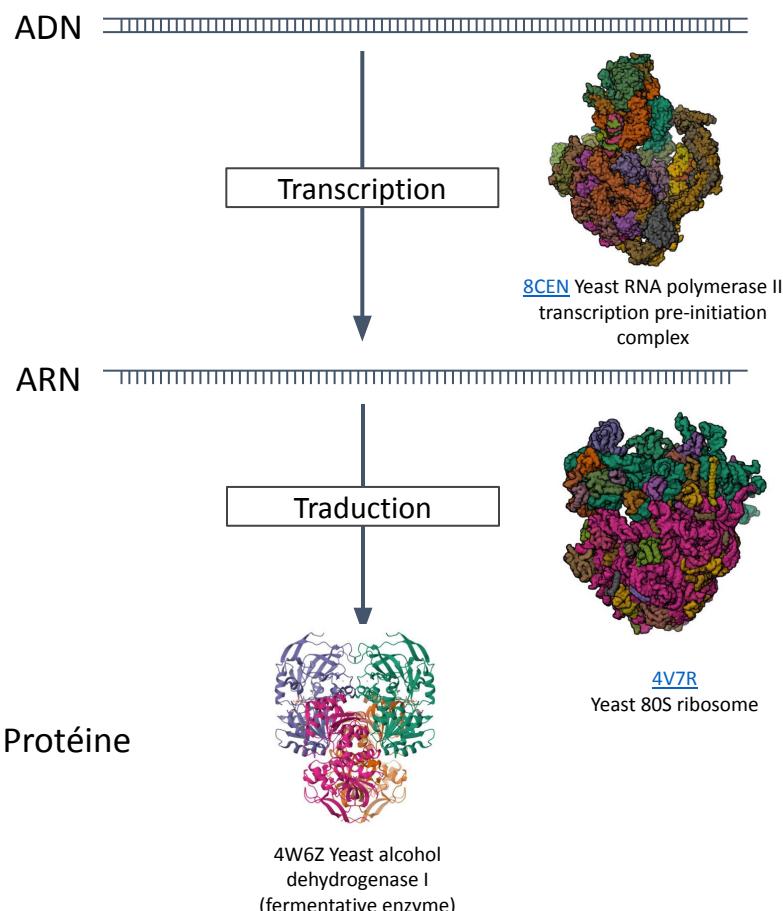
Ces gènes sont transcrits, et produisent des ARN non codants avec différentes fonctions:

- ARN de transfert (**tRNA**), 20 gènes sur le chromosome V
- snRNA: small nuclear RNA
- régulation d'autres gènes



Revenons au cas simple : l'ADN fait l'ARN fait la protéine

- Revenons au modèle de base (un peu trop simpliste)
 - **Transcription** : synthèse d'une molécule d'ARN sur modèle, à partir d'une région de l'ADN
 - **Traduction** : synthèse d'un polypeptide à partir de l'ARN messager (mRNA)



Pas si simple : l'ADN fait l'ARN primaire fait l'ARN mature fait la protéine

- D'après Uniprot, la myoglobine compte 154 acides aminés (Uniprot [MYG_HUMAN](#)).
- En principe il suffirait donc d'un ARN de $154 \text{ codons} = 154 \times 3 = 459$ nucléotides pour fournir l'information nécessaire à la traduction.
- Cependant, le UCSC genome browser indique que le gène occupe $\sim 17\text{kb}$ (piste [UCSC RefSeq](#))
 - Comment expliquer la différence ?
 - Comment lire et interpréter les informations du navigateur de génome ?

P02144 · MYG_HUMAN

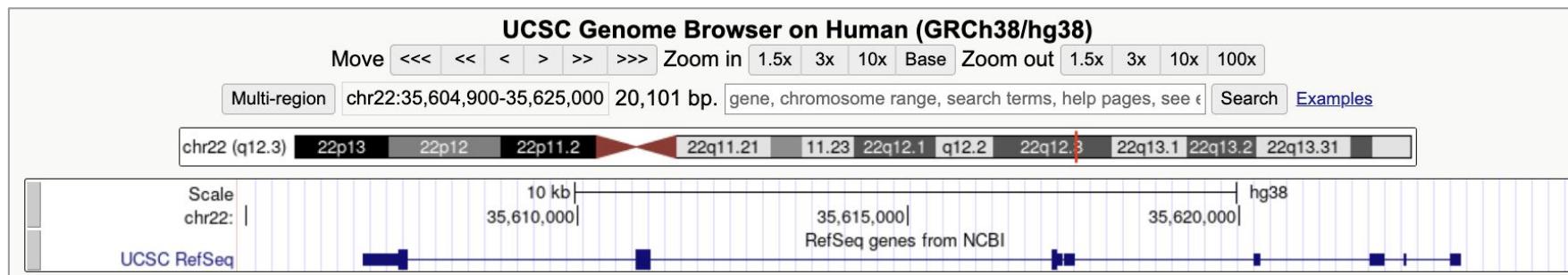
Protein ⁱ	Myoglobin	Amino acids	154 (go to sequence)
Gene ⁱ	MB	Protein existence ⁱ	Evidence at protein level
Status ⁱ	UniProtKB reviewed (Swiss-Prot)	Annotation score ⁱ	55
Organism ⁱ	Homo sapiens (Human)		

Entry Variant viewer Feature viewer Genomic coordinates Publications External links His

Tools Download Add Add a publication Entry feedback

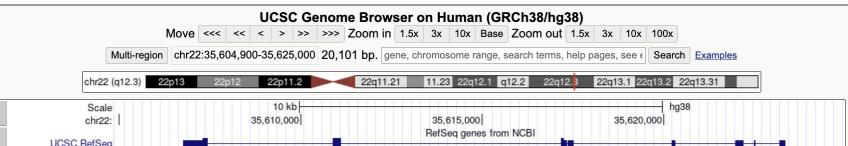
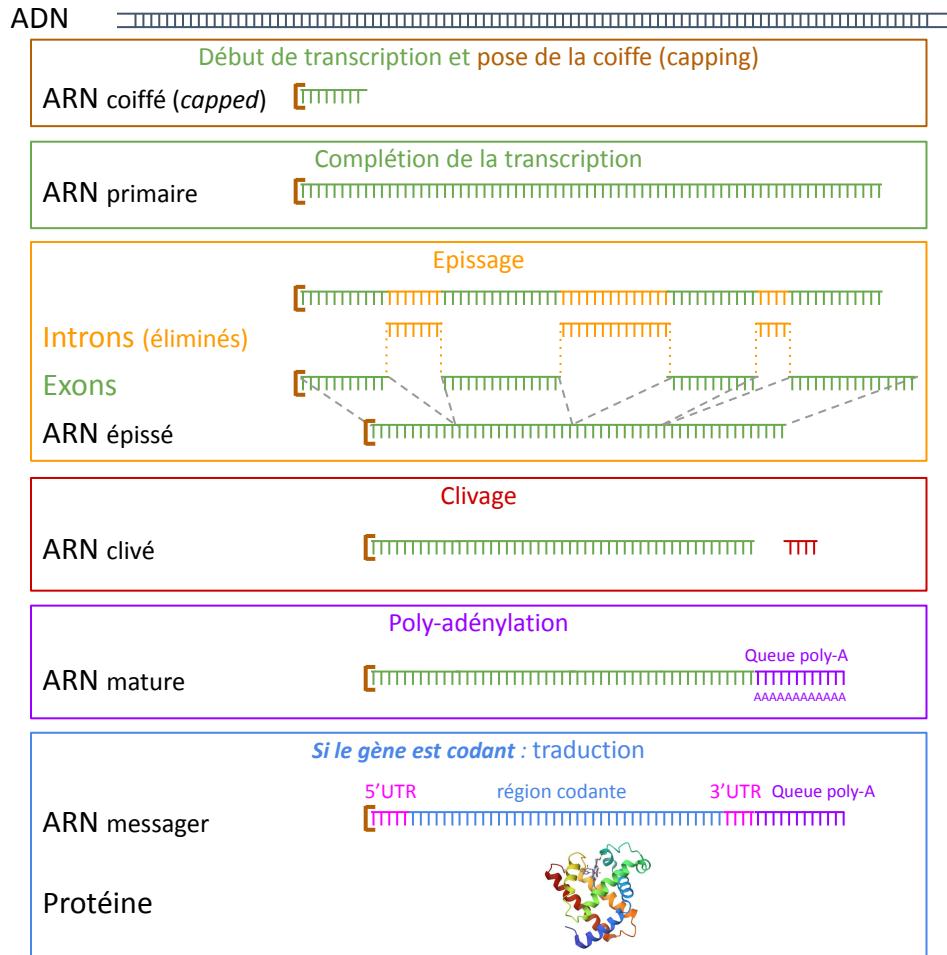
Functionⁱ

Monomeric heme protein which primary function is to store oxygen and facilitate its diffusion within muscle tissues. Reversibly binds oxygen through a pentacoordinated heme iron and enables its timely and efficient release as needed during periods of heightened demand (PubMed:30918256, PubMed:34679218). Depending on the oxidative conditions of tissues and cells, and in addition to its ability to bind oxygen, it also has a nitrite reductase activity whereby it regulates the production of bioactive nitric oxide (PubMed:32891753). Under stress conditions, like hypoxia and anoxia, it also protects cells against reactive oxygen species thanks to its pseudoperoxidase activity (PubMed:34679218). 



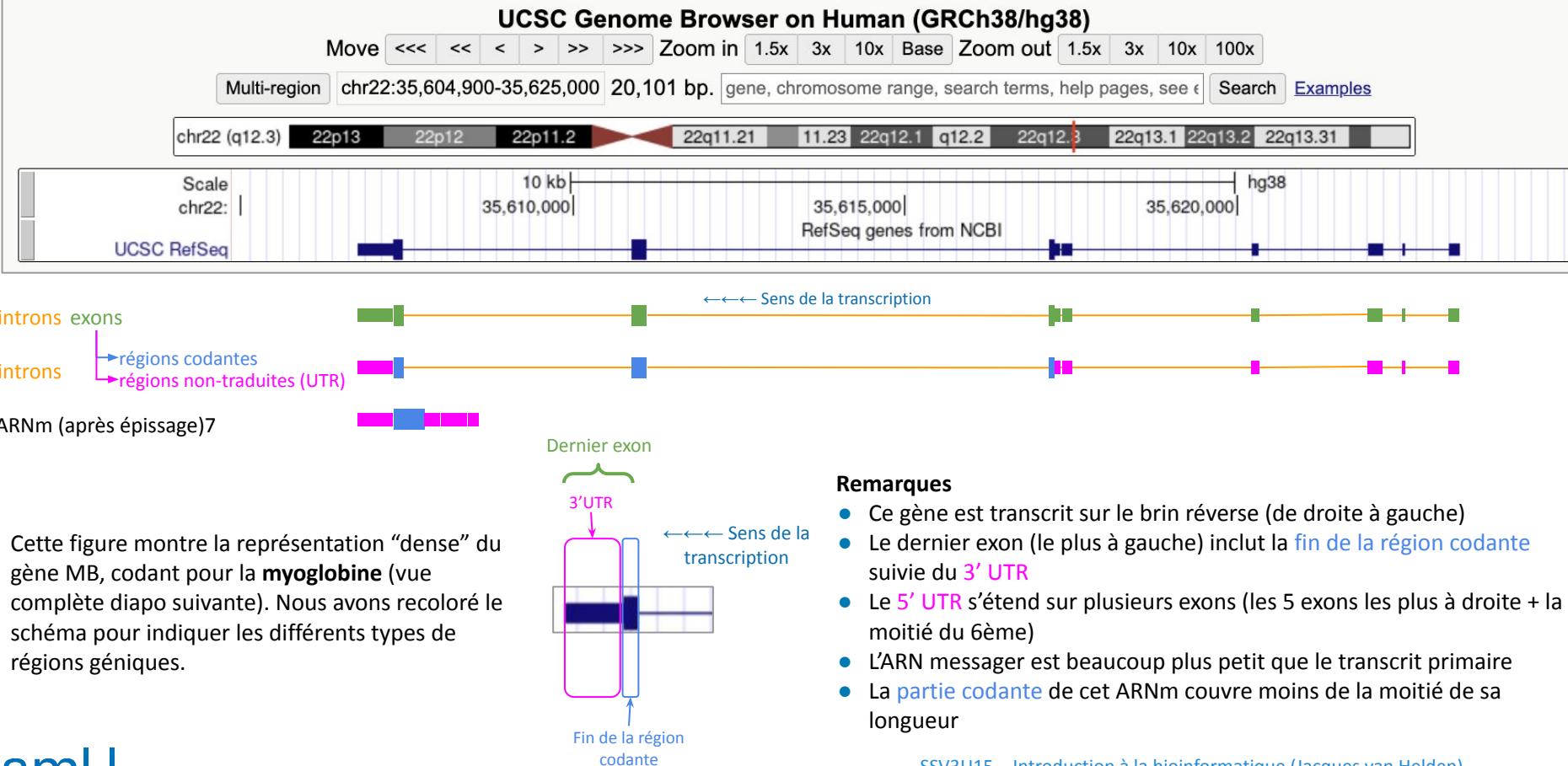
Pas si simple : l'ADN fait l'ARN primaire fait l'ARN mature fait la protéine

- Schéma adapté en incluant la **maturation de l'ARN**
- **Transcription** : synthèse d'une molécule d'ARN sur modèle, à partir d'une région de l'ADN.
 - Sites alternatifs d'initiation et de terminaison → transcrits multiples pour un gène
- **Epissage** : élimination de certains segments de l'ARN ("introns") et rabotage des autres segments ("exons").
 - Sites alternatifs d'épissage → transcrits multiples pour un gène
- **Clivage et poly-adénylation** : dans la région 3', l'ARN primaire est clivé (coupé), et une queue poly-A y est ajoutée (stabilisation de l'ARN). Cette queue polyA stabilise l'ARN.
- **Traduction** : synthèse d'un polypeptide à partir de la **partie codante** de l'ARN messager (mRNA).
- Note: les **régions non traduites (untranslated regions, UTR)** aux extrémités 5' et 3' de l'ARNm jouent un rôle dans la stabilité de l'ARN et dans la régulation de la traduction.

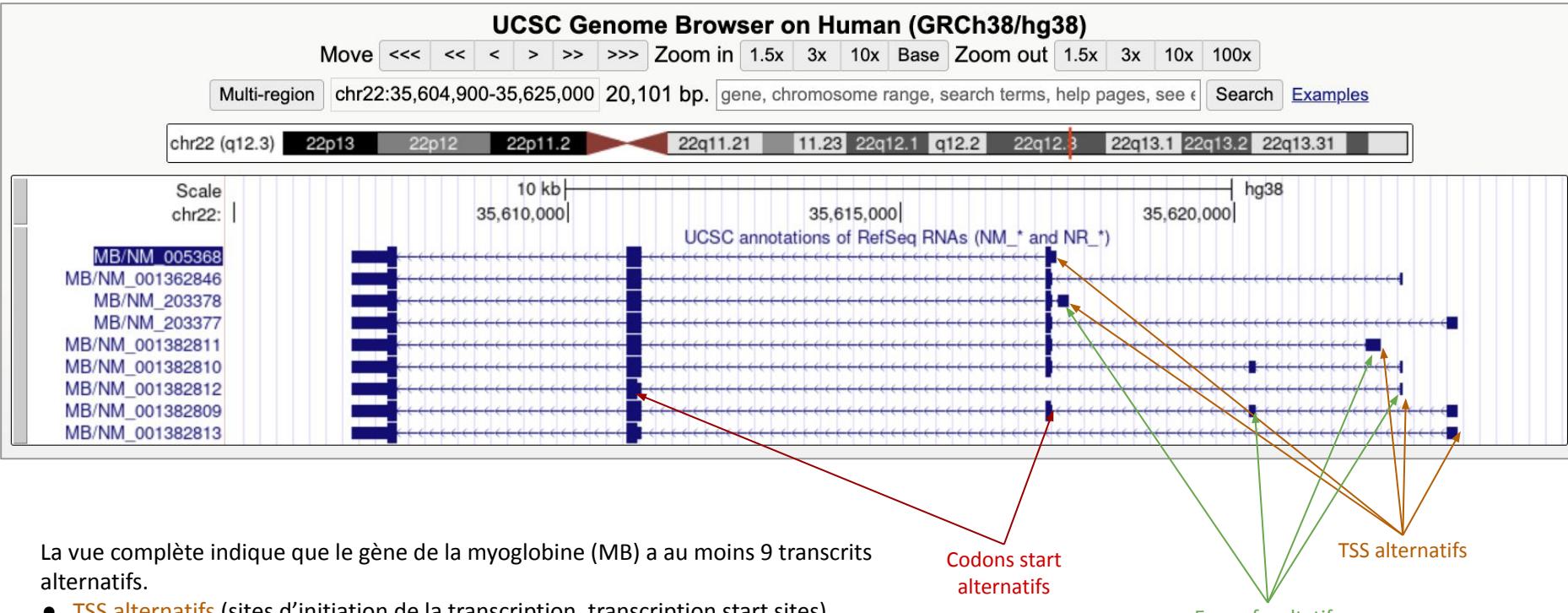


Pour aller plus loin : Bentley, D. L. Coupling mRNA processing with transcription in time and space. Nat Rev Genet 15, 163–175 (2014). doi.org/10.1038/nrg3662

Pas si simple : l'ADN fait l'ARN primaire fait l'ARN mature fait la protéine



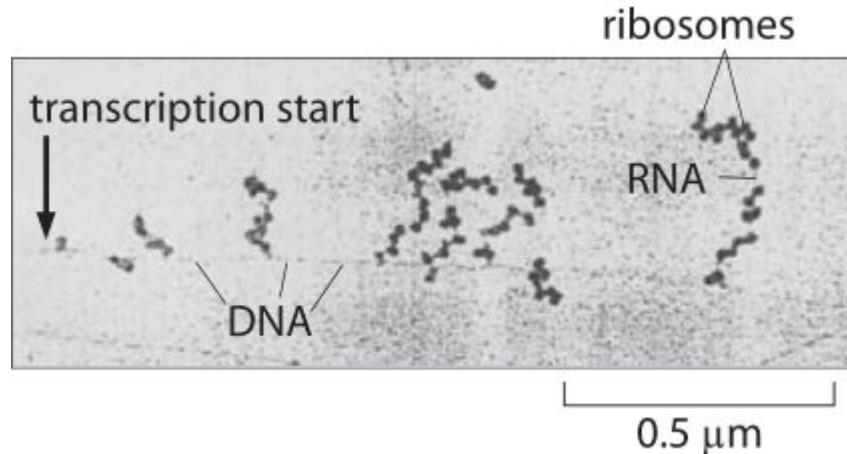
Pas si simple : transcrits alternatifs



Le transcrit du haut est majoritaire.

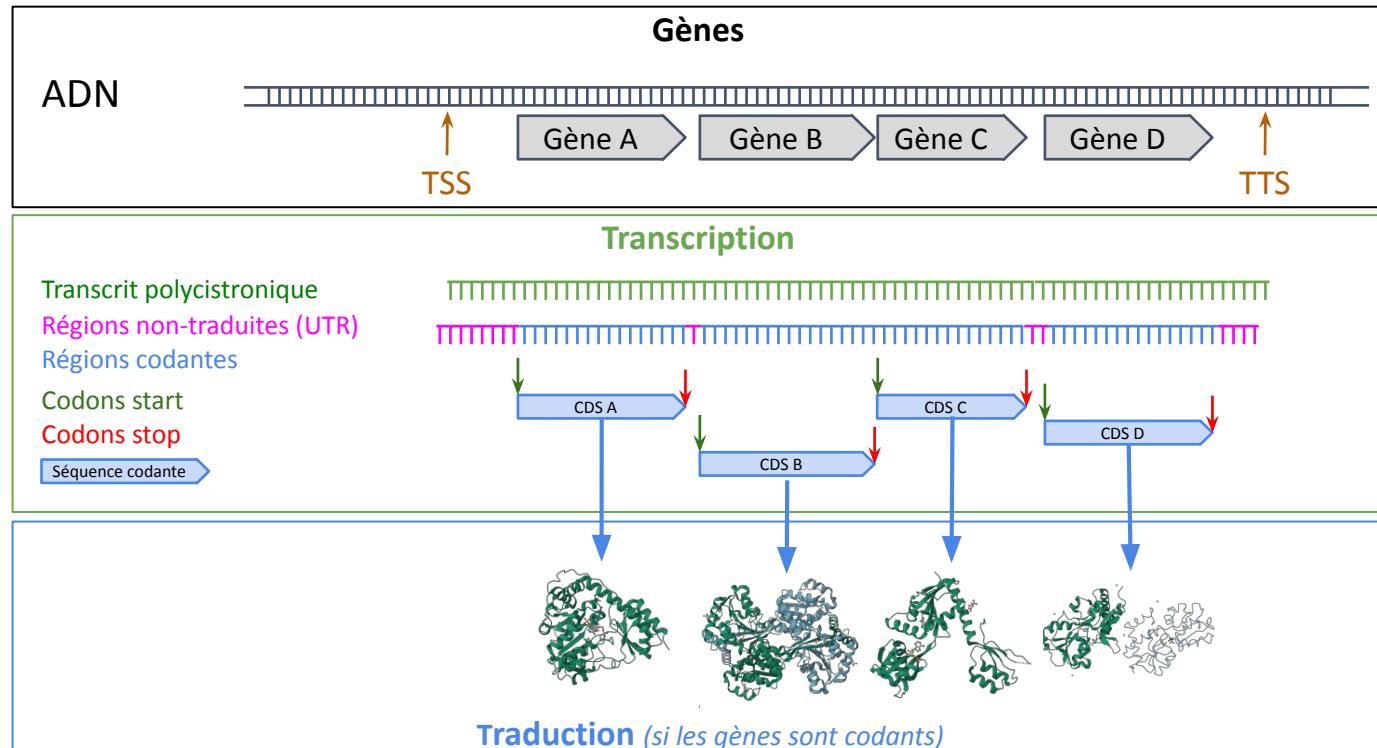
Transcription et traduction simultanées chez les bactéries

- Chez les eucaryotes, la transcription et la traduction se font séparément: dans le noyau pour la traduction, et dans le cytoplasme pour la transcription.
- Chez les prokaryotes, la transcription et la traduction se passent au même endroit, et simultanément.
- Figure: photo en microscopie électronique d'un morceau de génome bactérien (DNA) avec
 - plusieurs sites de transcription active (RNA),
 - sur chaque ARN, plusieurs sites de traduction active (ribosomes)



Génomes bactériens – Opéron

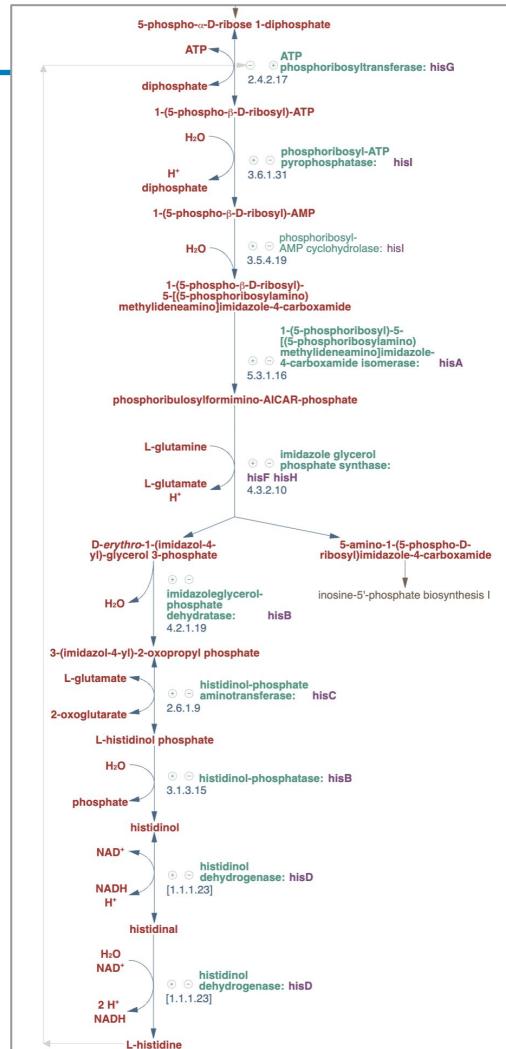
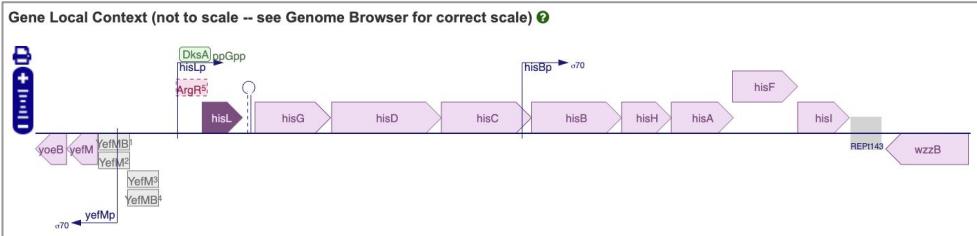
- Chez les prokaryotes, une unité de transcription peut couvrir un ou plusieurs gènes.
- **Transcrit polycistronique :** molécule d'ARN englobant plusieurs gènes
- **Opéron:** ensemble de gènes co-transcrits sur un même ARN



Exemple: l'opéron histidine d'*Escherichia coli*

Figure ci-dessous : structure de l'opéron histidine d'*Escherichia coli* extraite de la base de connaissances EcoCyc (ecocyc.org).

Figure de droite: voie métabolique de biosynthèse de la L-histidine



- Ecocyc histidine operon: ecocyc.org/gene?orgid=ECOLI&id=EG11269#TU

Exemple: l'opéron histidine d'*Escherichia coli*

Figure du haut: structure d'un opéron d'*Escherichia coli* extraite de la base de connaissances EcoCyc (ecocyc.org).

Figure du bas: localisation (mapping) des fragments de lecture d'ARN (RNA-seq transcriptomique) dans la région génomique correspondante.

- La hauteur des profils est proportionnelle au nombre de fragments de lecture localisés à chaque position.
- La couleur et l'orientation verticale indiquent le brin de lecture direct (vert, haut) ou réverse (violet, bas).
- On note un continuum de lectures sur toute la longueur de l'opéron (avec des disparités quantitatives).
- Noter aussi le gène b3207 (yrbL), transcrit séparément.

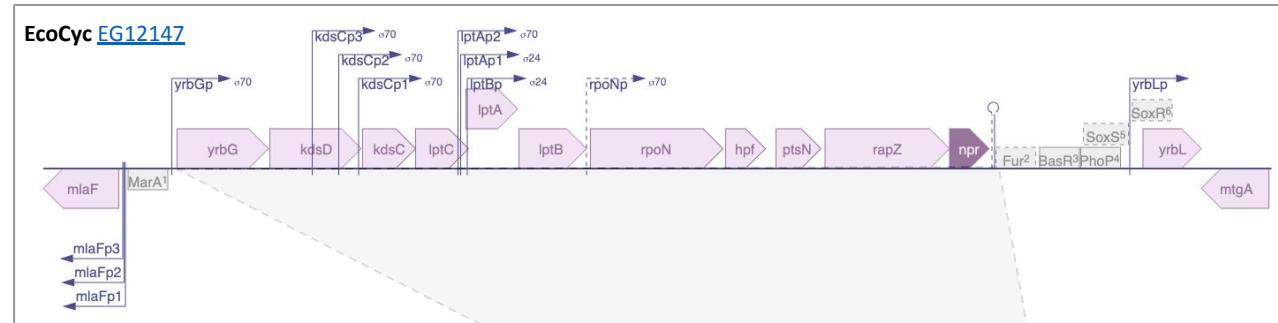
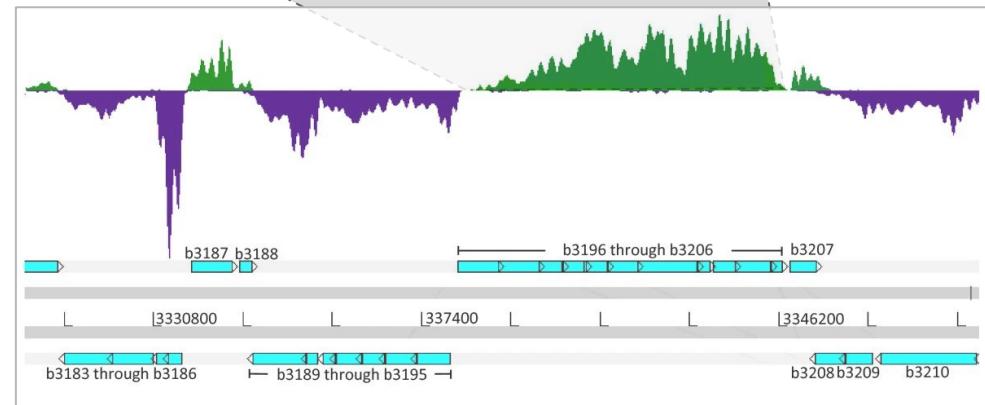


Figure de [Giannoukos et al.](#) (2012).

- Vert: régions transcrtes sur le brin positif
- Violet: régions transcrtes sur le brin négatif
- Cyan: gènes (la flèche indique l'orientation)
- "b3196 through b3206" : identifiants des gènes délimitant l'opéron



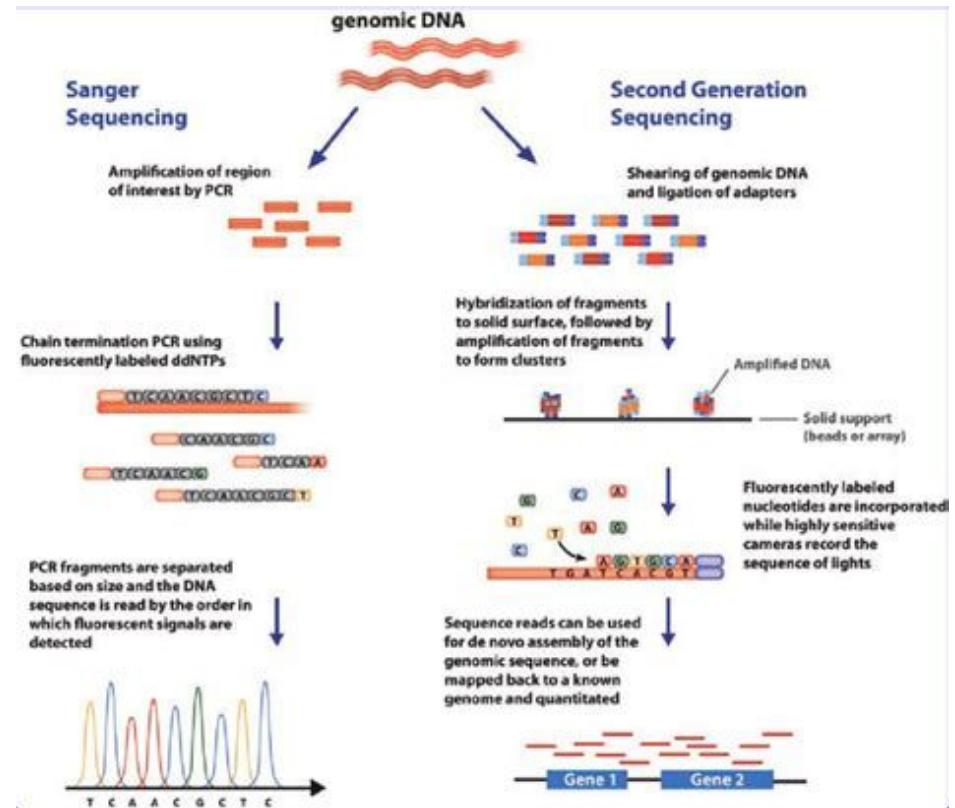
• Ecocyc : ecocyc.org/gene?orgid=ECOLI&id=EG12147#TU

• Giannoukos, G. et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol* 13, r23 (2012). doi.org/10.1186/gb-2012-13-3-r23

Disponibilité des génomes

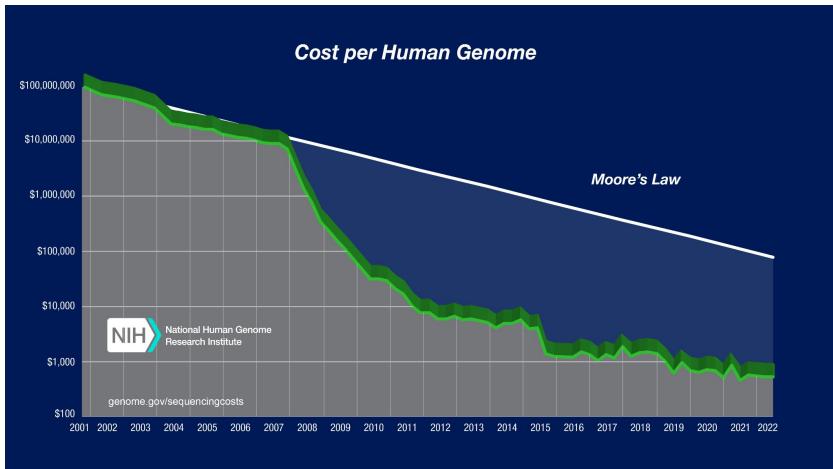
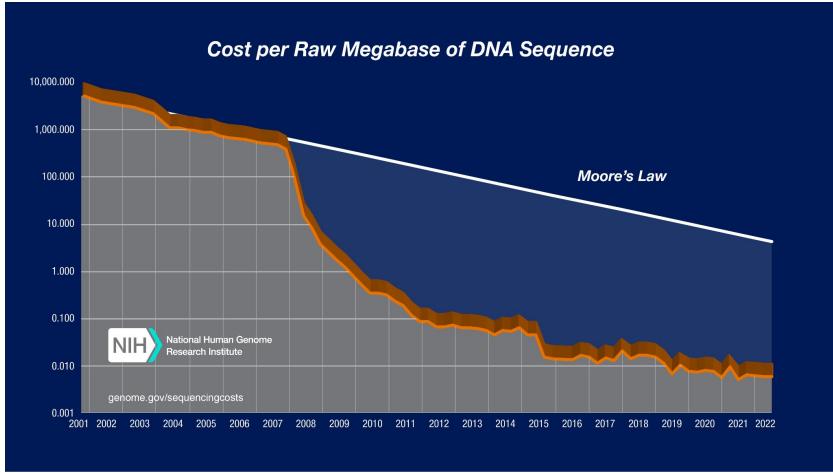
Séquençage massivement parallèle (“Next Generation Sequencing”)

- De 1977 à 2007, la méthode de Sanger était la seule façon de séquencer l'ADN (partie gauche de la figure)
- Durant les années 1990-2000, cette méthode a été utilisée pour les premiers projets de séquençage génomique, qui ont suscité des améliorations techniques (robotisation, informatisation)
- En 2007, plusieurs compagnies proposent une stratégie radicalement différente: le **séquençage massivement parallèle**, également appelée “*Next Generation Sequencing*”.
- Cette approche produit des millions de petits fragments de séquences (typiquement 36 à 300bp), qu'il faut ensuite analyser, avec différentes approches possibles
 - Localisation sur un génome de référence s'il existe
 - Assemblage de novo s'il n'y a pas de génome de référence



Du gène au génome

- 1990-2000 : premiers projets de séquençage du génome d'organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabe, et ... “le” génome humain
- 2001 : première publication d'un génome humain
- 2007 : technologies de séquençage massivement parallèle (“Next Generation Sequencing”, NGS)
 - De 2001 à 2007: les coûts diminuent en suivant la loi de Moore (décroissance exponentielle)
 - 2008; diminution brutale des coûts du séquençage
 - Depuis 2011: réduction plus modérée des coûts

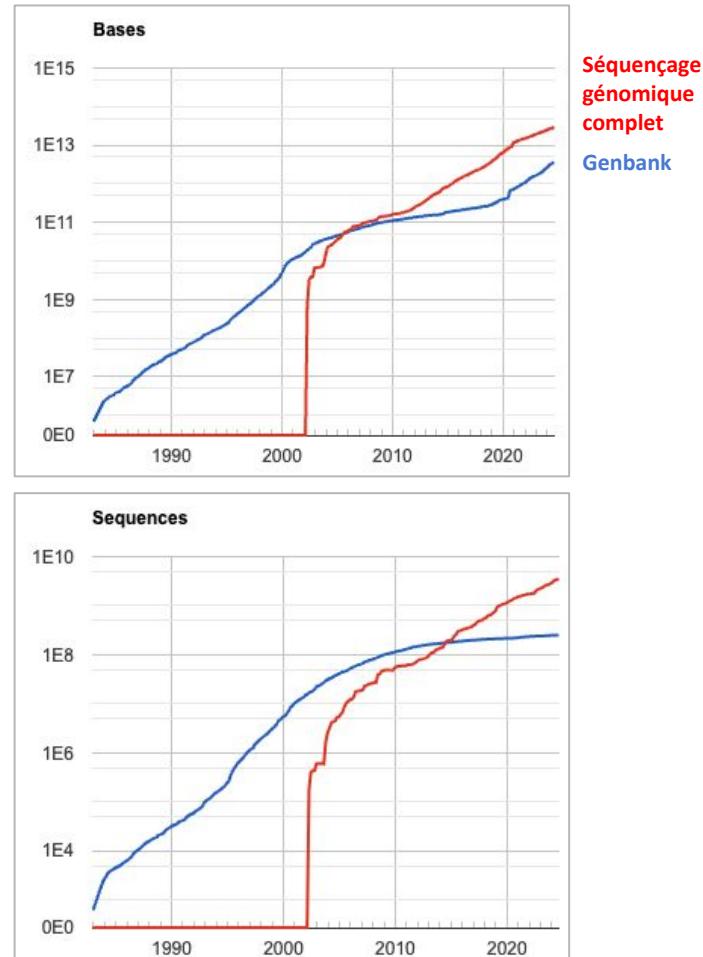


Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 2024-09-04.

<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

Disponibilité des séquences d'ADN

- Les séquences de macromolécules qui font l'objet de publications scientifiques sont systématiquement déposées dans des entrepôts de données internationaux, et rendues accessibles au public
 - Une exception: les séquences génomiques associées à des échantillons humains (voir cours sur la médecine génomique)
- Le nombre de séquences disponibles depuis 1980 montre une croissance exponentielle (linéaire sur un axe logarithmique).
 - Taux d'augmentation: de 1990 à 2020, $\times 1.48/\text{an}$
- Avant 2002, il s'agissait de séquences individuelles de gènes ou de fragments génomiques (courbe bleue, Genbank).
- A partir de 2002, le séquençage de génomes complets prend le pas (courbe rouge).



Disponibilité des génomes

- Avant les années 1990, le séquençage de l'ADN représentait un travail important. Un doctorant pouvait passer une partie significative de sa thèse à séquencer quelques kilobases afin de caractériser un seul gène.
- Les « projets génomes » ont stimulé le développement de méthodes de séquençage automatique, qui ont suscité des progrès technologiques impressionnants.
- Nous disposons aujourd’hui (septembre 2024) de plusieurs centaines de milliers de génomes complètement séquencés, en libre accès.

Remarques

- Le degré de finition de ces génomes varie d'un groupe à l'autre
- Un grand nombre de génomes additionnels ont été séquencés par des compagnies, et ne sont pas accessibles au public.

Génomes complets disponibles au NCBI

Date 31/07/2024
Source <https://ftp.ncbi.nlm.nih.gov/genomes/>

Nombre	Groupe
367 675	Bactéries
14 975	Virus
2 171	Archées
588	Fungi (levures, champignons)
404	Métazoaires – Invertébrés
399	Métazoaires – Autres vertébrés
220	Métazoaires – Mammifères
171	Plantes
83	Protozoaires
386 686	Total

Composition et organisation des génomes

De la génomique à la génomique fonctionnelle

Le séquençage ne constitue qu'une toute première étape pour l'analyse des génomes.

Au terme d'un projet de séquençage, on obtient un "texte" formé des 4 lettres A, C, G, T (une par nucléotide), et il reste un énorme travail de décryptage pour pouvoir interpréter ce texte.

L'exemple ci-dessous montre un fragment de 1000 nucléotides du génome humain.

```
....CGATGCTAAACATTCAATTAGTCAAAATGCCCTAGGTTAGCACAGCAATGTAGGTGCCAAACTC  
ATCGCAATGGAATTGCAAGCGGGAGAACAAAGGACGCCCTGCCTCCCTTGCAATAGTCGATTGA  
AAAGGGACCCACAGAGACACAAAATGCAACGCCACATCTTTACCCGCAATGGGTAAGACTGTC  
AACAGGCAGGCCACCTCGCAGCGTCCGGAGTTGCAGGGCCGGCCGGGGCGAGCGGAGGGAGTGA  
TGGGGGGGGAGGGGGCGCCGGCCGGAGGGGGCGGGGGGGGGGGAGCGGAGGGAGTGA  
GGACGCGTAGACGCCCGCGGTCCCCCGCTGCCGCTGTCGCCGCACTGCAGCTCCAGTCTATCCGCACTAGGA  
ACAGCCCAGGGCGAGACGGTCCCCGATGTCTGCCGATGAGGGAGAGGGTCTGACCGGTTCTGCACGAGAA  
GAAGCTCATGACTGACCTCTGGCAAGCTGAGGGAAAACCGGGCTGAACAGGACTCATCGCTCTGGTGGGT  
GGCCGGGGTGGCCGGCTGGTAGGGCACGGGAGCCGCTGGCCAGCTGCTGGGAAGGAAGCAGGGAGAGG  
ACTCGGGAAAGGTGGAGTCGGAGACAGACGGGACAAGCAGCATATTAGGGATCAGGCTGGCTCCCGAAAGCGTG  
GGCATCGAGGACCCCGGGGGCTCCAGGCTGAGGGTGGGGCTGGAGGGCAGCTGCCGCGCCGGCGCTGG  
CAGCTGGAAGGGCAGCGCTGACGTATGTCGCCCGGGCCGCGCTATTCTGCTGCTGCCGCGGTGGCG  
CGGACCGGGGGCCCTGGGGCGGGCGTGAACGGAGGTACCCGCTTACCCGACCCCTCGTGGAGCTCCGCC  
GGAG....
```

Le génome complet comporte 3 milliards de nucléotides, 3 millions de fois plus grand.

Les premières questions qui se posent au terme du séquençage =

- 1. Où sont localisés les gènes ?**
- 2. Quelle est la fonction de ces gènes ?**



[Drew Sheneman, New Jersey -- The Newark Star Ledger](#)

Annotation des génomes : où sont les gènes ?

Méthodes mises à contribution pour localiser les gènes dans un génome

A partir de la séquence « brute » d'un génome comment prédire la position des gènes ?

- Présence de **phases ouverte de lecture (longues régions sans codon stop)** indiquent des régions vraisemblablement codantes.
- **Fréquences de codons** sont caractéristiques des régions codantes.
- Fréquences des oligonucléotides.
 - Par exemple, les fréquences d'hexanucléotides diffèrent entre régions codantes et non-codantes.
- Présence de **signaux**
 - Chez les procaryotes: juste avant une région codante, on trouve parfois un motif appelé « boîte de Shine-Delgarno » (AGGAGGU), qui favorise la liaison du ribosome à l'ARN
 - Chez les eucaryotes, on peut détecter des signaux d'épissage qui indiquent les débuts et fins des exons
- **Recherche de similarité** avec des gènes connus.
 - Comparaison d'une séquence génomique avec tout ce qui a été préalablement séquencé → détection de correspondances avec séquences déjà connues.
- **Génomique comparative** : comparaison entre génomes entiers
- **Transcriptome** : localisation (“mapping”) de toutes les régions génomiques transcris dans différents tissus

Cadres ouverts de lecture (open reading frame)

Une séquence nucléique (ADN ou ARN) peut être parcourue en avançant de triplet en triplet de nucléotide, selon trois **cadres de lecture**, ou **phases de lecture**, selon qu'on parte du premier, du deuxième ou du troisième nucléotide de la séquence.

Pour les séquences d'ADN, il y a donc 6 cadres de lecture (3 sur chaque brin).

Un **cadre de lecture ouvert (open reading frame, ORF)** est un segment de séquence nucléique qui n'est pas interrompu par un **codon stop** (TAG, TGA ou TAA) dans une phase de lecture donnée, et est donc "ouvert" à la traduction (Sieber et al. 2018).

Quand on dispose d'un génome ou d'un fragment de génome, les séquences codantes (**coding sequences, CDS**) peuvent être identifiées en cherchant le cadre ouvert de lecture le plus long à partir d'un **codon start** potentiel (ATG) et du prochain codon stop.

Difficultés

- Tous les codons ATG ne sont pas des codons start, il existe des méthionines internes aux protéines. On prend donc généralement en compte le **plus long cadre de lecture** (depuis le codon start le plus éloigné en amont du codon stop)
- Chez les eucaryotes, les introns n'ont pas forcément une longueur multiple de 3, une protéine peut donc combiner des **cadres ouverts de lecture situés sur différentes phases** de la séquence génomique.
- Il existe des **codons start alternatifs** (exemple, chez Escherichia coli, ATG=85%, GTG=7,6%, TTG=1.2%, ...)

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

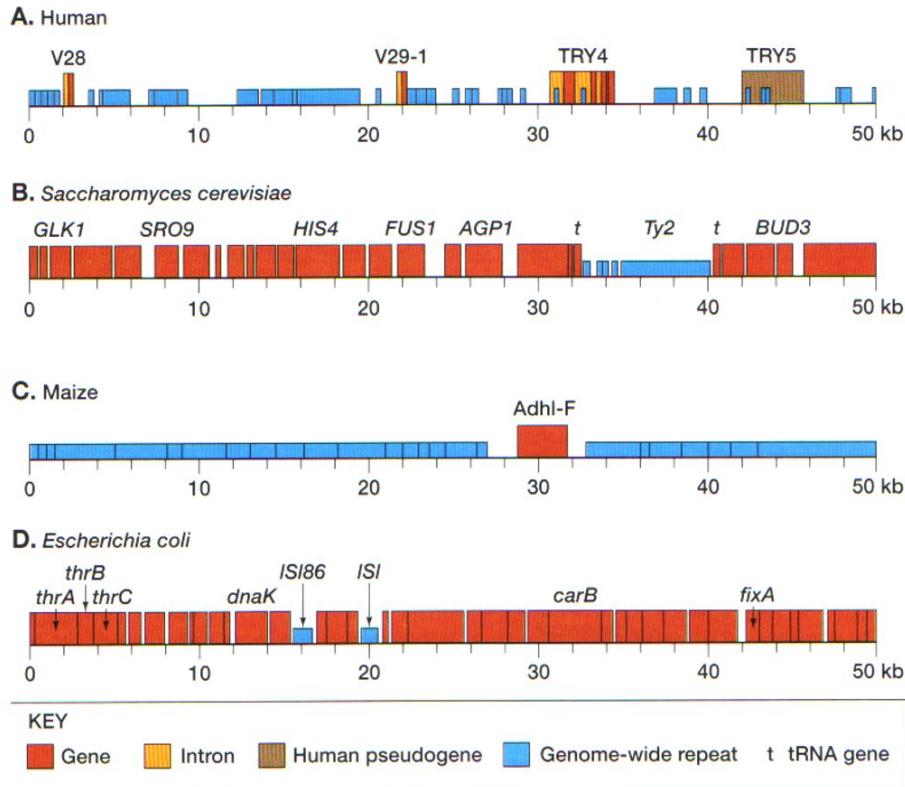
Taille et composition des génomes

Nom d'espèce	Nom commun	Année de publication	Taille du génome Mb	Nombre de gènes	Distance moyenne Kb	Fraction codante %	Fraction non-codante %	Fraction répétitive %	Fraction transcrive %
Bactéries									
<i>Mycoplasma genitalium</i>	Mycoplasma	1995	0,6	481	1,2	90	10		
<i>Haemophilus influenzae</i>		1995	1,8	1 717	1,0	86	14		
<i>Escherichia coli</i>	Entérobactéries	1997	4,6	4 289	1,1	87	13		
Levures									
<i>Saccharomyces cerevisiae</i>	Levure du boulanger	1996	12	6 286	1,9	72	28		
Animaux									
<i>Caenorhabditis elegans</i>	Ver nématode	1998	97	19 000	5	27	73		
<i>Drosophila melanogaster</i>	Mouche à vinaigre	2000	165	16 000	10	15	85		
<i>Ciona intestinalis</i>			174	14 180	12				
<i>Danio rerio</i>	Poisson zèbre		1 527	18 957	81				
<i>Xenopus laevis</i>	Xénopé (amphibiens)		1 511	18 023	84				
<i>Gallus gallus</i>	Poule		2 961	16 736	177				
<i>Ornithorynchus anatinus</i>	Ornithorynque		1 918	17 951	107				
<i>Mus musculus</i>	Souris	2002	3 421	23 493	146				
<i>Pan troglodytes</i>	Chimpanzé		2 929	20 829	141				
<i>Homo sapiens</i>	Humain	2001	3 200	21 528	149	2	98	46	28
Plantes									
<i>Arabidopsis thaliana</i>	Arabette	2001	120	27 000	4	30	70		
<i>Oryza sativa</i>	Riz		390	37 544	10				
<i>Zea mays</i>	Maïs		2 500	50 000	50			50	
<i>Triticum aestivum</i>	Blé		16 000						
<i>Lilium</i>	Lys		120 000						
<i>Psilotum nudum</i>			250 000						

Structuration des génomes

La structure des génomes dépend fortement du groupe taxonomique

- Bactéries (*Escherichia coli*)
 - génomes compacts
 - majorité codante
 - Organisation en opérons
- Levures (*Saccharomyces cerevisiae*)
 - Régulation séparée pour chaque gène
 - Exons / introns occasionnels ou fréquents selon espèce
- Métazoaires – animaux pluricellulaires (ex: humain)
 - Majorité non codante
 - Éléments répétitifs
 - Structure complexe des gènes (exons / exons, éléments de régulation)
- Plantes (ex: maïs)
 - Même type de complexité que chez les métazoaires



Annotations fonctionnelles : que font les gènes ?

Attribution de fonction par similarité de séquences

Alignements globaux (Needleman-Wunsch) versus locaux (Smith-Waterman)

- Alignement global

- +Approprié, par exemple, pour les protéines homologues qui sont conservées sur toute leur longueur.
- L'alignement final inclut obligatoirement les deux séquences complètes.

LQGPSK**TGKGS**-SRSWDN

| - - - | - | | | - - - | - - | -

LN-ITKAG**KGAIMRLGDA**

- Algorithme: **Needleman-Wunsch** (1970).
- Outil web EMBOSS : **needle** ([nucleic acids](#) or [proteins](#)).

- Alignement local

- Approprié, par exemple, pour les protéines qui partagent un domaine commun, restreint à un segment de chaque séquence.

LQGPSS**KTGKGS**-SSRIWDN

| - | | |

LN-ITK**KAGKG**AIMRLGDA

- L'alignement final est restreint aux segments conservés.

- KTGKG

- | - | | |

- KAGKG

- Algorithme: **Smith-Waterman** (1981).

- Outil Web EMBOSS : **water** ([nucleic acids](#) or [proteins](#))

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-53. [doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-7.

Alignement de séquences – Gènes S de SARS-CoV-2 et RaTG13c

```
# Aligned_sequences: 2
# 1: Human_SARS-CoV-2_BetaCoV/Wuhan/IPBCAMS-WH-01/2019
# 2: Bat_RaTG13
#
# Length: 3822
# Identity: 3549/3822 (92.9%)
# Similarity: NA/3822 (NA%)
# Gaps: 12/3822 (0.3%)
# Score: 5435.624
```

		Identités	
Human_SARS-CoV-2	1	ATGTTTGT TTTCTT ATTGCCACTAGTCCTCTAGTCAGTGTTAA	50
Bat_RaTG13	21545	ATGTTTGT TTTCTT ATTGCCACTAGTCCTCTAGTCAGTGTTAA	21594
...			
Human_SARS-CoV-2	2001	TGCAGGTATATGCGCTAGTTATCAGACTCAGACTAATTCTCCTCGGCGGG	2050
Bat_RaTG13	23545	TGCAGGAATATGCGC CA AGTTATCAGACTCAAACTAATT C -	23583
		Indel	
Human_SARS-CoV-2	2051	CACGTAGTGTAGCTAGTCAATCCATCATTGCCCTACACTATGTCAC TTGGT	2100
Bat_RaTG13	23584	-ACGTAGTGTGGCCAGTCAATCTATTATGCCCTACACTATGTCAC TTGGT	23632

Note

- “Indel” signifie “Insertion ou délétion”
- Sur base de ce seul résultat, one ne peut pas déterminer si la différence observée provient d'une insertion chez un ancêtre de SARS-CoV-2, ou d'une délétion chez un ancêtre de RaTG13

Alignement d'une paire de séquences protéiques

- Protéines metL et thrA d'E.coli
- Algorithme : Needleman-Wunsch.
- Barres verticales « | »
 - **Identité:** les deux résidus alignés sont identiques.
- Doubles points « : »
 - **Substitution « conservative »**
 - Les deux résidus alignés sont différents mais *similaires* (la paire de résidus a un score positif dans la matrice de substitution utilisée (ici, BLOSUM62). Voir plus loin pour comprendre ces matrices.
- Points « . »
 - **Substitution non-conservative**
 - Cette paire de résidus (distincts) a un score négatif dans la matrice de substitution.
- Espace: « »
 - **Gap:** les résidus d'une des deux séquences ne correspondent à aucun résidu sur l'autre.
 - Le gap peut provenir soit d'une délétion, soit d'une insertion, on parle donc d'*indel*, pour désigner l'événement évolutif d'où provient ce gap.

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 867
# Identity: 254/867 (29.3%)
# Similarity: 423/867 (48.8%)
# Gaps: 104/867 (12.0%)
# Score: 929.0
```

	metL	thrA	
1	MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDM-MVVSA	MRVLKFGGTSVANAERFLRVADILESNDARQGVATVLSA	49
50	AGSTTNQLINWLK-----LSQTDRLSAHQVQQTLRRYQCDLISG::: ...: ...:: ...:	88
40	PAKITNHLVAMIEKTISGQDALPNISDAERIFA-----ELLTG	40 PAKITNHLVAMIEKTISGQDALPNISDAERIFA-----ELLTG	77
89	LLPAEEADSL--ISAFV-SDLERLAALLDSGIN-----DAVYAEVVGHG	129
78	LAAAQPGFPLAQLKTFVDQEFAQIKHVL-HGISLLGQCPDSINAALICRG	78 LAAAQPGFPLAQLKTFVDQEFAQIKHVL-HGISLLGQCPDSINAALICRG	126
130	EVWSARLMSAVLNQQGLPAAWLD-AREFLRAERAAPQVD--EGLSYPLL	176
127	EKMSIAIMAGVLEARGNVTIDPVEKLLAVGHYLESTVDIAESTRIAA	127 EKMSIAIMAGVLEARGNVTIDPVEKLLAVGHYLESTVDIAESTRIAA	176
177	QQLLVQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRV	226
177	SRIPADH---MVLMAFGTAGNEKGELVVLGRNGSDYSAAVLAACLRADCC	177 SRIPADH---MVLMAFGTAGNEKGELVVLGRNGSDYSAAVLAACLRADCC	223
227	TIWSDVAGVYSADPRKVKDACLPLRLDEASELARLAAPVLHARTLQPV	. : . : : . : . : . : . : . : . :	276
224	EIWTDVDGVYTCDPRQPDARLLKSMSYQEAMELSYFGAKVLHPRTITPI	224 EIWTDVDGVYTCDPRQPDARLLKSMSYQEAMELSYFGAKVLHPRTITPI	273

Matrice de substitutions

- Une **matrice de substitution** associe un score à chaque paire de résidus qu'on peut trouver dans un alignement.
 - Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acide aminés).
 - La diagonale correspond aux identités.
 - Le triangle inférieur correspond à des substitutions.
 - Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.
 - Les **scores négatifs** sont considérés comme des pénalités associées à certaines substitutions qu'on n'observe que rarement dans les alignements. Les algorithmes d'alignements tenteront donc d'éviter ces substitutions.
 - Les **scores positifs** correspondent à deux types d'alignements
 - les **identités** ont toujours un score positif
 - certaines substitutions qu'on observe plus souvent que prévu, dans les alignements d'un grand nombre de séquences. Ceci suggère que ces substitutions particulières sont moins dommageable que d'autres, et on les qualifie donc de « **substitutions conservatives** » ou encore de « **mutations ponctuelles acceptées** » (**PAM**).
 - Au sein d'un alignement, le terme **similarité** désigne les positions où se superposent des résidus ayant un score positif dans la matrice de substitution (identité ou substitution conservative).

Matrice de substitutions entre nucléotides

	A	C	G	T	Scores
A	1	-2	-1	-2	1 -1 -2
C	-2	1	-2	-1	identité transition
G	-1	-2	1	-2	
T	-2	-1	-2	1	transversion

Matrice de substitutions entre acides aminés

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i de 1 à L), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
 - Pénalité d'ouverture de gap (**go**)
Valeur typique : -10
 - Pénalité d'extension de gap (**ge**)
valeur typique: -1

Exercice

Dans l'alignement ci-dessous,

- identifiez les identités, les transitions et les transversions
- en vous basant sur la matrice de substitution, calculez le score de l'alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores

1	identité
-1	transition
-2	transversion

$$S = \sum_{i=1}^L s_{r_{1,i} r_{2,i}}$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	T	A	A	C	T	T	A	G	G	-	-	-	-	C	C	C	A	T	G		
A	T	G	C	C	T	G	A	G	G	A	T	T	A	C	C	A	G	T	G		

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (*i* de 1 à *L*), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).

$$S = \sum_{i=1}^L s_{r_{1,i} r_{2,i}}$$

<i>i</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	A	A	C	T	T	A	G	G	-	-	-	-	C	C	C	A	T	G	
A	T	G	C	C	T	G	A	G	G	A	T	T	A	C	C	A	G	T	G	

+1 +1 -1

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (*i* de 1 à *L*), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).

$$S = \sum_{i=1}^L s_{r_{1,i} r_{2,i}}$$

<i>i</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	A	A	C	T	T	A	G	G	-	-	-	-	C	C	C	A	T	G	
A	T	G	C	C	T	G	A	G	G	A	T	T	A	C	C	A	G	T	G	
+1	+1	-1	-2	+1	+1	-2	+1	+1	+1											

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i de 1 à L), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
 - Pénalité d'ouverture de gap (**GO**)
Valeur typique : -10

$$S = \sum_{i=1}^L s_{r_{1,i} r_{2,i}}$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	A	A	C	T	T	A	G	G	-	-	-	-	C	C	C	A	T	G	
														GO						
A	T	G	C	C	T	G	A	G	G	A	T	T	A	C	C	A	G	T	G	

+1 +1 -1 -2 +1 +1 -2 +1 +1 +1 -10

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores
1
-1
-2

identité
transition
transversion

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i de 1 à L), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
 - Pénalité d'ouverture de gap (**GO**)
Valeur typique : -10
 - Pénalité d'extension de gap (**ge**)
valeur typique: -1

$$S = \sum_{i=1}^L s_{r_{1,i} r_{2,i}}$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	A	A	C	T	T	A	G	G	-	-	-	-	C	C	C	A	T	G	
														GO	ge	ge	ge			
A	T	G	C	C	T	G	A	G	G	A	T	T	A	C	C	A	G	T	G	

+1 +1 -1 -2 +1 +1 -2 +1 +1 +1 -10 -1 -1 -1

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i de 1 à L), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
 - Pénalité d'ouverture de gap (**GO**)
Valeur typique : -10
 - Pénalité d'extension de gap (**ge**)
valeur typique: -1

$$S = \sum_{i=1}^L s_{r_{1,i} r_{2,i}}$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	A	A	C	T	T	A	G	G	-	-	-	-	C	C	C	A	T	G	
										GO	ge	ge	ge							
A	T	G	C	C	T	G	A	G	G	A	T	T	A	C	C	A	G	T	G	
+1	+1	-1	-2	+1	+1	-2	+1	+1	+1	-10	-1	-1	-1	+1	+1	-2	-1	+1	+1	= -10

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

Ala	A	4																		
Arg	R	-1 5																		
Asn	N	-2 0 6																		
Asp	D	-2 -2 1 6																		
Cys	C	0 -3 -3 -3 9																		
Gln	Q	-1 1 0 0 -3 5																		
Glu	E	-1 0 0 2 -4 2 5																		
Gly	G	0 -2 0 -1 -3 -2 -2 6																		
His	H	-2 0 1 -1 -3 0 0 -2 8																		
Ile	I	-1 -3 -3 -3 -1 -3 -3 -4 -3 4																		
Leu	L	-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4																		
Lys	K	-1 2 0 -1 -3 1 1 -2 -1 -3 -2 5																		
Met	M	-1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5																		
Phe	F	-2 -3 -3 -3 -2 -3 -3 -1 0 0 -3 0 6																		
Pro	P	-1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7																		
Ser	S	1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4																		
Thr	T	0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 1 5																		
Trp	W	-3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11																		
Tyr	Y	-2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7																		
Val	V	0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4																		
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	

$$S = \sum_{i=1}^L S_{r_{1,i}, r_{2,i}}$$

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i de 1 à L), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
 - Pénalité d'ouverture de gap (GO)
Valeur typique : -10
 - Pénalité d'extension de gap (ge)
valeur typique: -1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
R	L	A	S	V	E	T	D	M	P	-	-	-	-	-	L	T	L	R	Q	H	
T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H	

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

Ala	A	4
Arg	R	-1 5
Asn	N	-2 0 6
Asp	D	-2 -2 1 6
Cys	C	0 -3 -3 -3 9
Gln	Q	-1 1 0 0 -3 5
Glu	E	-1 0 0 2 -4 2 5
Gly	G	0 -2 0 -1 -3 -2 -2 6
His	H	-2 0 1 -1 -3 0 0 -2 8
Ile	I	-1 -3 -3 -3 -1 -3 -3 -4 -3 4
Leu	L	-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4
Lys	K	-1 2 0 -1 -3 1 1 -2 -1 -3 -2 5
Met	M	-1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5
Phe	F	-2 -3 -3 -3 -2 -3 -3 -1 0 0 -3 0 6
Pro	P	-1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7
Ser	S	1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4
Thr	T	0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 1 5
Trp	W	-3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11
Tyr	Y	-2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7
Val	V	0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4
	Ala	
	Arg	
	Asn	
	Asp	
	Cys	
	Gln	
	Glu	
	Gly	
	His	
	Ile	
	Leu	
	Lys	
	Met	
	Phe	
	Pro	
	Ser	
	Thr	
	Trp	
	Tyr	
	Val	

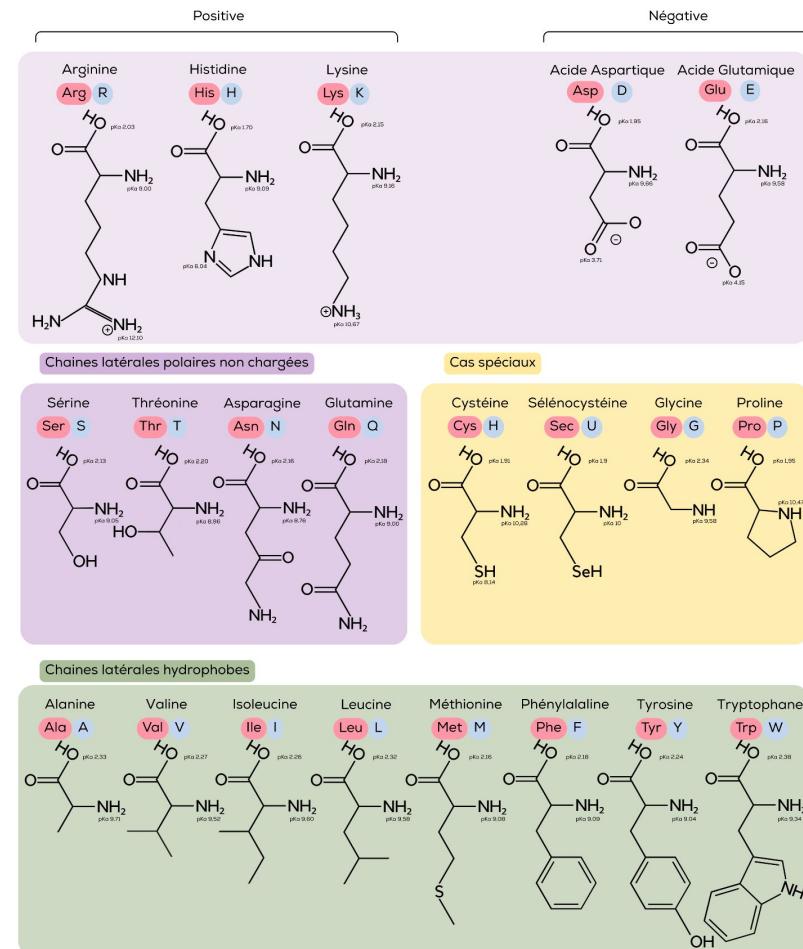
$$S = \sum_{i=1}^L S_{r_{1,i}, r_{2,i}}$$

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i de 1 à L), des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
 - Pénalité d'ouverture de gap (GO)
Valeur typique : -10
 - Pénalité d'extension de gap (ge)
valeur typique: -1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
R	L	A	S	V	E	T	D	M	P	-	-	-	-	-	L	T	L	R	Q	H	
.		.		:	:		.	:	.	GO	ge	ge	ge	ge		
T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H	
S	-1	+4	+0	+4	+1	+2	+5	-1	+2	-1	-10	-1	-1	-1	-1	-2	+4	-2	-1	+8	= 7

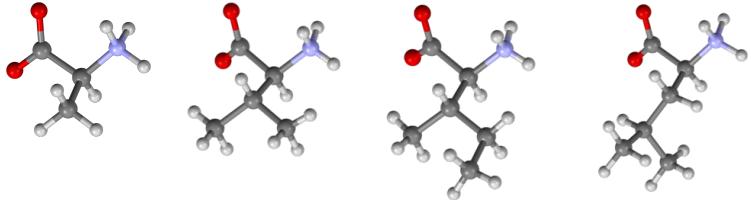
Rappel – Nomenclature et composition des acides aminés

Amino Acid	Abbrev	1-lettre	Codon(s)
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	CGA, CGC, CGG, CGT, AGA, AGG
Aspartic acid	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cysteine	Cys	C	TGC, TGT
Glutamic acid	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATA, ATC, ATT
Leucine	Leu	L	CTA, CTC, CTG, CTT, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Serine	Ser	S	TCA, TCC, TCG, TCT, AGC, AGT
Threonine	Thr	T	ACT, ACC, ACG, ACT
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA

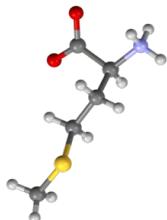


Similarités chimiques entre acides aminés et matrices de substitutions

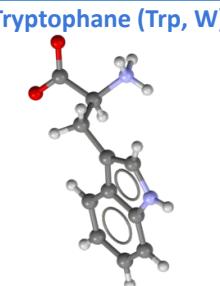
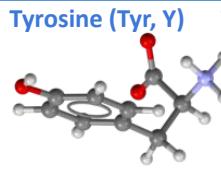
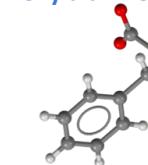
Alanine (Ala, A) Valine (Val, V) Isoleucine (Ile, I) Leucine (Leu, L) Méthionine (Met, M)



Méthionine (Met, M)

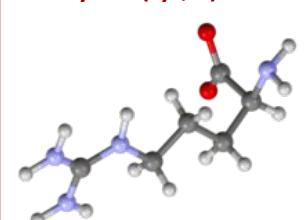


Phénylalanine (Phe, F)

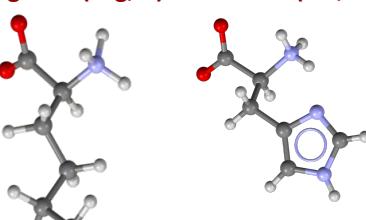


Hydrophobes

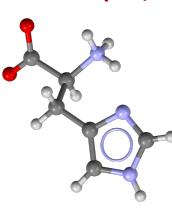
Lysine (Lys, K)



Arginine (Arg, R)

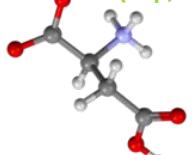


Histidine (His, H)

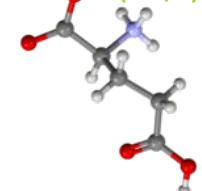


Chargés positivement

Acide aspartique (Asp, D)



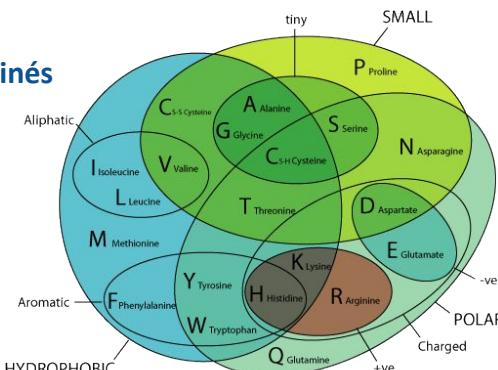
Acide glutamique (Glu, E)



Chargés négativement

Matrice de substitutions entre acides aminés

	Ala	A	V	I	L	M	F	T	Y	W	
Ala	4										A
Arg	-1	5									R
Asn	-2	0	6								N
Asp	-2	-2	1	6							D
Cys	0	-3	-3	-3	9						C
Gln	-1	1	0	0	-3	5					Q
Glu	-1	0	0	2	-4	2	5				E
Gly	0	-2	0	-1	-3	-2	-2	6			G
His	-2	0	1	-1	-3	0	0	-2	8		H
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	I
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	4	L
Lys	-1	2	0	-1	1	1	-2	-1	-3	-2	K
Met	-1	-1	-2	-3	-1	0	-2	-3	1	2	M
Phe	-2	-3	-3	-3	-2	-3	-3	-3	0	6	F
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-3	-3	P
Ser	1	-1	1	0	-1	0	0	-1	-2	-2	S
Thr	0	-1	0	-1	-1	-1	-2	-2	-1	-1	T
Trp	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	W
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	-1	-2	Y
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	1	V
	A	R	Z	N	D	S	T	W	Y	V	



Exemple de résultat de BLAST Requête peptidique vs DB de peptides

- La ligne entre les séquences "Query" et "Sbjct" indique les correspondances entre acides aminés.

- Identités**
- Substitutions "conservatives":** paires de résidus distincts mais dont la substitution est généralement moins délétère que pour d'autres paires de résidus.

- Substitutions non conservatives**

- Positives:** identités + substitutions conservatives.

- Gaps:** espaces (symboles -) insérés dans une séquence afin d'optimiser l'alignement des fragments avoisinants.

Note: les modalités de représentation des identités et substitutions conservatives ou non conservatives peuvent varier d'un outil logiciel à l'autre.

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I  
(N-terminal); homoserine dehydrogenase I (C-terminal)  
[Escherichia coli K12]  
Length = 820  
  
Score = 344 bits (882), Expect = 2e-95  
Identities = 247/821 (30%) Positives = 410/821 (49%) Gaps = 44/821 (5%)  
  
Query: 16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74  
KF GG+S+A+ + LRV A I+ ++ + V+SA TN L+ ++ + + + + +  
Sbjct: 5 KFGGTSVANAERELRVADILESMARQGVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64  
  
Query: 75 QQTLRRYQCCDLISGLLPAAEADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126  
R + +L++GL A+ L + FV + GI+ D++ A ++  
Sbjct: 65 SDAERIF-AELLTGLAAQPGFPLAQLKTFVDQEFAQIKHVLHGILLGQCPDSINAALI 123  
  
Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183  
GE S +M+ VL +G +D E L A + + E ++ H  
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183  
  
Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243  
+++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V  
Sbjct: 184 ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYCDPRQV 240  
  
Query: 244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ----GSTRI 298  
DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R  
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300  
  
Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358  
E L + +++ +++ + P + + + RA++ + + +  
Sbjct: 301 EDELPA---VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSEY 356  
  
Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411
```

Exemple de résultat de BLAST Requête peptidique vs DB de peptides

Exemple de résultat de recherche par similarité de séquences.

- Requête (**query**): metA
- Protéine identifiée dans la base de données: (**subject**): thrA.

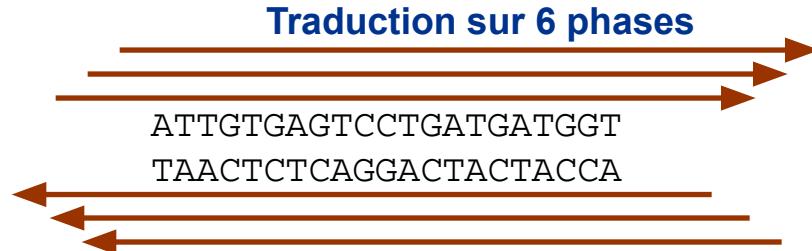
Le premier critère d'évaluation d'un résultat de BLAST:

- La **e-valeur (expect)** indique le nombre de faux-positifs attendus au hasard, si l'on plaçait le seuil au niveau du score observé (**344 bits** dans ce cas-ci).
- **Plus la e-valeur est faible, plus le résultat est statistiquement significatif.** Dans le cas présent, il est très significatif (**Expect = 2e-95**)
- **Si la e-valeur est ≥ 1 , le résultat n'est pas significatif** (on s'attendrait à trouver un alignement « aussi bon » avec des séquences aléatoires).

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I  
(N-terminal); homoserine dehydrogenase I (C-terminal)  
[Escherichia coli K12]  
Length = 820  
  
Score = 344 bits (882), Expect = 2e-95  
Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)  
  
Query: 16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74  
KFGG+S+A+ + +LRVA I+ ++ + V+SA TN L+ ++ + + + + +  
Sbjct: 5 KFGGTSVANAERFLRVADILESNDARQGVATVLSAPAKITNHVLVAMIEKTISGQDALPNI 64  
  
Query: 75 QQTLRRYQCQLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126  
R + +L++GL A+ L + FV + GI+ D++ A ++  
Sbjct: 65 SDAERIF-AELLTGLAAAQPGFPLAQQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123  
  
Query: 127 GHGEVWSARLMSAVLNQOGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183  
GE S +M+ VL +G +D E L A + + E ++ H  
Sbjct: 124 CRGEKMSIAIMAGVLEARGNVTIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183  
  
Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243  
+++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V  
Sbjct: 184 ---MVL MAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYCDPRQV 240  
  
Query: 244 KDACLLPLLRLEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298  
DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R  
Sbjct: 241 PDARLLKSMSSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300  
  
Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358  
E L + +++ +++ + P + + + RA++ + + +  
Sbjct: 301 EDEL P----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSEY 356  
  
Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGE LRLRQGLALVAMVGAGVTRNPLHCHRF 411
```

Traduction d'une séquence nucléique sur les 6 phases de lecture

- Si l'on dispose d'une séquence nucléique, on peut facilement déduire la séquence de la protéine qui pourrait être produite par sa traduction, sur chacun des 6 brins.
- Si cette séquence n'est pas codante, on s'attend à trouver des codons stop (noté *) assez fréquemment (3 codons sur 64).
- Les similarités entre une séquence traduite à partir d'ADN et des protéines connues constituent des indices pour la localisation de régions codantes, et pour la fonction potentielle des nouvelles séquences.



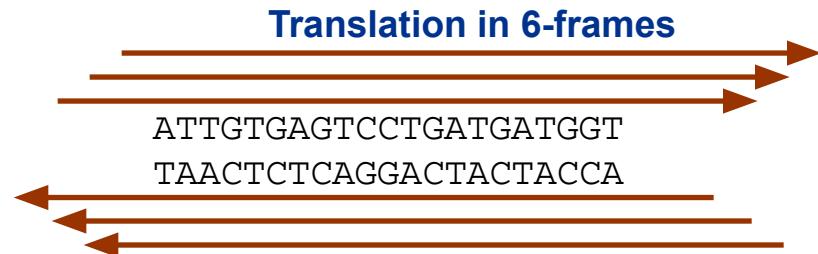
Résultat

F1	I	V	S	P	D	D	G
F2	L	*	V	L	M	M	V
F3	C	E	S	*	*	W	X
1	ATTG	TGA	GTCC	TGA	TGA	TGGT	21
	-----	:-----	-----	:-----	-----	-----	
1	TAAC	ACTCAGG	ACTACTACCA	21			
F6	X	T	L	G	S	S	P
F5	X	Q	S	D	Q	H	H
F4	N	H	T	R	I	I	T

Modalités de BLAST

Le logiciel BLAST présente 5 modalités différentes en fonction du type des séquences (peptidique ou nucléotidique) de requête et de la base de données.

Pour les comparaisons entre séquences nucléotidiques et peptidiques, la séquence nucléotidique est traduite dans les 6 phases de lecture (3 par brin), et on lance ensuite une recherche de similarité “protéine *versus* protéine”.



Séquence requête	Base de données	Outil	Exemples d'applications
peptidique	peptidique	blastp	En partant d'une protéine de fonction connue, collecter les protéines similaires dans la base de données Uniprot afin de constituer la famille de protéine supposées homologues.
nucléique	nucléique	blastn	Comparer les séquences d'ARNm aux séquences génomiques. Aligner un ARN d'interférence (ARNi) sur un génome pour détecter ses cibles potentielles.
nucléique (traduite dans les 6 cadres)	peptidique	blastx	Après avoir séquencé un morceau d'ADN, chercher des fragments potentiellement codants (susceptibles de produire un polypeptide similaire à des protéines connues) dans cette séquence même si on ne connaît pas la position des gènes.
peptidique	nucléique (traduite dans les 6 cadres)	tblastn	Identifier une région génomique susceptible de coder pour un homologue d'une protéine d'intérêt. Identifier dans un génome les pseudo-gènes (gènes défectifs, qui peuvent contenir un ou plusieurs codons stop) correspondant à une protéine d'intérêt.
nucléique (traduite dans les 6 cadres)	nucléique (traduite dans les 6 cadres)	tblastx	A partir d'une séquence d'ADN, identifier des segments de régions codantes ayant une contrepartie dans un génome ou une base de données de référence

Exemple de recherche de similarité : blastp (protéine vs DB de protéines)

BLAST permet de chercher, dans une base de données, toutes les séquences similaires à une séquence d'intérêt ("query", requête). L'analyse peut se faire avec des séquences nucléiques (blastn) ou peptidiques (blastp).

La figure ci-contre affiche le début de la liste des résultats, triés par significativité statistique de la similarité de séquence (les séquences les plus similaires viennent en premier).

La significativité est estimée par la E-valeur et par score de l'alignements (bit score).

Note: une E-valeur de 0 signifie en théorie qu'il n'existe aucune probabilité d'obtenir un aussi bon alignement par hasard.

Cependant, en pratique cette valeur signifie que la E-valeur est inférieure à la limite de précision numérique de BLAST (cette limite est de l'ordre de 1e-186).

National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® » blastp suite » results for RID-EZVGUHB3016

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Edit Search Save Search Search Summary

Job Title P26367:RecName: Full=Paired box protein Pax-6;...

RID EZVGUHB3016 Search expires on 09-23 18:4 pm Download All

Program BLASTP Citation

Database nr_clustered(experimental) See details

Query ID P26367.2

Description RecName: Full=Paired box protein Pax-6; AltName: Full=A...
Molecule type amino acid

Query Length 422

Other reports Distance tree of results Multiple alignment MSA viewer

Filter Results

Percent Identity E value Query Coverage

From To From To From To

Filter Reset

Clusters Graphic Summary Alignments Taxonomy

Clusters producing significant alignments

Download Select columns Show 100

Max Score	Total Score	Query Cover	E value	Per Ident	Acc Len	Accession
1361	1361	100%	0.0	0.00%	422	NP_000271.1
1351	1351	100%	0.0	96.79%	499	EDL27748.1
1347	1347	99%	0.0	0.00%	503	NP_00135833.1
1347	1347	99%	0.0	0.00%	482	XP_011820972.1
1347	1347	99%	0.0	0.00%	484	XP_04753014.1
1347	1347	100%	0.0	94.41%	447	NP_00135848.1
1343	1343	99%	0.0	99.76%	530	KAH051664.1
1340	1340	99%	0.0	99.52%	473	XP_02119822.1
1337	1337	99%	0.0	96.76%	504	XP_034788232.1
1337	1337	99%	0.0	96.76%	496	XP_028371676.1
1337	1497	99%	0.0	96.76%	564	EPO11950.1
1337	1337	100%	0.0	98.58%	421	XP_043915455.1
1337	1337	99%	0.0	99.28%	512	XP_064019897.1
1333	1333	100%	0.0	98.10%	542	CAH2326109.1
1332	1332	100%	0.0	98.10%	422	XP_019065548.1
1331	1331	99%	0.0	96.30%	487	XP_02119821.1
1330	1330	99%	0.0	96.30%	492	XP_059577325.1
1330	1330	99%	0.0	92.89%	538	KAF382197.1
1327	1327	99%	0.0	96.06%	616	XP_057235966.1
1327	1327	99%	0.0	96.06%	570	XP_064571693.1
1327	1327	99%	0.0	96.06%	520	XP_061219598.1
1318	1318	99%	0.0	98.09%	494	XP_023502322.1
1313	1313	99%	0.0	98.16%	554	XP_064145512.1
1311	1311	99%	0.0	95.14%	815	KAK9409966.1

Attribution de fonction par similarité de séquences - intérêt et limitations

- L'attribution de fonction par similarité est la méthode principale d'annotation des génomes nouvellement séquencés.
- Elle est notamment mise à contribution pour l'annotation automatique des protéines dans la section Unreviewed (TrEMBL) d'Uniprot, qui contient l'énorme majorité des séquences connues (245 millions de séquences protéiques au 22 septembre 2024).
- Il s'agit cependant d'une inférence très approximative. On ne peut pas définir un seuil d'identité ou de similarité qui permettrait d'inférer sans ambiguïté que deux protéines ont la même fonction.
 - Exemple: quelques changements d'acides aminés dans le site actif d'une enzyme peuvent suffire à changer sa spécificité de substrat ou de produit, même si le taux global d'identité ou de similarité reste très élevé.
- On essaie donc de combiner cette première phase des annotations par des informations complémentaires fournies par d'autres approches.

Status

 Reviewed (Swiss-Prot)
(571,864)

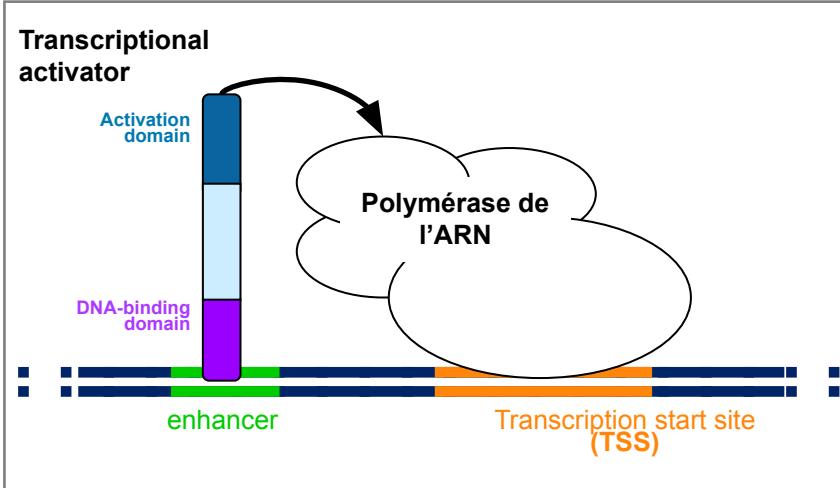
 Unreviewed (TrEMBL)
(245,324,902)

Un élément structurant des génomes: la régulation

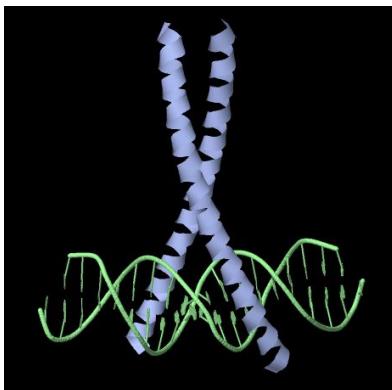
Pour pouvoir comprendre la structure des gènes et l'organisation des génomes, il est nécessaire de connaître quelques éléments concernant la régulation génétique.

Nous résumons ci-après les notions de base indispensables, sachant que ces concepts seront développés dans vos cours de génétique et de biologie moléculaire.

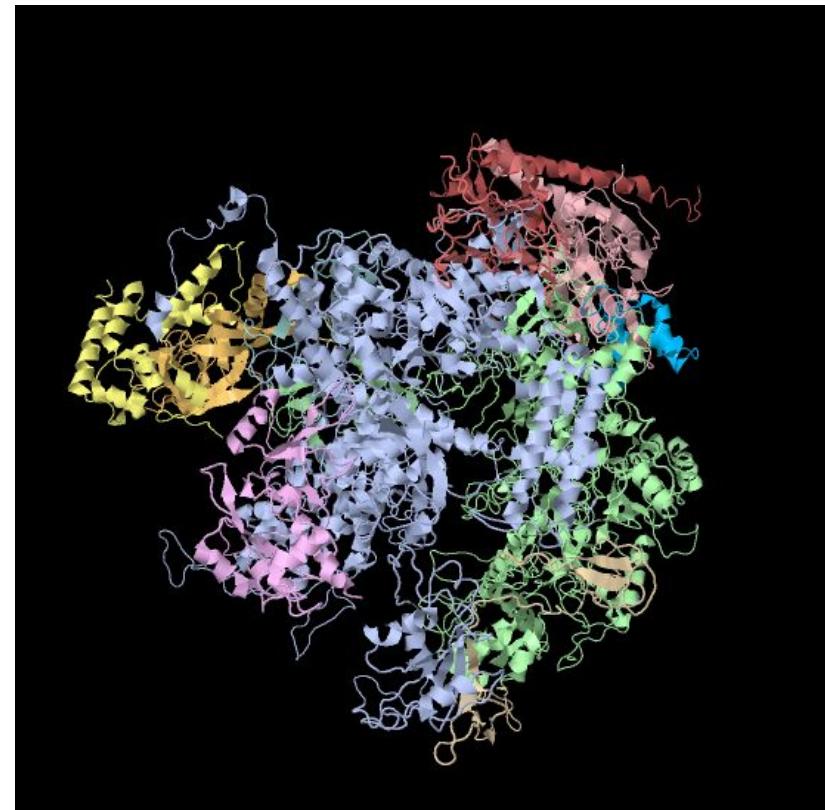
Activation de la transcription



Gcn4p from *Saccharomyces cerevisiae*
PDB 2DGC <http://www.rcsb.org/pdb/explore.do?structureId=2DGC>

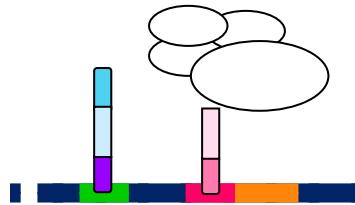


RNA polymerase II from *Schizosaccharomyces pombe*. (PDB 3H0G)
<http://www.rcsb.org/pdb/explore.do?structureId=3H0G>

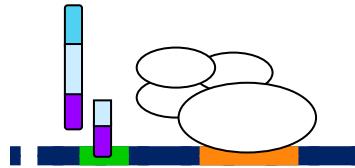


Répression transcriptionnelle

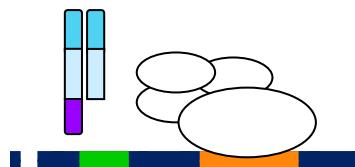
- The concept of transcriptional repression encompasses a variety of molecular mechanisms.



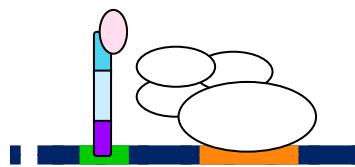
Promoter occupancy: prevent RNA polymerase from accessing DNA (e.g. many bacterial repressors)



Cis-regulatory element occupancy:
competition for factor binding site (e.g. yeast GATA factors)



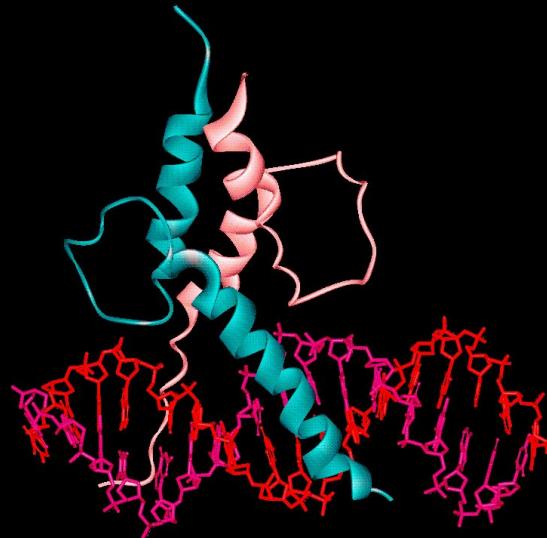
Titration of the activator: repressor forms dimer with activator, which prevents its binding to TFBS (e.g. Drosophila Helix-loop-helix)



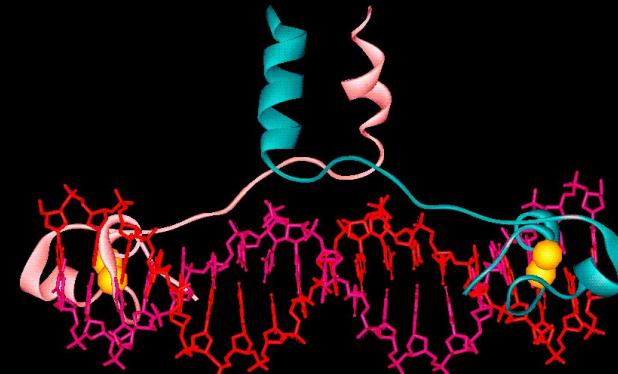
Allosteric regulation: repressor binds to activator, which alters activator conformation and prevents it from interacting with RNA-polymerase (e.g. yeast Gal80p)

Interfaces facteurs transcriptionnels / ADN

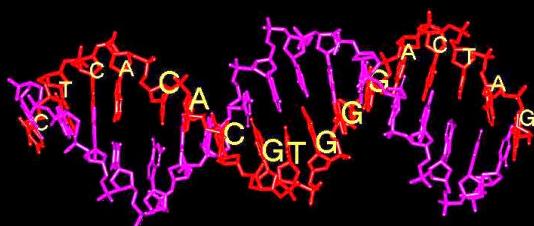
Pho4p (yeast)



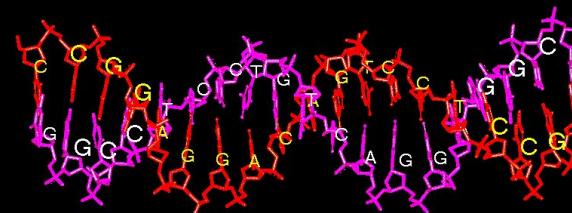
Gal4p (yeast)



Pho4p DNA binding site (oligonucleotide)



Gal4p DNA binding site (dyad)



Des génomes aux transcriptomes

Chez tous les êtres vivants l'expression des gènes fait l'objet d'un contrôle moléculaire à différents niveaux: transcription, maturation de l'ARN, traduction, post-traduction.

Une indication importante concernant la fonction des gènes est de savoir dans quelles conditions ils sont exprimés.

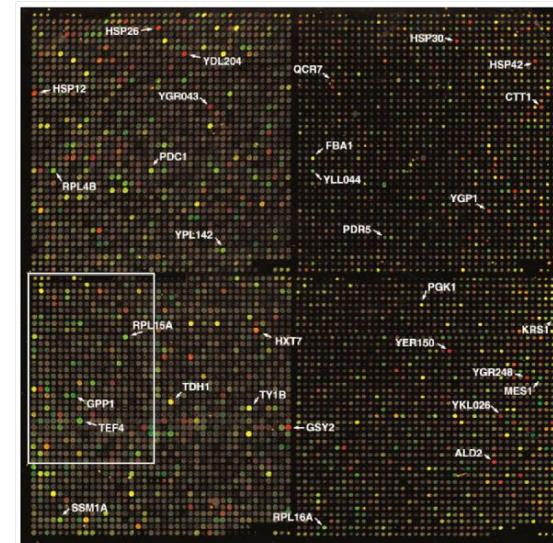
- Microbes: substrats disponibles, conditions environnementales, ...
- Multicellulaires: spécificité tissulaire, stades du développement, réponse aux conditions internes et externe de l'organisme

La transcriptomique consiste à mesurer simultanément l'expression de *tous* les gènes d'un échantillon prélevé sur un organisme dans des conditions particulières.

- 1997: premières approches de transcriptomiques par biopuces
- 2007: transcriptomique par séquençage massivement parallèle (RNA-seq)

La première biopuce transcriptomique (de Risi et al., 1997). Chacun des 6000 points lumineux correspond à un transcrit (ARN) de la levure du boulanger, *Saccharomyces cerevisiae*.

- L'intensité lumineuse est proportionnelle au niveau d'expression
- La couleur indique le sens de la régulation
 - Rouge: gènes sur-exprimés par rapport à l'échantillon témoin
 - Vert: gènes sous-exprimés
 - Jaune: gènes fortement exprimés dans les deux échantillons.
DeRisi et al. (1997), [10.1126/science.278.5338.680](https://doi.org/10.1126/science.278.5338.680)



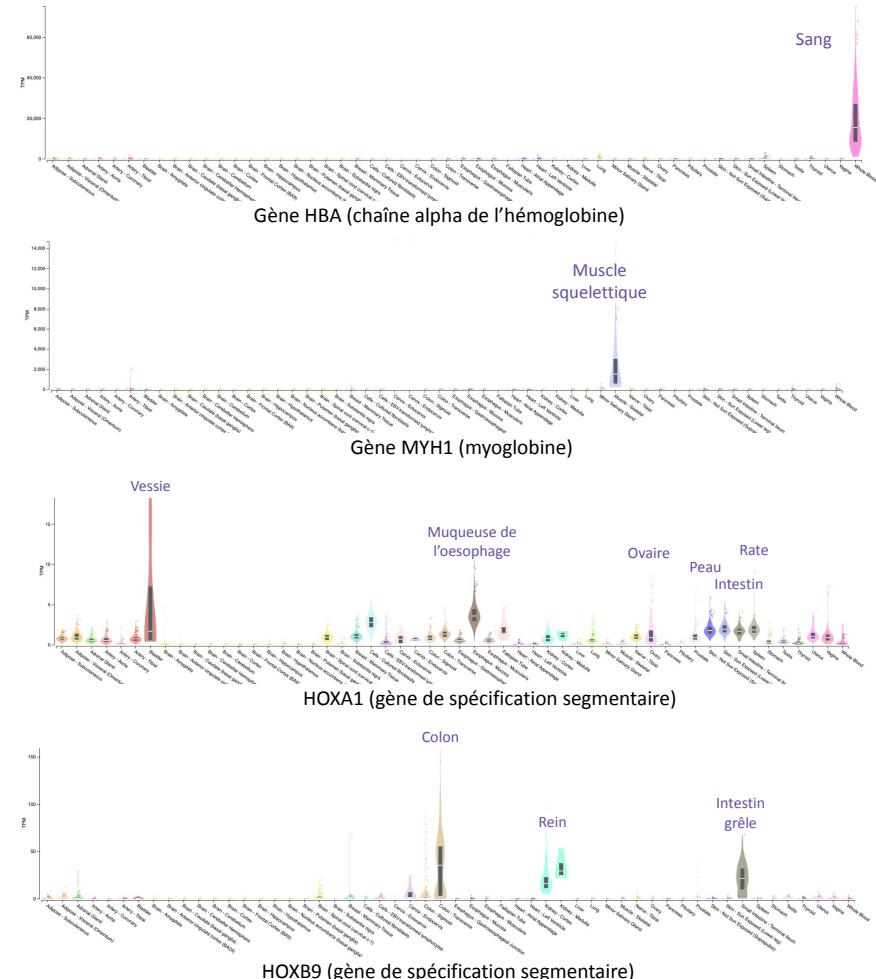
Dis-moi dans quels tissus tu t'exprimes, je te dirai qui tu es

Le projet GTEx (Adult Genotype Expression)

- Collecte d'échantillons de 54 tissus chez 1000 individus
 - Extraction de l'ARN
 - Séquençage et quantification dans chaque tissu (RNA-seq)

Exemples ci-contre: profils tissulaires d'expression pour quelques gènes illustratifs

- L'hémoglobine s'exprime uniquement dans le sang
 - La myoglobine s'exprime dans les muscles squelettiques
 - Les gènes HoxA1 et HoxB9, impliqués dans la différenciation entre segments lors du développement, sont exprimés dans des tissus différents.



Profil transcriptomique de la myoglobine

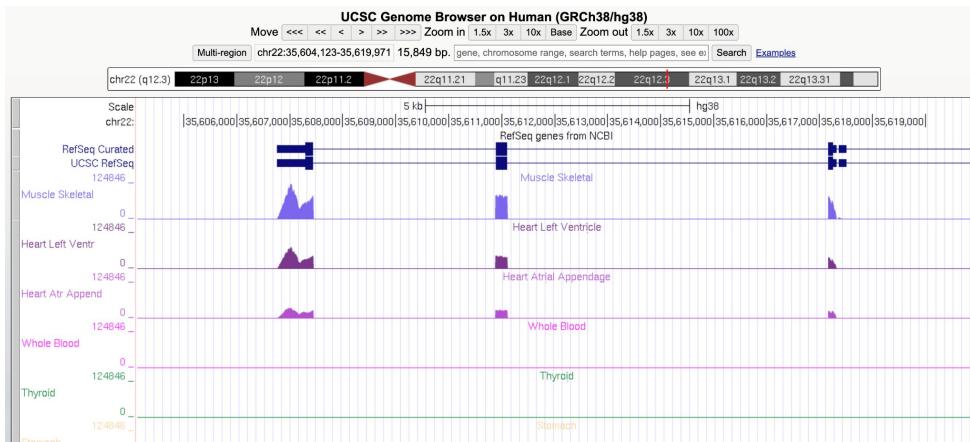
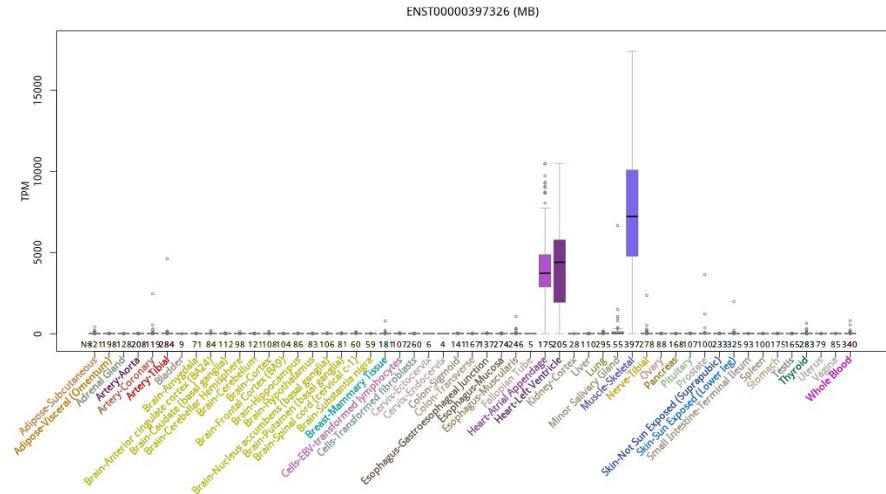
La base de données GTEx (gtexportal.org) contient les **profils transcriptomiques** (quantification de tous les ARN produits par un génome) à partir d'échantillons prélevés dans 54 tissus chez ~1000 individus.

UCSC Genome Browser (genome.ucsc.edu) permet d'afficher les données de GTEx au regard des annotations génomiques.

- **Figure du haut:** la myoglobine est fortement exprimée dans les muscles squelettiques et cardiaques. Ceci est parfaitement cohérent avec la fonction de la myoglobine.
- **Figure du bas:** dans ces tissus, la localisation génomique des fragments de lectures (short reads) correspond aux exons.

Intérêt des analyses transcriptomiques pour l'annotation des génomes : pour des gènes de fonction inconnue, les profils transcriptomiques peuvent apporter des indices concernant

- La localisation des exons
- Une fonction potentielle pour les gènes concernés



Génomique comparative

Cas d'étude : génomique comparative du gène PAX6

Le gène PAX6 code pour un facteur transcriptionnel (en violet sur l'image du haut), qui se lie à des sites spécifiques sur l'ADN génomique (vert et rose), et contrôle l'expression de gènes impliqués dans la formation de l'oeil. Les gènes cibles de PAX6 sont encore pour la plupart inconnus.

Phénotypes mutants chez la drosophile

- **Perte de fonction** : l'inactivation de *eyeless* (= PAX6) provoque une malformation ou une absence d'oeil.
- **Gain de fonction** : si on force le gène *eyeless* à s'exprimer dans les tissus larvaires précurseurs des antennes ou dans les ailes, les mouches développent à ces endroits des yeux à facettes (photo du bas).

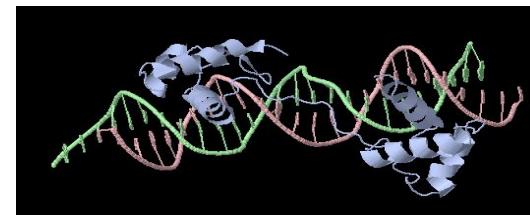
Ces phénotypes indiquent que PAX6 est le déterminant-clé de la formation de l'oeil au cours du développement de l'organisme.

Conservation du gène PAX6

- Le gène PAX6 est fortement conservé chez tous les animaux, des invertébrés aux vertébrés.
- Ceci est compréhensible étant donné son rôle crucial pour le développement de l'organe de la vision, qui est essentiel à la survie des métazoaires.

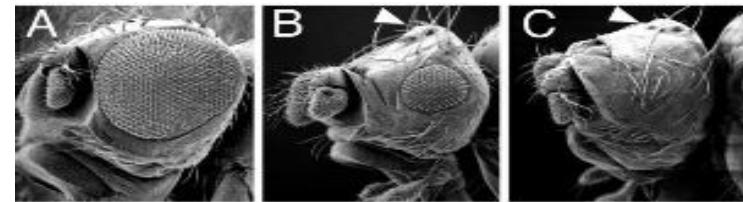
Halder et al. Eyeless initiates the expression of both *sine oculis* and *eyes absent* during Drosophila compound eye development. Development (1998) vol. 125 (12) pp. 2181-91.
Halder, Callaerts and Gehring (1995). Science, 267, 1788–1792.

Liaison PAX6 - ADN

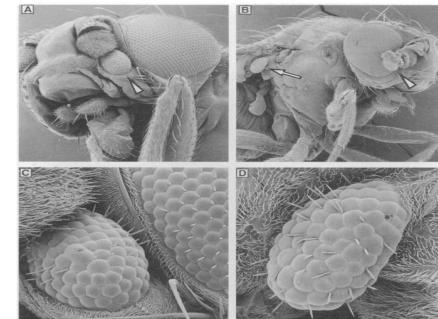


<http://www.rcsb.org/pdb/explore.do?structureId=6PAX>

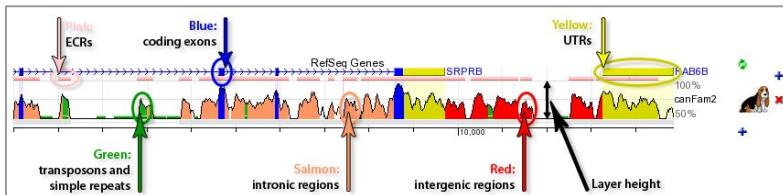
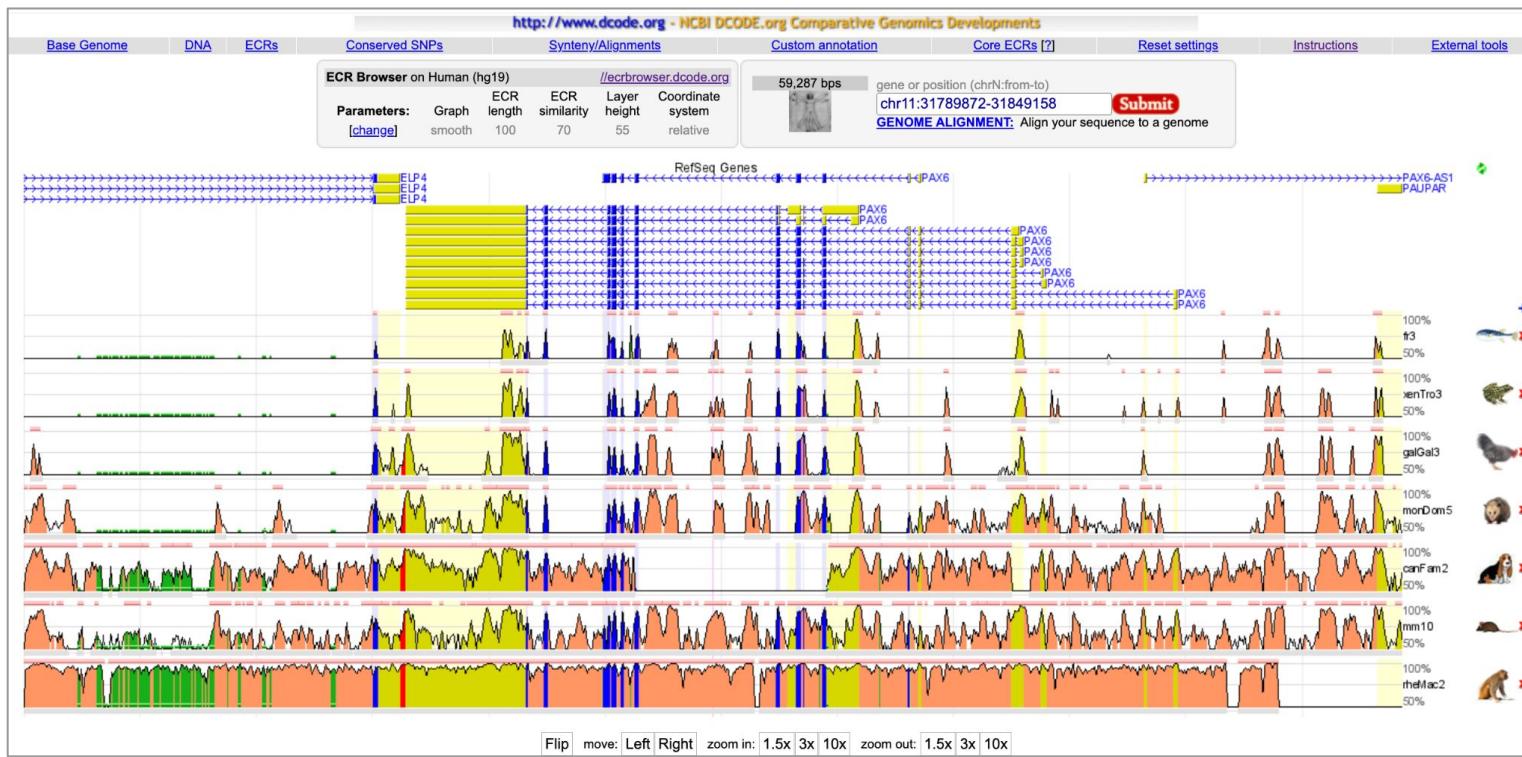
Phénotype de perte de fonction



Phénotype de gain de fonction



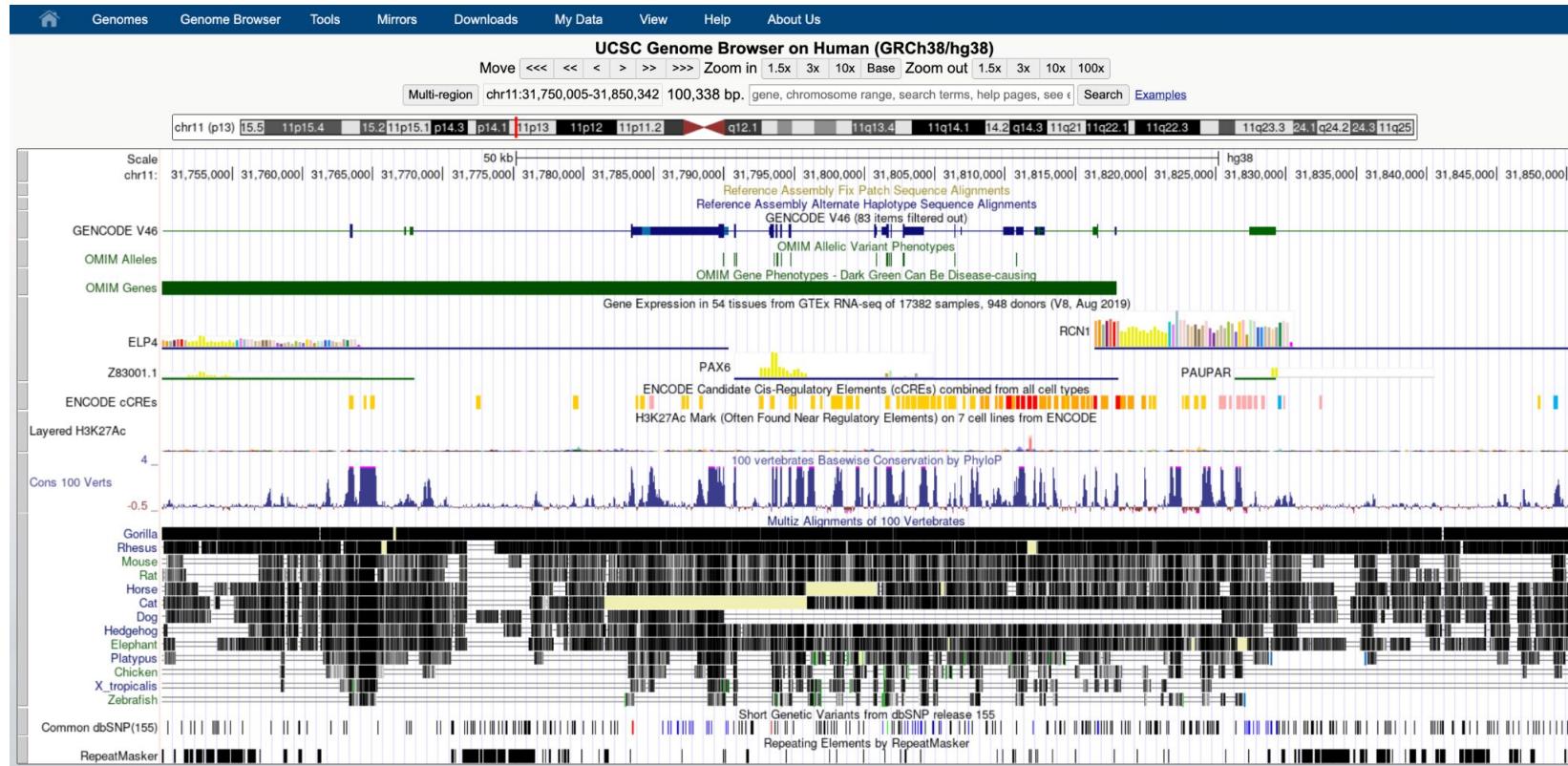
Génomique comparative : ECR genome browser (eukaryotic conserved regions)



<https://ecrbrowser.dcode.org/xB.php?db=hg19&location=chr11:31806340-31832690>

Une vue sur le gène PAX6 – UCSC Genome Browser

Exemple “brut”: vue sur le gène PAX6 au UCSC Genome Browser. Un peu indigeste, nous le présenterons progressivement



<https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A31750005%2D31850342&hgsid=2346774682> HANILICODtNRhZ5qhQaHwlcu7OR

Coupable par association

Des protéomes aux interactomes

Une protéine n'agit généralement pas seule: les protéines interagissent

- De façon stable, en formant des complexes multimériques (plusieurs polypeptides)
- De façon transitoire, en établissant des liaisons temporaires qui modifient leur niveau d'activité

Au début des années 2000, plusieurs méthodes sont mises au point pour déterminer l'**interactome**, c'est-à-dire l'ensemble des interactions entre protéines d'un système biologique (organisme, tissu, échantillon).

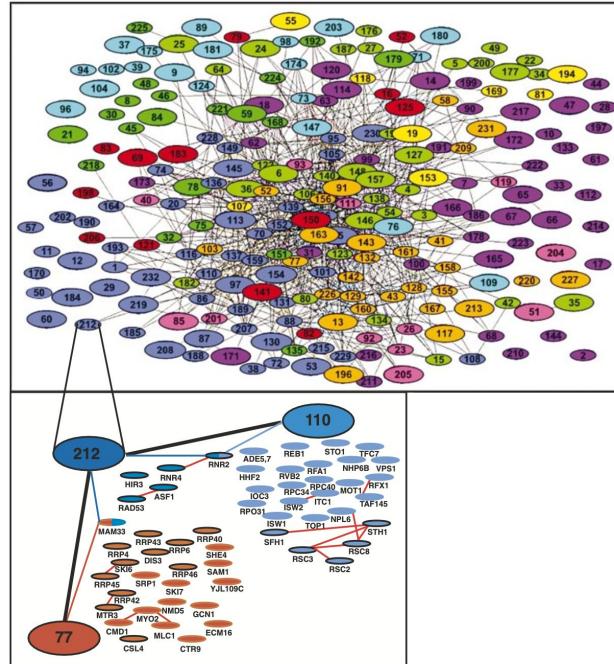


Figure 4 The protein complex network, and grouping of connected complexes. Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local minimum in the distribution of nodes by minimizing the distance of connected nodes and maximizing distance of unconnected nodes. In the upper panel, cellular roles of the individual complexes (ascribed in Supplementary Information Table S3) are colour coded: red, cell cycle; dark green, signalling; dark blue,

transcription, DNA maintenance, chromatin structure; pink, protein and RNA transport; orange, RNA metabolism; light green, protein synthesis and turnover; brown, cell polarity and structure; violet, intermediate and energy metabolism; light blue, membrane biogenesis and traffic. The lower panel is an example of a complex (yeast TAP-C212) linked to two other complexes (yeast TAP-C77 and TAP-C110) by shared components. It illustrates the connection between the protein and complex levels of organization. Red lines indicate physical interactions as listed in YPD²².

Le principe de culpabilité par association

- Le principe de culpabilité par association (*guilt by association*) en annotation fonctionnelle : si l'on ignore la fonction d'un gène ou d'une protéine, mais qu'on constate qu'elle est fréquemment associée à des gènes ou protéines de fonction connue, on suppose qu'ils peuvent participer à une même fonction.
- Les critères d'association peuvent être multiples
 - Interactions physiques entre protéines détectées dans les interactomes
 - Corrélation de présence / absence d'homologues dans les génomes / protéomes de différents organismes (profils phylogénétiques)
 - Corrélation entre profils transcriptomiques
 - Prokaryotes: inclusion dans le même opéron
 - ...
- La dénomination est ironique, car ce principe est bien entendu invalide en matière juridique : on ne peut pas condamner quelqu'un pour la seule raison qu'il a fréquenté des personnes qui ont commis un délit.

La Gene Ontology – Définir et structurer les termes d'annotation des gènes et de leurs produits

Gene Ontology (GO)

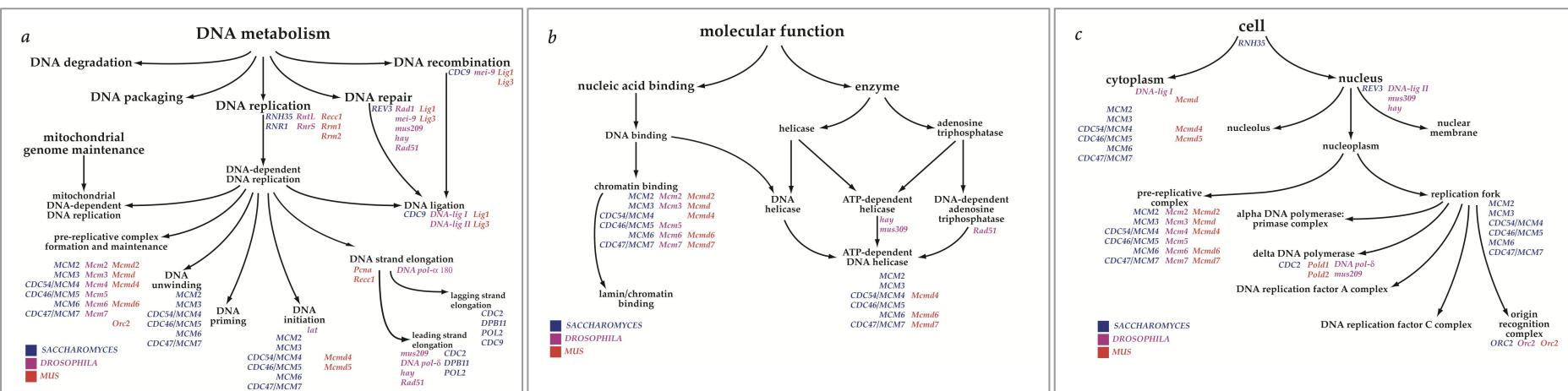
En 2000, Ashburner et collègues proposent à tous les projets de génomique d'adopter une "ontologie" pour annoter les fonctions des gènes (et des protéines qu'elles produisent).

Ils illustrent le concept avec trois organismes modèles.

- *Saccharomyces cerevisiae* (levure du boulanger)
- *Drosophila melanogaster* (mouche à vinaigre)
- *Mus musculus* (souris)

La Gene Ontology initiale définit 3 niveaux d'annotation

- a. Processus biologique (figure de gauche)
- b. Fonction moléculaire (milieu)
- c. Composante cellulaire (droite)



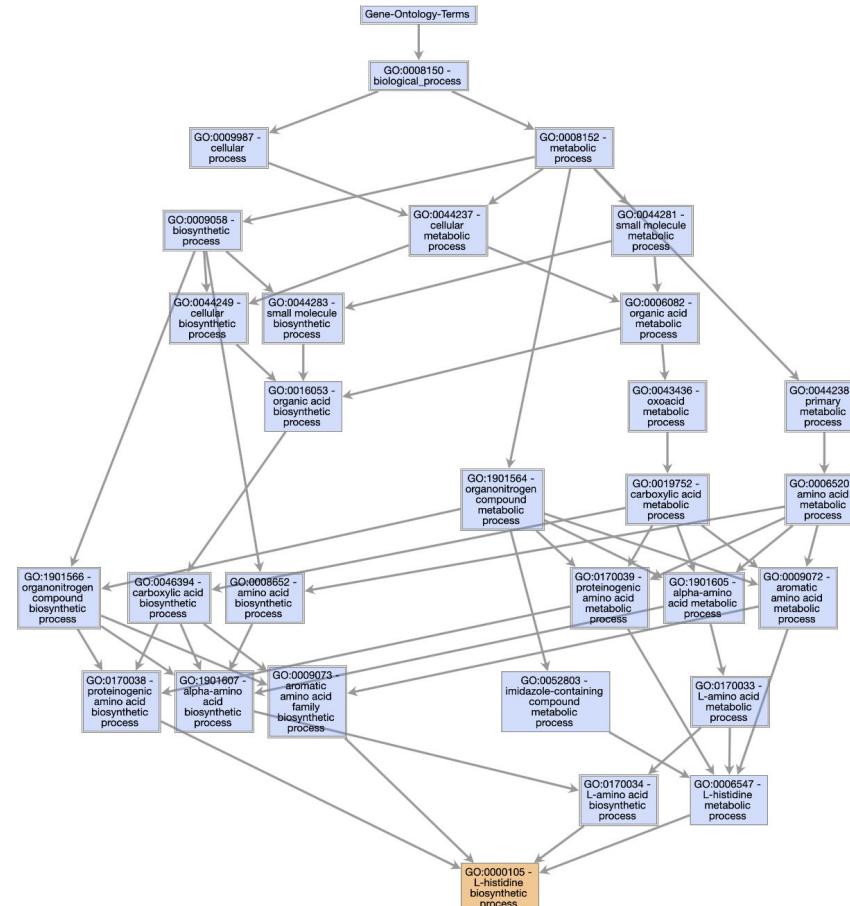
Exemple: diagramme GO du processus “biosynthèse de la L-histidine”

La voie métabolique de biosynthèse de l'histidine est rattachée à plusieurs processus parents :

- Métabolisme de la L-histidine
- Biosynthèse des acides aminés lévogyres
- Biosynthèse des acides aminés aromatiques
- Biosynthèse des acides aminés protogéniques (impliqués dans la composition des protéines)

Ces classes ontologiques ont à leur tour des classes parentes, avec certains entrecroisements.

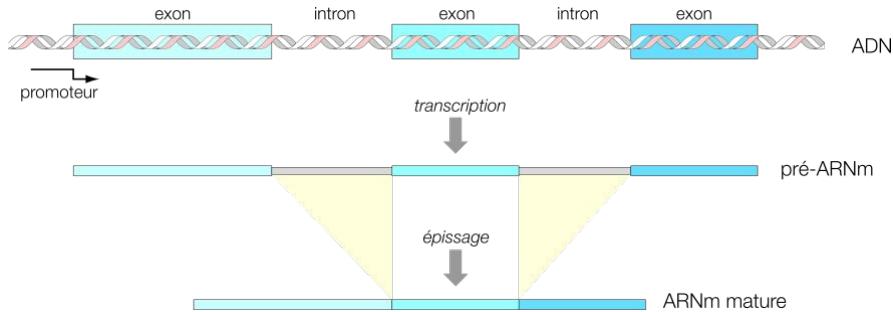
Cette structuration paraît complexe au premier abord, mais permet d'annoter chaque gène / protéine à un niveau plus ou moins détaillé de l'arborescence des termes de l'ontologie.



Information complémentaire

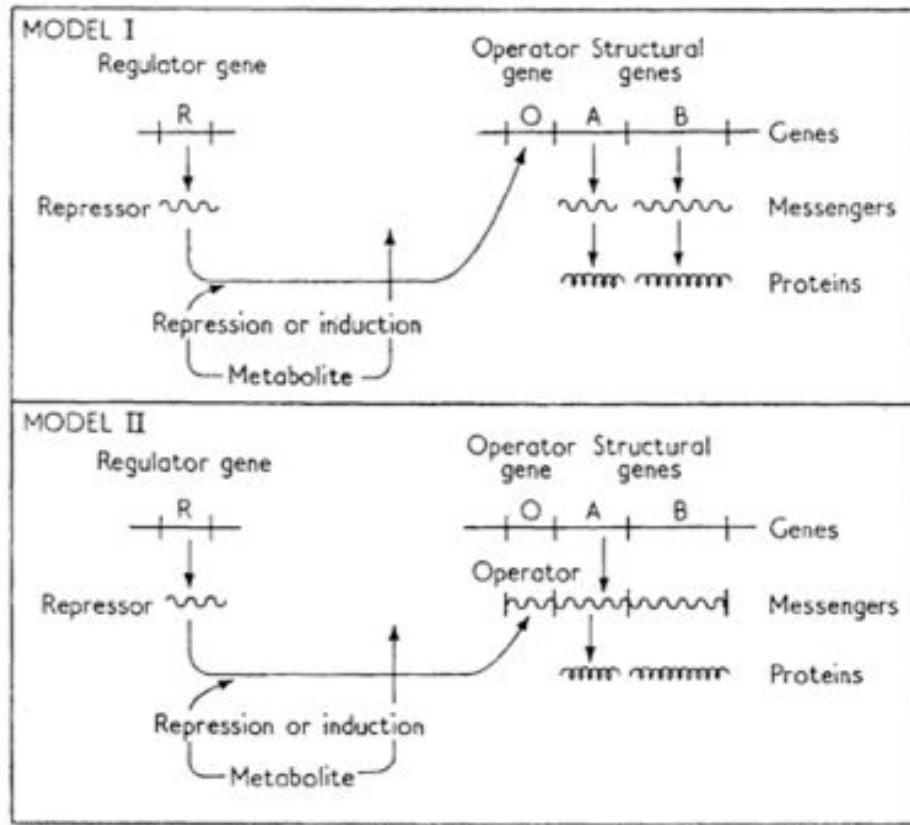
L'épissage

- Haut: ADN
- Milieu: pré-ARN = transcript primaire
 - Principale composante de la fraction nucléaire de l'ARN (extraite du noyau cellulaire)
- Bas: résultat de l'épissage: les exons sont
 - Principale composante de la fraction cytoplasmique de l'ARN (extraite du cytoplasme)
- Exons: parties de l'ADN qui se retrouvent dans l'ARN mature
- Introns: parties de l'ADN qui sont excisées entre ARN primaire et ARN mature
- **Attention:** les exons **ne correspondent pas** aux parties codantes des gènes
 - Il existe des ARN non-codants (ex: ARN de transfert, ribosomiques, ...)
 - Le concept d'**ARN messenger** ne concerne donc que les gènes codant pour des protéines
 - Même pour les gènes codants, l'ARN messenger inclut des parties non traduites à ses extrémités 3' et 5' (*UTR: untranslated regions*)



Structuration des gènes bactériens - La découverte de l'opéron

- Depuis les années 40, Jacques Monod entreprend de comprendre les mécanismes de régulation métabolique chez la bactérie *Escherichia coli*
- 1960: François Jacob and Jacques Monod proposent deux modèles alternatifs pour la régulation de l'opéron Lac
 - au niveau de la transcription
 - au niveau de l'ARN
- Le modèle de base sous-jacent à ces deux modèles est le contrôle négatif (répression) de l'expression des gènes.
- Dans les deux cas, ils soulignent l'importance des boucles de rétroaction



- Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960). [Operon: a group of genes with the expression coordinated by an operator.]. C R Hebd Seances Acad Sci 250, 1727-9.
- Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3, 318-56.
- Jacob, F. (1997). L'opéron, 25 ans après. C. R. Acad. Sci. PAris 320, 199-206.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3, 318-56.