

Introduction à la bioinformatique (UE SSV3U15)

TP3. Du gène au génome

Diaporama d'accompagnement du TP

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

Objectifs

- Utiliser des ressources bioinformatiques pour explorer les génomes d'organismes modèles, afin de comprendre la structuration et la composition de ces génomes.

Notions mises en pratique

- Structuration des gènes : transcrits, introns, exons, régions codantes, régions non traduites
- Organisation des génomes : régions géniques et intergéniques, régions répétitives, opérons bactériens
- Génomique comparative : conservation / divergence des séquences entre espèces
- Transcriptomique : expression différentielle des gènes dans différents tissus
- Annotation des séquences génomiques
- Détections des cadres ouverts de lecture
- Assignation de fonction par similarité de séquences

N'oubliez pas que vous pouvez à tout moment consulter le [glossaire du cours](#) pour obtenir une définition sommaire des principaux termes utilisés.

Etapes

- Cas d'étude 1. Contexte génomique et transcriptomique du gène humain PAX6
 - Annotations génomiques
 - Génomique comparative : régions conservées et divergentes chez les vertébrés
 - Profil tissulaire d'expression
- Cas d'étude 2. Annotation d'un fragment de séquence bactérienne
 - Recherche de cadres ouverts de lecture
 - Assignation de fonction par similarité
 - Consultation de la base de données RegulonDB

Complétion

- Tous les exercices doivent être réalisés par chaque étudiant.
- Les QCM de TP ne sont pas notés.

Éléments de contexte

Exemples traités

Analyse génomique 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

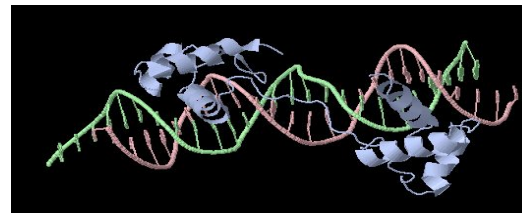
- **Perte de fonction** : l'inactivation de *eyeless* (= PAX6) provoque une malformation ou une absence d'oeil.
- **Gain de fonction** : si on force le gène *eyeless* à s'exprimer dans les tissus larvaires précurseurs des antennes ou dans les ailes, les mouches développent à ces endroits des yeux à facettes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

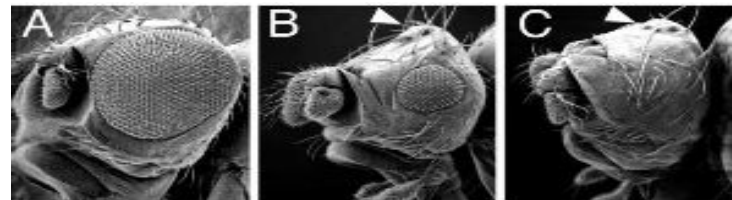
- Le gène PAX6 est fortement conservé chez tous les animaux, des invertébrés aux vertébrés.
- Ceci est compréhensible étant donné son rôle crucial pour le développement de l'organe de la vision, qui est essentiel à la survie des métazoaires.

Liaison PAX6 - ADN

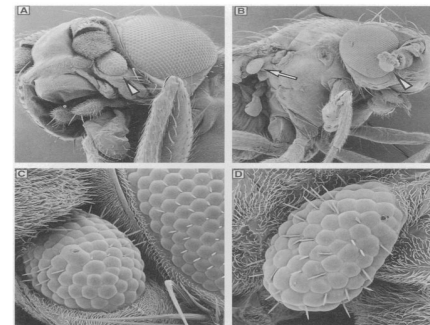


<http://www.rcsb.org/pdb/explore.do?structureId=6PAX>

Phénotype de perte de fonction



Phénotype de gain de fonction



Annotations du génome humain

Le tableau indique le nombre d'annotations pour différents types d'éléments du génome humain (gènes codants, non-codants de différents types, transcrits) dans différentes bases de données de référence.

Les nombres précis d'annotations varient d'une base de données à une autre, mais les ordres de grandeur sont indicatifs.

Constats :

- Au début du projet de séquençage du génome humain, on s'attendait à trouver ~100 000 gènes codants. Une vingtaine d'années plus tard, on en répertorie ~20 000
- Le séquençage de l'ARN révèle un nombre à peu près équivalent de "gènes" non-codants (plus précisément, régions transcrites dont on ignore généralement la fonction).
- Au total, on dénombre ~200 000 transcrits, qui incluent les transcrits alternatifs (variants d'épissage) pour ces gènes codants et non-codants.

Feature type	Gencode	Ensembl	RefSeq	CHES
Protein-coding genes	19 901	20 376	20 345	21 306
lncRNA genes	15 779	14 720	17 712	18 484
Antisense RNA	5 501		28	2 694
Miscellaneous RNA	2 213	2 222	13 899	4 347
Pseudogenes	14 723	1 740	15 952	
Total transcripts	203 835	203 903	154 484	323 827

Annotation d'un fragment de séquence génomique bactérienne

- L'annotation des génomes bactériens est relativement plus simple que celle des génomes d'eucaryotes (et en particulier des multicellulaires), car chez les procaryotes il n'y a pas d'épissage. On peut donc détecter une bonne partie des gènes codants par **recherche de cadres ouverts de lecture** (Open Reading Frames, **ORF**).
- On peut ensuite déduire la séquence des protéines potentiellement produites par ces ORF, en effectuant une **traduction informatique** (au sens moléculaire) sur base du **code génétique** (correspondance codons - acides aminés).
- Après avoir localisé les gènes, on tente de leur assigner une fonction. La première approche utilisée est l'**assignation de fonction par similarité de séquence** : on compare les séquences protéiques d'intérêt (celles potentiellement produites par les ORFs) avec une base de données de séquences protéiques de fonction connue. On retient ensuite les alignements significatifs (en veillant à appliquer un seuil de significativité assez sévère) et on considère en première approximation que la séquence protéique d'intérêt assure vraisemblablement la même fonction que l'homologue trouvé dans la base de donnée.
- Un autre élément qui intervient dans l'annotation des génomes bactérien est l'organisation des gènes : les gènes bactériens sont regroupés en **opérons**, qui consistent en une succession de gènes localisés sur le même brin, généralement peu espacés, et qui sont transcrits sur une molécule d'ARN unique (unité transcriptionnelle poly-cisdronique).

Codons start pour la détection d'ORF

Chez les procaryotes comme chez les eucaryotes, la traduction peut occasionnellement démarrer sur un autre codon que le "canonique" ATG.

A titre d'exemple, la table ci-contre indique le nombre et le pourcentage de gènes d'*Escherichia coli* qui commencent par différents codons start (souche de référence *Escherichia coli* str. K-12 substr. MG1655 GCF_000005845.2_A SM584v2).

ATG ne représente que 85% des codons. Les deux codons alternatifs les plus fréquents sont **GTG** (7.6%) et **TTG** (1.8%). Les autres codons sont marginaux (<1%).

L'existence de ces codons alternatifs peut être prise en compte lors de la recherche de cadres ouverts de lecture.

Codon	Occurrences	Frequency
ATG	3955	85.26%
GTG	354	7.63%
TTG	85	1.83%
GGG	19	0.41%
GCG	17	0.37%
GGT	15	0.32%
ATT	12	0.26%
GTC	10	0.22%
GCC	9	0.19%
GAT	8	0.17%
TGC	8	0.17%
AAA	8	0.17%
GCT	7	0.15%
GGC	7	0.15%
GCA	7	0.15%
GTT	7	0.15%
TCC	6	0.13%
TTT	6	0.13%
GGA	6	0.13%
AGG	5	0.11%
ACA	5	0.11%

Scores d'alignement

Un algorithme d'alignement de séquences arrivera toujours à aligner un certain nombre de résidus identiques en insérant les gaps aux endroits propices. Ceci est particulièrement vrai pour les alignements de séquences nucléiques, qui comportent seulement 4 résidus. A priori, même sans insérer le moindre gap, on s'attend déjà à observer 25% d'identité au hasard. Pour éviter cet écueil, les logiciels d'alignement de séquence calculent plusieurs scores informatifs qu'il est important de savoir interpréter.

- **Longueur de l'alignement** (qui dépend notamment des gaps insérés dans les deux séquences)
- **Nombre et pourcentage d'identité**
- **Nombre et pourcentage de positifs** (positifs = identités + similarités)
- **Nombre et pourcentage de gaps**
- **Score brut** (exprimé en bits d'information, que vous pouvez considérer comme une boîte noire à ce stade)
- **L'espérance aléatoire**, notée "**E-value**", "**expect**", est la clé principale d'évaluation de la significativité statistique de l'alignement (et indirectement, de sa pertinence biologique)

Comment interpréter une e-valeur ?

Dans un résultat de BLAST, l'espérance aléatoire (score **expect**) indique, le nombre d'alignements au moins aussi bons (avec un score brut au moins égal à celui obtenu) si on avait effectué la même recherche avec une requête et/ou une base de données de séquences aléatoires.

Ce nombre est par définition

- Toujours positif
- Éventuellement supérieur à 1

Par conséquent, une valeur **expect**

- ≥ 1 : résultat pas du tout significatif
- $\leq \leq 1$ (par exemple $1.4e-95$): résultat très significatif
- Faiblement inférieur à 1 : faiblement significatif, risques de faux-positifs (considérer comme significatif un résultat qui résulte du hasard)
- 0: littéralement, signifie qu'un tel serait impossible avec des séquences aléatoires. En pratique, signifie que la e-value est inférieure à la limite de précision du calcul informatique (ce qui correspond à peu près à $< 1e-320$)

Cas d'étude 1.

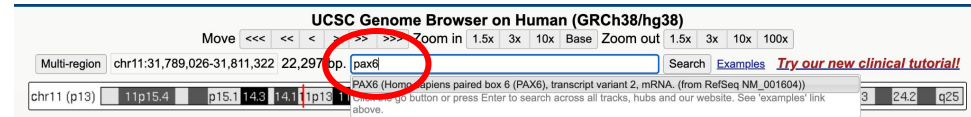
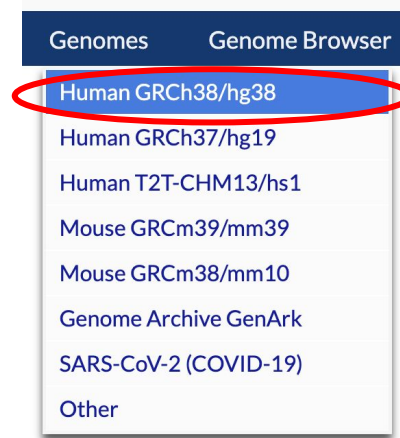
Contexte génomique et transcriptomique du gène humain PAX6

Tutoriel illustré

Annotations génomiques dans la région du gène humain PAX6

PAX6 sur UCSC genome browser

- Connectez-vous au [UCSC Genome Browser](#)
- Dans le menu **Genomes**, sélectionnez la version **hg38** du **génomme humain**.
- Entrez le nom du gène d'intérêt (**PAX6**) dans la boîte de recherche et cliquez sur **Search**.



- Connectez-vous au [UCSC Genome Browser](#)
- Sélectionnez la version **hg38** du génome humain
- Entrez le nom du gène d'intérêt (**PAX6**) dans la boîte de recherche et cliquez sur **Search**.

La page de résultat affiche une série d'annotations de PAX6 dans différentes bases de données de référence pour le génome humain. Comment choisir ? En première instance, le mieux est de se fier aux annotations du consortium international HUGO, responsable de la nomenclature des gènes humains.

- Sous le titre "HUGO Gene Nomenclature", cliquez sur le lien **PAX6 - chr11:31789026-31817960**

Search Results on hg38 (Human Dec. 2013 (GRCh38/hg38))

MANE Select Plus Clinical: Representative transcript from RefSeq & GENCODE:

HUGO Gene Nomenclature:

- **PAX6** - chr11:31789026-31817960
- PAX6-AS1 - chr11:31789026-31887040

Gencode Genes:

- **PAX6** (ENST00000640368.2) - chr11:31789026-31811322 - **PAX6** ENST00000640368.2 Homo sapiens paired box 6 PAX6 transcript variant
- **PAUPAR** (ENST00000644607.1) - chr11:31816266-32002405 - PAUPAR ENST00000644607.1 **PAX6** upstream antisense RNA from HGNC PAUPAR BX64896 uc285izg.1 uc285izg.1
- **BCL2L15** (ENST00000393316.8) - chr11:113876816-113887581 - ... Q5TBC7 P50222 MEOX2 NbExp 3 IntAct EBI-10247136 EBI-748397 Q5TBC7 P26367 **PAX6** NbExp 3 IntAct EBI-10247136 EBI-747278 Q5TBC7 P62487 POLR2G NbExp
- **LYSMD1** (ENST00000368908.10) - chr11:151159748-151165902 - ... Q96S90 Q5JR59 MTUS2 NbExp 3 IntAct EBI-10293291 EBI-742948 Q96S90 P26367 **PAX6** NbExp 3 IntAct EBI-10293291 EBI-747278 Q96S90 Q6NUQ1 RINT1 NbExp
- **CCDC103** (ENST00000417826.3) - chr17:44899729-44905390 - ... Q8IW40 Q6FHY5 MEOX2 NbExp 3 IntAct EBI-10261970 EBI-16439278 Q8IW40 P26367 **PAX6** NbExp 3 IntAct EBI-10261970 EBI-747278 Q8IW40 Q9NRD5 PICK1 NbExp
- **SLC12A8** (ENST00000469902.6) - chr3:125082644-125212748 - ... a role in the control of keratinocyte proliferation A0A0AV2 P26367 **PAX6** NbExp 3 IntAct EBI-11737524 EBI-747278 Membrane Multi-pass membrane protein
- **TCP11L1** (ENST00000334274.9) - chr11:33039572-33073550 - ... Q9NUJ3 P50221 MEOX1 NbExp 3 IntAct EBI-2555179 EBI-2864512 Q9NUJ3 P26367 **PAX6** NbExp 3 IntAct EBI-2555179 EBI-747278 Q9NUJ3 Q5SXH7-1 PLEKHS1 NbExp
- **ZNF513** (ENST00000327033.11) - chr2:27377235-27380734 - ... Binds DNA Can associate with the proximal promoter regions of **PAX6** and SP4 and their known targets including ARR3 RHO
- **SPDYC** (ENST00000377185.3) - chr11:65170233-65173374 - ... Q5MJ68 Q9Y250 LZTS1 NbExp 3 IntAct EBI-12162209 EBI-1216080 Q5MJ68 P26367 **PAX6** NbExp 3 IntAct EBI-12162209 EBI-747278 Cytoplasm Note Colocalizes with
- **CXorf38** (ENST00000327877.10) - chrX:40626921-40647561 - ... Q8TB03 Q02548 PAX5 NbExp 3 IntAct EBI-12024320 EBI-296331 Q8TB03 P26367 **PAX6** NbExp 3 IntAct EBI-12024320 EBI-747278 Q8TB03 Q9Y3C5 RNF11 NbExp

Show 98 more matches for Gencode Genes

NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*):

- NR_033971.1 - chr11:31816566-31887041
- **NM_001258463.2** - chr11:31789026-31812203
- **NM_001258464.2** - chr11:31789026-31811322
- **NM_001368926.2** - chr11:31789026-31811322
- **NM_001368914.2** - chr11:31789026-31811322
- **NM_001604.6** - chr11:31789026-31811322
- **NM_001368893.2** - chr11:31789026-31811322
- **NM_001368917.2** - chr11:31789026-31811322
- **NM_000280.6** - chr11:31789026-31811322
- **NM_001368927.2** - chr11:31789026-31811322

Show 44 more matches for NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*)

RefSeq Genes:

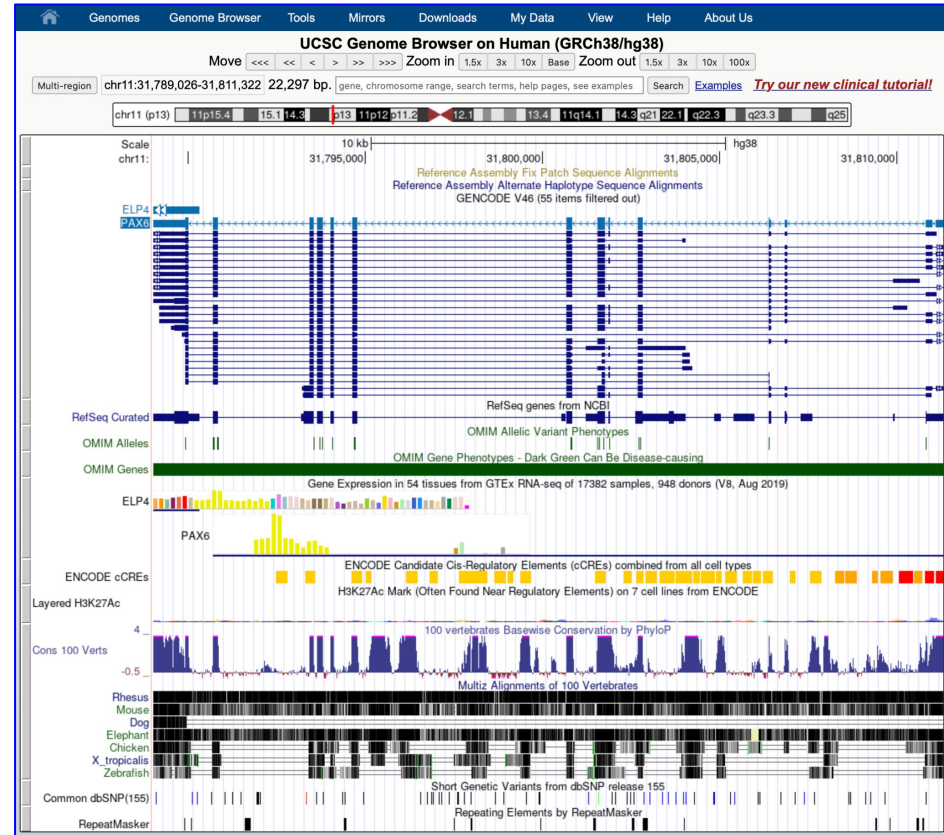
- **PAX6** - chr11:31789026-31811121 - (NM_000280) paired box protein Pax-6 isoform a
- **PAX6** - chr11:31789026-31811322 - (NM_001258464) paired box protein Pax-6 isoform a
- **PAX6** - chr11:31789026-31812203 - (NM_001310158) paired box protein Pax-6 isoform b
- **PAX6** - chr11:31789026-31817961 - (NM_001127612) paired box protein Pax-6 isoform a
- **PAX6** - chr11:31789026-31811322 - (NM_001604) paired box protein Pax-6 isoform b
- **PAX6** - chr11:31789026-31817961 - (NM_001258462) paired box protein Pax-6 isoform b
- **PAX6** - chr11:31789026-31804059 - (NM_001310161) paired box protein Pax-6 isoform d
- **PAX6** - chr11:31789026-31811121 - (NM_001258465) paired box protein Pax-6 isoform a
- **PAX6** - chr11:31789026-31812203 - (NM_001258463) paired box protein Pax-6 isoform b
- **PAX6** - chr11:31793205-31806925 - (NM_001310159) paired box protein Pax-6 isoform c

Show 44 more matches for RefSeq Genes

Choix de pistes d'annotations du UCSC Genome Browser

Le navigateur de génomes [UCSC Genome Browser](#) affiche un vaste choix de pistes d'annotation. La carte génomique en affiche un sous-ensemble, qui s'adaptent en fonction de vos consultations précédentes.

Nous allons restreindre la visualisation aux pistes d'annotations utilisées pour ce TP.



PAX6 sur UCSC genome browser

- Descendez sous la carte génomique pour afficher les choix de pistes d'annotations.
- Entre la carte et les options, cliquez **Hide all** pour masquer les pistes génomiques par défaut.
- Dans la catégorie "**Mapping and Sequencing**", sélectionnez le mode d'affichage "**dense**" pour la piste d'annotation "**Base position**".
- Dans la catégorie "**Genes and Gene Prediction**", sélectionnez le mode "**pack**" pour les pistes "**Gencode_V46**" et **HGNC**.
 - HGNC indique les limites des gènes, tandis que Gencode_V46 fournit des informations plus détaillées sur la structure des gènes (introns, exons, transcrits alternatifs, ...).
- Cliquez "**Refresh**" à droite d'une des catégories.
- Cliquez "**Resize**" sous la carte pour ajuster la largeur à celle de votre écran.

The screenshot displays the UCSC Genome Browser interface for the PAX6 gene region on chromosome 11 (GRCh38/hg38). The main view shows a genomic map with various tracks including Gencode V46, HGNC, and PAX6. The track configuration panel at the bottom is visible, showing the following settings:

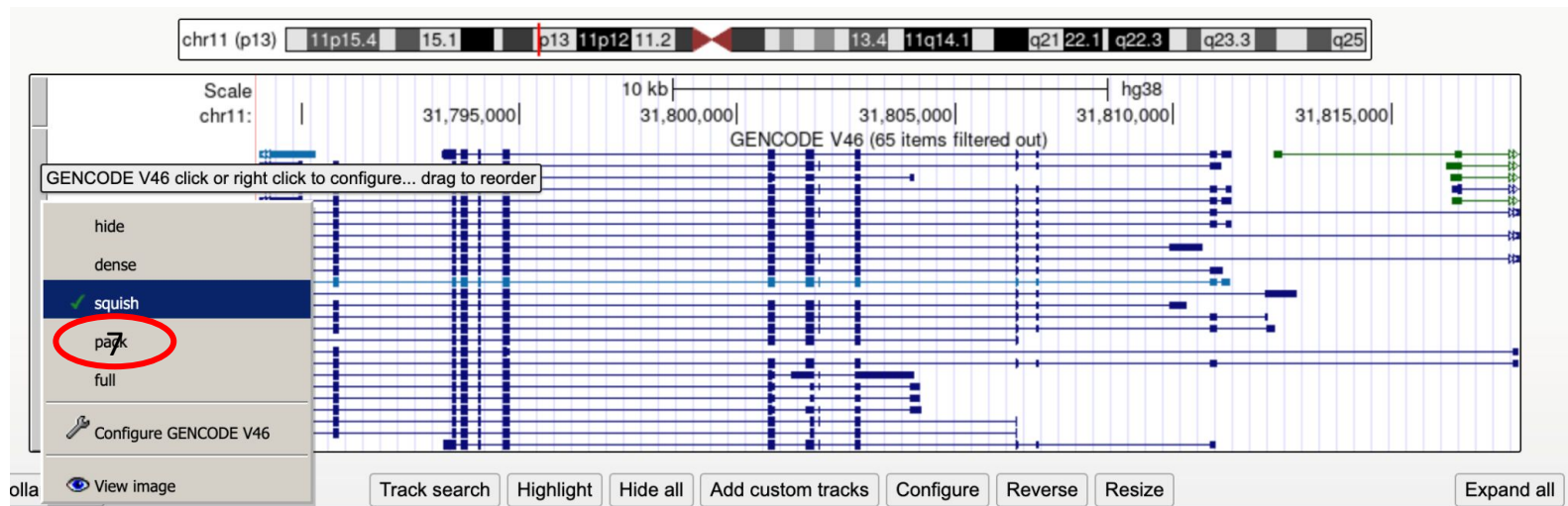
- Mapping and Sequencing:**
 - Base Position: dense
- Genes and Gene Predictions:**
 - Gencode_V46: pack
 - HGNC: pack

The 'Refresh' button is visible for each category.

Reconfigurer le mode d'affichage

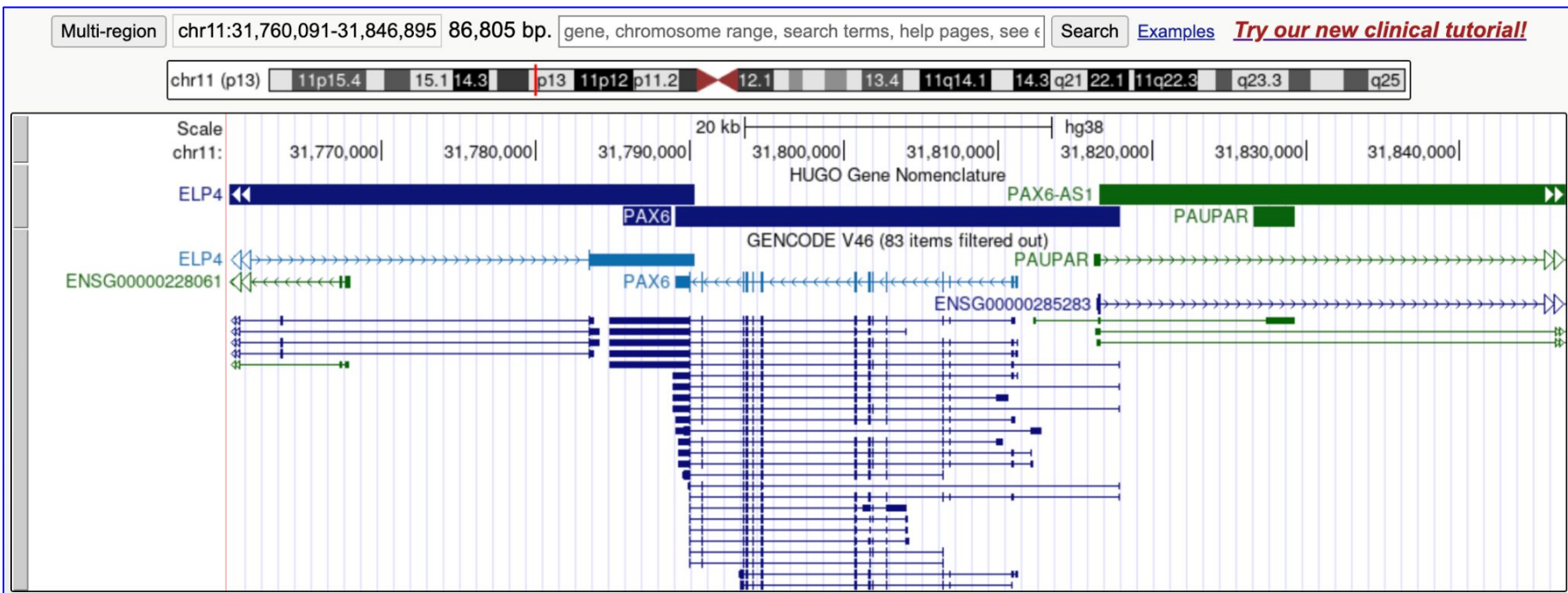
Vous pouvez à tout moment reconfigurer le mode d'affichage d'une piste d'annotation, en cliquant droit (**contrôle-clic**) sur la figure. Ceci vous affichera un menu avec des modes d'affichages de plus en plus détaillés : hide, dense, squish, pack, full.

- Testez les différents niveaux de détail avec la piste **GENCODE_V46**, puis sélectionnez le mode **pack**, qui vous permet généralement de visualiser les transcrits alternatifs en occupant une place raisonnable.



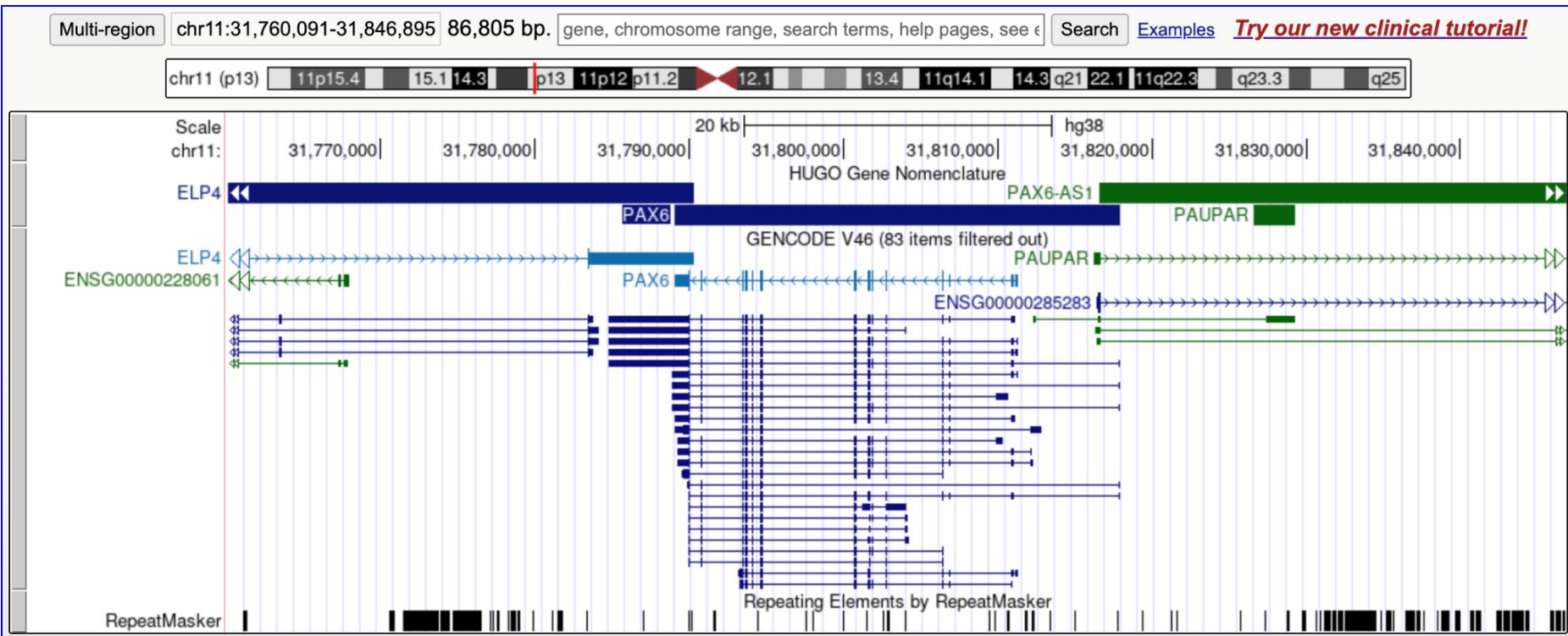
Voisinage du gène PAX6

- Dézoomez (**Zoom out**) d'un facteur **x3** pour voir les environs du gène
- **Déplacez** la piste HGNC au-dessus de la piste GENCODE (en positionnant la souris vers le coin supérieur gauche d'une piste, une flèche apparaît qui permet de la déplacer)
- Observez la disposition du gène PAX6. Notez qu'il chevauche ses voisins de gauche (ELP4) et de droite (PAX6-AS1, où AS indique qu'il s'agit d'un gène antisens).



Régions répétées

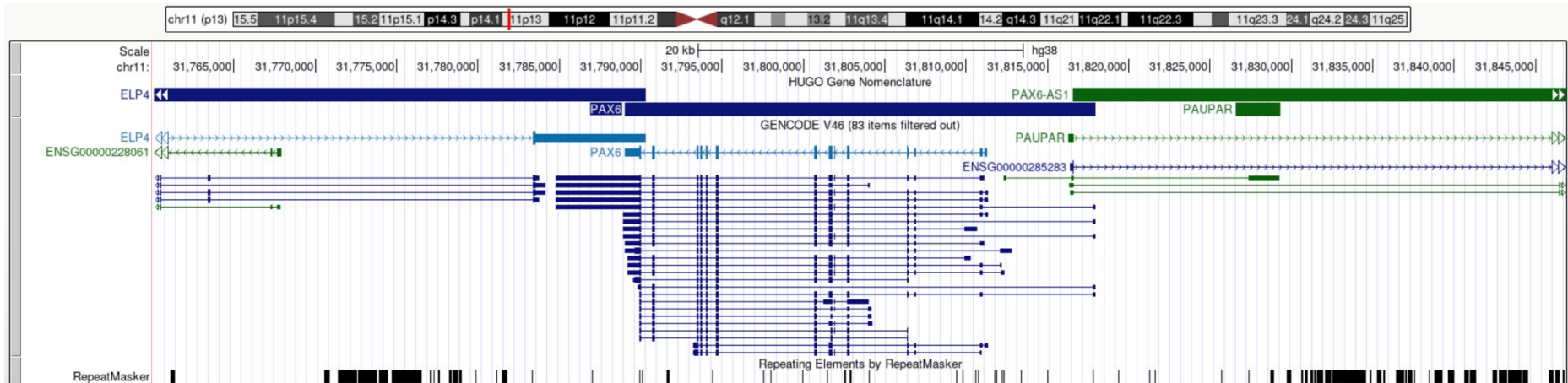
- Dans la catégorie "**Repeats**", activez l'affichage de "**Repeatmasker**" en format "**dense**" et cliquez "**Refresh**".
- Comparez les localisations des éléments répétés et la structure du gène PAX6 et des régions qui l'entourent. Pour mieux voir des détails, vous pouvez zoomer et parcourir ces régions progressivement.



Exercice 1. Annotations génomiques dans la région du gène humain PAX6

Sur Ametice, ouvrez le questionnaire du TP3 et répondez aux questions de l'**Exercice 1**.

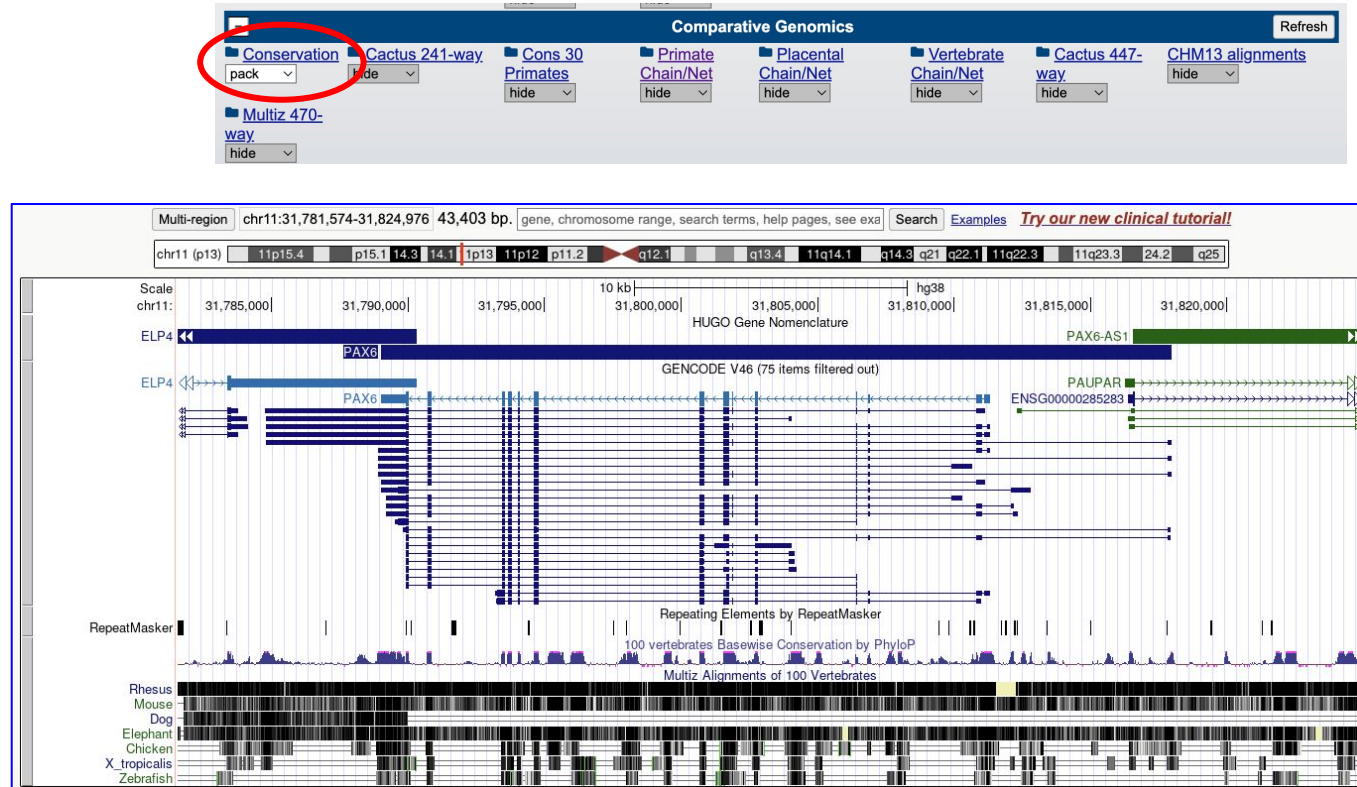
1. Sur quel chromosome est situé le gène PAX6 ? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y
2. Sur quel brin chromosomique se situe le gène PAX ? +, -
3. Sur quel bras chromosomique se trouve PAX6 ? Gauche, droite
4. Sur quelle région chromosomique se trouve le gène PAX6 ? p15.4, 15.1, 11p13, p11.2, q2.1, q23.3
5. Quelle est sa longueur en kilobases ? 12, 22, 29, 31 789, 31 817
6. Combien de régions répétitives distinguez-vous sur la région du gène PAX6 ? aucune, 15, 22, 30, 50
7. Quelle est la densité approximative (nombre de régions répétitives / kilobase) ? 0.1, 0.8, 1, 1.2, 1.3, 2, 5, 10
8. Sur la longueur du gène PAX6, les régions répétitives coïncident généralement avec les introns, régions exoniques codantes, régions exoniques non-codantes



Conservation du gène PAX6 dans les génomes de vertébrés

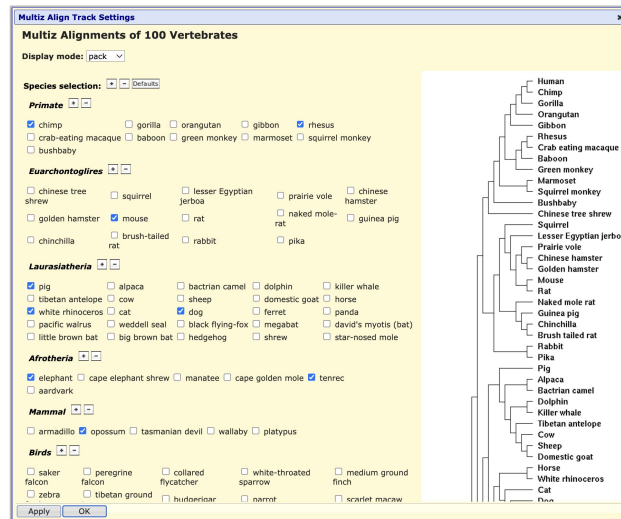
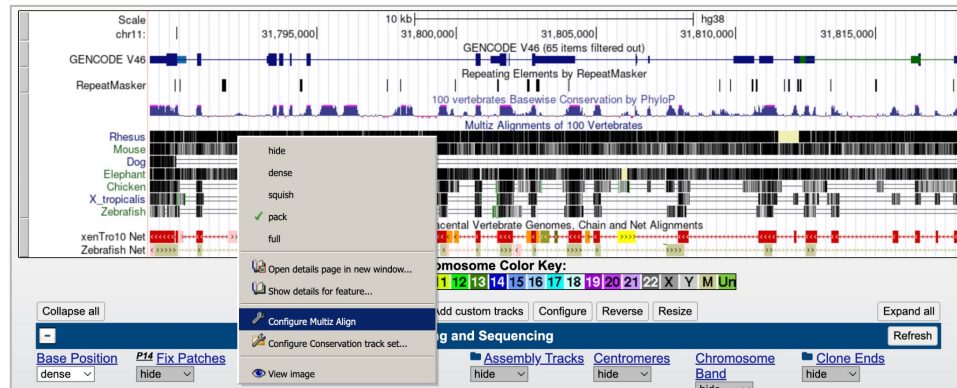
Génomique comparative : régions conservées chez les vertébrés

- Dans la catégorie **Comparative genomics**, activez l'affichage **pack** de la piste **Conservation**. Cette piste s'affiche entre les annotations GENCODE_V46 et les régions répétitives
- Faites remonter la piste **RepeatMasker** pour la placer entre les pistes GENCODE_V46 et Conservation



Configuration de la piste d'annotation Conservation

- Cliquez droit (contrôle-clic) sur l'image de conservation à la hauteur où s'affichent les espèces et sélectionnez **Configure MultiZ Align**.
- Dans la fenêtre d'options qui apparaît, **cochez quelques espèces de votre choix**
 - **Veillez à panacher** (essayez d'avoir une ou deux espèces de chaque groupe plutôt qu'un tas d'espèces du même groupe).
 - Pour une raison technique, le génome du chien présente des lacunes à cet endroit du génome. **Désactivez l'affichage du chien (dog)** et activez celui d'un ou deux autres mammifères du même groupe.
 - Dans la catégorie **Mammal**, **cochez toutes les espèces**. Notez que les catégories précédentes contiennent également des mammifères (Primates, Euarchontoglires, Laurasiatheria). La catégorie Mammal présente des espèces plus éloignées (marsupiaux, monotrèmes), qui sont utiles pour visualiser les régions les plus conservées entre mammifères.
- Cliquez **Apply**.
- **Dézoomez d'un facteur 1.5** pour observer le contexte aux alentours du gène.



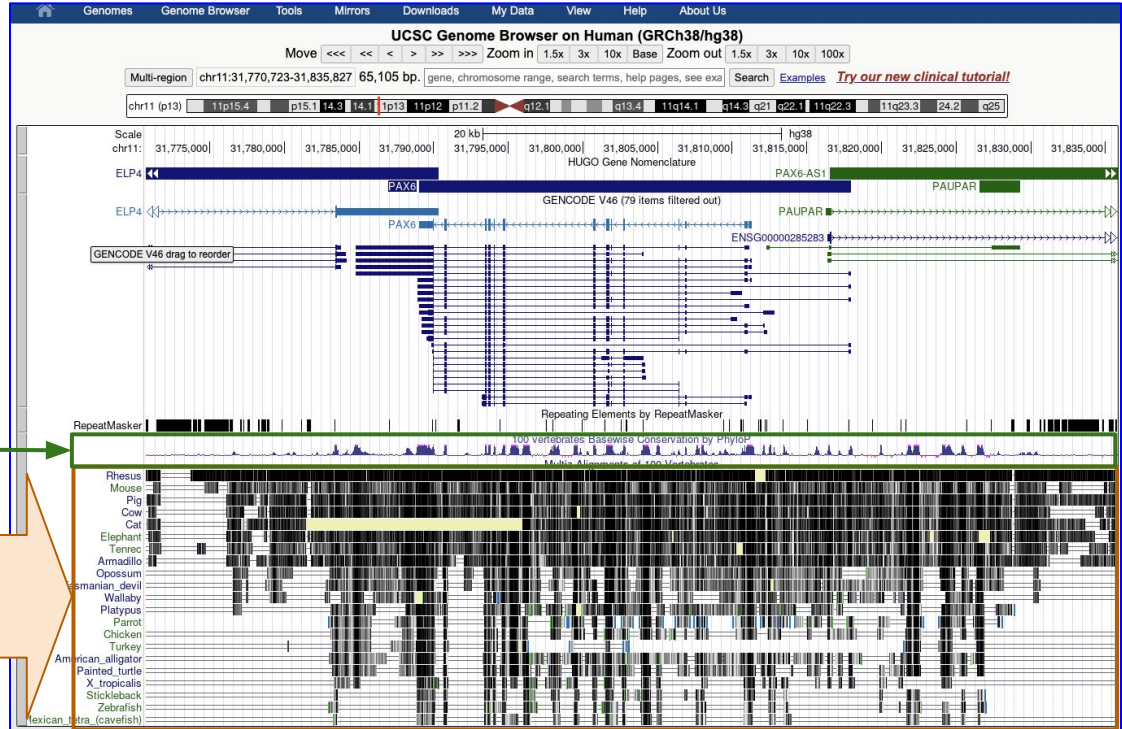
Conservation de la région génomique PAX6 chez les vertébrés

La carte de conservation génomique comporte deux parties.

1. La partie supérieure affiche un **profil de conservation dans 100 génomes de vertébrés**. La hauteur du profil indique le pourcentage de positions identiques (PPI) à chaque position du génome. Notez que l'échelle verticale va de 50% à 100%, pour mieux faire ressortir les régions conservées.
2. La partie inférieure indique, sous forme d'une échelle de gris, le **profil de conservation par espèce** (pour celles que vous avez sélectionnées).
3. Les zones marquées en jaune correspondent à des trous de séquençage (on ne dispose pas de la séquence).

Conservation chez
100 vertébrés

Profils par
espèce



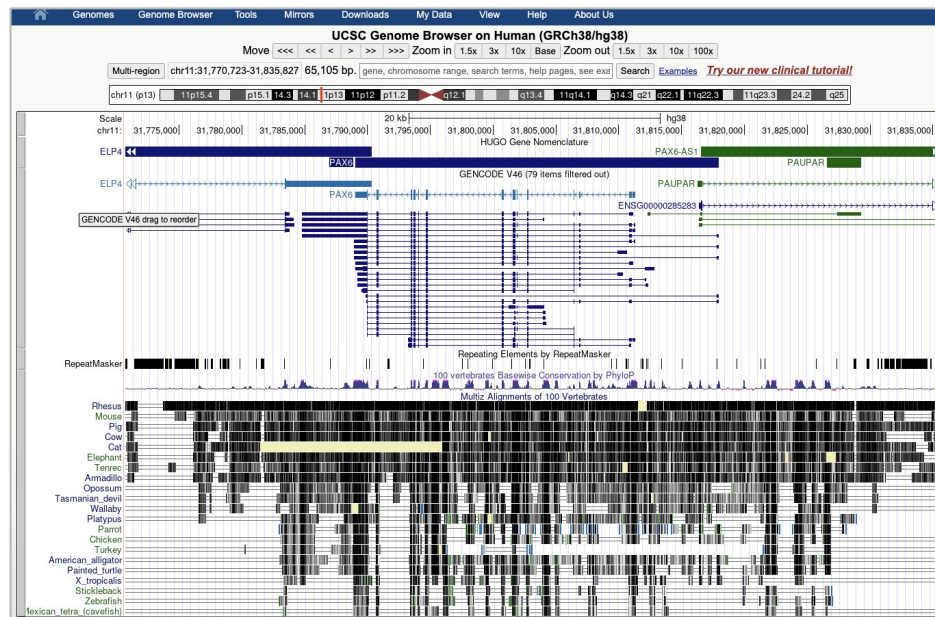
Exercice 2. Conservation de la région génomique PAX6 chez les vertébrés

Sur Ametice, ouvrez le questionnaire du TP3 et répondez aux questions de l'**Exercice 2**.

1. Zoomez sur la carte génomique que vous avez générée et parcourez le gène PAX6 sur toute sa longueur. Sur le profil de conservation génomique 100 vertébrés, dans quelles régions du gène retrouve-t-on généralement les régions conservées (plusieurs réponses possibles):

- 5'UTR
- partie codante des exons
- Introns
- 3'UTR

2. Sur la carte des profils par espèces, quel est l'ordre décroissant des degrés de conservation (une seule réponse)
 - Primates > Poissons > Mammifères non primates
 - Poissons > Mammifères non primates > Primates
 - Primates > Mammifères non primates > Poissons
 - Poissons > Primates > Mammifères non primates

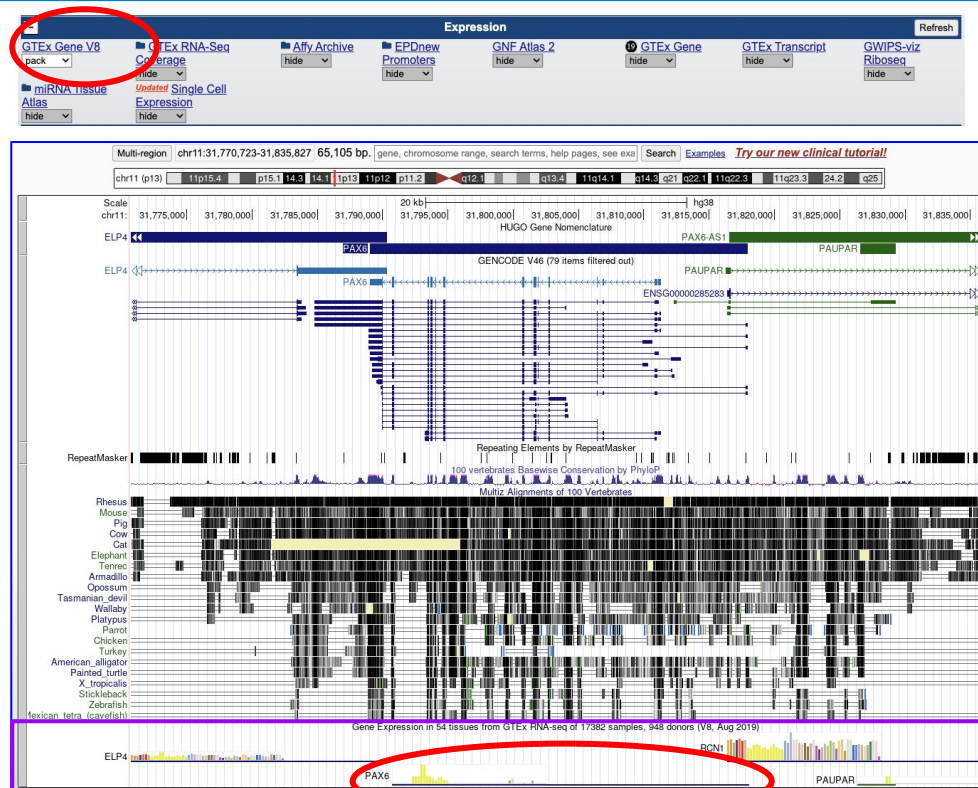


Expression tissulaire du gène PAX6

Profils tissulaires de transcription dans GTEx (Genotype-Tissue Expression)

Nous allons maintenant ajouter à notre carte génomique une piste d'annotation de la base de données [GTEx](#) (Genotype-Tissue Expression). GTEx contient des données de transcriptome (mesure quantitative de tous les transcrits produits par un génome) dans des échantillons de 54 tissus prélevés chez 948 personnes adultes.

- Dans la catégorie **Expression**, activez l'affichage **pack** de **GTEx_Gene_V8**.
- Cliquez sur l'icône du gène **PAX6** sur la piste **GTEx_Gene_V8**, et examinez le profil d'expression tissulaire.



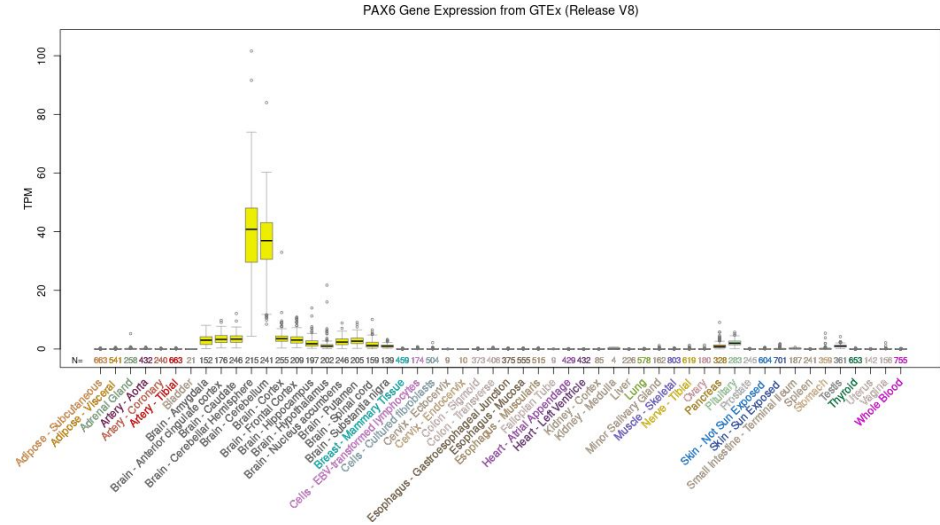
Exercice 3. Profil d'expression tissulaire de PAX6

Sur Ametice, ouvrez le questionnaire du TP3 et répondez aux questions de l'**Exercice 3**.

1. Indiquez le niveau d'expression de PAX6 dans les tissus suivants (fort/faible/indétectable)

- Cervelet
- Autres tissus du cerveau
- Muscles
- Poumons
- Sang

2. GTEx se base sur des échantillons adultes. D'après la fonction de PAX6, on s'attend à observer également un profil d'expression très spécifique durant le développement embryonnaire. Vrai / Faux



Informations complémentaires

- **Page "training" de UCSC Genome Browser (genome.ucsc.edu/training)**
 - Guides d'utilisation (en anglais)
 - Courtes vidéos pour apprendre à manipuler les nombreuses fonctionnalités du site Web

Cas d'étude 2.

Annotation d'un fragment de séquence génomique bactérienne

Une séquence bactérienne à annoter

Nous disposons d'un fragment chromosomique bactérien, qu'on peut récupérer en cliquant ici.

[seq_bact_a-annoter.fasta.txt](#)

Ouvrez ce fichier dans un onglet séparé. Pour l'étape suivante, vous pourrez soit le copier à partir de cet onglet, soit le sauvegarder sur votre ordinateur et l'ouvrir avec un éditeur de texte de votre choix.

Nous allons utiliser quelques outils bioinformatiques pour annoter ce fragment d'ADN chromosomique. La première étape consiste à localiser les gènes sur ce fragment d'ADN. Il faudra ensuite essayer de trouver la fonction assurée par ces gènes.

Recherche de cadres ouvertes de lecture avec ORFinder du NCBI

Afin de localiser les gènes sur ce fragment d'ADN chromosomique, nous allons effectuer une recherche de cadres ouverts de lecture (open reading frames, ORFs), en utilisant l'outil ORFinder du NCBI.

- Connectez-vous à l'outil [ORFinder du NCBI](#).
- Collez la séquence du fragment chromosomique bactérien dans l'encadré **Enter Query Sequence**.
- Dans la section, **Choose Search Parameters** :
 - Fixez la longueur minimale des ORFs recherchés à **300 pb**
 - Laissez le **code génétique** à sa valeur initiale (**standard** *)
 - Pour le codon start à utiliser pour la recherche, choisissez **ATG only** *
- Cliquez **Submit**.

* **Note** : le menu "Genetic code" permet de choisir des codes alternatifs spécifiques de bactéries, qui comportent des codons alternatifs. La prise en compte de ces codons est généralement pertinente quand on annote un génome entier. Cependant, ceci crée aussi des ORFS surnuméraires, qui n'apporteraient pas d'information pour ce segment-ci de séquences. Dans le cadre de ce TP, nous utilisons donc le code génétique standard, et n'acceptons comme codon start que la séquence ATG.

National Library of Medicine
National Center for Biotechnology Information

Log in

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:

```
>results for 4641652 residue sequence "NC_000913.3 Escherichia coli O157:H7 genomic DNA"
TTTCACTCAATGATGCTCTTTCCGTTCCCTTTCGCTGATTCAGGCTATCGATTGAGTCC
ATCAATCTCCGGGCGTTAGCGGGGAGCGCAGTAGATAGCCGCTCTTCCAGCGAGTTG
TATTCTTCCGATGACATCAGAACAAGCCCTCTCATTCTGACGAGTAATAAGGATCGGG
GCATGATCTTCAACGGCTTTCATCATTGTTGCCGACAAATTTCTGACGCGCTTCGCTGAG
CTAATGTTACGCATGCTCAATCTCTCTTTGTACAGTTCATTGTACAATGATGAGCGTTA
ATTAACATTTATTAATAGTTTGTAGATCAAGGTATTGTCAAGTACGAGAAAATCCAGG
```

From: To:

Choose Search Parameters

Minimal ORF length (nt):

Genetic code:

ORF start codon to use:

"ATG" only

"ATG" and alternative initiation codons

Any sense codon

Ignore nested ORFs:

Start Search / Clear

Submit Clear

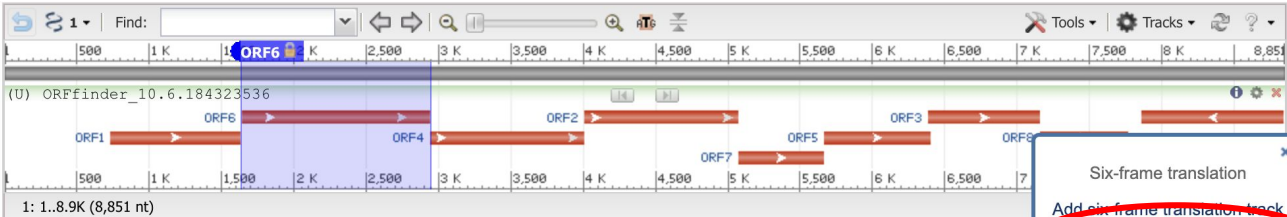
Recherche des cadres ouverts de lecture

- Sur la fenêtre de résultats de la recherche d'ORFs, Cliquez sur **Six-frame translation**, puis sur **Display six-frame translation**.

Open Reading Frame Viewer

Sequence

ORFs found: 10 Genetic code: 11 Start codon: 'ATG' only



1: 1..8.9K (8,851 nt)

ORF6 (434 aa) [Display ORF as...](#) [Mark](#)

```
>lc|ORF6
MSFNTIIDWNSCTAEQQRQLLMRPAISASESITRTVNDILDNVKARGDEA
LREYSAKFKTTVTALKVSAEEIAAASERLSDELKQAMAVAVKNIETFHT
AQLKPPVDVETQPGRVRCQVTRPVASVGLYIPGGSAPLFTVLMLATPAS
IAGCKKVVLCSPPIADEILYAAQLCGVQDVFNVGGQAIAALAFGTESV
PKVDKIFGPGNAFVTEAKRQVSQRLDGAAIDMPAGPSEVLVIADSGATPD
FVASDLLSQAEHGPDSQVILLTPAADMARRVAEAVERQLAELPRAETARQ
ALNASRLIVTKDLAQVEISNQYGEPEHLIIQTRNARELVDSITSAGSVFL
GDWSPESAGDYASGNTNHLPTYGYTATCSSLGLADFQKRMTVQELSKEGF
SALASTIETLAAAERLTAHKNAVTLRVNALKEQA
```

ORF6 [SmartBLAST](#) [BLAST](#)

Marked set (0) [SmartBLAST best hit titles...](#)

BLAST Database: [UniProtKB/Swiss-Prot \(swissprot\)](#)

Mark subset... Marked: 0 [Download marked set](#) as [Protein FASTA](#)

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	1638	2942	1305 434
ORF4	+	2	2939	4009	1071 356
ORF2	+	1	4009	5076	1068 355
ORF9	-	1	8845	7862	984 327
ORF1	+	1	733	1632	900 299
ORF3	+	1	6385	7161	777 258
ORF5	+	2	5666	6403	738 245
ORF8	+	3	7155	7766	612 203
ORF7	+	3	5076	5666	591 196
ORF10	-	2	7890	7369	522 173

Exercice 4. Traduction sur 6 phases

Questions

1. Parcourez la fenêtre de résultat de la traduction sur 6 cadres de lecture. Pourquoi y a-t-il 3 lignes de lettres décalées au dessus de la séquence d'ADN, et 3 en-dessous ? Cochez les affirmations correctes.
 - La place au-dessus ne suffit pas pour les 6 phases
 - La place au-dessous ne suffit pas pour les 6 phases
 - Les séquences protéiques au-dessus de la séquence d'ADN correspondent aux traductions sur les 3 phases directes
 - Les séquences protéiques au-dessous de la séquence d'ADN correspondent aux traductions sur les 3 phases réverse complémentaires
 - Les 3 séquences du dessus se lisent de gauche à droite
 - Les 3 séquences du dessus se lisent de droite à gauche
 - Les 3 séquences du dessous se lisent de gauche à droite
 - Les 3 séquences du dessous se lisent de droite à gauche

2. A quoi correspondent les lettres rouges?
 - UTR
 - Codons start
 - Introns
 - Exons
 - Régions codantes
 - Codons stop
 - Gaps d'alignement
3. A quoi correspondent les astérisques ? (mêmes options)
4. A quoi correspondent les lettres bleues (descendez dans la fenêtre pour les voir) (mêmes options)

Six-frame translation

Display: Six-frame translation Mark stop codon as * Download

```
F H S M M S F S V P L P D F R L S I E S I N L R A L A G E R
F T Q * C P F P F L C L I S G Y R L S P S I S G R * R G S A
S L N D V L F R S F A * F O A I D * V H Q A S P G V S G G A
1 TTTCACCTCAATGATGTCCTTTCCGTTCTTTGCTGATTTCAGGCTATCGATTGAGTCCATCAATCTCCGGGCGTTAGCGGGGAGCGC
K V * H H G K G N R Q R I E P * R N L G D I E P R * R P L A
E S L S T R K R E K A Q N * A I S Q T W * D G P T L P P A C
* E I I D K E T G K G S K L S D I S D M L R R R A N A P S R

S R * A V S S S E L Y S S S D I R T Q A S P F * R V I R I G
V D K P S L P A S C I L R V T S E H K P L H S D E * * G S G
Q * I S R L F Q R V V F F E * H Q N T S L S I L T S N K D R
91 AGTAGATAAGCCGTCTCTCCAGCGAGTTGTATTCTTCGAGTGACATCAGAACAAGCCTCTCCATTCTGACGAGTAATAAGGATCGGG
T S L G D R G A L Q I R R T V D S C L G R W E S S Y Y P D P
Y I L R R K W R T T N K S H C * F V L R E M R V L L L S R P
L L Y A T E E L S N Y E E L S M L V C A E G N Q R T I L I P

A * S S T A F I I V A D K F * R A S L * L I V R M S I S S F
H D L Q R L S S L L P T N S D A L R C S * L Y A C Q S P L L
G M I F N G F H H C C R Q I L T R F A V A N C T H V N L L F
181 GCATGATCTTCAACGGCTTTCATCATTGTTGCCGACAAATTCGACGGCCTTCGCTGTAGCTAATGTAGCGATGTCAATCTCTCTTTT
C S D * P S E D N M G V F E S A S D O I * N V A H * D C P K
```

Ajout de la traduction sur 6 phases de lecture

Fermez cette fenêtre, puis relancez la traduction sur 6 phases avec une option alternative:

- Cliquez sur **Six-frame translation** puis sur **Add six-frame translation track**.

NIH National Library of Medicine
National Center for Biotechnology Information

ORFfinder submitting page

Open Reading Frame Viewer

Sequence

ORFs found: 10 Genetic code: 11 Start codon: 'ATG' only

Find: [input] Tools Tracks

(U) ORFfinder_10.6.184323536

1: 1..8.9K (8,851 nt)

Six-frame translation
Add six-frame translation track
Display six-frame translation
Six-frame translation...

ORF6 (434 aa) Display ORF as... Mark

```
>cl|ORF6
MSFNTIIDWNSCTAEQQRQLLRPAISASESITRTVNDILDNVKARGDEA
LREYSAFKDITVTALKVSAEILAAASERLSDLEKQMAVAIVQETTFHT
AQKLPYVDVETPGVRCQVTRPVASVGLYIPGGSAPLFSVLMLATPAS
IAGCKKVVLCSPPTADEILYAAQLCGVDVFNVGGAAIAALAFGTESV
PKVDKIFGPGNAFVTEAKRQVSORLDGAAIDMPAGPSEVLVIADSGATPD
FVASDLLSQAEHGPSQVILLTPAADMARRVAEAVERQLAELPRAETARQ
ALNASRLIVTKDLAQCVEISNQYGEHLIIQTRNARELVDSITSAGSVFL
GDWSPESAGDYASGNTNHLPTYGYTATCSSLGLADFQKRMVTQELSKEGF
SALASTIETLAAAERLTAHKNNAVTLRVNALKEQA
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	1638	2942	1305 434
ORF4	+	2	2939	4009	1071 356
ORF2	+	1	4009	5076	1068 355
ORF9	-	1	8845	7862	984 327
ORF1	+	1	733	1632	900 299
ORF3	+	1	6385	7161	777 258
ORF5	+	2	5666	6403	738 245
ORF8	+	3	7155	7766	612 203
ORF7	+	3	5076	5666	591 196
ORF10	-	2	7890	7369	522 173

ORF6

SmartBLAST

BLAST

Marked set (0)

SmartBLAST best hit titles...

BLAST

BLAST Database:

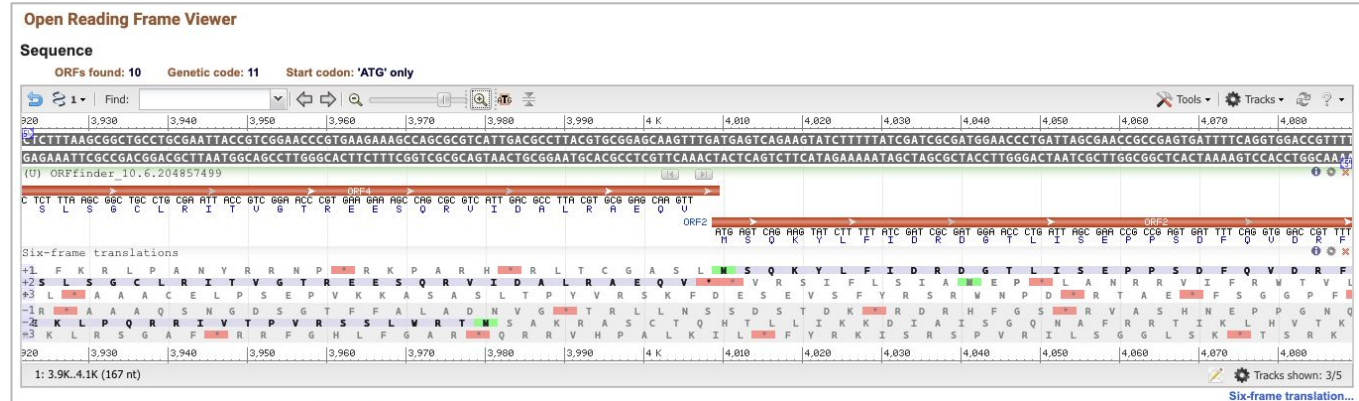
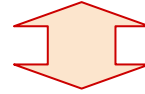
UniProtKB/Swiss-Prot (swissprot)

Exercice 5. Pistes de traduction sur les 6 phases de lecture

Qu'observez-vous dans la fenêtre qui s'affiche ?

Astuce : pour répondre aux questions 2 et 3, zoomez sur la carte jusqu'à faire apparaître l'enchaînement des résidus (acides aminés et nucléotides).

1. A quoi correspondent les pistes marquées +1, +2, +3, -1, -2, -3 ?
2. A quoi correspondent les traits verticaux verts ?
3. A quoi correspondent les traits verticaux rouges ?
4. A quoi correspondent les plages grises marquées d'une flèche ?
5. Y a-t-il le même nombre de plages grises que d'ORFs marqués en rouge ?
6. D'où vient la différence ?

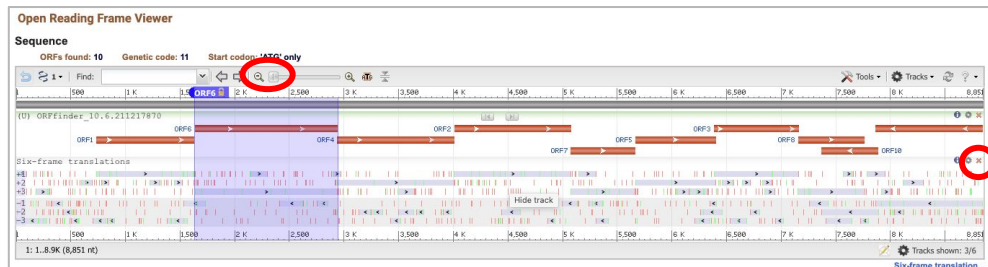


Tailles des régions intergéniques

Vous allez maintenant déterminer la taille des **régions intergéniques (RI)** entre ces ORFs.

Astuce: sous la carte des ORF, ORFfinder affiche un tableau indiquant les coordonnées génomiques et la taille des ORFs détectés. Vous pouvez récupérer les valeurs de ce tableau pour calculer la taille des régions intergéniques.

- **Dézoomez** complètement.
- **Refermez la piste six-frame translation** en cliquant sur la petite croix rouge en haut à droite (attention, ne fermez pas la piste des ORFs produits par ORFfinder).
- Sur le tableau de coordonnées des ORFs, cliquez sur l'en-tête de colonne **Start** pour **trier les ORFs par position**.
- **Notez les positions** de fin de l'ORF1 et de début de l'ORF6.
- **Calculez la taille de la RI** entre ORF1 et ORF6.
- **Zoomez sur la région intergénique entre ORF1 et ORF6** et vérifiez si la taille calculée correspond à ce que vous voyez, au nucléotide près.
- Faites de même pour les autres ORFs, pour déterminer la taille des RI qui les séparent.



Mark subset... Marked: 0 Download marked set as Feature table

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF1	+	1	733	1632	900 299
ORF6	+	3	1638	2942	1305 434
ORF4	+	2	2939	4009	1071 356
ORF2	+	1	4009	5076	1068 355
ORF7	+	3	5076	5666	591 196
ORF5	+	2	5666	6403	738 245
ORF3	+	1	6385	7161	777 258
ORF8	+	3	7155	7766	612 203
ORF10	-	2	7890	7369	522 173
ORF9	-	1	8845	7862	984 327

Exercice 6. Tailles des régions intergéniques

Questions (réponses numériques)

1. Quelle est la taille de la RI entre ORF1 et ORF6 ?
2. Quelle est la taille de la RI entre ORF2 et ORF7 ?
3. Quelle est la taille de la RI entre ORF5 et ORF3 ?

RI: région intergénique

QCM : que peut-on conclure de la taille de ces RI ?

1. Quelle structure serait présente sur ce fragment d'ADN chromosomique (une seule réponse) ?
 - UTR, intron, exon, opéron, site d'épissage
2. Quels ORFs seraient inclus dans cette structure (une ou plusieurs réponses) ?
 - ORF1, ORF6, ORF4, ORF5, ORF8, ORF10
3. Quels éléments vous permettent de conclure sur le nombre d'ORFs inclus dans cette structure (une ou plusieurs réponses) ?
 - Distances intergéniques courtes ou nulles
 - Chevauchements entre ORFs
 - Orientation des ORFs
 - Longueur des ORFs

Assignation de fonction par recherche de similarité

Vous allez maintenant vous intéresser à l'annotation fonctionnelle de ces ORFs détectés dans le fragment d'ADN chromosomique étudié.

Pour cela, le plus simple est de faire une recherche par similarité dans une base de données (outil BLAST), afin de comparer les ORFs identifiés aux séquences déjà connues et répertoriées dans les bases de données.

Vous allez ainsi vérifier à quel gène pourraient correspondre les ORF1 et 10.

- Cliquez sur l'**ORF1**. La traduction de l'ORF en protéine s'affiche dans l'encadré du dessous.
- Cliquez sur le bouton **BLAST**. Vous lancez ainsi une recherche de similarité en comparant la séquence protéique traduite de l'ORF1 avec chacune des séquences d'une base de données.

A partir de la page de résultats du BLAST, répondez aux questions suivantes

The screenshot displays the 'Open Reading Frame Viewer' interface. At the top, it shows 'Sequence' information: 'ORFs found: 10', 'Genetic code: 11', and 'Start codon: 'ATG' only'. A sequence viewer shows a genomic region from 1.500 to 8.950 Kbp. Several ORFs are highlighted with red arrows, with ORF1 and ORF10 circled in red. Below the viewer, a window for 'ORF1 (299 aa)' is open, showing the amino acid sequence:

```
>Lc1|ORF1
MTDNRLLPIAMQKSGRLSDSRELLARCGTKINLHRLQIAMAENMPIDI
LRVRDDIPGLVMDGVVDLGIIGENVLEELNRRAGDEPRYFTLRRLD
FGGCRSLATPVDEAWDGPLSLNKGRIATSYPHLLKRYLDKIGISFKSCL
LNSVSEVAPRAQLADADCOLVYTGATLEAWLREVEVYRSKACLIQDDG
EMEEKQLIDKLLTRIGVIGAESKYIMHAPTERLDEVIALPGAER
PTTLPLLAGDQQRVAHMVSSSETLFWETMEKLLKAGASSLVLPTEKME
```

 Below the sequence, there are buttons for 'SmartBLAST' and 'BLAST', with the 'BLAST' button circled in red. A 'Marked set (0)' section is also visible. On the right, a table titled 'Mark subset...' shows the following data:

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	1638	2942	1305 434
ORF4	+	2	2939	4009	1071 356
ORF2	+	1	4009	5076	1068 355
ORF9	-	1	8845	7862	984 327
ORF1	+	1	733	1632	900 299
ORF3	+	1	6385	7161	777 258
ORF5	+	2	5666	6403	738 245
ORF8	+	3	7155	7766	612 203
ORF7	+	3	5076	5666	591 196
ORF10	-	2	7890	7369	522 173

Exercice 7. Assignment de fonction par recherche de similarité

Questions

1. Quelle est la modalité de BLAST utilisée (une seule réponse) ?
 - o blastn, blastp, blastx, tblastn
2. En quoi consiste une recherche par BLASTP (une seule réponse) ?
 - o Requête protéine versus base de données de protéines
 - o Requête nucléique versus base de données de protéines
 - o Requête protéine versus base de données nucléique
3. Quelle base de données avez-vous interrogée lors de cette requête (une seule réponse) ?
 - o UniprotKB TREMBL, UniprotKB complet, Swiss-prot
4. Combien de résultats obtenez-vous (réponse numérique) ?
5. Pour la séquence cible la plus similaire à la requête soumise, quel est le pourcentage d'identité obtenu (réponse numérique) ?
6. Quel est son pourcentage de couverture par rapport à la séquence requête (réponse numérique) ?

The screenshot displays the NCBI BLAST search results for a query protein sequence. The top section shows the query details: Job Title, Protein Sequence (RID: G5W9X5S2013), Program (BLASTP), Database (swissprot), Query ID (Ic|Query_5931811), and Description (Ic|ORF1:733:1632 unnamed protein product). A 'Filter Results' panel on the right allows filtering by Organism, Percent Identity, E value, and Query Coverage.

The 'Sequences producing significant alignments' table is shown below, with columns for Description, Scientific Name, Max Score, Total Score, Query Cover, E value, Per. Ident, Acc. Len, and Accession. The table lists 10 results, all with 100% query coverage and E values near 0.0.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	607	607	100%	0.0	100.00%	299	A1ACN1.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Shigella dys...	607	607	100%	0.0	99.67%	299	Q32EE8.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	606	606	100%	0.0	99.67%	299	B7NQH1.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	605	605	100%	0.0	99.67%	299	B7LUF0.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	605	605	100%	0.0	99.33%	299	B5YU75.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	604	604	100%	0.0	99.33%	299	B7NC59.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	604	604	100%	0.0	99.67%	299	B7MWT8...
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Shigella flex...	603	603	100%	0.0	99.67%	299	Q0T3A8.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Salmonella ...	589	589	100%	0.0	96.32%	299	B4SX40.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Salmonella ...	588	588	100%	0.0	95.99%	299	B4TMR4.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Salmonella ...	587	587	100%	0.0	95.99%	299	B5BFC1.1

Exercice 7. Assignment de fonction par recherche de similarité

Questions (suite)

- Quelle est la E-value obtenue (réponse numérique) ?
- Que signifie une e-value de 0.0 (une seule réponse) ?
 - Similarité non significative
 - Ressemblance nulle
 - Similarité extrêmement significative
 - Impossibilité statistique d'obtenir un alignement aussi bon par hasard
- Est-ce que ces deux séquences alignées sont vraisemblablement homologues ? (Oui / Non)
- Quelle fonction peut-on ainsi associer à l'ORF1 (une seule réponse) ?
 - Histidinol dehydrogenase
 - ATP phosphoribosyltransferase
 - Histidinol-phosphate aminotransferase
 - ATP-PRT
 - Polysaccharide antigen chain regulator

The screenshot displays a search results page for a protein sequence. The top section shows the protein details for 'BLASTP' search against the 'swissprot' database. The protein is identified as 'ORF1:733:1632 unnamed protein product' with a length of 299 amino acids. A 'Filter Results' panel on the right allows filtering by organism, percent identity, E-value, and query coverage.

The main section shows a table of 'Sequences producing significant alignments'. The table has columns for Description, Scientific Name, Max Score, Total Score, Query Cover, E value, Per. Ident, Acc. Len, and Accession. All sequences shown have an E-value of 0.0, indicating high significance.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	607	607	100%	0.0	100.00%	299	A1ACN1.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Shigella dys...	607	607	100%	0.0	99.67%	299	Q32EE8.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	606	606	100%	0.0	99.67%	299	B7NQH1.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	605	605	100%	0.0	99.67%	299	B7LUF0.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	605	605	100%	0.0	99.33%	299	B5YU75.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	604	604	100%	0.0	99.33%	299	B7NC59.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Escherichia ...	604	604	100%	0.0	99.67%	299	B7MWT8...
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Shigella flex...	603	603	100%	0.0	99.67%	299	Q0T3A8.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Salmonella ...	589	589	100%	0.0	96.32%	299	B4SX40.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Salmonella ...	588	588	100%	0.0	95.99%	299	B4TMR4.1
RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRT...	Salmonella ...	587	587	100%	0.0	95.99%	299	B5BFC1.1

Identification du gène correspondant à ORF1

Vous allez maintenant rechercher le nom du gène correspondant à l'ORF1.

- Dans la dernière colonne du tableau de résultats de BLAST, cliquez sur le "**numéro d'accèsion**" de la séquence la plus similaire à l'ORF1. Ceci ouvre dans un nouvel onglet la fiche de la séquence protéique correspondante.
- Dans la section "**FEATURES**" (annotations de la séquence), trouvez le premier objet de type gène, et consultez son nom (attribut `/gene` de l'objet `"gene"`).

Exercice 7b : Identification du gène correspondant à ORF1

1. Quel est le nom du gène correspondant à l'ORF1 (une seule réponse) ?
 - HIS1_ECOK1
 - A1ACN1
 - Glycosyltransferase
 - hisG
 - HisG
 - ATP-PRT
 - ATP-PRTase
 - ATP phosphoribosyltransferase

RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRTase

UniProtKB/Swiss-Prot: A1ACN1.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: [C](#)

LOCUS HIS1_ECOK1 299 aa linear BCT 02-OCT-2024

DEFINITION RecName: Full=ATP phosphoribosyltransferase; Short=ATP-PRT; Short=ATP-PRTase.

ACCESSION A1ACN1
VERSION A1ACN1.1

DBSOURCE UniProtKB: locus HIS1_ECOK1, accession [A1ACN1](#);

class: standard,
created: Jan 15, 2008,
sequence updated: Jan 23, 2007,
annotation updated: Oct 2, 2024.

xrefs: CP000468.1, AB01421.1, WP_000131782.1
xrefs (non-sequence databases): AlphaFoldDB:A1ACN1, SMR:A1ACN1,
GeneID:86946973, KEGG:ecv:APEC01_1116, HOGONOM:CLU_038115_1_0_6,
UniPathway:UPA00031, Proteomes:UP00000216, GO:0005737, GO:0005524,
GO:000379, GO:0000287, GO:0000105, CCD:cd13592,
Gene3D:3.90.120, Gene3D:3.40.190.10, HAMAP:MF_00079,
InterPro:IPR030621, InterPro:IPR013820, InterPro:IPR019198,
InterPro:IPR001348, InterPro:IPR013115, InterPro:IPR011322,
InterPro:IPR015867, NCBIfam:TIGR00070, NCBIfam:TIGR03455,
PANTHER:PTHR21403:SF8, PANTHER:PTHR21403, Pfam:PF01634,
Pfam:PF08029, SUPFAM:SSF54913, SUPFAM:SSF53850, PROSITE:PS01316

KEYWORDS Amino-acid biosynthesis; ATP-binding; Cytoplasm;
Glycosyltransferase; Histidine biosynthesis; Magnesium;
Metal-binding; Nucleotide-binding; Reference proteome; Transferase.

SOURCE Escherichia coli APEC 01

ORGANISM Escherichia coli APEC 01
Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales;
Enterobacteriaceae; Escherichia.

REFERENCE 1 (residues 1 to 299)

AUTHORS Johnson,T.J., Kariyawasam,S., Wannemuehler,Y., Mangiamle,P.,
Johnson,S.J., Doetkott,C., Skyberg,J.A., Lynne,A.H., Johnson,J.R.
and Nolan,K.

TITLE The genome sequence of avian pathogenic Escherichia coli strain
01:K1:H7 shares strong similarities with human extraintestinal
pathogenic E. coli genomes

JOURNAL J Bacteriol 189 (8), 3228-3236 (2007)

PUBMED 17293413

REMARK NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
Erratum:J Bacteriol. 2007 Jun;189(12):4554

COMMENT [FUNCTION] Catalyzes the condensation of ATP and 5-phosphoribose
1-diphosphate to form N⁵-(5'-phosphoribosyl)-ATP (PR-ATP). Has a
crucial role in the pathway because the rate of histidine
biosynthesis seems to be controlled primarily by regulation of HisG
enzymatic activity. (ECO:0000255|HAMAP-Rule:MF_00079).

[CATALYTIC ACTIVITY] Reaction=1-(5-phospho-beta-D-ribose)-ATP +
diphosphate = 5-phospho-alpha-D-ribose 1-diphosphate + ATP;
Xref=rhea:RHEA:18473, CHEBI:CHEBI:30616, CHEBI:CHEBI:33019,
CHEBI:CHEBI:58017, CHEBI:CHEBI:73183; EC=2.4.2.17;
Evidence=(ECO:0000255|HAMAP-Rule:MF_00079).

[COFACTOR] Name=Mg(2+); Xref=CHEBI:CHEBI:18420;

Evidence=(ECO:0000255|HAMAP-Rule:MF_00079).

ACTIVITY REGULATION: Feedback inhibited by histidine.

(ECO:0000255|HAMAP-Rule:MF_00079).

[PATHWAY] Amino-acid biosynthesis; L-histidine biosynthesis;
L-histidine from 5-phospho-alpha-D-ribose 1-diphosphate: step 1/9.
(ECO:0000255|HAMAP-Rule:MF_00079).

[SUBUNIT] Equilibrium between an active dimeric form, an inactive
hexameric form and higher aggregates. Interconversion between the
various forms is largely reversible and is influenced by the
natural substrates and inhibitors of the enzyme.
(ECO:0000255|HAMAP-Rule:MF_00079).

[SUBCELLULAR LOCATION] Cytoplasm (ECO:0000255|HAMAP-Rule:MF_00079).
[SIMILARITY] Belongs to the ATP phosphoribosyltransferase family.
Long subfamily. (ECO:0000255|HAMAP-Rule:MF_00079).

FEATURES

source Location/Qualifiers

1..299

/organism="Escherichia coli APEC 01"

/db_xref="taxon:405955"

gene 1..299

/gene="hisG"

/locus_tag="EcoK1_19270"

/gene_synonym="APEC01_1116"

1..299

Protein /product="ATP phosphoribosyltransferase"

/EC_number="2.4.2.17"

/note="ATP-PRT; ATP-PRTase"

/UniProtKB_evidence="Inferred from homology"

Fonction et identification de l'ORF10

L'ORF10 chevauche étonnamment l'ORF8 de manière importante, sur une grande partie de sa longueur. Afin de tenter de déterminer à quel gène pourrait correspondre cet ORF10, vous allez donc faire, pour l'ORF10, la même manipulation que celle faite pour l'ORF1.

- Sur la page de résultat de la recherche d'ORF, commencez par trouver la taille en nt de l'ORF10, ainsi que la taille en aa de la protéine potentielle codée par cet ORF10. Pour cela, vous avez deux possibilités :
 - Trouver ces informations dans l'encadré qui récapitule les ORFs trouvés sur le fragment d'ADN.
 - Placer le curseur sur l'ORF10 (sans cliquer) : la fenêtre qui s'ouvre contient les caractéristiques de l'ORF10.
- Cliquez ensuite sur l'ORF10, afin de le sélectionner. La traduction de l'ORF en protéine s'affiche dans l'encadré du dessous.
- Cliquez sur le bouton BLAST pour lancer la recherche par similarité de séquences avec cette séquence protéique.

The screenshot displays a BLAST search interface. The top section shows job details for 'Job Title' (Protein Sequence), 'RID' (G609K7EB013), 'Program' (BLASTP), and 'Database' (swissprot). A 'Filter Results' panel on the right allows filtering by organism and setting identity and coverage thresholds. Below the job details, a table titled 'Sequences producing significant alignments' is shown. The table has columns for Description, Scientific Name, Max Score, Total Score, Query Cover, E value, Per. Ident, Acc. Len, and Accession. Six sequences are listed, all with high scores and low E values, indicating significant similarity.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin);...	Bacteroides ...	31.2	31.2	19%	4.6	48.57%	626	Q5L7W2
RecName: Full=4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin);...	Bacteroides ...	31.2	31.2	19%	4.7	48.57%	626	Q64N34
RecName: Full=Dynein axonemal heavy chain 1; AltName: Full=Axonemal beta dyn...	Mus musculus	31.2	31.2	50%	5.5	32.26%	4250	E9Q8T7
RecName: Full=Splicing factor U2af large subunit A; AltName: Full=NpU2AF65a; Alt...	Nicotiana gl...	30.8	30.8	32%	7.0	33.93%	555	Q9ZR39
RecName: Full=Uncharacterized protein YuaQ [Escherichia coli K-12]	Escherichia ...	30.8	30.8	28%	7.5	30.61%	1371	Q9JMS3
RecName: Full=4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin);...	Bacteroides ...	30.4	30.4	19%	9.0	42.86%	613	Q8A4T0

Fonction et identification de l'ORF10

Exercice 7c : Fonction et identification de l'ORF10

1. Combien de résultats obtenez-vous (réponse numérique) ?
2. Quelle fonction pouvez-vous assigner à l'ORF10 sur base du résultat (une ou plusieurs réponses) ?
 - 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin)
 - Dynein axonemal heavy chain 1
 - Splicing factor U2af large subunit A
 - Uncharacterized protein YuaQ
 - Aucune: les 6 résultats obtenus ne sont pas significatifs, car les e-values obtenues sont trop élevées (toutes > 1)
 - Aucune : ces 6 séquences cibles ne sont vraisemblablement pas homologues de l'ORF10
 - Aucune: l'ORF10 est vraisemblablement un faux positif, une fausse prédiction d'ORFfinder.

Job Title

Protein Sequence

RID

[G609K7EB013](#) Search expires on 10-08 05:29 am

[Download All](#) ▼

Program

BLASTP [?](#) [Citation](#) ▼

Database

swissprot [See details](#) ▼

Query ID

Icl|Query_4741941

Description

Icl|ORF10:7890:7369 unnamed protein product

Molecule type

amino acid

Query Length

173

Other reports

[Distance tree of results](#) [Multiple alignment](#)

[MSA viewer](#) [?](#)

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to E value to Query Coverage to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download ▼ Select columns ▼ Show 100 ▼ [?](#)

select all 6 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	RecName: Full=4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin);...	Bacteroides ...	31.2	31.2	19%	4.6	48.57%	626	Q5L7W2
<input checked="" type="checkbox"/>	RecName: Full=4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin);...	Bacteroides ...	31.2	31.2	19%	4.7	48.57%	626	Q64N34
<input checked="" type="checkbox"/>	RecName: Full=Dynein axonemal heavy chain 1; AltName: Full=Axonemal beta dyn...	Mus musculus	31.2	31.2	50%	5.5	32.26%	4250	E9Q8T7
<input checked="" type="checkbox"/>	RecName: Full=Splicing factor U2af large subunit A; AltName: Full=NpU2AF65a; Alt...	Nicotiana gl...	30.8	30.8	32%	7.0	33.93%	555	Q9ZR39
<input checked="" type="checkbox"/>	RecName: Full=Uncharacterized protein YuaQ [Escherichia coli K-12]	Escherichia ...	30.8	30.8	28%	7.5	30.61%	1371	Q9JMS3
<input checked="" type="checkbox"/>	RecName: Full=4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin);...	Bacteroides ...	30.4	30.4	19%	9.0	42.86%	613	Q8A4T0

Recherche d'information sur RegulonDB

- Connectez-vous à la base de connaissances [RegulonDB](#).
- Sur ce site, effectuez une recherche avec le nom de gène que vous avez trouvé précédemment pour l'ORF1. Pour cela, entrez simplement le nom de gène dans la barre de recherche.
- Cliquez sur l'unique résultat qui apparaît dans la section Gene.
- Dézoomez et recadrez la carte avec les flèches.

RegulonDB Search

Gene: **hisG** Product: **ATP phosphoribosyltransferase**

Synonyms: EG10449 RegulonDB: RDBECOLIGNC00442
Length: 900 bp ECOCYC: EG10449
Position: 2090192 -> 2091091 REFSEQ: b2019
Location: cytosol

Search Tools Releases & Downloads API & software Help

hisG

2089192 2092091

yeoB yefM yefMn hisLp hisL hisG hisD

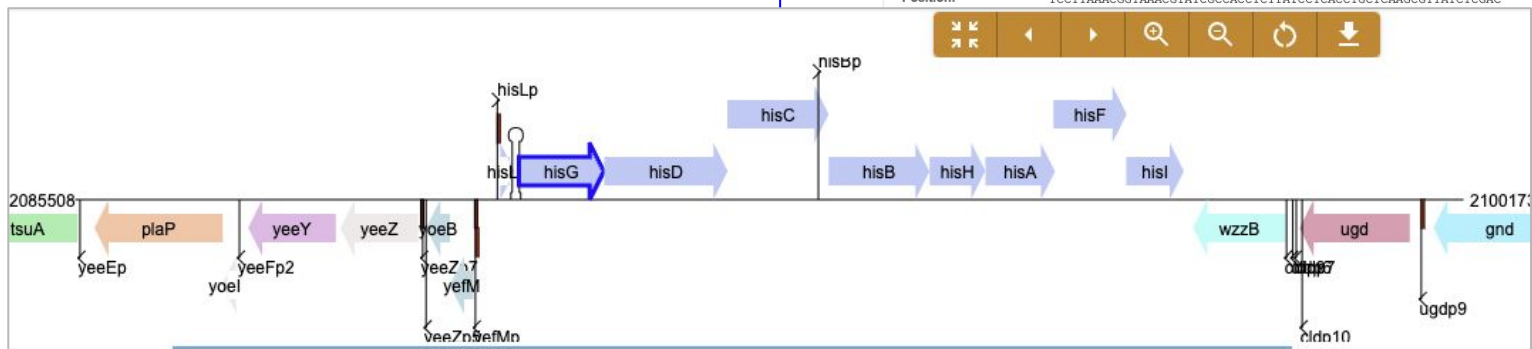
yeoZn5

Description

Description Synonyms: EG10449 Sequence: Format: Fasta OPTIONS DOWNLOAD
MultifunTerms Bnumber: b2019
Regulation Position: 2090192 -> 2091091 RegulonDB|RDBECOLIGNC00442|gene: hisG|product: ATP phosphoribosyltransferase|size: 900
Product: ATP phosph... Size: 900 bp
Strand: forward TCACGCCAATTTCCTGGCGCCTGTGGCATTAATAATTAATCTTCACACCCAGCGCTGATC
GC 54.11% GCATTCGCAGAAACAATGCCGATTGATATCTCGCGGGTGGTGCACGACACATCCCGT
content: CTGGTAATGGATGGCGGTGAGACTTGGGATATCCGGAAACCTCTGGAAAGAG
Centisome 45.031208 CTGCTTAACCGCCGCCAGGGTGAAGATCCACGCTACTTACCCTCGCTGTCTGAT
Position: TCCTTAAACGGTAAACGATCGCCACCTCTATCCTCACCTGCTCAACGGTATCTCGAC

Navigation
Operon
HT-Datasets
Related Tools
Download Options
External Cross References
FeedBack

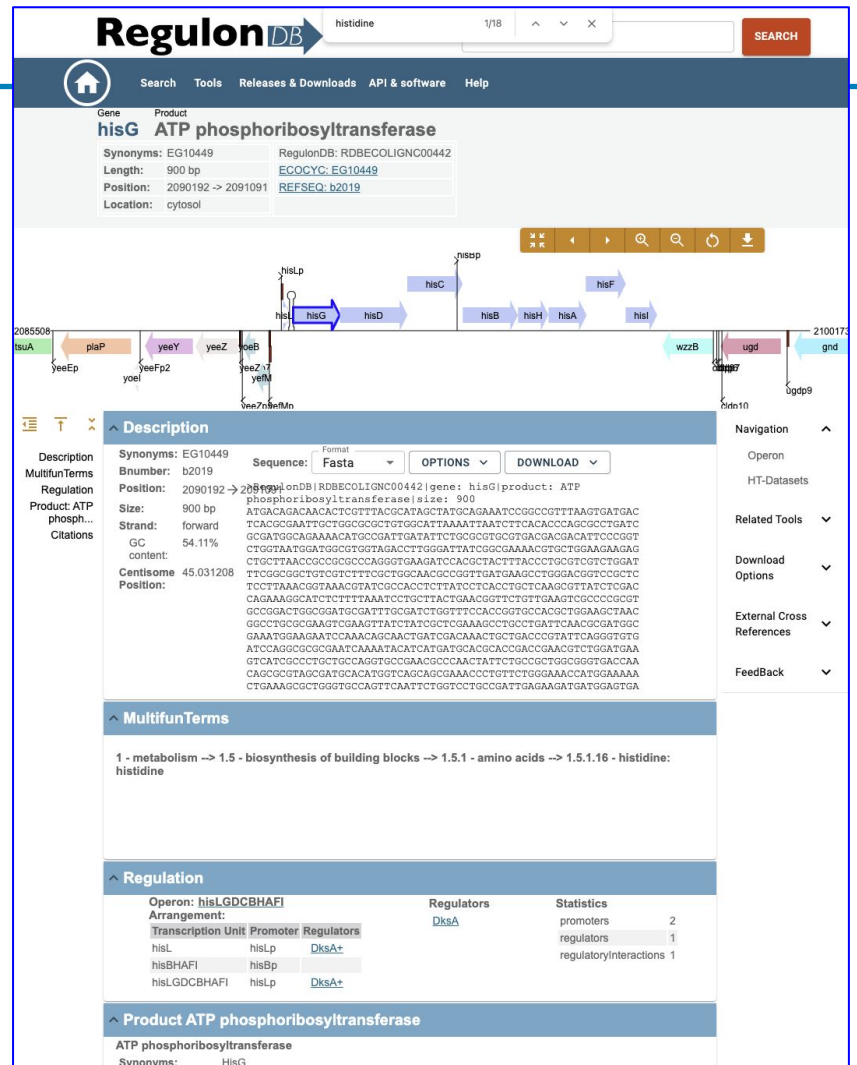
Operon: hisLGDCBHAFI Regulators Statistics



Exercice 8. Consultation de RegulonDB

Questions

1. D'après la carte des gènes sur RegulonDB, le gène étudié se trouve-t-il bien dans la structure supposée précédemment ? (oui / non)
2. Dans les informations qui apparaissent sous la carte, trouvez le nom de l'opéron. Cliquez sur le nom de l'opéron, puis répondez aux questions suivantes. Sur quel brin ce trouve cet opéron ? (forward / reverse)
3. Combien y-a-t-il de gènes dans cet opéron ? 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
4. Combien y-a-t'il de promoteurs dans cet opéron ? 0, 1, 2, 3
5. Quel est le type de terminateur de la transcription présent dans cette unité de transcription ?
 - Dépendant de Rho
 - Indépendant de Rho
6. Quelle est la fonction de cet opéron ?
 - Catalyser la voie de biosynthèse de l'histidine
 - Réguler l'hystérie métabolique
 - Réguler la terminaison de la transcription
 - Activer en cascade la transcription d'une série d'autres gènes



The screenshot shows the RegulonDB website interface for the gene **hisG**. The main header includes the site name "RegulonDB", a search bar, and navigation links for Search, Tools, Releases & Downloads, API & software, and Help. Below the header, the gene information is displayed, including Synonyms (EG10449), RegulonDB ID (RDCECOLIGNC00442), Length (900 bp), Position (2090192 -> 2091091), and Location (cytosol). The main content area features a genomic map of the **hisG** operon, showing the gene structure with arrows indicating the direction of transcription. The operon includes genes **hisLp**, **hisL**, **hisG**, **hisD**, **hisC**, **hisB**, **hisH**, **hisA**, and **hisI**. Other genes shown include **plpP**, **yeeY**, **yeeZ**, **yeeE**, **yeeFp2**, **yeeZ**, **yeeI**, **yeeK**, **wzzB**, **ugd**, and **gnd**. The **hisG** gene is highlighted in blue. Below the map, the "Description" section provides detailed information about the gene, including its synonyms, sequence (in FASTA format), and size (900 bp). The "Regulation" section shows the operon name **hisLGDCBHAFI** and lists regulators such as **hisLp**, **hisBhafi**, and **DksA+**. The "Product" section identifies the gene as **ATP phosphoribosyltransferase**.

Gene: Product
hisG ATP phosphoribosyltransferase

Synonyms: EG10449 | RegulonDB: RDCECOLIGNC00442
Length: 900 bp | ECOYC: EG10449
Position: 2090192 -> 2091091 | REFSEQ: b2019
Location: cytosol

Description
Synonyms: EG10449
Bnumber: b2019
Position: 2090192 -> 2091091
Size: 900 bp
Strand: forward
GC content: 54.11%
Content: 45.031208
Position: 2090192 -> 2091091

Regulation
Operon: hisLGDCBHAFI
Arrangement:
Transcription Unit Promoter Regulators
hisL hisLp DksA+
hisBHAFI hisBp
hisLGDCBHAFI hisLp DksA+

Product ATP phosphoribosyltransferase
ATP phosphoribosyltransferase
Synonyms: HisG

Exercice 8. Consultation de RegulonDB

- Comparez votre prédiction d'ORFs avec ORFfinder à la carte de l'opéron sur RegulonDB. Avez-vous détecté l'ensemble des gènes de l'opéron avec ORFfinder ? (oui / non)
- Quel est le gène manquant dans votre résultat d'ORFfinder (une seule réponse) ?
 - yefM, hisLp, hisL, wzzB, hisA
- Cliquez sur ce gène, puis trouvez sa taille (réponse numérique)
- Pourquoi ce gène n'est-il pas détecté lors de la recherche avec ORFfinder (une ou plusieurs réponses) ?
 - Ce gène n'est pas codant
 - La taille minimale d'ORFfinder était de 75bp
 - La taille minimale d'ORFfinder était de 300bp
 - Ce gène est situé sur le brin complémentaire

Open Reading Frame Viewer

Sequence

ORFs found: 10 Genetic code: 11 Start codon: 'ATG' only

ORF9 (327 aa) Display ORF as... Unmark

```
>|c1|ORF9
MRRVEMVVSQNHDPFEIDLLDLVLRGQHTIIISVIAIALATGYL
AVAKEKWTSTAIITFPDVGQIAGYNNMVIYGGAAKPKVSLDETLIGRP
SSAFSALAEITLNDREEREKLTIEPSVNDLPLTVSYVGTAGAQNKLA
QYTESQKQKNSLEKPKLKNLALGNLQDLSRTESVAGKQKQLEIRG
LDEALQYANQAVYTPQIQOTGEDITDPTFLGSEALSMIKHEATRPL
VFSFNYQTRNLLDIESLAVGDLSDHVRVYKQMLPDRSDPKKATLL
ILAVLLGHVGGAGVLGRNALRNYNAK
```

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	1638	2942	1305 434
ORF4	+	2	2939	4009	1071 356
ORF2	+	1	4009	5076	1068 355
ORF9	-	1	8845	7862	984 327
ORF1	+	1	733	1632	900 299
ORF3	+	1	6385	7161	777 258

Gene Product

hisG ATP phosphoribosyltransferase

Synonyms: EG10449 RegulonDB: RDBECOLIGNC00442
ECOCYC: EG10449

Length: 900 bp

Position: 2090192 -> 2091091 REFSEQ: b2019

Location: cytosol

Genomic map showing genes: hisG, hisD, hisB, hisH, hisA, hisI, hisLp, hisC, hisF, hisL, yefM, yeeZ, yeeB, yeeFp2, yeeI, wzzB, ugd, gnd, plaP, yefM, yeeZ, yeeB, yeeFp2, yeeI, wzzB, ugd, gnd, yefM, yeeZ, yeeB, yeeFp2, yeeI, wzzB, ugd, gnd.

Exercice 8. Consultation de RegulonDB

11. D'après la carte de l'opéron, à quel gène correspondrait l'ORF9 détecté avec ORFfinder (une seule réponse) ?
 - o *yeeZ, hisI, wzzB, ugd*
12. Comment s'appelle le gène directement en amont de l'opéron ?
 - o *wzzB, ugd, hisL, hisLp, yefM*
13. Pourquoi ce gène n'est-il pas détecté lors de la recherche avec ORFfinder (une seule réponse) ?
 - o Il n'est pas codant
 - o Il est situé sur le brin complémentaire
 - o Il est en amont de la séquence fournie en début d'exercice.
 - o La taille minimale d'ORFfinder était de 300bp

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	1638	2942	1305 434
ORF4	+	2	2939	4009	1071 356
ORF2	+	1	4009	5076	1068 355
ORF9	-	1	8845	7862	984 327
ORF1	+	1	733	1632	900 299
ORF3	+	1	6385	7161	777 258

Gene *hisG* **Product** ATP phosphoribosyltransferase
Synonyms: EG10449 **RegulonDB:** RDBECOLIGNC00442
Length: 900 bp **ECOCYC:** EG10449
Position: 2090192 -> 2091091 **REFSEQ:** b2019
Location: cytosol

Operon map showing genes: *plpA*, *yeeY*, *yeeZ*, *yeeB*, *yeeFp2*, *yeeI*, *yeeZ*, *yefM*, *hisLp*, *hisL*, *hisG*, *hisD*, *hisC*, *hisB*, *hisH*, *hisA*, *hisI*, *wzzB*, *ugd*, *gnd*. Promoters are indicated by arrows pointing to the right.