

Introduction à la bioinformatique (UE SSV3U15)

TP2. Du gène à la protéine

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 (éléments structurels des chromosomes, régions géniques et intergéniques, opérons bactériens).
- Quelques éléments de génomique comparative (conservation / divergence, réarrangements chromosomiques, synténie)
- Les principaux types d'homologie : orthologie, paralogie
- Annotation fonctionnelle des gènes.
- Transcriptomique : expression différentielle des gènes dans différents tissus
- Protéomique : analyse de l'ensemble des protéines codées par un génome

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

Etapes

- Annotations génomiques dans la région du gène humain PAX6
- Génomique comparative : homologues de PAX6 chez les métazoaires
- Organisation des gènes en intron chez les bactéries

Complétion

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

Éléments de contexte

Annotations génomiques

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.

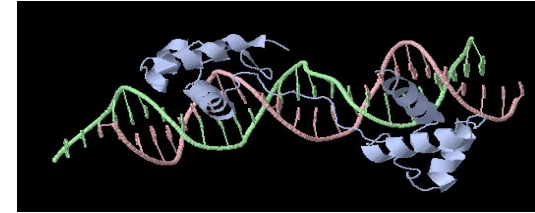
Feature type	Gencode	Ensembl	RefSeq	CHES
Protein-coding genes	19 901	20 376	20 345	21 306
lncRNA genes	15 779	14 720	17 712	18 484
Antisense RNA	5 501		28	2 694
Miscellaneous RNA	2 213	2 222	13 899	4 347
Pseudogenes	14 723	1 740	15 952	
Total transcripts	203 835	203 903	154 484	323 827

Exemples traités

Facteur transcriptionnel – PAX6

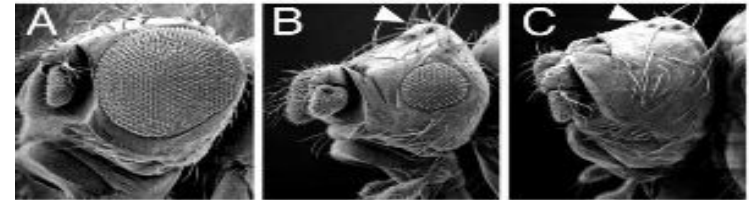
- La protéine PAX6 (violet sur l'image du haut) est un facteur transcriptionnel, qui se lie à des sites spécifiques sur l'ADN génomique (vert et rose).
- PAX6 contrôle l'expression de gènes impliqués dans la formation de l'oeil. Les gènes-cibles de PAX6 sont encore pour la plupart inconnus à ce jour.
- **Perte de fonction:** chez la drosophile, l'inactivation de *eyeless* (= PAX6) provoque une malformation ou une absence d'oeil.
- **Gain de fonction:** des mouches chez lesquelles on force le gène *eyeless* à s'exprimer dans les antennes développent des structures optiques à la place des antennes.

Liaison PAX6 - ADN

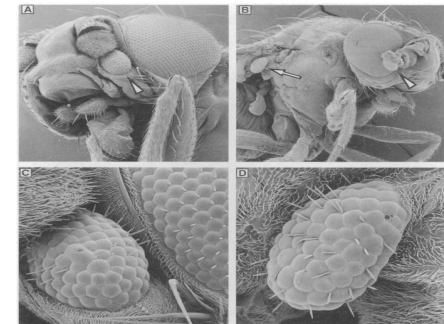


<http://www.rcsb.org/pdb/explore.do?structureId=6PAX>

Phénotype de perte de fonction



Phénotype de gain de fonction

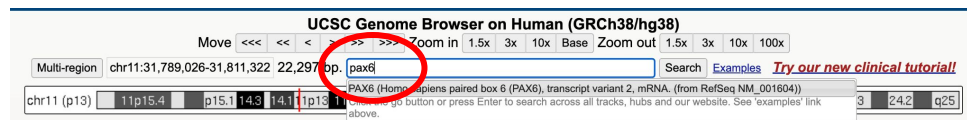
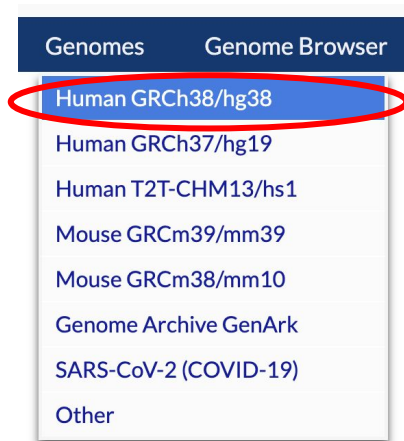


Tutoriel illustré

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 **génom**e **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** - chr11:31817960-31887040

☐ Gencode Genes:

- PAX6** (ENST00000604368.2) - chr11:31789026-31811322 - **PAX6** ENST00000604368.2 Homo sapiens paired box 6 PAX6 transcript variant
- PAUPAR** (ENST00000644607.1) - chr11:31816266-32002405 - PAUPAR ENST00000644607.1 PAX6 upstream antisense RNA from HGNc PAUPAR BX648962 uc285izg.1 uc285izg.1
- BCL2L15** (ENST00000393316.8) - chr11:13876816-13887581 - ... Q5TCB7 P50222 MEOX2 NbExp 3 IntAct EBI-10247136 EBI-748397 Q5TCB7 P26367 **PAX6** NbExp 3
- IntAct EBI-10247136 EBI-747278 Q5TCB7 P62487 POLR2G NbExp
- LYSM1D1** (ENST0000038980.10) - 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
- CDC1103** (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 Q6NRD5 PICK1 NbExp
- SLC12A8** (ENST00000469902.6) - chr3:125082644-125212748 - ... a role in the control of keratinocyte proliferation A0AV02 P26367 **PAX6** NbExp 3 IntAct EBI-11737524 EBI-747278 Membrane Multi-pass membrane protein
- TCF11L1** (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
- ZNFS13** (ENST00000337070.1) - 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 RNFI1 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:

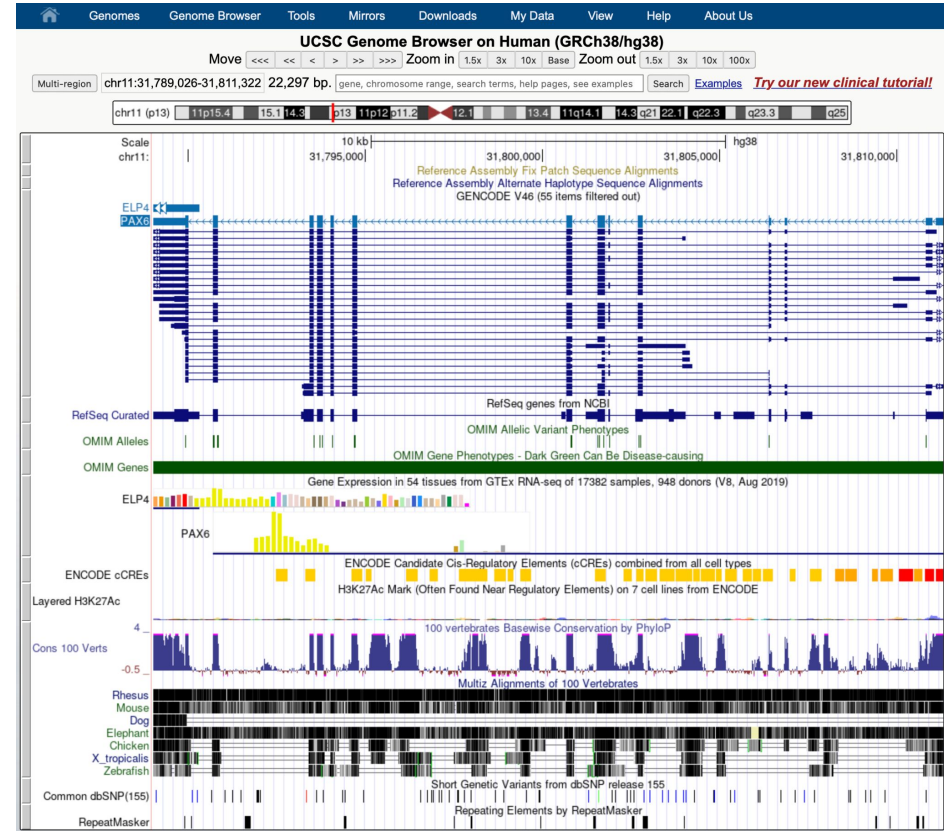
- PAX6** - chr11:31789026-31811121 - (NM_000280) paired box protein Pax-6 isoform a
- PAX6** - chr11:31789026-31811322 - (NM_001258464) paired box protein Pax-6 isoform a
- PAX6** - chr11:31789026-31812203 - (NM_001310158) paired box protein Pax-6 isoform b
- PAX6** - chr11:31789026-31817961 - (NM_001127612) 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** - chr11:31793205-31806929 - (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

- Connectez-vous au [UCSC Genome Browser](#)
- Sélectionnez la version **hg38** du **génom humain**
- Entrez le nom du gène d'intérêt (**PAX6**) dans la boîte de recherche et cliquez sur **Search**.
- Sous le titre "HUGO Gene Nomenclature", cliquez sur le lien **PAX6** - chr11:31789026-31817960
- Descendez sous la carte génomique pour afficher les choix de pistes d'annotations.
- 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 Predictions**", sélectionnez le mode "**pack**" pour la piste "**GENCODE_V46**".
- 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 on human chromosome 11 (hg38). The main view shows a genomic map with various tracks. Two tracks are highlighted with red circles:

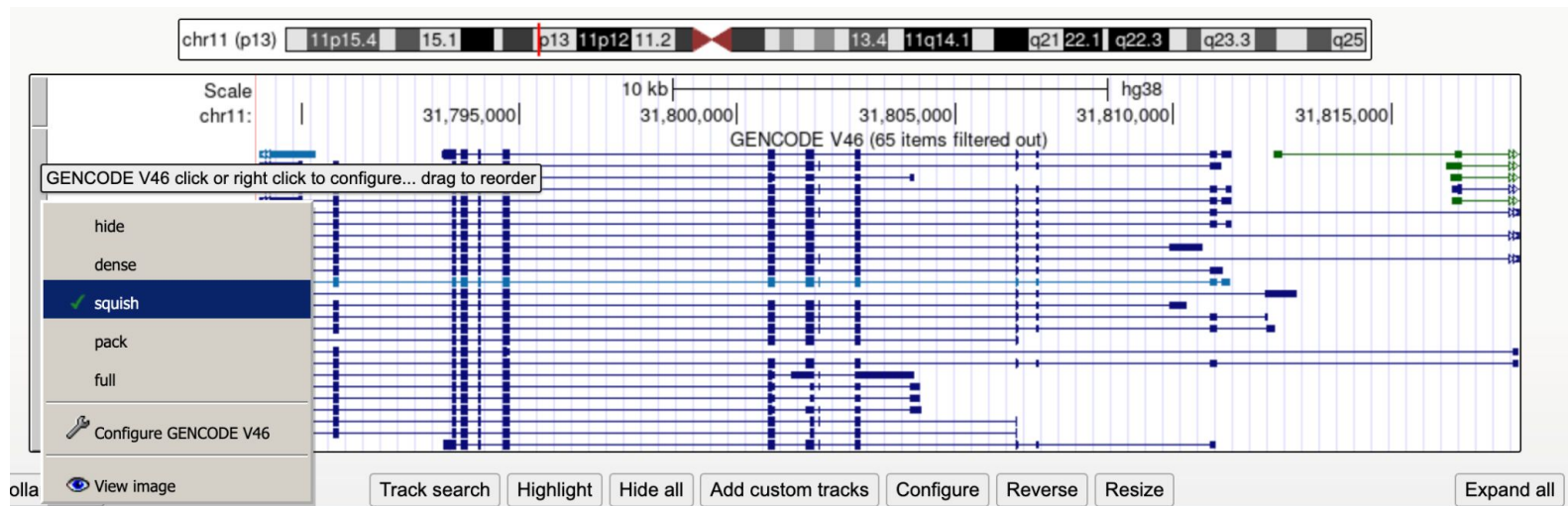
- Base Position** (under Mapping and Sequencing): Set to 'dense' mode.
- GENCODE_V46** (under Genes and Gene Predictions): Set to 'pack' mode.

The interface also includes a search bar, a genomic map, and various tracks for annotations. The 'Refresh' button is visible next to the 'Genes and Gene Predictions' 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-click) sur la barre grise à sa gauche. 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 **squish**, qui vous permet généralement de visualiser les transcrits alternatifs en occupant une place raisonnable.



Régions répétitives

- Dans la catégorie "**Repeats**", activez l'affichage de "**Repeatmasker**" en format "**dense**" et cliquez "**Refresh**".
- Modifiez l'affichage de la piste **GENCODE_V46** pour l'afficher en "**dense**".

Questions

- 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
- Sur quel bras chromosomique (gauche / droite) ? Gauche, droite
- Sur quelle région chromosomique ? p15.4, 15.1, p13, p11.2, q2.1, q23.3
- Quelle est sa longueur en kilobases ? 12, 22, 29, 31 789, 31 817
- Combien de régions répétitives distinguez-vous sur la région du gène PAX6 ? aucune, 15, 21, 22, 23, 25, 30
- Quelle est la densité approximative (nombre de régions répétitives / kilobase) ? 0.1, 0.8, 1, 1.2, 1.3, 2, 5, 10
- Sur la longueur du gène PAX6, les régions répétitives coïncident généralement avec les introns, régions exoniques codantes, régions exoniques non-codantes

