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

Objectifs et notions

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.

Déroulé du TP

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.

Eléments 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	CHESS
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

Source: Salzberg, S. L. Open questions: How many genes do we have? BMC Biol 16, 94 (2018). https://doi.org/10.1186/s12915-018-0564-x

Exemples traités

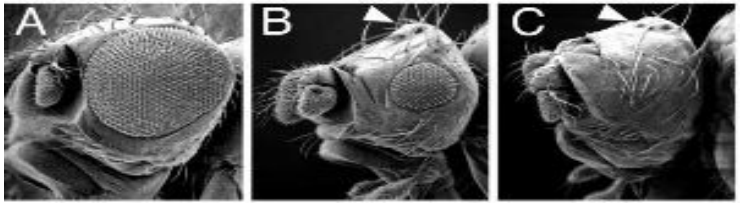
Facteur transcriptionnel – PAX6

- 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 impliquées 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.

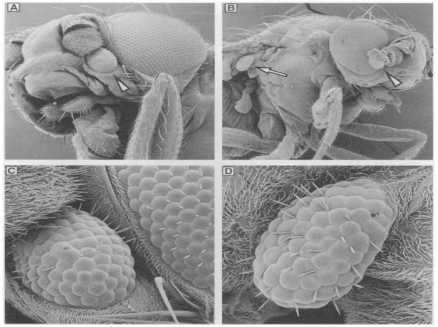
Liaison PAX6 - ADN

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

Phénotype de perte de fonction



Phénotype de gain de fonction

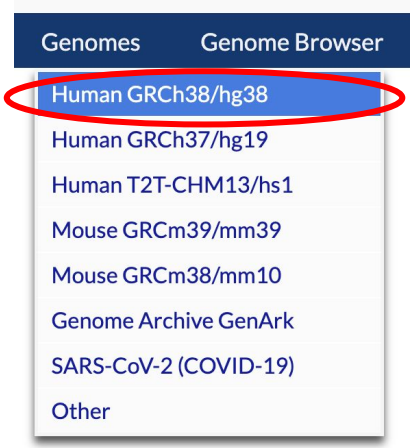


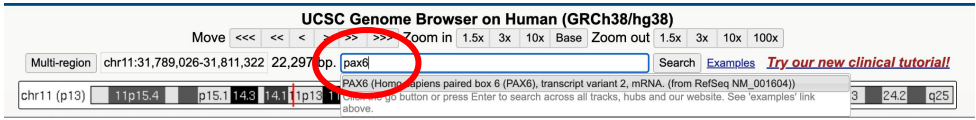
Halder et al. Eyeless initiates the expression of both sine oculis and eyes absent during Drosophila compound eye development. Development (1998) vol. 125 (12) pp. 2181-91.
Halder, Callaerts and Gehring (1995). Science, 267, 1788–1792.

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 génome humain.**
- Entrez le nom du gène d'intérêt (**PAX6**) dans la boîte de recherche et cliquez sur **Search**.



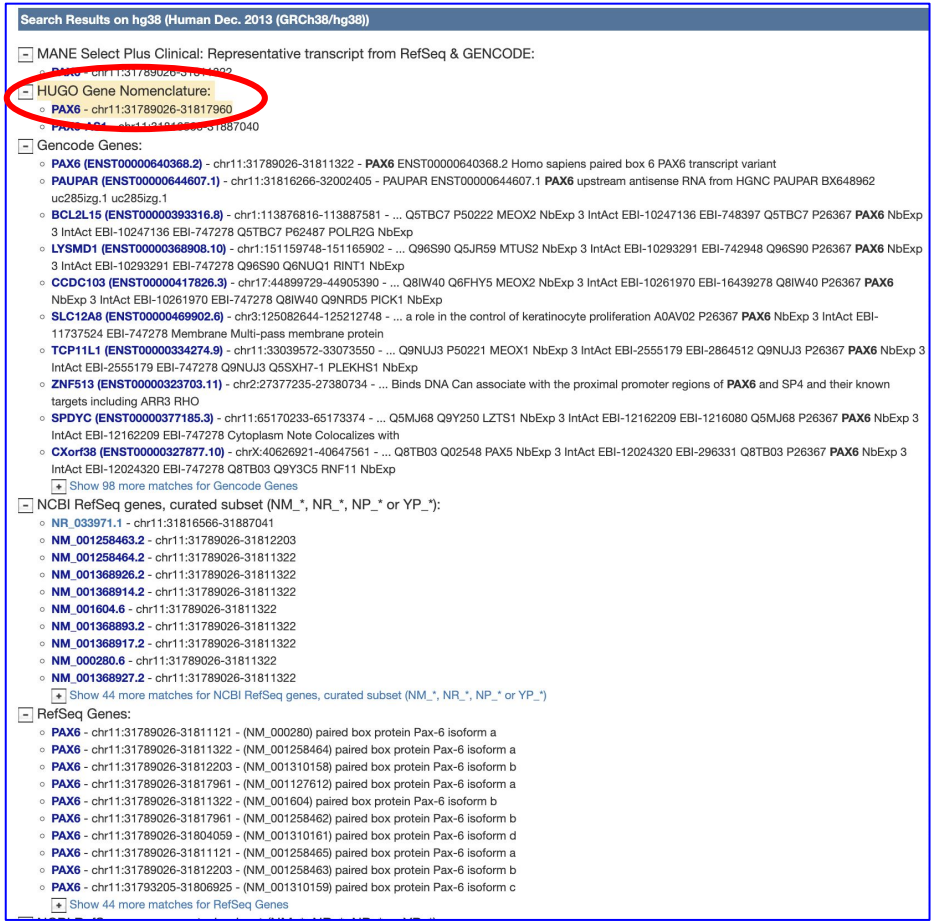


PAX6 sur UCSC genome browser

- 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.

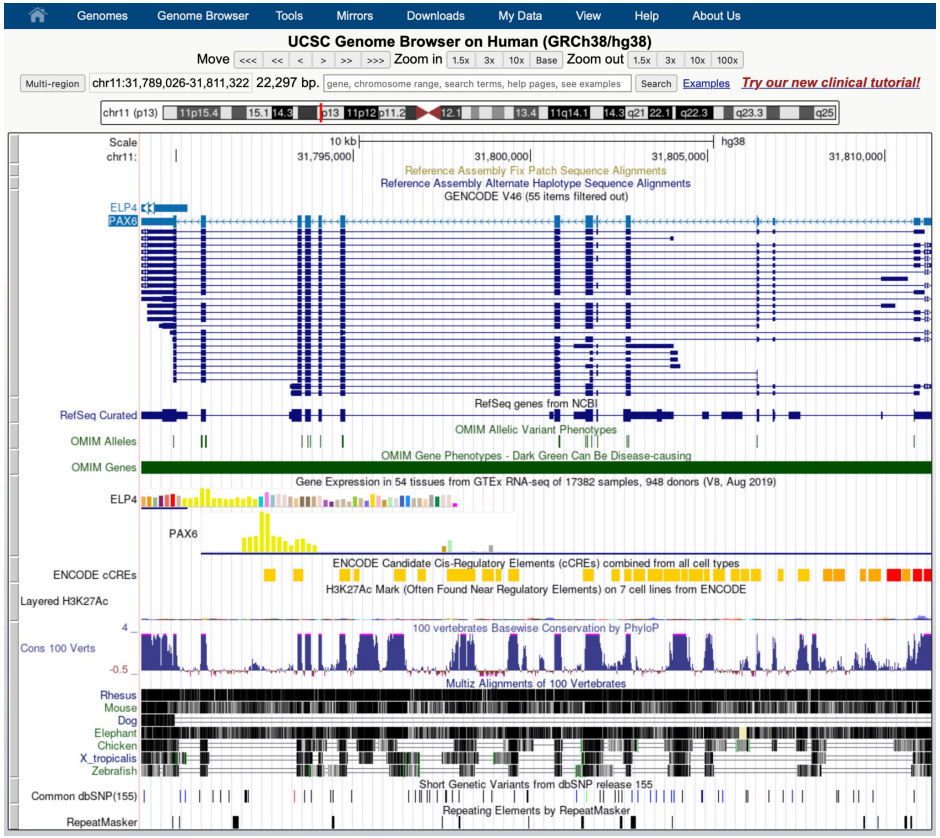
- Sous le titre "HUGO Gene Nomenclature", cliquez sur le lien **PAX6** - chr11:31789026-31817960



Choix de pistes d'annotations du UCSC Genome Browser

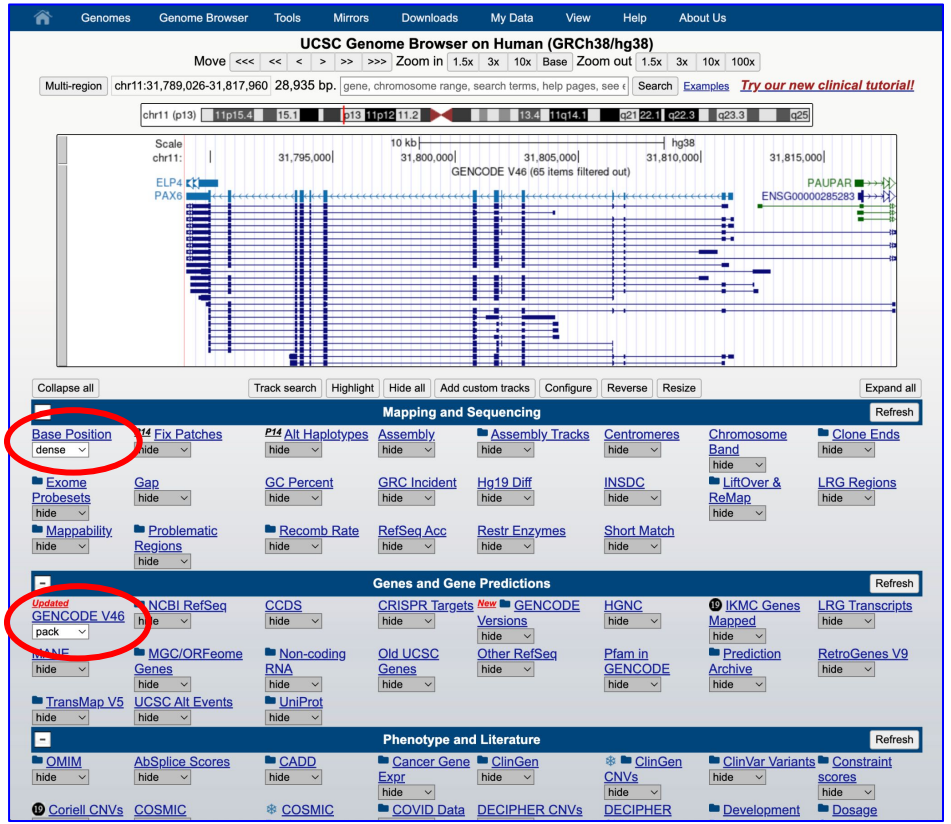
Le navigateur de génomes **UCSC Genome Browser** affiche un vaste choix de pistes d'annotation. La carte génomique en affiche un sous-ensemble, qui s'adaptent en fonction de vos consultations précédentes.

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



PAX6 sur UCSC genome browser

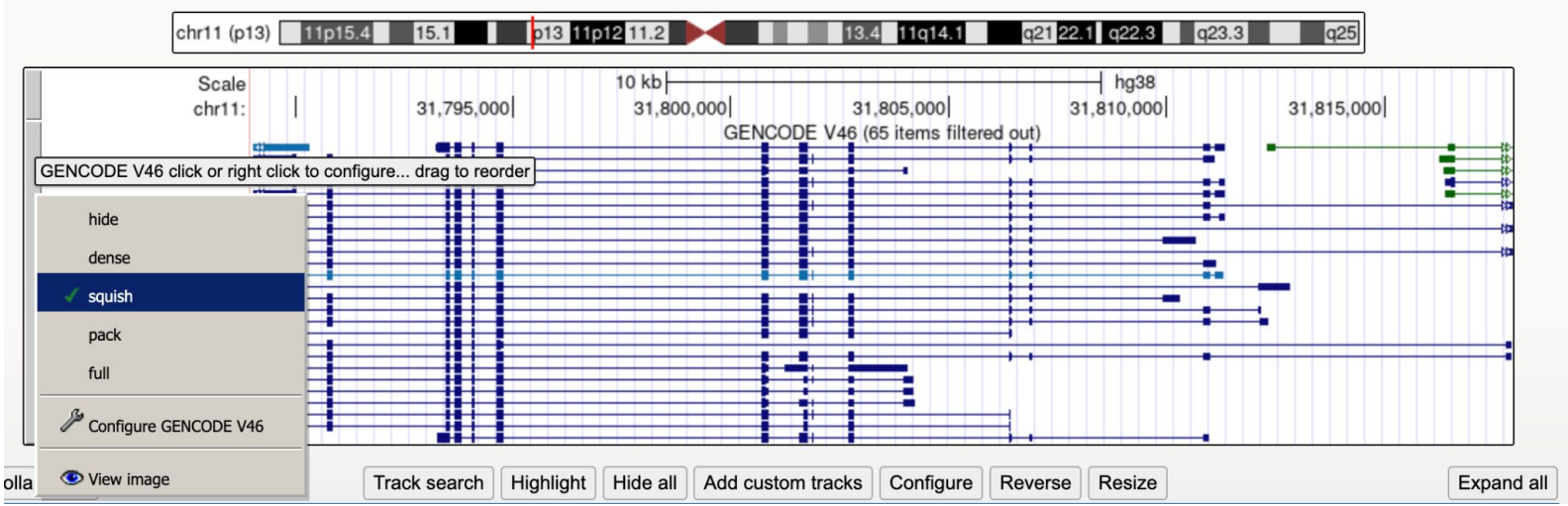
- Connectez-vous au [UCSC Genome Browser]
- Sélectionnez la version **hg38 du génome humain**
- Entrez le nom du gène d'intérêt (**PAX6**) dans la boîte de recherche et cliquez sur **Search**.
- Sous le titre "HUGO Gene Nomenclature", cliquez sur le lien **PAX6** - chr11:31789026-31817960
- Descendez 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 Prediction**", 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.



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