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

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

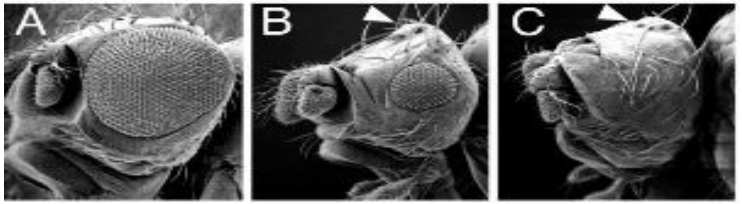
Facteur transcriptionnel – PAX6

- 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.

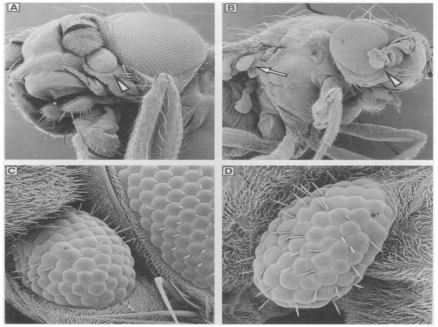
Liaison PAX6 - ADN

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

Phénotype de perte de fonction



Phénotype de gain de fonction

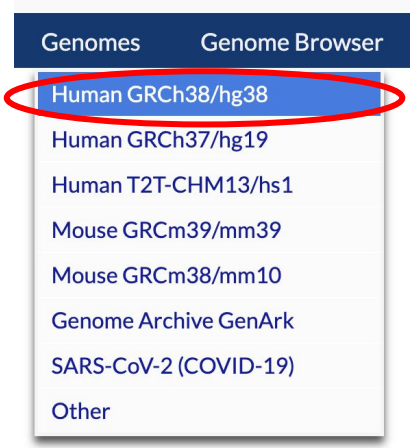


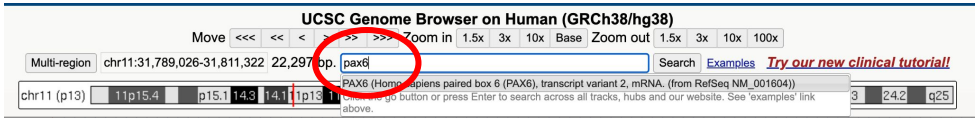
Halder et al. Eyeless initiates the expression of both sine oculis and eyes absent during Drosophila compound eye development. Development (1998) vol. 125 (12) pp. 2181-91.
Halder, Callaerts and Gehring (1995). Science, 267, 1788–1792.

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**.



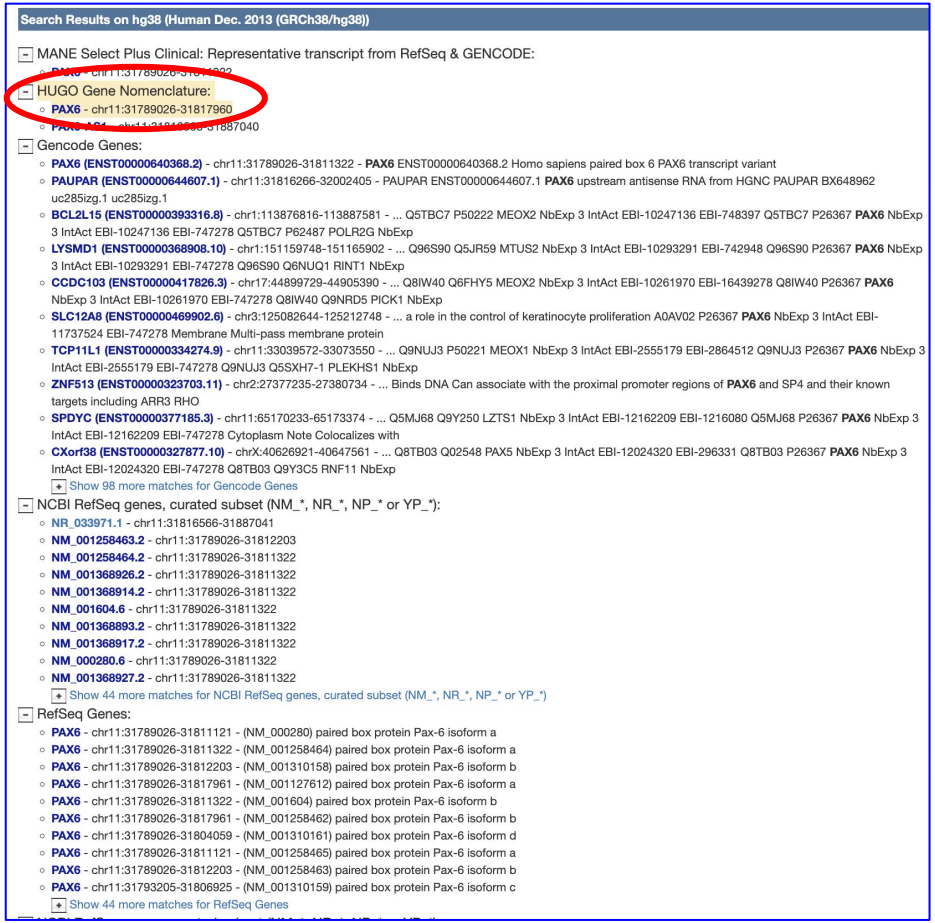


PAX6 sur UCSC genome browser

- 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.

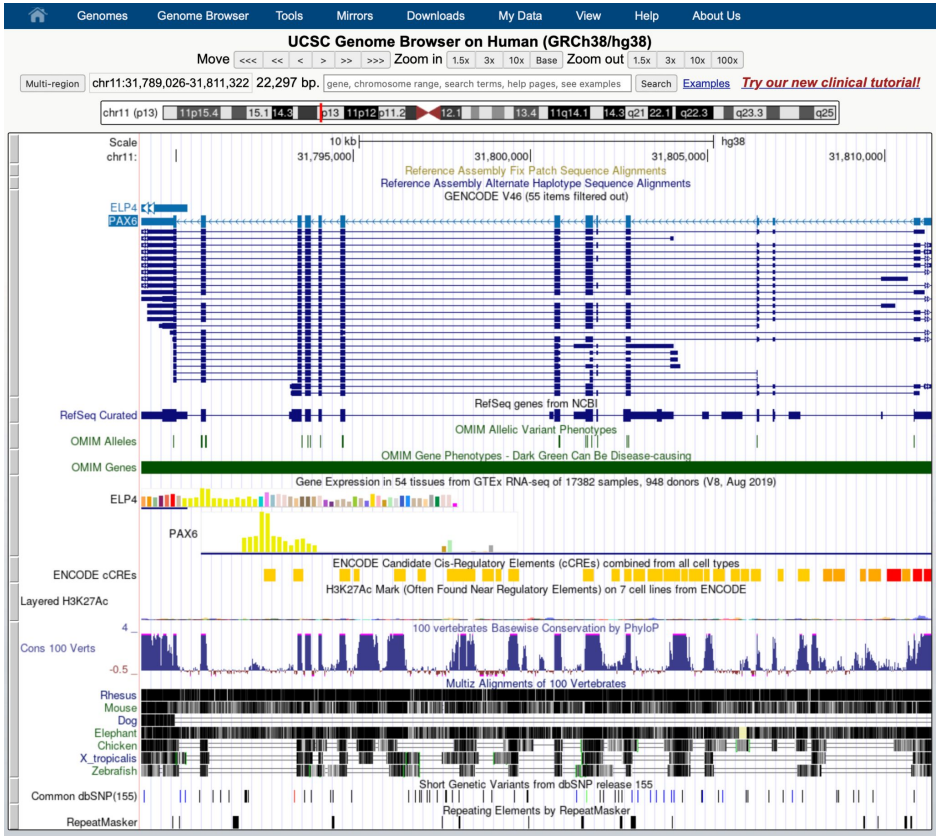
- Sous le titre "HUGO Gene Nomenclature", cliquez sur le lien **PAX6** - chr11:31789026-31817960



Choix de pistes d'annotations du UCSC Genome Browser

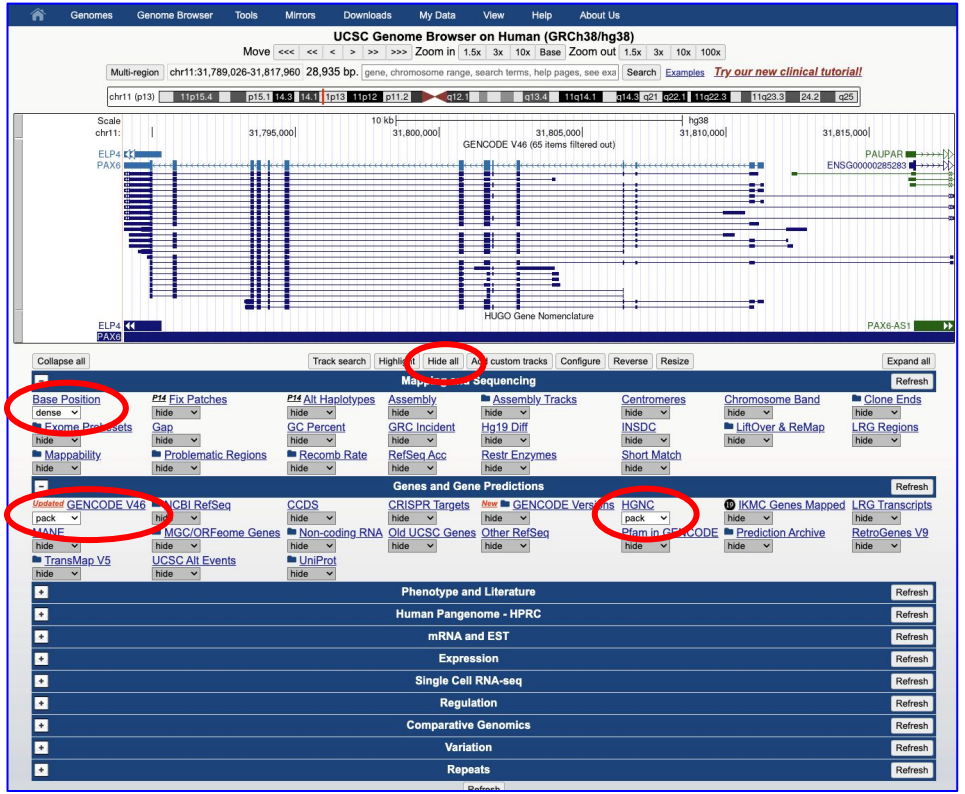
Le navigateur de génomes **UCSC Genome Browser** affiche un vaste choix de pistes d'annotation. La carte génomique en affiche un sous-ensemble, qui s'adaptent en fonction de vos consultations précédentes.

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



PAX6 sur UCSC genome browser

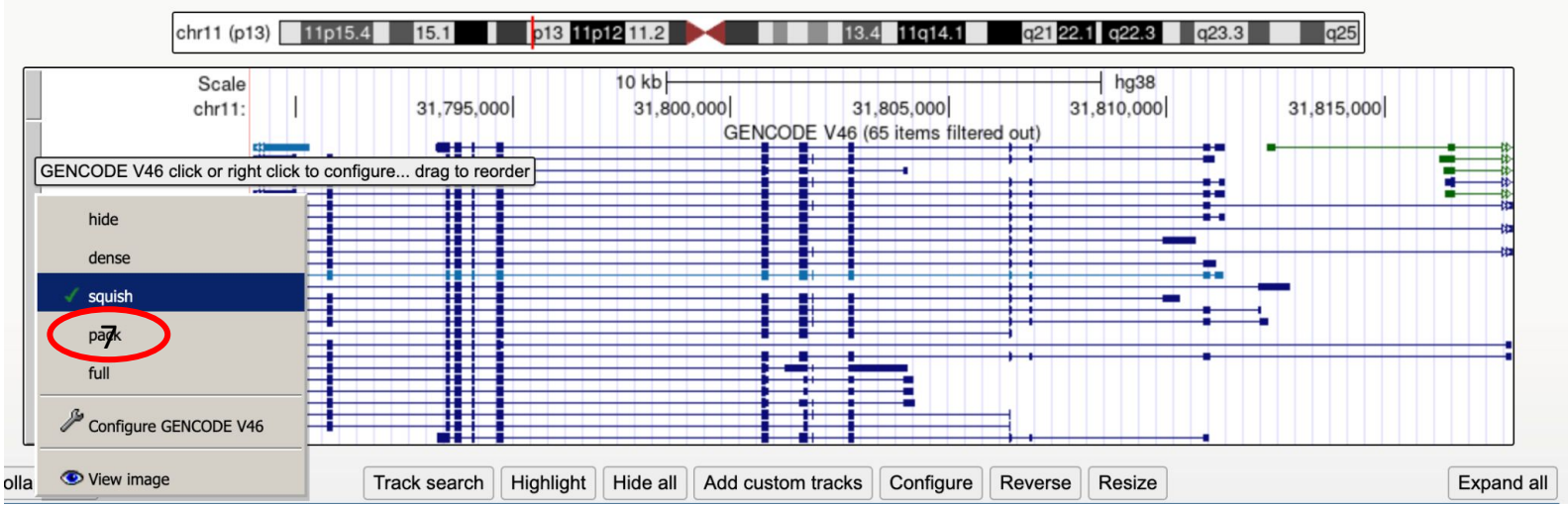
- Connectez-vous au [UCSC Genome Browser]
- Dans le menu **Genomes**, sélectionnez la version **hg38 du génome humain**.
- Entrez le nom du gène d'intérêt (**PAX6**) dans la boîte de recherche et cliquez sur **Search**.
- Sous le titre "HUGO Gene Nomenclature", cliquez sur le lien **PAX6** - chr11:31789026-31817960.
- Descendez sous la carte génomique pour afficher les choix de pistes d'annotations.
- 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.



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.

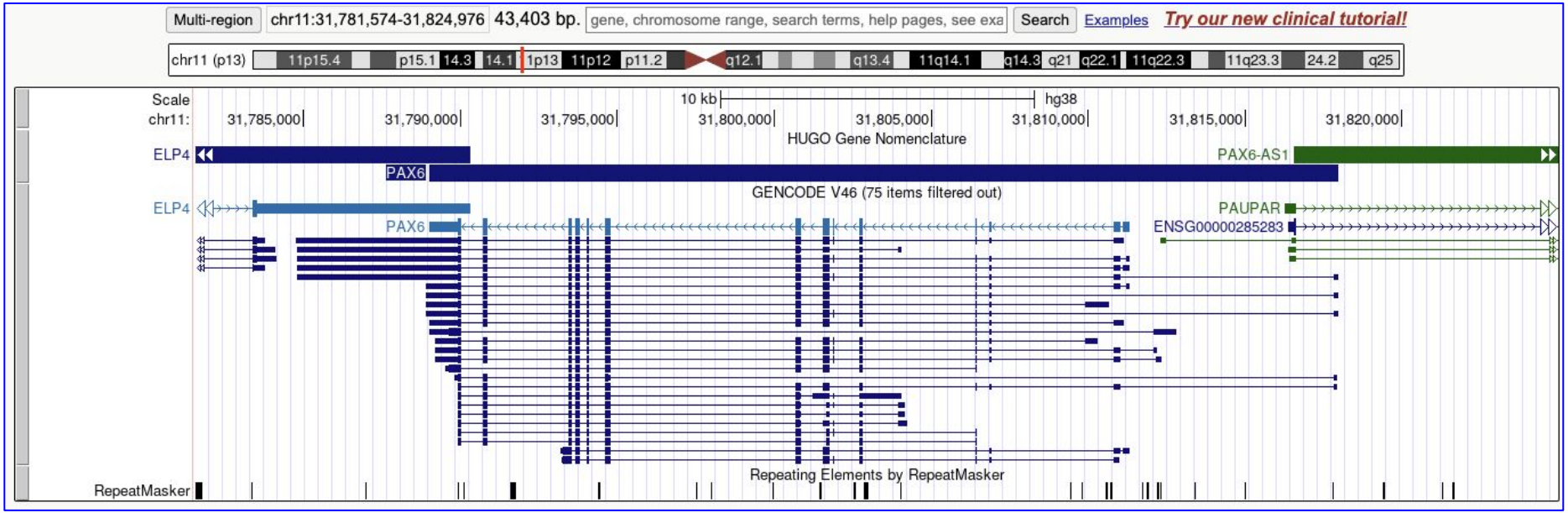
- Testez les différents niveaux de détail avec la piste **GENCODE_V46**, puis sélectionnez le mode **pack**, qui vous permet généralement de visualiser les transcrits alternatifs en occupant une place raisonnable.



Régions répétitives

- Dans la catégorie "**Repeats**", activez l'affichage de "**Repeatmasker**" en format "**dense**" et cliquez "**Refresh**".
- Dézoomez (**Zoom out**) d'un facteur **x1.5** pour voir les environs du gène
- **Déplacez** la piste HGCN 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)

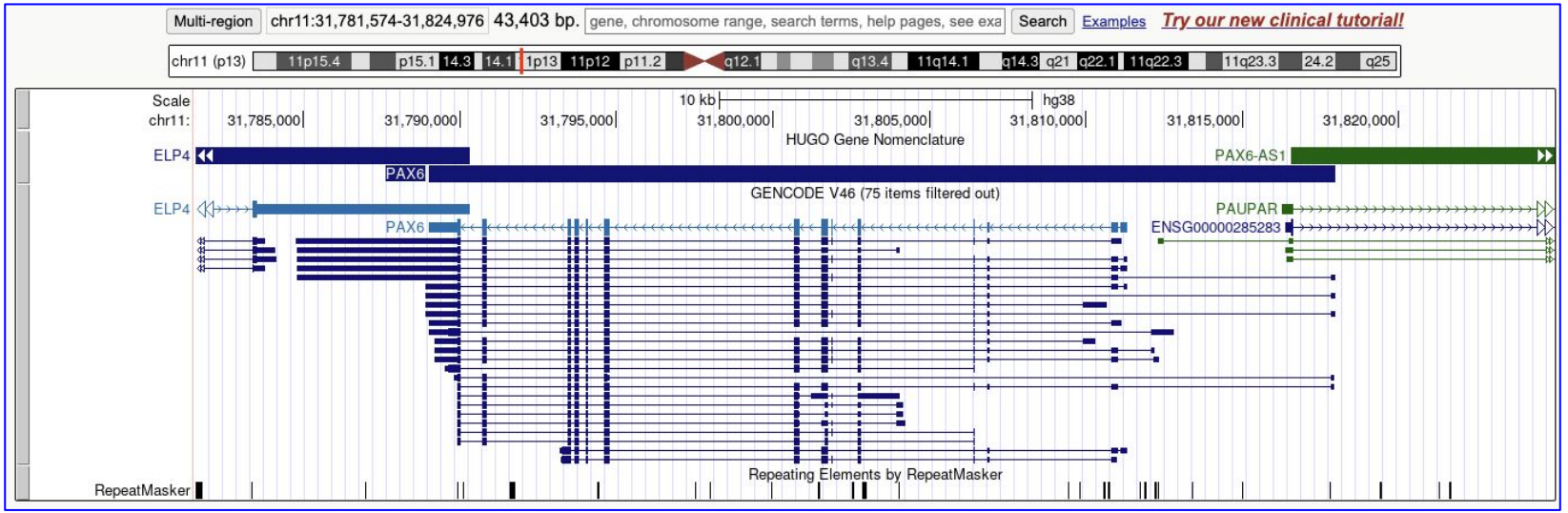
Observez la disposition du gène PAX6. Notez qu'il chevauche ses voisins de gauche (ELP4) et de droite (PAX6-AS1, où AS indique qu'il s'agit d'un gène antisens).



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**.

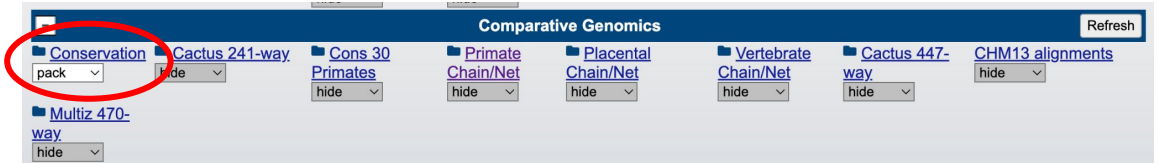
- Sur quel brin chromosomique se situe le gène PAX ? +, -
- Sur quel chromosome est situé le gène PAX6 ? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y
- Sur quel bras chromosomique (gauche / droite) ? Gauche, droite
- Sur quelle région chromosomique ? p15.4, 15.1, 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

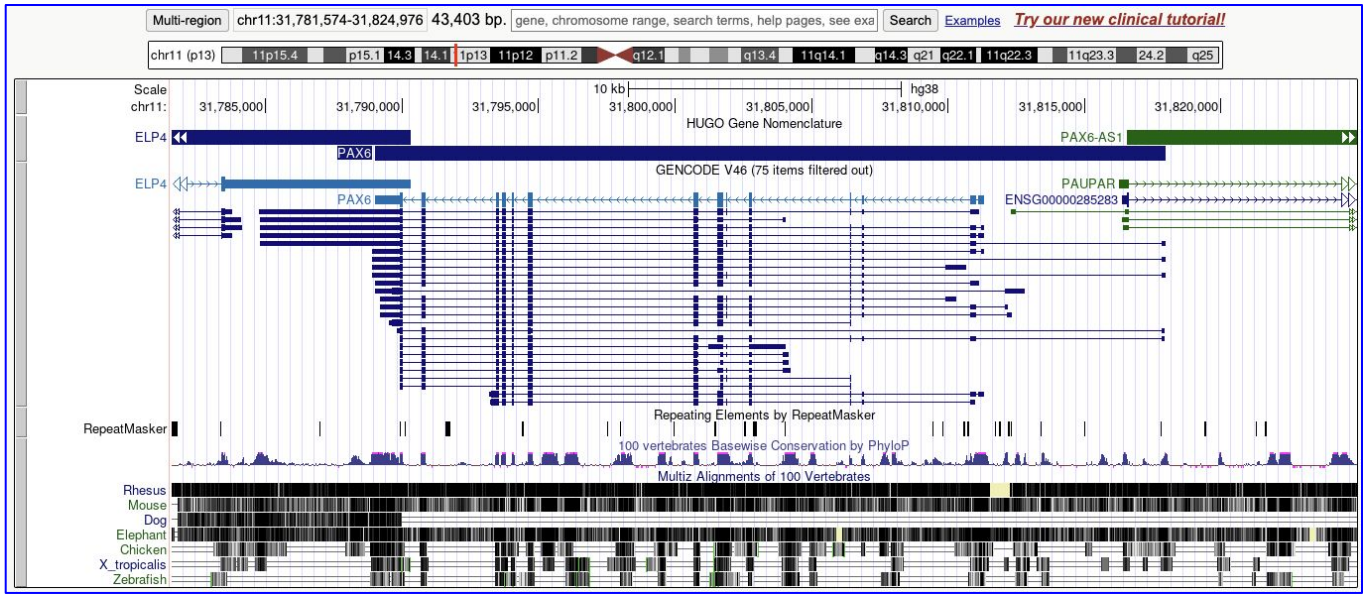


Conservation du gène PAX6 dans les génomes de vertébrés

Génomique comparative : régions conservées chez les vertébrés

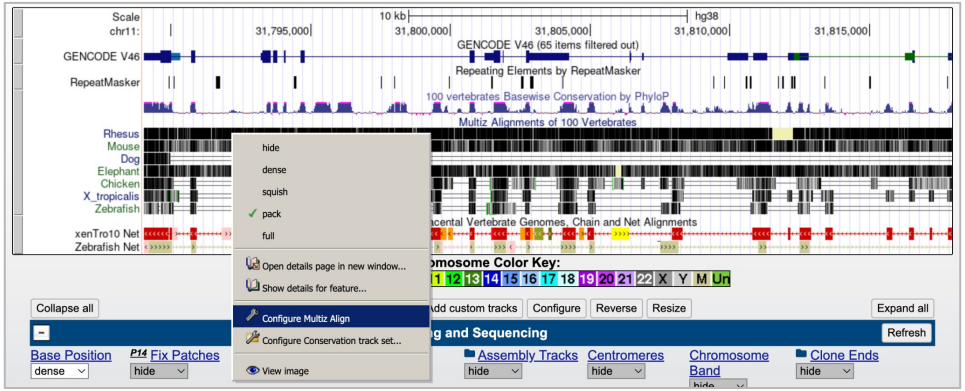
- Dans la catégorie **Comparative genomics**, activez l'affichage **pack** de la piste **Conservation**. Cette piste s'affiche entre les annotations GENCODE_V46 et les régions répétitives
- Faites remonter la piste **RepeatMasker** pour la placer entre les pistes GENCODE_V46 et Conservation

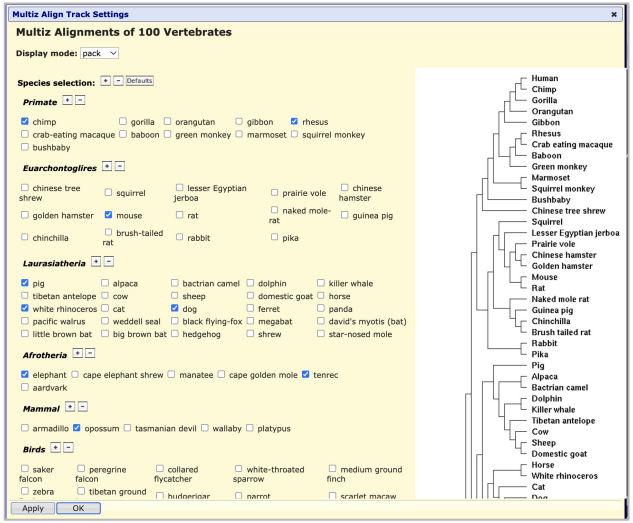




Configuration de la piste d'annotation Conservation

- 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 catéogrie **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.

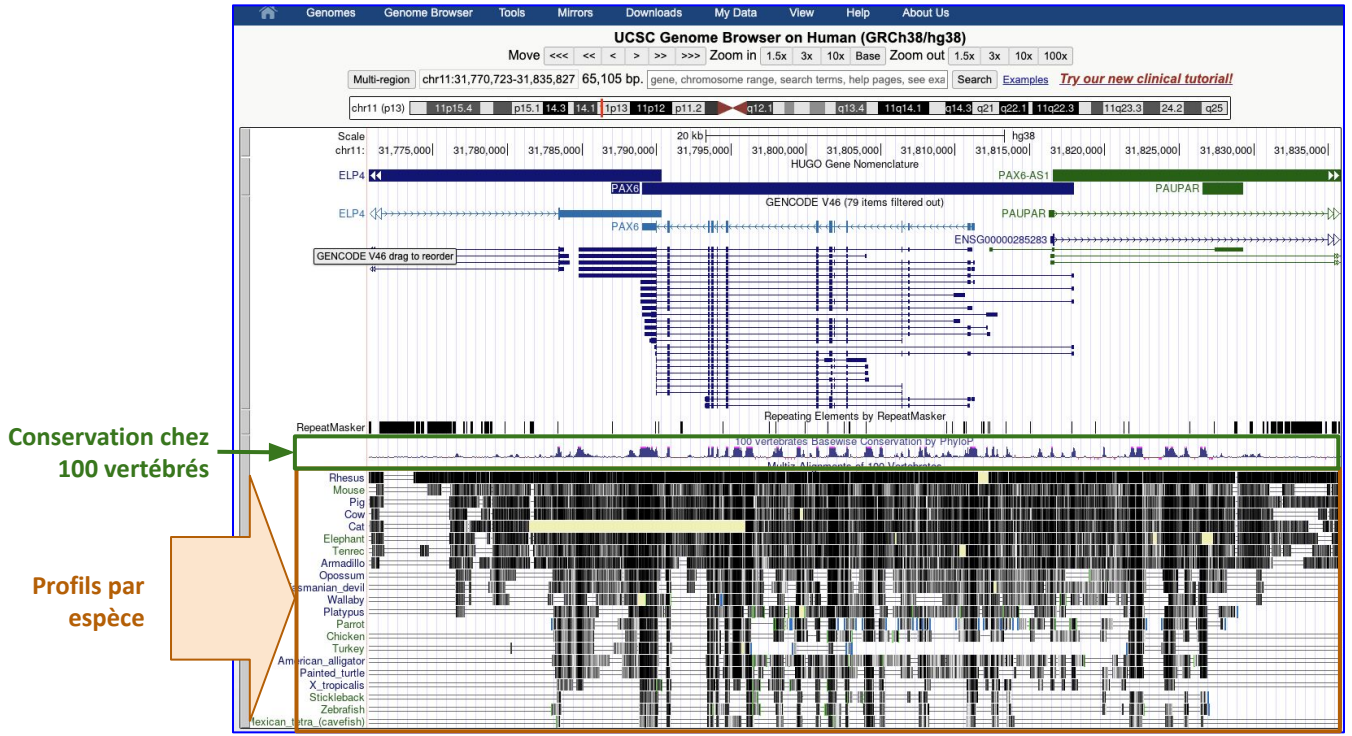




Exercice 2. Conservation de la région génomique PAX6 chez les vertébrés

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).

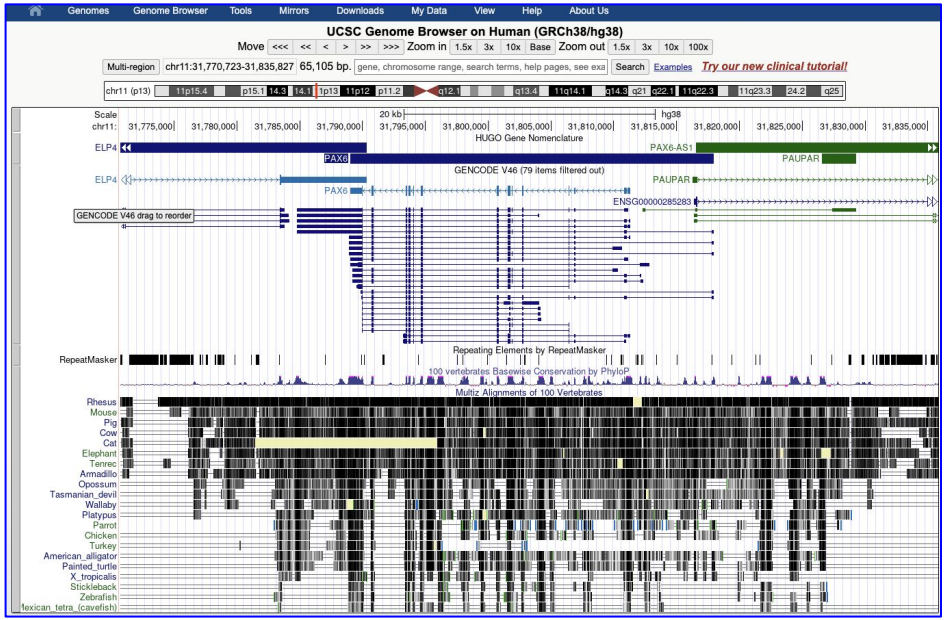


Exercice 2. Conservation de la région génomique PAX6 chez les vertébrés

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 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

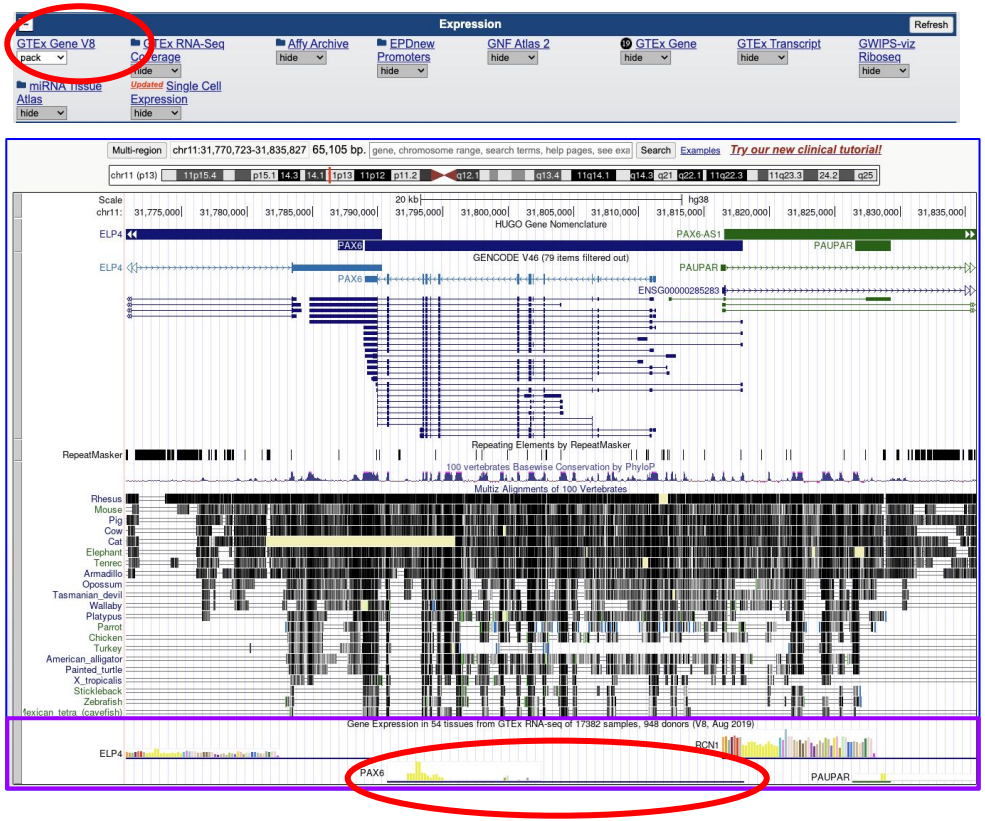


Expression tissulaire du gène PAX6

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.

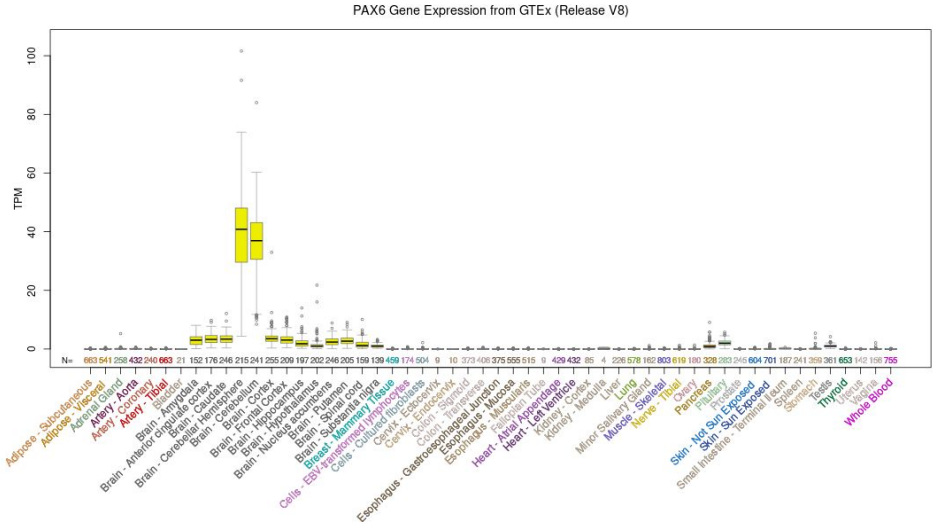


Exercice 3. Profil d’expression tissulaire de PAX6

Interprétation du graphique

Les profils sont affichés sous forme de "boîte à moustaches" (box plot en anglais) pour chaque tissu.

- L’axe horizontal indique les tissus échantillonnés dans GTEx.
- L’axe vertical indique le taux de transcription.
- La barre horizontale épaisse indique le niveau médian (50% des échantillons ont un niveau inférieur, et 50% un niveau supérieur)
- Le bas du rectangle indique le 1er quartile (25% des échantillons ont un niveau inférieur)
- Le haut du rectangle indique le 3ème quartile (75% des échantillons ont un niveau supérieur)
- La barre verticale indique un intervalle de confiance
- Les points isolés sont des "outliers" (valeurs exceptionnellement hautes ou basses, non représentatives de l’ensemble des échantillons).

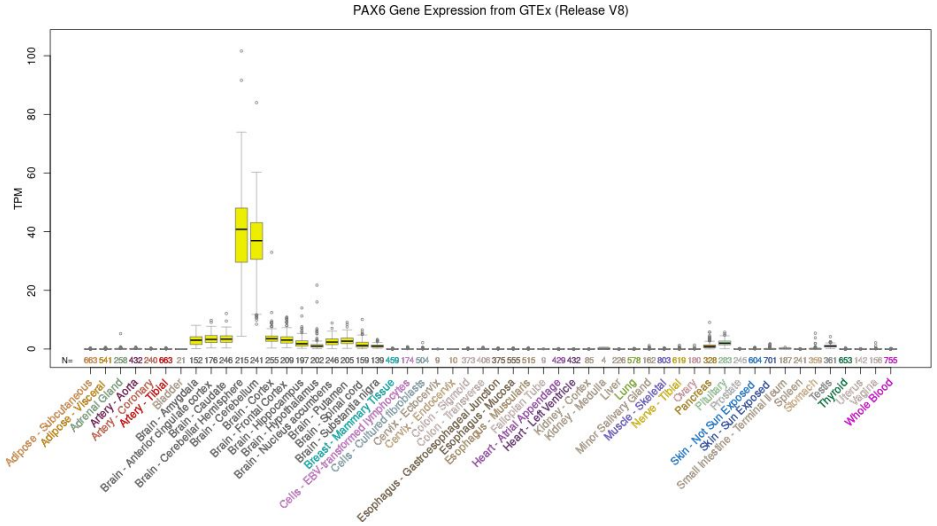


Exercice 3. Profil d'expression tissulaire de PAX6

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

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

Pour aller plus loin (facultatif)

- **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