Introduction

Le séquençage de l'ARN (RNA-seq) est une technique moderne et puissante utilisée pour analyser l'expression de l'ensemble des ARN dans une cellule. Son principal objectif est de quantifier et comparer les niveaux d'ARN dans différents échantillons, permettant ainsi de découvrir des gènes et des transcrits différentiellement exprimés sous diverses conditions expérimentales ou états pathologiques.

Le processus de RNA-seq commence par l'extraction de l'ARN total d'un échantillon biologique, suivi de la conversion des ARN en ADN complémentaire (cDNA). Ce cDNA est ensuite séquencé, générant des millions de courtes séquences d'ADN, appelées "reads". Ces reads sont alignés sur un génome de référence pour identifier et quantifier les transcrits exprimés. L'un des principaux avantages du RNA-seq est sa capacité à détecter des transcrits à faible abondance et à révéler des variantes d'épissage alternatif, offrant ainsi une vision détaillée de l'expression génique. Dans ce rapport, nous appliquons le RNA-seq pour comparer les profils d'expression des ARN entre le cerveau et le rein chez la souris. L'objectif est d'identifier les gènes et transcrits différentiellement exprimés entre ces deux tissus et d'analyser leurs fonctions potentielles.

Matériel & méthodes

Visualisation initiale des comptages

Dans cette étude, un prétraitement des données de comptage d'ARN a été effectué pour améliorer la fiabilité des analyses.

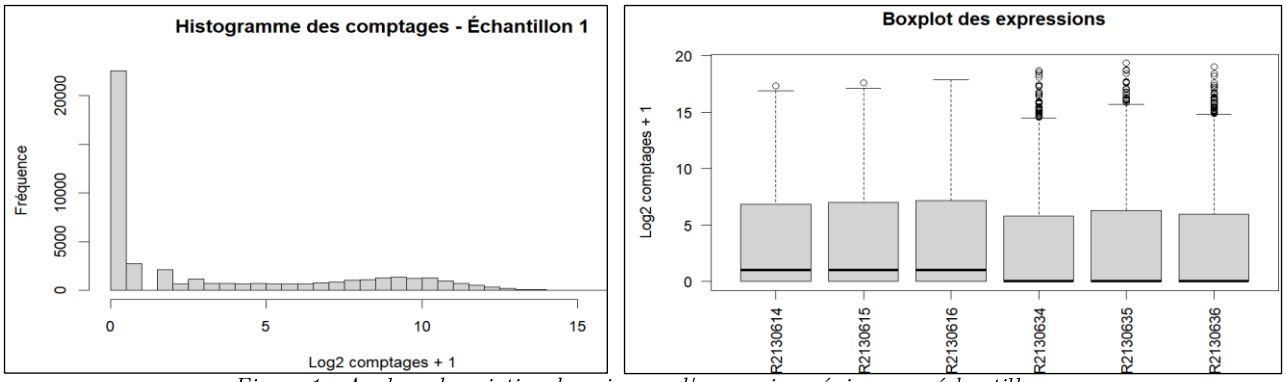

Figure 1 : Analyse descriptive des niveaux d'expression génique par échantillon

Le boxplot des expressions montre que la médiane des valeurs de comptage est basse pour tous les échantillons, reflétant un faible niveau d'expression généralisé parmi les gènes. Les outliers indiquent une variabilité notable, représentant des gènes spécifiquement régulés. L'histogramme de l'échantillon 1 confirme que la plupart des gènes ont de très faibles niveaux d'expression, avec une minorité de gènes exprimés à des niveaux élevés. Ces observations suggèrent une importante hétérogénéité dans l'expression génique. Ces visualisations préliminaires servent à évaluer la qualité des données et à ajuster les processus de normalisation et de filtrage pour améliorer la fiabilité des résultats de l'analyse différentielle. Tous les comptages nuls ont été remplacés par des valeurs manquantes (NA) et ces lignes ont été retirées du jeu de données. De plus, un filtrage des gènes a été réalisé pour retenir seulement ceux avec un niveau d'expression minimal (plus d'un comptage par million dans au moins deux échantillons).

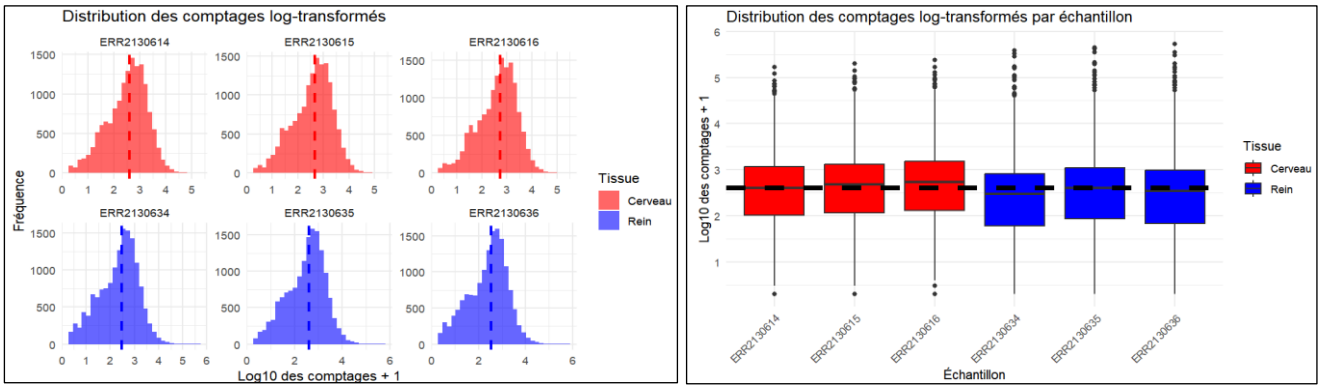

Figure 2 : Analyse de la distribution des expressions géniques dans les échantillons de cerveau et de rein

Les graphiques de la figure 2 montrent la distribution des comptages d'expression génique log-transformés pour les échantillons de cerveau et de rein. Les histogrammes détaillent la fréquence des comptages pour chaque échantillon, illustrant clairement les différences de distribution entre les échantillons de deux tissus différents. On observe que, malgré quelques différences entre réplicats du même tissu, la variabilité intra-échantillon est préservée. Cela indique que même si les médianes ne sont pas parfaitement alignées, ce qui suggère l'absence de normalisation complète, les profils généraux de comptage sont relativement stables au sein de chaque groupe de tissu. Cette analyse préliminaire est essentielle pour préparer les données à des analyses plus complexes comme la multidimensional scaling (MDS), qui utilisera ces données pour examiner la ressemblance entre les échantillons.

Analyse Comparative de l'Expression Génique des Échantillons

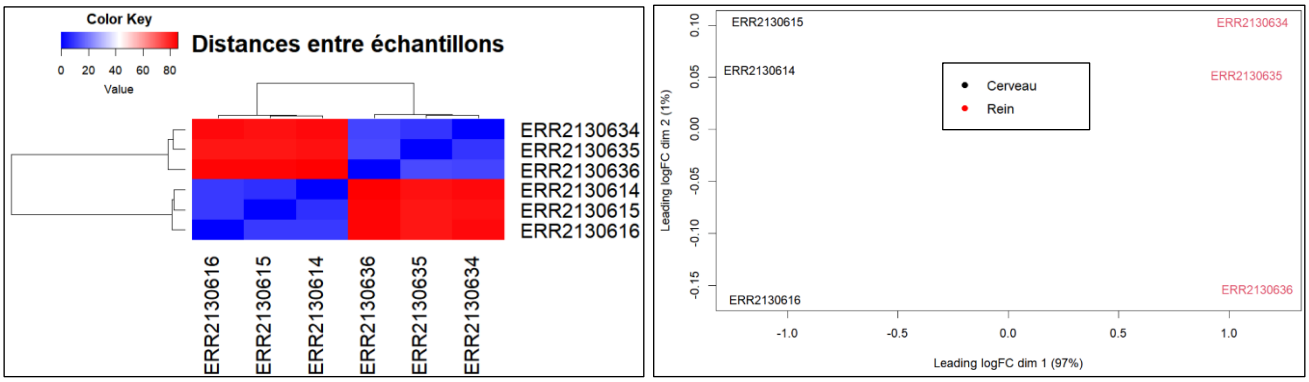

Figure 3 : Analyse des proximités inter et intra-tissulaires par techniques de clustering et MDS

Les figures présentées illustrent la répartition des proximités entre échantillons de RNA-seq basée sur leurs profils d'expression génique. La première figure, une heatmap de proximités, montre clairement que les échantillons issus du même tissu (cerveau ou rein) partagent des proximités plus élevées entre eux, indiquant une similarité plus importante dans leur expression génique comparée à celle entre échantillons de tissus différents. Cette observation est corroborée par le plot MDS, où les échantillons de cerveau et de rein se regroupent en deux clusters distincts sur la première dimension, qui capture la majorité de la variance (97%). Les duplicats sont très similaires entre eux, soulignant que la séparation est principalement influencée par le type de tissu. La seconde dimension, bien que moins informative (1%), aide à distinguer les nuances au sein des groupes de tissus. Si une troisième dimension était visualisée, on pourrait potentiellement mieux observer les proximités, comme suggéré par la heatmap. Ces résultats confirment que la variance génétique est principalement influencée par le type de tissu.

Dans notre étude, nous avons utilisé la méthode TMM (Trimmed Mean of M-values) pour normaliser les comptages de séquence des échantillons de cerveau et de rein, puis réalisé une analyse différentielle pour identifier les gènes exprimés de manière variable entre ces tissus. TMM calcule un facteur de normalisation pour chaque échantillon en comparant les rapports d'expression de tous les gènes à un échantillon de référence, tout en excluant les gènes extrêmes qui pourraient représenter des événements spécifiques ou des erreurs de séquençage. Le tableau des résultats inclut :

- LogFC (Log Fold Change) : mesure le changement d'expression entre les groupes.
- LogCPM (Log Counts Per Million) : reflète l'abondance générale des transcriptions.
- PValue : évalue la probabilité que les différences observées soient dues au hasard.
- FDR (False Discovery Rate) : ajuste les p-valeurs pour limiter les faux positifs dûs à de multiples comparaisons.

Dans les études de RNA-seq, une p-valeur est calculée pour chaque gène afin d'évaluer si les différences d'expression entre les groupes sont statistiquement significatives. Comme de multiples tests sont réalisés simultanément — un pour chaque gène —, cela augmente le risque de détecter faussement une différence significative (faux positifs) par pure chance. Pour contrer ce problème, on applique une correction pour tests multiples, comme l'ajustement du taux de fausse découverte (FDR). Les paramètres choisis pour l'identification des gènes différentiellement exprimés, un FDR inférieur à 1e-3 et une variation de l'expression logarithmique (logFC) supérieure à 3, visent à cibler les cas les plus pertinents et significatifs d'un point de vue biologique. Ce seuil de logFC signifie que les variations d'expression doivent être au moins huit fois supérieures entre les groupes comparés pour être considérées comme significatives.

Le volcan plot présenté ci-dessus offre une visualisation efficace des résultats de l'analyse différentielle des expressions géniques entre deux groupes de tissus. Sur cet axe horizontal, les variations d'expression logarithmique (logFC) sont représentées, indiquant une augmentation d'expression à droite et une diminution à gauche par rapport au groupe de référence (cerveau). Verticalement, l'axe des -log10(FDR) mesure la significativité statistique de ces différences d'expression. Les points en hauteur sur ce graphique signalent une forte improbabilité que les différences d'expression observées soient attribuables au hasard.

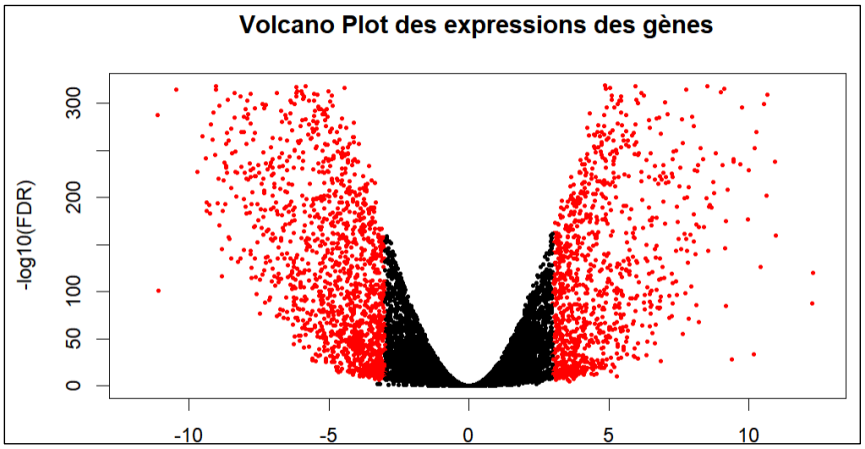

Figure 4 : Volcano plot des expressions des gènes

Les points rouges en position élevée, dispersés loin du centre, identifient les gènes avec des changements significatifs d'expression et une forte validation statistique. Ces gènes, fortement régulés, méritent une attention particulière car ils peuvent être cruciaux pour comprendre les mécanismes biologiques sous-jacents aux différences entre les tissus étudiés.

Nous avons d'abord sélectionné des gènes différentiellement exprimés en se basant sur des critères stricts pour assurer leur pertinence biologique. Ensuite, nous avons utilisé l'analyse d'enrichissement Gene Ontology (GO) pour explorer les fonctions biologiques et les processus associés à ces gènes. Cette analyse nous aide à comprendre l'impact des changements génétiques sur des phénomènes biologiques complexes.

La figure 5 illustre les voies biologiques enrichies parmi les gènes dont l'expression diffère significativement entre les échantillons de cerveau et de rein. Les catégories sont classées par le nombre de gènes impliqués (axe horizontal) et sont colorées selon la signification de leur enrichissement. Les gènes sélectionnés ont des niveaux d'expression variant d'au moins 8 fois entre les deux tissus.

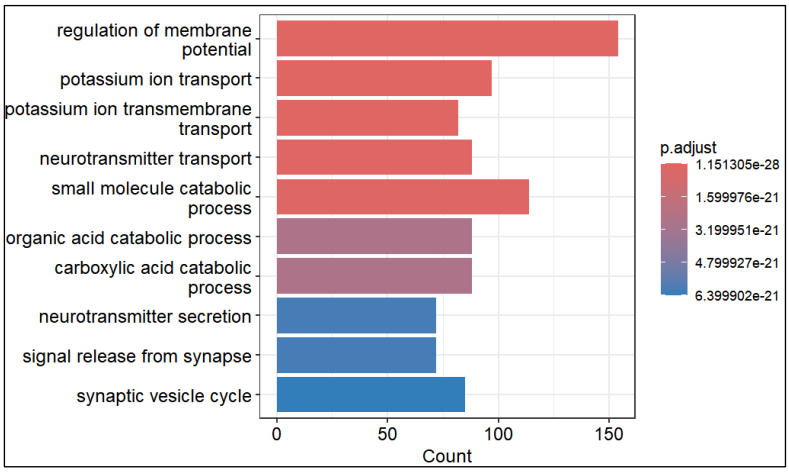

Figure 5 : Enrichissement fonctionnel des gènes différentiellement

L'analyse des gènes différentiellement exprimés montre une prépondérance des processus biologiques liés au métabolisme des petites molécules et à la neurotransmission, en particulier ceux impliquant le transport du potassium et la régulation du potentiel de membrane. Ces activités sont fortement représentées dans les tissus nerveux, indiquant une expression significative de ces gènes dans le cerveau par rapport au rein. Ce profil d'expression suggère que ces gènes sont cruciaux pour les fonctions neuronales spécifiques et qu'ils jouent un rôle moindre ou sont moins exprimés dans les tissus rénaux.

Extraction et intégration des données génomiques

Pour cette analyse, nous avons sélectionné les gènes basés sur leur expression différentielle significative entre les tissus cérébral et rénal. Le tableau suivant présente les cinq premiers gènes identifiés par leur expression différentielle la plus marquée. Pour identifier ces gènes, nous avons utilisé un test différentiel spécifique, prenant en compte le tissu cérébral comme référence et le tissu rénal comme comparaison. Ainsi, un logFC positif indique une sur-expression dans le cerveau, tandis qu'un logFC négatif signale une expression plus élevée dans le rein. Les gènes ont été filtré de sorte que leur logFc soit supérieur à 3 et un FDR inférieur à 0.05.

#	Gene ID	LogFC	LogCPM	PValue	FDR	Gene Name	Description
1	ENSMUSG00000000049	3.2	1.1	2.8e-14	6.4e-14	Apoh	apolipoprotein H
2	ENSMUSG00000000093	4.9	5.4	3.2e-127	3.1e-126	Tbx2	T-box 2
3	ENSMUSG00000000142	-4.0	3.8	1.7e-69	1.0e-68	Axin2	axin 2
4	ENSMUSG00000000159	7.2	5.3	5.7e-243	1.2e-241	Igsf5	immunoglobulin superfamily, member 5
5	ENSMUSG00000000216	6.5	4.6	8.9e-181	1.2e-179	Scnn1g	sodium channel, nonvoltage-gated 1 gamma

Tableau 1

En utilisant les bases de données NCBI et Ensembl, nous avons récupéré des informations détaillées pour chaque gène, tels que Apoh, Igsf5, Scnn1g, Tbx2, et Axin2, pour comprendre leurs fonctions biologiques spécifiques et leurs rôles dans les processus physiologiques ou pathologiques :

- Apoh (apolipoprotein H) : LogFC de 3.2 indiquant une forte sur-expression dans le cerveau. Ce gène joue un rôle important dans la coagulation et l'inflammation, et pourrait être impliqué dans des processus neuroprotecteurs ou des réponses inflammatoires dans le cerveau.
- Tbx2 (T-box 2) : LogFC de 4.9, également plus exprimé dans le cerveau. Tbx2 est impliqué dans le développement embryonnaire et pourrait influencer le développement ou la maintenance neuronale.
- Axin2 : LogFC de -4.0, ce qui suggère une expression prédominante dans le rein. Axin2 régule la voie de signalisation Wnt, essentielle à de nombreux processus de développement et de régénération tissulaire.
- Igsf5 (immunoglobulin superfamily, member 5) : Avec un LogFC de 7.2, ce gène montre une expression extrêmement élevée dans le cerveau. Il pourrait participer à la modulation de l'interaction cellulaire et de la signalisation dans le système nerveux.
- Scnn1g (sodium channel, nonvoltage-gated 1 gamma) : LogFC de 6.5, significativement exprimé dans le cerveau. Ce canal sodique joue un rôle crucial dans le maintien du potentiel de repos et la régulation du volume cellulaire.

Ces résultats de l'analyse de RNA-seq reflètent des adaptations ou des conditions spécifiques aux tissus qui pourraient être cruciales pour comprendre non seulement la physiologie normale mais aussi les pathologies liées au cerveau et au rein.

Conclusion

Dans notre étude RNA-seq sur des échantillons de souris, nous avons examiné les différences d'expression génique entre le cerveau et le rein. Après avoir normalisé les données via la méthode TMM pour équilibrer les variations inter-échantillons, nous avons identifié des gènes significativement régulés à l'aide de tests statistiques rigoureux, permettant de détecter des variations d'expression d'au moins 8 fois. Cinq gènes principaux ont été explorés pour leur expression différentielle : Apoh, Tbx2, Axin2, Igsf5, et Scnn1g, avec un intérêt particulier pour leur rôle biologique potentiel dans les tissus étudiés. Les informations de NCBI ont confirmé et enrichi nos connaissances sur leurs fonctions, telles que l'implication d'Apoh dans la coagulation et de Tbx2 dans le développement embryonnaire.

Pour aller plus loin dans notre étude, nous pourrions étendre notre analyse à des modèles de maladies spécifiques pour mieux comprendre comment les variations d'expression de ces gènes influencent la pathophysiologie. Par exemple, en utilisant des modèles de souris génétiquement modifiées pour simuler des conditions neurodégénératives ou rénales, nous pourrions observer l'impact direct de ces gènes sur la progression de la maladie.