UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODES BIO-INFORMATIQUES POUR LES ANALYSES DE DONNÉES SCRNA-SEQ POUR ÉTUDIER LA MALADIE D'HIRSCHSPRUNG

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

ÉMILIA AÏSHA COLEMAN

SEPTEMBRE 2025

REMERCIEMENTS

Je voudrai premièrement remercier mon directeur de recherche, le professeur Abdoulaye Baniré Diallo, qui a accepté de m'accueillir dans son laboratoire. Il m'a aussi aidée à mieux m'orienter sur mes intérêts de recherche et m'a proposé ce projet.

De plus, je remercie également le professeur Nicolas Pilon, car il m'a offert l'occasion de contribuer dans son laboratoire en me proposant de participer à ses recherches sur l'analyse de données de séquençage de l'ARN de cellule unique.

Je remercie aussi le professeur Rodolphe Soret, avec qui j'ai travaillé et qui m'a aidée sur le plan de l'aspect biologique de ce projet. Il m'a guidée à travers les étapes de recherche et d'analyses de données afin de m'aider à mieux comprendre et interpréter les résultats issus de ce projet.

Je souhaiterais aussi remercier Farzaneh Rahmdani et Golrokh Kiani, qui m'ont conseillée sur les méthodes d'analyses de séquençage. Je voudrais également remercier Thomas Gisier et Alexandre Michaud, qui m'ont beaucoup aidée lors de la révision du texte de mon mémoire.

Enfin, je voudrais remercier ma mère, qui m'a toujours soutenue tout au long de mes études, ainsi que mes tantes et mes cousines, qui m'ont motivée tout au long de mes recherches.

Un grand merci à toutes et à tous !

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

ATAC-seq : *Assay for Transposase-Accessible Chromatin* - Séquençage par la transposase accessible à la chromatine

dNTP : Désoxyribonucléotide triphosphate

DSS : *Dextran Sulfate Sodium* - Dextran sulfate de sodium

FACS : *Fluorescence Activated Cell Sorting* - Tri cellulaire activé par fluorescence

GDNF : *Glial cell line-derived neurotrophic factor* - Facteur neurotrope dérivé des cellules gliales

GESA : *Gene Set Enrichisment Analysis* - Analyse d'enrichissement des ensembles de gènes

HH−GDNF : Holstein−Homozygote traité au GDNF

HH : *Holstein*−Homozygote

HVGs : *Highly variable genes* - Gènes hautement variables

IDE : *Integrated development environment* - Environnement de développement intégré

MACS : *Magnetic Activated Cell Sorting* - Tri magnétique des cellules activées

NGS : *Next Generation Sequencing* - Séquençage de nouvelle génération

PCA : *Principal Conponent Analysis* - Analyse en composantes principales

PCR : *Polymerase Chain reaction* - Réaction en chaîne par polymérase

RNA-seq : *RNA sequencing* - Séquençage de l'ARN

scATAC-seq : *Single Cell Assay for Transposase-Accessible Chromatin* - Séquençage de cellule unique de la transposase accessible à la chromatine

SCP : *Secure Copy Protocol* - Protocol SCP

scRNA-seq : *Single Cell RNA-seq* - Séquençage de l'ARN de cellule unique

snRNA-seq : *Single-nucleus RNA-sequencing* - Séquençage de l'ARN nucléaire de cellule unique

SPLiT-seq : *Split-pool ligation-based transcriptome sequencing* - Séquençage du transcriptome basé sur la ligature par groupements séparés

SSH : *Secure Shell Protocol* - Protocol SSH

t-SNE : *T-distributed Stochastic Neighbor Embedding* - Incorporation de voisins stochastiques distribués en T

UMAP : *Uniform Manifold Approximation and Projection* - Approximation et projection uniforme de variétés

UMI : *Unique Molecular Identifier* - Identifiant moléculaire unique

WT : *Wild Type* - Type sauvage

RÉSUMÉ

Le séquençage de l'ARN de cellule unique a permis de mieux comprendre le transcriptome sur le plan cellulaire. Puisque cette méthode génère beaucoup de données, il y a eu un développement de plusieurs méthodes d'analyses sous forme d'outils. Cependant, ils sont trop spécialisés et ils ne sont pas complètement automatisés. De plus, ils requièrent d'avoir des compétences en programmation afin d'adapter leur code ainsi que leurs paramètres. Enfin, les données produites ne sont pas toujours sauvegardées dans un format qui est compatible entre les outils ainsi que pour l'interprétation des résultats. Afin de résoudre ces problèmes, nous proposons une méthode d'analyse sous une forme de flux de travail. L'objectif de notre méthode est de rendre les analyses plus accessibles pour les biologistes et les bio-informaticiens. Nous avons implémenté plusieurs solutions qui incluent l'automatisation de toutes les étapes d'analyses et de transferts de données entre les outils, de l'utilisation de paramètres, ainsi que la sauvegarde automatique des résultats pertinents. Nous avons testé l'efficacité de notre méthode sur un jeu de données de cellules provenant de tissus intestinaux de souris *Holstein*, un modèle murin utilisé pour la maladie d'Hirschsprung. Notre flux de travail a permis de détecter et d'isoler les différents types de cellules d'intérêt, telles que des cellules gliales et neuronales. Il a aussi décelé des phénomènes biologiques, tels que l'inflammation, la différenciation, la transdifférenciation, ainsi que la prolifération cellulaire. Le principal avantage de notre flux de travail est que nous avons intégré tous les outils nécessaires afin d'effectuer toutes les étapes essentielles d'analyses. Ces dernières peuvent être exécutées successivement de manière automatisée et reproductible. De plus, nous avons aussi implémenté l'option d'exécuter les outils individuellement, ainsi que des paramètres personnalisables.

Mots clés : bio-informatique, flux de travail, maladie d'Hirschsprung, outils, séquençage de l'ARN de cellule unique

INTRODUCTION

0.1 Mise en contexte

Le développement de la méthode de séquençage de l'ARN de cellule unique a permis de mieux comprendre les phénomènes qui sont impliqués dans la transcription chez les cellules et de les classifier en fonction de leurs caractéristiques, telles que leur expression génétique. Cela a aussi mené au développement de nouveaux outils permettant de mieux interpréter les résultats générés. Chacun des outils pour l'analyse de ces types de données utilise des méthodes différentes afin d'interpréter les résultats et, dans certains cas, ils pourraient mener à des conclusions complémentaires ou divergentes.

0.2 Problématique et motivation

L'une des principales problématiques de ces outils est qu'ils sont souvent trop spécialisés dans certaines étapes des analyses de données issues du séquençage de l'ARN de cellule unique. De plus, la plupart d'entre eux ne sont pas adaptés pour faire des analyses de manière intuitive pour les utilisateurs qui ne travailleraient pas dans le domaine de la bio-informatique. Ces derniers doivent souvent investir du temps pour apprendre la programmation et demander de l'aide à un bio-informaticien pour traiter leurs résultats, ce qui peut prendre plus de temps avant de pouvoir les interpréter. De plus, le code de ces outils n'est souvent pas adapté afin qu'ils soient automatisés et reproductibles.

Cela a mené à la création de ce projet de recherche afin d'implémenter un flux de travail qui intègre plusieurs outils permettant de suivre toutes les étapes essentielles d'analyses de données issues du séquençage de l'ARN de cellule unique de manière plus simplifiée, automatique et efficace tout en étant reproductible.

0.3 Présentation du projet de recherche

Notre flux de travail a été développé avec le flux de travail *Script of Script* (Wang et Peng, 2019) dans un bloc-notes Jupyter (Kluyver *et al.*, 2016). Cet environnement a pour particularité d'être polyglotte et il peut inclure plusieurs langages de programmation qui seront nécessaires pour utiliser le flux de travail tels que Python, R et Bash (Peng *et al.*, 2018) (Wang et Peng, 2019). Il comprend l'utilisation d'outils populaires, tels que Cell Ranger (Zheng *et al.*, 2017), Seurat (Hao *et al.*, 2021) et Monocle 3 (Cao *et al.*, 2019). Il inclut aussi les versions R des outils d'annotations des types de cellules tels que ScType (Ianevski *et al.*, 2022), EasyCellType (Li *et al.*, 2023), Enrichr (Jawaid, 2023a) et scMayoMap (Yang *et al.*, 2023b).

Ce mémoire est divisé en 6 chapitres :

⁻ Le premier chapitre est une mise en contexte du jeu de données ayant été utilisé pour notre flux de travail. Cela inclut le développement du système nerveux entérique impliqué dans le système intestinal ainsi que de la maladie d'Hirschsprung. Il inclut aussi les termes qui sont en lien avec le jeu de données que nous avons utilisé pour développer et valider notre flux de travail ainsi que l'importance de la méthode de séquençage de l'ARN de cellule unique afin d'étudier la maladie d'Hirschsprung.

⁻ Le deuxième chapitre est une revue de la littérature au sujet des méthodes de séquençage traditionnelles, des méthodes de séquençage de nouvelle génération, du séquençage de l'ARN de cellule unique ainsi que les enjeux qui sont associés aux données issues de cette technique de séquençage.

⁻ Le troisième chapitre traite des méthodes d'analyses de données issues de ce type de séquençage, notamment une revue des outils les plus utilisés ainsi qu'une

description des outils qui ont été implémentés dans le flux de travail.

– Le quatrième chapitre comporte une description plus détaillée du flux de travail, de sa structure ainsi que des outils qui ont été implémentés. Il inclut aussi la méthodologie qui a été utilisée lors de son développement, les justifications pour son utilisation, les solutions implémentées pour résoudre les problèmes ainsi que les paramètres à utiliser afin d'exécuter le flux de travail.

– Le cinquième chapitre présente l'utilisation de ce flux de travail afin d'analyser notre jeu de données ainsi que les résultats qui ont été générés avec celui-ci.

– Enfin, le sixième chapitre contient la discussion des résultats générés par les outils, une conclusion, une critique de la méthode utilisée ainsi qu'une comparaison entre les outils sur le plan de leurs performances.

En résumé, ce mémoire sera principalement une révision et une critique des méthodes biologiques et informatiques menant aux analyses de données de séquençage de l'ARN de cellule unique, tout en proposant une solution sous forme de flux de travail.

CHAPITRE I

MISE EN CONTEXTE POUR LE JEU DE DONNÉES UTILISÉ

Puisque notre jeu de données provient d'échantillons biologiques, il est important d'expliquer dans ce chapitre leur contexte et leur provenance. L'application de notre flux de travail a pour but de détecter et d'extraire les cellules gliales et neuronales, de les identifier à l'aide des outils d'annotation et de déterminer s'il y a eu une transition entre les cellules gliales et neuronales par le processus biologique de la différenciation, de la transdifférenciation et de la prolifération.

1.1 Le système intestinal et entérique

Le système intestinal est principalement régulé par le système nerveux entérique. Ce dernier est responsable de nombreuses fonctions essentielles, telles que la motilité intestinale et la régulation sur le plan hormonal. Cette dernière inclut la sécrétion de mucus, des enzymes responsables de la digestion des aliments (Mueller et Goldstein, 2022) ainsi que l'absorption des nutriments via les muqueuses (Mowat et Viney, 1997). Les muqueuses intestinales constituent la plus grande surface du corps humain et celles qui sont présentes dans le tube digestif représentent à elles seules environ une surface de 400 m^2. Si on le compare à la surface de la peau chez l'être humain, cela équivaut à 200 fois celle-ci (Mowat et Viney, 1997).

1.1.1 Structure du système intestinal chez le fœtus

Le système digestif chez les fœtus humains est divisé en trois segments. La première partie inclut l'œsophage, l'estomac, une partie du duodénum ainsi que la vésicule biliaire. Ensuite, il y a la partie médiale qui comprend le petit intestin et la partie supérieure du gros intestin. Enfin, il y a la partie distale qui inclut le reste du gros intestin, ainsi que le rectum. Ces segments sont variables au niveau de de la quantité de vaisseaux sanguins (Diposarosa *et al.*, 2021).

1.1.2 Développement du système nerveux entérique

Afin d'effectuer le transit intestinal, il est nécessaire qu'il y ait aussi la présence de cellules gliales et neuronales (Mueller et Goldstein, 2022). Selon Furness et al., (2014), le système nerveux entérique abrite un réseau d'environ 200 à 600 millions de neurones et, selon Grubišić et Gulbransen (2017), il contient aussi environ 7 fois plus de cellules gliales. Ces neurones et ces cellules gliales sont responsables de la motilité intestinale, de la sécrétion des hormones, du réseau sanguin qui tapisse les tissus tels que les muscles lisses, ainsi que les contractions musculaires effectuées par ces dernières (Diposarosa *et al.*, 2021).

Selon Mueller et Goldstein (2022), tout débute dans la crête neurale lors du développement du système nerveux entérique dans les premiers stages de développement chez l'embryon. Elle est responsable du développement du cerveau et de la moelle épinière. Durant le développement du système digestif, les cellules dérivées de la crête neurale migrent tout le long des tissus du tube digestif jusqu'au rectum. Lorsqu'elles atteignent leur destination, ces cellules colonisent les parois et commencent à proliférer et à se différencier, notamment en cellules gliales et en neurones. S'il y a des cellules qui n'ont pas été différenciées avant leur entrée dans le début du système digestif ou qu'elles restent non différenciées après

leur migration, elles commencent à exprimer les marqueurs SOX10 et PHOX2B MASH1 (Ascl1), qui sont tous les trois des facteurs de transcription. De plus, ces cellules vont aussi exprimer le récepteur de l'endothéline de type B (EDNRB), le récepteur à faible affinité pour le facteur de croissance des nerfs (P75) et le récepteur transmembranaire de la tyrosine kinase (RET) (Goldstein *et al.*, 2013). Le facteur neurotrope dérivé de la glie (GDNF) ainsi que son corécepteur alpha de la famille du GDNF (GDNF alpha-1) sont responsables de l'activation du récepteur de RET (Mueller et Goldstein, 2022). Lorsque les cellules progénitrices se différencient en neurones, elles deviennent positives aux marqueurs spécifiques aux neurones (RET, PHOX2B, TUBB3, PGP9, NF et TH) (Diposarosa *et al.*, 2021). De plus, le récepteur transmembranaire de la tyrosine kinase (RET) est pro-oncogène quand le facteur neurotrope dérivé de la glie (GDNF) se lie à ses récepteurs. Cela envoie un signal inhibiteur de l'apoptose, ce qui favorise la survie et la prolifération des cellules. Dans le cas contraire, cela cause une inhibition et provoque l'apoptose des cellules (Pan et Li, 2012).

Chez la souris, les cellules dérivées de la crête neurale atteignent la partie proximale entre les jours 9,5 et 11,5 (Durbec *et al.*, 1996) et atteignent la partie distale vers la semaine 14,5 (Anderson *et al.*, 2006). Chez l'être humain, les cellules commencent à migrer tout le long de l'iléon à partir de la 7e semaine et, vers la 8e semaine, elles atteignent la partie médiale du côlon. Enfin, vers la 12e semaine de la gestation, elles atteignent la partie distale du côlon (Wallace et Burns, 2005).

Selon Mueller et Goldstein (2022), la migration des cellules dérivée de la crête neurale représente une étape clé lors du développement du système intestinal. S'il y a une mauvaise migration, il risque d'avoir une répartition incomplète de cellules nerveuses dans le système digestif. Il est donc important que, lors de la migration de ces cellules, ces dernières doivent arriver au bon endroit ainsi qu'au bon moment. Dans le cas contraire, par exemple, si elles arrivent à leur destination

au mauvais moment, il se peut que l'environnement soit entre-temps modifié et qu'il ne soit plus favorable à leur survie. De plus, s'il y a des défauts au niveau de la prolifération des cellules, il peut y avoir des régions qui sont appauvries en cellules neuronales. Enfin, s'il y a des lacunes en matière de différenciation cellulaire, par exemple, si elles ont été différenciées trop tôt ou trop tard, elles risquent de ralentir, voire arrêter leur migration dans les régions du côlon en raison d'une différenciation précoce. Ces anomalies dans ces processus peuvent causer un développement incomplet ou anormal des neurones dans le côlon et développer des maladies inflammatoires, telles que la maladie d'Hirschsprung (Mueller et Goldstein, 2022).

1.2 Maladie d'Hirschsprung

La maladie d'Hirschspurng est caractérisée par une défaillance au niveau des cellules neuronales dans le système intestinal. Cette défaillance peut être soit due à une mauvaise répartition de cellules neuronales dans l'intestin ou à une aganglionose, c'est-à-dire une absence de cellules nerveuses, en particulier dans la partie distale du système intestinal (Mueller et Goldstein, 2022).

Dans environ 80% des cas, cette aganglionose est présente dans la partie rectosigmoïde du côlon. De plus, elle a une incidence d'un nouveau-né sur 5000 (Amiel, 2001). Dans environ 10 à 15% des cas, les patients sont diagnostiqués après leur première année suivant leur naissance et très peu de cas sont enregistrés chez les adultes (Kapur, 1999). Cette absence de cellules gliales et neuronales cause une absence de motilité intestinale, ce qui empêche l'élimination des selles. Cela va provoquer une accumulation de ces dernières dans l'intestin, ce qui va causer une occlusion intestinale (Diposarosa *et al.*, 2021).

Selon Butler Tjaden et Trainor (2013), cette maladie était souvent diagnostiquée

dans les premières 24 heures chez les nouveau-nés. De plus, les nouveau-nés avaient pour symptômes un gonflement au niveau de l'abdomen, des vomissements ainsi qu'un délai tardif pour l'élimination des selles. Chez les enfants, les symptômes étaient des vomissements ainsi que la constipation. Dans environ 10% des cas, il y avait des infections, y compris la fièvre ainsi que des inflammations des muqueuses, telles que l'entérocolite et de septicémie (Butler Tjaden et Trainor, 2013).

1.2.1 Principaux gènes impliqués dans la maladie d'Hirschprung

Dans des études de la littérature réalisées par Alves et al. (2013) au sujet de la pathogénicité de la maladie d'Hirschprung, ils ont mentionné qu'il y avait environ 15 gènes qui étaient impliqués dans cette maladie, y compris les gènes RET, GDNF, son corécepteur Gfra1, EDNRB, PHOX2B et SOX10. Ces gènes étaient tous impliqués dans le développement du système nerveux entérique régulant le système intestinal (Mueller et Goldstein, 2022).

1.2.2 Diagnostic

Afin de diagnostiquer cette maladie, il faut effectuer une biopsie d'aspiration rectale avec une coloration de l'hématoxyline, l'éosine et la calrétinine (Friedmacher et Puri, 2015). Elle est principalement utilisée au cours des 48 premières heures postnatales, lorsque les patients n'ont pas encore éliminé leurs selles. Par contre, si les patients ont une constipation ou une inflammation causée par une entérocolite suite à leur naissance, ce procédé est beaucoup plus difficile, car il faut effectuer une biopsie rectale sous anesthésie générale (Mueller et Goldstein, 2022).

1.2.3 Traitements

Le principal traitement de cette maladie est l'ablation de la partie du côlon qui présente une aganglionose totale (Urla *et al.*, 2018). Cependant, ce type d'intervention peut causer d'autres complications, telles que l'incontinence, la constipation, des infections et de l'entérocolite (Niramis *et al.*, 2008). D'autres études mentionnent d'autres complications, telles que des abcès, des dommages au côlon qui causent une incontinence permanente, ainsi qu'une fermeture complète du côlon due à l'atrésie (Bischoff *et al.*, 2011). D'autres traitements potentiels alternatifs impliquent la transplantation de cellules souches dans les parties du côlon ayant de l'aganglionose et de les traiter afin qu'elles prolifèrent et se différencient en cellules neuronales et gliales (Mueller et Goldstein, 2022).

1.2.4 Utilisation du modèle murin *Holstein* pour la maladie d'Hirschsprung

Bondurand et Southard-Smith (2016) ont fait un résumé de la littérature et mentionné les différents types de modèles de souris qui ont été utilisés afin de mieux comprendre la maladie. Les modèles utilisés avaient tous des modifications au niveau des gènes ayant été mentionnés par Alves et al. (2013).

Soret et al. (2015) ont créé un nouveau modèle de souris appelé *Holstein*. Cette dernière était utilisée comme modèle pour la maladie d'Hirschprung associée à la trisomie 21. Ils ont créé ce modèle de souris en effectuant des insertions transgéniques à différentes régions du gène Col6a4. Cela a causé une surexpression de ce dernier ainsi qu'une augmentation de la production du collagène VI dans les cellules. Cela a induit un ralentissement de la migration des cellules neurales entériques dérivées de la crête neurale dans le côlon et réduit leur colonisation graduellement dans ces régions. Enfin, cela a causé une aganglionose partielle qui est graduellement devenue totale dans la partie distale du côlon, ce qui a repro-

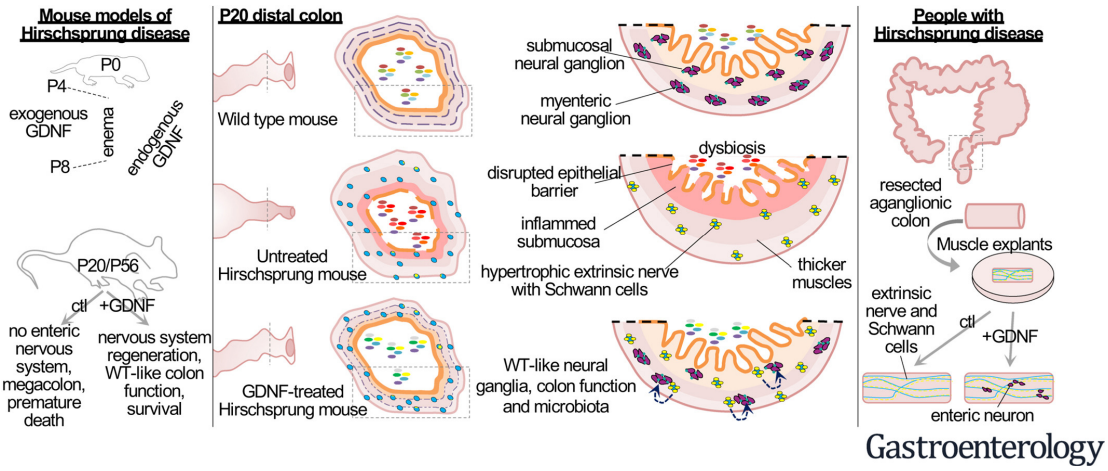

FIGURE 1.1 – Souris *Holstein* ($Hol^{Tg/Tg}$) et traitements avec le GDNF.

(À gauche) : Modèle de souris *Holstein* ($Hol^{Tg/Tg}$) utilisée comme modèle pour la maladie d'Hirschsprung. **(Au centre) :** Représentation du côlon au jour P20 pour la souris de type sauvage, la souris *Holstein* sans traitement ainsi que la souris *Holstein* traitée au GDNF. **(À droite) :** Représentation du côlon chez un patient atteint de la maladie d'Hirschsprung ayant subi un traitement au GDNF. Tiré de [Glial Cell-Derived Neurotrophic Factor Induces Enteric Neurogenesis and Improves Colon Structure and Function in Mouse Models of Hirschsprung Disease] (Soret *et al.*, 2020), CC BY-NC-ND 4.0 DEED.

duit les phénotypes de la maladie d'Hirschprung. Les souris $Hol^{Tg/Tg}$ avaient une dépigmentation au niveau de leur fourrure (albinisme) due à une réduction de la production de la mélanine. Chez les souris hétérozygotes ($Hol^{Tg/+}$), il y avait des taches très caractéristiques provoquées par le mini gène de la tyrosinase *Tyr* ayant été inséré dans le gène. Cela causait un rétablissement partiel de la pigmentation due à la mélanine sous forme de taches très caractéristiques et il permettait de reconnaître visuellement les souris transgéniques homozygotes *Holstein*. Comparées à celles qui étaient hétérozygotes, elles avaient une accumulation très prononcée de matière fécale dans leur côlon, ce qui causait un taux de mortalité de 100% chez ces souris (Soret *et al.*, 2015).

Soret et al. (2015) ont ensuite extrait les tissus au jour 12,5 après la naissance. Ensuite, ils avaient fait le génotypage du transcriptome par RT-PCR avec les cel-

lules dérivées de la crête neurale ayant été isolées par le tri cellulaire activé par fluorescence (FACS) en utilisant le marqueur de coloration rouge DsRed2. Cela avait pour but de démontrer la surexpression du gène Col6a4 dans les cellules, comparées aux souris de type sauvage. Ils ont aussi utilisé l'immunofluorescence afin de démontrer une augmentation de la production du collagène VI dans les cellules. Enfin, ils ont étudié les processus de différenciation et de prolifération des cellules en neurones et en cellules gliales, ainsi que leur apoptose. Ils ont découvert une réduction de la quantité de cellules progénitrices qui permettaient cette différenciation. Enfin, ils ont effectué des analyses afin de mieux comprendre la raison de la réduction de la vitesse de migration des cellules dérivées de la crête neurale due aux effets des mutations du gène Col6a4. Ces analyses impliquaient la culture des cellules des tissus ayant été traitées avec du GDNF. Ils ont découvert que s'il était utilisé conjointement avec la fibronectine, ils allaient tous les deux permettre de promouvoir la migration des cellules dans la partie médiale de l'intestin chez les souris de type sauvage, mais son effet était moins présent dans les souris homozygotes *Holstein* (Soret *et al.*, 2015).

Néanmoins, le fait que le GDNF ait eu un effet en augmentant la migration des cellules s'est avéré très prometteur pour eux et ils l'ont utilisé dans leurs prochaines études afin de développer de nouvelles thérapies alternatives à la chirurgie pour le traitement de la maladie d'Hirschprung (Soret *et al.*, 2015). Dans une autre étude faite par ces auteurs (Soret *et al.*, 2020), ils ont administré un lavement avec une solution composée de GDNF dans le côlon des modèles de souris couramment utilisés en plus de leurs propres modèles *Holstein*. Ces modèles de souris *TashT* ($TashT^{\mathrm{Tg/Tg}}$) avaient des insertions *Tyr* dans le chromosome 10 et les souris mâles avaient des phénotypes similaires à la maladie d'Hirschsprung (Bergeron *et al.*, 2015). De plus, ils ont utilisé des modèles *Piebald-lethal* avec une mutation sur le gène Ednrb ($Ednrb^{\mathrm{s\text{-}l//s\text{-}l}}$) (Cantrell, 2004) et qui avaient comme phénotype une

aganglionose dans la partie distale du côlon. Enfin, ils ont aussi utilisé des modèles de souris $Ret^{9\,/-}$, qui avaient comme phénotype une aganglionose du côlon (Uesaka *et al.*, 2008). Ils ont constaté que le traitement restaurait à des niveaux presque normaux la concentration en cellules gliales et neuronales dans la partie distale du côlon qui présentait une aganglionose chez les modèles de souris utilisés, sauf pour le modèle avec une mutation sur le gène Ret ($Ret^{9\,/-}$) (Soret *et al.*, 2020) (figure 1.1).

1.2.5 Importance de la cellule pour le séquençage de l'ARN de cellule unique afin de mieux comprendre la maladie Hirschsprung

Plusieurs processus biologiques sont nécessaires lors du développement du système nerveux entérique, notamment sur le plan de la migration des cellules progénitrices issues de la crête neurale vers le colon, ainsi que leur expression de gènes spécifiques afin de se différencier en cellules gliales et neuronales matures (Goldstein *et al.*, 2013) (Mueller et Goldstein, 2022).

Les méthodes de séquençage traditionnelles n'étaient pas appropriées afin de déterminer de manière spécifique les processus biologiques qui étaient impliqués dans la maladie d'Hirschsprung. En effet, il était nécessaire de pouvoir isoler et identifier les types de cellules d'intérêt qui étaient impliquées dans cette maladie et de pouvoir comparer les populations de cellules ainsi que leur expression des gènes d'intérêt avec les cellules issues des patients sains (Tarapcsak *et al.*, 2025).

Les nouvelles méthodes de séquençage ont permis de mieux étudier cette maladie, notamment le séquençage de l'ARN de cellule unique (scRNA-seq). Cette méthode a permis d'isoler les différentes populations de cellules qui sont impliquées dans les processus biologiques associés à cette maladie, notamment sur le plan des cellules progénitrices et des cellules neurales matures (Tarapcsak *et al.*, 2025). En effet,

puisque cette maladie implique une aganglionose du colon (Mueller et Goldstein, 2022), il est important d'analyser les populations de cellules présentes dans ces tissus afin de déterminer s'il y a la présence de cellules progénitrices. Ces dernières ont le potentiel de se différencier en cellules gliales et neuronales, ce qui ouvre des voies pour des traitements afin de restaurer les cellules gliales et neuronales à des niveaux qui sont similaires aux patients ne souffrant pas de cette maladie (Tarapcsak *et al.*, 2025).

De récentes études (Tarapcsak *et al.*, 2025) ont démontré en utilisant la méthode de séquençage d'ARN de cellule unique que, même s'il y a une aganglionose dans le colon des patients souffrant cette maladie, il avait quand même la présence de cellules progénitrices dans les tissus. Cependant, il n'y avait pas la présence de cellules neuronales qui sont matures. Ils ont aussi découvert une réduction de l'expression des gènes mentionnés précédemment dans ce chapitre qui sont associés à la différenciation des cellules progénitrices en cellules neurales matures, malgré le fait qu'elles avaient le potentiel de pouvoir quand même migrer dans les tissus du colon (Tarapcsak *et al.*, 2025).

Dans le cadre de ce mémoire, la méthode de séquençage d'ARN de cellule unique a aussi été utilisée afin d'isoler les différentes populations de cellules et de comparer leur expression des gènes associés aux cellules gliales et neuronales. Les échantillons proviennent de souris de type sauvage, de souris de type sauvage ayant été traitées au dextran sulfate de sodium afin de simuler une réponse inflammatoire (Chassaing *et al.*, 2014), de souris *Holstein* ayant des phénotypes similaires à la maladie d'Hirschsprung (Soret *et al.*, 2015) ainsi que de souris *Holstein* ayant été traitées au GDNF, un facteur de transcription qui a le potentiel de restaurer les populations de cellules gliales et neuronales à des niveaux similaires aux échantillons de souris de type sauvage (Soret *et al.*, 2015) (Soret *et al.*, 2020). Plus de détails sur le séquençage de l'ARN de cellule unique seront décrits dans le chapitre

2, ainsi que la méthodologie dans le chapitre 5.

1.3 Synthèse du chapitre

En résumé, le système nerveux joue un rôle très important dans le système intestinal. Il permet de faire la régulation sur le plan hormonal et d'assurer la motilité intestinale. C'est aussi grâce aux cellules gliales et neuronales provenant de la crête neurale qui jouent un rôle clé dans le développement du système intestinal (Mueller et Goldstein, 2022).

Lors de l'embryogenèse, les cellules dérivées de la crête neurale migrent à partir de cette dernière sur tout le côlon jusqu'au rectum. Par la suite, elles prolifèrent et se différencient en cellules gliales et neuronales afin d'assurer la motilité intestinale. Une défaillance dans une de ces étapes peut causer une aganglionose partielle ou totale dans le côlon (Mueller et Goldstein, 2022). Cela va aussi causer des déficits au niveau de la motilité intestinale et une accumulation des selles (Diposarosa *et al.*, 2021) ainsi qu'un risque de septicémie (Butler Tjaden et Trainor, 2013), ce qui est un phénotype caractéristique de la maladie d'Hirschsprung (Mueller et Goldstein, 2022).

Ce type de maladie intestinale qui est diagnostiquée chez les nouveau-nés peut être mortelle si elle n'est pas traitée rapidement dans les premiers jours (Mueller et Goldstein, 2022). De plus, les traitements actuels sont principalement l'ablation de la partie affectée via la chirurgie (Urla *et al.*, 2018), ce qui peut causer beaucoup d'effets secondaires chez les patients (Niramis *et al.*, 2008).

C'est pourquoi plusieurs études ont été réalisées afin d'étudier cette maladie. La méthode de séquençage de l'ARN de cellule unique est la plus appropriée pour l'étude de la maladie d'Hirschsprung, car elle permet d'isoler les cellules progénitrices, gliales et neuronales qui sont impliquées dans cette maladie et elle permet

aussi de déterminer leur évolution sur le plan de leur expression des gènes qui ont une incidence sur leur migration ainsi que sur leur différenciation en cellules matures (Tarapcsak *et al.*, 2025). Les études de cette maladie comportent le développement de modèles de souris *Holstein* qui reproduisent des phénotypes similaires à la maladie d'Hirschsprung (Soret *et al.*, 2015). De plus, ces études incluent le développement de traitements alternatifs qui sont moins invasifs, tels que l'utilisation du GDNF sur des souris *Holstein* afin de restaurer les niveaux de cellules gliales et neuronales dans le côlon (Soret *et al.*, 2020). Afin de mieux comprendre ce phénomène, les chercheurs ont utilisé des méthodes de séquençages qui seront expliquées dans le chapitre suivant.

CHAPITRE II

REVUE DES MÉTHODES DE SÉQUENÇAGE

Afin de mieux comprendre les processus biologiques tels que ceux du chapitre précédent, il y a eu un développement de nouvelles méthodes de séquençage de l'ADN et de l'ARN. Chacune de ces méthodes de séquençage utilise un type d'échantillon particulier et comporte ses avantages et ses faiblesses.

Le besoin de comprendre le génome des espèces a causé le développement des méthodes de séquençage afin d'atteindre ces objectifs. Autrefois, ils étaient plus axés pour cartographier le génome de toutes les espèces vivantes dans le monde. Pour les espèces animales, les méthodes de séquençages avaient permis de cartographier pour la première fois le génome du ver *Caenorhabditis elegans* en 1998 et il était d'une taille de 97 mégapaires de bases (The C. elegans Sequencing Consortium*, 1998). Ensuite, l'avancement des méthodes de séquençage a permis de séquencer d'autres espèces végétales et animales plus complexes, telles que la souris en 2002 (Mouse Genome Sequencing Consortium, 2002).

Le génome humain est extrêmement complexe, bien plus que la plupart des espèces animales et végétales. Selon Collins et Fink (1995), il a été évalué à plus de 3 milliards de nucléotides dans le génome humain. Autrefois, les objectifs des méthodes de séquençage étaient de cartographier l'entièreté du génome de référence de l'être humain afin que les futurs chercheurs puissent utiliser ces informations

pour mieux identifier les gènes qui sont impliqués dans les maladies. Le projet *Human Genome* a été établi afin d'atteindre cet objectif. De plus, des normes éthiques et légales ont aussi été établies sur le plan de la méthodologie et il y avait aussi un aspect social qui avait pour but de mieux éduquer la population. Ce projet a constitué le fondement de nombreux développements de nouvelles méthodes de séquençage afin de mieux comprendre la composition du génome humain ainsi que le fonctionnement du transcriptome (Collins et Fink, 1995).

2.1 Méthodes de séquençage traditionnelles

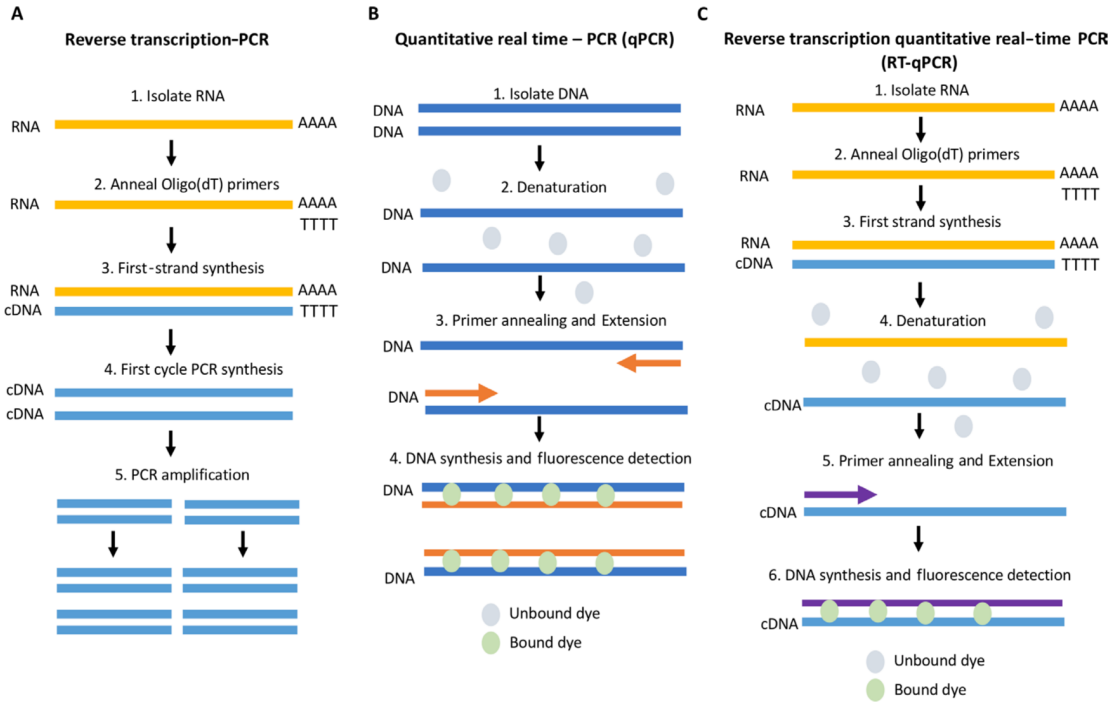

FIGURE 2.1 – Exemple de méthode de séquençage par PCR.

(A) PCR à l'aide de la transcriptase inverse (RT-PCR). **(B)** PCR quantitative en temps réel (qPCR). **(C)** PCR quantitative en temps réel à l'aide de la transcriptase inverse (RT-qPCR). Tiré de [A beginner's guide to RT-PCR, qPCR and RT-qPCR] (Adams, 2020), CC BY 4.0 DEED.

Selon Ribarska et al. (2022), les anciennes méthodes traditionnelles de séquen-

çage impliquaient l'utilisation d'enzymes afin de fragmenter l'ADN. Cependant, ces méthodes pouvaient comporter des biais, car il y avait des régions où les enzymes allaient préférablement cliver au lieu des régions d'intérêts. Une autre méthode utilisait la tagmentation, ce qui consistait à ajouter des séquences partielles d'adaptateurs sur plusieurs parties de l'ADN pour la fragmenter. Ces fragments d'ADN gardaient ainsi leurs adaptateurs à leurs deux extrémités qui allaient être complétées lors de l'amplification de l'ADN à l'aide de la réaction en chaîne par la polymérase (PCR) (Ribarska *et al.*, 2022). Enfin, il y avait la sonication, dont la méthode consistait à utiliser des ultrasons afin de fragmenter l'ADN et cette méthode est présentement la méthode la plus populaire (Sun *et al.*, 2022).

La base des méthodes de séquençage traditionnelles impliquait l'amplification de l'ADN par PCR. Cela permettait de mesurer la quantité de transcrits dans les échantillons (Kukurba et Montgomery, 2015) (figure 2.1). Cependant, il y avait plusieurs limitations dans ce type d'analyse. Selon Okoniewski et Miller (2006), il pouvait avoir une hybridation entre des séquences très similaires. S'il y avait de la contamination, ces derniers allaient être amplifiés par PCR, ce qui allait causer un biais dans les résultats en induisant de faux positifs. De plus, s'il y avait des amorces non spécifiques, elles risquaient d'amplifier plusieurs parties de l'ADN qui n'étaient pas pertinentes pour les recherches (Okoniewski et Miller, 2006).

2.1.1 Séquençage de Sanger

Selon Zhang et al, (2021), ce type de séquençage impliquait un mélange de ddNTP (didésoxyribonucléotides), qui sont des nucléotides n'ayant pas un groupement hydroxyle dans les extrémités 2' et 3'. Cette absence de groupement hydroxyle les empêchait de créer des liaisons phosphates lors de la réplication. Ces ddNTP étaient étiquetés par un marqueur fluorescent et ils étaient ajoutés dans le mélange

de nucléotides dans quatre réactions séparées. Chaque réaction contenait un type de nucléotide particulier, étant de l'adénine, de la cytosine, de la guanine ainsi que de la thymine. Lors de la réplication, ces nucléotides étaient incorporés à la chaîne terminale de la séquence, ce qui allait générer plusieurs différentes séquences qui seraient détectables par la fluorescence. Ensuite, ces chaînes d'ADN étaient triées et analysées par électrophorèse en fonction de leur taille (Zhang *et al.*, 2021) pour être ensuite alignées au génome de référence (Shaibu *et al.*, 2021). Par contre, si les séquences étaient très courtes avec des tailles de moins de 100 nucléotides, ils étaient plus difficiles à assembler dans le génome de référence, en particulier s'il y avait des séquences répétitives dans celui-ci (Dewey *et al.*, 2012). De plus, il avait pour désavantage d'être très dispendieux (Kozińska *et al.*, 2019) et il pouvait seulement séquencer une seule séquence d'ADN à la fois. En outre cette séquence ne devait pas être plus de 1000 nucléotides (Petersen *et al.*, 2017). Malgré cela, la méthode de Sanger reste à ce jour une méthode courante pour le séquençage de l'ADN, vu à sa précision de plus de 99,99% (Shaibu *et al.*, 2021).

2.1.2 Séquençage «fusil de chasse»

Le séquençage «fusil de chasse» est un dérivé de la méthode de Sanger qui consiste à séquencer de l'ADN d'environ 500 à 600 paires de bases en parallèle et qui sont réparties sur plusieurs régions de l'ADN. Ces séquences sont ensuite rassemblées en se superposant au niveau de leurs extrémités en séquences continues d'ADN appelées contigs, puis elles sont ensuite alignées sur le génome de référence en se superposant à leurs extrémités (Dewey *et al.*, 2012). Il est aussi très utilisé pour le séquençage métagénomique afin d'étudier la diversité sur le plan de la taxonomie (Meslier *et al.*, 2022). Il est aussi possible de l'utiliser pour le séquençage *de novo* afin de découvrir de nouvelles séquences (Chapman *et al.*, 2015). Il se révèle avantageux pour faire du séquençage plus quantitatif des régions d'intérêt et afin

d'extraire des données sur le plan des échantillons. Par contre, cette technique a le désavantage d'être très dispendieuse et d'exiger des séquences de haute qualité en très grande quantité. De plus, il est nécessaire que les bases de données du génome de référence soient assez complètes (Bell *et al.*, 2021).

2.1.3 Séquençage à l'aide d'un standard quantitatif

Selon Kukurba et Montgomery, (2015), ce type de séquençage utilisait un standard d'étalonnage appelé «pointe», qui était des séquences d'ARN synthétiques servant de contrôles positifs lors du profilage du transcriptome. Ces «pointes» de standard d'étalonnage étaient ajoutées à la librairie à différentes concentrations afin de faire un contrôle de la qualité en mesurant les variations techniques comparées aux variations biologiques lors du profilage du transcriptome. Elles étaient constituées de 96 plasmides composés d'une séquence standard d'ADN qui avait été insérée dans un vecteur. Ces plasmides étaient ensuite ajoutés à la librairie. Cette méthode était très utilisée dans une approche d'évaluation quantitative, qualitative en matière de sensibilité, ainsi que sur le plan des analyses de la cartographie des séquences (Kukurba et Montgomery, 2015).

2.2 Séquençage de nouvelle génération (NGS)

Dans un éditorial de Dahui (2019), il expliquait que cette méthode comportait quatre grandes étapes. La première étape était la fragmentation de l'ADN en séquences de 100 à 300 paires de bases en utilisant des méthodes de clivage de l'ADN. Ensuite, ces fragments étaient isolés en utilisant la méthode de capture par hybridation en utilisant des sondes complémentaires qui allaient s'hybrider avec ces fragments. L'isolation de ces fragments pouvait aussi se faire par la méthode d'amplification par PCR en utilisant des amorces qui allaient s'hybrider dans les

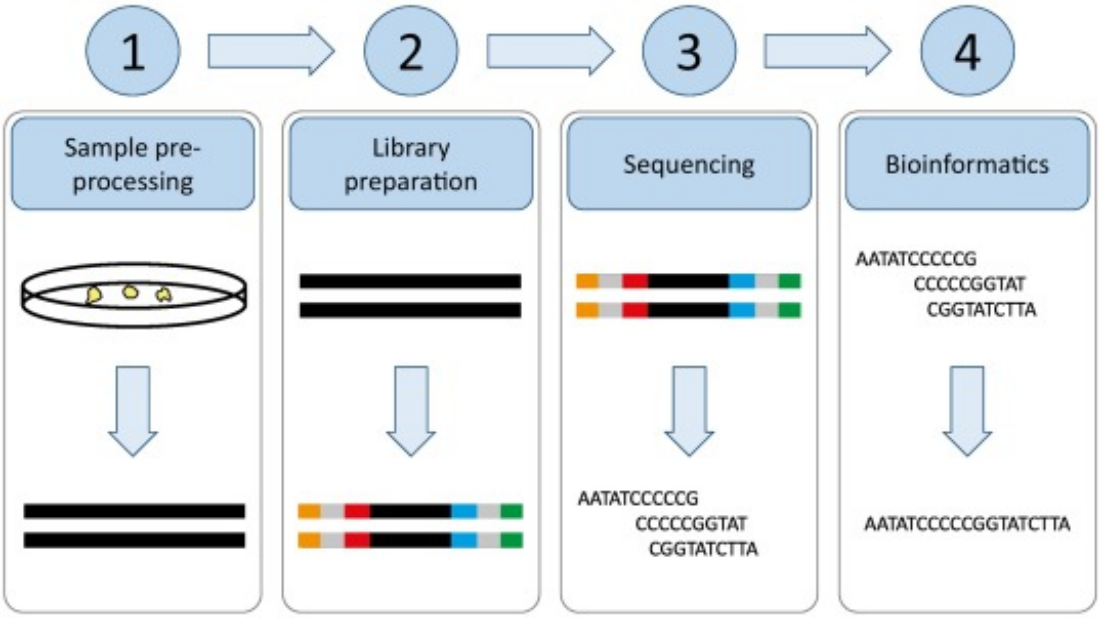

FIGURE 2.2 – Principales étapes du séquençage de nouvelle génération. **(1)** Prétraitement de l'échantillon. **(2)** Préparation de la librairie des séquences. **(3)** Séquençage des échantillons. **(4)** Analyses bio-informatiques des résultats du séquençage. Tiré de [Library preparation for next generation sequencing: A review of automation strategies] (Hess *et al.*, 2020), CC-BY-NC-ND.

régions d'intérêts et ces dernières allaient être amplifiées à l'aide de la polymérase (Dahui, 2019).

La prochaine étape était la préparation de la librairie en modifiant chaque segment en y ajoutant des adaptateurs afin de les identifier et de les indexer. Ces adaptateurs étaient universels en permettant aux amorces de s'hybrider afin d'amplifier toutes les séquences (Dahui, 2019).

Ensuite, la librairie des séquences était chargée dans une matrice afin que celles-ci soient séquencées en parallèle et les informations étaient stockés sous forme de lecture (Dahui, 2019).

Enfin, ces lectures étaient analysées à l'aide d'outils en bio-informatique en les identifiant, en les alignant sur le génome de référence et en effectuant des annota-

tions. Ces annotations permettaient de comparer les régions amplifiées au génome de référence afin de vérifier s'il y a des variations génétiques ou des mutations ou pour d'autres analyses selon les besoins du chercheur (Dahui, 2019) (figure 2.2).

2.2.1 Séquençage nanopore

Ce type de séquençage utilise une surface solide avec de petits pores à travers lesquels les fragments d'ADN peuvent passer. Cela génère un signal ionique qui peut ensuite être détecté. Cette méthode a pour avantage de ne nécessiter que de très petites quantités et elle ne requiert pas beaucoup de préparation des échantillons (Dewey *et al.*, 2012). Cependant, il peut y avoir des biais si les séquences ont une forte présence de guanine et de cytosine et s'il y a présence de substitutions (Delahaye et Nicolas, 2021). Une des méthodes d'optimisation est l'utilisation d'enzymes, telles que des exonucléases qui vont cliver les séquences d'ADN en nucléotides tout en les identifiant afin qu'elles puissent être reconnues dans le bon ordre lorsqu'elles traversent les pores (Clarke *et al.*, 2009).

2.2.2 Séquençage de l'ARN (RNA-seq)

Le séquençage d'ARN (RNA-seq) permet de mieux comprendre le transcriptome en étudiant l'expression des gènes, de l'épissage génétique, ainsi que l'expression de ces gènes dans les allèles spécifiques dans le chromosome (Kukurba et Montgomery, 2015). Il permet aussi d'identifier les variants dans les gènes qui sont souvent dus à des mutations génétiques, telles que des insertions, des délétions, des duplications, des inversions des transposons, ainsi que des substitutions. Tous ces variants peuvent affecter l'expression des gènes (Dewey *et al.*, 2012).

Au cours des dernières années, plusieurs méthodes de séquençage ont été développées et perfectionnées afin de permettre la réduction de la quantité d'ARN

nécessaire dans un échantillon. Dans certains cas, elle avait comme inconvénient d'être beaucoup plus coûteuse. Par contre, les avantages étaient qu'ils généraient beaucoup plus d'informations sur le plan du transcriptome (Kukurba et Montgomery, 2015).

Une manière de réduire les coûts était de sélectionner des régions spécifiques d'intérêts et de faire le multiplexage des échantillons en constituant une seule voie de séquençage. Cette méthode réduisait grandement les coûts reliés à la préparation de librairies de séquences (Kukurba et Montgomery, 2015).

Selon Kukurba et Montgomery (2015), lors de l'extraction des cellules pour les séquençages de l'ARN, il était important de considérer l'hétérogénéité des tissus et de sélectionner le bon type de tissu. Un exemple de l'hétérogénéité du tissu dans un seul échantillon était qu'il pouvait y avoir des cellules mortes et des cellules vivantes. Dans un autre cas, certaines cellules étaient saines et d'autres étaient atteintes de maladie, telle que le cancer. C'est pourquoi il existait des méthodes de capture et de purification des cellules à l'aide de la microdissection par laser ou par l'observation des cellules à l'aide d'un microscope (Kukurba et Montgomery, 2015).

De plus, ces auteurs mentionnaient qu'il fallait isoler l'ARN messager des échantillons et qu'il devait être en quantité suffisante. De plus, sa qualité et la pureté devaient aussi être très élevées en ayant le moins de contaminant et de dégradation possible. Un exemple de méthode afin d'évaluer la qualité de l'ARN messager était d'utiliser un bioanalyseur de la compagnie *Agilent*, dont les résultats pouvaient avoir un score allant de la plus mauvaise qualité (1) à la meilleure (10) (Kukurba et Montgomery, 2015).

Ensuite, il fallait convertir les échantillons d'ARN en ADN complémentaire à l'aide de la transcriptase inverse, préparer la librairie de séquences et les amplifier par

PCR. Ainsi, les lectures de ces séquences pouvaient être alignées sur le génome de référence ou par un assemblage *de novo* (Kukurba et Montgomery, 2015).

Enfin, les techniques de séquençage de nouvelle génération permettaient de mieux faire le profilage dans l'ensemble du transcriptome, mais il était limité à l'échantillon de tissus. De plus, il ne permettait pas d'analyser ce dernier sur le plan de la variabilité des cellules qui était présente dans les échantillons. C'est pourquoi d'autres techniques, telles que le séquençage de l'ARN de cellule unique, ont été développées par la suite (Kukurba et Montgomery, 2015).

2.3 Le séquençage de l'ARN de cellule unique (scRNA-seq)

Le séquençage de l'ARN de cellule unique a été développé en 2009 par Tang et al. Ils ont utilisé cette méthode pour le séquençage du transcriptome de blastomère et oocytes de souris (Tang *et al.*, 2009). Cette méthode de séquençage a permis d'analyser le transcriptome dans chaque cellule individuellement, de les comparer entre elles et de les regrouper en population ayant des expressions génétiques similaires (Anaparthy *et al.*, 2019).

2.3.1 Avantages et applications

Selon Anaparthy et al. (2019), cette méthode s'est avérée particulièrement utile afin d'analyser l'hétérogénéité des populations des cellules dans un tissu et même des microorganismes impliqués dans l'immunité. Elle a permis aussi d'analyser l'expression génétique d'une cellule en fonction du temps, mais aussi en fonction de son environnement. Traditionnellement, les cellules étaient classées en fonction de leur forme et de leur position dans l'organisme par les chercheurs, mais cette méthode a permis de mieux les classifier en population de cellules par des techniques de regroupement en sous-populations. Ces nouvelles méthodes ont permis

de mieux comprendre le fonctionnement des cellules lors du développement embryonnaire et dans le vieillissement des cellules. Cela a aidé aussi à mieux évaluer les variations phénotypiques, par exemple sur le plan des différentes cellules immunitaires. Une autre application était dans les recherches contre le cancer, car il permet d'analyser l'hétérogénéité des cellules dans les tissus tumoraux. Enfin, d'autres applications pour cette méthode sont l'étude des populations de microorganismes dans le microbiome (Anaparthy *et al.*, 2019).

2.3.2 Préparation de la librairie

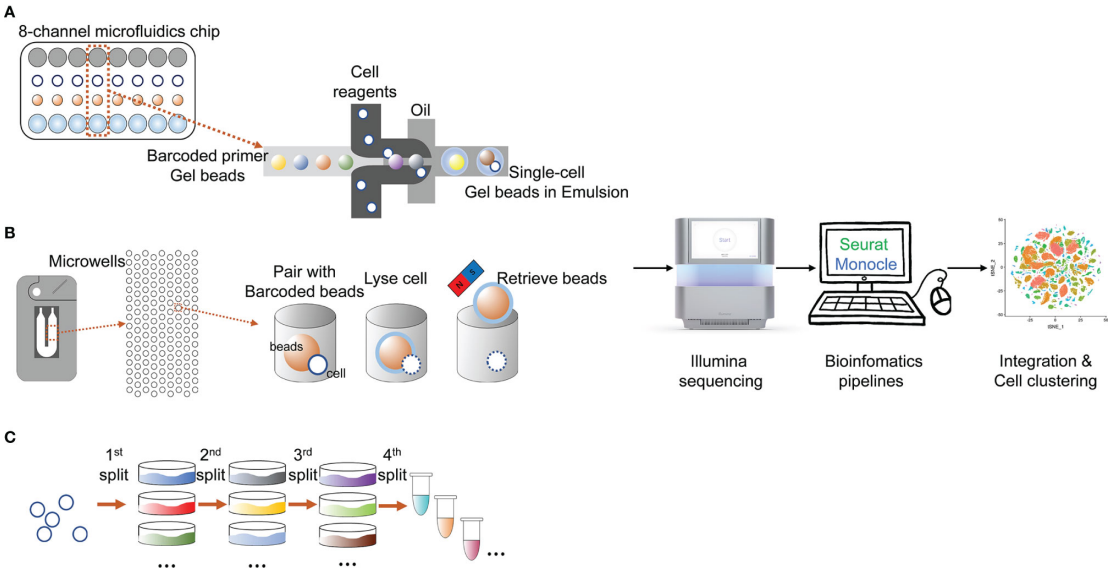

FIGURE 2.3 – État de l'art des méthodes de séquençage de l'ARN de cellule unique.

(A) Méthode utilisant les microfluides. **(B)** Méthode utilisant les micropuces. **(C)** Méthode de séquençage SPLiT-seq. Tiré de [Single-Cell Sequencing on Marine Life: Application and Future Development] (Li *et al.*, 2022), CC BY 4.0 DEED.

Les cellules peuvent être séparées en conservant tout leur contenu, seulement leur noyau ou seulement certaines de leurs organelles. Elles sont ensuite isolées en fonction des marqueurs grâce au tri cellulaire activé par fluorescence (FACS). D'autres techniques peuvent être utilisées, telles que le triage des cellules a l'air

de magnétisme (MACS), ou de l'utilisation de la microdissection à l'aide d'un laser ou d'un système microfluidique (Zeb *et al.*, 2019) (figure 2.3).

Anaparthy et al. (2019) expliquent que chaque cellule ayant été isolée est ensuite encapsulée dans un micro-environnement sous forme de gouttelettes. Cette gouttelette contient aussi un oligonucléotide qui sert de code-barre unique, ainsi que tous les réactifs permettant la synthèse de l'ADN complémentaire qui est alors hybridé avec cet oligonucléotide. Ensuite, ces gouttelettes sont brisées et tous les brins d'ADN complémentaires hybridés sont regroupés et amplifiés par PCR à l'aide d'amorces spécifiques à ces oligonucléotides. Une librairie est ensuite préparée avec les brins qui auront été séquencés et ces dernières sont analysées par des outils en bio-informatique (Anaparthy *et al.*, 2019).

Dans la méthode de séquençage du transcriptome basée sur la ligature par groupements séparés (SPLiT-seq), les cellules ou les noyaux sont traités successivement et aléatoirement avec 4 solutions de milieux réactifs contenant les codes-barres et les amorces afin d'avoir des combinaisons de codes-barres différents (Rosenberg *et al.*, 2018).

Jovic et al (2022) ont aussi mentionné une variante appelée le séquençage de l'ARN nucléaire de cellule unique (snRNA-seq). La différence avec la version traditionnelle est qu'au lieu d'utiliser l'ARN messager présent dans le cytoplasme de la cellule, seulement les ARN messagers qui étaient présents dans le noyau sont capturés et séquencés. Cette variante peut être utilisée avec des cellules provenant de divers types de tissus d'organes. Ce type de séquençage peut aussi être utilisé avec différentes espèces de la famille des eucaryotes. Cette technique s'avère utile pour évaluer le transcriptome uniquement dans le noyau, mais elle comporte le désavantage de ne pas pouvoir détecter les autres ARN ayant subi d'autres processus biologiques, tels que leur maturation (Jovic *et al.*, 2022).

Une autre méthode est le séquençage par la transposase accessible à la chromatine (ATAC-seq) ainsi que sa variante, le séquençage de cellule unique de la transposase accessible à la chromatine (scATAC-seq). Elles permettent d'analyser les régions de la chromatine qui sont accessibles (Ji *et al.*, 2020) grâce à la méthode de la tagmentation par la transposase Tn5. Cette transposase clive les régions ayant été exposées lors de du déroulement de la chromatine autour des histones et ces régions peuvent ensuite être séquencées (Buenrostro *et al.*, 2015).

2.3.3 Analyses des données de séquençage de l'ARN de cellule unique

Dans un article écrit par Jovic et al. (2022), ils expliquent que, selon le type de la plateforme de séquençage, les données brutes peuvent être sous forme de fichiers en format FASTQ ou BCL. Les fichiers en format FASTQ peuvent directement être analysés pour le contrôle de la qualité. Si les données sont en format BCL, il faut le convertir en données FASTQ à l'aide du logiciel *cellranger mkfastq*, qui est un encapsuleur du logiciel *bcl2fastq*. Dans les cas des fichiers BCL, il faut fournir un fichier CSV qui contient une matrice contenant les informations, telles que l'échantillon, la voie et l'index. Ensuite, il faut faire le contrôle de la qualité qui peut être fait avec le logiciel *FastQC*. Les séquences sont ensuite alignées à l'aide de différents outils, tels que *TopHat* et *STAR* dans le cas de Cell Ranger. Ce dernier effectue les alignements des lectures de séquences et il fait les annotations des régions grâce à un fichier GTF. Il comprend aussi d'autres étapes, telles que le filtrage et le comptage des identifiants uniques moléculaires (UMI). Ensuite, il faut faire une étape de normalisation afin d'éliminer les bruits causés par l'effet de lot qui est un phénomène causé par des variations techniques lors de la préparation de la librairie ou l'isolation de l'ARN. Ces variations peuvent être biologiques, telles que l'état des cellules, de leur expression génétique, des dommages au niveau de la membrane cellulaire ou lorsqu'elles s'agglutinent entre elles. Si ces effets ne sont

pas éliminés dans les premières étapes des analyses, cela peut causer des biais dans les résultats subséquents (Jovic *et al.*, 2022).

Ensuite, Jovic et al (2022) indiquent que, lorsqu'il y a des gènes hautement variables (HVGs) de bonne qualité, ils peuvent être sélectionnés et ils permettent de mieux différencier chaque type de cellule. Puisque la dimension est très élevée due à des centaines de milliers de cellules qui peuvent à elles seules exprimer plusieurs milliers de gènes, une réduction de cette dimension est nécessaire. En plus de la réduction de la dimension, une analyse en composantes principales (PCA) est aussi effectuée. Cette méthode est un algorithme de réduction linéaire qui permettait de transformer des variables multiples en une relation linéaire. Elle permet donc d'analyser les données à une dimension moins élevée et elle est aussi utilisée afin de réduire les bruits techniques. Ensuite, une projection UMAP ou t-SNE est générée afin de visualiser les regroupements des cellules et de les annoter. Normalement, les annotations se font manuellement à l'aide de l'expertise d'un biologiste ou de la littérature, ce qui peut être très laborieux. Ils peuvent aussi se faire automatiquement à l'aide d'algorithmes, de bases de données ainsi qu'une liste de marqueurs connue par type de cellule. Ces résultats peuvent ensuite être validés avec des expériences en laboratoire. Le principal avantage des annotations automatiques des cellules est qu'elles sont reproductibles et beaucoup plus rapides. Par contre, il ne permet pas d'identifier les cellules ayant des caractéristiques rares ou de nouveaux types de cellules. Les étapes subséquentes permettent de faire les analyses au niveau de l'expression différentielle des gènes entre chaque cellule, la trajectoire des cellules sur le plan temporel, de la sélection ainsi que les manipulations des populations de cellules (Jovic *et al.*, 2022).

2.3.4 Désavantages

Jovic et al (2022) mentionnent qu'un des principaux désavantages de la méthode de séquençage de l'ARN de cellule unique est qu'il y a une perte de l'information sur le plan temporel due à la préparation des cellules à partir des tissus. De plus, le fait d'isoler et de conserver les cellules en dehors de leur environnement biologique habituel peut altérer l'expression de leurs gènes. Les variations techniques et biologiques peuvent biaiser grandement les résultats si elles ne sont pas traitées adéquatement. De plus, cette méthode de séquençage s'avère très dispendieuse et les analyses ne sont pas automatisées. Enfin elle ne comporte pas une interface utilisable par un utilisateur qui n'aurait pas d'expérience en bio-informatique et qui ne posséderait pas les ressources nécessaires pour faire ce type d'analyse (Jovic *et al.*, 2022).

2.3.5 Enjeux des données associées au séquençage de l'ARN de cellule unique

Il existe plusieurs enjeux sur le plan des données issues du séquençage de l'ARN de cellule unique. Ces derniers peuvent grandement influencer la méthodologie lors des traitements des données, les résultats qui sont générés, ainsi que leur interprétation.

Un des enjeux est que les données provenant des études ne sont pas toujours accessibles aux chercheurs, ce qui rend très difficile la comparaison des résultats avec la littérature ou avec d'autres études (Lähnemann *et al.*, 2020). Un autre enjeu est que, lors du traitement des données, il est important d'éliminer l'effet de lot qui est dû à des variations techniques et biologiques lors de la préparation de la librairie, en particulier lorsqu'il y a plusieurs échantillons à analyser et que ces derniers contiennent une très grande quantité de cellules (Jovic *et al.*, 2022).

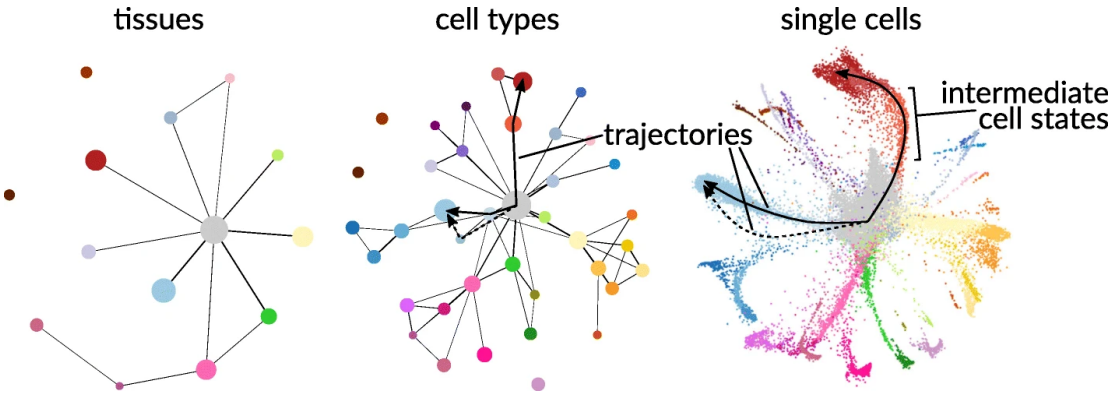

FIGURE 2.4 – Différents niveaux de résolutions d'intérêt.
(A) Tissus. **(B)** Types de cellules. **(C)** Cellules uniques. Tiré de [Eleven grand challenges in single-cell data science] (Lähnemann *et al.*, 2020), CC BY 4.0 DEED.

De plus, les données sont générées sont à très grande échelle, ce qui augmente leur dimensionnalité, en particulier si leur résolution est augmentée. Il est donc nécessaire de les diminuer lors des traitements des données grâce à une analyse en composantes principales (PCA) afin de faire une réduction sur le plan linéaire (Jovic *et al.*, 2022). Cependant, si la résolution est diminuée, il y aura une perte de l'information. Différents niveaux de la résolution permettent de voir les liens entre les tissus, les trajectoires de développement entre les populations de cellules, ainsi que les états intermédiaires entre les cellules lors de leur différenciation (figure 2.4). Il est donc nécessaire d'ajuster cette résolution afin de maximiser le gain d'information et de minimiser la perte d'information lors des annotations. Cet ajustement de la résolution est plus facile si les outils incluent une interface interactive qui permet de visualiser directement les projections (Lähnemann *et al.*, 2020).

Enfin, lors des analyses de l'expression différentielle des gènes entre les regroupements de cellules, il est nécessaire de bien comprendre quels phénomènes différencient les populations de cellules les unes par rapport aux autres (figure 2.5). Par

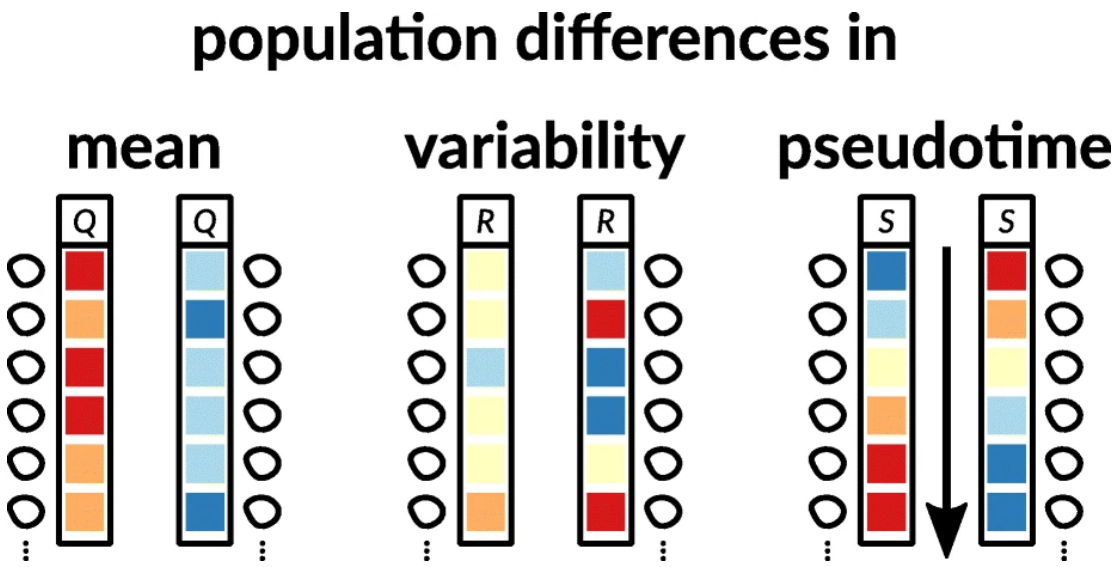

FIGURE 2.5 – Expression différentielle d'un gène ou transcrit entre les populations de cellules.

(A) Moyenne. **(B)** Variabilité. **(C)** Pseudotemps. Tiré de [Eleven grand challenges in single-cell data science] (Lähnemann *et al.*, 2020), CC BY 4.0 DEED.

exemple, s'il y a une différence au niveau de la moyenne entre les populations, c'est-à-dire si l'expression des gènes est constante entre toutes les cellules d'une population. Cette différence peut aussi se faire au niveau de la variabilité entre les groupes, c'est-à-dire que toutes les cellules dans un groupe ont une expression qui est différente par rapport à celles dans un autre groupe. Enfin cette différence peut se faire au niveau du pseudotemps, c'est-à-dire qu'il y a un changement de l'expression des gènes dans une population qui évolue de manière temporelle, par exemple selon une trajectoire de développement (Lähnemann *et al.*, 2020). Le pseudotemps est utile lorsque les échantillons ont été prélevés à un temps fixe au lieu de se faire à différents temps. Ce type d'analyse va générer une prédiction sous forme de trajectoire de l'évolution des différents groupes de cellules sur le plan des différents niveaux de l'expression des gènes. Cette méthode est très utile afin d'éviter de répéter les expériences à différents moments (Trapnell *et al.*, 2014) (Trapnell, 2022).

Tous ces enjeux doivent être pris en compte lors de l'élaboration de la méthodologie pour cette technique de séquençage et ils doivent être corrigés lors des traitements des données.

2.4 Synthèse du chapitre

En résumé, plusieurs méthodes ont été développées afin de séquencer l'ADN, mais aussi l'ARN. Les méthodes traditionnelles incluent l'utilisation d'enzymes qui clivent les régions d'intérêt dans l'ADN, mais elles ont pour désavantage de cliver des régions non spécifiques. Ensuite il y a la méthode de l'amplification par PCR qui permet de multiplier la quantité d'ADN (Ribarska *et al.*, 2022), mais elle a aussi comme désavantage d'amplifier des séquences contaminantes et de causer des biais dans les résultats (Okoniewski et Miller, 2006).

Il y a ensuite la méthode de séquençage de Sanger, qui utilise des nucléotides marqués par la fluorescence afin de trier et analyser les séquences par électrophorèse (Zhang *et al.*, 2021) avant de les aligner au génome de référence (Shaibu *et al.*, 2021). Ensuite, il y a sa variante appelée le séquençage «fusil de chasse» (Dewey *et al.*, 2012), qui peut être utilisée conjointement avec la méthode *de novo* (Chapman *et al.*, 2015) en alignant des séquences plus petites sur le génome de référence (Dewey *et al.*, 2012). Ces méthodes ont pour avantage d'être efficaces sur le plan quantitatif, mais c'est une méthode qui peut être très dispendieuse et elle requiert un génome de référence très complète (Kozińska *et al.*, 2019) (Bell *et al.*, 2021). Ensuite, il y a le séquençage à l'aide d'un standard quantitatif qui utilise un standard d'étalonnage à différentes concentrations en tant que contrôle positif. Il est utilisé lors du profilage du transcriptome et a comme avantage de détecter les variations biologiques (Kukurba et Montgomery, 2015).

Les méthodes de séquençage de nouvelle génération incluent le séquençage nano-

pore, qui permet de détecter les séquences d'ADN qui passent à travers les pores d'une surface ionisée. Elle a pour avantage de seulement avoir besoin d'une petite quantité d'ADN (Dewey *et al.*, 2012), mais elle a pour désavantage de risquer de comporter des biais si les séquences sont riches en guanine et en cytosine (Delahaye et Nicolas, 2021). Cependant, ce biais peut être réduit en utilisant des enzymes, tels que des exonucléases (Clarke *et al.*, 2009).

D'autres méthodes de séquençage ont été développées, telles que la méthode ATAC-seq, ainsi que sa variante scATAC-seq, qui utilise la transposase Tn5 afin de cliver les régions qui sont accessibles à la chromatine avant de les séquencer (Ji *et al.*, 2020). Enfin, il y a la méthode SPLiT-seq, qui utilise 4 milieux de réaction qui contiennent les codes-barres et les amorces (Rosenberg *et al.*, 2018).

Ensuite, il y a eu le développement de la méthode de séquençage RNA-seq afin de mieux étudier le transcriptome (Kukurba et Montgomery, 2015) en permettant de détecter les variations de l'expression des gènes ainsi que leurs variants qui contiennent des mutations génétiques (Dewey *et al.*, 2012). Elle utilise la transcriptase inverse afin d'amplifier et de convertir le brin d'ARN messager en brin d'ADN complémentaire lors de l'amplification par PCR avant d'être séquencée et alignée sur le génome de référence. Elle a pour avantage d'être très efficace lors du profilage du transcriptome, mais elle peut avoir des biais lorsqu'il y a différents types de cellules dans les tissus qui sont utilisés comme échantillon (Kukurba et Montgomery, 2015).

C'est pourquoi la méthode de séquençage scRNA-seq (Tang *et al.*, 2009) a été développée afin de pouvoir détecter les différentes populations de cellules dans les échantillons, ainsi que leur variation sur le plan de leur phénotype (Anaparthy *et al.*, 2019). Par contre, elle a comme désavantage d'être très coûteuse et il y a une perte de l'information sur le plan temporel des cellules. De plus, cette méthode est

très sensible aux variations techniques et biologiques. D'autres variantes ont aussi été développées, telles que la méthode snRNA-seq qui utilise seulement l'ARN contenu dans les noyaux des cellules (Jovic *et al.*, 2022).

Elle nécessite d'isoler chaque cellule à l'aide de méthodes utilisant la fluorescence (FACS), le magnétisme (MACS) ou de la microdissection afin d'isoler chaque cellule avant d'être encapsulée dans un micro-environnement sous forme de gouttelettes (Zeb *et al.*, 2019). Chacune de ces dernières contient des oligonucléotides qui sont utilisés comme codes-barres afin d'identifier chaque séquence avant de les amplifier par PCR (Anaparthy *et al.*, 2019). La préparation de la librairie de séquences inclut le séquençage de ces dernières avant d'être compilée dans un fichier FASTQ ou BCL. Ces séquences sont ensuite alignées sur le génome de référence à l'aide d'outils tels que Cell Ranger. Les fichiers générés par Cell Ranger peuvent ensuite être utilisés dans des analyses subséquentes qui incluent le prétraitement et le traitement des données, telles que la normalisation des résultats, ainsi qu'une réduction de la dimension sous une forme linéaire à l'aide de la méthode PCA. D'autres analyses comprennent la visualisation des différentes sous-populations de cellules sous la forme d'une projection UMAP ou t-SNE, ainsi que des analyses de l'expression différentielle des gènes afin de pouvoir annoter les différents types de cellules à l'aide d'un biologiste ou d'outils (Jovic *et al.*, 2022).

Enfin, plusieurs enjeux sont à considérer, tels que l'accès aux données, l'élimination des biais dus à l'effet de lot, le choix de la résolution et de la dimension, ainsi que les différences entre les populations de cellules lors des analyses de l'expression différentielle des gènes.

Chacun de ces outils a sa propre expertise sur le plan des types d'analyses et d'annotation qui seront expliqués dans le chapitre suivant.

CHAPITRE III

ÉTAT DE L'ART POUR LES OUTILS D'ANALYSE DE DONNÉES DE SÉQUENÇAGE DE L'ARN DE CELLULE UNIQUE

Le séquençage de l'ARN de cellule unique génère beaucoup de données et il requiert différentes méthodes d'analyses et d'outils afin de pouvoir les traiter et d'interpréter les résultats. Ce chapitre présentera les différents outils d'analyse de ce type de données, leurs forces et leurs faiblesses.

De nombreux outils ont été développés afin d'analyser les données provenant du séquençage de l'ARN de cellule unique et leur nombre est en constante évolution. Selon une étude réalisée par Zappia et Theis (2021), il y avait déjà 1059 outils qui ont été répertoriés dans leur site web de leur base de données *scRNA-tools.org*, site web qu'ils ont eux-mêmes développé afin de servir de référence (Zappia *et al.*, 2018a). Juste à la fin de cette année (décembre 2024), le nombre est déjà rendu à 1796 outils qui sont classés dans plus de 30 catégories d'analyses (Zappia *et al.*, 2018b).

3.1 Outils d'alignement des séquences sur le génome de référence

3.1.1 Cell Ranger

Cell Ranger a été développée par (Zheng *et al.*, 2017) de la compagnie 10x Genomics. C'est un logiciel dans un environnement Linux (10x Genomics, 2020a) qui permet d'analyser les données de séquençage faites par *Chromium Single Cell* et de générer des données sous la forme de fichiers FASTQ. Le logiciel va ensuite les quantifier, les filtrer et les aligner sur le génome de référence. Il offre 5 flux de travail d'analyses en fonction du type de fichiers bruts issues des séquençages de l'ARN de cellule unique (10x Genomics, 2020c).

Premièrement, si le séquençage a été effectué par la compagnie *Illumina*, ces derniers vont être dans le format BCL. Il est donc nécessaire d'utiliser *cellranger mkfastq*, qui est un encapsuleur de *bcl2fastq*. Il va démultiplexer les données brutes des fichiers BCL et de les convertir en fichiers FASTQ qui seront compatibles avec *cellranger count*. De plus, les fichiers qui ont été générés par 10x Genomics sont déjà sous le format FASTQ, donc cette étape peut aussi être optionnelle (10x Genomics, 2020c).

Ensuite, *cellranger count* va aligner ces fichiers au génome de référence (10x Genomics, 2020b). Il effectue aussi une étape de contrôle de la qualité en filtrant les données, une analyse de l'expression des gènes, ainsi que la quantification des codes-barres et UMI (10x Genomics, 2020c). Il va ensuite compiler le tout et les enregistrer dans des fichiers de sortie contenant une matrice, la liste des gènes ainsi que la liste des codes-barres. Il va aussi produire un rapport dans un fichier HTML. Ce fichier contient une estimation des regroupements de cellules, la quantification des codes-barres, les identifiants moléculaires uniques (UMI), une estimation de la qualité des lectures, ainsi qu'un fichier CLOUPE permettant de

visualiser les résultats des regroupements de cellules à l'aide de Loupe Browser. Il produit aussi d'autres fichiers qui peuvent être au format BAM, H5 ainsi que d'autres fichiers optionnels qui pourraient être utilisés dans d'autres flux de travail (10x Genomics, 2020b).

Ensuite, il y a *cellranger multi* qui permet d'analyser les données qui ont été multiplexées et il est particulièrement utile lorsque plusieurs échantillons ont été combinés lors de la préparation de la librairie. Dans le cas où il y aurait plusieurs échantillons et qu'il faudrait les combiner pour des analyses comparatives, *cellranger aggr* permet de combiner les données issues de *cellranger count* et *cellranger multi*. Il va ensuite faire la normalisation et produire les mêmes types de fichiers que ces derniers. Enfin, *cellranger reanalyze* peut être utilisé s'il est nécessaire de faire des analyses des résultats issues des flux de travail cités précédemment (10x Genomics, 2020c).

3.1.2 Plateforme nuagique de 10x Genomics

Dans la situation où il n'est pas possible d'avoir accès à un environnement Linux, ni de posséder les ressources computationnelles nécessaires ou d'avoir accès à des grappes de calculs dans d'autres plateformes nuagiques, il est possible de lancer les analyses avec Cell Ranger sur cette plateforme de Calcul Canada (l'Alliance). L'avantage de celle-ci est qu'elle contient toutes les versions de Cell Ranger et offre aussi une interface qui est interactive pour télécharger les données et lancer les analyses. Cependant, il n'offre pas d'options pour visualiser les résultats directement sur le site et il ne peut pas faire des analyses subséquentes de données de séquençage de l'ARN de cellule unique. De plus, ce service est payant afin de faire plus d'analyses et pour le stockage des données qu'il faudra ensuite télécharger (10x Genomics, 2023).

3.2 Flux de travail de traitement des données

Pour les méthodes de prétraitement et de traitement des données à partir des fichiers FASTQ générés par Cell Ranger, Seurat et Scanpy sont les outils les plus populaires (Zappia et Theis, 2021). Scanpy peut être lancé dans un environnement Python et comporte aussi un outil d'annotation appelé AnnData (Wolf *et al.*, 2018). Dans le cas de Seurat, il est codé en langage R (Hao *et al.*, 2021), mais il n'a pas d'outils d'annotation intégrée. Donc, les annotations doivent se faire manuellement à l'aide de la littérature (Hoffman *et al.*, 2023). De plus, il y a aussi Scater, qui est aussi codé en langage R et offre des étapes très similaires aux deux outils précédents, mais il est en partie codé en langage C++ afin d'optimiser sa performance computationnelle. De plus, il n'a pas de fonction d'annotation des types de cellule (McCarthy *et al.*, 2017).

3.2.1 Seurat

Développé par Hao et al. (2021) et étant disponible sur CRAN, Seurat est un des outils les plus populaires et il consiste en un flux de travail comportant plusieurs étapes de prétraitement et de traitement des données. De plus, il est compatible avec les fichiers générés par Cell Ranger.

Dans une vignette écrite par Hoffman et al., (2023) et décrivant toutes les étapes d'analyses avec Seurat, la première étape des analyses est de créer un objet Seurat à partir des fichiers générés par Cell Ranger. Ensuite, il y a plusieurs étapes de prétraitement des données. La première consiste à faire un contrôle de la qualité en retirant les cellules de mauvaise qualité, par exemple, celles qui ont une très faible expression de gènes, qui peuvent être des doublons dans les gouttelettes. Ces dernières peuvent aussi être vides lors de la quantification, ce qui pourrait biaiser les résultats. Il évalue aussi le nombre de cellules exprimant un certain

pourcentage de gènes mitochondriaux, de les retirer de l'échantillon et de garder les cellules exprimant un certain nombre de gènes uniques. Ensuite, il y a une étape de normalisation de l'expression des gènes dans chaque cellule par rapport à celles de la totalité des cellules. Les gènes ayant une haute variation au niveau de leur expression sont ensuite sélectionnés. Seurat applique ensuite une transformation linéaire de l'expression de chaque gène à une échelle de 0 à 1 et il fait ensuite une réduction de la dimension linéaire avec une analyse en composantes principales (PCA). Les résultats de cette réduction linéaire peuvent être visualisés à l'aide d'une carte de chaleur afin de voir quels gènes sont les plus exprimés et moins exprimées en fonction des cellules. Cette méthode permet de bien évaluer l'hétérogénéité du jeu de données. Il permet aussi de déterminer la dimension du jeu de données afin d'éliminer le bruit technique. Cette méthode peut aussi être supervisée et donc il est possible de choisir la dimension manuellement en se basant sur les résultats de l'analyse par PCA. La réduction de la dimension peut aussi se faire en utilisant des statistiques basées dans le modèle aléatoire nul ou il peut être heuristique. Ensuite vient l'étape des regroupements des cellules en se basant sur le modèle des voisins les plus proches afin de regrouper les cellules ayant une expression similaire de gènes en utilisant l'algorithme de Louvain afin de l'optimiser. L'étape suivante est la réduction non linéaire afin de visualiser les résultats en regroupant les cellules ayant une expression similaire de gènes sous forme de projection UMAP ou t-SNE. Ensuite, il y a l'étape de l'analyse de l'expression différentielle des gènes, qui permet d'identifier l'expression (positive) ou l'inhibition (négative) des gènes dans chaque regroupement de cellules et de les comparer avec les autres. Il permet ainsi de faire la sélection de gènes d'intérêt et de visualiser leur expression dans chaque regroupement de cellules par un diagramme en violon, par des diagrammes par points ou par une carte de chaleur. Il est aussi possible de visualiser le nombre de gènes les plus exprimés dans chaque regroupement de cellules et ces dernières pourront ensuite être annotées. Malheu-

reusement, Seurat ne permet pas d'identifier automatiquement les types de cellules ni de faire des analyses sur le plan temporel. Il est donc nécessaire d'utiliser la littérature afin d'identifier les gènes ou des outils d'identification pour ces types d'analyses (Hoffman *et al.*, 2023).

3.2.2 Monocle 3

Développé par Cao et al (2019), Monocle 3 est très similaire à Seurat en matière de flux de travail. L'objet Seurat généré par ce dernier peut être converti en objet Monocle afin de continuer les analyses avec cet outil (Cao *et al.*, 2020b). Il a la particularité de pouvoir faire des analyses de trajectoire de cellules, c'est-à-dire de suivre l'évolution des cellules sur une base temporelle. Dans le cas où les données n'ont pas été collectées en fonction du temps, il permet de faire une estimation sur le plan temporel de l'évolution des cellules ainsi que de l'expression de leurs gènes et de les regrouper selon une trajectoire grâce à une analyse de pseudotemps (Trapnell *et al.*, 2014).

3.3 Outils d'annotations des cellules

Pour les outils d'annotation des cellules extraites des tissus, les exemples les plus cités sont scCATCH, qui utilise sa propre base de données spécialisée appelée *CellMatch* (Shao *et al.*, 2020). Il y a aussi l'outil SCINA, qui utilise un algorithme semi-supervisé et qui prend en entrée une matrice de l'expression des gènes ainsi qu'une liste de gènes spécifique aux types de cellules (Zhang *et al.*, 2019). Enfin, il y a SingleR, qui offre aussi des options de regroupement de cellules en sous-populations (Aran *et al.*, 2019).

3.3.1 ScType

Développé par Ianevski et al. (2022), et étant disponible en langage R et sur le web (Ianevski, 2023b), il s'agit d'un logiciel d'identification de cellules de manière automatique en se basant sur la spécificité des gènes exprimés dans chaque cellule. Il peut prendre en entrée les fichiers de données brutes provenant du séquençage de l'ARN de cellule unique ou il peut aussi prendre des données qui ont déjà été prétraitées, par exemple avec Seurat. Il utilise sa propre base de données, qui est l'intégration des données provenant de *CellMarker* et de *Panglaodb*. Cet outil s'avère aussi utile pour distinguer les cellules normales des cellules de tumeur maligne (Ianevski *et al.*, 2022).

3.3.2 EasyCellType

Développé par Li et al. (2023) et étant accessible dans un package R et sur le web (Li, 2022b), il permet de faire des annotations des types de cellules dans les données de séquençage de l'ARN de cellule unique. Il utilise trois bases de données : *panglaodb*, *clustermole* et *cellmarker*. Il prend en entrée la liste des marqueurs et il utilise deux types de tests de statistiques. Le premier est une analyse d'enrichissement des ensembles de gènes (GSEA) qui donne une liste des cinq types de cellules les plus probables. Le deuxième type d'analyse est le test de Fisher, qui donne une seule estimation du type de cellule qui est le plus probable. Enfin, les auteurs recommandent la méthode GSEA ainsi que la base de données CellMarker en se basant sur leurs données de simulation de la précision des annotations (Li *et al.*, 2023).

3.3.3 Enrichr

Enrichr est une interface web qui permet d'analyser l'enrichissement des gènes dans les cellules (Kuleshov *et al.*, 2016) (Chen *et al.*, 2013) (Xie *et al.*, 2021). À l'époque de la publication, il y avait seulement 35 librairies, dont certaines étaient accessibles uniquement sur Enrichr et d'autres étaient accessibles dans d'autres plateformes (Chen *et al.*, 2013). Cependant, le nombre continue d'augmenter à mesure qu'ils publient (Kuleshov *et al.*, 2016). Il prend en entrée une liste des gènes et il fait des recherches dans toutes les bases de données. Ensuite, il classifie les différents termes présents dans les bases de données dans les 10 précictions les plus significatives et il offre une interface interactive afin de les visualiser sous forme de graphiques de tableaux (Chen *et al.*, 2013). Dans le cas des librairies, les gènes peuvent être associés à la transcription à des voies de signalisation, à des ontologies et même à des types de cellules (Kuleshov *et al.*, 2016). Une version d'Enrichr disponible en langage R a aussi récemment été développée (Jawaid, 2023a).

3.3.4 scMayoMap

Développé par Yang et al., (2023b), ScMayoMap est un outil disponible sous un module R (Yang *et al.*, 2023a) et il permet l'annotation de cellules à partir de sa propre base de données appelée *scMayoMapDatabase*. Celle-ci contient les annotations provenant de la littérature et des bases de données existantes et il est possible de l'intégrer comme référence dans d'autres outils. Son algorithme tient compte des gènes qui sont fortement exprimés dans certains types de cellules en excluant les autres types de cellules qui ne sont pas de la même famille. Sa base de données contient les données pour l'être humain et la souris et contient 340 types de cellules qui ont été prélevés dans 28 tissus différents. Par contre, les données

provenant de tissus atteints du cancer ou qui sont inconnus ne sont pas présentes dans la base de données (Yang *et al.*, 2023b).

3.4 Outils offrant une interface interactive pour les analyses de données

3.4.1 Loupe Browser

Loupe Browser est un logiciel de visualisation de données du fichier CLOUPE ayant été généré par Cell Ranger. Il permet de faire une analyse de l'expression des gènes, de faire des regroupements de cellules, de sélectionner les cellules exprimant les gènes d'intérêt et de refaire des regroupements de cellules (10x Genomics, 2020d).

L'avantage de ce logiciel est qu'il comporte une interface interactive qui est adaptée à l'utilisateur, mais son principal désavantage est qu'il ne permet pas de faire des analyses plus poussées sur les données de séquençage de l'ARN de cellule unique. Il permet seulement de produire et d'exporter un fichier CSV contenant les coordonnées des regroupements de cellules dans les projections UMAP et t-SNE (10x Genomics, 2021b).

3.4.2 Azimuth

Pour les flux de travail interactifs, le plus populaire est Azimuth, qui est disponible sur le web et qui utilise l'algorithme de Seurat. Il possède une interface qui est interactive et qui permet de faire les mêmes analyses que Seurat. Il a aussi une base de données intégrée pour l'annotation des cellules, mais il est limité pour une sélection de tissus humains (Hao *et al.*, 2021).

3.4.3 Bases de données d'identification des types de cellules

Dans le cas des bases de données, les plus populaires pour l'annotation des cellules sont CellMarker (Zhang *et al.*, 2019) (ainsi que CellMarker2.0 (Hu *et al.*, 2023a)) et PanglaoDB (Franzén *et al.*, 2019b), qui offrent des annotations qui sont basées sur la littérature.

3.5 Synthèse du chapitre

En résumé, plusieurs outils ont été développés afin d'analyser les données de séquençage de cellule unique. Ils possèdent tous leur spécialité, leurs forces et leurs faiblesses. Premièrement il y a les outils d'alignement des séquences sur le génome de référence, tels que Cell Ranger (10x Genomics, 2020c). Cet outil est aussi disponible sur la plateforme nuagique de 10x Genomics ainsi que sur Calcul Canada (l'Alliance) (Alliance, 2023a).

La plateforme nuagique de 10x Genomic permet de faire les analyses avec toutes les versions de Cell Ranger. Il a pour avantage d'offrir une interface interactive et de pouvoir utiliser les ressources computationnelles de la plateforme à distance. Par contre, il requiert un abonnement afin de pouvoir faire des analyses multiples et de pouvoir stocker les données de manière prolongée (10x Genomics, 2023). La plateforme de Calcul Canada offre aussi l'option d'utiliser Cell Ranger à distance, mais elle n'offre pas d'interface interactive et elle nécessite d'être membre ou d'être parrainé afin de pouvoir utiliser les ressources.

Cell Ranger prend en entrée les fichiers FASTQ contenant les séquences (10x Genomics, 2020c) et il les aligne sur le génome de référence (10x Genomics, 2020b). Il génère ensuite des fichiers en sortie, tels que le rapport d'analyses dans le format HTML qui peut ensuite être visualisé dans un navigateur web, ainsi qu'un

fichier CLOUPE qui peut être visualisé avec l'outil Loupe Browser (10x Genomics, 2020b). Ce dernier offre une interface interactive, permet de faire des regroupements de cellules sous forme de projection UMAP ou t-SNE, ainsi que des analyses de l'expression différentielle des gènes. Par contre, seulement les coordonnées de ces projections sont exportables dans un fichier CSV.

Cell Ranger génère aussi d'autres fichiers, tels que les matrices, la liste des gènes exprimés ainsi que la liste des codes-barres. Ces fichiers vont ensuite être utilisés par des outils d'analyses de données, tels que Seurat et Monocle 3 (10x Genomics, 2020b).

Seurat est l'un des outils les plus utilisés pour le prétraitement et le traitement des données provenant de Cell Ranger. Son flux de travail inclut un contrôle de la qualité, une normalisation, une réduction de la dimension par la méthode PCA ainsi qu'un regroupement des différentes sous-populations de cellules sous forme d'une projection UMAP ou t-SNE. Cet outil peut aussi faire une analyse de l'expression différentielle des gènes, mais elle n'inclut pas d'outil ou de base de données d'identification de type de cellules (Hoffman *et al.*, 2023). Seurat est aussi disponible sur Azimuth, qui a pour avantage d'offrir une interface interactive, mais ces outils d'annotation sont limités à certains tissus humains (Hao *et al.*, 2021).

Monocle 3 permet de faire des analyses qui sont très similaires à Seurat, mais il permet aussi de faire des analyses de la trajectoire et du pseudotemps afin de suivre l'évolution des cellules de manière temporelle. Il est aussi compatible avec les données provenant de Seurat, mais il recrée une conversion de l'objet Seurat en objet Monocle avant de pouvoir les utiliser. Par contre, il n'offre pas d'option d'annoter les types de cellules (Trapnell *et al.*, 2014).

Enfin, les outils d'annotation des types de cellules incluent ScType (Ianevski *et al.*, 2022), EasyCellType (Li *et al.*, 2023), Enrichr (Kuleshov *et al.*, 2016) (Chen *et al.*,

2013) (Xie *et al.*, 2021) et scMayoMap (Yang *et al.*, 2023b). Les trois premiers outils ont aussi une version interactive sur leur site web et utilisent chacun des bases de données. Sctype et scMayoMap utilisent leurs propres bases de données (Ianevski *et al.*, 2022) (Yang *et al.*, 2023b), tandis que EasyCelltype utilise les bases de données *panglaodb*, *clustermole* et *cellmarker* (Li *et al.*, 2023) et Enrichr offre l'option d'utiliser des centaines de bases de données différentes (Kuleshov *et al.*, 2016) (Chen *et al.*, 2013) (Xie *et al.*, 2021).

Enfin, en ce qui concerne les bases de données, il y a CellMarker (Zhang *et al.*, 2019) (Hu *et al.*, 2023a) et PanglaoDB (Franzén *et al.*, 2019b) qui sont les plus utilisés.

Tous ces outils sont spécialisés dans leurs domaines et chacun peut seulement faire une portion différente des analyses de données provenant de séquençage d'ARN de cellule unique. De plus, chacun d'entre eux requiert des transferts et des conversions de données en entrée et en sortie. Puisqu'aucun d'entre eux n'est des outils d'analyses à part entière de manière automatisée, il a été nécessaire de développer un flux de travail qui inclut tout les étapes de ces outils qui seront expliquées dans le chapitre suivant.

CHAPITRE IV

MÉTHODE PROPOSÉE

Le chapitre précédent traitait des différentes méthodes d'analyses de données de séquençage de cellule unique, ainsi que des outils qui étaient spécialisés à chacune des étapes. Cependant, aucun outil ne permettait de faire toutes les analyses de manière automatisée sous la forme d'un flux de travail.

4.1 Problématique

4.1.1 Limitations des outils

Jovic et al. (2022) mentionnaient le besoin que les flux de travail automatisés aient une interface qui soit accessible pour les personnes n'ayant pas de compétences en programmation. Cependant, ceux cités dans le chapitre précédent proposaient tous des vignettes qui requéraient un certain niveau en programmation pour adapter les codes en fonction des besoins de l'utilisateur. Des exemples étaient les vignettes en langage Python pour Scanpy (Wolf *et al.*, 2018). Pour le langage R, il y avait les vignettes de Seurat (Hoffman *et al.*, 2023), Monocle 3 (Trapnell, 2022) (Cao *et al.*, 2020b), ScType (Ianevski, 2023b), EasyCellType (Li, 2022b), Enrichr (Jawaid, 2023a), et ScMayomap (Yang *et al.*, 2023a). Dans le cas de Cell Ranger, les vignettes étaient en langage Bash (10x Genomics, 2020a).

Nous avons testé ces outils et nous avons constaté que, même si chacun avait ses forces et ses faiblesses, ils étaient trop spécialisés dans leurs domaines et ils manquaient certaines étapes essentielles des analyses de données afin d'être considérés comme des outils complètement automatisés (tableau 4.1).

Dans le cas de Cell Ranger, il pouvait seulement être installé et lancé dans un environnement Linux et requérait beaucoup de ressources computationnelles (10x Genomics, 2020a). Pour la plupart des utilisateurs, cela les obligerait à le lancer soit dans Calcul Canada (Alliance, 2023a) ou avec la plateforme nuagique de 10x Genomics (10x Genomics, 2023), qui n'offraient pas d'option supplémentaire afin de télécharger automatiquement les résultats ou de les visualiser directement sur leur plateforme.

Même si certains outils offraient une interface interactive sur un site web (Cell Ranger dans la plateforme infonuagique de 10x Genomics (10x Genomics, 2023), ScType (Ianevski *et al.*, 2022), EasyCellType (Li *et al.*, 2023), Enrichr (Chen *et al.*, 2013)) et Seurat (via Azimuth (Hao *et al.*, 2021)), cela nous obligeait à jongler entre les plateformes et à convertir les données en entrée pour afin de lancer les analyses.

Enfin, même si Seurat et Monocle 3 permettaient de faire une grande partie des analyses de données à partir des résultats de Cell Ranger, ils n'incluaient pas d'outils d'annotation des types de cellules. Il était donc nécessaire de demander l'avis d'un biologiste ou d'utiliser des outils complémentaires qui sont compatibles avec leurs données afin de compléter les analyses.

4.2 Présentation du flux de travail

C'est avec cette problématique que nous avons décidé d'implémenter un flux de travail qui serait entièrement automatisé en y intégrant les outils mentionnés ci-

dessus afin de réduire le nombre de manipulations de données requises entre chaque analyse. Ce flux de travail peut faire toutes ces étapes d'analyses offertes par ces outils de manière automatisée. De plus, nous avons aussi choisi de l'implémenter d'une façon à les utiliser individuellement ou en partie dans l'éventualité de refaire des analyses avec des outils spécifiques. Notre principal objectif lors de la conception de ce flux de travail était de minimiser la nécessité de l'utilisateur à ajouter des lignes de codes supplémentaires directement dans le code des outils, en rassemblant tous les paramètres en un seul endroit au début du code afin que l'utilisateur puisse les modifier plus facilement et rapidement (figures 4.1, 4.2 et 4.3).

4.3 Méthodologie

4.3.1 Choix du flux de travail

La première étape lors de la conception était de trouver un format de flux de travail qui serait approprié afin d'y intégrer les outils. Nextflow (Di Tommaso *et al.*, 2017) et Snakemake (Köster et Rahmann, 2012) ont été considérés, mais nous avons favorisé un autre flux de travail appelé *Script of Script* (Peng *et al.*, 2018) (Wang et Peng, 2019). Ce dernier offrait la possibilité d'implémenter le code dans un bloc-notes Jupyter, ce qui comportait plus d'originalité que les deux premiers flux de travail. Selon les auteurs (Wang et Peng, 2019), le langage de programmation associé aux types de flux de travail mentionnés précédemment n'est pas très efficace. En effet, ils mentionnent que l'adaptation des codes ainsi que les corrections à apporter en cas de problèmes ne sont pas très productives en matière de temps et d'effort (Wang et Peng, 2019). Ils ont donc conçu ce type de flux de travail, qui intègre des fonctionnalités assez similaires à ces deux flux de travail, tout en offrant une interface beaucoup plus simple, notamment en l'intégrant au

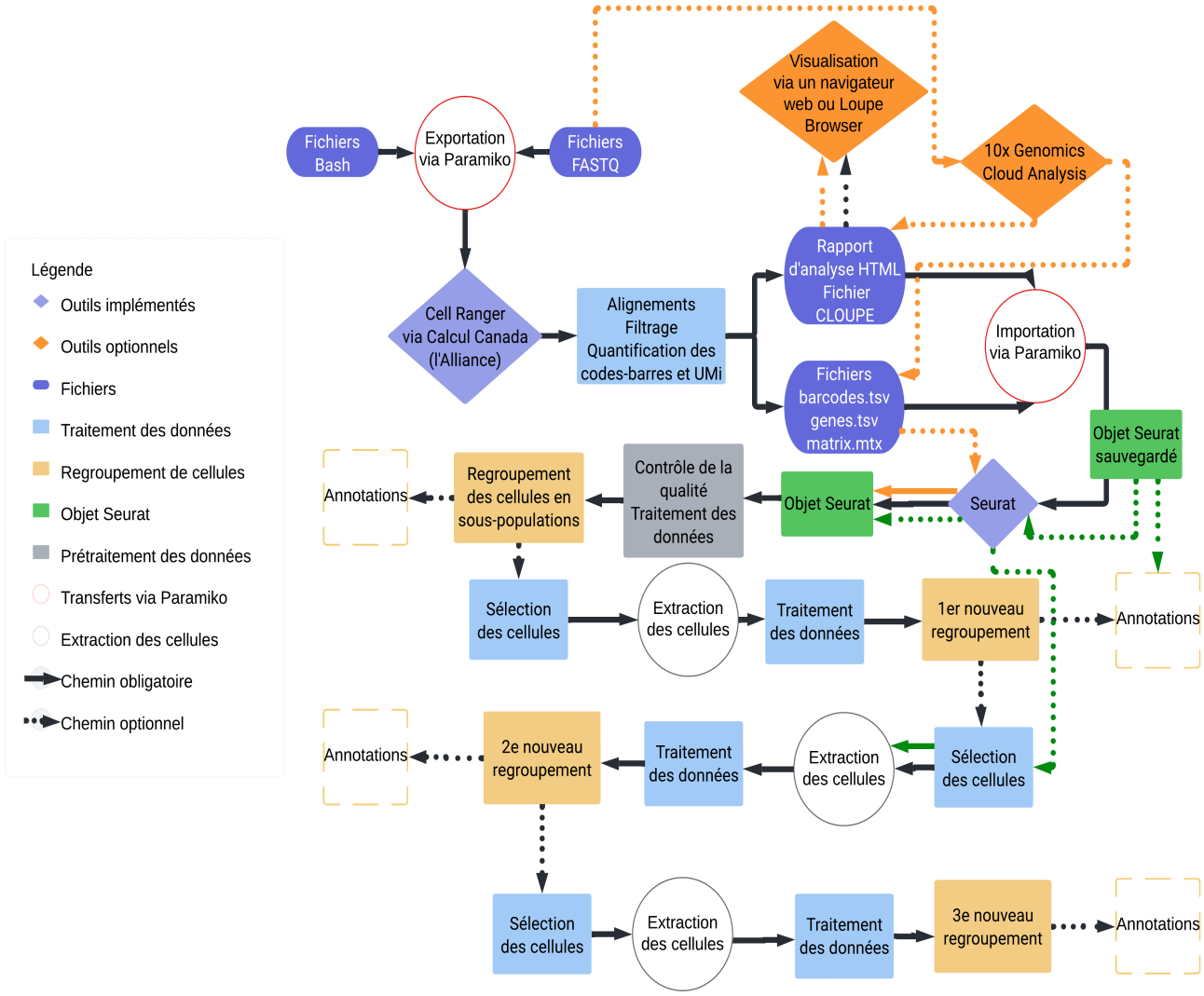

FIGURE 4.1 – Méthodologie et étapes du flux de travail

Les flèches ayant un trait continu représentent les chemins obligatoires pour une utilisation complète du flux de travail. Les flèches ayant un trait pointillé représentent les chemins optionnels pour d'outrepasser certaines étapes.

bloc-notes Jupyter (Peng *et al.*, 2018) (Kluyver *et al.*, 2016). De plus, ce type de flux de travail permet de coder en différent langage de programmation localement (Wang et Peng, 2019) en modifiant le type de noyau dans chaque cellule. De plus, il a la fonctionnalité de pouvoir échanger les données ainsi que les variables

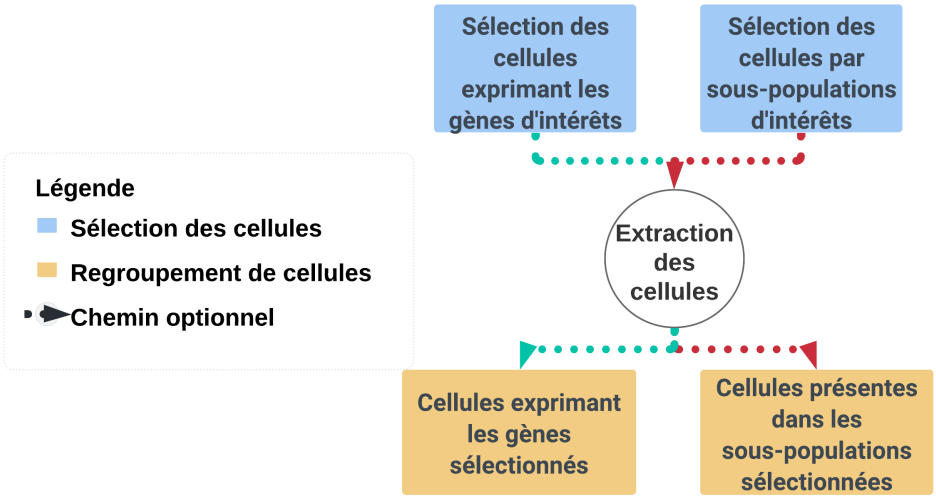

FIGURE 4.2 – Méthodes de sélection des cellules à l'aide du flux de travail. Représentation des deux possibilités pour extraire les cellules avant de continuer les étapes de traitement des données ainsi que les annotations. Ces extractions sont optionnelles dans le flux de travail.

entre ces cellules plus facilement, (Peng et of Texas MD Anderson Cancer Center, 2018b) (Peng *et al.*, 2018), ce qui est plus avantageux pour notre approche. De plus, le fait de pouvoir rassembler tous les codes des outils dans un seul fichier de flux de travail dans le bloc-notes Jupyter permet de les compartimenter dans des cellules séparées et de pouvoir les lancer individuellement ou successivement. Cela facilite grandement la maintenance, ainsi que la résolution des problèmes (Wang et Peng, 2019). Afin de pouvoir mieux déplacer les outils dans le flux de travail, nous avons intégré les codes des outils dans des fonctions qui prennent en entrée les paramètres définis par l'utilisateur. En effet, le fait de transformer les codes des outils en y ajoutant des paramètres et de rassembler ces derniers au début du code simplifie l'utilisation du code du flux de travail, car l'utilisateur aura seulement besoin de modifier ces paramètres au début du code au lieu de les modifier directement dans le code. Cela permet de rendre le flux de travail plus flexible et réutilisable (Shevlin, 2020). La liste des modules associée aux outils,

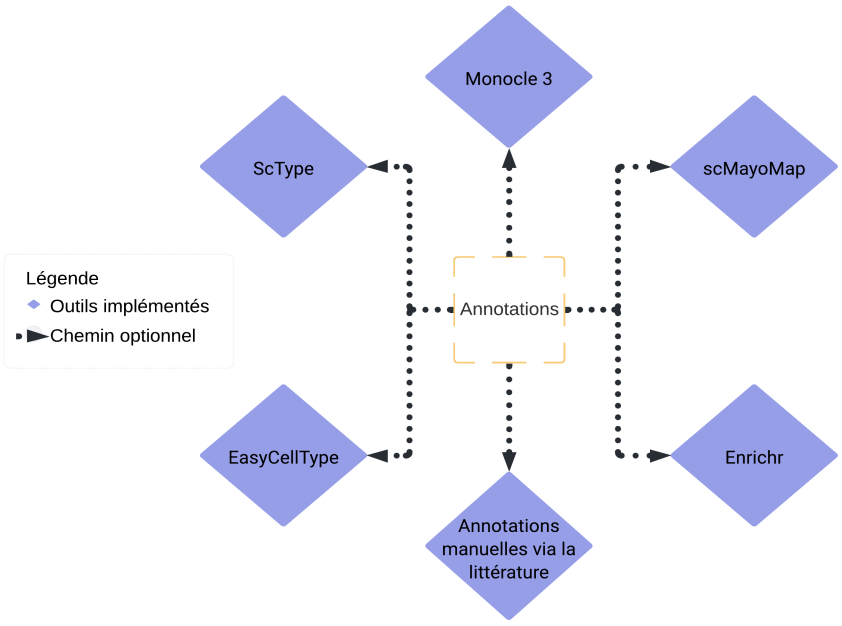

FIGURE 4.3 – Outils implémentés dans le flux de travail.
Tous les outils permettent de faire les annotations des différents regroupements de cellules. Ces annotations sont optionnelles dans le flux de travail.

ainsi que leur version et leurs références sont disponibles en annexe (tableau A.1).

4.3.2 Intégration de Cell Ranger dans le flux de travail.

La deuxième étape était de trouver une façon d'intégrer Cell Ranger dans ce flux de travail, tout en évitant son installation dans une machine ou un serveur. De plus, il fallait trouver une façon de pouvoir communiquer avec l'environnement où il effectuerait les analyses. Nous avons décidé d'utiliser la plateforme de Calcul Canada (l'Alliance), car il contient déjà l'outil Cell Ranger sur sa plateforme et qu'il était possible de le lancer à distance via SLURM (Alliance, 2023a). Afin de pouvoir établir la communication entre la machine locale et la plateforme, nous avons intégré le module Paramiko, qui instaurerait une connexion sécurisée entre la machine locale et la plateforme. Il peut même être utilisé avec un autre

serveur à distance qui contiendrait Cell Ranger (Forcier, 2023b). Cet outil prend en paramètre les informations, telles que le nom de l'utilisateur, le numéro de port, ainsi que la clé SSH (Forcier, 2023a). Tous ces paramètres ont été ajoutés dans la liste des paramètres à déterminer dans le flux de travail (figure 4.6). Afin de pouvoir communiquer avec SLURM pour de voir le statut des travaux, des fonctions ont été implémentées (Alliance, 2023a). La figure 4.4 explique toutes les étapes ainsi que les fonctions implémentées dans le flux de travail.

Essentiellement, ces fonctions permettent de lancer les analyses de Cell Ranger via SLURM, de vérifier le statut des analyses en cours ou en attente, les détails de celles-ci, de vérifier si ces analyses sont terminées et de les annuler au besoin en utilisant comme paramètre le numéro du travail sur SLURM. (Alliance, 2023a) (Forcier, 2023a). De plus, nous avons intégré des fonctions qui permettent d'exporter les fichiers FASTQ sur la plateforme. Nous avons aussi intégré une fonction permettant de créer le fichier SLURM avec les paramètres établis par l'utilisateur et qui sera automatiquement importé avec les fichiers FASTQ, ainsi qu'une fonction qui va exporter les fichiers des résultats de Cell Ranger, qui s'avèrent essentiels pour le flux de travail lorsque les analyses seront terminées (Forcier, 2023a).

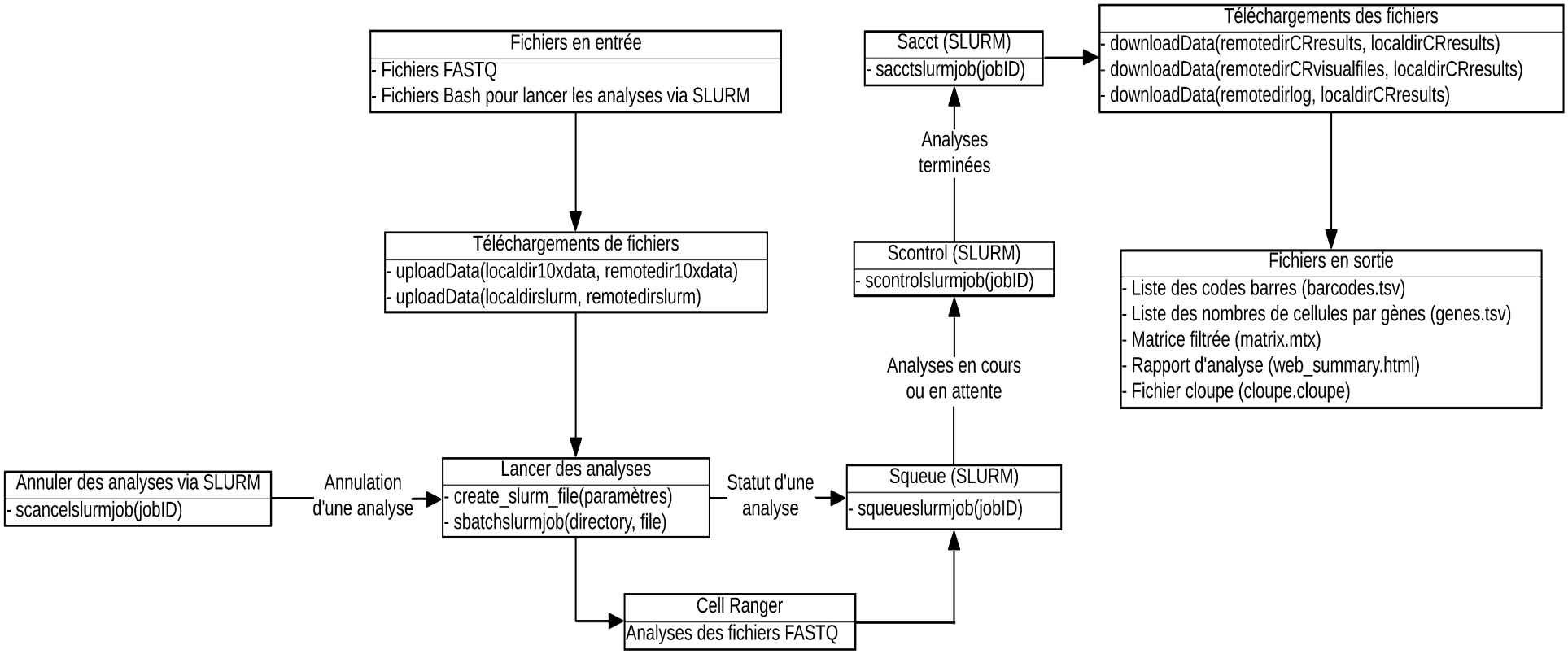

FIGURE 4.4 – Fonctions implémentées dans le flux de travail pour les analyses avec Cell Ranger via Paramiko et SLURM.

Ce flux de travail inclut toutes les étapes pour les analyses avec Cell Ranger. Toutes les fonctions ont été écrites avec le langage de programmation Python.

4.3.3 Intégration de Seurat dans le flux de travail

Pour la troisième étape, nous avons décidé d'implémenter l'outil Seurat (Hao *et al.*, 2021) dans le flux de travail, car cet outil est l'un des plus utilisés dans le langage R (Zappia *et al.*, 2018a), mais aussi parce que les autres outils que nous avons implémentés étaient compatibles avec ce dernier (Cao *et al.*, 2020b) (Ianevski, 2023b) (Li, 2022b) (Yang *et al.*, 2023a) (Jawaid, 2023a).

Essentiellement, le flux de travail de Seurat a été divisé en plusieurs parties afin de mieux faciliter sa reproductibilité. Le début consiste à initialiser le flux de travail en chargeant tous les modules nécessaires (tableau A.1 en annexe). De plus, l'allocation de la mémoire est augmentée pour le stockage des données (R Core Team, 2021b) et le thème de la mise en forme des figures est chargé (Wickham, 2016).

Ensuite, les fichiers générés par Cell Ranger sont lus et chargés dans un objet Seurat. Ces fichiers peuvent aussi provenir de la plateforme nuagique de 10x Genomics (10x Genomics, 2023) si l'utilisateur n'a pas accès à Cell Ranger sur Calcul Canada (l'Alliance) (Alliance, 2023a). Ensuite, il y a les étapes de prétraitement, de traitements de données et des regroupements par l'outil Seurat (Hoffman *et al.*, 2023). Ces étapes peuvent être optionnelles, car nous avons aussi implémenté la possibilité d'utiliser un objet Seurat déjà sauvegardé qui proviendrait d'analyses antérieures. Par la suite, des figures de la projection UMAP des regroupements des cellules et des tableaux contenant la liste des gènes sont produits et sauvegardés (Hoffman *et al.*, 2023). Puis, il est possible d'extraire les cellules d'intérêts à partir des marqueurs exprimés ou du numéro des regroupements des cellules et de refaire les analyses de traitement des données par Seurat (figure 4.5).

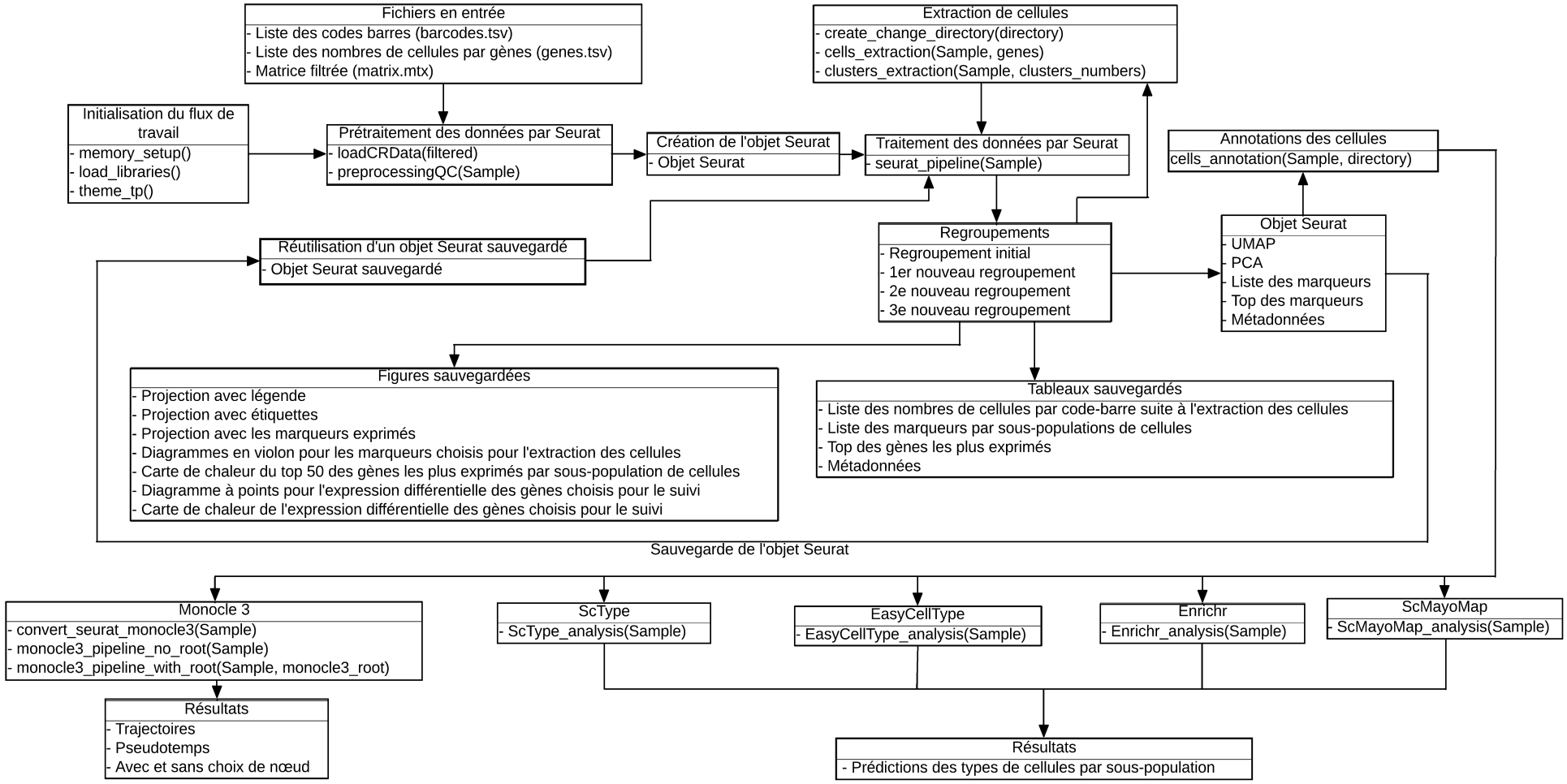

FIGURE 4.5 – Fonctions implémentées dans le flux de travail pour les analyses avec Seurat et les outils d'annotations.

Ce flux de travail inclut toutes les étapes pour les analyses avec Seurat ainsi que les outils d'annotation Monocle 3, ScType, EasyCellType, Enrichr et scMayoMap. Toutes les fonctions ont été écrites avec le langage de programmation R.

4.3.4 Intégration d'outils d'annotations dans le flux de travail

Enfin, la dernière étape était d'annoter les regroupements avec les outils d'annotations de cellules (ScType (Ianevski, 2023b), EasyCellType (Li *et al.*, 2023), Enrichr (Jawaid, 2023a) et ScMayomap (Yang et Zhang, 2023), Enrichr (Jawaid, 2023a)) et Monocle 3 (Cao *et al.*, 2019) pour les analyses de la trajectoire et du pseudotemps. La figure 4.5 explique toutes les étapes ainsi que les fonctions implémentées dans le flux de travail.

4.4 Fonctionnement du flux de travail

Le flux de travail débute par une liste des paramètres à définir par l'utilisateur (figure 4.6). Il inclut notamment l'option d'activer la partie associée aux analyses avec Cell Ranger sur Calcul Canada (l'Alliance) via Paramiko, ou l'option de le désactiver pour lancer les analyses via Seurat. Il peut aussi choisir de simplement lancer les annotations des cellules et du pseudotemps avec les autres outils implémentés. La liste des paramètres se trouve dans la figure 4.6.

Chaque choix d'activer une analyse ou non se fait par un booléen ('True' ou 'False'), suivi par les paramètres associés à l'outil. Ces paramètres peuvent être, par exemple, une liste de gènes, le choix des tissus et les bases de données à utiliser. Elles peuvent aussi être des valeurs, par exemple, le numéro de l'analyse via SLURM ou elles peut être des emplacements des fichiers nécessaires pour lancer les analyses (figure 4.6). Chaque outil va générer des résultats sous forme de fichiers, tels que des figures et des tableaux qui vont automatiquement être sauvegardés dans un dossier associé à l'outil.

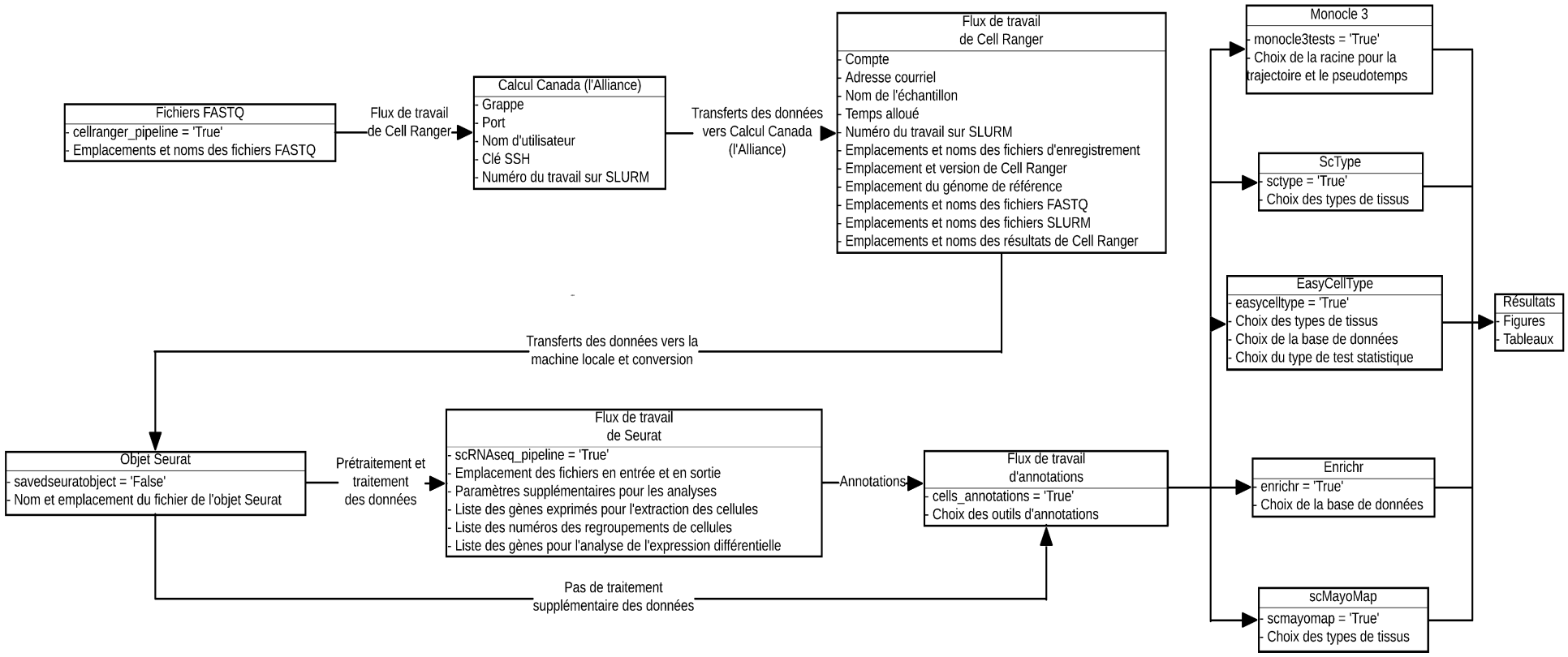

FIGURE 4.6 – Paramètres à définir par l'utilisateur dans le flux de travail.

Chaque outil a ses propres paramètres à définir par l'utilisateur. Ce dernier peut aussi choisir d'utiliser les outils à l'aide de booléens 'True' pour activer ou 'False' pour désactiver les analyses faites par les outils.

4.5 Paramètres à utiliser dans le flux de travail

Les flux de travail de Cell Ranger, de Seurat et des annotations peuvent être exécutés indépendamment et ils ont chacun des paramètres qui leur sont propres. Tous les paramètres à définir sont présents au début du code et ils peuvent tous être modifiés avant l'exécution du flux de travail. Enfin, il y a aussi des paramètres qui permettent d'exécuter les flux de travail individuellement ou successivement en fonction des besoins de l'utilisateur.

4.5.1 Flux de travail de Cell Ranger

Certains paramètres nécessitent des booléens, tels que le choix d'activer ou de désactiver ce flux de travail. De plus, certains d'entre eux permettent de télécharger ou non les fichiers en entrée pour lancer les analyses ainsi que les fichiers en sortie qui contiennent les résultats. Enfin, il y a des paramètres qui permettent de consulter le statut ainsi que le numéro du travail des analyses sur SLURM. Ce numéro est aussi inclus en tant que paramètre afin d'annuler les analyses au besoin (Alliance, 2023a).

Afin de pouvoir se connecter avec Paramiko (Forcier, 2023a), il est aussi nécessaire de définir les paramètres tels que le nom de la grappe sur la plateforme de Calcul Canada (l'Alliance) (Alliance, 2023a), le numéro du port pour la connexion, le nom de l'utilisateur, ainsi que le nom et l'emplacement du fichier contenant la clé privée pour la connexion SSH (Forcier, 2023b).

Afin de pouvoir lancer ce flux de travail, il est nécessaire de définir les noms ainsi que les emplacements des fichiers FASTQ contenant les données brutes du séquençage de l'ARN à cellule unique ainsi que les fichiers Bash contenant les instructions afin de lancer les analyses via SLURM. Pour ce dernier, des para-

mètres doivent aussi être défini tels que le nom du compte utilisant les ressources computationnelles sur la plateforme, l'adresse courriel de l'utilisateur afin de recevoir les statuts des analyses, le nom des échantillons, le temps alloué pour faire les analyses, l'emplacement du logiciel de Cell Ranger ainsi que le génome de référence. Ces informations permettent au flux de travail de repérer les fichiers nécessaires à son fonctionnement, de les télécharger vers la plateforme de Calcul Canada (l'Alliance) (Alliance, 2023a) et de pouvoir lancer les analyses. Il est aussi nécessaire de définir le paramètre de l'emplacement d'un dossier sur la machine locale de l'utilisateur afin de télécharger les résultats qui ont été générés par le flux de travail. Les paramètres pour les noms des fichiers à utiliser pour le flux de travail de Seurat sont déjà définis par défaut, mais ils peuvent être modifiés au besoin.

Les figures 4.4 et 4.6 illustrent tous les paramètres implémentés dans ce flux de travail.

4.5.2 Flux de travail de Seurat

Ce flux de travail peut être activé ou désactivé grâce à un booléen. De plus, ce type de paramètre permet aussi de choisir de faire un prétraitement des données et de traiter les données. Enfin, si des analyses avec Seurat ont déjà été faites, il est aussi possible d'utiliser un objet Seurat ayant été généré préalablement (Hoffman *et al.*, 2023). Cette fonctionnalité est aussi utile afin de lancer directement les analyses des annotations sans devoir passer par les flux de travail cités précédemment.

D'autres paramètres à définir sont le choix de l'organisme associé au génome utilisé, le nombre pour les gènes les plus exprimés, le choix des dimensions pour les projections, la liste des gènes pour la visualisation des résultats sous forme de diagrammes à points et en violon (Hoffman *et al.*, 2023). Cette liste de gènes peut

aussi être utilisée afin d'isoler les cellules d'intérêts exprimant ces gènes et de les regrouper lors des analyses subséquentes.

Enfin, il y a le choix entre deux méthodes de regroupements des cellules (Satija et Lab, 2023). La première possibilité est de choisir les noms des gènes à utiliser pour isoler les cellules d'intérêt avant de faire les regroupements. Cette méthode est très utile si les cellules exprimant les gènes d'intérêts sont dispersées sur plusieurs regroupements et qu'elles sont mélangées avec des cellules à éliminer. L'autre possibilité est de choisir les numéros de regroupements pour extraire et isoler les cellules. Cette méthode est pertinente si les cellules d'intérêts sont déjà bien isolées dans des regroupements plus spécifiques et qu'elles y sont très concentrées dans ces derniers. Ces regroupements de cellules pourront ensuite être annotés avec le flux de travail des annotations.

Les figures 4.5 et 4.6 illustrent tous les paramètres implémentés dans ce flux de travail.

4.5.3 Flux de travail des annotations

Les paramètres nécessitant un booléen permettent de choisir quels outils d'annotations à utiliser pour les regroupements des cellules. Pour chacun des outils, il est possible de choisir les types de tissus (pour ScType (Ianevski, 2023b), EasyCellType (Li, 2022a) et ScMayoMap (Yang *et al.*, 2023a)) ainsi que les bases de données (pour EasyCelType (Li, 2022a) et Enrichr (Jawaid, 2023b)) à utiliser pour les analyses. Pour EasyCellType, il est aussi possible de choisir le type de test (GSEA ou Fisher) pour les analyses (Li, 2022a).

Pour Monocle 3, il est possible de choisir le point de départ (la racine) afin de tracer la trajectoire (Trapnell, 2022) (Cao *et al.*, 2020b). Chaque outil va récupérer les résultats produits par le flux de travail de Seurat tels que les projections des

regroupements de cellules et les tableaux des gènes exprimés, et il va générer ses propres résultats pour les annotations.

Les figures 4.5 et 4.6 illustrent tous les paramètres implémentés dans ce flux de travail.

4.5.4 Recommandations pour le bon fonctionnement des flux de travail

Il est très important que tous les paramètres pour chacun des flux de travail soient bien définis avant de lancer les analyses, en particulier si ces flux de travail sont lancés de manière successive. Il est recommandé de lire au préalable les différentes documentations des outils qui sont mentionnées dans ce mémoire et d'utiliser les exemples dans les codes fournis lors des premiers essais des analyses afin de bien comprendre quels types de paramètres à utiliser. De plus, il est recommandé de lancer les flux de travail séparément, en particulier lors des premières analyses et de seulement les utiliser de manière successive lors des répétitions des analyses avec des échantillons issus de la même expérience.

L'implémentation de différents outils dans les flux de travail a permis de générer des résultats qui sont complémentaires entre les outils afin de combler les lacunes de ces derniers. Cependant, certains outils exigeaient des données générées par d'autres outils qui n'étaient pas nécessairement compatibles ou visuellement interprétables. Nous avons dû implémenter des solutions pour combler ces lacunes en faisant des conversions pour les données et les résultats.

4.6 Solutions implémentées dans le flux de travail

Suite à nos constatations au sujet des lacunes de ces outils, nous avons implémenté des solutions à notre flux de travail. Par exemple, ces outils exigeaient d'être lancés

individuellement avec des données en entrée très spécifiques. Même si certains étaient déjà compatibles avec les données provenant de l'objet Seurat (par exemple ScType (Ianevski, 2023b) et scMayomap (Yang *et al.*, 2023a)), d'autres requéraient une extraction et une conversion de certaines données. Il était nécessaire d'effectuer une conversion de l'objet Seurat pour Monocle 3 (Cao *et al.*, 2020b). Des lignes de codes supplémentaires ont été ajoutées afin de faire la conversion.

De plus, le site web de EasyCellType exigeait un tableau contenant la liste des gènes ainsi que de leur expression (Li, 2022b) (Li *et al.*, 2023). L'interface en ligne de Enrichr exigeait aussi une liste de gènes (Chen *et al.*, 2013) (Jawaid, 2023a). Des lignes de codes supplémentaires ont été ajoutées pour extraire les données nécessaires pour chacun des outils.

De plus, certains outils, tels que Enrichr et les analyses de l'expression différentielle de Seurat, donnaient des résultats qui ne pouvaient pas être interprétés visuellement. La solution était de faire une extraction des tableaux des résultats et de les convertir en diagrammes à points afin de faciliter visuellement leur interprétation.

Dans le cas de Cell Ranger, si on le lançait sur la plateforme de Calcul Canada via SLURM, une des façons de savoir si les analyses étaient terminées était d'attendre le courriel attestant la fin de celles-ci. Il y avait aussi une autre façon qui était de se connecter directement sur la plateforme et de lancer des lignes de commandes via SLURM afin de voir leur statut (Alliance, 2023a). Cela voulait dire qu'il n'y avait pas de façon de vérifier le statut d'un travail sur SLURM et d'automatiquement télécharger les résultats de Cell Ranger. Nous avons implémenté des fonctions dans notre flux de travail permettant de vérifier le statut des analyses de Cell Ranger via SLURM. Lorsque le statut indique que les analyses sont terminées, notre flux de travail va automatiquement télécharger les résultats à l'aide de Paramiko. La figure 4.4 indique les fonctions implémentées afin de vérifier le statut d'une

analyse via SLURM et de télécharger automatiquement les données lorsqu'elles sont terminées.

Enfin, les codes des vignettes ne sauvegardaient pas automatiquement les résultats, ce qui nous obligeait aussi à écrire du code supplémentaire afin de les sauvegarder. La figure 4.7 explique les modifications apportées au flux de travail pour corriger les limitations des outils (tableau 4.1).

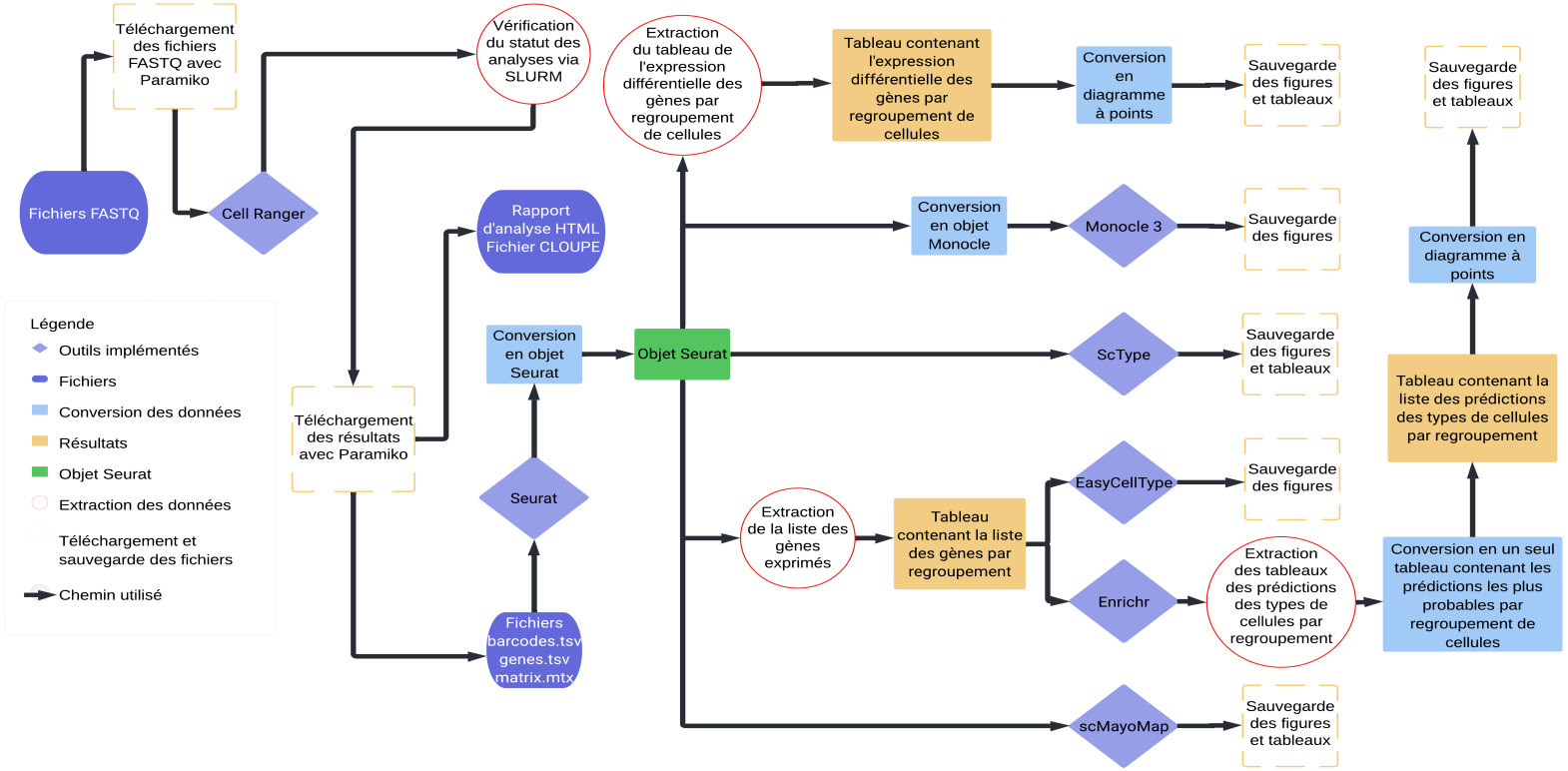

FIGURE 4.7 – Étapes ajoutées au flux de travail pour résoudre les limitations des outils.

Solutions implémentées dans le flux de travail pour corriger les lacunes des outils. Paramiko est utilisé pour les téléchargements des fichiers de Cell Ranger. SLURM est utilisé pour lancer les analyses avec Cell Ranger et pour vérifier leur statut sur la plateforme de Calcul Canada (l'Alliance). L'objet Seurat est généré à partir des résultats de Cell Ranger et il est compatible avec ScType et scMayoMap. Une conversion de l'objet Seurat est nécessaire pour Monocle 3. Une extraction de la liste des gènes exprimés est nécessaire pour EasyCellType et Enrichr. Les figures et les tableaux des résultats issus des outils sont ensuite sauvegardés. Une conversion des résultats en diagramme à points est nécessaire pour l'analyse de l'expression différentielle des gènes faite par Seurat ainsi que pour les résultats des annotations par Enrichr avant de faire la sauvegarde des résultats.

4.7 Synthèse du chapitre

En résumé, chacun des outils était disponible dans différents langages de programmation et certains d'entre eux n'offraient aussi une interface interactive pour les utilisateurs n'ayant pas de compétences en programmation ni les ressources computationnelles afin de lancer leurs analyses (Jovic *et al.*, 2022). De plus, chacun d'entre eux avait ses forces et ses faiblesses sur le plan des analyses.

Nous avons décidé d'utiliser le flux de travail Script of Script, car il offrait une interface plus interactive sous la forme d'un bloc-notes Jupyter (Peng *et al.*, 2018) (Wang et Peng, 2019), comparé aux autres flux de travail traditionnels, tels que Nextflow (Di Tommaso *et al.*, 2017) et Snakemake (Köster et Rahmann, 2012). De plus, il permettait de rassembler tous les codes des outils à un seul endroit et de modifier plus facilement. Enfin, notre méthodologie était de modifier les codes afin de déplacer tous les paramètres pertinents au début du code afin de faciliter leur modification et pour rendre le code du flux de travail plus reproductible.

Le fonctionnement de notre flux de travail se résume à une liste de paramètres qui sont associés aux outils que l'utilisateur peut modifier en fonction de ses besoins. De plus, il est aussi possible de lancer chacun des outils individuellement ou successivement de manière automatisée. De plus, notre flux de travail permet aussi à un utilisateur plus expérimenté de modifier le code des outils directement dans le bloc-notes Jupyter, ce qui facilite grandement la maintenance du code.

Dans le cas de Cell Ranger (10x Genomics, 2020c), puisqu'il requiert beaucoup de ressources computationnelles et n'offre pas d'interface interactive sur Calcul Canada (l'Alliance) (Alliance, 2023a), nous avons intégré des fonctions qui permettaient de vérifier le statut des analyses via SLURM et qui allaient télécharger automatiquement les données lorsqu'elles étaient terminées.

Pour Seurat (Hoffman *et al.*, 2023) (Hao *et al.*, 2021), il n'offrait pas d'option d'analyses de la trajectoire et du pseudotemps, donc nous avons intégré Monocle 3 (Trapnell *et al.*, 2014) pour pallier ce déficit. De plus, puisque Seurat ne permettait pas d'annoter les différentes sous-populations de cellules à l'aide de bases de données, nous avons intégré les outils ScType (Ianevski *et al.*, 2022), EasyCellType (Li *et al.*, 2023), Enrichr (Kuleshov *et al.*, 2016) (Chen *et al.*, 2013) (Xie *et al.*, 2021) et scMayoMap (Yang *et al.*, 2023b).

Tous ces outils n'enregistraient pas automatiquement les résultats, donc nous avons aussi ajouté des lignes de codes qui vont automatiquement les sauvegarder sous forme de figures et de tableaux.

De plus, certaines analyses, telles que l'expression différentielle des gènes, ainsi que les annotations d'Enrichr ne permettaient pas de bien visualiser les résultats, donc nous avons aussi ajouté des lignes de codes afin de convertir ces résultats en diagramme de points.

De plus, afin de pouvoir transférer les données entre les outils, par exemple entre Seurat, Monocle 3, ainsi que les outils d'annotations, nous avons aussi ajouté des lignes de codes pour faire les conversions des données entre les outils.

Enfin, chaque flux de travail peut être exécuté de manière indépendante ou successive grâce à des paramètres contenant des booléens. De plus, chacun d'entre eux a des paramètres qui leur sont propres et qui doivent être définis avant de lancer les analyses afin de pouvoir automatiquement chercher les informations nécessaires pour générer leurs résultats.

Afin de vérifier l'efficacité de notre flux de travail, nous l'avons testé sur des jeux de données qui ont été mentionnés dans le premier chapitre. La méthodologie que nous avons utilisée avec notre flux de travail afin d'analyser ces données sera

expliquée dans le chapitre suivant.

TABLEAU 4.1 – Limitations des outils intégrés au flux de travail.

	Seurat	**Monocle 3**	**ScType**	**EasyCellType**	**Enrichr**	**scMayoMap**	**Cell Ranger**
Analyse des données brutes du séquençage de l'ARN de cellule unique	Non	Non	Non	Non	Non	Non	Oui
Analyse des données provenant de Cell Ranger	Oui	Oui	Non	Non	Non	Non	NA
Compatible avec l'objet Seurat	Oui	Conversion	Oui	Non	Non	Oui	NA
Requiert une conversion des résultats de Seurat en entrée	NA	NA	Non	Tableaux	Tableaux	Non	NA

Prétraitement et traitement des données provenant de Cell Ranger	Oui	Oui	Non	Non	Non	Non	Loupe Browser
Regroupement des cellules	Oui	Oui	Non	Non	Non	Non	Rapport HTML, Loupe Browser
Analyse de l'expression différentielle des gènes	Oui	Oui	Non	Non	Non	Non	Loupe Browser
Analyse de la trajectoire et du pseudotemps	Non	Oui	Non	Non	Non	Non	Non
Bases de données d'identification des types de cellules	Azimuth	Non	Oui	Oui	Oui	Oui	Non
Interface interactive	Azimuth	Shiny	Version web	Version web	Version web	Non	Rapport HTML, Loupe Browser
Analyses à distance	Azimuth	Non	Version web	Version web	Version web	Non	Calcul Canada, 10x Genomics

Requiert une conversion des résultats en sortie	Diagramme à points	Non	Non	Non	Diagramme à points	Non	Non
Sauvegarde automatique des résultats en figures et en tableaux	Non	Non	Non	Non	Non	Non	Rapport HTML, Loupe Browser

Forces et faiblesses des outils qui ont été intégrés dans le flux de travail. Ces éléments ont été pris en considération lors du développement du flux de travail et les outils ont été implémentés de manière à ce qu'ils soient compatibles les uns par rapport aux autres et pour que leurs analyses se complémentent.

CHAPITRE V

APPLICATIONS DE LA MÉTHODE PROPOSÉE

5.1 Application du flux de travail sur des jeux de données

Le chapitre précédent traitait des outils essentiels afin de réaliser toutes les principales analyses de données de séquençage de cellule uniques de manière automatisée. Nous les avons intégrés à notre flux de travail et avons corrigé les problèmes et les déficits associés à ces outils. La prochaine étape était de vérifier l'efficacité de notre flux de travail sur des jeux de données.

Nous avons utilisé notre flux de travail sur des jeux de données provenant de Soret et al. (2021). Nous avons commencé par analyser les populations des cellules dans les regroupements de cellules. Ensuite, nous avons fait une extraction de cellules exprimant des marqueurs spécifiques à des sous-populations de cellules. Enfin, nous avons comparé les résultats des annotations des outils avec la littérature (tableau A.2). Un exemple standard de la méthodologie est dans la figure 5.1.

5.1.1 Jeux de données

Après avoir découvert que le GDNF pouvait restaurer les niveaux de cellules gliales et neuronales dans le côlon suite à un traitement au GDNF (Soret *et al.*, 2020) (Soret *et al.*, 2021), les auteurs ont retenté l'expérience, mais en utilisant des

souris triples transgéniques *Hol$^{Tg/Tg}$;Sox10-CreERT2$^{Tg/+}$;R26$^{YFP/+}$*. Les cellules avaient pour caractéristique de posséder un marqueur fluorescent (YFP+). Ils ont récupéré les cellules marquées de la fluorescence grâce au tri cellulaire activé par fluorescence (FACS) et ils les ont séquencées à l'aide de l'appareil *Chromium Next GEM technology* de 10x Genomics en utilisant la méthode de séquençage par paire l'appareil *Novaseq 6000* d'Illumina (N. Pilon, communication personnelle, le 29 septembre 2021).

Cela a produit 4 jeux de données pour notre flux de travail (figure 5.2). Le premier jeu de données provient de souris qui ont été traitées avec du dextran sulfate de sodium (DSS), un produit utilisé pour induire de l'inflammation dans le côlon qui est similaire à de la colite inflammatoire et des ulcères chez les souris (Chassaing *et al.*, 2014). Le deuxième provient de souris de type sauvage (*wild type (WT)*) n'ayant subi aucun traitement. Le troisième provient de souris *Holstein Homozygote (HH)* n'ayant subi aucun traitement. Enfin, le dernier jeu de données est constitué des souris *Holstein Homozygote (HH)* ayant été traitées avec du GDNF pendant les jours P5 à P8 afin de pouvoir régénérer le système entérique dans le côlon (HH-GDNF) (Soret, 2021) (R. Soret, communication personnelle, le 10 octobre 2021).

5.1.2 Recherche des marqueurs associés aux types de cellules

Nous avons utilisé les bases de données CellMarker2.0 (Hu *et al.*, 2023a) (Hu *et al.*, 2023b), PanglaoDB (Franzén *et al.*, 2019b)(Franzén *et al.*, 2019a), ainsi qu'une liste de gènes provenant de la littérature afin de compiler une liste de marqueurs connus qui sont spécifiques aux cellules gliales, aux neurones et aux cellules de Schwann (tableau A.2).

5.1.3 Méthode utilisée avec le flux de travail

Pour les analyses des jeux de données, les fichiers FASTQ ont été alignés sur le génome de référence (*refdata-gex-mm10-2020-A*) avec la version 6.1.1 de Cell Ranger sur la plateforme de Calcul Canada (l'Alliance). Ensuite, les jeux de données ont subi des étapes de prétraitement de données avec Seurat en ne gardant que les cellules exprimant moins de 10% de gènes mitochondriaux et en imposant des limites d'un minimum de 100 ainsi qu'un maximum de 10000 gènes uniques exprimés pour les cellules filtrées. Ensuite, une analyse en composantes principales (PCA) a été effectuée et les regroupements des cellules ont été faits avec une réduction de la dimension fixée à 10 pour les 4 jeux de données. Une projection UMAP a ensuite été générée et les sous-populations de cellules ont été annotées avec les outils implémentés dans le flux de travail. De plus, une analyse de l'expression différentielle des gènes avec la liste de marqueurs que nous avons compilée précédemment (tableau A.2) a été faite. Enfin, des analyses ont été réalisées sur les regroupements de cellules avec les outils d'annotation intégrés au flux de travail et nous avons ensuite comparé les résultats avec la littérature.

Ensuite, les cellules exprimant des marqueurs spécifiques aux cellules gliales (Sox10 (May-Zhang *et al.*, 2021)), ainsi que des neurones (Tubb3 (Lau *et al.*, 2019)) ont été extraits. Une autre série de traitement de données par Seurat a été menée en gardant les mêmes paramètres que lors des premières analyses et une analyse de l'expression différentielle des gènes a été effectuée. Ensuite, les nouvelles populations de cellules ont été de nouveau annotées avec les outils intégrés dans le flux de travail et nous avons aussi comparé les résultats avec la littérature. Toutes les étapes d'analyses associées à notre jeu de données sont énumérées à la figure 5.2.

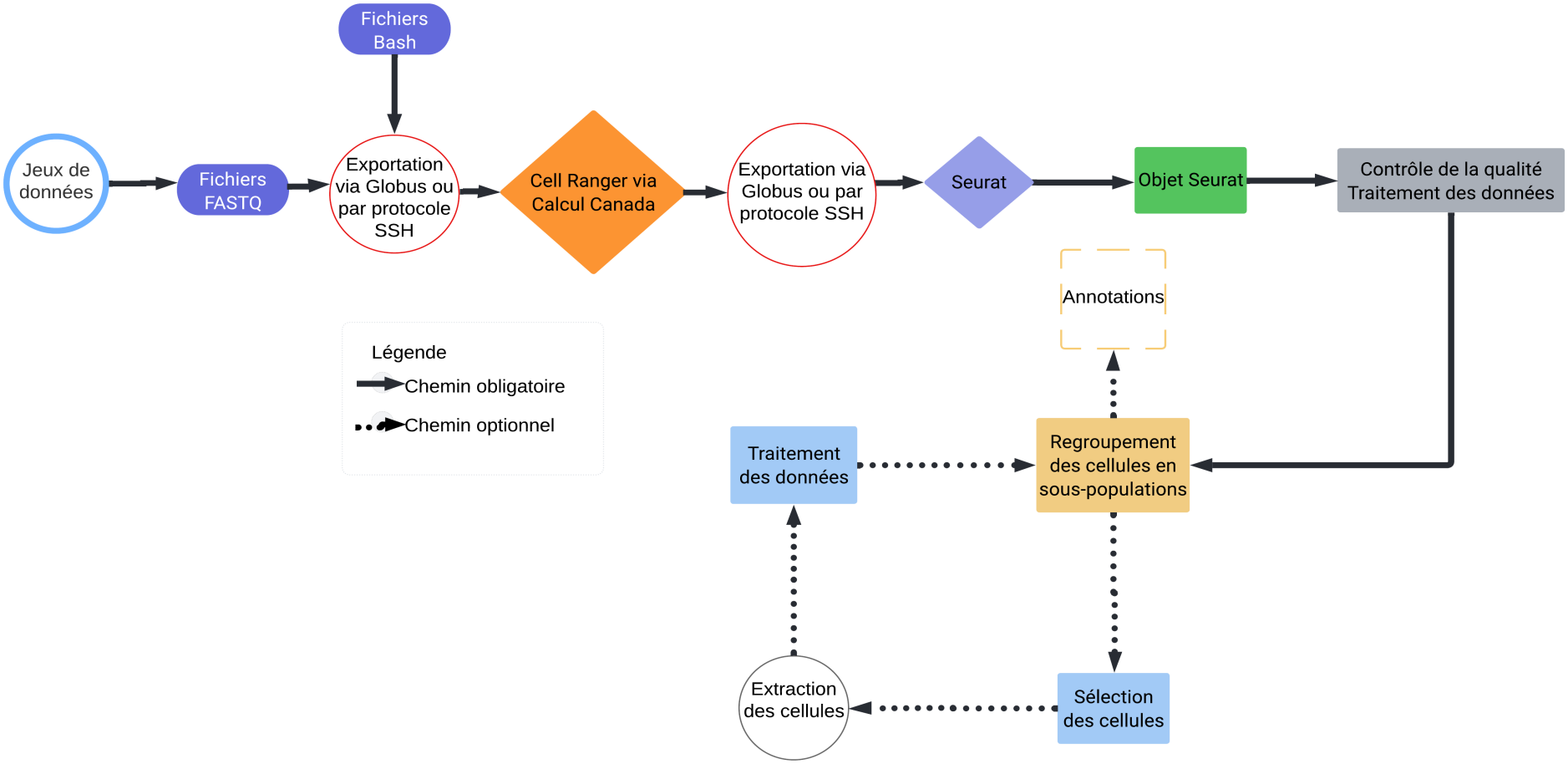

FIGURE 5.1 – Exemple simplifié de la méthodologie du flux de travail.

Cet exemple peut être utilisé pour des analyses de base avec le flux de travail. Les flèches ayant un trait continu représentent les chemins obligatoires pour le flux de travail. Les flèches ayant un trait pointillé représentent les chemins optionnels pour des analyses supplémentaires.

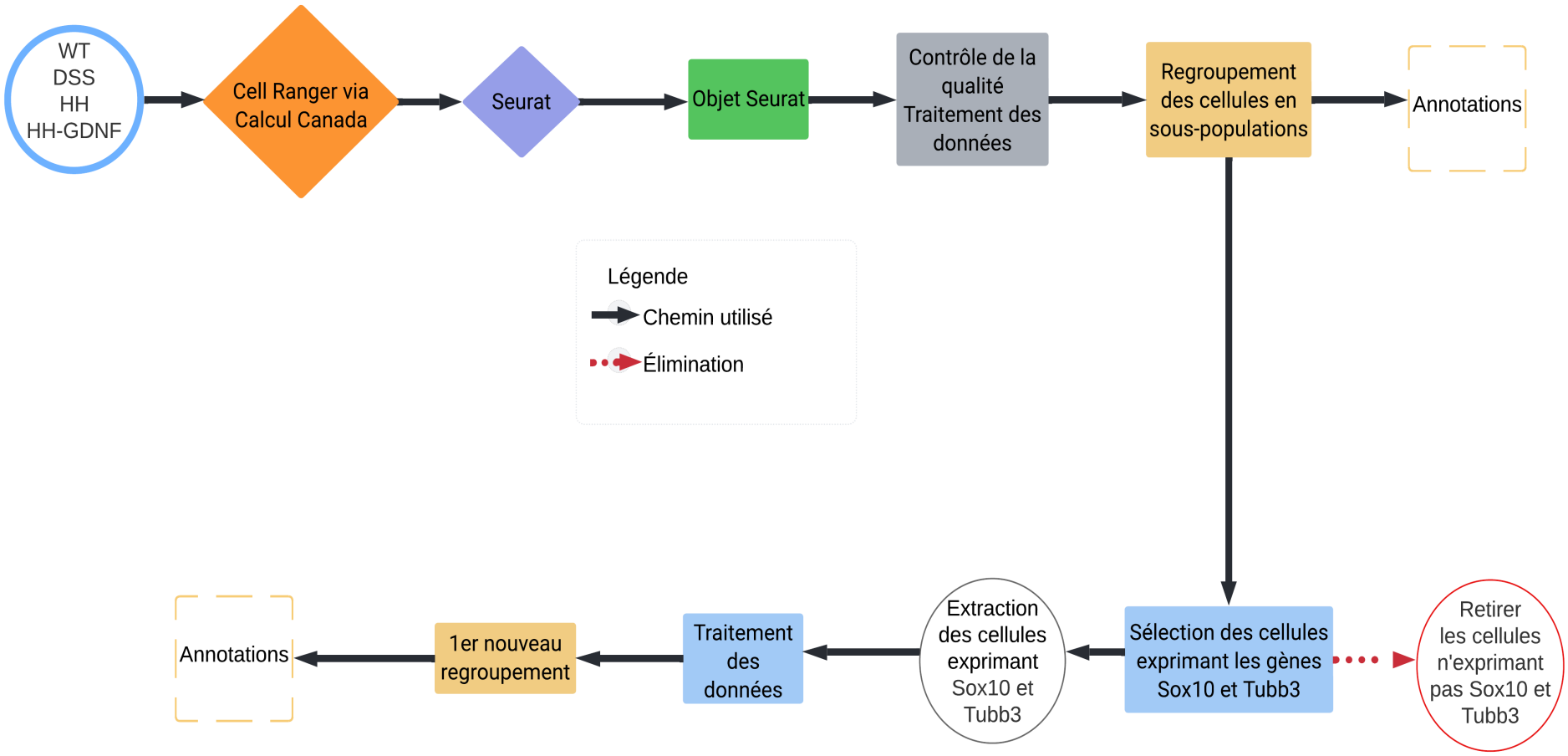

FIGURE 5.2 – Méthodologie avec les 4 jeux de données.

Cette figure représente la méthodologie pour analyser les 4 jeux de données avec le flux de travail. Les flèches ayant un trait continu représentent le chemin complet utilisé avec le flux de travail. La flèche ayant un trait pointillé représente l'élimination des cellules qui ne sont pas d'intérêt (qui n'expriment pas Sox10 et Tubb3).

5.2 Exécution des flux de travail

5.2.1 Exécution du flux de travail de Cell Ranger

Pour chaque jeu de données, les fichiers FASTQ ont été téléchargés sur Calcul Canada (l'Alliance) et Cell Ranger a effectué les alignements sur le génome de référence. Il a ensuite produit des résultats sous forme de fichiers contenant les codes-barres des cellules, la liste des gènes exprimés, ainsi que les matrices des nombres de cellules. De plus, il a aussi produit un fichier en format CLOUPE pouvant être visualisé avec Loupe Browser, ainsi qu'un rapport sous le format HTML pouvant être visualisé dans un navigateur web. Ces fichiers ont ensuite été téléchargés vers la machine locale.

5.2.2 Exécution du flux de travail de Seurat

Les fichiers contenant les codes-barres, les gènes et les matrices ont ensuite été analysés par Seurat et ce dernier a produit un objet Seurat pouvant être utilisé dans les analyses subséquentes. Les analyses de l'expression différentielle des gènes par Seurat ont ensuite produit des fichiers contenant les projections UMAP, les diagrammes en violon des gènes Sox10 (cellules gliales) et Tubb3 (cellules neuronales), des diagrammes à points et des cartes de chaleur pour les gènes associés aux différents types de cellules gliales et neuronales. Des tableaux contenant l'expression des gènes par regroupement de cellules ont aussi été produits.

5.2.3 Exécution du flux de travail des annotations

L'objet Seurat a ensuite été converti en objet Monocle 3 pour les analyses de la trajectoire et du pseudotemps qui font leurs annotations sur la projection UMAP faite par Seurat. Les outils d'identification des types de cellules, tels que ScType

et scMayoMap, ont réutilisé l'objet Seurat pour leurs analyses. ScType a produit un tableau et une figure contenant la projection UMAP faite par Seurat en y ajoutant ses annotations, tandis que scMayoMap a produit un diagramme à points avec ses annotations. Enfin, les listes des gènes exprimés ont été extraites de l'objet Seurat avant d'être utilisées pour les analyses avec EasyCellType et Enrichr. EasyCellType a produit des figures telles que des diagrammes à points contenant ses annotations, tandis que Enrichr a produit des tableaux contenant ses annotations qui ont ensuite été convertis en diagrammes à points.

5.3 Résultats

5.3.1 Souris de type sauvage sans traitement

Pour les souris de type sauvage (figures 5.3, 5.4 et 5.5), les cellules gliales sont détectées dans les sous-populations 0, 3 et 6 par ScType et Enrichr. Les analyses de l'expression différentielles confirment ces résultats au niveau des marqueurs associés aux cellules gliales avec une forte expression des marqueurs Sox10 (May-Zhang *et al.*, 2021), Fabp7 et Plp1 (Lau *et al.*, 2019) dans les sous-populations 3 et 6, ainsi qu'une plus forte expression des autres marqueurs associés aux cellules gliales (Ednrb (Sasselli *et al.*, 2012), Gfap (He *et al.*, 2021), S100b (Brügger *et al.*, 2020), Apoe (Morarach *et al.*, 2021) et Sox2 (Han *et al.*, 2019) dans le cluster 0. De plus, seulement ScType confirme une présence de cellules gliales dans la sous-population 8 (exprimant aussi les mêmes marqueurs). Pour les neurones, trois des outils d'annotation sur quatre confirment qu'ils sont présents dans la sous-population 2 (par ScType, Enrichr et scMayoMap), ce qui concorde avec une expression plus élevée des gènes associés aux neurones (en particulier Tubb3, Elavl4 (Lau *et al.*, 2019), Ret (Sasselli *et al.*, 2012) , Vip et Calb2 (May-Zhang *et al.*, 2021)). Pour EasyCellType, les sous-populations mentionnées correspondent aux

cellules entéroendocrines (seulement pour les sous-populations 0 et 2). Pour les cellules exprimant les marqueurs associés aux précurseurs des cellules de Schwann, elles sont fortement présentes dans la sous-population 6 (Dhh, Mal et Mpz (Morarach *et al.*, 2021)), mais il y a aussi la présence de cellules exprimant Gap43 (Carrió *et al.*, 2019) dans la sous-population 2, ainsi que la sous-population 3 pour les cellules de Schwann présynaptiques (Ajap1 (Castro *et al.*, 2020)) (aussi confirmée par Enrichr). En ce qui concerne la trajectoire avec Monocle 3, elle connecte toutes les sous-populations de cellules.

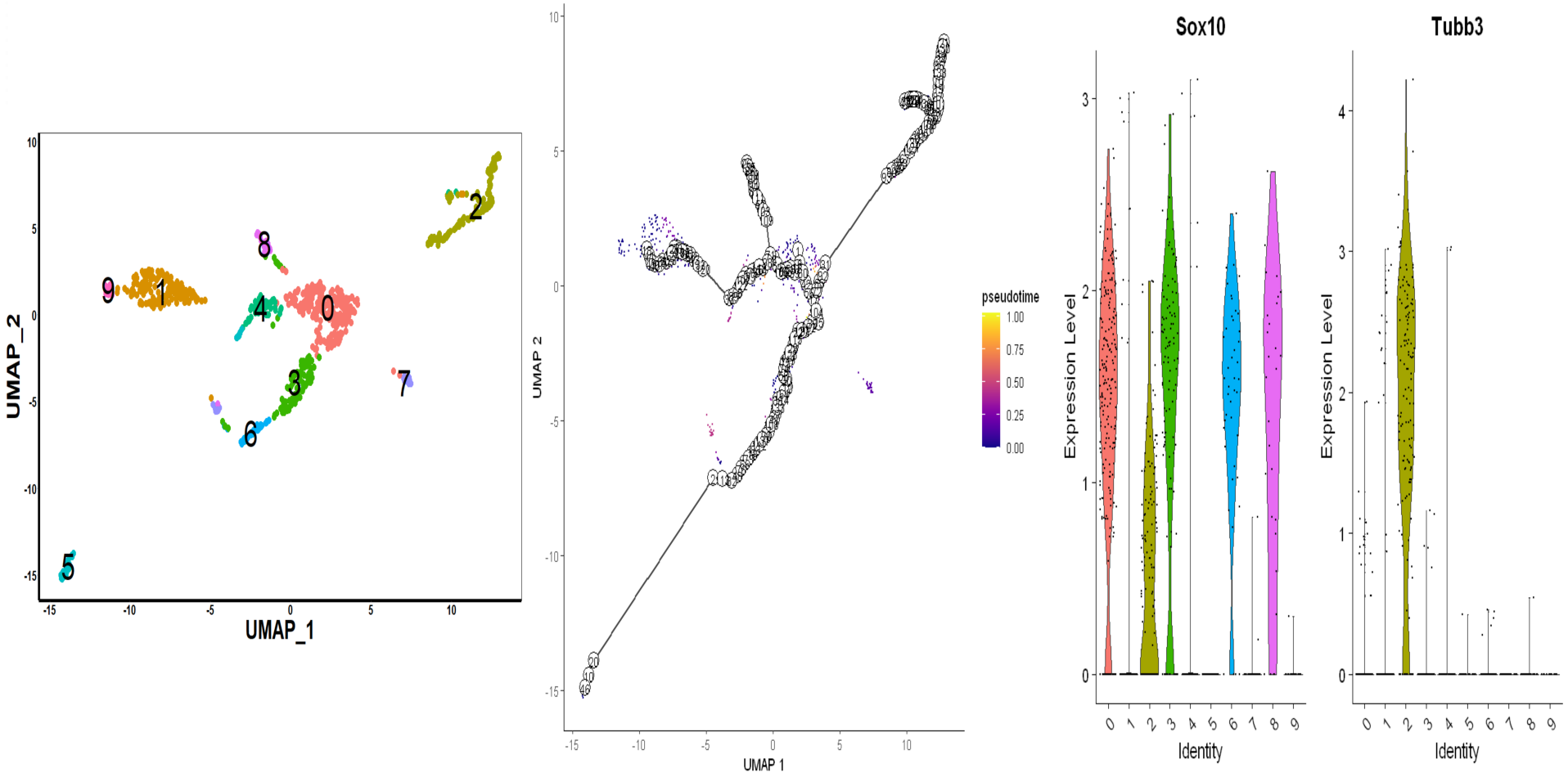

FIGURE 5.3 – Souris de type sauvage sans traitement – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

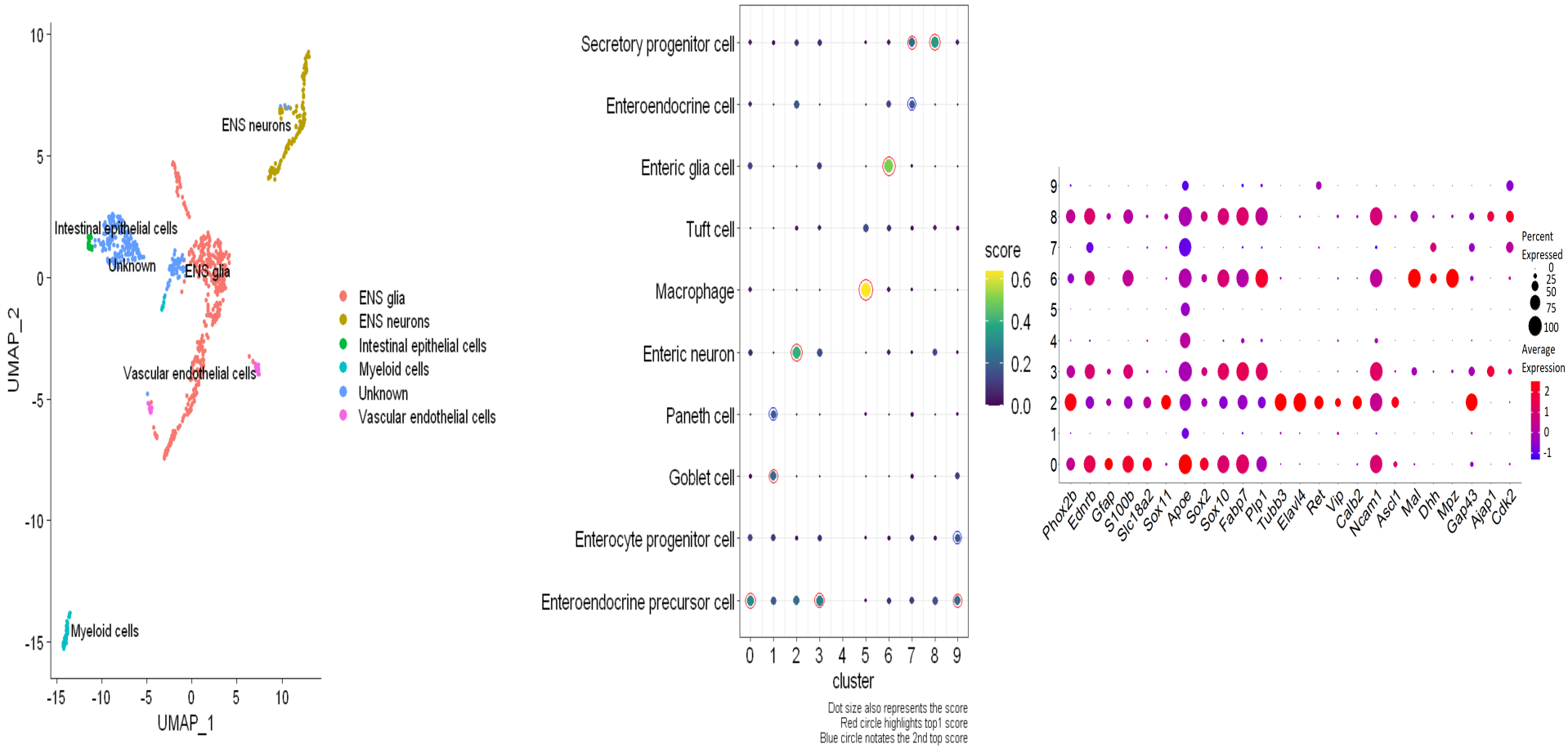

FIGURE 5.4 – Souris de type sauvage sans traitement – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

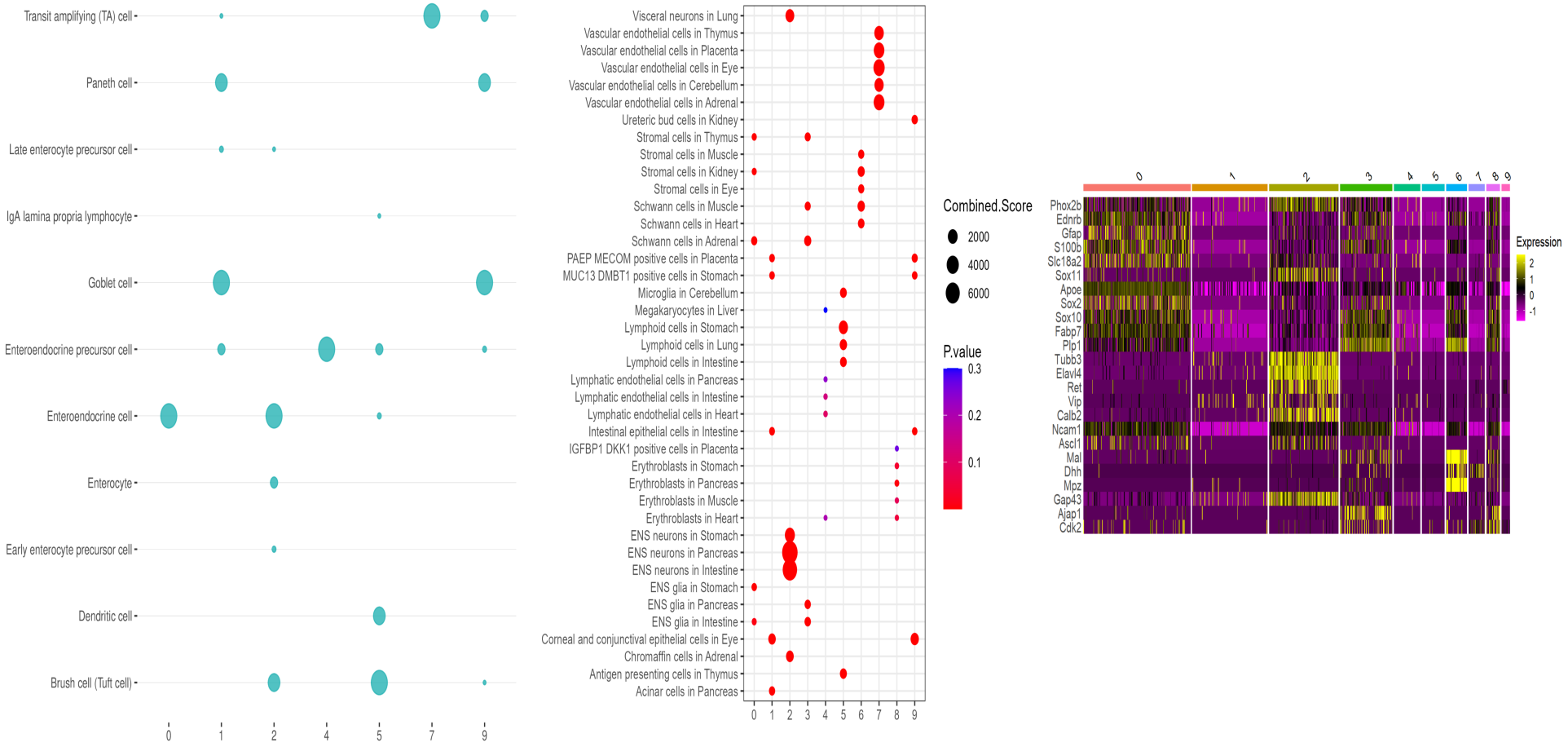

FIGURE 5.5 – Souris de type sauvage sans traitement — Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.2 Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage

En ce qui concerne les cellules Sox10 et Tubb3 provenant des souris de type sauvage non traitées (figures 5.6, 5.7 et 5.8), les cellules gliales sont seulement présentes dans les sous-populations 1, 3 et 5 pour Sctype, mais scMayoMap confirme seulement une très forte présence de ce type de cellule dans la sous-population 3. Dans le cas de Enrichr, elles sont présentes dans la sous-population 1. Les analyses de l'expression différentielle des gènes indiquent une plus forte expression du gène Sox11 (Morarach *et al.*, 2021) dans la sous-population 4, comparée à une expression presque nulle dans les autres sous-populations. Pour les cellules neuronales, les trois outils confirment leur présence dans la sous-population 4 (aussi confirmé par les marqueurs (Tubb3, Elavl4 (Lau *et al.*, 2019), Ret(Sasselli *et al.*, 2012) , Calb2 (May-Zhang *et al.*, 2021)). En ce qui concerne les marqueurs associés aux précurseurs des cellules de Schwann (Mal, Dhh et Mpz (Morarach *et al.*, 2021)), elles sont fortement présentes dans la sous-population 3. Il y a aussi une forte présence de cellules exprimant Gap43 dans la sous-population 4, la présence de gènes associés aux cellules de Schwann périsynaptiques (Ajap1 (Castro *et al.*, 2020)) et celles subissant le cycle cellulaire (Cdk2 (Tikoo *et al.*, 2000)) dans la sous-population 5. Enrichr confirme aussi leur présence dans les sous-populations 1 et 3. Pour EasyCellType la sous-population 4 est probablement des cellules entéroendocrines et des entérocolites dans la sous-population 2. En ce qui concerne la trajectoire avec Monocle 3, elle traverse toutes les sous-populations à l'exception de la sous-population 5.

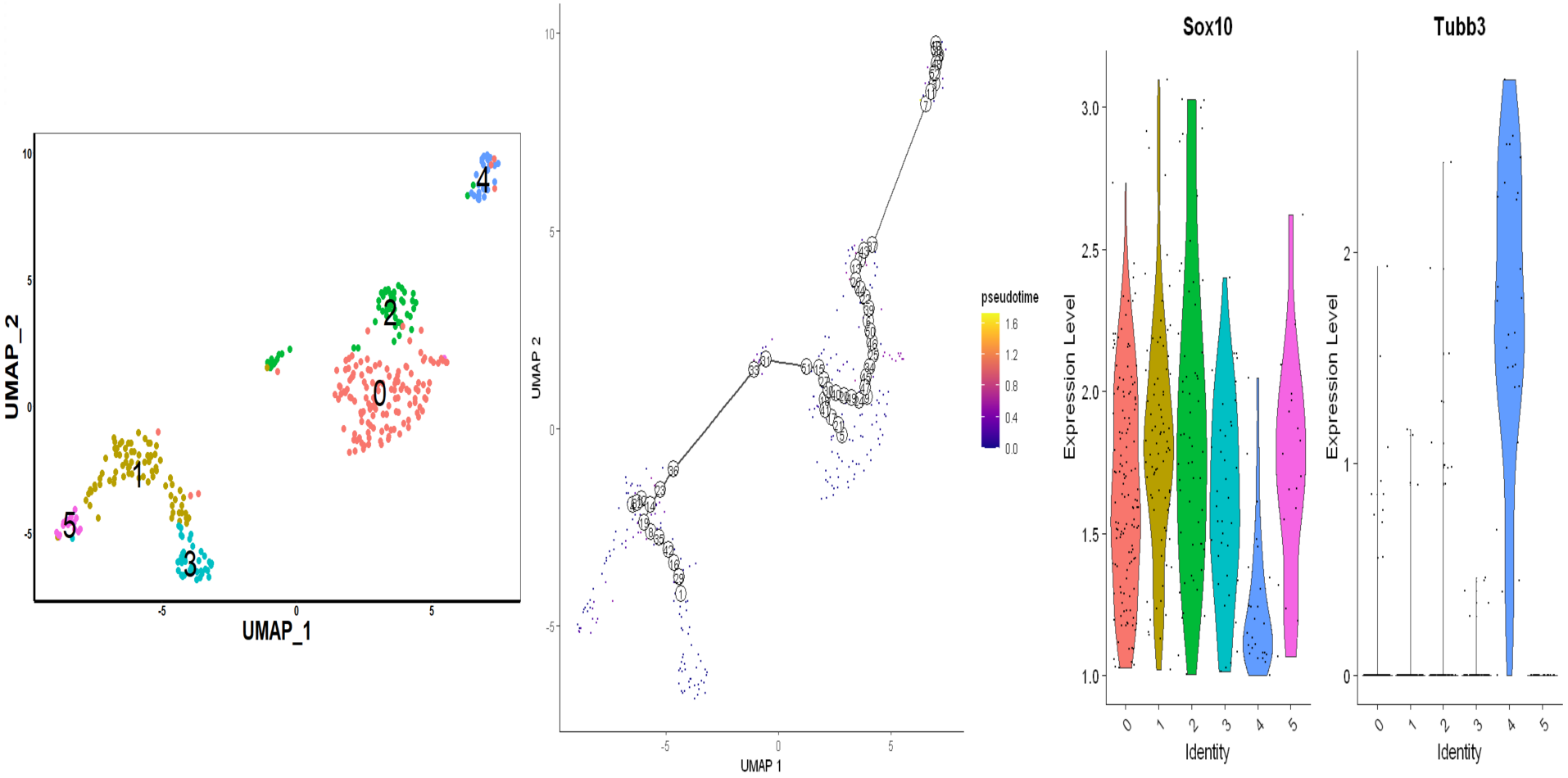

FIGURE 5.6 – Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage sans traitement – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

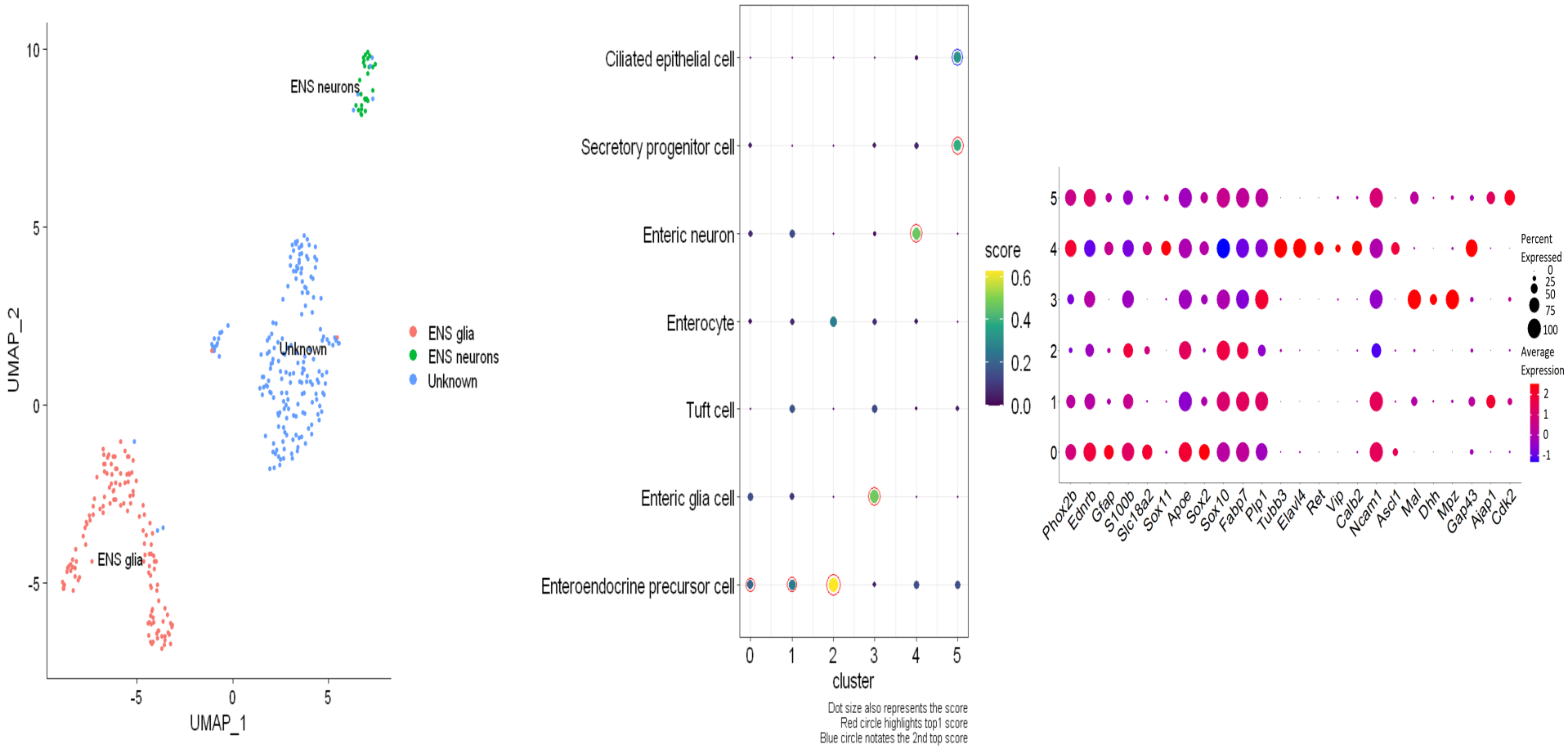

FIGURE 5.7 – Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage sans traitement – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

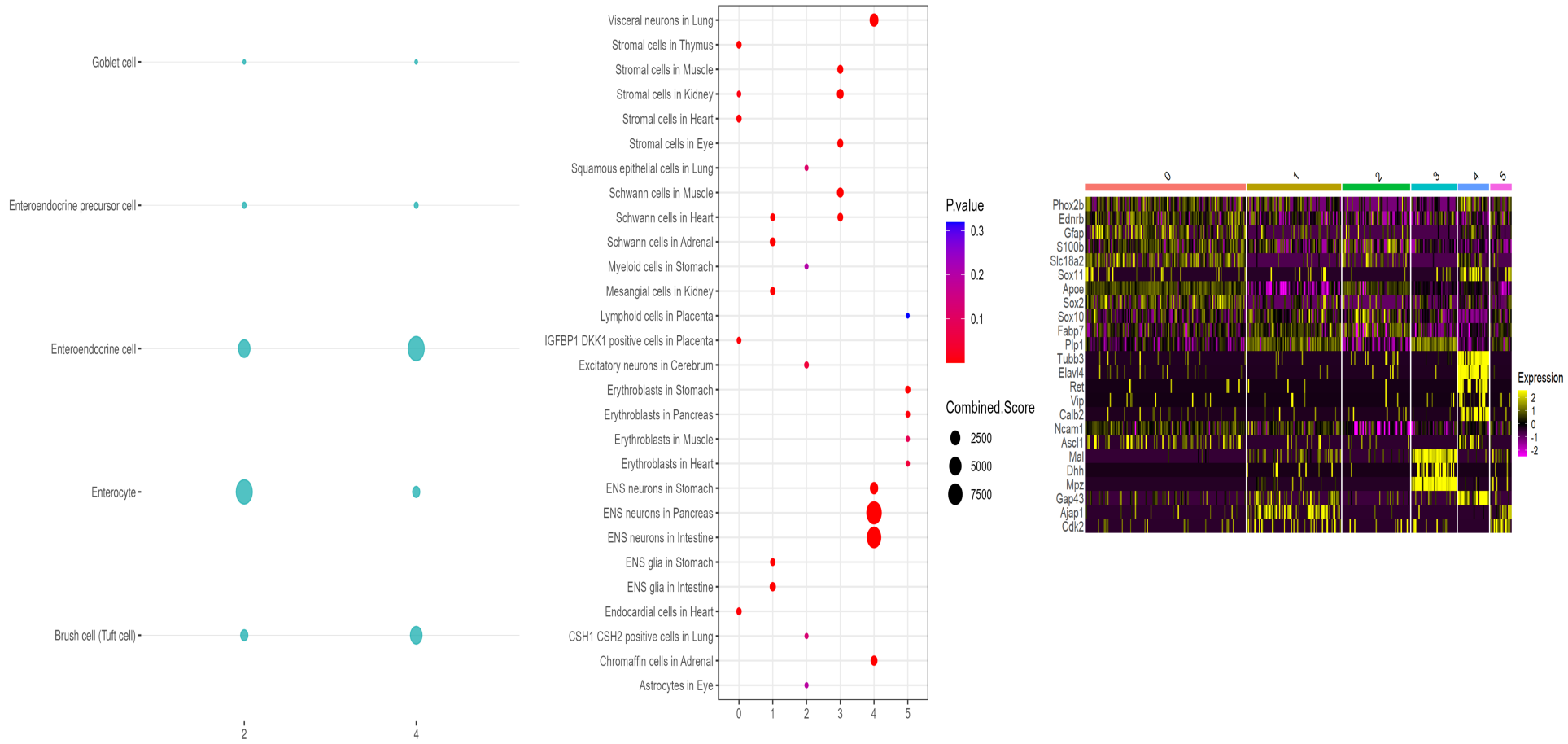

FIGURE 5.8 – Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage sans traitement – Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.3 Souris de type sauvage traitées avec du dextran sulfate de sodium

Pour les souris traitées avec du dextran sulfate de sodium (figures 5.9, 5.10 et 5.11), les cellules gliales ont été détectées dans les sous-populations 0 et 1 par les outils ScType, ScMayoMap et Enrichr. L'analyse de l'expression différentielle des gènes confirme ces résultats avec l'expression des gènes Gfap (He *et al.*, 2021) , S100b (Brügger *et al.*, 2020), Scl18a2 (Drokhlyansky *et al.*, 2020) , Apoe (Morarach *et al.*, 2021), Sox2 (Han *et al.*, 2019) Sox10 (May-Zhang *et al.*, 2021), Fabp7 et Plp1 (Lau *et al.*, 2019). De plus, ces marqueurs sont présents dans les sous-populations 3, 4, 8 et 10, ainsi que des neurones dans les sous-populations 3 et 10, ce qui concorde aussi avec la présence de cellules exprimant Tubb3 (Lau *et al.*, 2019) dans les deux sous-populations. Pour EasyCellType, ces sous-populations (3 et 10) correspondaient principalement aux cellules entéroendocrines. En ce qui concerne les résultats pour Monocle 3, il y a une trajectoire qui traverse toutes les sous-populations qui ont été détectées en tant que cellules gliales avec une trajectoire qui traverse aussi celles qui ont été détectées en tant que neurones. En ce qui concerne l'analyse différentielle des gènes, la sous-population 10 exprimait beaucoup plus de marqueurs associés aux neurones (Tubb3, Elavl4 (Lau *et al.*, 2019), Ret (Sasselli *et al.*, 2012), Vip, Calb2 (May-Zhang *et al.*, 2021)) que celles présentes dans la sous-population 3. Il y a aussi une présence de cellules exprimant les marqueurs des précurseures des cellules de Schwann (Mal, Dhh, Mpz (Morarach *et al.*, 2021) et Gap43 (Carrió *et al.*, 2019)) dans la sous-population 4 et dans la sous-population 1 avec Enrichr.

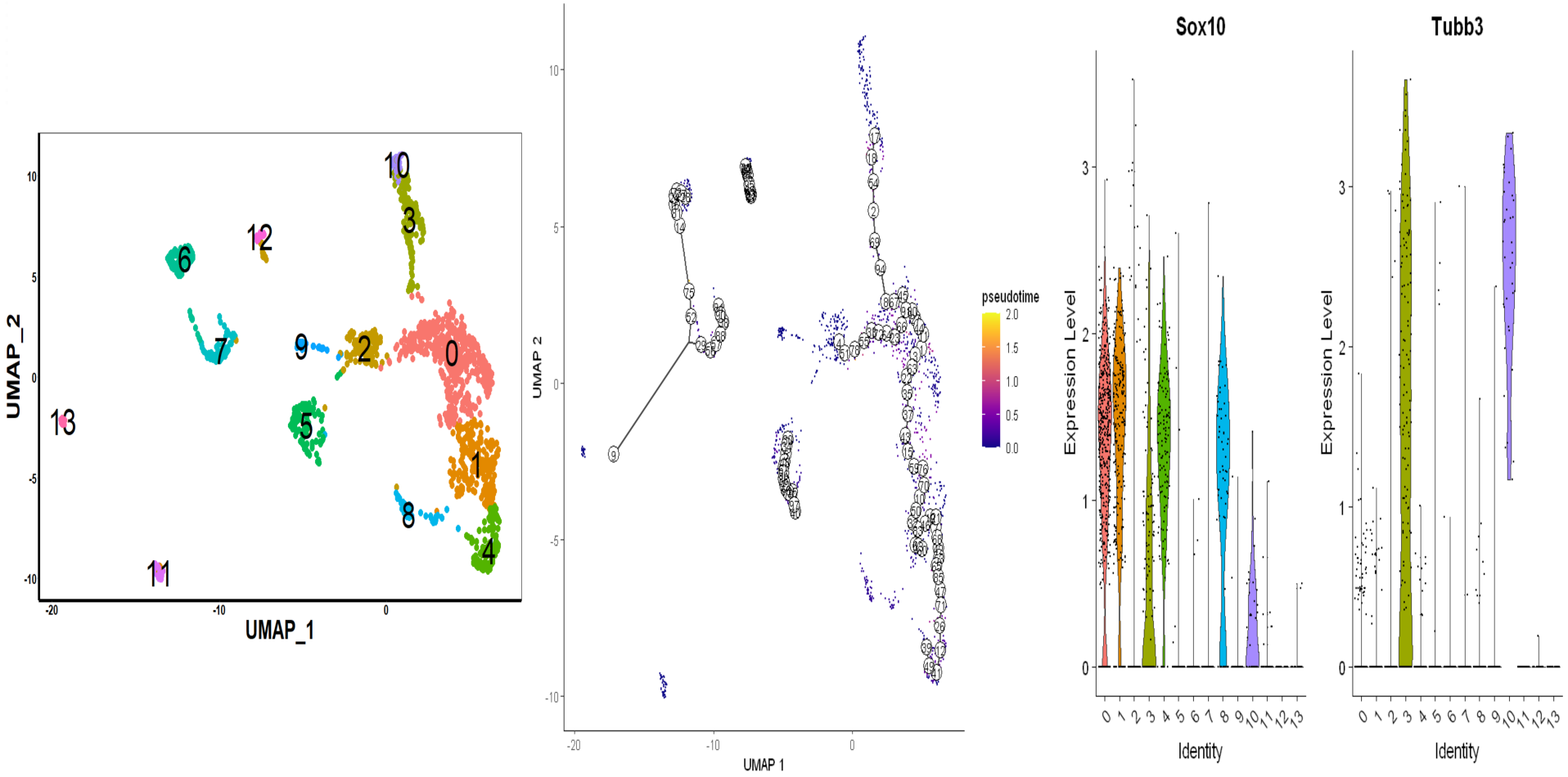

FIGURE 5.9 – Souris de type sauvage traitées avec du dextran sulfate de sodium – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

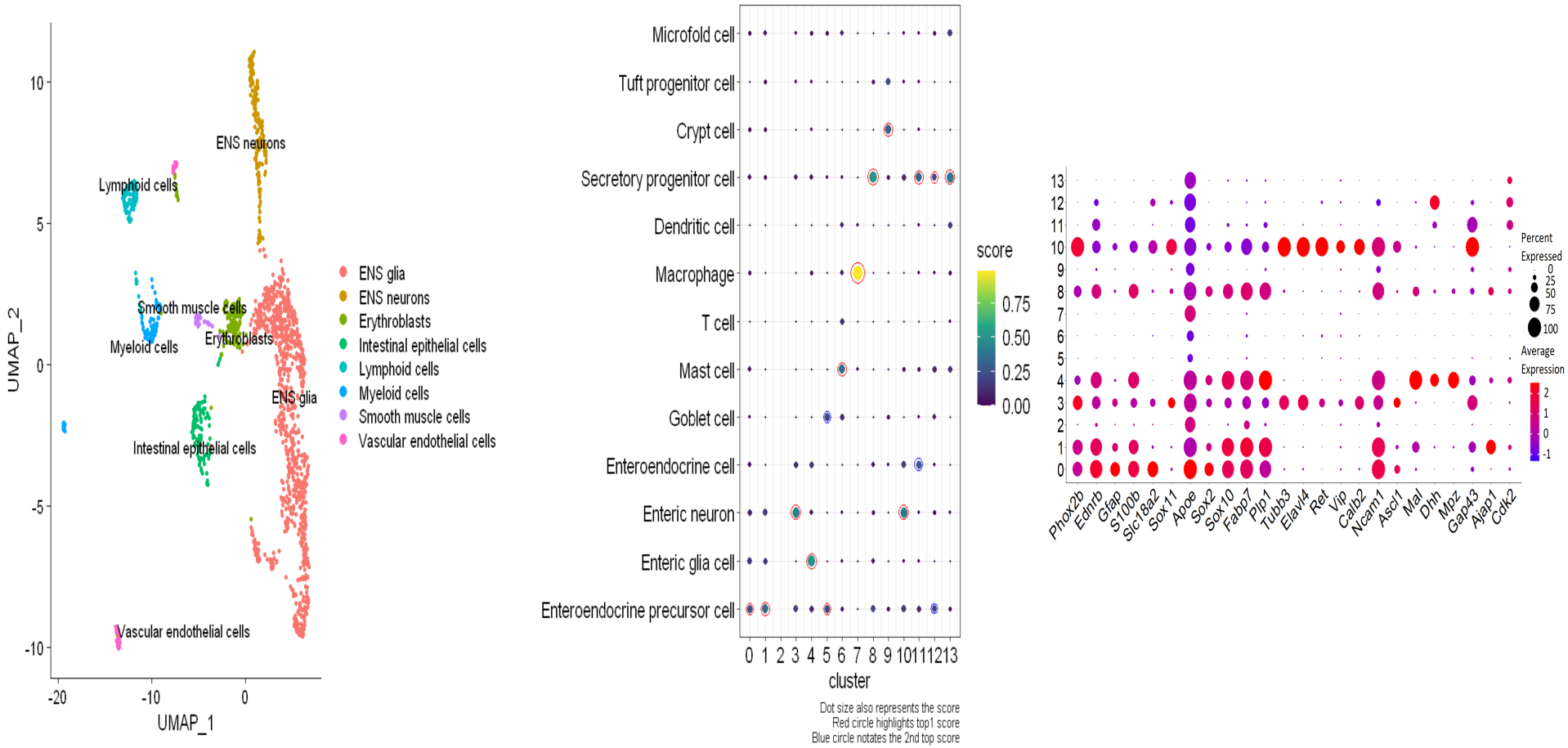

FIGURE 5.10 – Souris de type sauvage traitées avec du dextran sulfate de sodium — Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

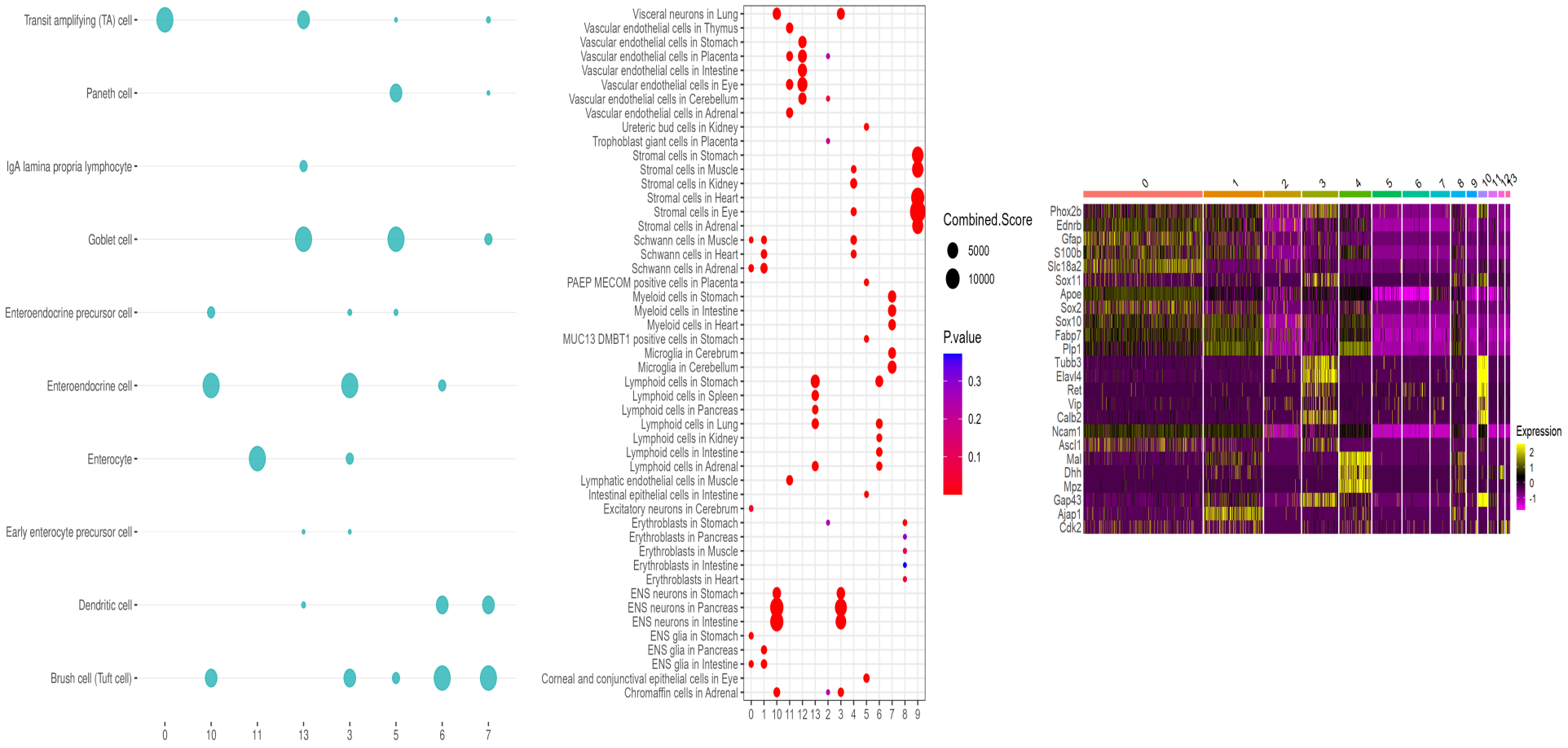

FIGURE 5.11 – Souris de type sauvage traitées avec du dextran sulfate de sodium – Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.4 Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage traitées avec du dextran sulfate de sodium

Suite à l'extraction des cellules exprimant Sox10 et Tubb3 pour les souris traitées avec du dextran sulfate de sodium (figures 5.12, 5.13 et 5.14), les cellules gliales sont fortement présentes dans les sous-populations 1 et 3, mais ScMayoMap détecte une plus forte présence de celles-ci dans la sous-population 3. Par contre, Enrichr le détecte plus dans la sous-population 1. Les analyses de l'expression différentielle confirment ces résultats, mais elles indiquent aussi une légère expression du gène Ascl1 (cellules précurseurs des neurones) (Castro *et al.*, 2020). En ce qui concerne les cellules neuronales, les trois outils détectent leur présence dans la sous-population 4, qui sont aussi confirmées par la présence de Tubb3 (Lau *et al.*, 2019). De plus, les analyses de l'expression différentielle des gènes confirment ces résultats (Tubb3, Elavl4 (Lau *et al.*, 2019), Ret (Sasselli *et al.*, 2012), Vip, Calb2 (May-Zhang *et al.*, 2021)) dans cette sous-population de cellules. Pour les précurseurs des cellules de Schwann (Mal, Dhh et Mpz (Morarach *et al.*, 2021)) elles sont fortement présentes dans la sous-population 3, mais Enrichr confirme aussi leur présence dans la sous-population 1. En ce qui concerne EasyCellType, il détecte aussi des entérocolites dans la sous-population 1 et des cellules entéroendocrines dans la sous-population 4. En ce qui concerne la trajectoire avec Monocle 3, il traverse seulement les sous-populations 0, 1 et 6.

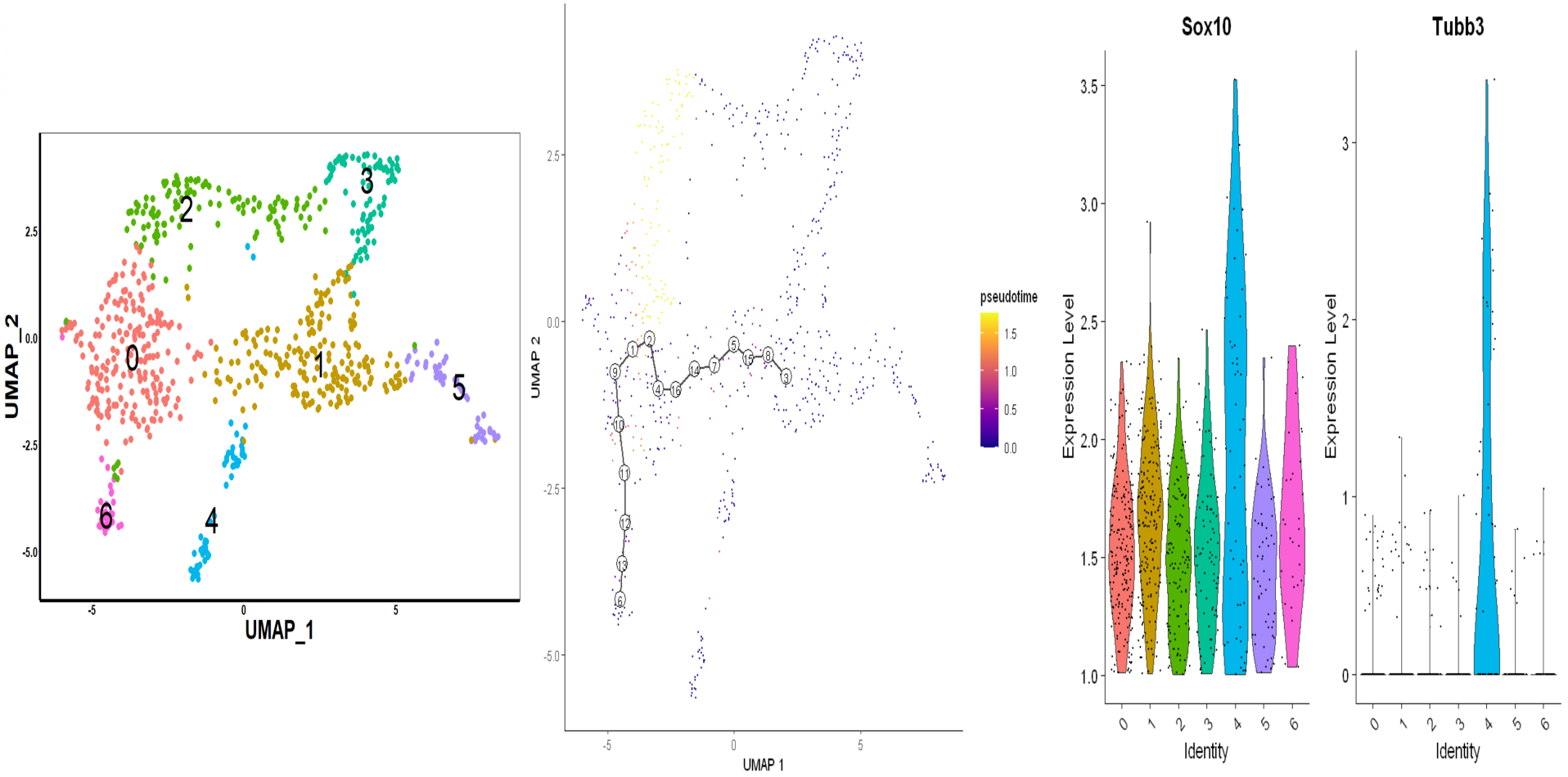

FIGURE 5.12 – Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage traitées avec du dextran sulfate de sodium – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

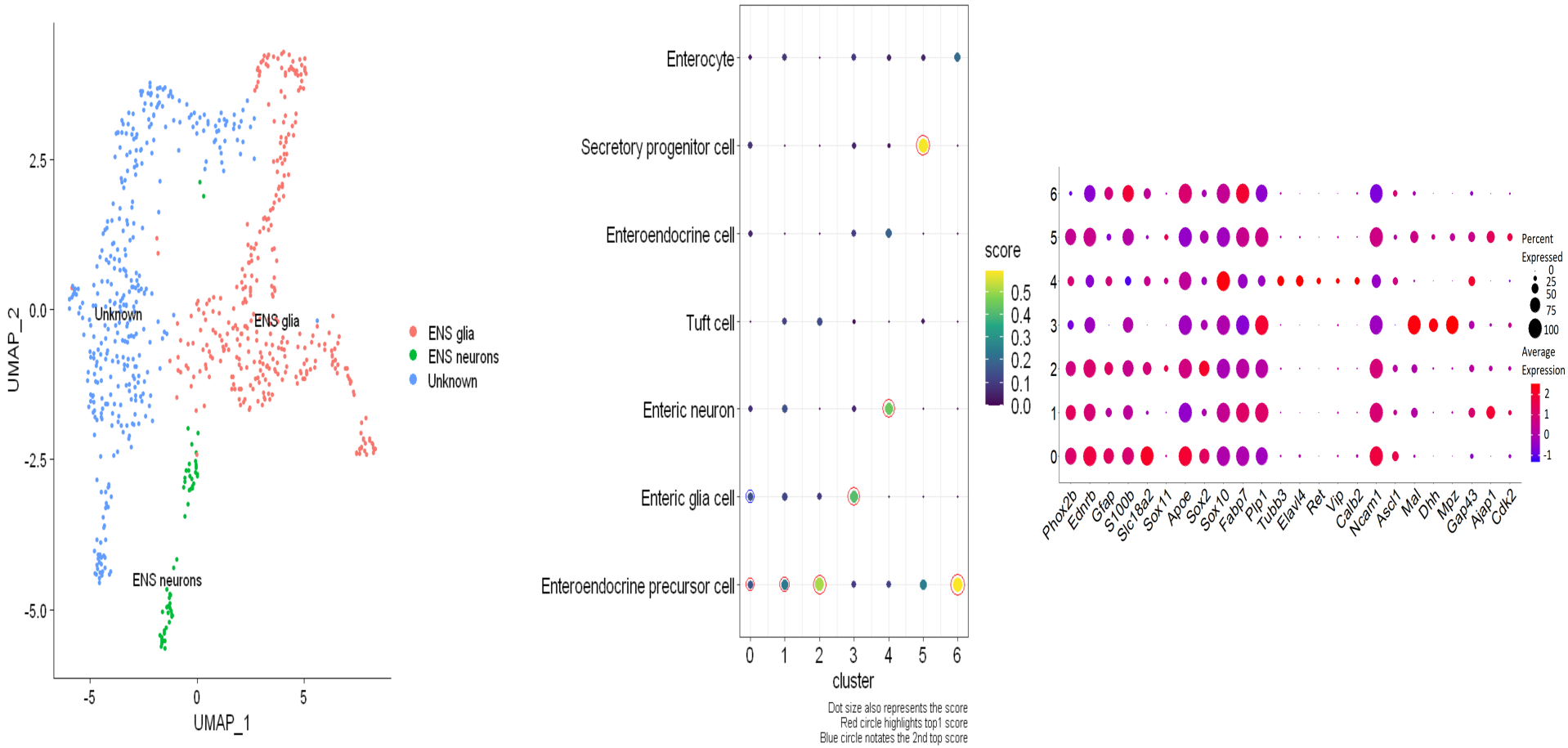

FIGURE 5.13 – Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage traitées avec du dextran sulfate de sodium – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

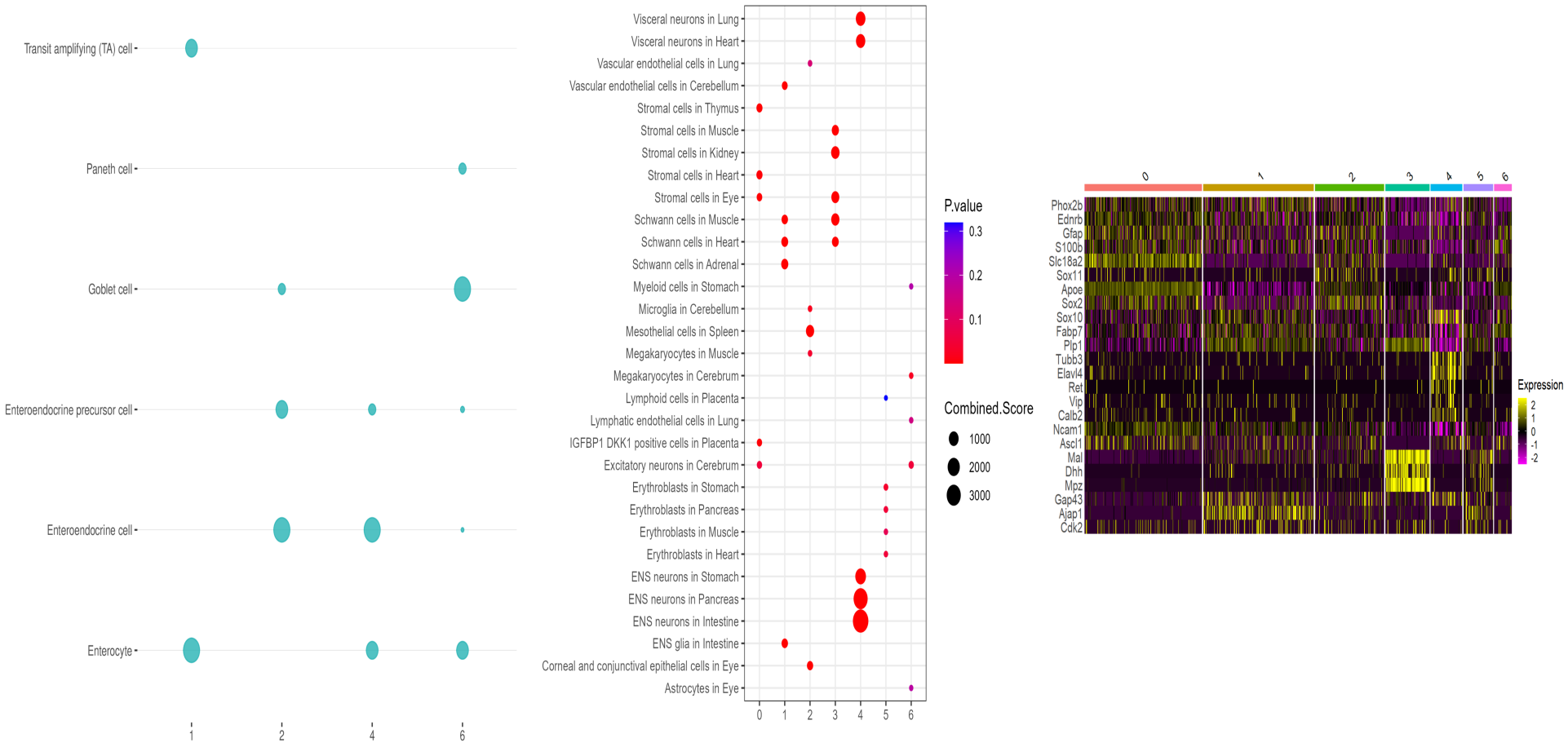

FIGURE 5.14 – Cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage traitées avec du dextran sulfate de sodium – Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.5 Souris *Holstein* sans traitement

Pour les souris *Holstein* n'ayant subi aucun traitement (figures 5.15, 5.16 et 5.17), ScType confirme que les cellules gliales sont présentes dans les sous-populations 1 (confirmé par Enrichr), 2 (confirmé par ScMayoMap) et 8 (confirmé par Enrichr). Les analyses de l'expression différentielle des gènes confirment que l'expression des gènes associés aux cellules gliales est beaucoup plus présente dans la sous-population 8. En ce qui concerne les cellules neuronales, les trois outils confirment qu'elles sont présentes dans les sous-populations 5 et 11, mais les analyses de l'expression différentielle des gènes confirment une plus haute expression dans la sous-population 11. Pour les précurseures des cellules de Schwann, elles sont détectées par Enrichr dans les sous-populations 1, 2 et 8, mais l'expression différentielle des gènes confirme une plus haute expression dans la sous-population 2. EasyCellType confirme la présence de cellules entéroendocrines pour les sous-populations mentionnées précédemment, mais détecte aussi des cellules amplificatrices de transit. En ce qui concerne la trajectoire avec Monocle 3, il y en a une qui traverse les sous-populations exprimant les cellules gliales seulement dans les sous-populations 1 et 2.

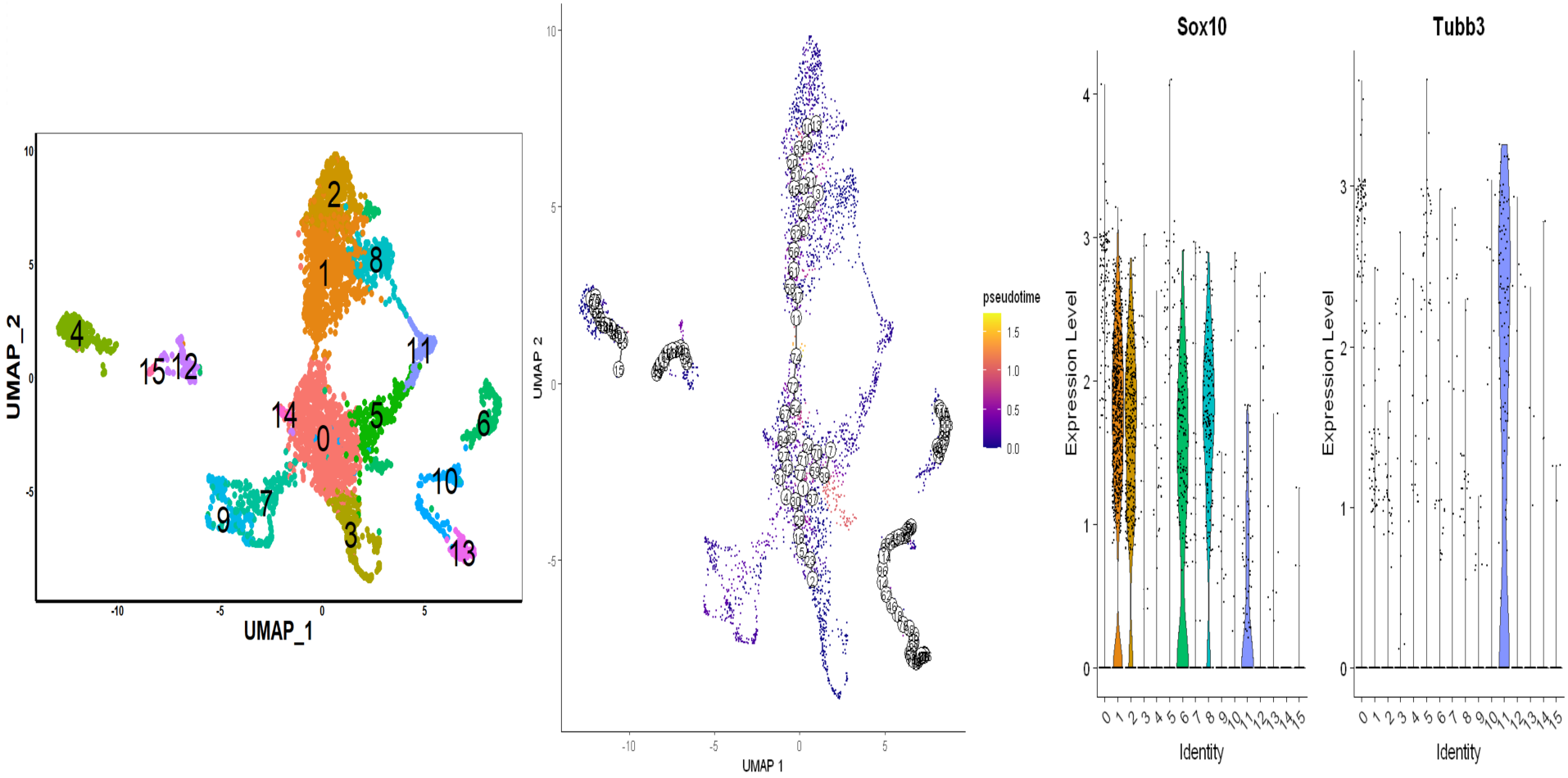

FIGURE 5.15 – Souris *Holstein* sans traitement — Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

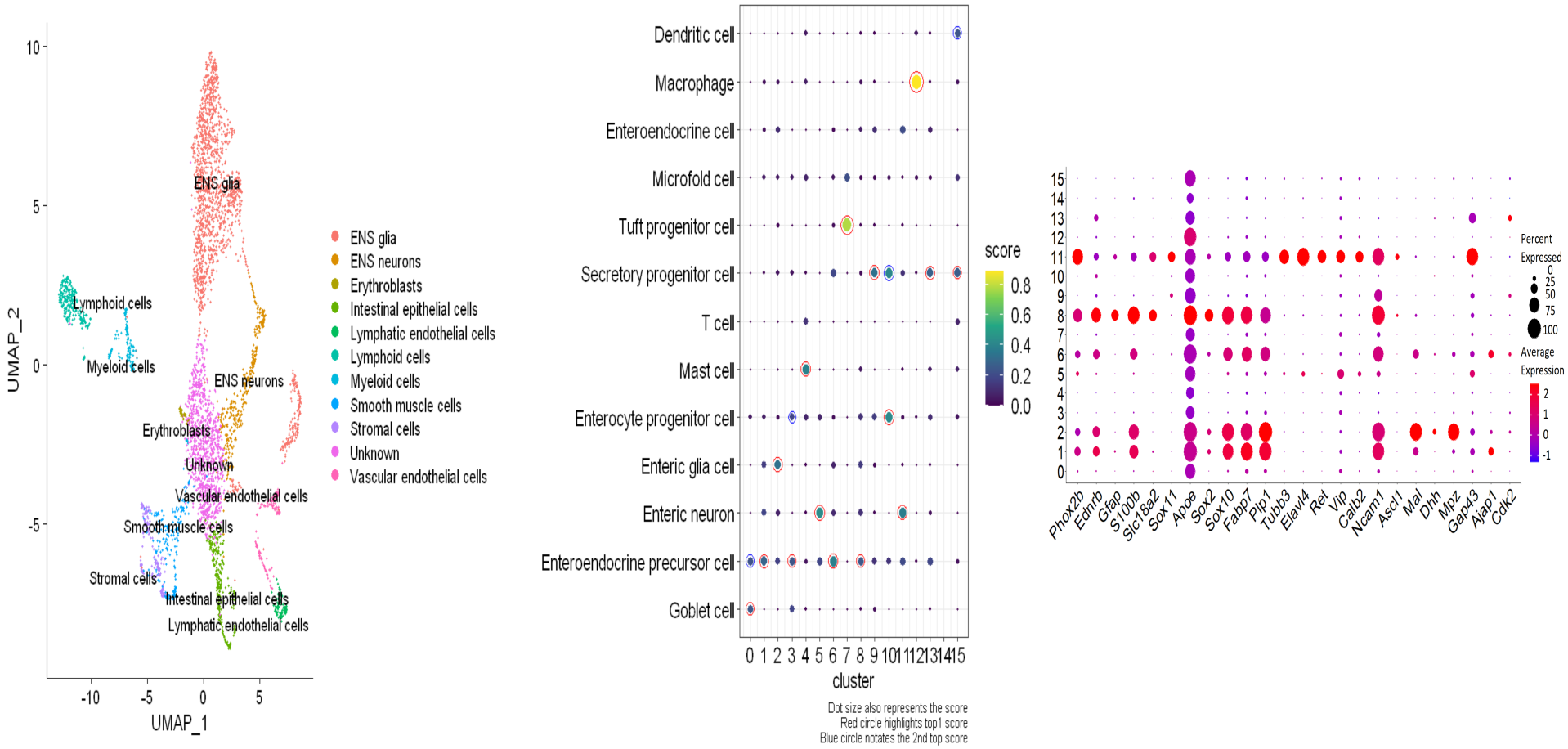

FIGURE 5.16 – Souris *Holstein* sans traitement – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

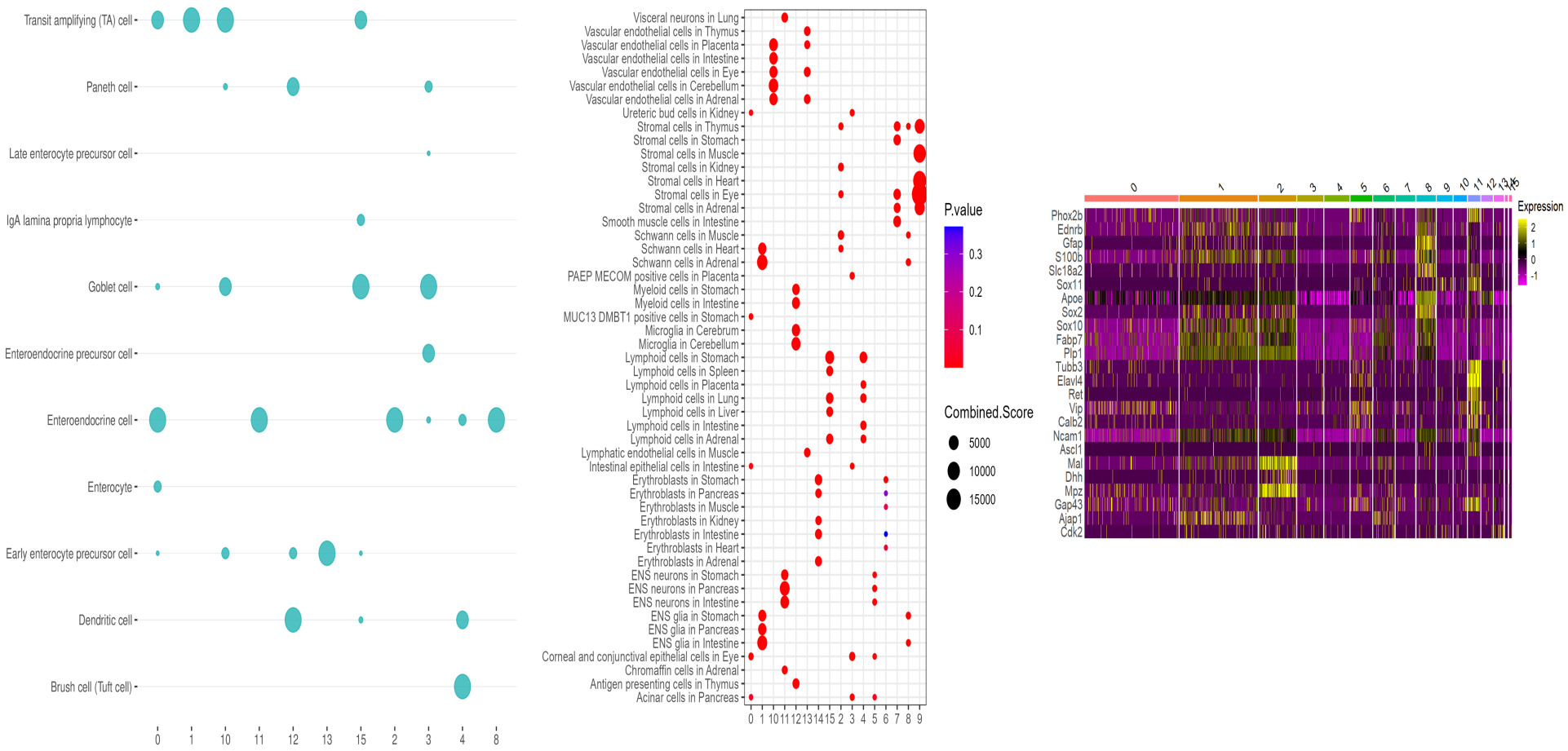

FIGURE 5.17 – Souris *Holstein* sans traitement — Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.6 Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* sans traitement

Pour les cellules exprimant les marqueurs Sox10 et Tubb3 provenant des souris *Holstein* (figures 5.18, 5.19 et 5.20), les cellules gliales sont détectées dans les sous-populations 0, 1, 5 et 7 (pour ScType), mais ScMayoMap détecte aussi leur présence dans la sous-population 3. Les analyses de l'expression différentielle des gènes confirment leur présence seulement dans les sous-populations 3, ainsi qu'une forte expression de Sox10 (May-Zhang *et al.*, 2021) dans la sous-population 4. En ce qui concerne les neurones, elles sont confirmées par ScType dans les sous-populations 4 et 8, mais aussi dans la population 5 par ScMayoMap. En ce qui concerne les analyses différentielles des gènes, elles sont fortement exprimées dans la sous-population 8. Pour les précurseurs des cellules de Schwann, les marqueurs sont surexprimés dans la sous-population 1 (Mal, Dhh et Mpz (Morarach *et al.*, 2021)), 8 (Gap43)(Carrió *et al.*, 2019) et 0 (Ajap1) (Castro *et al.*, 2020). EasyCellType confirme la présence de cellules entéroendocrines pour les sous-populations mentionnées, mais détecte aussi des cellules amplificatrices de transit. Enfin, la trajectoire avec Monocle 3 est circulaire en passant dans les sous-populations 1 et un peu dans la sous-population 3, traverse la sous-population 7 et il y a un autre cycle entre les sous-populations 0, 3 et 5 avec une trajectoire allant jusqu'aux cellules neuronales dans sous-population 4.

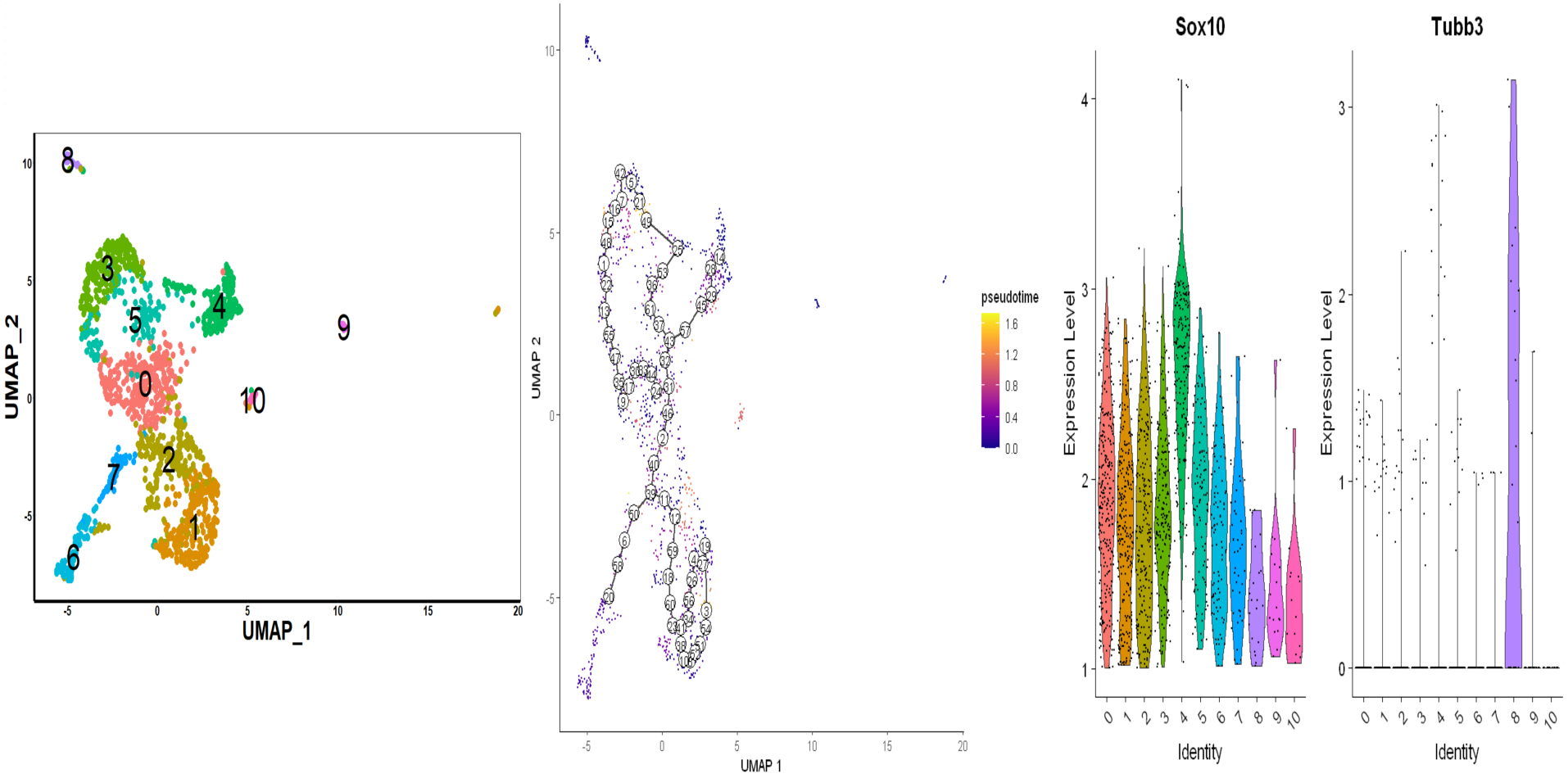

FIGURE 5.18 – Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* sans traitement – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

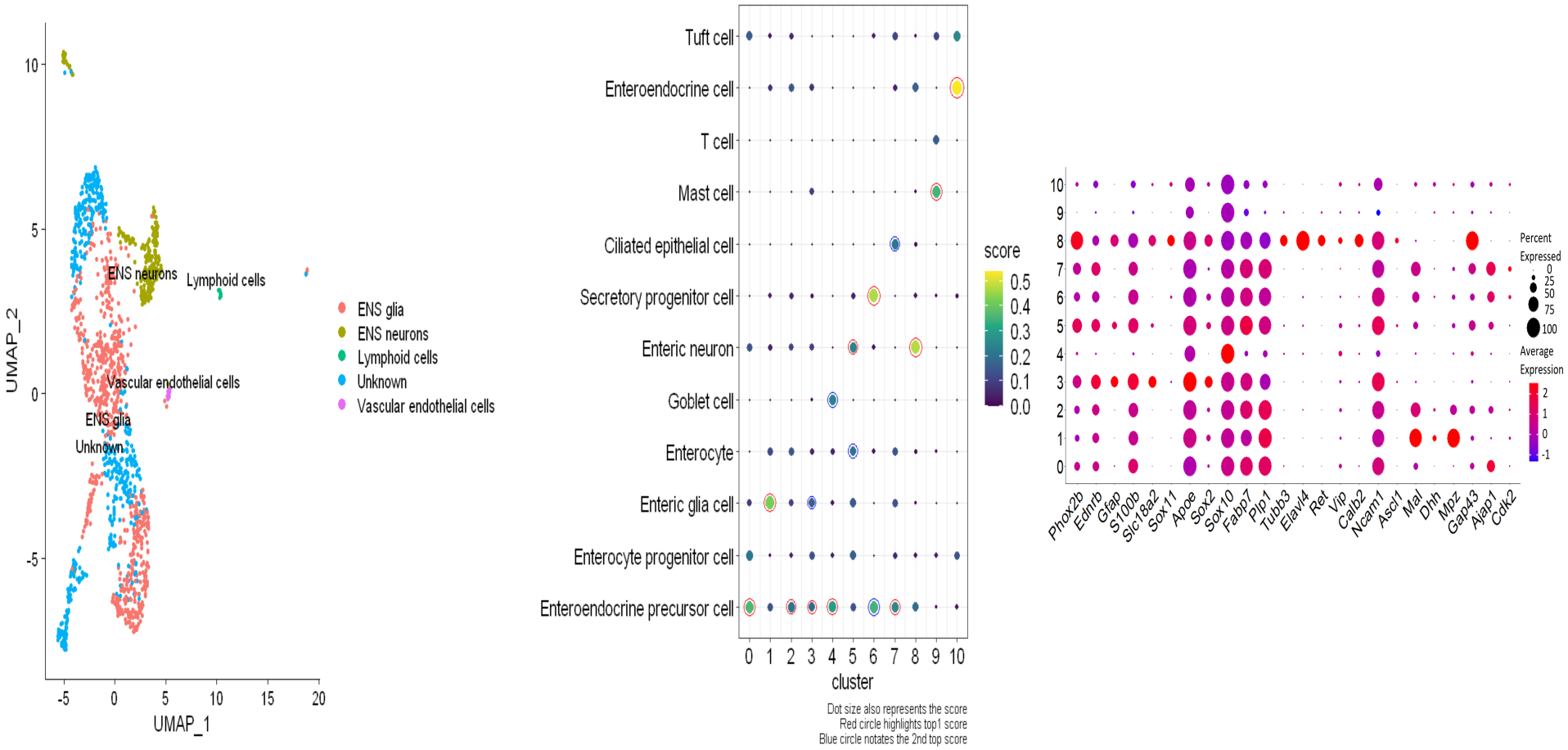

FIGURE 5.19 – Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* sans traitement – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

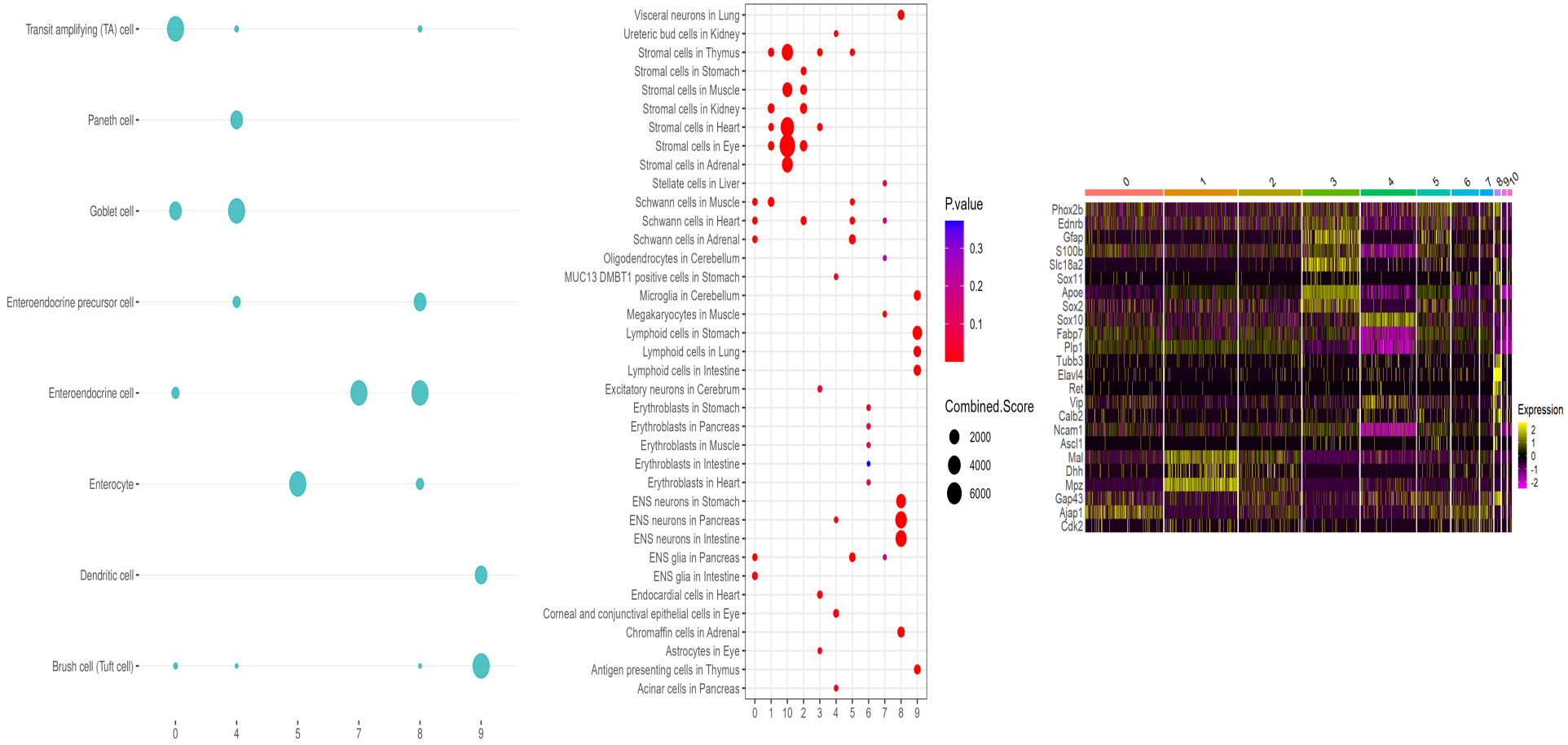

FIGURE 5.20 – Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* sans traitement – Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.7 Souris *Holstein* traitées avec du GDNF

En ce qui concerne les souris *Holstein* ayant été traitées avec du GDNF (figures 5.21, 5.22 et 5.23) ScType et ScMayoMap confirment la présence des cellules gliales dans les sous-populations 0, 1 et 3. Cependant, Enrichr confirme seulement leur présence dans les deux premières mentionnées ci-dessus. Les analyses de l'expression différentielle des gènes confirment leur présence seulement dans les sous-populations 0 et 1, ainsi que la présence de cellules de la crête neurale entérique (Phox2b (Sasselli *et al.*, 2012)) pour les sous-populations 6 et 8, ainsi que Sox11 dans cette dernière (Morarach *et al.*, 2021). Pour les cellules neurales, elles sont confirmées dans les sous-populations 6 et 8 et les gènes associés sont aussi fortement exprimés. Pour les précurseurs des cellules de Schwann, elles sont fortement présentes dans la sous-population 3 (Mal, Dhh et Mpz (Morarach *et al.*, 2021)) et cela est confirmé par Enrichr ainsi que 6 (Gap43). EasyCellType confirme la présence de cellules entéroendocrines pour les sous-populations mentionnées, mais détecte aussi des cellules amplificatrices de transit. Pour la trajectoire avec Monocle 3, elle est aussi fractionnée, mais il y en a une qui traverse les sous-populations gliales et neuronales.

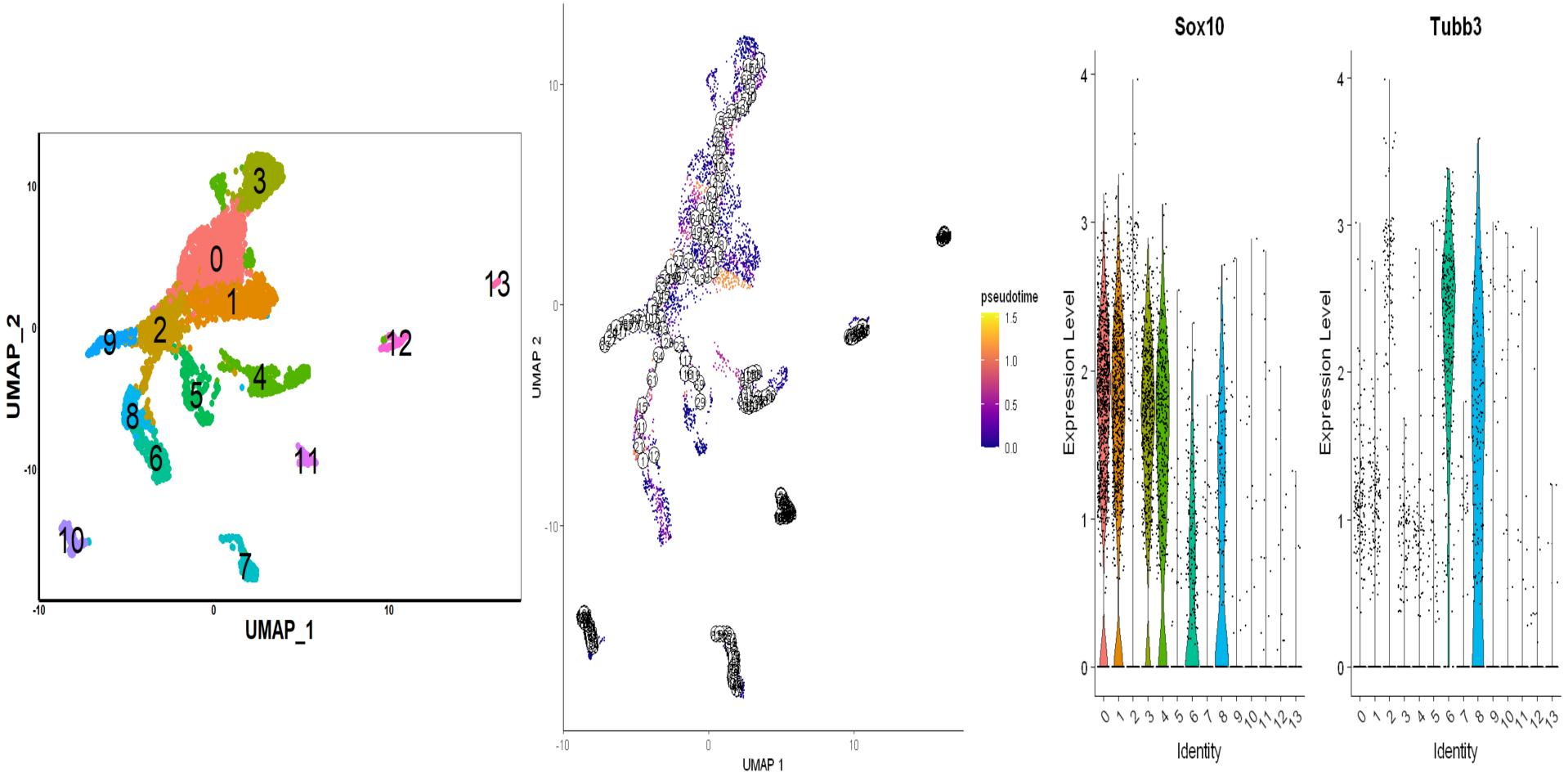

FIGURE 5.21 – Souris *Holstein* traitées avec du GDNF – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

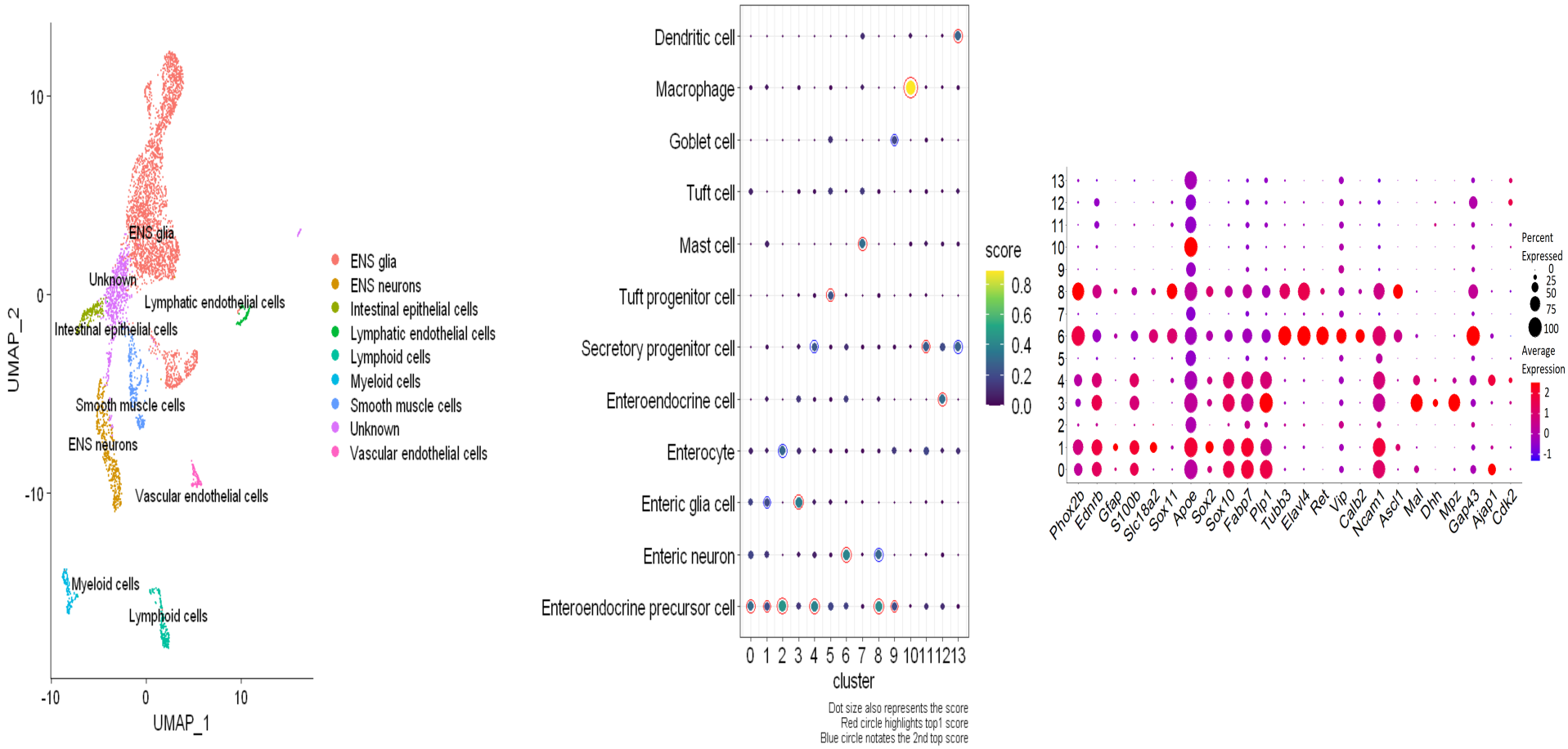

FIGURE 5.22 – Souris *Holstein* traitées avec du GDNF – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

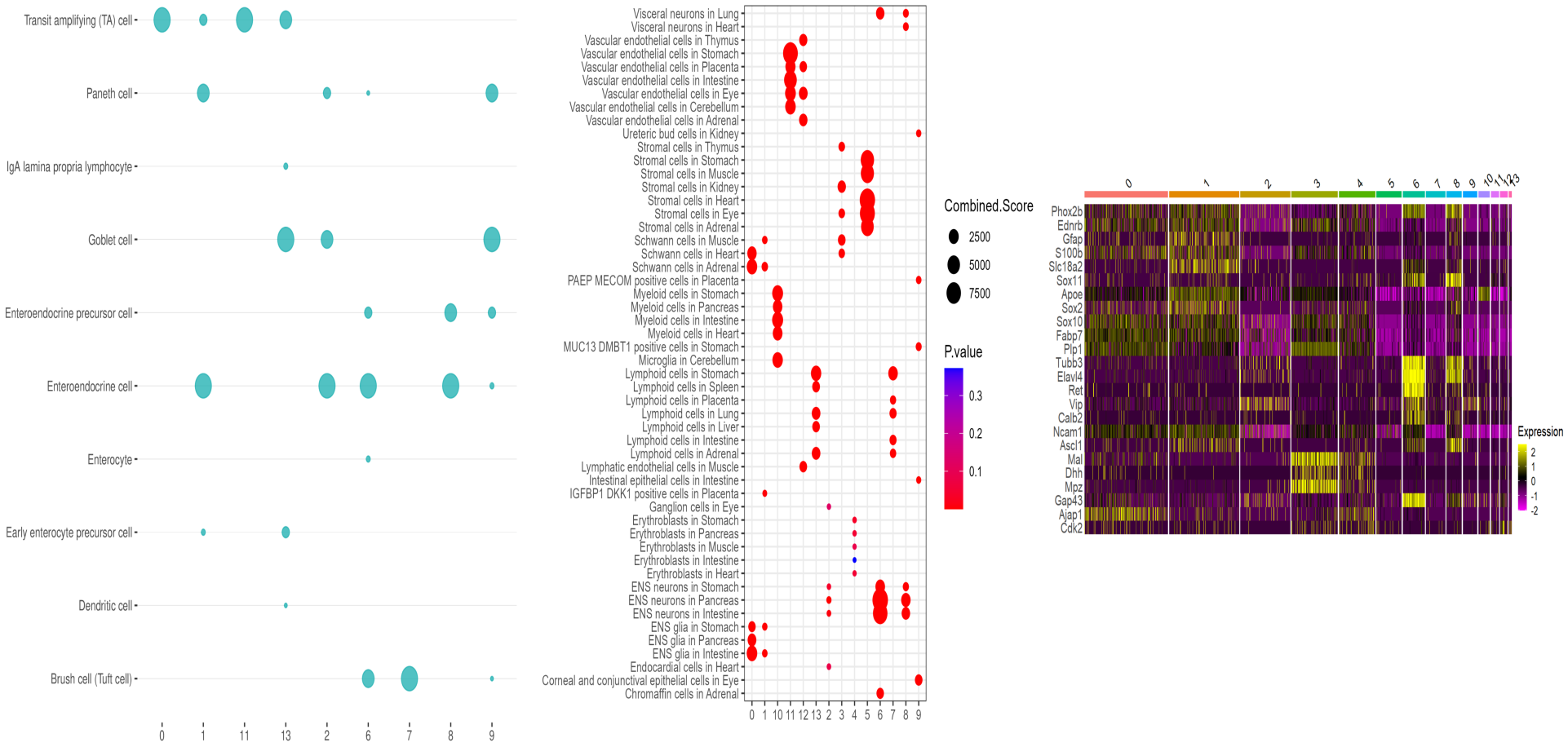

FIGURE 5.23 – Souris *Holstein* traitées avec du GDNF – Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.3.8 Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* traitées avec du GDNF

Enfin, en ce qui concerne les cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* ayant été traitées avec du GDNF (figures 5.24, 5.25 et 5.26), ScType confirme leur présence dans les sous-populations 0, 2 et 5, mais ScMayoMap confirme aussi leur présence dans les sous-populations 1 et 8. Enrichr confirme seulement leur présence dans la sous-population 2. Les analyses de l'expression différentielle des gènes montrent une expression plus élevée de ces marqueurs dans la sous-population 1. En ce qui concerne les cellules neuronales, elles sont présentes dans les sous-populations 7 et 8 (aussi confirmé par Enrichr), mais ScMayoMap confirme aussi leur présence dans les sous-populations 2, 3 et un peu dans la sous-population 1 (aussi confirmé par Enrichr). Cependant, il y a des gènes qui sont sous-exprimés dans la sous-population 8, notamment Fabp7 et Plp1 (cellules gliales entériques (May-Zhang *et al.*, 2021)). De plus, le marqueur Sox10 (cellules gliales entériques (May-Zhang *et al.*, 2021)) y est plus fortement exprimé, ainsi que Sox11 (Morarach *et al.*, 2021) dans la sous-population 7. En ce qui concerne les marqueurs associés aux cellules de Schwann, ils sont plus fortement exprimés dans les sous-populations 0 (Mal, Dhh et Mpz (May-Zhang *et al.*, 2021)), ainsi que Gap43 (Carrió *et al.*, 2019) dans la sous-population 7. Enrichr confirme la présence de cellules de Schwann dans les sous-populations de cellules 0, et 3. EasyCellType confirme la présence de cellules entéroendocrines pour les sous-populations mentionnées, mais détecte aussi des cellules amplificatrices de transit dans la sous-population 0. En ce qui concerne les trajectoires avec Monocle 3, il y a plusieurs cycles qui traversent les sous-populations 0, 1, 2, 3 et 5, ainsi qu'une trajectoire traversante les sous-populations des cellules gliales (3 mais aussi un peu dans le 1), jusqu'à la sous-population 7 (cellules neuronales).

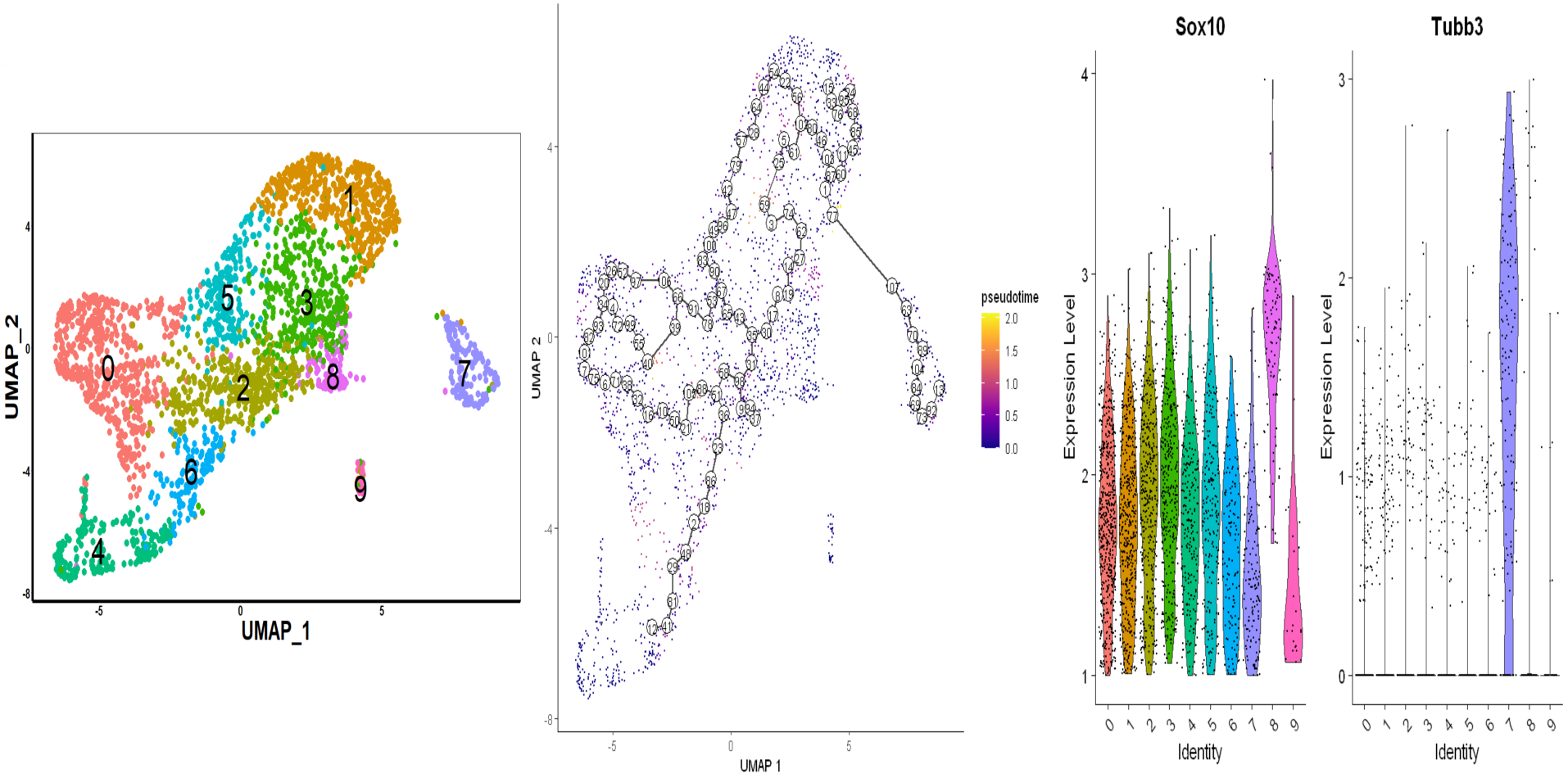

FIGURE 5.24 – Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* traitées avec du GDNF – Première partie.

(À gauche) : Projection UMAP des regroupements de cellules par Seurat. **(Au centre) :** Projection UMAP avec l'analyse de la trajectoire et du pseudotemps par Monocle 3. **(À droite) :** Diagramme en violon pour l'expression des gènes Sox10 et Tubb3 par Seurat.

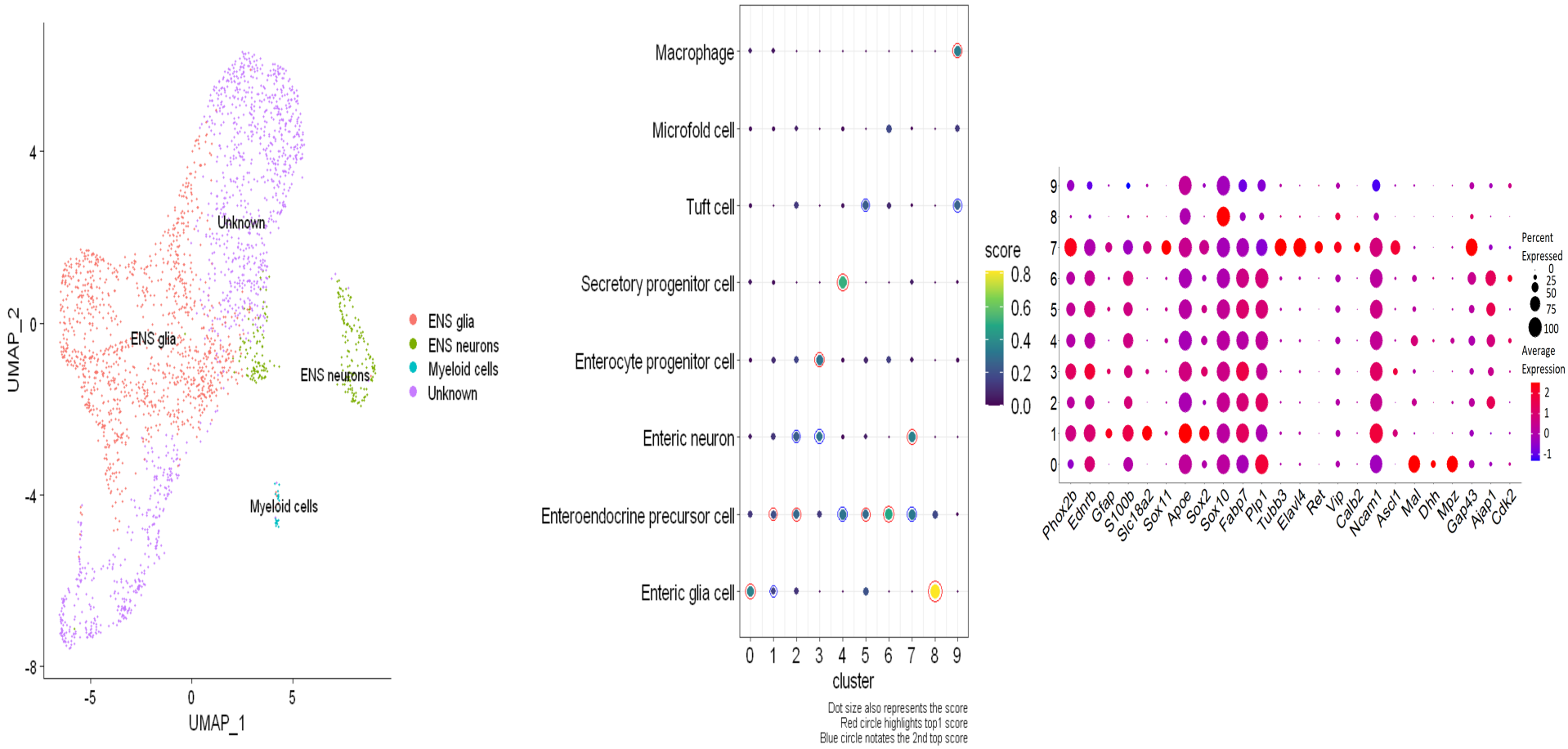

FIGURE 5.25 – Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* traitées avec du GDNF – Deuxième partie.

(À gauche) : Projection UMAP des prédictions des types de cellules par ScType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par ScMayoMap. **(À droite) :** Diagramme à points de l'analyse de l'expression différentielle des gènes par Seurat.

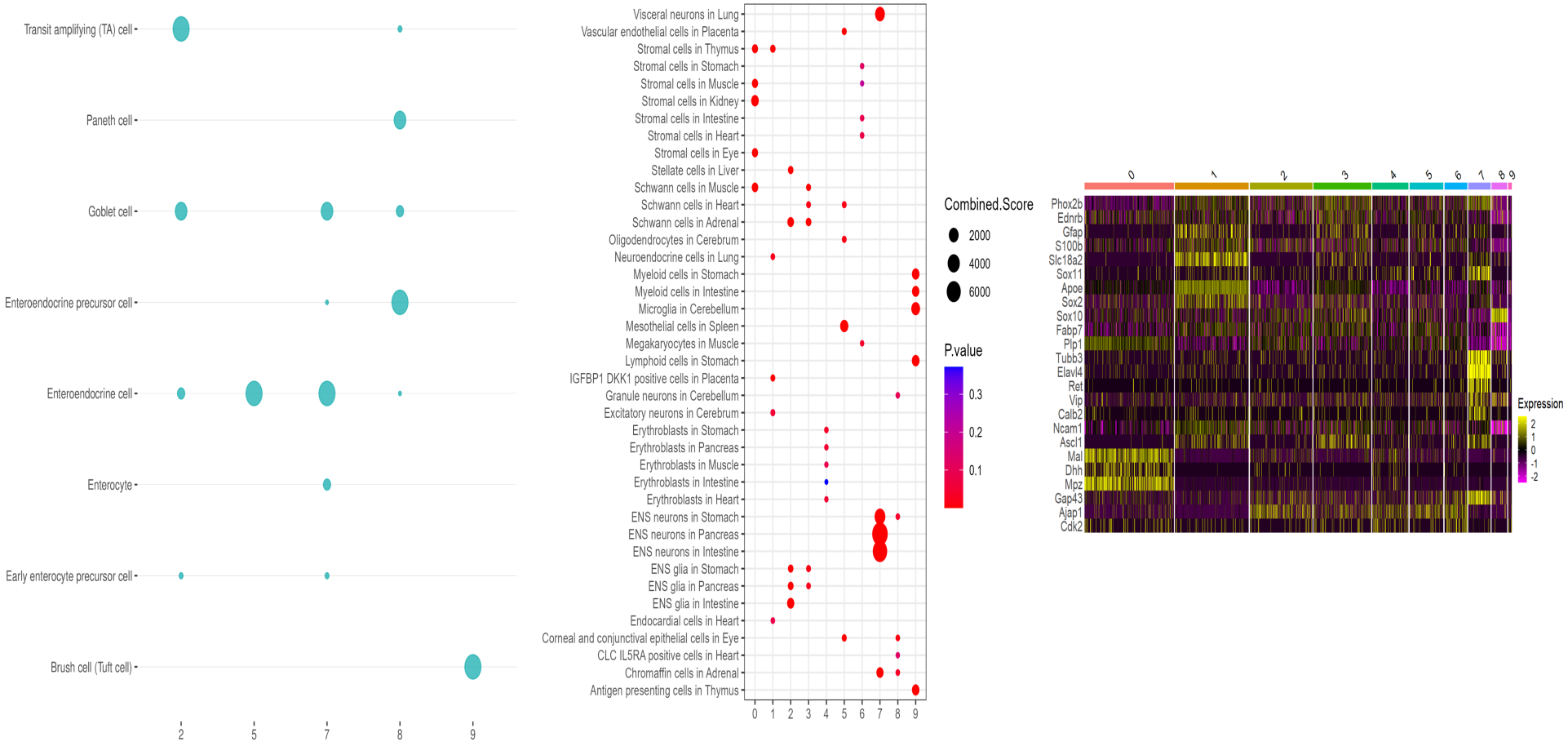

FIGURE 5.26 – Cellules exprimant Sox10 et Tubb3 provenant des souris *Holstein* traitées avec du GDNF – Troisième partie.

(À gauche) : Diagramme à points des prédictions des types de cellules par EasyCellType. **(Au centre) :** Diagramme à points des prédictions des types de cellules par Enrichr. **(À droite) :** Carte de chaleur de l'analyse de l'expression différentielle des gènes par Seurat.

5.4 Synthèse du chapitre

Les jeux de données que nous avons utilisés provenaient d'expériences en laboratoire réalisées par Soret et al. (2021). Ces données provenaient de souris de type sauvage avec et sans traitement au dextran sulfate de sodium (qui induit une colite inflammatoire), ainsi que des souris *Holstein* avec et sans traitement au GDNF (qui restaure les niveaux de cellules gliales et neuronales). Les séquences ont été alignées sur le génome de référence *refdata-gex-mm10-2020-A* grâce à Cell Ranger sur la plateforme de Calcul Canada (l'Alliance). Les résultats ont ensuite été téléchargés et ils ont subi un prétraitement ainsi qu'un traitement avec Seurat. Ce dernier aussi fait une analyse de l'expression différentielle avec une liste de gènes spécifiques aux différentes sous-populations de cellules gliales et neuronales. Ensuite, Monocle 3 a analysé la trajectoire et le pseudotemps de ces différentes sous-populations et ces dernières ont ensuite été annotées avec les différents outils d'identifications de types cellulaires. Par la suite, les cellules exprimant Sox10 (gliales) et Tubb3 (neurones) ont été extraites et elles ont été analysées avec la même méthodologie citée précédemment.

Dans leur ensemble, les outils ont pu facilement détecter et identifier les différents types de cellules et ils avaient en général tous des résultats qui étaient similaires et qui concordaient entre eux. De plus, les analyses de l'expression différentielle avec les gènes spécifiques aux différentes sous-populations de cellules gliales et neuronales confirmaient ces résultats. Ces derniers seront interprétés et discutés dans le prochain chapitre à l'aide d'une comparaison avec la littérature afin de faire une critique de l'efficacité des outils ayant été intégrés dans notre flux de travail.

CHAPITRE VI

DISCUSSION, CONCLUSION ET CRITIQUE

Les résultats des analyses effectuées par notre flux de travail démontrent un consensus général entre les outils, notamment sur le plan de la détection des cellules gliales et neuronales lorsqu'on les compare avec la liste des gènes qui sont spécifiques à différentes sous-populations de cellules gliales et neuronales. Cependant, il est nécessaire d'interpréter ces résultats de manière plus approfondie et de bien les expliquer à l'aide de la littérature.

6.1 Discussion et conclusion pour les analyses

Nous avons remarqué une grande hétérogénéité dans les regroupements des cellules dans toutes les expériences avant l'extraction de cellules exprimant Sox10 et Tubb3, telles que la présence de cellules immunitaires, musculaires, vasculaires et épithéliales. Cela reflète les différents processus biologiques dans l'intestin et des études mentionnent la possibilité de communication entre les cellules immunitaires, neuronales et épithéliales.

Par exemple (Stakenborg *et al.*, 2020) ont mentionné que les mastocytes, les macrophages et les cellules lymphoïdes jouent des rôles essentiels dans la régulation du système intestinal et qu'elles peuvent communiquer avec les cellules gliales et neuronales. Par exemple, s'il y a une inflammation ou la présence d'agents in-

fectieux dans la barrière épithéliale de l'intestin, les cellules gliales et neuronales vont activer ces cellules immunitaires afin de réduire l'inflammation (lymphocytes) et phagocyter des agents infectieux (macrophages). En retour, ces cellules immunitaires encouragent la survie des neurones et à activer le transit intestinal (macrophages). C'est aussi présent dans le processus de la détection de la douleur par les nerfs sensitifs lorsqu'il y a une inflammation (mastocytes) ainsi que dans l'homéostasie dans la barrière intestinale (cellules lymphoïdes). De plus, même s'il y a la présence de cellules immunitaires dans les cellules ayant été extraites, c'est possiblement parce que ces types de cellules peuvent aussi interagir avec des facteurs neurotropes tels que le GDNF, VIP et RET (Stakenborg *et al.*, 2020).

Pour les autres types de cellules, d'autres études de séquençage de l'ARN de cellule unique révèlent aussi la présence de celles-ci (cellules stromales, entéroendocrines, vasculaires, endothéliales, muscles lisses et immunitaires) lors de l'extraction des cellules de la crête neurale, due à leur interaction avec les cellules du système nerveux entérique via leur ligand et récepteur (Drokhlyansky *et al.*, 2020).

En sachant cela, il est normal de constater que, dans tous les jeux de données n'ayant pas encore subi d'extraction de cellules exprimant Sox10 et Tubb3, les sous-populations de cellules sont très hétérogènes avec la présence de cellules gliales, neuronales et de Schwann, ainsi que d'autres types de cellules. De plus, leur présence dans les cellules exprimant Sox10 et Tubb3 ayant été extraites est aussi due à ces phénomènes expliqués par (Stakenborg *et al.*, 2020) et aussi détectée par (Drokhlyansky *et al.*, 2020). Cela pourrait aussi expliquer la raison de la présence de trajectoires détectées par Monocle 3 qui interconnectent quasiment toutes les sous-populations de cellules.

Pour les cellules provenant des souris *Holstein* avec et sans traitement, ainsi que pour les souris de type sauvages ayant été traitées avec du DSS, nous avons aussi

constaté un plus grand nombre de sous-populations de cellules comparées à celles des souris de type sauvage sans traitement. Cela est surement dû au fait que les trois premières étaient toutes dans des conditions anormales comparées à celles des souris de type sauvage sans traitement. De plus, lorsqu'on compare les projections UMAP des jeux de données des souris *Holstein* avec et sans traitement au GDNF, nous avons constaté que les projections étaient assez similaires, même après l'extraction de cellules exprimant les marqueurs Sox10 et Tubb3. Le même phénomène s'appliquait aussi en effectuant la comparaison entre les souris de type sauvage avec et sans traitement au DSS. Ce phénomène pourrait s'expliquer parce que chaque comparaison se faisait entre des souris du même type, donc les regroupements de cellules auraient été assez similaires.

Par exemple, pour les souris de type sauvages ayant été traitées avec du DSS, il est connu que ce dernier induit une colite inflammatoire dans les intestins, car il induit des dommages dans la barrière intestinale. Ce type de traitement est très utilisé dans les expériences afin d'étudier la mobilisation de cellules immunitaires, telles que les cellules T et les macrophages dans l'épithélium intestinal lors de l'inflammation (Chassaing *et al.*, 2014). Cela expliquerait une plus grande hétérogénéité des populations de cellules quand on la compare à celles de type sauvage sans traitement et qui serait comparable à celles provenant des souris *Holstein* avec et sans traitement, en particulier sur le plan du nombre de sous-populations de ces cellules immunitaires.

De plus, pour les souris *Holstein* avec et sans traitement, la raison qui explique la présence de populations plus hétérogènes, en particulier sur le plan des cellules immunitaires, est sûrement le fait que, comme dans le cas avec les souris traitées avec du DSS, il y a une inflammation dans les muqueuses dues à une aganglionose dans le côlon qui causerait une accumulation de matières fécales, comme il avait été mentionné par Soret et al. (2020).

Pour les cellules exprimant Sox10 et Tubb3 provenant des souris de type sauvage n'ayant subi aucun traitement, il est normal de voir une interconnexion des trajectoires entre les sous-populations de cellules dues à une migration de cellules dérivées de la crête neurale dans le côlon qui vont ensuite se différencier en cellules gliales et neuronales (Uesaka *et al.*, 2015), ainsi qu'une proportion de cellules précurseures des cellules de Schwann (provenant de la sous-population 3 (Mal, Dhh et Mpz (Morarach *et al.*, 2021)) qui vont aussi se différencier en neurones (Uesaka *et al.*, 2015).

Étrangement, pour les souris traitées avec du DSS, la trajectoire aurait dû traverser la sous-population 4 qui a une forte signature de neurones (Tubb3, Elavl4 (Lau *et al.*, 2019), Ret (Sasselli *et al.*, 2012), Vip et Calb2 (May-Zhang *et al.*, 2021)) en passant par la sous-population 1 (gliale), car des études ont démontré qu'un traitement avec du DSS va induire des dommages aux neurones et induire un processus de différenciation et de transdifférenciation afin de les régénérer (Belkind-Gerson *et al.*, 2017). Au lieu de cela, elle traverse les sous-populations 0 et 6, qui sont plus ambiguës sur le plan du type de cellules (ScType). Toutefois, en observant les résultats avec Enrichr, on peut voir qu'il y a aussi une légère présence de neurones (un peu plus dans la sous-population 1 pour ScMayoMap), ce qui expliquerait cette trajectoire. De plus, en voyant que la sous-population 1 a aussi une expression des gènes associés aux précurseurs des cellules de Schwann (Mal, Dhh et Mpz (Morarach *et al.*, 2021)) et que la sous-population 0 a une expression des gènes associés aux précurseurs des neurones (Ascl1 (Memic *et al.*, 2016)), il serait possible qu'il y ait un début d'une reprogrammation des cellules de Schwann en neurones dans ces sous-populations, car elles commencent à exhiber des phénotypes typiques aux cellules neuronales (Milichko et Dyachuk, 2020), ce qui expliquerait aussi la trajectoire faite par Monocle 3.

Pour les souris *Holstein* sans traitement, nous avons remarqué un fractionnement

sur le plan des sous-populations de cellules, comparées à celles ayant été traitées avec du GDNF. De plus, Monocle 3 a tracé deux cycles entre les sous-populations 0 et 5 (gliales selon les outils) et 3 (qui expriment plus Gfap (He *et al.*, 2021) et S100b (Brügger *et al.*, 2020), qui sont associées aux cellules gliales), ainsi qu'entre les sous-populations 1 (précurseur des cellules de Schwann (Mal, Dhh et Mpz (Morarach *et al.*, 2021)). Pour la sous-population 2, ce serait aussi probablement ces mêmes cellules qui seraient en transition pendant leur prolifération et qui auraient une modification de leurs phénotypes (Milichko et Dyachuk, 2020). De plus, Monocle 3 a indiqué une trajectoire vers la sous-population 4, qui a été détectée comme étant des neurones par ScType et Enrichr, sûrement en raison de la légère expression du marqueur Vip associé aux neurones (May-Zhang *et al.*, 2021).

Pour les cellules *Holstein* traitées avec le GDNF, il y a manifestement des cellules de Schwann en cycle cellulaire étant donné la présence de plusieurs cycles qui ont été tracés par Monocle 3 entre les sous-populations 0 et 2, ainsi qu'une forte expression de marqueurs associée aux précurseurs des cellules de Schwann dans la sous-population 0 (Mal, Dhh et Mpz (Morarach *et al.*, 2021)). De plus, il y a une trajectoire traversant la sous-population 1, qui est aussi ambiguë selon ScType. Cependant, il y a aussi une signature de cellules gliales selon les analyses de l'expression différentielle des gènes et que la sous-population 7 a été détectée en tant que neurones par les outils, un phénomène qui s'expliquerait par un processus de transdifférenciation des cellules gliales en neurones, comme mentionné dans les études (Belkind-Gerson *et al.*, 2017).

6.2 Critique de la méthode proposée

En ce qui concerne l'efficacité des outils utilisés dans ce flux de travail, les outils comportaient chacun des forces et des faiblesses. Dans le cas de Cell Ranger, sa principale faiblesse était son grand besoin en ressources computationnelles (10x Genomics, 2023), ainsi que le fait qu'il pouvait seulement être lancé dans un environnement Linux ou dans un serveur (10x Genomics, 2020a). Néanmoins, l'implémentation de Paramiko afin de se connecter à Calcul Canada (qui contient déjà ce logiciel (Alliance, 2023a)) facilite son lancement sans devoir sacrifier les ressources computationnelles de la machine locale. De plus, ce flux de travail a été conçu afin d'offrir la possibilité de lancer les analyses sans devoir passer par ces étapes, par exemple en le faisant par la plateforme nuagique de 10x Genomics (10x Genomics, 2023), ce qui le rend beaucoup plus versatile.

En ce qui concerne Seurat, son principal défaut est qu'il n'offre pas la possibilité d'annoter automatiquement les sous-populations de cellules. Son tutoriel implique une annotation manuelle des cellules à partir de la littérature (Hoffman *et al.*, 2023), ce qui demanderait beaucoup de temps afin de faire la recherche pour trouver les marqueurs pour chaque type de cellules, comme dans le cas de ce projet où il a fallu chercher dans les bases de données telles que CellMarker2.0 (Hu *et al.*, 2023a) (Hu *et al.*, 2023b) et PanglaoDB (Franzén *et al.*, 2019b) (Franzén *et al.*, 2019a) et de faire la recherche dans la littérature afin de valider la pertinence des outils utilisés pour les annotations. C'est pourquoi plusieurs outils d'annotation ont été implémentés dans le flux de travail afin de pouvoir faire des comparaisons des résultats.

Pour la trajectoire et le pseudotemps, le plus grand défaut de Monocle 3 est qu'il est nécessaire de choisir le nœud lors des analyses de pseudotemps en utilisant une interface interactive avec Shiny (Trapnell, 2022). Cela ne le rend pas pertinent

pour ce flux de travail dont le but était d'être automatisé. Heureusement, dans une vignette rédigée par un des auteurs (Cao *et al.*, 2020b), il offre la possibilité de choisir directement le nœud directement dans le code de la vignette, donc ce paramètre avait aussi été déplacé dans le code du flux de travail (figure 4.6).

En ce qui concerne ces derniers, ScType avait des annotations assez limitées sur le plan des types de cellules. Il détectait directement les cellules gliales et neuronales dans les sous-populations de cellules, mais ses prédictions comportaient des lacunes, car il ne pouvait pas détecter tous les types de cellules. En effet, puisque les cellules de Schwann sont une sous-population de cellules gliales (Milichko et Dyachuk, 2020), l'outil avait souvent tendance à les prédire en tant que cellules gliales.

Pour ScMayoMap, sa variété d'annotations était plus élevée que ScType, car il pouvait aussi détecter les cellules gliales, les neurones, ainsi que d'autres types de cellules, telles que les cellules immunitaires, épithéliales, entéroendocrines, mais il ne pouvait pas détecter les cellules de Schwann. Néanmoins, sa performance était mieux que celle de ScType, car les auteurs (Yang *et al.*, 2023b) avaient intégré d'autres bases de données, telles que celles utilisées lors de la recherche manuelle des marqueurs les plus exprimés. Ces bases de données étaient la première version de CellMarker (Zhang *et al.*, 2019) (nous avions utilisé la version web de la version 2.0 de cette base de données (Hu *et al.*, 2023a) (Hu *et al.*, 2023b)) et PanglaoDB (Franzén *et al.*, 2019b) (Franzén *et al.*, 2019a) (Yang *et al.*, 2023b).

Dans le cas de Enrichr, il offrait des prédictions plus exhaustives sur le plan des types de cellules que les deux premiers outils, mais il avait tendance à confondre la provenance de ces cellules, par exemple, en prédisant des organes n'ayant pas du tout de lien avec le système intestinal. Les auteurs (Xie *et al.*, 2021) recommandaient un minimum de 20 gènes afin d'avoir de bons résultats, donc le fait

d'avoir utilisé le top 50 des gènes les plus exprimés était pertinent afin d'avoir les meilleures prédictions. Cependant, le fait que d'autres auteurs (Kuleshov *et al.*, 2016) aient aussi mentionné le fait que les bases de données pour les humains et les souris ont été fusionnées sur Enrichr et que cet outil n'offre pas l'option de choisir spécifiquement l'organisme peut influencer positivement ou négativement les prédictions par cet outil. Néanmoins, la plus grande force de cet outil est qu'il a intégré une très grande variété de bases de données que les trois autres outils (Xie *et al.*, 2021) et que la version R de cet outil permet aussi de choisir la base de données la plus pertinente pour les recherches (Jawaid, 2023b). Cette possibilité de choisir la base de données a aussi été prise en compte lors de son implémentation dans le flux de travail en modifiant le code de la version R de cet outil de manière à déplacer le choix de ce paramètre au début du code dans le bloc-notes Jupyter au lieu de le modifier directement dans le code rédigé par l'auteur de l'implémentation R de cet outil (Jawaid, 2023b).

Enfin, en ce qui concerne EasyCellType, sa performance s'est avérée médiocre comparée aux autres outils, car il n'a détecté aucune cellule gliale, neurone, ni de cellule de Schwann. Ses prédictions étaient souvent des cellules entéroendocrines, des cellules amplificatrices de transit, ainsi que ces cellules caliciformes. Pourtant, les auteurs mentionnaient que les meilleurs choix étaient CellMarker, ainsi que le test GSEA pour effectuer les prédictions (Li *et al.*, 2023) (ces paramètres ont été utilisés lors des analyses du jeu de données). Même si ces cellules sont toutes présentes dans ce côlon (Umar, 2010) dans cette expérience, cet outil n'était pas adapté. Cependant, il reste pertinent dans le flux de travail, dans l'éventualité de son utilisation pour des analyses avec des jeux de données provenant d'autres sources, telles que des organes différents ou des espèces différentes. Ces choix de paramètres ont été considérés lors de l'implémentation de la version R de cet outil (Li, 2022b), en déplaçant aussi ces paramètres au début du code.

La grande force de ce flux de travail est que tous ces outils se complémentent sur le plan de leurs prédictions et, dans la grande majorité des cas, les prédictions étaient assez similaires sur le plan des cellules d'intérêt. Les prédictions des cellules concordaient majoritairement avec la liste des marqueurs que nous avons compilée à partir des bases de données ainsi que la littérature. De plus, le fait que ce flux de travail offre un grand choix de paramètres et de cheminements possible le rend plus personnalisable. Enfin, le fait de sauvegarder automatiquement l'objet Seurat et de pouvoir recommencer les analyses directement avec celui-ci rend ces dernières beaucoup plus rapides, car il n'est pas nécessaire de les répéter dans leur entièreté, par exemple, en repassant par les analyses avec Cell Ranger, ce qui rend cette méthode plus reproductible.

Pour ce qui est des faiblesses de cette méthode, il y a un grand risque de bogue, par exemple, si les paramètres ont été mal définis ou mal formatés (ex. : un mauvais format de fichier, de répertoire ou du nom de la base de données ou des types de tissus). Il est donc important de consulter la littérature de ces outils afin d'avoir une liste exhaustive des choix possibles de leurs paramètres. De plus, le fait que ce flux de travail ait été rédigé dans un bloc-notes Jupyter (Kluyver *et al.*, 2016) au lieu d'un flux de travail plus spécialisé, telle que Nextflow (Di Tommaso *et al.*, 2017) et Snakemake (Köster et Rahmann, 2012), peut le rendre moins attrayant pour quelqu'un qui rechercherait un flux de travail qui serait plus purement automatisé dans son entièreté. Néanmoins, son implémentation avec le flux de travail Script of Script sous la forme d'un bloc-notes permet de lancer toutes les analyses les unes à la suite des autres, ou séparément en utilisant des fonctions (Wang et Peng, 2019). Enfin, ce flux de travail requiert l'installation de beaucoup d'outils, dont les versions pourraient possiblement changer et ne plus être compatibles avec le code. Néanmoins, le fait que les versions des outils soient archivées, par exemple dans GitHub (en particulier pour EasyCellType, dont la version 0.99 est requise

pour le flux de travail (Li, 2022b) (tableau A.1)) les rend quand même disponibles pour leur installation.

Malgré tout cela, ce flux de travail demeure pertinent en tant qu'outil d'assistance sur le plan de l'identification des types de cellules, avec ou sans l'aide de la littérature. Nous recommandons quand même l'utilisation de tous les outils en même temps lors des analyses afin d'avoir une vue d'ensemble de toutes les possibilités concernant les types de cellules.

6.2.1 Comparaison sur le plan des performances des outils ayant été implémentés dans le flux de travail

Chacun des outils ayant été implémentés dans le flux de travail a offert des performances différentes. Sur le plan du temps, le flux de travail de Cell Ranger est l'outil qui a été le plus lent, pouvant aller jusqu'à trois jours pour faire les analyses sur la plateforme de Calcul Canada (l'Alliance). Cela nous obligeait de nous déconnecter de Paramiko et de revenir quelques jours plus tard suite à la réception de la notification de la fin des analyses. De plus, il était nécessaire de relancer le pipeline afin de vérifier le statut des analyses et pour télécharger les résultats. Cet outil générait plusieurs dizaines de fichiers dont la plupart étaient soit seulement interprétables avec d'autres outils qui ne sont pas implémentés dans le flux de travail, soit qu'ils n'étaient pas pertinents ou qu'ils étaient beaucoup trop lourds à télécharger. C'est pourquoi nous avions limité le choix des fichiers à télécharger à seulement ceux qui étaient nécessaires pour le flux de travail de Seurat ainsi que pour la visualisation sur un navigateur web ou avec Loupe Browser. Enfin, ces facteurs nous ont obligés à ne pas l'inclure lors des exécutions des flux de travail de manières successives pour nos analyses comparées aux autres outils.

Pour Seurat, certains échantillons requéraient jusqu'à plusieurs dizaines de mi-

nutes pour les analyses et pour le flux de travail des annotations, cela durait quelques minutes en fonction des outils, ce qui facilitait grandement les analyses de manière successive. Les fichiers générés étaient aussi moins nombreux et moins lourds à télécharger sur la machine locale. Cependant, les résultats générés par le flux de travail de Seurat n'étaient pas forcément compatibles pour les outils d'annotation, donc nous avions dû ajouter des lignes de codes pour extraire quelques données à utiliser en entrée pour le flux de travail des outils.

Enfin, certaines données générées par le flux de travail de Seurat et des annotations des outils n'étaient pas visuellement interprétables, donc nous avions dû ajouter des lignes de codes pour convertir ces résultats en graphiques. Néanmoins les performances des outils étaient satisfaisantes pour générer assez de résultats pour valider des analyses ainsi que pour leur interprétation.

6.3 Synthèse du chapitre

En résumé, les résultats de notre flux de travail ont pu détecter certains phénomènes biologiques dans nos jeux de données. En effet, pour les 4 conditions avant d'extraire les cellules exprimant Sox10 et Tubb3, notre flux de travail a su démontrer qu'il y avait une grande hétérogénéité en matière des types de cellules, notamment en détectant des cellules autres que les cellules gliales et neuronales telles que des cellules immunitaires endothéliales et musculaires, ce qui reflète les types de cellules qui sont présents dans les tissus intestinaux (Stakenborg *et al.*, 2020). Ce phénomène est particulièrement présent chez les souris de type sauvage sans traitement.

De plus, il a aussi pu détecter une augmentation des cellules immunitaires chez les souris ayant été traitées avec du dextran sulfate de sodium étant donné l'inflammation causée par celui-ci (Chassaing *et al.*, 2014). De plus, les projections

UMAP était assez similaires à celles des souris *Holstein* avec et sans traitement au GDNF puisque ce sont tous les trois des conditions anormales.

Suite à l'extraction des cellules exprimant Sox10 et Tubb3, il a pu identifier les cellules en tant que cellules gliales et neuronales pour chacun des jeux de données. On a pu constater une augmentation du nombre de cellules gliales et neuronales chez les souris *Holstein* ayant été traitées avec du GDNF, ce qui confirme l'efficacité de ce traitement.

Seurat a pu générer des projections UMAP permettant de bien séparer et identifier les sous-populations de cellules à l'aide des analyses de l'expression différentielle et des annotations. Les analyses de la trajectoire ainsi que du pseudotemps par Monocle 3 ont aussi pu illustrer une différenciation ainsi qu'une transdifférenciation des cellules gliales vers les cellules neuronales chez les souris *Holstein* ayant été traitées avec du GDNF. Il a aussi pu détecter la présence d'un cycle cellulaire chez les cellules de Schwann (Belkind-Gerson *et al.*, 2017).

En ce qui concerne les outils d'annotation, ScType, Enrichr et scMayoMap ont offert une performance satisfaisante pour l'identification des types de cellules. Chacun d'entre eux se complémentait en détectant les types de cellules que les autres outils ne pouvaient pas détecter. Par contre, EasyCellType n'a pas réussi à détecter les types de cellules d'intérêt, mais il reste pertinent s'il est utilisé pour d'autres types d'échantillon, tel que des cellules immunitaires (Li *et al.*, 2023).

Enfin, le flux de travail de Cell Ranger est celui qui a pris le plus de temps pour faire les analyses et qui a généré le plus de résultats qui n'étaient pas pertinents comparé aux flux de travail de Seurat et des annotations. Pour ces derniers, il a fallu ajouter des lignes de codes pour convertir les données afin qu'elles soient compatibles entre les outils et qu'elles soient visuellement interprétables.

Notre flux de travail offre plusieurs avantages, tels que la possibilité de lancer les outils séparément et successivement, et de pouvoir le personnaliser à l'aide des paramètres ou directement dans le code à l'aide de Script of Script dans le bloc-notes Jupyter. Par contre, il n'est pas sans défaut, puisqu'il requiert une bonne connaissance des paramètres associés aux outils afin qu'il puisse bien fonctionner. De plus, il requiert l'installation de plusieurs outils ainsi que de leurs dépendances. Néanmoins sa performance avec les jeux de données le rend pertinent pour des utilisations dans des recherches ultérieures.

ANNEXE

TABLEAU A.1 – Liste des outils utilisés.

Type	**Nom**	**Version**	**Références**
Système d'exploitation	Microsoft Windows 10 Famille	22H2	(Microsoft, 2022)
IDE	Jupyter Notebook	6.4.12	(Kluyver *et al.*, 2016)
	PyCharm Community Edition	2020.1	(JetBrains, 2020)
	RStudio	2021.09.0	(R Core Team, 2021b)
Langage	Bash (Unix shell)	5.2	(Project, 2022)
	Python	3.7	(Van Rossum et Drake Jr, 2009)
			(Van Rossum et Drake, 2009)
	R	4.1	(R Core Team, 2021b)
Module Python	jupyterlab	3.4.7	(Kluyver *et al.*, 2016)
	notebook	6.4.12	(Kluyver *et al.*, 2016)
	paramiko	2.12.0	(Forcier, 2023a)
			(Forcier, 2023b)
	sos	0.23.4	(Peng *et al.*, 2018)
			(Wang et Peng, 2019)
	sos-notebook	0.23.4	(Peng *et al.*, 2018)
			(Wang et Peng, 2019)
	sos-python	0.18.4	(Peng *et al.*, 2018)
			(Wang et Peng, 2019)
	sos-r	0.19.6	(Peng *et al.*, 2018)
			(Wang et Peng, 2019)
	jupyterlab-sos	0.8.8	(Peng *et al.*, 2018)
			(Wang et Peng, 2019)
Package R	data.tree	1.0.0	(Glur, 2020)
	dplyr	1.1.0	(Wickham *et al.*, 2023b)
	EasyCellType	0.99.0	(Li, 2022b)
			(Li *et al.*, 2023)
			(Li, 2023)
			(Li, 2022a)
			(Center, 2023)
	enrichR	3.2	(Jawaid, 2023a)
			(Kuleshov *et al.*, 2016)

	ggplot2	3.4.1	(Wickham, 2016)
	ggraph	2.1.0	(Pedersen, 2022)
	HGNChelper	0.8.1	(Waldron et Riester, 2019)
	igraph	1.4.2	(Csardi et Nepusz, 2006)
	leidenbase	0.1.4	(Ewing, 2022)
	magrittr	2.0.3	(Bache et Wickham, 2022)
	monocle3	1.0.0	(Trapnell *et al.*, 2014)
			(Qiu *et al.*, 2017a)
			(Qiu *et al.*, 2017b)
			(Cao *et al.*, 2019)
			(Trapnell, 2022)
	openxlsx	4.2.5.2	(Schauberger et Walker, 2023)
	org.Hs.eg.db	3.14.0	(Carlson, 2021a)
	org.Mm.eg.db	3.14.0	(Carlson, 2021b)
	parallel	4.1.2	(R Core Team, 2021b)
	patchwork	1.1.1	(Pedersen, 2020)
	readxl	1.4.2	(Wickham et Bryan, 2023)
	scMayoMap	0.2.0	(Yang et Zhang, 2023)
			(Yang *et al.*, 2023b)
			(Yang *et al.*, 2023a)
	ScType	1.0	(Ianevski *et al.*, 2022)
			(Ianevski, 2023b)
	Seurat	4.1.0	(Hao *et al.*, 2021)
	SeuratData	0.2.1	(Satija *et al.*, 2019)
	SeuratWrappers	0.3.0	(Satija *et al.*, 2020)
	tidyverse	2.0.0	(Wickham *et al.*, 2019)
			(Wickham *et al.*, 2023a)
	xlsx	0.6.5	(Dragulescu et Arendt, 2020)
Plateforme	10x Cloud Analysis	N/A	(10x Genomics, 2023)
	Calcul Canada (Cedar)	N/A	(Alliance, 2022)
			(Alliance, 2023b)
			(Alliance, 2023c)
			(Alliance, 2023a)

	Globus	N/A	(Foster, 2011)
			(Allen *et al.*, 2012)
Logiciel	Cell Ranger	6.1.1	(Zheng *et al.*, 2017)
			(10x Genomics, 2021a)
			(10x Genomics, 2020c)
			(10x Genomics, 2020b)
	Loupe Browser	6.0.0	(10x Genomics, 2021b)
	PuTTY	0.78	(Tatham, 2022)
Base de données	CellMarker2.0	2.0	(Hu *et al.*, 2023a)
			(Hu *et al.*, 2023b)
	PanglaoDB	N/A	(Franzén *et al.*, 2019b)
			(Franzén *et al.*, 2019a)
	scRNA-tools	N/A	(Zappia *et al.*, 2018a)
			(Zappia *et al.*, 2018a)

La liste des marqueurs spécifiques à chaque type de cellule a été compilée à l'aide des bases de données CellMarker (Zhang *et al.*, 2019) (Hu *et al.*, 2023a), PanglaoDB (Franzén *et al.*, 2019b) ainsi que la littérature. De plus, elle a été utilisée pour faire les analyses de l'expression différentielle des gènes par Seurat afin de faire une identification manuelle des sous-populations de cellules. Les résultats de ces analyses sont sous la forme de diagramme à point et de carte de chaleur.

TABLEAU A.2 – Liste des marqueurs par type de cellules.

Cellules	**Marqueurs**	**Source**
Cellules de la crête neurale entérique	Phox2b	(Sasselli *et al.*, 2012)
	Ednrb	(Sasselli *et al.*, 2012)
Cellules gliales	Gfap	(He *et al.*, 2021)
	S100b	(Brügger *et al.*, 2020)
	Slc18a2	(Drokhlyansky *et al.*, 2020)
	Ncam1	(He *et al.*, 2021)
Cellules gliales entériques	Sox11	(Morarach *et al.*, 2021)
	Apoe	(Morarach *et al.*, 2021)
	Sox2	(Han *et al.*, 2019)
	Sox10	(May-Zhang *et al.*, 2021)
	Fabp7	(Lau *et al.*, 2019)
	Plp1	(Lau *et al.*, 2019)
Neurones	Tubb3	(Lau *et al.*, 2019)
	Elavl4	(Lau *et al.*, 2019)
Neurones entériques	Ret	(Sasselli *et al.*, 2012)
	Vip	(May-Zhang *et al.*, 2021)
	Calb2	(May-Zhang *et al.*, 2021)
Cellules précurseures des neurones entériques	Ascl1	(Memic *et al.*, 2016)
Précurseurs des cellules de Schwann	Mal	(Morarach *et al.*, 2021)
	Dhh	(Morarach *et al.*, 2021)
	Mpz	(Morarach *et al.*, 2021)
	Gap43	(Carrió *et al.*, 2019)
Cellules de Schwann présynaptiques	Ajap1	(Castro *et al.*, 2020)
Cycle cellulaire des cellules de Schwann	Cdk2	(Tikoo *et al.*, 2000)

Ceci est la liste des outils ainsi que leur version qui a été installée pour le flux de travail. Il inclut la liste des langages de programmation ainsi que les versions des environnements de développement intégrés (IDE) utilisés pour le développement du flux de travail.

RÉFÉRENCES

10x Genomics (2020a). Installing Cell Ranger -Software -Single Cell Gene Expression -Official 10x Genomics Support, https ://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/tutorial_in. Récupéré le 2024-01-08 de https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/tutorial_in

10x Genomics (2020b). Single-Library Analysis with Cell Ranger count -Software -Single Cell Gene Expression . Official 10x Genomics Support. https ://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/count. Récupéré le 2024-01-03 de https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/count

10x Genomics (2020c). What is Cell Ranger ? -Software -Single Cell Gene Expression. Official 10x Genomics Support. https ://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger. Récupéré le 2023-12-19 de https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

10x Genomics (2020d). What is Loupe Browser ? - Software -Single Cell Gene Expression, Official 10x Genomics Support, https ://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/what-is-loupe-cell-browser. Récupéré le 2024-01-03 de https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/what-is-loupe-cell-browser

10x Genomics (2021a). Cell Ranger, (Version 6.1.1), 10x Genomics, [Pipeline d'analyse de données de single cell], https ://www.10xgenomics.com/support/software/cell-ranger. Récupéré de https://www.10xgenomics.com/support/software/cell-ranger

10x Genomics (2021b). Loupe Browser, (Version 6.0.0). [Logiciel de visualisation]. 10x Genomics. https ://www.10xgenomics.com/support/software/loupe-browser. Récupéré de https://www.10xgenomics.com/support/software/loupe-browser

10x Genomics (2023). Cloud Analysis. [Plateforme nuagique], 10x Genomics, https ://www.10xgenomics.com/products/cloud-analysis. Récupéré le 2023-12-19 de https://www.10xgenomics.com/products/cloud-analysis

Adams, G. (2020). A beginner's guide to RT-PCR, qPCR and RT-qPCR. *The Biochemist*, *42*(3), 48–53. http://dx.doi.org/10.1042/BIO20200034. Récupéré le 2024-01-05 de https://portlandpress.com/biochemist/article/42/3/48/225280/A-beginner-s-guide-to-RT-PCR-qPCR-and-RT-qPCR

Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K. et Tuecke, S. (2012). Software as a service for data scientists. *Communications of the ACM*, *55*(2), 81–88. http://dx.doi.org/10.1145/2076450.2076468. Récupéré le 2023-12-19 de https://dl.acm.org/doi/10.1145/2076450.2076468

Alliance, T. (2022). Calcul Canada (L'Alliance), [Plateforme de recherche], The Alliance, https ://alliancecan.ca. Récupéré de https://alliancecan.ca/en

Alliance, T. (2023a). Cellranger, Alliance Doc - Digital Research Alliance of Canada, https ://docs.alliancecan.ca/wiki/Cellranger. Récupéré le 2023-12-18 de https://docs.alliancecan.ca/wiki/Cellranger

Alliance, T. (2023b). Running jobs, Alliance Doc - Digital Research Alliance of Canada, https ://docs.alliancecan.ca/wiki/Running_jobs. Récupéré le 2023-12-18 de https://docs.alliancecan.ca/wiki/Running_jobs

Alliance, T. (2023c). Technical documentation, Alliance Doc - Digital Research Alliance of Canada, https ://docs.alliancecan.ca/wiki/Technical_documentation. Récupéré le 2023-12-18 de https://docs.alliancecan.ca/wiki/Technical_documentation

Alves, M. M., Sribudiani, Y., Brouwer, R. W., Amiel, J., Antiñolo, G., Borrego, S., Ceccherini, I., Chakravarti, A., Fernández, R. M., Garcia-Barcelo, M.-M., Griseri, P., Lyonnet, S., Tam, P. K., Van IJcken, W. F., Eggen, B. J., Te Meerman, G. J. et Hofstra, R. M. (2013). Contribution of rare and common variants determine complex diseases—Hirschsprung disease as a model. *Developmental Biology*, *382*(1), 320–329. http://dx.doi.org/10.1016/j.ydbio.2013.05.019. Récupéré le 2023-12-31 de https://linkinghub.elsevier.com/retrieve/pii/S0012160613002649

Amiel, J. (2001). Hirschsprung disease, associated syndromes, and genetics : a review. *Journal of Medical Genetics*, *38*(11), 729–739. http://dx.doi.org/10.1136/jmg.38.11.729. Récupéré le 2023-12-31 de https://jmg.bmj.com/lookup/doi/10.1136/jmg.38.11.729

Anaparthy, N., Ho, Y.-J., Martelotto, L., Hammell, M. et Hicks, J. (2019). Single-Cell Applications of Next-Generation Sequencing. *Cold Spring Harbor Perspectives in Medicine*, *9*(10), a026898. http://dx.doi.org/10.1101/cshperspect.a026898. Récupéré le 2023-08-25 de http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a026898

Anderson, R. B., Stewart, A. L. et Young, H. M. (2006). Phenotypes of neural-crest-derived cells in vagal and sacral pathways. *Cell and Tissue Research*, *323*(1), 11–25. http://dx.doi.org/10.1007/s00441-005-0047-6. Récupéré le 2023-12-31 de http://link.springer.com/10.1007/s00441-005-0047-6

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J. et Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163–172. http://dx.doi.org/10.1038/s41590-018-0276-y. Récupéré le 2024-01-04 de https://www.nature.com/articles/s41590-018-0276-y

Bache, S. M. et Wickham, H. (2022). *magrittr : A forward-pipe operator for R*. manual

Belkind-Gerson, J., Graham, H. K., Reynolds, J., Hotta, R., Nagy, N., Cheng, L., Kamionek, M., Shi, H. N., Aherne, C. M. et Goldstein, A. M. (2017). Colitis promotes neuronal differentiation of Sox2+ and PLP1+ enteric cells. *Scientific Reports*, *7*(1), 2525. http://dx.doi.org/10.1038/s41598-017-02890-y. Récupéré le 2024-01-07 de https://www.nature.com/articles/s41598-017-02890-y

Bell, K. L., Petit, R. A., Cutler, A., Dobbs, E. K., Macpherson, J. M., Read, T. D., Burgess, K. S. et Brosi, B. J. (2021). Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and Evolution*, *11*(22), 16082–16098. http://dx.doi.org/10.1002/ece3.8281. Récupéré le 2024-01-02 de https://onlinelibrary.wiley.com/doi/10.1002/ece3.8281

Bergeron, K.-F., Cardinal, T., Touré, A. M., Béland, M., Raiwet, D. L., Silversides, D. W. et Pilon, N. (2015). Male-Biased Aganglionic Megacolon in the TashT Mouse Line Due to Perturbation of Silencer Elements in a Large Gene Desert of Chromosome 10. *PLOS Genetics*, *11*(3), e1005093. http://dx.doi.org/10.1371/journal.pgen.1005093. Récupéré le 2023-04-03 de https://dx.plos.org/10.1371/journal.pgen.1005093

Bischoff, A., Levitt, M. A. et Peña, A. (2011). Total colonic aganglionosis : a surgical challenge. How to avoid complications ? *Pediatric Surgery International*, *27*(10), 1047–1052. http://dx.doi.org/10.1007/s00383-011-2960-y. Récupéré le 2023-12-31 de http://link.

springer.com/10.1007/s00383-011-2960-y

Bondurand, N. et Southard-Smith, E. M. (2016). Mouse models of Hirschsprung disease and other developmental disorders of the enteric nervous system : Old and new players. *Developmental Biology*, *417*(2), 139–157. http://dx.doi.org/10.1016/j.ydbio.2016.06.042. Récupéré le 2023-12-31 de https://linkinghub.elsevier.com/retrieve/pii/S0012160616301336

Brügger, M. D., Valenta, T., Fazilaty, H., Hausmann, G. et Basler, K. (2020). Distinct populations of crypt-associated fibroblasts act as signaling hubs to control colon homeostasis. *PLOS Biology*, *18*(12), e3001032. http://dx.doi.org/10.1371/journal.pbio.3001032. Récupéré le 2023-12-30 de https://dx.plos.org/10.1371/journal.pbio.3001032

Buenrostro, J. D., Wu, B., Chang, H. Y. et Greenleaf, W. J. (2015). ATAC-seq : A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, *109*(1). http://dx.doi.org/10.1002/0471142727.mb2129s109. Récupéré le 2024-01-03 de https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/0471142727.mb2129s109

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. et Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. http://dx.doi.org/10.1038/nbt.4096. Récupéré le 2023-04-03 de http://www.nature.com/articles/nbt.4096

Butler Tjaden, N. E. et Trainor, P. A. (2013). The developmental etiology and pathogenesis of Hirschsprung disease. *Translational Research*, *162*(1), 1–15. http://dx.doi.org/10.1016/j.trsl.2013.03.001. Récupéré le 2023-12-31 de https://linkinghub.elsevier.com/retrieve/pii/S1931524413000716

Cantrell, V. A. (2004). Interactions between Sox10 and EdnrB modulate penetrance and severity of aganglionosis in the Sox10Dom mouse model of Hirschsprung disease. *Human Molecular Genetics*, *13*(19), 2289–2301. http://dx.doi.org/10.1093/hmg/ddh243. Récupéré le 2024-01-02 de https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddh243

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., Zager, M. A., Aldinger, K. A., Blecher-Gonen, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D., Steemers, F. J., Glass, I. A., Trapnell, C. et Shendure, J. (2020a). A human cell atlas of fetal gene expression. *Science*, *370*(6518), eaba7721. http://dx.doi.org/10.1126/science.aba7721. Récupéré le 2023-08-29 de https://www.science.org/doi/10.1126/

science.aba7721

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C. et Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502. http://dx.doi.org/10.1038/s41586-019-0969-x. Récupéré le 2023-12-19 de https://www.nature.com/articles/s41586-019-0969-x

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C. et Shendure, J. (2020b). Calculating Trajectories with Monocle 3 and Seurat, Github, https ://htmlpreview.github.io/ ?https ://github.com/satijalab/seurat-wrappers/blob/master/docs/monocle3.html. Récupéré le 2023-12-20 de https://htmlpreview.github.io/?https://github.com/satijalab/seurat-wrappers/blob/master/docs/monocle3.html

Cardinal, T., Bergeron, K.-F., Soret, R., Souchkova, O., Faure, C., Guillon, A. et Pilon, N. (2020). Male-biased aganglionic megacolon in the TashT mouse model of Hirschsprung disease involves upregulation of p53 protein activity and Ddx3y gene expression. *PLOS Genetics*, *16*(9), e1009008. http://dx.doi.org/10.1371/journal.pgen.1009008. Récupéré le 2023-04-03 de https://dx.plos.org/10.1371/journal.pgen.1009008

Carlson, M. (2021a). *org.Hs.eg.db : Genome wide annotation for Human.* manual.

Carlson, M. (2021b). *org.Mm.eg.db : Genome wide annotation for Mouse.* manual.

Carrió, M., Mazuelas, H., Richaud-Patin, Y., Gel, B., Terribas, E., Rosas, I., Jimenez-Delgado, S., Biayna, J., Vendredy, L., Blanco, I., Castellanos, E., Lázaro, C., Raya, et Serra, E. (2019). Reprogramming Captures the Genetic and Tumorigenic Properties of Neurofibromatosis Type 1 Plexiform Neurofibromas. *Stem Cell Reports*, *12*(2), 411–426. http://dx.doi.org/10.1016/j.stemcr.2019.01.001. Récupéré le 2023-12-29 de https://linkinghub.elsevier.com/retrieve/pii/S2213671119300025

Castro, R., Taetzsch, T., Vaughan, S. K., Godbe, K., Chappell, J., Settlage, R. E. et Valdez, G. (2020). Specific labeling of synaptic schwann cells reveals unique cellular and molecular features. *eLife*, *9*, e56935. http://dx.doi.org/10.7554/eLife.56935. Récupéré le 2023-12-30 de https://elifesciences.org/articles/56935

Center, M. A. C. (2023). EasyCellType : tutorial (Version : V1.2.1.0) - Welcome to EasyCellType !, Md Anderson Cancer Center, https ://biostatis-

tics.mdanderson.org/shinyapps/EasyCellType/. Récupéré le 2023-12-22 de https://biostatistics.mdanderson.org/shinyapps/EasyCellType/

Chapman, J. A., Mascher, M., Buluç, A., Barry, K., Georganas, E., Session, A., Strnadova, V., Jenkins, J., Sehgal, S., Oliker, L., Schmutz, J., Yelick, K. A., Scholz, U., Waugh, R., Poland, J. A., Muehlbauer, G. J., Stein, N. et Rokhsar, D. S. (2015). A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology*, *16*(1), 26. http://dx.doi.org/10.1186/s13059-015-0582-8. Récupéré le 2024-01-02 de https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0582-8

Chassaing, B., Aitken, J. D., Malleshappa, M. et Vijay-Kumar, M. (2014). Dextran Sulfate Sodium (DSS)-Induced Colitis in Mice. *Current Protocols in Immunology*, *104*(1). http://dx.doi.org/10.1002/0471142735.im1525s104. Récupéré le 2023-04-03 de https://onlinelibrary.wiley.com/doi/10.1002/0471142735.im1525s104

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R. et Ma'ayan, A. (2013). Enrichr : interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 128. http://dx.doi.org/10.1186/1471-2105-14-128. Récupéré le 2023-08-31 de https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-128

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S. et Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265–270. http://dx.doi.org/10.1038/nnano.2009.12. Récupéré le 2024-01-02 de https://www.nature.com/articles/nnano.2009.12

Collins, F. S. et Fink, L. (1995). The Human Genome Project. *Alcohol Health and Research World*, *19*(3), 190–195.

Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Récupéré de https://igraph.org

Dahui, Q. (2019). Next-generation sequencing and its clinical application. *Cancer Biology & Medicine*, *16*(1), 4–10. http://dx.doi.org/10.20892/j.issn.2095-3941.2018.0055. Récupéré le 2023-12-12 de http://www.cancerbiomed.org/lookup/doi/10.20892/j.issn.2095-3941.2018.0055

Delahaye, C. et Nicolas, J. (2021). Sequencing DNA with nanopores : Troubles and biases. *PLOS ONE*, *16*(10), e0257521. http://dx.doi.org/10.1371/journal.pone.0257521. Récupéré le 2024-01-02 de https://dx.plos.org/10.1371/journal.pone.0257521

Dewey, F. E., Pan, S., Wheeler, M. T., Quake, S. R. et Ashley, E. A. (2012). DNA Sequencing :

Clinical Applications of New DNA Sequencing Technologies. *Circulation*, *125*(7), 931–944. http://dx.doi.org/10.1161/CIRCULATIONAHA.110.972828. Récupéré le 2023-08-26 de https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.110.972828

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. et Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. http://dx.doi.org/10.1038/nbt.3820. Récupéré le 2024-01-08 de https://www.nature.com/articles/nbt.3820

Diposarosa, R., Bustam, N., Sahiratmadja, E., Susanto, P. et Sribudiani, Y. (2021). Literature review : enteric nervous system development, genetic and epigenetic regulation in the etiology of Hirschsprung's disease. *Heliyon*, *7*(6), e07308. http://dx.doi.org/10.1016/j.heliyon.2021.e07308. Récupéré le 2023-08-22 de https://linkinghub.elsevier.com/retrieve/pii/S2405844021014110

Dragulescu, A. et Arendt, C. (2020). *xlsx : Read, write, format excel 2007 and excel 97/2000/XP/2003 files*. manual

Drokhlyansky, E., Smillie, C. S., Van Wittenberghe, N., Ericsson, M., Griffin, G. K., Eraslan, G., Dionne, D., Cuoco, M. S., Goder-Reiser, M. N., Sharova, T., Kuksenko, O., Aguirre, A. J., Boland, G. M., Graham, D., Rozenblatt-Rosen, O., Xavier, R. J. et Regev, A. (2020). The Human and Mouse Enteric Nervous System at Single-Cell Resolution. *Cell*, *182*(6), 1606–1622.e23. http://dx.doi.org/10.1016/j.cell.2020.08.003. Récupéré le 2023-12-30 de https://linkinghub.elsevier.com/retrieve/pii/S0092867420309946

Durbec, P. L., Larsson-Blomberg, L. B., Schuchardt, A., Costantini, F. et Pachnis, V. (1996). Common origin and developmental dependence on *c-ret* of subsets of enteric and sympathetic neuroblasts. *Development*, *122*(1), 349–358. http://dx.doi.org/10.1242/dev.122.1.349. Récupéré le 2023-12-31 de https://journals.biologists.com/dev/article/122/1/349/38888/Common-origin-and-developmental-dependence-on-c

Ewing, B. (2022). *leidenbase : R and C wrappers to run the Leiden findartition function*. manual.

Forcier, J. (2023a). Paramiko, (Version 2.12.0), [Implémentation Python (Module)], Paramiko documentation, https ://docs.paramiko.org/en/latest/. Récupéré de https://docs.paramiko.org/en/latest/

Forcier, J. (2023b). Welcome to Paramiko's documentation !, https ://-docs.paramiko.org/en/latest/. Récupéré de https://docs.paramiko.org/en/latest/

Foster, I. (2011). Globus Online : Accelerating and Democratizing Science through Cloud-Based

Services. *IEEE Internet Computing*, *15*(3), 70–73. http://dx.doi.org/10.1109/MIC.2011.64. Récupéré le 2023-12-19 de http://ieeexplore.ieee.org/document/5755602/

Franzén, O., Gan, L.-M. et Björkegren, J. L. M. (2019a). PanglaoDB - A Single Cell Sequencing Resource For Gene Expression Data, PanglaoDB, https ://panglaodb.se/. Récupéré de https://panglaodb.se/

Franzén, O., Gan, L.-M. et Björkegren, J. L. M. (2019b). PanglaoDB : a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, *2019*. http://dx.doi.org/10.1093/database/baz046. Récupéré le 2023-12-29 de https://academic.oup.com/database/article/doi/10.1093/database/baz046/5427041

Friedmacher, F. et Puri, P. (2015). Rectal suction biopsy for the diagnosis of Hirschsprung's disease : a systematic review of diagnostic accuracy and complications. *Pediatric Surgery International*, *31*(9), 821–830. http://dx.doi.org/10.1007/s00383-015-3742-8. Récupéré le 2023-12-31 de http://link.springer.com/10.1007/s00383-015-3742-8

Furness, J. B., Callaghan, B. P., Rivera, L. R. et Cho, H.-J. (2014). The Enteric Nervous System and Gastrointestinal Innervation : Integrated Local and Central Control. In M. Lyte et J. F. Cryan (dir.), *Microbial Endocrinology : The Microbiota-Gut-Brain Axis in Health and Disease*, volume 817 39–71. New York, NY : Springer New York. Series Title : Advances in Experimental Medicine and Biology

Glur, C. (2020). *data.tree : General purpose hierarchical data structure*. manual

Goldstein, A., Hofstra, R. et Burns, A. (2013). Building a brain in the gut : development of the enteric nervous system. *Clinical Genetics*, *83*(4), 307–316. http://dx.doi.org/10.1111/cge.12054. Récupéré le 2023-12-29 de https://onlinelibrary.wiley.com/doi/10.1111/cge.12054

Goldstein, A. M., Thapar, N., Karunaratne, T. B. et De Giorgio, R. (2016). Clinical aspects of neurointestinal disease : Pathophysiology, diagnosis, and treatment. *Developmental Biology*, *417*(2), 217–228. http://dx.doi.org/10.1016/j.ydbio.2016.03.032. Récupéré le 2023-12-31 de https://linkinghub.elsevier.com/retrieve/pii/S0012160616300227

Grubišić, V. et Gulbransen, B. D. (2017). Enteric glia : the most alimentary of all glia. *The Journal of Physiology*, *595*(2), 557–570. http://dx.doi.org/10.1113/JP271021. Récupéré le 2023-12-31 de https://physoc.onlinelibrary.wiley.com/doi/10.1113/JP271021

Han, S., Fink, J., Jörg, D. J., Lee, E., Yum, M. K., Chatzeli, L., Merker, S. R., Josserand, M., Trendafilova, T., Andersson-Rolf, A., Dabrowska, C., Kim, H., Naumann, R., Lee, J.-H., Sasaki, N., Mort, R. L., Basak, O., Clevers, H., Stange, D. E., Philpott, A., Kim, J. K.,

Simons, B. D. et Koo, B.-K. (2019). Defining the Identity and Dynamics of Adult Gastric Isthmus Stem Cells. *Cell Stem Cell*, *25*(3), 342–356.e7. http://dx.doi.org/10.1016/j.stem.2019.07.008. Récupéré le 2023-12-30 de https://linkinghub.elsevier.com/retrieve/pii/S1934590919303066

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. et Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29. http://dx.doi.org/10.1016/j.cell.2021.04.048. Récupéré le 2023-04-03 de https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833

He, X., Smith, S. E., Chen, S., Li, H., Wu, D., Meneses-Giles, P. I., Wang, Y., Hembree, M., Yi, K., Zhao, X., Guo, F., Unruh, J. R., Maddera, L. E., Yu, Z., Scott, A., Perera, A., Wang, Y., Zhao, C., Bae, K., Box, A., Haug, J. S., Tao, F., Hu, D., Hansen, D. M., Qian, P., Saha, S., Dixon, D., Anant, S., Zhang, D., Lin, E. H., Sun, W., Wiedemann, L. M. et Li, L. (2021). Tumor-initiating stem cell shapes its microenvironment into an immunosuppressive barrier and pro-tumorigenic niche. *Cell Reports*, *36*(10), 109674. http://dx.doi.org/10.1016/j.celrep.2021.109674. Récupéré le 2023-12-30 de https://linkinghub.elsevier.com/retrieve/pii/S2211124721011189

Hess, J., Kohl, T., Kotrová, M., Rönsch, K., Paprotka, T., Mohr, V., Hutzenlaub, T., Brüggemann, M., Zengerle, R., Niemann, S. et Paust, N. (2020). Library preparation for next generation sequencing : A review of automation strategies. *Biotechnology Advances*, *41*, 107537. http://dx.doi.org/10.1016/j.biotechadv.2020.107537. Récupéré le 2024-12-12 de https://linkinghub.elsevier.com/retrieve/pii/S0734975020300343

Hoffman, P., Lab, S. et Collaborateurs (2023). Seurat - Guided Clustering Tutorial (Version 4.3), Satija Lab, https ://satijalab.org/seurat/archive/v4.3/pbmc3k_tutorial. Récupéré le 2023-12-20 de https://satijalab.org/seurat/archive/v4.3/pbmc3k_tutorial

Hotaling, S., Kelley, J. L. et Frandsen, P. B. (2021). Toward a genome sequence for every animal : Where are we now ? *Proceedings of the National Academy of Sciences*, *118*(52), e2109019118. http://dx.doi.org/10.1073/pnas.2109019118. Récupéré le 2024-01-02 de https://pnas.org/doi/full/10.1073/pnas.2109019118

Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P. et Zhang, Y. (2023a). CellMarker 2.0 : an updated database of manually curated

cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, *51*(D1), D870–D876. http://dx.doi.org/10.1093/nar/gkac947. Récupéré le 2023-12-29 de https://academic.oup.com/nar/article/51/D1/D870/6775381

Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P. et Zhang, Y. (2023b). CellMarker2.0, http ://bio-bigdata.hrbmu.edu.cn/CellMarker/. Récupéré de http://bio-bigdata.hrbmu.edu.cn/CellMarker/

Ianevski, A. (2023a). ScType : Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. original-date : 2019-06-05T10 :02 :31Z. Récupéré le 2023-12-20 de https://github.com/IanevskiAleksandr/sc-type

Ianevski, A. (2023b). ScType : Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data, (Commit 72e3a39), Github, https ://github.com/IanevskiAleksandr/sc-type. Récupéré de https://github.com/IanevskiAleksandr/sc-type

Ianevski, A., Giri, A. K. et Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature Communications*, *13*(1), 1246. http://dx.doi.org/10.1038/s41467-022-28803-w. Récupéré le 2023-08-30 de https://www.nature.com/articles/s41467-022-28803-w

Jawaid, W. (2023a). *enrichR : Provides an R interface to 'enrichr'*. manual

Jawaid, W. (2023b). An R interface to the Enrichr database, (Commit 54aaff4), Github, https ://github.com/wjawaid/enrichR. Récupéré le 2023-12-20 de https://cran.r-project.org/web/packages/enrichR/vignettes/enrichR.html

JetBrains (2020). PyCharm Community Edition, (Version 2020.1), [Python IDE]. JetBrains. https ://www.jetbrains.com/pycharm/. Récupéré de https://www.jetbrains.com/pycharm/

Ji, Z., Zhou, W., Hou, W. et Ji, H. (2020). Single-cell ATAC-seq signal extraction and enhancement with SCATE. *Genome Biology*, *21*(1), 161. http://dx.doi.org/10.1186/s13059-020-02075-3. Récupéré le 2023-12-13 de https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02075-3

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F. et Luo, Y. (2022). Single-cell RNA sequencing technologies and applications : A brief overview. *Clinical and Translational Medicine*,

12(3), e694. http://dx.doi.org/10.1002/ctm2.694. Récupéré le 2023-08-26 de https://onlinelibrary.wiley.com/doi/10.1002/ctm2.694

Kapur, R. P. (1999). Hirschsprung Disease and Other Enteric Dysganglionoses. *Critical Reviews in Clinical Laboratory Sciences*, *36*(3), 225–273. http://dx.doi.org/10.1080/10408369991239204. Récupéré le 2023-12-31 de http://www.tandfonline.com/doi/full/10.1080/10408369991239204

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S. et Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. Dans F. Loizides et B. Schmidt (dir.). *Positioning and power in academic publishing : Players, agents and agendas*, 87 – 90. IOS Press.

Kozińska, A., Seweryn, P. et Sitkiewicz, I. (2019). A crash course in sequencing for a microbiologist. *Journal of Applied Genetics*, *60*(1), 103–111. http://dx.doi.org/10.1007/s13353-019-00482-2. Récupéré le 2024-01-02 de http://link.springer.com/10.1007/s13353-019-00482-2

Kukurba, K. R. et Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb.top084970. http://dx.doi.org/10.1101/pdb.top084970. Récupéré le 2023-08-24 de http://www.cshprotocols.org/lookup/doi/10.1101/pdb.top084970

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W. et Ma'ayan, A. (2016). Enrichr : a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. http://dx.doi.org/10.1093/nar/gkw377. Récupéré le 2023-08-30 de https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw377

kythol, L. (2022). Solution proposée par Lisa (kythol) dans : duplicate vertex names" error when visualizing a bubble plot by ScType #6 (Issue), Github. https ://github.com/IanevskiAleksandr/sc-type/issues/6#issuecomment-1180619645. Récupéré de https://github.com/IanevskiAleksandr/sc-type/issues/6#issuecomment-1180619645

Köster, J. et Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. http://dx.doi.org/10.1093/bioinformatics/bts480. Récupéré le 2024-01-08 de https://academic.oup.com/bioinformatics/article/28/

19/2520/290322

Lau, S.-T., Li, Z., Pui-Ling Lai, F., Nga-Chu Lui, K., Li, P., Munera, J. O., Pan, G., Mahe, M. M., Hui, C.-C., Wells, J. M. et Ngan, E. S.-W. (2019). Activation of Hedgehog Signaling Promotes Development of Mouse and Human Enteric Neural Crest Cells, Based on Single-Cell Transcriptome Analyses. *Gastroenterology*, *157*(6), 1556–1571.e5. http://dx.doi.org/10.1053/j.gastro.2019.08.019. Récupéré le 2023-12-30 de https://linkinghub.elsevier.com/retrieve/pii/S001650851941233X

Li, J., Wang, H. et Li, C. (2022). Single-Cell Sequencing on Marine Life : Application and Future Development. *Frontiers in Marine Science*, *9*, 906267. http://dx.doi.org/10.3389/fmars.2022.906267. Récupéré le 2023-12-31 de https://www.frontiersin.org/articles/10.3389/fmars.2022.906267/full

Li, R. (2022a). EasyCellType : an example workflow, (Commit dd51a8f), Github, https ://github.com/rx-li/EasyCellType/blob/main/vignettes/my-vignette.Rmd. Récupéré de https://github.com/rx-li/EasyCellType/blob/main/vignettes/my-vignette.Rmd

Li, R. (2022b). EasyCellType (version 0.99.0), [Package R], Github, https ://github.com/rx-li/EasyCellType/tree/1278319714f573cfa3d3cf93bffb8a57627b261c. Récupéré de https://github.com/rx-li/EasyCellType/tree/1278319714f573cfa3d3cf93bffb8a57627b261c

Li, R. (2023). *EasyCellType : Annotate cell types for scRNA-seq data.* manual.

Li, R., Zhang, J. et Li, Z. (2023). EasyCellType : marker-based cell-type annotation by automatically querying multiple databases. *Bioinformatics Advances*, *3*(1), vbad029. http://dx.doi.org/10.1093/bioadv/vbad029. Récupéré le 2023-08-30 de https://academic.oup.com/bioinformaticsadvances/article/doi/10.1093/bioadv/vbad029/7085606

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. D., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Rączkowska, A., Reinders, M., Ridder, J. D., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P. et Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*,

21(1), 31. http://dx.doi.org/10.1186/s13059-020-1926-6. Récupéré le 2025-08-29 de https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1926-6

May-Zhang, A. A., Tycksen, E., Southard-Smith, A. N., Deal, K. K., Benthal, J. T., Buehler, D. P., Adam, M., Simmons, A. J., Monaghan, J. R., Matlock, B. K., Flaherty, D. K., Potter, S. S., Lau, K. S. et Southard-Smith, E. M. (2021). Combinatorial Transcriptional Profiling of Mouse and Human Enteric Neurons Identifies Shared and Disparate Subtypes In Situ. *Gastroenterology*, *160*(3), 755–770.e26. http://dx.doi.org/10.1053/j.gastro.2020.09.032. Récupéré le 2023-12-30 de https://linkinghub.elsevier.com/retrieve/pii/S0016508520352136

McCarthy, D. J., Campbell, K. R., Lun, A. T. L. et Wills, Q. F. (2017). Scater : pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, *33*(8), 1179–1186. http://dx.doi.org/10.1093/bioinformatics/btw777. Récupéré le 2024-01-04 de https://academic.oup.com/bioinformatics/article/33/8/1179/2907823

Memic, F., Knoflach, V., Sadler, R., Tegerstedt, G., Sundström, E., Guillemot, F., Pachnis, V. et Marklund, U. (2016). Ascl1 Is Required for the Development of Specific Neuronal Subtypes in the Enteric Nervous System. *The Journal of Neuroscience*, *36*(15), 4339–4350. http://dx.doi.org/10.1523/JNEUROSCI.0202-16.2016. Récupéré le 2024-01-04 de https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0202-16.2016

Meslier, V., Quinquis, B., Da Silva, K., Plaza Oñate, F., Pons, N., Roume, H., Podar, M. et Almeida, M. (2022). Benchmarking second and third-generation sequencing platforms for microbial metagenomics. *Scientific Data*, *9*(1), 694. http://dx.doi.org/10.1038/s41597-022-01762-z. Récupéré le 2024-01-02 de https://www.nature.com/articles/s41597-022-01762-z

Microsoft (2022). Microsoft Windows 10 Famille (Version 22H2), [Système d'exploitation], https ://learn.microsoft.com/en-us/windows/release-health/status-windows-10-22h2. Récupéré de https://learn.microsoft.com/en-us/windows/release-health/status-windows-10-22h2

Milichko, V. et Dyachuk, V. (2020). Novel Glial Cell Functions : Extensive Potency, Stem Cell-Like Properties, and Participation in Regeneration and Transdifferentiation. *Frontiers in Cell and Developmental Biology*, *8*, 809. http://dx.doi.org/10.3389/fcell.2020.00809. Récupéré le 2024-01-07 de https://www.frontiersin.org/article/10.3389/fcell.2020.00809/full

Morarach, K., Mikhailova, A., Knoflach, V., Memic, F., Kumar, R., Li, W., Ernfors, P. et Marklund, U. (2021). Diversification of molecularly defined myenteric neuron classes revealed by single-cell RNA sequencing. *Nature Neuroscience*, *24*(1), 34–46. http://dx.doi.org/10.1038/s41593-020-00736-x. Récupéré le 2023-12-30 de https://www.nature.com/articles/s41593-020-00736-x

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. http://dx.doi.org/10.1038/nature01262. Récupéré le 2024-01-02 de https://www.nature.com/articles/nature01262

Mowat, A. M. et Viney, J. L. (1997). The anatomical basis of intestinal immunity. *Immunological Reviews*, *156*(1), 145–166. http://dx.doi.org/10.1111/j.1600-065X.1997.tb00966.x. Récupéré le 2023-08-24 de https://onlinelibrary.wiley.com/doi/10.1111/j.1600-065X.1997.tb00966.x

Mueller, J. L. et Goldstein, A. M. (2022). The science of Hirschsprung disease : What we know and where we are headed. *Seminars in Pediatric Surgery*, *31*(2), 151157. http://dx.doi.org/10.1016/j.sempedsurg.2022.151157. Récupéré le 2023-08-21 de https://linkinghub.elsevier.com/retrieve/pii/S105585862200018X

Niramis, R., Watanatittan, S., Anuntkosol, M., Buranakijcharoen, V., Rattanasuwan, T., Tongsin, A., Petlek, W. et Mahatharadol, V. (2008). Quality of Life of Patients with Hirschsprung's Disease at 5 - 20 Years Post Pull-Through Operations. *European Journal of Pediatric Surgery*, *18*(1), 38–43. http://dx.doi.org/10.1055/s-2008-1038325. Récupéré le 2023-12-31 de http://www.thieme-connect.de/DOI/DOI?10.1055/s-2008-1038325

of Chicago, T. U. (2020). Globus, https ://www.globus.org/. Récupéré le 2023-12-19 de https://www.globus.org

of Chicago, T. U. (2022). Globus Connect Personal, (Version 3.2.1), [Service de transfert de fichiers], https ://www.globus.org/globus-connect-personal. Récupéré de https://www.globus.org/globus-connect-personal

Okoniewski, M. J. et Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, *7*(1), 276. http://dx.doi.org/10.1186/1471-2105-7-276. Récupéré le 2024-01-01 de https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-276

Pan, Z. et Li, J. (2012). Advances in Molecular Genetics of Hirschsprung's Disease. *The Anatomical Record*, *295*(10), 1628–1638. http://dx.doi.org/10.1002/ar.22538. Récu-

péré le 2023-12-31 de https://anatomypubs.onlinelibrary.wiley.com/doi/10.1002/ar.22538

Pedersen, T. L. (2020). *patchwork : The composer of plots*. manual

Pedersen, T. L. (2022). *ggraph : An implementation of grammar of graphics for graphs and networks*. manual

Peng, B. et of Texas MD Anderson Cancer Center, U. (2018a). SoS - Script of Script, vatlab.github.io, https ://vatlab.github.io/sos-docs/. Récupéré de https://vatlab.github.io/sos-docs/

Peng, B. et of Texas MD Anderson Cancer Center, U. (2018b). Using multiple kernels in one Jupyter notebook, vatlab.github.io, https ://vatlab.github.io/sos-docs/doc/user_guide/multi_kernel_notebook.html. Récupéré le 2023-12-28 de https://vatlab.github.io/sos-docs/doc/user_guide/multi_kernel_notebook.html

Peng, B., Wang, G., Ma, J., Leong, M. C., Wakefield, C., Melott, J., Chiu, Y., Du, D. et Weinstein, J. N. (2018). SoS Notebook : an interactive multi-language data analysis environment. *Bioinformatics*, *34*(21), 3768–3770. http://dx.doi.org/10.1093/bioinformatics/bty405. Récupéré le 2023-12-19 de https://academic.oup.com/bioinformatics/article/34/21/3768/5001386

Petersen, B.-S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D. et Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, *18*(1), 14. http://dx.doi.org/10.1186/s12863-017-0479-5. Récupéré le 2024-01-02 de http://bmcgenet.biomedcentral.com/articles/10.1186/s12863-017-0479-5

Project, G. (2022). GNU Bash manual, (Version 5.2), [Langage de programmation], GNU PRoject., https ://www.gnu.org/software/bash/. Récupéré de https://www.gnu.org/software/bash/

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. et Trapnell, C. (2017a). Single-cell mRNA quantification and differential analysis with Census. *Nature Methods*, *14*(3), 309–315. http://dx.doi.org/10.1038/nmeth.4150. Récupéré le 2023-12-19 de https://www.nature.com/articles/nmeth.4150

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. et Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, *14*(10), 979–982. http://dx.doi.org/10.1038/nmeth.4402. Récupéré le 2023-12-19 de https://www.nature.com/articles/nmeth.4402

R Core Team (2021a). *R : A language and environment for statistical computing*. manual, R Foundation for Statistical Computing, Vienna, Austria

R Core Team (2021b). *R : A language and environment for statistical computing, R (Version 4.1)*. manual, R Foundation for Statistical Computing, Vienna, Austria

Ribarska, T., Bjørnstad, P. M., Sundaram, A. Y. M. et Gilfillan, G. D. (2022). Optimization of enzymatic fragmentation is crucial to maximize genome coverage : a comparison of library preparation methods for Illumina sequencing. *BMC Genomics*, *23*(1), 92. http://dx.doi.org/10.1186/s12864-022-08316-y. Récupéré le 2023-08-27 de https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-022-08316-y

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B. et Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182. http://dx.doi.org/10.1126/science.aam8999. Récupéré le 2024-01-03 de https://www.science.org/doi/10.1126/science.aam8999

Sasselli, V., Pachnis, V. et Burns, A. J. (2012). The enteric nervous system. *Developmental Biology*, *366*(1), 64–73. http://dx.doi.org/10.1016/j.ydbio.2012.01.012. Récupéré le 2023-12-29 de https://linkinghub.elsevier.com/retrieve/pii/S0012160612000280

Satija, R., Butler, A., Hoffman, P. et Stuart, T. (2020). *SeuratWrappers : Community-provided methods and extensions for the seurat object*. manual.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. et Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. http://dx.doi.org/10.1038/nbt.3192. Récupéré le 2023-04-03 de http://www.nature.com/articles/nbt.3192

Satija, R., Hoffman, P. et Butler, A. (2019). *SeuratData : Install and manage seurat datasets*. manual.

Satija, R. et Lab, S. (2023). Seurat Command List, Satija Lab, https ://satija-lab.org/seurat/articles/essential_commands.html. Récupéré le 2023-12-20 de https://satijalab.org/seurat/articles/essential_commands.html

Schauberger, P. et Walker, A. (2023). *openxlsx : Read, write and edit xlsx files*. manual

Shaibu, J. O., Onwuamah, C. K., James, A. B., Okwuraiwe, A. P., Amoo, O. S., Salu, O. B., Ige, F. A., Liboro, G., Odewale, E., Okoli, L. C., Ahmed, R. A., Achanya, D., Adesesan, A., Amuda, O. A., Sokei, J., Oyefolu, B. A. O., Salako, B. L., Omilabu, S. A. et Audu,

R. A. (2021). Full length genomic sanger sequencing and phylogenetic analysis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in Nigeria. *PLOS ONE*, *16*(1), e0243271. http://dx.doi.org/10.1371/journal.pone.0243271. Récupéré le 2023-08-27 de https://dx.plos.org/10.1371/journal.pone.0243271

Shao, X., Liao, J., Lu, X., Xue, R., Ai, N. et Fan, X. (2020). scCATCH : Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience*, *23*(3), 100882. http://dx.doi.org/10.1016/j.isci.2020.100882. Récupéré le 2024-01-04 de https://linkinghub.elsevier.com/retrieve/pii/S2589004220300663

Shevlin, K. (2020). Encapsulation or the Primary Purpose of Functions, kyleshevlin.com, https ://kyleshevlin.com/encapsulation. Récupéré le 2024-01-08 de https://kyleshevlin.com/encapsulation

Soret, R. (2021). RS_wt_ney1135a6-23_s61, RS_dss_ney1135a6-22_s57, RS_holstein-Homozygote_ney1135a6-24_s65, RS_holstein-Homozygote-GDNF_ney1135a6-25_s69, [Fichiers FASTQ] [Non publiés], CERMO-FC - UQAM.

Soret, R., Lassoued, N., Bonnamour, G., Bernas, G., Barbe, A., Pelletier, M., Aichi, M. et Pilon, N. (2021). Genetic Background Influences Severity of Colonic Aganglionosis and Response to GDNF Enemas in the Holstein Mouse Model of Hirschsprung Disease. *International Journal of Molecular Sciences*, *22*(23), 13140. http://dx.doi.org/10.3390/ijms222313140. Récupéré le 2023-04-03 de https://www.mdpi.com/1422-0067/22/23/13140

Soret, R., Mennetrey, M., Bergeron, K. F., Dariel, A., Neunlist, M., Grunder, F., Faure, C., Silversides, D. W., Pilon, N. et for the Ente-Hirsch study group (2015). A collagen VI–dependent pathogenic mechanism for Hirschsprung's disease. *Journal of Clinical Investigation*, *125*(12), 4483–4496. http://dx.doi.org/10.1172/JCI83178. Récupéré le 2023-04-03 de https://www.jci.org/articles/view/83178

Soret, R., Schneider, S., Bernas, G., Christophers, B., Souchkova, O., Charrier, B., Righini-Grunder, F., Aspirot, A., Landry, M., Kembel, S. W., Faure, C., Heuckeroth, R. O. et Pilon, N. (2020). Glial Cell-Derived Neurotrophic Factor Induces Enteric Neurogenesis and Improves Colon Structure and Function in Mouse Models of Hirschsprung Disease. *Gastroenterology*, *159*(5), 1824–1838.e17. http://dx.doi.org/10.1053/j.gastro.2020.07.018. Récupéré le 2023-04-03 de https://linkinghub.elsevier.com/retrieve/pii/S0016508520349441

Stakenborg, N., Viola, M. F. et Boeckxstaens, G. E. (2020). Intestinal neuro-immune in-

teractions : focus on macrophages, mast cells and innate lymphoid cells. *Current Opinion in Neurobiology*, *62*, 68–75. http://dx.doi.org/10.1016/j.conb.2019.11.020. Récupéré le 2024-01-07 de https://linkinghub.elsevier.com/retrieve/pii/S095943881930131X

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. et Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888–1902.e21. http://dx.doi.org/10.1016/j.cell.2019.05.031. Récupéré le 2023-04-03 de https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598

Stuart, T. et Satija, R. (2019). Integrative single-cell analysis. *Nature Reviews Genetics*, *20*(5), 257–272. http://dx.doi.org/10.1038/s41576-019-0093-7. Récupéré le 2023-08-27 de https://www.nature.com/articles/s41576-019-0093-7

Sun, L., Liu, Y., Lehnert, T., Gijs, M. A. M. et Li, S. (2022). The enhancement of DNA fragmentation in a bench top ultrasonic water bath with needle-induced air bubbles : Simulation and experimental investigation. *Biomicrofluidics*, *16*(4), 044103. http://dx.doi.org/10.1063/5.0101740. Récupéré le 2024-01-01 de https://pubs.aip.org/bmf/article/16/4/044103/2835446/The-enhancement-of-DNA-fragmentation-in-a-bench

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. et Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. http://dx.doi.org/10.1038/nmeth.1315. Récupéré le 2023-08-27 de https://www.nature.com/articles/nmeth.1315

Tarapcsak, S., Huang, X., Qiao, Y., Farrell, A., Mammen, L., Lovichik, A., Khanderao, G. D., Musci, T., Moos, P. J., Firpo, M. A., Rollins, M. et Marth, G. T. (2025). Single-cell RNA sequencing in Hirschsprung's disease tissues reveals lack of neuronal differentiation in the aganglionic colon segment. http://dx.doi.org/10.1101/2025.07.01.662516. Récupéré le 2025-08-29 de http://biorxiv.org/lookup/doi/10.1101/2025.07.01.662516

Tatham, S. (2022). PuTTY, (Version 0.78), [Client SSH et Telnet]. https ://www.chiark.greenend.org.uk/~sgtatham/putty/. Récupéré de https://www.chiark.greenend.org.uk/~sgtatham/putty/

The C. elegans Sequencing Consortium* (1998). Genome Sequence of the Nematode *C. elegans* : A Platform for Investigating Biology. *Science*, *282*(5396), 2012–2018. http://dx.doi.org/10.1126/science.282.5396.2012. Récupéré le 2024-01-02 de https://www.science.org/doi/10.1126/science.282.5396.2012

Tikoo, R., Zanazzi, G., Shiffman, D., Salzer, J. et Chao, M. V. (2000). Cell Cycle Control of Schwann Cell Proliferation : Role of Cyclin-Dependent Kinase-2. *The Journal of Neuroscience*, *20*(12), 4627–4634. http://dx.doi.org/10.1523/JNEUROSCI.20-12-04627.2000. Récupéré le 2023-12-29 de https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.20-12-04627.2000

Trapnell, C. (2022). Constructing single-cell trajectories, Monocle 3, https ://cole-trapnell-lab.github.io/monocle3/docs/trajectories/. Récupéré de https://cole-trapnell-lab.github.io/monocle3/docs/trajectories/

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. et Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. http://dx.doi.org/10.1038/nbt.2859. Récupéré le 2023-12-19 de https://www.nature.com/articles/nbt.2859

Uesaka, T., Nagashimada, M. et Enomoto, H. (2015). Neuronal Differentiation in Schwann Cell Lineage Underlies Postnatal Neurogenesis in the Enteric Nervous System. *Journal of Neuroscience*, *35*(27), 9879–9888. http://dx.doi.org/10.1523/JNEUROSCI.1239-15.2015. Récupéré le 2024-01-07 de https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1239-15.2015

Uesaka, T., Nagashimada, M., Yonemura, S. et Enomoto, H. (2008). Diminished Ret expression compromises neuronal survival in the colon and causes intestinal aganglionosis in mice. *Journal of Clinical Investigation*, *118*(5), 1890–1898. http://dx.doi.org/10.1172/JCI34425. Récupéré le 2024-01-02 de http://www.jci.org/articles/view/34425

Umar, S. (2010). Intestinal Stem Cells. *Current Gastroenterology Reports*, *12*(5), 340–348. http://dx.doi.org/10.1007/s11894-010-0130-3. Récupéré le 2024-01-08 de http://link.springer.com/10.1007/s11894-010-0130-3

Urla, C., Lieber, J., Obermayr, F., Busch, A., Schweizer, R., Warmann, S. W., Kirschner, H.-J. et Fuchs, J. (2018). Surgical treatment of children with total colonic aganglionosis : functional and metabolic long-term outcome. *BMC Surgery*, *18*(1), 58. http://dx.doi.org/10.1186/s12893-018-0383-6. Récupéré le 2023-12-31 de https://bmcsurg.biomedcentral.com/articles/10.1186/s12893-018-0383-6

Van Rossum, G. et Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA : CreateSpace.

Van Rossum, G. et Drake Jr, F. L. (2009). Python, (Version 3.7), Langage de programmation,

CreateSpace. Récupéré de https://www.python.org/doc/

Waldron, L. et Riester, M. (2019). *HGNChelper : Identify and correct invalid HGNC human gene symbols and MGI mouse gene symbols*. manual

Wallace, A. S. et Burns, A. J. (2005). Development of the enteric nervous system, smooth muscle and interstitial cells of Cajal in the human gastrointestinal tract. *Cell and Tissue Research*, *319*(3), 367–382. http://dx.doi.org/10.1007/s00441-004-1023-2. Récupéré le 2023-12-31 de http://link.springer.com/10.1007/s00441-004-1023-2

Wang, G. et Peng, B. (2019). Script of Scripts : A pragmatic workflow system for daily computational research. *PLOS Computational Biology*, *15*(2), e1006843. http://dx.doi.org/10.1371/journal.pcbi.1006843. Récupéré le 2023-12-19 de https://dx.plos.org/10.1371/journal.pcbi.1006843

Wickham, H. (2016). *ggplot2 : Elegant graphics for data analysis*. Springer-Verlag New York. Récupéré de https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. et Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. http://dx.doi.org/10.21105/joss.01686

Wickham, H. et Bryan, J. (2023). *readxl : Read excel files*. manual

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H. et Dunnington, D. (2023a). ggplot2, Tidyverse, https ://ggplot2.tidyverse.org/. Récupéré le 2023-12-30 de https://ggplot2.tidyverse.org/index.html

Wickham, H., François, R., Henry, L., Müller, K. et Vaughan, D. (2023b). *dplyr : A grammar of data manipulation*. manual

Wolf, F. A., Angerer, P. et Theis, F. J. (2018). SCANPY : large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. http://dx.doi.org/10.1186/s13059-017-1382-0. Récupéré le 2024-01-04 de https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1382-0

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M. et Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3), e90. http://dx.doi.org/10.1002/cpz1.90. Récupéré le 2023-08-31 de https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpz1.90

Yang, L., Ng, Y. E. et Sun, H. (2023a). scMayoMap (Commit 993e81a), GitHub, https ://-github.com/chloelulu/scMayoMap. Récupéré de https://github.com/chloelulu/scMayoMap

Yang, L., Ng, Y. E., Sun, H., Li, Y., Chini, L. C. S., LeBrasseur, N. K., Chen, J. et Zhang, X. (2023b). Single-cell Mayo Map (scMayoMap) : an easy-to-use tool for cell type annotation in single-cell RNA-sequencing data analysis. *BMC Biology*, *21*(1), 223. http://dx.doi.org/10.1186/s12915-023-01728-6. Récupéré le 2023-12-24 de https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-023-01728-6

Yang, L. et Zhang, X. (2023). *scMayoMap : Single-cell mayo Map(scMayoMap)*. manual.

Zappia, L., Phipson, B. et Oshlack, A. (2018a). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology*, *14*(6), e1006245. http://dx.doi.org/10.1371/journal.pcbi.1006245. Récupéré le 2024-01-03 de https://dx.plos.org/10.1371/journal.pcbi.1006245

Zappia, L., Phipson, B. et Oshlack, A. (2018b). scRNA-tools. Récupéré le 2024-01-03 de https://www.scrna-tools.org/

Zappia, L. et Theis, F. J. (2021). Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biology*, *22*(1), 301. http://dx.doi.org/10.1186/s13059-021-02519-4. Récupéré le 2024-01-03 de https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02519-4

Zeb, Q., Wang, C., Shafiq, S. et Liu, L. (2019). An Overview of Single-Cell Isolation Techniques. In *Single-Cell Omics* 101–135. Elsevier

Zhang, Luo, Zhong, Choi, Ma, Wang, Mahrt, Guo, Stawiski, Modrusan, Seshagiri, Kapur, Hon, Brugarolas et Wang (2019). SCINA : Semi-Supervised Analysis of Single Cells in Silico. *Genes*, *10*(7), 531. http://dx.doi.org/10.3390/genes10070531. Récupéré le 2024-01-04 de https://www.mdpi.com/2073-4425/10/7/531

Zhang, L., Chen, F., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y., Hao, H., Yi, W., Li, M. et Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Frontiers in Microbiology*, *12*, 766364. http://dx.doi.org/10.3389/fmicb.2021.766364. Récupéré le 2024-01-01 de https://www.frontiersin.org/articles/10.3389/fmicb.2021.766364/full

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., Ping, Y., Li, F., Shi, A., Bai, J., Zhao, T., Li, X. et Xiao, Y. (2019). CellMarker : a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*,

47(D1), D721–D728. http://dx.doi.org/10.1093/nar/gky900. Récupéré le 2024-01-08 de https://academic.oup.com/nar/article/47/D1/D721/5115823

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. et Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. http://dx.doi.org/10.1038/ncomms14049. Récupéré le 2023-08-30 de https://www.nature.com/articles/ncomms14049

biblatex