

Safiya ATIA  
M2 AMI2B  
2022-2023



**Rapport de stage**  
15 mars 2023 - 25 août 2023

**Sujet : Profilage cellule unique du transcriptome de gliomes pour la caractérisation de l'hétérogénéité métabolique tumorale en biologie spatiale multiplexée**

Encadrant : Dr Pierre-François ROUX

Introduction.....	3
Méthodes.....	10
A. Les données RNA-seq.....	10
B. Langages de programmation.....	10
C. Environnement de travail.....	10
a. Ressources.....	10
b. Containers singularity.....	11
D. Processing des données.....	11
a. Les paramètres.....	11
b. Le pipelines nf-core.....	11
E. Analyse du RNA-seq cellule unique.....	12
a. Cycle cellulaire et pourcentage de gènes mitochondriaux.....	12
b. Filtrage.....	12
c. Normalisation.....	13
d. Intégration des données.....	13
e. Méthode de réduction d'effet batch.....	14
f. Le UMAP.....	14
g. Méthode de clustering.....	14
h. Analyse différentielle.....	15
i. Focalisation sur le métabolisme.....	15
j. Annotations des types cellulaires.....	15
Résultats.....	16
A. Présentation de la stratégie analytique.....	16
B. Caractérisation générale du jeu de données.....	17
a. Caractéristiques des échantillons.....	17
b. Normalisation et réduction de dimensionnalité.....	19
c. Caractérisations des lignées.....	21
C. Caractérisation biologique des lignées de gliomes étudiées.....	22
a. Caractérisation sur la base du transcriptome complet.....	22
b. Focalisation sur le métabolisme.....	24
c. Comparaison entre les classifications pan-transcriptome et.....	
métabolisme-centrique.....	26
C. Sélection de variables pour le développement d'un panel métabolique.....	27
Discussion.....	30
A. Choix de développement.....	30
a. Normalisation.....	30
b. Le clustering.....	30
B. Caractérisation du gliome.....	31
C. Suite envisageable.....	32
a. Sélection de gènes.....	32

b. DEXOM.....	32
c. 10X.....	32
D. Biais possibles.....	33
a. La normalisation.....	33
b. Les échantillons.....	33
c. La méthode d'analyse.....	33
Bilan.....	34
Annexes.....	35
Bibliographie.....	38

## Introduction

En étudiant l'expression des ARN produits par l'ensemble des cellules constitutives d'un tissu sain ou tumoral, il est possible de déterminer son profil d'expression génique - que l'on appelle transcriptome - et qui reflète au moins en partie sa nature, son état physiologique, son identité, et sa réponse à d'éventuels stimuli ou stress environnementaux. Aussi, dans le cadre de la biologie du cancer, l'étude du transcriptome offre des informations cruciales lorsqu'il s'agit de comprendre les mécanismes biologiques sous-jacents au développement tumoral ou la réponse thérapeutique.

L'Institut de Recherche en Cancérologie de Montpellier (IRCM) est une entité de recherche renommée réunissant 17 équipes dédiées à l'étude exhaustive du cancer. Les activités des équipes qui le composent s'étendent de la recherche fondamentale, axée sur l'exploration des mécanismes moléculaires et cellulaires de la cancérogenèse, à la recherche clinique, visant l'évaluation et le développement de nouvelles approches thérapeutiques pour les patients. L'IRCM s'implique également dans le développement de biotechnologies innovantes, traduisant ainsi ses découvertes scientifiques en applications cliniques au bénéfice des patients. Le thème fédérateur de l'institut - "Cibles moléculaires et thérapie des cancers, Découverte, Biologie et Applications Cliniques" - illustre la mission de l'IRCM : adopter une approche intégrée et multidisciplinaire pour l'identification de nouvelles cibles thérapeutiques anti-cancer, en élucider la biologie et en catalyser la transition vers des applications cliniques novatrices.

L'équipe "oncogenèse moléculaire" dirigée par le Dr Laurent Le Cam au sein de laquelle j'ai effectué mon stage a pour thème de recherche principal la compréhension des mécanismes de la tumorigenèse, se focalisant notamment sur l'étude des mécanismes de régulation de la voie p53, perturbée dans presque toutes les tumeurs humaines. Le facteur de transcription p53, parfois qualifié de « gardien du génome », est l'effecteur aval d'un point de contrôle qui empêche la prolifération de cellules dont le génome est altéré, en régulant l'expression de gènes cibles impliqués dans la réparation de l'ADN, le contrôle de la prolifération ou de la mort cellulaire, ou encore de la sénescence. Il empêche ainsi l'accumulation de mutations et réduit le risque de néoplasie [1].

Les tumeurs solides présentent en général une composition complexe, intégrant des populations parfois diverses de cellules cancéreuses, des cellules saines constituant le stroma ainsi que pour certaines un infiltrat immunitaire. Ces composants prolifèrent de manière anarchique, engendrant une hétérogénéité marquée au sein de la masse tumorale. Cette hétérogénéité peut se manifester à plusieurs niveaux : génétique, épigénétique, phénotypique ou encore métabolique, reflétant des variations considérables dans les caractéristiques, la nature et la biologie des cellules tumorales. Au niveau génétique, l'accumulation de mutations somatiques dans les oncogènes ou les gènes suppresseurs de tumeur peut conférer à certaines cellules un avantage sélectif, se traduisant par une capacité accrue à proliférer ou par une résistance augmentée aux thérapies anticancer. Sur le plan épigénétique, différentes sous-populations cellulaires peuvent émerger au sein de la tumeur, en raison de variations dans l'expression génique, et pouvant se traduire par des modifications distinctes de la méthylation de l'ADN, de l'acétylation des histones ou d'autres marques épigénétiques. Ces différences génétiques et épigénétiques peuvent ultérieurement influencer les propriétés phénotypiques des cellules, impactant par exemple leur morphologie, leur taille ou encore leur métabolisme. En outre, l'environnement tumoral joue un rôle critique dans la modulation de cette hétérogénéité. L'écosystème tumoral,

comprenant des cellules non cancéreuses, des tissus conjonctifs, des vaisseaux sanguins et des cellules immunitaires, peut ainsi soit faciliter la prolifération tumorale, en favorisant les apports en oxygène et en nutriments, soit l'inhiber, en créant un environnement déficitaire, anoxique et pauvre en nutriments, pour les cellules tumorales. Ainsi, les interactions dynamiques entre les cellules tumorales et leur microenvironnement contribuent significativement à la complexité et à l'évolution des tumeurs solides [2].

Le métabolisme cellulaire constitue l'ensemble des réactions chimiques qui se produisent au sein d'une cellule assurant la transformation de nutriments en énergie, la construction et le maintien des structures cellulaires, et la régulation des déchets produits par ces processus. Il implique environ 3000 gènes et joue un rôle déterminant dans la zonation et l'hétérogénéité des tumeurs solides. Il est en effet profondément impacté dans de nombreux types tumoraux et de nombreuses voies métaboliques peuvent ainsi être détournées afin de répondre aux besoins d'adaptation, d'échappement et de prolifération des cellules tumorales [3]. Ainsi, la glycolyse - une voie métabolique anaérobie qui convertit une molécule de glucose en deux molécules de pyruvate, produisant de l'énergie sous forme de deux molécules d'ATP et deux molécules de NADH - est le processus métabolique privilégié dans le contexte de certains cancers tels que les carcinomes du col utérin : bien que procurant moins d'énergie que la respiration aérobie, elle répond aux besoins en biomasse nécessaire à la division accrue des cellules constitutives de ces néoplasies. Ainsi, même en présence d'oxygène, une partie du glucose est transformé en pyruvate qui pourra lui-même être dégradé en lactate et acidifier le microenvironnement tumoral propice à la prolifération : on parle de l'effet Warburg [4]. Par ailleurs, le métabolisme lipidique peut également être perturbé comme c'est le cas dans certaines formes de cancer de la prostate où les cellules tumorales stimulent la lipogenèse, qui fournit les acides gras essentiels comme éléments constitutifs des membranes cellulaires et comme substrats énergétiques, favorisant ainsi la prolifération et la progression de la tumeur [5]. Ces reprogrammations métaboliques s'opèrent ainsi toujours dans le but d'assurer l'adaptabilité des cellules tumorales et peuvent se manifester au niveau histologique par la formation de nouveaux vaisseaux sanguins à travers un processus nommé l'angiogenèse, stimulé par les cellules tumorales, stromales et immunitaires.

Aussi, le concept d'hétérogénéité tumorale peut être décliné sous différents prismes : l'hétérogénéité fonctionnelle et l'hétérogénéité spatiale - les deux étant largement intriquées. Les différents types cellulaires constitutifs de la tumeur solide vont pouvoir être caractérisés par leurs rôles et fonctions (cellules tumorales, stromales ou immunitaires) mais aussi selon une zonation en territoires tumoraux présentant des caractéristiques moléculaires et métaboliques singuliers dépendant notamment des apports en nutriments et oxygène, et de façon plus globale, de leur micro-environnement. Mieux caractériser l'hétérogénéité tumorale au sens large constitue un des enjeux majeurs de la recherche moderne en oncologie et constitue un prérequis pour le développement de thérapies innovantes et améliorer la prise en charge des patients dans l'optique d'une médecine personnalisée.

Jusqu'alors, après biopsie, les cliniciens se reposaient majoritairement sur les techniques d'anatomopathologie pour établir le portrait des tumeurs, poser le diagnostic, évaluer le stade du cancer et orienter la thérapeutique. Parmi ces techniques, celles reposant sur la microscopie tiennent une place prépondérante.

La coloration à l'hématoxyline et à l'éosine (H&E), est la technique de coloration histologique la plus couramment utilisée en anatomo-cytopathologie pour évaluer la morphologie des tissus. Elle est utilisée de manière routinière pour le diagnostic de divers types de cancer. Elle met ainsi en évidence les noyaux des cellules en bleu foncé (hématoxyline), et les structures cytoplasmiques et extracellulaires en rose (éosine), permettant au pathologiste d'évaluer les caractéristiques architecturales et cytologiques des tissus, telles que la taille et la forme des cellules, la présence de mitoses, la polynucléation, le degré de différenciation des cellules tumorales, et l'invasion des structures environnantes. Aussi, dans le cadre du cancer du sein, la coloration H&E est utilisée pour évaluer la morphologie des cellules tumorales et du tissu conjonctif environnant permettant de distinguer les caractéristiques des différentes catégories histologiques du cancer du sein, telles que le carcinome canalaire *in situ*, le carcinome lobulaire *in situ*, le carcinome canalaire invasif et le carcinome lobulaire invasif.

L'immunofluorescence repose sur l'utilisation d'anticorps pour marquer des protéines d'intérêt dans un échantillon biologique. Des anticorps primaires, qui n'ont aucune propriétés fluorescentes, vont se lier spécifiquement à une protéine cible. Ensuite sont ajoutés des anticorps dits secondaires, conjugués à des marqueurs fluorescents, qui vont pouvoir se lier aux anticorps primaires. Les échantillons peuvent à présent être observés à l'aide d'un microscope à fluorescence pour pouvoir localiser précisément la présence des protéines d'intérêt et leur distribution. Dans le cadre de l'anatomo-cytopathologie appliquée au cancer, les anticorps anti-cytokeratine 7 (CK7) et anti-cytokeratine 20 (CK20) sont fréquemment utilisés pour aider à déterminer l'origine primaire des carcinomes métastatiques.

L'hybridation *in situ* en fluorescence (FISH) ne permet pas la détection de protéines, mais permet de détecter et localiser des séquences spécifiques d'acides nucléiques (ADN ou d'ARN) dans les coupes histologiques. Des sondes fluorescentes ciblant ces séquences d'intérêt vont s'y lier de manière complémentaire pour pouvoir visualiser des régions spécifiques du génome ou des ARN afin de détecter, par exemple, des anomalies génétiques. Cette technique est par exemple utilisée en routine dans le diagnostic du carcinome pulmonaire non à petites cellules, pour identifier le réarrangement du gène ALK (Anaplastic Lymphoma Kinase), orientant de manière déterminante la stratégie thérapeutique.

Bien qu'elles apportent des informations extrêmement précieuses aujourd'hui en clinique, ces techniques restent très limitées car seules quelques caractéristiques moléculaires peuvent être analysées en même temps, limitant drastiquement les informations déduites sur une coupe tumorale. Pour un type histologique de cancer donné correspondent bien souvent plusieurs sous-types moléculaires, et l'hétérogénéité fonctionnelle et spatiale ne peuvent être étudiées avec un nombre si réduit de marqueurs. Aussi, disposer de stratégies holistiques s'avère indispensable pour aborder ces problématiques.

Dans ce contexte, les récentes avancées technologiques en matière de séquençage du transcriptome entier (RNA-seq) ont progressivement changé la donne, apportant d'ores et déjà de précieuses informations en clinique. En quantifiant la population d'ARN présents dans un échantillon sous forme de fragments de petite taille (oligonucléotides), il est possible d'évaluer le niveau de

transcription de l'ensemble des gènes exprimés par les cellules constitutives de la tumeur. Il y a deux manières d'analyser un échantillon par la technique de RNA-seq (Fig. 1).

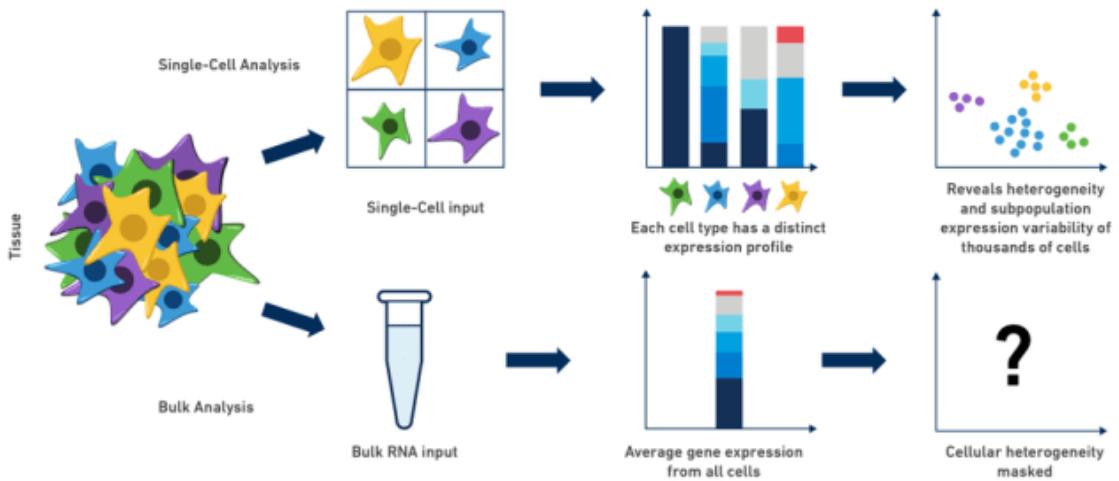


Figure 1 : Le RNA-seq en cellule unique révèle une hétérogénéité masquée en bulk RNA-seq.(de <https://www.10xgenomics.com/>)

Pour le *bulk* RNA-seq, après avoir broyé un fragment tissulaire contenant des milliers voire des millions de cellules, l'ARN est extrait et converti en ADN complémentaire (ADNc). L'ADNc est ensuite fragmenté en courtes séquences puis séquencé grâce à des technologies comme Illumina, pour quantifier le niveau d'expression des gènes. Cette technique permet ainsi d'obtenir la moyenne d'expression des populations de cellules constitutives de l'échantillon et va pouvoir mettre en avant les différences moyennes d'expression lorsque différentes conditions sont étudiées (par exemple un tissu tumoral comparé au tissu sain adjacent). En revanche, les informations plus précises et notamment relatives aux différentes sous-populations cellulaires présentes dans l'échantillon et à l'hétérogénéité fonctionnelle et spatiale sont perdues.

C'est pourquoi les approches cellules uniques (*scRNA-seq*) sont extrêmement prometteuses, ne consistant non pas en un séquençage global de la population d'ARN constitutive de l'échantillon homogénéisé, mais à un séquençage précis cellule par cellule. Plutôt que de broyer l'échantillon, les technologies *scRNA-seq* recourent à différentes méthodes d'isolation des cellules telles que la dissociation mécanique, la dissociation enzymatique ou le tri cellulaire par cytométrie en flux, en amont des étapes de préparation des librairies d'ARN.

La première technologie qui a révolutionné les analyses RNA-seq en ciblant le séquençage sur des cellules uniques est Smart-seq/C1. Publiée en 2012 par une équipe de recherche suédoise de Ludwig Institute for Cancer Research, cette technique utilise le dispositif appelé puce à cellules C1, développé par Fluidigm, pour capturer individuellement les cellules dans une microchambre. Elles sont lysées pour capturer leurs ARN à l'aide d'amorces oligo-dT spécifiques qui se lient à la partie polyA des ARN messagers, ensuite amplifiés et convertis en ADNc pour être séquencés. Cette technologie permet un séquençage complet du transcriptome, capable d'analyser de longs fragments pour garder l'information de la diversité des transcrits avec l'expression des différentes isoformes.

Smart-seq/C1 est limité cependant, notamment par les puces C1 qui ne peuvent traiter qu'un nombre de cellules limité en une seule fois.

Pour des études à plus grande échelle, les technologies à haut débit comme Chromium Single cell de 10X Genomics est plus adaptée, permettant de traiter des milliers de cellules simultanément, au détriment d'une perte de sensibilité sur la détection des ARN. Chaque cellule est isolée via une partition microfluidique qui les emprisonne une à une dans des gouttelettes distinctes, chacune d'elle étant associée à un *Unique Molecular Identifier* (UMI). Après séquençage, ces UMI sont utilisés pour trier les oligonucléotides identifiés et les assigner à la cellule d'origine.

Ces techniques de *scRNA-seq*, beaucoup plus coûteuses, sont particulièrement intéressantes dans le contexte oncologique pour caractériser des sous-populations qui seraient totalement négligées en étudiant le profil global d'expression. Toutefois il ne faut pas négliger que la faible quantité de matériel génétique de départ peut mener à des biais plus importants comparé au *bulk RNA-seq*, avec des dégradations et contaminations possibles qui expliquent des étapes de filtrations importantes au niveau analytique pour tenter de les éliminer.

Jusque très récemment, alors que l'hétérogénéité fonctionnelle pouvait être étudiée avec les techniques de *bulk* et de *scRNA-seq*, il n'était pas possible d'accéder à l'information concernant l'hétérogénéité spatiale. Depuis 2016, des techniques de transcriptomique spatiale ont été développées et rendent possible la visualisation des molécules d'ARN et cellules tout en conservant leur emplacement spatial dans les échantillons tissulaires. Cette information sur la localisation est primordiale pour étudier l'hétérogénéité spatiale des tumeurs solides.

En transcriptomique spatiale, 10X Genomics propose entre autres 2 technologies qui peuvent être utilisées pour comprendre et caractériser l'hétérogénéité spatiale des tumeurs.

Sur une coupe de tissu, la technologie Visium commence par une étape d'identification de zones de 50 µm, pouvant comprendre une à dix cellules, qui seront associées à un UMI avant d'être séquencées de manière classique. La plus-value de cette technique est l'ajout d'un code-barre pour chaque zone, correspondant à la localisation initiale sur l'échantillon. L'analyse de l'expression des gènes par zone peut donc être faite tout en combinant leur information spatiale avec une visualisation sur une reconstruction de la coupe histologique. Bien qu'exhaustive (cette technologie permet d'interroger l'ensemble du transcriptome sans a priori lié au développement de sondes), cette technique est peu résolutive, et ne permet d'accéder à l'information d'hétérogénéité spatiale à l'échelle cellulaire et subcellulaire.

Toujours sur une coupe histologique, la technologie Xenium qui ne repose pas sur le séquençage d'ARN, peut aussi être utilisée en transcriptomique spatiale. L'expression des gènes va pouvoir être détectée et quantifiée grâce à une technique multiplexée de FISH, le MERFISH (Multiplexed Error-Robust Fluorescence In Situ Hybridization).

Comme pour l'hybridation *in situ* classique, des sondes d'ARN spécifiques aux gènes d'intérêt couplés à des fluorochromes vont s'hybrider avec les ARN messagers présents dans l'échantillon. Ces sondes sont associées à des codes-barres qui permettent de différencier les sondes entre elles.

En combinant l'information de la fluorescence et des code-barres, il est possible de distinguer les signaux provenant des différentes sondes et donc des algorithmes de correction sont ensuite appliqués pour améliorer la précision. Seulement 500 transcrits d'intérêt au maximum peuvent être étudiés via cette technique [6].

Comparé au Visium qui donne un aperçu de tout le transcriptome mais à une résolution de 50µm, Xenium est non exhaustive en ciblant 500 transcrits, mais est particulièrement adaptée à la recherche ciblée et “thématique” sur le transcriptome dans les tumeurs solides, de par sa résolution subcellulaire. Elle est donc particulièrement adaptée à l'étude de l'hétérogénéité métabolique spatiale.

Cependant, comment choisir ces 500 gènes contre lesquels les sondes d'ARN de la technique Xenium vont être ciblées ? C'est dans ce cadre que s'inscrit mon stage dont le but est de **développer un pipeline d'analyse qui permettra d'identifier de manière objective un panel de 500 gènes métaboliques en me reposant sur des données RNA-seq, single cell et bulk appariés**, générées à partir de gliomes isolés de patients, grâce auxquelles nous étudierons ultérieurement l'hétérogénéité métabolique spatiale dans ce type de tumeur.

In fine, les gènes choisis seront des gènes impliqués dans le métabolisme pour tenter de comprendre la reprogrammation qui a lieu au sein de la tumeur, connaître les voies métaboliques activées ou réprimées selon le type cellulaire mais aussi selon la localisation spatiale pour définir des territoires tumoraux. Cela permettra également de comprendre dans quelles mesures le métabolisme peut à lui seul permettre de caractériser les différentes populations de cellules constitutives de la tumeur. Il me fallait donc développer un pipeline d'analyse, reproductible et réutilisable, permettant d'identifier les gènes représentatifs de populations et métabolismes importants pour un cancer donné. Comme le nombre de gènes à déterminer pour l'étude en transcriptomique spatiale est limité à 500, il faut que les gènes sélectionnés soient les plus pertinents possible pour permettent de recouvrir le plus de sous-populations cellulaires identifiées par les approches de transcriptomiques holistiques (scRNA-seq).

Aussi, des données RNA-seq de gliome provenant de l'Institut de Génomique Fonctionnelle (IGF) m'ont été fournies. Ces données correspondent à du séquençage *single cell* et *bulk* RNA-seq générés à partir des mêmes échantillons, l'objectif étant de combiner les informations que l'on peut déduire de ces deux types de données et tenter de ne manquer aucune sous-population, en cherchant à retrouver des groupes de cellules aux mêmes caractéristiques fonctionnelles et/ou morphologiques avec le scRNA-seq.

Les gliomes sont des tumeurs cérébrales causées par la prolifération anormales des cellules gliales, qui constituent le tissu de soutien des neurones. Ils sont catégorisés en différents grades, selon la classification de l'Organisation Mondiale de la Santé, en fonction de la densité cellulaire, de l'atypie nucléaire, de l'indice mitotique et de la nécrose définissant les grades pathologiques. Les gliomes diffus de bas grade (GDBG ou gliome de grade II) représentent 15% de l'ensemble des gliomes et évoluent inéluctablement vers des gliomes de haut grade : les gliomes anaplasiques (grade III) et les glioblastomes (grade IV) [7]. Les traitements sont limités en raison d'une compréhension incomplète des CDBG et l'ablation de la tumeur est presque impossible dû à la nature diffuse des régions sensibles

où ces tumeurs se développent. Le pronostic quant à l'évolution de ces tumeurs est variable, peu prédictible d'un patient à l'autre, avec une survie globale entre 5 ans et 15 ans.

Deux sous-populations cellulaires non chevauchantes sont retrouvées dans les GDBG : les astrocytomes, avec une morphologie étoilée, dont un des marqueurs est le facteur de transcription Sox9 qui est sur-exprimé, et les oligodendrocytes avec Olig1 comme marqueur, lui aussi sur-exprimé dans ce type cellulaire. [8]

Les gliomes dérivés des astrocytes sont les astrocytomes, et les gliomes dérivés des oligodendrocytes, moins fréquents, sont les oligodendrogiomes.

Une mutation sur l'enzyme Isocitrate déshydrogénase 1 (IDH1) est retrouvée dans ces deux types de cancer. Cette mutation conduit à un phénotype d'hyperméthylation, à cause de la production de l'oncométabolique 2-OH-glutarate, bloquant la différenciation cellulaire des cellules mutées.

Il est parfois possible d'identifier les progéniteurs d'oligodendrocytes, des cellules immatures qui ont le potentiel de se différencier en oligodendrocytes fonctionnels.

Un troisième type cellulaire retrouvés dans les tumeurs solides sont les cellules souches cancéreuses, capables de se différencier en un autre type cellulaire comme les oligodendrocytes ou les astrocytes [9].

C'est donc sur ce type de cancer que mes analyses sont effectuées afin de caractériser au mieux les sous-populations cellulaires des CDBG avec des données RNA-seq, qui me permettront de a) caractériser l'hétérogénéité fonctionnelle dans le gliome à partir de données scRNAseq et b) identifier un panel de 500 gènes métaboliques qui serviront à développer un panel de sondes MERFISH pour des analyses en transcriptomique spatiale qui viseront à étudier l'hétérogénéité métabolique spatiale.

# Méthodes

Dans l'optique de respecter les principes d'une recherche scientifique FAIR (Findable, Accessible, Interoperable, Reusable), l'ensemble des lignes de codes développées dans le cadre de mon stage sont disponibles sur le répertoire Github : [https://github.com/soeurwiki/m2\\_internship](https://github.com/soeurwiki/m2_internship)

## A. Les données RNA-seq

Les fichiers fastq des données RNA-seq m'ont été fournis par le Pr. Jean-Philippe Hugnot de l'Institut de Génomique Fonctionnelle de Montpellier, co-dirigeant l'équipe de "plasticité cérébrale, cellules souches et tumeurs gliales de bas grade" avec le Pr Hugues Duffau, dont un des thèmes de recherche principal est l'hétérogénéité des CDBG.

Huit lignées cellulaires (LGG275, LGG336, BT1, BT2, BT54, BT88, LGG85 et LGG349) dérivées de GDBG de patients cultivées sur court-terme ont fait l'objet de séquençage en *scRNA-seq* et *bulk RNA-seq*.

Chaque lignée est étudiée sous deux conditions : avec facteurs de croissance (en prolifération) ou bien sans facteur de croissance (en différenciation/quiescence).

Les facteurs de croissance ajoutés poussent les cellules vers l'état souche et induisent l'expression ubiquitaire de gènes permettant d'identifier les sous-populations de CDBG.

En retirant ces facteurs, le transcriptome des cellules est plus proche de celui retrouvés *in situ* chez les patients, et il est possible d'étudier la différenciation des cellules souches.

Ces lignées proviennent de tumeurs présentant différents grades (II, III ou IV) et sont dérivées soit d'astrocytes, soit d'oligodendrocytes et portaient initialement toutes la mutation IDH1. Les lignées BT1 et BT88 l'ont néanmoins perdues.

Il y a ainsi 8 échantillons en prolifération et 8 autres en différenciation. Au final, 16 fichiers fastq pour le *single cell RNA-seq* et 16 autres pour le *bulk RNA-seq* sont analysés.

## B. Langages de programmation

Pour le développement du pipeline d'analyse *single cell* et *bulk RNA-seq*, il m'a fallu utiliser R, Singularity, Nextflow et le langage shell GNU bash.

Les différentes librairies utilisées ainsi que leur version peuvent être retrouvées sur le répertoire Github.

Pour le *single cell*, le package Seurat, qui propose une multitude de fonctions afin d'analyser et représenter les données de milliers de cellules individuelles, a été majoritairement utilisé en suivant les recommandations proposées par le laboratoire de Rahul Satija - porteur du package Seurat.

## C. Environnement de travail

### a. Ressources

Afin de pouvoir lancer toutes les analyses qui nécessitent un stockage et une puissance de calcul importantes, un compte sur le cluster Genotoul m'a été créé. Il s'agit d'un service proposé par Genotoul, réseau toulousain de plateformes de recherche faisant partie de Genotoul GIS, une équipe de

l'INRAe MIAT (département MathNum). Depuis 2009, c'est une des 13 plateformes bioinformatiques IBISA, mettant à disposition des ressources de calcul à la communauté scientifique. Genotoul met à disposition 5000 cœurs au total, du stockage, des banques de données et plus de 200 logiciels bioinformatiques [10].

En tant qu'utilisateur, un stockage de 1To, 64 coeurs maximum et un temps de calcul équivalent à 100000h CPU me sont accordés.

Ce cluster fonctionne sous SLURM (*Simple Linux Utility for Resource Management*), un gestionnaire de ressources et ordonnanceur de tâches. Il permet de répartir les ressources de calculs entre utilisateurs selon la demande et par ordre de priorité.

Pour chaque calcul, une requête d'allocation de ressource est faite en précisant la mémoire et le nombre de cœurs requis. Les temps d'attente sont variables selon les calculs déjà en cours.

#### b. Containers singularity

De nombreux programmes sont déjà installés sur le cluster (Singularity, Nextflow, R, ...) et il est possible d'envoyer des demandes afin d'en installer de nouveaux (ou seulement pour une librairie en particulier).

Puisque pour mes analyses j'allais avoir besoin de plusieurs librairies R dont certaines nécessitent un statut *admin* pour être installées, j'ai préféré créer un *container* Singularity, contenant R et toutes les librairies dont j'aurais besoin pour ne pas avoir à demander l'installation de chaque librairie une à une. Pour l'utilisation de la librairie cellid, le docker cellid (version 0.1.0) de thugenomefacility a été utilisé. Créer ces *containers* permet également d'assurer la reproductibilité des résultats, s'inscrivant dans les pratiques FAIR.

### D. Processing des données

nf-core est un effort communautaire qui vise à créer des pipelines de bioinformatique construits sur la base de Nextflow, hautement standardisés et reproductibles, facilitant l'analyse omique à grande échelle de manière cohérente et fiable dans le respect des principes d'une recherche scientifique FAIR. Les données RNA-seq ont été traitées avec des pipelines nf-core.

#### a. Les paramètres

Le génome de référence utilisé pour les données de RNA-seq est le hg38 récupéré sur ensembl.org, de même pour l'annotation hg38 v109.

Les données de *single cell* ont été analysées avec le pipeline nf-core/scrnaseq (version 2.2.0), avec l'aligneur cellranger et adapté pour le protocol 10XV3.

Les données de bulk ont été analysées avec le pipeline nf-core/rnaseq (version 3.11.1).

#### b. Le pipelines nf-core

Le pipeline nf-core/scrnaseq permet de réaliser toutes les étapes classiques de pré-traitement de fichiers fastq provenant de données scRNA-seq issues de la technologie 10x Genomics.

A partir des fichiers fastq, un contrôle qualité est effectué pour chaque échantillon pour évaluer la qualité des données après séquençage, puis les étapes d'alignements des séquences sur le génome de référence et de comptage sont faites avec cellranger (qui inclut STAR). Un multiQC qui résume diverses informations sur la qualité des données est produit à la fin de ces étapes.

En sortie, trois fichiers par échantillon sont utilisés pour l'analyse sous R avec Seurat: barcodes.tsv qui contient la liste des code-barres de toutes les cellules présentes, features.tsv qui contient la liste des gènes quantifiés et matrix.mtx qui est une matrice de comptage avec en ligne les gènes et en colonnes, les cellules. Ces fichiers sont récupérés sans aucun filtrage pour pouvoir le gérer sous R avec Seurat.

Le pipeline nf-core/rnaseq commence également par un contrôle qualité des fichiers fastq issus du séquençage, puis les étapes d'alignement avec STAR et de quantification avec Salmon sont effectuées pour obtenir les matrices de comptage.

En sortie, des fichiers quant.sf contenant la quantification pour chaque gène, pour chaque échantillon et un fichier salmon\_tx2gene.tsv avec les correspondances entre les ID et les noms de gènes et de transcripts vont permettre de construire la matrice de comptage sur R et poursuivre les analyses.

## E. Analyse du RNA-seq cellule unique

### a. Cycle cellulaire et pourcentage de gènes mitochondriaux

Pour associer chaque cellule à une des phases du cycle cellulaire (G0/G1, S, G2/M), la fonction CellCycleScoring() va associer un score à chaque cellule selon son expression des gènes signatures des phase S ou G2/M, récupéré de Tirosh et al, 2015 [11] et prédire ensuite leur appartenance à l'une des trois phases.

Le pourcentage de gènes mitochondriaux au sein de chaque cellule peut être calculé sur la matrice de comptage brut avec PercentageFeatureSet(seurat\_object, pattern = "MT-"), qui permet de'évaluer l'expressions de tous les gènes dont le nom commence par "MT-", c'est-à-dire les gènes mitochondriaux humains d'après la convention de notation Ensembl, pour calculer le pourcentage correspondant.

### b. Filtrage

Les 16 échantillons ont été filtrés individuellement en se basant sur le pourcentage d'ARN mitochondrial contenu dans chaque cellule, ainsi que le nombre de séquences générées par cellule et de gènes considérés comme exprimés par celles-ci.

Dans un premier temps, suivant les recommandations du Satija Lab, ont été retenues les cellules exprimant au minimum 220 gènes et les gènes ayant été comptabilisés au moins une fois dans 3 cellules différentes. Ensuite, les cellules ne présentant pas un nombre de gènes considérés comme exprimés supérieur ou égal à 500 ont été écartées de l'étude. De même pour les cellules présentant plus de 15% de séquences provenant de gènes mitochondriaux car un trop grand pourcentage est souvent un indicateur d'une cellule en apoptose (mort cellulaire).

`DoubletFinder()` , de la librairie DoubletFinder, permet un deuxième filtrage en retirant les cellules qui semblent être en réalité la combinaison de 2 cellules, dû à un problème technique lors de la séparation des cellules par microfluidique, avant le séquençage. Lorsque chaque cellule est capturée dans une gouttelette en scRNA-seq, il arrive que deux cellules soient emprisonnées dans la même gouttelette.

### c. Normalisation

La normalisation est une des étapes les plus importantes pour correctement analyser les données scRNA-seq afin de limiter les biais et récupérer une information qui représente une réalité biologique. Deux stratégies sont possibles.

La première consiste à utiliser la fonction `NormalizeData()` qui effectue une normalisation Log. Le principe repose sur la division de l'expression de chaque gène par l'expression totale de la cellule, multiplié par un facteur (10000 par défaut) auquel est appliqué le logarithme naturel  $\log_{10}$ . Avec `ScaleData()`, l'expression de chaque gène est décalée pour que l'expression moyenne entre cellules soit égale à 0 et la variance égale à 1, pour que les gènes fortement exprimés ne dominent pas sur les autres. On obtient ainsi des données centrées-réduites.

La méthode SCT qui combine la normalisation et la stabilisation de la variance peut aussi être utilisée avec `sctransform()`. Elle est basée sur le postulat que le nombre de gènes et le nombre d'UMI détectés dans une cellule présentent une relation presque linéaire [12]. Un modèle linéaire généralisé avec le nombre d'UMI comme variable de réponse est calculé pour décrire le bruit technique à partir des données et le retirer. Chaque UMI est transformé en résidu de Pearson qui sera utilisé comme comptage normalisé.

### d. Intégration des données

Les seize échantillons sont combinés pour pouvoir les étudier comme un seul objet Seurat. A nouveau, deux méthodes sont possibles: les matrices de comptage brut de tous les échantillons peuvent être simplement combinées avec `merge()` ou bien ce sont les matrices de comptage normalisées qui vont être combinées, ce qui nécessite une intégration particulière des données. Pour les intégrer, des *anchors* sont déterminés grâce à la fonction `FindIntegrationAnchors()`. Elle repose sur l'hypothèse qu'il y a au moins un petit groupe de cellules partageant le même état biologique entre deux jeux de données.

Les jeux de données sont projetés sur un espace défini par leur structure de corrélation partagées. Les paires de cellules entre les deux jeux de données qui partagent leurs caractéristiques seront utilisées comme *anchors*.

L'intégration des données se fait avec la fonction `IntegrateData()`, en s'appuyant sur les *anchors* définis précédemment pour supprimer au mieux les différences non biologiques entre les données (effet *batch*) possibles.

#### e. Méthode de réduction d'effet *batch*

Une analyse en composantes principales (ACP) avec `runpca()` sur l'objet Seurat normalisé est faite pour expliquer la variabilité du jeu de données. Ce sont les variables de variances maximales obtenues avec les combinaisons linéaires des variables d'origine qui permet d'obtenir les composantes principales.

Pour supprimer au mieux les effets *batch* possibles, Harmony est appliqué sur l'objet Seurat et va utiliser un partitionnement en  $k$ -moyennes, pour former  $k$  clusters dans un espace utilisant toutes les composantes principales des données, calculé avec l'ACP. Chaque cluster doit être équilibré dans sa diversité cellulaire, basé sur les lignée et les conditions. Une correction est appliquée sur ces clusters pour recentrer les cellules au sein de leur cluster respectif pour corriger les composantes principales. Cette étape de création de clusters et de réajustement des composantes principales est réitérée jusqu'à stabilisation [13].

#### f. Le UMAP

Le Uniform Manifold Approximation and Projection (UMAP) est une méthode de réduction de dimensionnalité non linéaire basé sur la géométrie riemannienne et la topologie algébrique, avec laquelle il est possible de représenter les données sur un graphe bidimensionnel en conservant autant que possible la structure globale des données, avec la possibilité d'interpréter la distance entre les cellules (plus elles sont proches sur l'UMAP, plus elles ont des caractéristiques similaires) [14]. Cette représentation est obtenue en utilisant la fonction `runUMAP()` et en précisant 'Harmony' en réduction pour travailler avec les données corrigées.

#### g. Méthode de *clustering*

Pour définir des sous-populations de cellules, il est possible de former des clusters, soit des groupes de cellules - sur la base de la similarité de leur profil d'expression génique, grâce aux fonctions `FindNeigbors()` suivi de `FindClusters()`.

Seurat utilise une approche de clustering de graphes basée sur l'algorithme KNN (K-nearest neighbor) où les sommets représentent les cellules et les arêtes relient celles ayant des profils d'expression similaires. Par défaut, le nombre de cellules voisines pour un sommet donné est égal à 20 dans `FindNeigbors()` et ils sont déterminés selon la distance euclidienne dans l'espace ACP. Le poids des arêtes entre deux cellules est attribué à l'aide de l'indice et la distance de Jacquard qui permet de comparer la similarité et la diversité entre les cellules.

`FindClusters()` sépare ensuite les cellules en groupes de manière itérative selon leur similarité en utilisant l'algorithme de Louvain, un algorithme glouton dont le paramètre de résolution impactant la taille des clusters formés est à préciser par l'utilisateur. Plus la résolution donnée sera élevée et plus le nombre de clusters augmentera.

Afin d'avoir une idée de la cohérence de la résolution choisie pour l'objet Seurat, le librairie `clustree` est utilisé pour tracer le devenir des cellules réparties en clusters selon les résolutions sous la forme d'un arbre.

En visualisant la répartition des cellules entre clusters selon les résolutions, il peut être plus facile de déterminer les résolutions où il y a un over-clustering par exemple, si des clusters stables sur plusieurs

résolutions forment tout d'un coup des clusters avec des cellules qui ne provenaient pas du même noeud parent.

#### h. Analyse différentielle

Une fois les données séparées en clusters de cellules, il faut trouver les gènes les plus intéressants pour définir ces derniers. Pour cela, la fonction `FindConservativeMarkers()` de Seurat est utilisée pour trouver les gènes différentiellement exprimés entre les clusters.

Puisque nous sommes dans un cas où deux conditions (prolifération *vs* différenciation) sont présentes dans nos données, il faut éviter de trouver des gènes considérés comme intéressants alors qu'il ne résulte que d'une différence entre conditions, un biais qu'on évite grâce à cette fonction, en précisant le paramètre '`grouping.var=condition`'.

Les cellules des deux conditions sont étudiées séparément. Si des gènes sont différentiellement exprimés dans un cluster comparé à tous les autres mais que leur expression est similaire entre toutes les cellules de ce-dit cluster, indépendamment de la condition, ils sont alors retenus.

Pour considérer un gène comme différentiellement exprimé (positivement ou négativement) c'est le test statistique de Wilcoxon-Mann-Whitney qui permet de vérifier si les valeurs moyennes des deux clusters testés diffèrent significativement l'une de l'autre. Par défaut, les gènes ne sont testés que s'ils sont détectés à au moins 0,1% dans un des groupes de cellules et qu'ils présentent un  $\log_2(\text{fold-change})$  d'au moins 0.25.

#### i. Focalisation sur le métabolisme

Pour pouvoir définir l'hétérogénéité métabolique dans les données scRNA-seq, l'objet Seurat est filtré de tous les gènes ne faisant pas partie du métabolisme. Toutes les étapes de clustering et d'analyse différentielle sont à nouveau effectuées sur cet objet réduit.

Ce filtrage se base sur une liste de 2921 gènes classés comme étant impliqués dans le métabolisme, récupérée de l'équipe "Métabolisme et xénobiotiques" du laboratoire Toxalim situé à l'INRAE de Toulouse, comprenant le Dr Nathalie Poupin, le Dr Fabien Jourdan et Maximilian Stingl avec qui ce projet est en collaboration.

Il est ainsi possible de comparer la quantité d'information perdue ou ajoutée par rapport à l'analyse plus générale, en reprenant les étapes de clustering et d'analyses différentielles.

#### j. Annotations des types cellulaires

La librairie SCINA utilise un algorithme semi-supervisé de détection automatique de type cellulaire selon les gènes signatures sur-exprimé (ou sous-exprimés) spécifiques de chaque type pour pouvoir annoter les cellules.

Dans cette analyse, deux types cellulaires sont précisés pour l'annotation: les gènes SLC1A3, NFIA, SOX9, GFAP, APOE, AQP4, ALDH1L1, FABP7 et TNC forment la signature des astrocytes et les gènes SOX8, SOX11, CLDN11, MBP, SOX10, SOX4, MOG, MYT1, CNP, PLP1, OLIG1, OLIG2, NKX2-2, ERBB3, UGT8, SOX17, GPR17 et TNR, celle des oligodendrocytes. Cette liste non exhaustive m'a été fournie par le Pr Jean-Philippe Hugnot de l'IGF et représente les gènes majoritairement sur-exprimés dans ces populations cellulaires.

# Résultats

L'objectif de mon projet a été de concevoir et de développer un ensemble de routines analytiques robustes et réutilisables destiné à l'analyse de données issues du séquençage d'ARN cellule unique (*scRNA-seq*), pour **a)** assurer la caractérisation de l'hétérogénéité métabolique fonctionnelle au sein des tumeurs solides à partir de données *scRNA-seq* et **b)** automatiser et simplifier l'identification d'un panel de gènes d'intérêt focalisé sur le métabolisme pour l'étude de l'hétérogénéité métabolique spatiale par des approches de transcriptomique spatiale multiplexée. Ces outils informatiques ont été élaborés et éprouvés sur des données *scRNA-seq* générées à partir de gliomes humains.

## A. Présentation de la stratégie analytique

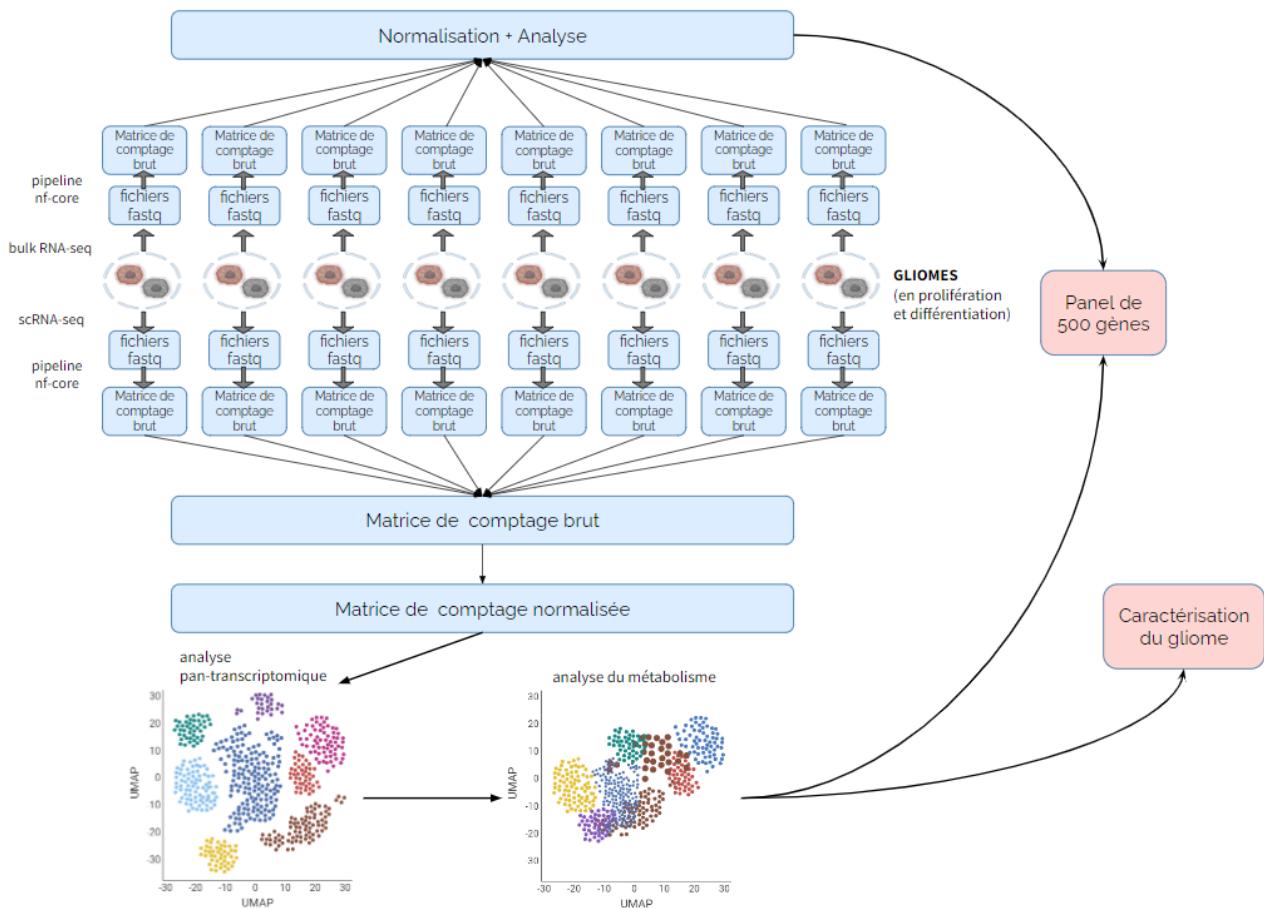


Schéma 1 : Schéma présentant la stratégie d'analyse des données single cell et bulk RNA-seq de gliomes.

Afin d'analyser les données de séquençage d'ARN, tant au niveau de cellules uniques (*single-cell*) que de prélèvements globaux (*bulk*), j'ai élaboré une stratégie analytique après avoir procédé à une lecture critique et approfondie de la littérature - visant à identifier les outils les plus adaptés et les plus robustes [15, 16]. Cette démarche a été conçue avec pour objectif final l'identification d'une liste restreinte de gènes impliqués dans le métabolisme, qui se révèlent pertinents pour la caractérisation des tumeurs, et plus spécifiquement, des gliomes.

Puisque les données RNA-seq vont subir toute une série d'analyses pour déterminer cette liste de gènes, il est possible par la même occasion de proposer une caractérisation biologique et métabolique de celles-ci, vierges de toutes analyses encore.

Pour la première étape de pré-traitement des données, cellranger et STAR étaient les aligneurs recommandés pour des données de *scRNA-seq* d'après l'étude de Brüning et al.(2021) [17]; c'est pourquoi dans le pipeline d'analyse nf-core, cellranger a été choisi plutôt que d'autres aligneurs [Méthodes D].

Ensuite, une fois les matrices de comptages obtenues, les analyses sous R ont pu être effectuées avec Seurat, package regroupant un ensemble de fonctions et méthodes conçus pour le contrôle qualité, l'analyse et l'exploration du *scRNA-seq*, en suivant les recommandations et bonnes pratiques explicitées dans le manuel de l'outil [18].

## B. Caractérisation générale du jeu de données

### a. Caractéristiques des échantillons

Les données *scRNA-seq* de 8 lignées de gliomes dérivés de patients obtenus sous 2 conditions (prolifération et différenciation) sont analysées. Comme détaillé sur le tableau 1, ces dernières correspondent à différents grades de gliome et portent la mutation IDH1 (mutation de l'enzyme Isocitrate déshydrogénase 1), sauf pour BT1 et BT88 qui l'ont perdu.

Lignée	IDH1	Grade	Type
<b>LGG275</b>	muté	II	Astrocytomes
<b>LGG336</b>	muté	II	Astrocytomas
<b>BT1</b>		III	Oligodendrogiome
<b>BT2</b>	muté	III	Oligodendrogiome
<b>BT88</b>		III	Oligodendrogiome
<b>BT54</b>	muté	III	Oligodendrogiome
<b>LGG85</b>	muté	IV	Astrocytomes
<b>LGG349</b>	muté	IV	Astrocytomes

Table 1 : Caractéristiques des lignées dérivées de gliomes de patients

Dans les analyses *scRNA-seq*, la qualité des cellules considérées *in fine* dans l'analyse est cruciale pour minimiser les artefacts et les biais. En *scRNA-seq*, les critères d'évaluation de la qualité incluent classiquement le nombre de gènes considérés comme exprimés par cellule, servant de proxy pour évaluer la richesse du transcriptome. Une proportion élevée d'ARN mitochondrial par rapport à l'ARN total est souvent indicative de cellules stressées ou apoptotiques. Par ailleurs, l'identification et l'élimination des droplets vides et des doublets (droplets contenant deux cellules) sont essentielles pour prévenir les contaminations. Une proportion élevée d'ARN ribosomal peut quant à lui être un marqueur

de dégradation de l'ARN, compromettant la fiabilité des données. Enfin, un nombre de comptage par cellule trop bas est souvent significatif d'une cellule endommagée ou en cours de lyse ayant ainsi moins d'ARN disponible pour le séquençage, ce qui se traduit par des comptages plus bas.

Ces critères sont utilisés pour effectuer un filtrage rigoureux des cellules avant les analyses en aval, assurant ainsi l'intégrité et la robustesse des interprétations biologiques. Ainsi, les 16 échantillons sont filtrés un à un pour retirer les cellules considérées comme étant de mauvaise qualité [détaillé dans Méthodes E-a].

	Condition	Nb de gènes	Nb de cellules	Cellules filtrées		Nb de cellules après filtrage
				nb	%	
<b>LGG275</b>	diff.	25626	2239	298	13%	1941
	prolif.	26360	2629	249	9%	2380
<b>LGG336</b>	diff.	29657	6026	825	14%	5201
	prolif.	28528	4860	1137	23%	3723
<b>BT1</b>	diff.	28186	6502	1869	29%	4633
	prolif.	25726	2569	462	18%	2107
<b>BT2</b>	diff.	26817	3984	765	19%	3219
	prolif.	25000	1800	192	11%	1608
<b>BT88</b>	diff.	28238	5393	1781	33%	3612
	prolif.	28799	4298	652	15%	3646
<b>BT54</b>	diff.	26841	2607	1628	<b>62%</b>	979
	prolif.	28556	12625	7028	<b>56%</b>	5597
<b>LGG85</b>	diff.	28541	3258	787	24%	2471
	prolif.	25121	1586	140	9%	1446
<b>LGG349</b>	diff.	29131	3366	543	16%	2823
	prolif.	29084	3256	470	14%	2786

Tableau 2 : Nombre de gènes exprimés et nombre de cellules avant et après filtrage par échantillon.

Le tableau 2 met en évidence une grande disparité dans le pourcentage de cellules filtrées par échantillon. Certains d'entre eux ont dû être filtrés d'énormément de leurs cellules, comme la lignée BT54 où plus de la moitié est filtrée (62% de cellules en condition de différenciation et 56% en condition de prolifération). La filtration concernant les autres lignées restent plutôt comprises entre 10 et 30% de leurs cellules.

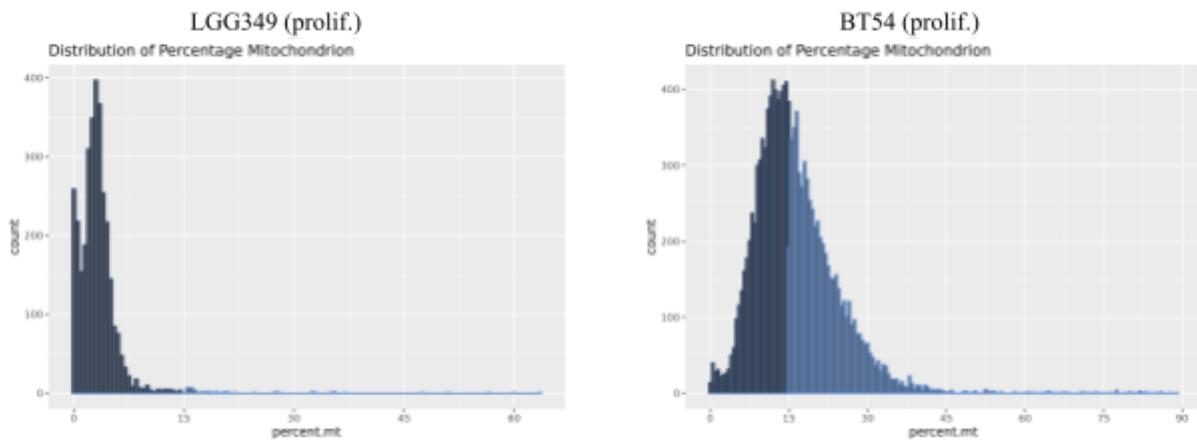


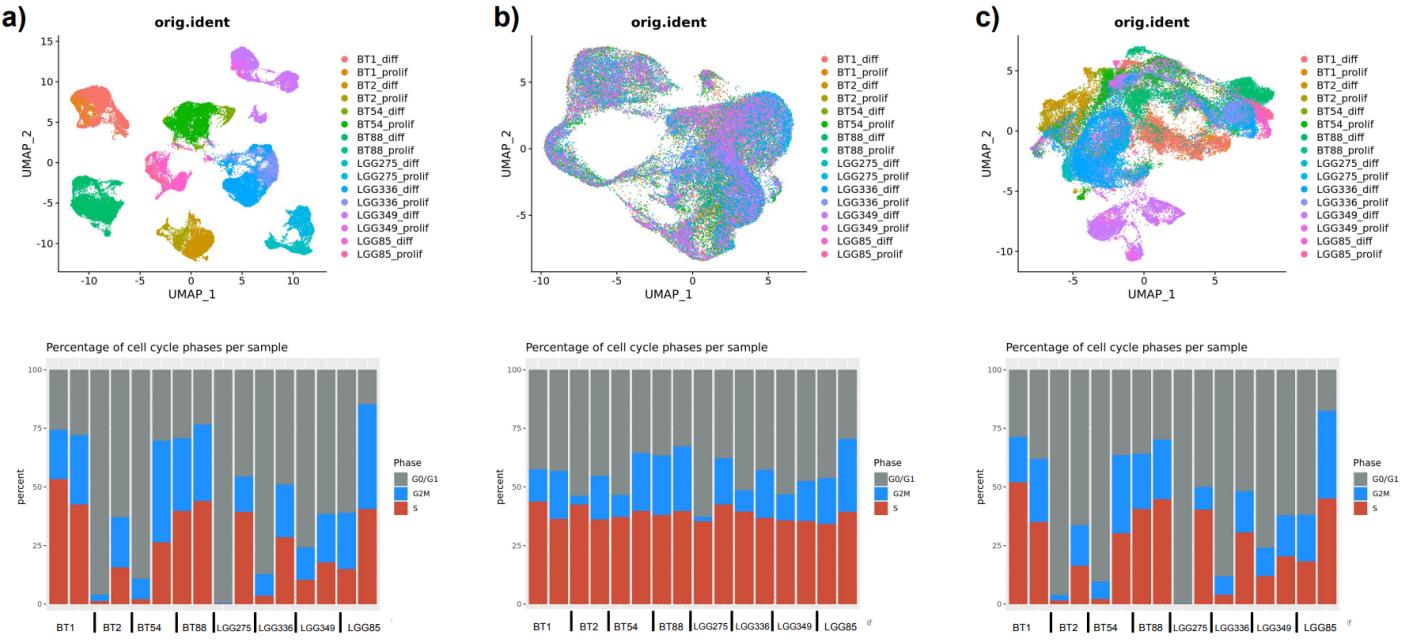
Figure 2: Pourcentage de gènes mitochondriaux avant filtrage dans la lignée LGG349 en condition de prolifération (représentant le profil moyen retrouvé dans les autres lignées) contre la lignée BT54 (aussi en condition de prolifération).

Ce filtrage important de BT54 s'explique par son pourcentage de gènes mitochondriaux. Au lieu d'avoir une moyenne aux alentours de 5 à 10% exprimés dans une cellule, elle possède une moyenne comprise entre 15 et 20% (Fig. 2). Plusieurs raisons pourraient expliquer cette sur-expression mitochondriale; une surexpression de gènes mitochondriaux peut indiquer une augmentation de l'activité métabolique, généralement associée à une demande accrue en énergie; il peut s'agir d'une réponse aux exigences de croissance et de prolifération rapides que demande une tumeur ou encore un artefact technique lors de la préparation des échantillons et/ou le séquençage pourrait expliquer ce pourcentage mitochondriale.

Une fois tous les échantillons filtrés, ils vont pouvoir être combinés pour la suite des analyses.

### b. Normalisation et réduction de dimensionnalité

Lors du séquençage, chaque échantillon, autrement dit chaque lignée pour chaque condition, a été séquencé séparément ce qui introduit de nombreuses sources d'effet *batch* qu'il faut minimiser. Pour analyser toutes ces lignées, plusieurs stratégies ont été explorées. La première consistait à combiner toutes les lignées et les normaliser avec la méthode SCT (détalée dans Méthode E-c). Sur le UMAP tracé dans la figure 3-a, il est clair que les sources de variation des données ne sont pas traitées, c'est-à-dire qu'on se rend compte d'un effet *batch* majeur, toutes les lignées étant clairement séparées les unes des autres.



(Pour chaque lignée, la condition différentiation est présentée avant celle de prolifération.)

*Figure 3: Représentation des UMAP obtenus pour le jeu de données après différentes stratégies de normalisation et représentations des pourcentages des différentes phases du cycle cellulaires pour chaque lignée de chaque condition. a) échantillons combinés (comptages bruts) puis normalisation SCT, b) échantillons traités un à un avec la méthode SCT puis intégrer en une seule matrice de comptage, c) échantillons combinés (comptage bruts) puis normalisation LogNormalize suivi de l'algorithme Harmony.*

Il est recommandé par Choudhary et Satija (2021) [19] de normaliser chaque échantillon séparément à cause des trop grandes variations entre les jeux de données qu'il peut y avoir. En suivant la vignette Seurat “Introduction to SCTransform, v2 regularization”, chaque échantillon a donc été normalisé, toujours avec la normalisation SCT, avant d'être combiné aux autres grâce à des méthodes d'intégration [Méthodes E-c]. Avec le UMAP généré à partir de ces données(Fig. 3-b), il ne semble pas y avoir d'effet *batch*, aucune lignée ou condition n'étant séparée des autres. Cependant, si l'on s'intéresse à la distribution des cellules dans les différentes phases du cycle cellulaire, par échantillon, alors que l'on s'attendrait à ce que les lignées en différenciation soient majoritairement enrichies en cellules en phase G0/G1 par rapport à celle en prolifération, les distributions sont homogènes. Aussi, cette stratégie de normalisation est beaucoup trop agressive et retire la variabilité technique du jeu de données, au détriment de la variabilité biologique.

Il fallait donc trouver une méthode de normalisation des données qui soit plus conservatrice des variations biologiques tout en corrigeant les variabilités techniques.

Une troisième méthode est donc testée. Comme pour la première stratégie, les échantillons sont finalement tous combinés et normalisés, mais avec la méthode LogNormalize cette fois-ci. Ensuite, pour retirer les variations liées aux différentes lignées et conditions, la correction Harmony est appliquée [20]. Sur le UMAP (Fig. 3-c), la distribution des cellules dans les différentes phases du cycle cellulaire se rapproche grandement de ce qu'on attendait : une différence notable entre les conditions de différenciation et de prolifération est retrouvée pour presque toutes les lignées.

C'est donc avec cette dernière stratégie, compromis entre les deux premières méthodes, que les analyses ont été poursuivies.

### c. Caractérisations des lignées

A présent que toutes les données sont combinées et normalisées, il est possible de travailler avec un seul objet Seurat pour mieux analyser le gliome dans son ensemble et tenter de comprendre les spécificités des données étudiées.

Differentes caractéristiques relatives aux cellules analysées sont projetées sur un UMAP en figure 4, comme leur lignée d'origine, le grade du gliome dont elles sont issues, la condition dans laquelle elles ont été cultivées, leur statut mutationnel pour IDH1, la phase du cycle cellulaire dans laquelle elles se trouvent [Méthodes E-a] et le type cellulaire auquel elles sont associées (astrocytes ou oligodendrocytes) [Méthodes E-j].

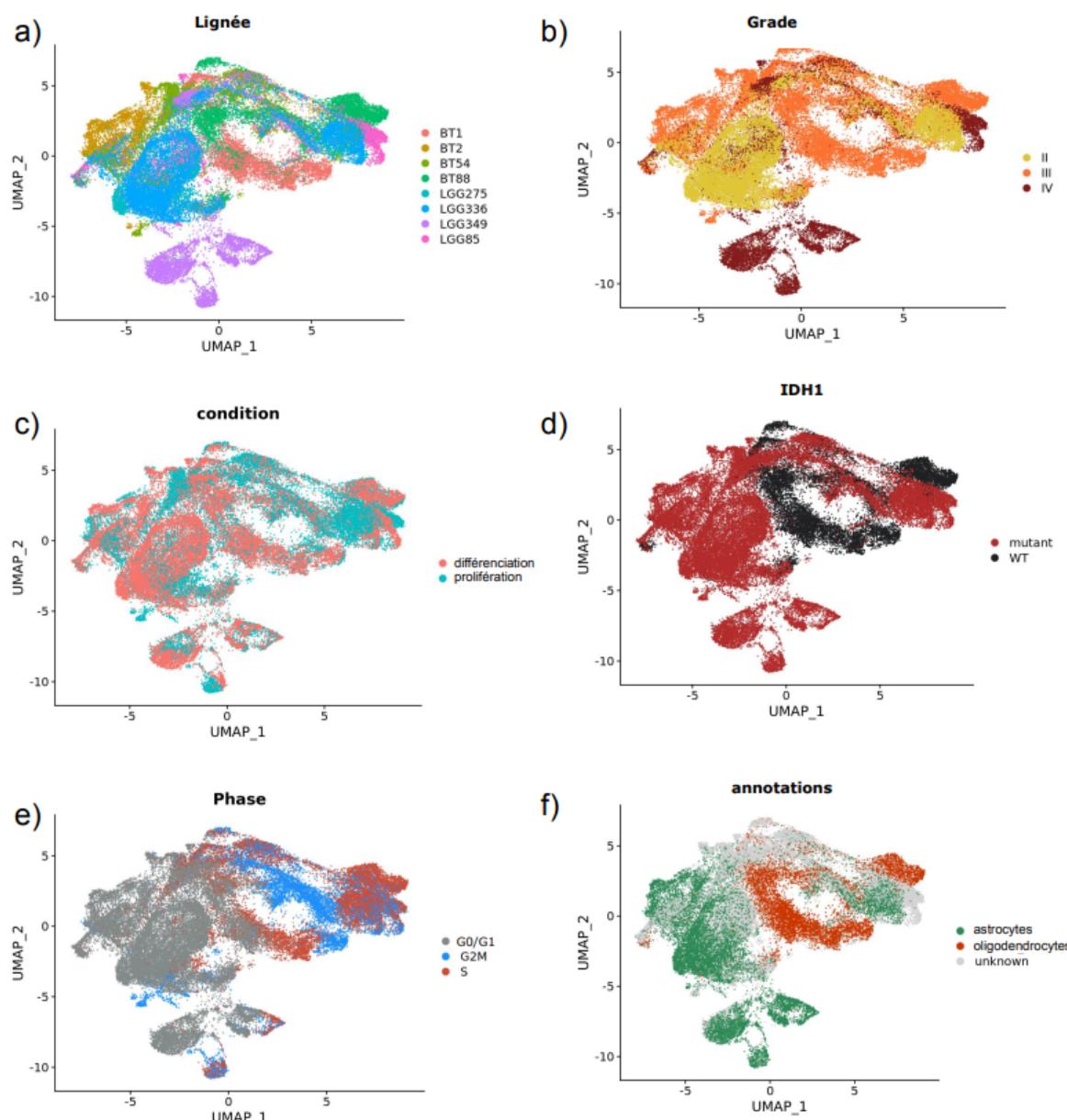


Figure 4: Représentation UMAP des cellules de gliome étudiées, colorées selon différentes caractéristiques : a) la lignée dont elles proviennent, b) le grade du gliome duquel elles font parties (grade II, III o IV), c) la condition (prolifération ou différenciation), d) la présence de la mutation IDH1 dans la lignée (WT : wild-type, non-muté), e) la phase du cycle cellulaire G0 ou G1, G2 ou M et S, f) le type cellulaire (astrocytes et oligodendrocytes) déterminé par l'algorithme SCINA [Méthodes E-j]

La lignée BT54, qui a un pourcentage mitochondrial élevé comparée aux autres lignées, ne se démarque pas spécialement des autres, elle sera donc gardée pour les analyses et non considérée comme un *outlier* à cause d'erreurs de préparation de librairies ou de séquençage..

La lignée LGG349 se démarque particulièrement des autres en étant presque totalement isolée. Elle fait partie du stade les plus avancés du gliome, le grade IV. En revanche, LGG85, un grade IV aussi, n'est pas isolé de manière aussi drastique.

Toutes les lignées possèdent au moins une partie de leur cellules qui se regroupent uniquement entre elles.

Les conditions (prolifération et différentiation) ne jouent pas une place majeure justifiant la répartition des cellules sur la UMAP.

Avec les marqueurs majoritaires des populations oligodendrocytes et astrocytes [Méthode E-j], on retrouve que BT1 et BT88 (grade III) sont les deux lignées dont la population oligodendrocytes est le mieux reconnue. Ce sont aussi les deux seuls dont la caractéristique est de ne pas porter la mutation IDH1.

Pour les autres grades II et IV, on retrouve une population majoritaire d'astrocytes.

BT2 et BT54 n'ont pas la population d'oligodendrocytes que l'on se serait attendu à voir, mais plutôt une population mixte, voir même majoritaire en astrocyte pour BT2. Ces lignées sont issues d'oligodendrogiomes (des gliomes dérivés d'oligodendrocytes) et d'après les études d'anatomo-cyto-pathologie, ce type de gliome est majoritairement composé de ce type cellulaire [21].

Le cycle cellulaire pourrait avoir joué un rôle important dans la séparation des cellules. Un nombre important des phases G0/G1 sont regroupés ensemble, bien qu'une partie d'entre elles chevauchent des cellules d'une phase S. Les phases G2/M et S possèdent un groupe de cellules isolé des autres phases, mais forment également des groupes mixtes.

## C. Caractérisation biologique des lignées de gliomes étudiées

### a. Caractérisation sur la base du transcriptome complet

Pour caractériser les sous-populations du gliome, 18 clusters sont formés avec une résolution choisie de 0.4 qui correspond à un *clustering* où les cellules d'un cluster ne correspondent pas un mélange de cellules provenant de cluster parents différents sur l'arbre tracé en figure 5-a [Méthodes E-g].

Les clusters comportent tous un ensemble de cellules qui se trouvaient proches sur le UMAP, excepté pour le cluster 7 qui est composé de deux groupes séparés de cellules.

Huit de ces clusters sont majoritairement formés de cellules en phase G0/G1.

A ce stade du projet, il serait prématuré de proposer une caractérisation biologique exhaustive des différentes populations cellulaires identifiées tant les informations obtenues sont riches et complexes. Cependant, mes analyses révèlent d'ores et déjà plusieurs résultats importants. En plus de révéler une grande hétérogénéité fonctionnelle - 18 populations cellulaires ont été identifiées - elles mettent en effet en évidence qu'il existe des populations cellulaires communes et d'autres spécifiques des patients dont sont originaires les tumeurs étudiées.

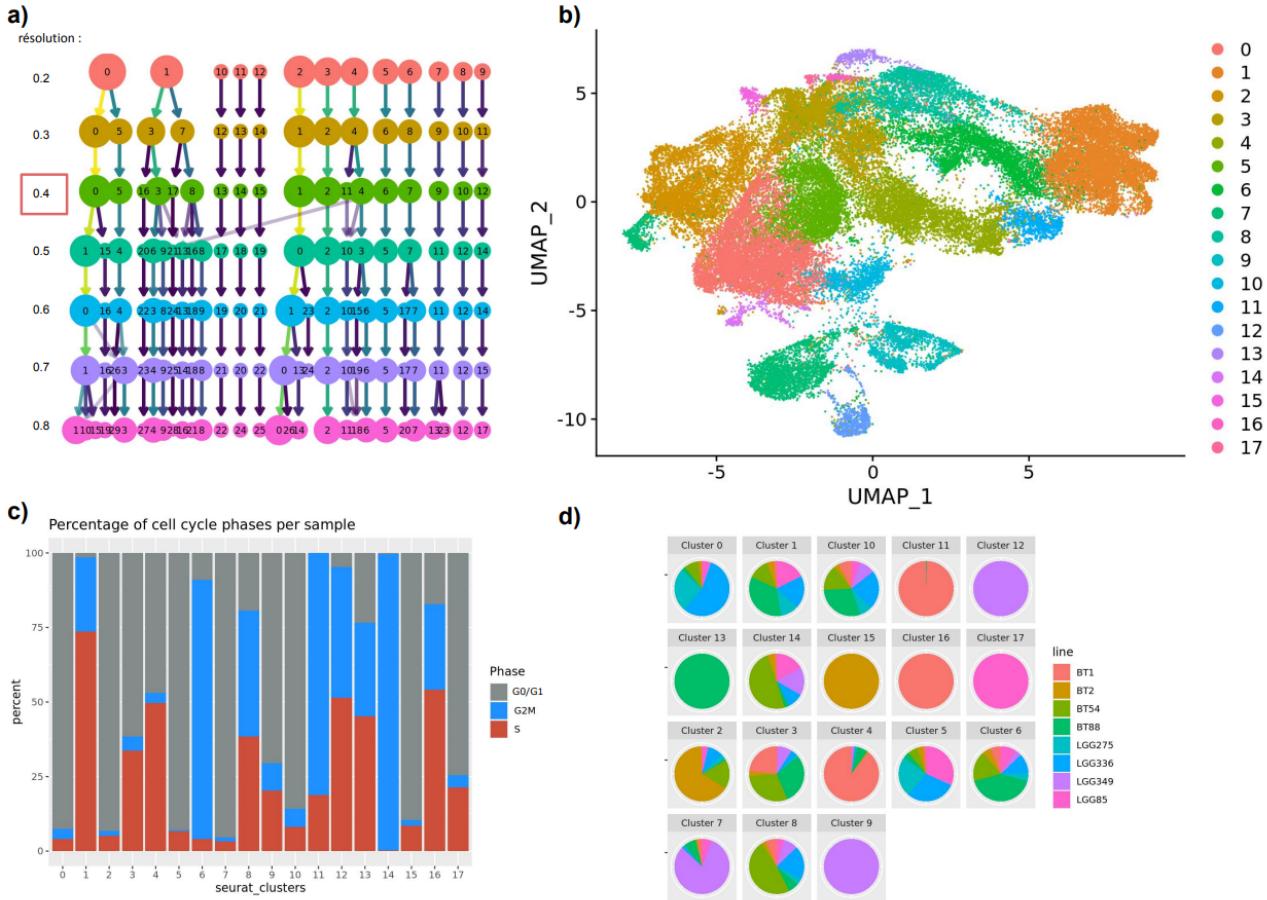


Figure 5: Détermination et caractérisation des clusters à partir des données de gliomes. a) arbre obtenu avec clustree, traçant le devenir des cellules en fonction des résolutions choisies [Méthodes E-g], b) UMAP coloré selon les clusters identifiés, c) Pourcentage des phases de cycle cellulaires trouvées dans chaque cluster; d) pourcentage des lignées retrouvées dans chaque cluster.

Ainsi le cluster 10 - qui n'est caractéristique d'aucun grade, et ne dépend ni du statut mutationnel IDH1 ni du statut de différenciation ou de prolifération - comprend des cellules issues de l'ensemble des tumeurs. Ces cellules constituent donc un cœur commun à l'ensemble des gliomes étudiés. Il apparaît que la majorité des cellules présentes dans ce cluster sont en phase G0/G1 et ne se divisent donc pas. Ces résultats dérivés des analyses d'enrichissement [Annexe 1] en gènes signature du cycle cellulaire sont corroborés par les analyses de surreprésentation en termes biologiques qui montrent que l'expression des gènes du cycle cellulaire sont réprimés alors que ceux de la voie p53 sont sur-exprimés - verrouillant ces cellules dans un état de sénescence ou les orientant vers l'apoptose.

Les clusters 12 et 9 - quant à eux - sont complètement spécifiques de l'échantillon LGG349, un astrocytome de grade IV. Ils sont composés de cellules à l'activité mitotique et proliférative relativement importante. L'analyse fonctionnelle des gènes qu'expriment les cellules constitutives de ces clusters révèle une activité accrue des gènes impliqués dans la synthèse des protéoglycans et dans le réseau de la protéine ECM1 qui jouent un rôle majeur dans la transition épithélio-mésenchymateuse et le processus métastatique.

## b. Focalisation sur le métabolisme

Afin de sélectionner les gènes qui représentent le mieux les différentes populations présentes dans ces échantillons de gliome, des méthodes de clustering sont appliquées en ne se focalisant que sur les gènes du métabolisme: sur les 31722 gènes exprimés dans les lignées, seuls 2814 sont impliqués dans le métabolisme et sont conservés [Méthode E-i].

Le UMAP obtenu avec les gènes du métabolisme est bien différent de l'étude générale incluant tous les gènes (Fig. 6). Il n'y a pas de groupes très distincts selon la lignée ou le grade, bien que LGG349 garde une grande partie de ses cellules isolées des autres. La séparation selon le cycle cellulaire est en partie retrouvée, avec les phases G0/G1 regroupées ensemble en grande majorité. L'annotation des cellules selon le type cellulaire est gardée de celle définie dans l'étude globale.

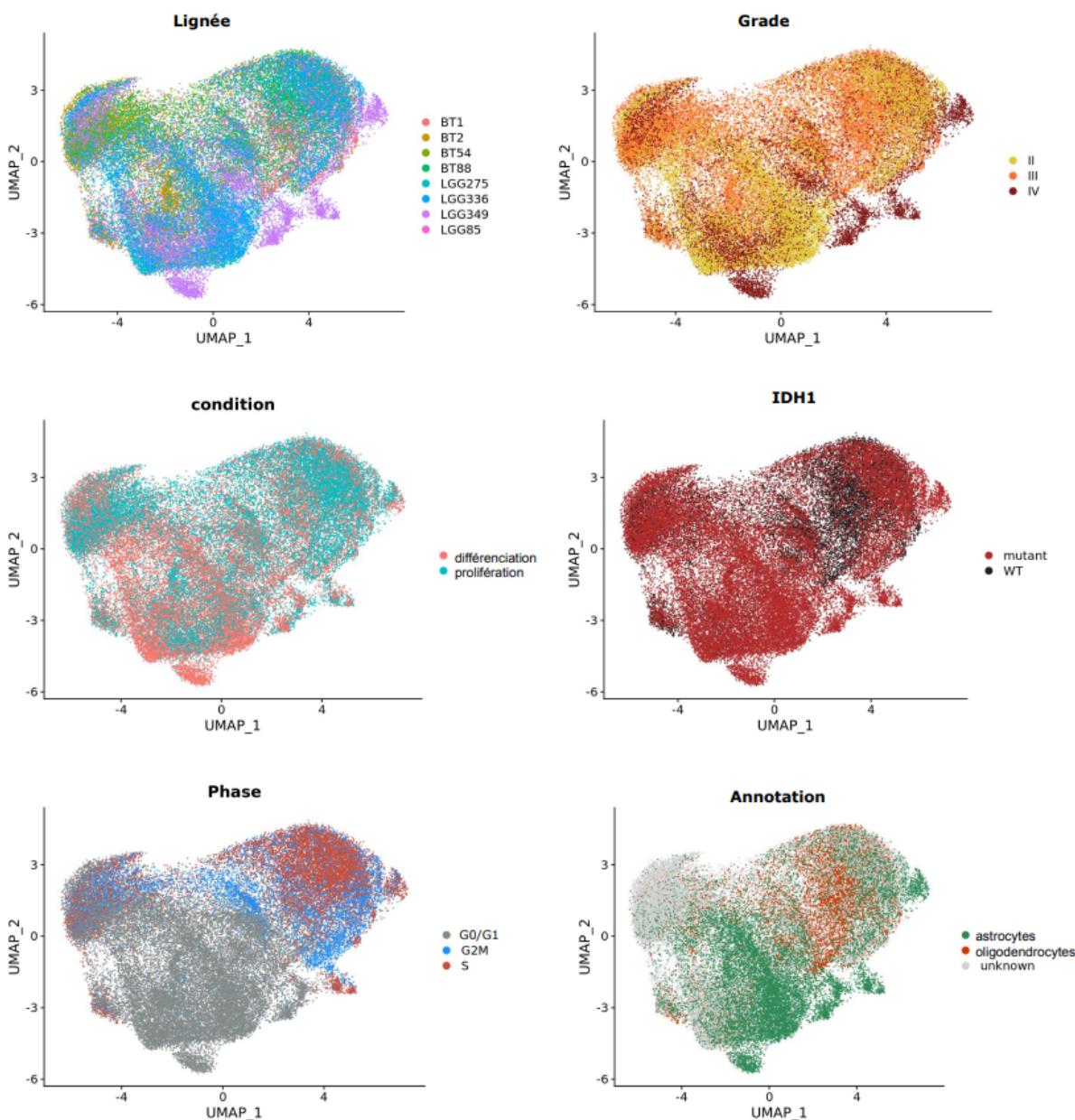


Figure 6: Représentation UMAP des cellules de gliome étudiées avec seulement les 2814 gène du métabolisme, colorées selon différentes caractéristiques : a) la lignée dont elles proviennent, b) le grade du gliome duquel elles font parties (grade II, III o IV), c) la condition (prolifération ou différenciation), d) la présence de la mutation IDH1 dans la lignée (WT : wild-type, non-muté), e) la phase du cycle cellulaire G0 ou G1, G2 ou M et S, f) le type cellulaire (astrocytes et oligodendrocytes) déterminé par l'algorithme SCINA [Méthodes E-j]

Avec ces 2814 gènes, il est encore possible de séparer les données en 11 clusters. La résolution qui a été choisie est 0.4. Il est compliqué de déduire la résolution où les clusters sont les plus stables car dès les plus petites, des échanges de cellules sont faits entre clusters (Fig. 7-a).

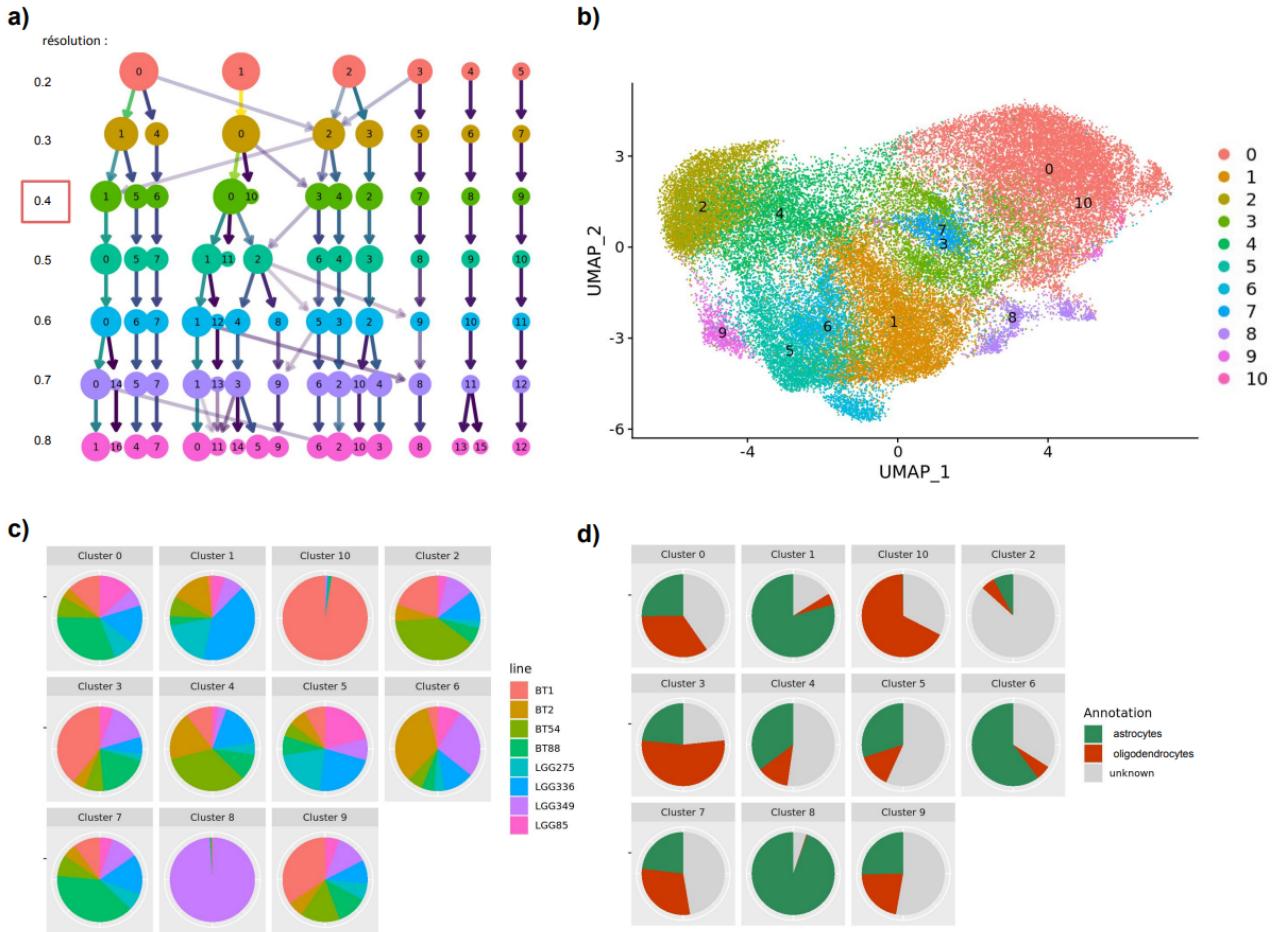


Figure 7: Détermination et caractérisation des clusters à partir des données de gliomes. a) arbre obtenu avec clustree, traçant le devenir des cellules en fonction des résolutions choisies [Méthodes E-g], b) UMAP coloré selon les clusters identifiés, c) pourcentage des lignées retrouvées dans chaque cluster, d) Pourcentage des types cellulaires (astrocytes et oligodendrocytes) retrouvés dans chaque cluster.

Sur le UMAP en figure 7-c, seulement 2 clusters (8 et 10) ne sont composés que d'une lignée en très grande majorité, LGG349 et BT1 respectivement. Elles présentent donc un métabolisme spécifique de leur lignée qui permet leur séparation en des clusters distincts. L'analyse fonctionnelle révèle que la population caractéristique de l'échantillons LGG349 (cluster 8) présente un métabolisme ré-axé autour du catabolisme des sphingolipides et la synthèse des glycane, alors que celle caractéristique de l'échantillons BT1 (cluster 10) est orientée sur la synthèse des lipides insaturés et le métabolisme des monocarbone [Annexe 2]. A noter que le cluster 8 est caractéristique des astrocytomes.

L'ensemble des autres clusters comportent des cellules constitutives de toutes les tumeurs analysées. Ces résultats mettent donc en évidence qu'il existe une grande hétérogénéité fonctionnelle métabolique au sein des différents gliomes étudiés. Alors que certaines populations présentent une phosphorylation oxidative importante (cluster 2 et 4) laissant présager que les ressources en oxygène dont elles disposent ne sont pas limitantes, d'autres populations présentes dans la même tumeur

présentent un statut OxPhos réduit et un métabolisme centré sur les purines (cluster 0), les acides gras (cluster 4) ou encore les acides aminés à chaîne latérale ramifiée (cluster 1).

### c. Comparaison entre les classifications pan-transcriptome et métabolisme-centrique

Avec la séparations en clusters des cellules basées sur tous les gènes puis focalisé sur le métabolisme uniquement, une comparaison de ces deux résultats peut être faite afin de déterminer à quel point le métabolisme peut à lui seul caractériser l'hétérogénéité tumorale.

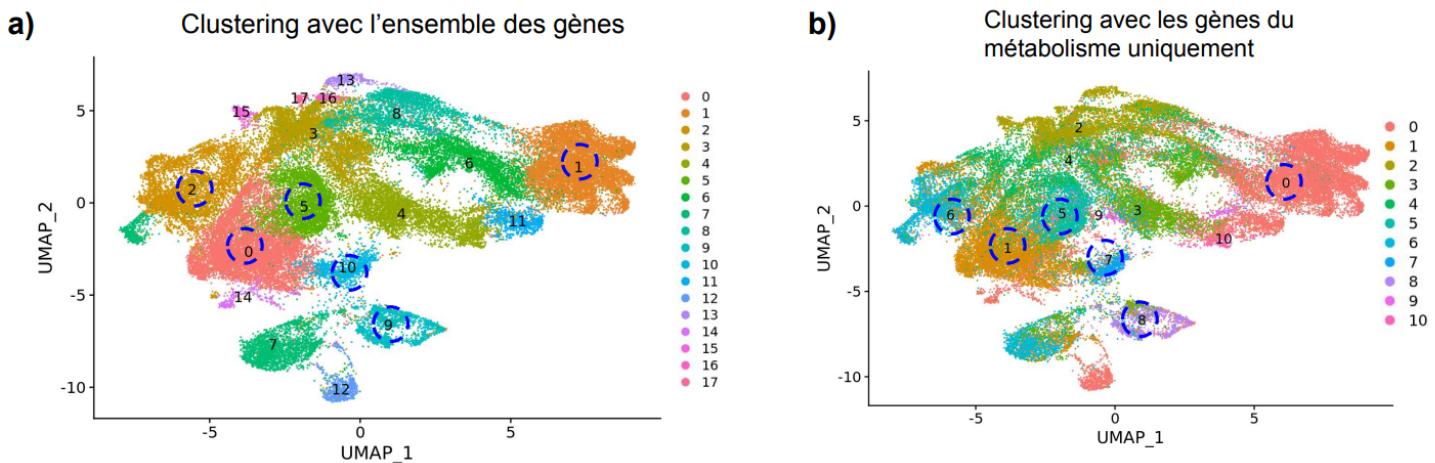


Figure 8: Sur le UMAP obtenu avec l'étude de tous les gènes, la coloration correspond a) aux clusters obtenus lorsque les cellules sont étudiés avec tous leurs gènes, b) aux clusters obtenus lorsque les cellules ne sont étudiés qu'avec les gènes participant au métabolisme. Les cercles bleu entourent les clusters dont la composition est similaire selon ces deux études.

Sur la figure 8, le UMAP de l'étude globale est représenté. Il est coloré soit par les clusters obtenus avec tous les gènes (Fig. 8-a), soit coloré avec les clusters qui ont été obtenus uniquement basé sur le métabolisme (Fig. 8-b). Ainsi, il est possible de visualiser comment les cellules se répartissent selon ces deux analyses.

Des sous-populations sont perdues lorsqu'on les étudie sous le prisme du métabolisme. Les clusters 13, 15, 16 et 17 ne se répartissent plus qu'en 1 cluster, le 2 (Annexe 3). Les clusters 1, 11, 12 et 14 se combinent pour devenir le cluster 0. Ces populations présentent donc un métabolisme similaire bien qu'elles aient des caractéristiques assez différentes pour être séparées dans l'étude globale.

Certains groupes de cellules gardent une composition très similaire selon les 2 analyses, comme les clusters 0, 1, 2, 5, 9 et 10 dont les cellules se retrouvent dans les clusters 1, 0, 6, 5, 7 et 8 respectivement. La figure annexe 4. met en évidence la différence d'expression du métabolisme, plus élevé que tous les autres gènes, ce qui a donc pu jouer une part non négligeable dans le clustering de l'étude globale. Cela expliquerait qu'il est possible de retrouver des clusters de l'étude globale en focalisant sur le métabolisme. Ainsi même avec l'information de seulement 10% des gènes exprimés dans les cellules, il est toujours possible de conserver une partie de l'information après cet immense filtrage.

Une séparation qui forme de nouveaux gradients dans la population est observée, avec les clusters 4 et 11 par exemple, qui se répartissent en majorité dans les clusters 3 et 10 selon leur métabolisme. La séparation des deux clusters n'est pas la même selon l'étude globale avec tous les gènes et celle focalisée sur le métabolisme.

Le gliome présente donc une hétérogénéité métabolique qui peut en partie expliquer l'hétérogénéité globale du gliome mais qui ajoutent une nouvelle information avec de nouvelles sous-populations définies. Ces différents *clustering* mettent en avant les différences entre hétérogénéité fonctionnelle et hétérogénéité fonctionnelle métabolique : des cellules de même nature pourraient être regroupées sur l'étude global mais présenter des états métaboliques différents et se retrouver séparées basées sur le métabolisme.

### C. Sélection de variables pour le développement d'un panel métabolique

A présent que les cellules ont pu être caractérisées de manière générale et métabolique, seuls 500 des 2814 gènes métaboliques pourront être utilisés pour une prochaine analyse afin de caractériser l'hétérogénéité spatiale du gliome via la technologie Xenium - 10X Genomics. Pour choisir les gènes les plus représentatifs des 11 clusters métaboliques du gliome, une liste de gènes signatures par clusters est obtenue, la méthode est détaillée en E-h.

<b>Cluster</b>	0	1	2	3	4	5	6	7	8	9	10	<b>Total</b>
<b>Gènes signatures</b>	105	161	303	102	159	174	134	91	360	574	495	2658

Table 3: Nombre de gènes signatures de chaque cluster métabolique.

Le tableau 3 résume le nombre de gènes différentiellement exprimé pour un cluster comparé aux autres. Basé sur cette méthode uniquement, il faudrait 2658 gènes pour caractériser les clusters métaboliques; c'est bien trop comparé au panel de 500 qu'il faut développer.

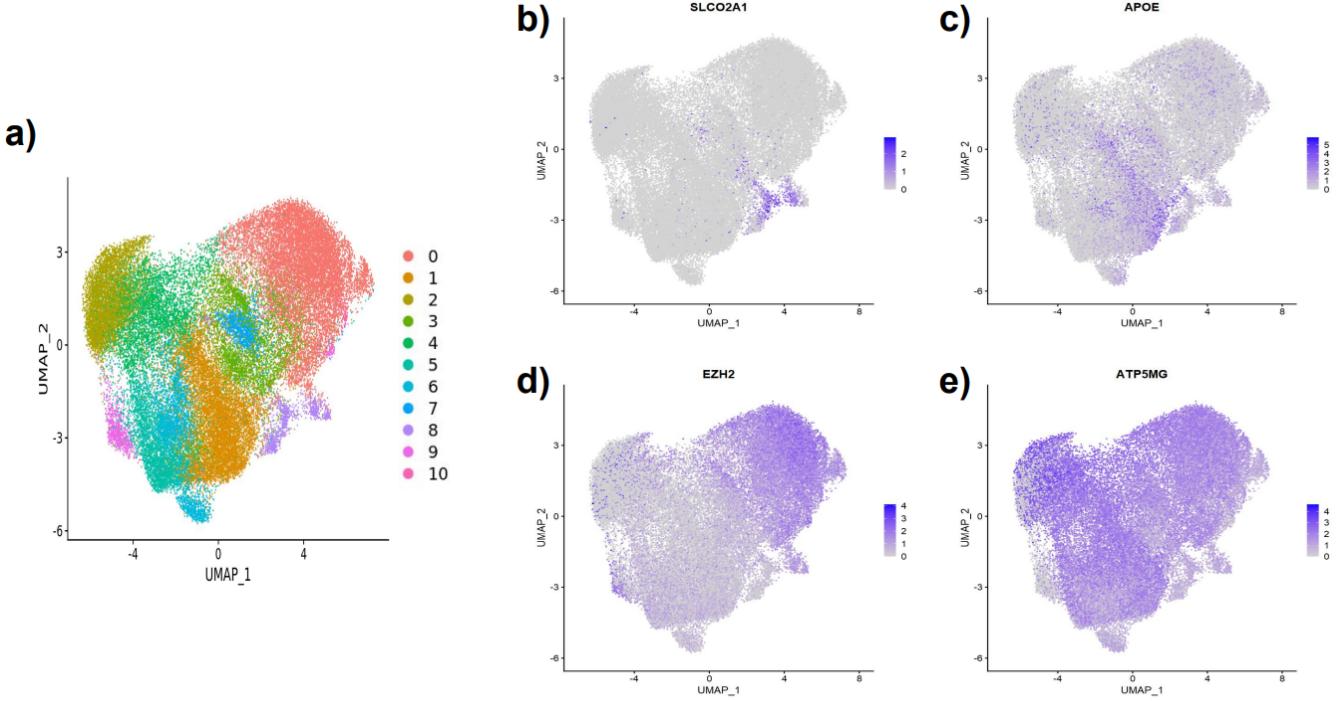


Figure 9: Représentation UMAP obtenus après étude du gliome en focalisant sur le métabolisme avec des expressions de gènes selon un gradient violet. En a) UMAP de référence avec les clusters représentés, b) SLCO2A1 (gène signature du cluster 8), c) APOE (gène signature du cluster 1), d) EZH2 (gène signature du cluster 0) et e) ATP5MG (gène signature du cluster 4).

Si l'on prend pour chaque cluster le gène le plus significatif de ce dernier, il est possible de trouver des gènes qui à eux seuls permettent de définir un cluster métabolique. Sur la figure 9, l'expression de 4 gènes est représentée selon un gradient. Les gènes EZH2, APOE et SLC02A1 sont surexprimés pour les clusters 0, 1 et 8 respectivement mettant en évidence leur spécificité face au cluster auquel ils sont associés. En revanche, beaucoup de ces gènes signatures ne peuvent définir un cluster qu'en combinaison avec d'autres. Par exemple, le gène ATP5MG est le gène signature le plus significatif du cluster 4 et pourtant il n'est pas exprimé de manière spécifique.

Il faut donc plus de gènes pour redéfinir les clusters métaboliques avec des combinaisons significatives.

Afin de réduire la liste des 2658 gènes, les gènes signatures sont triés par cluster selon le Log2FC et 50 gènes les plus différemment exprimés par cluster sont retenus. Ensuite, de ces 550 gènes, les gènes qui sont considérés importants pour plus d'un cluster sont retirés, réduisant le nombre de gènes à 248 (Table 4).

Cluster	0	1	2	3	4	5	6	7	8	9	10	Total
Gènes signatures	15	26	26	10	22	15	29	37	27	21	20	248

Table 4: Nombre de gènes signatures de chaque cluster métabolique.

Pour tester ce premier panel de gènes obtenus, les cellules ne vont pas être annotées selon leur type cellulaire mais selon le cluster pour lequel elles sont reconnues grâce à l'algorithme semi-supervisé SCINA [Méthode E-j].

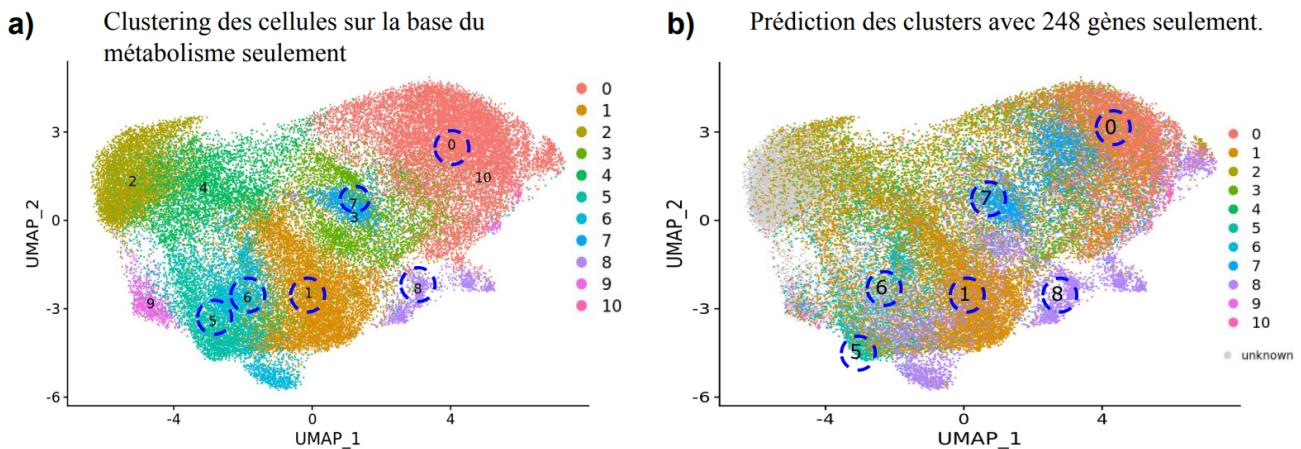


Figure 10: Représentation du clustering métabolique sur le UMAP: a) clusters définis à l'aide des 2814 gènes du métabolisme, b) prédictions de ces clusters avec seulement 248 gènes.

Le résultat de l'annotation est représenté dans la figure 10. Plusieurs clusters sont retrouvés avec une résolution plus ou moins précise.

Le cluster 7 et 8 sont correctement annoté mais ils englobent trop de cellules. A l'inverse, les clusters 5 et 6 ne sont pas totalement reconnus. On devine les clusters 2 et 3 dont les zones commencent à se définir mais c'est encore bien trop léger. Enfin, le cluster 1 avec 26 gènes est le mieux délimité suivi du cluster 0 avec seulement 15 gènes (bien que ce ne soit pas totalement suffisant puisque une partie des cellules sont associées à d'autres clusters).

Le panel de 500 gènes métaboliques n'est pas encore totalement défini, il faut encore affiner les méthodes de sélection afin de capturer au maximum l'hétérogénéité métabolique retrouvées dans le gliome.

# Discussion

## A. Choix de développement

Seurat est une librairie qui est très bien documentée avec plusieurs vignettes explicatives dédiées aux différents types d'analyses qu'il est possible d'effectuer sur les données de type scRNA-seq. Malgré tout, la personne en charge de l'analyse est amenée à faire de nombreux choix qui dépendent de la qualité des données dont elle dispose, du design expérimental, et de la question biologique soulevée..

### a. Normalisation

La normalisation est une étape cruciale pour analyser les données de scRNaseq mais à l'heure actuelle, aucune méthode ne s'impose comme méthodologie standard, en particulier lorsqu'il s'agit d'intégrer des jeux de données multiples, présentant un effet *batch* significatif.

Dans le cadre de mon projet de stage, j'ai opté pour une LogNormalisation suivi d'une correction Harmony pour retirer les effets *batch* liés aux lignées et conditions. En effet, sur la base des nombreux tests et comparaisons que j'ai effectués, cette combinaison est apparue comme un bon compromis entre réalité biologique et précision mathématique, permettant de maintenir à fois les singularités liés aux patients dont les tumeurs ont été extraits, mais également d'identifier des populations cellulaires communes à l'ensemble des tumeurs

### b. Le clustering

Selon la résolution choisie, il est possible de définir plus ou moins de clusters. Bien qu'il existe certaines méthodologies qui permettent d'objectiver un minimum la définition de ce paramètre, il revient néanmoins à la personne en charge de l'analyse des données de le fixer. Il est donc indispensable de définir une stratégie la moins subjective possible.

Dans le cadre de mes analyses, la résolution pour l'étude générale - sur le transcriptome complet - a été choisie de telle sorte que les cellules qui ont été regroupées dans un même cluster à une résolution donnée, ne se divisent plus en son sein lorsque l'on incrémentale la résolution.

Pour l'étude qui ne se focalise que sur les gènes liés aux métabolismes, même avec les plus petites résolutions, beaucoup de clusters ne restaient pas stables et les cellules passaient d'un cluster à un autre. Pour choisir le paramètre de résolution, j'ai tenté de me rapprocher du nombre de clusters trouvés dans l'étude avec tous les gènes, tout en permettant de garder une traçabilité du devenir des cellules entre les deux analyses lisible et interprétable. Cette approche empirique pourrait être améliorée en implémentant des méthodologies telles que SINCERA, RaceID ou encore SHARP [22].

Ces *clustering* sont censés représenter des regroupements de cellules aux caractéristiques similaires ou proches du point de vue de leur état transcriptionnel, et de fait de par leur nature, leur fonction, leur état moléculaire et métabolique. La prise en compte de données issues de l'annotation fonctionnelle des gènes dont le niveau d'expression est caractéristique de chaque cluster pourrait ainsi aider à la définition du nombre de clusters, comme proposé via la méthode ASURAT [23].

## B. Caractérisation du gliome

Les marqueurs utilisés pour reconnaître les populations astrocytes et oligodendrocytes ne sont pas exhaustifs. Ils ne sont peut-être pas assez précis pour reconnaître ces deux populations, et cela pourrait être la raison pour laquelle les cellules ne sont pas assez bien définies pour les oligodendrogiomes BT2 et BT54. Une analyse en anatomo-cytopathologie permettrait de confirmer ou infirmer cette absence d'oligodendrocytes qu'on remarque avec ces marqueurs. Si ces lignées n'ont réellement que très peu d'oligodendrocytes au détriment d'astrocytes, cela pourrait signifier que les oligodendrocytes étaient entrés en sénescence, ne laissant qu'une majorité d'astrocytes dans la lignée ou alors, comme une grande population de cellules en phase G0/G1 est constaté surtout pour la lignée BT2, une grande plasticité cellulaire pourrait être à l'origine d'un changement d'identité cellulaire, perturbant l'identification des cellules.

La lignée LGG349, un grade IV, se démarque des autres lignées, autant dans l'analyse globale que lorsqu'on se concentre sur le métabolisme. Son stade avancé pourrait en être la raison: les cellules de glioblastomes présentent une grande hétérogénéité en raison de la nature agressive et le potentiel invasif du grade IV. Pourtant la lignée LGG85, grade IV également, ne se différencie pas spécialement des autres. Cela pourrait s'expliquer par un état trop avancé de LGG85 où la plupart des cellules seraient entrées en apoptose, résultant en leur filtrage lors de la pré-analyse des données. Ainsi, seules les cellules les plus saines seraient sélectionnées et elles seraient donc confondues avec les grades moins avancés (II et III), non représentatif de l'état réel de la lignée.

A l'inverse, il est possible que ce soit LGG349 qui soit réellement différente de toutes les lignées, y compris même celle de grade identique. Dans ce cas, il faudrait une analyse plus poussée pour déterminer les types cellulaires qui la composent et le degré de différenciation de ces cellules.

Il ne faut pas négliger le fait qu'il puisse simplement s'agir d'un problème technique lors de la préparation de l'échantillon.

BT1, un oligodendrogiome de grade III, possède également un métabolisme distinct, orienté sur la synthèse des lipides insaturés et le métabolisme des monocarbones, qui pourrait être spécifiques d'une population d'oligodendrocytes. BT1 est la seule lignée qui a été cultivée *in vitro* sous forme de sphère ce qui rend l'accès aux nutriments et à l'oxygène inégale entre les cellules. Cette condition de culture aurait pu ajouter un stress aux cellules permettant l'identification cette population d'oligodendrocytes qui se regroupent entre eux avec ce métabolisme qui pourrait leur être spécifique.

Avec les analyses pan-transcriptomiques et metabo-centrées, des cellules de différentes lignées et de différents grades se retrouvent regroupées ensemble. Cela met en évidence des groupes cellulaires communs à tous les gliomes indépendamment de leur stade de développement.

Il faut encore poursuivre les analyses pour mieux définir ces résultats et les combiner à d'autres analyses (protéiques, ...). Cela pourrait aider aux développements de nouveaux traitements contre le gliome; en traitant de manière non-spécifique ces cellules retrouvées dans tous les grades et patients et le ré-adapter en fonction des groupes cellulaires plus grade-spécifiques comme les populations du grade IV de LGG349 qui pourrait être trouvés.

## C. Suite envisageable

### a. Sélection de gènes

Les données ont été séparées en 11 clusters sur la base de l'expression des gènes du métabolisme uniquement. Des clusters ont été définis et caractérisés.

A présent, il faut réduire la liste des gènes permettant de les retrouver à une liste de 500 seulement. Pour cela, avec Seurat, il a été possible de définir les gènes caractéristiques de chaque cluster par une approche d'analyse différentielle, il serait intéressant de comparer les résultats obtenus avec cette approche avec ceux que l'on obtiendrait avec d'autres stratégies. Des méthodes de sélection de variables, comme la sPLSDA (Sparse Partial Least Square Discriminant Analysis) - une méthode basée sur l'apprentissage machine qui permet de sélectionner le nombre minimal de variable indispensable à la discrimination de groupe d'individus - peuvent en effet être utilisées afin de ne pas être limitées et biaisées par le nombre de gènes signatures qu'il faut récupérer par cluster. Ils ne nécessitent pas tous le même nombre de gènes signatures pour être retrouvés.

Les 16 échantillons de *bulk* RNA-seq ont déjà été pré-traité et combinés en une seule matrice de comptage. Une fois qu'une liste réduite de gènes signatures sera défini, le *bulk* RNA-seq sera utile pour se rendre compte du niveau d'expression des gènes dans l'échantillon, permettant d'éliminer tous les gènes qui sont trop exprimés selon une limite imposé par les machines de transcriptomique spatiale pour que la lecture de la fluorescence ne soit pas trop perturbées au risque d'avoir une sursaturation des fluorochromes et leur quantification serait mal effectuées. De plus, les gènes trop peu exprimés sont également à éliminer de l'analyse car leur signaux pourrait être confondu avec du bruit de fond. Cela réduira encore cette liste de quelques gènes pour atteindre les 500 maximum en écartant les plus exprimés dans le *bulk* RNA-seq ainsi que ceux qui sont le moins exprimés.

### b. DEXOM

L'équipe de l'INRAE de Toulouse a développé une méthode de calcul DEXOM qui est utilisée pour prédire l'activité des voies métaboliques à partir de données d'expression [24].

Après avoir sélectionné les 500 gènes, il serait intéressant d'utiliser cet algorithme afin d'évaluer s'ils sont suffisants pour reconstruire des réseaux métaboliques et évaluer leur activité de manière suffisamment précise, en comparaison de ce que l'on obtiendrait à partir de données pan-transcriptome.

Cela permettra d'affiner cette liste de 500 gènes et de sélectionner des gènes qui avait été écartés au détriment d'autres pour que ce panel soit le plus informatif possible non seulement pour retrouver les clusters métaboliques mais également déterminer les voies qui sont activés ou non.

### c. 10X

Après avoir étudié les profils d'expression d'ARN, il faudrait vérifier leur expression protéique et comparer les différences qu'on observe car un gène fortement exprimé ne signifie pas qu'il est fortement traduit et que sa protéine associée sera retrouvée en grande quantité.

Si les moyens techniques et financiers du laboratoire et la disponibilité des matériels biologiques le permettent, ajouter cette nouvelle couche d'information sur la compréhension du gliome serait un axe de travail intéressant.

Par exemple, la technologie de cytométrie de masse spatiale Cytof/Hyperion présente à l'IRCM, permettrait d'ajouter de nouvelles informations sur l'hétérogénéité du gliome, en cartographiant le microenvironnement cellulaire à l'échelle protéique cette fois-ci. Sur une seule coupe histologique, il est possible de détecter jusqu'à 40 anticorps couplés à des métaux, dont le marquage est révélé par spectrométrie de masse; une image de l'échantillon sera ensuite reconstituée avec les marquages associés.

## D. Biais possibles

### a. La normalisation

Différentes méthodes de normalisation ont été testées et j'ai dû en choisir une qui reflètent au mieux les données biologiques auxquelles on s'attendait. Malgré tout, ce n'est pas parce-qu'elle semblait correcte, que toutes les variations techniques ont été retirées. Il pourrait donc y avoir encore des biais qui n'ont pas été traités.

### b. Les échantillons

Les procédures selon lesquelles les lignées ont été traitées peuvent être source de biais. Elles sont en effet étudiées *ex vivo*, sous deux conditions (prolifération et différenciation) qui s'éloignent probablement de la réalité physiologique *in situ*.

Les informations qui en sont tirées pourraient être très différentes de ce qu'on retrouverait chez après analyse immédiate des biopsies isolées de patients.

De multiples biais techniques sont possibles dû à la préparation des échantillons et aux méthodes de séquençage. De plus, les échantillons n'ont pas tous été séquencés le même jour, ce qui ajoute encore un biais technique.

### c. La méthode d'analyse

Au cours de mes analyses, j'ai consciencieusement utilisé le package Seurat, après avoir effectué une recherche approfondie sur les possibilités qu'il offrait et sur ses limitations. Chaque décision, que ce soit pour le filtrage, la normalisation ou le *clustering*, a été prise après une réflexion approfondie et une consultation étendue de la littérature pertinente. En adaptant ces fonctions aux données scRNA-seq spécifiques à mon étude, je reconnaissais que chaque choix apporte son propre ensemble d'avantages et de limitations. La nature dynamique de la bio-informatique appliquée aux données *cellule unique* implique qu'aujourd'hui encore, aucune méthode ne fait référence.

## Bilan

Ce stage de fin d'étude que j'ai pu effectué à l'IRCM encadré par le Dr Pierre-François ROUX a été une expérience dont j'ai pu apprendre énormément et ceux à différents niveaux.

Le sujet de mon stage était de caractériser le gliome via des analyses de *single cell* RNA-seq. Or, pendant ce master AMI2B, nous avons pu apprendre à analyser des données *bulk* RNA-seq; il s'agissait donc d'une nouvelle technique à laquelle il me fallait me familiariser. Beaucoup de documentations ont été faites afin d'être assez compétentes pour faire ces analyses. J'ai bien sûr eu à utiliser de nombreuses connaissances acquises lors du master notamment sur le filtrage des données, l'utilisation et la création de pipeline et de container, le choix des méthodes statistiques par rapport aux résultats souhaités, et bien d'autres compétences encore.

En collaboration avec plusieurs équipes pour ce projet, il m'a fallu présenter mes résultats tant à des bio-informaticiens, avec qui il était possible de détailler les algorithmes et avoir un échange critique sur la manière de procéder aux analyses, qu'à des biologistes où à l'inverse, les résultats devaient être plus visuels et les échanges concernaient la signification biologique que reflétait ces données.

De ces échanges avec des personnes de divers milieux d'expertise, j'ai pu accroître mes connaissances sur des domaines variés et porter un regard plus critique sur mes analyses.

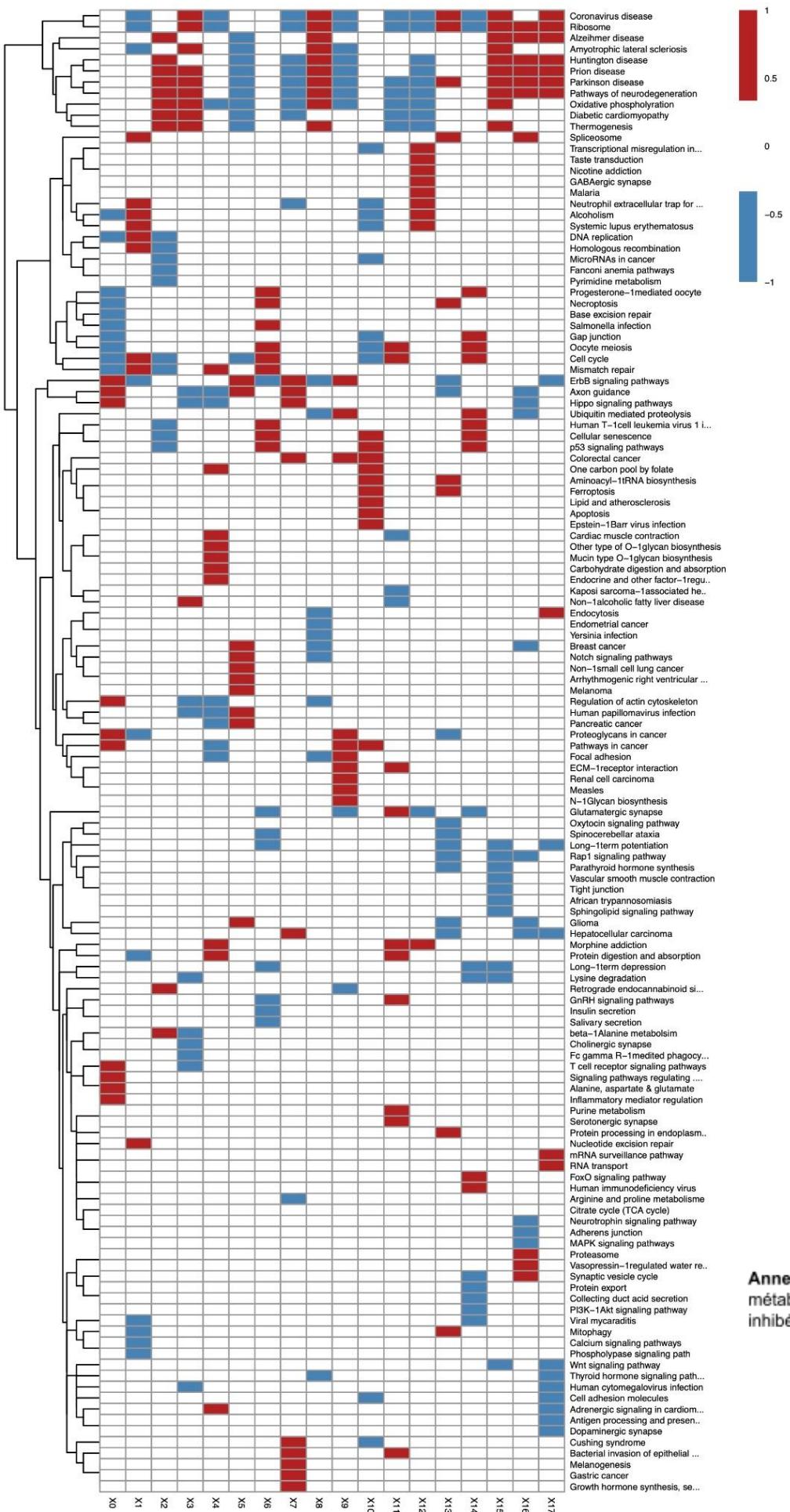
Aussi, c'était une réelle chance d'être entourée de biologistes qui pouvaient m'aiguiller sur la direction que devait prendre mes analyses, selon ce qu'ils pouvaient interpréter de la signification des résultats, chose que je ne pouvais faire seul tant le sujet était précis.

Être entouré d'une majorité de biologistes plutôt que d'informaticiens m'a forcé à prendre confiance en moi et à être capable de justifier tous mes choix puisque, à certaines réunions, personnes n'avait de compétence bio-informatique qui leur permettaient de critiquer mes choix de développement. Je ne pouvais donc pas me permettre d'oublier une justification qui les aurait erronés sur la suite de leur analyse.

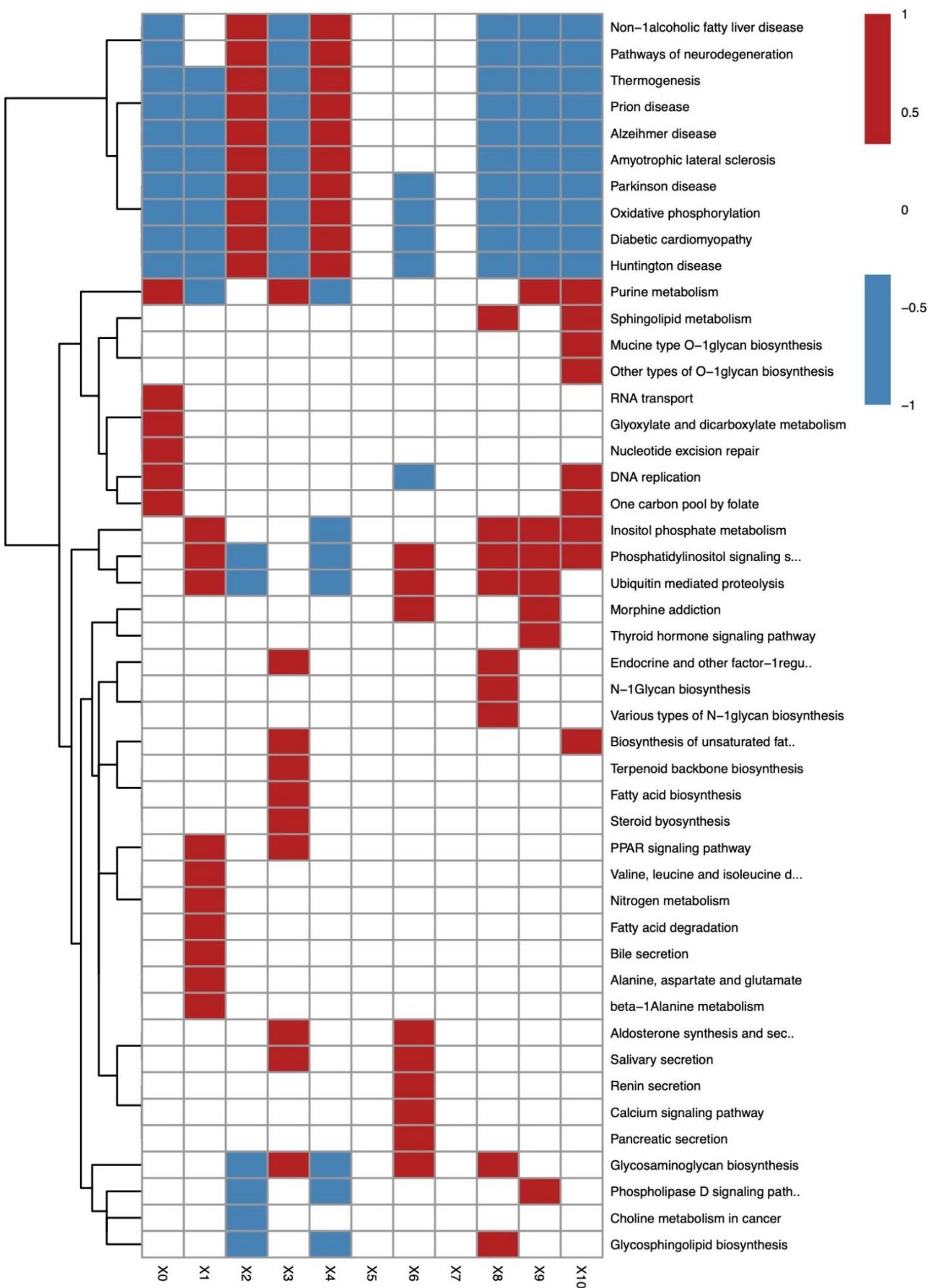
Travailler dans cet environnement où tous les corps de métiers entre la biologie, le médical et l'informatique se rencontraient était une réelle opportunité qui m'était offerte pour apprendre et m'améliorer.

Enfin, j'étais persuadée après mon stage de M1 (développement d'une application web) que finalement je serais capable d'apprécier un métier qui ne requiert que peu voir pas de liens avec la biologie. Ce stage m'a fait changer d'avis. Être constamment entre ces deux disciplines, biologie et informatique, m'a énormément stimulé intellectuellement. Avoir la satisfaction de mettre en relief les résultats bio-informatiques et la réalité biologiques et discuter directement avec des spécialistes des domaines concernant les données me poussent à m'orienter plutôt vers la recherche et des analyses de données omiques.

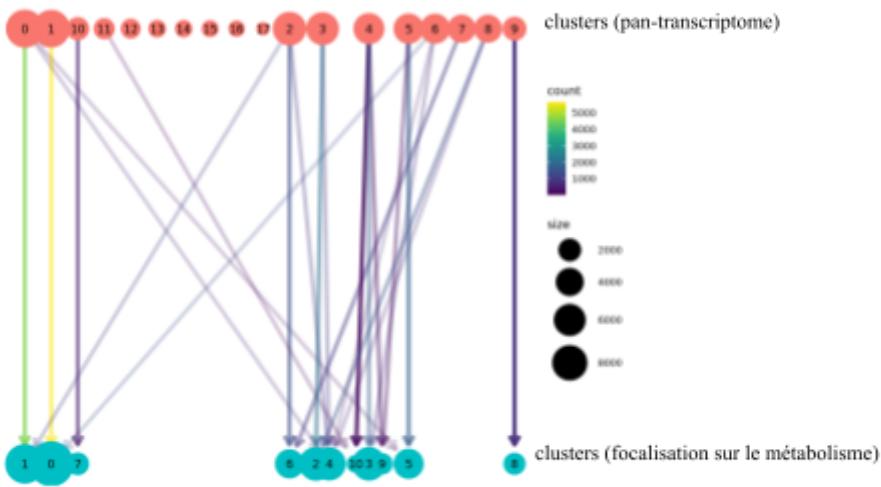
## Annexes



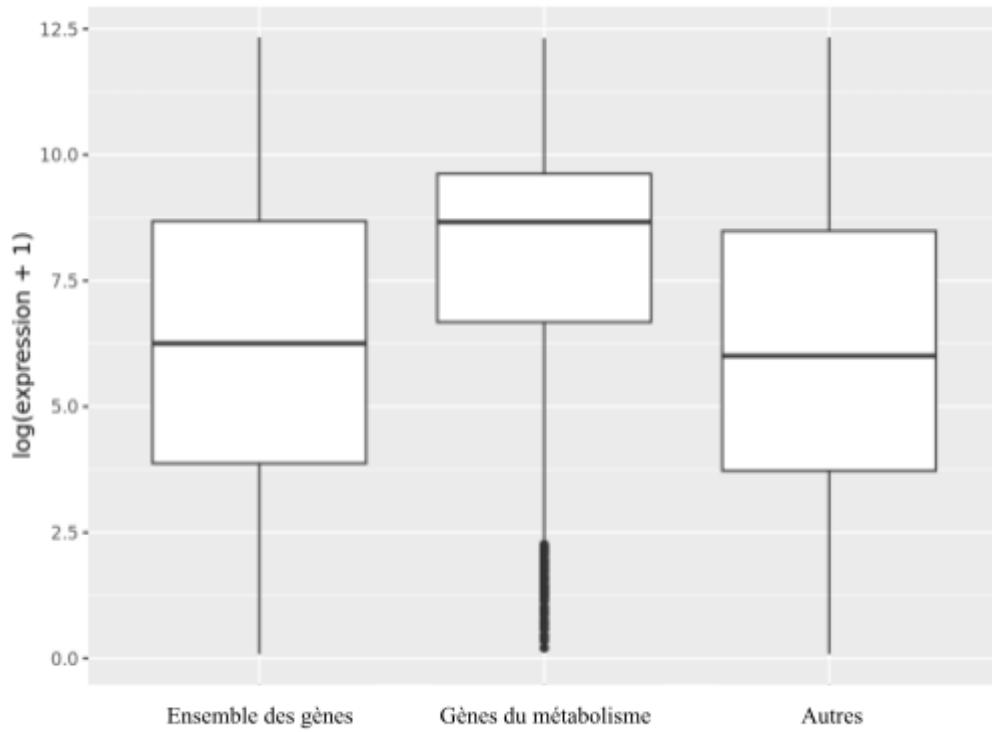
**Annexe 1 :** représentation des voies métaboliques activées (rouge) ou inhibées (bleu) par cluster



Annexe 2 : représentation des voies métaboliques activées (rouge) ou inhibées (bleu) par cluster pour l'étude sur le métabolisme



**Annexe 3 :** représentation à l'aide de clustree du devenir des cellules entre les clusters formés à partir de tous les gènes, et les clusters formés à partir des gènes du métabolisme.



**Annexe 4:** Boxplots des niveaux d'expression de l'ensemble des gènes, des gènes du métabolisme et des gènes ne faisant pas partie du métabolisme.

## Bibliographie

1. *Recherche*, Oncogenèse moléculaire: L. Le Cam, IRCM. <https://www ircm.fr/index.php?pagindx=243>
2. Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W., & Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*, 114(5), 547–547. <https://doi.org/10.1007/s00401-007-0278-6>
3. Faubert, B. (2020). Metabolic reprogramming and cancer progression. *Science*, 368(6487). <https://doi.org/10.1126/science.aaw5473>
4. Yan, H., Parsons, D.W., Jin, G., McLendon, R., Rasheed, B.A., Yuan, W., Kos, I., BatinicHaberle, I., Jones, S., Riggins, G.J., et al. (2009). IDH1 and IDH2 mutations in gliomas. *N Engl J Med* 360, 765-773. <https://doi.org/10.1056/nejmoa0808710>
5. Santos, C. R., & Schulze, A. (2012). Lipid metabolism in cancer. *The FEBS Journal*, 279(15), 2610–2623. <https://doi.org/10.1111/j.1742-4658.2012.08644.x>
6. *Xenium platform page*. (2023). 10x Genomics. <https://www.10xgenomics.com/platforms/xenium>
7. Jooma, R., Waqas, M., & Khan, I. (2019). Diffuse low-grade glioma - Changing concepts in diagnosis and management: A review. *Asian Journal of Neurosurgery*, 14(2), 356–363. [https://doi.org/10.4103/ajns.AJNS\\_24\\_18](https://doi.org/10.4103/ajns.AJNS_24_18)
8. Augustus K. (2020), Characterization of cellular heterogeneity in Diffuse Low Grade Glioma. *Human health and pathology* [Université Montpellier]. Retreived August 16, 2023, from <https://theses.hal.science/tel-03156473/>
9. Augustus, M., Pineau, D., Aimond, F., Azar, S., Lecca, D., Scamps, F., Muxel, S., Darlix, A., Ritchie, W., Gozé, C., Rigau, V., Duffau, H., & Hugnot, J.-P. (2021). Identification of CRYAB+ KCNN3+ SOX9+ astrocyte-like and EGFR+ PDGFRA+ OLIG1+ oligodendrocyte-like tumoral cells in diffuse idh1-mutant gliomas and implication of NOTCH1 signalling in their genesis. *Cancers*, 13(9), 2107
10. *Home*. (2016, May 4). Genotoul-Bioinfo. <https://bioinfo.genotoul.fr/>
11. Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., & Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research*, 25(12), 1860–1872. <https://doi.org/10.1101/gr.192237.115>
12. Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1). <https://doi.org/10.1186/s13059-019-1874-1>
13. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12), 1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>

14. McInnes, L., Healy, J., & Melville, J. (2018, February 9). *UMAP: Uniform manifold approximation and projection for dimension reduction*. arXiv.Org. <https://arxiv.org/abs/1802.03426>
15. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücke, M. D., Strobl, D. C., Henao, J., Curion, F., Schiller, H. B., & Theis, F. J. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8), 550–572. <https://doi.org/10.1038/s41576-023-00586-w>
16. Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, 15(6). <https://doi.org/10.15252/msb.20188746>
17. Brüning, Tombor, Schulz, Dimmeler, & John. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *GigaScience*, 11. <https://doi.org/10.1093/gigascience/giac001>
18. *Seurat example*. (2022). Babraham Bioinformatics. [https://www.bioinformatics.babraham.ac.uk/training/10XRNASeq/seurat\\_workflow.html](https://www.bioinformatics.babraham.ac.uk/training/10XRNASeq/seurat_workflow.html)
19. Choudhary, S., & Satija, R. (2021). *Comparison and evaluation of statistical error models for scRNA-seq*. Cold Spring Harbor Laboratory. <http://dx.doi.org/10.1101/2021.07.07.451498>
20. Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1), 1–32. <https://doi.org/10.1186/s13059-019-1850-9>
21. Themes, U. (2017, April 14). Models. Veteran Key. <https://veteriankey.com/models-4/>
22. Yu, L., Cao, Y., Yang, J. Y. H., & Yang, P. (2022a). Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biology*, 23(1), 1–21. <https://doi.org/10.1186/s13059-022-02622-0>
23. Iida, Kondo, Wibisana, Inoue, & Okada. (2022). ASURAT: Functional annotation-driven unsupervised clustering of single-cell transcriptomes. *Bioinformatics*, 38(18), 4330–4336. <https://doi.org/10.1093/bioinformatics/btac541>
24. Rodríguez-Mier, P., Poupin, N., de Blasio, C., Cam, L. L., & Jourdan, F. (2021). DEXOM: Diversity-based enumeration of optimal context-specific metabolic networks. *PLOS Computational Biology*, 17(2). <https://doi.org/10.1371/journal.pcbi.1008730>