UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE (ED414)

Laboratoire d'Innovation Thérapeutique (UMR 7200)

THÈSE présentée par :

Noémie ROBIL

soutenue le : **8 octobre 2015**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Bioinformatique et biologie des systèmes

Recherche d'antigènes spécifiques de tumeurs et analyse des cellules souches de glioblastomes

THÈSE dirigée par :
M HAIECH Jacques — Professeur, Université de Strasbourg

RAPPORTEURS :
M DE REYNIES Aurélien — Docteur, Ligue Nationale Contre le Cancer
M MOREAU Marc — Directeur de Recherches, Université Paul Sabatier

---

AUTRES MEMBRES DU JURY :
M HEINRICH Christian — Professeur, Université de Strasbourg

Remerciements

A l'origine, le projet de ce travail de thèse était une collaboration entre la société Transgène, l'Université de Strasbourg par le biais du Pr Jacques Haiech et la Ligue Nationale Contre le Cancer. Je tiens à remercier l'ensemble des personnes ayant participé à la mise en place du projet. En effet, même si cette collaboration s'est terminée un peu brutalement, il n'y aurait pas eu de projet de thèse sans elle. Je tiens à remercier personnellement Jacqueline Godet, présidente de la Ligue pour avoir soutenu mon projet, y compris à la fin de la collaboration avec Transgène.

Je souhaite remercier les membres du jury, Aurélien De Reyniés, Marc Moreau et Christian Heinrich pour avoir consacré du temps à l'examen de mon travail de thèse.

Un grand merci à Jacques Haiech pour m'avoir encadrée pendant ces trois années et quelques mois, merci pour tous les conseils et discussions enrichissantes dont j'ai pu profiter, que ce soit en lien direct avec mon travail ou bien sur des sujets scientifiques plus larges.

De même, merci Fabien pour ton accompagnement au quotidien et ton soutien. Tu es toujours disponible pour répondre à mes questions, ou m'aider lorsque j'ai un problème.

Un grand merci à l'ensemble de personnes qui ont collaboré à ce travail. Je vais commencer par le début et la période de collaboration avec Transgène, merci Ronald Rooke de m'avoir encadrée et fait découvrir ce qu'était l'immunothérapie des cancers. J'ai particulièrement apprécié l'ensemble des *Data-club* avec Benoit Grellier et Philippe Ancien. Ces premiers moments d'échanges entre biologistes, bioinformaticiens et statisticiens furent très enrichissants et passionnant. Merci à Hervé Chneiweiss, Sarah Cianférani et Leslie Muller pour la deuxième partie de ma thèse. Grâce à vous, j'ai découvert le monde de la protéomique et des cellules souches cancereuses. J'ai beaucoup appris au cours de nos réunions, et j'ai particulièrement appréciée de travailler avec vous. Je pense que le plus enrichissant sur l'ensemble de travail vient de toutes ces rencontres et réunions avec vous tous, venant de différents domaines et ne parlant pas toujours exactement le même langage que moi.

Si l'ensemble de cette aventure s'est déroulée dans de si bonnes conditions, c'est aussi grâce à tous les membres de l'équipe CIT et je vous en remercie. Que ce soit au niveau scientifique, pour l'ensemble de vos conseils, réponses à mes questions techniques, et pour toutes ces réunions qui m'ont permis de comprendre de mieux en mieux la génomique des cancers ; ou bien au niveau personnel parce qu'il n'y a rien de plus important qu'une bonne ambiance dans l'équipe pour se motiver à venir au labo tous les matins. Merci à Aurélien pour tes conseils et ta disponibilité quand j'en ai eu besoins. Merci à celles et ceux qui organisent « la vie » de CIT : Laetitia pour les FTWE (très important pour la cohésion, et j'ai découvert plein

de nouveaux restos sympa !) et les séminaires, Sylvie pour les réunions d'équipe et Mira pour la gestion globale de tout le reste ! Merci aussi pour tout ce que vous avez pu m'apporter en dehors du travail, que ce soit la découverte de nouveaux sports : sans toi Sylvie je n'aurais jamais su le bonheur de courir dans une forêt pleine d'épines pour trouver des balises ; les découvertes culinaires, surtout pour ce qui concerne la cuisine méditerranéenne (merci Nabila et Aurélie, j'ai dû prendre quelques kilos grâce à tous ces gâteaux), ou bien les discussions cinéma et expos de Jacqueline. Finalement, merci aux deux nouveaux, Rémi et Yuna d'avoir ramené un peu de sang neuf dans l'équipe. Grâce à vous, on n'est passé d'un certain nombre de discussions sur le thème bébé/enfant/école à aller boire un verre (ou plus) après le boulot, c'est bien aussi !

Je remercie aussi tous mes amis qui m'accompagnent depuis plusieurs années maintenant et qui m'ont permis de décompresser quand j'en avais besoins. Merci à vous, Mél & Julien, Virginie, Alice, Emeric, Marielle, Laetitia, Camille D, Julien, Caroline, Jonathan, Ombeline & Charles, Eric & Diane, Camille L & Karen.

Je souhaite enfin remercier ma famille qui m'a toujours soutenue dans mes choix. Je sais que je vous dois beaucoup, merci pour tout ce que vous avez fait pour moi au quotidien. Merci, papa, maman et Manue et promis j'arrête là les études ! Merci aussi à ma deuxième famille, Caroline, Eloise, Marie, Julie, Galder, Abel, Tessa, Ari, Mathieu, Ony et Victor pour vos encouragements. J'ai évidemment une pensée particulière pour Christiane qui aurait voulu voir l'aboutissement de ce travail. Vous et Daniel m'avez accueilli dans votre famille et je vous en serais toujours reconnaissante. Finalement, merci Adri pour ton amour et ton soutien tout au long de ces années. Elles auraient été bien difficiles et tristes sans toi.

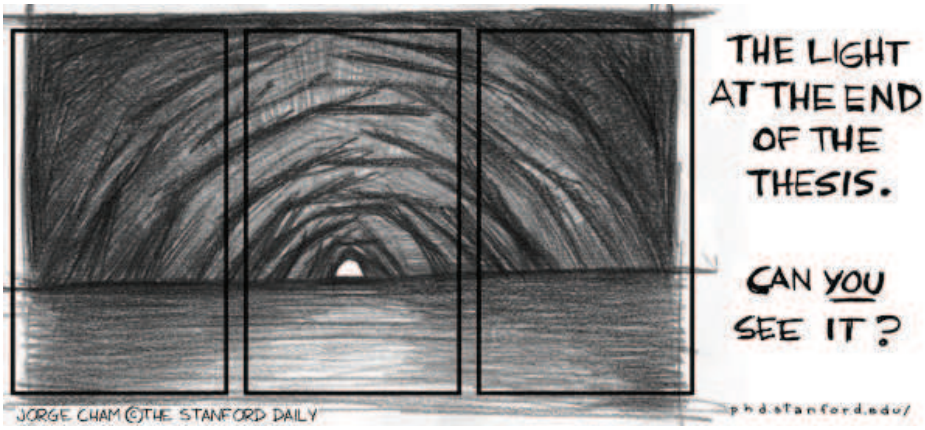

Table des matières

Table des figures

Liste des tableaux

Abréviations

ACP	Analyse en Composantes Principales
ADN	Acide desoxyribonucléique
ARN	Acide ribonucléique
ARNm	ARN messager
ATP	Adénosine Triphosphate
CD	Cluster de différenciation
CRAC	Canal activé par le relargage du Ca2+ de l'ER
CSCs	Cellules Souches Cancéreuses
EGFR	Epidermal Growth Factor Receptor
EMT	Transition Epithélio-Mésenchymateuse
ER	Reticulum Endoplasmique
ESC	Cellules Souches Embryonnaires
FDA	Food and Drug Administration
G-CIMP	Glioma-CpG Island Methylator Phenotype
gCSCs	Cellules Souches Cancéreuses de Glioblastomes
HA	lignées d'Astrocytes Humains
LOH	Perte d'Hétérozygotie
MDR	MultiDrug Resistance
miARN	micro ARN
OMS	Organisation Mondiale de la Santé
PCR	Polymerase Chain Reaction
RCPG	Récepteur Couplé aux Protéines G
RMA	Robust Multi-array Average
RPKM	Read Per Kilobase per Million mapped reads
RPPA	Reverse Phase Protein Array
SHH	Sonic Hedgehog
shRNA	Petit ARN en épingle à cheveu （Small Hairpin RNA）
TCGA	The Cancer Genome Atlas
TM	Transmembranaire
TMZ	Témozolomide

Avant-Propos

Les glioblastomes sont les tumeurs du système nerveux central les plus fréquentes et les plus agressives. Les thérapies actuelles, basées sur la chirurgie, la chimiothérapie et la radiothérapie, sont mises en échec par ce cancer, avec un taux de survie à 5 ans qui n'excède pas 10%. Cet échec pourrait s'expliquer en partie par l'existence de cellules particulières, les cellules souches cancéreuses.

Longtemps considérée comme une hypothèse, l'existence de cellules souches cancéreuses au sein des tumeurs a été mise en évidence dans différents cancers au cours des dernières années, comme les cancers du sein, du système nerveux central, du pancréas, du cou, de la prostate, du colon et de la peau. Ces cellules ont plusieurs propriétés communes aux cellules souches : une capacité d'auto-renouvèlement, la capacité de se différencier et la quiescence[1]. Cette dernière propriété pourrait expliquer certains des mécanismes de résistance aux traitements usuels. Il est donc important de les identifier et de les cibler pour espérer éliminer totalement la tumeur.

Bien que différents marqueurs des cellules souches aient été proposés ces dernières années, aucun d'entre eux ne permet de sélectionner de manière spécifique les cellules souches de glioblastomes.

A la fin des années 1990, un nouveau type de traitement ciblé a émergé, les anticorps monoclonaux. Il s'agit de molécules se liant spécifiquement à un antigène tumoral à la surface des cellules cancéreuses, et qui permettent donc de cibler les cellules à détruire. Depuis le succès du Trastuzumab (Herceptin ou anti-HER2) dans le cancer du sein, le nombre d'anticorps approuvés comme traitements en oncologie par la *Food and Drug Administration* (FDA) est de plus en plus important.

L'identification de marqueurs spécifiques de cellules cancéreuses ou de cellules souches cancéreuses ouvre la voie au ciblage thérapeutique de ces cellules grâce à l'utilisation d'anticorps marqués ou couplés à des molécules cytotoxiques.

L'objectif de ce travail de thèse est de déterminer des biomarqueurs des cellules souches de glioblastomes (gCSCs) permettant de cibler ces cellules et de les caractériser. Pour cela, le travail a été organisé sur deux axes principaux :

[1] Temps durant lequel une cellule arrête de se diviser et sort du cycle cellulaire

- Le développement d'une méthode générique permettant de prédire les antigènes spécifiques de cancer à partir de données d'expression.
- L'analyse des cellules souches de glioblastomes, que ce soit par l'identification de protéines sur-exprimées à la surface des gCSCs ou bien par l'étude des modifications du signal calcium, dérégulé dans de nombreux cancers.

L'ensemble du travail s'est déroulé en collaboration entre le Laboratoire d'Innovation Thérapeutique (UMR 7200) de l'Université de Strasbourg et le programme 'Cartes d'Identité des Tumeurs' (CIT®) de la Ligue Nationale Contre le Cancer.

Le **chapitre 1** de ce manuscrit est divisé en quatre parties. La première partie est un état des lieux des connaissances générales sur le cancer, notamment sur la notion d'hétérogénéité. La deuxième partie s'attache particulièrement à la description des glioblastomes, leur classification et les traitements recommandés. L'état actuel des connaissances sur les cellules souches de glioblastomes, des marqueurs connus aux voies de signalisations dérégulées, ainsi que les caractéristiques leur conférant leur résistance aux traitements classiques y seront également décrits. La troisième partie présente l'évolution des technologies à haut et moyen débit utilisées pour l'analyse du protéome[2] et du transcriptome[3], ainsi que le processus d'analyse de ces données. La dernière partie est une réflexion sur les différentes sources de bruit et d'erreur dans une expérience « omique » (analyse du protéome, transcriptome ou autre « omique ») et sur la manière de les prendre en compte.

Dans **le chapitre 2,** nous nous intéresserons au développement de KANT, une méthode permettant d'identifier des antigènes spécifiques de tumeurs à partir de données de puces d'expression. Les critères biologiques d'une protéine d'intérêt (ou antigène putatif) sont l'accessibilité de l'antigène par un anticorps circulant, et la spécificité d'expression dans une tumeur donnée. Ceci implique une annotation fiable des protéines membranaires basée sur leur topologie, et la comparaison de données transcriptomiques des cancers avec celles de tissus sains. Nous allons donc d'abord prédire les protéines membranaires à partir de leur séquence en utilisant et en améliorant des algorithmes d'apprentissage supervisé existants. Ensuite, nous nous intéresserons à la mise en place d'un algorithme de recherche de sur-expression prenant en compte l'hétérogénéité tumorale. Afin de valider la méthode, deux cancers seront étudiés : le cancer du sein et le lymphome,

[2] Ensemble des protéines exprimées dans une cellule ou un groupe de cellules dans des conditions données et à un moment donné

[3] Ensemble des ARNs transcrits dans une cellule ou un groupe de cellules dans des conditions données et à un moment donné

afin de retrouver des gènes connus comme étant spécifiques de ces cancers, et d'identifier de nouvelles cibles potentielles jusqu'alors inconnues. Le travail de ce chapitre fait l'objet d'un article soumis à *BMC bioinformatics*.

L'objectif du **chapitre 3** est d'identifier et de caractériser de nouveaux biomarqueurs spécifiques des gCSCs. Pour cela, nous allons analyser les données de spectrométrie de masse[4] de plusieurs gCSCs afin d'identifier les protéines sur-exprimées, puis nous validerons les protéines d'intérêt sur un ensemble de données publiques de puces d'expression. La validation transcriptomique de données protéomiques n'est pas habituelle, mais la rareté des gCSCs ne nous permet pas d'accéder à d'autres données de protéomique. L'analyse de données de séquençage ARN et la vérification de l'expression de ces protéines dans le glioblastome nous permet d'aller plus loin dans la caractérisation de nos cibles. Le travail de ce chapitre constitue une partie d'un article en préparation avec le Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO) de l'Institut Pluridisciplinaire Hubert Curien (IPHC) de Strasbourg.

Dans le **chapitre 4**, nous analyserons l'ensemble des gènes impliqués dans le signal calcium et leur dérégulation dans les glioblastomes et gCSCs. Avant d'analyser un ensemble de données publiques conséquent, nous nous intéresserons à la prise en compte du bruit dans les annotations de ces données, ainsi que dans l'analyse des différents échantillons. Finalement, l'utilisation d'une méthode basée sur l'algorithme de sur-expression du chapitre 2 nous permettra d'émettre des hypothèses sur la dérégulation de deux mécanismes impliqués dans la signalisation calcique. Le travail de ce chapitre fait l'objet d'un article sous presse à *l'International Journal of Developmental Biology*.

[4] Technique permettant l'analyse du protéome

Chapitre 1 : Introduction

1.1 Cancer et hétérogénéité tumorale

1.1.1 Généralités sur les cancers

Le dernier rapport de l'Institut National du Cancer (édition 2014) fait état de 355 000 nouveaux cancers par an et 155 000 décès (chiffres de l'année 2012) ; ce qui fait du cancer, en France, la première cause de mortalité chez les hommes et la deuxième chez les femmes. Néanmoins, grâce à l'amélioration des diagnostics et l'optimisation des traitements, la mortalité tous cancers confondus diminue (Figure 1 a et b). Dans le même temps, l'incidence diminue chez les hommes (-1.3% par an entre 2005 et 2012) et ralentit sa progression chez les femmes (+0.2% par an entre 2005 et 2012 au lieu de +1.6% entre 1980 et 2005). Le taux d'incidence correspond au nombre de cas pour 100 000 personnes/an standardisé sur la structure d'âge de la population mondiale, il permet ainsi des comparaisons entre pays mais cache l'augmentation du nombre de cas de nouveaux cancers par an en France, dû à l'augmentation et au vieillissement de la population.

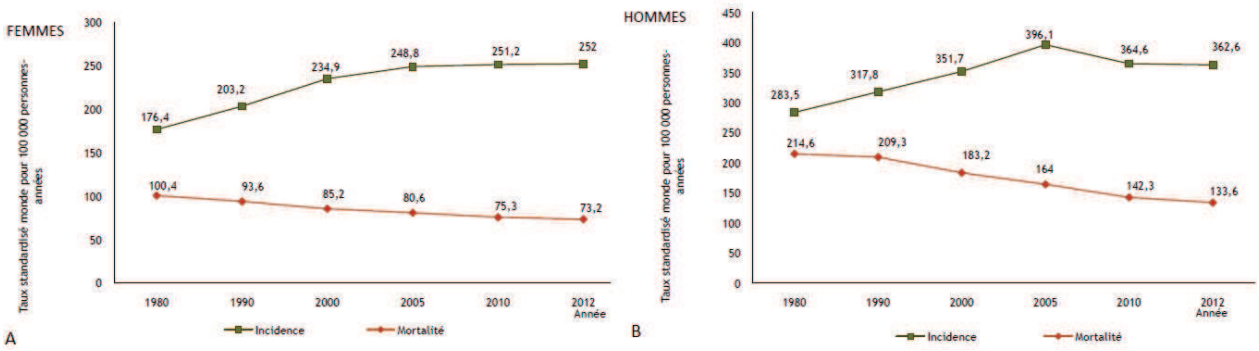

Figure 1 : Evolution (en %) de l'incidence (A) et de la mortalité (B) « tous cancers » en France métropolitaine de 1980 à 2012 selon le sexe

Sources : Binder-Foucard F, 2013. Traitement : INCa 2013

Les cancers les plus fréquents sont ceux de la prostate chez l'homme avec 56 840 nouveaux cas en 2012, et du sein chez la femme avec 48 800 nouveaux cas (Figure 2). La majorité des décès sont imputables au cancer du poumon chez l'homme (21 300 décès) et au cancer du sein chez la femme (11 900 décès) (Figure 3).

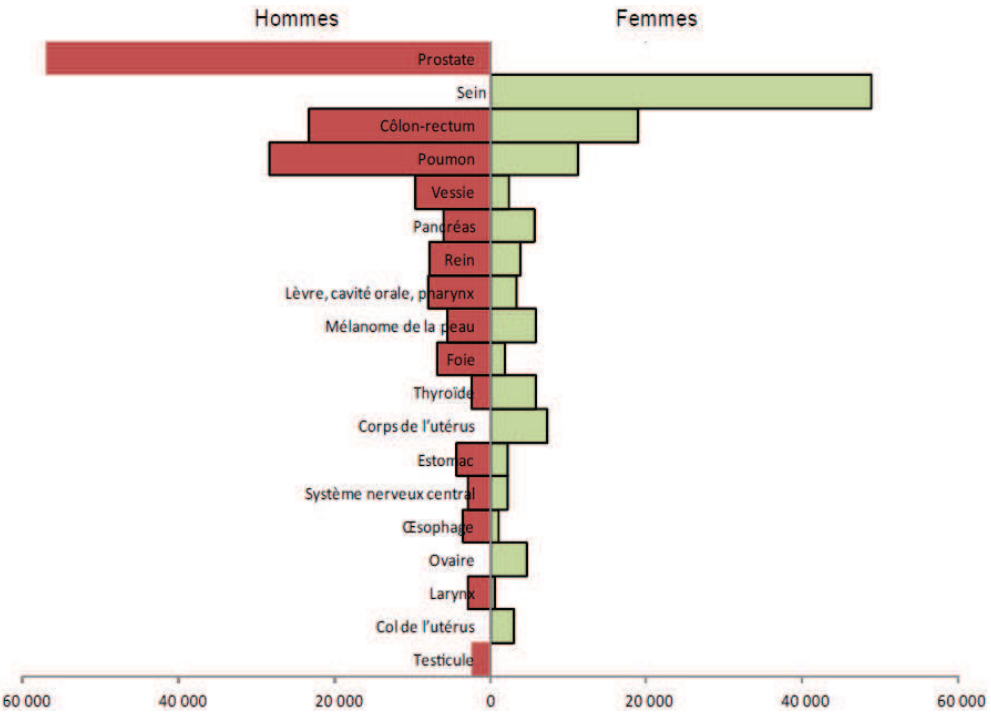

Figure 2 : Classement des cancers par incidence estimée en 2012 en France métropolitaine par localisations selon le sexe

Source : Binder-Foucard F, 2013. Traitement : INCa 2014

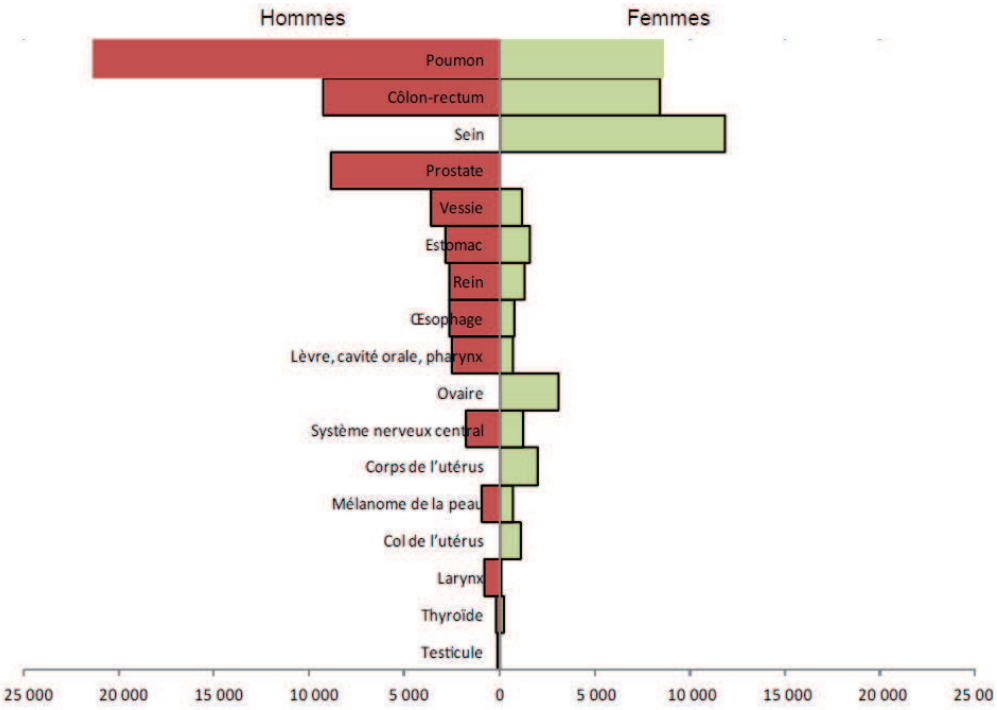

Figure 3 : Classement des cancers par mortalité estimée en 2012 en France métropolitaine par localisations selon le sexe

Source : Binder-Foucard F, 2013. Traitement : INCa 2014

1.1.1. Caractéristiques

Même s'il est souvent associé à des facteurs de risque contemporains, le cancer semble avoir de tout temps existé. Les traces les plus anciennes proviennent du fossile d'un individu néandertalien vieux de 120 000 ans, dans lequel on a détecté la présence d'une tumeur osseuse (Monge et al., 2013). Historiquement, les premières descriptions remontent à 3000 av J.-C., avec la description de 8 cas de cancers du sein, dans le papyrus d'Edwin Smith, mais c'est Hippocrate en 460 av JC qui a nommé la maladie en référence au crabe (carcinos en grec).

Le cancer est un terme général qui regroupe un grand nombre de maladies présentant six caractéristiques fondamentales communes décrites en 2000 par Hanahan et Weinberg (Hanahan and Weinberg, 2000). Ce sont les caractéristiques que doit acquérir une cellule normale afin de devenir cancéreuse et qui résument l'ensemble des modifications à l'origine du cancer (Figure 4) :

- **Autosuffisance en signaux de croissance** : Les cellules normales sont régulées de manière précise par la production et la sécrétion de facteurs de croissance. Les cellules cancéreuses se caractérisent par leur capacité à soutenir une prolifération chronique sans prendre en compte ces facteurs de croissances. Trois stratégies moléculaires sont possibles : l'altération des signaux extracellulaires, l'altération des récepteurs ou l'altération des voies de transduction intracellulaires[5].
- **Evitement des signaux inhibiteurs de la croissance** par la perturbation des voies de signalisation ayant un effet antiprolifératif, comme TGF-ß.
- **Résistance à l'apoptose[6]** : par exemple par la perte du suppresseur de tumeur TP53 ou par sur-expression de facteurs anti-apoptotiques.
- **Capacité réplicative infinie** : la réplication des cellules cancéreuses, contrairement à celle des cellules normales, n'est pas limitée en nombre grâce au maintien de leurs télomères.
- **Induction de l'angiogenèse[7]** : Les cellules tumorales, comme les cellules normales ont besoin de se nourrir, de respirer et d'évacuer leurs déchets. Durant l'embryogenèse, de nouveaux vaisseaux sanguins sont formés afin de remplir cette fonction mais le processus devient quiescent une fois les tissus formés. Dans le cas de cellules cancéreuses, l'angiogenèse est activée pour permettre la croissance des tumeurs.
- **Capacité à migrer, à envahir de nouveaux tissus et à créer des métastases** qui sont à l'origine de 90% des décès.

[5] Mécanisme par lequel une cellule répond à l'information qu'elle reçoit

[6] Mort cellulaire

[7] Processus de croissance de nouveaux vaisseaux sanguins

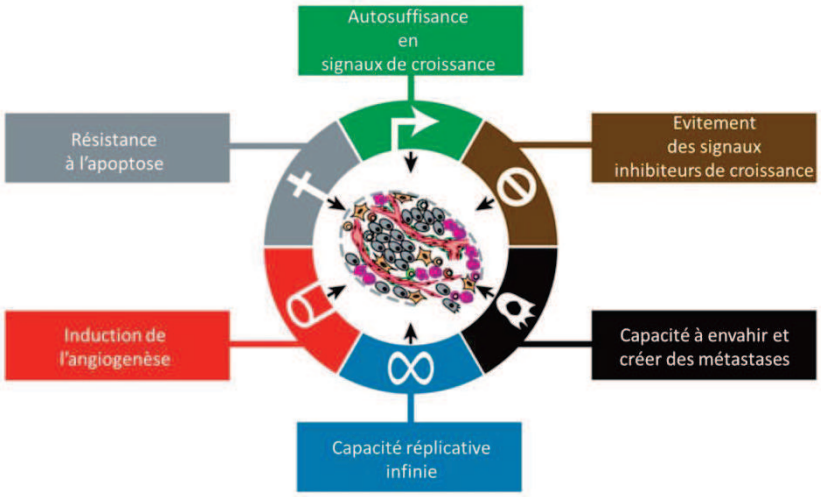

Figure 4 : Capacités acquises par l'ensemble des cellules cancéreuses : "Hallmarks of Cancer"
Source : adaptée de Hanahan and Weinberg (Hanahan and Weinberg, 2000)

En 2011, l'article de Hanahan and Weinberg a été réactualisé (Figure 5) et deux nouvelles caractéristiques dites « émergentes » sont venues s'ajouter aux précédentes (Hanahan and Weinberg, 2011) :

- La dérégulation de la gestion énergétique des cellules tumorales.
- L'échappement des cellules tumorales au système immunitaire.

Deux autres caractéristiques, dites « permissives », ont aussi été ajoutées : l'instabilité génomique et l'inflammation promotrice de la tumeur.

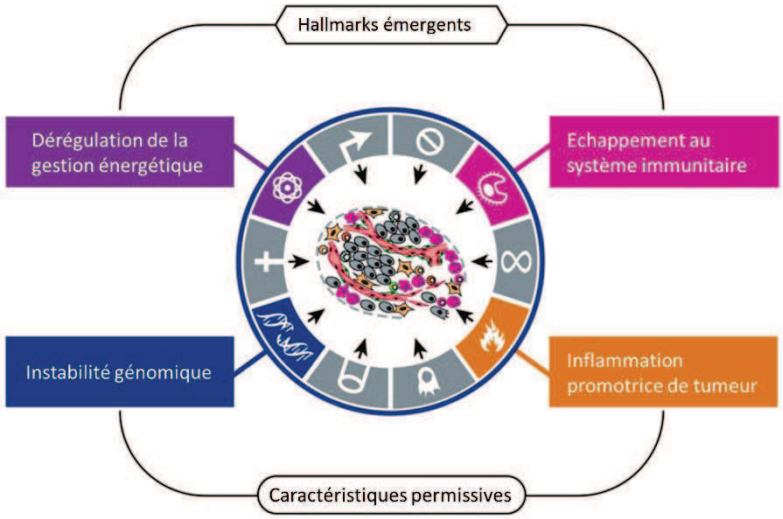

Figure 5 : Hallmarks emergents et caractéristiques permisives
Source : adaptée de Hanahan and Weinberg (Hanahan and Weinberg, 2011)

Tous les cancers présentent les huit hallmarks, ou caractéristiques mais leur ordre d'apparition est variable. L'acquisition de ces caractéristiques se fait en grande partie par des modifications

sur le génome : mutations, altérations chromosomiques (gain ou perte de segments chromosomiques), translocations (passage d'un fragment de chromosome sur une autre région de ce chromosome ou sur un autre chromosome) et altérations épigénétiques (changements dans la fonction des gènes, ayant lieu sans altération de la séquence ADN).

1.1.2. Traitements actuels

Les traitements actuels des tumeurs se divisent en trois grandes classes (Tiwari and Roy, 2012) : chirurgie, radiothérapie et chimiothérapie.

La chirurgie consiste à enlever l'intégralité de la tumeur. L'exérèse doit si possible être suffisamment large pour s'assurer qu'il ne reste pas de cellules tumorales.

La radiothérapie bloque la division cellulaire des cellules cancéreuses par rayonnements à haute énergie. La dose et le fractionnement des séances doivent être adaptés à chaque type de tumeur et de zone afin d'éviter de détruire trop de cellules saines, tout en irradiant le maximum de cellules tumorales (Bernier et al., 2004; Hellevik and Martinez-Zubiaurre, 2014).

Les chimiothérapies se basent sur l'utilisation de molécules chimiques (Chabner and Roberts, 2005; DeVita and Chu, 2008). La plus ancienne fut développée à partir du gaz moutarde utilisé pendant la seconde guerre mondiale. La majorité des substances chimiothérapeutiques fonctionnent par arrêt de la mitose en ciblant les cellules se divisant plus rapidement, elles sont dites cytotoxiques. Les premiers essais cliniques ont été menés en 1943 sur des patients atteints de lymphome et ont montré une brève régression de la maladie. De nombreuses autres molécules ont été développées par la suite. Le principal inconvénient de cette technique vient de la non-spécificité du traitement qui entraine des effets secondaires importants.

Depuis la fin des années 1990, de nouveaux traitements ciblés ont été développés ; on a d'abord vu l'émergence des anticorps monoclonaux, puis de petites molécules capables de traverser la membrane cellulaire. Ces nouveaux traitements ciblent des voies métaboliques spécifiques du cancer du patient. Les anticorps thérapeutiques sont des molécules se liant spécifiquement à un antigène tumoral à la surface des cellules cancéreuses, voire deux antigènes différents (anticorps bi-spécifiques). En 1997/1998, deux premiers anticorps ont été approuvés par la FDA (*Food and Drug Administration*), il s'agit du Rituximab (anti-CD20 pour le lymphome B) et du Trastuzumab (anti-HER2 pour le cancer du sein). Depuis, 19 anticorps monoclonaux ont été approuvés dans différents cancers, comme le montre le Tableau 1.

Plusieurs modes de fonctionnement sont possibles pour les anticorps monoclonaux (l'anticorps cible l'antigène d'une cellule). Ils peuvent activer une voie de signalisation, bloquer un récepteur afin de provoquer l'apoptose ou encore bloquer un ligand. Ils peuvent aussi être liés à une drogue (ADC : *Antibody Dependent Cytotoxicity*) ou à un élément radioactif (RIT : Radioimmunothérapie) qui va agir sur la cellule cancéreuse pour la détruire.

Les traitements de type anticorps monoclonaux font partie de ce qu'on appelle l'immunothérapie 'passive'. D'autres traitements se sont développés récemment qui stimulent le système immunitaire des patients pour détruire la tumeur : il s'agit d'une immunothérapie 'active'.

Tableau 1 : Anticorps monoclonaux approuvés par la FDA comme traitement contre le cancer

Nom de l'anticorps	Cible	Cancer	Date d'approbation (FDA)	Méthode d'action
Trastuzumab	HER2	Sein, gastrique	1997	
Rituximab	CD20	Lymphone non Hodgkinien (LNH), Leucémie lymphoïde chronique (LLC)	1997	
90Y-labeled ibritumomab tiuxetan	CD20	Lymphome folliculaire bas grade, NHL	2002	RIT
131I-labeled tositumomab	CD20	Lymphome folliculaire bas grade, NHL	2003	RIT
Cetuximab	EGFR	Colon, voies aérodigestives	2004	
Panitumumab	EGFR	Colon	2006	
Bevacizumab	VEGF	Sein	2006	
Alemtuzumab	CD52	LLC	2007	
Ofatumumab	CD20	LLC	2009	
Ipilimumab	CTLA4	Mélanome	2011	
Brentuximab vedotin	CD30	Lymphome Hodkinien	2011	ADC
Pertuzumab	HER2	Sein	2012	
Ado trastuzumab entansine	HER2	Sein	2013	ADC
Obinutuzumab	CD20	LLC	2013	
Ramucirumab	VEGFR2	gastrique	2014	
Pembrolizumab	PD1	Mélanome	2014	
Blinatumomab	CD19,CD3	Leucémie lymphoblastique aigue	2014	Bi-spécifique
Nivolumab	PD1	Mélanome	2014	
Dinutuximab	GD2	Neuroblastome	2015	

1.1.3. Hétérogénéité des cancers

Ces nouveaux traitements ciblés ont permis d'améliorer la survie globale des patients. Toutefois, ils ne permettent pas encore de traiter l'ensemble des cancers. Et s'ils se sont montrés très efficaces dans certains cas, ils ne bénéficient qu'à un sous-groupe de patients au sein d'un

même type histologique de cancer. L'utilisation et le développement de ces traitements nécessitent une très bonne connaissance de la maladie et de son hétérogénéité. Par exemple, l'utilisation du Trastuzumab est liée à l'amplification de HER2 que l'on retrouve chez environ 15% des patients atteints d'un cancer du sein. La découverte de nouveaux traitements et l'amélioration de leur utilisation ne peut se faire que grâce à la compréhension de la maladie au niveau moléculaire. De grands programmes (Chin et al., 2011) comme le TCGA (*The Cancer Genome Atlas*) aux Etats-Unis, le programme CIT (Carte d'Identité des Tumeurs) en France, ou encore le programme ICGC (*International Cancer Genome Consortium*) au niveau international, se sont donnés pour objectif l'analyse à différentes échelles des altérations du génome à l'origine des cancers et leur classification moléculaire détaillée. L'objectif est de favoriser l'émergence d'une médecine dite personnalisée, c'est-à-dire une médecine permettant à chaque patient d'être traité de façon individualisée, en fonction des spécificités génétiques et biologiques de sa tumeur. Il est important pour les cliniciens de connaitre les sous-types histologiques et moléculaires de la maladie, car ils vont conditionner diagnostic et pronostic, ainsi que la prise en charge thérapeutique la plus adéquate (Allison and Sledge, 2014).

Malgré les différents efforts de classification, la majorité des thérapies ciblées ne suffisent pas à assurer une guérison au malade. Très souvent, le traitement est efficace pendant un temps, permettant une rémission du patient, mais des phénomènes de résistance limitent l'efficacité sur le long terme. L'une des hypothèses expliquant cette résistance aux traitements est basée sur l'hétérogénéité intra-tumorale. Ainsi, la thérapie ciblée éliminerait une majeure partie de la tumeur permettant la rémission du patient mais les clones minoritaires résistants au traitement initial pourraient ensuite reprendre leur expansion tumorale. L'hétérogénéité intra-tumorale pourrait aussi influer sur la diffusion des métastases en privilégiant certains clones selon les organes cibles des métastases. L'émergence de technologie permettant d'analyser le transcriptome au niveau d'une seule cellule (*single-cell RNAseq*) a permis d'étudier ces phénomènes au sein d'une tumeur. Une équipe a ainsi étudié l'hétérogénéité des glioblastomes au sein des tumeurs (Patel et al., 2014) et a montré que plusieurs sous-types définis précédemment cohabitaient.

La définition de l'hétérogénéité des cancers est donc double (Figure 6) : on parle d'abord d'hétérogénéité inter-tumorale pour expliquer l'existence de différents types de cancers bien qu'ils aient pour origine le même organe, on parle d'hétérogénéité intra-tumorale car au sein même d'une tumeur plusieurs clones coexistent.

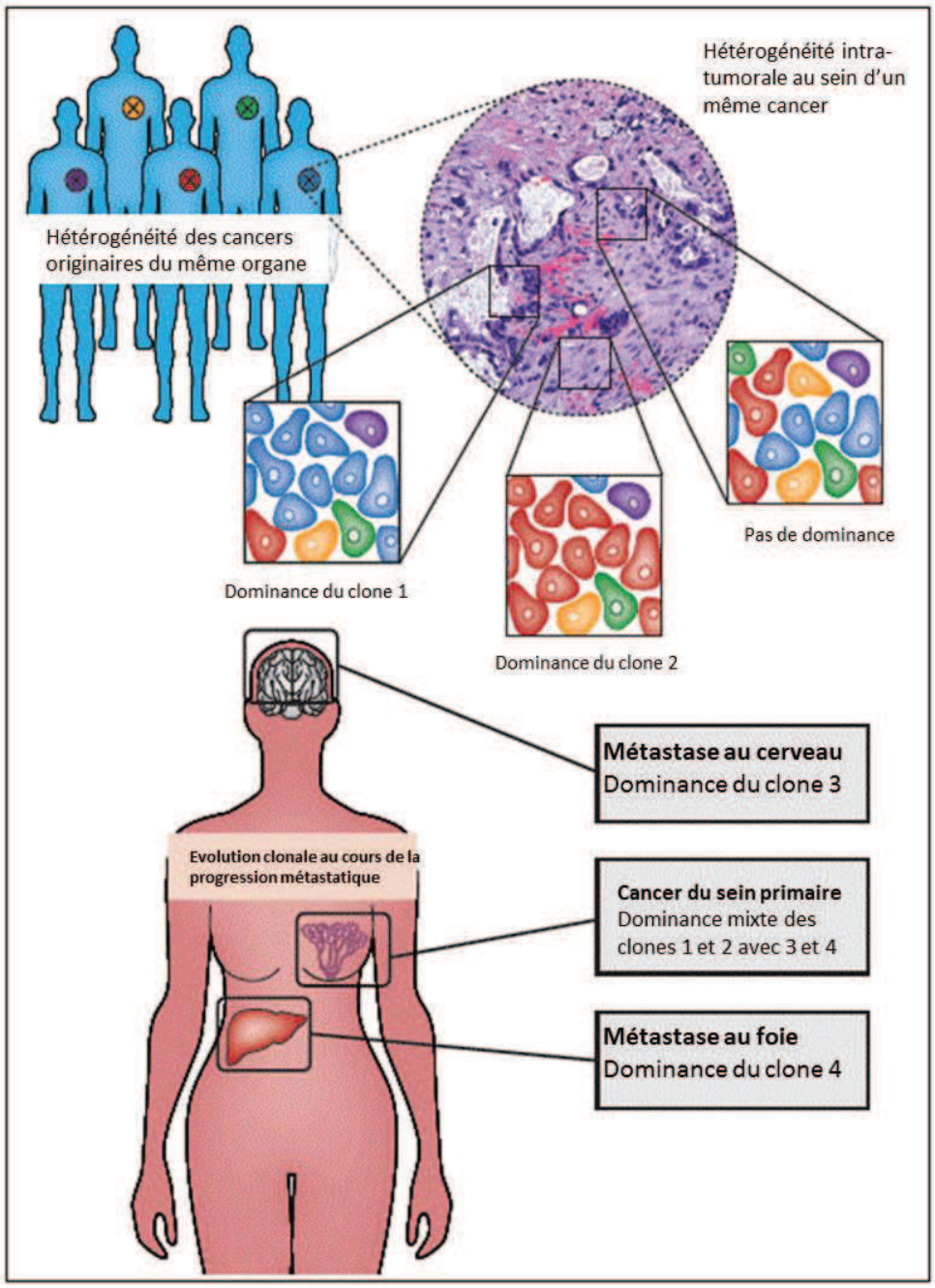

Figure 6 : Hétérogénéité des cancers à différents niveaux

Source : adaptée de Allison and Sledge (Allison and Sledge, 2014)

Une deuxième hypothèse pour expliquer la résurgence des tumeurs, y compris après chirurgie, porte sur l'existence de cellules souches cancéreuses (ce point sera approfondi dans la partie 1.1.4).

L'hétérogénéité intracellulaire montre la nécessité d'attaquer la tumeur à l'aide de différentes thérapies afin d'éliminer les différents clones. Pour cela, il faut définir les différentes populations cellulaires des tumeurs et développer des traitements les ciblant spécifiquement.

1.2 Glioblastomes et cellules souches de glioblastomes

1.2.1 Les glioblastomes

1.2.1.1. Description des gliomes

Les gliomes sont l'ensemble des tumeurs cérébrales issues des cellules gliales, cellules qui assurent le soutien des neurones.

La classification de l'OMS (Organisation Mondiale de la Santé), la plus utilisée, distingue trois types de gliomes selon leur composition cellulaire (Louis et al., 2007) : les astrocytomes, les oligodendrogliomes et les oligo-astrocytomes. Ces tumeurs sont classées de I à IV en fonction de leur degré de malignité. Le grade I (tumeurs bénignes) correspond à des tumeurs ayant une croissance lente et bien délimitées par rapport au tissu sain. Le grade II correspond à des tumeurs ayant une croissance lente mais avec un caractère infiltrant, c'est-à-dire des tumeurs mal limitées, envahissantes. Le grade III (tumeurs malignes) correspond à des tumeurs anaplasiques, c'est-à-dire ayant perdu une partie de leurs caractères propres ; leur évolution est plus rapide que celle des grades inférieurs. Le grade IV correspond à des tumeurs malignes, se multipliant rapidement, avec une forte tendance à la nécrose et très infiltrantes.

Parmi les astrocytomes, on trouve les astrocytomes de grade IV ou glioblastomes, représentant plus de 50% des gliomes (Ostrom et al., 2014), il s'agit du plus commun des gliomes. Son taux d'incidence est d'environ 3 pour 100 000.

1.2.1.1. Classifications des glioblastomes

Les glioblastomes (GBMs) se divisent en deux parties ayant la même histologie mais une origine différente : les glioblastomes primaires, ou de novo, qui représentent 85% des GBMs, et les glioblastomes secondaires, issus d'une évolution des grades inférieurs (Agnihotri et al., 2013). Les glioblastomes secondaires sont issus d'astrocytomes de bas grades et ont tendance à apparaitre chez des patients jeunes (moins de 45 ans), contrairement aux glioblastomes primaires. Les deux types de GBMs présentent des caractéristiques génétiques très différentes, qui sont résumées Figure 7.

Les GBMs primaires sont caractérisés par l'amplification et la mutation du gène EGFR (*Epidermal Growth Factor Receptor*), la perte d'hétérozygotie[8] (LOH) du chromosome 10q (contenant le gène PTEN (*Phosphatase and TENsin homolog*)), la sur-expression de MDM2 (*Murine Double Minute 2*) et la délétion de p16. Le gène IDH1 (*Isocitrate Dehydrogenase 1*) est muté dans la majorité des cas.

[8] Perte d'une des deux copies d'un ensemble de gènes

Les GBMs secondaires sont caractérisés par la mutation de TP53 (*Tumor Protein P53*) et RB (*Retinoblastoma 1*), la sur-expression de PDGF A (*Platelet-Derived Growth Factor Alpha Polypeptide*) et PDGFRa (*Platelet-Derived Growth Factor Receptor, Alpha Polypeptide*) et la perte d'hétérozygotie du chromosome 19q.

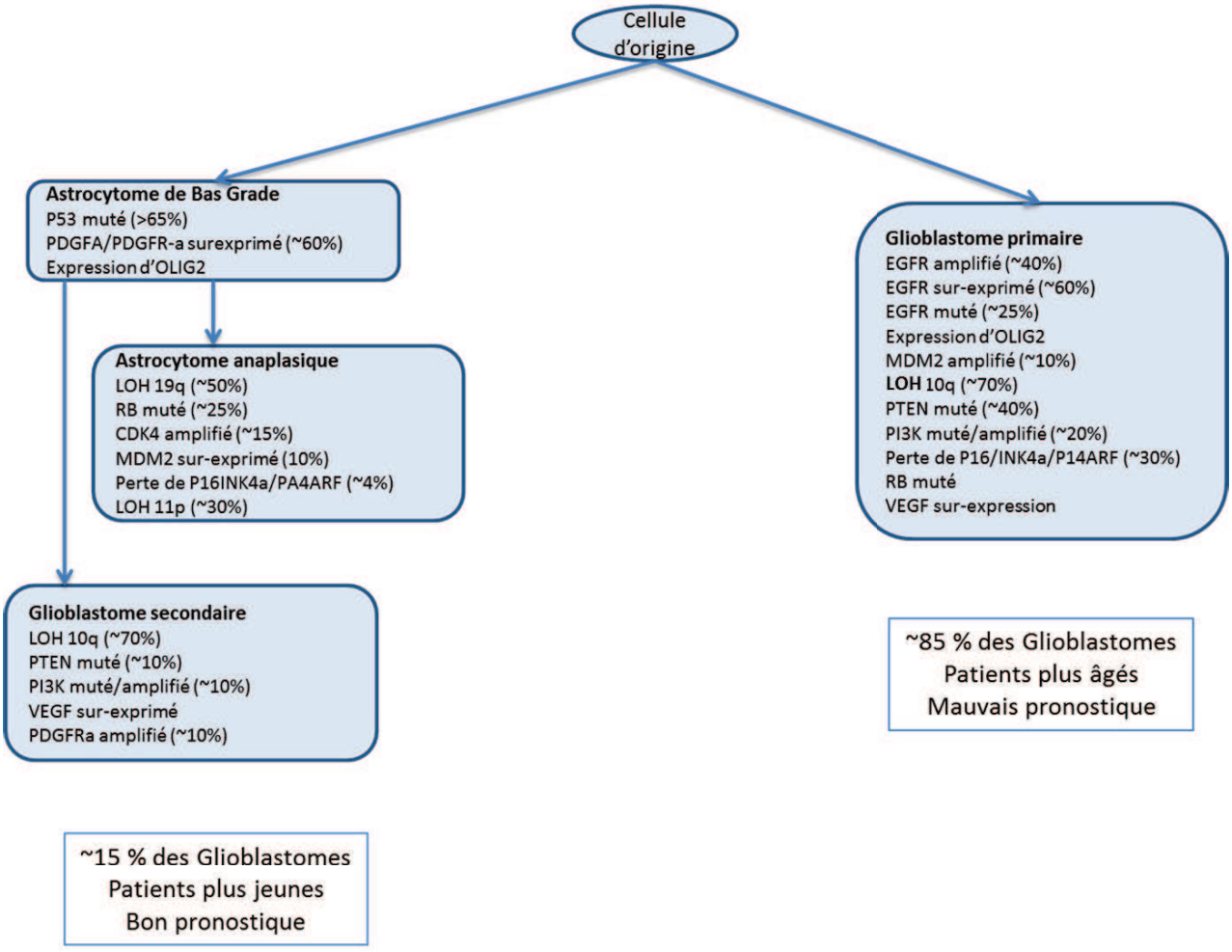

Figure 7: Caractéristiques génétiques des glioblastomes primaires et secondaires
Source : adapté de Agnihotri et al (Agnihotri et al., 2013)

The Cancer Genome Atlas (TCGA) a regroupé près de 600 échantillons de glioblastomes et les a analysés à différents niveaux : séquençage ADN, nombre de copies et méthylation de gènes, transcriptome (séquençage ARN ou puces d'expression), micro-ARN. Cette étude à grande échelle (Brennan et al., 2013; Cancer Genome Atlas Research Network, 2008) a permis de mettre en évidence trois voies de signalisation dérégulées dans la plupart des GBMs : les voies activées par les récepteurs à tyrosine kinase (RTKs) et celles des suppresseurs de tumeurs TP53 et RB.

Ces deux groupes de GBMs montrent la limite de la classification histologique de l'OMS. Néanmoins, les GBMs sont très hétérogènes et nécessitent une caractérisation et classification plus précise, permettant de prédire la survie et la réponse au traitement. Ce travail a été réalisé

à partir d'une classification non supervisée du transcriptome par le TCGA (Verhaak et al., 2010). Ils ont identifiés quatre sous-types moléculaires (voir Figure 8) : glioblastomes classiques, mésenchymateux, proneuraux et neuraux.

Les glioblastomes proneuraux sont définis par des altérations des gènes PDGFRA (amplifications, mutations) et IDH1 (mutations). On y trouve aussi la majorité des mutations TP53. Ce groupe contient le pourcentage de jeunes patients le plus élevé, surement à cause de l'enrichissement en mutations IDH1, qui est associé à des patients plus jeunes. Les glioblastomes classiques sont caractérisés par l'amplification d'EGFR et la perte de PTEN. Ce sous-type est aussi associé à EGFRvIII, le mutant d'EGFR ayant une délétion des exons 2-7. Le sous-type mésenchymateux est associé avec une espérance de vie très faible. Au niveau moléculaire, il est enrichi en mutations NF1 (*Neurofibromin 1*) et perte d'hétérozygotie pour TP53 et CDKN2A (*Cyclin-Depedent Kinase Inhibitor 2A*).

Les glioblastomes neuraux ont des marqueurs neuraux élevés comme NEFL (*Neurofilament, Light Polypeptide*), et un taux de mutation d'ERBB2 (*Erb-B2 Receptor Tyrosine Kinase 2*) important.

Avec cette classification, on observe un effet significatif du traitement sur la survie des patients des groupes classiques et mésenchymateux mais peu de bénéfices pour les patients des groupes neuraux et proneuraux.

La classification a été reprise en 2013 (Brennan et al., 2013) avec l'ajout de nouveaux échantillons et l'utilisation de nouvelles technologies telles que le séquençage de l'ARN, l'analyse protéique par RPPA (*Reverse Phase Protein Array*), le séquençage d'exomes, ainsi qu'une analyse de l'épigénétique du glioblastome. En plus d'identifier de nouvelles mutations, cette analyse montre la division du sous-type proneural en deux, l'une ayant un phénotype CIMP (*CpG island methylator phenotype*), c'est-à-dire une hyperméthylation[9] au niveau des promoteurs de gènes, l'autre pas. Ce phénotype a déjà été observé dans différents cancers. Dans le glioblastome, où il est nommé G-CIMP, il semble associé à la mutation d'IDH1. Ce phénotype est associé avec une meilleure survie et pourrait être utilisé comme biomarqueur, ainsi que la méthylation du promoteur de MGMT (*O-6-Methylguanine DNA Methyltransferase*), qui peut être utilisé comme marqueur de la réponse au témozolomide. En effet, MGMT est une protéine de réparation de l'ADN qui, lorsqu'elle est active (et donc non méthylée) confère aux cellules une résistance aux agents alkylants comme le témozolomide.

[9] Modification chimique de l'ADN (ajout d'un groupe methyl) impactant l'expression des gènes ciblés

A

	Proneural	Neural	Classique	Mésenchymateux
Age median	51.8	63.8	55.7	57.7
Amplification (%)				
EGFR	54	96	100	95
CDK6	46	96	92	89
MET	54	92	86	91
Forte amplification (%)				
EGFR	17	67	95	29
PDGFRA	35	13	5	9
LOH				
NF1	6	17	5	38
PTEN	69	96	100	87
CDKN2A/CDKN2B	56	71	95	67
RB1	52	46	16	53
Mutations				
TP53	54	21	0	32
PTEN	16	21	23	32
NF1	5	16	5	37
EGFR	16	26	32	5
IDH1	30	5	0	0
PIK3R1	19	11	5	0
RB1	3	5	0	13
ERBB2	5	16	5	3
EGFRvIII	3	0	23	3
PIK3CA	8	5	5	3
PDGFRA	11	0	0	0

B

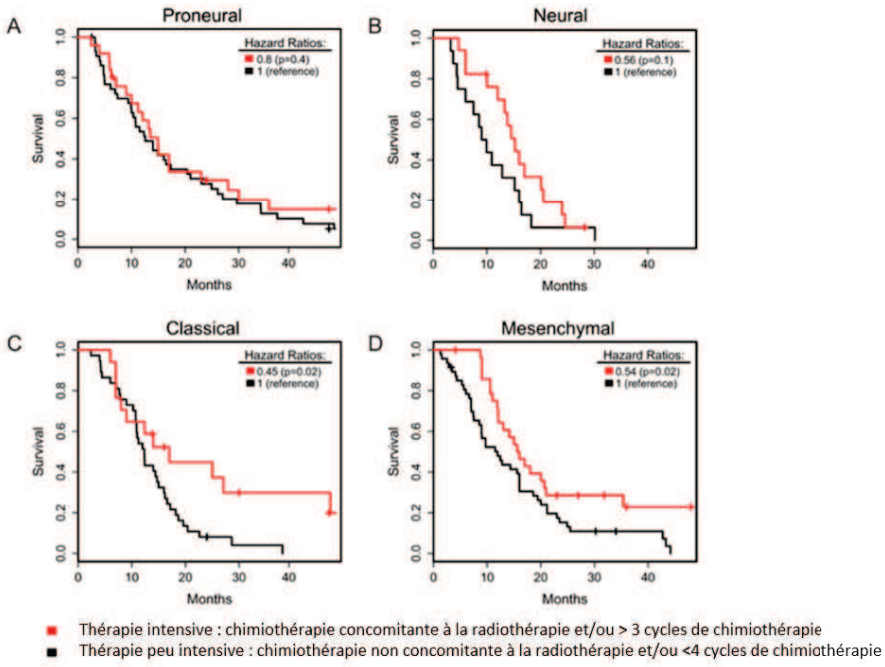

Figure 8 : Classification de Verhaak

La classification distingue 4 sous-groupes de glioblastomes : Proneural, neural, Classique, Mésenchymateux. A : Principales altérations génétiques. B : Courbes de survie des patients selon le sous-groupe et le traitement reçu. *Source : (Verhaak et al., 2010)*

D'autres classifications ont été publiées, comme celle de Sturm (Sturm et al., 2012) se basant sur des données épigénétiques (

Figure 9). Il s'agit d'une classification plus globale puisqu'elle prend en compte les glioblastomes adultes et pédiatriques. Six sous-groupes de glioblastomes ont été identifiés à partir de données épigénétiques et génétiques (profil de méthylation, mutations, gain ou perte

chromosomique), immuno-histochimiques (pour les protéines OLIG2 (*Oligodendrocyte Lineage Transcription Factor 2*) et FOXG1 (*Forkhead Box G1*)), épidémiologiques (âge), cliniques (localisation de la tumeur) et pronostiques. Les 6 sous-groupes de glioblastomes sont :

- IDH : Les glioblastomes du jeune adulte associés à la mutation IDH1,
- K27 : Les glioblastomes de l'enfant associés à la mutation K27 de l'histone H3F3A,
- G34 : Les glioblastomes de l'adolescent associés à la mutation G34 de l'histone H3F3A,
- RTK1 : Les glioblastomes de l'enfant et de l'adulte, caractérisé par l'amplification de PDGFRa (Récepteur à Tyrosine Kinase),
- Mésenchymateux : Les glioblastomes de l'enfant et de l'adulte sans spécificité (peu de perte ou de gain chromosomique, ressemble au profil d'un tissu cérébral non tumoral),
- RTK2 : Les glioblastomes de l'adulte, présentant une amplification d'EGFR (Récepteur à Tyrosine Kinase).

Cette classification est moins précise concernant les glioblastomes de l'adulte mais elle permet une comparaison avec les glioblastomes pédiatriques.

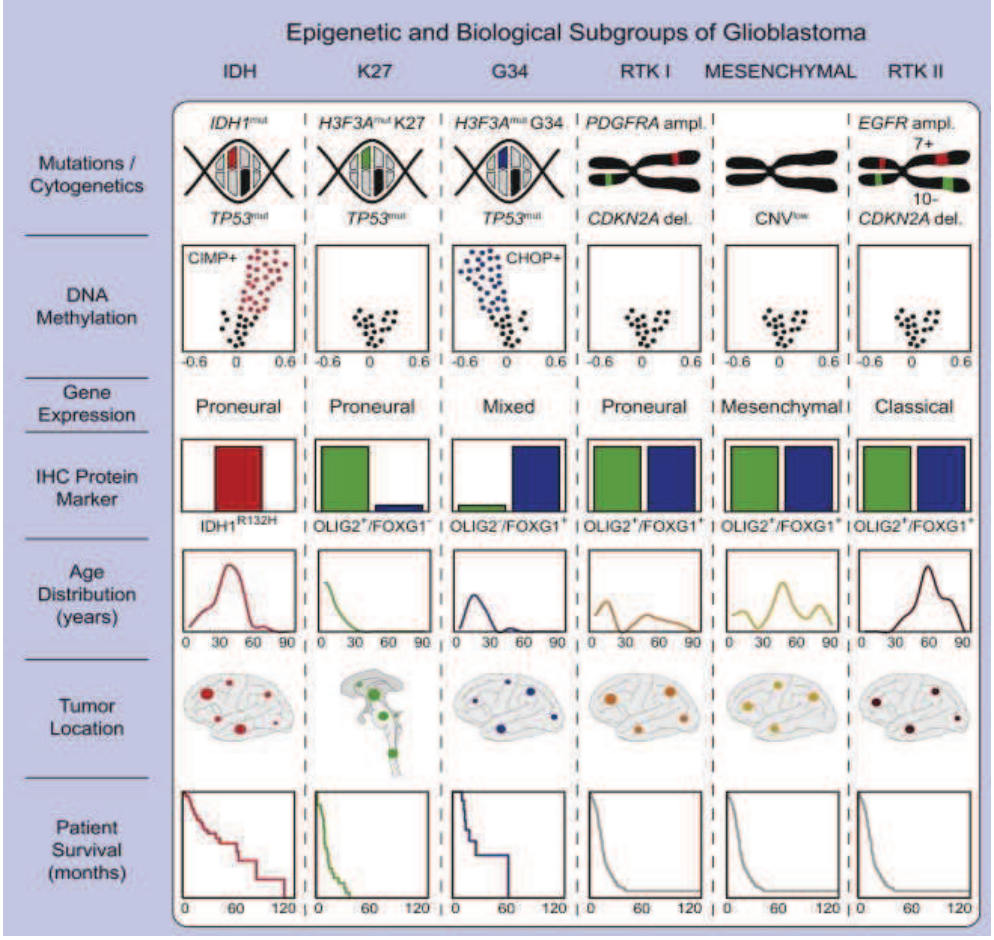

Figure 9 : Classification des glioblastomes adultes et pédiatriques de Sturm

La classification distingue 6 sous-groupes caractérisés par des données épigénétiques et génétiques

Source : (Sturm et al., 2012)

Il semble que les glioblastomes de l'enfant et de l'adulte soient différents au niveau épigénétiques puisque quatre groupes sur six sont liés à l'âge (pédiatrique ou adulte). Cette classification montre la différence entre des glioblastomes pédiatriques et adultes, tout en réussissant à les comparer. La question à se poser est : faut-il étudier ces deux types de tumeurs ensemble ou bien sont-elles trop différentes ?

1.2.1.2. Traitement des glioblastomes

La première phase du traitement des glioblastomes consiste en l'exérèse chirurgicale de la tumeur. Cette chirurgie permet de faire le diagnostic histologique et constitue la première étape de la prise en charge thérapeutique. D'après la littérature, une résection de plus de 90% du volume tumorale permet d'augmenter de manière significative la durée de vie du patient (Stummer et al., 2011 ; Orringer et al., 2012 ; Chaichana et al., 2014).

Actuellement, le traitement complémentaire suit le protocole décrit par Stupp (Stupp et al., 2005), il associe radiothérapie adjuvante et témozolomide (TMZ) comme chimiothérapie de référence. Les résultats de l'essai cliniques de Stupp montrent une augmentation de la médiane de survie des patients traités par TMZ par rapport aux patients traités seulement avec la radiothérapie de 12,1 mois à 14,6 mois, ainsi qu'une augmentation de la médiane de survie sans progression de 5 mois à 6,9 mois.

Une autre option thérapeutique approuvée par la FDA dans le traitement des glioblastomes est l'utilisation d'implants de Carmustine, mis en place au sein de la cavité de résection chirurgicale. Il s'agit d'un agent alkylant de l'ADN ayant une action plus faible que le Témozolomide mais prolongée dans le temps. Cependant, sur les glioblastomes, l'action du Témozolomide et de la Carmustine ne permettent pas d'augmenter la survie des patients (Noël et al., 2012).

Le Bevacizumab a été autorisé en 2009 en complément du protocole de Stupp dans le cas de récidives (Cohen et al., 2009). Ce traitement anti-angiogénique permet d'augmenter la survie médiane de 6 mois (Nagpal et al., 2011). Des études sont en cours pour son utilisation dans le cas de glioblastomes nouvellement diagnostiqués.

De nombreux essais cliniques sur les glioblastomes sont en cours. La majorité des molécules testées ciblent les voies de signalisation sur-exprimées. Malheureusement, les résultats sont décevants et le glioblastome reste une maladie incurable.

L'une des hypothèses expliquant les rechutes serait l'existence de « cellules souches cancéreuses », résistantes aux thérapies actuelles. Les traitements actuels ciblent les cellules tumorales « normales » mais ces cellules, résistantes, sont capables de régénérer la tumeur (Pattabiraman and Weinberg, 2014; Reya et al., 2001).

1.1.4. Cellules Souches Cancéreuses

1.2.1.3. Généralités

Les glioblastomes, de mêmes que de nombreuses autres tumeurs, sont constitués d'un ensemble cellulaire hétérogène, sur le plan histologique et moléculaire. La majeure partie de ces cellules est différenciée mais un petit ensemble de cellules présente des capacités d'auto-renouvèlement et de différenciation. Ces cellules ont été nommées cellules souches cancéreuses pour leurs propriétés communes avec les cellules souches ou bien cellules propagatrices de tumeurs. Ces cellules ont d'abord été mises en évidences dans les leucémies aigües myéloïdes (Bonnet and Dick, 1997), puis dans les tumeurs solides, notamment cancers du sein (Al-Hajj et al., 2003), gliomes (Singh et al., 2003), cancers de la prostate (Collins et al., 2005), des ovaires (Bapat et al., 2005), du colon (O'Brien et al., 2007), du foie (Yang et al., 2008), du poumon (Eramo et al., 2008) et du pancréas (Hermann et al., 2007).

1.2.1.4. Caractérisation des Cellules Souches de Glioblastomes (ou gCSCs)

Marqueurs

Historiquement, le marqueur le plus utilisé pour isoler et cibler les cellules souches de glioblastomes est le marqueur CD133 bien qu'on ne connaisse pas son rôle. Il a été observé comme marqueur de cellules souches dans différents cancers. Néanmoins, l'utilisation de CD133 est sujette à controverse, étant donné que des cellules CD133- (ne possédant pas le marqueur CD133) ont montré des caractéristiques de cellules souches cancéreuses dans différentes études (Beier et al., 2007; Joo et al., 2008). Ce marqueur n'est donc pas spécifique.

La Nestine est le deuxième marqueur majoritairement décrit pour les gCSCs (Bexell et al., 2009 ; Zhang et al., 2008a). Il s'agit d'une protéine des filaments intermédiaires du cytosquelette exprimée par les cellules souches neuro-épithéliales.

Beaucoup d'autres marqueurs ont été décrits ces dernières années, comme A2B5 (Ogden et al., 2008 ; Tchoghandjian et al., 2010), CD44 (Anido et al., 2010), CD171 (L1CAM, *L1 Cell Adhesion Molecule*)(Bao et al., 2008), CD15 (SSEA1, *Stage Specific Embryonic Antigen 1*) (Read et al., 2009; Son et al., 2009; Ward et al., 2009), CD49f (integrin alpha 6)(Lathia et al., 2010), nanog (Guo et al., 2013; Mathieu et al., 2011; Niu et al., 2011), oct4 (*Octamer binding transcription factor 4*) (Guo et al., 2011; Ikushima et al., 2011; Mathieu et al., 2011), musashi (Thon et al., 2010) ou encore Sox2 (*Sex Determinining Region Y Box 2*) (Ge et al., 2010 ; Guo et al., 2011 ; Hägerstrand et al., 2011).

Parmi ces marqueurs, un certain nombre sont des marqueurs des cellules souches neurales et se retrouve sur-exprimé dans d'autres cellules souches cancéreuses, comme CD133 ou CD44 (Pattabiraman and Weinberg, 2014).

Aucun n'est universel, ce qui nous montre l'hétérogénéité des cellules souches de glioblastomes. Aujourd'hui, l'utilisation de marqueurs multiples est nécessaire pour caractériser complètement les gCSCs. Dans l'idéal, les cellules souches sont caractérisées de manière fonctionnelle (par exemple la xénogreffe de cellules souches donne une tumeur).

Voies de signalisation

Les voies de signalisation Notch, Hedgehog et Wnt sont les trois voies de signalisation principales altérées dans les gCSCs, qui sont communes aux cellules souches normales. Ces voies ont initialement été décrites dans le cadre du développement embryonnaire. Elles participent à l'auto-renouvèlement et la prolifération des gCSCs (Takebe et al., 2011).

La voie de signalisation Notch joue différents rôles dans le développement du système nerveux : elle favorise l'auto-renouvèlement et empêche la différenciation des cellules souches neurales (Zhong et al., 1997). Notch est activé par contact avec les ligands DLL (*Delta-like*) ou JAG (*Jagged*), ce qui entraine la libération cytoplasmique de la partie intracellulaire de Notch (NICD). NICD migre jusqu'au noyau et permet la transcription de ses gènes cibles (Figure 10). La voie Notch est dérégulée dans de nombreux cancers, dont les GBMs (Kanamori et al., 2007). Dans les gCSCs, la sur-expression de Notch permet le maintien d'un état indifférencié, pluripotent et d'auto-renouvèlement (Zhang et al., 2008b) et favorise la Transition Epithélio-Mésenchymateuse (EMT) (Wang et al., 2010b).

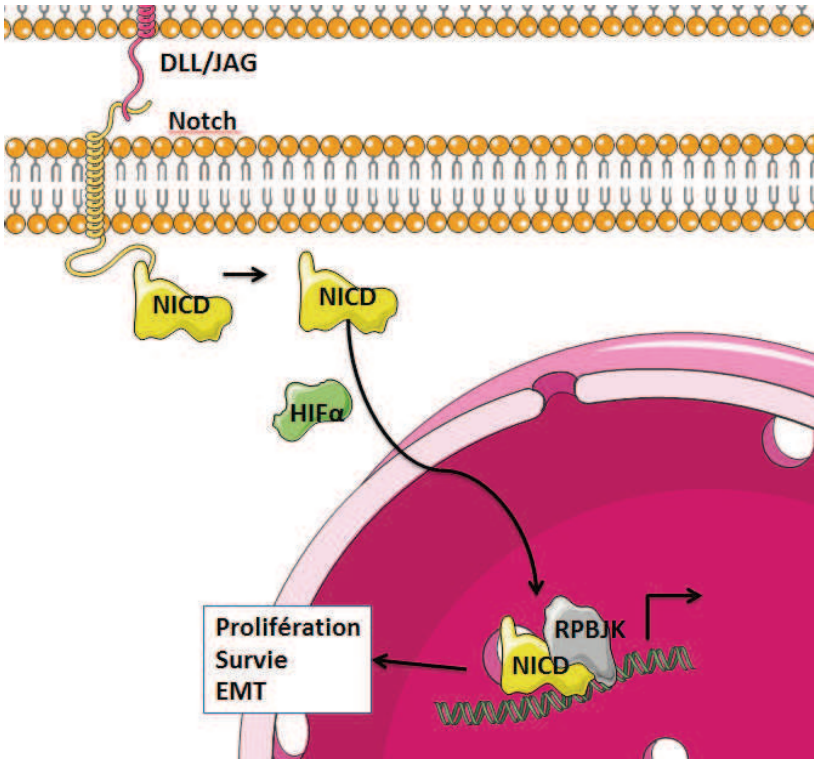

Figure 10 : Schéma simplifié de la voie de régulation Notch
Figure réalisée à l'aide de Servier Medical Art

La fixation du ligand sonic hedgehog (SHH) sur le récepteur PTCH1 (*Protein patched homolog 1*) permet de lever l'inhibition du récepteur SMO (*Smoothened*); ce qui entraine la libération des facteurs Gli qui migrent dans le noyau et activent la transcription de gènes cibles (Figure 11). Cette voie de signalisation est impliquée dans la progression des tumeurs et maintient les gCSCs dans un état indifférencié et invasif (Clement et al., 2007).

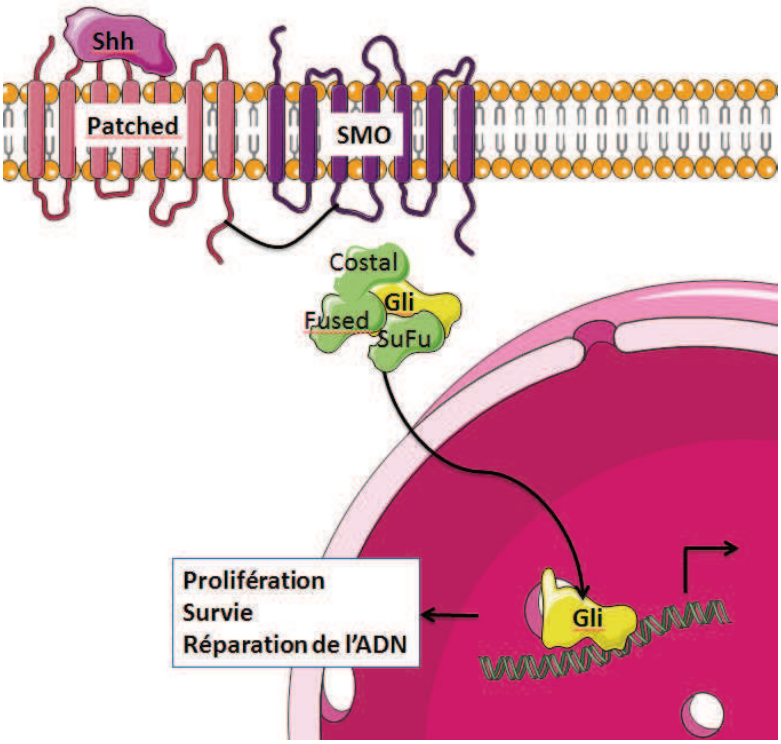

Figure 11 : Schéma simplifié de la voie Sonic Hedgehog.
Figure réalisée à l'aide de Servier Medical Art.

L'activation de la voie Wnt se fait par la fixation de Wnt à son récepteur FZD (*Frizzled*) et aux co-récepteurs LRP (*Low Density Lipoprotein Receptor*). Aujourd'hui 19 membres de la famille Wnt ont été identifiés (Katoh and Katoh, 2007), ainsi que 10 récepteurs Frizzled. En absence de Wnt, la ß-caténine est associée à un complexe multi-protéiques de destruction, comprenant APC (*Adenomatous Polyposis Coli*), GSK3 (*Glycogen Synthase Kinase 3*) et AXIN, phosphorylée ; ubiquitinée et dégradée par le protéasome (Figure 12). La liaison de Wnt à son récepteur libère la ß-caténine, qui migre dans le noyau où elle s'associe à p300 ou CBP (*Cyclique AMP response element Binding Protein*) et agit en tant que facteur de transcription sur des gènes impliqués dans la prolifération, l'auto-renouvèlement et la tumorigénicité (Kahn, 2014 ; Rampazzo et al., 2013 ; Rheinbay et al., 2013 ; Sandberg et al., 2013).

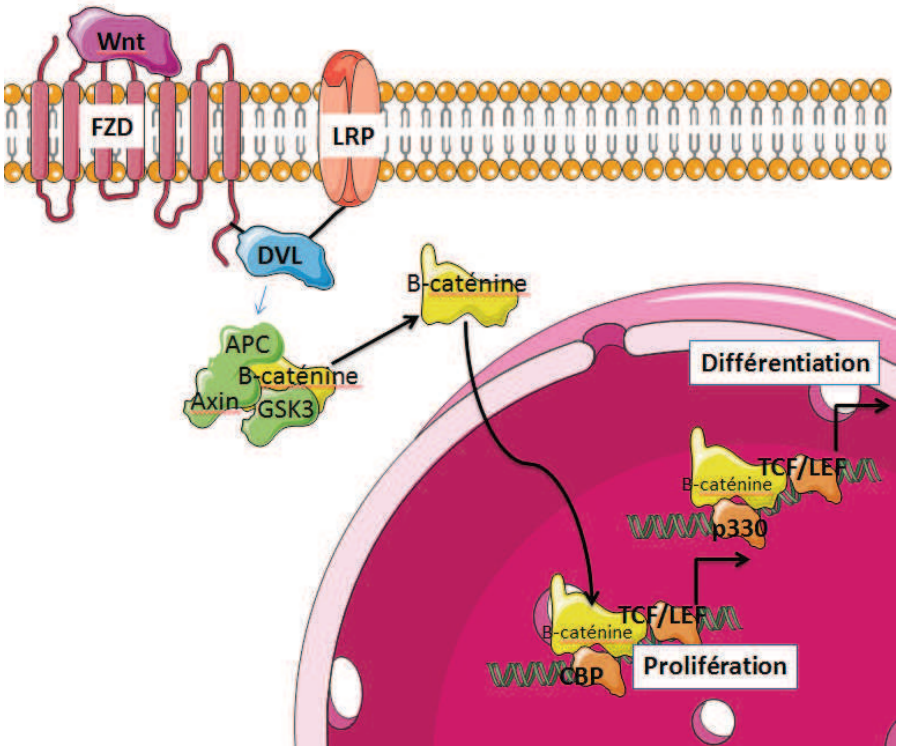

Figure 12 : Schéma simplifié de la voie Wnt/β-caténine
Figure réalisée à l'aide de Servier Medical Art

Transition épithéliamo-mésenchymateuse (EMT)

L'EMT est le processus qui permet à une cellule épithéliale de modifier sa composition et l'organisation de ses protéines afin de perdre ses caractéristiques épithéliales (adhésion, manque de mobilité) et d'acquérir les caractéristiques des cellules mésenchymateuses lui donnant des capacités migratoires et invasives, ainsi qu'une bonne résistance à l'apoptose (Thiery, 2002). L'EMT a un rôle crucial au cours du développement embryonnaire mais dans le contexte du cancer, elle permet aux cellules cancéreuses d'envahir les organes distants où elles initient les métastases.

L'EMT dépend du microenvironnement de la tumeur et peut être induite par plusieurs facteurs, dont le TGF-β et WNT (Pattabiraman and Weinberg, 2014). Il semble que le caractère souche de certaines cellules tumorales soit lié à l'EMT. En effet, plusieurs équipes ont montré que dans certains cas l'EMT peut induire des capacités souches telles que l'auto-renouvèlement, la différentiation et l'invasion (Mani et al., 2008 ; Morel et al., 2008).

1.2.1.5. Résistance des gCSCs aux thérapies classiques du GBM

Tout porte à croire que l'inefficacité des traitements actuels seraient dû à la résistance des gCSCs aux thérapies des glioblastomes. Les glioblastomes sont traités par chimiothérapie avec le témolozomide (TMZ) et par radiothérapie.

Concernant la chimiorésistance, le TMZ, en tant qu'agent alkylant abime l'ADN et induit l'apoptose. Il existe plusieurs mécanismes de résistance des gCSCs. L'une des premières explications est la sur-expression du gène codant pour l'enzyme MGMT, qui est une enzyme de réparation de l'ADN. Les gCSCs limitent ainsi l'impact du TMZ (Qiu et al., 2014). Un autre mécanisme, le phénotype MDR (MultiDrug Resistance) explique cette résistance. Il est dû à la sur-expression de transporteurs membranaires ABC (*ATB Binding Cassette*) transportant de façon active et non spécifique les molécules chimiques en-dehors de la cellule. Les gCSCs montrent une sur-expression des transporteurs ABCB1 (glycoprotéine P, MDR-1), ABCC1 (MRP1) et ABCG2 (BCRP1) (Dean et al., 2005). Enfin, les agents chimiothérapeutiques ciblent l'activité du cycle cellulaire. Or les gCSCs sont dormantes ou ont des cycles cellulaires lents, ces cellules sont donc peu touchées par des molécules agissant sur des cellules en prolifération rapide.

La radiothérapie induit la mort des cellules cancéreuses en abimant l'ADN par rayonnement. Or les gCSCs ont des capacités importantes de réparation de l'ADN. Dans l'étude de (Bao et al., 2006), les auteurs ont montré un enrichissement en cellules CD133+ après radiothérapie. D'après leurs résultats, la résistance des gCSCs à la radiothérapie serait liée à la hausse de l'activité de réparation de l'ADN de la voie CHK1/CHK2 (checkpoint kinase ½). Une autre étude a montré que L1CAM (CD171), marqueur de gCSCs, augmentait la résistance des gCSCs en renforçant la réparation de l'ADN (Cheng et al., 2011) par le biais de NBN (*Nibrin*), composante du complexe MRN (MRE11-RAD50-NBN) qui active la kinase ATM et la réparation de l'ADN. L'utilisation de shRNA (Small Hairpin RNA, petits ARN en épingles à cheveux permettant de reconnaitre et dégrader un ARNm cible) pour diminuer l'expression de L1CAM a pour conséquence la diminution de la réparation de l'ADN et la sensibilisation des gCSCs à la radiation. Un autre mécanisme de radiorésistance par le biais de BMI1 (*BMI1 Proto-Oncolgene, Polycomb Ring Finger*) a été rapporté dans (Facchino et al., 2010). Après un traitement de radiothérapie, BMI1 était davantage présent au niveau de la chromatine et associé à des protéines impliquées dans la réparation des cassures double brin de l'ADN. L'utilisation de shRNA pour inactiver BMI1 a eu pour conséquence une sensibilisation des gCSCs à la radiation.

Puisque les traitements classiques sont inefficaces, il faut trouver des alternatives pour les cibler spécifiquement et éviter la récurrence de la tumeur. Pour cela, plusieurs stratégies thérapeutiques peuvent être utilisées : cibler les voies impliquées dans leur auto-renouvèlement et leur prolifération, induire leur différenciation ou sensibiliser les gCSCs à la radiothérapie et à la chimiothérapie. Les trois voies principales impliquées dans l'auto-renouvèlement et la prolifération sont les voies de signalisation Notch, SHH et Wnt. Ces voies sont des cibles de choix pour supprimer les caractéristiques souches des cellules et les rendre sensibles aux traitements

classiques. Plusieurs essais cliniques sont en cours en utilisant des inhibiteurs de ces voies de signalisation. Il est possible de combiner l'inhibition de ces différentes voies de signalisation afin d'augmenter l'efficacité des traitements. Par exemple, Ulasov et al ont montré que l'inhibition des voies Hedgehog et Notch permet d'augmenter la sensibilité des gCSCs au TMZ (Ulasov et al., 2011).

Les thérapies futures devront pouvoir cibler les GBMs et les gCSCs pour avoir une chance de guérir le patient. De même que les GBMs sont hétérogènes, les gCSCs le sont aussi, comme le montrent les différents marqueurs qui ne sont spécifiques que de sous-populations. Il faut donc d'abord étudier ces sous-groupes pour pouvoir les cibler spécifiquement afin d'éviter la récurrence de la tumeur à partir d'un clone non traité.

1.3 Technologies haut et moyen débit utilisées

Nous avons vu dans les paragraphes précédents que des anomalies génomiques sont à l'origine des caractéristiques qui transforment une cellule saine en cellule cancéreuse. Ces anomalies sont étudiées à grande échelle grâce aux technologies « omiques ». Elles se basent sur l'étude des différents niveaux d'informations entre l'ADN et les métabolites, en passant par l'ARN messager et les protéines comme expliqué dans la Figure 13.

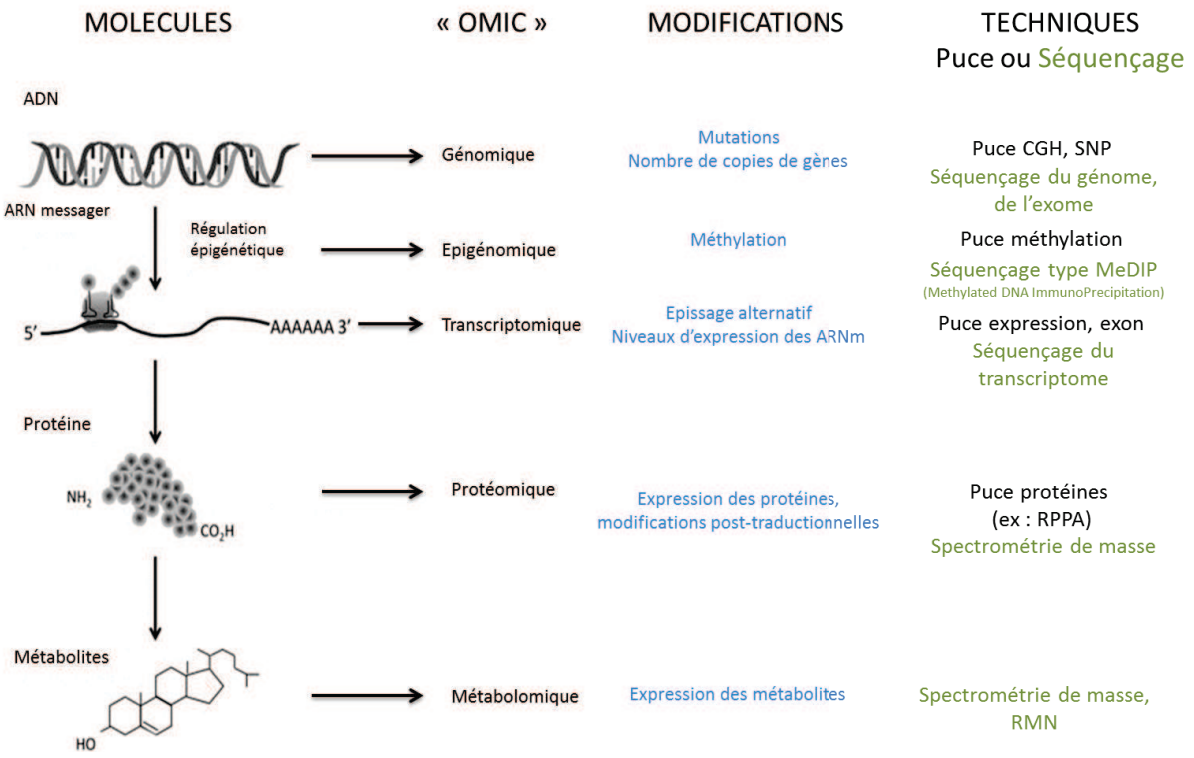

Figure 13 : Analyses puce et séquençage des "omiques"
d'après (Wood et al., 2014)

La génomique est l'analyse constitutionnelle de l'ADN. Plusieurs types d'altérations peuvent avoir lieu, que ce soient des mutations, des insertions ou délétions de quelques nucléotides, ou bien une variation du nombre de copies de certains gènes.

L'épigénomique est l'analyse de l'ensemble des modifications que subit l'ADN sans en modifier sa séquence, qui sont transmissibles de la cellule mère aux cellules filles et qui ont un caractère réversible. Les deux principaux composants du code épigénétique sont d'une part les mécanismes de méthylation de l'ADN (ajout d'un groupement méthyl), d'autre part les modifications des queues d'histones, protéines autour desquelles l'ADN s'enroule et qui sont liées à sa compaction, que ce soit par méthylation ou acétylation (ajout d'un groupement acetyl). Ces modifications ont un impact sur la transcription des gènes, en l'activant ou la réprimant. L'épigénome d'un individu évolue au cours du temps, contrairement à son génome.

La transcriptomique concerne l'analyse des ARNs messagers, en particulier des modifications de leur niveau d'expression entre différentes conditions, ainsi que des évènements d'épissage alternatifs.

Epissage alternatif (Figure 14) : Processus par lequel les ARN transcrits à partir de l'ADN génomique peuvent subir des étapes de coupure et ligature qui conduisent à l'élimination de certaines régions dans l'ARN final. Les segments conservés sont les exons et ceux qui sont éliminés les introns. Le transcrit final peut contenir plus ou moins d'exons, conduisant à l'existance de plusieurs isoformes (ou ARNm différents mais du même gène), et donc de plusieurs protéines.

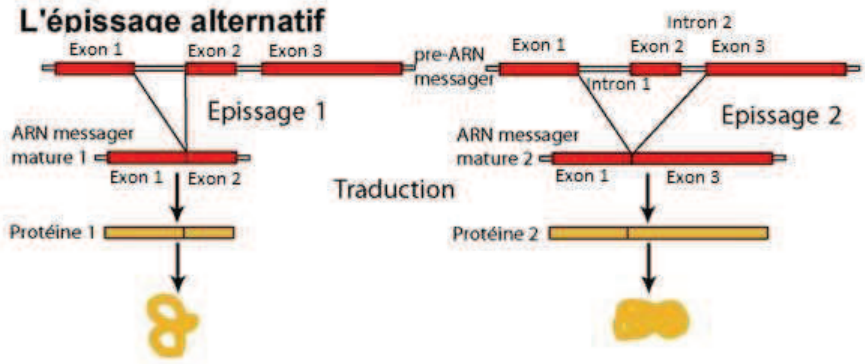

Figure 14 : L'épissage alternatif

Source : http://www.snv.jussieu.fr/vie/dossiers/evolution/evol/traits_2.html

La protéomique est l'analyse des protéines. Il s'agit de connaitre le protéome d'un échantillon donné, c'est-à-dire l'ensemble des protéines présentes au moment de l'analyse, ainsi que les modifications post-traductionnelles qu'elles subissent.

L'ensemble de ces analyses peut se faire à partir de deux grandes techniques : les puces et le séquençage (Figure 13).

1.3.1 Evolution technologique

Les technologies « omiques » ont fortement évolué ces dernières années. Nous allons nous intéresser particulièrement aux puces d'expression et aux analyses de séquençage de l'ARNm pour la partie transcriptomique et aux analyses de types spectrométrie de masse pour la partie protéomique, technologies utilisées dans les analyses présentées dans les chapitres suivants.

1.3.1.1. Transcriptomique

Le principe général des puces d'expression et des analyses de séquençage repose sur la propriété d'hybridation d'un brin d'ADN sur son complémentaire. L'utilisation de cette propriété pour la détection spécifique d'ADN a été décrite par Ed Southern (Southern, 1975) et est à l'origine des Southern et Northern blot, l'ancêtre des puces à ADN (Lander, 1999).

Puces d'expression

Les puces sont faites de sondes ADN déposées ou synthétisées par photolithographie sur une surface solide (verre, plastique ou silicone). Par hybridation de cibles (ADN ou ADN complémentaire (ADNc) obtenus à partir de l'ARN messager extrait des cellules et rétro-transcrit) marquées par fluorescence sur les sondes, on peut mesurer les niveaux d'expression de ces sondes dans les échantillons étudiés.

Il existe trois technologies différentes :

- *Spotted array* : les sondes ADN sont synthétisées puis déposées à la surface de la puce. Les probes sont des fragments courts ou longs d'ADN (ou oligonucléotides). Dans ce type de puce, on hybride deux échantillons d'ADN complémentaires marqués par un fluorochrome : le tissu analysé et un tissu contrôle. Un scanner lit l'intensité relative des fluorochromes et permet d'obtenir une expression relative des gènes.
- *Puces à oligonucléotides* : les probes correspondent à une partie de la séquence d'ARNm connue ou prédite à partir du génome. Ces puces se composent de probes de 50 à 60 mers (*Long Oligonucleotide Arrays*) ou 25 à 30 mers (*Short Oligonucleotide Arrays*). Plusieurs probes ciblent le même gène et forment un probeset. Les compagnies Affymetrix ou Agilent ont créé des puces permettant d'analyser le génome humain. Ces puces fournissent une estimation de l'expression absolue des gènes.
- *Bead arrays* : Développée par Illumina et sortie en 2002, cette technologie se base sur des billes de silices de 3 microns, recouvertes d'environ 100 000 copies d'un oligonucléotide spécifique et d'un Tag de 29 bases permettant de les identifier. Ces billes sont localisées de manière aléatoire sur la puce. Chaque oligonucléotide est représenté par 20 à 30 billes. Ce processus a l'avantage de limiter les biais spatiaux de l'analyse.

L'évolution des puces, chez Affymetrix, s'est faite par la prise en compte de l'épissage alternatif dans le dessin des sondes. Après la puce Exon Array, contenant des sondes spécifiques

pour chaque exon, l'entreprise a sorti la puce Human Trandscriptome Array (HTA) qui possède des sondes au niveau des exons et des jonctions, permettant la détection des événements d'épissage alternatifs.

Séquençage ARNm

Nous allons parler de séquençage ADN car, y compris pour le séquençage de l'ARN messager, le produit final séquencé est une molécule d'ADN complémentaire (obtenue par rétro-transcription de l'ARN messager).

Depuis le développement de la première technique de séquençage en 1977 par Sanger (Sanger and Coulson, 1975) d'une part et Maxam et Gilbert d'autre part (Maxam and Gilbert, 1977), les technologies n'ont pas cessé d'évoluer. On parle de séquençage de première génération pour la méthode Sanger, qui a permis le séquençage du premier génome humain en 13 ans pour un coût total de 2,7 milliards de dollars (Consortium, 2004). La deuxième génération concerne les technologies dites « NGS » pour *Next Generation Sequencing* avec les technologies de séquençage haut débit permettant le séquençage de milliers de molécules ADN en une à deux semaines pour moins de 2500$. Les nouvelles générations de séquenceurs arrivent et permettent de séquencer une molécule unique pour 1000$ en quelques jours. Les différentes technologies sont présentées dans le Tableau 2 de la plus ancienne (Sanger) à la plus récente (nanopore).

La méthode de Sanger se base sur une étape d'amplification générant des fragments d'ADN de différentes longueurs et se terminant par un fluorochrome spécifique de chaque base. Après électrophorèse, les fragments se regroupent et s'ordonnent en fonction de leur longueur. La lecture des couleurs permet ainsi d'obtenir la séquence dans son ensemble.

Les méthodes de deuxième génération, apparues en 2005, ont été développées par différentes entités telles que Roche, Illumina ou Life Technologies mais se basent sur la même philosophie : amplification de l'ADN, succession de cycles de lavage et incorporation puis identification de nucléotides, que ce soit par fluorescence ou bien par pyroséquençage.

Ion Torrent procède de la même façon au niveau de l'amplification de l'ADN et des cycles, mais la détection du nucléotide se fait par la mesure d'ions hydrogène relargués lors de l'incorporation d'une base dans le brin d'ADN.

Les dernières méthodes ne nécessitent pas d'amplification, la molécule est « lue ». Chez Pacific Biosciences, la synthèse de l'ADN par l'ADN polymerase est observée en temps réel par fluorescence. Quant à la technologie nanopore, c'est un signal électrique qui permet d'obtenir l'information des 4 nucléotides incorporés.

Tableau 2 : Evolution des technologies pour le séquençage,
à partir de (Feng et al., 2015; Liu et al., 2012b; Quail et al., 2012)

	Technologie	Précision	Longueur des lectures (en bases)	Nombre de lectures par exécution	Temps / exécution	Coût pour 1 million de bases	Avantages	Inconvénients
Sanger 3730xl	Terminaison de chaine	99,9%	400 à 900	NA	20 mn à 3 h	2400 $	Qualité élevée, lectures longues	Prix
Illumina HiSeq 2000/2500	Par synthèse	98%	de 50 à 300	jusqu'à 6 millions	2 à 11 jours	0.05 à 0.15 $	Haut débit	Lectures courtes, Nécessite des concentrations élevées en matériel
Applied Biosystems Solid v4	Séquençage réversible par ligation, encodage de 2 bases	99,9%	50+35 ou 50+50	1,2 à 1,4 milliards	1 à 2 semaines (Single-end / Paired-end)	0.13 $	Peu cher	Plus lent que d'autres méthodes. Problèmes avec des séquences palindromiques
Roche 454 GS FLX	Pyroséquençage	99,9%	700	jusqu'à 6 millions	24h	10$	Lectures longues, rapide	Prix
Ion torrent PGM	Détection des ions H+ après réaction enzymatique (pH)	98%	jusqu'à 400	jusqu'à 5 millions	2h	1$	Prix, rapide	Erreurs dans les homopolymers
Pacific Biosciences PacBio RS	Observation directe de la synthèse de l'ADN en temps réel	87%	moyenne 15 000, jusqu'à 40 000	~55 000 par cellule	30 mn à 4h	0,13 à 0,60$	Lectures longues, rapides	Erreurs
Oxford nanopore Technology MinION	Mesure la translocation des nucléotides par mesure de concentrations ioniques	~80 %	5 000-10 000		48 à 72h		Lectures longues	Erreurs

L'évolution des technologies de séquençage va vers un séquençage de plus en plus rapide, bon marché et des appareils miniaturisés. Ainsi le MinION est sous forme de clé USB. Pour le moment, les dernières générations ne sont pas assez précises pour être utilisée en recherche et en clinique bien qu'une version beta du MinION soit actuellement en test à l'hôpital, mais l'amélioration des techniques devraient permettre d'augmenter la fiabilité des analyses sur ces

technologies. La technique de séquençage la plus utilisée en recherche aujourd'hui est l'HiSeq 2500 d'Illumina.

Puces d'expression ou séquençage ?

Les puces d'expression ont été la technologie la plus utilisée pour l'analyse transcriptomique dans les études à grande échelle. Néanmoins, le séquençage, du fait de la baisse des prix, remplace de plus en plus les puces. La question est de savoir quelles sont les avantages et les limites de ces deux technologies.

Un premier sujet de comparaison porte sur la gamme dynamique d'expression détectée. Dans une cellule, les ARNm peuvent être présents à raison de quelques copies jusqu'à plus de 10 000 copies, ce qui représente une gamme importante.

Zhao et al ont comparé des données de séquençage (HiSeq 2000) et des données de puces (Affymetrix HT-HG U133) sur des cellules T dans différentes conditions. La corrélation entre les deux technologies est plutôt bonne, autour de 0.89 selon les conditions. Cependant, le graphique représentant les données d'une technologie en fonction de l'autre ne semble pas uniforme (voir Figure 15A). Ceci est dû aux différences de gamme dynamique, qui dans le cas du séquençage peut varier en fonction de la profondeur de séquençage. Dans le cas d'un séquençage avec 50 millions de séquences alignées (minimum de profondeur dans les analyses récentes), la gamme dynamique est supérieure à 10^5, ce qui est supérieure à celle des puces (10^3-10^4), comme représenté Figure 15B. Les puces sont saturées lorsque l'expression d'ARNm est trop importante.

Afin de vérifier la précision de l'expression des gènes, et notamment des gènes différentiellement exprimés, on utilise une analyse quantitative de *Polymerase Chain Reaction* (qPCR). C'est ce qu'ont fait Wang et al en comparant les gènes différentiellement exprimés détectés avec les puces d'expression ou le séquençage (Wang et al., 2014). D'après la Figure 16A, lorsque le niveau d'expression des gènes est important (supérieur à la médiane), les gènes détectés comme différentiellement exprimés à partir des données de puces ou de séquençage sont validés en qPCR, alors que cette validation est très faible pour les gènes avec un niveau d'expression plus faible détectés par les puces. La corrélation du *fold change (*ratio d'expression entre deux conditions) entre qPCR et séquençage est meilleure (0.97) par rapport à celle entre qPCR et puce (0.85) (Figure 16B).

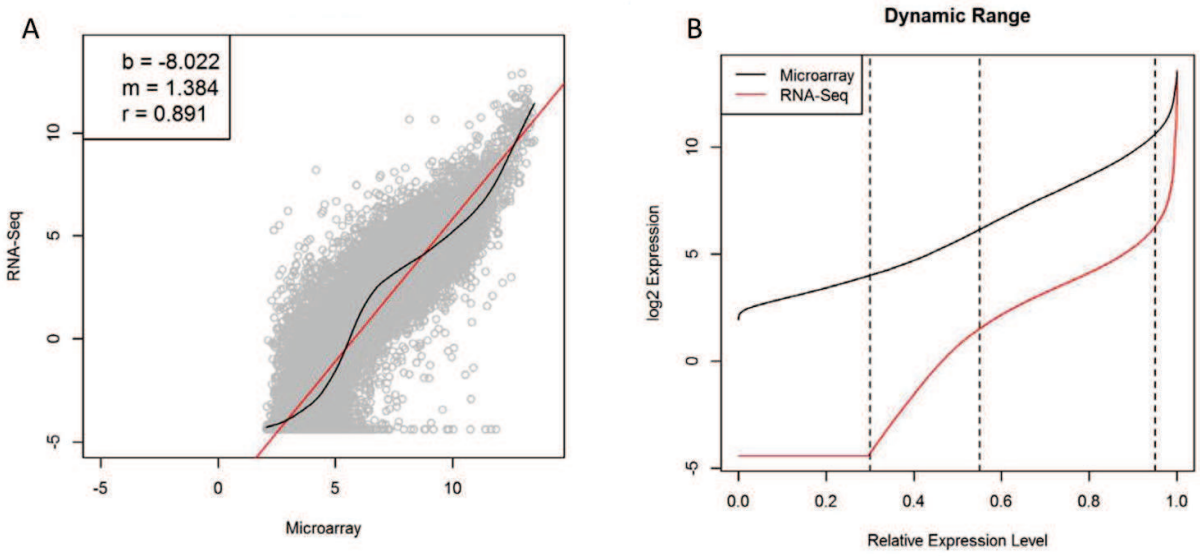

Figure 15 : Comparaison des données de puce et séquençage.
A) Expression transformée en log2 entre les deux technologies, la régression linéaire entre les deux profils d'expression est représentée en rouge et le lissage par spline en noir. B) Gamme dynamique dans les deux technologies représentée par l'expression en log2 des gènes en fonction de leur niveau d'expression relatif. D'après (Zhao et al., 2014).

L'une des limites des puces d'expression est le niveau de bruit. Il n'existe pas de niveau d'expression nul, ce qui rend difficile l'évaluation des gènes peu exprimés.

Le séquençage fournit davantage d'informations qu'une puce. En effet, en plus du niveau d'expression des gènes, on peut obtenir des informations sur les évènements d'épissage alternatifs, qui, même si ces informations commencent à être accessibles avec les nouvelles puces, sont plus limitées ; ainsi que des informations sur les SNPs (*Single Nucleotide Polymorphisms*) à partir de la lecture des séquences de l'ARN.

Malgré ces avantages, le séquençage a encore quelques limites. Pendant longtemps la question du coût était la principale limite, mais il est aujourd'hui presque aussi couteux de faire une analyse de puce que de séquençage. Néanmoins, le recul sur le séquençage est moins important que sur les puces d'expression, les processus d'analyse ne sont pas encore bien définis et restent coûteux en terme de ressources informatiques (puissance et mémoire).

Aujourd'hui, pour une analyse simple et rapide, la puce d'expression reste une meilleure alternative, mais le développement des analyses bioinformatiques et la généralisation du séquençage devraient limiter son intérêt.

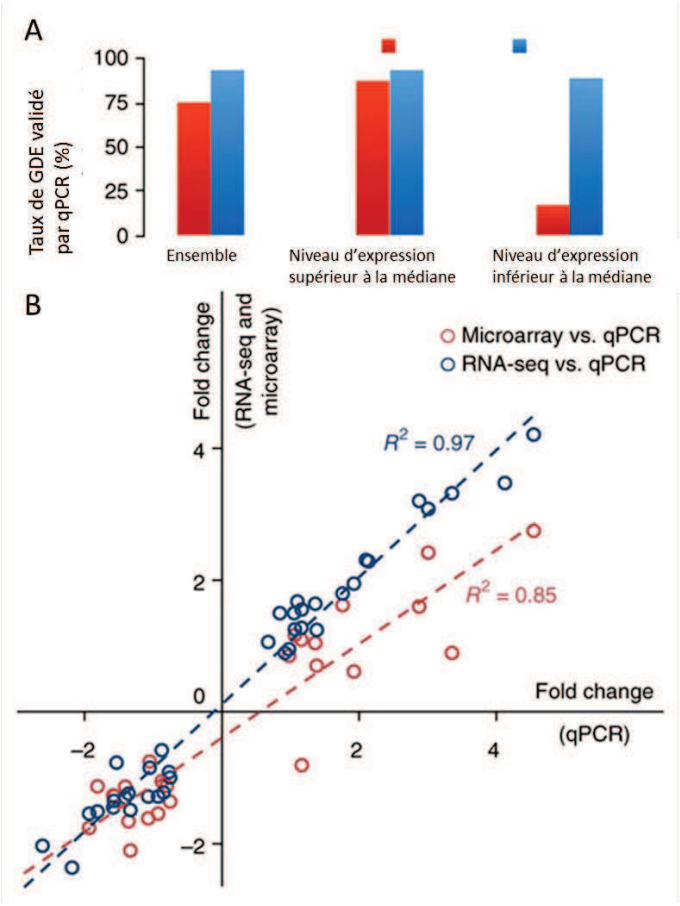

Figure 16 ***: D'après (Wang et al., 2014)*** **, gènes différentiellement exprimés (GDE) détectés.**
A) GDE validés par qPCR, détectés par puce en rouge et par séquençage en bleu, selon leur niveau d'expression. B) Fold Change d'expression sur puce ou séquençage en fonction du Fold Change en qPCR.

1.3.1.2. Protéomique

Pour l'analyse des protéines aussi il existe deux techniques semblable aux puces et séquençage de la transcriptomique : la technique RPPA (*Reverse Phase Protein Array*) qui consiste en l'utilisation d'une puce à protéines, où des anticorps fluorescents permettent d'évaluer la présence de protéines pour un nombre important d'échantillons ; l'utilisation de la spectrométrie de masse pour pouvoir caractériser un grand nombre de protéines présentes dans un échantillon. La technique RPPA est limitée par le nombre d'anticorps validés, capables de reconnaitre spécifiquement une protéine. Nous n'avons étudié que des données de spectrométrie de masse, c'est donc la seule technique que nous allons aborder dans la suite de cette partie.

Les débuts de la protéomique remontent à la mise au point de l'électrophorèse bidimensionnelle sur gel de polyacrylamide ou 2D-PAGE en 1975 (O'Farrell, 1975). Depuis, les différents éléments formant le processus d'une analyse de spectrométrie de masse actuelle se sont développés.

Le nombre de protéines présentes dans un échantillon humain s'avère trop conséquent à analyser tel quel, surtout si nous nous intéressons aux protéines les moins abondantes. Pour solutionner ce problème, il faut simplifier les mélanges avant l'analyse par spectromètre de masse, en effectuant une séparation des protéines à analyser. Cette séparation peut se faire par exemple par l'utilisation d'un gel d'électrophorèse SDS-PAGE, qui consiste à faire migrer les protéines dans un gel, sous l'influence d'un champ électrique, permettant ainsi la séparation des protéines en fonction de leur poids moléculaire. Le découpage du gel en plusieurs bandes permet ensuite d'analyser un sous-ensemble de ces protéines à la fois. Une fois séparées, les protéines sont réduite, alkylées et digérées par la trypsine afin d'obtenir des peptides de masse plus faible, et plus facilement analysables.

L'analyse de spectrométrie de masse est une analyse LC-MS/MS pour « Liquid chromatography coupled to tandem mass spectrometry ». Les peptides sont dans un premier temps séparés par chromatographie liquide en phase inverse suivant leur hydrophobicité, puis ils sont envoyés dans la source du spectromètre de masse où ils sont ionisés. Dans un deuxième temps, les ions sont séparés et identifiés en fonction du rapport masse/charge de l'ion, il s'agit de l'analyse MS simple. L'appareil isole ensuite les ions les plus intenses, sélectionne l'un des ions, le fragmente partiellement et expulse les ions fils. C'est la mesure des masses des ions fils qui va nous permettre de remonter aux peptides et protéines.

L'identification des protéines est faite en comparant les listes de masses expérimentales obtenues en sortie du spectromètre avec les listes obtenues par digestion in-silico de toutes les protéines contenues dans la banque *Swissprot*, ce qui permet d'identifier la protéine obtenue.

La dernière étape est l'étape de quantification. Il existe deux approches différentes :

- une approche de type marquage (comme SILAC) consistant à marquer deux échantillons différents avant d'être mélangés et analysés en LC-MS/MS. On peut ainsi obtenir un ratio de signal entre les deux échantillons et supprimer les biais techniques, mais le nombre d'échantillons pouvant être analysés simultanément est limité.
- une approche de type 'label-free' sans marquage. On distingue trois méthodes de quantification différentes : le compte du nombre de peptides ayant permis d'identifier la protéine, le compte du nombre de spectres MS/MS générés par les peptides de la protéine ou bien l'extraction des courants d'ions (méthode XIC) qui consiste à intégrer les pics de chaque peptide. La dernière méthode est plus sensible mais plus lourde à analyser.

Au niveau historique, ces différents éléments ont été développés au cours du temps : le premier séquençage par MS/MS date de 1986, le développement des techniques d'ionisation des

protéines de 1988. La complexité des mélanges protéiques a nécessité le développement de techniques de plus en plus précises pour pouvoir analyser un échantillon humain complexe (Mann et al., 2013).

1.3.1.3. Corrélations entre Transcriptomique et Protéomique

Le dogme central de la biologie moléculaire, tel que formulé par Francis Crick en 1958 est que l'ADN est transcrit en ARN messager, n'ayant qu'une vie temporaire, qui est lui-même traduit en protéine. La réalité est plus complexe car de nombreux évènements de régulation interviennent au cours du processus. Ces évènements sont résumés sur la Figure 17. En particulier, le niveau d'expression des protéines est régulé par différents mécanismes, ce qui implique que l'analyse de l'expression des ARN messagers ne permet pas d'accéder au niveau d'expression des protéines, qui est à l'origine du phénotype cellulaire.

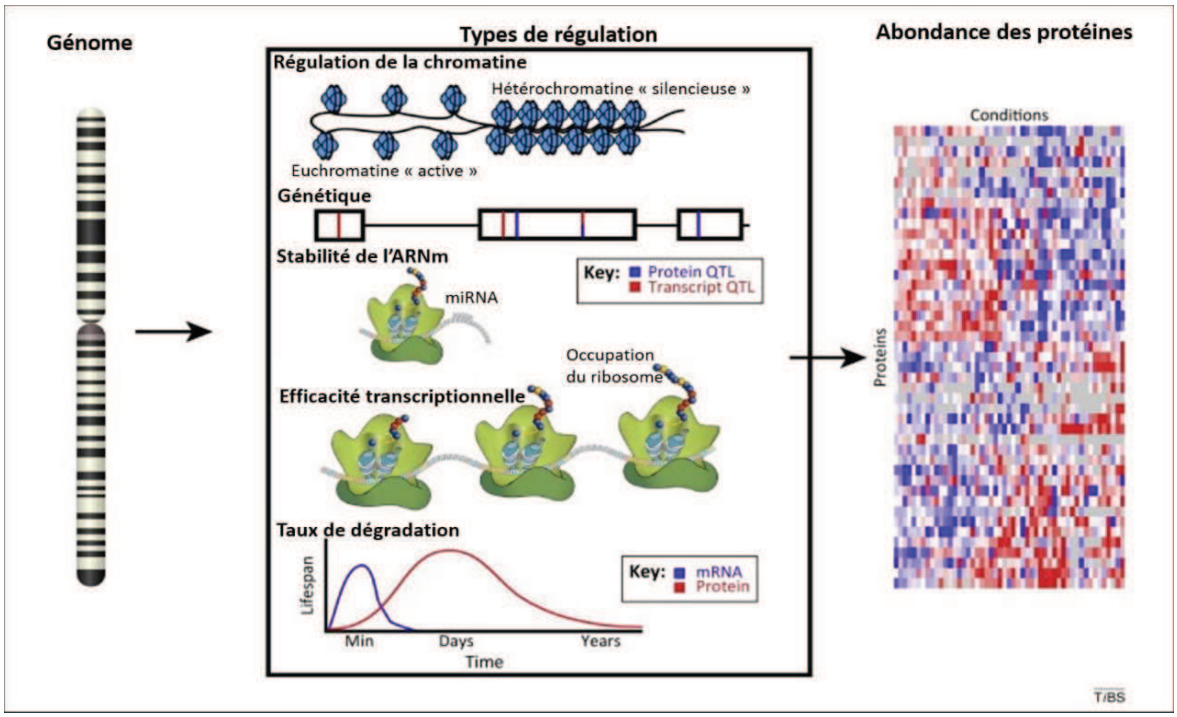

Figure 17 : Diversité des processus de régulation entre l'ARNm et les proteins

Source : d'après (Payne, 2015)

Une étude publiée par Wu et al dans Nature (Wu et al., 2013) a montré que sur l'étude d'une cohorte de 95 individus, les loci[10] contrôlant l'expression des ARNs messagers ne sont communs qu'à 50% des loci contrôlant l'expression des protéines. L'ARN messager peut être affecté par des micro-ARNs (ou miARNs), qui en s'hybridant spécifiquement sur les ARNs

[10] Emplacement physique sur un chromosome

messagers peuvent limiter leur traduction ou bien provoquer la dégradation des ARNs (Baek et al., 2008).

La régulation de l'abondance des protéines passe aussi par la régulation du mécanisme de la traduction. Enfin les taux de synthèse et de dégradation, ainsi que les durées de vie relatives des ARNm et des protéines ne sont pas à la même échelle : quand l'ARN messager a une durée de vie de quelques minutes, celle de la protéine peut être de quelques heures à quelques années. Ces différents événements limitent la corrélation entre l'abondance des ARNs messagers et l'abondance des protéines. Avec le développement des technologies multi-omiques, il est intéressant d'analyser à différents niveaux pour comprendre les événements de régulation. Dans une récente étude portant sur l'analyse de 90 tumeurs de colons, Zhang et al. ont intégré des données de protéomique, génomique et transcriptomique (Zhang et al., 2014). Ils ont notamment montré que l'impact du nombre de copies des gènes est plus corrélé avec l'expression de l'ARN qu'avec l'expression des protéines, ce qui montre l'importance des événements de régulation post-transcriptionnel dans le maintien du phénotype. Ils ont aussi analysé la corrélation entre l'abondance des ARNm et des protéines : dans chaque échantillon la corrélation est positive avec une moyenne à 0.47.

L'intégration des différents « omiques » est nécessaire pour pouvoir comprendre le fonctionnement cellulaire au niveau global.

1.3.2 Analyse des données

1.3.2.1. Processus général

Quel que soit le type de données à analyser (puces d'expression, séquençage ARNm, protéomique), les grandes étapes d'analyse sont les mêmes (Figure 18).

A la sortie de l'instrument effectuant le processus d'analyse biologique, on obtient un fichier numérique à analyser. Les premières étapes du processus peuvent être effectuées par les logiciels vendus avec l'instrument, il s'agit de lire les données dans leur format spécifique et d'effectuer les premières étapes de prétraitement permettant de passer de l'analyse d'un signal à une variable quantitative. L'étape suivante consiste à vérifier la qualité des données et à normaliser l'ensemble des données à analyser. L'objectif est de supprimer/diminuer les biais entre les échantillons. On peut ensuite passer à l'étape d'analyse des données, puis enfin la visualisation des résultats. L'analyse des données peut se faire de façon non supervisée, c'est-à-dire sans à priori, sans utilisation d'annotations sur les échantillons pour diriger les analyses statistiques ou bien de façon supervisée.

Nous allons développer les étapes de prétraitement et contrôle qualité dans la suite.

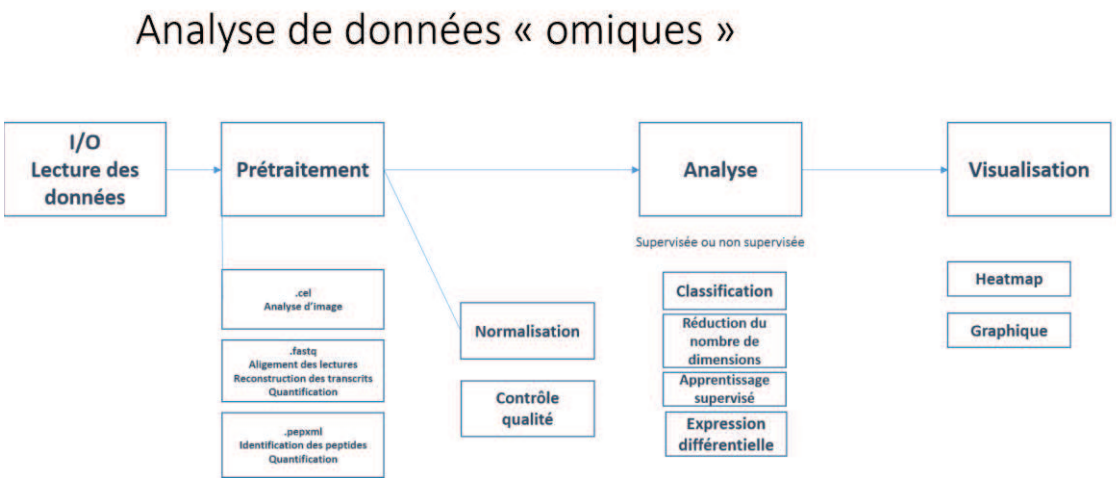

Figure 18 : Processus d'analyse de données "omiques"

1.3.2.2. Prétraitement

Cette première étape diffère selon les technologies « omiques » étudiées.

Pour les puces, il s'agit en général d'analyser une image de points lumineux afin d'extraire la quantité de signal proportionnel à l'expression de ces points et de faire la correspondance entre des probesets (et donc des gènes) et cette information.

Pour le séquençage, le fichier de départ est un fichier contenant plusieurs millions de séquences, ainsi que des informations sur la qualité des lectures de ces séquences. Les étapes de prétraitement vont dépendre de l'analyse effectuée. Si on s'intéresse à une analyse de niveau d'expression des gènes, le prétraitement comporte deux étapes principales : l'alignement des lectures sur un génome de référence, puis la reconstruction des transcrits exprimés et l'assignement des lectures aux transcrits/gènes permettant d'obtenir une valeur d'expression par transcrits (étape de quantification) ou bien par gènes ; ou bien l'assemblage de novo des transcrits puis l'alignement sur le génome de référence, bien que cette solution soit peu utilisée (Figure 19).

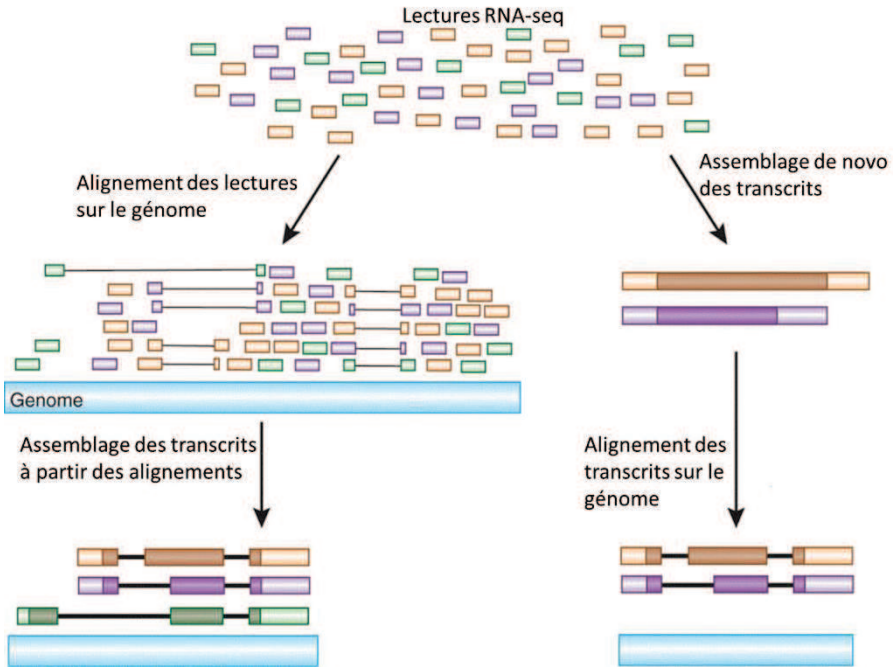

Figure 19: Deux stratégies possibles pour l'alignement des lectures, alignement sur le génome et reconstruction des transcrits ou assemblage des transcrits, puis alignement sur le génome
Source : d'après (Haas and Zody, 2010)

Pour ces deux étapes, de nombreux programmes ont été développés. Afin de choisir les plus efficaces, plusieurs groupes ont évalué les principaux logiciels et ont publié leur résultat. Ainsi, concernant l'étape d'alignement, le RGASP consortium (*RNA-seq Genome Annotation Assessment Project*) a publié un article en 2013 qui compare les différents outils sur des données Illumina (Engström et al., 2013). Ils en concluent que les meilleurs programmes d'alignements semblent être GSNAP (Wu and Nacu, 2010), GSTRUCT, MapSplice (Wang et al., 2010a) et STAR (Dobin et al., 2013). Néanmoins STAR est beaucoup plus rapide (30 à 40 fois plus rapide que Tophat2 (Dobin et al., 2013)) (Williams et al., 2014).

Le calcul du nombre de lectures par gène ou exon ne présente pas de difficulté particulière mais au niveau de l'isoforme, l'analyse pose davantage de problèmes. En effet, le comptage simple du nombre de lectures alignées sur chaque isoforme surestime l'expression de l'isoforme lorsque plusieurs isoformes sont présents (comptage double des lectures alignées sur les exons communs). Pour optimiser cette analyse, les algorithmes doivent distribuer les lectures entre les isoformes. Comme précédemment, une publication présente les résultats de la comparaison de différents outils (Williams et al., 2014) : Cufflinks (Trapnell et al., 2012), RSEM (Li and Dewey, 2011), TIGAR (Nariai et al., 2013) et MISO (Katz et al., 2010). Les auteurs ont calculé les corrélations de quantification des isoformes entre deux réplicas. Ils concluent que Cufflinks et RSEM présentent des résultats légèrement supérieurs aux autres algorithmes.

Concernant l'analyse protéomique, à la sortie du spectromètre de masse, on obtient des spectres MS/MS. Les étapes d'analyses sont l'identification des protéines à partir de ces spectres, puis la validation des protéines identifiées et enfin la quantification. Les étapes sont très similaires à l'analyse de séquençage ARN. Les spectres MS/MS sont d'abord interprétés en terme de listes de masses et de précurseurs associés, qui sont ensuite comparés aux banques protéiques afin d'identifier les peptides et protéines associés. Cette étape d'analyse peut être effectuée par le logiciel *Mascot* (Matrix Science). La validation des protéines identifiées est nécessaire pour limiter le taux de faux positifs et est recommandée. Plusieurs approches sont possibles, dont l'approche « Target-Decoy »(Elias and Gygi, 2010) ou « Cible-Leurre ». L'idée est de réitérer l'étape précédente d'identification dans une banque protéique classique (Cible) combinée à une banque contenant les séquences inversées (Leurre). Les faux positifs correspondent au nombre de séquences assignées dans la banque « Leurre ». Cette analyse peut se faire à l'aide du logiciel *Scaffold* (Proteome Software).

La dernière étape de l'analyse est la quantification. En nous plaçant dans une étude de type « Label-free » sans marquage, deux stratégies sont possibles : sommer le nombre de spectres MS/MS par peptide en partant du principe que plus le peptide est présent dans l'échantillon plus le nombre de spectres est élevé (Bantscheff et al., 2012) ou bien extraire les courants d'ions sur la trace MS, ce qui consiste à intégrer l'aire sous les pics des peptides sur la trace MS. Le logiciel *Skyline* (MacCoss Lab) effectue cette partie de l'analyse. Néanmoins, une validation manuelle peut être nécessaire car le logiciel n'identifie pas toujours le bon pic à intégrer.

1.3.2.3. Contrôle qualité

Le contrôle qualité est une étape cruciale qui peut se faire avant/pendant ou après les étapes de prétraitements selon la variable contrôlée et le type de données analysées.

Dans le cas des puces, il s'agit principalement de vérifier si les annotations des échantillons correspondent bien et s'il n'y a pas de biais technique entre les échantillons (voir partie 1.4). Pour cela, on peut vérifier si les échantillons de même type (même tissu, mêmes conditions expérimentales) sont proches entre eux et éloignés de ceux ayant une annotation différente. Cette analyse peut se faire par calcul de distance ou corrélation entre les échantillons ou bien par des méthodes de classification non supervisée ou de réduction de dimension (telle que l'analyse en composante principale) permettant de visualiser en deux ou trois dimensions la répartition des échantillons.

Dans le cas du séquençage, le contrôle qualité se situe à différentes étapes du prétraitement. La première étape consiste à vérifier la qualité des lectures lues par le séquenceur. En utilisant

le séquenceur Illumina, un score de qualité est associé à chaque nucléotide lu (*Q phred*). Ce score correspond à une probabilité d'erreur de séquençage (Ewing and Green, 1998). La valeur minimale pour considérer le nucléotide comme correct est généralement fixée à 30, ce qui correspond à une probabilité d'erreur de 1/1000. Les erreurs de séquençage ne sont pas uniformément réparties le long des lectures mais sont de plus en plus importantes le long de la lecture de la séquence.

L'outil majoritairement utilisé pour visualiser la qualité des lectures à l'échelle de l'échantillon est FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Cet outil fournit un rapport html constitué de plusieurs graphiques, comprenant notamment, la qualité des séquences par base sous forme de boxplot, la répartition du nombre de séquences par score et le nombre de nucléotides indéterminé par base. Il permet d'avoir une idée rapide de la qualité des séquences par échantillon.

La qualité des séquences se dégradant au fur et à mesure de la lecture, différents outils sont utilisés pour supprimer les nucléotides probablement mal lus, on appelle cette étape le *trimming*. Les logiciels comme ConDe Tri, Trimmomatic, Cutadapt (Bolger et al., 2014; Martin, 2011; Smeds and Künstner, 2011) effectuent cette étape. La comparaison des différents outils et des seuils de Q-phred utilisés (Del Fabbro et al., 2013) montre que les résultats biologiques diffèrent entre les différentes méthodes. Une autre possibilité est de ne pas appliquer de *trimming* aux données. Les séquences de mauvaises qualités ne seront pas alignées sur le génome si les paramètres des algorithmes d'alignement ne le permettent pas.

Une deuxième étape de contrôle qualité consiste à vérifier le taux de lectures alignées sur le génome. Ce résultat est fourni par l'outil utilisé pour l'alignement. Ce taux de lecture est normalement proportionnel à la qualité des séquences. Il s'agit donc plutôt de vérifier que l'alignement s'est bien déroulé, qu'il n'y pas eu d'erreur dans le génome de référence utilisé.

Dans le cas de la protéomique, le contrôle qualité se déroule aussi au cours des étapes de prétraitement. Le calcul du taux de faux positif par l'utilisation d'une méthode de « Cible-Leurre » peut être considéré comme une étape de contrôle qualité, de même que la vérification manuelle de l'intégration des pics du logiciel *Skyline* (MacCoss Lab) lors de la quantification *XIC.*

1.3.2.4. Normalisation

L'étape de normalisation permet de diminuer les différents biais (voir partie 1.4) entre échantillons que l'on voudrait comparer. Les méthodes sont spécifiques des types de données.

Dans le cas des puces, les algorithmes dépendent du type de puce utilisé. Pour les puces Affymetrix (technologie majoritaire), les plus utilisés sont MAS 5.0, l'algorithme développé par

Affymetrix (Hubbell et al., 2002) mais surtout RMA (*Robust Multi-array Average*) (Irizarry et al., 2003) et frozenRMA (McCall et al., 2010). Dans l'algorithme RMA, le bruit de fond est calculé de manière globale sur la puce et est soustrait à l'ensemble des probes. Les données sont normalisées par quantile, puis l'algorithme « median polish » est appliqué à chaque probeset pour prendre en compte le fait que l'affinité des probes et leur variance devrait être constant pour un même probeset. Les données sont ensuite transformées en log2 et le résultat final est donné par probeset. RMA nécessite l'analyse simultanée de toutes les puces. Ceci limite son utilisation, car, à chaque ajout de données, il faut relancer l'algorithme. C'est pourquoi frozenRMA a été développé. Il fonctionne de la même manière que RMA mais utilise des données externes pour pré-calculer l'affinité des probesets et leur variance, et ainsi utiliser ces résultats « gelés ». Il devient possible d'ajouter des puces au fur et à mesure de l'analyse en ne lançant l'algorithme que sur ces dernières données.

Dans le cas du séquençage de l'ARN, deux biais sont à prendre en compte : la longueur des transcrits dû à la quantification par addition des lectures alignées sur un transcrit, la profondeur de séquençage de la librairie (nombre total de séquences lu pour un échantillon) qui peut être différente entre les échantillons. La longueur des transcrits n'est pas forcément un biais limitant selon l'objectif de l'analyse. En effet, dans la plupart des expériences de séquençage ARN, l'objectif est de comparer un même transcrit entre échantillons.

Le deuxième biais est dû à la profondeur de séquençage de la librairie, la quantité de lectures séquencées par échantillon peut varier.

Pour répondre à ces différents biais, des méthodes de normalisation ont été développées. Elles se basent sur des hypothèses et des principes différents :

- Ajustement de distribution : on part du principe que le nombre de lectures est proportionnel au niveau d'expression et à la profondeur de séquençage et on définit un facteur d'échelle pour chaque échantillon.

  Méthode de calcul du facteur d'échelle : Nombre total de lecture : TC (Marioni et al., 2008), Quartile supérieur : UQ (Bullard et al., 2010), Mediane
- Méthode prenant en compte la taille du fragment : RPKM (nb de Reads Per Kilobase per Million mapped reads) (Mortazavi et al., 2008).

$$RPKM = \frac{\frac{Nombre\ de\ lectures\ sur\ le\ transcrit}{Taille\ du\ transcript\ \times 10^3}}{Nombre\ total\ de\ lectures\ \times\ 10^6}$$

Il a été montré que le RPKM introduit un biais dans la variance du gène, ce qui peut être très gênant pour des analyses de type classification (Oshlack and Wakefield, 2009).

- Méthode prenant en compte la profondeur de séquençage de la librairie et partant de l'hypothèse que peu de gènes sont différentiellement exprimés : DESeq (Anders and Huber, 2010), TMM (Trimmed Mean of M-values) (Robinson and Oshlack, 2010), disponibles dans le package edgeR.
- Voom (Law et al., 2014), disponible dans le package limma : estime la relation moyenne-variance et fournit un poids de précision pour chaque observation.

En 2012 le *French StatOmique Consortium* a publié un papier comparant différentes méthodes de normalisation, dont TC, UQ, Médiane, DESeq, TMM, l'utilisation du RPKM, ainsi que d'autres méthodes non décrites ici. Ils en concluent que le RPKM et l'utilisation du Total Count (le nombre de lectures est divisé par la taille de la librairie et multiplié par la taille moyenne de la librairie de l'ensemble des échantillons) sont inefficaces pour prendre en compte la variation entre échantillons et les différences de taille de librairie ; et que seuls DESeq et TMM sont robustes aux variations typiques d'une expérience de séquençage ARN. La méthode Voom n'a pas été testée dans cette analyse.

En ce qui concerne les données de protéomique, contrairement aux analyses transcriptomiques, la nécessité de la normalisation des données, et le type de normalisation à utiliser ne semble pas encore fixé. Dans le cas d'une analyse de type « comptage de spectres », dans l'analyse protéogénomique du cancer du colon (Zhang et al., 2014), les auteurs ont comparé différentes méthodes de normalisation : pas de normalisation, des méthodes empruntées aux analyses de puce (méthode globale et quantile) et une méthode développée pour la protéomique NSAF (*Normalized Spectal Abundance Factor*). Ils ont montré que pour l'analyse de classes, les méthodes des puces étaient plus efficaces, avec la méthode quantile ayant des résultats légèrement supérieurs.

1.3.2.5. Analyse des données

L'analyse des données dépend de la question biologique posée et de la taille du jeu de données. Partons d'un ensemble de données suffisamment important, les étapes d'une analyse seraient les suivantes:

- Description de données sans à priori sur les différences entre échantillons : analyse non supervisée
    - Réduction de dimension des données : par exemple, l'Analyse en Composantes Principales (ou ACP) ayant pour objectif de revenir à un espace de dimension réduite en déformant le moins possible la réalité. Permet d'étudier l'homogénéité ou l'hétérogénéité des données dans un espace « visualisable », c'est-à-dire en

deux ou trois dimensions, ainsi que les variables principales participant à la variabilité entre échantillons.

- Classification : pour étudier la division en classes homogènes de l'ensemble des échantillons.

- Identification des caractéristiques d'une classe
  - Analyse différentielle : sur les variables (gènes, protéines) ou sur des ensembles des variables (voies de signalisation)
- Classement d'une tumeur parmi des classes connues, c'est-à-dire prédire le diagnostic ou pronostic de nouveaux patients s'il existe une classification
  - Analyse supervisée

1.4 Méta-analyse et bruit

1.4.1 Introduction générale : expérience « omique »

Les étapes entre le patient et les données « omiques » prêtes à être analysées sont nombreuses (voir Figure 20). Chacune de ces étapes, du fait de l'utilisation de technologies différentes par des manipulateurs différents et du fait de l'erreur humaine est susceptible de générer un bruit supplémentaire. Nous allons voir dans la suite les différentes sources de bruit et leur conséquence, ainsi que la manière de les évaluer et de les prendre en compte dans l'analyse de données.

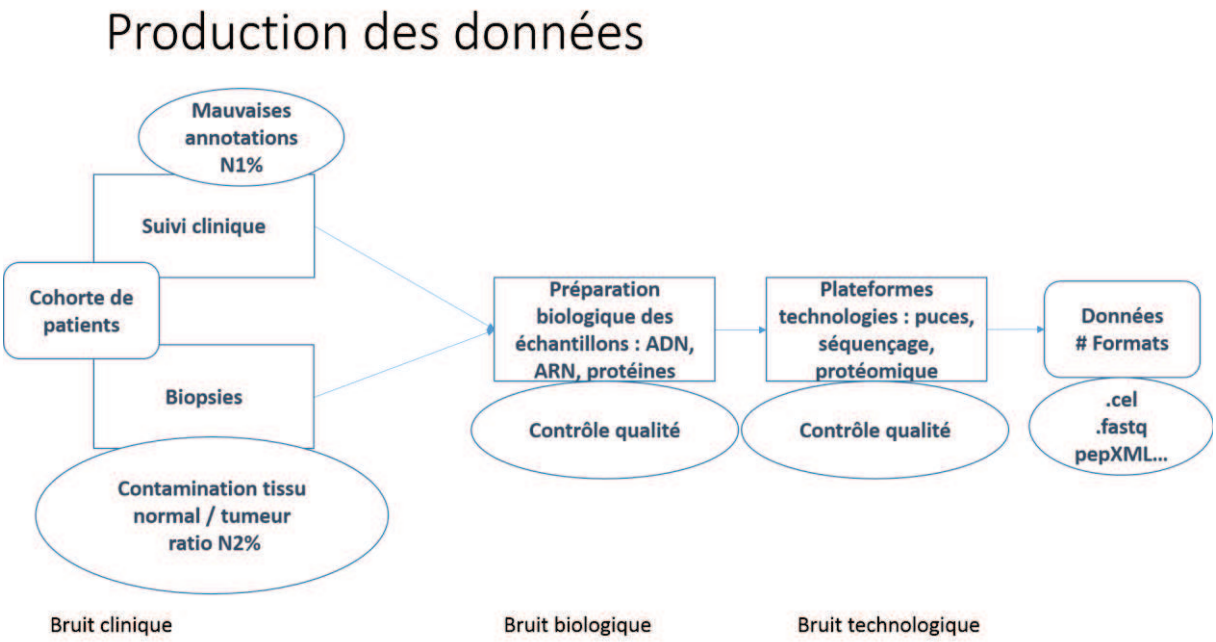

Figure 20 : Etapes de production de données "omiques"

1.4.2 Détail des sources de bruits

Comme vu dans Figure 20, une analyse de « données omiques » peut se diviser en trois phases : préparation de l'échantillon (biopsie de la tumeur sur le patient, récolte des informations cliniques par le médecin), extraction et préparation biologique des échantillons, analyse sur la plateforme technologique d'intérêt. Chacune de ces étapes est source de bruit.

Concernant la première étape, on parle de bruit clinique. Il inclut la qualité de l'échantillon issu de la biopsie. En effet, quel que soit l'échantillon d'intérêt (tumeur, tissu), il est important que la biopsie soit le moins possible contaminée par d'autres types de cellules qui masqueraient le signal que l'on cherche à obtenir. Le bruit clinique inclut aussi la qualité des informations cliniques qui vont suivre l'échantillon. Les problèmes de bruit peuvent venir d'informations non exactes ou à jour, voire d'informations n'appartenant pas au bon patient.

Concernant la seconde étape, on parle de bruit biologique. L'étape de préparation des échantillons et d'extraction de la matière d'intérêt (ADN, ARN, protéines) peut se faire selon différents protocoles, susceptibles de modifier le signal de sortie ; de plus, des manipulateurs différents peuvent générer des résultats différents. Le bruit biologique pose problème lors des méta-analyses, c'est-à-dire lorsque l'on souhaite intégrer des données de différentes provenances. Le résultat peut être une plus grande ressemblance sur les données d'échantillons traités au même laboratoire que sur ceux partageant les mêmes caractéristiques biologiques.

Concernant la dernière étape, on parle de bruit technologique. Il s'agit notamment de la difficulté de comparer des données issues de plateformes technologiques différentes. Il est par exemple très difficile de comparer des données de différents types de puces car elles n'ont pas toutes la même gamme dynamique, les probesets ne sont pas forcément les mêmes pour un gène, ce qui peut modifier l'affinité des ARNm pour ces probesets, et le niveau d'expression établi pour une puce donnée.

Dans la situation idéale, l'ensemble des données étudiées viendrait d'un même chirurgien, ayant fourni le suivi clinique de ces patients, extrait par le même manipulateur et analysé sur la même plateforme, avec un contrôle qualité effectué à chaque étape (contrôle du pourcentage de cellules tumorales de la biopsie par l'anatomo-pathologiste, contrôle de la qualité des ADNs/ARNs/protéines par le biologiste, contrôle de la qualité des données issues de la plateforme technologique par le bio-informaticien).

La situation n'est pas toujours idéal et il faut vérifier que l'ensemble des données n'est pas trop bruité, voire supprimer les données qui posent problèmes.

1.4.3 Prise en compte du bruit

Les différentes sources de bruit peuvent être analysées et traitées à plusieurs niveaux.

Concernant le bruit technologique, le mieux est de traiter ces données de manière indépendante. Concernant les puces par exemple, puisque les gammes dynamiques ne sont pas les mêmes, et que la définition des gènes n'est pas tout à fait identique non plus (différents *probesets*), il est plus simple de travailler par plateforme technologique, en faisant les mêmes analyses sur les différentes plateformes pour vérifier que l'on retrouve bien les mêmes résultats biologiques.

Concernant le bruit clinique et biologique provenant de l'analyse sur une même plateforme, la première étape consite à vérifier l'existence d'un effet *batch*, c'est-à-dire une situation dans laquelle la variabilité des échantillons est davantage due à une variabilité technique entre les jeux de données que biologique. Les mêmes échantillons biologiques venant de différents ensembles de données vont être plus différents que des échantillons biologiquement éloignés mais venant d'un même ensemble de données. Cet effet *batch* peut venir du bruit accumulé lors des différentes étapes d'analyse. Pour évaluer ce bruit, on peut vérifier la proximité entre échantillons et regarder si les regroupements sont dus à la nature biologique des échantillons ou à d'autres variables (http://www.molmine.com/magma/global_analysis/batch_effect.html).

Différentes méthodes permettent d'estimer la proximité des échantillons : visuellement on peut faire une analyse en composante principale (ACP) ou bien une analyse de type classification hiérarchique. En ajoutant les annotations biologiques et/ou les sources de bruit sur le graphique, il est possible de se rendre compte de l'existence d'un effet *batch*. Ce type d'analyse permet aussi de vérifier le bruit issu d'erreurs d'échantillons. En effet un échantillon associé à une variable clinique très éloigné de tous les autres échantillons ayant la même variable est très certainement mal annoté. Il est aussi possible d'utiliser d'autres méthodes basées sur l'analyse de proximité des échantillons, en calculant par exemple les distances entres échantillons ou entre barycentres de groupes d'échantillons biologiquement semblables et les échantillons ou en utilisant la corrélation à la place de la distance euclidienne.

Principe de l'ACP : l'objectif est de réduire la dimension des données initiales (environ 40 000 *probesets* sur la puce HG U133 Plus 2 par exemple) en un nombre beaucoup plus faible de facteurs. Ces facteurs sont des moyennes pondérées des variables initiales, choisis pour maximiser la dispersion entre échantillons ; c'est-à-dire que les facteurs sont choisis en maximisant la variance). On obtient ainsi deux ou trois facteurs, représentant au mieux la dispersion inter-échantillons. Il devient possible de visualiser ces deux ou trois facteurs.

Principe du *clustering* hiérarchique (ou classification) : l'objectif est de classifier les échantillons ayant un comportement similaire. Pour cela on définit deux méthodes : une distance (distance euclidienne, corrélation ou autre) et une méthode d'agrégation des données (simple lien : on prend le minimum entre deux points de chaque classe, lien complet : on prend le maximum

entre deux points de chaque classe, moyenne, ward : permet de trouver à chaque étape l'agrégation qui minimise la perte d'information). Le principe est de chercher les deux points les plus proches selon la distance et de les regrouper dans une classe ; puis de remplacer ces points par leur centre. On recommence cette étape de manière itérative en utilisant une méthode d'agrégation à chaque calcul de la distance entre deux classes. Le résultat est un arbre, coupable à différents niveaux, selon la taille et le nombre de classes recherchées.

En cas d'*outliers* (ou échantillons très éloignés des autres échantillons ayant les mêmes propriétés biologiques), il est plus simple de ne pas les prendre en compte pour la suite des analyses. Si ce n'est pas possible, il faut bien garder à l'esprit que ces échantillons sont potentiellement différents de l'annotation biologique indiquée.

Concernant les problèmes d'effet *batch*, il est possible de laisser les données ainsi et de les analyser sans oublier le problème, mais il existe aussi des algorithmes permettant de supprimer la variabilité technique. Les plus utilisés et les plus efficaces (Chen et al., 2011) sont Combat (pour *Combating Batch Effects When Combining Batches of Gene Expression Microarray*) (Johnson et al., 2007), basé sur une estimation empirique bayesienne des paramètres de localisation et d'échelle à ajuster pour chaque *batch* pour chaque gène indépendamment ; et SVA (Surrogate variable analysis) (Leek and Storey, 2007), qui combine la décomposition en valeurs singulière et un modèle d'analyse linéaire pour estimer les vecteurs propres d'une matrice d'expression résiduelle après avoir enlevé la variation biologique.

En tentant de réduire ainsi les effets *batch*, les données sont modifiées, ce qui peut limiter la découverte de gènes dont l'expression est faiblement modifiée entre les groupes biologiques.

Objectifs

Ce chapitre introductif laisse en suspens un certain nombre de questions auxquelles nous allons tenter de répondre dans la suite du manuscrit :

- Nous avons vu l'importance de l'hétérogénéité inter-tumorale dans le cancer. Afin d'éradiquer totalement la maladie, il est important de cibler les différents clones présents et diminuer le risque de résistance aux traitements. Pour cela, il faut identifier et cibler ces clones. **Comment définir facilement des biomarqueurs de sous-population à partir de données accessibles ?**
- Les cellules souches de glioblastomes sont peu connues et caractérisées, or elles sont probablement en partie responsables de la résurgence du cancer car non sensibles aux thérapies actuelles. Il faut donc trouver de nouvelles voies pour les cibler spécifiquement. **Peut-on identifier des biomarqueurs spécifiques des cellules souches de glioblastomes ?**

Comment peut-on les caractériser ? Quels autres mécanismes pourraient être dérégulés et pourraient donc être l'objet de nouvelles voies thérapeutiques ?

- Au niveau technique, nous avons vu l'importance du bruit généré aux différentes phases d'une expérience « omique **». Comment prendre en compte ce bruit lors d'une méta-analyse à partir de données publiques (et donc sur lesquelles nous n'avons aucun contrôle) ?**

Le travail décrit dans cette thèse cherche à répondre ces questions. Pour cela, il a été organisé sur deux axes principaux :

- Le développement d'une méthode générique permettant de prédire les antigènes spécifiques de cancer à partir de données de puces d'expression.
- L'analyse des cellules souches de glioblastomes, que ce soit par l'identification de protéines sur-exprimées à la surface des gCSCs ou bien par l'étude des modifications du signal calcium, dérégulé dans de nombreux cancers.

Chapitre 2 : Détection et priorisation d'antigènes spécifiques de tumeurs (KANT)

2.1 Contexte

Comme évoqué dans le chapitre 1, le cancer est une maladie très hétérogène, que ce soit entre les patients ou entres les différentes cellules d'un même patient. C'est pourquoi, il est important de pouvoir développer des traitements personnalisés, qui viseront spécifiquement le type de cellules cancéreuses dont est atteint le patient. Le travail présenté dans ce chapitre s'est fait en partenariat avec l'entreprise Transgène, société biomarphaceutique qui conçoit, développe et fabrique des produits d'immunothérapie ciblée contre les cancers et les maladies infectieuses. L'objectif de la société était de développer une méthode permettant de détecter et prioriser des antigènes spécifiques de tumeurs, puis d'utiliser les collections de données du programme Carte d'Identité de Tumeurs (CIT) de la Ligue contre le cancer pour identifier des antigènes putatifs de cancer d'intérêt. Ainsi, ces antigènes pourront être utilisés comme cible de nouveaux traitements immunothérapiques ou bien comme biomarqueurs.

Pour cela, deux critères biologiques ont été définis en collaboration avec les immunobiologistes de Transgène : les protéines d'intérêt doivent être accessibles aux anticorps et sur-exprimées dans les cellules cancéreuses par rapport à toutes les autres cellules humaines normales. Pour être accessibles, nous avons choisi de travailler sur les protéines transmembranaires. En effet, les protéines transmembranaires ont une partie externe à la cellule et il est possible de prédire de manière bioinformatique si une protéine est transmembranaire ou pas. Concernant la sur-expression, nous avons choisi de travailler sur les gènes, donc au niveau transcriptomique car les techniques d'analyses haut-débit du transcriptome sont plus en avance que celles du protéome, les données disponibles permettent donc d'analyser d'importantes cohortes sur de nombreux cancers.

L'identification de ces cibles d'intérêt pour être découpée en quatre étapes : (a) l'identification des protéines transmembranaires, (b) la sélection des protéines transmembranaires codées par des gènes sur-exprimés dans le cancer étudié, (c) la priorisation de ces protéines, (d) la validation biologique des résultats (Figure 21).

Nous nous sommes concentrés sur la partie bioinformatique, et donc les étapes a, b et c.

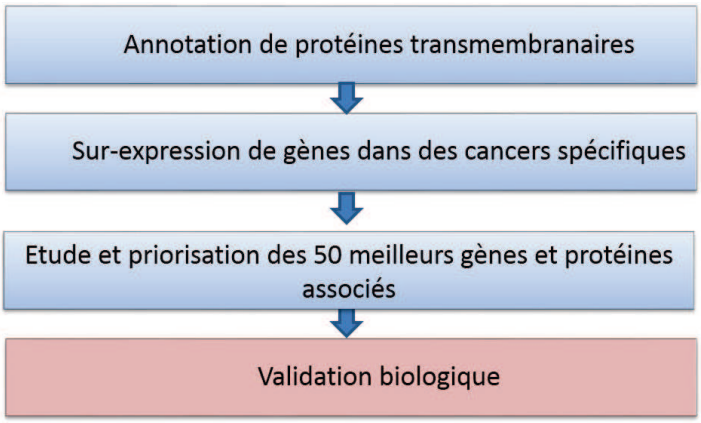

Figure 21 : Etapes pour l'identification d'antigènes putatifs de tumeurs

Première étape

L'exploration de la structure en trois dimensions d'une protéine se fait par diffractions des rayons X sur les cristaux de cette protéine dans un synchrotron. La cristallisation des protéines, et particulièrement celle des protéines transmembranaires est une étape difficile, c'est pourquoi le nombre de protéines transmembranaires avec une structure 3D connue est faible.

La structure d'une protéine s'étudie sur quatre niveaux : la structure primaire ou la séquence d'acides aminés formant la protéine, la structure secondaire qui correspond aux formations périodiques prises par des portions d'une protéine donnée (hélice alpha, feuillet bêta, coude), la structure tertiaire, c'est-à-dire la formation tridimensionnelle stable, la structure quaternaire qui correspond à l'assemblage des sous-unités pour les protéines complexes.

Pour pallier à la difficulté d'obtenir expérimentalement la structure tertiaire d'une protéine, de nombreux algorithmes ont été développés durant ces trente dernières années pour prédire la structure secondaire à partir de la structure primaire des protéines.

La première méthode a été développée par Chou et Fasman en 1974 (Chou and Fasman, 1974). Les auteurs ont calculé pour chaque acide aminé la probabilité de se trouver dans une structure d'hélice alpha, de feuillet beta ou de coude (les trois grandes structures protéiques) à partir de la structure cristalline de 29 protéines déterminées par rayons X. Pour chaque protéine à analyser, la séquence est parcourue par une fenêtre glissante de quatre acides aminés (AA) sur laquelle les probabilités de chaque AA pour chaque structure sont additionnées. Selon le score, on obtient ainsi la structure des acides aminés, qui est étendue jusqu'à trouver un acide aminé avec une probabilité trop faible pour cette structure. Cette méthode présente une efficacité de 50 à 60%.

L'évolution de cet algorithme est la méthode GOR (Garnier, Osguthorbe et Robson) développée en 1978 (Garnier et al., 1978). La fenêtre de travail est de 17 résidus et le score associé est calculé à partir d'une matrice 20*17 prédisant un type de structure secondaire. La position de

l'acide aminé dans la fenêtre est prise en compte pour le calcul du score. L'efficacité de cette méthode dans ses différentes versions est de 55 à 70 %.

L'efficacité de ces méthodes s'est améliorée grâce à l'alignement multiple de séquences. Ces algorithmes reposent sur l'idée que les séquences protéiques ayant subi la même évolution ont en commun quelques sites conservant la même structure secondaire. L'utilisation du programme PSI-BLAST (Altschul et al., 1997) permet ainsi d'accroitre l'efficacité des algorithmes.

Pour traverser la bicouche lipidique formant la membrane, les protéines transmembranaires possèdent des hélices alpha ou des feuillets beta, hydrophobiques et liés entre eux par des coudes extra-membranaires. La structure secondaire des protéines peut donc nous permettre de savoir si les protéines sont transmembranaires ou pas. La majeure partie des protéines membranaires adoptent une conformation de type hélice alpha alors que les feuillets beta se retrouvent exclusivement dans les membranes externes de bactéries à Gram négative, des mitochondries et des chloroplastes et dans certaines toxines formant des pores (Wimley, 2003).

La prédiction des portions transmembranaires (TM) des protéines peut se faire à partir de la prédiction des hélices alpha, tout en s'appuyant sur d'autres propriétés : (i) les segments TM ont une taille limitée pour traverser la bicouche lipidique (environ 25 acides aminés) (Cuthbertson et al., 2005) ; (ii) les acides aminés chargés positivement tels que Arg et Lys se retrouvent préférentiellement du côté du cytoplasme (« positive-inside rule ») (Heijne, 1986).

La première génération d'algorithme se basait sur ces propriétés. La deuxième génération utilise des algorithmes d'apprentissage supervisée, en plus de prendre en compte les propriétés précédentes, tels que les chaines de Markov cachées (ou HMM : Hidden Markov Model) comme HMMTOP (Tusnády and Simon, 1998), TMHMM (Sonnhammer et al., 1998) puis beaucoup d'autres ; les réseaux de neurone (NN : Neural Network) comme (Rost et al., 1995) MEMSAT3 (Jones, 2007) et les machines à vecteur de support (SVM : Support Vector Machine) comme Octopus (Viklund and Elofsson, 2008), MEMSAT-SVM (Nugent and Jones, 2009). Le Tableau 3 est une sélection de quelques algorithmes de type apprentissage supervisée.

En plus d'utiliser des algorithmes d'apprentissage supervisé, l'évolution des algorithmes s'est faite par l'utilisation d'informations d'évolution grâce à l'alignement multiple de séquences, puis la prédiction des hélices « ré-entrantes » qui ne traversent une partie de la membrane avant de retourner en arrière, ainsi que la prédiction des peptides signaux, très similaires en terme de propriétés physico-chimiques et de taille aux segments TM. Finalement, des algorithmes de type consensus ont aussi été développé ; le principe est de se baser sur différentes méthodes de prédiction avant d'en faire un consensus, c'est le cas de TOPCONS (Bernsel et al., 2009).

Différentes études ont eu pour objectif de comparer ces algorithmes (Chen and Rost, 2002; Cuthbertson et al., 2005), ces études ont montré que les performances de ces algorithmes sont à peu près semblables.

Notre objectif, pour pouvoir prédire l'ensemble du protéome est d'avoir un algorithme dont les sources sont disponibles et avec les meilleures performances possibles. Se basant sur ces critères, nous avons choisi d'étudier MEMSAT-SVM et TOPCONS, qui semble avoir les meilleurs résultats d'après leurs propres données et de les tester sur un ensemble de données pour garder le meilleur.

Tableau 3 : Sélection d'algorithmes de type machine learning pour la prédition des protéines transmembranaires
NN : neural network (réseaux de neurones); HMM : hidden Markov model (chaine de Markov caché); SVM : support vector machine (machine à vecteurs de support); TM : région transmembranaire; PS : peptide signal; RE : région ré-entrante

Méthodes	Type d'algorithme	Structures prédites
PHDhtm (Rost et al., 1995)	NN	TM
HMMTOP (Tusnády and Simon, 1998)	HMM	TM
TMHMM (Sonnhammer et al., 1998)	HMM	TM
ENSEMBLE (Martelli et al., 2003)	NN+HMM	TM
PHOBIUS (Käll et al., 2004)	HMM	TM + PS
PRODIV-TMHMM (Viklund and Elofsson, 2004)	HMM	TM
ZPRED (Granseth et al., 2006)	NN + HMM	TM
MEMSAT3 (Jones, 2007)	NN	TM + PS + RE
SPOCTOPUS (Viklund and Elofsson, 2008)	NN + HMM	TM + PS + RE
SVMtop (Lo et al., 2006)	SVM	TM
MEMSAT-SVM (Nugent and Jones, 2009)	SVM	TM + PS + RE

Deuxième étape

Les méthodes traditionnelles de comparaison de niveau d'expression des gènes entre groupes, telles que le t-test ou le Mann-Whitney test se basent sur l'expression d'un gène dans un groupe test par rapport à un groupe contrôle. Cette approche convient pour des groupes homogènes, ce qui n'est pas le cas dans le cancer comme vu dans le chapitre 1.

Pour résoudre ce problème, plusieurs méthodes ont été développées récemment pour détecter des gènes différentiellement exprimés y compris dans des groupes hétérogènes, comme COPA (Cancer Outlier Profile Analysis) (MacDonald and Ghosh, 2006), OS (Outlier Sum) statistic (Tibshirani and Hastie, 2007), ORT (Outlier Robust T-statistic) (Wu, 2007) et DIDs (de Ronde et al., 2013). COPA et OS-statistic sont dérivés du t-test mais utilisent la médiane et la déviation absolue de la médiane au lieu de la moyenne et l'écart-type. ORT base sa détection de gènes sur-exprimés dans un groupe par rapport à un autre groupe sans contraindre les gènes sur-exprimés dans le groupe test à avoir une expression supérieure à l'ensemble des échantillons du groupe contrôle ; ce qui ne correspond pas à ce que l'on recherche. L'algorithme DIDs a été développé pour détecter des biomarqueurs dans des groupes hétérogènes. Cet algorithme prend donc en compte l'hétérogénéité du cancer, et recherche les gènes sur-exprimés dans un sous-ensemble du groupe test par rapport à l'ensemble du groupe contrôle. Nous avons proposé un algorithme similaire en termes de méthode mais plus efficace pour notre analyse.

Troisième étape

Après avoir obtenu une liste d'antigènes putatifs, il faut pouvoir les prioriser pour minimiser les cibles à tester au niveau expérimental. La difficulté de cette priorisation est d'évaluer les critères et leur poids respectif qui donnent une information sur la réalité biologique des analyses sachant que tous les antigènes trouvés sont sur-exprimés dans le cancer et transmembranaires. Le *National Cancer Institute* (Cheever et al., 2009) a fait cette analyse. Ils ont étudié les antigènes de cancer connus et en test pré-clinique, donc bien caractérisés et documentés, afin d'évaluer les critères donnant de « bons antigènes ». Dans notre cas, les antigènes putatifs intéressants sont peu étudiés et donc peu documentés. Nous ne pouvons donc pas utiliser directement leurs critères mais nous nous en sommes inspirés et les avons adaptés à notre problème.

2.2 Matériels & Méthodes

2.2.1 Données

Protéines

Pour comparer les algorithmes de prédiction des protéines transmembranaires, nous avons sélectionnés plusieurs familles de protéines, connues pour être transmembranaires : les RCPG (Récepteurs Couplés aux Protéines G), les canaux ioniques et les clusters de différenciation (CDs). Les deux premières listes proviennent de la base de données IUPHAR (International Union of Basic and Clinical Pharmacology) (Sharman et al., 2011) ; quant aux CDs, nous avons utilisé la liste publiée par Uniprot à http://www.uniprot.org/docs/cdlist. Pour tester la prédiction des structures, nous sommes partis d'un ensemble de données de protéines ayant leur structure résolue, récupérée depuis l'ensemble d'apprentissage du logiciel Octopus (Viklund and Elofsson, 2008). Pour les protéines non transmembranaires, nous avons construit une liste de protéines difficiles à prédire, à partir des récepteurs nucléaires hormonaux (depuis la base données IUPHAR), et de protéines possédant un peptide signal (ensemble s0 utilisé par le logiciel Phobius (Käll et al., 2004)). Nous avons au final une liste de 1088 protéines transmembranaires (TM) et 175 non transmembranaires difficiles à prédire.

Nous avons utilisé un deuxième ensemble de données en tant qu'ensemble test pour vérifier l'amélioration de MEMSAT-SVM. En tant que protéines TM, nous avons sélectionné la famille des transporteurs SLC (*Solute Carrier Transporter*) depuis SwissProt (123 protéines) et en tant que protéines non transmembranaires avec un peptide signal, nous avons utilisé l'ensemble s2 des données de Phobius (127 protéines).

Le travail sur l'ensemble du protéome s'est fait à partir des séquences et annotations d'Uniprot Swissprot (Magrane and Consortium, 2011) qui est vérifiée manuellement. La version de Novembre 2014 que nous avons utilisé contient 20193 protéines humaines.

Puces d'expression

Les données transcriptomiques proviennent de la base de données ArrayExpress (Rustici et al., 2012) et contiennent deux séries de données de cancer du sein : E-MTAB-365 et E-TABM-854, publiés par le programme Cartes d'Identités des Tumeurs (CIT) de La Ligue Nationale contre le Cancer et préparées dans les mêmes conditions. Ces données se composent, respectivement de 537 puces d'expression utilisées pour définir une classification du cancer du sein (Guedj et al., 2012) ; et 74 échantillons de patients avec un cancer du sein familial (Banneau et al., 2010), hybridés sur les puces HG-U133 Plus 2.0 d'Affymetrix.

Notre algorithme utilise deux types de contrôle : le premier CTRL1 est constitué de tissus normaux du même organe que la tumeur étudiée ; le second CTRL2 est constitué d'un ensemble de tissus normaux provenant des autres organes humains. Pour ces contrôles, nous avons utilisé deux séries de données hybridées sur la même puce que précédemment : E-GEOD-10780, qui contient 143 tissus normaux provenant du sein en tant que CTRL1 (Chen et al., 2010) et GSE7307, qui contient 500 échantillons de tissus normaux ayant plus de 90 annotations différentes en tant que CTRL2.

Concernant, les données de lymphome T, les jeux de données utilisés comprennent les séries E-TABM-702 (Huang et al., 2010), E-MTAB-638 (Travert et al., 2012), E-TABM-783 (de Leval et al., 2007). L'ensemble des données provient du programme CIT et est composé (après suppression des doublons) de : 19 échantillons de lymphome T angio-immunoblastique, dont 17 tissus et 2 cellules triées ; 16 échantillons de lymphomes T périphériques sans autre précision ; 9 échantillons de lymphomes T/NK extranodal, dont 2 lignées cellulaires et 7 tissus ; 10 échantillons de lymphome T hépatosplénique, dont 6 tissus, 3 cellules triées, 1 lignée cellulaire ; un ensemble de 26 cellules normales utilisées comme CTRL1 : cellules B naïves, cellules B mémoire, centrocytes, centroblastes, lymphocytes NK activées ou non.

Les données ont été normalisées en utilisant la fonction « justRMA » du package Bioconductor.

2.2.2 Algorithmes de prédictions des protéines transmembranaires

Les algorithmes étudiés sont : MEMSAT-SVM et TOPCONS. La prédiction de MEMSAT-SVM se fait en deux étapes : l'algorithme prédit d'abord si la protéine est transmembranaire ou globulaire à partir d'un algorithme de type SVM, puis, dans le cas d'une protéine TM, il prédit la topologie de la protéine, en incluant la présence ou non d'un peptide signal et d'hélices ré-entrantes à l'aide de 4 SVMs. Lors de ces deux étapes, un score de prédiction est calculé. Nous avons utilisé ces deux scores pour filtrer les protéines transmembranaires de manière plus stringeante que l'algorithme initial et ainsi réduire le nombre de faux-positifs. TOPCONS est un algorithme consensus qui analyse les résultats d'autres prédicteurs. Les prédicteurs utilisés dans notre cas sont ceux proposés dans les sources de TOPCONS : Spoctopus (Viklund et al., 2008), Scampi_msa et Scampi (Bernsel et al., 2008), Prodiv et Pro tmhmm (Viklund and Elofsson, 2004).

Pour améliorer les résultats, nous avons utilisé les scores de prédiction de MEMSAT-SVM. Le premier score (« transmembrane score ») donne une information sur le statut TM alors que le second (« topology score ») qualifie la topologie de l'algorithme. Nous avons calculé des scores minimaux pour pouvoir être prédit comme TM (voir la section résultats).

2.2.3 Traduction des identifiants des protéines Uniprot en identifiant de gènes geneID

Les identifiants Uniprot ont été traduits en Identifiants GeneID en deux étapes : nous avons d'abord utilisé la fonction « ID Mapping » d'Uniprot, puis avons vérifié manuellement les identifiants non traduits et ceux ayant plusieurs résultats en utilisant la base de données du NCBI.

En partant de 5113 protéines TM, nous avons obtenus 4965 GeneID, dont 4367 sont présents sur les puces Affymetrix HG-U133 Plus 2.

2.2.4 Algorithme pour détecter les gènes sur-exprimés

Algorithme développé

L'objectif de cet algorithme est d'identifier les gènes sur-exprimés dans un ensemble de tumeurs par rapport à du tissu normal. Pour cela, nous avons les contraintes suivantes : (i) travailler à partir des probesets Affymetrix associés aux gènes codant pour des protéines transmembranaires, (ii) 100% de spécificité, (iii) une sensitivité maximale, (iv) maximiser la différence d'expression entre les tumeurs et les échantillons contrôle. Cet algorithme est utilisé pour chaque probeset associé à un gène codant pour une protéine TM. Il fonctionne en deux étapes, en utilisant les différents contrôles : le premier, CTRL1, constitué de tissus histologiquement normaux du même organe que les tumeurs, le second, CTRL2 constitué de tissus normaux d'autres organes. Le résultat est le calcul de deux scores (score1 et score2) pour chaque probeset.

Soit l'expression des n_1 CTRL1 échantillons normaux pour un probeset donné par

$$\{x_{1i}\}_{i\in\{1,2,\ldots,n_1\}}$$

L'expression des n_2 CTRL2 échantillons normaux pour un probeset donné par

$$\{x_{2i}\}_{i\in\{1,2,\ldots,n_2\}}$$

Et l'expression des n_3 échantillons tumoraux pour un probeset donné par

$$\{x_{3i}\}_{i\in\{1,2,\ldots,n_3\}}$$

Etape 1 :

Nous commençons par calculer le maximum d'expression pour un probeset pour les échantillons CTRL1 en prenant en compte les outliers potentiels.

Soit Q3 le quartile supérieure et Q1 le quartile inférieur, nous avons défini comme outliers les 2.5% des valeurs supérieures à

$$Q3\{x_{1i}\}_{i\in\{1,2,\ldots,n_1\}} + 3(Q3\{x_{1i}\}_{i\in\{1,2,\ldots,n_1\}} - Q1\{x_{1i}\}_{i\in\{1,2,\ldots,n_1\}})$$

Ces valeurs ne sont pas prises en compte pour calculer le maximum d'expression des CTRL1.

$$\hat{x}_1 = \max\{X_{1i} - \text{outliers}\}$$

$$t_1 = 0.5$$

$$pop_1 = \{x_{3i} > \hat{x}_1 + t_1\}$$

$$n_{1p} = \text{size}\{pop_1\}$$

$$\Delta_1 = \text{median}\{pop_1\} - \hat{x}_1$$

$$\text{Score}_1 = 2^{\Delta_1} \frac{n_{1p}}{n_3}$$

Etape 2 :

En reprenant les mêmes définitions

$$\hat{x}_2 = \max\{X_{2i} - \text{outliers}\}$$

$$t_2 = 0.2$$

$$pop_2 = \{x_{3i} > \hat{x}_2 + t_2\}$$

$$n_{2p} = \text{size}\{pop_2\}$$

$$\Delta_2 = \text{median}\{pop_2\} - \hat{x}_2$$

$$\text{Score}_2 = 2^{\Delta_2} \frac{n_{2p}}{n_3}$$

Ces scores sont absolus et peuvent être comparés entre expériences.

Finalement, deux scores sont assignés à chaque probeset, permettant de les classer. Plus précisément, le Score1 est utilisé pour classer les probesets alors que le Score2 permet de s'assurer que le niveau d'expression d'un gène dans le tissu tumoral est plus élevé que dans les autres tissus.

Nous avons évalué la sensibilité du seuil utilisé pour identifier le meilleur pour chaque score. Pour cela, l'algorithme a été testé avec de seuils de 0 ; 0,1 ; 0,2 jusqu'à 1 en utilisant le jeu de données de cancers du sein et avons choisi des scores de 0.5 pour le seuil du Score1 et 0.2 pour le seuil du Score2.

La sélection de probesets avec un niveau d'expression supérieur à celui des échantillons contrôle additionné d'un seuil nous assure la sur-expression réelle des probesets. La liste finale contient le top 100 des probesets selon le Score1.

Algorithme DIDs

L'algorithme DIDs est assez similaire au nôtre. Il ne prend pas en compte les outliers en calculant le maximum d'expression des contrôles et les différentes fonctions utilisées pour le calcul du score dans la publication sont différentes, la plus appropriée étant la fonction tanh, c'est celle-là que nous allons utiliser à titre de comparaison.

Avec les mêmes définitions que précédemment :

$$\hat{x}_1 = \max\{X_{1i}\}$$

$$score\ (DIDS) = \sum_{i=1}^{n_3} f(|x_{3i} - \hat{x}_1|^+)$$

où

$$|x|^+ = \begin{cases} x\, if\ x \geq 0 \\ 0\, if\ x < 0 \end{cases}$$

$$f(x) = 1 + \tanh(3x - 3)$$

Pour pouvoir comparer DIDs (tanh) et notre algorithme, nous avons utilisé DIDs en deux étapes, avec l'ensemble contrôle CTRL1, puis avec CTRL2.

2.2.5 Priorisation

Les protéines sélectionnées par l'algorithme doivent être validées expérimentalement. C'est pourquoi nous avons créé une méthode de priorisation pour s'assurer que les premières protéines testées sont les cibles les plus prometteuses pour le développement d'un nouveau médicament basé sur des anticorps. Cette méthode part des résultats générés par l'algorithme précédent et se base sur une analyse approfondie de la littérature. Nous l'avons défini à partir de celle développée par le NCI, ainsi que d'une analyse déjà utilisée par Transgène, après discussion entre biologistes et bioinformaticiens sur les résultats obtenus sur le cancer du sein.

Les 50 meilleurs probesets, triés selon le Score1 sont d'abord évalués et sélectionnés si leur Score2 est supérieur à 0.4. Les gènes, représentés par ces probesets, s'ils codent pour une protéine déjà ciblées par un anticorps utilisé en clinique ou en développement ; ou bien s'ils sont exprimés de manière trop importante dans d'autres tissus ; sont supprimés de la liste. Dans le premier cas, l'objectif étant de trouver de nouvelles cibles, nous ne nous intéressons pas à ce qui est déjà développé. Dans le second cas, malgré les différentes précautions, certains

gènes/certaines protéines peuvent être fortement exprimés dans d'autres organes bien qu'ils soient sur-exprimés dans les tumeurs. L'objectif est de ne pas cibler d'autres tissus sains. A la fin de ces étapes de pré-analyse et en prenant en compte le fait qu'un gène est souvent représenté par différents probesets, il reste une liste d'environ 10 gènes codant pour une protéine transmembranaire.

Il est alors possible de faire une revue littéraire approfondie pour les protéines sélectionnées en se basant sur quatre thèmes clés (Tableau 4) : l'accessibilité des anticorps, la connaissance de la fonction et des effets toxiques potentiels, la connaissance du fonctionnement de la protéine dans le cas d'une tumeur, des modèles animaux et/ou d'une preuve de concept préclinique. A chacun de ces thèmes, on attribue un score entre 0 et 20 ; 0 correspondant à la « situation idéale » ; 10 signifiant « pas de donné » et 20 la « mauvaise situation, dangereux pour le patient ».

Chaque thème a un poids différent dans le calcul du score final car ils n'ont pas tous la même importance. En effet l'accessibilité est primordiale, la connaissance de la toxicologie est importante (surtout si elle est négative), et les deux derniers thèmes peuvent être étudiés a posteriori. Finalement, 1000 points sont attribués à chaque protéine, et pour chaque thème, le produit du score obtenu et du poids du thème est soustrait au score total. Chaque protéine finit avec un score, permettant de les trier. Lorsque le score est négatif, la protéine est éliminée.

Tableau 4 : Thèmes et poids utilisés pour la priorisation

Poids	Critère
50	Accessibilité à un anticorps
10	Connaissance de la fonction / Toxicologie
5	Implication connue dans le cancer
5	Modèles animaux + Preuve de concept préclinique

2.2.6 Lien avec les sous-groupes moléculaires et la survie

Les sous-groupes moléculaires utilisés proviennent des annotations des données (sous-groupes de la classification CIT). On considère qu'un échantillon sur-exprime un gène selon la définition de KANT, c'est-à-dire que l'expression du gène est supérieur à celle du maximum des normaux (mêmes tissus), additionné du seuil (0.5). Pour chaque gène, l'analyse est effectuée sur un seul probeset, celui ayant obtenu les meilleurs résultats de l'algorithme. Le test statistique utilisé est un test du chi2.

Les données de survie sont représentées en utilisant l'estimateur de Kaplan-Meier, à l'aide du package *survival* du logiciel R, et les p-valeurs sont estimées par un test du log-rank.

2.3 Résultats

2.3.1 Choix de l'algorithme de prédiction des protéines transmembranaires

Pour choisir entre les deux algorithmes précédemment sélectionnés, TOPCONS et MEMSAT-SVM, nous avons testés leur performance sur un ensemble de protéines composé de 1088 protéines transmembranaires et 175 protéines non transmembranaires difficiles à prédire (car ayant des similarités avec les protéines transmembranaires). Les résultats sont montrés dans les Tableau 5 et Tableau 6. La précision des deux algorithmes est proche de 100% pour les protéines TM mais beaucoup plus faible pour les protéines non TM. Le taux de faux-positifs est de 76,6% pour TOPCONS et 29,1% pour MEMSAT-SVM, ce qui est trop élevé pour servir de référence dans la suite de notre étude. Ces protéines ne sont pas représentatives de l'ensemble des protéines non TM car elles sont difficile à prédire, ce qui pourrait expliquer ces mauvaises performances.

Pour surmonter cette limitation, nous avons proposé une légère modification de MEMSAT-SVM en utilisant les scores TM et de topologie générés par l'algorithme (Figure 22). Nous avons choisi de fixer de nouveaux seuils à ces scores pour qu'une protéine puisse être définie comme transmembranaire. L'objectif est de diminuer le nombre de faux-positifs. Ces scores ont été choisis en maximisant la précision de la prédiction sur l'ensemble de données précédent. On obtient des résultats optimaux en fixant le score TM minimum à 1 et le score de topologie à 0.6. Ces scores permettent d'augmenter la performance de l'algorithme en diminuant le taux de faux positif à 3.4% (voir Tableau 6) et en diminuant le taux de vrais positifs de 99.1% à 97.2%.

Pour valider ces seuils, nous avons utilisés un deuxième ensemble de protéines. Sur les 123 protéines TM, 122 ont été annotées comme TM avec et sans seuils. La protéine mal annotée, SLC7A6OS semble ne pas être TM comme les autres récepteurs SLC d'après Uniprot ; ce qui explique cette erreur. Concernant l'ensemble de données avec un peptide signal (non TM), 41% sont bien annotées sans les seuils et 91% avec les seuils. Ces résultats confirment l'intérêt des seuils pour diminuer le nombre de faux positifs et augmenter la précision globale de l'algorithme.

La topologie exacte prédite par les algorithmes a aussi été comparée car la connaissance des parties externes de la protéine peuvent être utile pour la validation des protéines à la fin de l'étude. Pour cela, nous avons utilisé un ensemble de protéines, dont la structure 3D est connue. Pour ces données, nous savons pour chaque acide aminé s'il est situé dans la cellule, à l'extérieur ou bien dans la membrane, information qui est prédite par MEMSAT-SVM et TOPCONS. Nous avons évalué le pourcentage d'acides aminés correctement prédits pour chaque protéine et avons calculé la moyenne sur l'ensemble des protéines. TOPCONS a correctement annoté 83.34%

des acides aminés et MEMSAT-SVM 85.69%. La majorité des acides aminés mal annotés sont dus à un décalage de deux/trois acides aminés dans la prédiction de la partie transmembranaire.

Aux vues de ces différents résultats, nous avons choisi d'utiliser MEMSAT-SVM avec les nouveaux seuils pour la suite des analyses pour ses meilleurs résultats de prédiction, ainsi que sa facilité d'utilisation et ses performances en termes de temps de calcul.

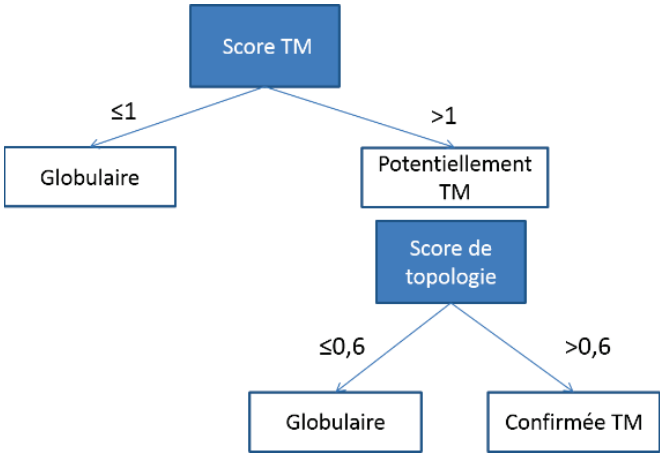

Figure 22 : Modification de MEMSAT-SVM, ajout de seuils sur les scores TM et topologie

Tableau 5 : Prédiction des protéines transmembranaires, résultats obtenus sur le jeu de données test pour Topcons et MEMSAT-SVM

	Nombre de protéines	Proteines annotatées TM par TOPCONS	Proteines annotatées TM par MEMSAT-SVM
CD	389	383	381
Données d'Octopus	124	124	123
RCPG a	282	282	282
RCPG b	49	48	49
RCPG c	22	22	22
RCPG 'crépus – grésillés »	11	11	11
Canaux ioniques voltage dépendant	140	140	140
Canaux ioniques ligand dépendant	71	70	70
Récepteurs nucléaires hormonaux	48	21	0
Données de Phobius « Signal Peptide)	127	113	51

Tableau 6 : Résultats pour TOPCONS, MEMSAT-SVM, MEMSAT-SVM + seuils pour les deux ensembles de données

Données	Algorithme	Vrais Positifs	Faux positifs	Vrais négatifs	Faux négatifs	Taux de vrais positifs (%)	Taux de faux positifs (%)	Précision (%)
Ensemble d'apprentissage	Topcons	1080	134	41	6	99.4	76.6	88.9
	MEMSAT-SVM	1078	51	124	10	99.1	29.1	95.2
	MEMSAT + seuils	1057	6	170	30	97.2	3.4	97.1
Ensemble test	MEMSAT-SVM	122	74	52	1	99.2	58.7	69.9
	MEMSAT + seuils	122	11	115	1	99.2	8.7	95.2

2.3.2 Prédiction de l'ensemble du protéome

Nous avons utilisé MEMSAT-SVM, avec les nouveaux seuils, sur l'ensemble du protéome humain, à partir de la base de données Swissprot contenant 20193 protéines en Novembre 2014. 5113 protéines ont été annotées comme transmembranaires, correspondant à 25% du protéome. Ces résultats correspondent aux chiffres trouvés dans la littérature, en 2006, Ahram et al ont testé quatre algorithmes de prédiction, ils estiment qu'entre 15 et 39 % du protéome humain est situé au niveau des membranes cellulaires (Ahram et al., 2006).

2.3.3 Application 1 : cancer du sein

2.3.3.1. Cancer du sein

Il s'agit d'un cancer qui se développe dans le sein, principalement dans les canaux galactophores (qui transportent le lait) dans 75% des cas, ou bien dans les lobules (glandes qui produisent le lait) pour 10% des cas (Bertos and Park, 2011), (voir Figure 23). Le cancer du sein peut toucher les femmes et les hommes, même si ceux-ci sont très minoritaires.

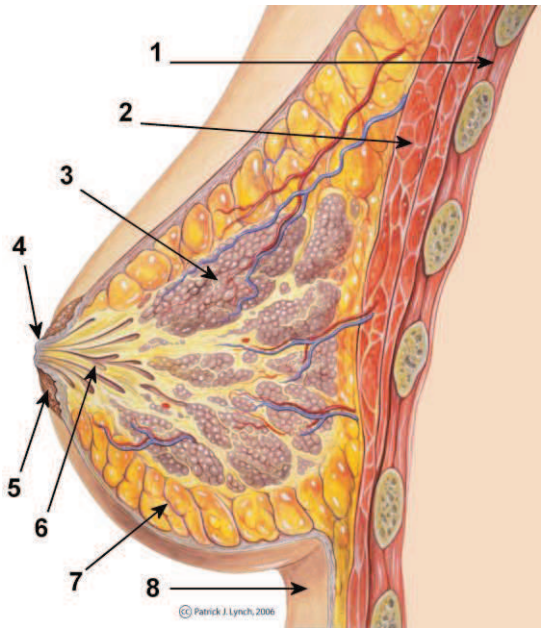

Figure 23 : Schéma anatomique du sein.
1-Cage thoracique, 2 - Muscles pectoraux, 3- Lobules, 4- Mamelon, 5 - Aréole, 6- Canaux galactophores, 7-Tissu adipeux, 8-Peau

Comme vu dans le Chapitre 1, le cancer du sein est le cancer ayant la plus grande incidence chez la femme en France et responsable du plus grand nombre de décès par cancer.

Le cancer du sein est très hétérogène. L'expression d'un certain nombre de protéines est testée en routine clinique pour pouvoir adapter le traitement aux patientes. En effet, le tamoxifène (traitement de type hormonothérapie) n'est administré qu'aux femmes dont les cellules tumorales sur-expriment les récepteurs aux œstrogènes (RE) et progestérones (RP), de même l'herceptin, anticorps monoclonal anti HER2/NEU n'est prescrit qu'à la petite proportion de cancers positifs à ERBB2/HER2. Plusieurs études ont été publiées ces dernières années pour établir une classification moléculaire. Le premier travail, publié par Sorlie et al. (Sorlie et al., 2003) a permis la définition de 5 sous-types moléculaires : *Basal*, *ERBB2 +*, *Luminal A*, *Luminal B*, *Normal breast-like*. Ces sous-types ont été retrouvés en utilisant différentes plateformes technologiques (Hu et al., 2006; Parker et al., 2009). En clinique, les sous-types utilisés se basent sur le statut de RE, RP et ERBB2. On trouve ainsi, les cancers dits triples-négatifs (RE-/RP-/ERBB2-), les cancers luminal (RE+/PR+/ERBB2-) et les tumeurs ERBB2 +. L'équipe CIT a publié une nouvelle classification en 2012 (Guedj et al., 2012) redéfinissant les sous-types moléculaires en six groupes : *basal*, *molecular apocrine*, *luminal A*, *luminal B*, *luminal C* et *normal-like*, nommés d'après la précédente classification. Ces groupes ont été définis à partir de la classification non supervisée des données de transcriptomique. Ces sous-groupes montrent des différences spécifiques de nombre de copies de gènes et d'activations de voies de signalisation. Le groupe triple négatif (RE-/PR-/ERBB2-) est aussi négatif aux récepteurs des androgènes (AR-), c'est le groupe basal. Quatre groupes sont RE+ : deux peu prolifératifs (gènes du cycle cellulaire peu exprimés),

il s'agit de *Luminal A* et *Normal breast-like* ; et deux très prolifératifs, *Luminal B* et *Luminal C*. Un groupe RA-/ER-/PR- est nommé *Molecular Apocrine.* Le groupe ERBB2+ est réparti entre le groupe *Luminal C* et *Molecular Apocrine*.

L'utilisation de KANT sur un ensemble de tumeurs appartenant à ces différents sous-groupes devrait nous permettre de trouver des gènes sur-exprimés dans un ou plusieurs sous-groupes.

2.3.3.2. Application de KANT et DIDs (tanh)

Les résultats de KANT, triés selon le score1 sont montrés dans le Tableau 7. Certains gènes sont représentés par différents probesets (comme BMPR1B *(Bone morphogenetic Protein Receptor, Type 1B)* ou CEACAM6 (*Carcinoembryonic Antigen-Related Cell Adhesion Molecule 6)*). La sensibilité maximale n'atteint pas 100%, à cause de l'hétérogénéité des échantillons et des critères drastiques définis pour qu'un échantillon soit considéré sur-exprimé pour un gène. Néanmoins, les gènes candidats avec une sensibilité plus faible sont intéressants. L'objectif du score2 est de s'assurer que le niveau d'expression d'un gène cible dans les tumeurs est plus élevé que dans les autres tissus normaux. L'idée générale étant qu'une molécule cible ne doit être présente que dans les cellules affectées par la pathologie afin de minimiser le risque d'effets secondaires si ces molécules sont utilisées comme cibles de nouveaux traitements. Le score2 minimal pour les probesets sélectionnés par la première partie de l'algorithme est 0,4.

Parmi les gènes sélectionnés, certains sont connus pour être sur-exprimés dans le cancer du sein (Tableau 8). MUC1 (*Mucin 1*) est la cible de nombreux traitements de type immunothérapeutiques en cours de développement à différentes phases cliniques pour le traitement du cancer du sein, mais aussi d'autres cancers épithéliaux comme la prostate, le pancréas, le colon ou le poumon. D'autres gènes, comme SQLE *(Squalene Epoxidase)*, TBC1D9 (*TBC1 Domain Family, Member 9)* et SLC40A1 *(Solute Carrier Family 40 (Iron-Regulated Transporter) Member 1)* sont connus comme indicateurs de résultats cliniques (Andres et al., 2013; Helms et al., 2008; Miller et al., 2011).

Nous avons aussi utilisé la fonction de score de l'algorithme DIDs(tanh) en remplacement de KANT dans notre méthode, une première fois pour obtenir un score1 (en utilisant les tumeurs comparées aux échantillons CTRL1), une deuxième fois pour obtenir un score2 (en utilisant les tumeurs comparées aux échantillons CTRL2). Nous avons choisi le seuil minimum pour le score2 en sélectionnant le même nombre de probesets que KANT sur ce jeu de données. Un minimum de 0.4 permet de sélectionner un total de 174 probesets avec KANT (sans prendre en compte le score1), ce qui correspond à un score de 39 avec DIDs (tanh). Sur la méthode complète, 27 gènes sont sélectionnés en utilisant KANT et 26 avec DIDs(tanh), dont 18 en commun (voir Tableau 8). Pour évaluer la pertinence des gènes sélectionnés par les 2 algorithmes, nous avons regardé le nombre de publications ressortant d'une recherche Pubmed avec le nom du gène et « breast cancer ». Nous avons aussi regardé les raisons pour lesquelles un gène était

sélectionné par un algorithme mais pas par le deuxième (score1 ou score2) (Tableau 6). Parmi les gènes connus (beaucoup de publications Pubmed) DIDs (tanh) ne sélectionne pas MUC1 et ABO (*ABO Blood Group*) et KANT PRLR *(Prolactin Receptor)*. Les gènes non sélectionnés par KANT à cause du Score1 sont proches (103^{e} rang à 131^{e} rang) alors que ceux non sélectionnés par DIDs (tanh) sont beaucoup plus éloignés (134^{e} rang à 635^{e} rang pour ABO). Au final, même si KANT et DIDs (tanh) sont des algorithmes très proches, KANT donne de meilleurs résultats pour notre méthode.

Tableau 7 : Gènes sélectionnés par KANT d'après les Scores1 et Scores2 sur les données de cancer du sein

Probeset	Entrez Gene	Gene Symbol	Score1	Score2
229975_at	658	BMPR1B	3.05816344	0.88707622
222906_at	28982	FLVCR1	2.99410871	0.59180496
210523_at	658	BMPR1B	2.51705498	1.15385813
242579_at	658	BMPR1B	2.30811787	0.42998616
242517_at	84634	KISS1R	2.29885462	0.74914923
213562_s_at	6713	SQLE	2.05833411	0.56002527
239983_at	169026	SLC30A8	1.76126233	1.44118269
207142_at	3760	KCNJ3	1.67705449	0.71825344
222379_at	23704	KCNE4	1.62492641	0.90421885
1555274_a_at	85465	EPT1	1.53686491	0.97968386
212960_at	23158	TBC1D9	1.38814068	1.34602923
235976_at	84189	SLITRK6	1.36690925	1.0342263
217787_s_at	2590	GALNT2	1.35276835	0.48160234
233123_at	30061	SLC40A1	1.29765077	0.44067042
213693_s_at	4582	MUC1	1.29271625	0.93586082
218989_x_at	64924	SLC30A5	1.26446136	0.43377187
238635_at	64417=$$	C5orf28	1.24283971	0.66223472
1555460_a_at	25800	SLC39A6	1.22530472	1.7369967
213909_at	131578	LRRC15	1.14926896	3.25927515
212008_at	23190	UBXN4	1.14836376	0.52120313
216929_x_at	28	ABO	1.13939537	0.52720015
208716_s_at	54499	TMCO1	1.13384789	0.44230849
215088_s_at	6391	SDHC	1.1191709	0.50653395
218073_s_at	55706	TMEM48	1.10349314	1.50953385
203108_at	9052	GPRC5A	1.10300631	0.40529415
242019_at	253782	LASS6	1.10156335	1.68185453
1552508_at	23704	KCNE4	1.05716505	0.40503406
207549_x_at	4179	CD46	1.02880073	0.44426047
235463_s_at	253782	LASS6	1.01649211	2.3117165
229500_at	10463	SLC30A9	1.00535197	0.77729866
1555201_a_at	55005	RMND1	0.99369345	0.46047998

Tableau 8 : Comparaison des résultats entre KANT et DIDs (tanh).

Pour chaque gène, le nombre de publication avec les mots clés "breast cancer" et le nom du gène est indiqué, ainsi que la raison pour laquelle le deuxième algorithme n'a pas sélectionné le gène lorsqu'il n'est sélectionné que par un algorithme.

Algorithme	Gènes	Pubmed "breast cancer"	Pourquoi le gène n'a pas été sélectionné par le deuxième algorithme ?
KANT et DIDs (tanh)	BMPR1B	12	
	FLVCR1	1	
	KISS1R	17	
	KCNE4	0	
	EPT1	0	
	TBC1D9	2	
	SLITRK6	0	
	SLC30A5	0	
	SLC39A6	20	
	UBXN4	0	
	TMCO1	0	
	SDHC	5	
	TMEM48	0	
	LASS6	2	
	CD46	29	
	SLC30A9	0	
	SLC40A1	4	
	RMND1	1	
KANT	SQLE	6	Score2 trop faible
	SLC30A8	1	219[e] Score1
	KCNJ3	2	134[e] Score1
	GALNT2	0	Score2 trop faible
	MUC1	767	Score2 trop faible
	C5orf28	0	Score2 trop faible
	LRRC15	5	423[d] Score1
	ABO	152	635[e] Score1
	GPRC5A	6	344[e] Score1 and Score2 trop faible
DIDs (tanh)	C1orf43	0	Score2 trop faible
	DNAJC1	0	Score2 trop faible
	DEGS2	0	103[e] Score 1 and Score2 trop faible
	SELT	2	Score2 trop faible
	SLC2A10	2	110[e] Score1
	TAP1	11	131[e] Score1 and Score2 trop faible
	MFN1	2	Score2 trop faible
	PRLR	116	126[e] Score1

2.3.3.3. Priorisation des protéines

La priorisation des protéines a été faite à partir des 50 premiers probesets selon le score1. L'objectif est d'avoir un nombre réduit de protéines à étudier. On s'intéresse donc à BMPR1B, FLVCR1 *(Feline Leukemia Virus Subgroup C Cellular Receptor 1)*, KISS1R *(KISS1 Receptor)*, SQLE, SLC30A8 *(Solute Carrier Family 30 (Zinc Transporter), Member 8)*, KCNJ3 *(Potassium Channel, Inwardly Rectifying Subfamily J, Member 3)*, KCNE4 *(Potassium Channel, Voltage Gated Subfamily E Regulatory Beta Subunit 4)*, EPT1 *(Ethanolaminephosphotransferase 1)*, TBC1D9, SLITRK6 *(SLIT and NTRK-Like Family, Member 6)*, GALNT2 *(Polypeptide N-Acetylgalactosaminyltransferase 2)*, SLC40A et MUC1.

La première étape consiste à supprimer les protéines déjà connues, trop exprimées ou bien non localisées sur la membrane cellulaire d'après la littérature. Nous avons déjà vu que des traitements ciblant MUC1 sont en cours de développement, il est donc retiré de la liste. Nous avons ensuite supprimé BMPR1B car il est fortement exprimé dans les muscles squelettiques. KCNE4 et SLC40A1 ont aussi un niveau d'expression trop élevé dans d'autres tissus humains. Les informations sur les niveaux d'expression viennent du jeu de données CTRL2 utilisé, ainsi que du *Gene Atlas Datasets* du site *bioGPS* (Su et al., 2004). SQLE et GALNT2 sont localisés sur d'autres membranes que la membrane cellulaire (réticulum endoplasmique et appareil de Golgi). Il ne reste plus que sept protéines à étudier. Après une analyse approfondie de la littérature sur ces protéines, nous avons affecté un score à chacun pour ces thèmes : Accessibilité / Connaissance de la fonction et des effets toxiques potentiels / Connaissance du fonctionnement de la protéine dans le cas d'une tumeur / Connaissance des modèles animaux et/ou de preuves de concepts précliniques. A partir de ces données nous avons calculé un score final (Tableau 9).

Au final, dans l'ordre des cibles les plus prometteuses aux moins intéressantes, on trouve FLVCR1, SLC30A8, SLITRK6, EPT1 et TBC1D9. KCNJ3 et KISS1R ont obtenu des scores trop faibles (proches de zéro) et ne devrait donc pas être considérés.

Tableau 9: Priorisation des protéines à partir d'un petit nombre de critères.
Les scores par critère sont compris entre 0 et 20. Le score final est calculé en soustrayant à 1000 la somme des produits des scores par leurs poids

Gène	Accessibilité de la protéine à un anticorps	Connaissance de la function / Toxicologie	Lien avec le cancer	Modèles animaux /POC précliniques	Score final
Poids	*50*	*20*	*5*	*5*	
FLVCR1	0	0	10	17	865
SLC30A8	0	15	10	0	650
SLITRK6	5	10	5	10	475
EPT1	10	2	10	10	360
TBC1D9	10	10	10	10	200
KCNJ3	10	20	10	0	50
KISS1R	10	20	20	10	-50

2.3.3.4. Lien avec les sous-groupes moléculaires et la survie

Nous avons voulu savoir si les gènes/protéines sur-exprimés mis en évidence précédemment étaient associés à un sous-type moléculaire. Pour cela, nous avons utilisé les annotations de la classification CIT, en considérant les données du cancer du sein familial comme un groupe à part (kfam). Nous avons donc 6 groupes : Basal (basL), Luminal A (lumA), Luminal B (lumB), Luminal C (lumC), Molecular Apocrine (mApo), Normal breast-like (normL) et cancer familial (kfam).

Nous avons regardé pour tous les gènes sélectionnés avant priorisation (voir Tableau 10). FLVCR1 et GALNT2 sont sur-exprimés dans tous les sous-groupes et par conséquent, la p-valeur associée au chis-square test est élevée. BMPR1B est sur-exprimé dans les *Luminal A* mais pas du tout dans le groupe *basal* ; on le retrouve en partie dans les autres groupes. KISS1R est sur-exprimé par les groupes *Luminal A* et *B* majoritairement. SLC30A8 est sur-exprimé par peu d'échantillons mais aucun dans les groupes *basal* et *molecular apocrine* ; de même pour KCNJ3. KCNE4 est sur-exprimé dans le groupe *Luminal A* mais pas dans les groupes *basal* et *molecular apocrine*. EPT1 est sur-exprimé partout sauf dans le groupe *normal-like* (61% des échantillons le sur-expriment). TBC1D9 est sur-exprimé dans les groupes *Luminal A*, *B*, *normal-like* mais pas du tout dans les groupes *basal* et *molecular apocrine.* SLITRK6 est peu sur-exprimé, il ne l'est pas du tout dans les groupes *basal* et *molecular apocrine*. SLC40A1 est sur-exprimé dans les *normal-like* mais pas dans le groupe *basal*. Finalement MUC1 est sur-exprimé dans les groupes *luminal A* et *normal-like*.

Tableau 10 : Résultat de l'association entre sur-expression des gènes identifiés par KANT et sous-groupes moléculaires

La p-valeur a été calculée par un test du chi2. Les échantillons sont considérés sur-exprimés ou sous-exprimés selon les critères de KANT. Lorsque plusieurs probesets pour un même gène faisaient partie des probesets sélectionnés, seul le premier a été noté dans le tableau mais nous avons vérifié que les résultats étaient équivalents pour les autres probesets. Les groupes correspondent à basal, cancers familiaux, luminal A, luminal B, luminal C, molecular apocrine, normal-like.

	p-valeur	basL	kfam	lumA	lumB	lumC	mApo	normL
Total		53	74	118	103	91	40	132
BMPR1B +	2.6E-30	2 (4%)	44 (59%)	97 (82%)	62 (60%)	27 (30%)	1 (3%)	62 (47%)
BMPR1B -		51 (96%)	30 (41%)	21 (18%)	41 (40%)	64 (70%)	39 (97%)	70 (53%)
FLVCR1 +	0.3	52 (98%)	72 (97%)	109 (92%)	101 (98%)	88 (97%)	37 (93%)	126 (95%)
FLVCR1 -		1 (2%)	2 (3%)	9 (8%)	2 (2%)	3 (3%)	3 (7%)	6 (5%)
KISS1R +	9.6E-13	11 (21%)	44 (59%)	92 (78%)	76 (74%)	57 (63%)	16 (40%)	90 (68%)
KISS1R -		42 (79%)	30 (41%)	26 (22%)	27 (26%)	34 (37%)	24 (60%)	42 (32%)
SQLE +	5.8E-11	48 (91%)	69 (93%)	106 (90%)	103 (100%)	83 (91%)	38 (95%)	93 (70%)
SQLE -		5 (9%)	5 (7%)	12 (10%)	0 (0%)	8 (9%)	2 (5%)	39 (30%)
SLC30A8 +	6.6E-07	0 (0%)	20 (27%)	42 (36%)	37 (36%)	20 (22%)	4 (10%)	26 (20%)
SLC30A8 -		53 (100%)	54 (73%)	76 (64%)	66 (64%)	71 (78%)	36 (90%)	106 (80%)
KCNJ3 +	3.6E-20	0 (0%)	27 (36%)	63 (53%)	51 (50%)	21 (23%)	1 (3%)	20 (15%)
KCNJ3 -		53 (100%)	47 (64%)	55 (47%)	52 (50%)	70 (77%)	39 (97%)	112 (85%)
KCNE4 +	6.67E-25	1 (2%)	40 (54%)	87 (74%)	42 (41%)	25 (27%)	1 (3%)	70 (53%)
KCNE4 -		52 (98%)	34 (46%)	31 (26%)	61 (59%)	66 (73%)	39 (97%)	62 (47%)
EPT1 +	4.08E-14	45 (85%)	73 (99%)	108 (92%)	95 (92%)	79 (87%)	35 (88%)	81 (61%)
EPT1 -		8 (15%)	1 (1%)	10 (8%)	8 (8%)	12 (13%)	5 (12%)	51 (39%)
TBC1D9 +	3.06E-55	0 (0%)	56 (76%)	113 (96%)	93 (90%)	41 (45%)	4 (10%)	105 (80%)
TBC1D9 -		53 (100%)	18 (24%)	5 (4%)	10 (10%)	50 (55%)	36 (90%)	27 (20%)
SLITRK6 +	1.88E-11	5 (9%)	40 (54%)	66 (56%)	32 (31%)	30 (33%)	3 (7%)	52 (39%)
SLITRK6 -		48 (91%)	34 (46%)	52 (44%)	71 (69%)	61 (67%)	37 (93%)	80 (61%)
GALNT2 +	0.045	50 (94%)	63 (85%)	100 (85%)	82 (80%)	82 (90%)	39 (98%)	115 (87%)
GALNT2 -		3 (6%)	11 (15%)	18 (15%)	21 (20%)	9 (10%)	1 (2%)	17 (13%)
SLC40A1 +	4.17E-24	2 (4%)	44 (59%)	66 (56%)	25 (24%)	48 (53%)	10 (25%)	112 (85%)
SLC40A1 -		51 (96%)	30 (41%)	52 (44%)	78 (76%)	43 (47%)	30 (75%)	20 (15%)
MUC1 +	5.04E-24	7 (13%)	47 (64%)	96 (81%)	57 (55%)	57 (63%)	17 (43%)	117 (89%)
MUC1 -		46 (87%)	27 (36%)	22 (19%)	46 (45%)	34 (37%)	23 (57%)	15 (11%)

En plus de l'association aux sous-groupes moléculaires, nous avons regardé si la sur-expression de ces gènes d'intérêt avait un effet sur la survie, que ce soit positif ou négatif. Pour cela, nous avons analysé les courbes de survie des patients qui sur-expriment les gènes par rapport à ceux qui ne les sur-expriment pas (Figure 24). L'expression de 5 gènes semble avoir un effet significatif sur la survie (p-valeur <0.05) : BMPR1B, KISS1R, SLITRK6, SLC40A1 et MUC1. L'utilisation des sous-groupes définis par KANT pour diviser les patients en sur-exprimant et ne sur-exprimant pas le gène limite cette analyse lorsque l'un des deux groupes contient beaucoup plus de patients que l'autre (comme pour FLVCR1) mais cela reste la meilleure manière de définir les patients plutôt que de prendre un seuil sans réalité biologique.

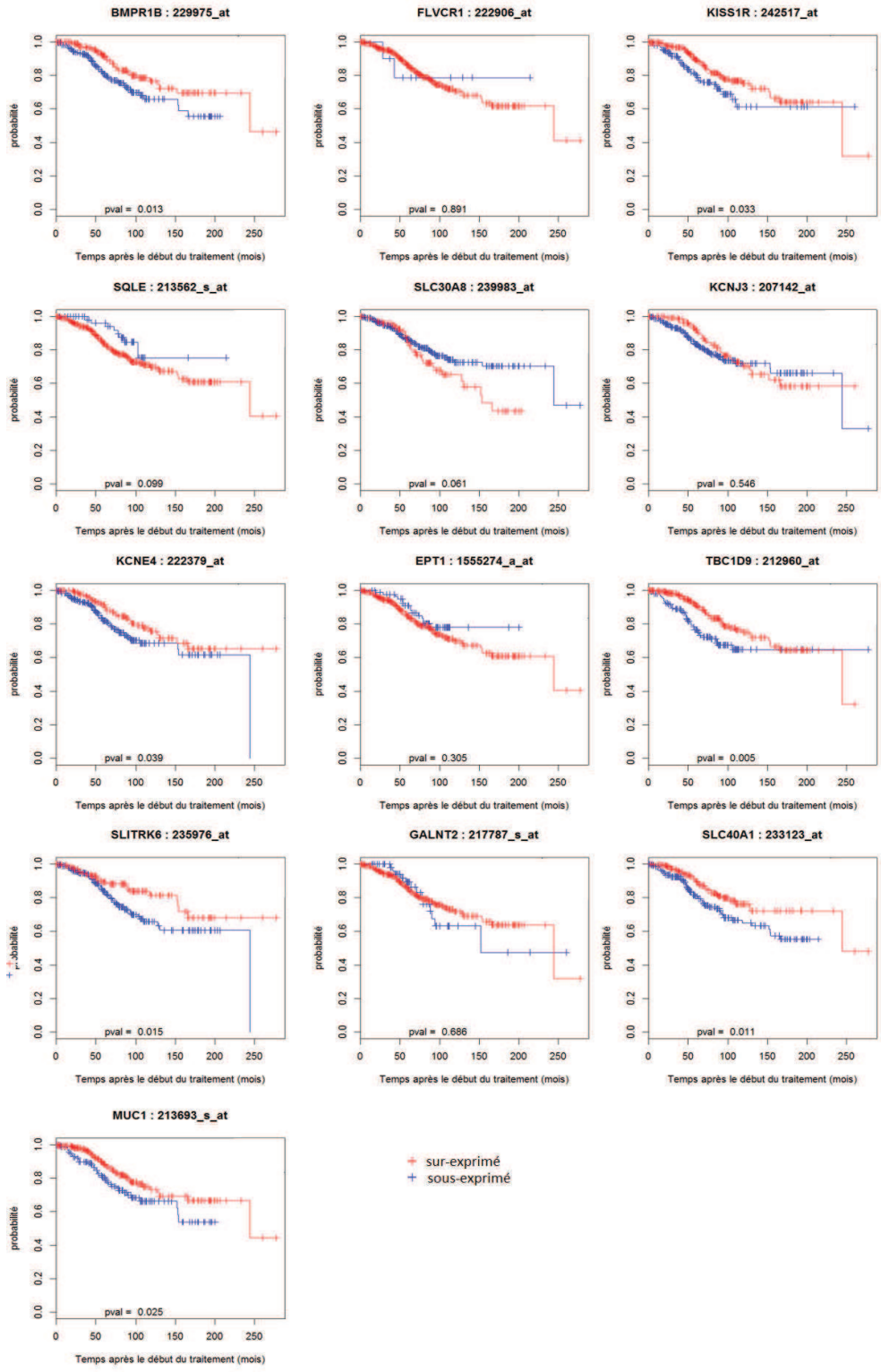

Figure 24 : Courbes de survie des patients atteints de cancer du sein en fonction de l'expression des marqueurs trouvés

La p-valeur a été calculée par un test du logrank. Les échantillons sont considérés sur-exprimés ou sous-exprimés selon les critères de KANT. Lorsque plusieurs probesets pour un même gène faisait partie des probesets sélectionnés, seul le premier a été noté dans le tableau mais nous avons vérifié que les résultats étaient équivalents pour les autres probesets.

2.3.4 Application 2 : lymphome T

2.3.4.1. Lymphome T

Le lymphome est le cancer du système lymphatique. Les lymphomes se divisent en deux types : les lymphomes hodgkiniens (nommé ainsi d'après le Dr Thomas Hodgkin, qui fut le premier à les décrire) caractérisés par la présence de cellules anormales particulières, les cellules de Sternberg et les lymphomes non hodgkiniens. Ces derniers se divisent en deux catégories en fonction des cellules atteintes : les lymphomes B qui touchent les lymphocytes B et les lymphomes T qui touchent les lymphocytes T. Les lymphomes T ne représentent que 10 à 15% des lymphomes non-hodgkiniens. Les sous-types de lymphomes B et T sont nombreux. Les lymphomes T se divisent entre ceux issus des précurseurs des cellules T et ceux issus des cellules T. Les lymphomes T comptent 19 sous-types d'après la classification OMS de 2008 (voir Figure 25). Dans la suite de cette étude, l'ensemble de données étudié contient des lymphomes T angio-immunoblastiques, des lymphomes T périphériques sans autre précision (tous deux de type nodal) et des lymphomes T/NK extranodals. Il s'agit donc de trois formes différentes, plus ou moins rares, agressives.

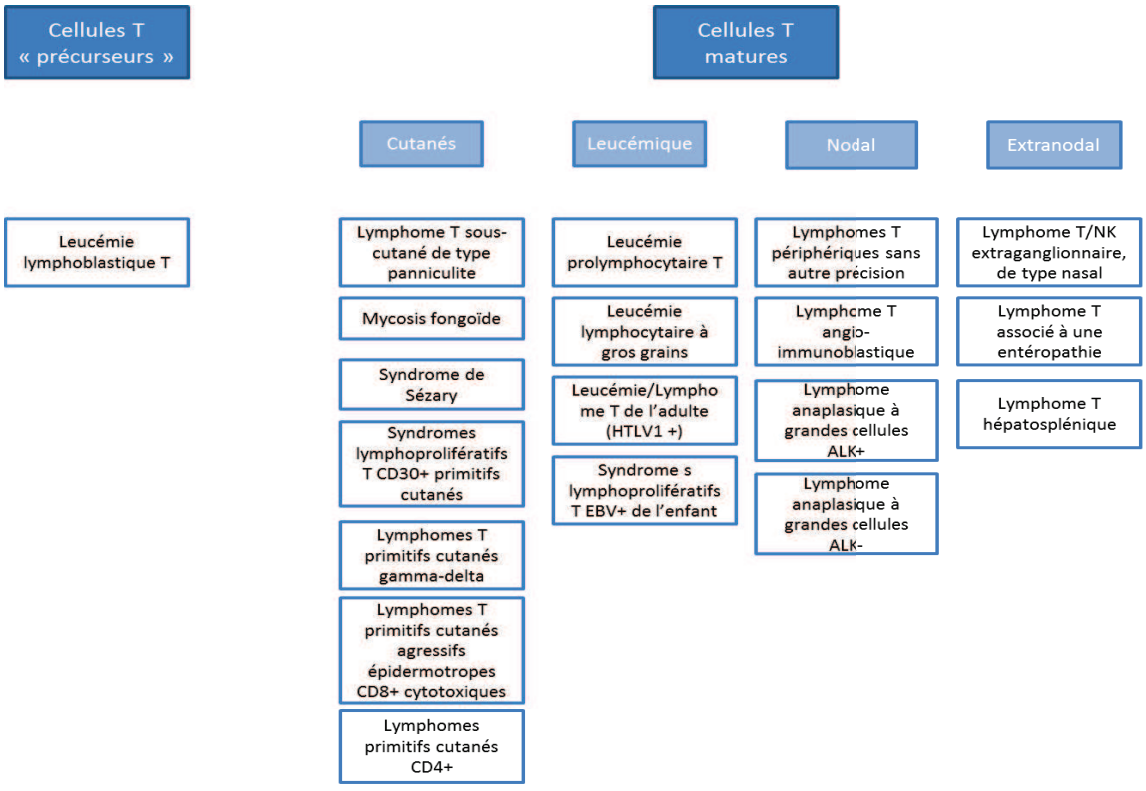

Figure 25 : Classification des Lymphomes T, *Source : d'après (de Leval et al., 2009)*

2.3.4.2. Application de KANT

Les résultats de KANT, triés selon le score1 sont donnés dans le Tableau 11. Le nombre d'échantillons tumoraux est de 54. On observe une sensibilité plus importante des gènes sur-exprimés par rapport aux résultats du cancer du sein, qui s'explique par une hétérogénéité des échantillons beaucoup plus faibles.

Parmi les gènes/protéines ressortant, on trouve CTLA-4 *(Cytotoxic T-Lymphocyte-Associated Protein 4)*, qui est ciblé par l'ipilimumab dans le cas du mélanome (voir chapitre 1) et est en étude clinique de Phase 1 pour les rechutes de cancers hématologiques, incluant les lymphomes T (essai clinique NCT01822509). ENTPD1 *(Ectonucleoside Triphosphate Diphosphohydrolase 1)* a été noté comme marqueur pronostique dans le cas des leucémies lymphocytaires chroniques (lymphocytes de type B) (Abousamra et al., 2015). On trouve d'autres résultats pour TLR8 *(Toll-Like Receptor 8)* (Huen and Rook, 2014). Dans (Sekulic et al., 2015), les auteurs ont montré l'existence d'un gène de fusion CTLA4-CD28 chez des patients atteint du syndrome de Sézary. L'utilisation de l'ipilimumab chez ces patients a résulté en une réponse clinique rapide.

Comme dans le cas du cancer du sein, l'utilisation de l'algorithme KANT permet de retrouver un certain nombre de protéines, cibles de traitements de type anticorps et d'autres non connues.

Tableau 11 : Résultats de KANT selon les Score1 et Score2 sur les lymphomes T

Probeset	Entrez Gene	Gène	Score1	Score2
214511_x_at	2210	FCGR1B	8.5143416	0.91859452
236341_at	1493	CTLA4	6.79424034	7.36693315
220005_at	53829	P2RY13	5.92577905	0.68365953
1555728_a_at	51338	MS4A4A	5.81925535	1.2350735
206420_at	10261	IGSF6	5.50276993	2.04922533
203547_at	920	CD4	4.88890478	0.88238497
223501_at	10673	TNFSF13B	4.8865512	1.86070154
229560_at	51311	TLR8	4.65038549	3.06800511
223502_s_at	10673	TNFSF13B	4.55347873	1.1976272
1553043_a_at	146722	CD300LF	4.45139995	2.40584312
1560686_at	3681	ITGAD	3.84084741	2.47438706
208894_at	3122	HLA-DRA	3.69648554	1.19575071
209474_s_at	953	ENTPD1	3.61179585	0.95637045
203923_s_at	1536	CYBB	3.57941949	2.04518417
207085_x_at	1438	CSF2RA	3.51971929	1.95732683
206545_at	940	CD28	3.51596875	5.51258874
210982_s_at	3122	HLA-DRA	3.50865406	0.92892772
230550_at	64231	MS4A6A	3.45037119	1.58548569
1552280_at	91937	TIMD4	3.13113396	0.90850562
219385_at	56833	SLAMF8	3.07066988	2.07064868
211991_s_at	3113	HLA-DPA1	3.02040514	0.93500736

2.4 Discussion

L'utilisation de KANT sur deux jeux de données différents, l'un très hétérogène et l'autre plus homogène, a permis d'identifier dans les deux cas un ensemble d'antigènes putatifs pour des traitements immunothérapeutiques contre le cancer. Certains sont déjà connus, ciblés par des médicaments commercialisés ou en phase d'essai clinique (MUC1, CTLA-4) et d'autres inconnus ; ce qui nous montre l'intérêt de cette méthode. Cependant, plusieurs points sont à discuter.

Nous avons pu voir que la gestion des probesets n'est pas aisée. En effet, pour un même gène, les différents probesets n'obtiennent pas les mêmes résultats, que ce soit parce que certains probesets ne sont pas spécifiques, parce que la partie ciblée par le probeset dans le transcrit est dégradée et donc ne détecte pas le transcrit ou bien parce que le gène subit un phénomène d'épissage alternatif détecté par certains probesets. Concernant les probesets non spécifiques, nous avons utilisé le site ADAPT (Leong et al., 2005), qui pour chaque probeset Affymetrix a effectué l'alignement des probes sur le génome, afin de vérifier la spécificité des probesets utilisés. Ce service ne fonctionne plus aujourd'hui. Néanmoins, deux autres groupes de recherche ont effectué le même type d'étude et mis les résultats en libre accès : The Weizmann Institute of Science avec GeneAnnot (http://genecards.weizmann.ac.il/geneannot/index.shtml), faisant partie de l'ensemble Genecards (Rebhan et al., 1998) et le Center for Biological Sequence Analysis (CBS) avec Jetset (Li et al., 2011). Le premier fournit un score de sensitivité (fraction de probes dans un probeset alignés sur le bon gène) et de spécificité (somme des probes alignés sur le gène divisé par le nombre total de gènes sur lesquels elles s'alignent, le tout divisé par le nombre de probes) ; le second fournit un score de spécificité (fraction de probes détectant spécifiquement le gène), un score de couverture (fraction de transcrits détectés par le probeset), un score de robustesse (probabilité que la synthèse de la cible à détecter se finisse sans interruption jusqu'à la fin de la séquence s'hybridant sur le probeset) et un score total étant le produit des 3 précédents. Les informations de la deuxième méthode sont donc plus complètes. Les résultats obtenus, se basant sur des critères différents, ne sont pas les mêmes. Par exemple pour le gène BMPR1B, qui apparaissait dans nos résultats de cancer du sein sous 3 probesets différents : pour la première méthode le meilleur probeset est 210523_at car il a une meilleur sensitivité mais pour la deuxième méthode le meilleur est 229975_at car bien que sa spécificité soit plus faible obtient le meilleur score de robustesse, et le meilleur score total. C'est ce-dernier que nous avons utilisé pour les analyses de survie et d'appartenance aux sous-groupes moléculaires. Selon les critères utilisés, le meilleur probeset n'est pas toujours le même par gène mais les probesets non spécifiques obtiendront dans tous les cas de mauvais scores.

L'ensemble de la méthode n'est pas entièrement automatisé : pour pouvoir aller jusqu'à bout de la priorisation, il faut passer par une étape de revue approfondie de la littérature. En effet, un certain nombre d'informations ne sont pas prises en compte dans l'algorithme. Par exemple, la localisation précise de la protéine n'est pas considérée, une protéine transmembranaire peut être située au niveau d'une autre membrane que la membrane cellulaire et donc ne pas être accessible. De même, l'expression de la protéine dans certains tissus peut être problématique bien que nous ayons vérifié que le gène est sur-exprimé dans les cellules tumorales par rapport à l'ensemble des tissus, il peut avoir un niveau d'expression important dans des tissus « vitaux », rendant létal tout traitement la prenant pour cible.

Une autre limite de la méthode concerne son utilisation sur des données transcriptomiques, qui ne sont pas toujours associées aux mêmes résultats en protéomique. Néanmoins, les données protéomiques étant moins accessible, les premiers résultats restent intéressants. Cette limite souligne la nécessité d'une étape de validation biologique, que ce soit par l'utilisation de données de protéomique dans un premier temps, voire de *tissue array*. De plus, l'algorithme est utilisable pour d'autres types de données, que ce soit des données de séquençage ARN ou bien des résultats de spectrométrie de masse pour les protéines, à condition d'utiliser les mêmes gammes de valeurs d'expression ou bien de changer les seuils utilisés. Dans la définition initiale du projet, la validation des cibles trouvées sur le cancer du sein et le lymphome devait être faite par l'entreprise Transgène mais suite à une réorganisation interne, Transgène a arrêté la collaboration sans pouvoir aller au bout du projet.

La méthode de priorisation a été développée en prenant en compte les résultats des données de cancer du sein et du lymphome et en cherchant les meilleurs critères possibles pour définir un bon antigène. Elle pourrait être enrichie en développant de nouveaux poids par critère pour obtenir une priorisation des meilleurs biomarqueurs. En effet, une protéine peut avoir des effets secondaires importants si elle est ciblée par un traitement cytotoxique mais rester intéressante en tant que marqueur de la tumeur, en imagerie par exemple.

Nous avons utilisé des données publiques mais toutes générées par l'équipe CIT (pour les données de tumeurs), en utilisant les mêmes méthodes et avec des étapes de contrôles qualité régulières. C'est pourquoi, l'analyse de la qualité des données a été rapide et n'a pas été décrite. Néanmoins, dans le cas d'une méta-analyse à partir de données publiques de divers laboratoires, sur lesquelles nous n'avons aucun contrôle, il est important de s'assurer de la qualité des données et des annotations afin de ne pas biaiser les résultats de l'algorithme. Cette question sera plus approfondie au Chapitre 4 : Analyse du signal calcium des GBMs et gCSCs.

Chapitre 3 : Biomarqueurs des gCSCs

3.1 Objectifs

Nous avons vu dans le premier chapitre que les marqueurs connus des cellules souches de glioblastomes ne sont pas suffisants pour les caractériser. Ce problème est sans doute en partie dû à l'hétérogénéité des cellules souches qui, contrairement à celle des glioblastomes, a été peu étudiée. Quelques travaux ont tenté de répondre à cette question, comme l'étude de (Zorniak et al., 2012) définissant trois classes de glioblastomes à partir de l'expression de plusieurs protéines de différentes lignées neurales. Ces classes ont été définies à partir de cinq échantillons, ce qui limite la puissance des analyses et la robustesse des résultats.

Au-delà de la caractérisation de sous-groupes utiles au diagnostic et au pronostic, ces marqueurs pourraient être utilisés en imagerie moléculaire pour évaluer la réponse aux traitements. Il est également possible d'imaginer le développement de nouveaux médicaments de type 'anticorps conjugués' (cf chapitre 1) qui cibleraient spécifiquement les cellules souches en utilisant les protéines sur-exprimées à leur surface. Associés aux traitements conventionnels qui ciblent les cellules tumorales, ces médicaments pourraient ainsi augmenter l'efficacité de la prise en charge thérapeutique.

L'objectif de notre étude est de trouver de nouveaux biomarqueurs des gCSCs à partir d'un petit ensemble d'échantillons (4 cellules souches) analysés à la fois en spectrométrie de masse et transcriptomique (puces d'expression et séquençage ARN), ce qui nous donne accès à des informations plus complètes que la plupart des études précédentes.

Pour cela, la stratégie d'analyse s'est déroulée en 4 étapes :

- Description et définition des cellules souches étudiées : nous sommes partis d'un ensemble de données de cellules souches de glioblastomes adultes et pédiatriques sur puce HumanExon d'Affymetrix pour pouvoir vérifier si l'expression de ces cellules était différente, ce qui est le cas et nous a donc amené à nous concentrer sur les gCSCs issues de patients adultes. Nous avons ensuite étudié au niveau transcriptomique et protéomique l'hétérogénéité des 4 gCSCs étudiées.
- Recherche de biomarqueurs : à partir des données de protéomique et en utilisant l'algorithme KANT, développé dans cet objectif et présenté au chapitre 2, nous avons défini les protéines sur-exprimées dans nos 4 cellules souches par rapport aux contrôles HA (lignée d'astrocytes humains) et U87-MG (lignée cellulaire de glioblastome)

- Validation de ces biomarqueurs : nous avons vérifié la sur-expression des biomarqueurs potentiels sur des données de puces d'expression sur ces mêmes cellules et sur un ensemble public conséquent afin de valider leur expression à plus grande échelle
- Caractérisation des biomarqueurs : nous avons étudié l'expression des biomarqueurs trouvés dans des GBMs, ainsi que leur effet sur la survie de patient et avons analysé l'existence de différents transcrits dans nos cellules (données de séquençage ARN)

L'ensemble de cette étude s'est faite en partenariat avec les laboratoires du Dr Hervé Chneiweiss pour l'extraction et la mise en culture des gCSCs, et du Dr Sarah Cianférani pour l'analyse protéomique des échantillons.

3.2 Matériels & Méthodes

Echantillons analysés

- HA : culture primaire d'astrocytes humains (ATCC)
- U87-MG : lignée cellulaire de glioblastome (ATCC)
- CSN : cellules souches neurales de fœtus, préparées par le laboratoire du Dr Hervé Chneiweiss, Inserm U752, Hôpital Sainte-Anne, Paris (Patru et al., 2010)
- TG30, TG29, TG1, OB1, TG10, TG16 : cellules souches cancéreuses issues de patients atteints de glioblastomes, cultures réalisées au laboratoire du Dr Hervé Chneiweiss (Patru et al., 2010)
- TP54, TP59, TP80, TP83, TP84 : cellules souches cancéreuses issues de patients pédiatriques atteints de glioblastomes, cultures réalisées au laboratoire du Dr Hervé Chneiweiss (Thirant et al., 2011)
- TG1 dans différentes conditions : TG1 prolifératif jour 2 / TG1 prolifératif jour 2 + témolozomide / TG1 quiescent jour 2 / TG1 quiescent jour 2 + témolozomide / TG1 prolifératif jour 7 / TG1 prolifératif jour 7 + témolozomide / TG1 quiescent jour 7 / TG1 quiescent jour 7 + témolozomide

Les échantillons HA et U87-MG ont été cultivés selon les instructions du fournisseur.

Pour les gCSCs, des biopsies de 1mm^3 de tumeurs de patients atteints d'un glioblastome ont été réalisées et ont ensuite été placées en culture dans un milieu sans sérum avec des facteurs de croissance tels qu'EGF (Epidermal Growth Factor). Deux fois par semaine, les neurosphères formées sont dissociées mécaniquement et remises en culture dans les mêmes conditions. Environ cinquante passages sont nécessaires pour obtenir une population homogène de cellules souches cancéreuses. Ces cellules souches ont été précédement décrites dans (Patru et al., 2010).

Les cellules quiescentes sont obtenues en laissant les cellules dans l'incubateur sans renouvellement du milieu pendant 9 jours après un passage normal, comme décrit dans (Zeniou et al., 2015). La quiescence des cellules est vérifiée avec le test d'incorporation d'ethynyl desoxyuridine (EdU), test qui mesure la synthèse d'ADN dans les cellules. L'ajout de témolozomide est de 100 µM pendant 72h.

Le caractère 'souche' des cellules cancéreuses a été vérifié à l'aide de plusieurs tests sur la capacité à s'auto-renouveler, la clonogénicité, et la capacité à générer une nouvelle tumeur si ces cellules sont greffées dans le cerveau d'une souris immunodéficiente.

Puces d'expression

Les échantillons CSN, TG30, TG29, TG1, OB1, TG10, TG16, TP54, TP59, TP80, TP83 et TP84 ont été hybridés sur puce Affymetrix HumanExon.

Les échantillons OB1, TG10, TG16, HA et U87-MG ont été hybridés sur puce Affymetrix HG-U133 Plus 2. TG1 a également été hybridé sur puce Affymetrix HG-U133 Plus 2 mais dans différentes conditions (prolifératif/quiescent/après utilisation de témolozomide).

Toutes les puces ont été hybridées et scannées selon le protocole de la plateforme Biopuces et Séquençage de l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC).

Les jeux de données publics utilisés sont énumérés dans le Tableau 12 (seules les cellules souches de glioblastomes et les lignées cellulaires d'astrocytes humains ont été utilisées).

Tableau 12 : Jeux de données publics utilisés pour l'identification de biomarqueurs de gCSCs

Jeux de données	**Publication**	Echantillons utilisés dans l'analyse
GSE7181	(Beier et al., 2007)	6 cellules souches de glioblastomes
GSE23806	(Schulte et al., 2011)	36 lignées cellulaires de gliomes, 27 cellules souches de glioblastomes, 12 glioblastomes (GBM)
GSE18015	(Garcia et al., 2010)	Cellules de gliomes isolées, 8 CD133+ and 8 CD133-
GSE21514	(Moser and Fritzler, 2010)	2 astrocytes humains
GSE44841	(Aldaz et al., 2013)	8 cellules souches de glioblastomes
GSE46016	(Rheinbay et al., 2013)	10 cellules souches de glioblastomes
GSE46531	(Ye et al., 2013)	12 cellules souches de glioblastomes
GSE51822	(Zorniak et al., 2015)	2 cellules souches de glioblastomes

Les données des puces HG U133 Plus 2 ont été normalisées sous R à l'aide de l'algorithme justRMA du package affy (Gautier et al., 2004). Une étape de contrôle qualité des données et des annotations a été effectuée, elle est détaillée dans le chapitre suivant sur l'ensemble des gCSCs utilisées. Cette étape a abouti à la suppression de plusieurs échantillons : le jeu de données GSE18015, ainsi que quatre autres échantillons, laissant un total de 73 gCSCs en comptant les données publiques et les nôtres.

Les puces HumanExon ont été normalisé sous R à l'aide de l'algorithme RMA du package oligo (Carvalho and Irizarry, 2010) ; nous avons utilisé les données normalisées à partir des *core probeset* au niveau des transcrits.

Séquençage ARN

Les échantillons TG1, TG10, TG16, OB1, HA et U87-MG ont été séquencés par la plateforme Biopuces et Séquençage de l'IGBMC sur Illumina Hiseq 2500. Il s'agit d'un séquençage 'single-end' de 50 bases de long. Les librairies vont de 56 millions de lectures pour U87-MG, à 77 millions pour TG16. Le contrôle qualité a été effectué avec le package FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Les lectures ont été alignées par le logiciel STAR en deux passages (Dobin et al., 2013) sur le génome UCSC hg19, puis assemblées par RSEM (Li and Dewey, 2011). L'expression par gène et par transcrit du comptage des lectures a été normalisée à l'aide du package DESeq2 du logiciel R.

Les données du TCGA ont été téléchargées depuis cBioPortal (http://www.cbioportal.org/public-portal/) le 31 octobre 2014. Nous avons utilisé les données RNAseqV2 normalisées par gène. L'ensemble contient 174 échantillons, dont 170 ayant l'annotation classification.

Spectrométrie de masse

Les analyses de spectrométrie de masse ont été effectuées à l'Institut Pluridisciplinaire Hubert Curien (IPHC), UMR 7178 par le Dr Sarah Cianférani et Leslie Muller.

Les 6 échantillons (TG10, TG16, OB1, TG1, U87-MG, HA) ont été enrichis en protéines membranaires par lyse mécanique de 4.10^7 cellules conduisant à la formation de 2 à 3 mg de *ghosts*. 50 µg de la préparation sont déposés sur un gel d'électrophorèse 1D SDS-PAGE. Les protéines sont migrées de manière à les concentrer sur une seule bande, qui est ensuite traitée (réduction, alkylation des protéines présentes dans la bande), puis les protéines sont digérées à la trypsine pour générer les peptides. Ces peptides sont extraits du gel, puis analysés par nano-chromatographie liquide couplée à la spectrométrie de masse en tandem. Le couplage utilisée

est un couplage nono-LC (Nano-Acquity-WATERS, Mancherster, UK) nonoESI (ElectroSpray Ionisation)-Q (Quatripöle)-TOF (IMPACT-HD-BRUKER, Madison, USA). Les recherches de protéines ont été effectuées dans la banque de données Swiss-Prot à l'aide des moteurs Mascot (Matrix Science) et Omssa (NCBI). Puis, les résultats de ces algorithmes ont été combinés dans le logiciel Scaffold (Proteome Software) et des filtres de sélection ont été appliqués afin d'obtenir un taux de faux positifs inférieur à 1% (FDR<1%). Une liste de protéines a ainsi été validée pour chaque échantillon. Par ailleurs, à chaque identification est associée une donnée de quantification basée sur l'intégration des aires des pics chromatographiques du spectre MS de chaque peptide, obtenue à l'aide du logiciel Skyline. Pour chaque échantillon, l'analyse de spectrométrie est répétée sur 3 réplicats de préparation.

L'intégration des pics des peptides est vérifiée manuellement et les protéines ne sont considérées que si au moins 5 peptides sont identifiés dans au moins un échantillon. Pour calculer les valeurs d'expression des protéines par échantillon, on utilise la médiane d'expression des peptides pour une même protéine. Pour chaque échantillon, l'expression utilisée est la moyenne des 3 réplicats. Les données sont ensuite normalisées à l'aide de l'algorithme de normalisation par quantile du package *limma* du logiciel R. On soustrait ensuite le maximum d'expression sur l'ensemble des données, considéré comme le bruit de fond, puis l'ensemble est transformé en log2.

Analyse des données

Toutes les analyses ont été effectuées sous R.

L'analyses en composantes principales a été faite à l'aide du package ade4.

Concernant les *heatmap,* la classification hiérarchique des échantillons est faite par la fonction *hclust* en utilisant une distance de *pearson* (1-corrélation de pearson entre échantillons).

Les analyses de survie sont effectuées à l'aide du package *survival*, et les p-valeurs sont estimées par un test du log-rank.

Les analyses de sur-expression se basent sur l'algorithme KANT décrit dans le chapitre 2. Seule la première phase de KANT est utilisée, l'ensemble des tissus du corps humain n'est pas utilisé en deuxième contrôle car cette analyse a pour objectif de valider les résultats de protéomiques, nous avons donc utilisé les mêmes contrôles en protéomique et transcriptomique, c'est-à-dire HA.

Immunohistochimie

Les analyses ont été effectuées par Jihu Dong, doctorant au Laboratoire d'Innovation Thérapeutique, Strasbourg.

Les cellules sont incubées avec les anticorps primaires (anti-CD205, Abcam, 1-50 ; anti-CD56, Biolegend, 1-50 ; anti-CD97, Merck-Milipore, 1-50) pendant une nuit à 4°C, avant d'être incubées avec l'anticorps secondaire correspondant (Jackson ImmunoResearch) pendant 20 minutes à température ambiante. Après une dernière incubation pendant 20 minutes à température ambiante avec ExtrAvidin-Peroxidase (Sigma-Aldrich, 1 :500), le signal est détecté avec un kit DAB substrate (BD Biosciences).

3.3 Résultats

3.3.1 Description des cellules souches étudiées

Comme vu dans le Chapitre 1, les glioblastomes touchent les adultes et les enfants. La classification de Sturm, qui se base sur des données génétiques et épigénétiques, établit les groupes de glioblastomes en prenant les deux en compte, mais quatre groupes sur six sont liés à l'âge.

Disposant de cellules souches de glioblastomes adultes et pédiatriques, nous avons d'abord souhaité savoir si ces cellules sont différentes au niveau transcriptomique. Pour cela, nous avons effectué une analyse en composantes principales des échantillons disponibles sur puce Affymetrix HumanExon (Figure 26). Chaque cellule souche de glioblastome est présente en trois échantillons, qui correspondent chacun à un passage différent des cellules. On remarque une nette séparation entre tumeurs adultes et pédiatriques, avec les cellules souches neurales entre les deux. Seuls les échantillons pédiatriques TP54 et adultes TG16 semblent être entre les deux zones. On remarque aussi que les trois passages des cellules représentent une certaine hétérogénéité.

La différence entre cellules souches adultes et pédiatriques, en accord avec la classification de Sturm nous amène à ne pas étudier ces types de cellules ensemble. Nous nous focaliserons par la suite sur les cellules souches adultes. De plus, vu les résultats sur les passages cellulaires, **nous considérerons dans la suite différents passages comme différents échantillons de cellules souches.**

Nous cherchons à identifier des biomarqueurs des cellules souches à partir des données de protéomique. En effet, comme vu dans le chapitre 1, il existe différentes étapes de régulation entre l'expression d'un ARN messager et l'expression de la protéine. Pour l'identification de biomarqueurs qui soient utilisables en clinique, les données de protéomiques sont donc plus pertinentes que les données de transcriptomique.

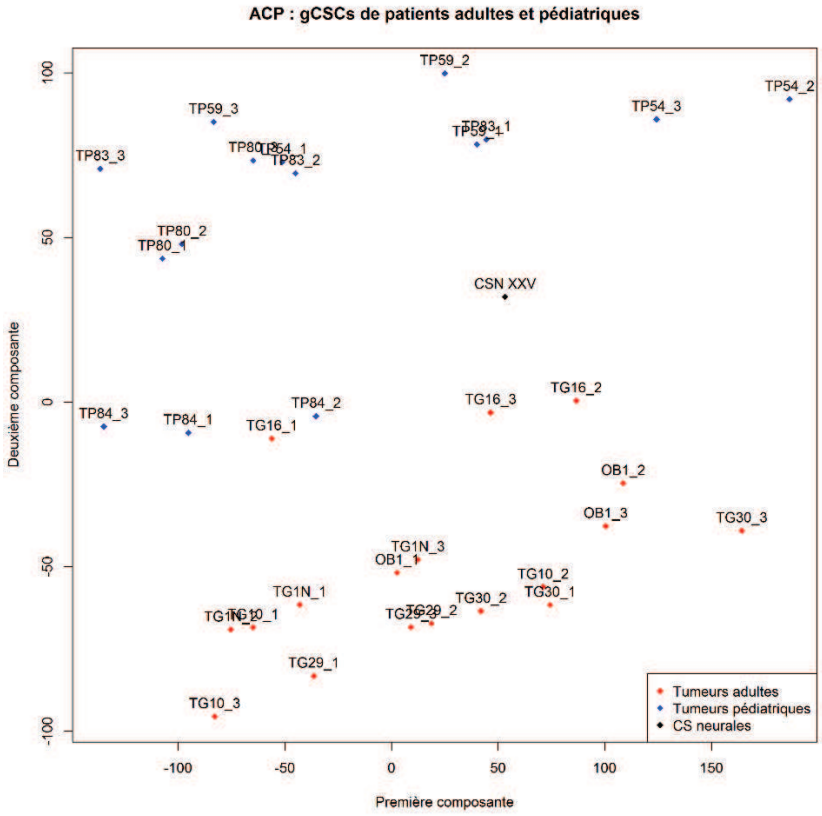

Figure 26 : ACP de données transcriptomique (puce HumanExon) de cellules souches adultes et pédiatriques de glioblastomes, et de cellules souches neurales (CSN).
Chaque cellule souche de glioblastome est présente en trois exemplaires, correspondant à trois passages différents des cellules (notés « _x » dans le nom des échantillons).

Seules quatre cellules souches adultes ont été analysées en protéomique : TG10, TG16, OB1 et TG1, ainsi qu'U87-MG (lignée cellulaire de glioblastome) et HA (lignée cellulaire d'astrocytes humains). Ces données ont aussi été analysées sur puce Affymetrix HG-U133 Plus 2 et par séquençage ARN. Nous n'avons donc pas utilisé d'avantage les données HumanExon, pour différentes raisons : il semble que l'échantillon TG10 sur puce HumanExon soit différent de celui utilisé sur l'ensemble des autres plateformes (expression des gènes, proximité avec TG16 plutôt qu'avec les autres échantillons) et il y a peu de données publiques hybridées sur cette puce.

Nous avons donc cherché à caractériser davantage ces quatre cellules. Pour cela, la première étape a été de faire une analyse de classification hiérarchique et de la visualiser sous forme de *heatmap* à partir des données de transcriptomique (séquençage de l'ensemble des ARNm) et protéomique (données de spectrométrie de masse des Clusters de Différenciations (CDs)).

Ces analyses (Figure 27) nous montrent la séparation des cellules souches en deux groupes distincts : d'un côté on trouve TG10/TG16 et de l'autre TG1/OB1. U87-MG et HA sont plus éloignés des cellules souches.

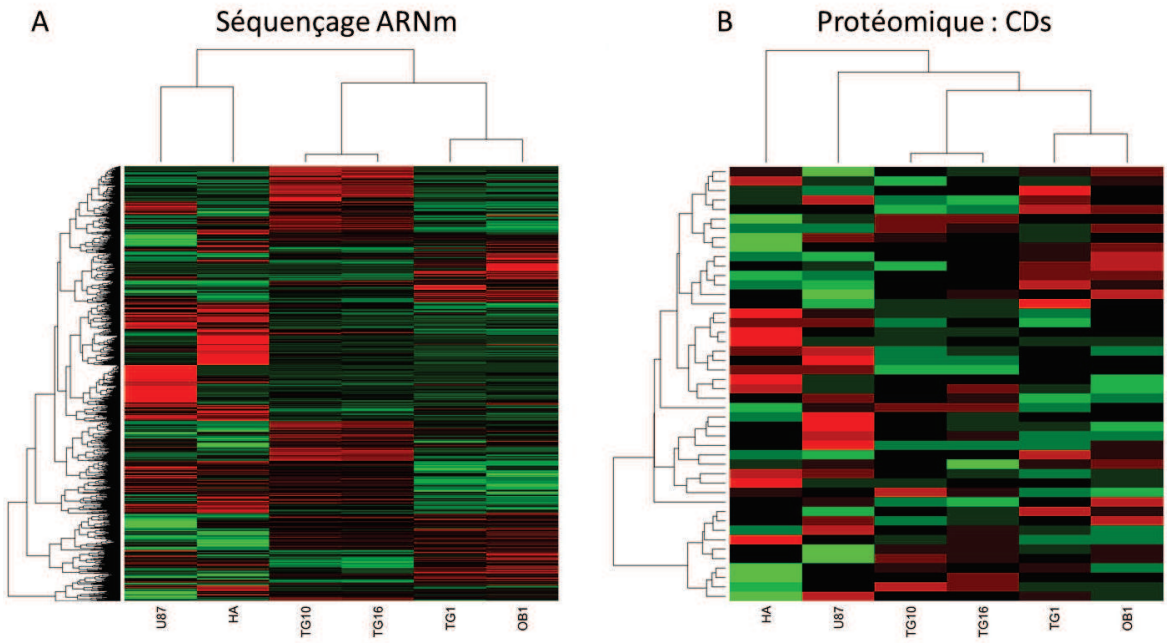

Figure 27 : Heatmap des quatre gCSCs, HA et U87-MG. A) à partir des 5000 gènes les plus variants en séquençage ARNm. B) à partir des CDs exprimés en protéomique

Nous avons vu dans la première partie qu'un certain nombre de marqueurs des gCSCs sont connus, dont des CDs tels que CD133, CD44, CD171 (ou L1CAM), CD15 (ou SSEA1), CD49f (ou ITGA6). Nos cellules souches n'expriment pas le marqueur principal CD133, mais parmi l'ensemble des marqueurs précédents, on retrouve CD44, C171 et CD49f (Tableau 13). L'expression de ces CDs ne semble pas caractéristique des cellules souches de glioblastomes en général, car on retrouve des valeurs plus élevée dans U87-MG (pour CD44) et HA (pour ITGA6 et L1CAM). Néanmoins, il semble que leur expression permette de différencier les deux sous-groupes OB1/TG1 et TG10/TG16, le premier exprimant davantage ITGA6 et un peu moins CD44 par rapport au deuxième. **Ces résultats confirment la non spécificité des marqueurs de gCSCs connus et la nécessité d'en définir de nouveaux.**

Tableau 13 : Expression des CDs marqueurs des gCSCs parmi nos données de spectrométrie de masse

	OB1	TG1	TG10	TG16	U87-MG	HA
ITGA6	7,17	8,94	6,11	6,27	5,26	7,37
L1CAM	5,97	5,73	6,72	5,91	5,82	6,08
CD44	8,58	8,77	9,05	9,16	9,50	8,49

Une autre caractéristique de TG10/TG16 par rapport à OB1/TG1 est la mutation de TP53, plutôt caractéristique du sous-type proneural, il semble donc que les gCSCs étudiés soient à l'origine de GBMs de différentes classes.

3.3.2 Biomarqueurs potentiels

Afin d'identifier des biomarqueurs potentiels des cellules souches de glioblastomes, nous avons utilisé l'algorithme de sur-expression de KANT sur les données de protéomique en comparant les quatre cellules souches à HA. Puisque l'algorithme travaille par sous-groupe, il devrait nous permettre de trouver les biomarqueurs de l'ensemble des gCSCs et ceux spécifiques des deux sous-groupes.

Sur l'ensemble des Cluster de Différenciations (CDs), 46 ont été trouvés exprimés selon nos critères (vérification manuelle de l'intégration des bons « pics » sur le spectre MS et détection d'au moins 5 peptides par protéine dans au moins un échantillon). D'après l'algorithme, 18 sont sur-exprimés par rapport à HA (voir Tableau 14). Nous avons choisi HA comme référence, plutôt qu'U87-MG en se basant sur l'hypothèse qu'U87-MG, lignée cellulaire de glioblastomes, ne représente plus vraiment les cellules de glioblastomes du fait de sa mise en culture mais a du acquérir un certain nombre de caractéristiques propres aux cellules souches, et donc aux cellules souches cancéreuses.

L'algorithme nous permet de ressortir en premier les CDs avec le delta d'expression le plus grand entre protéines sur-exprimées et HA.

Nous avons éliminé les protéines non-spécifiques car très exprimées dans U87-MG : TFRC (Transferrin receptor protein 1 ou CD71), SCARB2 (Lysosome membrane protein 2 ou CD36), LAMP1 (Lysosome-associated membrane protein 1 ou CD107A), LAMP2 (*Lysosome-associated membrane protein 2* ou CD107B), et CD44. On remarque que la moitié des protéines sont sur-exprimées dans les quatre échantillons.

Il nous reste 13 protéines potentiellement intéressantes en tant que biomarqueurs des gCSCs, dont 3 sont spécifiquement sur-exprimées dans le groupe OB1/TG1 (CD63, CD109 et CD276).

Il n'existe pas de données publiques de protéomique qui nous permettraient de valider ces protéines sur un ensemble plus important. C'est pourquoi nous avons décidé d'utiliser les données publiques d'expression de gènes (puce HG-U133 Plus 2.0) afin de vérifier la sur-expression des gènes associés aux protéines d'intérêt de nos quatre gCSCs. Les résultats sont donnés dans le Tableau 15.

Tableau 14: CDs sur-exprimés dans les gCSCs (TG1, TG10, TG16, OB1) par rapport à HA sur les données de protéomique.

L'expression d'HA et U87-MG est donnée pour information, ainsi que le Delta d'expression entre la médiane des gCSCs et HA

Nom du gène	CD	Score	Echantillons sur-exprimés	HA	U87-MG	Delta
F11R	CD321	10.70	TG10, TG16, OB1	3.98	7.23	3.83
NCAM1	CD56	3.49	TG1, TG10, TG16, OB1	6.23	6.12	1.80
LY75	CD205	3.39	TG1, TG10, TG16, OB1	5.16	5.48	1.76
CD97	CD97	3.30	TG1, TG10, TG16, OB1	5.69	6.89	1.72
TFRC	*CD71*	*3.16*	*TG1, TG10, TG16, OB1*	*7.75*	*9.10*	*1.66*
IGF2R	CD222	2.81	TG1, TG10, TG16, OB1	6.52	7.12	1.49
ATP1B3	CD298	2.48	TG1, TG10, TG16, OB1	6.12	7.95	1.31
IGF1R	CD221	1.96	TG1, TG10, TG16, OB1	4.85	4.58	0.97
DDR1	CD167A	1.93	TG1, TG10, TG16, OB1	5.41	5.11	0.95
SCARB2	*CD36*	*1.84*	*TG1, TG10, TG16*	*8.32*	*9.41*	*1.30*
LAMP1	*CD107A*	*1.65*	*TG1, TG10, TG16, OB1*	*9.01*	*10.32*	*0.72*
LAMP2	*CD107B*	*1.53*	*TG10, TG16*	*8.57*	*9.31*	*1.62*
SLC44A1	CD92	1.28	TG1, TG16, OB1	6.17	6.15	0.77
BCAM	CD239	1.27	TG16, OB1	6.29	4.42	1.35
CD63	CD63	1.21	TG1, OB1	4.91	6.03	1.27
CD109	CD109	1.11	TG1, OB1	7.12	6.84	1.16
CD276	CD276	1.00	TG1, OB1	7.39	7.51	0.99
CD44	*CD44*	*0.74*	*TG10, TG16*	*8.50*	*9.54*	*0.56*

Il n'y a pas de probeset pour DDR1, nous n'avons donc pas pu vérifier son expression dans les données publiques.

Les gènes IGF2R (*Insulin-Like Growth Factor 2 Receptor*), ATP1B3 (*ATPase, Na+/K+ Transporting, Beta 3 Polypeptide*), BCAM (*Basal Cell Adhesion Molecule*), CD63 et CD109 ne sont pas sur-exprimés dans les 4 gCSCs par rapport à HA, bien que les protéines associées le soient. Pour ces gènes/protéines, il semble que l'expression des transcrits ne soit pas corrélée à celle de la protéine. Différentes raisons peuvent expliquer ce phénomène, comme le fait que la durée de vie des transcrits et des protéines ne soient pas la même (voir Chapitre 1, partie 4). On ne peut donc pas utiliser les données transcriptomiques sur ces protéines. Nous n'allons pas les considérer dans l'identification de biomarqueurs potentiels car il nous est impossible de savoir si leur sur-expression est une spécificité de nos 4 gCSCs ou s'il s'agit d'une propriété commune à un ensemble de gCSCs.

Tableau 15 : Résultats de l'algorithme de sur-expression KANT sur les puces U133 Plus 2 pour les protéines sélectionnées en tant que biomarqueurs potentiels.

On retrouve les résultats sur les 4 gCSCs précédents (OB1, TG1, TG10 et TG16) par rapport à HA, et les résultats sur l'ensemble des données publiques par rapport à 3 échantillons d'HA différents.

Probeset	Gène	Score (4 gCSCs)	gCSC sur-exprimés	Score total	Echantillons sur-exprimés (73)
223000_s_at	F11R	8,93	OB1, TG10, TG16	0,11	2
212843_at	NCAM1	5,41	OB1, TG1, TG10, TG16	3.28	54
205668_at	LY75	24,30	OB1, TG1, TG10, TG16	5,21	35
202910_s_at	CD97	1,02	OB1, TG1	0,43	18
201392_s_at	IGF2R	0,00		0,00	0
208836_at	ATP1B3	0,00		0,00	0
203628_at	IGF1R	2,83	OB1, TG1, TG10, TG16	0,08	3
	DDR1				
222364_at	SLC44A1	5,02	OB1, TG1, TG10, TG16	0,72	25
203009_at	BCAM	0,00		0,00	0
200663_at	CD63	0,00		0,00	0
226545_at	CD109	0,00		0,06	3
1559583_at	CD276	0,80	OB1, TG1	0,71	23

Quant aux gènes F11R (*F11 Receptor*) et IGF1R (*Insulin-Like Growth Factor 1 Receptor*), bien qu'étant sur-exprimés en transcriptomique et protéomique dans nos gCSCs, on ne les retrouve pas dans l'ensemble de données publiques (dans seulement 2 échantillons sur 77). Deux explications sont possibles : soit cela vient du contrôle utilisé (HA) car pour l'ensemble public 3 échantillons de HA différents font office de contrôle au lieu d'un seul dans nos données, soit ce sont nos gCSCs qui expriment particulièrement cette protéine. Pour F11R, les 2 échantillons sur-exprimés sont TG10, TG16. Ces résultats se retrouvent sur les 4 probesets représentant F11R. Il semble donc que la deuxième hypothèse soit la bonne. Pour IGF1R, il existe 5 probesets montrant des résultats très différents : le premier (243358_at) est sur-exprimé dans 17 échantillons publics mais pas nos 4 gCSCs de départ, le deuxième (225330_at) est sur-exprimé dans 5 échantillons publics, faisant tous partis des gCSCs de départ (plusieurs passages de TG1 dans différentes conditions font partis des données), le troisième (2036628_at) est sur-exprimé dans 3 échantillons publics mais pas nos gCSCs, c'est celui qui était sur-exprimé par rapport à « notre » HA dans la première partie de l'algorithme, les quatrième et cinquième ne sont pas sur-exprimé. Il semble que pour IGF1R, la valeur du contrôle HA ne soit pas la même dans les données publiques, de même que les résultats sur l'ensemble des gCSCs par rapport à nos 4 données. Puisqu'aucun probeset n'est sur-exprimé pour nos gCSCs et des données publiques, et que nous pensons que la protéine est sur-exprimée dans nos gCSCs, il est difficile de conclure sur l'ensemble plus large des données publiques.

Finalement, les gènes/protéines les plus intéressants sont NCAM1 (*Neural Cell Adhesion Molecule 1*), LY75 (*Lymphocyte Antigen 75*), CD97 et SLC44A1 (*Solute Carrier Family 44 (Choline Transporter), Member 1*). Le dernier, CD276 présente un score faible en protéomique et transcriptomique, nous avons donc décidé de ne pas le prendre en compte dans la suite des analyses. NCAM1 est représenté par 4 *probesets* différents sur la puce, tous les 4 sur-exprimés par un ensemble de données publiques important (de 22 à 57 échantillons). LY75 et CD97 ne sont représentés que par un probeset. SLC44A1 est représenté par 5 probesets différents, tous sur-exprimés par un sous-ensemble des échantillons publics (9 à 35 échantillons).

Nous avons 4 protéines sur-exprimées dans nos gCSCs d'après les analyses de spectrométrie de masse, et qui semblent l'être aussi dans des ensembles plus importants de données publiques d'après les analyses de transcriptomiques.

3.3.3 Caractérisation des biomarqueurs

Au niveau de nos quatre échantillons qui semblent se répartir en deux groupes distincts, bien que les quatre protéines d'intérêt soient sur-exprimées sur l'ensemble des données (pour SLC44A1, TG10 n'est pas noté comme sur-exprimé bien que son expression soit supérieure à celle de HA car l'algorithme considère comme sur-exprimé un échantillon supérieur à l'expression du contrôle plus un seuil de 0.5), nous avons voulu savoir s'il y avait une différence d'expression entre les deux groupes TG1/OB1 et TG10/TG16. Le Tableau 16 récapitule l'expression pour chacune des quatre protéines sur l'ensemble des données. On remarque pour NCAM1 une expression plus importante dans les cellules OB1 et TG10 et pour LY75 une expression plus importante dans TG1 et OB1. Seule cette dernière semble donc être caractéristique de l'un des deux sous-groupes.

Tableau 16 : Expression des protéines d'intérêt dans les cellules souches, U87 et HA.
L'expression est donnée en log2.

	TG1	OB1	TG10	TG16	U87-MG	HA
CD97	7.51	7.75	7.31	7.05	6.89	5.69
SLC44A1	6.95	7.17	6.58	6.70	6.15	6.17
NCAM1	6.84	8.32	8.42	7.75	6.12	6.23
LY75	7.39	7.51	6.45	6.01	5.48	5.16

3.3.3.1. CD97

CD97 est une protéine de la famille des RCPGs (Récepteurs Couplés aux Protéines G), et plus précisément de la sous-famille EGF-TM7, représentant des protéines hybrides possédant un domaine transmembranaire traversant 7 fois la membrane cellulaire (TM7) et une partie extracellulaire ayant plusieurs domaines EGF-like (*Epidermal Growth Factor like*). Le nombre de

domaines dépend de l'isoforme de CD97 : l'isoforme principal a 5 domaines EGF (notés 1 à 5), le deuxième isoforme en possède 4 (1,2,3,5), le plus petit en possède 3 (1,2 et 5).

CD97 est exprimé à la surface des lymphocytes, monocytes, macrophages, cellules dendritiques, granulocytes et cellules des muscles lisses (Eichler et al., 1997).

CD97 a été trouvé exprimé dans différents cancers tels que la thyroïde, l'estomac, l'œsophage, le pancréas, le colon, ainsi que le glioblastome (Aust et al., 2002; Safaee et al., 2015; Steinert et al., 2002). Dans le cancer, différentes études ont montré que CD97 conférait un phénotype invasif et stimulait l'angiogenèse (Galle et al., 2006; Wang et al., 2005) et il a été décrit comme associé à l'invasion et la migration dans les glioblastomes (Safaee et al., 2013). Une étude plus récente, du même groupe (Safaee et al., 2013) a montré que CD97 était exprimé dans les GBMs et gCSCs mais pas dans les gliomes de bas grade. De plus, les auteurs ont trouvé que seuls 2 isoformes, EGF (1,2,5) et EGF(1,2,3,5) étaient exprimaient dans les tumeurs. Or l'isoforme EGF(1,2,5) est connu pour favoriser la croissance des tumeurs, aider à la migration et promouvoir l'invasion métastique dans différents cancers (Galle et al., 2006; Liu et al., 2012a). Enfin, à partir des données du TCGA, ils ont montré que l'expression de CD97 était plus élevée dans les groupes classiques et mésenchymateux que dans les autres.

A partir de nos données de séquençage ARNm sur nos 4 gCSCs, nous avons pu vérifier l'expression des différents isoformes de CD97 (Tableau 17). 3 isoformes semblent exprimé dans l'ensemble des données (gCSCs, U87 et HA) mais EGF(1,2,3,4,5) est très faiblement exprimé partout (en terme de pourcentage). Les isoformes majoritaires sont donc, comme décrit dans la littérature, EGF(1,2,3,5) et EGF(1,2,5) qu'on retrouve dans nos 6 échantillons. Même si l'expression de CD97 est plus importante pour TG1/OB1 que TG10/TG16, la répartition des isoformes ne semblent pas être différente. HA exprime davantage EGF(1,2,5), l'isoforme lié à la croissance et l'invasion des tumeurs, bien que non cancéreux.

Tableau 17 : Expression des isoformes de CD97, d'après les données de séquençage ARNm

L'expression est donnée en log2, le % indiqué correspond au pourcentage d'expression par isoforme par échantillon.

Transcrit	Protéine	TG1	OB1	TG10	TG16	U87-MG	HA
ENST00000242786.5	EGF(1,2,3,4,5)	7,64	7,75	5,90	5,82	6,04	2,91
		7%	7%	5%	5%	4%	0%
ENST00000357355.3	EGF(1,2,3,5)	9,86	10,32	8,64	8,30	9,18	7,17
		34%	43%	32%	32%	39%	11%
ENST00000358600.3	EGF(1,2,5)	10,53	10,41	9,31	9,10	9,55	9,99
		60%	50%	57%	62%	57%	88%
ENST00000587728.1	Pas de protéine	0,00	0,00	3,71	0,00	0,00	0,00
		0%	0%	6%	0%	0%	0%

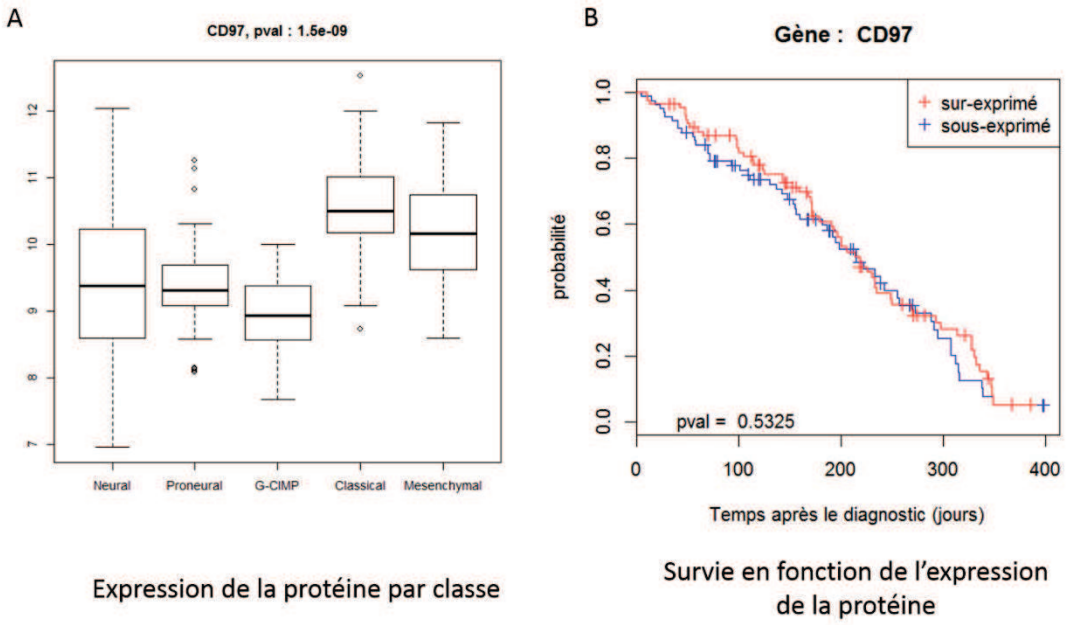

Figure 28 : Expression de CD97 dans les classes du glioblastomes (A) et effet sur la survie (B)

En reprenant les données du TCGA sur le glioblastome (données de séquençage ARNm), nous retrouvons la sur-expression de CD97 dans les sous-types classiques et mésenchymateux mais l'analyse de survie sur l'ensemble des GBMs ne semblent pas indiquer que CD97 soit associé à un mauvais pronostic (Figure 28).

3.3.3.2. SLC44A1

SLC44A1 (ou CTL1, *Choline Transport-Like Protein 1*) fait partie de la famille des transporteurs de choline. La choline a un rôle essentiel pour la synthèse du phospholipide le plus abondant dans les membranes cellulaires, la phosphatidylcholine (PC), du donneur de méthyl bétaine et du neurotransmetteur acetylcholine (Ach).

Différentes études ont montré une dérégulation des niveaux de choline des cellules cancéreuses (Hara et al., 2003; Kwee et al., 2006), qui serait corrélée avec la progression maligne de ces cellules (Glunde et al., 2004, 2006). L'utilisation de la Tomographie à Emission de Positions (TEP) SCAN avec la ^{11}C-choline ou ^{18}F-choline permet de détecter différentes tumeurs, telles que les gliomes, cancers de la prostate, du poumon ou de l'œsophage (Hara et al., 1997, 1998, 2000, 2003).

Le transporteur SLC44A1 est présent dans différents tissus, dont le cerveau et le colon. On le retrouve anormalement exprimé dans des lignées cellulaires cancéreuses de poumon et gliomes (Machová et al., 2009; Wang et al., 2007).

L'inhibition de l'expression d'ARNm de SLC44A1 par *silencing RNA* inhibe la viabilité cellulaire dans les cellules du carcinome du poumon à petites cellules (SCLC), ce qui en ferait une bonne cible thérapeutique (Inazu et al., 2013). De plus, l'inhibition du transporteur par des cations peut promouvoir la mort cellulaire par apoptose dans les SCLC et cellules leucémiques. Ces résultats montrent qu'une inhibition fonctionnelle de SLC44A1 pourrait entrainerait la mort cellulaire, ce qui fait donc de SLC44A1 une cible de choix dans le traitement du cancer.

A partir de nos données de séquençage ARNm, nous avons pu vérifier l'expression des différents transcrits de SLC44A1 (Tableau 18). Deux transcrits sont exprimés majoritairement et un troisième est très peu présent. Il ne semble pas y avoir de différence en terme de proportions entre transcrits entre les échantillons, que ce soit au niveau des gCSCs ou de U87 et HA.

Tableau 18 : Expression des isoformes de SLC44A1, d'après les données de séquençage ARNm
L'expression est donnée en log2, le % indiqué correspond au pourcentage d'expression par isoforme par échantillon.

Transcrit	TG1	OB1	TG10	TG16	U87-MG	HA
ENST00000343170.7	11.08	12.45	10.42	10.51	9.84	10.36
	61%	60%	59%	61%	66%	64%
ENST00000374720.3	0.00	0.00	0.00	0.00	0.00	0.00
	0%	0%	0%	0%	0%	0%
ENST00000374723.1	0.00	0.00	0.00	0.00	0.00	0.00
	0%	0%	0%	0%	0%	0%
ENST00000374724.1	10.77	12.18	10.17	10.22	9.22	9.91
	38%	38%	39%	38%	33%	36%
ENST00000607692.1	3.32	5.61	3.86	2.77	2.26	1.21
	1%	2%	3%	1%	1%	0%

Nous avons aussi regardé l'expression de SLC44A1 dans les glioblastomes, à partir des données du TCGA (Figure 29). On remarque, que l'expression de SLC44A1 est davantage associée aux groupes neuraux et proneuraux, y compris aux échantillons présentant un phénotype G-CIMP. L'analyse de survie ne donne pas de résultat significatif.

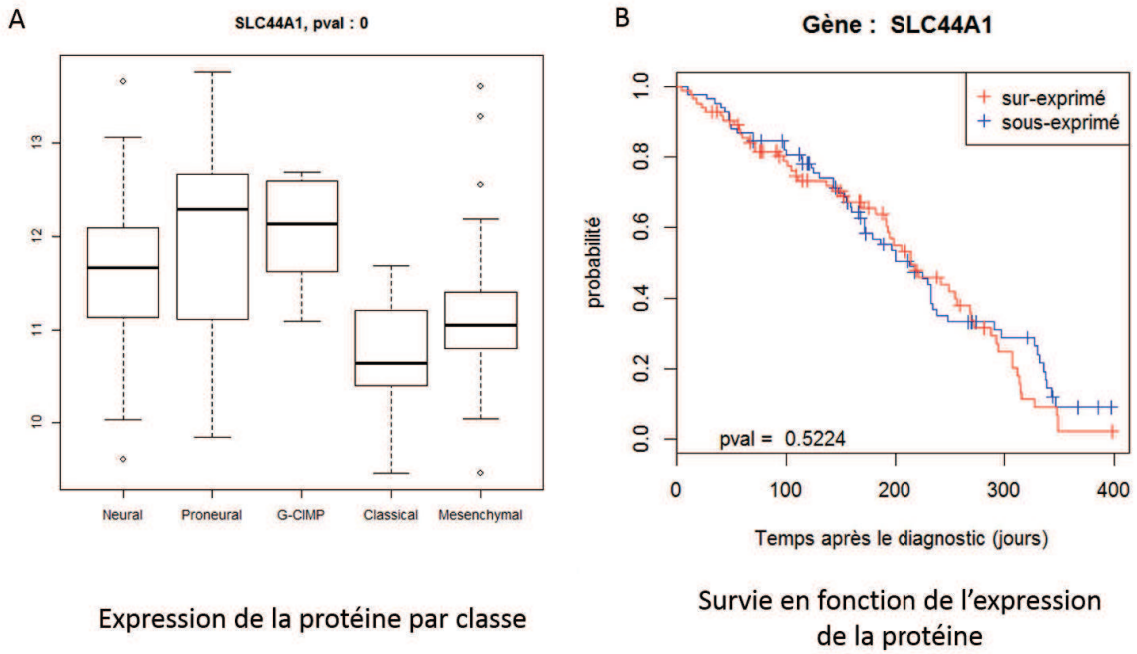

Figure 29 : Expression de SLC44A1 dans les classes du glioblastomes (A) et effet sur la survie (B)

3.3.3.3. NCAM1

NCAM (ou CD56) est un membre de la famille des glycoprotéines situées à la surface des cellules. Cette protéine joue un rôle majeur dans le développement et la plasticité du système nerveux et est impliquée dans les mécanismes d'apprentissage et de mémorisation (Cambon et al., 2004). Dans le cancer, CD56 est spécifiquement exprimée dans les tumeurs neuroendocrines (médulloblastomes, astrocytomes) (Etzell et al., 2006), on la retrouve aussi dans les tumeurs du rein (Ronkainen et al., 2010). La sur-expression de NCAM1 est associée à des tumeurs agressives et une survie plus faible dans différents types de tumeurs, tels que les leucémies (Raspadori et al., 2001, 2002; Ravandi et al., 2002) , les mélanomes malins (Abbott et al., 2004; Johnson, 1999) et différents carcinomes (poumon, ovaire, prostate, sein, colon, rein) (Cho et al., 2006; Choi et al., 2004; Daniel et al., 2003; Evans et al., 2006; Pujol et al., 1993; Zołtowska et al., 2001).

A partir de nos données de séquençage ARNm, nous avons pu vérifier l'expression des différents transcrits de NCAM1 (Tableau 19). Les deux premiers isoformes ne sont exprimés dans aucun échantillon mais les deux derniers semblent exprimés dans des proportions semblables quelque soient les données.

Tableau 19 : Expression des isoformes de NCAM1, d'après les données de séquençage ARNm

L'expression est donnée en log2, le % indiqué correspond au pourcentage d'expression par isoforme par échantillon.

Transcrits	TG1	OB1	TG10	TG16	U87-MG	HA
ENST00000316851.7	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
ENST00000397957.4	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
ENST00000401611.2	7.29 86%	9.95 82%	9.68 80%	8.44 79%	5.24 86%	6.21 77%
ENST00000533760.1	5.96 14%	9.06 18%	8.92 20%	7.82 21%	3.90 14%	5.74 22%

Quant à l'expression de NCAM1 dans le glioblastome, on observe une différence entre les classes de glioblastome, avec une expression plus importante dans les classes proneurales et G-CIMP. Sur l'ensemble des glioblastomes, l'expression de NCAM1 ne semble pas liée à la survie des patients (Figure 30).

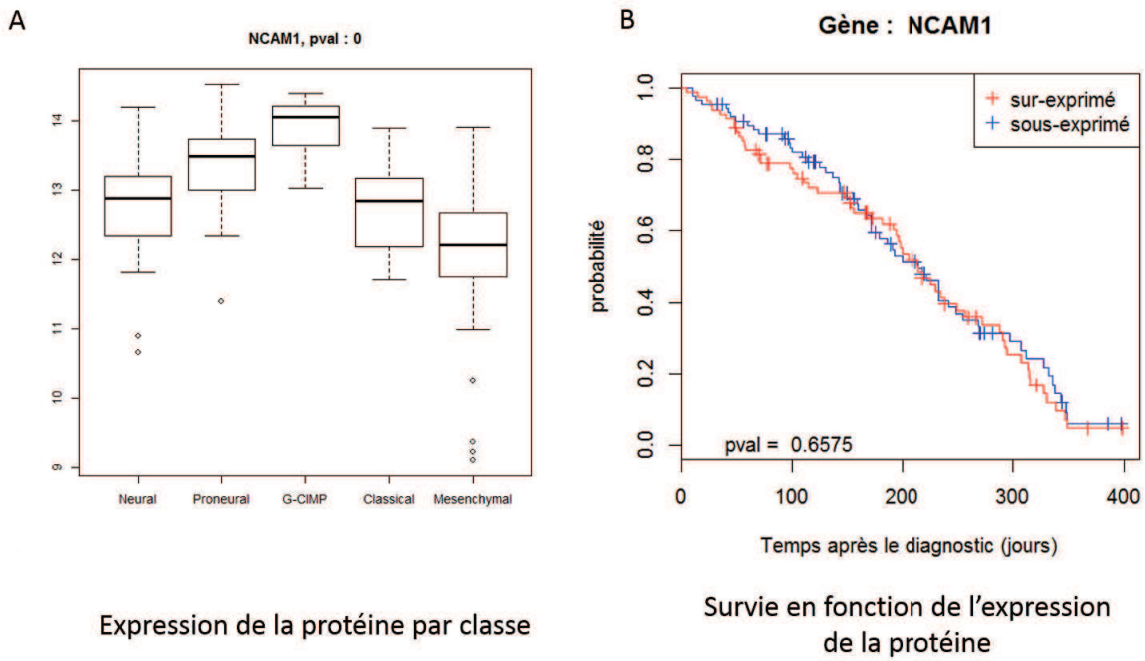

Figure 30 : Expression de NCAM1 dans les classes du glioblastomes (A) et effet sur la survie (B)

3.3.3.4. LY75

LY75 (ou DEC-205) fait partie de la famille des protéines réceptrices de mannose. Cette protéine est impliquée dans le processus de présentation des antigènes de classe I et II permettant d'initier la réponse immunitaire, et dans la reconnaissance des cellules apoptotiques ou nécrotiques

(Shrimpton et al., 2009). Ce récepteur a été reporté comme très exprimé dans les cellules dendritiques myéloïdes et monocytes, modérément exprimé dans les cellules B, et exprimé à un niveau faible dans les cellules NK, T et les cellules dendritiques plasmacytoïdes (Witmer-Pack et al., 1995). L'expression de CD205 a été observée dans le cancer des ovaires, dans des cellules stimulée par l'expression de l'interleukine 6 (Il6), montrant son rôle dans l'adhésion à la matrice extracellulaire et la formation de métastases (Giridhar et al., 2011).

Concernant les transcrits exprimés dans nos données (Tableau 20), il semble que leur expression soit spécifique des échantillons. En effet U87 et HA expriment majoritairement le transcrit n°3 alors qu'on ne le retrouve qu'à un faible pourcentage dans TG1 et pas du tout dans les autres cellules souches. TG1 et OB1, qui expriment davantage le gène, expriment à 95% le transcrit 1, alors que TG10/TG16 l'expriment à 60%, 20% pour le transcrit 2 et 20% pour le transcrit 4.

Concernant l'expression de la protéine dans les glioblastomes (Figure 31), on remarque une expression plus importante dans les sous-groupes mésenchymateux et neural.

Tableau 20 : Expression des isoformes de LY75 d'après les données de séquençage ARNm

L'expression est donnée en log2, le % indiqué correspond au pourcentage d'expression par isoforme par échantillon.

Transcrits	TG1	OB1	TG10	TG16	U87-MG	HA
ENST00000263636.4	12.60	12.57	9.29	9.28	1.82	2.91
	95%	95%	62%	62%	12%	13%
ENST00000492955.1	5.44	4.57	5.38	5.11	0.00	1.85
	3%	2%	20%	17%	0%	27%
ENST00000553424.1	6.33	0.00	0.00	0.00	4.02	4.62
	2%	0%	0%	0%	88%	60%
ENST00000554112.1	3.86	7.32	7.24	7.41	0.00	0.00
	0%	3%	18%	21%	0%	0%

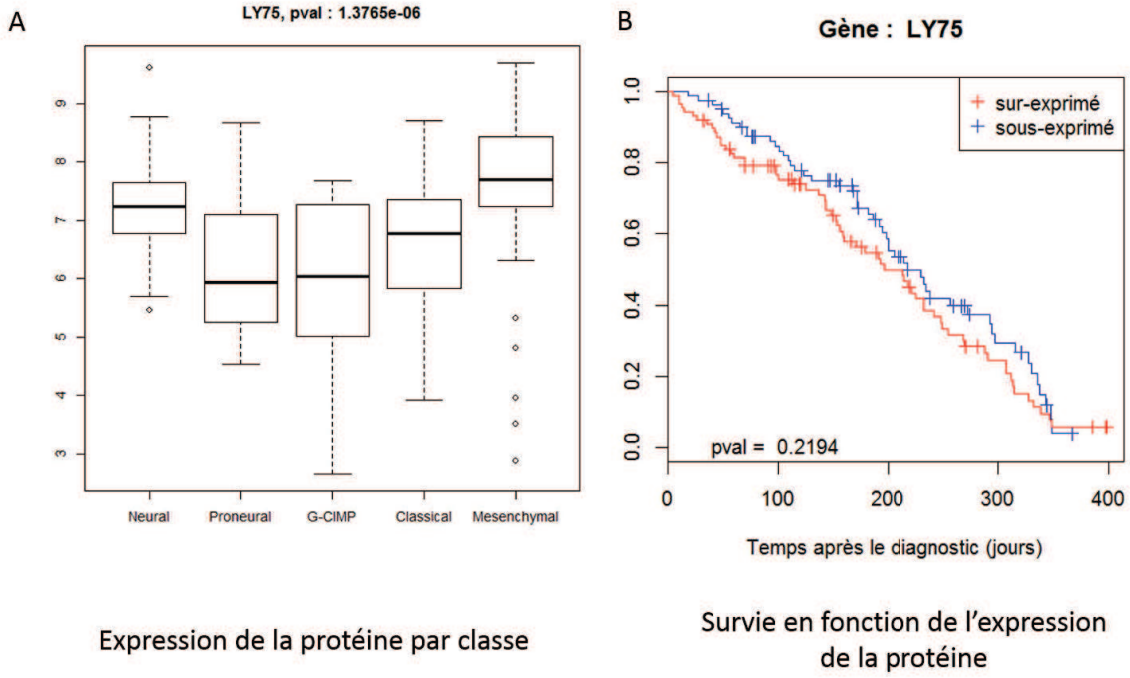

Figure 31 : Expression de LY75 dans les classes du glioblastomes (A) et effet sur la survie (B)

L'expression de CD205, CD56 et CD97 a été vérifiée par immunofluorescence, technique moins précise en terme de quantification, mais qui a l'avantage d'être facilement utilisée en clinique. Les résultats sont montrés Figure 32. On observe la présence des trois CDs au niveau des membranes cellulaires de TG1 et OB1 mais une expression faible ou peu détectable pour TG10 et TG16. Quant aux contrôles, U87-MG et HA, ils sont positifs à CD97. Ces résultats ne semblent pas corrélés avec ceux de la spectrométrie de masse qui montre une expression plus faible de CD205 (ou LY75) dans TG10/TG16 par rapport à TG1/OB1 mais pas des protéines CD56 (NCAM1) et CD97. Ces résultats ne semblent pas non plus s'expliquer par des transcrits alternatifs, et donc des protéines différentes, qui pourraient ne pas être reconnus de la même manière par l'anticorps ; en effet, pour CD56, il n'y a pas de différence d'expression des transcrits entre les échantillons, et on observe pour CD97 une différence dans HA mais pas dans les autres échantillons.

Il est difficile d'interpréter ces résultats mais on peut émettre différentes hypothèses :

- des modifications post-traductionnelles de sites reconnus par les anticorps pour TG10/TG16 ; ces modifications ne changeraient pas la reconnaissance de peptides situés en dehors de cette zone en spectrométrie de masse.
- les protéines de ces cellules sont masquées et ne sont pas reconnues par nos anticorps ou la technique utilisée dans le cas des échantillons TG10/TG16 ; ces protéines seraient néanmoins identifiées dans l'analyse classique de spectrométrie de masse.

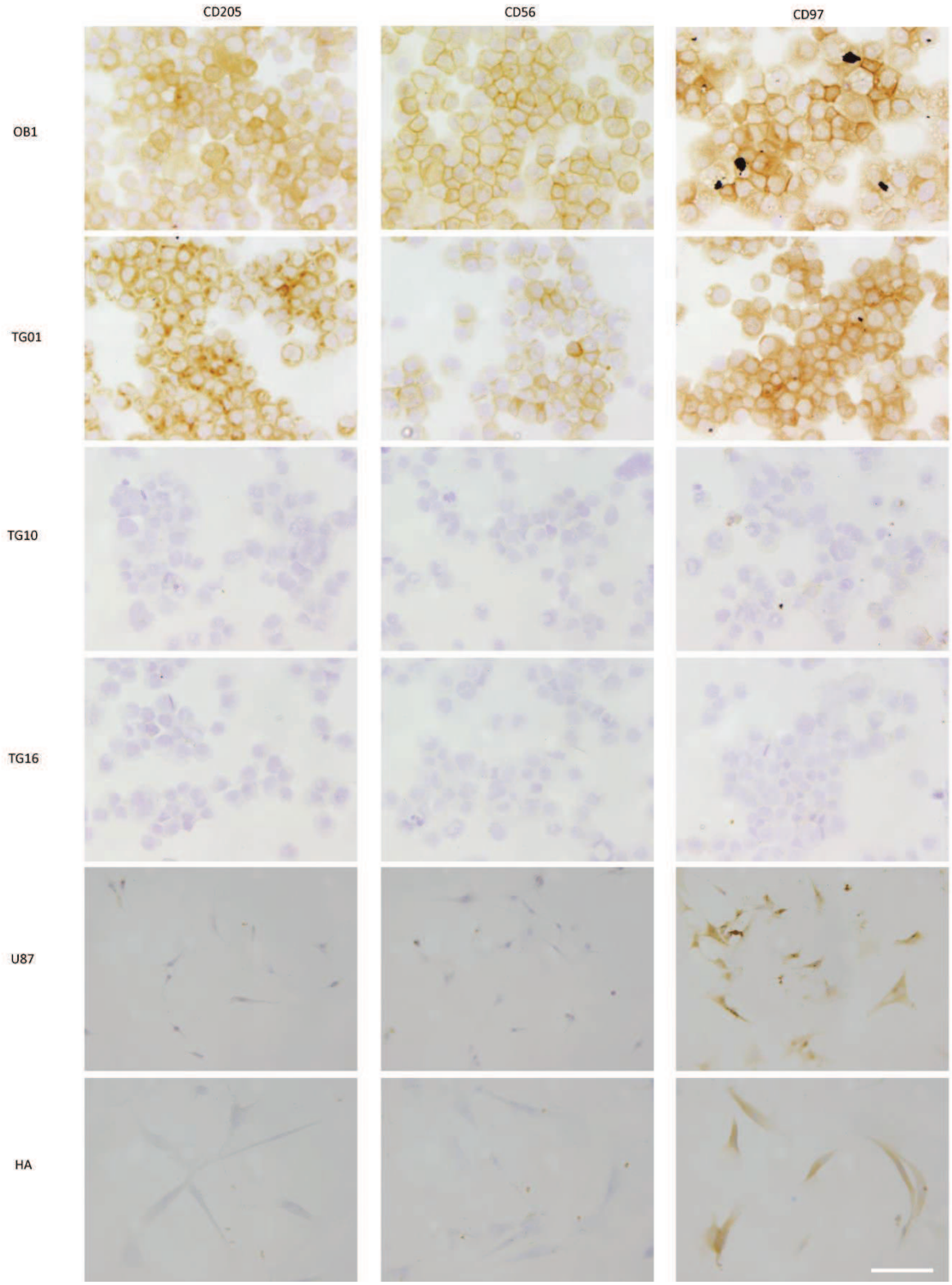

Figure 32 : Analyse immuno-histochimique des CD56, CD205 et CD97 sur nos six échantillons. On observe un marquage positif des 3 CDs pour TG1 et OB1, et de CD97 pour U87 et HA. TG10 et TG16 sont négatifs aux 3 CDs.

3.4 Conclusion

Notre étude de l'expression des gènes des cellules souches confirme la différence mentionnée dans la littérature entre glioblastomes adultes et pédiatriques. De même, nous avons pu vérifier l'hétérogénéité des cellules souches de glioblastomes, la mise en évidence de sous-groupes distincts, et l'intérêt d'identifier de nouveaux marqueurs quand ceux utilisés en routine, par exemple CD133, s'avère en fait non-spécifique.

Sur 46 CDs exprimés dans nos échantillons, 13 sont potentiellement intéressants en tant que biomarqueurs, car sur-exprimés dans les cellules souches par rapport à HA et U87. La sur-expression en transcriptomique a pu être validée pour 4 d'entre eux sur un ensemble public de plus de 70 échantillons de cellules souches. Il s'agit de LY75, plutôt caractéristique des cellules TG1 et OB1, SLC44A1, NCAM1 et CD97.

On trouve dans la littérature des études sur l'importance de l'épissage alternatif de CD97, et les différents rôles des protéines qui en résultent. Nous retrouvons l'expression de ces différents transcrits mais sans identifier de différence majeure entre les cellules souches et U87, seul HA présente une différence d'expression notable. Par contre, dans le cas de la protéine LY75, le transcrit exprimé majoritairement dans les cellules souches ne l'est pas dans U87 ni dans HA, et inversement. Il s'agit donc d'un marqueur très spécifique. Pour aller plus loin dans l'analyse, nous souhaiterions par la suite caractériser les protéines correspondant à ces différents transcrits et réaliser des études fonctionnelles.

L'analyse immunohistologique n'a pas été concluante pour TG10 et TG16 car aucun des anticorps testés (pour LY75, NCAM1 et CD97) ne s'est révélé positif. Ce problème justifie des analyses supplémentaires et nous souhaiterions vérifier dans un deuxième temps que les cellules n'ont pas internalisé les protéines, rendant ainsi nulle la possibilité de les cibler. Par contre, nous avons obtenus des résultats immunohistochimiques probants pour TG1 et OB1.

L'expression transcriptomique de ces marqueurs a été vérifiée à l'aide des données de séquençage ARN du TCGA sur le glioblastome. Aucun des marqueurs ne semble avoir un effet sur la survie globale mais ils sont tous associés à des sous-groupes particuliers : CD97 est associé aux groupes classiques et mésenchymateux, SLC44A1 aux groupes neuraux, proneuraux et G-CIMP, NCAM1 au groupe G-CIMP et LY75 au groupe mésenchymateux.

Afin de valider l'utilisation des marqueurs, des analyses de *tissue array* sont en cours sur ces quatre CDs. Bien que les cellules souches de glioblastomes ne représentent qu'environ 10% des cellules totales dans une tumeur, nous devrions pouvoir les observer.

En plus de répondre à notre question initiale en permettant la caractérisation de biomarqueurs spécifiques des cellules souches de glioblastomes, cette étude nous permet de valider l'utilisation de la méthode KANT sur des données de puces d'expression ainsi que sur des données de spectrométrie de masse. Nous avons aussi montré l'intérêt d'une méthode prenant en compte l'hétérogénéité des données sur des petits nombres d'échantillons.

Chapitre 4 : Analyse du signal calcium des GBMs et gCSCs

4.1 Qu'est-ce que le signal calcium ?

4.1.1 Définition de la signalisation calcium

Le calcium est le 4^{e} élément le plus important dans l'océan. Il est aussi présent en quantités importantes dans le plasma sanguin (environ 1 mM). Bien que nécessaire à la vie, une concentration en Ca^{2+} trop importante est toxique pour la cellule car le phosphate de calcium précipite. La concentration cytoplasmique en Ca2+ oscille entre 100 nM et 1 µM, soit 1 000 à 10 000 fois moins que dans le milieu extracellulaire. Cette compartimentation des ions calcium crée d'importants gradients de concentration calcique entre les différents organites, mais son rôle vital impose une régulation très fine des flux. Trois groupes de protéines permettent cette régulation : les transporteurs, les canaux calciques et les protéines senseurs.

Les transporteurs régulent directement les influx et efflux de calcium entre les différents compartiments cellulaires, il s'agit des pompes calciques, qui utilisent l'énergie de l'hydrolyse de l'ATP (*Adénosine triphosphate*), et des échangeurs ioniques, qui utilisent l'énergie du gradient électrochimique de sodium (Na^{+}). Ces protéines permettent d'expulser le calcium.

Les canaux calciques font entrer le calcium. Ces protéines sont présentes dans les différentes membranes de la cellule, que ce soit la membrane plasmique ou les membranes des organites.

Les protéines senseurs sont des protéines qui se lient au calcium dans une gamme de concentration spécifique. Ces protéines permettent de décoder un signal calcium et/ou de moduler la concentration calcique spatialement et temporellement.

Ces trois groupes de protéines permettent à la cellule de réguler l'homéostasie calcique et en réponse à un stimulus, d'encoder une information dans un signal calcium. Les étapes du signal calcium sont les suivantes : un stimulus externe module l'entrée et la sortie du calcium, ce qui va modifier la concentration calcique intracellulaire et donc activer les protéines senseurs ; ces dernières vont pouvoir, par le biais d'une cascade d'événements moléculaires, modifier l'activité de différentes voies de signalisation, et générer ainsi un événement cellulaire en réponse au stimulus initial. Le calcium est de ce fait un excellent messager secondaire.

On appelle toolbox calcium l'ensemble des gènes (environ 300) permettant de coder les protéines impliquées dans le signal calcium. Chaque cellule exprime une fraction de la toolbox

calcium, le signalosome qui est traduit en protéines s'assemblant en complexes macromoléculaires formant les calcisomes (Haiech et al., 2011).

Le signal calcium permet ainsi une communication entre les cellules par l'intermédiaire des canaux et récepteurs situés sur la membrane plasmique. Il existe 3 types de canaux :

- Les canaux calciques « voltage dépendants » (VOCs) qui sont activés par une dépolarisation membranaire. Ces canaux se retrouvent essentiellement dans les cellules excitables, comme les cellules musculaires et neuronales.
- Les canaux calciques dépendants de l'activation d'un récepteur (ROCs) par un ligand extracellulaire comme l'ATP, la sérotonine, le glutamate ou bien l'acétylcholine. Ils sont principalement exprimés dans les cellules sécrétrices et les terminaisons nerveuses.
- Les canaux activés mécaniquement suite à une déformation de la cellule.

Pour contrôler l'homéostasie calcique entre les différents organites et ressources de calcium (mitochondries, réticulum endoplasmique ...), la communication issue du signal calcium doit aussi se faire au niveau intracellulaire, à l'aide de différents mécanismes :

- Les canaux SOCs (*Stock operated channel*) ou canaux capacitifs qui sont activés en réponse au vidage d'un stock intracellulaire au cours du processus SOCE (*Store Operated Calcium Entry*), correspondant à l'entrée du calcium stimulée par le vidage des stocks internes. Plus précisément, la vidange en Ca^{2+} du Réticulum Endoplasmique (ER) induit la multimérisation de STIM1 qui s'accumule dans la membrane de l'ER, proche de la membrane plasmique. ORAI1 s'accumule à proximité de STIM1 et la formation de l'ensemble ORAI1/STIM1 entraine la multimérisation d'ORAI1, formant ainsi un pore du canal actif CRAC, sélectif de Ca^{2+}.
- Les canaux SMOCs, canaux sensibles à des seconds messagers intracellulaires.
- Les canaux SCaMPER, sensibles à la concentration intracellulaire en sphingolipides.
- TPs (*Two pore channel*): les canaux sensibles au métabolisme cellulaire, qui sont activés par les messagers synthétisés en fonction de la concentration d'ATP intracellulaire.
- Les TRPs (*Transient receptor potential*), qui sont activés par plusieurs types de stimulus.

4.1.2 Fonctions du signal calcium

Le signal calcium est impliqué dans la régulation de nombreux mécanismes comme la prolifération, la différenciation, l'apoptose, le métabolisme énergétique, toutes les formes de mobilités cellulaires ou intracellulaires ou les phénoménes de sécrétion pour ne citer que les plus importants.

Prenons pour exemple la régulation de la transcription. Elle se fait généralement à partir de senseurs, tels que la calmoduline, qui peut induire l'activation de kinases et ainsi la phosphorylation

du facteur de transcription CREB (*C-AMP Response Element-binding protein*) résultant en une augmentation de la transcription par la stabilisation de l'ARN polymerase II sur les promoteurs de gènes (van Haasteren et al., 1999). De plus, la calmoduline peut induire la déphosphorylation du facteur de transcription NFAT (*Nuclear factor of activated T-cells*), via l'activation de la calcineurine (phosphatase 2B), qui peut ainsi accéder au noyau et induire la transcription de gènes liés au cycle cellulaire (Crabtree, 2001). Une dérégulation du signal calcium peut avoir un effet sur la prolifération cellulaire (Prevarskaya et al., 2014).

Un autre exemple concerne l'apoptose. Le calcium peut induire l'apoptose cellulaire par une surcharge massive des ions calciques, que ce soit par une entrée importante ou bien par le relargage des stocks intracellulaires. Ainsi, un remodelage des canaux d'entrée ou des protéines impliquées dans l'efflux du calcium (pompes et échangeurs) peut jouer un rôle anti-apoptotique (Dubois et al., 2013).

Les mutations des canaux calciques sont à l'origine d'un groupe hétérogène de maladies héréditaires appelées canalopathies calciques et comprenant paralysie périodique, migraine, ataxie, rétinite pigmentaire, épilepsie et autisme (Kim, 2014), alors qu'une mutation de la calmoduline conduit à des pathologies cardiaques (Arnáiz-Cot et al., 2013). Ceci nous montre l'importance et la diversité des fonctions du signal calcium. Dans le cas du cancer, la dérégulation de nombreux canaux et pompes calciques a été observée (Monteith et al., 2012). Ils peuvent participer à la mise en place des différents *hallmarks of cancer* décrits par Hanahan et Weinberg (Prevarskaya et al., 2014).

4.2 Analyse du signal calcium

4.2.1 Objectif et étapes de l'analyse

Comme nous l'avons vu précédemment, le signal calcium est dérégulé dans différents cancers, avec des implications probables pour la mise en place des *hallmarks of cancer.* La signalisation calcique des cellules normales est suffisamment étudiée pour pouvoir interpréter nos résultats bioinformatiques.

L'objectif de cette analyse est de comparer le signalosome calcique des glioblastomes et des cellules souches de glioblastomes à celui de tissus normaux du cerveau. Pour prendre en compte la variabilité biologique (hétérogénéité des tumeurs et cellules), les erreurs d'annotations (erreur humaine) et la variabilité technique (protocoles et expériences), nous avons développé un processus d'analyse en quatre étapes successives :

- Nettoyage des données : Nous avons utilisé une analyse en composantes principales pour détecter les éventuels *outliers* de nos différents jeux de données et constituer des ensembles de données homogènes.
- Signal calcium : à partir de la toolbox calcium, défini précédemment comme étant l'ensemble de gènes codant pour des protéines impliquées dans la génération et la modulation du signal calcium (entrée et sortie du calcium, protéines se liant au calcium impliquées dans la modulation du flux de calcium). Nous avons cherché à classer les échantillons selon l'expression de l'ensemble des gènes, et également selon l'expression des gènes calcium de la toolbox, afin de savoir si la signature calcium apporte une information différente du transcriptome total. Habituellement, en utilisant une analyse transcriptomique, on associe un profil génétique (ou signature) à un groupe d'échantillons. Nous avons obtenu des couples ensembles d'échantillons/ signatures transcriptomiques différents selon que l'on utilise l'ensemble des gènes ou la toolbox calcium. Cela suggère que la signature calcium distingue des sous-ensembles d'échantillons qui apparaissent très similaires lorsque l'on regarde le profil transcriptomique global.
- Signature calcium et cerveau : à partir de données transcriptomiques publiques des différentes zones du cerveau humain, nous avons développé une méthode permettant d'extraire une signature calcium transcriptomique spécifique pour chaque zone. Les signatures obtenues sont cohérentes avec la connaissance actuelle de la biologie du cerveau, ce qui nous a conforté dans l'utilisation de notre méthode d'analyse.
- Signalosome calcium du glioblastome et des cellules souches : après avoir utilisé une méthode de classification afin de définir des groupes homogènes (glioblastomes, gliomes, lignées cellulaires de gliomes, cellules souches de glioblastomes et cellules souches embryonnaires), nous avons utilisé le processus d'analyse défini à l'étape précédente afin de trouver une signature transcriptomique élaborée à partir de la *toolbox* calcium pour les différentes classes définies. Nous avons cherché les gènes qui s'exprimaient pratiquement dans tous les tissus et ceux spécifiques d'un sous-ensemble.

Les résultats de cette méta-analyse sont un point de départ pour la compréhension des différences du signal calcium entre tissus, et notamment l'altération des cascades de signalisation calcium dans les cellules tumorales.

4.2.2 Matériels & Méthodes

4.2.2.1. Données utilisées

Nous avons utilisé les puces d'expression issues de 14 jeux de données Affymetrix indépendants, ainsi que des échantillons de notre laboratoire (voir Tableau 21), tous hybridés sur puce HG-U133 Plus 2.0.

Tableau 21 : Jeux de données utilisés pour l'analyse du signal calcium

Jeu de données	Publication	Echantillons utilisés dans l'analyse
GSE7181	(Beier et al., 2007)	6 cellules souches de glioblastomes
GSE23806	(Schulte et al., 2011)	36 lignées cellulaires de gliomes, 27 cellules souches de glioblastomes, 12 glioblastomes (GBMs)
GSE18015	(Garcia et al., 2010)	Cellules de gliomes isolées, 8 CD133+ and 8 CD133-
GSE7307		677 échantillons de tissus normaux et maladies. Seuls les tissus normaux du cerveau ont été utilisés (229 échantillons)
GSE4290	(Sun et al., 2006)	157 tumeurs, incluant 26 astrocytomes, 50 oligodendrogliomes et 81 GBMs. Les échantillons de patients épileptiques n'ont pas été utilisés.
GSE21514	(Moser and Fritzler, 2010)	2 lignées cellulaires d'astrocytes humains
GSE17312	(Bernstein et al., 2010)	4 lignées de cellules souches embryonnaires
GSE20126	(Fong et al., 2011)	4 lignées de cellules souches embryonnaires
GSE34200	(Mallon et al., 2013)	12 lignées de cellules souches embryonnaires
GSE39762	(Aldaz et al., 2013)	3 lignées de cellules souches embryonnaires
GSE44841	(Aldaz et al., 2013)	8 cellules souches de glioblastomes
GSE46016	(Rheinbay et al., 2013)	10 cellules souches de glioblastomes
GSE46531	(Ye et al., 2013)	12 cellules souches de glioblastomes
GSE51822	(Zorniak et al., 2015)	2 cellules souches de glioblastomes
		1 lignée d'astrocytes humains (HA)
Nos échantillons		1 lignée de gliomes (U87-MG)
		4 cellules souches de glioblastomes (TG1, OB1, TG10, TG16)
		8 cellules souches de glioblastomes (TG1) dans différentes conditions (quiescent, prolifératif, après traitement au temozolomide), voir Chapitre 3 : Biomarqueurs des gCSCs

L'ensemble des données a été normalisé avec la fonction justRMA du package affy (Gautier et al., 2004) sous R.

La toolbox calcium utilisée contient 260 gènes (voir Annexe 1 : Toolbox calcium), dont quelques-uns annotés comme des pseudogènes. Elle inclut 75 canaux calcium, 25 pompes ou échangeurs, 160 protéines se liant au calcium, appartenant à la famille EF-domaine (Haiech et al., 2011). Cette toolbox calcium n'est pas exhaustive mais facilite l'interprétation biologique.

4.2.2.2. Nettoyage des données

Les échantillons ont été classés en 7 groupes distincts d'après les annotations : tissu normal neural, gliomes, glioblastomes (GBMs), cellules souches de glioblastomes (gCSCs), astrocytes humains (HA), cellules souches embryonnaires (ESCs), lignées cellulaires de gliomes.

A partir de la matrice d'expression Affymetrix, le barycentre de chaque groupe d'échantillon a été calculé, ainsi que la distance des échantillons à chaque barycentre. Les échantillons plus proches d'un autre barycentre que celui de leur groupe ont été définis comme *outliers* et supprimés.

La même analyse a été faite sur les tissus normaux neuraux, classés en 38 groupes différents d'après leurs annotations. Nous avons supprimé les échantillons seuls dans leur groupe, supprimé ceux plus proches d'un autre barycentre que le leur si la distance était différente d'au moins 20% et ré-annoté hémisphère du cervelet et vermis du cervelet (seuls dans leur groupe) en cervelet (barycentre le plus proche).

4.2.2.3. Définition de sous-groupes homogènes

Pour définir les sous-groupes, nous avons appliqué une méthode de *clustering consensus* à l'aide de la librairie du logiciel R consensusClusterPlus (Wilkerson and Hayes, 2010) afin de trouver la partition la plus stable. Pour ne pas inclure de bruits dans les données, nous avons filtré les probesets ayant une expression moyenne inférieure à 6 ou supérieure à 12. Nous avons travaillé sur des groupes de taille raisonnable pour les analyses (une analyse sur trois échantillons n'est pas très utile) mais les plus homogènes possibles. Pour cela, nous avons choisi le nombre de classes en fonction des groupes qui semblaient stables, quitte à ne pas prendre en compte les singletons présents si nous choisissions un nombre de classes important. Ce choix est détaillé pour les différentes classifications effectuées. Nous montrons ici les informations de deux graphiques fournis par le programme : le graphique de la matrice consensus du nombre d'échantillons choisi, montrant la fréquence de co-classification des échantillons lors du re-calcul de la classification après les différentes étapes de ré-échantillonnages, et le *tracking plot* montrant la distribution des échantillons par classe selon le nombre choisi de classes (k), les échantillons se situent sur

l'axe des abscisses, l'axe des ordonnées représente le nombre de groupes et chaque couleur du graphique est associée à une classe. Dans tous les cas, les paramètres suivants ont été utilisés : test de 1 à 10 classes pour la classification sur l'ensemble des échantillons et 1 à 20 pour la classification des sous-groupes ; 1000 répétitions avec 80% des échantillons. La distribution des échantillons dans les classes pour chaque classification est donnée en Annexe 2 : Echantillons utilisés pour l'analyse du signal calcium et classifications.

Pour la représentation sous forme de heatmap, nous avons utilisé le package Heatplus du logiciel R, en prenant le résultat de la classification consensus comme dendrogramme.

Classification de l'ensemble des échantillons à partir de tous les gènes.

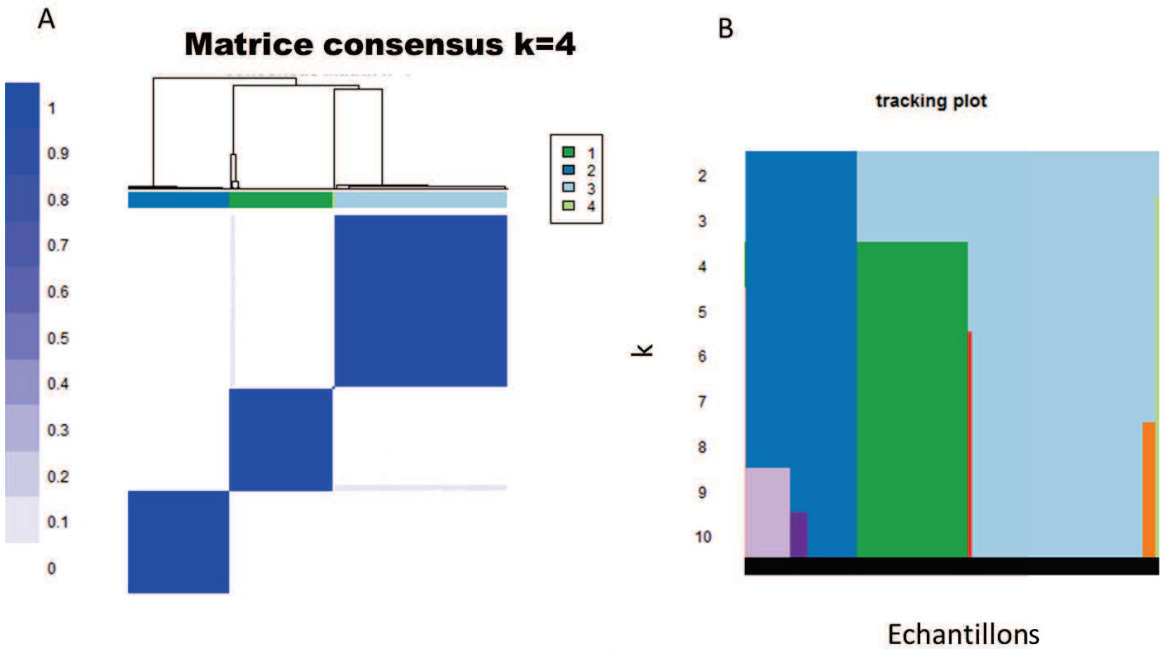

Figure 33 : Classification de l'ensemble des échantillons à partir de tous les gènes.
A) Matrice consensus. B) Distribution des échantillons dans les classes en fonction du nombre de classes. La classification est très stable en 4 groupes.

La classification est très stable en 4 groupes même si on obtient 3 groupes principaux et les 3 échantillons « glande pituitaire » à part (Figure 33).

Classification de l'ensemble des échantillons à partir des gènes calcium

La classification obtenue (Figure 34) est moins stable que dans le cas précédent, d'après la matrice consensus. Néanmoins l'objectif étant de comparer avec la classification basée sur l'ensemble des gènes, nous n'avons pas voulu partir sur un nombre de groupes trop importants,

qui pourront être étudiés par la suite. On obtient ainsi 4 groupes principaux (1 groupes gliomes avec les gliomes et glioblastomes, 1 groupe lignées cellulaires contenant les lignées cellulaires de gliomes, les cellules souches de glioblastomes, les lignées d'astrocytes et les cellules souches embryonnaires, 2 groupes de tissus neuraux normaux), les 3 échantillons « glande pituitaire » ensemble et l'échantillon « cerveau fœtal » seul.

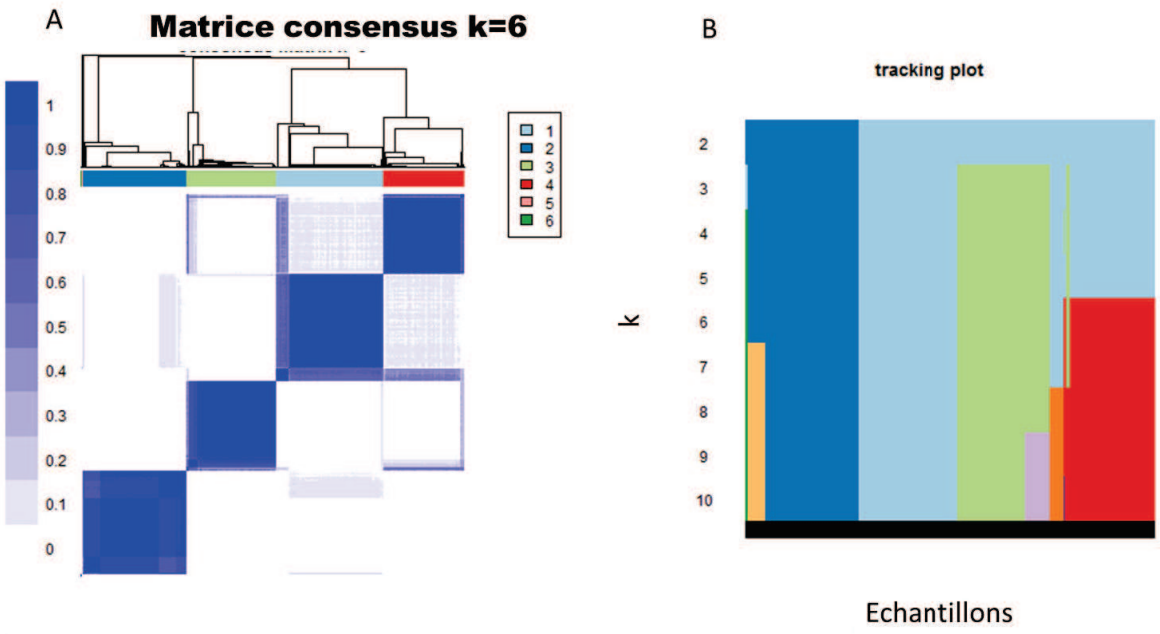

Figure 34 : Classification de l'ensemble des échantillons à partir des gènes calcium.
A) Matrice consensus. B) Distribution des échantillons dans les classes en fonction du nombre de classes. La classification est stable en 6 groupes.

Classification du sous-groupe « gliomes » à partir des gènes calcium

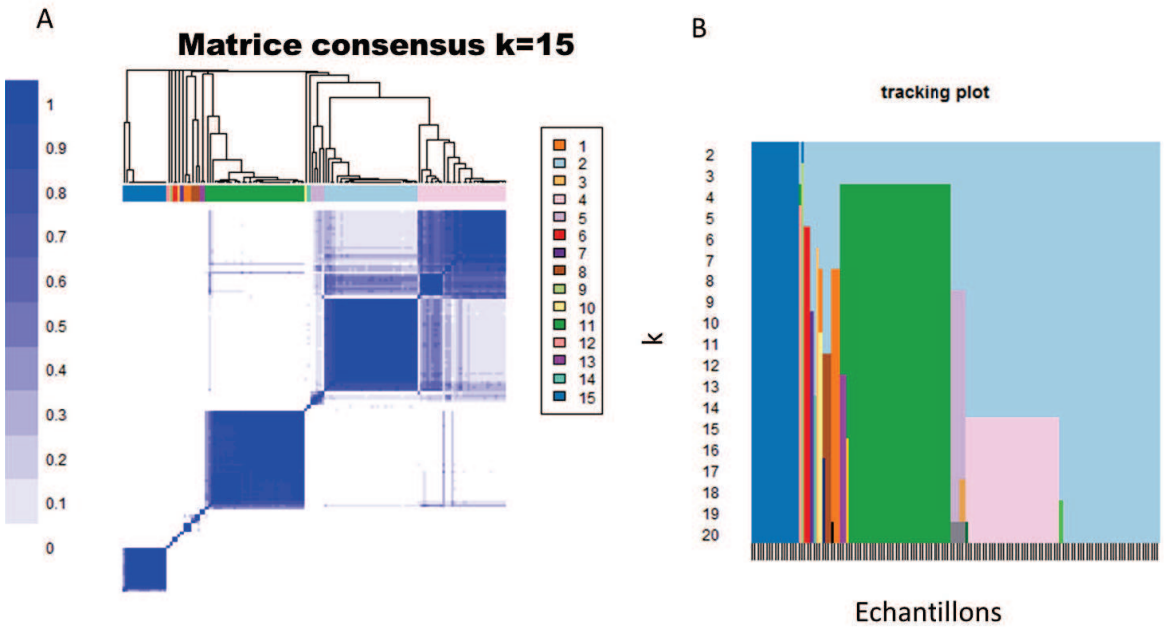

Figure 35 : Classification des échantillons du sous-groupe gliomes (défini précédemment) à partir des gènes calcium. A) Matrice consensus. B) Distribution des échantillons dans les classes en fonction du nombre de classes

On obtient 4 classes principales, une de tissus normaux, une de gliomes, deux de glioblastomes, avec quelques gliomes et divers échantillons dans des petites classes, qui ne seront pas étudiées. Nous avons vérifié quels étaient ces échantillons « à part ». Ils viennent des 2 ensembles de données ayant des gliomes, ce n'est donc pas un problème de variabilité technique. En regardant la distance au barycentre calculée précédemment, on remarque qu'elle est plus importante pour ces échantillons (moyenne 122 et écart-type 19 au lieu de moyenne 102 et écart-type 23 pour les échantillons dans les classes principales); ce qui va dans le sens d'échantillons non homogènes, différents des autres.

Classification du sous-groupe « lignées cellulaires »

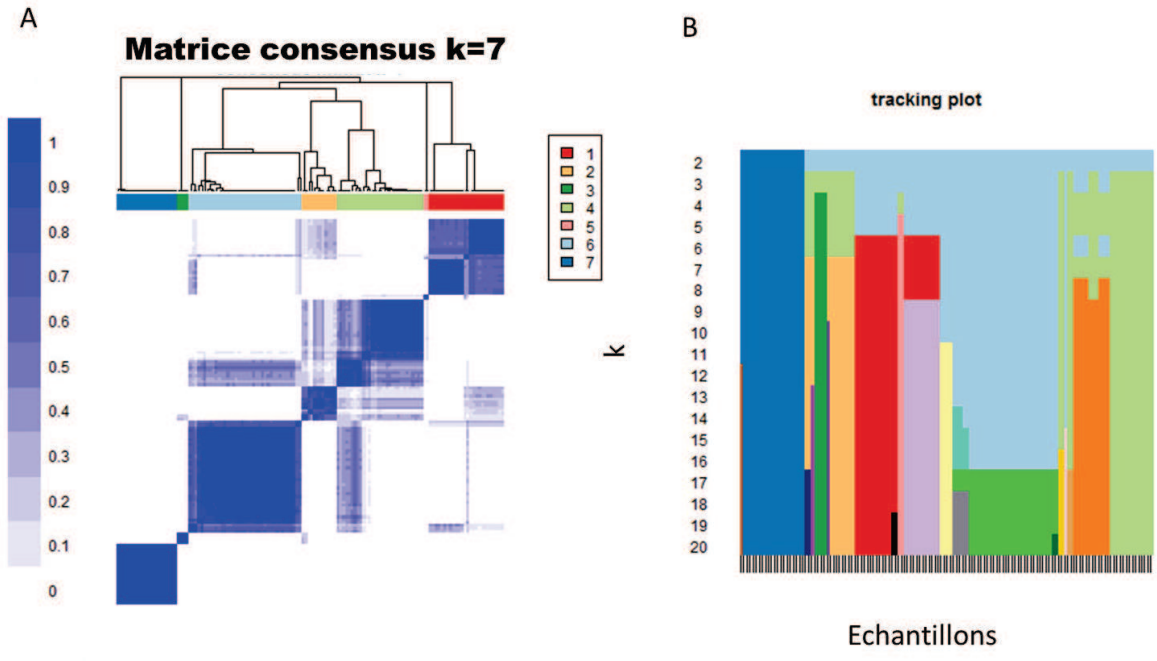

Figure 36 : Classification des échantillons du sous-groupe lignées cellulaires (défini précédemment) à partir des gènes calcium.

A) Matrice consensus. B) Distribution des échantillons dans les classes en fonction du nombre de classes

On obtient 5 classes principales (Figure 36) : une contenant les cellules souches embryonnaires, une contenant la majorité des lignées cellulaires de gliomes et HA et 3 classes de gCSCs, ainsi qu'un ensemble de 4 gCSCs et un autre de 2 gCSCs. Ces échantillons ont une distance au barycentre plus élevée que les autres lignées cellulaires (moyenne 180 et écart-type 33 contre une moyenne de 150 et un écart-type de 32 pour les autres échantillons). On observe aussi une lignée de glioblastomes dans un groupe de gCSCs (échantillon GSM587157).

4.2.2.4. Méthode de détection des gènes sur-exprimés

L'algorithme utilisé (ALGO) est dérivé de celui de Kant (chapitre 2). Un gène est considéré sur-exprimé dans un tissu si son expression est supérieure à celle du maximum des échantillons contrôles additionné de 0,5 ; ce qui correspond au seuil défini pour des données de puces d'expression du même type de tissu.

ALGO :

Soit l'expression des n_1CTRL échantillons normaux pour un probeset donné par

$$\{x_{1i}\}_{i\in\{1,2,..,n_1\}}$$

Et l'expression des n_2 échantillons tumoraux pour un probeset donné par

$$\{x_{2i}\}_{i\in\{1,2,..,n_3\}}$$

Nous commençons par calculer le maximum d'expression pour un probeset pour les échantillons CTRL1 en prenant en compte les outliers potentiels.

Soit Q3 le quartile supérieure et Q1 le quartile inférieur, nous avons défini comme outliers les 2.5% des valeurs supérieures à

$$Q3\{x_{1i}\}_{i\in\{1,2,..,n_1\}} + 3(Q3\{x_{1i}\}_{i\in\{1,2,..,n_1\}} - Q1\{x_{1i}\}_{i\in\{1,2,..,n_1\}})$$

Ces valeurs ne sont pas prises en compte pour calculer le maximum d'expression des CTRL1.

$$\hat{x}_1 = \max\{X_{1i} - \text{outliers}\}$$

$$t = 0.5$$

$$pop = \{x_{2i} > \hat{x}_1 + t\}$$

$$n_p = \text{size}\{pop\}$$

$$\Delta = \text{median}\{pop\} - \hat{x}_1$$

$$\text{Score} = 2^{\Delta}\ \frac{n_p}{n_2}$$

4.2.2.5. Détection des gènes spécifiques d'une population

L'objectif de l'analyse est de vérifier si un calcium signalosome est spécifique d'un tissu. Pour cela, nous avons utilisé l'algorithme de sur-expression (ALGO) de façon récursive sur le jeu de données des tissus neuraux normaux (dans un premier temps) en utilisant une fois chaque annotation comme contrôle et l'ensemble des autres annotations comme test. Cet algorithme trouve les gènes sur-exprimés dans des sous-ensembles. Lorsque tous les échantillons ayant la

même annotation sur-expriment un *probeset*, ce probeset est considéré comme sur-exprimé pour le groupe d'échantillons considéré. Ainsi, nous avons obtenu l'ensemble des probesets sur-exprimés pour un groupe par rapport à un autre. L'objectif est de détecter les gènes spécifiques pour chaque groupe. Nous avons d'abord agrégé les probesets par gène, puis sélectionné, pour chaque annotation, les gènes sur-exprimés dans la moitié des cas (par rapport à la moitié des contrôles).

Le résultat final est une liste de gènes spécifiques d'un groupe.

4.2.3 Résultats

4.2.3.1. *Outliers* et description des différents groupes

A partir des ensembles de données sélectionnés dans la base de données GEO (Voir Table 1 Matériel & Méthodes), nous avons défini 7 groupes différents :

1) Tissus normaux du cerveau (n=229)
2) Gliomes : tous les cancers du cerveau sauf les glioblastomes (n=76)
3) Glioblastomes (GBMs) (n=93)
4) Lignées cellulaires de gliomes, c'est-à-dire lignées cellulaires dérivées de tumeurs (n=37)
5) Cellules souches de glioblastomes, c'est-à-dire cellules souches isolées de biopsies de GBMs (gCSCs) (n=93)
6) Lignées primaires d'astrocytes humains obtenues d'ATTC (HA) (n=3)
7) Cellules souches embryonnaires (ESCs) (n=21)

Afin de vérifier la cohérence entre les annotations des échantillons venant de différents ensembles de données, nous avons utilisé une ACP sur les données transcriptomiques nous permettant de mieux visualiser les groupes et de détecter les *outliers* putatifs.

Les *outliers* sont définis comme des échantillons plus proches du barycentre d'un autre groupe que du leur.

En utilisant cette définition, 15 gliomes ont été trouvés plus proches des GBMs et 3 des échantillons normaux du cerveau. Sur les 93 GBMs, 13 étaient plus proches des gliomes et 3 des tissus normaux. Concernant les GBMs et les gliomes proches des tissus normaux, on peut avancer l'hypothèse d'une contamination importante des biopsies tumorales par le tissu normal. Il n'est pas possible d'obtenir des informations pouvant confirmer ou infirmer les annotations des données publiques. Si ces annotations sont erronées, cela impliquerait que l'erreur d'annotations dans les gliomes serait de 15 à 20%. Nous avons supprimé les *outliers* de l'analyse sans essayer de les ré-annoter.

Aucun *outlier* n'a été trouvé parmi les lignées cellulaires de gliomes et les ESCs. Concernant les gCSCs, 16 ont été défini comme *outliers*. Parmi ces-dernières, 11 appartiennent à l'ensemble de données GSE18015, qui contient 16 échantillons. Nous avons décidé de supprimer l'ensemble du jeu de données. Sur les 5 derniers *outliers*, 4 étaient proches des lignées cellulaires de gliomes et un des GBM. Si on ne prend pas en compte les données de l'ensemble GSE18015, le taux d'erreur est d'environ 6%, montrant une forte homogénéité dans le groupe des gCSCs malgré les différents protocoles utilisés dans les différents laboratoires pour l'isolation des cellules souches. Celles isolées par le département de médecine de l'Université de Salamanca (GSE18015) semblent au niveau transcriptomique différentes de toutes les autres mais il est difficile de trouver une explication technique. Pour mieux visualiser l'homogénéité du groupe, nous avons effectué une ACP des gCSCs seuls (voir Figure 37) après avoir supprimé les *outliers*. Les échantillons des différents jeux de données ne sont pas totalement regroupés ensemble dans l'ACP, ce qui pourrait s'expliquer par les différences dans les milieux de culture utilisés. Nous avons choisi de considérer les différents clones de gCSCs comme des échantillons différents, en partant de l'hypothèse qu'ils sont représentatifs de la diversité des gCSCs.

Nous avons vérifié par une ACP sur l'ensemble des échantillons l'homogénéité des annotations les unes par rapport aux autres (Figure 38). Quelques échantillons semblent un peu éloignés de leur groupe mais nous ne voulons pas supprimer l'hétérogénéité biologique des échantillons. Les tissus normaux sont particulièrement « étalés », ceci s'explique par la diversité des tissus composant cet ensemble. On peut aussi remarquer que les lignées cellulaires se retrouvent ensemble par opposition aux tissus, que ce soit les cellules souches de glioblastomes ou embryonnaires, les lignées de gliomes ou bien HA. Ceci montre, notamment que les lignées de gliomes ou glioblastomes ont un transcriptome éloigné des gliomes et glioblastomes. Or ces lignées sont souvent utilisées pour des tests biologiques (traitements, marqueurs) visant les tumeurs mais il semble que leur mise en culture leur confère des propriétés proches des cellules souches sur le plan transcriptomique.

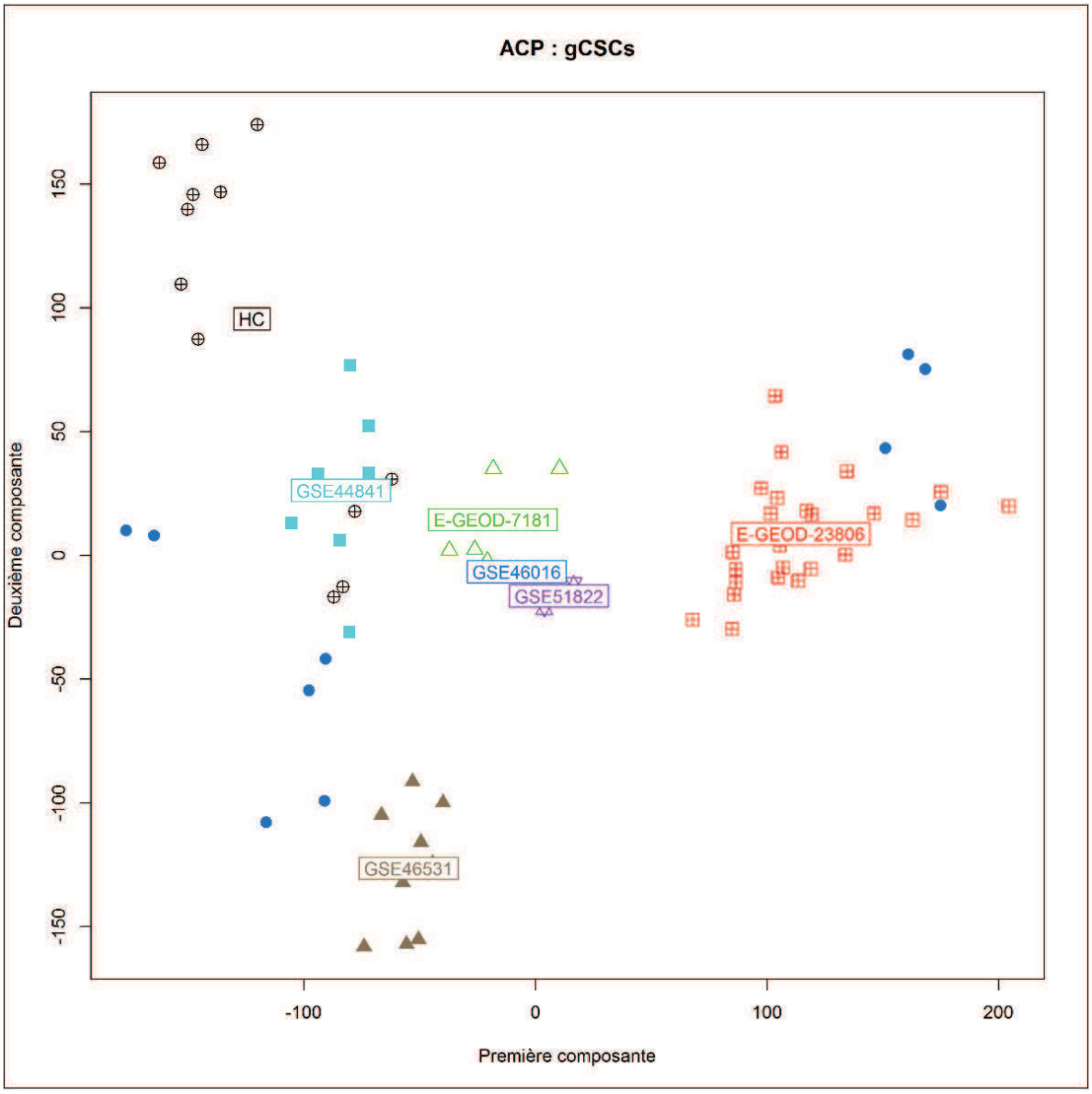

Figure 37 : Analyse en Composantes Principales des cellules souches de glioblastomes, après suppression des outliers.

Les différents ensembles de données ne sont pas totalement mélangés. Chaque jeu de données est représenté par une couleur, l'ensemble HC désigne nos échantillons, cultivés par Hervé Chneiweiss.

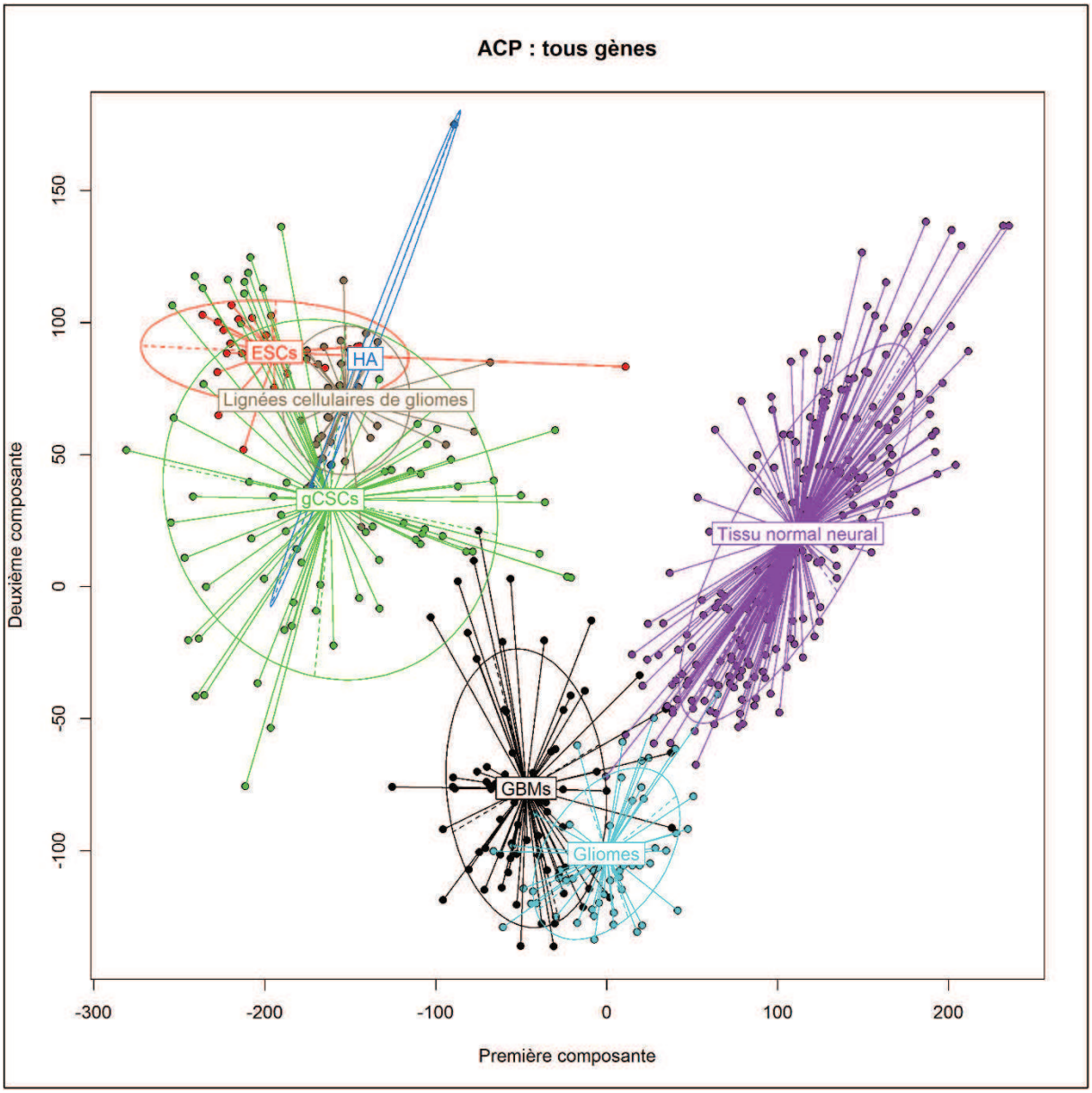

Figure 38 : ACP de l'ensemble des échantillons après suppression des outliers.
On remarque que les lignées cellulaires, que ce soit HA, les lignées de gliomes, les lignées de cellules souches embryonnaires et les cellules souches de glioblastomes sont proches.

Pour la suite des analyses, les différents échantillons sont considérés homogènes, nous n'allons donc pas étudier de manière plus approfondie les problèmes de bruit pouvant provenir d'erreurs d'annotations ou d'effet batch.

4.2.3.2. Comparaison des sous-groupes d'échantillons définis à partir de l'ensemble des gènes et de ceux définis à partir des gènes calcium

Notre hypothèse de travail est que la signature calcium est caractéristique d'une cellule ou bien d'un tissu donné. Cette signature dépend du signalosome calcium, c'est-à-dire de l'ensemble des gènes de la toolbox calcium exprimés.

Nous avons donc restreint notre ACP aux gènes de la toolbox calcium et comparé le graphique obtenu par rapport à celui obtenu avec l'ensemble des gènes. Comme on peut le remarquer dans la Figure 39, les groupes d'échantillons par annotation sont similaires En se restreignant aux gènes de la toolbox calcium, nous perdons probablement une partie des informations. Cependant, en étudiant des voies de signalisation spécifiques, l'interprétation des résultats sera facilitée au niveau biologique.

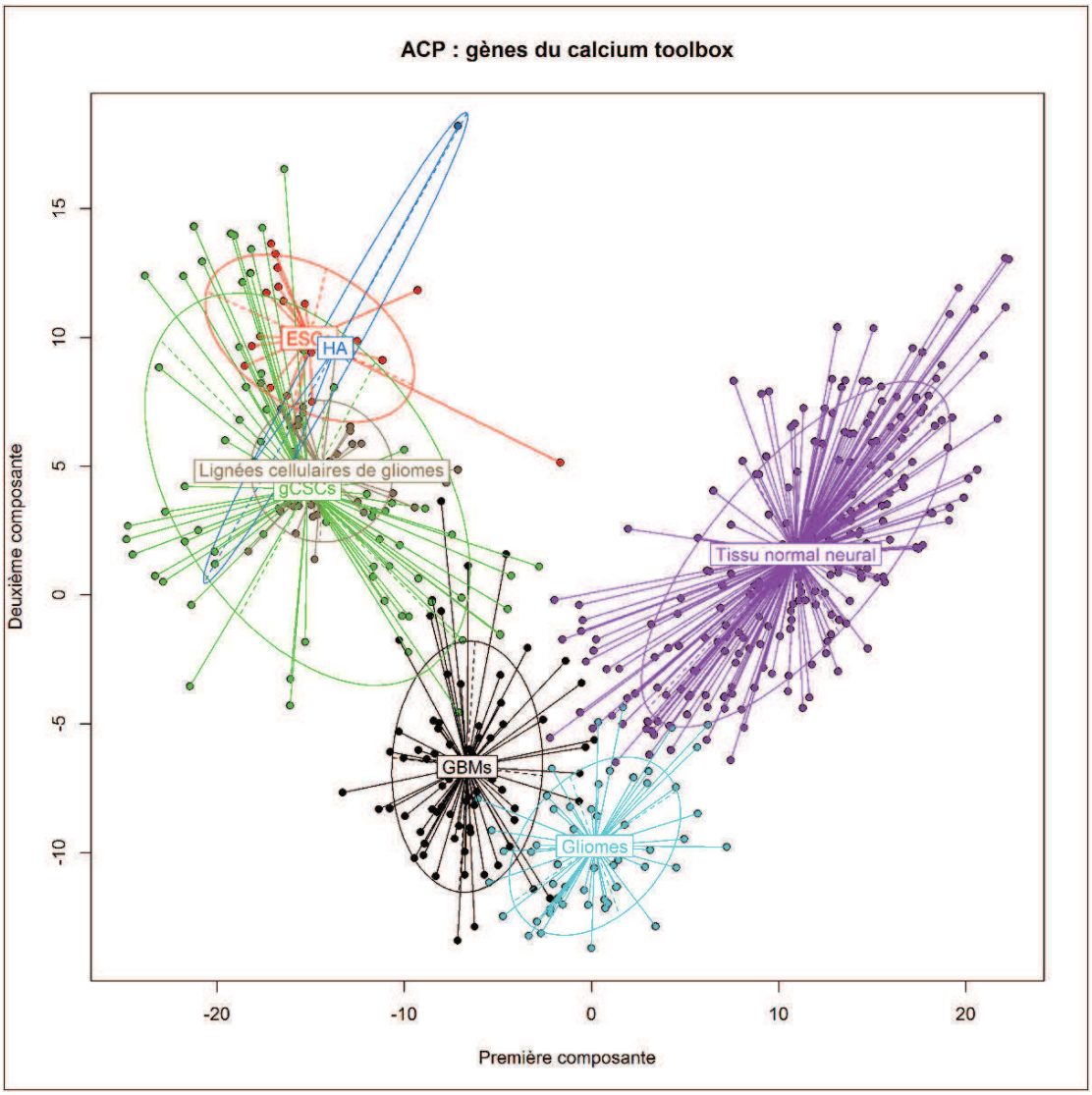

Figure 39 : ACP à partir des gènes de la toolbox calcium.
On retrouve le regroupement par annotation, même si les groupes sont légèrement plus proches qu'en prenant l'ensemble des genes.

Pour aller plus loin, nous avons fait une analyse de classification sur l'ensemble des échantillons, à partir de l'ensemble des gènes ou bien uniquement des gènes de la toolbox calcium.

Nous avons utilisé une classification consensus sur les probesets filtrés (voir la partie Définition de sous-groupes homogènes). Après filtrage, 26429 probesets sont utilisés pour l'ensemble « tous gènes ». La classification consensus a été effectuée en considérant 2 à 10 clusters afin de trouver la partition la plus stable, ce qui a été obtenu en quatre clusters. Dans ce cas, les échantillons sont divisés en trois classes principales, corrélées avec les annotations, l'une comprenant les tissus neuraux normaux, la deuxième les gliomes (GBMs inclus), la troisième les lignées cellulaires (HA, gliomes, gCSCs et ESCs) ; et une petite classe comprenant les trois tissus normaux "glande pituitaire" (Figure 40).

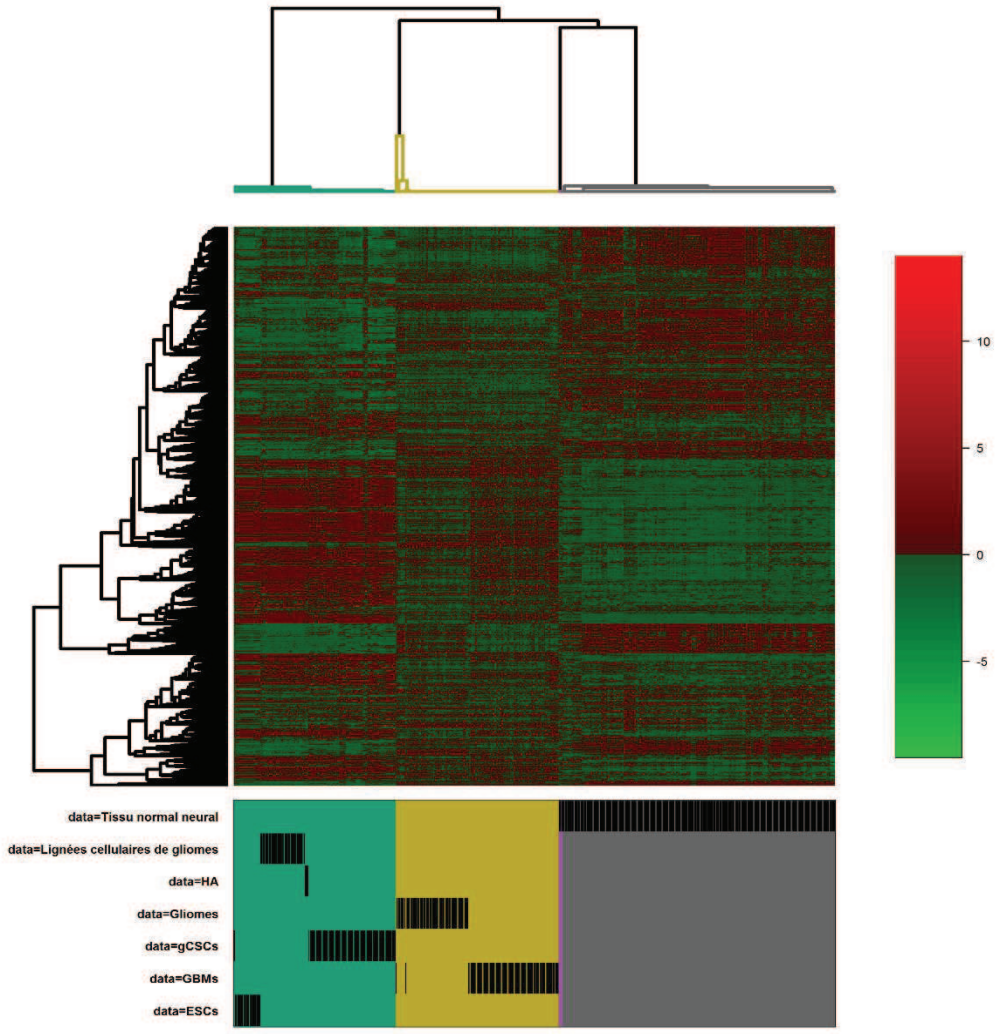

Figure 40 : Classification de l'ensemble des éhcantillons à partir des gènes filtrés et visualisation sous forme de heatmap avec les annotations.

On observe 3 groupes principaux : les tissus normaux, les gliomes et les lignées cellulaires.

En appliquant la même analyse à partir des gènes de la toolbox calcium, nous travaillons à partir de 254 probesets. La meilleure classification semble être composée de six classes, dont quatre principales et deux petites. La division est différente de la précédente (voir Figure 41). Les classes « lignées cellulaires » et « gliomes » sont conservées. La différence principale concerne les tissus neuraux normaux, en utilisant l'ensemble des gènes ils restent ensemble alors qu'en travaillant sur la toolbox calcium ils sont sous-divisés en quatre parties : les échantillons annotés glande pituitaire (hypophyse) seuls, quelques échantillons avec les gliomes (ganglions trigéminés et rachidiens) et deux autres classes indépendantes. L'échantillon normal annoté cerveau foetal se retrouve aussi seul. De plus quelques gliomes se retrouvent dans les classes "tissue normal neural", 10 gliomes dans une classe et un dans la deuxième classe. La repartition de l'ensemble des échantillons dans les classes est disponible en Annexe 2 : Echantillons utilisés pour l'analyse du signal calcium et classifications. Les annotations cliniques ne nous permettent pas d'interpréter ces résultats mais ils sont en accord avec l'ACP sur les gènes de la toolbox calcium qui semble rapprocher les gliomes des tissus neuraux normaux.

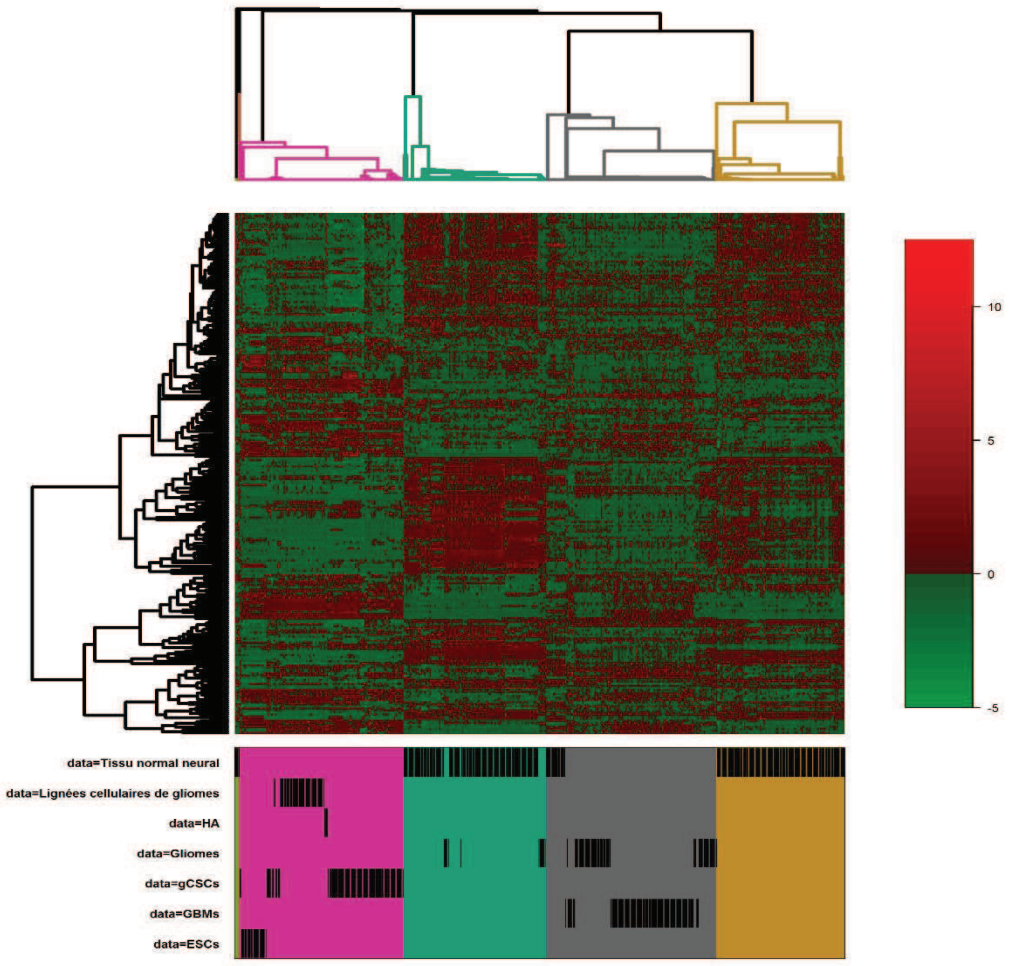

Figure 41 : Classification de l'ensemble des échantillons à partir des genes de la toolbox calcium filtrés et visualisation sous forme de heatmap avec les annotations.

On observe 4 groupes principaux : 2 groupes de tissus normaux, les gliomes et les lignées cellulaires.

Les gènes de la toolbox calcium ne contiennent pas exactement les mêmes informations que l'ensemble des gènes puisqu'ils modifient la classification des échantillons. L'intérêt est de pouvoir obtenir un point de vue différent, plus facile à interpréter car nous pouvons avoir l'information biologique faisant le lien entre la classification et l'expression spécifique de gènes impliqués dans des fonctions connues des tissus neuraux.

4.2.3.3. Signature calcium des tissus du cerveau

L'objectif de cette partie est de savoir s'il est possible de trouver une signature spécifique du signalosome calcium pour les différentes annotations des tissus neuraux.

L'ensemble de données public contient 229 échantillons ayant 38 annotations différentes, qui correspondent aux différentes zones du cerveau et de la moelle épinière.

Nous avons défini les *outliers* de la même manière que précédemment à partir des barycentres.

Sur les 38 annotations, sept ne concernent qu'un seul échantillon, il s'agit de l'hémisphère du cervelet, du vermis du cervelet, cerveau fœtal, globus pallidus, pont, cortex préfrontal et noyau thalamique latéral et une annotation concerne deux échantillons, le cortex frontal. Ces échantillons ont été supprimés sauf les premier et deuxième qui ont été ré-annotés en cervelet.

Sur les 222 échantillons restants, 55 ont été trouvés plus proches d'un autre barycentre que le leur. Nous avons émis l'hypothèse que les tissus du cerveau, bien que différents n'étaient pas si éloignés que cela les uns des autres et avons décidé de garder l'annotation des échantillons quand la différence entre la distance du plus proche barycentre et la distance à leur barycentre était inférieure à 12% et que la distance à leur barycentre était inférieur à 100.

Concernant l'annotation cortex cérébral, six échantillons sur neuf ont été définis comme *outliers*, nous avons donc décidé de supprimer la totalité des échantillons avec l'annotation cortex cérébral.

Au final, 35 échantillons ont été supprimés, dont 7 correspondant à des annotations uniques et les 9 échantillons du cortex cérébral. Notre jeu de données final est composé de 194 échantillons, annotés dans 29 différents groupes de tissus (voir Annexe 2 : Echantillons utilisés pour l'analyse du signal calcium et classifications). Cela suggère qu'il existe 16% de bruit dans l'annotation de ces échantillons. Nous avons considéré ce pourcentage comme acceptable.

Pour analyser l'expression différentielle des gènes de la toolbox calcium dans les différents tissus neuraux normaux, nous avons utilisé un algorithme nous permettant de trouver un gène sur-exprimé dans un tissu par rapport aux autres (voir Matériels & Méthodes). Nous avons considéré qu'un gène est exprimé dans un tissu lorsque la moyenne d'expression de ce gène sur les échantillons du tissu est supérieure à 8.

Quatre gènes impliqués dans l'entrée du calcium, CACFD1, CACNG4, ORAI2 et TPCN1 sont exprimés dans tous les tissus neuraux et deux autres (CACNA1A ou Cav2.1 et CACNG6) dans

plus de 80 % des tissus. Ceci indique que peu de gènes de la toolbox calcium impliqués dans l'entrée du calcium sont exprimés en commun dans l'ensemble du cerveau (moins de 10%).

Dans les gènes calcium exprimés dans les tissus, CACFD1 encode une protéine membranaire, de type canal calcique, et il a été rapporté que ce canal était lié à l'exocytose/l'endocytose neuronal (Yao et al., 2009). TPCN1 semble être l'un des principaux canaux calciques au niveau des membranes du lysosome et de l'endosome (Neely Kayala et al., 2012).

CACNA2D1 est présent dans tous les tissus sauf un, il s'agit d'une sous-unité du canal calcium dépendant du voltage Cav 2.1. Ce-dernier est aussi impliqué dans l'exocytose (Weiss, 2010). CACNG4 et CACNG6 sont connus en tant que modulateurs des récepteurs AMPA (récepteurs membranaires ionotropes de glutamate qui permettent une transmission synaptique rapide dans le système nerveux central) et sont exprimés dans tous les tissus sauf deux pour CACNG6. Finalement ORAI2 mais ni ORAI1 ou ORAI3 sont exprimés dans tous les tissus, montrant l'importance d'un mécanisme SOCE spécifique (Heo et al., 2015; Kito et al., 2015) soit dans la microglie ou dans les cellules endothéliales capillaires du cerveau.

Un autre ensemble de gènes exprimés dans tous les tissus du cerveau code pour des pompes et échangeurs. Parmi eux, Serca2 est impliqué dans la régulation des vésicules endoplasmiques (Baba-Aissa et al., 1998), deux gènes codent pour des pompes localisées sur la membrane plasmique (PMCA2 et PMCA4 (Baba-Aissa et al., 1998)), un ensemble de gènes est impliqué dans l'homéostasie calcique des mitochondries (MICU1, MICU2, SLC25A23 (Hoffman et al., 2014)) et un ensemble d'échangeurs calciques (SLC24A2 (Li et al., 2002) et SLC25A12 (Rueda et al., 2014)). Le nombre de gènes exprimés dans 25 tissus représentent 32% du nombre total de gènes de la toolbox calcium impliqués dans le repompage du calcium.

Finalement, 17 gènes codant pour des protéines liant le calcium (CAB39, CABP4, CALCOCO1, CALCOCO2, CALR, CHP1, EFCAB14, HPCAL1, MYL12B, MYL6, MYL6B, RASEF, RCN2, S100A13, S100A16, S100A6, SRI/sorcin) sont exprimées dans tous les tissus ce qui représente 15% des gènes codant pour des protéines liant le calcium. Six autres gènes sont exprimés dans 25 tissus (CETN2, EFHD1, S100B, S100A1, VSNL1, NCS1).

Maintenant que nous avons caractérisé les gènes exprimés dans l'ensemble des tissus et appartenant à la toolbox calcium, nous voulant connaitre ceux qui sont spécifiquement exprimés dans un tissu ou un ensemble de tissus neuraux.

Concernant l'entrée du calcium, la spécificité réside dans l'expression des sous-unités alpha et beta des canaux calciques dépendant du voltage, des récepteurs dépendant du phosphate d'inositol, des récepteurs de ryanodine et dans un ensemble de canaux TRP.

Concernant les pompes et échangeurs, la spécificité des tissus est surtout donnée par la différence d'expression des échangeurs sodium/calcium.

Concernant les protéines se liant au calcium, un ensemble de 36 gènes différents permet de définir la signature de chacun des tissus.

4.2.3.4. Signature calcium des tumeurs et cellules souches de glioblastomes

Pour pouvoir appliquer le même type d'analyse et établir une signature transcriptomique de la toolbox calcium pour les gliomes, notre ensemble de tumeurs doit d'abord être classifié en classes homogènes.

Nous avons repris la classification précédente, définie à partir de la toolbox calcium afin de le raffiner. Nous avions six classes : quatre de tissus neuraux normaux (deux principales et deux petites), une de GBMs et gliomes (appelé classe « gliomes ») et la dernière de lignées cellulaires. Notre objectif est de reprendre et raffiner les classes « gliomes » et « lignées cellulaires » en utilisant la même méthode que précédemment.

Concernant la classe « gliomes », elle se divise en 15 sous-classes (Figure 42) : quatre principales et onze petites classes composées de un à quatre échantillons. Ces derniers ne sont pas pris en compte dans notre analyse car trop petits (plus d'explications dans la partie Matériels & Methodes). Concernant les classes principales, une contient les échantillons normaux (ganglions trigéminés et rachidiens), le deuxième la majorité des gliomes et quelques GBMs, les troisième et quatrième des GBMs et quelques gliomes. Puisque nous voulons des groupes les plus homogènes possibles et qu'il est impossible de vérifier les annotations, nous utiliserons seulement les échantillons de l'annotation majoritaire des classes (donc suppression des quelques GBMs de la deuxième sous-classe et des quelques gliomes des troisièmes et quatrièmes sous-classes), ainsi que les échantillons neuraux normaux. Au final, trois groupes sont définis, nommés gliomes, GBM1, GBM2 et contenant respectivement 36, 27 et 34 échantillons.

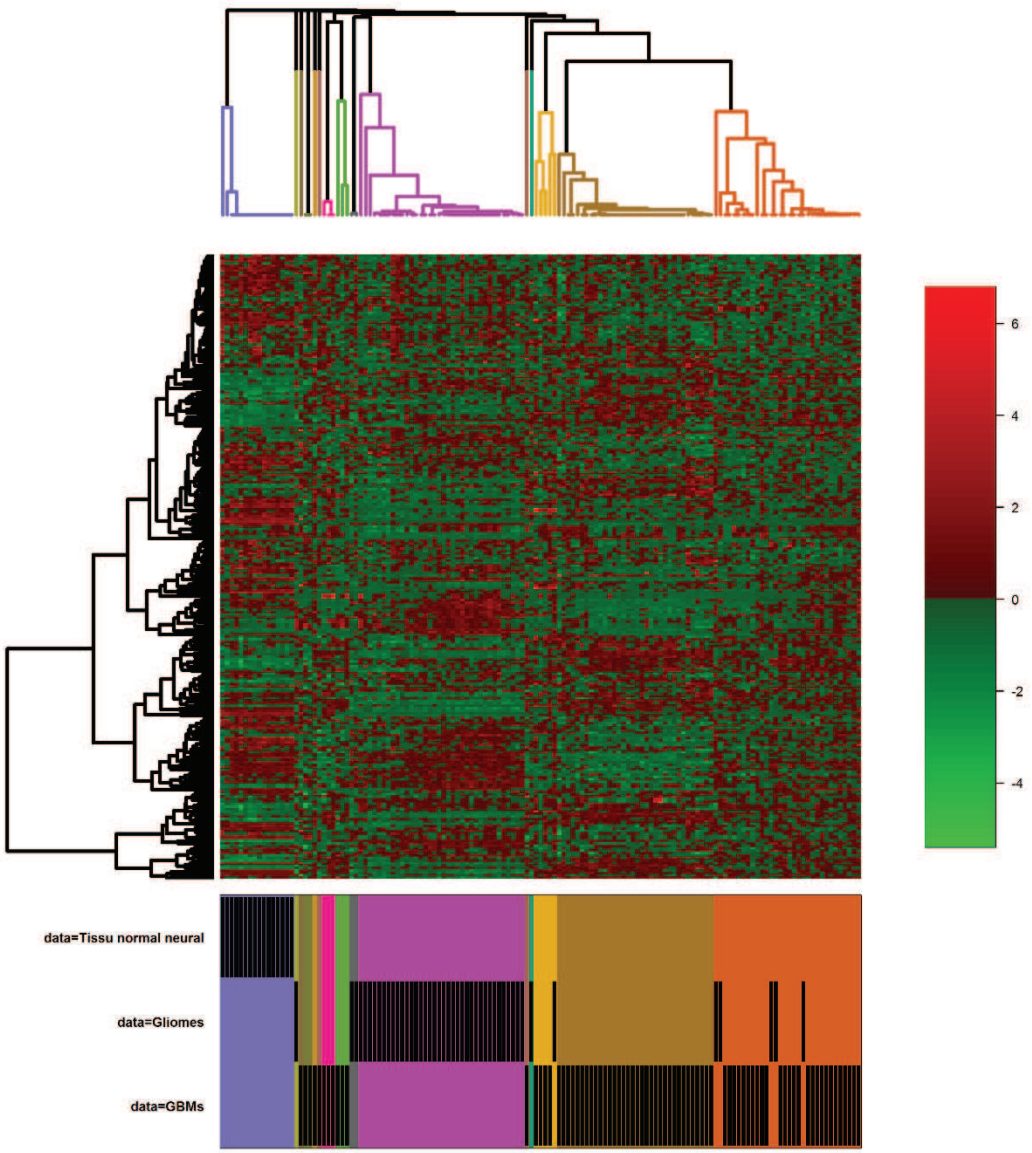

Figure 42 : Classification des échantillons du groupe des gliomes.
On obtient quatre classes principales et de nombreux échantillons à part.

La classe « lignées cellulaires » se divise en sept sous-classes (Figure 43), dont cinq principales composées pour la première de lignées cellulaires de gliomes et HA, de gCSCs pour les deuxièmes, troisièmes et quatrièmes, de cellules souches embryonnaires (ESCs) pour le dernier. Nous considérerons ces cinq sous-classes pour la suite, appelées lignées cellulaires, gCSC1, gCSC2, gCSC3 et ESCs et contenant respectivement 39, 12, 29, 26 et 21 échantillons. Les deux autres classes contiennent respectivement 4 et 2 gCSCs. Si on reprend l'analyse de distance au barycentre, ces 6 échantillons sont plus éloignés que ceux classés (voir partie Materiels & Methodes, Définition de sous-groupes homogènes).

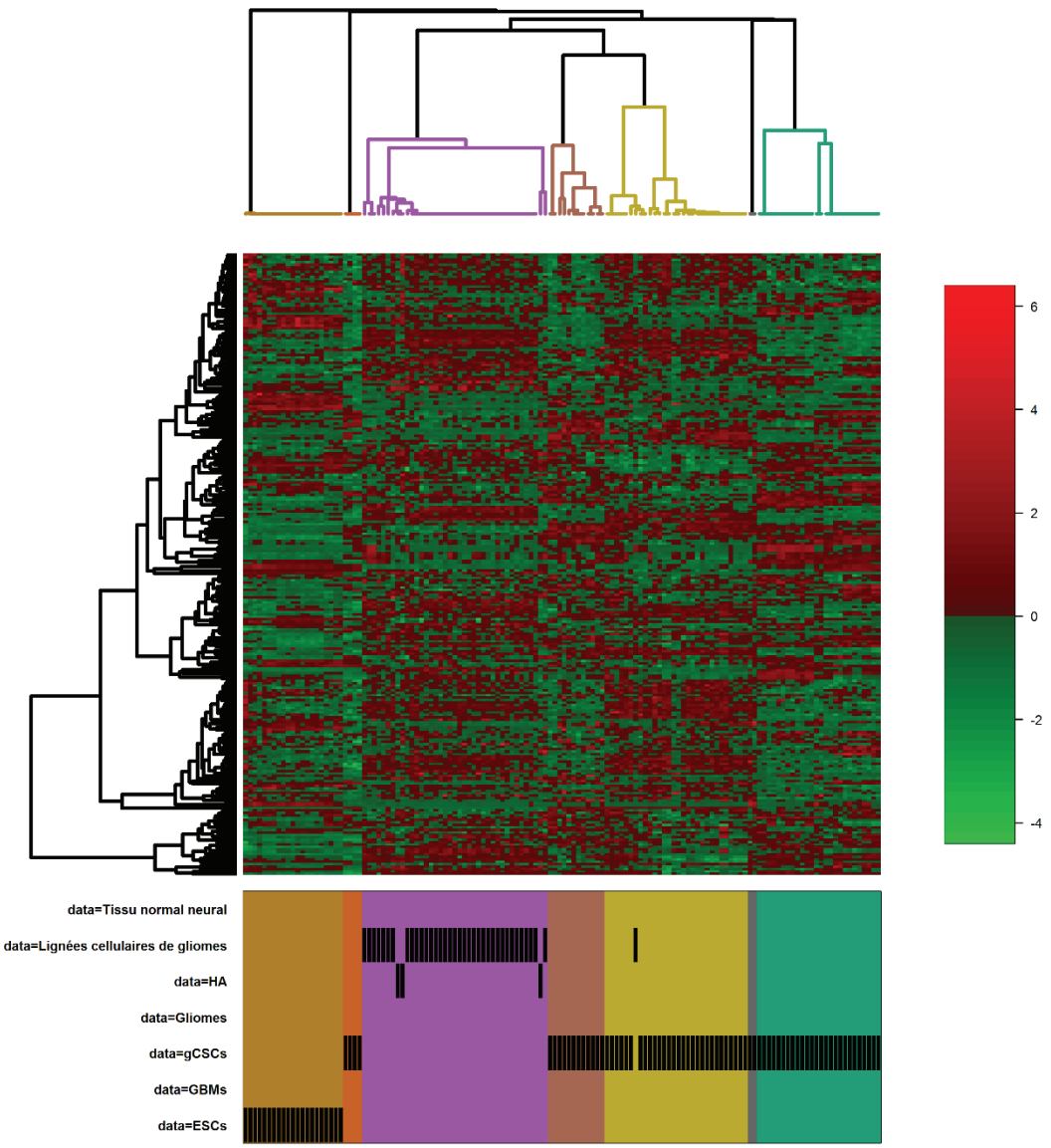

Figure 43 : Classification des échantillons du groupe "lignées cellulaires".
On obtient cinq classes principales et quelques échantillons à part.

Nous pouvons donc reprendre l'analyse précédente en ajoutant nos huit nouveaux groupes : gliomes, GBM1, GBM2, gCSC1, gCSC2, gCSC3, lignées cellulaires, ESCs.

Nous avons regardé les gènes exprimés dans les gliomes ou lignées cellulaires cancéreuses mais pas dans les tissus normaux (soient exprimés, soient sur-exprimés selon la définition de l'algorithme). On trouve six gènes impliqués dans l'entrée du calcium (STIM2, ORAI1, TRPA1, TRPP2/PKD2, TRPM7 et ITPR3), deux dans la récupération du calcium (MCU et SLC25A13) et onze codant pour des protéines se liant au calcium, putativement impliquées dans la modulation des flux de calcium (STC1, S100A11, S100A2, S100A4, S100A8, EFCAB7, EFCAB11, CETN3, RCN1, SRI et CALU).

TRPM7 a récemment été décrit comme modulant la prolifération et les métastases dans les gCSCs (Liu et al., 2014). Nous avons trouvé que ce gène est spécifiquement exprimé dans l'un des trois groupes de gCSCs. Ce résultat semble confirmer l'hétérogénéité des gCSCs isolées dans les différents laboratoires et indique que TRPM7 pourrait être un marqueur pour classer ces populations.

STIM2, ORAI1 et TRPP2 sont probablement impliqués dans le mécanisme SOCE (Store operated calcium entry) bien que le rôle de TRPP2 reste vague (Saul et al., 2014). Cependant, nos résultats suggèrent que les GBMs et gCSCs pourraient avoir des mécanismes moléculaires différents pour contrôler le processus SOCE.

Concernant la récupération du calcium, deux principaux gènes sont très exprimés dans les gCSCs, ESCs et lignées cellulaires de glioblastomes et à un niveau plus faible dans les tissus normaux, il s'agit de MCU codant pour le transporteur de calcium mitochondrial et SLC25A13 codant pour une protéine intervenant dans la régulation du calcium dans les mitochondries. En prenant en compte l'expression du récepteur de triphosphate d'inositol ITPR3 et des changements probables dans le canal CRAC dans les cellules et tissus cancéreux, nous émettons l'hypothèse que lors de l'ouverture des canaux pour remplir les stocks, une partie du calcium passe dans le cytoplasme et que pour maintenir l'homeostasie calcique du cytoplasme, les mitochondries jouent un rôle en repompant le calcium venant de l'extérieur de la cellule.

Le rôle des protéines se liant au calcium et sur-exprimées dans les tumeurs ne sera pas discuté. Nous pouvons juste remarquer que l'ensemble des protéines S100 sur-exprimé dans les gliomes et cellules cancéreuses (S100A2, S100A4, S100A8 et S100A11) a été rapporté comme marqueurs de tumeurs (Rand et al., 2008).

De plus, la concentration dans le plasma de certaines de ces protéines pourrait être utilisé comme biomarqueurs avec une utilité potentielle dans la classification et le pronostic des GBMs (Sreekanthreddy et al., 2010).

4.3 Conclusion et discussion

Nous avons proposé un processus général en quatre étapes permettant d'extraire de données transcriptomiques les gènes exprimés et sur-exprimés codant pour des protéines impliquées dans la génération et la modulation du signal calcium dans des cellules et tissus normaux et tumoraux.

Dans cette méthode, la première étape consistait à diminuer le bruit et l'hétérogénéité dans les données publiques utilisées. Le niveau de bruit estimé varie entre 10 et 20%. Les *outliers* ont été identifiés et supprimés de manière rigoureuse.

Les étapes suivantes nous ont amenés à étudier un ensemble de gènes impliqués dans la génération et la modulation du signal calcium intracellulaire en réponse à un stimulus externe. En utilisant un ensemble de gènes bien définis, la cohérence biologique de notre processus d'analyse peut être contrôlée plus facilement et la fixation des différents seuils utilisés est plus appropriée. De plus, bien qu'en général ce type d'analyse résulte en une corrélation entre un ensemble de gènes et un phénotype spécifique défini par les annotations, en mettant l'accent sur la toolbox calcium, il devient possible de proposer des hypothèses biologiques testables.

En suivant cette approche, et en se restreignant à l'analyse des gènes de la toolbox calcium impliqués dans les flux de calcium (entré et sortie), nous proposons que :

- Des modifications importantes apparaissent dans la régulation du processus SOCE dans les glioblastomes et cellules souches de glioblastomes par rapport aux tissus normaux du cerveau,
- Dans un sous-ensemble de cellules souches de glioblastomes, les mitochondries pourraient jouer un rôle important dans le mécanisme de récupération du calcium. Ce rôle pourrait être corrélé avec les modifications moléculaires du mécanisme SOCE.

Puisque nous n'avons pas accès à des données transcriptomiques de cellules souches de glioblastomes cultivées dans différentes conditions, nous ne pouvons dire si les modifications dans la gestion du signal calcium sont dues à la tumorigénicité ou bien s'ils sont intrinsèques aux cellules cultivées.

Pour aller plus loin dans l'analyse, nous avons cherché à savoir si le niveau d'expression des gènes de la toolbox calcium sur-exprimés dans les gliomes, lignées cellulaires ou cellules souches de glioblastomes (TRPM7, STIM2, ORAI1, PKD2L1 (encode TRPP3), TRPA1, CCDC109A (MCU), ITPR3, SLC8A3, STC1, S100A11, S100A2, S100A4, S100A8, EFCAB7, C14orf143, CETN3, RCN1, CALU) avaient un effet sur la survie de patients atteints de glioblastomes. Pour cela, nous avons utilisé les données de séquençage ARN du TCGA et avons regardé la survie des patients ayant une expression supérieure à la médiane pour chaque gène par rapport à ceux ayant une expression inférieure à la médiane. Nous n'avons trouvé aucune différence significative de survie à partir de ces données.

L'ensemble des résultats montrent que l'utilisation de l'algorithme de sur-expression développé pour la méthode KANT permet de trouver les gènes sur-exprimés dans des sous-groupes d'un ensemble d'échantillon.

L'ensemble de cette analyse a d'abord été effectuée sur un ensemble de données plus réduit, ne comprenant pas nos gCSCs (soit 12 échantillons de gCSCs en moins). Les résultats finaux et conclusions étaient les mêmes mais on a pu observer quelques différences : la classification des sous-classes gliomes et lignées cellulaires était différente. Au final, on obtient le même

nombre de groupes, et les mêmes "petits groupes" mais dans les 2 classes glioblastomes, certains échantillons ont changés. Concernant la classe lignées cellulaires, le fait d'ajouter de nouveaux gCSCs modifie la classification; nous n'avions que 2 classes gCSCs au lieu de 3 avec ces nouveaux échantillons. Les résultats de l'analyse calcium ne montrent pas de différence majeure, mais l'hétérogénéité des gCSCs est davantage mise en valeur. Au final, ces résultats montrent l'importance du concept de méta-analyse, notamment lorsque l'étude porte sur des échantillons hétérogènes.

Conclusion générale

Ce travail de thèse s'est articulé autour de deux objectifs principaux :

- Le développement d'une méthode générique permettant de prédire les antigènes spécifiques de cancer à partir de données de puces d'expression.
- L'analyse des cellules souches de glioblastomes, que ce soit par l'identification de protéines sur-exprimées à la surface des cellules souches de glioblastomes ou bien par l'étude des modifications du signal calcium, dérégulé dans de nombreux cancers.

Dans une première partie, nous avons cherché à développer une méthode de prédiction d'antigènes et/ou biomarqueurs putatifs de tumeurs à partir de données de puces d'expression. Pour cela, nous avons d'abord prédit l'ensemble des protéines transmembranaires, qui ont donc une partie extracellulaire accessible à un anticorps, puis avons développé un algorithme de sur-expression prenant en compte l'hétérogénéité tumorale. La méthode globale est nommée KANT pour Kancer ANtigene Tracker.

Cette étude nous a permis de créer une base de données de protéines transmembranaires, qui représentent 25% du protéome humain.

Afin de valider la méthode, nous l'avons testée sur deux jeux de données : un de cancer du sein et un de lymphome T. Nous avons obtenu l'identification de protéines déjà connues comme spécifiques, ciblées par des anticorps commercialisés ou en phase clinique (MUC1, CTLA-4), et d'autres protéines non connues dans le cas de la pathologie, potentiellement très intéressantes. Néanmoins, nous n'avons pas pu valider les cibles, que ce soit par immunohistochimie ou par *tissue array* suite au désengagement du laboratoire partenaire, chargé de cette étape.

Finalement, nous avons mis en place une méthode de priorisation des cibles, basée sur une étude approfondie de la littérature, permettant d'attribuer différents poids à ce qui ferait un « bon antigène» pour anticorps anti-cancéreux. La méthode de prédiction est disponible sous forme de package R, nommé KANT sur le CRAN.

Nous avons ensuite tenté d'identifier des biomarqueurs spécifiques des cellules souches de glioblastomes. Avant de commencer l'étude principale, nous avons regardé l'expression de cellules souches de glioblastomes de patients adultes par rapport à celles de glioblastomes de patients jeunes. Conformément à la littérature concernant les glioblastomes, les deux types de cellules souches sont différents. Nous nous sommes donc concentrés sur les échantillons de patients

adultes. Pour cela, nous avions 4 cellules souches TG1, OB1, TG10 et TG16, et 2 contrôles, HA, une lignée cellulaire d'astrocytes humains et U87 une lignée cellulaire de glioblastome. Pour ces 6 échantillons, nous avions des données de spectrométrie de masse sur les Clusters de Différenciation (CDs), des données de transcriptomiques venant de puces d'expression et également de séquençage ARN. Pour des questions pratiques, nous ne pouvions étudier l'ensemble des protéines. En se focalisant sur les CDs, nous avons fait le choix d'étudier des protéines idéales en tant que biomarqueurs : en effet, elles ont une partie extracellulaire et il existe des anticorps les ciblant, ce qui simplifie les validations et leur utilisation potentielle en tant que cible thérapeutique.

L'utilisation de l'algorithme de sur-expression de KANT sur les données de protéomique nous a permis d'identifier 13 biomarqueurs potentiels. Afin de les valider, nous avons utilisé un ensemble public de 73 échantillons de cellules souches de glioblastomes hybridés sur puce HG-U133 Plus 2. En utilisant KANT sur cet ensemble, nous avons pu valider l'expression des gènes associés aux protéines d'intérêt sur un ensemble plus significatif. Cette méthode nous a permis d'identifier 4 biomarqueurs/cibles thérapeutiques d'intérêt : LY75, SLC44A1, NCAM1 et CD97. Afin de les caractériser, nous avons étudié leurs différents transcrits dans nos 4 cellules souches, ainsi que leur expression dans le glioblastome, que ce soit l'expression par sous-type de la classification du TCGA ou bien la corrélation entre durée de vie et niveau d'expression. Notre analyse montre que LY75 est spécifique de TG1/OB1 dans nos données et qu'il est associé au groupe mésenchymateux des glioblastomes ; de plus, il semble que les transcrits exprimés dans nos cellules souches et contrôles ne soient pas les mêmes. SLC44A1 est associé aux groupes neuraux, pro-neuraux et G-CIMP des glioblastomes ; NCAM1 au groupe G-CIMP. CD97 est associé aux groupes classiques et mésenchymateux. Il est décrit dans la littérature comme sur-exprimé dans les glioblastomes et cellules souches de glioblastomes et possédant plusieurs isoformes, dont un connu pour favoriser la croissance des tumeurs et promouvoir l'invasion métastique dans différents cancers. Nous avons retrouvé les isoformes dans nos données mais ce-dernier ne semble pas plus exprimé dans nos cellules souches que dans HA.

Afin de valider nos résultats, des analyses immunohistochimiques ont été effectuées pour 3 CDs : LY75, NCAM1 et CD97. Nous avons observé un marquage d'OB1 et TG1 pour les 3 anticorps utilisés mais rien n'apparait pour TG10 et TG16. HA et U87 sont positifs à CD97. Ces résultats sont inattendus. A ce stade, nous n'avons pu qu'émettre des hypothèses pour essayer de les interpréter. Soit une modification post-traductionnelle des sites reconnus par les anticorps pour TG10/TG16 ne permet pas leur reconnaissance, soit les protéines de ces cellules sont masquées et ne sont pas reconnues par nos anticorps ou la technique utilisée. Afin d'aller plus loin dans la validation, des analyses de *tissue array* sont en cours sur des glioblastomes en collaboration avec le groupe de Thierry Virolle à Nice.

Ces résultats nous ont donc permis de proposer 4 nouveaux biomarqueurs/cibles thérapeutiques potentielles. En plus de valider leur utilisation en détectant la présence de cellules souches dans des glioblastomes, il faudrait aller plus loin dans l'analyse des différents isoformes au niveau fonctionnel, notamment pour LY75, qui est une protéine peu étudiée mais déjà détectée dans une étude portant sur le cancer des ovaires.

Cette analyse nous a permis de nous rendre compte de la différence d'expression au niveau gènes et protéines, conformément aux derniers résultats de la littérature. Ces 10 dernières années, la recherche dans le domaine de la cancérologie (et de la biologie en général) s'est beaucoup focalisée sur la génomique, grâce à l'évolution des technologies, mais la protéomique conserve un très fort intérêt biologique car beaucoup plus proche du phénotype observé. L'évolution des technologies de spectrométrie de masse devrait bientôt permettre d'atteindre le même niveau d'expérience et d'analyse en protéomique qu'en génomique (en terme de rapidité et taille des cohortes analysées) et l'utilisation conjuguée du séquençage ARN et de la protéomique (protéogénomique) promet des avancées significatives, par exemple en utilisant les variants détectés en génomique pour prédire et identifier les nouvelles protéines.

Nous avons cherché à étudier un ensemble de gènes impliqués dans la génération et la modulation du signal calcium intracellulaire, appelé toolbox calcium sur un ensemble de données publiques composé de gliomes, glioblastomes, cellules souches de glioblastomes et lignées cellulaires de gliomes. L'objectif était d'analyser les modifications de ce signal impliqué dans la régulation de nombreux mécanismes cellulaires.

La première étape de cette analyse consistait à diminuer le bruit et l'hétérogénéité dans les données publiques utilisées issues de 14 ensembles de données différents. La mise en place d'une méthodologie générale nous a permis d'estimer un niveau de bruit de 10 à 20% dans les annotations et la qualité des échantillons.

Nous avons ensuite comparé la classification de l'ensemble de nos échantillons à partir de l'ensemble des gènes par rapport à celle faite à partir de la toolbox calcium, dans le but de savoir si l'information portée par la signature calcium était différente ou identique de celle portée par l'ensemble des gènes. Nous avons obtenu des classifications différentes, ce qui nous a amené à étudier plus précisément la signature calcium.

Afin de valider notre méthode, nous avons d'abord étudié la signature calcium sur les différentes zones du cerveau humain. Nous avons obtenu des signatures par zone cohérentes avec les connaissances actuelles de la biologie du cerveau. Puis nous avons appliqué cette méthode sur les échantillons de glioblastomes et de cellules souches.

Les résultats de cette approche nous ont amené à proposer les hypothèses suivantes :

- Des modifications importantes apparaissent dans la régulation du processus SOCE dans les glioblastomes et cellules souches de glioblastomes par rapport aux tissus normaux du cerveau.
- Dans un sous-ensemble de cellules souches de glioblastomes, les mitochondries pourraient jouer un rôle important dans le mécanisme de récupération du calcium. Ce rôle pourrait être lié aux modifications moléculaires du mécanisme SOCE.

Le niveau d'expression des différents gènes impliqués dans la signalisation calcium et sur-exprimés dans les gliomes et/ou cellules souches de gliomes ne semblent pas être corrélé à la survie globale des patients atteints de glioblastomes. L'étape suivante est la validation biologique de ces hypothèses, à partir d'une étude fonctionnelle sur cellules souches et/ou glioblastomes.

L'ensemble de cette analyse nous a permis de proposer de nouvelles voies d'études sur la dérégulation des mécanismes de la signalisation calcique dans les glioblastomes et cellules souches mais aussi d'établir une méthodologie globale pour la gestion du bruit dans une méta-analyse. Cette dernière partie prend toute son importance dans le contexte de vastes bases de données publiques qui permettent de valider in silico des hypothèses biologiques à condition de prendre en compte le bruit généré par les différents traitements informatiques, analyses biostatistiques, protocoles expérimentaux et expériences de biologie moléculaire.

Ce travail de thèse nous a permis de développer une méthode générique d'identification de biomarqueurs à partir de données transcriptomiques et de valider l'utilisation de cette méthode, ainsi que d'identifier 4 biomarqueurs/ cibles thérapeutiques potentiels des cellules souches de glioblastomes, et deux mécanismes du signal calcium qui semblent dérégulés.

Bibliographie

Abbott, J.J., Amirkhan, R.H., and Hoang, M.P. (2004). Malignant melanoma with a rhabdoid phenotype: histologic, immunohistochemical, and ultrastructural study of a case and review of the literature. Arch. Pathol. Lab. Med. *128*, 686–688.

Abousamra, N.K., Salah El-Din, M., Hamza Elzahaf, E., and Esmael, M.E. (2015). Ectonucleoside triphosphate diphosphohydrolase-1 (E-NTPDase1/CD39) as a new prognostic marker in chronic lymphocytic leukemia. Leuk. Lymphoma *56*, 113–119.

Agnihotri, S., Burrell, K.E., Wolf, A., Jalali, S., Hawkins, C., Rutka, J.T., and Zadeh, G. (2013). Glioblastoma, a brief review of history, molecular genetics, animal models and novel therapeutic strategies. Arch. Immunol. Ther. Exp. (Warsz.) *61*, 25–41.

Ahram, M., Litou, Z.I., Fang, R., and Al-Tawallbeh, G. (2006). Estimation of membrane proteins in the human proteome. In Silico Biol. (Gedrukt) *6*, 379–386.

Aldaz, B., Sagardoy, A., Nogueira, L., Guruceaga, E., Grande, L., Huse, J.T., Aznar, M.A., Díez-Valle, R., Tejada-Solís, S., Alonso, M.M., et al. (2013). Involvement of miRNAs in the differentiation of human glioblastoma multiforme stem-like cells. PLoS ONE *8*, e77098.

Allison, K.H., and Sledge, G.W. (2014). Heterogeneity and cancer. Oncology (Williston Park, N.Y.) *28*, 772–778.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Andres, S.A., Brock, G.N., and Wittliff, J.L. (2013). Interrogating differences in expression of targeted gene sets to predict breast cancer outcome. BMC Cancer *13*, 326.

Anido, J., Sáez-Borderías, A., Gonzàlez-Juncà, A., Rodón, L., Folch, G., Carmona, M.A., Prieto-Sánchez, R.M., Barba, I., Martínez-Sáez, E., Prudkin, L., et al. (2010). TGF-β Receptor Inhibitors Target the CD44(high)/Id1(high) Glioma-Initiating Cell Population in Human Glioblastoma. Cancer Cell *18*, 655–668.

Arnáiz-Cot, J.J., Damon, B.J., Zhang, X.-H., Cleemann, L., Yamaguchi, N., Meissner, G., and Morad, M. (2013). Cardiac calcium signalling pathologies associated with defective calmodulin regulation of type 2 ryanodine receptor. J. Physiol. (Lond.) *591*, 4287–4299.

Aust, G., Steinert, M., Schütz, A., Boltze, C., Wahlbuhl, M., Hamann, J., and Wobus, M. (2002). CD97, but not its closely related EGF-TM7 family member EMR2, is expressed on gastric, pancreatic, and esophageal carcinomas. Am. J. Clin. Pathol. *118*, 699–707.

Baba-Aissa, F., Raeymaekers, L., Wuytack, F., Dode, L., and Casteels, R. (1998). Distribution and isoform diversity of the organellar Ca2+ pumps in the brain. Mol. Chem. Neuropathol. *33*, 199–208.

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71.

Banneau, G., Guedj, M., MacGrogan, G., de Mascarel, I., Velasco, V., Schiappa, R., Bonadona, V., David, A., Dugast, C., Gilbert-Dussardier, B., et al. (2010). Molecular apocrine differentiation is a common feature of breast cancer in patients with germline PTEN mutations. Breast Cancer Research *12*, R63.

Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem *404*, 939–965.

Bao, S., Wu, Q., McLendon, R.E., Hao, Y., Shi, Q., Hjelmeland, A.B., Dewhirst, M.W., Bigner, D.D., and Rich, J.N. (2006). Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. Nature *444*, 756–760.

Bao, S., Wu, Q., Li, Z., Sathornsumetee, S., Wang, H., McLendon, R.E., Hjelmeland, A.B., and Rich, J.N. (2008). Targeting cancer stem cells through L1CAM suppresses glioma growth. Cancer Res. *68*, 6043–6048.

Bapat, S.A., Mali, A.M., Koppikar, C.B., and Kurrey, N.K. (2005). Stem and progenitor-like cells contribute to the aggressive behavior of human epithelial ovarian cancer. Cancer Res. *65*, 3025–3029.

Beier, D., Hau, P., Proescholdt, M., Lohmeier, A., Wischhusen, J., Oefner, P.J., Aigner, L., Brawanski, A., Bogdahn, U., and Beier, C.P. (2007). CD133(+) and CD133(-) glioblastoma-derived cancer stem cells show differential growth characteristics and molecular profiles. Cancer Res. *67*, 4010–4015.

Bernier, J., Hall, E.J., and Giaccia, A. (2004). Radiation oncology: a century of achievements. Nat. Rev. Cancer *4*, 737–747.

Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., and Elofsson, A. (2008). Prediction of membrane-protein topology from first principles. Proc. Natl. Acad. Sci. U.S.A. *105*, 7177–7181.

Bernsel, A., Viklund, H., Hennerdal, A., and Elofsson, A. (2009). TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res. *37*, W465–W468.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.

Bertos, N.R., and Park, M. (2011). Breast cancer — one term, many entities? J Clin Invest *121*, 3789–3796.

Bexell, D., Gunnarsson, S., Siesjö, P., Bengzon, J., and Darabi, A. (2009). CD133+ and nestin+ tumor-initiating cells dominate in N29 and N32 experimental gliomas. Int. J. Cancer *125*, 15–22.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bonnet, D., and Dick, J.E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat. Med. *3*, 730–737.

Brennan, C.W., Verhaak, R.G.W., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H., et al. (2013). The somatic genomic landscape of glioblastoma. Cell *155*, 462–477.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Cambon, K., Hansen, S.M., Venero, C., Herrero, A.I., Skibo, G., Berezin, V., Bock, E., and Sandi, C. (2004). A synthetic neural cell adhesion molecule mimetic peptide promotes synaptogenesis, enhances presynaptic function, and facilitates memory consolidation. J. Neurosci. *24*, 4197–4204.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Carvalho, B.S., and Irizarry, R.A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics *26*, 2363–2367.

Chabner, B.A., and Roberts, T.G. (2005). Timeline: Chemotherapy and the war on cancer. Nat. Rev. Cancer *5*, 65–72.

Chaichana, K.L., Jusue-Torres, I., Navarro-Ramirez, R., Raza, S.M., Pascual-Gallego, M., Ibrahim, A., Hernandez-Hermann, M., Gomez, L., Ye, X., Weingart, J.D., et al. (2014). Establishing percent resection and residual volume thresholds affecting survival and recurrence for patients with newly diagnosed intracranial glioblastoma. Neuro-Oncology *16*, 113–122.

Cheever, M.A., Allison, J.P., Ferris, A.S., Finn, O.J., Hastings, B.M., Hecht, T.T., Mellman, I., Prindiville, S.A., Viner, J.L., Weiner, L.M., et al. (2009). The prioritization of cancer

antigens: a national cancer institute pilot project for the acceleration of translational research. Clin. Cancer Res. *15*, 5323–5337.

Chen, C.P., and Rost, B. (2002). State-of-the-art in membrane protein prediction. Appl. Bioinformatics *1*, 21–35.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS One *6*.

Chen, D.-T., Nasir, A., Culhane, A., Venkataramu, C., Fulp, W., Rubio, R., Wang, T., Agrawal, D., McCarthy, S.M., Gruidl, M., et al. (2010). Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. Breast Cancer Res Treat *119*, 335–346.

Cheng, L., Wu, Q., Huang, Z., Guryanova, O.A., Huang, Q., Shou, W., Rich, J.N., and Bao, S. (2011). L1CAM regulates DNA damage checkpoint response of glioblastoma stem cells through NBS1. EMBO J. *30*, 800–813.

Chin, L., Andersen, J.N., and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. Nat. Med. *17*, 297–303.

Cho, E.Y., Choi, Y., Chae, S.W., Sohn, J.H., and Ahn, G.H. (2006). Immunohistochemical study of the expression of adhesion molecules in ovarian serous neoplasms. Pathol. Int. *56*, 62–70.

Choi, Y.-L., Xuan, Y.H., Shin, Y.K., Chae, S.W., Kook, M.C., Sung, R.H., Youn, S.J., Choi, J.W., and Kim, S.H. (2004). An immunohistochemical study of the expression of adhesion molecules in gallbladder lesions. J. Histochem. Cytochem. *52*, 591–601.

Chou, P.Y., and Fasman, G.D. (1974). Prediction of protein conformation. Biochemistry *13*, 222–245.

Clement, V., Sanchez, P., de Tribolet, N., Radovanovic, I., and Ruiz i Altaba, A. (2007). HEDGEHOG-GLI1 signaling regulates human glioma growth, cancer stem cell self-renewal, and tumorigenicity. Curr. Biol. *17*, 165–172.

Cohen, M.H., Shen, Y.L., Keegan, P., and Pazdur, R. (2009). FDA drug approval summary: bevacizumab (Avastin) as treatment of recurrent glioblastoma multiforme. Oncologist *14*, 1131–1138.

Collins, A.T., Berry, P.A., Hyde, C., Stower, M.J., and Maitland, N.J. (2005). Prospective identification of tumorigenic prostate cancer stem cells. Cancer Res. *65*, 10946–10951.

Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Crabtree, G.R. (2001). Calcium, calcineurin, and the control of transcription. J. Biol. Chem. *276*, 2313–2316.

Cuthbertson, J.M., Doyle, D.A., and Sansom, M.S.P. (2005). Transmembrane Helix Prediction: A Comparative Evaluation and Analysis. Protein Engineering, Design and Selection *18*, 295–308.

Daniel, L., Bouvier, C., Chetaille, B., Gouvernet, J., Luccioni, A., Rossi, D., Lechevallier, E., Muracciole, X., Coulange, C., and Figarella-Branger, D. (2003). Neural cell adhesion molecule expression in renal cell carcinomas: relation to metastatic behavior. Hum. Pathol. *34*, 528–532.

Dean, M., Fojo, T., and Bates, S. (2005). Tumour stem cells and drug resistance. Nat. Rev. Cancer *5*, 275–284.

DeVita, V.T., and Chu, E. (2008). A history of cancer chemotherapy. Cancer Res. *68*, 8643–8653.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dubois, C., Vanden Abeele, F., and Prevarskaya, N. (2013). Targeting apoptosis by the remodelling of calcium-transporting proteins in cancerogenesis. FEBS J. *280*, 5500–5510.

Eichler, W., Hamann, J., and Aust, G. (1997). Expression characteristics of the human CD97 antigen. Tissue Antigens *50*, 429–438.

Elias, J.E., and Gygi, S.P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol. Biol. *604*, 55–71.

Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigó, R., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods *10*, 1185–1191.

Eramo, A., Lotti, F., Sette, G., Pilozzi, E., Biffoni, M., Di Virgilio, A., Conticello, C., Ruco, L., Peschle, C., and De Maria, R. (2008). Identification and expansion of the tumorigenic lung cancer stem cell population. Cell Death Differ. *15*, 504–514.

Etzell, J.E., Keet, C., McDonald, W., and Banerjee, A. (2006). Medulloblastoma simulating acute myeloid leukemia: case report with a review of "myeloid antigen" expression in nonhematopoietic tissues and tumors. J. Pediatr. Hematol. Oncol. *28*, 703–710.

Evans, A.J., Humphrey, P.A., Belani, J., van der Kwast, T.H., and Srigley, J.R. (2006). Large cell neuroendocrine carcinoma of prostate: a clinicopathologic summary of 7 cases of a rare manifestation of advanced prostate cancer. Am. J. Surg. Pathol. *30*, 684–693.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F.M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE *8*, e85024.

Facchino, S., Abdouh, M., Chatoo, W., and Bernier, G. (2010). BMI1 confers radioresistance to normal and cancerous neural stem cells through recruitment of the DNA damage response machinery. J. Neurosci. *30*, 10096–10111.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. Genomics, Proteomics & Bioinformatics *13*, 4–16.

Fong, C.-Y., Chak, L.-L., Biswas, A., Tan, J.-H., Gauthaman, K., Chan, W.-K., and Bongso, A. (2011). Human Wharton's jelly stem cells have unique transcriptome profiles compared to human embryonic stem cells and other mesenchymal stem cells. Stem Cell Rev *7*, 1–16.

Galle, J., Sittig, D., Hanisch, I., Wobus, M., Wandel, E., Loeffler, M., and Aust, G. (2006). Individual cell-based models of tumor-environment interactions: Multiple effects of CD97 on tumor invasion. Am. J. Pathol. *169*, 1802–1811.

Garcia, J.L., Perez-Caro, M., Gomez-Moreta, J.A., Gonzalez, F., Ortiz, J., Blanco, O., Sancho, M., Hernandez-Rivas, J.M., Gonzalez-Sarmiento, R., and Sanchez-Martin, M. (2010). Molecular analysis of ex-vivo CD133+ GBM cells revealed a common invasive and angiogenic profile but different proliferative signatures among high grade gliomas. BMC Cancer *10*, 454.

Garnier, J., Osguthorpe, D.J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. *120*, 97–120.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315.

Ge, Y., Zhou, F., Chen, H., Cui, C., Liu, D., Li, Q., Yang, Z., Wu, G., Sun, S., Gu, J., et al. (2010). Sox2 is translationally activated by eukaryotic initiation factor 4E in human glioma-initiating cells. Biochem. Biophys. Res. Commun. *397*, 711–717.

Giridhar, P.V., Funk, H.M., Gallo, C.A., Porollo, A., Mercer, C.A., Plas, D.R., and Drew, A.F. (2011). Interleukin-6 receptor enhances early colonization of the murine omentum by upregulation of a mannose family receptor, LY75, in ovarian tumor cells. Clin. Exp. Metastasis *28*, 887–897.

Glunde, K., Jie, C., and Bhujwalla, Z.M. (2004). Molecular causes of the aberrant choline phospholipid metabolism in breast cancer. Cancer Res. *64*, 4270–4276.

Glunde, K., Jacobs, M.A., and Bhujwalla, Z.M. (2006). Choline metabolism in cancer: implications for diagnosis and therapy. Expert Rev. Mol. Diagn. *6*, 821–829.

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., et al. (2012). A refined molecular taxonomy of breast cancer. Oncogene *31*, 1196–1206.

Guo, J., Hammar, M., Öberg, L., Padmanabhuni, S.S., Bjäreland, M., and Dalevi, D. (2013). Combining Evidence of Preferential Gene-Tissue Relationships from Multiple Sources. PLoS ONE *8*, e70568.

Guo, Y., Liu, S., Wang, P., Zhao, S., Wang, F., Bing, L., Zhang, Y., Ling, E.-A., Gao, J., and Hao, A. (2011). Expression profile of embryonic stem cell-associated genes Oct4, Sox2 and Nanog in human gliomas. Histopathology *59*, 763–775.

Haas, B.J., and Zody, M.C. (2010). Advancing RNA-Seq analysis. Nat. Biotechnol. *28*, 421–423.

Van Haasteren, G., Li, S., Muda, M., Susini, S., and Schlegel, W. (1999). Calcium signalling and gene expression. J. Recept. Signal Transduct. Res. *19*, 481–492.

Hägerstrand, D., He, X., Bradic Lindh, M., Hoefs, S., Hesselager, G., Ostman, A., and Nistér, M. (2011). Identification of a SOX2-dependent subset of tumor- and sphere-forming glioblastoma cells with a distinct tyrosine kinase inhibitor sensitivity profile. Neuro-Oncology *13*, 1178–1191.

Haiech, J., Audran, E., Fève, M., Ranjeva, R., and Kilhoffer, M.-C. (2011). Revisiting intracellular calcium signaling semantics. Biochimie *93*, 2029–2037.

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proc. Natl. Acad. Sci. U.S.A. *100*, 3983–3988.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hara, T., Kosaka, N., Shinoura, N., and Kondo, T. (1997). PET imaging of brain tumor with [methyl-11C]choline. J. Nucl. Med. *38*, 842–847.

Hara, T., Kosaka, N., and Kishi, H. (1998). PET imaging of prostate cancer using carbon-11-choline. J. Nucl. Med. *39*, 990–995.

Hara, T., Inagaki, K., Kosaka, N., and Morita, T. (2000). Sensitive detection of mediastinal lymph node metastasis of lung cancer with 11C-choline PET. J. Nucl. Med. *41*, 1507–1513.

Hara, T., Kondo, T., Hara, T., and Kosaka, N. (2003). Use of 18F-choline and 11C-choline as contrast agents in positron emission tomography imaging-guided stereotactic biopsy sampling of gliomas. J. Neurosurg. *99*, 474–479.

Heijne, G. (1986). The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J. *5*, 3021–3027.

Hellevik, T., and Martinez-Zubiaurre, I. (2014). Radiotherapy and the tumor stroma: the importance of dose and fractionation. Front Oncol *4*, 1.

Helms, M.W., Kemming, D., Pospisil, H., Vogt, U., Buerger, H., Korsching, E., Liedtke, C., Schlotter, C.M., Wang, A., Chan, S.Y., et al. (2008). Squalene epoxidase, located on chromosome 8q24.1, is upregulated in 8q+ breast cancer and indicates poor clinical outcome in stage I and II disease. Br. J. Cancer *99*, 774–780.

Heo, D.K., Lim, H.M., Nam, J.H., Lee, M.G., and Kim, J.Y. (2015). Regulation of phagocytosis and cytokine secretion by store-operated calcium entry in primary isolated murine microglia. Cell. Signal. *27*, 177–186.

Hermann, P.C., Huber, S.L., Herrler, T., Aicher, A., Ellwart, J.W., Guba, M., Bruns, C.J., and Heeschen, C. (2007). Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. Cell Stem Cell *1*, 313–323.

Hoffman, N.E., Chandramoorthy, H.C., Shanmughapriya, S., Zhang, X.Q., Vallem, S., Doonan, P.J., Malliankaraman, K., Guo, S., Rajan, S., Elrod, J.W., et al. (2014). SLC25A23 augments mitochondrial Ca^{2+} uptake, interacts with MCU, and induces oxidative stress-mediated cell death. Mol. Biol. Cell *25*, 936–947.

Hu, Z., Fan, C., Oh, D.S., Marron, J., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics *7*, 96.

Huang, Y., de Reyniès, A., de Leval, L., Ghazi, B., Martin-Garcia, N., Travert, M., Bosq, J., Brière, J., Petit, B., Thomas, E., et al. (2010). Gene expression profiling identifies emerging oncogenic pathways operating in extranodal NK/T-cell lymphoma, nasal type. Blood *115*, 1226–1237.

Hubbell, E., Liu, W.-M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585–1592.

Huen, A.O., and Rook, A.H. (2014). Toll receptor agonist therapy of skin cancer and cutaneous T-cell lymphoma. Curr Opin Oncol *26*, 237–244.

Ikushima, H., Todo, T., Ino, Y., Takahashi, M., Saito, N., Miyazawa, K., and Miyazono, K. (2011). Glioma-initiating cells retain their tumorigenicity through integration of the Sox axis and Oct4 protein. J. Biol. Chem. *286*, 41434–41441.

Inazu, M., Yamada, T., Kubota, N., and Yamanaka, T. (2013). Functional expression of choline transporter-like protein 1 (CTL1) in small cell lung carcinoma cells: a target molecule for lung cancer therapy. Pharmacol. Res. *76*, 119–131.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Johnson, J.P. (1999). Cell adhesion molecules in the development and progression of malignant melanoma. Cancer Metastasis Rev. *18*, 345–357.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat *8*, 118–127.

Jones, D.T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics *23*, 538.

Joo, K.M., Kim, S.Y., Jin, X., Song, S.Y., Kong, D.-S., Lee, J.-I., Jeon, J.W., Kim, M.H., Kang, B.G., Jung, Y., et al. (2008). Clinical and biological implications of CD133-positive and CD133-negative cells in glioblastomas. Lab. Invest. *88*, 808–815.

Kahn, M. (2014). Can we safely target the WNT pathway? Nat Rev Drug Discov *13*, 513–532.

Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. *338*, 1027–1036.

Kanamori, M., Kawaguchi, T., Nigro, J.M., Feuerstein, B.G., Berger, M.S., Miele, L., and Pieper, R.O. (2007). Contribution of Notch signaling activation to human glioblastoma multiforme. J. Neurosurg. *106*, 417–427.

Katoh, M., and Katoh, M. (2007). WNT signaling pathway and stem cell signaling network. Clin. Cancer Res. *13*, 4042–4045.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods *7*, 1009–1015.

Kim, J.-B. (2014). Channelopathies. Korean J Pediatr *57*, 1–18.

Kito, H., Yamamura, H., Suzuki, Y., Yamamura, H., Ohya, S., Asai, K., and Imaizumi, Y. (2015). Regulation of store-operated Ca2+ entry activity by cell cycle dependent up-regulation of Orai2 in brain capillary endothelial cells. Biochem. Biophys. Res. Commun. *459*, 457–462.

Kwee, S.A., Wei, H., Sesterhenn, I., Yun, D., and Coel, M.N. (2006). Localization of primary prostate cancer with dual-phase 18F-fluorocholine PET. J. Nucl. Med. *47*, 262–269.

Lander, E.S. (1999). Array of hope. Nat Genet *21*, 3–4.

Lathia, J.D., Gallagher, J., Heddleston, J.M., Wang, J., Eyler, C.E., Macswords, J., Wu, Q., Vasanji, A., McLendon, R.E., Hjelmeland, A.B., et al. (2010). Integrin alpha 6 regulates glioblastoma stem cells. Cell Stem Cell *6*, 421–432.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.

Leong, H.S., Yates, T., Wilson, C., and Miller, C.J. (2005). ADAPT: a database of affymetrix probesets and transcripts. Bioinformatics *21*, 2552–2553.

De Leval, L., Rickman, D.S., Thielen, C., de Reynies, A., Huang, Y.-L., Delsol, G., Lamant, L., Leroy, K., Brière, J., and Molina, T. (2007). The gene expression profile of nodal peripheral T-cell lymphoma demonstrates a molecular link between angioimmunoblastic T-cell lymphoma (AITL) and follicular helper T (TFH) cells. Blood *109*, 4952–4963.

De Leval, L., Bisig, B., Thielen, C., Boniver, J., and Gaulard, P. (2009). Molecular classification of T-cell lymphomas. Crit. Rev. Oncol. Hematol. *72*, 125–143.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, Q., Birkbak, N.J., Gyorffy, B., Szallasi, Z., and Eklund, A.C. (2011). Jetset: selecting the optimal microarray probe set to represent a gene. BMC Bioinformatics *12*, 474.

Li, X.-F., Kraev, A.S., and Lytton, J. (2002). Molecular cloning of a fourth member of the potassium-dependent sodium-calcium exchanger gene family, NCKX4. J. Biol. Chem. *277*, 48410–48417.

Liu, D., Trojanowicz, B., Ye, L., Li, C., Zhang, L., Li, X., Li, G., Zheng, Y., and Chen, L. (2012a). The invasion and metastasis promotion role of CD97 small isoform in gastric carcinoma. PLoS ONE *7*, e39989.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012b). Comparison of Next-Generation Sequencing Systems. J Biomed Biotechnol *2012*.

Liu, M., Inoue, K., Leng, T., Guo, S., and Xiong, Z. (2014). TRPM7 channels regulate glioma stem cell through STAT3 and Notch signaling pathways. Cell. Signal. *26*, 2773–2781.

Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W., and Kleihues, P. (2007). The 2007 WHO Classification of Tumours of the Central Nervous System. Acta Neuropathol *114*, 97–109.

MacDonald, J.W., and Ghosh, D. (2006). COPA--cancer outlier profile analysis. Bioinformatics *22*, 2950–2951.

Machová, E., O'Regan, S., Newcombe, J., Meunier, F.-M., Prentice, J., Dove, R., Lisá, V., and Dolezal, V. (2009). Detection of choline transporter-like 1 protein CTL1 in neuroblastoma x glioma cells and in the CNS, and its role in choline uptake. J. Neurochem. *110*, 1297–1309.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database *2011*, bar009–bar009.

Mallon, B.S., Chenoweth, J.G., Johnson, K.R., Hamilton, R.S., Tesar, P.J., Yavatkar, A.S., Tyson, L.J., Park, K., Chen, K.G., Fann, Y.C., et al. (2013). StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. Stem Cell Res *10*, 57–66.

Mani, S.A., Guo, W., Liao, M.-J., Eaton, E.N., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M., et al. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell *133*, 704–715.

Mann, M., Kulak, N.A., Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. Mol. Cell *49*, 583–590.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, pp. 10–12.

Mathieu, J., Zhang, Z., Zhou, W., Wang, A.J., Heddleston, J.M., Pinna, C.M.A., Hubaud, A., Stadler, B., Choi, M., Bar, M., et al. (2011). HIF induces human embryonic stem cell markers in cancer cells. Cancer Res. *71*, 4640–4652.

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci U S A *74*, 560–564.

McCall, M.N., Bolstad, B.M., and Irizarry, R.A. (2010). Frozen robust multiarray analysis (fRMA). Biostatistics *11*, 242–253.

Miller, L.D., Coffman, L.G., Chou, J.W., Black, M.A., Bergh, J., D'Agostino, R., Jr, Torti, S.V., and Torti, F.M. (2011). An iron regulatory gene signature predicts outcome in breast cancer. Cancer Res. *71*, 6728–6737.

Monge, J., Kricun, M., Radovčić, J., Radovčić, D., Mann, A., and Frayer, D.W. (2013). Fibrous dysplasia in a 120,000+ year old Neandertal from Krapina, Croatia. PLoS ONE *8*, e64539.

Monteith, G.R., Davis, F.M., and Roberts-Thomson, S.J. (2012). Calcium Channels and Pumps in Cancer: Changes and Consequences. J. Biol. Chem. *287*, 31666–31673.

Morel, A.-P., Lièvre, M., Thomas, C., Hinkal, G., Ansieau, S., and Puisieux, A. (2008). Generation of breast cancer stem cells through epithelial-mesenchymal transition. PLoS ONE *3*, e2888.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Moser, J.J., and Fritzler, M.J. (2010). The microRNA and messengerRNA profile of the RNA-induced silencing complex in human primary astrocyte and astrocytoma cells. PLoS ONE *5*, e13445.

Nagpal, S., Harsh, G., and Recht, L. (2011). Bevacizumab improves quality of life in patients with recurrent glioblastoma. Chemother Res Pract *2011*, 602812.

Nariai, N., Hirose, O., Kojima, K., and Nagasaki, M. (2013). TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. Bioinformatics *29*, 2292–2299.

Neely Kayala, K.M., Dickinson, G.D., Minassian, A., Walls, K.C., Green, K.N., and Laferla, F.M. (2012). Presenilin-null cells have altered two-pore calcium channel expression and lysosomal calcium: implications for lysosomal function. Brain Res. *1489*, 8–16.

Niu, C.-S., Li, D.-X., Liu, Y.-H., Fu, X.-M., Tang, S.-F., and Li, J. (2011). Expression of NANOG in human gliomas and its relationship with undifferentiated glioma cells. Oncol. Rep. *26*, 593–601.

Noël, G., Schott, R., Froelich, S., Gaub, M.-P., Boyer, P., Fischer-Lokou, D., Dufour, P., Kehrli, P., and Maitrot, D. (2012). Retrospective comparison of chemoradiotherapy followed by adjuvant chemotherapy, with or without prior gliadel implantation (carmustine) after initial surgery in patients with newly diagnosed high-grade gliomas. Int. J. Radiat. Oncol. Biol. Phys. *82*, 749–755.

Nugent, T., and Jones, D.T. (2009). Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics *10*, 159.

O'Brien, C.A., Pollett, A., Gallinger, S., and Dick, J.E. (2007). A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature *445*, 106–110.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. *250*, 4007–4021.

Ogden, A.T., Waziri, A.E., Lochhead, R.A., Fusco, D., Lopez, K., Ellis, J.A., Kang, J., Assanah, M., McKhann, G.M., Sisti, M.B., et al. (2008). Identification of A2B5+CD133- tumor-initiating cells in adult human gliomas. Neurosurgery *62*, 505–514; discussion 514–515.

Orringer, D., Lau, D., Khatri, S., Zamora-Berridi, G.J., Zhang, K., Wu, C., Chaudhary, N., and Sagher, O. (2012). Extent of resection in patients with glioblastoma: limiting factors, perception of resectability, and effect on survival. J. Neurosurg. *117*, 851–859.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct *4*, 14.

Ostrom, Q.T., Gittleman, H., Liao, P., Rouse, C., Chen, Y., Dowling, J., Wolinsky, Y., Kruchko, C., and Barnholtz-Sloan, J. (2014). CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007–2011. Neuro Oncol *16*, iv1–iv63.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Patru, C., Romao, L., Varlet, P., Coulombel, L., Raponi, E., Cadusseau, J., Renault-Mihara, F., Thirant, C., Leonard, N., Berhneim, A., et al. (2010). CD133, CD15/SSEA-1, CD34 or side populations do not resume tumor-initiating properties of long-term cultured cancer stem cells from human malignant glio-neuronal tumors. BMC Cancer *10*, 66.

Pattabiraman, D.R., and Weinberg, R.A. (2014). Tackling the cancer stem cells — what challenges do they pose? Nat Rev Drug Discov *13*, 497–512.

Payne, S.H. (2015). The utility of protein and mRNA correlation. Trends Biochem. Sci. *40*, 1–3.

Prevarskaya, N., Ouadid-Ahidouch, H., Skryma, R., and Shuba, Y. (2014). Remodelling of Ca2+ transport in cancer: how it contributes to cancer hallmarks? Philosophical Transactions of the Royal Society of London B: Biological Sciences *369*, 20130097.

Pujol, J.L., Simony, J., Demoly, P., Charpentier, R., Laurent, J.C., Daurès, J.P., Lehmann, M., Guyot, V., Godard, P., and Michel, F.B. (1993). Neural cell adhesion molecule and prognosis of surgically resected lung cancer. Am. Rev. Respir. Dis. *148*, 1071–1075.

Qiu, Z.-K., Shen, D., Chen, Y.-S., Yang, Q.-Y., Guo, C.-C., Feng, B.-H., and Chen, Z.-P. (2014). Enhanced MGMT expression contributes to temozolomide resistance in glioma stem-like cells. Chin J Cancer *33*, 115–122.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 341.

Rampazzo, E., Persano, L., Pistollato, F., Moro, E., Frasson, C., Porazzi, P., Della Puppa, A., Bresolin, S., Battilana, G., Indraccolo, S., et al. (2013). Wnt activation promotes neuronal differentiation of glioblastoma. Cell Death Dis *4*, e500.

Rand, V., Prebble, E., Ridley, L., Howard, M., Wei, W., Brundler, M.-A., Fee, B.E., Riggins, G.J., Coyle, B., Grundy, R.G., et al. (2008). Investigation of chromosome 1q reveals differential expression of members of the S100 family in clinical subgroups of intracranial paediatric ependymoma. Br. J. Cancer *99*, 1136–1143.

Raspadori, D., Damiani, D., Lenoci, M., Rondelli, D., Testoni, N., Nardi, G., Sestigiani, C., Mariotti, C., Birtolo, S., Tozzi, M., et al. (2001). CD56 antigenic expression in acute myeloid leukemia identifies patients with poor clinical prognosis. Leukemia *15*, 1161–1164.

Raspadori, D., Damiani, D., Michieli, M., Stocchi, R., Gentili, S., Gozzetti, A., Masolini, P., Michelutti, A., Geromin, A., Fanin, R., et al. (2002). CD56 and PGP expression in acute myeloid leukemia: impact on clinical outcome. Haematologica *87*, 1135–1140.

Ravandi, F., Cortes, J., Estrov, Z., Thomas, D., Giles, F.J., Huh, Y.O., Pierce, S., O'Brien, S., Faderl, S., and Kantarjian, H.M. (2002). CD56 expression predicts occurrence of CNS disease in acute lymphoblastic leukemia. Leuk. Res. *26*, 643–649.

Read, T.-A., Fogarty, M.P., Markant, S.L., McLendon, R.E., Wei, Z., Ellison, D.W., Febbo, P.G., and Wechsler-Reya, R.J. (2009). Identification of CD15 as a marker for tumor-propagating cells in a mouse model of medulloblastoma. Cancer Cell *15*, 135–147.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1998). GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics *14*, 656–664.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Rheinbay, E., Suvà, M.L., Gillespie, S.M., Wakimoto, H., Patel, A.P., Shahid, M., Oksuz, O., Rabkin, S.D., Martuza, R.L., Rivera, M.N., et al. (2013). An aberrant transcription factor network essential for Wnt signaling and stem cell maintenance in glioblastoma. Cell Rep *3*, 1567–1579.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25.

De Ronde, J.J., Rigaill, G., Rottenberg, S., Rodenhuis, S., and Wessels, L.F.A. (2013). Identifying subgroup markers in heterogeneous populations. Nucleic Acids Res *41*, e200.

Ronkainen, H., Soini, Y., Vaarala, M.H., Kauppila, S., and Hirvikoski, P. (2010). Evaluation of neuroendocrine markers in renal cell carcinoma. Diagn Pathol *5*, 28.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. Protein Sci. *4*, 521–533.

Rueda, C.B., Llorente-Folch, I., Amigo, I., Contreras, L., González-Sánchez, P., Martínez-Valero, P., Juaristi, I., Pardo, B., del Arco, A., and Satrústegui, J. (2014). Ca(2+) regulation of mitochondrial function in neurons. Biochim. Biophys. Acta *1837*, 1617–1624.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., et al. (2012). ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Research *41*, D987–D990.

Safaee, M., Clark, A.J., Oh, M.C., Ivan, M.E., Bloch, O., Kaur, G., Sun, M.Z., Kim, J.M., Oh, T., Berger, M.S., et al. (2013). Overexpression of CD97 confers an invasive phenotype in glioblastoma cells and is associated with decreased survival of glioblastoma patients. PLoS ONE *8*, e62765.

Safaee, M., Fakurnejad, S., Bloch, O., Clark, A.J., Ivan, M.E., Sun, M.Z., Oh, T., Phillips, J.J., and Parsa, A.T. (2015). Proportional upregulation of CD97 isoforms in glioblastoma and glioblastoma-derived brain tumor initiating cells. PLoS ONE *10*, e0111532.

Sandberg, C.J., Altschuler, G., Jeong, J., Strømme, K.K., Stangeland, B., Murrell, W., Grasmo-Wendler, U.-H., Myklebost, O., Helseth, E., Vik-Mo, E.O., et al. (2013). Comparison of glioma stem cells to neural stem cells from the adult human brain identifies dysregulated Wnt- signaling and a fingerprint associated with clinical outcome. Exp. Cell Res. *319*, 2230–2243.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology *94*, 441–448.

Saul, S., Stanisz, H., Backes, C.S., Schwarz, E.C., and Hoth, M. (2014). How ORAI and TRP channels interfere with each other: interaction models and examples from the immune system and the skin. Eur. J. Pharmacol. *739*, 49–59.

Schulte, A., Günther, H.S., Phillips, H.S., Kemming, D., Martens, T., Kharbanda, S., Soriano, R.H., Modrusan, Z., Zapf, S., Westphal, M., et al. (2011). A distinct subset of glioma cell lines with stem cell-like properties reflects the transcriptional phenotype of glioblastomas and overexpresses CXCR4 as therapeutic target. Glia *59*, 590–602.

Sekulic, A., Liang, W.S., Tembe, W., Izatt, T., Kruglyak, S., Kiefer, J.A., Cuyugan, L., Zismann, V., Legendre, C., Pittelkow, M.R., et al. (2015). Personalized treatment of Sézary syndrome by targeting a novel CTLA4:CD28 fusion. Mol Genet Genomic Med *3*, 130–136.

Sharman, J.L., Mpamhanga, C.P., Spedding, M., Germain, P., Staels, B., Dacquet, C., Laudet, V., and Harmar, A.J. (2011). IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. Nucleic Acids Res. *39*, D534–D538.

Shrimpton, R.E., Butler, M., Morel, A.-S., Eren, E., Hue, S.S., and Ritter, M.A. (2009). CD205 (DEC-205): a recognition receptor for apoptotic and necrotic self. Mol. Immunol. *46*, 1229–1239.

Singh, S.K., Clarke, I.D., Terasaki, M., Bonn, V.E., Hawkins, C., Squire, J., and Dirks, P.B. (2003). Identification of a cancer stem cell in human brain tumors. Cancer Res. *63*, 5821–5828.

Smeds, L., and Künstner, A. (2011). ConDeTri - A Content Dependent Read Trimmer for Illumina Data. PLoS One *6*.

Son, M.J., Woolard, K., Nam, D.-H., Lee, J., and Fine, H.A. (2009). SSEA-1 is an enrichment marker for tumor-initiating cells in human glioblastoma. Cell Stem Cell *4*, 440–452.

Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol *6*, 175–182.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. U.S.A. *100*, 8418–8423.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Sreekanthreddy, P., Srinivasan, H., Kumar, D.M., Nijaguna, M.B., Sridevi, S., Vrinda, M., Arivazhagan, A., Balasubramaniam, A., Hegde, A.S., Chandramouli, B.A., et al. (2010). Identification of potential serum biomarkers of glioblastoma: serum osteopontin levels correlate with poor prognosis. Cancer Epidemiol. Biomarkers Prev. *19*, 1409–1422.

Steinert, M., Wobus, M., Boltze, C., Schütz, A., Wahlbuhl, M., Hamann, J., and Aust, G. (2002). Expression and regulation of CD97 in colorectal carcinoma cell lines and tumor tissues. Am. J. Pathol. *161*, 1657–1667.

Stummer, W., van den Bent, M.J., and Westphal, M. (2011). Cytoreductive surgery of glioblastoma as the key to successful adjuvant therapies: new arguments in an old discussion. Acta Neurochir (Wien) *153*, 1211–1218.

Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J.B., Belanger, K., Brandes, A.A., Marosi, C., Bogdahn, U., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N. Engl. J. Med. *352*, 987–996.

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D.-A., Jones, D.T.W., Konermann, C., Pfaff, E., Tönjes, M., Sill, M., Bender, S., et al. (2012). Hotspot Mutations in H3F3A

and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. Cancer Cell *22*, 425–437.

Sturm, D., Bender, S., Jones, D.T.W., Lichter, P., Grill, J., Becher, O., Hawkins, C., Majewski, J., Jones, C., Costello, J.F., et al. (2014). Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. Nat. Rev. Cancer *14*, 92–107.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U.S.A. *101*, 6062–6067.

Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., et al. (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell *9*, 287–300.

Takebe, N., Harris, P.J., Warren, R.Q., and Ivy, S.P. (2011). Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways. Nat Rev Clin Oncol *8*, 97–106.

Tchoghandjian, A., Baeza, N., Colin, C., Cayre, M., Metellus, P., Beclin, C., Ouafik, L., and Figarella-Branger, D. (2010). A2B5 cells from human glioblastoma have cancer stem cell properties. Brain Pathol. *20*, 211–221.

Thiery, J.P. (2002). Epithelial-mesenchymal transitions in tumour progression. Nat. Rev. Cancer *2*, 442–454.

Thirant, C., Bessette, B., Varlet, P., Puget, S., Cadusseau, J., Dos Reis Tavares, S., Studler, J.-M., Silvestre, D.C., Susini, A., Villa, C., et al. (2011). Clinical Relevance of Tumor Cells with Stem-Like Properties in Pediatric Brain Tumors. PLoS One *6*.

Thon, N., Damianoff, K., Hegermann, J., Grau, S., Krebs, B., Schnell, O., Tonn, J.-C., and Goldbrunner, R. (2010). Presence of pluripotent CD133+ cells correlates with malignancy of gliomas. Mol. Cell. Neurosci. *43*, 51–59.

Tibshirani, R., and Hastie, T. (2007). Outlier sums for differential gene expression analysis. Biostatistics *8*, 2–8.

Tiwari, A.K., and Roy, H.K. (2012). Progress against cancer (1971-2011): how far have we come? J. Intern. Med. *271*, 392–399.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562–578.

Travert, M., Huang, Y., de Leval, L., Martin-Garcia, N., Delfau-Larue, M.-H., Berger, F., Bosq, J., Brière, J., Soulier, J., Macintyre, E., et al. (2012). Molecular features of hepatosplenic T-cell lymphoma unravels potential novel therapeutic targets. Blood *119*, 5795–5806.

Tusnády, G.E., and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J. Mol. Biol. *283*, 489–506.

Ulasov, I.V., Nandi, S., Dey, M., Sonabend, A.M., and Lesniak, M.S. (2011). Inhibition of Sonic Hedgehog and Notch Pathways Enhances Sensitivity of CD133+ Glioma Stem Cells to Temozolomide Therapy. Mol Med *17*, 103–112.

Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.

Viklund, H., and Elofsson, A. (2004). Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci *13*, 1908–1917.

Viklund, H., and Elofsson, A. (2008). OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics *24*, 1662–1668.

Viklund, H., Bernsel, A., Skwark, M., and Elofsson, A. (2008). SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics *24*, 2928–2929.

Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotech *32*, 926–932.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010a). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, T., Li, J., Chen, F., Zhao, Y., He, X., Wan, D., and Gu, J. (2007). Choline transporters in human lung adenocarcinoma: expression and functional implications. Acta Biochim. Biophys. Sin. (Shanghai) *39*, 668–674.

Wang, Z., Li, Y., Kong, D., and Sarkar, F.H. (2010b). The role of Notch signaling pathway in epithelial-mesenchymal transition (EMT) during development and tumor aggressiveness. Curr Drug Targets *11*, 745–751.

Ward, R.J., Lee, L., Graham, K., Satkunendran, T., Yoshikawa, K., Ling, E., Harper, L., Austin, R., Nieuwenhuis, E., Clarke, I.D., et al. (2009). Multipotent CD15+ cancer stem cells in patched-1-deficient mouse medulloblastoma. Cancer Res. *69*, 4682–4690.

Weiss, N. (2010). Control of depolarization-evoked presynaptic neurotransmitter release by Cav2.1 calcium channel: old story, new insights. Channels (Austin) *4*, 431–433.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Williams, A.G., Thomas, S., Wyman, S.K., and Holloway, A.K. (2014). RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. Curr Protoc Hum Genet *83*, 11.13.1–11.13.20.

Wimley, W.C. (2003). The versatile beta-barrel membrane protein. Curr. Opin. Struct. Biol. *13*, 404–411.

Witmer-Pack, M.D., Swiggard, W.J., Mirza, A., Inaba, K., and Steinman, R.M. (1995). Tissue distribution of the DEC-205 protein that is detected by the monoclonal antibody NLDC-145. II. Expression in situ in lymphoid and nonlymphoid tissues. Cell. Immunol. *163*, 157–162.

Wood, S.L., Westbrook, J.A., and Brown, J.E. (2014). Omic-profiling in breast cancer metastasis to bone: Implications for mechanisms, biomarkers and treatment. Cancer Treatment Reviews *40*, 139–152.

Wu, B. (2007). Cancer outlier differential gene expression detection. Biostatistics *8*, 566–575.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature *499*, 79–82.

Yang, Z.F., Ho, D.W., Ng, M.N., Lau, C.K., Yu, W.C., Ngai, P., Chu, P.W.K., Lam, C.T., Poon, R.T.P., and Fan, S.T. (2008). Significance of CD90+ cancer stem cells in human liver cancer. Cancer Cell *13*, 153–166.

Yao, C.-K., Lin, Y.Q., Ly, C.V., Ohyama, T., Haueter, C.M., Moiseenkova-Bell, V.Y., Wensel, T.G., and Bellen, H.J. (2009). A synaptic vesicle-associated Ca2+ channel promotes endocytosis and couples exocytosis to endocytosis. Cell *138*, 947–960.

Ye, F., Zhang, Y., Liu, Y., Yamada, K., Tso, J.L., Menjivar, J.C., Tian, J.Y., Yong, W.H., Schaue, D., Mischel, P.S., et al. (2013). Protective properties of radio-chemoresistant glioblastoma stem cell clones are associated with metabolic adaptation to reduced glucose dependence. PLoS ONE *8*, e80397.

Zeniou, M., Fève, M., Mameri, S., Dong, J., Salomé, C., Chen, W., El-Habr, E.A., Bousson, F., Sy, M., Obszynski, J., et al. (2015). Chemical Library Screening and Structure-Function Relationship Studies Identify Bisacodyl as a Potent and Selective Cytotoxic Agent Towards Quiescent Human Glioblastoma Tumor Stem-Like Cells. PLoS ONE *10*, e0134793.

Zeromski, J., Szczepański, M., and Mozer-Lisewska, null (2005). [Prevalence of CD56 /NCAM molecule in nervous system immune system and endocrine glands--accidental coincidence?]. Endokrynol Pol *56*, 78–82.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhang, M., Song, T., Yang, L., Chen, R., Wu, L., Yang, Z., and Fang, J. (2008a). Nestin and CD133: valuable stem cell-specific markers for determining clinical outcome of glioma patients. J. Exp. Clin. Cancer Res. *27*, 85.

Zhang, X.-P., Zheng, G., Zou, L., Liu, H.-L., Hou, L.-H., Zhou, P., Yin, D.-D., Zheng, Q.-J., Liang, L., Zhang, S.-Z., et al. (2008b). Notch activation promotes cell proliferation and the formation of neural stem cell-like colonies in human glioma cells. Mol. Cell. Biochem. *307*, 101–108.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE *9*, e78644.

Zhong, W., Jiang, M.M., Weinmaster, G., Jan, L.Y., and Jan, Y.N. (1997). Differential expression of mammalian Numb, Numblike and Notch1 suggests distinct roles during mouse cortical neurogenesis. Development *124*, 1887–1897.

Zołtowska, A., Stepiński, J., Lewko, B., Serkies, K., Zamorska, B., Roszkiewicz, A., Izycka-Swieszewska, E., and Kruszewski, W.J. (2001). Neural cell adhesion molecule in breast, colon and lung carcinomas. Arch. Immunol. Ther. Exp. (Warsz.) *49*, 171–174.

Zorniak, M., Clark, P.A., Leeper, H.E., Tipping, M.D., Francis, D.M., Kozak, K.R., Salamat, M.S., and Kuo, J.S. (2012). Differential expression of 2',3'-cyclic-nucleotide 3'-phosphodiesterase and neural lineage markers correlate with glioblastoma xenograft infiltration and patient survival. Clin. Cancer Res. *18*, 3628–3636.

Zorniak, M., Clark, P.A., and Kuo, J.S. (2015). Myelin-forming cell-specific cadherin-19 is a marker for minimally infiltrative glioblastoma stem-like cells. J. Neurosurg. *122*, 69–77.

ANNEXES

Annexe 1 : Toolbox calcium

Annexe 2 : Echantillons, ré-annotations et différents niveaux de classification des jeux de données utilisés pour l'analyse calcium

Annexe 1 : Toolbox calcium

Gene ID	Symbol	Rôle	Localisation 1	Localisation 2	Localisation 3
487	ATP2A1	Ca2+ sortie	ER		
488	ATP2A2	Ca2+ sortie	ER		
489	ATP2A3	Ca2+ sortie	ER		
490	ATP2B1	Ca2+ sortie	Membrane		
491	ATP2B2	Ca2+ sortie	Membrane		
492	ATP2B3	Ca2+ sortie	Membrane		
493	ATP2B4	Ca2+ sortie	Membrane		
27032	ATP2C1	Ca2+ sortie	Golgi		
9914	ATP2C2	Ca2+ sortie	Golgi	Membrane	
51719	CAB39	Ca2+ protéine de liaison			
81617	CAB39L	Ca2+ protéine de liaison			
9478	CABP1	Ca2+ protéine de liaison			
51475	CABP2	Ca2+ protéine de liaison			
57010	CABP4	Ca2+ protéine de liaison			
56344	CABP5	Ca2+ protéine de liaison			
164633	CABP7	Ca2+ protéine de liaison			
85438	CABS1	Ca2+ protéine de liaison			
11094	CACFD1	Ca2+ entrée	Vésicule		
773	CACNA1A	Ca2+ entrée	Membrane	Noyau	
774	CACNA1B	Ca2+ entrée	Membrane		
775	CACNA1C	Ca2+ entrée	Membrane	Cytoskelette	
776	CACNA1D	Ca2+ entrée	Membrane		
777	CACNA1E	Ca2+ entrée	Membrane		
778	CACNA1F	Ca2+ entrée	Membrane		
8913	CACNA1G	Ca2+ entrée	Membrane		
8912	CACNA1H	Ca2+ entrée	Membrane		
8911	CACNA1I	Ca2+ entrée	Membrane		
779	CACNA1S	Ca2+ entrée	Membrane	ER	
781	CACNA2D1	Ca2+ entrée	Membrane	ER	
9254	CACNA2D2	Ca2+ entrée	Membrane		
55799	CACNA2D3	Ca2+ entrée	Membrane	Cytosol	Noyau
93589	CACNA2D4	Ca2+ entrée	Membrane	ER	Noyau
782	CACNB1	Ca2+ entrée	Membrane	ER	Cytosol
783	CACNB2	Ca2+ entrée	Membrane	Noyau	Cytosol
784	CACNB3	Ca2+ entrée	Membrane	Cytosol	Noyau
785	CACNB4	Ca2+ entrée	Membrane	Cytosol	Noyau
786	CACNG1	Ca2+ entrée	Membrane		
10369	CACNG2	Ca2+ entrée	Membrane		
10368	CACNG3	Ca2+ entrée	Membrane		
27092	CACNG4	Ca2+ entrée	Membrane		
27091	CACNG5	Ca2+ entrée	Membrane	Cytoskelette	

Gene ID	Symbol	Rôle	Localisation 1	Localisation 2	Localisation 3
59285	CACNG6	Ca2+ entrée	Membrane		
59284	CACNG7	Ca2+ entrée	Membrane		
59283	CACNG8	Ca2+ entrée	Membrane	Cytoskelette	
793	CALB1	Ca2+ protéine de liaison			
794	CALB2	Ca2+ protéine de liaison			
57658	CALCOCO1	Ca2+ protéine de liaison			
10241	CALCOCO2	Ca2+ protéine de liaison			
801	CALM1	Ca2+ protéine de liaison			
805	CALM2	Ca2+ protéine de liaison			
808	CALM3	Ca2+ protéine de liaison			
810	CALML3	Ca2+ protéine de liaison			
91860	CALML4	Ca2+ protéine de liaison			
51806	CALML5	Ca2+ protéine de liaison			
163688	CALML6	Ca2+ protéine de liaison			
83698	CALN1	Ca2+ protéine de liaison			
811	CALR	Ca2+ protéine de liaison			
125972	CALR3	Ca2+ protéine de liaison			
813	CALU	Ca2+ protéine de liaison			
828	CAPS	Ca2+ protéine de liaison			
84698	CAPS2	Ca2+ protéine de liaison			
133690	CAPSL	Ca2+ protéine de liaison			
79800	CARF	Ca2+ protéine de liaison			
23589	CARHSP1	Ca2+ protéine de liaison			
844	CASQ1	Ca2+ protéine de liaison			
845	CASQ2	Ca2+ protéine de liaison			
846	CASR	Ca2+ protéine de liaison			
117144	CATSPER1	Ca2+ entrée	Membrane	ER	
117155	CATSPER2	Ca2+ entrée	Membrane	ER	
440278	CATSPER2P1	Ca2+ entrée	Membrane	ER	
347732	CATSPER3	Ca2+ entrée	Membrane	ER	
378807	CATSPER4	Ca2+ entrée	Membrane	ER	
79820	CATSPERB	Ca2+ entrée	Membrane		
257062	CATSPERD	Ca2+ entrée	Membrane	ER	
57828	CATSPERG	Ca2+ entrée	Membrane		
1068	CETN1	Ca2+ protéine de liaison			
1069	CETN2	Ca2+ protéine de liaison			
1070	CETN3	Ca2+ protéine de liaison			
729338	CETN4P	Ca2+ protéine de liaison			
11261	CHP1	Ca2+ protéine de liaison			
63928	CHP2	Ca2+ protéine de liaison			
1047	CLGN	Ca2+ protéine de liaison			
79645	EFCAB1	Ca2+ protéine de liaison			
100130771	EFCAB10	Ca2+ protéine de liaison			
90141	EFCAB11	Ca2+ protéine de liaison			
90288	EFCAB12	Ca2+ protéine de liaison			

Gene ID	Symbol	Rôle	Localisation 1	Localisation 2	Localisation 3
124989	EFCAB13	Ca2+ protéine de liaison			
9813	EFCAB14	Ca2+ protéine de liaison			
84288	EFCAB2	Ca2+ protéine de liaison			
146779	EFCAB3	Ca2+ protéine de liaison			
283229	EFCAB4A	Ca2+ protéine de liaison			
84766	EFCAB4B	Ca2+ protéine de liaison			
374786	EFCAB5	Ca2+ protéine de liaison			
64800	EFCAB6	Ca2+ protéine de liaison			
84455	EFCAB7	Ca2+ protéine de liaison			
388795	EFCAB8	Ca2+ protéine de liaison			
285588	EFCAB9	Ca2+ protéine de liaison			
79825	EFCC1	Ca2+ protéine de liaison			
151651	EFHB	Ca2+ protéine de liaison			
114327	EFHC1	Ca2+ protéine de liaison			
80258	EFHC2	Ca2+ protéine de liaison			
80303	EFHD1	Ca2+ protéine de liaison			
79180	EFHD2	Ca2+ protéine de liaison			
2312	FLG	Ca2+ protéine de liaison			
388698	FLG2	Ca2+ protéine de liaison			
25801	GCA	Ca2+ protéine de liaison			
2978	GUCA1A	Ca2+ protéine de liaison			
2979	GUCA1B	Ca2+ protéine de liaison			
3208	HPCA	Ca2+ protéine de liaison			
3241	HPCAL1	Ca2+ protéine de liaison			
51440	HPCAL4	Ca2+ protéine de liaison			
3270	HRC	Ca2+ protéine de liaison			
388697	HRNR	Ca2+ protéine de liaison			
3708	ITPR1	Ca2+ entrée	ER	membrane	
3709	ITPR2	Ca2+ entrée	ER	membrane	
3710	ITPR3	Ca2+ entrée	ER	membrane	
3713	IVL	Ca2+ protéine de liaison			
30818	KCNIP3	TF			
3954	LETM1	Ca2+ sortie	Mitochondrie		
137994	LETM2	TF			
100288712	LOC100288712	Ca2+ protéine de liaison			
100291628	LOC100291628	Ca2+ protéine de liaison			
123430	LOC123430	TF			
137698	LOC137698	TF			
197023	LOC197023	TF			
342076	LOC342076	TF			
391722	LOC391722	Ca2+ protéine de liaison			
728549	LOC728549	Ca2+ protéine de liaison			
729603	LOC729603	Ca2+ protéine de liaison			
79772	MCTP1	Ca2+ protéine de liaison			
55784	MCTP2	Ca2+ protéine de liaison			

Gene ID	Symbol	Rôle	Localisation 1	Localisation 2	Localisation 3
90550	MCU	Ca2+ sortie	Mitochondrie		
10367	MICU1	Ca2+ sortie	Mitochondrie		
221154	MICU2	Ca2+ sortie	Mitochondrie		
286097	MICU3	Ca2+ sortie	Mitochondrie		
93408	MYL10	Ca2+ protéine de liaison			
10627	MYL12A	Ca2+ protéine de liaison			
103910	MYL12B	Ca2+ protéine de liaison			
4633	MYL2	Ca2+ protéine de liaison			
4634	MYL3	Ca2+ protéine de liaison			
4635	MYL4	Ca2+ protéine de liaison			
4636	MYL5	Ca2+ protéine de liaison			
4637	MYL6	Ca2+ protéine de liaison			
140465	MYL6B	Ca2+ protéine de liaison			
58498	MYL7	Ca2+ protéine de liaison			
10398	MYL9	Ca2+ protéine de liaison			
29895	MYLPF	Ca2+ protéine de liaison			
83988	NCALD	Ca2+ protéine de liaison			
23413	NCS1	Ca2+ protéine de liaison			
64168	NECAB1	Ca2+ protéine de liaison			
54550	NECAB2	Ca2+ protéine de liaison			
63941	NECAB3	Ca2+ protéine de liaison			
4924	NUCB1	TF			
4925	NUCB2	TF			
654231	OCM	Ca2+ protéine de liaison			
4951	OCM2	Ca2+ protéine de liaison			
84876	ORAI1	Ca2+ entrée	Membrane	ER	
80228	ORAI2	Ca2+ entrée	Membrane	ER	
93129	ORAI3	Ca2+ entrée	Membrane	ER	
5310	PKD1	Ca2+ entrée	Membrane	Noyau	Cytosol
114780	PKD1L2	Ca2+ entrée	Membrane		
5311	PKD2	Ca2+ entrée	Membrane	ER	Cytoskelette
9033	PKD2L1	Ca2+ entrée	Membrane	ER	
27039	PKD2L2	Ca2+ entrée	Membrane	Cytosol	
10343	PKDREJ	Ca2+ entrée	Membrane		
5535	PPP3R2	Ca2+ protéine de liaison			
5663	PSEN1	Ca2+ protéine de liaison			
5664	PSEN2	Ca2+ protéine de liaison			
5816	PVALB	Ca2+ protéine de liaison			
9727	RAB11FIP3	Ca2+ protéine de liaison			
401258	RAB44	Ca2+ protéine de liaison			
158158	RASEF	Ca2+ protéine de liaison			
10125	RASGRP1	Ca2+ protéine de liaison			
10235	RASGRP2	Ca2+ protéine de liaison			
25780	RASGRP3	Ca2+ protéine de liaison			
5954	RCN1	Ca2+ protéine de liaison			

Gene ID	Symbol	Rôle	Localisation 1	Localisation 2	Localisation 3
442234	RCN1P1	Ca2+ protéine de liaison			
728913	RCN1P2	Ca2+ protéine de liaison			
5955	RCN2	Ca2+ protéine de liaison			
57333	RCN3	Ca2+ protéine de liaison			
5957	RCVRN	Ca2+ protéine de liaison			
126638	RPTN	Ca2+ protéine de liaison			
6261	RYR1	Ca2+ entrée	ER	Membrane	Cytosol
6262	RYR2	Ca2+ entrée	ER	Membrane	
6263	RYR3	Ca2+ entrée	ER		
6271	S100A1	Ca2+ protéine de liaison			
6281	S100A10	Ca2+ protéine de liaison			
6282	S100A11	Ca2+ protéine de liaison			
729659	S100A11P1	Ca2+ protéine de liaison			
347701	S100A11P2	Ca2+ protéine de liaison			
645474	S100A11P3	Ca2+ protéine de liaison			
100506938	S100A11P4	Ca2+ protéine de liaison			
6283	S100A12	Ca2+ protéine de liaison			
6284	S100A13	Ca2+ protéine de liaison			
57402	S100A14	Ca2+ protéine de liaison			
140576	S100A16	Ca2+ protéine de liaison			
6273	S100A2	Ca2+ protéine de liaison			
6274	S100A3	Ca2+ protéine de liaison			
6275	S100A4	Ca2+ protéine de liaison			
6276	S100A5	Ca2+ protéine de liaison			
6277	S100A6	Ca2+ protéine de liaison			
6278	S100A7	Ca2+ protéine de liaison			
338324	S100A7A	Ca2+ protéine de liaison			
645922	S100A7L2	Ca2+ protéine de liaison			
127481	S100A7P1	Ca2+ protéine de liaison			
375027	S100A7P2	Ca2+ protéine de liaison			
6279	S100A8	Ca2+ protéine de liaison			
6280	S100A9	Ca2+ protéine de liaison			
6285	S100B	Ca2+ protéine de liaison			
795	S100G	Ca2+ protéine de liaison			
6286	S100P	Ca2+ protéine de liaison			
170591	S100Z	Ca2+ protéine de liaison			
10590	SCGN	Ca2+ protéine de liaison			
9187	SLC24A1	Ca2+ sortie	Membrane	Cytoskelette	
25769	SLC24A2	Ca2+ sortie	Membrane		
57419	SLC24A3	Ca2+ sortie	Membrane		
123041	SLC24A4	Ca2+ sortie	Membrane		
283652	SLC24A5	Ca2+ sortie	Membrane		
8604	SLC25A12	Ca2+ sortie	Mitochondrie	cytosol	
10165	SLC25A13	Ca2+ sortie	Mitochondrie	Membrane	Cytosol
79085	SLC25A23	Ca2+ sortie	Mitochondrie	cytosol	

Gene ID	Symbol	Rôle	Localisation 1	Localisation 2	Localisation 3
6546	SLC8A1	Ca2+ sortie	Membrane		
6543	SLC8A2	Ca2+ sortie	Membrane		
6547	SLC8A3	Ca2+ sortie	Membrane		
80024	SLC8B1	Ca2+ entrée	Mitochondrie		
132203	SNTN	Ca2+ protéine de liaison			
6717	SRI	Ca2+ protéine de liaison			
6781	STC1	Ca2+ protéine de liaison			
6786	STIM1	Ca2+ entrée	ER	Membrane	Cytoskelette
57620	STIM2	Ca2+ entrée	Membrane	ER	
7062	TCHH	Ca2+ protéine de liaison			
126637	TCHHL1	Ca2+ protéine de liaison			
54997	TESC	Ca2+ protéine de liaison			
7134	TNNC1	Ca2+ protéine de liaison			
7125	TNNC2	Ca2+ protéine de liaison			
53373	TPCN1	Ca2+ entrée	ER	membrane	
219931	TPCN2	Ca2+ entrée	ER	membrane	
8989	TRPA1	Ca2+ entrée	Membrane		
7220	TRPC1	Ca2+ entrée	Membrane	ER	
7222	TRPC3	Ca2+ entrée	Membrane	ER	
7223	TRPC4	Ca2+ entrée	Membrane	Cytoskelette	ER
7224	TRPC5	Ca2+ entrée	Membrane	cytosol	
7225	TRPC6	Ca2+ entrée	Membrane	ER	
57113	TRPC7	Ca2+ entrée	Membrane	Noyau	
4308	TRPM1	Ca2+ entrée	Membrane		
7226	TRPM2	Ca2+ entrée	Membrane	Noyau	cytosol
54795	TRPM4	Ca2+ entrée	Membrane	ion_monovalent	
29850	TRPM5	Ca2+ entrée	Membrane	ion_monovalent	
140803	TRPM6	Ca2+ entrée	Membrane		
54822	TRPM7	Ca2+ entrée	Membrane	Noyau	
79054	TRPM8	Ca2+ entrée	Membrane	ER	
7442	TRPV1	Ca2+ entrée	Membrane	cytosol	
51393	TRPV2	Ca2+ entrée	Membrane		
162514	TRPV3	Ca2+ entrée	Membrane		
59341	TRPV4	Ca2+ entrée	Membrane	Cytoskelette	
56302	TRPV5	Ca2+ entrée	Membrane	Peroxysome	
55503	TRPV6	Ca2+ entrée	Membrane		
7429	VIL1	Ca2+ protéine de liaison			
7447	VSNL1	Ca2+ protéine de liaison			
23140	ZZEF1	TF			

Annexe 2 : Echantillons utilisés pour l'analyse du signal calcium et classifications

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM537470_S803681.H52	E_GEOD_21514	HA		NA	2	2	NA	6	cell_line
GSM537471_S905515.K438	E_GEOD_21514	HA		NA	2	2	NA	6	cell_line
GSM587218	E-GEOD-23806	GBMs		NA	1	1	1	NA	NA
GSM587219	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587220	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587221	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587222	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587223	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587224	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587225	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587226	E-GEOD-23806	GBMs		NA	1	1	3	NA	NA
GSM587227	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587228	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587229	E-GEOD-23806	GBMs		NA	1	1	2	NA	GBM2
GSM587192	E-GEOD-23806	gCSCs		NA	2	2	NA	5	NA
GSM587193	E-GEOD-23806	gCSCs		NA	2	2	NA	5	NA
GSM587195	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587196	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587197	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587198	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587199	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587200	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587201	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587202	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587203	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587204	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587205	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587206	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587207	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587208	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587209	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587210	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587211	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587212	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587213	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587214	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587215	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587216	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587217	E-GEOD-23806	gCSCs		NA	2	2	NA	4	gCSC2
GSM587194	E-GEOD-23806	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM587191	E-GEOD-23806	gCSCs	outlier	NA	NA	NA	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM587155	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587156	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587157	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	4	NA
GSM587158	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587159	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587160	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587161	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587162	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587163	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587164	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587165	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587166	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587167	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587168	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587169	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587170	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587171	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587172	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587173	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587174	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587175	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587176	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587177	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM587178	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587179	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587180	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587181	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587182	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587183	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587184	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587185	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587186	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587187	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587188	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587189	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM587190	E-GEOD-23806	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line
GSM172063	E-GEOD-7181	gCSCs		NA	2	2	NA	2	gCSC1
GSM172064	E-GEOD-7181	gCSCs		NA	2	2	NA	2	gCSC1
GSM172065	E-GEOD-7181	gCSCs		NA	2	2	NA	2	gCSC1
GSM172066	E-GEOD-7181	gCSCs		NA	2	2	NA	2	gCSC1
GSM172067	E-GEOD-7181	gCSCs		NA	2	2	NA	2	gCSC1
GSM170898	E-GEOD-7181	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM981283_INPUT_p_CE_AM_101309_HG-U133_Plus_2_A01_514164	GSE17312	ESCs		NA	2	2	NA	7	ESC
GSM981284_INPUT_p_CE_AM_101309_HG-U133_Plus_2_A05_514058	GSE17312	ESCs		NA	2	2	NA	7	ESC
GSM981285_INPUT_p_CE_AM_101309_HG-U133_Plus_2_A06_514108	GSE17312	ESCs		NA	2	2	NA	7	ESC
GSM981286_INPUT_p_CE_AM_101309_HG-U133_Plus_2_A07_514004	GSE17312	ESCs		NA	2	2	NA	7	ESC
GSM450644	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450641	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450629	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450632	GSE18015	gCSCs		NA	NA	NA	NA	NA	NA
GSM450633	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450637	GSE18015	gCSCs		NA	NA	NA	NA	NA	NA
GSM450638	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM450639	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450642	GSE18015	gCSCs		NA	NA	NA	NA	NA	NA
GSM450630	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450631	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450634	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450635	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450640	GSE18015	gCSCs		NA	NA	NA	NA	NA	NA
GSM450643	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM450636	GSE18015	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM503594_HES3_R1	GSE20126	ESCs		NA	2	2	NA	7	ESC
GSM503595_HES3_R2	GSE20126	ESCs		NA	2	2	NA	7	ESC
GSM503596_BG01V_R1	GSE20126	ESCs		NA	2	2	NA	7	ESC
GSM503597_BG01V_R2	GSE20126	ESCs		NA	2	2	NA	7	ESC
GSM835796_SAMPLE_2	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835797_SAMPLE_7	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835798_SAMPLE_9	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835799_SAMPLE_4	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835800_SAMPLE_10	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835801_SAMPLE_6	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835802_SAMPLE_1	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835803_SAMPLE_3	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835804_SAMPLE_11	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM835805_SAMPLE_8	GSE34200	ESCs		NA	2	2	NA	7	ESC
GSM978836_Untreated-1	GSE39762	ESCs		NA	2	2	NA	7	ESC
GSM978839_Untreated-2	GSE39762	ESCs		NA	2	2	NA	7	ESC
GSM978842_Untreated-3	GSE39762	ESCs		NA	2	2	NA	7	ESC
GSM97794	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97796	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97797	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97798	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97801	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97806	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97808	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97813	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97814	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97818	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97826	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97829	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97836	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97839	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97847	GSE4290	GBMs		NA	1	1	5	NA	NA
GSM97851	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97852	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97856	GSE4290	GBMs		NA	1	1	1	NA	NA
GSM97859	GSE4290	GBMs		NA	1	1	4	NA	GBM1

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM97861	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97863	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97869	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97871	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97879	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97882	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97885	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97886	GSE4290	GBMs		NA	1	1	6	NA	NA
GSM97887	GSE4290	GBMs		NA	1	1	7	NA	NA
GSM97888	GSE4290	GBMs		NA	1	1	5	NA	NA
GSM97889	GSE4290	GBMs		NA	1	1	8	NA	NA
GSM97891	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97893	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97894	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97895	GSE4290	GBMs		NA	1	1	9	NA	NA
GSM97896	GSE4290	GBMs		NA	1	1	8	NA	NA
GSM97898	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97903	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97905	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97906	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97908	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97912	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97914	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97915	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97917	GSE4290	GBMs		NA	1	1	10	NA	NA
GSM97918	GSE4290	GBMs		NA	1	1	1	NA	NA
GSM97919	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97922	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97930	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97935	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97936	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97938	GSE4290	GBMs		NA	1	1	5	NA	NA
GSM97940	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97945	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97950	GSE4290	GBMs		NA	1	1	8	NA	NA
GSM97952	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97953	GSE4290	GBMs		NA	1	1	5	NA	NA
GSM97954	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97955	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97959	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97961	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97963	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97965	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97966	GSE4290	GBMs		NA	1	1	2	NA	GBM2
GSM97969	GSE4290	GBMs		NA	1	1	6	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM97971	GSE4290	GBMs		NA	1	1	4	NA	GBM1
GSM97924	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97931	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97858	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97942	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97946	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97892	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97819	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97844	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97832	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97967	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97948	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97968	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97870	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97877	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97821	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97926	GSE4290	GBMs	outlier	NA	NA	NA	NA	NA	NA
GSM97793	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97795	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97799	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97802	GSE4290	Gliomes		NA	1	1	12	NA	NA
GSM97810	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97815	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97823	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97824	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97830	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97831	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97835	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97838	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97841	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97842	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97843	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97857	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97860	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97862	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97866	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97867	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97868	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97872	GSE4290	Gliomes		NA	1	1	4	NA	NA
GSM97873	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97874	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97875	GSE4290	Gliomes		NA	1	1	13	NA	NA
GSM97876	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97880	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97881	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM97883	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97884	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97890	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97901	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97902	GSE4290	Gliomes		NA	1	1	4	NA	NA
GSM97904	GSE4290	Gliomes		NA	1	1	14	NA	NA
GSM97909	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97910	GSE4290	Gliomes		NA	1	1	4	NA	NA
GSM97911	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97920	GSE4290	Gliomes		NA	1	1	4	NA	NA
GSM97923	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97925	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97929	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97933	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97934	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97939	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97941	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97943	GSE4290	Gliomes		NA	1	1	4	NA	NA
GSM97944	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97947	GSE4290	Gliomes		NA	1	1	13	NA	NA
GSM97949	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97951	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97956	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97957	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97958	GSE4290	Gliomes		NA	1	1	5	NA	NA
GSM97960	GSE4290	Gliomes		NA	1	3	NA	NA	NA
GSM97962	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97964	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97970	GSE4290	Gliomes		NA	1	1	11	NA	brain_cancer
GSM97972	GSE4290	Gliomes		NA	1	4	NA	NA	NA
GSM97897	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97927	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97928	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97845	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97865	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97932	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97854	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97937	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97899	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97900	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97878	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97916	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97837	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97864	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97822	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM97913	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97921	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM97907	GSE4290	Gliomes	outlier	NA	NA	NA	NA	NA	NA
GSM1092481_G48_NS_12209	GSE44841	gCSCs		NA	2	2	NA	1	gCSC3
GSM1092482_G48_DF_4d_12209	GSE44841	gCSCs		NA	2	2	NA	1	gCSC3
GSM1092483_G52_NS_081008	GSE44841	gCSCs		NA	2	2	NA	2	gCSC1
GSM1092485_G59_NS_12209	GSE44841	gCSCs		NA	2	2	NA	2	gCSC1
GSM1092486_G59_DF_4d_12209	GSE44841	gCSCs		NA	2	2	NA	2	gCSC1
GSM1092487_G63_NS_12209	GSE44841	gCSCs		NA	2	2	NA	2	gCSC1
GSM1092488_G63_DF_4d_12209	GSE44841	gCSCs		NA	2	2	NA	2	gCSC1
GSM1092484_G52_DF_4d_081008	GSE44841	gCSCs	outlier	NA	NA	NA	NA	NA	NA
GSM1119322_MGG23_rep1	GSE46016	gCSCs		NA	2	2	NA	2	gCSC1
GSM1119323_MGG23_rep2	GSE46016	gCSCs		NA	2	2	NA	2	gCSC1
GSM1119324_MGG4_rep1	GSE46016	gCSCs		NA	2	2	NA	3	NA
GSM1119325_MGG4_rep2	GSE46016	gCSCs		NA	2	2	NA	3	NA
GSM1119326_MGG6_rep1	GSE46016	gCSCs		NA	2	2	NA	4	gCSC2
GSM1119327_MGG6_rep2	GSE46016	gCSCs		NA	2	2	NA	4	gCSC2
GSM1119328_MGG8_rep1	GSE46016	gCSCs		NA	2	2	NA	4	gCSC2
GSM1119329_MGG8_rep2	GSE46016	gCSCs		NA	2	2	NA	4	gCSC2
GSM1119330_MGG8_rep3	GSE46016	gCSCs		NA	2	2	NA	3	NA
GSM1119331_MGG8_rep4	GSE46016	gCSCs		NA	2	2	NA	3	NA
GSM1131806_D431-1	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131807_D431-2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131808_D431RC2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131809_D431RZC2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131810_E445-1	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131811_E445-2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131812_E445RC2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131813_E445RZC2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131814_E496-1	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131815_E496-2	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131816_S496RC1	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1131817_S496RZC1	GSE46531	gCSCs		NA	2	2	NA	1	gCSC3
GSM1253303_12.1_GSC	GSE51822	gCSCs		NA	2	2	NA	4	gCSC2
GSM1253304_22_GSC	GSE51822	gCSCs		NA	2	2	NA	4	gCSC2
GSM175825	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM175826	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM175827	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM175828	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM175829	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA
GSM175830	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA
GSM175831	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM175832	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA
GSM175842	GSE7307	Tissu normal neural		amygdale	3	4	NA	NA	NA
GSM175843	GSE7307	Tissu normal neural		amygdale	3	3	NA	NA	NA
GSM175844	GSE7307	Tissu normal neural		amygdale	3	3	NA	NA	NA
GSM175845	GSE7307	Tissu normal neural		amygdale	3	3	NA	NA	NA
GSM175846	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM175847	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM175848	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM175849	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM175850	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM175851	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM175852	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM175853	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM175854	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM175855	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM175856	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM175857	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM175858	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM175859	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM175860	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM175861	GSE7307	Tissu normal neural		hyppocampe	3	3	NA	NA	NA
GSM175862	GSE7307	Tissu normal neural		lobe pariétal	3	3	NA	NA	NA
GSM175863	GSE7307	Tissu normal neural		lobe pariétal	3	3	NA	NA	NA
GSM175864	GSE7307	Tissu normal neural		lobe pariétal	3	3	NA	NA	NA
GSM175865	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM175866	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM175867	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM175868	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM175869	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA
GSM175870	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM175871	GSE7307	Tissu normal neural		substance noire	3	4	NA	NA	NA
GSM175872	GSE7307	Tissu normal neural		substance noire	3	4	NA	NA	NA
GSM175873	GSE7307	Tissu normal neural		substance noire	3	4	NA	NA	NA
GSM175874	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM175875	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM175876	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM175877	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM175885	GSE7307	Tissu normal neural		thalamus	3	3	NA	NA	NA
GSM175886	GSE7307	Tissu normal neural		thalamus	3	3	NA	NA	NA
GSM175887	GSE7307	Tissu normal neural		thalamus	3	3	NA	NA	NA
GSM175888	GSE7307	Tissu normal neural		thalamus	3	3	NA	NA	NA
GSM175889	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM175890	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM175891	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM175892	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM175893	GSE7307	Tissu normal neural		noyau vestibulaire supérieur	3	4	NA	NA	NA
GSM175894	GSE7307	Tissu normal neural		noyau vestibulaire supérieur	3	4	NA	NA	NA
GSM175895	GSE7307	Tissu normal neural		noyau vestibulaire supérieur	3	4	NA	NA	NA
GSM175901	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM175902	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM175903	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM175904	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM175907	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM175909	GSE7307	Tissu normal neural		outlier	3	5	NA	NA	NA
GSM175936	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM175956	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM175957	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM175958	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM175959	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM175987	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM175988	GSE7307	Tissu normal neural		hyppocampe	3	3	NA	NA	NA
GSM175989	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM175990	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176017	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM176018	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176020	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176024	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176030	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM176031	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM176033	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176034	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176036	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176037	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176045	GSE7307	Tissu normal neural		amygdale	3	3	NA	NA	NA
GSM176046	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176047	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176048	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM176049	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176050	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM176051	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM176052	GSE7307	Tissu normal neural		hyppocampe	3	3	NA	NA	NA
GSM176053	GSE7307	Tissu normal neural		lobe pariétal	3	4	NA	NA	NA
GSM176054	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176055	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA
GSM176056	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176057	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176058	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA
GSM176059	GSE7307	Tissu normal neural		substance noire	3	4	NA	NA	NA
GSM176060	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM176061	GSE7307	Tissu normal neural		thalamus	3	3	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM176062	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM176063	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM176064	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA
GSM176065	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA
GSM176066	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176067	GSE7307	Tissu normal neural		hypothalamus	3	4	NA	NA	NA
GSM176068	GSE7307	Tissu normal neural		mésencéphale	3	4	NA	NA	NA
GSM176069	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM176070	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176071	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	3	NA	NA	NA
GSM176072	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA
GSM176073	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176074	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176075	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM176076	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM176077	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176116	GSE7307	Tissu normal neural		hypothalamus	3	4	NA	NA	NA
GSM176120	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM176124	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA
GSM176125	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176147	GSE7307	Tissu normal neural		amygdale	3	3	NA	NA	NA
GSM176148	GSE7307	Tissu normal neural		amygdale	3	3	NA	NA	NA
GSM176149	GSE7307	Tissu normal neural		amygdale	3	4	NA	NA	NA
GSM176150	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176151	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176152	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176153	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176154	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176156	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176157	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM176158	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM176159	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM176160	GSE7307	Tissu normal neural		cervelet	3	3	NA	NA	NA
GSM176161	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176162	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176163	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176164	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176165	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM176166	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM176167	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM176168	GSE7307	Tissu normal neural		corps calleux	3	4	NA	NA	NA
GSM176169	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176170	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM176171	GSE7307	Tissu normal neural		lobe occipital	3	3	NA	NA	NA
GSM176172	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM176173	GSE7307	Tissu normal neural		lobe frontal	3	3	NA	NA	NA
GSM176174	GSE7307	Tissu normal neural		hyppocampe	3	3	NA	NA	NA
GSM176175	GSE7307	Tissu normal neural		hyppocampe	3	3	NA	NA	NA
GSM176176	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176177	GSE7307	Tissu normal neural		hyppocampe	3	3	NA	NA	NA
GSM176178	GSE7307	Tissu normal neural		lobe pariétal	3	3	NA	NA	NA
GSM176179	GSE7307	Tissu normal neural		lobe pariétal	3	3	NA	NA	NA
GSM176180	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176181	GSE7307	Tissu normal neural		lobe pariétal	3	3	NA	NA	NA
GSM176182	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176183	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176184	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176185	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176186	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA
GSM176205	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM176206	GSE7307	Tissu normal neural		lobe occipital	3	3	NA	NA	NA
GSM176207	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176208	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176209	GSE7307	Tissu normal neural		moelle épinière	3	4	NA	NA	NA
GSM176210	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA
GSM176211	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176212	GSE7307	Tissu normal neural		substance noire	3	4	NA	NA	NA
GSM176213	GSE7307	Tissu normal neural		substance noire	3	4	NA	NA	NA
GSM176214	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM176215	GSE7307	Tissu normal neural		lobe temporal	3	3	NA	NA	NA
GSM176216	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176217	GSE7307	Tissu normal neural		thalamus	3	3	NA	NA	NA
GSM176218	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM176219	GSE7307	Tissu normal neural		ganglions rachidiens	3	1	15	NA	NA
GSM176220	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM176221	GSE7307	Tissu normal neural		ganglions trigéminés	3	1	15	NA	NA
GSM176222	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA
GSM176223	GSE7307	Tissu normal neural		aire tegmentale ventrale	3	4	NA	NA	NA
GSM176224	GSE7307	Tissu normal neural		noyau vestibulaire supérieur	3	4	NA	NA	NA
GSM176225	GSE7307	Tissu normal neural		noyau vestibulaire supérieur	3	4	NA	NA	NA
GSM176226	GSE7307	Tissu normal neural		hypothalamus	3	4	NA	NA	NA
GSM176233	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176293	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176344	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176345	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176346	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176347	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176348	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA
GSM176349	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA
GSM176350	GSE7307	Tissu normal neural		ganglion noueux	3	4	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM176352	GSE7307	Tissu normal neural		lobe occipital	3	3	NA	NA	NA
GSM176353	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176354	GSE7307	Tissu normal neural		lobe occipital	3	3	NA	NA	NA
GSM176355	GSE7307	Tissu normal neural		lobe occipital	3	3	NA	NA	NA
GSM176357	GSE7307	Tissu normal neural		outlier	3	4	NA	NA	NA
GSM176363	GSE7307	Tissu normal neural		noyau caudé	3	3	NA	NA	NA
GSM176364	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176365	GSE7307	Tissu normal neural		noyau caudé	3	3	NA	NA	NA
GSM176366	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176367	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176369	GSE7307	Tissu normal neural		noyau caudé	3	3	NA	NA	NA
GSM176370	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176371	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176372	GSE7307	Tissu normal neural		noyau caudé	3	3	NA	NA	NA
GSM176373	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176374	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176379	GSE7307	Tissu normal neural		accumbens	3	3	NA	NA	NA
GSM176380	GSE7307	Tissu normal neural		putamen	3	3	NA	NA	NA
GSM176381	GSE7307	Tissu normal neural		hypothalamus	3	3	NA	NA	NA
GSM176382	GSE7307	Tissu normal neural		hypothalamus	3	4	NA	NA	NA
GSM176383	GSE7307	Tissu normal neural		hypothalamus	3	4	NA	NA	NA
GSM176384	GSE7307	Tissu normal neural		hypothalamus	3	4	NA	NA	NA
GSM176393	GSE7307	Tissu normal neural		substance noire pars compacta	3	4	NA	NA	NA
GSM176394	GSE7307	Tissu normal neural		substance noire pars compacta	3	4	NA	NA	NA
GSM176395	GSE7307	Tissu normal neural		substantia nigra_reticulata	3	4	NA	NA	NA
GSM176397	GSE7307	Tissu normal neural		substance noire pars compacta	3	4	NA	NA	NA
GSM176398	GSE7307	Tissu normal neural		substantia nigra_reticulata	3	4	NA	NA	NA
GSM176401	GSE7307	Tissu normal neural		substance noire pars compacta	3	4	NA	NA	NA
GSM176402	GSE7307	Tissu normal neural		substantia nigra_reticulata	3	4	NA	NA	NA
GSM176403	GSE7307	Tissu normal neural		substance noire pars compacta	3	4	NA	NA	NA

Echantillons	Jeu de données	Annotation	Outliers	Annotations tissus neuraux	Classification tous gènes	Classification toolbox calcium	Classification "gliomes"	Classification "lignees"	Classes
GSM176404	GSE7307	Tissu normal neural		substantia nigra_reticulata	3	4	NA	NA	NA
GSM176411	GSE7307	Tissu normal neural		glande pituitaire	4	6	NA	NA	NA
GSM176412	GSE7307	Tissu normal neural		glande pituitaire	4	6	NA	NA	NA
GSM176413	GSE7307	Tissu normal neural		glande pituitaire	4	6	NA	NA	NA
GSM176436	GSE7307	Tissu normal neural		globus pallidus interne	3	3	NA	NA	NA
GSM176445	GSE7307	Tissu normal neural		globus pallidus interne	3	3	NA	NA	NA
GSM176446	GSE7307	Tissu normal neural		globus pallidus interne	3	4	NA	NA	NA
GSM176447	GSE7307	Tissu normal neural		globus pallidus externe	3	3	NA	NA	NA
GSM176448	GSE7307	Tissu normal neural		globus pallidus externe	3	3	NA	NA	NA
GSM176451	GSE7307	Tissu normal neural		noyau sous-thalamique	3	4	NA	NA	NA
GSM176452	GSE7307	Tissu normal neural		outlier	3	3	NA	NA	NA
GSM176453	GSE7307	Tissu normal neural		noyau sous-thalamique	3	3	NA	NA	NA
JHH_1_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_2_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_3_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_4_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_5BIS_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_6BIS_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_7_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_8_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_OB1_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_TG01A_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_TG10_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_TG16_U133_2	HC	gCSCs		NA	2	2	NA	1	gCSC3
JHH_HA_U133_2	HC	HA		NA	2	2	NA	6	cell_line
JHH_U87-MG_U133_2	HC	Lignées cellulaires de gliomes		NA	2	2	NA	6	cell_line

Publications

Glioblastoma and Calcium signaling. Analysis of the calcium toolbox expression.
Robil N., Petel F., Kilhoffer M.-C and Haiech J.
International Journal of Developmental Biology *in press*

KANT: A gene expression-based tool for detecting putative membrane cancer-specific antigens
Robil N., Grellier B., Petel F., Rooke R. and Haiech J.
BMC Bioinformatics, *submitted*

Publication en préparation :

Proteomic cell surface immunophenotyping of Glioblastoma Cancer Stem-like cells reveals CD205 and CD109 as differential biomarker
Muller L.*, Robil N.*, Dong J.*, Lennon S., Saliou J.-M, Audran E., Zeniou M., Carapito C., Van Dorsselaer A., Petel F., Rooke R., Accard N., Junier M.-P, Kilhoffer M.-C, Chneiweiss H., Haiech J. and Cianférani S.

Noémie ROBIL

Recherche d'antigènes spécifiques de tumeurs et analyse des cellules souches de glioblastomes

Résumé

Les glioblastomes sont les tumeurs du système nerveux central les plus fréquentes et agressives. Avec une survie médiane inférieure à 2 ans, les thérapies actuelles restent inefficaces. Cet échec pourrait être expliqué en partie par l'existence de cellules particulières, les cellules souches cancéreuses. Ces cellules ont plusieurs propriétés communes aux cellules souches, qui les rendent résistantes aux traitements des glioblastomes. Il est donc important de pouvoir les identifier et les cibler pour pouvoir éliminer totalement la tumeur.

L'objectif de ce travail de thèse est de déterminer des biomarqueurs des cellules souches de glioblastomes (gCSCs). Pour cela, nous avons d'abord développé une méthode générique permettant de prédire des antigènes spécifiques de cancer à partir de données de puces d'expression. Puis, nous avons travaillé sur les gCSCs, en identifiant des biomarqueurs potentiels, puis en étudiant les modifications du signal calcium, dérégulé dans de nombreux cancers.

Mots clés : Glioblastome, Cellule Souche Cancéreuse, Antigène, Biomarqueur, Protéine Transmembranaire, Calcium

Résumé en anglais

Glioblastoma are the most common and aggressive nervous system tumors. With a median overall survival smaller than 2 years, usual therapies remain inefficient. This failure could be explained in part by the existence of cancer stem cells. These cells share several properties with stem cells which make them resistant to glioblastoma treatments. This is why it is important to identify and target them to suppress the whole tumor.

The goal of this thesis work is to identify glioblastoma stem cells (gCSCs) biomarkers. To this end, we first developed a global method predicting cancer antigens from microarray data. Then, by studying gCSCs we identified several putative biomarkers and generated insights concerning the calcium signals which are deregulated in numerous cancers.

Keywords : Glioblastoma, Cancer Stem Cell, Antigen, Biomarker, Transmembrane Protein, Calcium