Etude de l'impact de la dérégulation transcriptionnelle liée à des transcrits chimères initiés à partir d'éléments répétés de type LINE-1 dans la tumorigenèse gliale

Marie-Elisa Pinson

Année 2017 N° d'ordre

ECOLE DOCTORALE
DES SCIENCES DE LA VIE ET DE LA SANTE

Thèse

Présentée à l'Université Clermont Auvergne

pour l'obtention du grade de DOCTEUR
(Décret du 5 juillet 1984)

Spécialité Génétique et Biologie Moléculaire

soutenue le 6 décembre 2017

PINSON Marie-Elisa

Etude de l'impact de la dérégulation transcriptionnelle liée à des transcrits chimères initiés à partir d'éléments répétés de type LINE-1 dans la tumorigenèse gliale.

Président :	Mme VAURY Chantal, Directeur de Recherche
Rapporteurs :	M. CRISTOFARI Gaël, Chargé de Recherche Mme LESAGE Pascale, Directeur de Recherche
Examinateur :	M. DE SMET Charles, Professeur
Directeurs de thèse :	Mme VAURS-BARRIERE Catherine, MCU M. ARNAUD Philippe, Directeur de recherche

Laboratoire Génétique Reproduction et Développement UMR CNRS 6293 – Clermont Université – INSERM U1103

Remerciements

Tout d'abord, je tiens à remercier les membres de mon jury d'avoir accepté de juger mes travaux de thèse et pour leur temps pour la lecture de ce manuscrit.

Merci à Chantal VAURY, directrice du GReD, de m'avoir accueillie au sein de son laboratoire, pour les voyages et les bons moments partagés lors de congrès. Merci d'avoir accepté de faire partie de mon jury.

Merci à l'ensemble des membres du laboratoire pour leur disponibilité et leurs conseils : tronc commun et services administratifs.

Un grand merci à Philippe de m'avoir recrutée au sein de son équipe. Pour ta bienveillance, ta bonne humeur, tes blagues. Mais surtout merci pour le savoir que tu m'as transmis, l'aide que tu m'as apporté dans les moments de doutes.

Catherine, merci est presque trop faible. Je te suis extrêmement reconnaissante pour ton engagement, tes conseils, ta présence. Nos discussions pour faire avancer le projet quand parfois les expériences ne fonctionnaient pas. Pour m'aider à structurer mes idées qui partent parfois (ou souvent) un peu dans tous les sens. Et merci dans un tout autre registre pour tes conseils culinaires et ma découverte de ton pounti qui reste à jamais un merveilleux souvenir.

Merci à tous les membres de l'équipe Arnaud, la « dream team » sans qui mon arrivée sur Clermont aurait été aussi noire que sa cathédrale. Pour les sorties restaurants, les randonnées, laser game (même si nous avons dû tricher pour gagner, n'est-ce pas Philippe !).

Merci à David avec qui j'ai partagé mon bureau pendant un temps, pour ta musique et nos discussions.

Merci à Mélanie, pour nos moments en pièce PCR !

Merci à Stéphanie, Bertille et Isabelle, les trois inséparables qui ont rendu les moments de manip au labo beaucoup plus ludiques !

Bertille, tu as été très présente pour moi à différents moments de ma thèse et je t'en remercie. Pour les trajets Clermont-Limoges qui passaient tellement vite quand nous étions ensemble. Et puis tu m'as beaucoup aidée pour les manips aussi surtout pour les expériences de RACE (même si elles ne sont toujours pas optimisées).

Stéphanie, tu es un rayon de soleil, toujours de bonne humeur, agréable ! Ton aide m'a été tellement précieuse pour les qAMP avec ton expertise sur la méthylation ADN.

Isabelle, grâce à nos pauses sur la passerelle, les journées étaient moins longues…

Anne, tu es une boule d'énergie positive qui emmène tout sur ton passage. C'est rafraichissant ! Tu es toujours optimiste et tu m'as toujours aidée à trouver des solutions quand j'étais dans une impasse. La qPCR n'a pas de secret pour toi et les stats non plus ! Merci pour le temps que tu trouves toujours à m'accorder, que ce soit pour m'aider à lancer un projet professionnel ou à discuter des possibilités de post-doc !

Francky le seul homme de l'équipe après le chef ! Tu es comme une mascotte ! Ta vision de la démarche scientifique, tes idées, tu es d'une aide précieuse en bio-info, en biologie, quand tu dois présenter un poster et que tu es en stress ! Pour ces moments de délires qui permettent d'avancer !

Céline même si tu as passé beaucoup de temps dans ton « aquarium », cela a été un plaisir de partager des petits moments de discussion autour d'un café !

A mes stagiaires, Ophélie et Marie, qui ont permis au projet d'avancer pour les L1PA7 et 8 ! Cela a été un plaisir de vous co-encadrer.

Jil et Elisa, vous êtes la nouvelle génération, ma bouffée d'oxygène. Surtout depuis le déménagement, je suis toujours heureuse de vous retrouver. D'arriver et d'entendre Jil crier « ****re » et tout son panel d'expression bien à elle qui la rend unique ! N'est-ce pas « Crousti » ! Et Elisa par ta douceur et ta gentillesse les journées sont merveilleuses.

Merci à Yoan pour son aide au cours de ma thèse non seulement pour le logiciel mais aussi pour les discussions, les bonbons. Je te souhaite beaucoup de réussite car tu es une belle personne.

Une dédicace spéciale à mes amis thésard présents et passés, avec qui j'ai partagé 3 ans de sourires, de bonheurs, de galères mais surtout de bière lors d'un quizz ! A Emilie, Cristiana, Manue, Lilia, Stephie, Guiguiette, Benjamin, Quentin, Marion, Aurélie, Sabine, Fabiana, Colin, Quentin :

« **M**es collègues thésards qui partageaient mon espoir,

Etre enfin docteur sera un vrai bonheur

Rengaine incessante et journées éreintantes

Ce long chemin est presque à son terme

Intégrant toutes les données, la réussite est à notre portée ! »

Merci à tous ceux qui croisent mon chemin tous les jours, avec qui je partage un repas ou une discussion dans un couloir, c'est ces petits moments avec vous qui font que la journée est plus belle et enrichissante.

Merci également à Loïc et Cédric pour les TP de biologie moléculaire et les cours qui sont beaucoup plus fun à préparer !

Je tiens particulièrement à remercier ma famille et mes amis pour leur soutien sans faille ! Vous avez été mon roc dans la tempête.

Mon amour, Guillaume, tu as réussi à me supporter pendant ces 3 ans, dans les moments de doutes, de pleurs, mais aussi de joies !

Ma petite Maman, merci pour la vache qui rit, ton amour inconditionnel et ta confiance qui me portent et qui font de moi qui je suis aujourd'hui ! Je suis fière d'être ta fille et d'avoir une maman aussi extraordinaire qui sait me faire rire, me faire parler quand il le faut, me tendre son épaule pour pleurer.

Je tenais à remercier particulièrement ma grand-mère. Mon rêve était que tu sois là pour me voir arriver où j'en suis aujourd'hui ! Toi qui m'a bercée, élevée, consolée quand maman ne pouvait pas être là. Cette dédicace me fait monter des larmes de joies, tellement je suis fière. Fière du chemin parcouru, fière de mes origines. Regarde Mamé, j'y suis…

A toi et à Papé, je vous dédie cette thèse.

Avant-propos

Les éléments répétés représentent 45% de la séquence du génome humain. Toutefois, de par leur grand nombre et leur forte homologie entre eux, la contribution de ces éléments au transcriptome reste difficile à appréhender. Notamment, l'implication des éléments LINE-1 récents (L1PA1 à 7) sur le profil transcriptionnel des cellules et tissus est probablement sous-évaluée. En effet, les éléments L1 récents possèdent, en plus de leur promoteur sens, un promoteur antisens (ASP) qui permet la transcription des séquences adjacentes (pouvant contenir un gène) sous la forme de transcrits dit « chimères » ou LCT (LINE-1 Chimeric Transcripts). De tels LCT ont été décrits dans la littérature mais la question reste entière de l'étendue pangénomique de cette activité transcriptionnelle.

Dans cette optique, un nouvel outil nommé CLIFinder (Chimeric Line Finder) a été développé et validé pour permettre l'identification de LCT à partir de données de RNA-seq paired-end orientées. Les RNA-seq de 13 gliomes, qui sont les cancers du cerveau les plus fréquents chez l'adulte, et de 3 tissus contrôles ont été étudiés avec CLIFinder. 2675 chimères ont été détectées dont 84% impliquent des L1 récents et 50% sont spécifiquement détectés dans les tumeurs. 78 chimères correspondent à des LCT déjà décrits dans la littérature. L'étude d'un groupe de chimères par marche en 5' par RT-PCR valide que 89% (56/63) des chimères impliquant des L1 récents sont initiés dans la région de l'ASP et correspondent à des LCT. Des études de RT-qPCR sur une cohorte plus large de 51 gliomes montrent que les 56 LCT testés, incluant des LCT tumeurs spécifiques, sont exprimés non seulement dans les tumeurs mais aussi dans les tissus contrôles. Pour autant, 70% des LCT tumeurs spécifiques possèdent une surexpression tumorale significative. Ces résultats suggèrent donc une transcription basale provenant de l'ASP dans les tissus normaux et que la dérégulation transcriptionnelle liée aux LCT dans les gliomes passe par une surexpression.

Par ailleurs, afin de déterminer le ou les mécanismes sous-jacents impliqués dans l'augmentation de l'activité transcriptionnelle de l'ASP, deux hypothèses ont été testées. La première implique une diminution de la méthylation ADN du promoteur de L1. Toutefois, mes résultats tendent à réfuter cette hypothèse puisqu'aucune diminution de la méthylation n'est retrouvée au niveau de la région promotrice des L1 impliqués dans la transcription des LCT surexprimés. En revanche, les gènes associés aux LCT dont l'expression est dérégulée en contexte tumoral présentent une dérégulation dans le même sens qui corrèle avec celle du LCT. Ceci suggère qu'une augmentation de l'activité transcriptionnelle aux loci des LCT serait responsable de la surexpression de ceux-ci.

Enfin, parmi les LCT surexprimés que nous avons identifiés, 2 LCT candidats pourraient jouer un rôle fonctionnel dans l'initiation, la progression et/ou l'agressivité tumorale.

Pour conclure, mes travaux valident que CLIFinder est un outil pertinent pour l'identification pangénomique de LCT exprimés dans différents types tumoraux à partir des données de RNA-seq paired-end orientées. L'observation d'une récurrence entre différents types tumoraux testés ainsi qu'une surexpression tumorale de certains LCT suggère qu'ils pourraient jouer un rôle fonctionnel dans les processus de tumorigenèse.

Table des matières

TABLE DES ILLUSTRATIONS

FIGURES

TABLES

ANNEXES

INTRODUCTION

Les éléments transposables sont des séquences d'ADN capables de se mobiliser dans un génome par un mécanisme appelé transposition. Ces éléments sont présents chez tous les organismes vivants et sont l'un des constituants les plus importants des génomes eucaryotes. Ces séquences mobiles d'ADN sont considérées comme des moteurs puissants de l'évolution et de la biodiversité des organismes vivants.

1. Génome humain et éléments transposables.

Grâce à l'avancée des techniques de biologies moléculaires et notamment au séquençage de Sanger, notre compréhension des génomes a progressé. En 2001, une première ébauche de la séquence complète du génome humain a été publiée et a montré qu'il contient 47% de séquences répétées (Lander *et al.*, 2001). Parmi ces éléments répétés, la majorité correspond à des éléments transposables (ET). Malgré les travaux de Barbara McClintock sur le génome du maïs, qui a émis l'hypothèse que ces ET jouent un rôle dans la régulation des génomes, ceux-ci étaient tout d'abord considérés comme de l'ADN « poubelle ». Grâce aux nombreuses études qui ont suivies et aux avancées des techniques de séquençage, son hypothèse a été validée et reconnue. Aujourd'hui, l'étude des ET représente un champ d'investigation prometteur.

Ces ET, qui ont été acquis et conservés au cours de l'évolution, peuvent avoir des rôles positifs, négatifs ou neutres au sein des génomes, seront détaillés dans le chapitre 2.

Les ET ont été classifiés selon leurs mécanismes de transposition en deux classes majeures (Wicker *et al.*, 2007).

Les éléments de classe I, appelés rétrotransposons, vont se mobiliser par un mécanisme de « copier-coller » *via* un ARN intermédiaire. Parmi les rétrotransposons sont distingués :

- Les rétrotransposons autonomes contenant les éléments LTR (Long Terminal Repeats) provenant de rétrovirus endogènes.
- Les rétrotransposons non LTR autonomes représentés par les LINE (Long Interspersed Nuclear Elements),
- Les rétrotransposons non LTR et non autonomes correspondent aux éléments SINE (Short Interspersed Nuclear Elements).

Les éléments de la classe II, ou transposons ADN, se mobilisent *via* un mécanisme de « couper-coller » sans ARN intermédiaire. Les caractéristiques de chacun de ces éléments vont être détaillées ci-après.

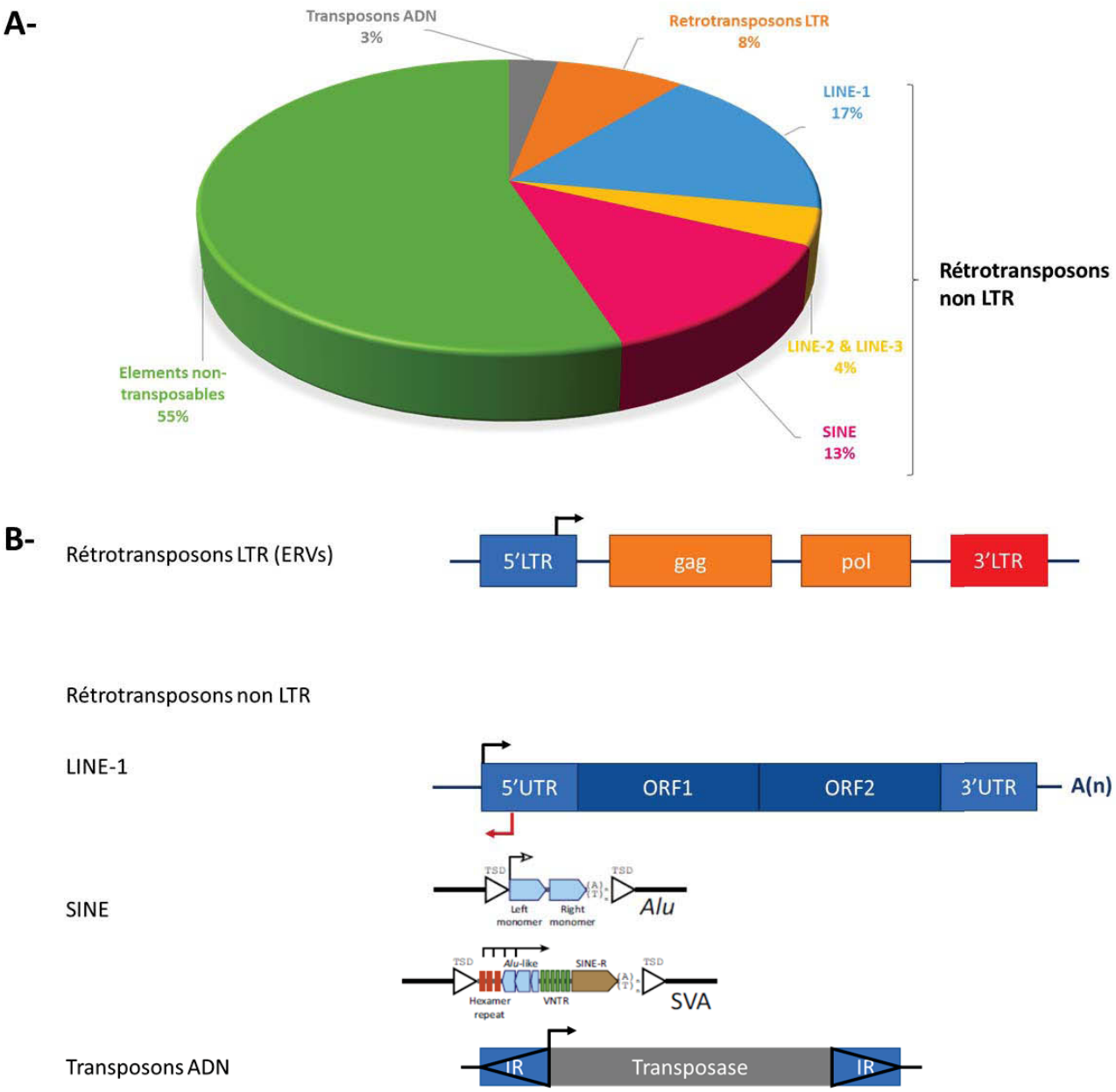


Figure 1 : Caractéristiques des éléments transposables du génome humain. *(Cordaux & Batzer, 2009; Faulkner & Garcia-Perez, 2017)*

A- **Composition en éléments transposables du génome humain.** Environ 45% du génome humain est constitué d'ET. La majorité est constituée des rétrotransposons non LTR (Long Terminal Repeat) comme les éléments LINE-1, SINE.

B- **Structure des différentes classes d'éléments transposable du génome humain.** Les différentes classes d'ET des mammifères sont représentées.

Les rétrotransposons LTR sont constitués de deux séquences LTR aux extrémités permettant leur transposition *via* un ARN intermédiaire grâce à l'activité des gènes *gag* (protéines de capside) et *pol* (transcriptase inverse, intégrase, protéase et RNAseH).

Les éléments LINE-1 (Long Interspersed Nuclear Element) possèdent une région 5'non-traduite (5'UTR) contenant un promoteur bidirectionnel (sens et antisens), deux cadres ouverts de lecture *ORF1* (protéine chaperonne) et *ORF2* (endonucléase et transcriptase inverse), ainsi qu'une région 3'non-traduite (3'UTR) et un signal de polyadénylation (An).

Les éléments SINE (Short Interspersed Nuclear Elements) ne sont pas capables de rétrotransposer seuls. Ils nécessitent la machinerie des LINE notamment la transcriptase inverse produite par l'*ORF2* des L1. De plus, ce sont de courtes séquences d'environ 300pb. Les deux sous-familles principales sont les éléments *Alu* et les éléments *SVA*. Les éléments *Alu* sont caractérisés par un monomère gauche et un monomère droit, encadrés par des séquences TSD (Terminal Side Duplication). Pour les éléments *SVA*, leur séquence est plus complexe avec une région répétée d'hexamères, des séquences *Alu*-like, une région VNTR et une région SINE-R, l'ensemble encadré par des séquences TSD.

Les transposons ADN sont caractérisés par des séquences répétées (IR) permettant leur transposition grâce à la transposase.

1.1. Les éléments transposables de classe I.

Les ET de classe I sont appelés rétrotransposons. Ils se mobilisent *via* un ARN intermédiaire qui sera rétro-transcrit en ADNc afin de permettre son insertion à un nouvel endroit du génome. Selon la famille de rétroéléments considérée, les intermédiaires de ce mécanisme de « copier-coller » varient. Ces ET de classe I sont constitués des rétrotransposons autonomes LTR, des rétrotransposons autonome non LTR : les LINE et les éléments non autonomes non LTR : les SINE.

1.1.1. Les rétrotransposons LTR.

Les rétrotransposons LTR représentent 8% du génome humain (Lander *et al.*, 2001) (Figure 1-A). Ils proviennent de rétrovirus qui se sont intégrés dans le génome humain, d'où le nom de rétrovirus endogènes. La famille majoritaire de rétrotransposons LTR chez l'Homme sont les rétrovirus endogènes HERV (HERV-L, HERV-K) (Cohen *et al.*, 2009). Leur structure est caractérisée par la présence de séquences LTR à chacune de leurs extrémités (Figure 1-B). Ces séquences LTR vont servir de bornes de reconnaissance afin de permettre la rétrotransposition de l'élément. Ces rétrotransposons LTR sont constitués des gènes *gag* codant une protéine de capside, *pol* codant une transcriptase inverse, intégrase, protéase, RNase H et *env* codant les protéines d'enveloppe. Ces rétrotransposons LTR ont souvent une perte totale du gène *env*. Ce gène sert à la biologie des virus pour la formation des virions infectieux. A partir de son promoteur situé dans la région 5'LTR, l'intermédiaire ARN du rétrotransposon sera transcrit grâce à l'ARN polymérase II. Celui-ci sera reconnu et pris en charge par la transcriptase inverse qui va assurer sa rétrotranscription et les autres protéines vont participer à son intégration au niveau du site de coupure généré par l'endonucléase (Rebollo *et al.*, 2012). C'est par ce mécanisme que les rétrotransposons LTR se répliquent au sein du génome et le modifient. Parmi les rôles que peuvent jouer ces rétrotransposons, on peut citer celui du rétrovirus *MER41B* intégré en amont du gène *AIM2* qui code une protéine de l'inflammasome et contribue à la défense antivirale. *MER41B* contient des sites de fixation pour le facteur de transcription STAT1 inductible par les interférons. Cette insertion permet de moduler la réponse immunitaire notamment en augmentant la transcription du gène *AIM2* en réponse à l'interféron *via* la voie de signalisation STAT (Chuong *et al.*, 2016). Par ailleurs, il a été décrit la présence de 2 rétrotransposons LTR dans 2 introns du gène *ERBB4* qui code un récepteur à activité tyrosine kinase. En contexte tumoral, les LTR de ces rétrotransposons sont anormalement activés. Ainsi, ils se comportent comme des promoteurs alternatifs conduisant à l'expression de 2 isoformes

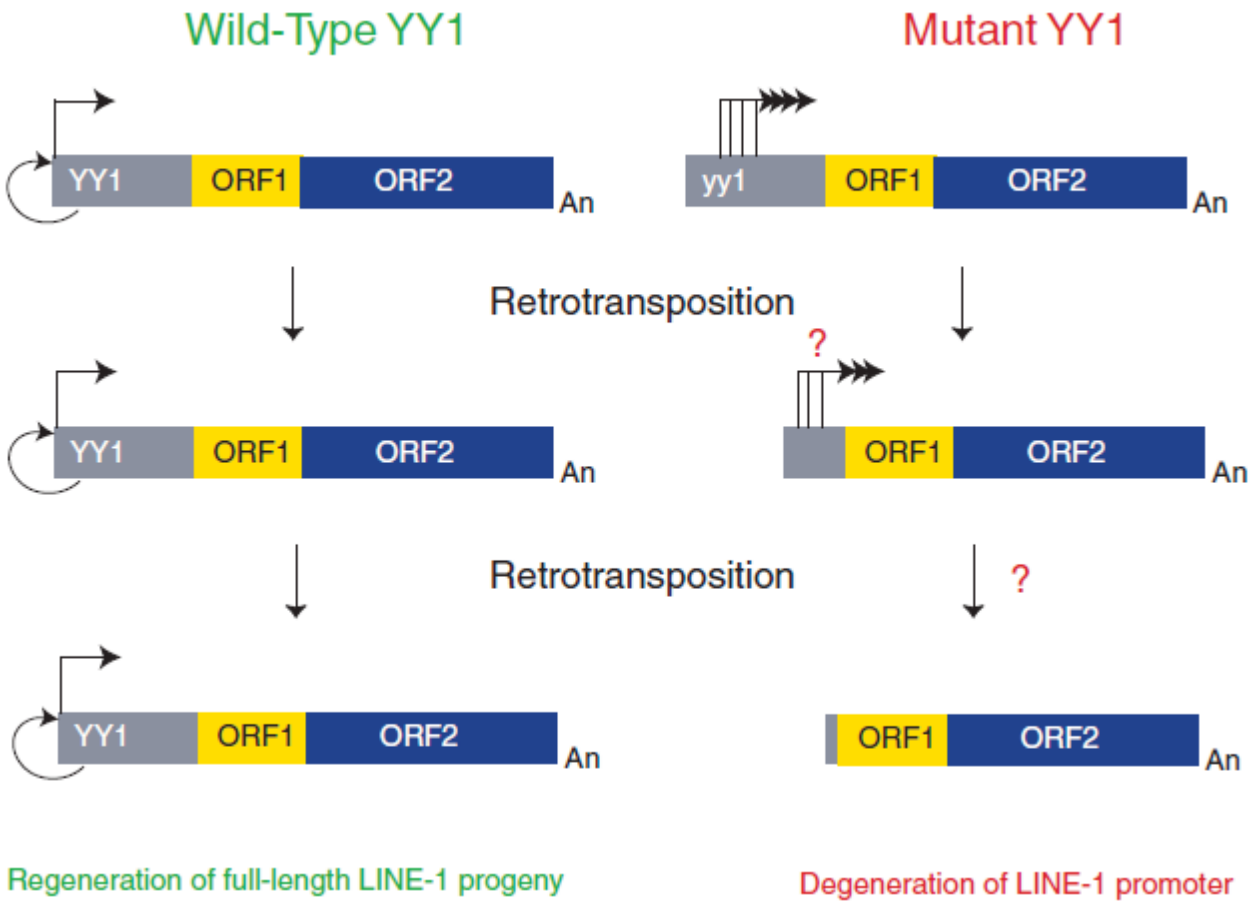


Figure 2 : Modèle du rôle des sites de liaisons du facteur de transcription YY1 dans la rétrotransposition des L1. *(Athanikar* et al.*, 2004)*

Ce modèle décrit la rétrotransposition provenant d'un progéniteur L1 pleine taille qui contient soit un site de fixation pour le facteur de transcription YY1 sauvage (partie gauche) soit muté (partie droite). Dans le scénario sauvage, le L1 pleine taille contient une extrémité 5'UTR intacte à partir de laquelle la transcription est initiée au niveau du site +1 par le promoteur sens interne qui peut être régénéré par la rétrotransposition. Ce progéniteur pleine taille peut produire un élément L1 pleine taille qui a rétrotransposé à un autre endroit du génome, de manière autonome. Celui-ci pourra à son tour rétrotransposer car il est pleine taille. Par comparaison, un élément L1 pleine taille ayant perdu la fonction de liaison au facteur de transcription YY1 ne peut pas initier la transcription à partir du site +1 du promoteur sens interne de la région 5'UTR. A la place, l'initiation de la transcription pourra avoir lieu à différentes positions de l'extrémité 5'UTR (représenté par les nombreuses flèches). Ensuite, les transcrits résultants vont devenir progressivement plus courts au fur et à mesure des rétrotranspositions jusqu'à conduire à la formation d'un élément qui ne sera plus autonome pour la rétrotransposition. Le point d'interrogation en rouge dans la partie droite indique une incertitude concernant la possibilité des L1 produits à être transcrits et leur capacité à être compétents pour la rétrotransposition. Il est admis que les séquences génomiques adjacentes peuvent influencer l'expression de ces L1 et leur capacité à rétrotransposer *in vivo*.

aberrantes tronquées dans leur partie N-terminale. Ces isoformes sont constitutivement actives et jouent un rôle promoteur dans l'étiologie du lymphome anaplasique à grandes cellules (Scarfò *et al.*, 2016). Ces 2 exemples, l'un en contexte physiologique, l'autre en contexte pathologique, montrent que selon le site d'intégration de ces rétrotransposons LTR et le contexte physiologique, les conséquences peuvent être soit positives, soit délétères.

1.1.2. Les éléments LINE (Long Interspersed Nuclear Element).

Les rétrotransposons autonomes non LTR de type LINE représentent plus de 17% du génome humain (Lander *et al.*, 2001) (Figure 1-A). Ils sont constitués de trois familles d'éléments : les LINE-1 (L1), les LINE-2 (L2) et les LINE-3 (L3). La famille la plus abondante dans le génome humain est la famille des L1. Les L2 et L3 sont structurellement distincts des L1, et ne sont plus compétents pour la rétrotransposition. Ces deux familles ne seront pas détaillées ici. Dans le génome humain, il existe 500 000 copies de L1 dont seulement 7 000 copies sont pleine taille (Lander *et al.*, 2001). Un L1 pleine taille mesure 6 Kb et est constitués de deux extrémités 5' et 3' non traduites (UTR), de deux cadres ouverts de lecture, *ORF1* et *ORF2*, ainsi qu'un signal de polyadénylation (Figure 1-B). La région 5'UTR du L1 est une région d'environ 910 pb contenant un promoteur sens interne dont la transcription est initiée à partir de l'ARN polymérase II (Swergold, 1990). Ce promoteur sens est caractérisé par une absence de boîte TATA (TATA box-less). Aussi une combinaison fine de facteurs de transcription va être nécessaire à l'activation de la transcription à partir de ce promoteur sens interne (Lavie *et al.*, 2004). Différents sites de liaison pour des facteurs de transcription sont présents dans cette région pour moduler la transcription du L1 selon les conditions de l'environnement :

- Un site de fixation pour le facteur de transcription YY1 localisé du +13 au +21 sur le brin non codant est nécessaire pour l'initiation de la transcription. En effet, dans un modèle de rétrotransposition, quand le site de fixation pour le facteur de transcription YY1 est muté, le L1 ainsi affecté va progressivement se raccourcir jusqu'à ne plus être compétent pour la rétrotransposition (Athanikar *et al.*, 2004) (Figure 2).
- Deux sites de fixation pour le facteur de transcription SOX, un situé de +472 à +477 et un de +572 à +577.
- Un site de fixation pour le facteur de transcription RUNX3 localisé de +83 à +101.

Ces sites de fixation pour les facteurs de transcription SOX et RUNX3 sont nécessaires à la rétrotransposition des L1 *in vitro* (Yang, 2003).

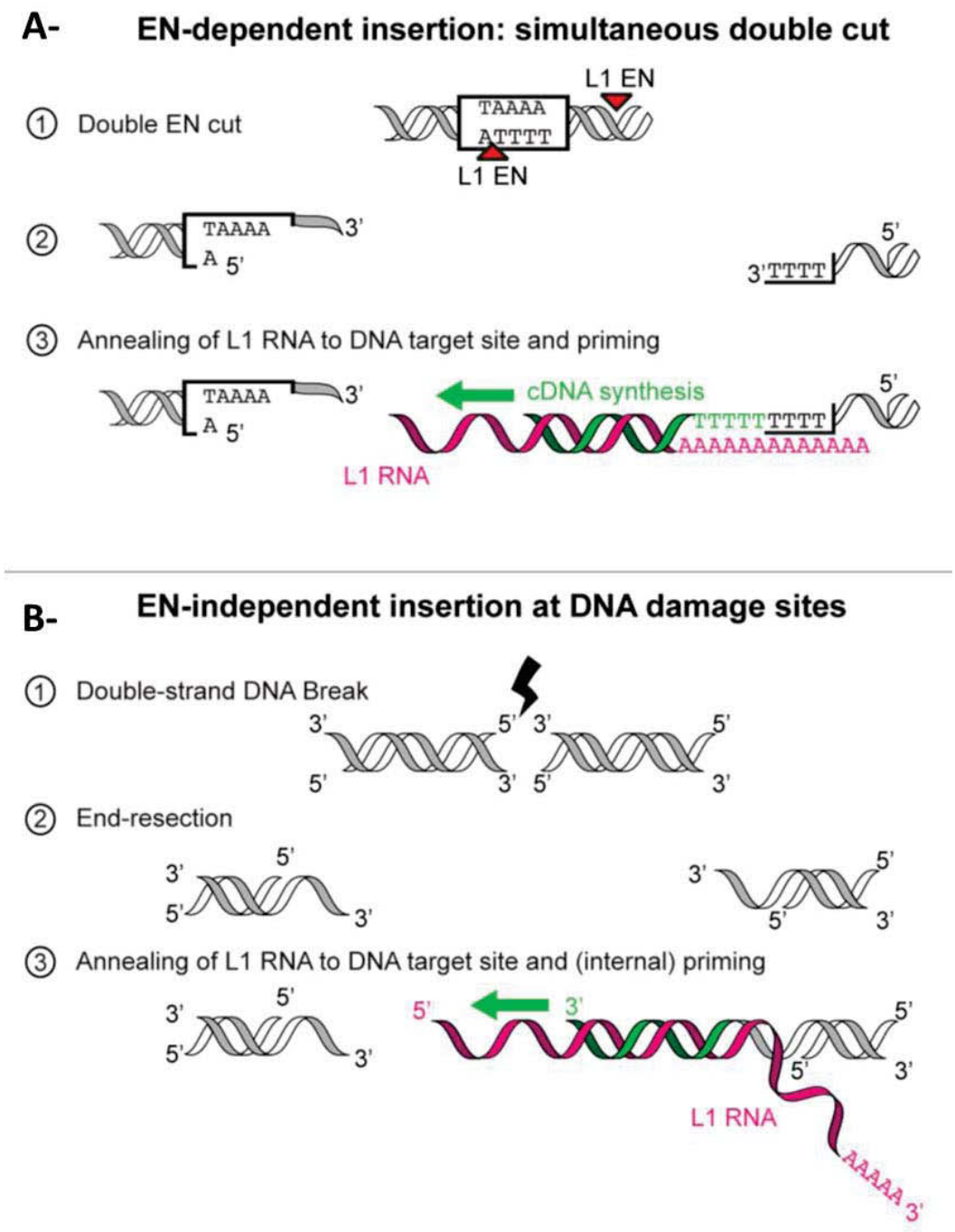


Figure 3 : Différentes hypothèses sur les mécanismes permettant à l'ARN du L1 de cibler l'ADN. *(Viollet* et al.*, 2014)*

A- La dimérisation de l'ORF2p conduit à la synchronisation de la coupure et de l'activité endonucléase, au cours de la rétrotransposition. L'extrémité résultante a une extrémité sortante en 3', qui permet d'amorcer la synthèse de l'ADNc du L1 en se servant de l'ARN du L1 comme matrice.

B- Après la cassure double brin de l'ADN, les facteurs de la réparation vont dégrader en partie cette zone en générant une extrémité sortante en 3'. Cette nouvelle extrémité ne possède pas nécessairement la séquence de reconnaissance pour la liaison de l'EN du L1. Par conséquent, un mécanisme d'appariement peut avoir lieu avec l'ARN d'un L1 au niveau de la région lésée et permettre ainsi l'insertion et la réparation.

(Flèche rouge : site de coupure de l'EN du L1 ; en vert : l'ADNc ; en rose : ARN du L1)

Chez certains éléments L1, cette extrémité 5'UTR est également caractérisée par la présence d'un promoteur antisens, nommé ASP, situé à 450 pb du début du L1 (Speek, 2001). Celui-ci est dépendant de l'ARN polymérase II pour sa transcription et sera détaillé dans le chapitre 2.

L'*ORF1* du L1 code pour une protéine de 40 KDa, contenant un domaine de liaison à l'ARN et ayant un rôle de protéine chaperonne. Cette protéine est nommée ORF1p. Celle-ci se lie à l'ARN du L1 dans le cytoplasme, avec une préférence pour l'ARN du L1 duquel elle provient (préférence en *cis*), formant ainsi des particules ribonucléoprotéiques (RNP). Cette stabilisation de l'ARN du L1 est nécessaire à la fixation de protéines de l'hôte afin de permettre la rétrotransposition. L'*ORF2* code pour une protéine de 150 KDa possédant des activités de transcriptase inverse (RT) et d'endonucléase (EN). C'est cette protéine qui conduit la dernière étape de la rétrotransposition nommée TPRT (Target-Primed Reverse Transcription). Suite à l'import du RNP du L1 dans le noyau de la cellule où il a été transcrit, l'EN du L1 va couper l'ADN génomique au niveau d'une séquence consensus de reconnaissance 5'-TTTT/A-3' (où le clivage s'effectue au niveau du « / ») en libérant un 3'-hydroxyl (3'-OH) qui servira d'amorce à la transcription inverse de l'ARN du L1 par la RT, générant ainsi l'ADNc à intégrer. L'ORF1p va faciliter cette étape grâce à son rôle de protéine chaperonne. Puis le second brin va être synthétisé en se servant de l'autre comme matrice et une ligase intervenant dans les mécanismes de réparation de l'ADN va permettre à la molécule de retrouver son intégrité (Hulme *et al.*, 2006) (Figure 3-A). Le mécanisme de TPRT est le principal utilisé pour la rétrotransposition des L1 mais il en existe un autre, décrit en cas de cassure double brin de l'ADN. L'ARN du L1 va reconnaître une homologie au niveau du site de coupure et s'intégrer afin de réparer la lésion. Ce mécanisme a été décrit comme un mécanisme de réparation de l'ADN ancestral avant l'apparition des systèmes de réparation (Figure 3-B). Pour être mobile, c'est-à-dire compétent pour la rétrotransposition, l'intégralité du L1 doit-être transcrite à partir du promoteur sens, c'est le cycle de transposition des L1 (Figure 4). L'insertion des L1 se fait préférentiellement dans les régions riches en A-T (Lander *et al.*, 2001).

L'extrémité 3'UTR quant à elle contient un signal de polyadénylation ou une région riche en A pour la terminaison de la transcription (Hulme *et al.*, 2006). En amont de ce signal se trouve une séquence permettant la liaison du facteur d'export nucléaire NXF1 (TAP) permettant la sortie de l'ARN du L1 du noyau vers le cytoplasme (Lindtner *et al.*, 2002).

Ces L1 ont participé à la modulation et à l'évolution des génomes (cf chapitre 2), mais ceux-ci peuvent également être à l'origine du développement de pathologies (cf chapitre 3).

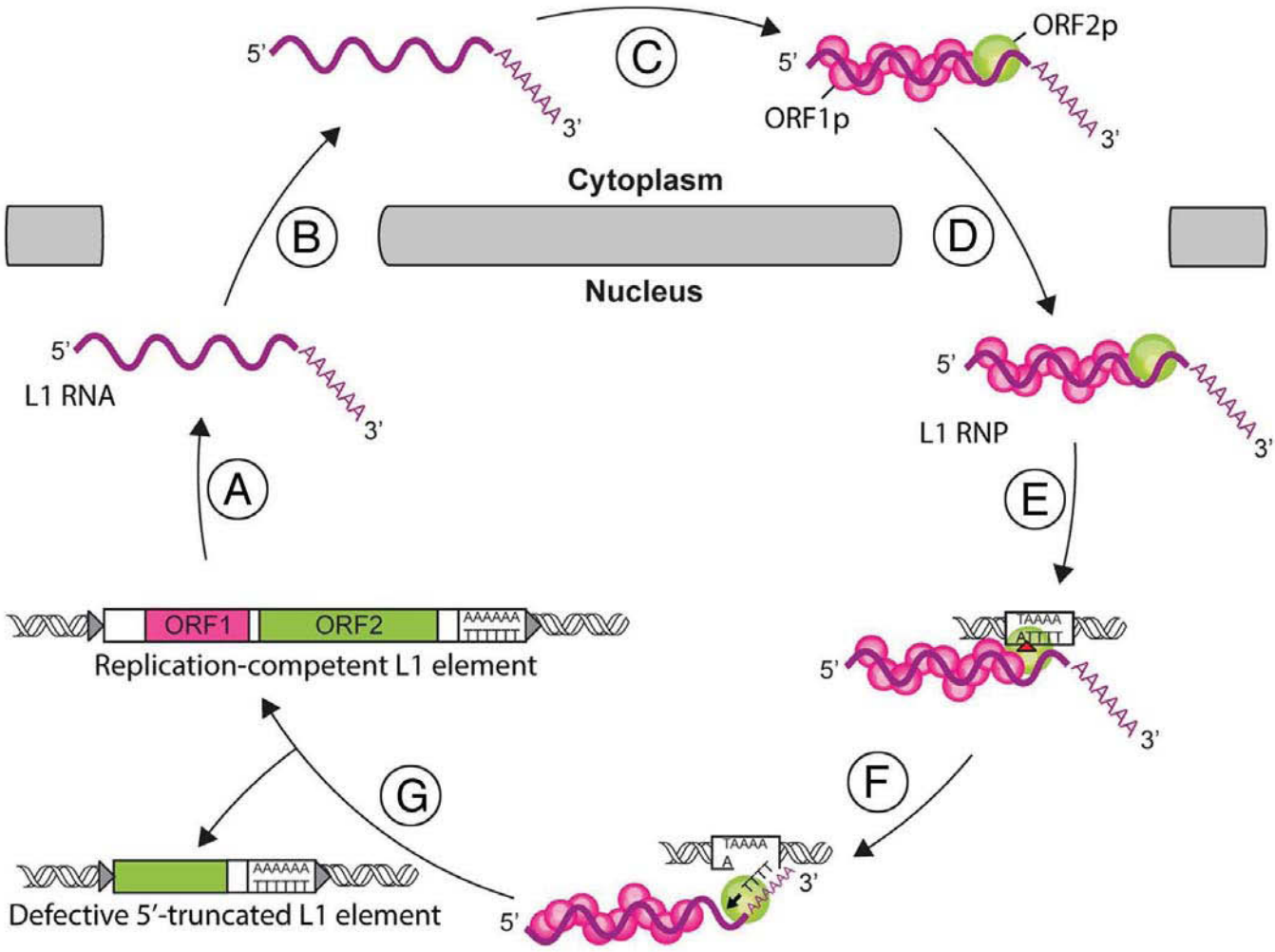


Figure 4 : Le cycle de transposition des L1 *(Viollet* et al*., 2014)*

(A) La réplication des L1 commence par la transcription d'un ARN biscistronique. (B) L'ARN du L1 est exporté dans le cytoplasme. (C) Les protéines ORF1p et ORF2p sont traduites et se lient à leur propre ARN pour former les particules ribonucléoprotéiques (RNP). (D) Les RNP du L1 sont importés dans le noyau. (E) L'intégration et la transcription inverse ont lieu au site d'intégration de l'ADN génomique. Premièrement, l'activité endonucléase du L1 (EN) coupe l'ADN cible (flèche rouge). (F) Puis la transcriptase inverse du L1 (RT) initie la transcription inverse de l'ARN du L1 en ADNc (Flèche noire). (G) Le mécanisme s'engage dans la phase finale de ce processus avec la résolution de l'intégration dont le mécanisme est encore inconnu. Une transcription inverse partielle peut donner lieu à des copies de L1 tronquées en 5'.

1.1.3. Les éléments SINE (Short Interspersed Nuclear Element).

Les rétroéléments non LTR non autonomes que sont les éléments SINE, représentent 13% du génome (Lander *et al.*, 2001) (Figure 1-A). Ces éléments sont constitués de différentes familles, tout d'abord la famille des éléments *Alu*, qui est la plus abondante, puis la famille des éléments *SVA* (SINE-VNTR-Alu). Les éléments SINE sont très abondants dans le génome humain puisqu'il en existe environ 1 000 000 copies, majoritairement insérées dans les régions riches en G-C (Lander *et al.*, 2001). Ces séquences sont relativement courtes (< 700 pb) et sont définies comme des rétrotransposons non autonomes. Ces éléments vont détourner les protéines produites par les L1 pour leur propre rétrotransposition. Ces éléments SINE sont constitués en général de 2 parties : une partie 5' dont l'origine dépend de la classe du SINE, par exemple un élément *Alu* a pour origine un ARN 7SL et une partie 3' constituée d'une séquence de simple répétition (CA_n) de longueur variable (Vassetzky & Kramerov, 2013) se terminant par une séquence polyA et sont bornés par des séquences TSD (Terminal Side Duplication) (Figure 1-B). Les SINE sont transcrits en ARN à partir de leur promoteur sens interne par l'ARN polymérase III, mais un promoteur antisens a également été décrit dépendant de l'ARN polymérase II (Lai *et al.*, 2009). Sauf pour les éléments *SVA* dont la transcription est initiée à partir d'un promoteur polymérase II. La petite taille des SINE a permis leur forte intégration dans le génome et ils ont largement contribué à l'évolution du génome humain (Elbarbary *et al.*, 2016). Ainsi l'insertion d'un *SVA* dans l'unique intron du gène *AMAC1L3* a permis de générer par transduction en 3' une nouvelle famille génique. Suite à la transcription de ce nouveau gène à partir du promoteur du *SVA*, cette séquence a pu s'insérer à d'autres endroits du génome : sur le chromosome 17 à l'origine du gène *AMAC1*, sur le chromosome 18 à l'origine du gène *AMAC1L1* et sur le chromosome 8 à l'origine du gène *AMAC1L2* (Figure 5). Ces 3 gènes supplémentaires existent chez l'Homme alors que chez les primates et les espèces les plus anciennes, seul le gène ancestral *AMAC1L3* est présent (Xing *et al.*, 2006). Ce gène *AMAC1* (Alternative Macrophage Activation associated CC Chemiokine 1) code pour la protéine CCL18 qui est produite et sécrétée par les cellules du système immunitaire inné pour le recrutement d'effecteurs du système immunitaire adaptatif. Des effets négatifs peuvent également découler de la néo transposition de ces éléments. La néo insertion somatique d'un élément *Alu* dans le gène *BRCA2* perturbe sa transcription. Cet évènement est ainsi impliqué dans l'étiologie d'un type de cancer du sein (Hancks & Kazazian, 2012).

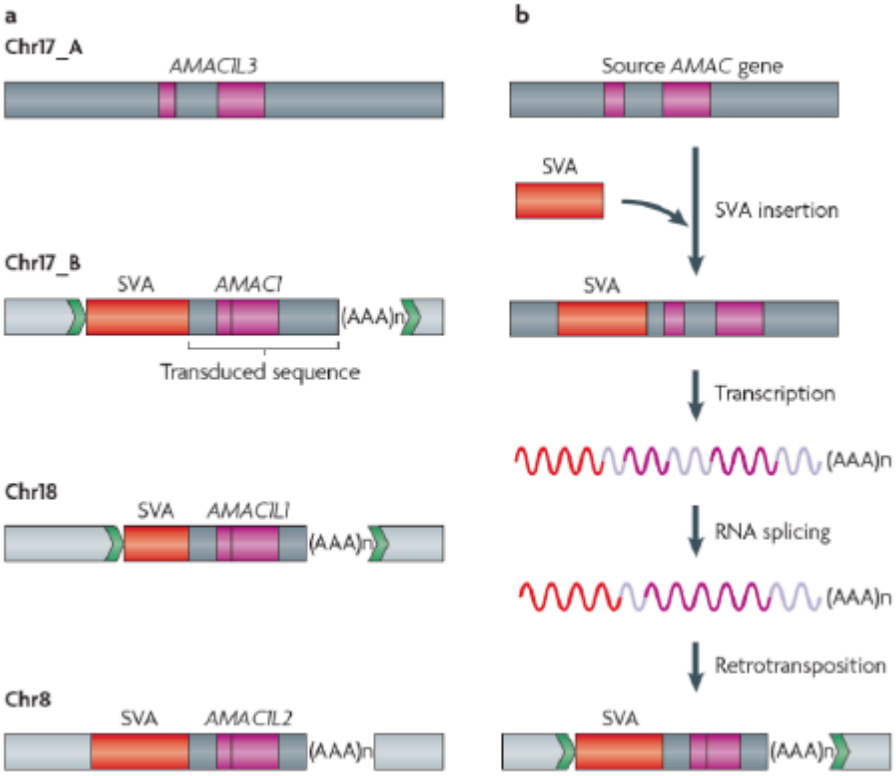


Figure 5 : Naissance d'une nouvelle famille de gènes médiée par la transduction d'un transposon. *(Cordaux & Batzer, 2009)*

L'insertion d'un *SVA* dans l'unique intron du gène *AMAC1L3* a permis de générer par transduction en 3' une nouvelle famille génique. Suite à la transcription d'un transcrit alternatif du gène *AMAC1L3* à partir du promoteur du *SVA*, cette séquence a pu s'insérer à d'autres endroits du génome : sur le chromosome 17 à l'origine du gène *AMAC1*, sur le chromosome 18 à l'origine du gène *AMAC1L1* et sur le chromosome 8 à l'origine du gène *AMAC1L2* (a). La version ancestrale du gène *AMAC1L3A* consiste en 2 exons séparés par un intron. Par contraste, les 3 copies transduites du gène ancestral *AMAC1L3* sont des versions sans intron qui résultent de l'épissage de l'intron au cours du processus de rétrotransposition.

1.2. Les éléments transposables de classe II : les transposons ADN.

Ils constituent près de 3% du génome humain (Lander *et al.*, 2001) (Figure 1-A). Il en existe plusieurs familles dont les plus étudiés sont les éléments *Mariner* et *Sleeping beauty*. Ces éléments possèdent à leurs extrémités de courtes séquences inversées répétées (IR) permettant leur transposition grâce à une enzyme appelée transposase selon un mécanisme de « couper-coller » (Figure 1-B). En effet, la séquence ADN de ces éléments va être excisée et intégrée à un nouvel endroit du génome sans intermédiaire ARN. Les transposons ADN ne sont plus actifs dans le génome humain mais il existe de nombreux exemples de leur domestication. Les gènes *RAG1* et *RAG2* requis pour la recombinaison V(D)J, ont pour origine un transposon ADN, *Transib*. Il a été démontré chez la drosophile et chez l'anophèle, dans lesquelles le transposon Transib a été mis en évidence, que ce transposon se serait inséré dans un gène ancestral de *RAG1*, il y a 500 millions d'années, et aurait été conservé au cours de l'évolution. En effet, des séquences IR témoignent de l'insertion de ce transposon dans un gène ancestral des récepteurs aux cellules B et T (BCR et TCR) suivi par des duplications de gènes, permettant ainsi l'émergence de la machinerie de la recombinaison V(D)J. Ces gènes impliqués dans cette recombinaison vont participer à la création des nombreuses combinaisons nécessaires à la diversité de notre répertoire d'anticorps (Kapitonov & Jurka, 2005).

De plus, l'élément *Sleeping Beauty* est actuellement étudié afin de servir de vecteur d'intégration en thérapie génique de par son innocuité pour les cellules (Turchiano *et al.*, 2014).

Ces ET participent, comme nous venons de le voir, à l'évolution et à la modulation de l'expression des génomes mais peuvent également induire des maladies *via* l'instabilité génomique qu'ils génèrent. Au cours de mes travaux de thèse, je me suis particulièrement intéressée aux ET de type L1 dans le cancer. Le rôle des L1 en conditions physiologiques puis pathologiques seront détaillés dans la suite de cette introduction.

2. LINE-1 et physiologie.

Afin de comprendre le rôle des L1 dans un contexte pathologique, il faut tout d'abord appréhender leur rôle en contexte physiologique.

2.1. Evolution et organisation des LINE-1.

Il existe différentes sous-familles de L1 qui ont évoluées dans le génome humain. En 1995, Smit et ses collaborateurs ont reconstitué grâce à des bases de données d'EST et des outils bio-

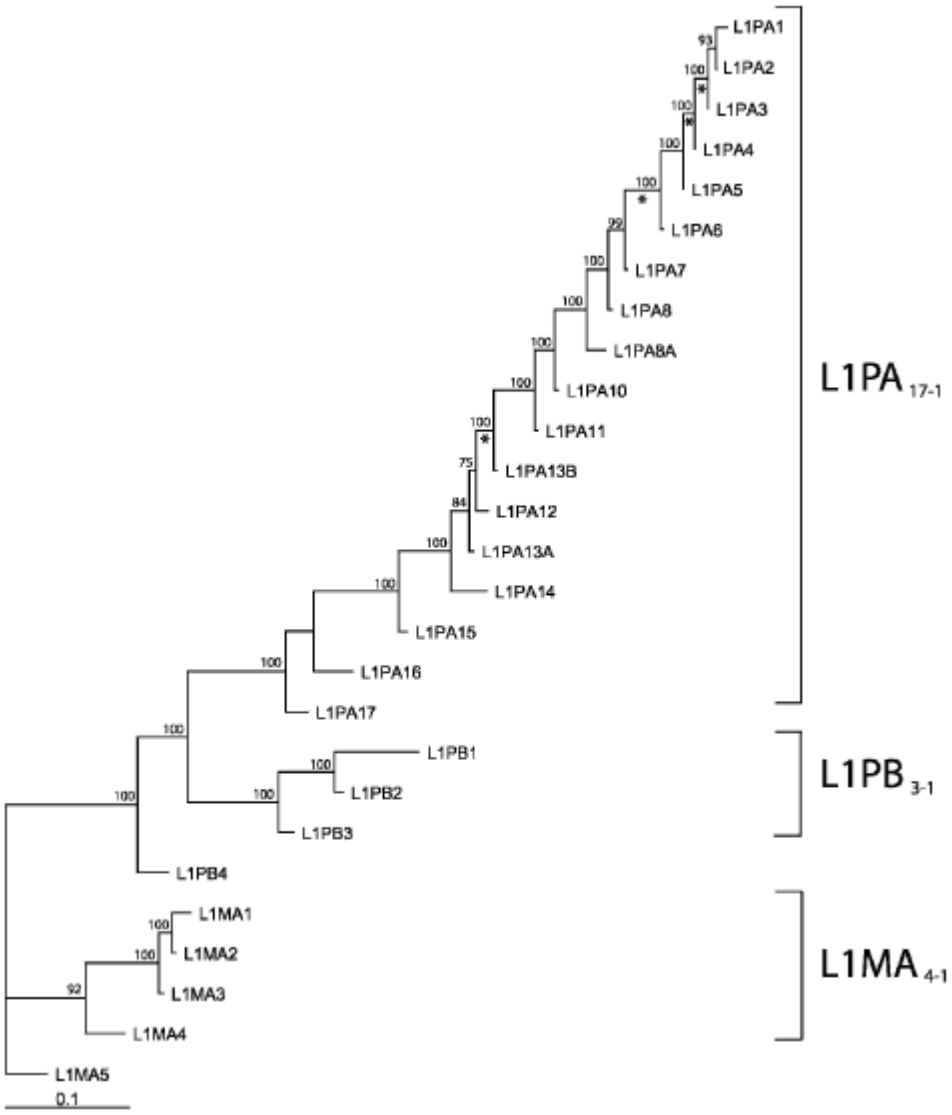


Figure 6 : Phylogénie des séquences consensus de L1. *(Khan, 2005)*

Cet arbre de maximum de vraisemblance est basé sur les séquences consensus des ORF1 et ORF2 de 27 sous-familles de L1. Les nombres au-dessus des nœuds indiquent le pourcentage de temps basé sur les données de 1000 bootstrap répliqués. Les astérisques indiquent les branches sur lesquelles un model ratio estime un ω>1.

informatiques, des séquences consensus de différentes sous-familles sur lesquelles des analyses phylogénétiques ont été réalisées. Leur hypothèse est que la majorité des L1 humains ont été acquis au cours des derniers 30 millions d'années et proviennent de gènes L1 source qui ont proliféré au sein du génome jusqu'à l'apparition d'une nouvelle sous-famille plus compétente, les anciens se figeant dans le génome. Ces analyses se basent sur les homologies de séquences au niveau de la région 3' UTR, mais aussi de la jonction entre la fin de l'*ORF2* et l'extrémité 3'UTR. Au cours de cette étude, ils ont standardisé le nom des différentes sous-familles et proposé une nomenclature encore utilisée aujourd'hui. Le nom L1 est suivi de deux lettres capitales indiquant le groupe majeur auquel appartient la sous-famille, sachant que la première lettre correspond à l'ordre phylogénétique (P pour primate et M pour mammifère), la lettre suivante correspond à une subdivision basée sur la structure globale de la séquence 3'UTR (A, B, C, D ou E). Enfin à chaque sous-famille a été attribué un nombre, allant de la sous-famille la plus récente à laquelle est attribué le chiffre 1, à la plus ancienne. Les analyses phylogénétiques et les alignements de séquences ont permis la reconstitution des séquences consensus pour les sous-familles L1PA1 à 16, L1PB1 à 3, L1MA1 à 10, L1MB1 à 8, L1MC1 et 2, L1MD1 et 2 et L1ME1 à 3 (Smit *et al.*, 1995).

En 2005, Khan et ses collaborateurs ont affiné ces séquences consensus grâce au séquençage du génome humain et au développement de nouveaux outils bio-informatiques. Ils ont confirmé les précédentes observations et établi un arbre phylogénétique (Figure 6) basé sur les séquences consensus des séquences *ORF1* et *ORF2* (plus conservées). Ils ont mis en évidence que les L1 ont été acquis il y a 150 millions d'années et que les sous-familles L1PB1 à 4 et L1MA1 à 5, les plus anciennes de cette étude, se sont éteints il y a 70-74 millions d'années. Seuls les éléments L1PA1 à 17 sont actifs dans les temps modernes, depuis 46 millions d'années, où une famille active est remplacée par une autre suite à des modifications ayant lieu notamment au niveau de la région 5'UTR. Ils ont ainsi identifié 7 types différents d'extrémités 5'UTR, chaque changement correspondant à l'activation et la mobilisation d'une nouvelle famille. L'extrémité 5'UTR la plus récente a été remplacée dans la sous-famille L1PA8, il y a environ 40 millions d'années, et de manière concomitante une région de l'*ORF1* a été modifiée, ce qui caractérise une différence entre les éléments « récents » plus anciens (L1PA17 à 8) des éléments plus récents (L1PA7 à 1) (Khan, 2005). Seuls les L1 possédant une structure intègre vont pouvoir se mobiliser et se multiplier dans le génome. Aujourd'hui, la majorité des copies de L1 sont inactives suite à des mutations, des troncations en 5' ou bien des insertions d'autres éléments, notamment *Alu*, en leur sein ce qui les rend incompétents pour la rétrotransposition. Dans le génome humain, il existe 500 000 copies de L1 dont seulement 7 000 sont pleine taille (Khan,

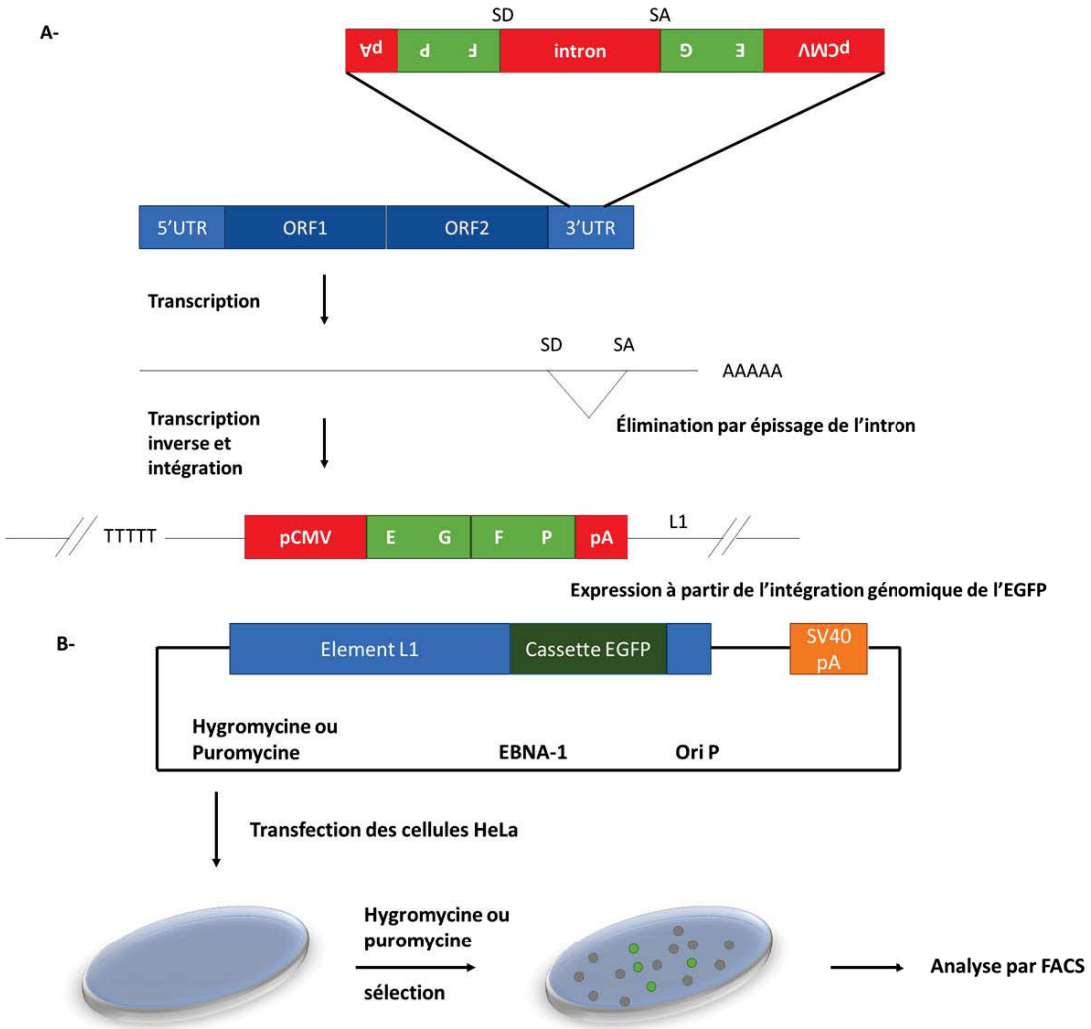


Figure 7 : Représentation schématique du test de rétrotransposition des L1 (retrotransposition assay) *(Ostertag* et al*., 2002)*

A- L'élément L1 consiste en deux extrémités non traduites en 5' et en 3' (5'UTR et 3'UTR) ainsi que deux cadres ouverts de lecture (ORF1 et ORF2). La cassette de rétrotransposition de l'EGFP est clonée dans la région 3'UTR du L1 en orientation antisens. Cette cassette consiste à mettre sous la dépendance du promoteur CMV (pCMV) le gène *EGFP* en antisens coupé par un intron de la *γ-globine* en orientation sens (intron) et un signal de polyadénylation (pA). Les sites donneur (SD) et accepteur (SA) d'épissage de l'intron sont indiqués. Les cellules exprimeront seulement l'*EGFP* quand un transcrit L1 contenant l'*EGFP* en antisens comme marqueur d'épissage sera rétrotranscrit et intégré à une nouvelle position chromosomique dans l'ADN. L'EGFP pourra alors être exprimée sous la dépendance du promoteur pCMV.

B- L'élément L1 marqué avec la cassette EGFP est cloné dans le vecteur d'expression mammifère pCEP, où est situé un signal de polyadénylation du virus SV40 (SV40pA) en aval du L1 marqué. Le vecteur pCEP est alors capable de se répliquer dans les cellules HeLa en utilisant l'origine de réplication des cellules eucaryotes (OriP/EBNA). Ces vecteurs peuvent contenir aussi bien le gène de résistance à l'hygromycine ou à la puromycine. Ces constructions contenant l'élément L1 marqué sont transfectées dans les cellules HeLa et une sélection *via* un antibiotique (hygromycine ou puromycine) est réalisée pendant 24h après la transfection. Après cette sélection, les cellules sont analysées par le FACS (Fluorescence Activated Cell Scanning) afin de quantifier les cellules exprimant l'EGFP.

2005; Lander *et al.*, 2001). Parmi ces 7 000 copies pleine taille, toutes ne sont pas compétentes pour la rétrotransposition. Par des analyses bio-informatiques sur la séquence du génome humain, Brouha et ses collaborateurs ont estimé qu'il y avait entre 80 et 100 L1 encore compétents pour la rétrotransposition. L'activité de ces éléments a été mesurée grâce à un test de rétrotransposition dans des cellules neurales progénitrices (NPC) qui ont été transfectées soit avec un $L1_{RP}$ servant de référence, soit avec un plasmide contenant le L1 à tester. Cette technique permet d'identifier dans une lignée cellulaire (NPC ou HeLa par exemple) si un L1 est capable de rétrotransposer. En effet, dans le plasmide qui est transfecté dans les cellules, le gène de l'EGFP en orientation antisens est coupé par une séquence intronique de la γ-globine en orientation sens flanquée de 2 sites accepteur et donneur d'épissage. L'ensemble se termine par une séquence polyA. Cette construction est insérée au niveau de la région 3'UTR du L1 à tester en orientation antisens (Figure 7). Les cellules ainsi transfectées exprimeront seulement l'EGFP quand un transcrit L1 contenant l'EGFP en antisens, comme marqueur d'épissage, sera rétrotranscrit et intégré à une nouvelle position chromosomique. Le FACS (Fluorescence-Activated Cell sorting) permet de quantifier la production d'EGFP dans les cellules transfectées. Ceci permet d'obtenir une activité par rapport à une référence, ici le $L1_{RP}$. L'activité de 82 L1 a été mesurée et 40 ont une activité supérieure d'au moins un tiers à celle du $L1_{RP}$, ce qui correspond à la définition d'un L1 actif ou « hot L1 ». Ils ont également mis en évidence que les éléments L1 jeunes ayant une petite divergence de séquences, sont généralement polymorphiques dans la population et sont actifs dans des cellules en culture. Contrairement aux L1 plus anciens dont les séquences sont hautement divergentes, qui sont fixés dans le génome et ne sont plus actifs. Ils ont pu classifier les L1 récents encore actifs dans le génome humain en classe du plus jeune au plus vieux : L1Ta-1d, L1Ta-1nd, L1Ta-0, L1pre-Ta, L1PA2 (Ta pour transcriptionnellement actif). Ensuite grâce à la séquence consensus de ces L1 actifs, une cartographie de ceux-ci à l'échelle du génome humain a été réalisée. Cette carte montre qu'il y a une distribution génomique similaire entre les L1 intacts actifs et non-actifs (Brouha *et al.*, 2003). Ces L1 actifs vont donc participer à l'évolution et à la diversité des génomes mais également à l'apparition de pathologies dues à des néo insertions.

2.2. LINE-1 et promoteur bidirectionnel.

L'expression des L1 dépend en grande partie de la structure de l'extrémité 5'UTR et de l'activité du promoteur interne. La nouvelle extrémité 5'UTR acquise dans les sous-familles récentes L1PA8 à 1 est une région complexe qui a fait l'objet de différentes études.

Tout d'abord, des analyses concernant le promoteur sens interne ont été réalisées afin d'identifier une zone critique nécessaire à son fonctionnement dans une lignée cellulaire de carcinome embryonnaire humain, les cellules NTera2D1. Ces cellules ont été transfectées avec un plasmide (p1LZ) contenant l'extrémité 5'UTR et le début de l'*ORF1* d'un L1HS qui a été fusionné avec le gène *LacZ*. Le gène *LacZ* est l'un des gènes de l'opéron lactose des bactéries, provenant ici d'*E. coli*, qui code pour l'enzyme β-galactosidase. L'expression du gène rapporteur *LacZ* est quantifiée dans les cellules transfectées par un test colorimétrique dosant l'activité de l'enzyme β-galactosidase. Différentes transfections de plasmides contenant des délétions de la région 5'UTR du L1HS ont permis l'identification de la région critique nécessaire au fonctionnement du promoteur sens localisée dans la région +1 à +101 (Swergold, 1990). Cette région contient le site de fixation pour le facteur de transcription YY1 qui est requis pour l'initiation de la transcription du L1 (Athanikar *et al.*, 2004).

Cette région 5'UTR contient non seulement un promoteur sens interne mais aussi un promoteur antisens, nommé ASP, qui a d'abord été mis en évidence par des études bio-informatiques sur des banques d'EST et d'ADNc. Après alignement sur la séquence consensus du L1HS, il a été observé que certains EST étaient orientés dans le sens inverse de la transcription sens. La région 5'UTR-*ORF1* d'un de ces L1HS a été clonée dans un plasmide contenant la luciférase comme gène rapporteur. Après avoir validé la présence d'une transcription provenant d'un promoteur antisens dans les cellules HeLa, des délétions ont été réalisées pour voir l'effet de celles-ci sur l'activité du promoteur ASP. La région critique nécessaire à l'activité de l'ASP identifiée est localisée dans la région +400 à +600. Dans cette région se trouve des sites pour la fixation de facteurs de transcription SOX et Sp1 qui pourraient avoir un rôle dans son activité (Speek, 2001). En 2006, une nouvelle étude caractérise cet ASP par des analyses de Northern blot sur des cellules HeLa transformées avec un plasmide contenant la séquence 5'UTR de $L1_{RP}$ en sens ou en antisens. Des expériences de 5'RACE montrent 2 régions majeures pour l'initiation de la transcription en antisens, une localisée de +378 à +431 et une de +480 à +497 (Figure 8). Les auteurs confirment également que la région critique, nécessaire à l'activité de l'ASP, se situe dans la région +400 à +600 (Yang & Kazazian, 2006). Par ailleurs, des analyses sur le rôle de ces facteurs sur l'activité des promoteurs ont montré que les facteurs de transcription YY1, SOX et RUNX3 seraient plus impliqués dans la régulation du promoteur sens (Swergold, 1990; Yang, 2003). Les facteurs RUNX3 et Sp1, quant à eux, seraient impliqués dans la régulation de l'ASP (Speek, 2001; Yang, 2003). En effet lorsque le 3ème site RUNX3 (+526 à +508) est muté, l'activité transcriptionnelle de l'ASP diminue d'au moins 5 à 10 fois par rapport au contrôle (Yang, 2003). Le facteur RUNX3 active chez l'Homme la transcription et la rétrotransposition

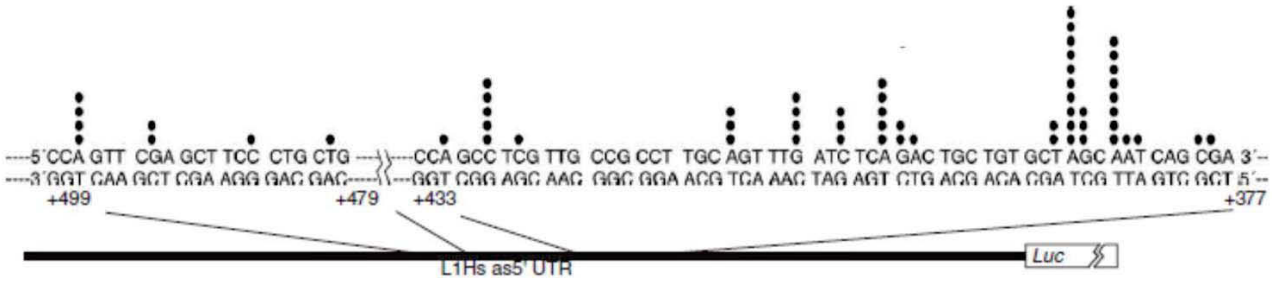


Figure 8 : Caractérisation des sites d'initiation de la transcription de l'ASP au niveau de la région 5'UTR de L1. *(Yang & Kazazian, 2006)*

Identification des différents sites d'initiation de la transcription de l'ASP par 5'RACE. Les points noirs indiquent les occurrences de l'initiation de la transcription à ce nucléotide.

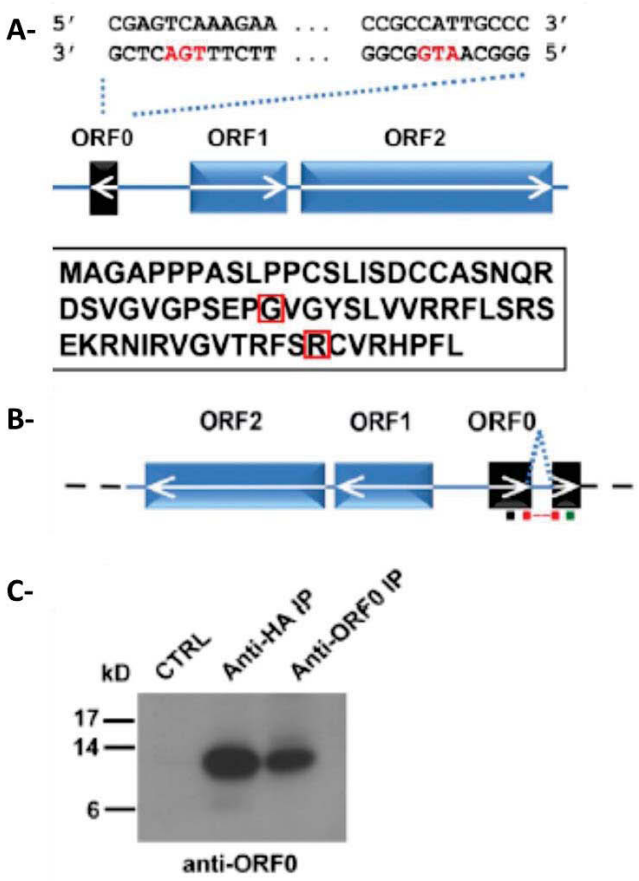


Figure 9 : Caractéristiques de l'ORF0. *(Denli* et al.*, 2015)*

A- Localisation de l'*ORF0* dans le promoteur de L1. Le codon d'initiation ATG et le codon stop TGA sont indiqués en rouge en orientation antisens. Les positions des sites donneurs d'épissage dans la séquence codante sont indiquées par les carrés rouges. La séquence consensus de la protéine ORF0 pleine taille est basée sur l'alignement de 781 loci potentiellement ORF0 du génome humain.

B- Description schématique du peptide ORF0 identifié. Le premier peptide identifié (rectangle noir) réside en aval du site SD2. Le second peptide (rectangle rouge) est formé grâce à l'épissage entre le site SD2 et un site SA1. Le troisième peptide (rectangle vert) est localisé en amont du site SA1 de la séquence L1.

C- Un anticorps fonctionnel dirigé contre l'ORF0 a été testé sur des extraits protéiques surexprimant l'ORF0, suivi d'une immunoprécipitation et d'un western blot.

du L1 par liaison au niveau de la région 5'UTR (+83 à +101). Il est possible que les 2 promoteurs, sens et antisens, soient régulés par le même facteur de transcription *via* des interactions à différents sites comme ce qui a été montré. Ces 2 sites sont très conservés chez l'Homme, des sous-familles L1PA8 à L1PA1, c'est-à-dire durant 40 millions d'années, suggérant ainsi leur fonction dans la régulation des L1 (Yang & Kazazian, 2006).

Cette transcription provenant de l'ASP de L1 peut influencer positivement l'expression en sens du L1 *via* le recrutement de la machinerie de transcription en favorisant la présence d'un état chromatinien ouvert ou *via* la formation d'ARN non-codant. D'autre part, l'expression de cet ARN antisens peut conduire à la formation d'ARN double brin impliqué dans les mécanismes d'ARN interférence qui sera détaillé dans le paragraphe 2.3 (Yang & Kazazian, 2006).

Par ailleurs, la présence d'un *ORF0* a été prédite par des modèles bio-informatiques puis caractérisée dans la région 5'UTR de L1 entre les nucléotides 452-236. Cet *ORF0,* dont l'initiation de la transcription se fait dans la région de l'ASP, est caractérisé par la présence d'une séquence Kozak permettant de promouvoir la traduction de la séquence en protéine. Cet *ORF0* est également caractérisé par la présence de 2 sites donneurs d'épissage SD1 et SD2 localisés respectivement aux positions +106 et +191 (Figure 9-A). Cet ORF0 code une petite protéine d'environ 10 KDa (Figure 9-B et 9-C). Chez l'Homme, une analyse bio-informatique sur le génome humain de référence avec la séquence consensus de l'*ORF0* montre que 3 528 loci possèdent un *ORF0* avec au moins un site donneur conservé. Parmi ces loci détectés, les sous-familles de L1 impliquées sont :

- L1PA1 et L1PA2 possédant l'*ORF0* avec les 2 sites donneurs d'épissage fonctionnels,
- L1PA3 et 4 possédant l'*ORF0* avec le 2ème site donneur conservé et le premier peut-être parfois muté,
- L1PA5 et 6 possèdent un *ORF0* avec le premier site donneur d'épissage qui est toujours muté et le deuxième intact.

En revanche, les sous-familles L1PA7 et L1PA8 ne possèdent pas d'*ORF0*.

Cet *ORF0* n'est pas essentiel pour l'activité du L1 mais semble impliqué dans l'augmentation de la mobilité des L1. En effet, un test de rétrotransposition de L1 dans des cellules HEK293T surexprimant la protéine *ORF0* montre une augmentation de 40% de la mobilité des L1 par rapport aux cellules contrôles. Cet *ORF0* peut également conduire à la formation de protéines chimères par épissage avec les régions géniques codantes adjacentes (Denli *et al.*, 2015).

Il a donc été démontré fonctionnellement dans la sous-famille la plus récente, L1HS, que cette région 5'UTR contient un promoteur bidirectionnel. Les alignements de séquences permettent

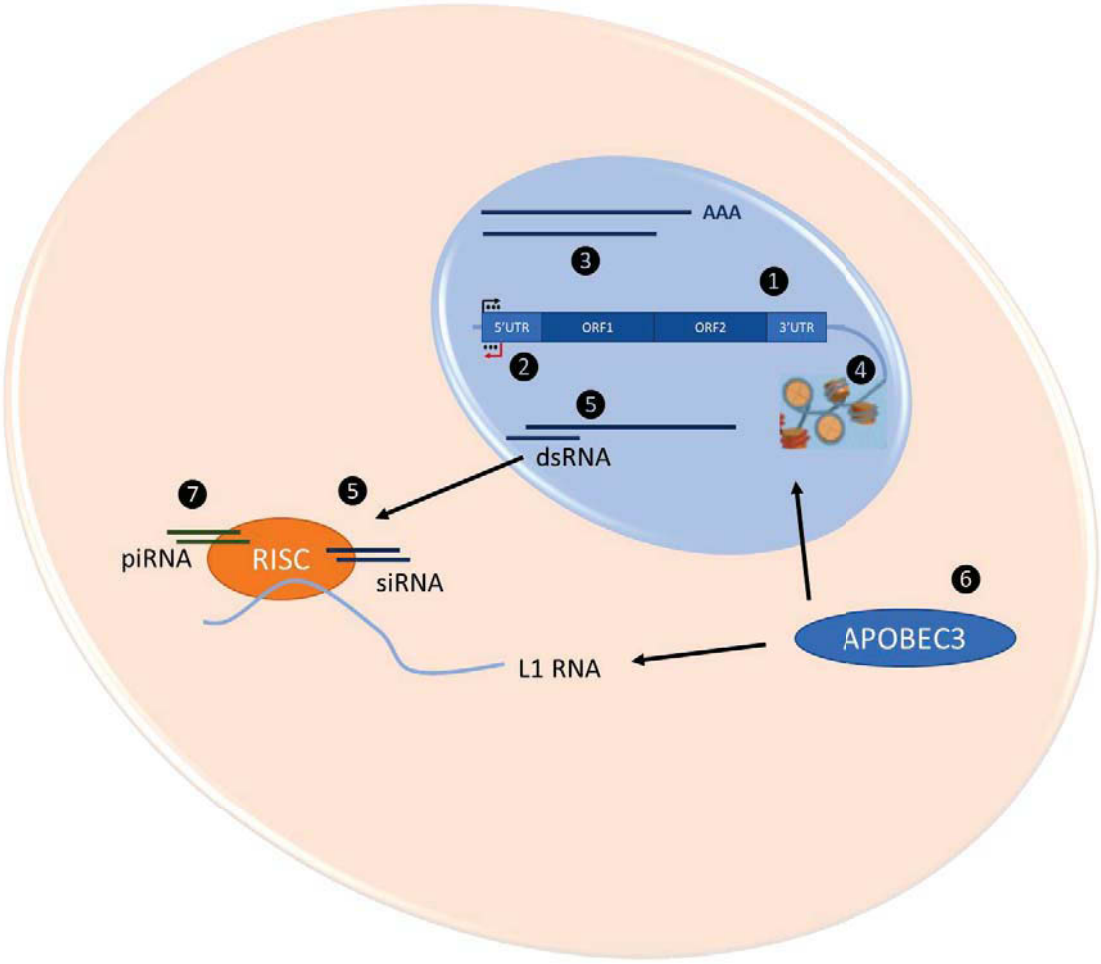


Figure 10 : Schéma récapitulatif des différents modes de régulation des L1

(1) Troncations, mutations, réarrangements inactivant le L1.
(2) La méthylation ADN de la région 5'UTR du L1 diminue son expression dans les cellules.
(3) La terminaison prématurée de la transcription du L1 ou son épissage inhibe son expression.
(4) L'hétérochromatisation des L1 et des éléments *Alu* supprime leur expression.
(5) La production d'ARN double brin par la transcription à partir des promoteurs sens et antisens (ASP) peut inhiber la rétrotransposition des L1 par un mécanisme d'ARN interférence, via la production de siRNA.
(6) De nombreuses protéines APOBEC3 réduisent la rétrotransposition des L1 dans des cultures cellulaires par un mécanisme inconnu qui n'implique pas la déamination des cytosines. Leurs rôles et leurs effets dans le noyau ou dans le cytoplasme ne sont pas encore très clairs.
(7) Rôle des piRNA dans la répression germinale des L1.

(Goodier & Kazazian, 2008)

d'estimer que cet ASP aurait été acquis au cours de l'évolution et conservé des sous-familles L1PA1 à 6 (Khan, 2005; Speek, 2001). En 2011, Macia et ses collaborateurs ont étudié la transcription en antisens des L1 à partir de promoteurs provenant des sous-familles L1PA1 à 10 dans des cellules souches embryonnaires humaines. Ils démontrent *via* des analyses de transfection de ces cellules par des séquences contenant la région 5'UTR de L1 en antisens, qu'un ASP fonctionnel serait présent dans les sous-familles L1PA1 à 6 mais aussi dans les sous-familles L1PA7, 8 et 10 qui sont plus anciennes (Macia *et al.*, 2011). Dans les cellules transfectées avec des constructions plasmidiques contenant le promoteur sens et antisens d'un L1HS, le promoteur sens est 3 à 13 fois plus actif que l'ASP pour toutes les lignées cellulaires testées (H13B, H7, 2102Ep) (Macia *et al.*, 2011). Cette transcription à partir de l'ASP permet la formation d'un ARN interférent qui réduit la stabilité des ARN de L1 *in vivo* (Yang & Kazazian, 2006). Cet ASP peut servir de signal d'autorégulation grâce à cette différence d'activité entre les 2 promoteurs ou participer à la rétrotransposition du L1 (Macia *et al.*, 2011). Les différents mécanismes permettant une régulation fine de ce promoteur bidirectionnel sont détaillés dans la partie 2.3.

2.3. Mécanismes de répression des LINE-1.

Sachant que les ET peuvent affecter le génome de bien des manières, il est nécessaire aux cellules de développer différents moyens de réguler leur activité. La question se pose donc de savoir comment sont régulés ces ET dans les cellules somatiques et germinales pouvant expliquer que ceux-ci ne perturbent pas l'intégrité et la stabilité du génome. Dans cette partie, je m'intéresserai aux mécanismes de régulation des L1 chez l'Homme (Figure 10).

2.3.1. Troncations, mutations et réarrangements inactivant les LINE-1.

De par le mécanisme de rétrotransposition des L1 nommé TPRT et détaillé dans ma partie 1.1.2 (Figures 3 et 4), les L1 peuvent être inactivés par la création de troncations, mutations ou réarrangements.

En effet, au cours de la rétrotransposition des L1, la reconnaissance par l'EN de la séquence d'ADN à cliver pour la néo insertion du L1 va libérer une extrémité 3'-OH qui servira d'amorce à la transcription inverse de l'ARN du L1 par la RT. Or des endonucléases de la cellule peuvent reconnaître cette extrémité 3'-OH et dégrader une partie de l'ADN au niveau du site de coupure avant l'intégration de l'ADNc du L1 ; ceci induit une modification de l'environnement génomique adjacent au L1 pouvant modifier l'expression du gène associé ou la séquence

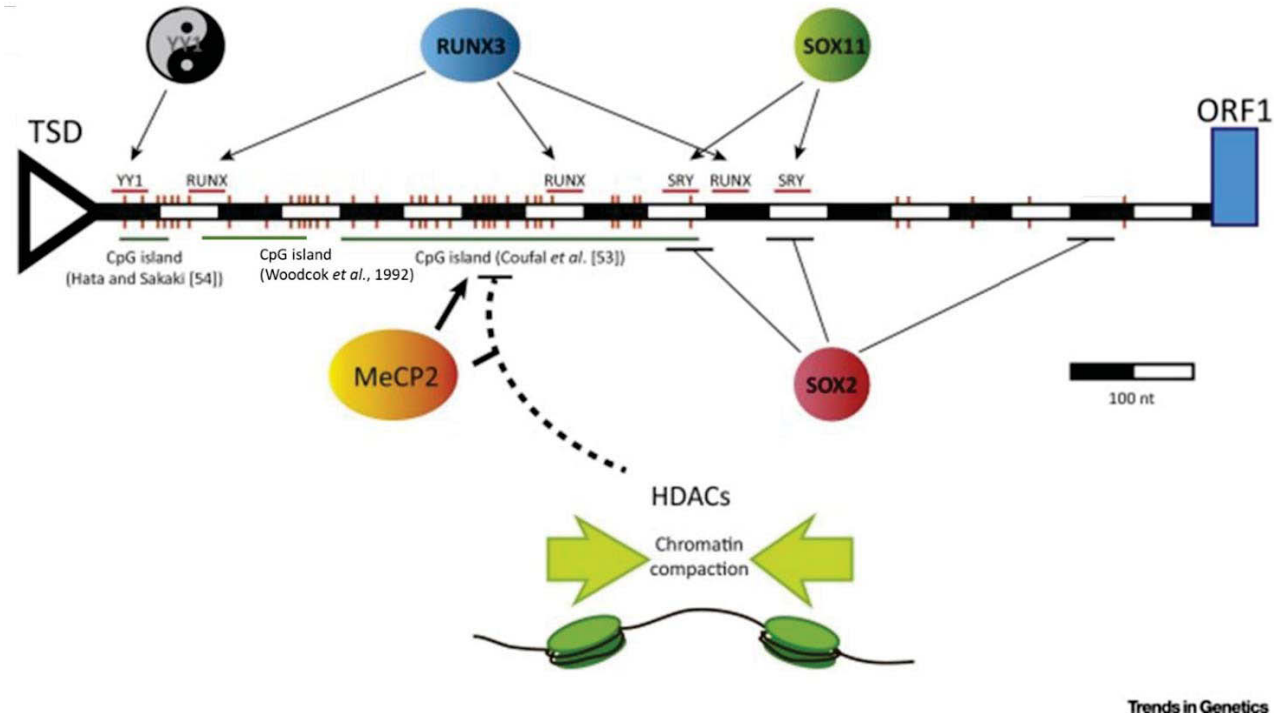


Figure 11 : Structure de la région 5'UTR des L1.

La région riche en dinucléotides CpG est représentée par des barres rouges. Cette région a été étudiée par différents auteurs. Hata et Sakaki ont étudiés la région +1 à +100, Woodcok et ses collaborateurs ont étudiés la région +161 à +210, Coufal et ses collaborateurs ont étudiés la région +232 à +491. Coufal et ses collaborateurs montrent également l'implication de MeCP2 et des protéines HDAC dans la formation d'une structure chromatinienne fermée nécessaire à la régulation négative de l'expression des L1. Les facteurs de transcription activateurs (YY1, RUNX3, SOX11) et les facteurs répresseurs (SOX2) sont représentés sur le schéma ainsi que les sites de liaison pour ces facteurs. Le site de fixation du facteur YY1 est localisé de +13 à +21 ; les sites de fixation pour le facteur RUNX3 sont localisés de +83 à +101, de +389 à +407, de +508 à +526 ; les sites de fixation pour les facteurs SOX sont localisés de +472 à +477, de +572 à +577. (Faulkner & Garcia-Perez, 2017)

intergénique d'ADN. De plus, au moment de la transcription inverse de l'ARN du L1 en ADNc, la RT peut faire des erreurs et ainsi induire des mutations. Ces mutations pourront aboutir :

- Soit à la formation d'un signal de terminaison prématurée de la transcription qui induira à sa prochaine rétrotransposition l'intégration d'un L1 tronqué à une nouvelle position génomique.
- Soit à la formation d'un codon stop induisant la terminaison prématurée de la traduction. Ce L1 deviendra alors inactif car il ne pourra pas être traduit intégralement et ne pourra donc pas rétrotransposer.

Le L1 peut devenir inactif si au sein de sa séquence un autre ET, notamment un élément *Alu*, s'insère en son sein induisant ainsi une modification de la séquence codante du L1 qui ne sera plus traduite dans son intégralité et ne pourra donc plus rétrotransposer (Goodier & Kazazian, 2008; Viollet *et al.*, 2014).

2.3.2. La méthylation ADN des LINE-1.

C'est le premier mécanisme de la répression des L1 qui a été mis en évidence. La méthylation ADN a lieu sur les bases cytosines dans un contexte de dinucléotides CpG et consiste en l'ajout d'un groupement méthyl (-CH3) sur le carbone en position 5 du cycle aromatique de la cytosine. Cette méthylation ADN est assurée par des enzymes nommées les DNMT (méthyltransférases ADN). La région promotrice des L1 est une région riche en dinucléotides CpG (Figure 11) auxquels différentes études se sont intéressées (Coufal *et al.*, 2009; Hata & Sakaki, 1997; Woodcock *et al.*, 1997). Le rôle de la méthylation dans la répression de l'expression des L1 a été mis en évidence par la reconnaissance des cytosines méthylées dans la région 5'UTR. Une première étude s'est intéressée aux dinucléotides CpG des positions +161 à +210 de la région 5'UTR des L1 grâce à un traitement au bisulfite et au séquençage de différents clones. Par des expériences de transfection de plasmides contenant un élément L1 dont la région promotrice en 5' a été traitée par l'agent déméthylant 5-azacytidine (5-aza) dans des cellules de fibroblastes embryonnaires humains, l'amplification PCR de ces régions montre une augmentation de l'expression des L1. Au cours de cette étude, ils observent une forte méthylation au niveau de cette région 5'UTR associée à une hémi méthylation dans la région +178 à +198 (Woodcock *et al.*, 1997). Par ailleurs, une autre étude s'est intéressée aux dinucléotides CpG au niveau de la région 5'UTR du L1. Grâce à des constructions plasmidiques, à la transfection de cellules HeLa et à la quantification d'un gène rapporteur (*CAT*), les auteurs ont d'abord identifié que la région se situant des positions +1 à +155 était importante dans la régulation de la transcription des L1

via la méthylation ADN. Selon la même méthode et par des mutations des sites CpG présents dans cette région, les auteurs identifient 4 sites CpG aux positions +52, +58, +61 et +70 dont la méthylation est suffisante et nécessaire à la répression transcriptionnelle de L1 (Hata & Sakaki, 1997). Ils émettent l'hypothèse que 3 mécanismes différents peuvent intervenir dans la répression transcriptionnelle *via* la méthylation des CpG :

1- La méthylation des CpG va bloquer la liaison de facteurs de transcription. Or grâce à des expériences de mobilité sur gel, ils mettent en évidence que la méthylation de la région promotrice du L1 ne va pas empêcher la fixation du facteur de transcription YY1 (Figure 11) important pour la transcription du L1. Peut-être que la répression de la transcription des L1 est due à un autre facteur qui reste à identifier.
2- La méthylation des CpG peut changer la conformation de la chromatine pour induire un état transcriptionnellement inactif. Avec leur modèle, ils n'ont pas pu prouver cette hypothèse car la structure des plasmides est trop courte pour permettre ce type de répression.
3- La méthylation des CpG peut inhiber la transcription *via* la liaison de protéines de reconnaissance des cytosines méthylées comme MMBP-1,2 qui reconnait une cytosine dans une séquence spécifique ou comme MeCP-1,2 et MDBP2 qui n'ont pas de séquences spécifiques de reconnaissance (Hata & Sakaki, 1997).

Une étude plus récente utilisant une conversion au bisulfite de l'ADN dérivé de tissus de cerveau et de peau montre que dans la région +232 à +491 de la région promotrice du L1, les 20 sites CpG de cette région ont une grande variation de la méthylation. Certains sont non méthylés dans le cerveau où une transcription active des L1 est observée. Par la suite des expériences de ChIP ont été réalisées pour savoir si l'affinité de la liaison des facteurs de transcription SOX2 et MeCP2, dont les sites de liaisons sont localisés dans cette région (Figure 11), sont modifiés par la méthylation de ces cytosines. Les résultats leur permettent d'émettre l'hypothèse que la diminution de la méthylation du promoteur de L1 dans le cerveau en développement corrèle avec l'augmentation de la transcription des L1 et peut-être avec la rétrotransposition des L1. De plus, l'interaction différentielle de SOX2 et de MeCP2 avec les séquences régulatrices du L1 peut moduler négativement l'activité du L1 dans différents types cellulaires neuronaux (Coufal *et al.*, 2009).

Ceci suggère que différents éléments et évènements vont influencer, en plus de la méthylation de la région promotrice du L1, l'expression de celui-ci. Au final, cette méthylation de la région

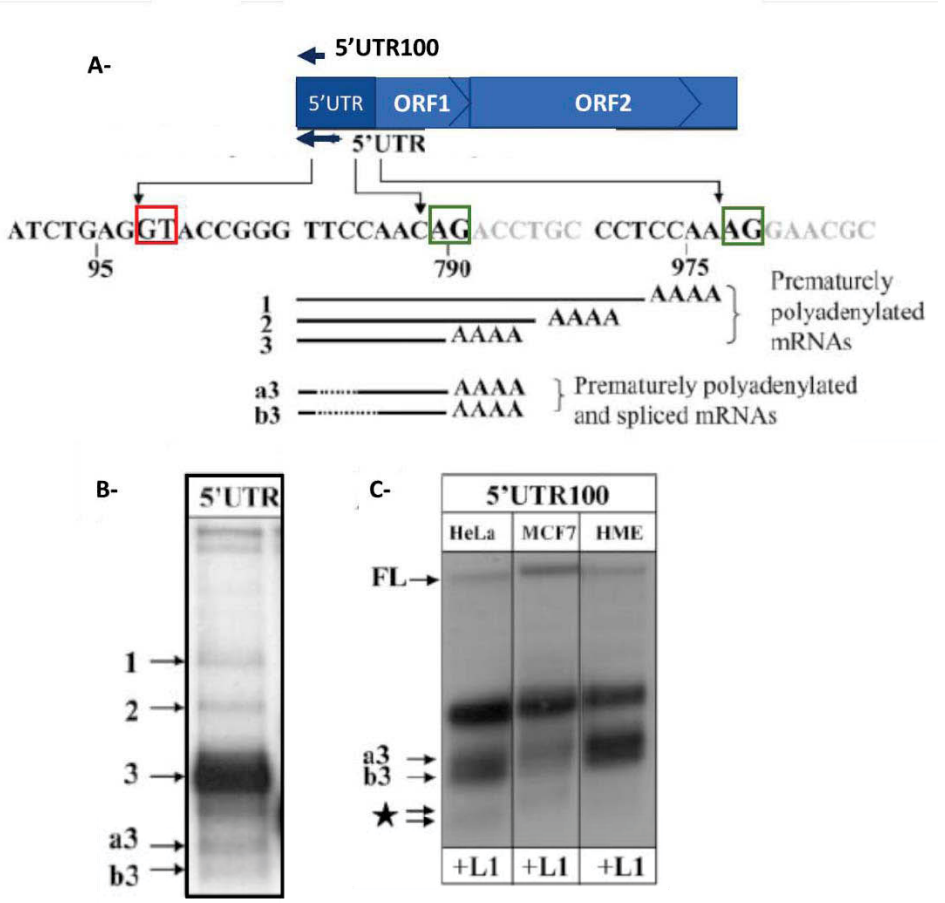


Figure 12 : De nombreux sites d'épissages sont identifiés dans la région 5'UTR de L1 et contribuent à la production de variants d'épissage.

A- Représentation schématique de l'extrémité 5'UTR, de l'ORF1 et de l'ORF2 du L1.3. La sonde brin spécifique utilisée pour le northern blot est positionnée au niveau de la région 5'UTR et indiqué par le flèche bleu foncé, nommée 5'UTR et détectant les 900 premières paires de bases du L1. En dessous est reportée la séquence nucléotidique d'une partie de cette région 5'UTR au niveau du site donneur d'épissage principal identifié à la position +97 (encadré rouge) et les sites de polyadénylation prématurée aux positions +789 et +977 (encadrés verts). Les transcrits provenant de la terminaison prématurée de la transcription *via* la présence d'un signal de polyadénylation sont schématiquement représentés par les numéros 1 à 3. Les transcrits prématurément polyadénylés et épissés sont schématiquement présentés par a3 et b3.

B- Northern blot réalisé sur des ARN polyadénylés extraits des cellules NIH3T3 transfectées avec le vecteur contenant le L1.3. Les 5 formes de transcrits prématurément polyadénylés et/ou épissés sont identifiés avec la sonde 5'UTR.

C- Northern blot réalisé sur les ARN polyadénylés extraits de cellules HeLa, MCF7 et HME transfectées avec le plasmide contenant le L1.3. Le transcrit L1 pleine taille est retrouvé ainsi que les 2 formes épissées et prématurément polyadénylées a3 et b3. (Belancio, 2006)

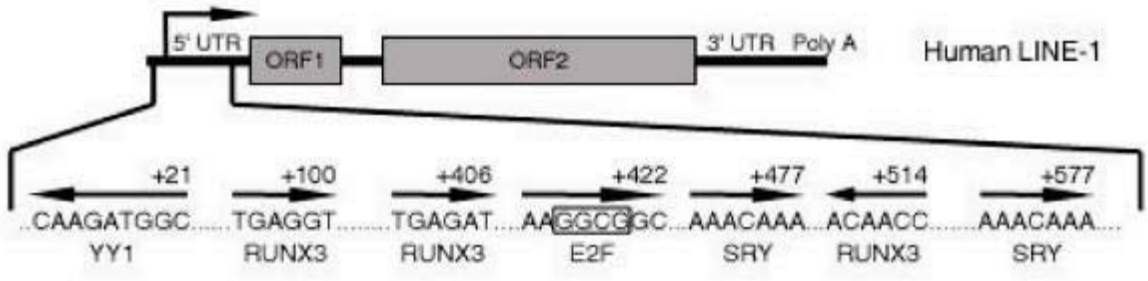


Figure 13 : Représentation schématique d'un L1 humain.

Sur cette représentation sont localisés les sites putatifs connus pour la fixation des facteurs de transcription YY1, RUNX3, E2F et SRY. Les flèches indiquent la direction de la séquence ADN par rapport à la région promotrice. Les séquences consensus de ces sites de fixation sont indiquées et les nombres représentent la localisation du dernier nucléotide de la séquence de reconnaissance par rapport au site +1 d'initiation de la transcription du L1.(Montoya-Durango *et al.*, 2009)

promotrice des L1 n'est plus considérée comme le seul mécanisme responsable de la répression des L1 mais associée à d'autres évènements de régulation.

2.3.3. La terminaison prématurée et les sites d'épissage modulent la transcription des L1.

Par analyse *in silico*, il a été mis en évidence que les L1 contiennent de nombreux sites donneur (SD) et accepteur (SA) d'épissage dans leur séquence en sens et en antisens (Figure 12-A) (Belancio, 2006). Différents transcrits produits à partir des L1 ont été détectés par Northern blot sur cellules de fibroblastes NIH3T3 mais également des cellules humaines HeLa, MCF7 et HME (Figure 12-B et 12-C). Ceci a permis la mise en évidence de sites fonctionnels d'épissage, dont le site donneur (SD) le plus utilisé lors du processus d'épissage est localisé à la position +97 (Figure 12-A). De plus, des études rapportent qu'il existe une compétition entre les différents sites d'épissage et les signaux de polyadénylation. Les auteurs estiment que le nombre de transcrits épissés est équivalent au nombre de transcrits prématurément stoppés. Entre la terminaison prématurée due à un signal de polyadénylation et un évènement d'épissage, différents ARNm sont produits. Ceux-ci peuvent être traduits en protéines ORF1p ou ORF2p, mais aussi produire des versions tronquées de ces protéines. La protéine ORF2p seule ne va pas être suffisante à la rétrotransposition des L1 mais va pouvoir être utilisée pour la rétrotransposition des éléments *Alu* (Belancio, 2006). Par ailleurs, les auteurs mettent en évidence la présence d'un transcrit variant de l'ORF2 qu'ils nomment *SpORF2*. Ce transcrit a une expression tissu-spécifique, principalement retrouvée dans les testicules et le muscle cardiaque. Grâce à un test de rétrotransposition ciblant des éléments *Alu*, ce transcrit est capable de mobiliser ces éléments dans des cellules HeLa. De plus, l'expression de ce transcrit *SpORF2* va induire des dommages à l'ADN des cellules dans lesquelles il s'exprime. En effet, un test COMET, qui détecte les cassures doubles brins de l'ADN, a été réalisé sur des cellules HeLa exprimant ce transcrit. Au final, ces expériences mettent en évidence que ce transcrit variant *SpORF2* va, dans la lignée germinale, favoriser la rétrotransposition des éléments *Alu* et ainsi participer à la diversité des génomes. Mais dans les cellules somatiques, ce transcrit va déstabiliser le génome en induisant des cassures doubles brin de l'ADN contribuant ainsi au développement de maladies, comme des cancers (Belancio *et al.*, 2010).

Ces deux mécanismes vont contribuer à limiter la rétrotransposition des L1 puisque 1) les protéines nécessaires ne pourront pas être produites et 2) de par la préférence en *cis* des L1, le mécanisme de transposition ne pourra aboutir (Belancio *et al.*, 2010). Ces évènements

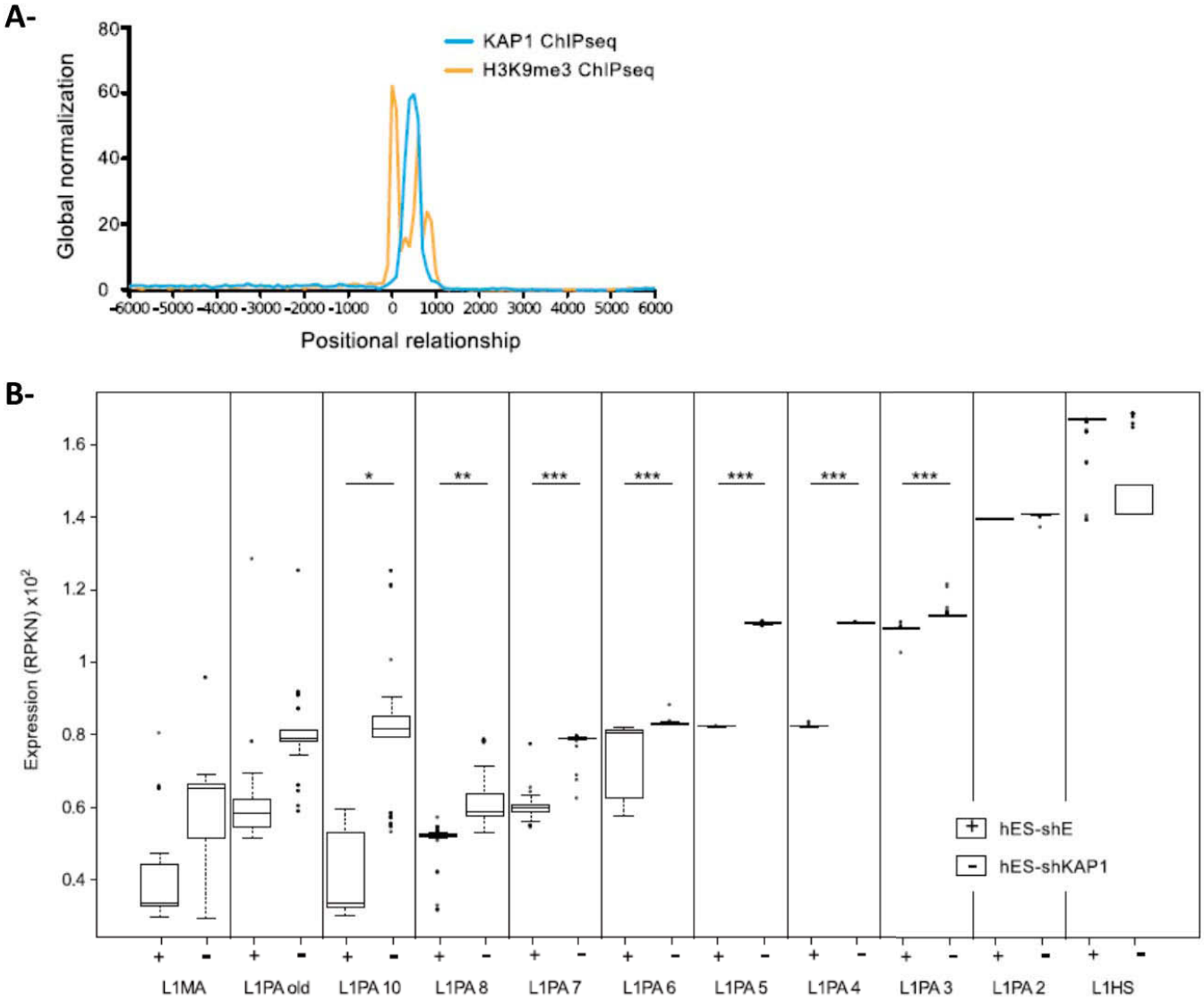


Figure 14 : Hétérochromatisation des L1.

A- La fixation de KAP1 coïncide avec la présence de la marque d'histone répressive H3K9me3 au niveau de la région 5'UTR d'un L1 pleine taille dans les cellules ES. Le chevauchement des pics de détection de KAP1 et de H3K9me3 est obtenu par observation des données de ChIP-seq de cellules ES au niveau de la région 5'UTR de L1 pleine taille.

B- Comparaison de l'expression de L1 pleine taille dans des cellules ES contrôles (hES-shE) par rapport à des cellules ES sous-exprimant KAP1 (hES-shKAP1). Chaque sous-famille de L1 est analysée séparément de la plus ancienne à la plus récente. La catégorie « L1MA » correspond aux sous-familles L1MA1 à 9. La catégorie « L1PA old » correspond aux sous-familles L1PA11 à 17. L'expression correspond aux valeurs de RPKN des RNA-seq sur les ARN totaux extraits des cellules hES-shE et hES-shKAP1 et les p-value ont été calculées avec le test non paramétrique de Wilcoxon. (Castro-Diaz *et al.*, 2014)

d'épissage peuvent également compromettre l'expression normale de gène ou leur épissage en induisant la formation de transcrits instables. Parmi les scénarios possibles, l'interférence du L1 avec l'expression d'un gène en incluant un de ses exons est possible *via* un épissage entre les sites SD d'un L1 intronique ou d'un L1 intergénique et un site SA d'épissage localisé dans un gène. Par exemple, un évènement d'épissage entre un L1 et un gène codant pour un récepteur aux estrogènes produit un transcrit tumeur spécifique codant une protéine ne possédant pas de domaine de liaison aux hormones (Dotzlaw *et al.*, 1992). Ainsi ce type d'épissage, en plus de moduler l'expression des L1, peut avoir des effets négatifs majeurs pour les cellules.

2.3.4. Hétérochromatisation des LINE-1.

Ce phénomène est induit par le recrutement de protéines impliquées dans le remodelage de la chromatine. Sachant que les L1 contiennent des dinucléotides CpG dans leur région promotrice, le complexe E2F/Rb, qui se lie aux régions riches en CpG du génome avec une haute affinité, pourrait se lier à la région promotrice L1 et ainsi moduler leur activité. En effet, l'extrémité 5'UTR des L1 contient des sites de liaison pour les protéines E2F et Rb (Figure 13). Par des analyses de ChIP sur des cellules humaines (cellules épithéliales, U2OS et HeLa), il a été montré que les protéines Rb, E2F1 et E2F4 se lient aux L1. Des cellules n'exprimant pas les protéines Rb (Rb null cell) présentent une augmentation de l'expression des L1. Les protéines Rb recrutent des histones désacétylases (HDAC1, HDAC2) ainsi que des histones méthyltransférases, notamment Suv39h1 et Suv39h2 impliquées dans la déposition de la marque H3K9me3 et des protéines Suv4-20h1 et Suv4-20h2 impliquées dans la déposition de la marque H4K20me3. Ces dépositions de marques d'histones répressives associées à la méthylation ADN de la région promotrice vont participer à la formation d'une structure chromatinienne fermée nommée hétérochromatine (Montoya-Durango *et al.*, 2009).

Les protéines E2F ont peu d'effet sur l'activité des L1. En revanche, ces protéines interagissent avec le récepteur AHR qui est impliqué dans la régulation des L1 en cas de stress des cellules (Montoya-Durango *et al.*, 2009).

D'autres protéines peuvent être impliquées dans l'hétérochromatisation des L1. En effet, des expériences de ChIP sur des cellules souches embryonnaires humaines (hES) montrent un enrichissement de la protéine KAP1 et de la marque d'histone H3K9me3 au niveau de la région promotrice en 5' des L1 (Figure 14-A). Quand l'expression des L1 est quantifiée par RNA-seq dans des cellules hES transformées avec un shARN contrôle (shE) et dans des cellules hES transformées avec un shARN dirigé contre le transcrit *KAP1* aboutissant à la diminution de

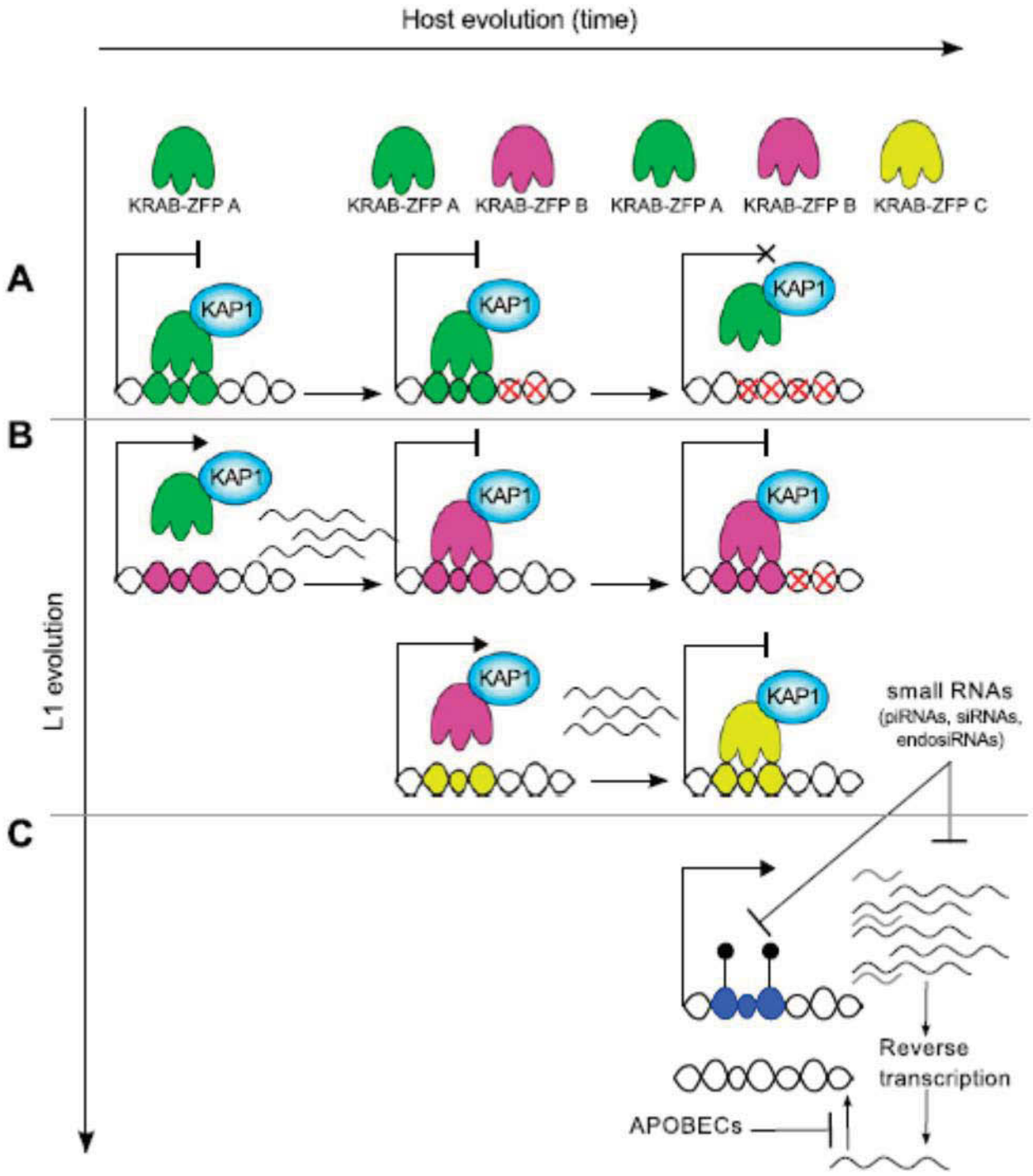


Figure 15 : Modèle de l'évolution du contrôle dynamique des L1.

(A) Les L1 très anciens possèdent une faible reconnaissance par le système KRAB/KAP1 et ont depuis accumulé des mutations (croix rouges). Ces mutations abolissent la possibilité de liaison de KRAB/KAP1 mais aussi leur capacité de transcription.

(B) Les sous-familles récentes recrutent KAP1 grâce à une séquence spécifique de reconnaissance par KRAB-ZFP mais il peut y avoir également des mutations permettant de dompter leur expression basale.

(C) Les éléments L1 les plus jeunes sont hautement transcrits et ne sont pas encore reconnus par aucun KRAB-ZFP mais produisent des petits ARN comme les piARN qui peuvent diminuer leur expression via le recrutement de protéines permettant la méthylation ADN et bloquer leur rétrotransposition grâce aux membres de la famille APOBEC3.

l'expression de la protéine KAP1 (shKAP1), une augmentation significative de l'expression est observée pour les L1 des sous-familles L1PA10 à L1PA3, suggérant ainsi une régulation par liaison du complexe KRAB-ZFP (Figure 14-B). Les différentes expériences menées ont permis d'aboutir à un modèle évolutif (Figure 15). Les L1 anciens (ayant plus de 26.8 millions d'années) peuvent être reconnus pour certains par le système KRAB/KAP1. Certains ayant accumulés trop de mutations dans leur séquence ne vont plus pouvoir être régulés par ce système de par l'absence de site de reconnaissance mais ces mutations vont également empêcher leur transcription. Les sous-familles les plus récentes recrutent le complexe KRAB/KAP1 au niveau d'un site de reconnaissance. Ce site peut accumuler des mutations qui ne vont pas gêner la liaison du complexe pour autant. Ce complexe va ainsi induire la répression de leur transcription. La liaison de ce complexe va permettre le recrutement d'histones méthyltransférases permettant la déposition de la marque H3K9me3 ainsi que la protéine HP1 et des ADN méthyltransférases. L'ensemble de ces protéines vont permettre la formation d'une structure hétérochromatinienne et ainsi réprimer la transcription des L1. Pour les L1 les plus récents, qui sont fortement transcrits (L1HS et +/- L1PA2) et qui ne sont pas encore reconnus par les protéines du système KRAB/ZFP, ceux-ci vont produire des petits ARN comme les piARN et leur rétrotransposition sera bloquée par les protéines APOBEC (Castro-Diaz *et al.*, 2014). Ces 2 derniers mécanismes seront détaillés dans les parties 2.3.5 et 2.3.6.

Grâce à la déposition de marques d'histones répressives et/ou activatrices, la transcription du L1 pourra être finement régulée pour soit empêcher sa transcription et ainsi être associée à de l'hétérochromatine ou bien, en réponse à une contrainte de l'environnement, au cours du développement ou de stress, ces éléments pourront être mobilisés *via* une structure chromatinienne plus permissive autorisant l'accès aux différents facteurs modulant leur activité.

2.3.5. Mécanisme d'ARN interférence.

Les mécanismes d'ARN interférence sont connus dans la lutte antivirale mais jouent également un rôle dans la régulation de l'expression des gènes. Suite à l'identification de transcrits initiés à l'ASP de L1 (Speek, 2001), des études se sont intéressées à l'implication de ces transcrits dans la régulation de l'expression des L1 *via* un mécanisme d'ARN interférence (Yang & Kazazian, 2006). Des ARN double brin sont produits à partir de séquences L1 et peuvent servir de mécanisme endogène pour la mise sous silence de la transcription des L1 et *in fine* de la rétrotransposition. Des expériences de northern blot sur des cellules humaines en utilisant des ribosondes dirigées contre les 500 premières paires de bases de la région 5'UTR de L1 en sens

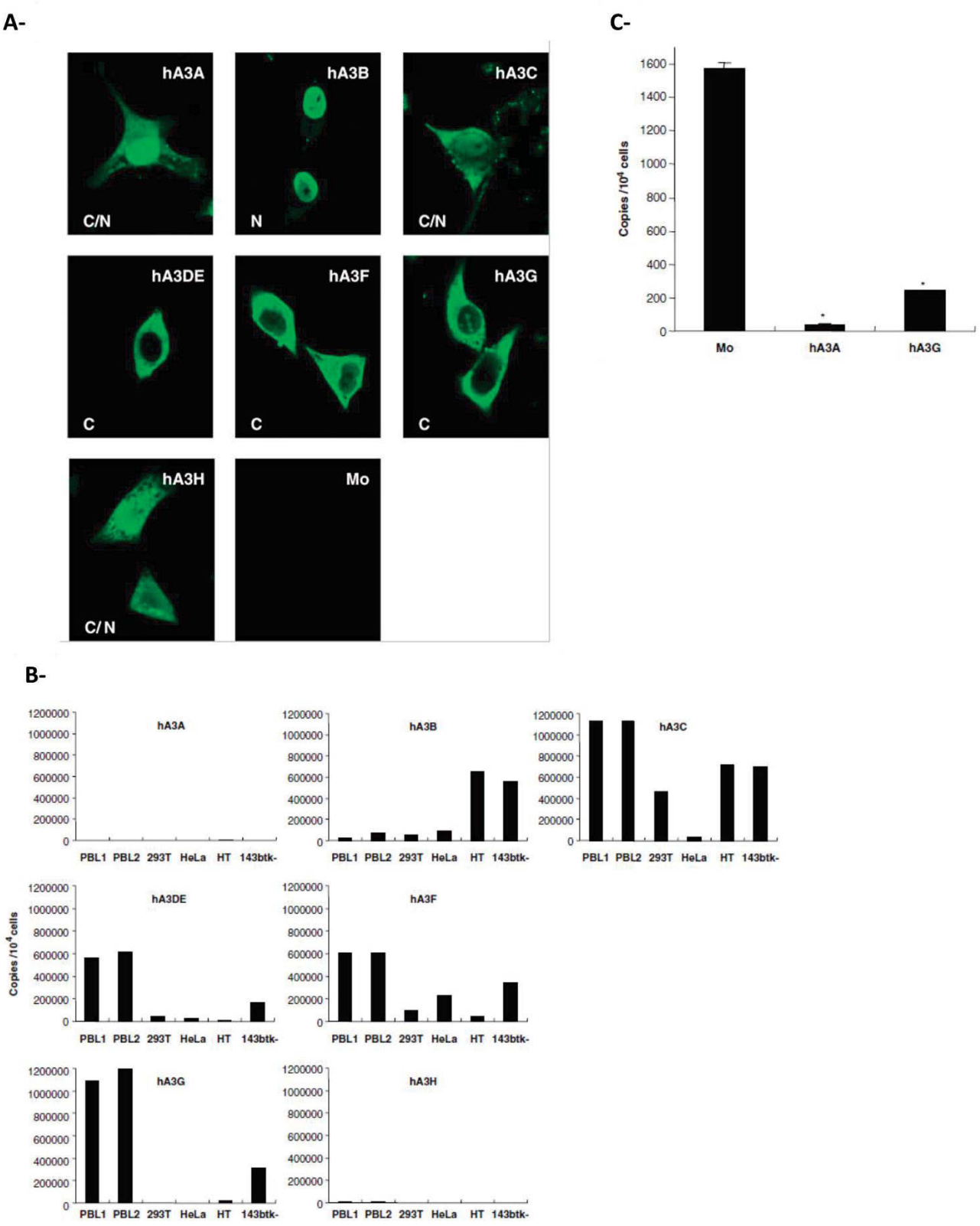


Figure 16 : Répression de la rétrotransposition par les protéines APOBEC3 (hA3).

A- Analyse de la localisation subcellulaire des protéines hA3 par microscopie en immunofluorescence en utilisant des cellules HeLa transfectées avec un plasmide d'expression hA3 taggé avec HA. L'anticorps monoclonal anti-HA et un anticorps Alexa488 conjugé à l'IgG anti-souris sont utilisés comme 1er et 2nd anticorps respectivement. Les lettres C, N ou C/N sur les figures indiquent la localisation cytoplasmique, nucléaire ou les deux.

B- Expression endogène des ARNm de hA3 des des cellules primaires ou des lignées cellulaires. Les ARN totaux sont extraits des cellules PBL (de 2 donneurs différents), 293T, HeLa, HT1080 et 143BTK-. Une RT-qPCR est réalisée pour déterminer le nombre de copies respectives d'ARNm hA3. Les niveaux d'ARNm hA3 sont normalisés par rapport aux niveaux d'ARNm de GAPDH.

C- PCR quantitative ciblant l'EGFP épissée. L'ADN total est extrait des cellules 293T co-transfectées avec le vecteur de rétrotransposition (contenant le L1 et l'EGFP à épisser) et un vecteur d'expression des hA3. Une PCR quantitative est réalisée. Les amorces utilisées permettent de détecter l'EGFP épissé. Le contrôle (Mock), hA3A et hA3G sont représentatives. *(Kinomoto* et al*., 2007)*

ou en antisens permettent la détection de nombreux siARN d'environ 21 nucléotides et de transcrits L1 dégradés dans les deux sens. Les siARN sont fortement détectés dans les cellules embryonnaires de reins (HEK293) alors que dans les cellules HeLa, il n'y a pas ou très peu de détection. Cet écart entre les abondances des siARN sens et antisens est probablement dû à une différence de clivage du siARN par la protéine Argonaute-2 du complexe RISC dans les cellules HeLa. A cause de l'expression à peine détectable des siARN provenant des L1 dans les cellules HeLa, ces cellules sont donc un bon modèle pour des expériences de transfection de siARN spécifiquement dirigés contre les L1. Aussi un test de rétrotransposition dans des cellules HeLa met en évidence que cette ARN interférence induit une diminution de 50% de la capacité de rétrotransposition des L1. La production des siARN de L1 est induite par la voie de biogenèse médiée par Dicer1. En effet, dans des cellules HEK293 où une diminution conditionnelle de la protéine Dicer1 (knockdown) a été réalisée, les L1 sont capables de rétrotransposer malgré la présence d'un plasmide transfecté dans ces mêmes cellules contenant la séquence de L1 en antisens. Ceci suggère que la production de siARN ne peut avoir lieu en absence de Dicer1 et que le mécanisme d'ARN interférence est donc déficient (Yang & Kazazian, 2006). Ce mécanisme d'ARN interférence n'est initié qu'à partir d'éléments pleine taille non tronqués dans la région 5' avec possiblement une préférence en *cis* à l'image de ce qui a été décrit dans la littérature pour le mécanisme de rétrotransposition qui va préférentiellement impliquer les protéines et ARN codés par sa propre séquence afin de former les RNP nécessaires à la rétrotransposition du L1. La vaste majorité des L1 présents dans le génome humain sont tronqués en 5'. Cette perte de la région 5'UTR les rend incapables de rétrotransposer. Seuls les L1 pleine taille possédant une région 5'UTR intacte et étant compétents pour la rétrotransposition doivent être contrôlés. Par conséquent, le mécanisme d'ARN interférence ciblant exclusivement les L1 pleine taille est un mécanisme de contrôle hautement économique qui a été probablement sélectionné au cours de l'évolution. Il a été proposé que la production de transcrits à partir de l'ASP de L1 peut induire une ARN interférence des gènes adjacents si un L1 pleine taille est inséré dans un gène (Speek, 2001; Yang & Kazazian, 2006).

2.3.6. Régulation de la rétrotransposition via les protéines APOBEC3.

Chez l'Homme, la famille des protéines APOBEC3 est composée de 7 membres (APOBEC3A, B, C, DE, F, G et H). Ces protéines jouent un rôle dans la défense antivirale, en induisant la déamination des cytosines en uraciles sur le brin + rétrotranscrit. Les protéines APOBEC3 ont une localisation cellulaire différente à savoir (Figure 16-A) :

- Soit uniquement nucléaire comme APOBC3B (hA3B).
- Soit uniquement cytoplasmique comme APOBEC3DE (hA3DE), APOBEC3F (hA3F) et APOBEC3G (hA3G).
- Soit nucléaire et cytoplasmique comme APOBEC3A (hA3A), APOBEC3C (hA3C) et APOBEC3H (hA3H).

Des analyses de RT-qPCR sur des lymphocytes sanguins périphériques, une lignée cellulaire humaine de rein embryonnaire et 3 lignées cancéreuses (col de l'utérus, fibrosarcome et ostéosarcome) mettent en évidence des expressions tissulaires différentes selon le type de transcrits. Ainsi les transcrits hA3C, hA3DE, hA3F et hA3G sont plus exprimés dans les cellules de donneurs sains, alors que hA3B est plus exprimé dans les lignées cancéreuses (Figure 16-B).

Des tests de rétrotransposition utilisant l'EGFP ont été réalisés en co-transfectant des plasmides contenant la séquence L1 fusionnée à l'EGFP ($L1_{RP}$-EGFP) et un plasmide contenant les séquences codant pour les différentes protéines APOBEC3 dans les cellules 293T. Ces tests mettent en évidence que les protéines hA3A, hA3B, hA3F et hA3G diminuent la rétrotransposition des L1. Afin de confirmer ces résultats et de voir si ceux-ci sont reproductibles, un test alternatif de rétrotransposition est réalisé. Le plasmide $L1_{RP}$-EGFP est remplacé par un plasmide contenant le L1 ainsi qu'une résistance à la néomycine. Puis les mêmes expériences de co-transfections sont réalisées dans les cellules HeLa. Les protéines hA3A, hA3B, hA3F et hA3G inhibent la rétrotransposition des L1. Ces activités contre les L1 ne corrèlent pas avec le profil de localisation des protéines APOBEC3 (Figure 16-A), ni avec leurs activités inhibitrices des rétrovirus, comme par exemple contre le HIV. Les résultats montrent que la protéine hA3G est capable d'inhiber la rétrotransposition des L1 ainsi que les protéines hA3A, hA3B et hA3F. Une quantification de l'EGFP épissée par qPCR sur l'ADN des cellules 293T (utilisées pour le 1er test de rétrotransposition) permet de détecter non seulement les L1 intégrés mais aussi les L1 reverse transcrits (Figure 16-C). Cette quantification montre une diminution significative des L1 dans les cellules en présence de hA3A et hA3G. Ceci suggère que les protéines APOBEC3 ou au moins hA3A et hA3G inhibent la synthèse ADN des L1 *de novo*. Enfin les séquences des L1 reverse transcrits sont analysées à partir de l'ADN extrait des cellules 293T cotransfectées. L'analyse de ces séquences montre qu'il y a peu de mutations C → T, ce qui suggère que l'effet inhibiteur de hA3G sur la rétrotransposition des L1 peut être indépendant de l'activité désaminase utilisée dans la défense antivirale. Une hypothèse peut être émise pour expliquer cette faible détection de la transition C → T, peut-être

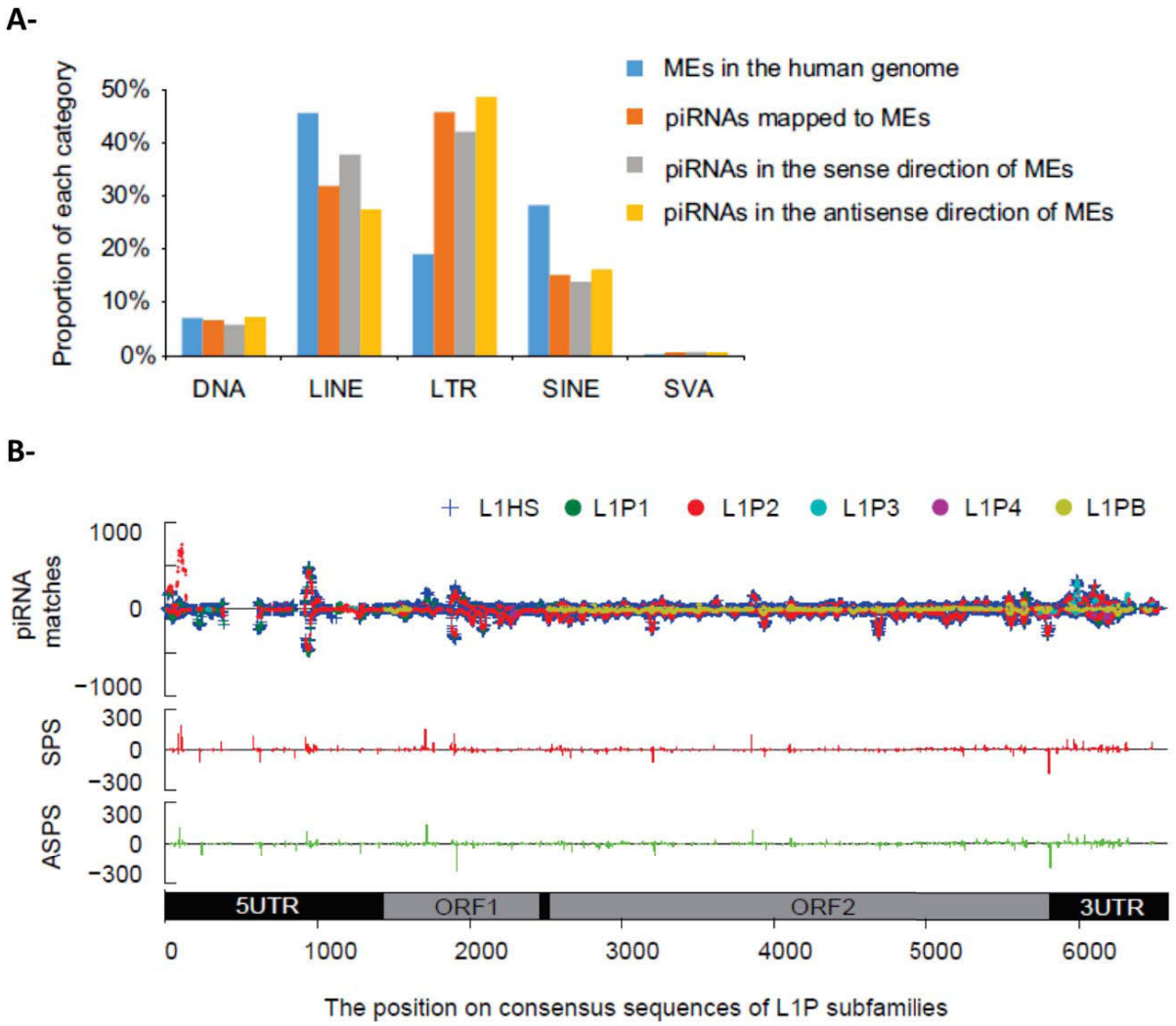


Figure 17 : Profil des piARN sur les ET.

A- Proportion de piARN se liant aux différentes familles d'ET. Les sous-familles sont les transposons ADN, les LINE, les LTR, les SINE et les éléments *SVA*. La proportion relative des ET dans le génome humain sert de référence. Il y a seulement une liaison significative des piARN au niveau des éléments LTR.

B- Profil de localisation des piARN sur les L1. Le premier graphique montre la densité en piARN se liant aux séquences consensus de 6 sous-familles de L1, avec un pic observé aux environs de 900 – 1000 pb de l'extrémité 5'UTR des L1. Le 2nd et le 3ème graphique représentent respectivement une signature de ping-pong en sens (SPS, en rouge) ou en antisens (ASPS, en vert). La faible intensité des pics indique que ce mécanisme ne semble pas impliqué dans la production de piARN se liant aux L1. Le schéma du bas représente la séquence consensus des L1.

que les protéines hA3G inhibent la rétrotransposition des L1 grâce à un nouveau mécanisme non encore élucidé. Une deuxième hypothèse peut être que cette protéine hA3G peut désaminer le L1 reverse transcrit et induire sa dégradation rapide par les enzymes cellulaires, comme l'ADN uracile glycosylase, rendant ainsi cette désamination indétectable (Kinomoto *et al.*, 2007). Mais le mécanisme sous-jacent reste à identifier.

Plus récemment, une autre étude a mis en évidence que les protéines APOBEC3B et DE peuvent se lier aux RNP *via* l'ORF1p dans des granules cytoplasmiques. Ces protéines peuvent également se lier à l'ORF2p et ainsi inhiber la rétrotransposition (Liang *et al.*, 2016).

Toutes ces protéines vont ainsi agir en intervenant soit au moment de la transcription du L1 soit aux cours des différentes étapes de la rétrotransposition du L1.

2.3.7. piARN et régulation germinale.

Les piARN sont une classe distincte de petits ARN, dont la taille est de 26 à 32 nucléotides. Ils ont un rôle dans la lignée germinale de mise sous silence des transposons et de régulation épigénétique. Ceux-ci ont été démontrés dans les organismes modèles tels que la drosophile ou la souris. Les piARN sont caractérisés par la phosphorylation du 1er nucléotide qui est généralement une base uridine. Les piARN sont produits par la transcription d'un long transcrit provenant d'un cluster de gènes piARN, puis ensuite pris en charge par différentes protéines qui vont assurer le clivage du piARN prémature. Le mécanisme impliqué dans la biogenèse de piARN est appelé mécanisme de ping-pong : le piARN généré va pouvoir aller cibler le long transcrit produit à partir du cluster et ainsi favoriser la production de piARN.

Chez l'Homme, la présence de piARN est caractérisée dans les gonades mâles et femelles durant la période fœtale avec une expression sexe spécifique. En revanche, il n'y a pas voire peu d'expression de piARN dans les testicules et les ovaires adultes. Des analyses de séquençage haut-débit sur 3 tissus testiculaires adultes humains montrent que 22% des piARN reconnaissent des ET chez l'Homme. Cette fixation se fait majoritairement sur les éléments LTR (Figure 17-A). Pour les L1, il y a une faible densité de fixation des piARN qui se fait principalement au niveau de la région 5'UTR du L1 à la position +900 à +1000 pb (Figure 17-B). Contrairement aux LTR, les piARN se fixant sur les L1 ne possèdent pas de signature du mécanisme de ping-pong (Ha *et al.*, 2014). La faible densité de reconnaissance des piARN et les rares signatures du mécanisme de ping-pong pour les éléments *Alu* et L1 suggèrent que ce mécanisme n'est pas le mécanisme primaire de la régulation de l'activité des éléments L1 et

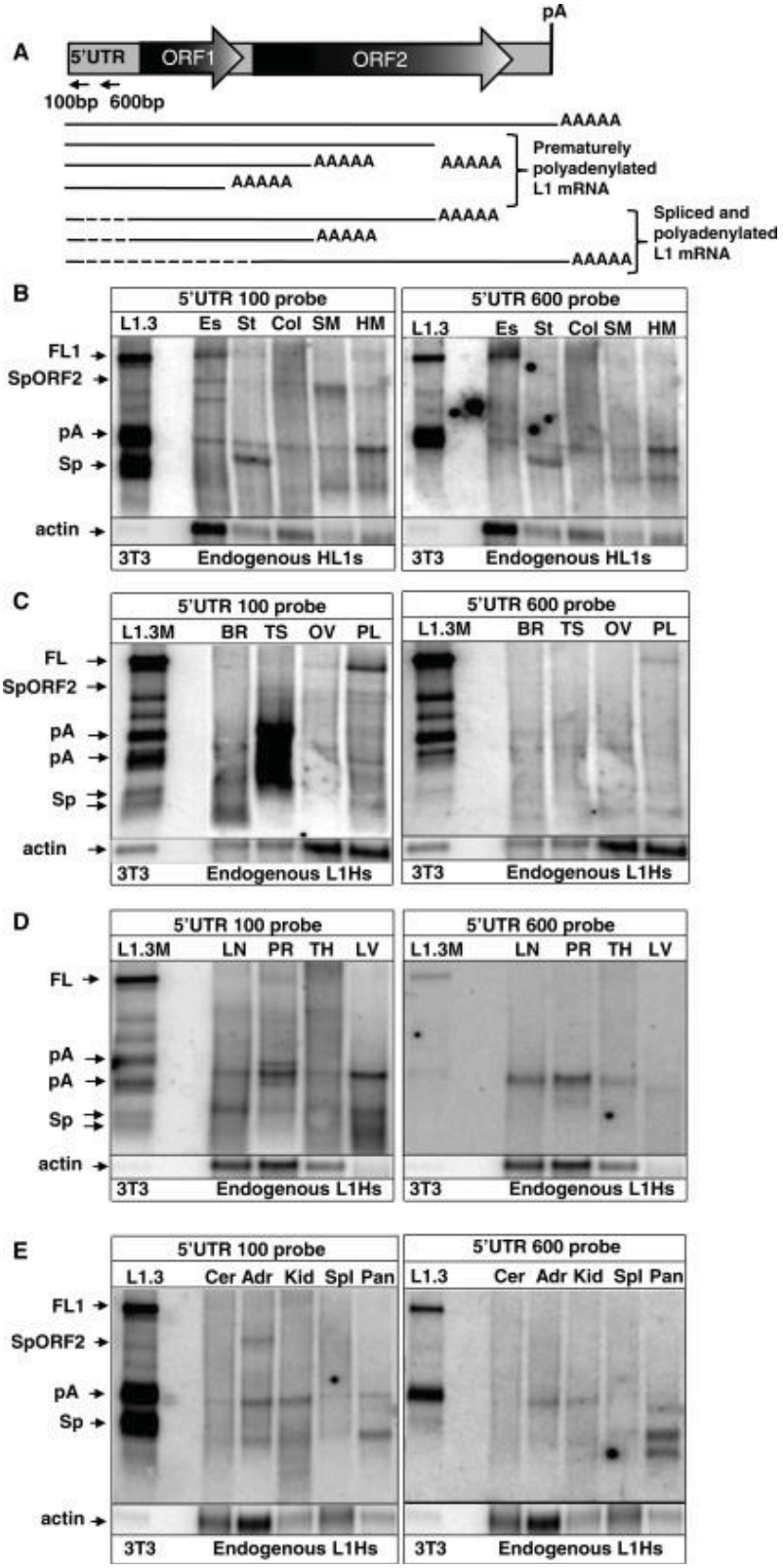


Figure 18 : Expression endogène des L1 dans les tissus normaux.

(A) Représentation schématique de la structure des L1 et des produits issus de la transcription dans les cellules normales. Un L1 pleine taille contient une région 5'UTR, 2 ORF (ORF1 et ORF2), une région 3'UTR et un signal de polyadénylation (PA). La transcription des L1 produit un ARNm pleine taille (FL1), des transcrits polyadénylés prématurément (pA) et des ARNm épissés et polyadénylés (Sp) dont un a le potentiel de coder la protéine ORF2p seule (SpORF2). Les pointillés correspondent aux séquences L1 éliminées au cours de l'épissage. Les flèches horizontales indiquent les positions relatives des sondes brin-spécifiques 5'UTR 100pb et 5'UTR 600pb utilisées pour les analyses de Northern blot.
(B-E) Analyse de l'expression endogène des L1 par Northern blot dans une variété de tissus humains adultes : Es, œsophage ; St, estomac ; Col, colon ; SM, muscle squelettique ; HM, muscle cardiaque ; Cer, col de l'utérus ; Adr, glande surrénale ; Kid, rein ; Spl, rate ; Pan, pancréas ; Br, cerveau ; Ts, testicules ; Ov, ovaires ; Pl, placenta ; LN, poumons ; PR, prostate ; TH, thymus ; LV, foie. La colonne de gauche montre les profils d'expression du L1.3 sauvage (L1.3) et mutant (L1.3M) contenant une mutation dans le site de polyadénylation dans des cellules NIH3T3 transfectées avec les vecteurs contenant ces L1 humains. La transcription des L1 produits des ARNm pleine taille (FL1), des ARNm prématurément polyadénylés (pA), des ARNm épissés et polyadénylés (Sp) dont un code pour la protéine ORF2p (SpORF2). L'actine sert de témoin, une sonde brin spécifique détecte l'ARNm de la β-actine. *(Belancio* et al*., 2010)*

Alu (Gainetdinov *et al.*, 2017; Ha *et al.*, 2014). La fonction des piARN chez l'Homme et leur rôle dans la régulation des ET restent encore à élucider (Williams *et al.*, 2015).

2.4. Expression et transposition en condition physiologique.

2.4.1. Expression des LINE-1 détectable dans les différents types cellulaires.

Différentes études se sont intéressées à caractériser l'expression des L1 à l'échelle du génome entier à différents stades de développement ou dans des cellules différenciées.

L'expression des L1 a été quantifiée dans les cellules somatiques humaines par des approches de Northern blot utilisant des sondes ciblant les extrémités 5'UTR (100 pb ou 600 pb) afin de discriminer la présence de variants d'épissage dans les différents tissus. Les résultats montrent une forte expression de L1 pleine taille dans l'œsophage, la prostate, l'estomac et le muscle cardiaque associé à des variants d'épissage (Figure 18). Il n'y a pas ou peu d'expression dans la glande surrénale, la rate, les reins et le col de l'utérus (Belancio *et al.*, 2010).

L'expression des L1 a également été quantifiée par RT-qPCR sur les ARN totaux des lignées de cellules souches embryonnaires humaines H7 et H9. Deux couples d'amorces ont été utilisés : un localisé dans la région 5'UTR et un dans la région de l'ORF2, dessinés à partir de la séquence consensus du L1HS. 22 loci exprimés ont été validés pour les cellules H7 et 26 pour les cellules H9, impliquant des L1 des sous-familles L1HS à L1PA7 et un L1PA15. L'utilisation de la séquence consensus du L1HS pour le dessin des amorces introduit un biais car l'amplification ciblera des L1 récents ayant une forte homologie de séquence avec le L1HS, ce qui sous-estime le nombre de loci pouvant être exprimés (Macia *et al.*, 2011).

Une autre étude quantifie l'expression en sens de L1 dans des lignées lymphoblastoïdes humaines en adaptant des protocoles de 5' et 3' RACE (Rapid Amplification of cDNA end) suivi de séquençage. Ainsi 692 loci ont été identifiés comme exprimés dans le génome humain, impliquant les sous-familles de L1HS à L1PA6, dont 410 correspondent à des éléments pleine taille (Rangwala *et al.*, 2009).

Par ailleurs, l'expression des protéines ORF1p et ORF2p a été quantifiée dans des tissus fœtaux et adultes de la lignée germinale par des analyses de Western blot. Aussi les protéines ORF1p er ORF2p sont exprimées dans les testicules adultes : dans les spermatocytes secondaires, les spermatides immatures, les cellules de Sertoli, de Leydig, de couverture et les cellules vasculaires endothéliales. Ces protéines sont également exprimées au stade fœtal dans les pré-

spermatogonies, les cellules de Leydig, les cellules vasculaires endothéliales, l'épididyme et l'épithélium colonnaire. Enfin le syncytiotrophoblaste et les cellules vasculaires endothéliales du placenta expriment ORF1p et ORF2p (Ergün *et al.*, 2004).

Grâce à des données de séquençage obtenues à partir de libraire de CAGE (Cap Analysis Gene Expression), l'expression des L1 dans différents tissus a été mise en évidence. Les L1 sont majoritairement exprimés dans l'embryon, les cellules souches embryonnaires mais aussi le tissu adipeux et le cerveau (Faulkner *et al.*, 2009).

De plus, une étude combinant des expériences de 5'RACE et un séquençage haut débit avec la technologie PacBio permet de développer une approche d'expression des L1 dans le génome entier, non seulement à partir du promoteur sens mais aussi de l'ASP. Ils montrent que l'expression des L1 proviendrait majoritairement d'une vingtaine de loci pleine taille impliquant les sous-familles L1HS à L1PA3. Une cinquantaine d'autres loci participeraient plus faiblement à cette expression. Ces expériences étant réalisées sur des fractions cytoplasmiques provenant des cellules HeLa en ciblant les ARN polyadénylés, cela peut induire une sous-estimation du nombre de copies impliquées dans l'expression des L1 (Deininger *et al.*, 2017), par rapport à la centaine estimée par Brouha (Brouha *et al.*, 2003).

Enfin, concernant les éléments de la sous famille L1HS, il a été mis en évidence l'existence d'une transcription d'une vingtaine de L1HS-Ta dans des lignées cellulaires somatiques normales et tumorales. Toutefois, ceux-ci diffèrent d'une lignée à l'autre. L'étude des loci associés aux L1HS-Ta transcrits versus ceux associés aux L1HS-Ta non transcrits a montré que les gènes associés aux L1 exprimés sont eux même plus exprimés que ceux contenant des L1 non exprimés. Cette observation suggère que les loci de gènes fortement exprimés pourraient représenter un environnement génomique favorable pour la réactivation des L1 (Philippe *et al.*, 2016).

Toutes ces études montrent une expression non seulement dans les cellules de la lignée germinale, mais aussi dans les cellules somatiques différenciées.

2.4.2. Transposition *de novo* en condition physiologique.

Comment les L1 ont-ils pu coloniser le génome humain jusqu'à en représenter 17% ? Pour ce faire, les insertions de L1 doivent être héritables et dans ce cas, les nouvelles insertions doivent avoir lieu dans la lignée germinale ou les cellules embryonnaires afin d'être transmises à la descendance.

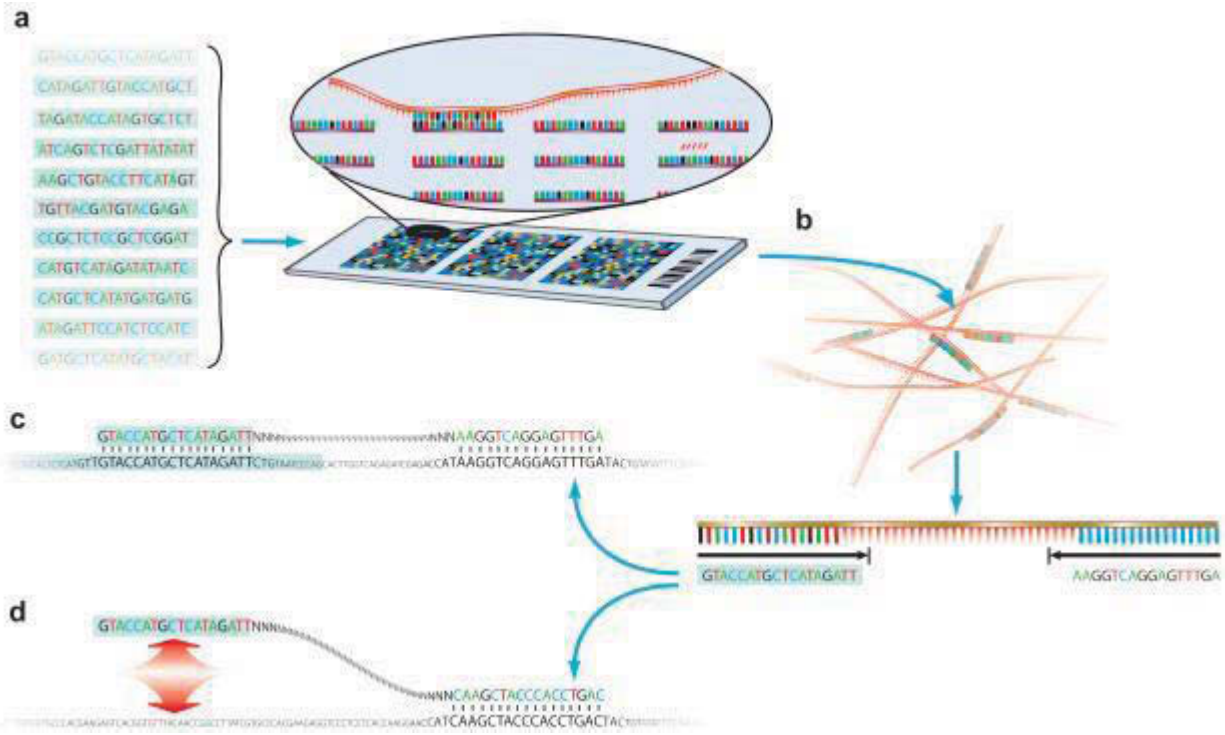


Figure 19 **: Méthodologie de la technique de RC-seq.**

a- Capture des rétrotransposons : l'ADN génomique fragmenté est hybridé sur une puce où se trouvent les sondes de capture.
b- Séquençage après hybridation des fragments d'ADN avec un séquenceur Illumina produisant ≈$2.5x10^7$ lectures paired-end par librairie. Les séquences obtenues sont ensuite alignées sur le génome humain.
c- Les lectures alignées par paire à un locus unique indiquent les insertions connues des rétrotransposons.
d- Pour les lectures non pairées, lorsqu'une l'extrémité d'une lecture s'aligne à un locus unique et l'autre s'aligne au niveau d'un rétrotransposon distant, ceci indique un nouvel évènement de rétrotransposition. *(Baillie* et al.*, 2011)*

La première mise en évidence de l'insertion des L1 dans la lignée germinale a été décrite en 1988 où un L1 s'est inséré dans le gène du facteur VIII de la coagulation provoquant ainsi l'hémophilie A. Ces résultats ont établi que des insertions *de novo* de L1 ont lieu chez l'Homme et peuvent provoquer des maladies (Partie 3). Néanmoins, l'origine développementale de la rétrotransposition *de novo* reste à élucider. Comme décrit précédemment, les protéines ORF1p et ORF2p s'expriment durant la spécification des cellules germinales. De plus, l'étude des maladies liées à l'X chez l'Homme induites par les L1 suggère une mobilisation des L1 dans la lignée germinale et l'embryon précoce (Brouha *et al.*, 2003; Kazazian *et al.*, 1988; Van Den Hurk *et al.*, 2007). Des néo-insertions de L1 ont été mises en évidence dans des cellules souches embryonnaires humaines H9 transfectées avec un plasmide rapporteur de la rétrotransposition des L1 (Garcia-Perez *et al.*, 2007).

Grâce aux avancées du séquençage, une nouvelle technique combinant une étape de capture des rétrotransposons suivi d'un séquençage haut débit, nommé RC-seq (Retrotransposon-Capture Sequencing) a été mise au point. Sur l'ADN génomique fragmenté, des sondes de capture ciblant les régions 5' et 3' de L1 pleine taille sont hybridées, puis une étape de séquençage Illumina a lieu. Enfin les lectures obtenues sont alignées sur le génome humain et analysées pour identifier des insertions de L1 connues mais également des nouvelles insertions (Figure 19). 7 743 insertions somatiques putatives de L1 sont identifiées dans l'hippocampe et le noyau caudé de 3 donneurs. La majorité de ces insertions somatiques sont dues pour 68% au L1-Ta, 12% aux L1 pré-Ta et 20% aux L1PA2 (Baillie *et al.*, 2011).

Grâce aux tests de rétrotransposition utilisant le système L1-EGFP et à l'utilisation d'un vecteur adénoviral, la mobilisation des L1 dans les progéniteurs neuronaux a été mise en évidence. Les auteurs différencient des progéniteurs neuronaux (NPC) modifiés avec le plasmide rapporteur L1-EGFP pendant 31 jours jusqu'à l'obtention de neurones en ajoutant un marqueur des cellules en division le BrdU (5-bromo-2'désoxyuridine) dans le milieu de culture. Les cellules neuronales EGFP+ ne sont pas marquées par l'anticorps anti-BrdU. Ceci suggère que les cellules neuronales qui ne se divisent pas possèdent une mobilisation *de novo* des L1 (Macia *et al.*, 2017).

Les L1 sont capables de rétrotransposer dans des cellules de la lignée progénitrice neuronale en conditions physiologiques, incluant également les neurones post-mitotiques.

Grâce à la technique d'ATLAS-seq qui a été adaptée afin d'identifier des éléments L1HS-Ta, 7 823 insertions putatives de L1HS-Ta ont été identifiées dans 12 lignées cellulaires de cellules normales de fibroblastes (BJ, MRC-5, IMR-90), de rein fœtal (HEK-293) et des cellules

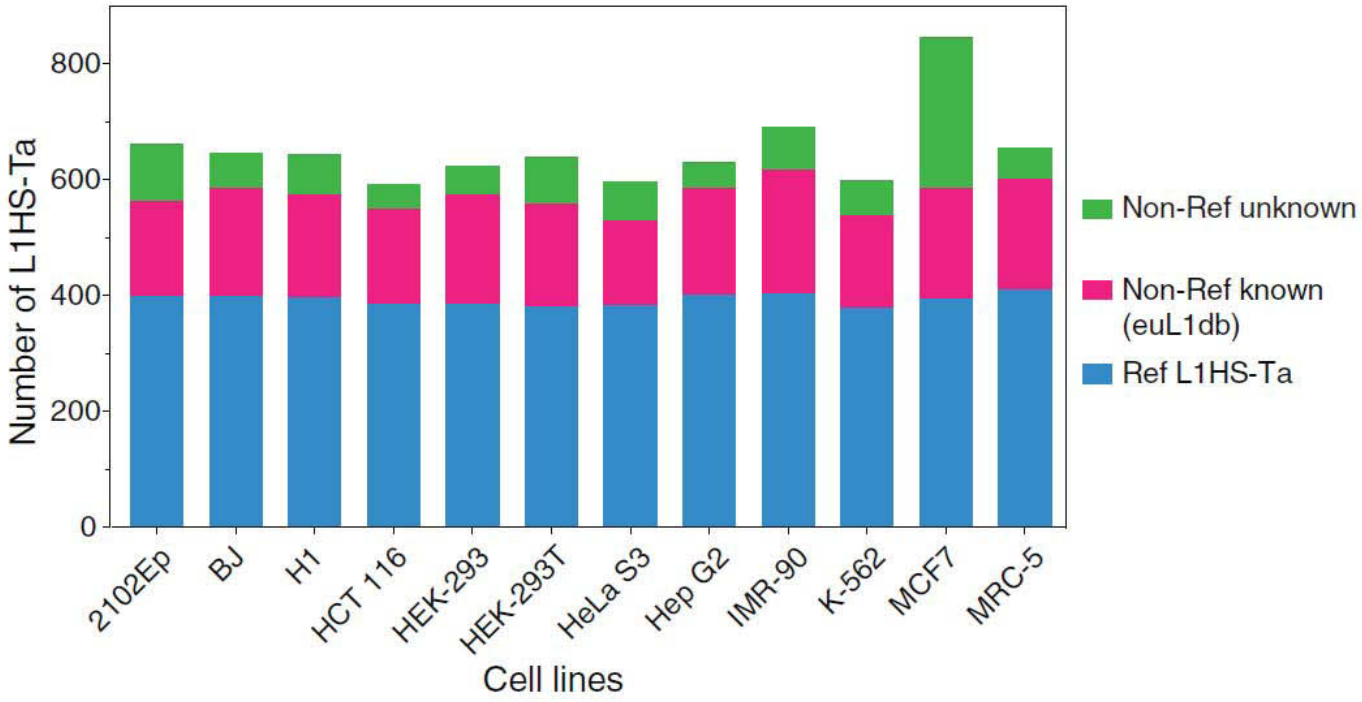


Figure 20 : Identification des insertions de L1HS-Ta polymorphiques dans 12 lignées cellulaires somatiques humaines.

Les insertions L1HS-Ta sont identifiées grâce à la technique d'ATLAS-seq dans le génome humain de différentes lignées cellulaires normales : de fibroblastes (BJ, MRC-5, IMR-90), de rein fœtal (HEK-293 et HEK-293T), de cellules souches embryonnaires (H1) ; et de lignées cellulaires tumorales : de carcinome embryonnaire (2102Ep), colorectal (HCT 116), du foie (HepG2), du sein (MCF 7), du col de l'utérus (HeLa) et de lymphomes (K562). La majorité des insertions des L1HS-Ta détectées sont référencées dans la littérature (en bleu), certaines sont non référencées (en rose) mais sont connues (euL1db) *(Mir* et al.*, 2015)* et de nouvelles insertions sont détectées (en vert).

souches embryonnaires (H1) ainsi que des lignées cellulaires cancéreuses. En moyenne chaque lignée contient 652 (±68) copies de L1HS-ta dont 178 (±12) sont pleine taille. Parmi ces copies, 393 (±10) sont des insertions référencées, 179 (±18) ne sont pas référencées mais avaient été identifiées comme polymorphiques et 80 (±60) correspondent à de nouvelles insertions (Figure 20). Afin de valider ces observations, 72 insertions non référencées ont été choisies au hasard. Les PCR réalisées sur les cellules HEK-293T sur 70 loci en valident 66. Au final, le ratio de vrais positifs identifiés avec cette méthode est de 94%. L'expression de 337 insertions de L1 a été analysée à partir de données de RNA-seq pour les lignées H1, HEK-293T, BJ et IMR-92 (ainsi que pour les lignées tumorales). Les résultats montrent que de nombreux L1HS-Ta sont présents à la même position génomique dans différents types cellulaires mais ceux-ci sont différentiellement exprimés indiquant une réactivation de L1HS-Ta dépendante du type cellulaire mais également copie-spécifique. Enfin afin de déterminer si ces L1HS-Ta exprimés sont compétents pour la rétrotransposition, différents critères ont été mis en place :

1) rétrotransposition de L1 décrite dans la littérature (Beck *et al.*, 2011; Brouha *et al.*, 2003) ;
2) comparaison des copies L1HS-Ta identifiées comme étant fortement exprimées avec ce qui a été décrit dans la littérature grâce à la détection de transductions en 3' (Tubio *et al.*, 2014) ;
3) identification de transduction en 3' dans les résultats de 3'ATLAS-seq.

Si une copie répond à au moins 2 de ces critères, elles sont considérées comme compétentes pour la rétrotransposition. Aussi 11 copies parmi les 20 fortement exprimées sont compétentes pour la rétrotransposition. Au final, la transcription des L1HS-Ta est majoritairement inactive dans les cellules somatiques, seulement quelques copies échappent aux mécanismes de répression mis en place par les cellules, permettant ainsi leur expression et dans un second temps, pour certaines copies leur rétrotransposition (Philippe *et al.*, 2016).

2.5. Rôles des L1 dans le génome humain.

Comme nous l'avons vu, les L1 sont importants pour le génome humain, puisque depuis leur apparition il y a 150 millions d'années, ceux-ci se sont amplifiés, ont co-évolué avec le génome pour en représenter 17% aujourd'hui. Les rôles des L1 sont divers.

Tout d'abord, les L1 sont une source importante de l'évolution et de l'innovation de notre génome. En effet, les protéines ORF2p, produites à partir des L1, ayant une activité d'endonucléase peuvent générer des cassures double brin de l'ADN (Gasior *et al.*, 2006). Il est

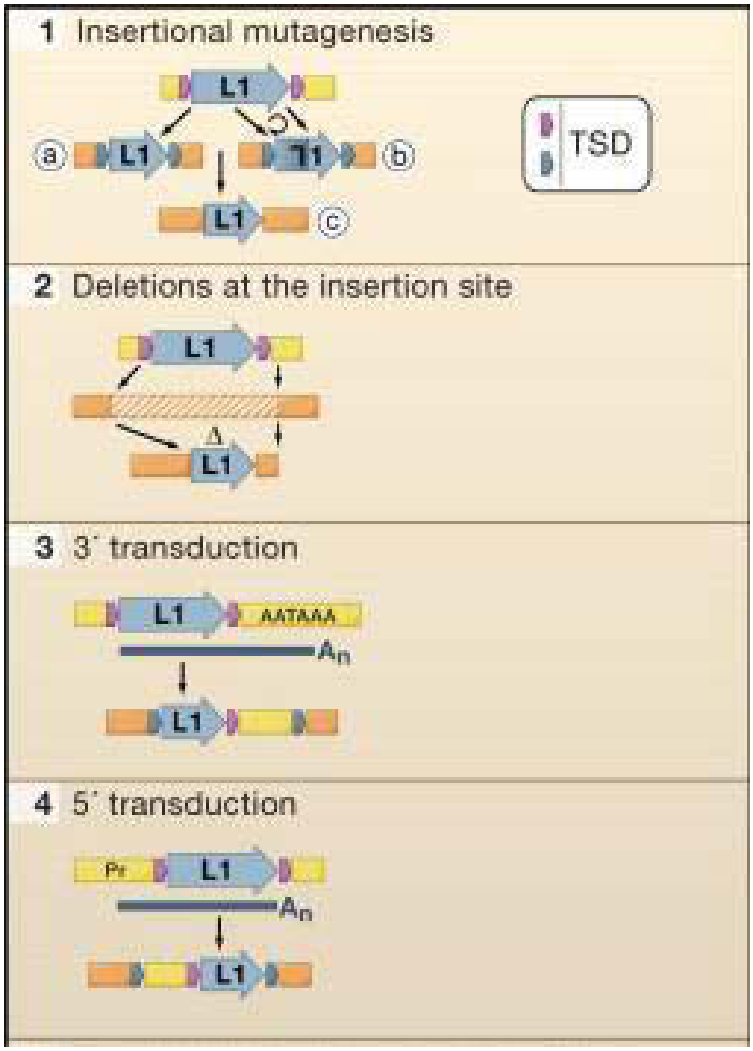


Figure 21 : Comment les insertions de L1 affectent les cellules ?

Les insertions à de nouvelles positions via le mécanisme de TPRT peuvent impliquer des L1 pleine taille ou tronqués (1a) ou bien des L1 contenant des inversions ou des délétions (1b). Les insertions indépendantes de l'endonucléase peuvent avoir lieu à une faible fréquence suite à une cassure double brin de l'ADN (1c). Les insertions de L1 peuvent s'accompagner d'une délétion de la séquence ADN (hachuré) au site d'insertion (2). Les régions flanquantes en 3' ou en 5' du L1 peuvent être associées au L1 au cours de la rétrotransposition du L1 donnant lieu à des transductions en 3' ou en 5' (3 et 4). (Goodier & Kazazian, 2008)

possible qu'une partie de l'instabilité génomique attribuée aux cassures double brin de l'ADN, qui sont hautement mutagéniques, soit en grande partie due à l'activité de l'ORF2p des L1 dans la lignée germinale et les cellules somatiques (HeLa et MCF7) (Gasior *et al.*, 2006).

Les L1 peuvent être associés à la réparation de ces cassures. En effet, des expériences de rétrotransposition utilisant un modèle cellulaire ayant une déficience dans le mécanisme de réparation non homologue (NHEJ) de l'ADN (lignées cellulaires XR-1 et V-3) ont démontré que des insertions de L1 pouvaient avoir lieu indépendamment de l'activité endonucléase et donc du mécanisme de TPRT. Dans ce cas, le L1 s'intègre au niveau de la lésion et la répare (Morrish *et al.*, 2002). Ce mécanisme de rétrotransposition indépendant de l'endonucléase du L1 pourrait être un mécanisme de réparation ancestral des cassures double brin médié par un ARN intermédiaire, avant que les cellules n'aient acquis des protéines possédant un domaine endonucléase (Morrish *et al.*, 2007).

Parmi les nombreux rôles pouvant être attribués aux L1, ceux-ci peuvent porter dans leur séquence, la séquence des régions génomiques adjacentes au cours du mécanisme de rétrotransposition (Figure 21). Ceci est nommé le mécanisme de transduction en 3' (qui est le plus fréquent), mais aussi en 5'. Dans le mécanisme de transduction en 3', la machinerie de transcription des ARN va sauter le signal de polyadénylation des L1 et terminer la transcription en utilisant un signal de polyadénylation alternatif localisé en aval dans la région génomique adjacente en 3'. De manière similaire, une transduction en 5' a lieu quand un promoteur localisé en amont du L1 est utilisé pour transcrire la séquence en amont et se poursuit dans le L1 (Hancks & Kazazian, 2012). Par ailleurs, les L1 possèdent de nombreux sites d'épissage comme nous l'avons vu dans la partie 2.3.3. Ces sites donneur et accepteur d'épissage vont pouvoir être utilisés par le splicesome lors de l'épissage du pré-ARNm quand un L1 est intégré dans une région intronique d'un gène (Kaer *et al.*, 2011).

A l'image des rétrotransposons non-autonomes tels que les SINE, la machinerie de rétrotransposition peut être détournée par les transcrits ARN de la cellule hôte. Cela va donner lieu à des duplications de gènes et leur insertion en une nouvelle localisation génomique. A cause de l'absence des séquences régulatrices permettant leur transcription, ces séquences seront nommées rétro-pseudogènes. Elles peuvent parfois donner lieu à la création de nouveaux gènes selon le lieu d'insertion (Kaessmann *et al.*, 2009).

D'autre part, les régulations épigénétiques régulant la région promotrice du L1 vont pouvoir s'étendre et ainsi modifier l'expression des gènes adjacents. Par exemple, la formation d'hétérochromatine au niveau de la région 5'UTR du L1 va se propager et ainsi empêcher

n°	Sous-famille	Maladie	gène affecté	Chr	Taille L1	sens/antisens	Position de l'insertion	Troncation L1	Mécanisme mutationnel au niveau du transcrit	Références
1	L1 Ta	Dystrophie musculaire de Duchenne	DMD	X	212	AS	Exon 67	troncation en 5'	Transduction en 3' : transcrit tronqué (saut exon)	Solyom et al., 2011
2	L1 Ta	Maladie chronique granulomateuse	CYBB	X	1722	S	Exon 4	troncation en 5'	Transduction en 3' formant un nouveau transcrit	Meischl et al., 1998 ; Brouha et al., 2002
3	L1 Ta	Choroideremie	CHM	X	6017	AS	Exon 6	non	Transduction en 5' avec saut de l'exon 6	Van den Hurk et al., 2003
4	L1 HS	Syndrome de Coffin-Lowry	RPS6KA3	X	2800	AS	Intron 3	troncation en 5' et inversion	Saut de l'exon 4	Martinez-Garay et al., 2003
5	L1 Ta	Dystrophie musculaire de Duchenne	DMD	X	608	AS	Exon 44	troncation en 5' (délétion de 2 nts)	Saut de l'exon 44	Narita et al., 1993
6	L1 Ta	Dystrophie musculaire de type Fukuyama	FKTN	9	1200	S	Intron 7 / Exon 8 (24 nt)	troncation en 5'	Saut des exon 7 et 8	Kondo-Iida et al., 1999
7	L1 preTa	Neurofibromatose I	NF1	17	1800	S	Exon 21	troncation en 5'	Saut de l'exon 22	Wimmer et al., 2011
8	L1 HS	Ataxie avec apraxie occulo-motrice 2	SETX	9	1300	S	Exon 12	troncation en 5' et inversion	Saut d'exon partiel et exonisation partielle du L1	Bernard et al., 2009
9	L1 Ta	Neurofibromatose I	NF1	17	6000	S	Exon 33		Saut d'une partie de l'exon 39 et insertion d'une partie du L1	Wimmer et al., 2011
10	L1 Ta	Maladie chronique granulomateuse	CYBB	X	836	S	Intron 5	troncation en 5' et inversion	Epissage dans le L1 et saut d'une partie de l'exon 6	Meischl et al., 2000
11	L1 HS	Syndrome de Chanarin-Dorfman (CDS)	ABHD5	3	6000	S	Intron 3		Exonisation du L1 par création d'un site d'épissage	Samuelov et al., 2011
12	L1 Ta-1d	Syndrome de Rotor	SLCO1B3	12	5989	S	Intron 5		Variant d'épissage	Kagawa et al., 2015
13	L1 Ta	Dystrophie musculaire de Duchenne	DMD	X	452	AS	Exon 44	troncation en 5' et inversion	Création d'un site polyA : terminaison prématurée	Musova et al., 2006
14	L1 Ta	Hémophilie B	FIX	X	163	S	Exon 7	troncation en 5'	Création d'un site polyA : terminaison prématurée	Mukherjee et al., 2004
15	L1 Ta	Dystrophie musculaire de Duchenne	DMD	X	1400	S	Exon 48	troncation en 5' et inversion	Terminaison prématurée	Holmes et al., 1994
16	L1 Ta	Rétinite pigmentaire (XLRP)	RP2	X	6000	S	Intron 1		Perte de l'ARNm par création d'un site polyA	Schwahn et al., 1998
17	L1 HS	Déficience en pyruvate déshydrogénase	PDHX	11	6086	S	Intron 2		Délétion de 46 kb	Miné et al., 2007
18	L1 Ta	Cardiomyopathie dilatée liée à l'X	DMD	X	530	AS	Exon 1	troncation en 5'	Perte de l'ARNm	Yoshida et al., 1998
19	L1 HS	Syndrome de BOR	EYA1	8	3756	AS	Intron 7	troncation en 3'	Délétion de 17 kb avec perte des exons 5 à 7 qui ont été remplacé par la séquence L1	Morisada er al., 2010
20	L1 Ta	Hémophilie B	FIX	X	463	S	Exon 5	troncation en 5'	Rétention de la séquence du L1	Li et al., 2001
21	L1 Ta	β-thalassémie	HBB	11	6000	AS	Intron 2		ND	Divoly et al., 1996
22	L1 Ta	Hémophilie A	FVIII	X	3800	S	Exon 14	troncation en 5'	ND	Kazazian et al., 1988
23	L1 preTa	Hémophilie A	FVIII	X	2300	AS	Exon 14	troncation en 5' et inversion	ND	Kazazian et al., 1988
24	L1 Ta	Cancer du colon	APC	5	520	S	Exon 16	troncation en 5' et inversion	Exonisation du L1 qui perturbe la séquence codante du gène APC	Miki et al., 1992
25	L1 HS	Carcinome endométrial	PTEN	10	90	S	Exon 6	troncation en 5'	Exonisation du L1 au niveau de l'exon 6	Helman et al., 2014
26	L1 Ta-1d	Rétinoblastome familial	RB1	13	6044	S	Intron 14		Variant d'épissage, exonisation d'une partie du L1	Rodrigez-Martin et al., 2016

Table 1 : Evènements de rétrotranspositions associées à des maladies humaines

Les résultats de cette table sont la compilation de différentes références (Ostertag & Kazazian, 2001 ; Chen et al., 2006 ; Belancio et al., 2008) ; Hancks et Kazazian, 2012)

l'expression des gènes se situant à proximité. C'est la raison pour laquelle, le chromosome X est riche en L1. En effet, les L1 vont servir de relais à l'hétérochromatisation du chromosome X chez la femme afin de permettre l'inactivation de l'un des deux chromosomes (Lyon, 2006).

De plus, ces L1 sont essentiels à l'établissement de la chromatine néo-centromérique *via* un processus similaire d'hétérochromatisation.

Par ailleurs, les L1 de par leur caractère répété peuvent être impliqués dans les recombinaisons homologues des chromosomes au cours de la méiose et ainsi participer au brassage génétique des génomes (Burwinkel & Kilimann, 1998).

3. Les LINE-1 et leurs implications en pathologie humaine.

Ces éléments ne vont pas avoir que des rôles positifs. En effet, ils contribuent à des réarrangements chromosomiques induisant des délétions, des duplications et des inversions, dont les effets vont être détaillés dans ce chapitre.

3.1. Transposition germinale et maladies monogéniques héréditaires.

Les maladies héréditaires liées à des évènements de rétrotransposition des L1 ont lieu soit au cours de la gamétogenèse chez les parents, notamment chez la mère pour les maladies liées à l'X, soit au cours du développement embryonnaire précoce.

Ces insertions vont pouvoir avoir lieu car les mécanismes de répression de la rétrotransposition des L1 sont déficients ou ont été détournés à des fins délétères. Vingt-trois cas de phénotypes monogéniques liés à une insertion de L1 ont été rapportés dans la littérature (Table 1). Parmi ces maladies, la première à avoir été décrite est l'hémophilie A. Cette maladie liée à l'X est causée par l'insertion d'un L1 dans l'exon 14 du facteur VIII de la coagulation induisant une augmentation de la taille du transcrit et la perte de l'activité de la protéine (Kazazian *et al.*, 1988).

Les maladies liées à l'insertion de L1 sont nombreuses et sont causées par différents mécanismes comme nous allons le voir.

- *Maladies induites par saut d'exon.*

Les insertions des L1 dans les gènes peuvent induire des sauts d'exons. Sept cas de maladies impliquant ce type d'évènement ont été caractérisés (Table 1).

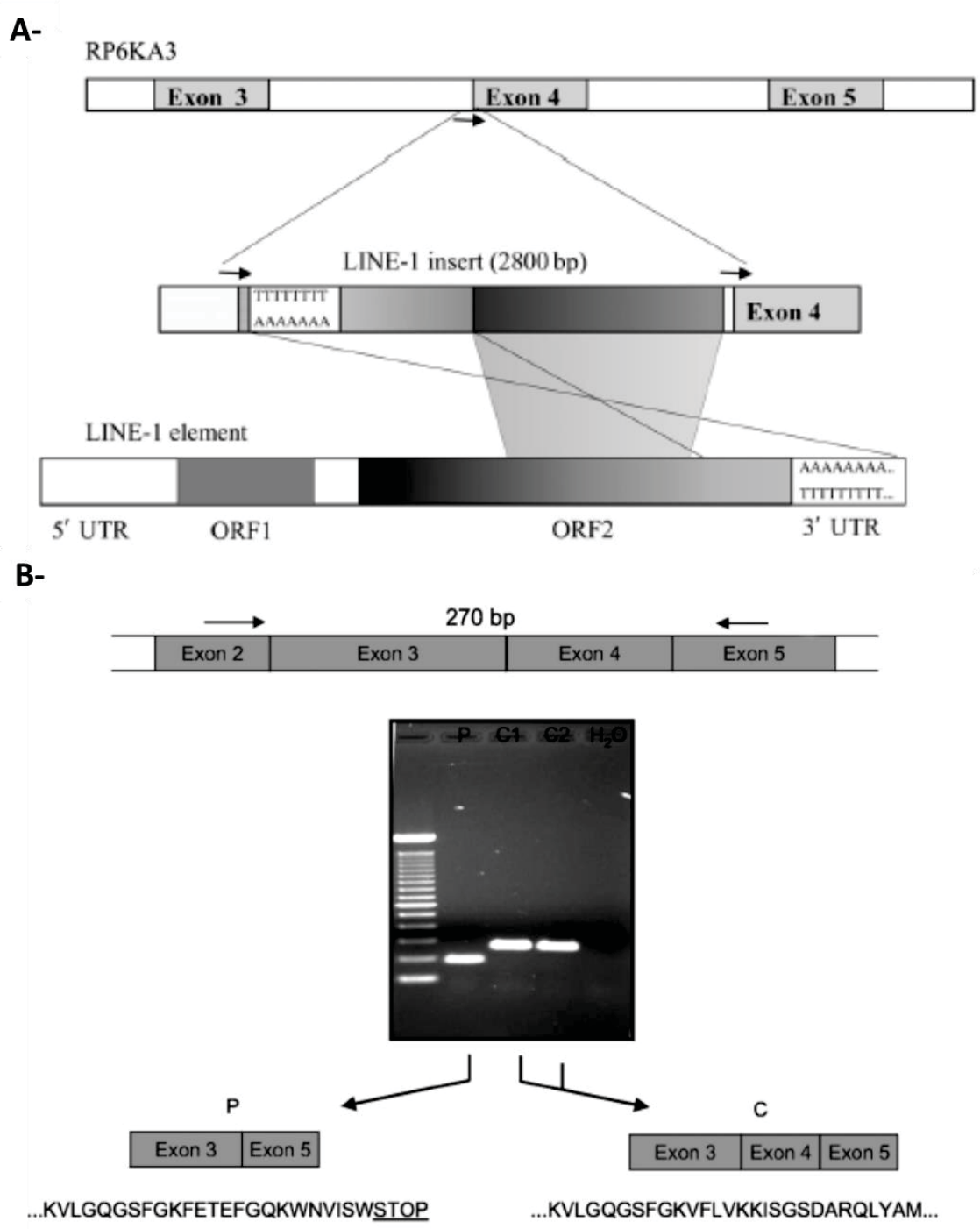


Figure 22 : Saut d'exon induit par l'insertion d'un L1 dans le gène RP6KA3.

A- Vue schématique de l'insertion du L1 identifiée chez le patient. Le L1 tronqué en 5' s'insère dans l'intron 3 du gène RP6KA3.

B- L'amplification par RT-PCR avec des couples d'amorces localisés dans l'exon 2 et l'exon 5 (flèches noires) montre un transcrit plus petit de 82 bases dû à la perte de l'exon 4 par rapport aux transcrits des deux individus contrôles. La séquence protéique produite à partir de ces transcrits est schématisée en dessous et montre la perte de la séquence codée par l'exon 4 induisant la maladie de Coffin-Lowry (Martínez-Garay *et al.*, 2003).

Par exemple, le syndrome de Coffin-Lowry est un trouble neurologique rare caractérisé par un retard psychomoteur et de croissance, une dysmorphie faciale, des anomalies digitales et des déformations squelettiques progressives. Un L1 (tronqué dans sa partie 5') a été identifié inséré dans l'intron 3 du gène *RPS6KA3* qui code une protéine kinase impliquée dans le contrôle de la croissance cellulaire et de la différenciation (Figure 22-A). Cette insertion va induire le saut de l'exon 4. Aussi les transcrits produits auront perdu l'exon 4 et se termineront dans l'exon 5 produisant un codon stop prématuré au niveau de la protéine (Figure 22-B). Par ailleurs, cette maladie est associée à une mutation ponctuelle hétérozygote dans l'exon 10 (833T>C) sur le deuxième allèle. Cela produit un changement d'acide aminé au niveau de la protéine (F268S), qui ne sera plus fonctionnelle. L'ensemble de ces évènements vont conduire au développement de cette maladie autosomique récessive (Martínez-Garay *et al.*, 2003).

Parfois, les insertions des L1 dans les gènes peuvent se produire avec des L1 transduits en 3', c'est-à-dire qu'au cours du mécanisme de rétrotransposition du L1 celui-ci a emporté une partie de la région génomique de son locus initial située juste en aval de son 3'UTR. Trois cas de maladies impliquant ce type d'évènement ont été identifiés (Table 1). Par exemple, la myopathie de Duchenne est caractérisée par une atrophie et une faiblesse musculaire progressive dues à une anomalie dans le gène *DMD* codant pour la dystrophine qui est impliquée dans le maintien de la fibre musculaire. Un cas de myopathie de Duchenne rapporte une insertion d'un L1 dans l'exon 67 du gène *DMD* sur le chromosome X. La taille de l'insertion du L1, provenant du chromosome 11, est de 327 pb et celui-ci contient 212 pb homologues à une séquence adjacente au L1 localisée en 3' de celui-ci sur le chromosome 11 (Figure 23). Cette insertion va induire le saut de l'exon 67 et va produire un transcrit *DMD* tronqué. Lors de sa traduction en protéine, la dystrophine tronquée ne pourra plus assurer son rôle dans le maintien de la fibre musculaire, induisant ainsi la faiblesse musculaire observée dans cette maladie (Solyom *et al.*, 2012).

De plus, les insertions des L1 dans les gènes peuvent se produire avec des L1 transduits en 5', c'est-à-dire qu'au cours du mécanisme de rétrotransposition du L1 celui-ci a emporté une partie de la région génomique de son locus originel située en amont de sa région 5'UTR. Un cas de maladie impliquant ce type d'évènement a été étudié (Table 1).

Par exemple, la choroïdérémie est une maladie récessive liée à l'X induisant une dégénération progressive du tissu choriorétinal de l'œil. Un cas de cette maladie est dû à l'insertion d'un L1 dans le gène *CHM* codant la protéine REP-1 (Figure 24). Cette protéine est essentielle à l'ancrage des protéines Rab aux vésicules de sécrétion et d'endocytose. Le L1 pleine taille

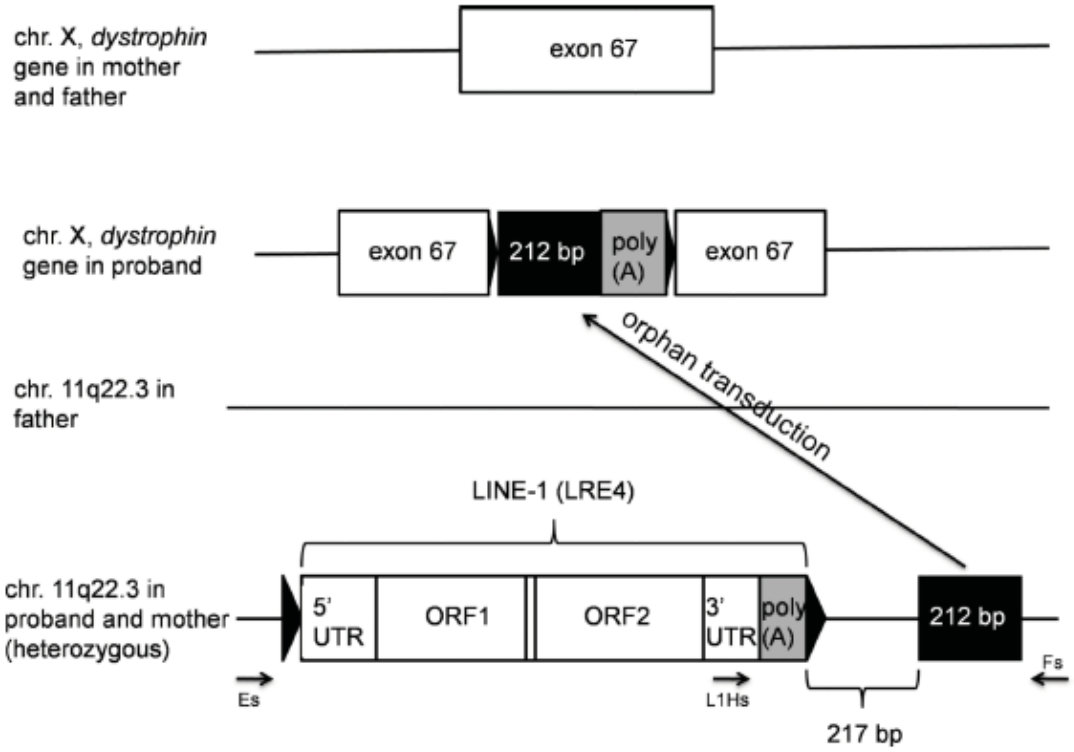


Figure 23 : Evènement de transduction en 3' de L1 causant la myopathie de Duchenne.

Un cas de transduction en 3' de L1 a été rapporté. La région transduite est en noir sur le schéma et les triangles noirs indiquent les sites de duplication (TSD). Le L1 va s'insérer dans l'exon 67 du gène de la dystrophine (*DMD*) induisant le saut de cet exon et la production d'une protéine tronquée aboutissant au phénotype de faiblesse musculaire observé dans la myopathie de Duchenne. (Solyom *et al.*, 2012)

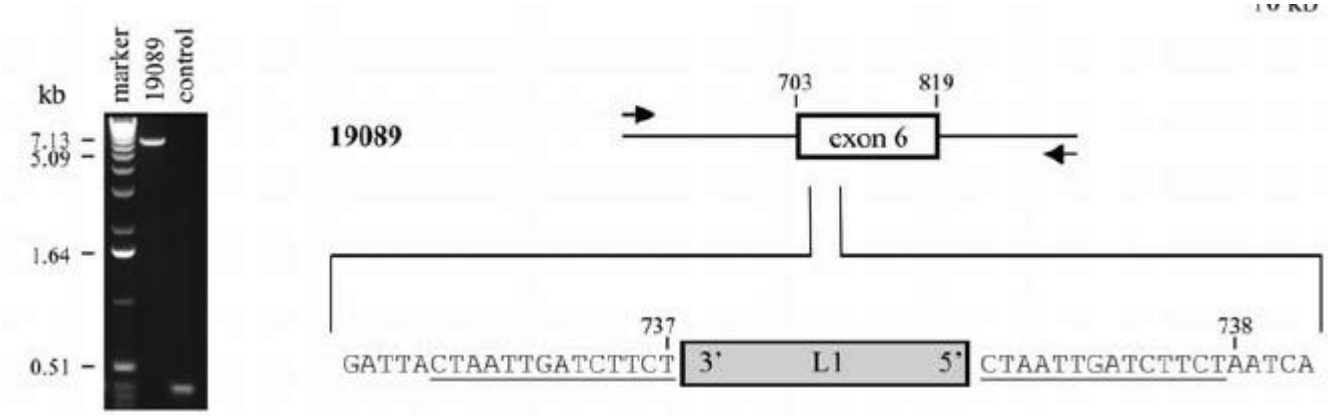


Figure 24 : Insertion d'un L1 dans l'exon 6 du gène CMH.

L'analyse par PCR de l'ADNg du patient 19089 et d'un individu contrôle en utilisant des amorces localisées de part et d'autre de l'exon 6 (flèches noires) montre qu'à la place d'un produit de 410 pb chez l'individu contrôle, un produit de 6 Kb est amplifié chez le patient 19089. A côté est schématisé l'insertion du L1 avec la transduction en 5' dans l'exon 6 du gène *CMH*. La maladie associée est le chroroïdérémie (Van Den Hurk *et al.*, 2007).

s'insère dans l'exon 6 du gène. Ceci va induire le saut de l'exon 6. La protéine résultante aura perdu les acides aminés 235 à 273 correspondant à une région conservée de la protéine (SRC) impliquée dans la liaison des protéines Rab. Ceci induit la dégénération des cellules de l'œil par défaut des voies impliquées dans la sécrétion (Van Den Hurk *et al.*, 2007).

- *Maladies induites par la création de variants d'épissage.*

Des variants d'épissage associés à l'exonisation de L1 est une autre conséquence de l'insertion dans une séquence génique. Deux cas de maladies impliquant ce type d'évènement ont été décrits (Table 1).

L'insertion d'un L1 dans le gène *ABHD5* qui est impliqué dans la biogenèse des triacylglycérols et la régulation de la lipolyse dans les adipocytes induit le syndrome de Chanarin-Dorfman, une maladie rare de surcharge en triglycérides. Cette insertion de 101 pb a lieu dans l'intron 3 du gène *ABHD5* (Figure 25). De par la présence de 2 sites d'épissage accepteur et donneur en début et en fin de cette séquence, le L1 sera retenu après épissage entre les exons 3 et 4 du gène. Ce variant ne pourra pas coder une protéine fonctionnelle induisant la surcharge en triglycérides observé chez ces patients (Samuelov *et al.*, 2011).

- *Maladies induites par terminaison prématurée du transcrit.*

Les insertions de L1 peuvent induire la terminaison prématurée de la transcription par l'introduction d'un signal de polyadénylation prématuré. Trois cas de maladies impliquant ce type d'évènement ont été rapportés (Table 1).

Par exemple, un cas d'hémophilie de type B est dû à l'insertion d'un L1 tronqué de 288 pb dans l'exon 7 du gène *FIX* codant pour le facteur IX de la coagulation. Ce L1 est tronqué en 5' et contient une séquence polyA de 163 pb induisant la terminaison prématurée du transcrit *FIX* (Figure 26). Ceci va induire la production d'un facteur IX de la coagulation tronqué qui ne sera plus fonctionnel (Mukherjee *et al.*, 2004).

- *Maladies induites par perte de matériel génétique.*

A cause des rétrotranspositions de L1, une partie du matériel génétique au site d'insertion peut être perdue. Quatre cas de maladies impliquant ce type d'évènement ont été identifiés (Table 1). L'insertion d'un L1 pleine dans le gène *PDHX* induit une perte de 46 Kb. L'insertion a été réalisée entre l'intron 2 et l'intron 9 en induisant la délétion des exons 3 à 9 (Figure 27). Cette

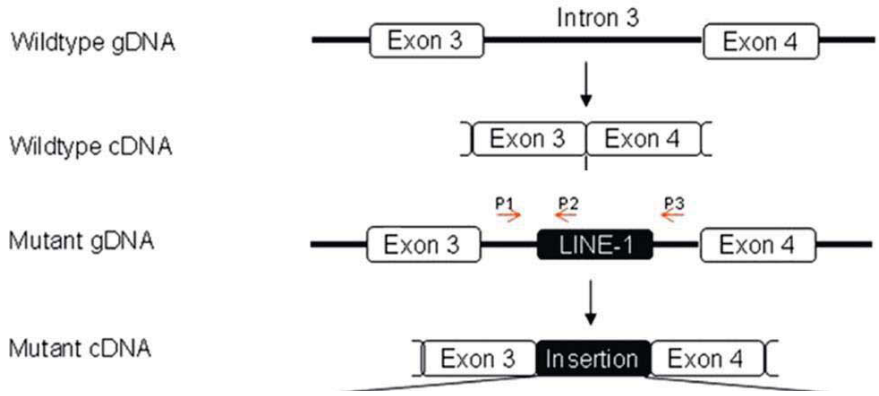


Figure 25 : Création d'un nouvel exon par insertion d'un L1 dans le gène ABHD5 induisant le syndrome de Chanarin-Dorfman.

Représentation schématique de l'insertion d'un L1 tronqué dans l'intron 3 du gène *ABHD5*. Le gène sauvage est schématisé au niveau des exons 3 et 4 séparés par l'intron 3. Le transcrit résultant de l'épissage exclu l'intron 3. Pour le gène mutant dans lequel s'en inséré le L1 au niveau de l'intron 3, de par la présence de sites accepteur et donneur d'épissage dans le L1, celui-ci va être retenu dans le transcrit entre les exons 3 et 4.*(Samuelov* et al.*, 2011)*

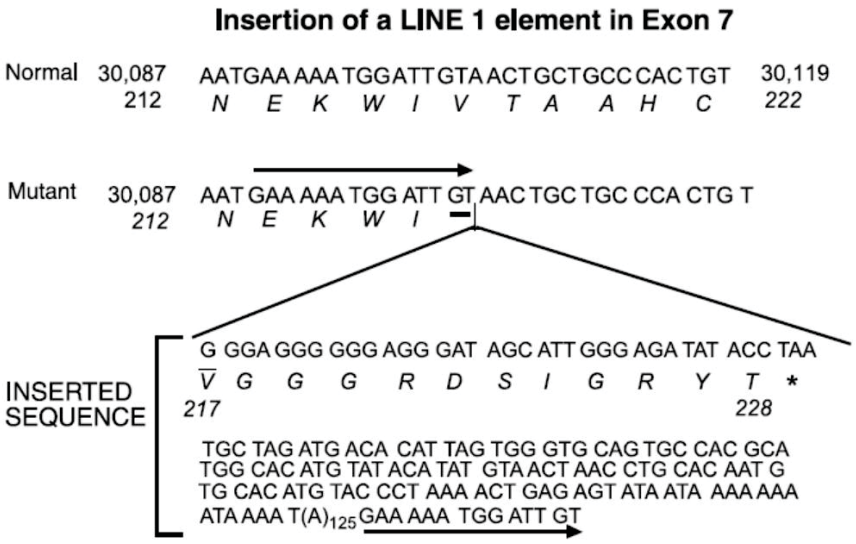


***Figure 26 :* Terminaison prématurée de la transcription par l'insertion d'un L1 dans le gène *FIX induisant l'Hémophilie B*.**

Insertion d'un L1 tronqué en 5' de 288 pb possédant une queue polyA en 3' de 125 nucléotides au niveau de l'exon 7 du gène *FIX*. Ce gène code pour le facteur IX de la coagulation, qui sera tronqué et non fonctionnel. *(Mukherjee* et al.*, 2004)*

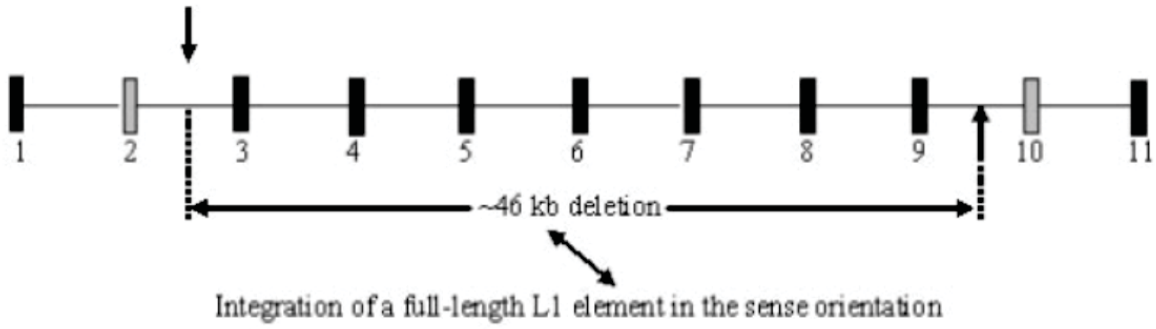


Figure 27 : Délétion de 46Kb due à l'insertion d'un L1 dans le gène PDHX.

Représentation schématique de l'allèle mutant du gène *PDHX*. Le point de cassure en 5' est localisé dans l'intron 2 et le point de cassure en 3' est localisé dans l'intron 9 (indiqués par les flèches verticales). Un L1 pleine taille s'est inséré induisant ainsi la perte des exons 3 à 9. Cet évènement est associé au développement d'une maladie impliquée dans un syndrome de déficience en pyruvate déshydrogénase (Miné *et al.*, 2007).

délétion empêche la production d'une sous-unité du complexe pyruvate déshydrogénase, E3BP, et est associé au développement d'une maladie induisant une déficience de ce complexe impliqué dans la chaine respiratoire mitochondriale (Miné *et al.*, 2007).

3.2. Transposition somatique des LINE-1 et cancer.

3.2.1. Méthylation ADN et cancer.

Au-delà d'un réseau complexe de facteurs de transcription, le programme du développement s'appuie sur un ensemble de modifications épigénétiques qui contribuent à la dynamique de l'expression des gènes ainsi qu'à la stabilité du génome.

L'une des marques épigénétique majeure est la méthylation de l'ADN. Dans les cellules de mammifères, la méthylation est distribuée, par défaut, sur l'ensemble du génome, couvrant les régions intra- et intergéniques. Paradoxalement, seuls des clusters de dinucléotides CpG (50 à 200 pb de long), nommés ilots CpG, sont non méthylés. Près de 29 000 de ces ilots sont dénombrés chez l'Homme, principalement associés à des promoteurs de gènes (60% des promoteurs ont des ilots CpG) ou des régions régulatrices. Pour la plupart, ces ilots sont maintenus déméthylés tout au long du développement somatique, quel que soit l'état d'expression du gène qu'il contrôle. Cette distribution exhaustive de la méthylation permet de maintenir le génome intègre, en particulier *via* son implication dans la régulation des ET.

Dans les cellules cancéreuses, le profil de méthylation est totalement remanié, aboutissant à un profil « miroir » de celui observé dans les cellules saines. Ainsi une diminution globale de la méthylation ADN est observée dans les cancers (Feinberg & Vogelstein, 1983) et elle affecte notamment les séquences des ET. Cette hypométhylation est associée à une hyperméthylation aberrante sur de nombreux ilots CpG. Ce phénotype « d'hyperméthylation des ilots CpG » (ou CpG Island Methylator Phenotype, CIMP) est décrit dans de nombreux types de cancers, définissant des sous-classes moléculaires de tumeurs (Noushmehr *et al.*, 2010). Ce phénotype correspond à un gain de méthylation vaste et coordonné à des ilots spécifiques.

Dans de nombreux types tumoraux, il a été montré que l'hypométhylation globale de l'ADN tumoral corrèle avec une hypométhylation au niveau des promoteurs de L1 (Piskareva *et al.*, 2011). De plus, celle-ci est associée généralement à un mauvais pronostic (Baba *et al.*, 2014). Cette hypométhylation au niveau de la région promotrice des L1 va pouvoir être en partie associée à l'activité des promoteurs et à la rétrotransposition des L1 par levée d'inhibition. Lee et ses collaborateurs ont mis en évidence dans le cancer colorectal où une hypométhylation

globale de l'ADN est observée que les sites d'insertion des L1 étaient significativement surreprésentés dans les régions hypométhylées (Lee *et al.*, 2012). Ceci est faveur d'une dérégulation concertée dans les tumeurs permettant la progression tumorale.

3.2.2. LINE-1, hypométhylation et rétrotransposition tumorale.

Les cancers sont des maladies où de nombreux remaniements ont lieu. Une hypométhylation est fréquemment observée dans les cancers notamment au niveau des L1. Celle-ci peut conduire à l'activation de la transcription des L1.

Aussi, l'expression de la protéine ORF1p est une marque des cancers. En effet, 90% des cancers du sein et des ovaires ainsi que des cancers pancréatiques, 50-60% des cancers du tractus gastro-intestinal (cancers du côlon et de l'œsophage), des poumons et 40% des cancers de la prostate ont une surexpression de la protéine ORF1p. En revanche, les glioblastomes et les lymphomes B de bas grade ne l'expriment que faiblement (Rodić *et al.*, 2014). Cette expression dépend de l'origine de la tumeur, de son profil d'expression des ET, de la mutation du gène *p53* ainsi que du profil de méthylation qui vont augmenter sa production. Aussi, il a été proposé que les protéines ORF1p mais aussi ORF2p peuvent servir de biomarqueurs pour une étude non-invasive (Burns, 2017).

La rétrotransposition des L1 dans les cancers est un évènement bien documenté malgré de nombreuses questions restantes sur sa cause et ses conséquences. Le premier cas de rétrotransposition somatique dans les cancers a été rapporté par Miki et ses collaborateurs dans le cancer colorectal. Ils observent une insertion d'une partie 3' de L1 dans le dernier exon du gène *APC* conduisant à sa perte de fonction. Ce gène *APC* est l'un des principaux gènes suppresseurs de tumeur et sa perte de fonction induit la transformation tumorale (Miki *et al.*, 1992). Par la suite, de nombreuses insertions ont été observées dans différents types de cancers, aussi bien colorectal, que des cancers des poumons, de la prostate et des cancers ovariens (Miousse & Koturbash, 2015). Grâce aux avancées des techniques de séquençage haut débit et au développement d'un nouvel outil bio-informatique, Lee et ses collaborateurs ont étudié des nouvelles insertions somatiques de L1 dans les cancers provenant de 43 patients atteints de cancers colorectaux, de la prostate, des ovaires, de myélome multiple et de glioblastomes. Les échantillons de sang correspondant à chaque patient ont été prélevés, afin d'extraire les cellules normales qui serviront de référence à l'identification de nouvelles insertions de L1 dans l'ADN tumoral. Ainsi, ils identifient 183 nouvelles insertions somatiques de L1. Le plus grand nombre de nouvelles insertions est observé dans les tumeurs d'origine épithéliale (colorectale, prostate

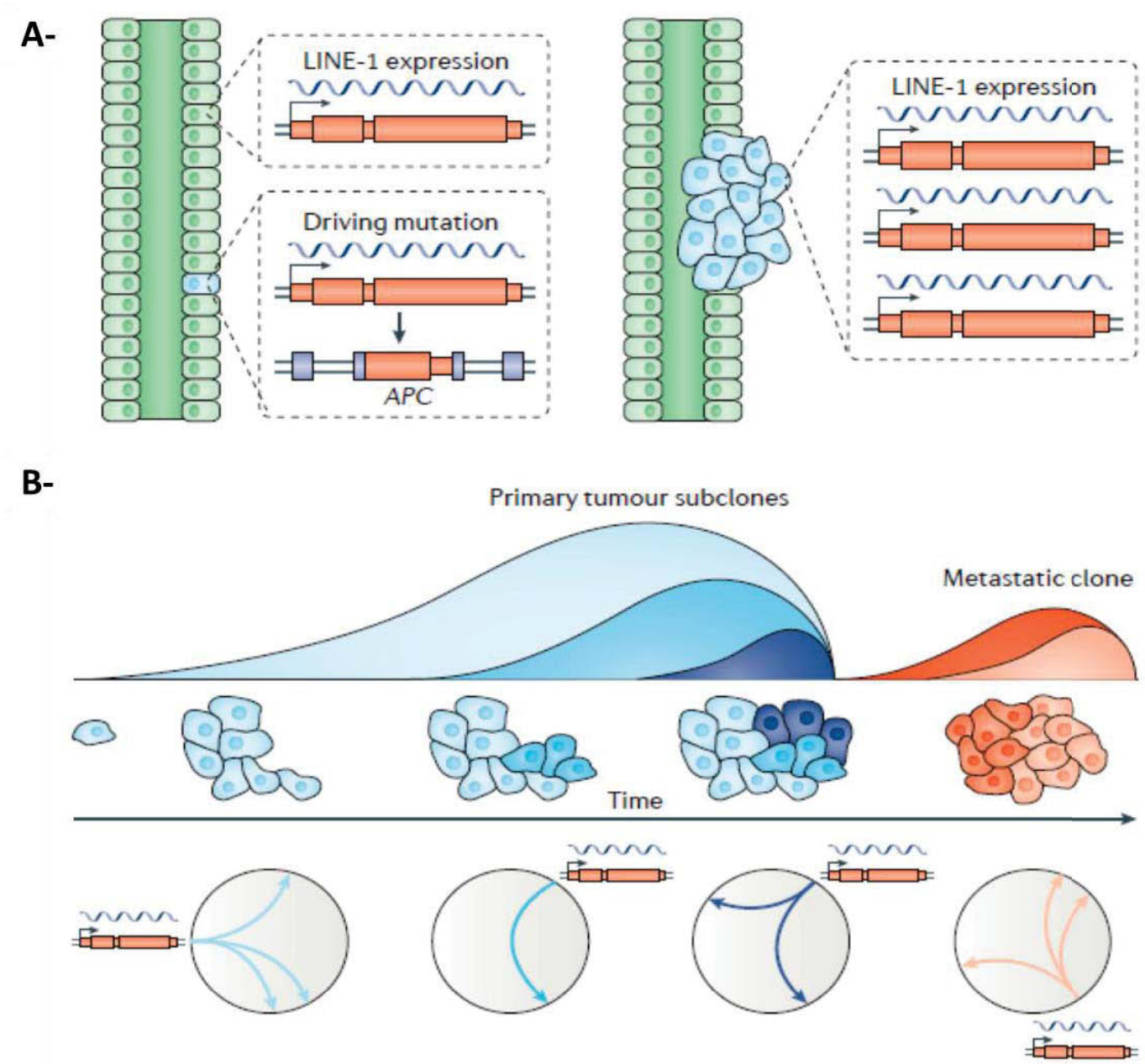


Figure 28 : Insertion accrue des L1 dans les cellules cancéreuses

A- **Les cellules précancéreuses et normales autorisent l'expression des L1**, bien que les mécanismes permettant ceci ne sont pas encore clairs. A gauche : le diagramme montre un L1 source qui échappe à la mise sous silence somatique dans les tissus épithéliaux normaux délimitant le tractus gastro-intestinal (indiqué par les cellules vertes). Dans la majorité des cellules, il n'y a pas d'effet. Dans une rare cellule (en bleu), cette expression va générer une insertion qui va promouvoir le développement cancéreux, comme identifié au locus *APC* (Adenomatous Polyposis Coli). A droite : Ces cellules qui ont initiées la transformation tumorale prolifèrent et l'expression des L1 est encore plus permissive dans ces cellules.

B- **Modèle montrant comment les L1 transitoirement exprimés peuvent contribuer à l'évolution génétique des cancers**. L'évolution clonale d'une tumeur primaire et de ses métastases sont représentées sur l'axe des x. Quatre L1 sources successifs se succèdent comme illustré et deviennent actifs avec le développement tumoral ; il y a un code couleur afin de déterminer quel L1 est actif dans quelle sous-population. Les cercles représentent schématiquement les localisations génomiques de la mobilisation des L1 entre leur L1 source et l'acquisition de nouvelles insertions (flèches). L'hétérogénéité génétique de la population tumorale, résultant d'acquisition de nouvelles insertions de L1, forment un cluster de cellules tumorales. (Burns, 2017)

et ovaires) alors qu'ils n'en détectent pas dans les cancers du cerveau. Une insertion préférentielle est observée dans les gènes suppresseurs de tumeurs comme par exemple *NELL1*, *DBC1*, *ROBO2* et *PARK2*, ainsi que les gènes associés à des fonctions d'adhésion cellulaire, comme *CDH12*, *ROBO2*, *NRXN3*, *FPR2*, *COL11A1*, *NEGR1*, *NTM* et *CTNNA2*. L'insertion des L1 va pouvoir induire des délétions de ces gènes ou bien leur mise sous silence *via* des mécanismes épigénétiques. La diminution de l'expression de ces gènes va ainsi favoriser la transformation tumorale (Lee *et al.*, 2012). Ceci suggère une potentielle contribution des insertions d'ET dans le développement des cancers. Par ailleurs, dans cette même étude, en observant les séquences de reconnaissance (TSD) de l'insertion des L1 par l'EN, les auteurs mettent en évidence que les néo-insertions tumorales ne sont pas seulement liées à la rétrotransposition conduite par l'EN du L1, mais également à une insertion au niveau d'un site de cassure double brin de l'ADN (Figure 3) (Lee *et al.*, 2012).

Au sein des tumeurs, ces évènements d'expression et de rétrotransposition des L1 semblent avoir lieu selon des restrictions spatiales et temporales et varient au cours du développement tumoral. Une étude utilisant les transductions en 3' pour identifier les éléments sources montre que ces « hot L1 » ne sont pas actifs simultanément dans la tumeur. A la place, les différentes phases de croissance de la tumeur vont servir à l'expression de différents éléments sources à l'origine de la progression tumorale (Tubio *et al.*, 2014). Par exemple, dans l'évolution des cancers pancréatiques, la tumeur primaire va acquérir un grand nombre d'insertions somatiques de L1 alors que dans les métastases associées, il n'y aura pas d'insertion supplémentaire. Ceci suggère une capacité hétérogène et épisodique de l'activité des L1 dans les cancers (Figure 28) (Rodić *et al.*, 2014).

Bien que la majorité de ces insertions aient lieu préférentiellement dans des régions introniques ou intergéniques, quand elles s'insèrent à proximité d'un gène, celle-ci vont pouvoir induire :

- Des délétions dans la séquence codante.
- La formation de nouveaux transcrits impliqués dans la transformation tumorale dûe par exemple à des transductions en 3'.
- Leur mise sous silence *via* des mécanismes épigénétiques, tels que par exemple le recrutement d'ADN méthyltransférases ou histones méthyltransférases.

Les insertions introniques ou intergéniques peuvent jouer un rôle dans la dérégulation des sites de liaisons pour des facteurs de transcription ou des séquences régulatrices. Ceci aboutit au même constat, c'est-à-dire que cette insertion induit la perturbation de la transcription des gènes adjacents promouvant ainsi le mécanisme de transformation tumorale.

3.3. Autres implications des LINE-1 en pathologie : les LCT et les cancers.

En contexte tumoral, suite à l'hypométhylation affectant notamment les ET, il peut se produire une levée d'inhibition au niveau de la région 5'-UTR du L1 conduisant ainsi non seulement à l'activation du promoteur sens, mais aussi à l'activation du promoteur antisens (Speek, 2001). Cette transcription à partir du promoteur antisens va pouvoir se poursuivre dans la région génomique adjacente produisant ainsi un transcrit chimère, nommé LCT (L1 Chimeric Transcript), qui sera constitué en partie de la région promotrice du L1 en 5', en antisens, associée à la région génomique unique adjacente. Cet ASP se comporte alors comme un promoteur alternatif pouvant perturber la transcription de la région dans laquelle il se trouve (Speek, 2001). Différentes études ont prouvé conceptuellement l'existence de tels transcrits.

Dans l'étude où un promoteur ASP est mis en évidence dans la région 5'UTR du L1, Speek identifie également dans la base de données GenBank, 4 ADNc ayant une configuration de LCT avec en 5' la séquence du L1 en antisens et en 3' une région génomique unique, se poursuivant dans les gènes *GAP-43*, *CHRM3*, *SPT3* et *OATP*. Grâce à une expérience de protection RNase, il met en évidence la présence de ces LCT dans des lignées cellulaires NTera2D1, JEG3 et HeLa (Speek, 2001).

Suite à cela, l'équipe de Speek met en place une approche bio-informatique pour identifier des LCT dans la base de données EST de GenBank (Figure 29-A). Cette approche bio-informatique utilise le programme WU-BLAST2 pour identifier des EST composés en 5' de la région 5'UTR du L1 en antisens à partir de la séquence consensus d'un L1HS et en 3' d'une région unique du génome. Ils identifient ainsi 42 EST ayant une configuration de transcrit chimère. Parmi eux, 23 EST dérivent de tissus normaux, à savoir, 10 proviennent de foie et de rate fœtaux ; 2 proviennent de placenta, de tissu mammaire, de la rétine fœtale, et des poumons ; 1 retrouvé dans le muscle, les testicules, la prostate, le cerveau et les neurones. 19 EST sont identifiés dans de nombreux types tumoraux à savoir 7 provenant de tératocarcinome ; 4 de carcinome du côlon ; 3 de néoplasie prostatique ; 1 retrouvé dans chacune des tumeurs suivantes : phéochromocytome, liposarcome, cellules-B, tumeurs génitales et ovariennes (Nigumann *et al.*, 2002). L'analyse de la séquence unique en 3' en utilisant les programme BLASTN et BLASTX a identifié 25 EST chimères possédant des ORF similaires avec des protéines connues. Parmi eux, ils en ont sélectionné 10 pour lesquels un ARNm était connu et pour 4 d'entre eux, un épissage a lieu et se poursuit dans les exons du gène dans lequel est situé le L1. Ces 4 EST sont localisés dans les gènes *MET* codant pour un récepteur tyrosine kinase aux facteurs de

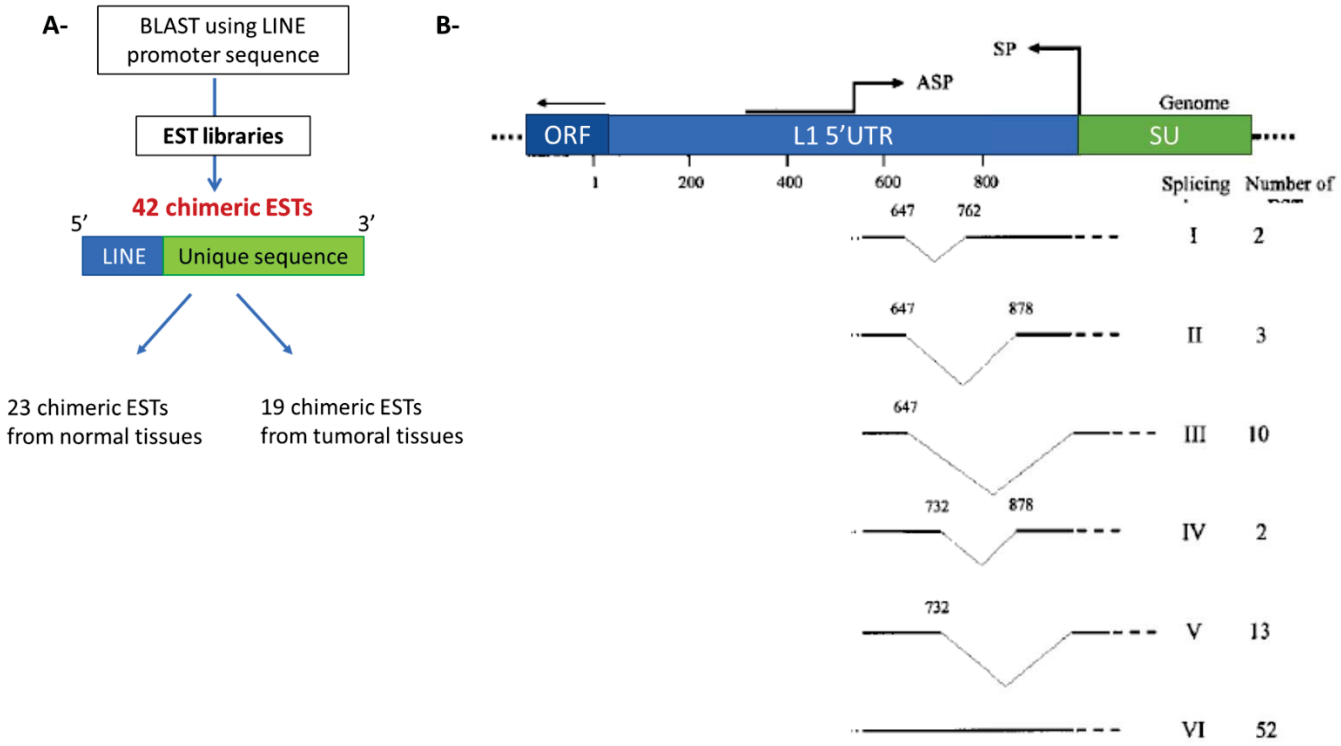


Figure 29 : Identification de transcrits chimères dans des banques d'EST et de variants d'épissage.

A- En utilisant le programme BLAST et RepeatMasker, des EST chimères sont identifiées dans les bases de données GenBank avec en 5' une séquence de L1 en antisens et en 3' une séquence unique du génome.

B- Les EST sont épissés selon 6 schémas différents (noté I à VI), représentés par la position des sites donneur et accepteur d'épissage localisé dans la région 5'UTR du L1.

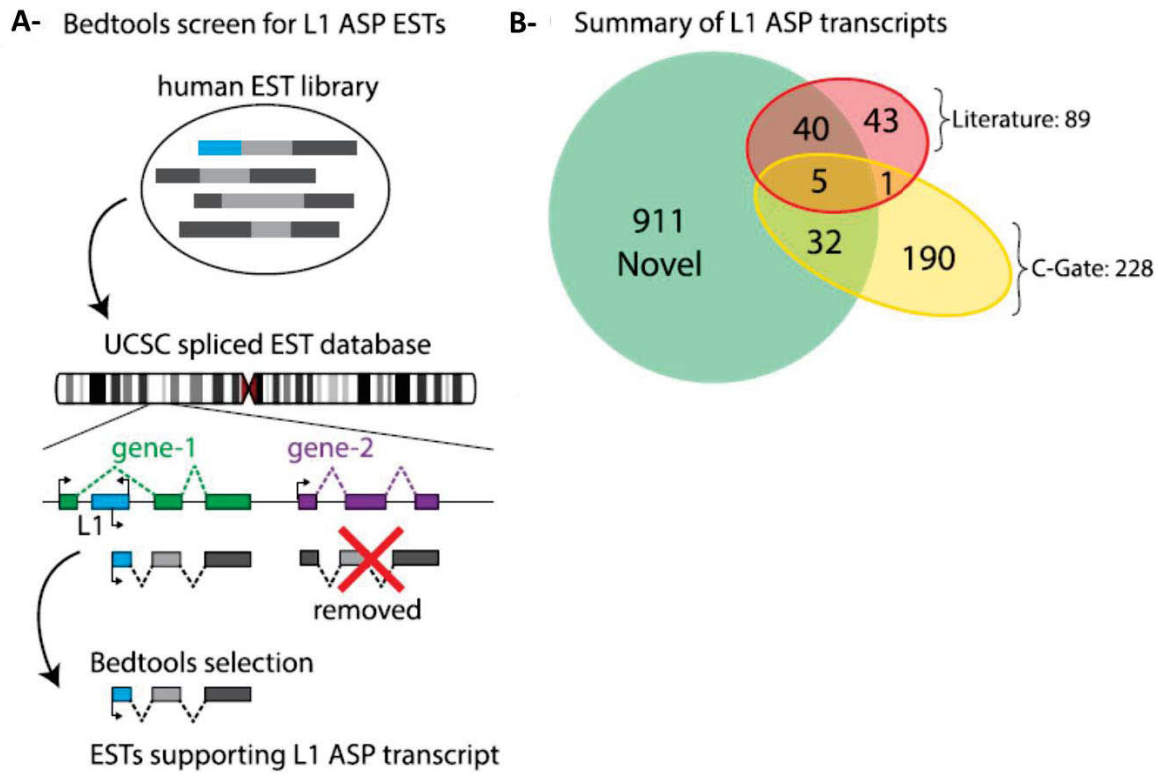


Figure 30 : Identification de nouveaux LCT utilisant un pipeline bio-informatique.

A- Représentation schématique de la méthode utilisée pour identifier des LCT. Les coordonnées des EST épissées chez l'Homme sont croisées avec les exons annotés des gènes (selon la base Refseq) pour identifier les EST avec un site d'initiation de la transcription en antisens du L1 et se poursuit dans l'exon d'un gène connu. Les LCT qui montrent des EST correspondant à l'exonisation de L1 sont éliminés.

B- Le nombre total de LCT identifié ici est de n=988 et est croisé avec les données de la littérature et la base de données C-GATE. (Criscione *et al.*, 2016)

croissance des hépatocytes, *DNCI-1* codant pour la chaine intermédiaire 1 de la dynéine, *SCAMP* (Secretory carrier-associated membrane protein) et *TACTILE* qui code pour une protéine de surface de lymphocyte T.

Grâce à l'analyse des séquences de ces EST ayant une configuration de LCT, différents mécanismes d'épissage ont été mis en évidence. Ces évènements d'épissage ont lieu entre 2 sites donneurs et 2 sites accepteurs localisés, en aval de l'ASP, aux positions +647, +732, +762 et +878 dont les positions sont calculées par rapport à un site +1 positionné en début d'ORF1 en antisens. Aussi pour 82 EST décrite dans cette étude, 6 schémas d'épissage sont identifiés (Figure 29-B). Le premier schéma, identifié dans 2 EST, correspond à un épissage entre le site donneur en position +647 avec le site accepteur en position +762 (scheme I). Le second schéma, identifié dans 3 EST, correspond à un épissage entre le site donneur en position +647 avec le site accepteur en position +878 (scheme II). Le schéma III, identifié dans 10 EST, correspond à un épissage entre le site donneur +647 du L1 et un site accepteur inconnu localisé dans la séquence unique adjacente (scheme III). Le quatrième schéma, identifié dans 2 EST, indique un épissage entre le site donneur +732 et le site accepteur en +878 (scheme IV). Le schéma V, identifié dans 13 EST, montre un évènement d'épissage entre le site donneur +732 du L1 et un site accepteur inconnu localisé dans la séquence unique adjacente (scheme V). Enfin le dernier schéma, identifié dans 52 EST, correspond à une absence d'épissage (scheme VI). Au final, la majorité des EST identifiés ne subissent pas d'épissage. Quand un évènement d'épissage a lieu, celui-ci implique un des deux sites donneurs localisés dans le L1 et un site accepteur inconnu présent dans la région génomique adjacente (Nigumann *et al.*, 2002).

Suite à cette étude, l'équipe de Speek en a réalisé une seconde en utilisant la même technique que précédemment (Mätlik *et al.*, 2006) sur les bases de données GenBank (mise à jour en 2004), EMBL et DDBJ. Ils ont ainsi identifié 49 EST chimères. Parmi elles, ils valident pour 6 EST que l'ASP servirait de promoteur alternatif dont la transcription se poursuit dans les gènes *KIAA1797*, *CLCN5*, *SLCO1A2*, *MET*, *COL11A1* et *BOLL.* Des approches de RT-PCR sur 16 tissus normaux (thymus, prostate, rate, intestin, côlon, ovaire, testicule, leucocytes du sang périphérique, placenta, muscle squelettique, cerveau, rein, cœur, poumon, pancréas et foie) valident l'expression de ces EST. Par exemple, le transcrit *L1-KIAA1797* est exprimé dans les poumons et le placenta, alors que le transcrit du gène *KIAA1797* est exprimé dans les testicules, le placenta et le foie, ou encore le *L1-CLCN5* exprimé uniquement dans le placenta alors que le transcrit du gène *CLCN5* est uniquement exprimés dans les poumons. De plus, des variants d'épissage sont détectés, comme par exemple la transcription à partir de l'ASP localisé dans le

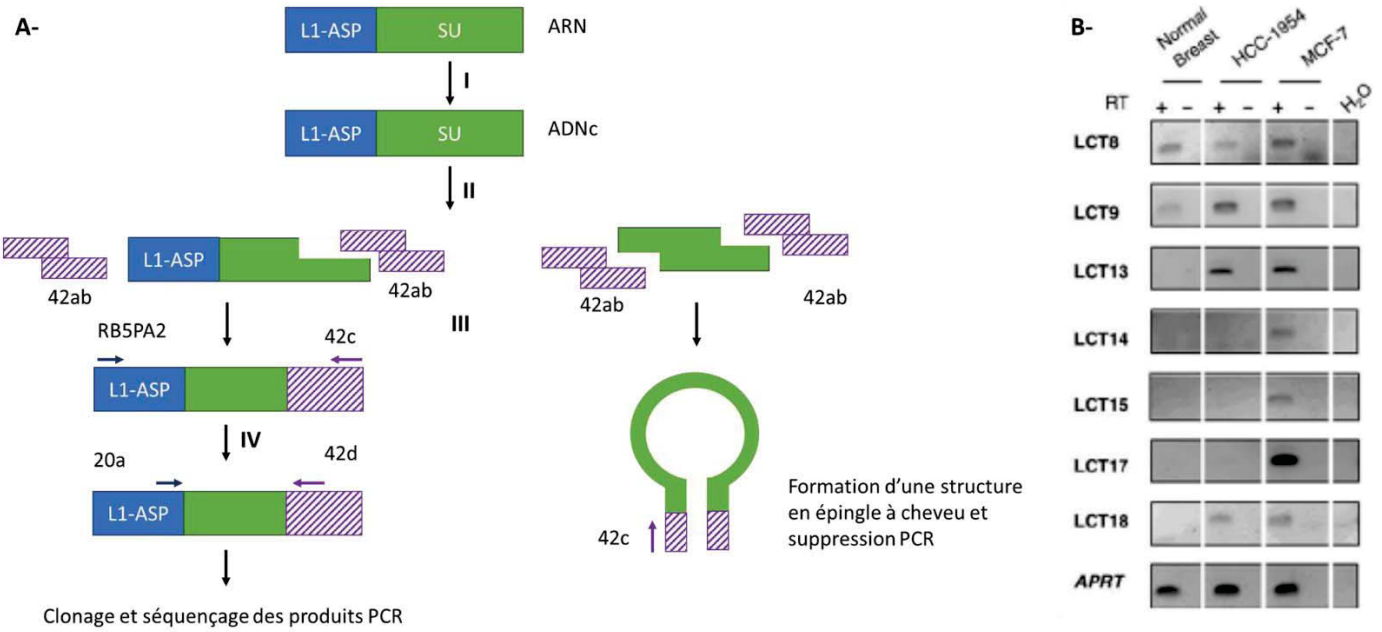


Figure 31 : Représentation schématique de la technique de LCD (L1 Chimera Display).

A- Après avoir converti l'ARN en ADNc (I), cet ADNc est digéré avec une enzyme de restriction (ici XbaI) qui ne coupe pas dans la région 5'UTR du L1 (L1-ASP) et des adaptateurs complémentaires (42ab) sont liés aux extrémités sortantes (II). Ces adaptateurs ont un fort contenu en dinucléotides C et G autorisant ainsi la formation de structures en épingles à cheveux si des adaptateurs sont liés aux 2 extrémités du même ADNc, ce qui va permettre l'amplification spécifique des LCT avec les amorces RB5PA2 localisée dans le L1 et 42c localisée dans l'adaptateur (III). Une PCR nichée est réalisée avec les amorces 20a localisée dans le L1 et 42d localisée dans l'adaptateur (IV). Les produits ainsi obtenus sont clonés et séquencés.

B- Validation de 7 LCT dans des lignées de cancer du sein (HCC-1954 et MCF-7) avec une RT-PCR brin spécifique sur l'ARN total extrait de cellules normales et des lignées cellulaires cancéreuses. Une RT- sert de contrôle pour exclure une contamination par l'ADNg et l'amplification du gène de ménage *APRT* sert de contrôle positif de RT. (Cruickshanks & Tufarelli, 2009)

gène *BOLL*, donnent deux variants selon le schéma III ou V (Figure 29-B). Cet ASP peut servir de promoteur alternatif pour de nombreux gènes dans les tissus normaux, comme dans les tissus tumoraux avec une régulation tissus-spécifique, puisque l'expression des gènes et des LCT ne se fait pas dans les mêmes tissus à l'image du gène *CLCN5* et du *L1- CLCN5*.

Des mécanismes d'ARN interférence font intervenir un ARN antisens qui participe à la dégradation de l'ARNm ciblé permettant ainsi de réguler l'expression cellulaire des gènes. Des mécanismes similaires peuvent servir à la régulation de la transcription des L1 à partir de l'ASP dans les cellules normales et tumorales (Mätlik *et al.*, 2006).

Enfin, Criscione et ses collaborateurs ont identifié par des approches bio-informatiques des LCT dans des banques d'EST, utilisant un pipeline où 1) la partie 5' de l'EST doit correspondre à la région 5'UTR du L1 en antisens, 2) la partie 3' doit correspondre à un évènement d'épissage avec une séquence annotée comme exon (RefSeq), 3) ces EST identifiés sont filtrés avec l'outil BEDtool pour éliminer ceux qui ne correspondraient pas à la définition d'un LCT mais serait dû à une exonisation partielle d'un L1 (Figure 30-A). Ainsi 2015LCT ont été identifiés correspondant à 988 gènes dont 911 n'ont jamais été décrites dans la littérature (Figure 30-B). Parmi ces EST_LCT, 609 sont représentés uniquement par 1 EST dans la base de données, ce qui suggère que les LCT sont transcrits à des taux relativement faibles, et 379 sont représentées par de multiples EST, comme par exemple ceux localisés dans les gènes *CPM* (99 EST), *BCAS3* (49 EST) et *DDX39B* (42 EST). Un grand nombre d'EST_LCT est retrouvé dans les tissus normaux, notamment le cerveau et les testicules (Criscione *et al.*, 2016). De plus, ces EST_LCT impliquent des L1 de nombreuses sous familles, incluant des sous familles L1 très anciennes. Ainsi, 18% sont associés à des L1 des sous familles récentes L1PA1 à L1PA8, 8% sont associés à des L1 primates plus anciens et 74% sont associés à des L1 mammifères très anciens. L'étude de l'initiation de la transcription des EST_LCT en fonction de l'origine du L1 associé montre que celle-ci se fait dans la région 5'UTR (où a été décrit le promoteur ASP) uniquement pour les L1 des sous familles récentes. Pour les EST_LCT des sous familles plus anciennes l'initiation de la transcription en antisens se produit alors à partir de la région 3' des éléments. Cette observation conforte l'idée que l'ASP dans la région 5'UTR est présent uniquement dans les sous familles récentes de L1.

Malgré ces preuves de concept, toutes les études précédentes sont basées sur l'analyse de base de données d'EST. Aussi, l'implication des transcrits chimères à l'échelle d'un échantillon tumoral n'a pas pu être appréhendée. En 2009, Cruickshancks et Tufarelli mettent au point une technique nommée LCD (L1 Chimera Display) qui a pour but d'identifier des LCT initiés à

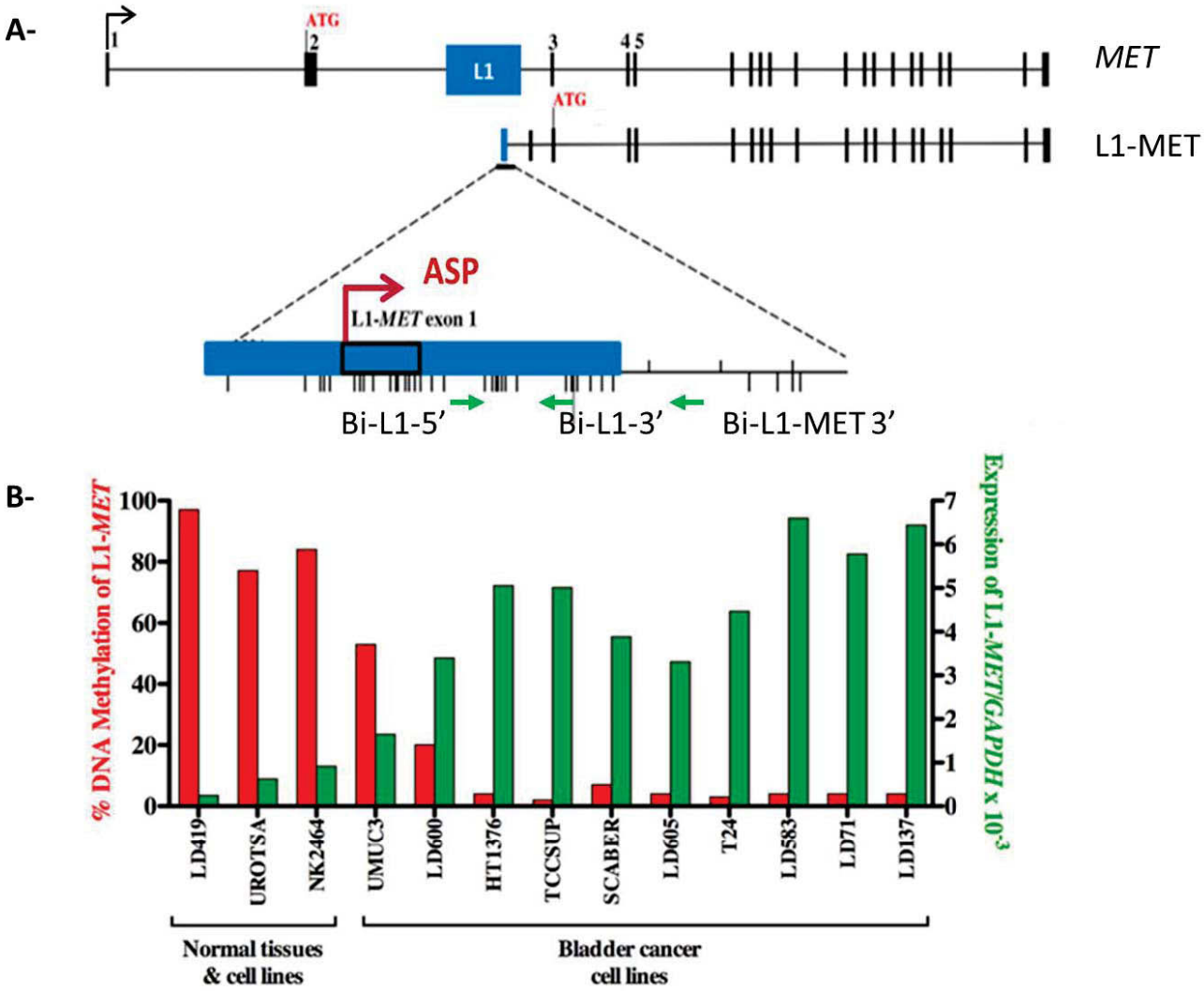


Figure 32 : Méthylation et expression du transcrit L1-MET corrèlent dans les lignées cellulaires de vessie.

A- Le L1PA2 est localisé dans l'intron 2 du gène *MET* avec le promoteur ASP orienté dans le sens de la transcription du gène. L'initiation de la transcription à partir de l'ASP permet la formation d'un transcrit tronqué L1-MET. Les barres noires sous le L1 représentent les CpG qui peuvent être méthylés et les flèches vertes correspondent aux amorces utilisées pour la quantification de la méthylation par séquençage bisulfite.

B- Corrélation inversement proportionnelle entre le pourcentage de méthylation au niveau du promoteur ASP du transcrit L1-MET en rouge et son expression en vert dans des tissus normaux de vessie et des lignées cellulaires du cancer de la vessie. (Wolff *et al.*, 2010)

l'ASP dans des échantillons humains (Figure 31-A). Cette technique utilise pour cibler l'extrémité L1 des LCT une amorce (RB5PA2) localisée en position +72 à +92 du promoteur L1 qui a été dessinée sur la base de la séquence des L1HS. Pour autant, cette amorce peut identifier des chimères impliquant d'autres sous-familles récentes qui possèdent une forte homologie de séquence avec le L1HS. De plus, ce positionnement ne permet pas de capturer des LCT ayant subi un mécanisme d'épissage mais permet la capture uniquement d'évènement de type schéma VI (Figure 29-B). Aussi la technique LCD présente des biais qui ne permettent pas une identification exhaustive de tous les transcrits chimères d'un échantillon. Malgré cela, les auteurs ont pu mettre en évidence la présence de 18 LCT dans des lignées cellulaires de tumeurs mammaires MCF-7 et HCC-1954 ainsi que l'expression de certains dans les tissus normaux correspondants. Leur initiation de la transcription à l'ASP a été validée par des expériences de 5'RACE. Parmi ces 18 LCT, 8 sont uniquement constitués de séquences répétées (L1, LTR, Alu…). Parmi les 10 restants, des amorces spécifiques ont pu être dessinées pour 7 d'entre eux pour leur amplification par RT-PCR (Figure 31-B). Ainsi 5 LCT sont exprimés spécifiquement dans les lignées cellulaires de cancer du sein et 2 sont exprimés dans les tumeurs et les contrôles (LCT8 et LCT9). La même expérience a été réalisée pour des lignées cellulaires de cancer colorectal HCT-116 et SW-480 et tous les LCT à l'exception d'un sont tumeurs spécifiques. Il semble donc qu'il y ait une régulation tissus-spécifique de l'activation de la transcription à partir de l'ASP.

Les auteurs se sont donc demandé ce qui pouvait réguler l'activation de cet ASP. Ils ont regardé la méthylation de la région 5'UTR des L1, sachant que l'hypométhylation est un évènement récurrent dans les cancers. Par des traitements au 5-aza de cellules MCF10 transfectées avec un plasmide contenant le promoteur ASP contrôlant l'expression d'un gène rapporteur, ils montrent une corrélation inverse entre la diminution de la méthylation et l'augmentation de la transcription à partir de l'ASP (Cruickshanks & Tufarelli, 2009).

Par la suite des analyses fonctionnelles ont mis en évidence une corrélation inverse entre la méthylation de l'ASP et l'expression d'un LCT. En effet, un LCT initié à partir d'un promoteur antisens localisé dans intron 2 du gène *MET*, qui avait été décrit par Speek dans les banques d'EST, a été retrouvé exprimé dans des cellules normales de vessie et des lignées cellulaires de cancer de la vessie (Figure 32-A). Une corrélation inverse a été établie entre la diminution de la méthylation de la région promotrice du L1 et l'augmentation de l'expression du transcrit *L1-MET* (Figure 32-B). En d'autres termes, dans les cellules normales la région promotrice du L1 est méthylée et associée à une faible expression du transcrit *L1-MET* alors que dans des lignées

cellulaires du cancer de la vessie, la méthylation de la région promotrice du L1 est diminuée associée à une augmentation du transcrit *L1-MET*. Par ailleurs, les auteurs associent également à la diminution de la méthylation, le recrutement de marques d'histone activatrices, telles que H3K4me3 et des acétylations au niveau de l'histone H3 dans les lignées cellulaires au niveau de l'ASP du transcrit L1-MET (Wolff *et al.*, 2010).

Ces études questionnent sur des propriétés de l'ASP impliqué dans la formation de ces EST chimères. Ces EST chimères impliquent des L1HS récents et actifs mais aussi des L1 des sous-familles L1PA2 à 8. Est-ce que ces ASP possèdent une signature particulière induisant leur activation ?

Pour répondre à cette question, Criscione et ses collaborateurs se sont intéressés à la présence du site de fixation du facteur de transcription YY1, qui est localisé dans la région +13 à +21 de la région 5'UTR d'un L1HS sur le brin antisens et qui régule la transcription du promoteur sens (Athanikar *et al.*, 2004). Ce facteur de transcription YY1 est aussi important pour la régulation de la transcription bidirectionnelle de nombreux promoteurs de gènes (Gaston & Fried, 1994). Aussi peut-être que celui-ci est impliqué dans la régulation non seulement du promoteur sens mais aussi du promoteur antisens. Premièrement, grâce à un outil de prédiction de site d'interaction potentiel avec les facteurs de transcription (Jaspar Scan), 3 sites de fixation pour YY1 ont été identifiés dans la région 5'UTR du L1 aux positions +90 à +95, +448 à +453 et +870 à +875, dont un se situe au niveau du promoteur antisens et n'avait jamais été décrit auparavant. Deuxièmement, avec des données de ChIP-seq, 2 pics sont identifiés pour le facteur de transcription YY1, un au niveau du site proche du promoteur sens et un aux alentours de l'ASP. Troisièmement les auteurs ont croisé les données de ChIP-seq pour YY1, et les marques d'histones activatrices H3K4me2 et H3K4me3 avec des données de GRO-seq dans les lignées cellulaires K562 et HeLa. Ils identifient que les EST chimères identifiées sont majoritairement transcrites par un ASP actif dû à la présence de ces marques. Ensuite, 4 EST chimères ont été sélectionnés afin de valider si ceux-ci correspondent à des LCT, par des approches de biologie moléculaire combinant des RT-qPCR et du séquençage, pour les EST chimères *L1-KIAA1324L*, *L1-UVRAG* sur les cellules H116 et les testicules respectivement. L'expression des transcrits *L1-NF1* et *L1-SEC22B*, qui pourraient avoir un rôle dans les processus de tumorigenèse, a été testée sur la peau, le placenta, le côlon, le poumon, la rate, le foie, la vessie, le rein et le muscle où une expression différentielle est observée dans chacun de ces tissus. Ainsi au cours de cette étude, ils ont mis en évidence la présence de LCT dans les tissus normaux et tumoraux, initiés à partir d'ASP fonctionnels et actifs issus de L1 des sous-familles PA1 à 8, avec la liaison de

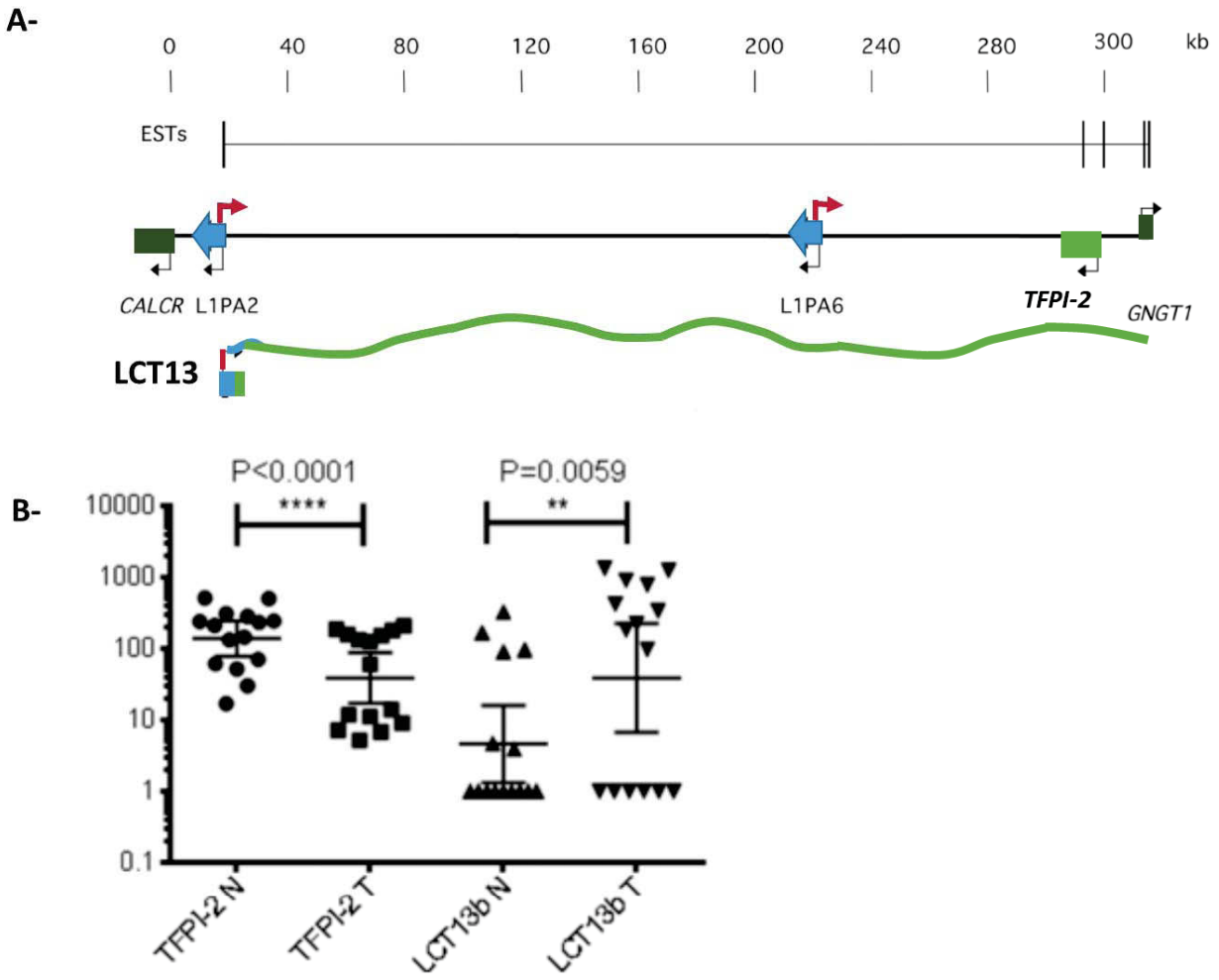


Figure 33 : Anti corrélation entre l'expression du LCT13 et du gène TFPI-2

A- Représentation schématique de la région de 300 Kb dans laquelle se situent le LCT13 et le gène *TFPI-2*. L'échelle est en kilobases. Deux éléments L1 sont présents dans cette région (L1PA2 et L1PA6). L'orientation de la transcription est indiquée par les flèches. Le LCT13 a été identifié précédemment (Cruickshanks & Tufarelli, 2009) et des expériences de 5'RACE ont montré l'initiation de sa transcription au niveau de l'ASP du L1PA2 intergénique.

B- Analyse de l'expression du LCT13b et du gène *TFPI-2* dans des cas sporadiques de cancers colorectaux provenant de patients et des tissus sains correspondants. Ceci montre une relation inverse entre l'augmentation de l'expression du LCT13b ($p<0.007$, test de Wilcoxon, n=15) et la diminution de l'expression du gène *TFPI-2* dans les tumeurs. Cette RT-qPCR a été normalisée avec les gènes de ménage *PGK1*, *GAPDH* et *HPRT*. (Cruickshanks *et al.*, 2013)

YY1 et la présence des marques d'histones activatrices H3K4me2 et H3K4me3. Ces LCT pourraient coder pour des protéines, comme ce qui a été décrit pour l'ORF0 mais n'a pas été mis en évidence ici (Criscione *et al.*, 2016).

Est-ce que ces LCT décrits peuvent jouer un rôle dans le développement tumoral ?

Ces LCT peuvent avoir un rôle fonctionnel dans la régulation de la transcription des gènes. Par exemple, un LCT nommé LCT13 localisé en position intergénique et dont l'initiation de la transcription se fait à l'ASP d'un L1PA2 forme un long transcrit non codant qui va induire l'inhibition de la transcription du gène *TFPI-2* situé 300Kb en aval (Cruickshanks *et al.*, 2013)(Figure 33-A). Suite à la mise en évidence de l'expression de ce LCT et la diminution d'expression du gène *TFPI-2* dans des tumeurs mammaires et colorectales, les auteurs ont corrélé et validé par des approches de siRNA ciblant ce LCT, que celui-ci inhibe la transcription du gène *TFPI-2.* De plus, l'analyse de l'expression du LCT13b et du gène *TFPI-2* dans des échantillons tumoraux de patients par RT-qPCR montre une relation inverse entre l'augmentation de l'expression du LCT13 et la diminution d'expression du gène *TFPI-2* dans les tumeurs (Figure 33-B). Le gène *TFPI-2* code pour un homologue d'inhibiteur de protéase de type Kunitz qui quand elle est sous-régulée dans les cancers, contribue à la tumorigénicité à cause de son rôle de remodelage de la matrice extracellulaire. Le LCT13 *via* la formation de ce long transcrit, va induire la déposition de marques d'histones répressives au niveau de la région promotrice de ce gène et ainsi induire sa sous-expression dans les tumeurs (Cruickshanks *et al.*, 2013). Cette étude démontre ainsi un autre rôle que peuvent avoir les LCT dans la modulation de l'expression des gènes et les cancers.

Malgré la validation de la présence des LCT dans les tumeurs mais aussi pour certains dans les échantillons contrôles, ces études ne permettent pas d'appréhender de façon globale l'étendue pangénomique de l'impact des LCT dans la tumorigenèse, de par des *a priori* sur les séquences identifiées. En effet, les analyses sur les banques d'EST identifient des LCT épissés mais au sein d'un ensemble de tissus. De plus, les outils bio-informatiques utilisés pour la détection de la partie L1 identifient majoritairement des L1 récents car la séquence consensus de L1HS est souvent utilisée comme référence. L'approche de LCD induit également un biais dans la détection de LCT non épissés et impliquant des L1 récents à cause du dessin des amorces. Ces études ont mis en évidence à la fois des LCT tumeurs spécifiques et des LCT exprimés dans les tissus normaux. Enfin ces études semblent montrer un lien entre la diminution de la méthylation au niveau de la région promotrice des L1 et l'augmentation d'expression des transcrits LCT dans les tumeurs. Aussi, il est supposé que le mécanisme impliqué dans l'expression des LCT

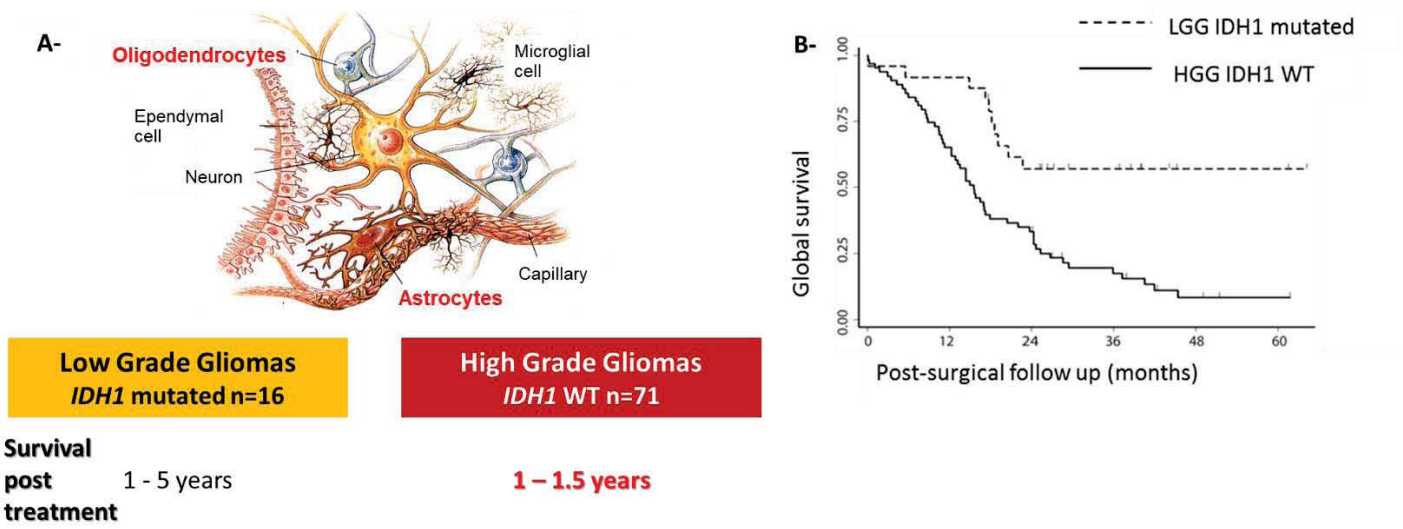


Figure 34 : Classification des gliomes

A- Les gliomes proviennent de la transformation tumorale des cellules de la glies (oligodendrocytes et astrocytes). On peut les classifier en gliomes de bas grade associés à la mutation du gène *IDH1* et à une meilleure survie, des gliomes de haut grade où le gène *IDH1* est sauvage et associé à une faible survie.

B- Courbe de Kaplan Meier de la survie des patients de 87 gliomes de la tumorothèque Auvergne Gliome selon le statut de la mutation *IDH1*.

serait dû à l'hypométhylation tumorale au niveau des L1, qui induirait l'activation de l'ASP et permettrait la fixation de facteurs de transcription produisant ainsi des LCT.

4. Modèle d'étude et objectifs du travail.

4.1. Modèle d'étude : les gliomes.

Les gliomes sont les tumeurs cérébrales primaires les plus fréquentes chez l'adulte avec une incidence de 3 cas sur 100 000 pour les gliomes les plus sévères. Ces tumeurs dérivent des cellules de la glie, notamment des astrocytes et des oligodendrocytes. Le système de classification de l'Organisation Mondiale de la Santé définit 4 grades de sévérité croissante selon des critères histopathologiques et cliniques strictes : du grade I, tumeurs bénignes, au grade IV, correspondant aux gliomes les plus sévères appelés aussi glioblastomes multiformes ou GBM (Palka *et al.*, 2012). Seulement 5% des GBM sont dits secondaires et dérivent de l'évolution des grades antérieurs. En effet, 95% des GBM surviennent *de novo* (Figure 34-A). Le traitement de ces tumeurs consiste en leur exérèse chirurgicale combinée à des traitements de chimio- et radiothérapies. A cause de leur prolifération et de leur invasion rapide, la médiane de survie des patients atteint d'un GBM est de 12 à 14 mois malgré les traitements (et de 5 ans au plus pour les long-survivants). Pour les grades II et III (GII et GIII), le pronostic est meilleur mais reste court de 2 à 5 ans et 2 ans , respectivement (Stupp & Weber, 2005).

En 2014, la classification de l'OMS des gliomes a été mise à jour. En plus des caractéristiques histopathologiques des tumeurs, des marqueurs moléculaires supplémentaires sont utilisés pour mieux caractériser ces tumeurs.

La méthylation du promoteur du gène *MGMT* est un outil pronostique intéressant. En effet, le traitement des tumeurs par un agent anticancéreux, le témozolomide qui est un agent alkylant, va induire la fixation d'un groupement alkyl sur l'oxygène 6 de la guanine, induisant ainsi des dommages à l'ADN des cellules cancéreuses et leur mort. Or le gène *MGMT* code pour une enzyme de réparation de l'ADN qui va enlever cette lésion et ainsi diminuer la réponse de la cellule tumorale à l'agent anti-cancéreux. Aussi quand le promoteur du gène *MGMT* est méthylé, à cause du phénotype G-CIMP par exemple, la cellule cancéreuse répondra mieux au traitement avec cet agent de chimiothérapie (Berghoff *et al.*, 2015).

D'autres marqueurs moléculaires permettent de discriminer les populations de gliomes, comme la co-délétion 1p19q. Les patients présentant cette délétion ont une survie plus longue suite à un traitement de radiothérapie (Wick *et al.*, 2009).

Enfin, le gène *IDH1* code pour une enzyme l'isocitrate déshydrogénase 1 qui va convertir l'isocitrate en α-cétoglutarate. Quand ce gène est muté, l'enzyme va catalyser la transformation de l'α-cétoglutarate en 2-hydroxyglutarate qui est un oncométabolite associé à l'instabilité génétique et à la transformation tumorale. La mutation *IDH1* est la plus commune et la plus précoce des altérations génétiques du gliome. Cette mutation est principalement observée dans les gliomes de bas grade (GII et GIII) et participerait à l'instabilité génomique (Wang *et al.*, 2015). De plus, cette mutation serait suffisante à induire le phénotype d'hyperméthylation des ilots CpG observé dans les gliomes et nommé G-CIMP (Turcan *et al.*, 2012). Ce phénotype G-CIMP est principalement identifié dans les gliomes de bas grade où *IDH1* est muté (Noushmehr *et al.*, 2010). Cette mutation permet de mieux discriminer les populations de gliomes de bas grade, où le gène *IDH1* est muté, des GBM, où le gène *IDH1* est sauvage et associé à une survie faible des patients (Figure 34-B).

Par ailleurs, les gliomes se comportent à l'image des autres tumeurs à savoir qu'une hypométhylation globale de leur génome est observée. Cette hypométhylation va majoritairement toucher les régions intergéniques et les séquences répétées notamment les L1. Elle est retrouvée dans les GBM où elle va induire une prolifération cellulaire accrue qui corrèle avec l'agressivité tumorale (Cadieux *et al.*, 2006). L'analyse de la méthylation plus spécifiquement des promoteurs de L1 dans les gliomes montre une diminution significative de celle-ci dans 80% des GBM *de novo* par rapport aux gliomes de bas grades et aux tissus sains. Cette étude établit que le maintien d'une méthylation normale des promoteurs de L1 est un facteur de pronostic favorable (Ohka *et al.*, 2011). Aussi les gliomes semblent être un bon modèle d'étude pour étudier l'impact des LCT au cours de la tumorigenèse gliale

4.2. Objectifs des travaux de thèse.

Les études précédentes ont permis de prouver conceptuellement l'existence de LCT et en ont validés certains. Mais ces études présentent des biais méthodologiques et ne permettent pas d'appréhender l'étendue pangénomique de la dérégulation transcriptionnelle liée aux LCT dans les tumeurs.

Dans les cancers, une hypométhylation globale est observée et affecte notamment les L1. Ces L1, de par la présence d'un promoteur bidirectionnel, vont soit pouvoir rétrotransposer à un nouvel endroit du génome participant ainsi au développement tumoral *via* l'instabilité génétique ; soit participer à la dérégulation transcriptionnelle du génome tumoral *via* la production de LCT à partir de leur promoteur ASP. C'est à ce dernier évènement que je

m'intéresse particulièrement. Aussi, le gliome a été choisi comme modèle d'étude car à l'image des autres cancers, une hypométhylation affectant les L1 est observée qui semble corrélé avec l'agressivité. Dans ce contexte, les objectifs de mes travaux de thèse sont :

1- De valider un outil bio-informatique dédié à la détection de LCT dans les données de séquençage ARN (RNA-seq), développé au sein de l'équipe par un ingénieur.
2- D'évaluer l'étendue pangénomique de la dérégulation transcriptionnelle liée aux LCT dans les gliomes,
3- De démontrer les bases mécanistiques de cette dérégulation,
4- D'évaluer si certains LCT peuvent correspondre à de nouveaux biomarqueurs voir jouer un rôle fonctionnel dans les processus tumoraux.

RESULTATS

Partie 1 : CLIFinder un outil bio-informatique dédié pour identifier des LCT potentiels dans les données de RNA-seq.

1. Objet de l'étude partie 1.

Plusieurs études ont apporté une preuve de concept de l'existence de transcrits LCT initiés par les ASP d'éléments L1 récents (Criscione *et al.*, 2016; Cruickshanks *et al.*, 2013; Cruickshanks & Tufarelli, 2009; Mätlik *et al.*, 2006; Nigumann *et al.*, 2002; Speek, 2001; Wolff *et al.*, 2010). Ceux-ci sont retrouvés dans des tissus tumoraux et également dans des tissus normaux avec, pour certains LCT, la démonstration d'une surexpression en contexte tumoral (Cruickshanks *et al.*, 2013; Cruickshanks & Tufarelli, 2009; Wolff *et al.*, 2010). De plus, un rôle fonctionnel dans le processus de tumorigenèse a été mis en évidence pour deux LCT. Ainsi, *L1-MET* contribuerait à la production d'un récepteur MET tyrosine kinase tronqué, constitutivement activé (Wolff *et al.*, 2010), et LCT13 induirait la mise sous silence épigénétique du gène suppresseur de tumeur *TFPI-2*, impliqué dans le processus d'invasion tumorale (Cruickshanks *et al.*, 2013). Les LCT se positionnent donc comme des transcrits dont la dérégulation pourrait participer fonctionnellement aux processus de tumorigenèse.

Malgré ces preuves de concept, ces études n'ont pas permis d'appréhender de façon globale l'étendue pangénomique de l'expression des LCT dans un tissu. En effet, les approches expérimentales utilisées dans ces études pour identifier des LCT présentent de nombreux biais ou *a priori*. Les méthodes bio-informatiques ont consisté à interroger des banques d'EST, tous types de tissus normaux et tumoraux confondus, qui restent bien souvent incomplètes en termes de représentativité de toutes les séquences transcrites. De plus, Criscione n'a retenu que les LCT-EST épissés incluant des régions exoniques. Quant à la technique de LCD développée par Cruickshanks et collaborateurs, de par l'utilisation de l'amorce d'amplification PCR localisée en position +72 à +92 du promoteur de L1, elle impose l'identification de LCT sans épissage dans la séquence promotrice du L1 (LCT de classe VI). Aussi, la question de l'implication au niveau pangénomique de la dérégulation transcriptionnelle liées aux LCT dans un tissu tumoral reste entière.

Avec l'avènement des techniques de séquençage haut débit telles que le séquençage ARN (RNA-seq), l'intégralité du transcriptome d'un échantillon est accessible. Toutefois, l'étude de la transcription liée aux éléments répétés à partir de ces données reste difficile, l'alignement en

une position unique du génome demeurant un challenge. Pour autant, les LCT sont composés non seulement d'une extrémité 5' correspondant à une séquence répétée du génome (séquence promoteur de L1 en antisens) mais aussi d'une extrémité 3' correspondant à une séquence unique du génome. Aussi, grâce à la présence de la séquence unique, il est envisageable de pouvoir aligner précisément sur le génome les lectures de RNA-seq localisées au niveau de la région chimérique caractérisant un LCT. En s'appuyant sur un RNA-seq particulier et sur les caractéristiques des LCT, un outil bio-informatique dédié à l'identification de séquences transcrites chimères pouvant correspondre à des LCT potentiels a été développé au sein de l'équipe par un ingénieur bio-informaticien. Celui-ci a été nommé CLIFinder pour Chimeric LINE Finder.

Dans ce contexte, le premier objectif de ma thèse a été de valider que l'outil CLIFinder est capable d'identifier des chimères correspondant à des LCT potentiels. Pour répondre à cette question, j'ai utilisé des données de la littérature correspondant à des RNA-seq paired-end orientés d'ARNm polyA+ provenant d'une métastase ovarienne humaine et d'une lignée de cancer du sein MCF7. Les résultats obtenus ont fait l'objet d'un article accepté, présenté dans la partie suivante.

2. Article

Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx671
Advance Access Publication Date: 23 October 2017
Applications Note

Sequence analysis

CLIFinder: identification of LINE-1 chimeric transcripts in RNA-seq data

Marie-Elisa Pinson†, Romain Pogorelcnik†, Franck Court, Philippe Arnaud and Catherine Vaurs-Barrière*

GReD, Université Clermont Auvergne, CNRS, INSERM, 63001 Clermont-Ferrand, France

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Ivo Hofacker

Received on June 27, 2017; revised on October 9, 2017; editorial decision on October 18, 2017; accepted on October 19, 2017

Abstract

Summary: L1 Chimeric Transcripts (LCTs) are initiated by repeated LINE-1 element antisense promoters and include the L1 5′UTR sequence in antisense orientation followed by the adjacent genomic region. LCTs have been characterized mainly using bioinformatics approaches to query dbEST. To take advantage of NGS data to unravel the transcriptome composition, we developed Chimeric LIne Finder (CLIFinder), a new bioinformatics tool. Using stranded paired-end RNA-seq data, we demonstrated that CLIFinder can identify genome-wide transcribed chimera sequences corresponding to potential LCTs. Moreover, CLIFinder can be adapted to study transcription from other repeat types.

Availability and implementation: The code is available at: https://github.com/GReD-Clermont/CLIFinder; and for Galaxy users, it is directly accessible in the tool shed at: https://toolshed.g2.bx.psu.edu/view/clifinder/clifinder/.

Contact: catherine.barriere@uca.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transposable elements represent 45% of the human genome. Due to the challenge of sequence alignment, their role in the transcriptional landscape is poorly studied. However, it is known that LINE-1 retro-elements (L1s) can drive transcription of adjacent genomic regions. For instance, in the 5′ untranslated region (5′UTR) of the most recent L1 subfamilies (from the oldest L1PA10 to the most recent L1PA1), an antisense promoter (ASP) has been identified (Speek, 2001; Macia *et al.*, 2011). This ASP can produce L1 Chimeric Transcripts (LCTs) that include the L1 5′UTR sequence in antisense orientation followed by the adjacent genomic region.

LCTs have been characterized mainly using bioinformatics approaches to query dbEST and correspond to spliced ESTs where the L1 5′UTR sequence in antisense orientation is associated with a gene exon (Speek, 2001; Nigumann *et al.*, 2002; Mätlik *et al.*, 2006; Wolff *et al.*, 2010; Criscione *et al.*, 2016). In addition, Cruickshanks and Tufarelli (2009) developed a dedicated biomolecular approach (called LCD) that allowed the identification of 18 LCTs. Nevertheless, these studies remain limited by the *a priori* definition of such sequences (position of the L1 primer in LCD and selection of only spliced exon-containing ESTs). To take advantage of the strong potential of next generation sequencing to unravel the transcriptome composition, we developed Chimeric LINE Finder (CLIFinder), a new bioinformatics tool that identifies chimeras corresponding to potential LCTs in RNA-seq data.

2 Materials and methods

CLIFinder is a Galaxy tool designed to identify chimeras from one or several samples. This tool was developed mainly to analyze stranded paired-end RNA-seq data. It can also analyze non-stranded paired-end RNA-seq data, but with a higher rate of false-positive chimeras, particularly from events that encompass L1 promoter sequences. Quality and adapter trimming must be done on Fastq files, before their use in CLIFinder.

2.1 Finding chimeras

The iterative procedure starts by finding chimeras for each sample (Supplementary Fig. S1). We consider here that Read 1 (R1, starting at the 5′-end) of the stranded pair must contain an L1 sequence. This parameter is customable according to the stranded or non-stranded sequencing and the repeat element considered. For stranded libraries:

- If the first read of the pair (R1) is in the sense of the transcript, R1 is considered to correspond to the transposable element part of the chimera (Supplementary Fig. S1)
- If the first read of the pair is in the opposite sense of the transcript, then the second read of the pair (R2) is considered to be the transposable element part of the chimera.

For non-stranded sequencing, both hypotheses are retained.

Step 1 consists in selecting pairs (R1 and R2) that include R1 with a minimum of X consecutive bp that correspond to repeat element sequences given by the user (for instance, 5′-end LINE-1 sub-families sequences) and an unmapped mate (R2). Only paired reads that respect this condition are retained. Mismatches (Y) can be tolerated or not for L1 mapping. In step 2, these paired reads are filtered using the RepeatMasker tool to retain only reads for which at least Z consecutive bp are not detected as a repeated sequence. To identify chimeras (step 3), the retained filtered paired sequences are aligned to the human reference genome with a flexible maximum insert size between paired reads (W).

Therefore, CLIFinder is a fully customizable tool to detect transcripts initiated by different types of repeat elements with X, Y, Z and W values that can be changed by the user to adapt the filter stringency. The parameter settings depend, notably, on the read length and the L1 reference sequence list given by the user in Step 1. In our hands, using one consensus sequence for each L1 element from the 27 sub-families (Khan *et al.*, 2006) as reference for analyzing 100–150 bp paired-end reads, the optimal conditions to identify the largest number of chimeras was $X = 30$ bp, $Y = 6$ mismatches for Step 1. For Step 2 only the less stringent condition of $Z = 30$ bp was used to allow at least the design of a primer in the unique sequence for later validation experiments. Finally, $W = 50\ 000$ bp was the maximum insert size between paired reads that allowed the identification of the largest number of spliced chimeras.

2.2 Merging chimeras from all samples and concatenation

For each sample, the chimera datasets are stored in two bed files that correspond to each part of the paired-end read alignments. For an overview of all data and comparison of different datasets, CLIFinder concatenates the different files in two unique bed files, one that corresponds to the L1 part (R1) and the other to the unique sequence part (R2) (step 4). At this stage, several potential chimeras can overlap in the first or/and second read dataset. Bedtools merge function is used to group the different reads corresponding to the same locus (±100 bp) (step 5) to eliminate repetitive information that might correspond to the same chimera. This task is done for both files. The Getfasta function from Bedtools is then employed to create a unique fasta file (step 6) that contains all the potential chimeras concatenated from all samples to generate the html report.

2.3 Visualizing and downloading the resulting chimeras

CLIFinder results can be visualized on a Galaxy interface. An html table is generated where each line corresponds to a chimera defined by its genomic coordinates (hg19), transcription strand, and number of reads detected in each sample. BLASTn searches against EST and RNA databases are performed for chimera annotation, if data are already available. Users can also download: (i) a tabular file with the same content as the html page; (ii) a final annotated file with additional information for each chimera: ID, gene name and strand (if it is localized in an intragenic region), and characteristics of the involved L1 element (family name, position, size, transcription strand) (step 7) (Supplementary Table S1) and (iii) a CLIFinder execution report with the number of pairs retained after sample filtering.

CLIFinder was first tested on publicly available metastatic high grade serous ovarian cancer mRNA-seq data (Illumina HiSeq2500, 100 bp stranded paired-end, 47 million reads) (Böhm *et al.*, 2016). The original dataset GSM2122741 (Ov1.1) was used to create: (i) Ov1.2, an implemented dataset similar to Ov1.1, but including also the *L1-MET* LCT sequence (positive control) and a L1PA2 5′UTR antisense sequence (negative control) and (ii) Ov1.3, a depleted dataset in which 30% of reads included in Ov1.1 were randomly eliminated and the *L1-MET* and L1PA2 5′UTR antisense control sequences were added. Two additional datasets that corresponded to mRNA-seq data for the MCF-7 (untreated breast adenocarcinoma) cell line (E-MTAB-3788, Illumina HiSeq2500, 150 bp stranded paired-end, mean 135 million reads) were also analyzed (Philippe *et al.*, 2016). The parameters used for this analysis are described in Supplementary Figure S1. On average, each analysis required few hours depending on sequencing depth (i.e. <3 h for the three Ov datasets and up to 8 h for the two MCF-7 replicates) using a Quad-Core workstation with 16Go RAM (Supplementary Table S1).

3 Results

All results are available in Supplementary Table S2. From Ov1.1, CLIFinder identified 37 chimeras that corresponded to potential LCTs. For Ov1.2, 38 chimeras were obtained: the previous 37 chimeras and *L1-MET* LCT (ID_27). Ov1.3 analysis highlighted only three chimeras: *L1-MET* and two other chimeras with a reduced number of reads. Some chimeras were associated with multiple L1s (ID_19). In this case, after visualization of the chimera position loaded from the html file on the genome browser UCSC, manual monitoring must be performed to retain the element presumed to initiate the chimera transcription. Attention must also be paid to remove the few chimeras transcribed in the same orientation as the L1 (ID_12_13). Consistent with the RNA-seq specificities, chimeras defined by only one read, displayed R1 and R2 sequences of 100 bp. When a chimera is defined by several overlapping reads, R1 and R2 size can be increased. Negative (in the case of R1-R2 overlapping) or null distance between R1 and R2 was often observed, suggesting a linear transcription from L1 that continued in the adjacent unique sequence. In some cases, the distance between R1 and R2 was larger than expected (ID_7, ID_30), demonstrating the occurrence of splicing events (Supplementary Fig. S2).

To assess whether chimeras detected by CLIFinder in Ov1.1 could correspond to LCTs initiated at L1 ASPs, the L1 elements involved in the 36 chimeras were compared to the list of 14 495 L1s containing a 5′UTR sequence (extracted from RepeatMasker for the 27 sub-families used by CLIFinder). This showed that 33 chimeras involved L1 elements with a 5′UTR. Then, analysis of the L1 sub-families involved in these 33 chimeras revealed that, although CLIFinder used the consensus sequences from the 27 L1 sub-families to identify chimeras, 31 chimeras (94%) involved L1s only from the more recent sub-families (i.e. L1PA1 to L1PA7) that possess ASP. Finally, comparison of the 33 chimeras and of known LCTs demonstrated that five chimeras identified in this metastatic ovarian cancer sample corresponded to EST-LCTs previously found in other tissues.

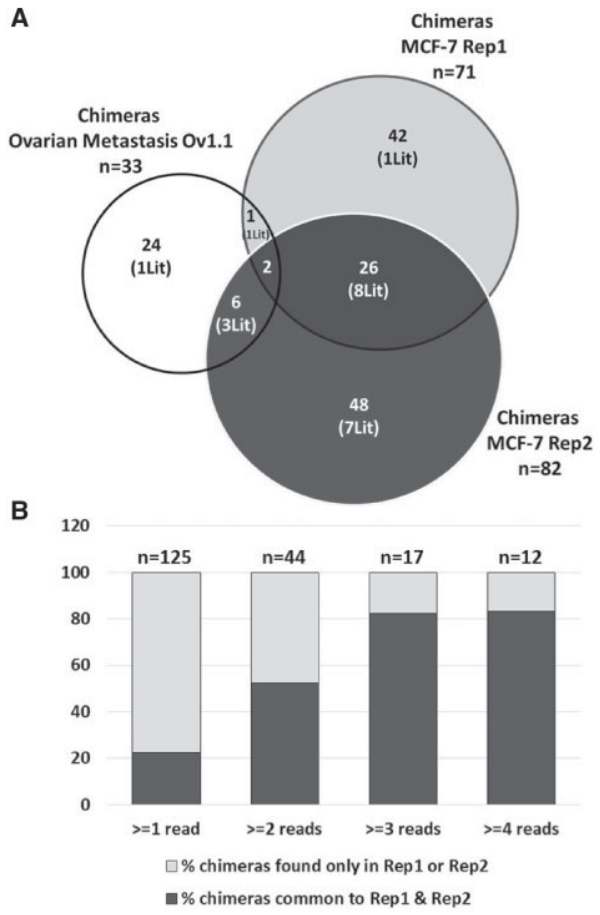


Fig. 1. Analysis of the chimeras identified in stranded paired-end RNA-seq datasets from different tissues by CLIFinder. (**A**) Comparison of chimeras identified in the Ov1.1 dataset and the two replicates (Rep1 and Rep2) from the untreated breast cancer cell line MCF-7. (Lit) indicates the number of chimeras corresponding to LCTs already described in the literature. Two to three times more chimeras were identified in the MCF-7 RNA-seq datasets than in the Ov1.1 dataset, in agreement with the higher sequencing depth. Nine chimeras were found both in the Ov1.1 and MCF-7 datasets. Already known LCTs were identified in each dataset. (**B**) Analysis of the concordance concerning the number of chimeras identified in the two MCF-7 replicates. Among all identified chimeras (detected by at least one read in one replicate), 22% were common to both replicates. When only chimeras depicted by at least 2, 3 or 4 reads in one replicate were considered, the overlap increased, respectively, to 52%, 82% and 83%, suggesting that CLIFinder analysis results are reproducible

Altogether, these observations suggest that most chimeras identified by CLIFinder from RNA-seq data are true LCTs.

To strengthen these results, two replicates from stranded paired-end untreated breast adenocarcinoma MCF-7 cells were also analyzed. After curation of the chimeras that did not meet the criteria for potential LCTs (i.e. chimeras in the same transcription sense of L1: $n = 39$; and chimeras associated with L1 but without 5′UTR ASP region: $n = 8$), in total, 125 chimeras were identified in the two replicates (Supplementary Table S2C). Among these chimeras, 118 (94.5%) involved recent L1s and 20 (16%) corresponded to already described LCTs (Fig. 1A). Comparison of the genomic position of the chimeras identified in the two MCF-7 replicates indicated that 28 chimeras (22%) were common to both samples supress (22%) (Fig. 1A). This moderate overlap can be explained by the fact that chimeras can be detected only if at least one cDNA fragment encompassing the L1/unique sequence junction is sequenced. When only chimeras defined by at least 2 or 3 reads in one MCF-7 replicate were considered, the overlap between replicates increased to 52% (two reads) and 82% (three reads) (Fig. 1B). This indicates that results are reproducible for highly represented chimeras and that the analysis of different samples together allows assessing precisely LCT expression genome-wide in a specific condition.

Two to three times more chimeras were identified in the MCF-7 RNA-seq datasets than in the Ov1.1 dataset, in agreement with the higher sequencing depth (Supplementary Table S1). Interestingly, among the identified chimeras, nine were common between the MCF-7 RNA-seq datasets and the Ov1.1 dataset (Fig. 1A). Among these chimeras, some corresponded to already known LCTs (n = 4), such as Ov1.1 ID_7 and MCF-7 ID_19 (Supplementary Fig. S2) that were previously found in thymus, retinoblastoma and tongue tumor samples. Others corresponded to new loci, such as Ov1.1 ID_30 that seems to be a good candidate because it was recurrently detected in MCF-7 Rep1 and Rep2 (chimeras defined by 2 and 4 reads, respectively) (Supplementary Fig. S2, Table S2C).

In summary, CLIFinder is a new bioinformatics tool to identify genome-wide transcribed chimera sequences corresponding to LCTs from stranded paired-end RNA-seq datasets. This tool will be useful for genome-wide analyses of LCT expression in different tissues, in normal or pathological conditions. Moreover, CLIFinder can be adapted to study transcription from other repeat types.

Acknowledgement

The authors acknowledge the support of Y Renaud from Byonet (byonet.fr).

Funding

This work was supported by the Plan Cancer-INSERM [CS14085CS, to P.A.], ARC [SFI20121205549, to C.V.B.], the Auvergne Region and the Fonds Européen de Développement Régional (FEDER).

Conflict of Interest: none declared.

References

Böhm,S. *et al.* (2016) Neoadjuvant chemotherapy modulates the immune microenvironment in metastases of tubo-ovarian high-grade serous carcinoma. *Clin. Cancer Res.*, **22**, 3025–3036.

Criscione,S.W., Theodosakis,N., and Micevic,G. (2016) Genome-wide characterization of human L1 antisense promoter-driven transcripts. *BMC Genomics*, **17**, 463.

Cruickshanks,H.A., and Tufarelli,C. (2009) Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. *Genomics*, **94**, 397–406.

Khan,H. *et al.* (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.*, **16**, 78–87.

Macia,A., Muñoz-Lopez,M., and Cortes,J.L. (2011) Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol. Cell. Biol.*, **31**, 300–316.

Mätlik,K. *et al.* (2006) L1 antisense promoter drives tissue-specific transcription of human genes. *J. Biomed. Biotechnol.*, **2006**, 71753.

Nigumann,P., Redik,K., and Mätlik,K. (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*, **79**, 628–634.

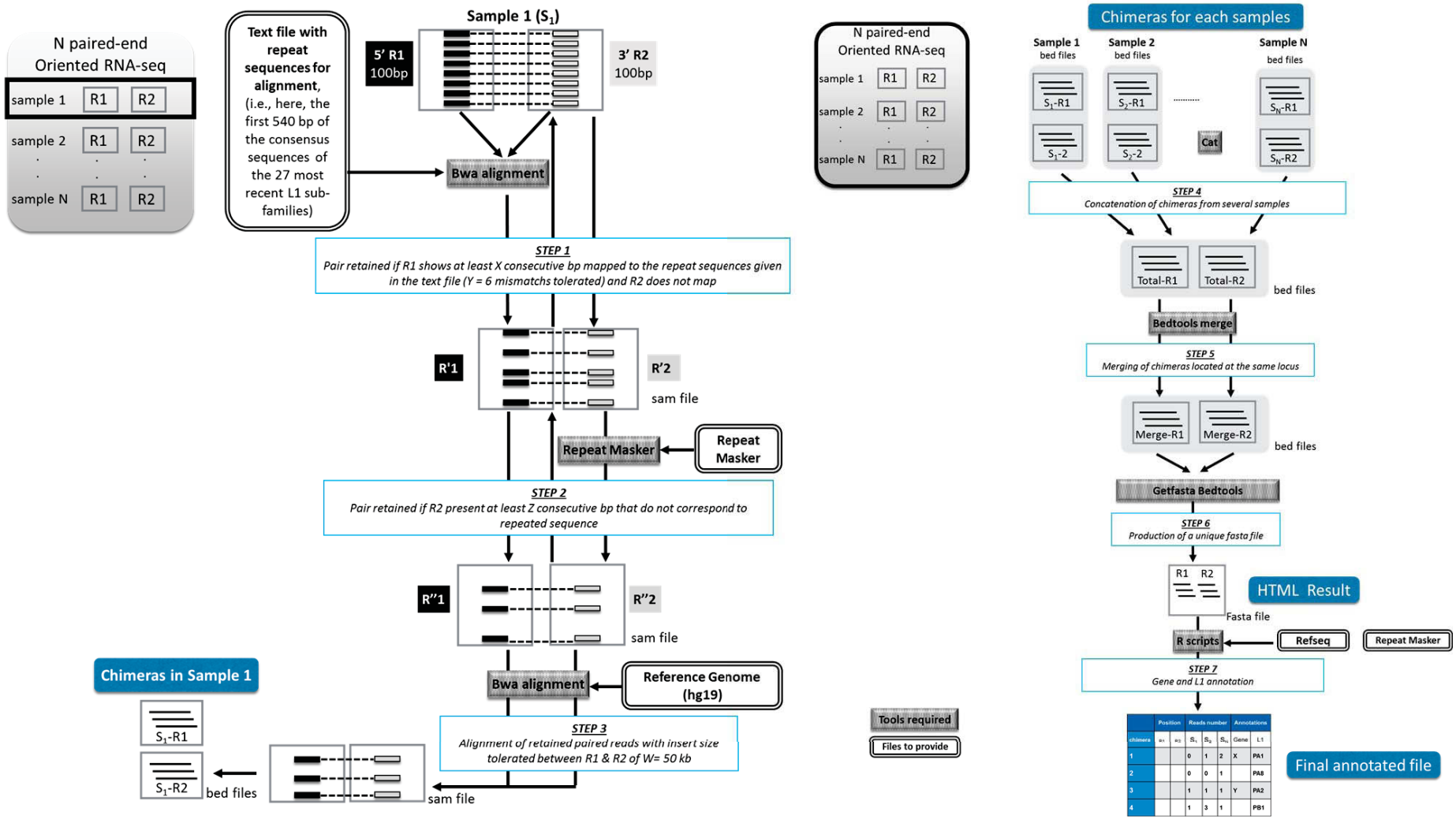
Philippe,C., Vargas-Landin,D.B., and Doucet,A.J. (2016) Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *Elife*, **5**, e13926.

Speek,M. (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.*, **21**, 1973–1985.

Wolff,E.M., Byun,H.-M., and Han,H.F. (2010) Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet.*, **6**, e1000917.

Supplementary Figure S1

A.



Overview of the CLIFinder process.

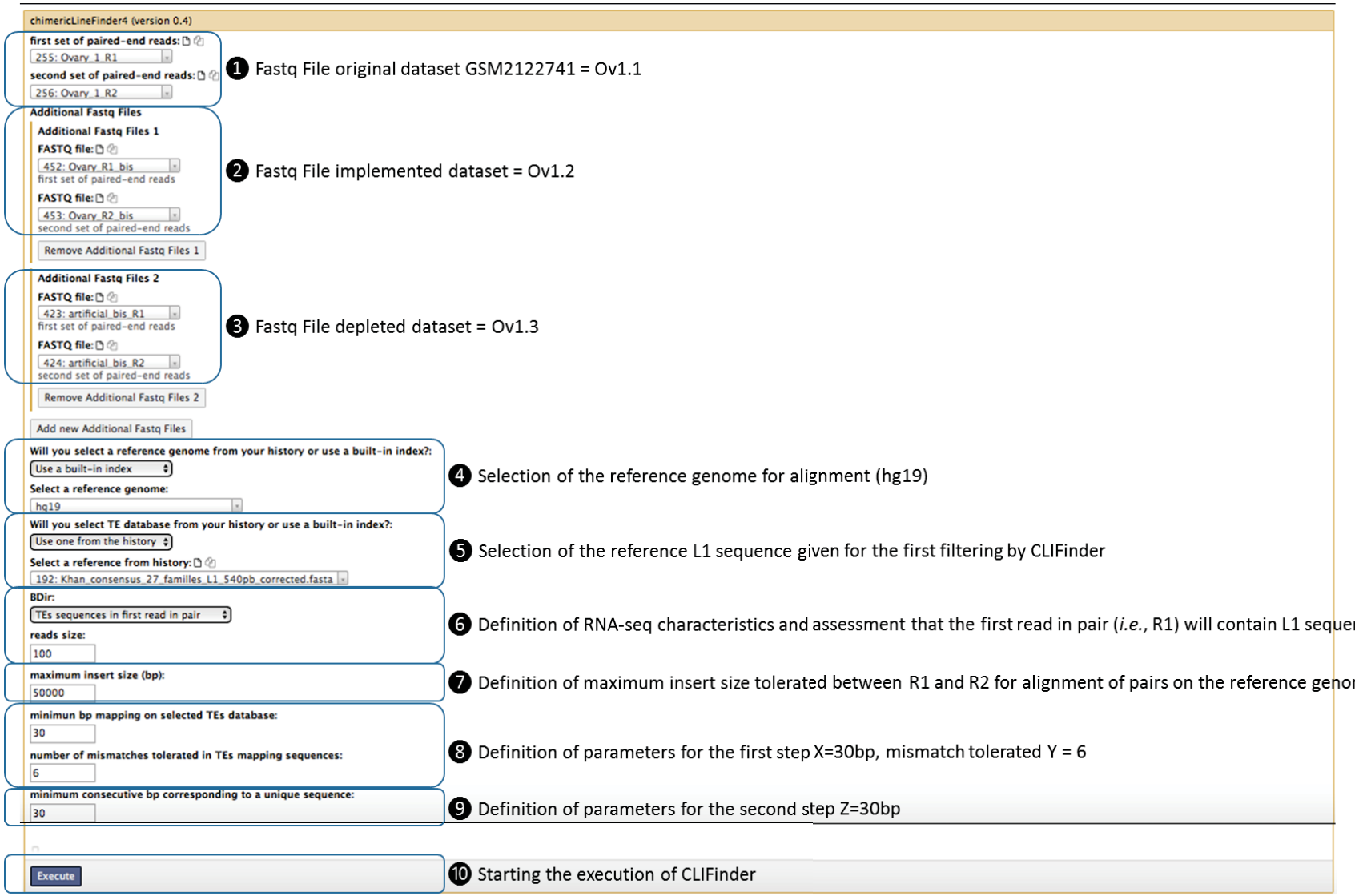
Steps 1 to 3 are performed for each sample individually.

The tools used for each step are listed in the grey boxes.

Files to be provided by the users are indicated in the white boxes.

Steps 4 to 7 are performed after the individual analysis of each sample in the series.

B.


chimericLineFinder4 (version 0.4)
first set of paired-end reads:
255: Ovary_1_R1
second set of paired-end reads:
256: Ovary_1_R2
1 Fastq File original dataset GSM2122741 = Ov1.1
Additional Fastq Files
Additional Fastq Files 1
FASTQ file:
452: Ovary_R1_bis
first set of paired-end reads
FASTQ file:
453: Ovary_R2_bis
second set of paired-end reads
2 Fastq File implemented dataset = Ov1.2
Remove Additional Fastq Files 1
Additional Fastq Files 2
FASTQ file:
423: artificial_bis_R1
first set of paired-end reads
FASTQ file:
424: artificial_bis_R2
second set of paired-end reads
3 Fastq File depleted dataset = Ov1.3
Remove Additional Fastq Files 2
Add new Additional Fastq Files
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Select a reference genome:
hg19
4 Selection of the reference genome for alignment (hg19)
Will you select TE database from your history or use a built-in index?:
Use one from the history
Select a reference from history:
192: Khan_consensus_27_familles_L1_540pb_corrected.fasta
5 Selection of the reference L1 sequence given for the first filtering by CLIFinder
BDir:
TEs sequences in first read in pair
reads size:
100
6 Definition of RNA-seq characteristics and assessment that the first read in pair (i.e., R1) will contain L1 sequence
maximum insert size (bp):
50000
7 Definition of maximum insert size tolerated between R1 and R2 for alignment of pairs on the reference genome W=50kb
minimun bp mapping on selected TEs database:
30
number of mismatches tolerated in TEs mapping sequences:
6
8 Definition of parameters for the first step X=30bp, mismatch tolerated Y = 6
minimum consecutive bp corresponding to a unique sequence:
30
9 Definition of parameters for the second step Z=30bp
Execute
10 Starting the execution of CLIFinder

Description of the CLIFinder screenshot from the Galaxy interface. All the parameters used for this analysis are indicated in bold in the legend.

(1, 2, 3) Uploading Fastq files from **the three datasets Ov1.1, Ov1.2 and Ov1.3**

(4) Selection of the **reference genome** for alignment (here, **hg19**)

(5) Selection of the reference L1 sequences that will be used by CLIFinder for step 1. Here, the reference **L1 sequences** corresponded to the first **540 bp of the consensus sequences of the 27 most recent L1 sub-families** (Khan *et al.*, 2006)

(6) Definition of the **RNA-seq characteristics**, here **oriented paired-end reads of 100 bp**

(7) Definition of the **maximum insert size tolerated between R1 and R2 for alignment** of pairs on the reference genome (here: **30 kb)**

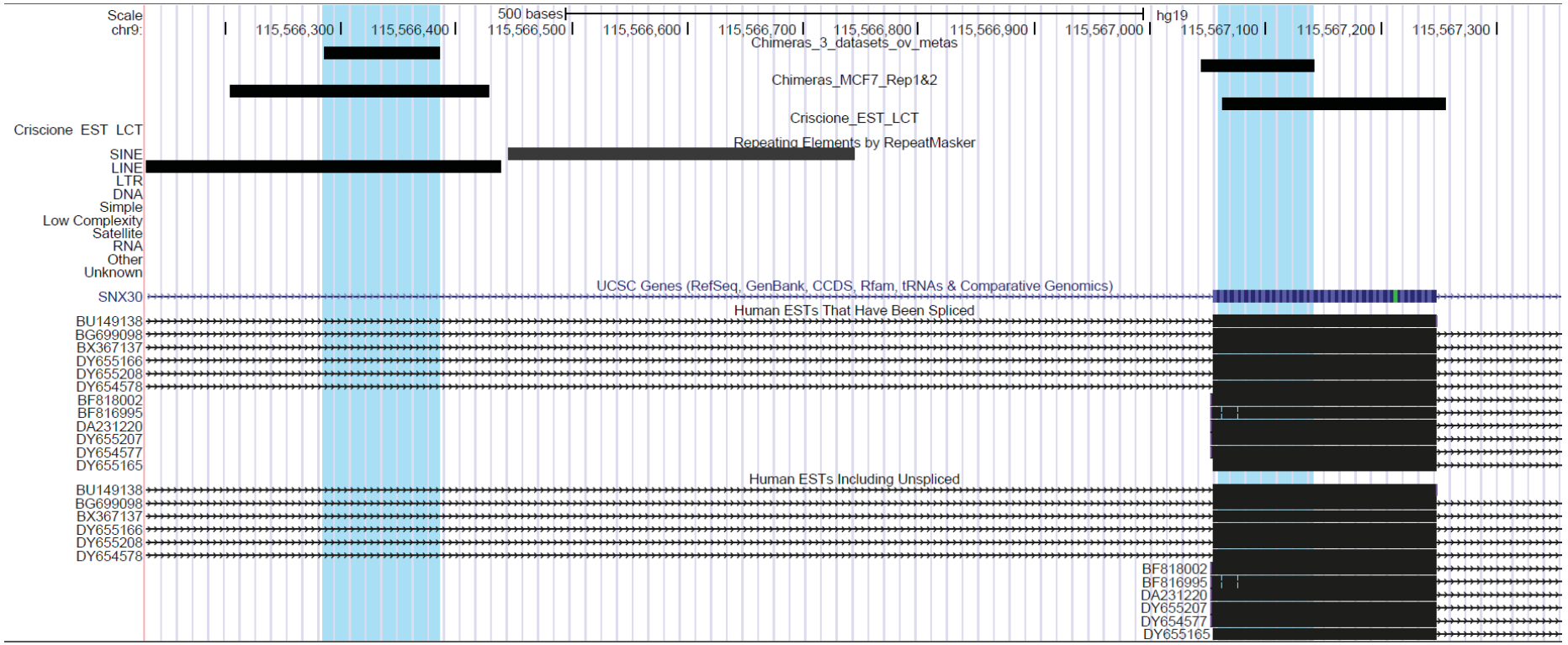
(8) Definition of the parameters for the **first step**; here: **at least 30 consecutive bp matching a reference L1 sequence with a maximum of 6 mismatches**

(9) Definition of the parameters for the **second step**: here, **at least 30 consecutive bp not annotated by RepeatMasker**, thus corresponding to a unique sequence in the human genome

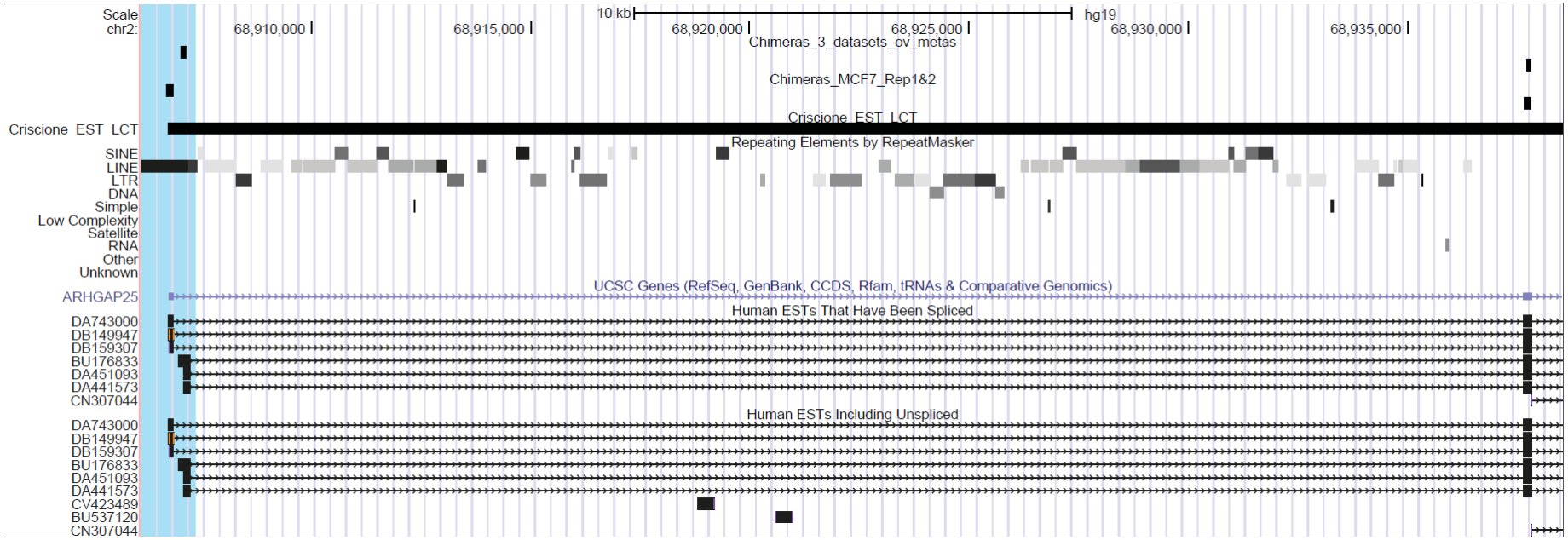
(10) Execution button

Supplementary Figure S2

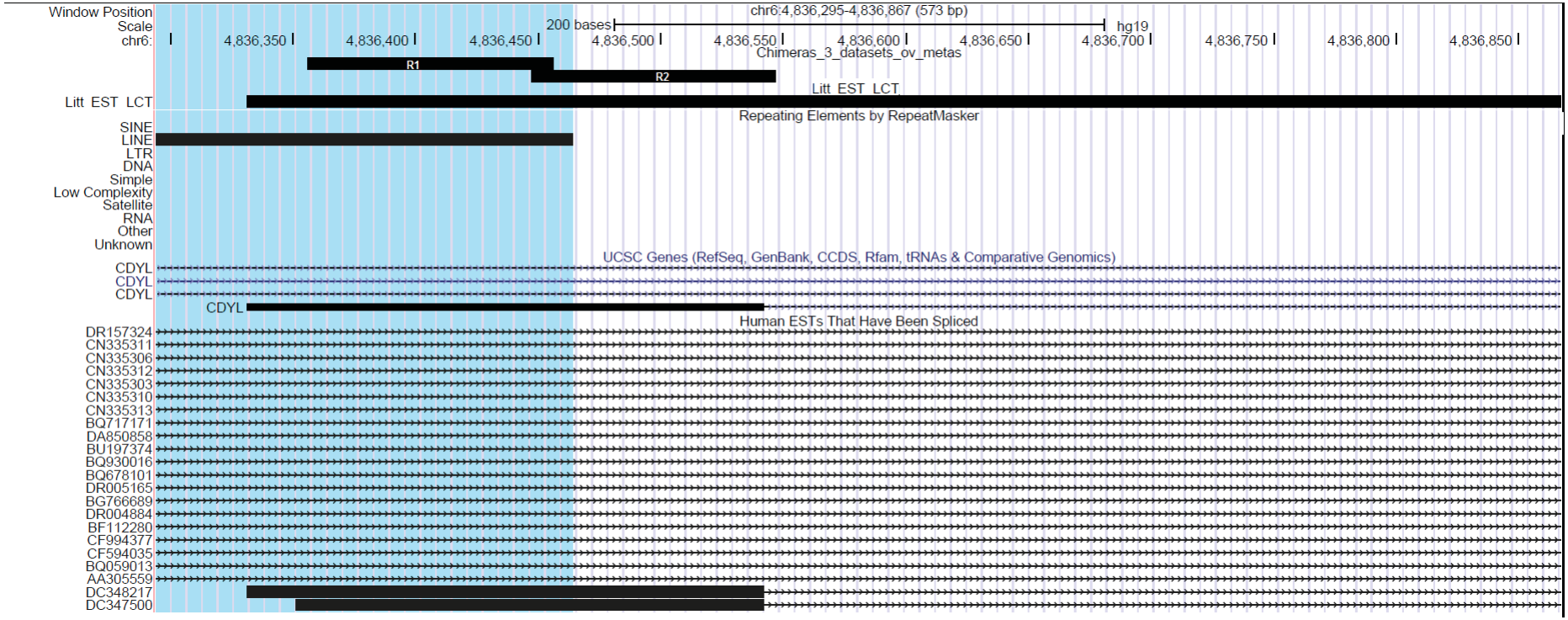
Ov1.1 ID_30 is a newly identified LCT described by 1 read, and confirmed in the MCF-7 datasets (ID_177) by 2 reads in Rep1 and 4 reads in Rep2. The minimal common sequences between the Ov1.1 and MCF-7 datasets are shaded in blue. The R1 and R2 sequences for Ov1.1 are 100bp-long, as expected for one read according to the RNA-seq conditions. As the distance between R1 and R2 is about 856 bp, whereas the cDNA size selection during library preparation was between 200 and 400 bp, a splice event between R1 and R2 could have occurred. Moreover, the R2 sequences overlap with exon 2 of the *SNX30* gene. L1-ASP in intron 1 seems to correspond to an alternative TSS for *SNX30*.



Ov1.1 ID_7 corresponds to an already identified EST-LCT in the *ARHGAP25* gene. It corresponds to MCF-7 Rep 1 ID_19. A splicing event occurred between the L1 5' sequence and the unique sequence in 3' (R1 - R2 distance >30 kb).



ID-22 corresponds to an already identified EST-LCT in the *CDYL* gene. The overlapping between R1 and R2 is the result of the concatenation of the two reads that describe this chimera.



	Data RNA-seq size	
Samples	# paired	# reads
MCF-7_rep1	119891128	239782256
MCF-7_rep2	151717228	303434456
Ovary Metastasis (Ov1.1)	47776441	95552882
Ovary Metastasis_bis (Ov1.2)	47776444	95552888
Ovarian Metastasis Artificial (Ov1.3)	3000003	6000006

	Data execution time
Ov1.1	88 minutes (1h28)
Ov1.1 + Ov1.2	158 mintes (2h38)
Ov1.1 + Ov1.2 + Ov1.3	174 minutes (2h54)
MCF-7 rep 1 + MCF-7 rep 2	488 minutes (8h08)

Supplementary Table S1

Chimera[1]				L1 (R1)[2]				Unique (R2)[3]				ID_final[4]	Reads#[5]			Gene[6]		Repeat[7]					Supplemental annotation			
Chimera.Chr	Chimera.Start	Chimera.End	Chimera.Strand	L1.Chr	L1.Start	L1.End	L1.Strand	Unique.Chr	Unique.Start	Unique.End	Unique.Strand		Ovary_1_R1.read...	Ovary_R1_bis.read...	artificial_bis_R1.read...	GeneName	GeneStrand	repName	repStart	repEnd	repWidth	repStrand	L1wi5UTR[8]	LCT_litterature[9]	R1-R2 Distce (bp)[10]	
chr1	100665031	100665292	-	chr1	100665192	100665292	-	chr1	100665031	100665131	+	Id_1	1	1	0	*DBT*	+	L1PA7	100665158	100671644	6487	+	TRUE	FALSE	61	
chr1	108710311	108710551	+	chr1	108710311	108710411	+	chr1	108710451	108710551	-	Id_2	1	1	0	*SLC25A24*	+	L1PA3	108704418	108710431	6014	-	TRUE	FALSE	40	
chr1	119563185	119563375	-	chr1	119563275	119563375	-	chr1	119563185	119563285	+	Id_3	2	2	0	no_gene	*	L1PA3	119563254	119569271	6018	+	TRUE	FALSE	-10	
chr2	1392520	1392712	+	chr2	1392520	1392620	+	chr2	1392612	1392712	-	Id_4	1	1	0	no_gene	*	L1PA5	1386506	1392621	6116	-	TRUE	FALSE	-8	
chr2	63062493	63062793	-	chr2	63062693	63062793	-	chr2	63062493	63062593	+	Id_5	2	2	0	*EHBP1*	-	L1PA2	63062613	63068633	6021	+	TRUE	FALSE	100	
chr2	68631761	68632187	-	chr2	68632087	68632187	-	chr2	68631761	68631861	+	Id_6	1	1	0	no_gene	*	L1PA7	68632040	68638471	6432	+	TRUE	FALSE	226	
chr2	68907039	68937807	+	chr2	68907039	68907139	+	chr2	68937707	68937807	-	Id_7	2	2	1	no_gene	*	L1PA3	68904993	68907206	2214	-	TRUE	Thymus, Retinoblastoma, Tongue tumor	30568	
chr2	71644297	71644747	+	chr2	71644297	71643397	+	chr2	71644647	71644747	-	Id_8	1	1	0	*ZNF638*	+	L1HS	71638604	71644631	6028	-	TRUE	Lung	250	
chr2	114234988	114235214	+	chr2	114234988	114235088	+	chr2	114235114	114235214	-	Id_9	1	1	0	*CWH32*	-	L1PA3	114229058	114235122	6065	-	TRUE	FALSE	26	
chr2	218216857	218217055	+	chr2	218216857	218216957	+	chr2	218216955	218217055	-	Id_10	2	2	0	*DIRC3*	-	L3PA13	218210252	218216987	6736	-	TRUE	FALSE	-2	
chr3	67998139	67998371	-	chr3	67998271	67998371	-	chr3	67998139	67998239	+	Id_11	2	2	0	no_gene	*	L1P2	67998240	68000000	1761	+	FALSE			
chr3	139108667	139138496	-	chr3	139138395	139138496	-	chr3	139108667	139109056	+	ID_12_13	2	2	0	*LOC1005072*	-	L1M1	139134333	139138647	4315	-	TRUE			
chr3	167728782	167729193	-	chr3	167729093	167729193	-	chr3	167728782	167728882	+	Id_14	1	1	0	*GOLIM4*	+	L1PA2	167728820	167734843	6024	+	TRUE	FALSE	211	
chr4	165556	165854	+	chr4	165556	165656	+	chr4	165754	165854	-	Id_15	1	1	0	*ZNF718*	-	L1PA4	1596 38	165778	6141	-	TRUE	FALSE	98	
chr4	53017310	53017474	-	chr4	53017374	53017474	-	chr4	53017310	53017410	+	Id_16	1	1	0	no_gene	*	L1PA7	53017370	53023777	6408	+	TRUE	FALSE	-36	
chr4	149244935	149245159	+	chr4	149244935	149245035	+	chr4	149245059	149245159	-	Id_17	1	1	0	*NR3C2*	-	L1PA2	149239026	149245056	6031	-	TRUE	FALSE	24	
chr5	1953498	1953740	-	chr5	1953640	1953740	-	chr5	1953498	1953598	+	Id_18	2	2	0	no_gene	*	L1P2	1953549	1955638	2090	+	FALSE			
chr5	114531644	114531928	-	chr5	114531828	114531928	-	chr5	114531644	114531744	+	Id_19	1	1	0	no_gene	*	L1PBa/L1PBa	114531749/114531927	114531930/114533318	182/1392	+/+	FALSE			
chr5	115526645	115526974	-	chr5	115526874	115526974	-	chr5	115526645	115526745	+	Id_20	1	1	0	*COMMD10*	+	L1PB1	115526812	115531414	4603	+	TRUE	FALSE	129	
chr5	126231868	126232078	+	chr5	126231868	126231968	+	chr5	126231978	126232078	-	Id_21	1	1	0	*MARCH3*	+	L1PA3	126225978	126232025	6048	-	TRUE	FALSE	10	
chr6	4836356	4836547	+	chr6	4836356	4836456	+	chr6	4836447	4836547	-	Id_22	2	2	0	*CDYL*	+	L1PA3	4831406	4836464	5059	-	TRUE	Tongue tumor	-9	
chr6	6191262	6191647	+	chr6	6191262	6191463	+	chr6	6191438	6191647	-	Id_23	3	3	0	*F13A1*	+	L1PA3	6185362	6191484	6123	-	TRUE	FALSE	-25	
chr6	6300825	6301045	+	chr6	6300825	6300925	+	chr6	6300945	6301045	-	Id_24	1	1	0	*F13A1*	+	L1PA3	6294976	6301001	6026	-	TRUE	FALSE	20	
chr6	116256140	116256315	-	chr6	116256215	116256315	-	chr6	116256140	116256240	+	Id_25	1	1	0	no_gene	*	L1PA7	116256217	116262668	6452	+	TRUE	FALSE	-25	
chr6	160091346	160092041	-	chr6	160091643	160092041	-	chr6	160091346	160091687	+	Id_26	7	7	0	no_gene	*	L1PA5	160091646	160097787	6142	+	TRUE	FALSE	-44	
chr7	***116364091***	***116371821***	***+***	***chr7***	***116364091***	***116364191***	***+***	***chr7***	***116371721***	***116371821***	***-***	***Id_27***	***0***	***1***	***1***	***MET***	-	***L1PA2***	***116358551***	***116364543***	***5993***	-	***TRUE***	***Tongue tumor, bladder carcinoma, placenta***	***7530***	
chr8	61357564	61357857	+	chr8	61357564	61357664	+	chr8	61357757	61357857	-	Id_28	1	1	0	no_gene	*	L1PA6	61354821	61357672	2852	-	TRUE	FALSE	93	
chr9	32448706	32448926	+	chr9	32448706	32448806	+	chr9	32448826	32448926	-	Id_29	2	2	0	*ACO1*	+	L1PA7	32444662	32448807	4146	-	TRUE	FALSE	20	
chr9	115566286	115567142	+	chr9	115566286	115566386	+	chr9	115567045	115567142	-	Id_30	1	1	0	*SNX30*	+	L1HS	115560407	115566439	6033	-	TRUE	FALSE	659	
chr11	19608656	19609364	+	chr11	19608656	19608756	+	chr11	19609264	19609364	-	Id_31	1	1	0	*NAV2*	+	L1PA4	19602918	19609066	6149	-	TRUE	FALSE	508	
chr11	86355771	86356055	-	chr11	86355955	86356055	-	chr11	86355771	86355871	+	Id_32	1	1	0	*ME3*	+	L1PA7	86355808	86358331	2524	+	TRUE	FALSE	84	
chr11	87056772	87057154	+	chr11	87056772	87056872	+	chr11	87057054	87057154	-	Id_33	3	3	0	no_gene	*	L1HS	87051073	87057103	6031	-	TRUE	FALSE	182	
chr14	31159850	31160082	+	chr14	31159850	31159950	+	chr14	31159982	31160082	-	Id_34	2	2	0	*SOD1*	-	L1HS	31154015	31160043	6029	-	TRUE	Bladder carcinoma	32	
chr15	43467844	43468006	-	chr15	43467906	43468006	-	chr15	43467844	43467944	+	Id_35	1	1	0	*TMEM62*	-	L1P1	43467898	43469402	1505	+	TRUE	FALSE	-38	
chr16	54075929	54076188	-	chr16	54076088	54076188	-	chr16	54075929	54076029	+	Id_36	1	1	0	*FTO*	-	L1HS	54076008	54082036	6029	+	TRUE	FALSE	59	
chr18	7966256	7966571	-	chr18	7966456	7966571	-	chr18	7966256	7966472	+	Id_37	4	4	1	*PTPRM*	+	L1PA2	7966440	7972472	6033	+	TRUE	FALSE	-16	
chr18	9264716	9265125	-	chr18	9265025	9265125	-	chr18	9264716	9264816	+	Id_38	1	1	0	*ANKRD12*	+	L1PA4	9264937	9270777	5841	+	TRUE	FALSE	209	
chr21	35303722	35304040	+	chr21	35303722	35303854	+	chr21	35303857	35304040	-	Id_39	3	3	0	*LINC00649*	+	L1PA2	35297926	35303910	5985	-	TRUE	Liver, spleen, embryonic stem cells, placenta, lung tumor, nasopharynx	3	
													Ov1.1	**Ov1.2**	**Ov1.3**											

Legend

1: Genomic position and transcription strand for the chimera (from R1 to R2).

2, 3: Same data, but specifically for the 5' R1 sequence containing the L1 sequence and for the 3' R2 sequence containing at least 30 consecutive bp not annotated by RepeatMasker.

4: Chimera ID number. IDs with more than one number, such as ID_12_13, indicate that this chimera corresponds to the compilation of two chimeras sharing the same position on one side (usually on the 5' R1 side with the L1 sequence) and two different non-overlapping positions on the other side (usually on the 3' R2 unique sequence)

5: These columns recapitulate the read numbers that correspond to the chimera in each sample

6: Gene annotation at the chimera position according to RefSeq

7: Repeat annotation (repeat name, width, transcription strand). In the case of multiple L1 annotations for one chimera (*i.e.* ID_12_13), the elements supposed to be involved in the chimera genesis must be retained by manual monitoring. Attention must be paid to remove the few chimeras with a transcription strand similar to the one of the associated L1 element. As in this case transcription cannot be initiated by an ASP, such chimeras do not correspond to potential LCTs.

8: Supplemental annotation after comparison of the chimera coordinates with those of the 14,495 L1 elements containing a 5'UTR (extracted via UCSC from RepeatMasker for the 27 sub-families used by CLIFinder).

9: Supplemental annotation after comparison of the chimera coordinates with those of the about 200 EST-LCTs already described (Speek, 2001; Ngumann *et al.*, 2002; Matlik *et al.*, 2006; Cruickshanks and Tufarelli, 2009; Criscione *et al.*, 2016). Five chimeras identified in the metastatic ovarian cancer sample were already known (ID_7, ID_8, ID_22, ID_34 and ID_39) and *L1-MET* LCT (ID_27) was the positive control in the Ov1.2 and Ov1.3 datasets.

10: Supplemental information on the distance between R1 and R2. As size selection in the RNA-seq library was between 200 and 400 bp, a distance between R1-R2 >300 bp suggests the occurrence of a splicing event. For instance, chimera ID_7, which is known to be spliced, displayed a distance >30 kb. Similarly, chimera ID_30, which was detected by only one read, had a distance of 659 bp. UCSC genome browser views of this chimera (Supplementary Figure S2A) suggest as splicing event between the L1 sequence and the downstream exon of the *SNX30* gene.

Chimera[1]				L1 (R1)[2]				Unique (R2)[3]				ID_final[4]	Reads#[5]			Gene[6]	Repeat[7]			Supplemental Annotations		
Chr	Start	End	Strand	Chr	Start	End	Strand	Chr	Start	End	Strand		Ov1.1	Ov1.2	Ov1.3	Name	Name	Width	Strand	L1wi5UTR[8]	Literature[9]	R1-R2 distce (bp)[10]
chr1	100665031	100665292	-	chr1	100665192	100665292	-	chr1	100665031	100665131	+	Id_1	1	1	0	DBT	L1PA7	6487	+	TRUE	FALSE	261
chr2	68907039	68937807	+	chr2	68907039	68907139	+	chr2	68937707	68937807	-	Id_7	2	2	1	no_gene	L1PA3	2214	-	TRUE	Thymus, Retinoblastoma, Tongue tumor	**30768**
chr3	67998139	67998371	-	chr3	67998271	67998371	-	chr3	67998139	67998239	+	Id_11	2	2	0	no_gene	L1P2	1761	+	FALSE		
chr3	139108667	139138496	-	chr3	139138395	139138496	-	chr3	139108667	139109056	+	ID_12_13	2	2	0	LOC1005072	L1M1	4315	-			
chr5	1953498	1953740	-	chr5	1953640	1953740	-	chr5	1953498	1953598	+	Id_18	2	2	0	no_gene	L1P2	2090	+	FALSE		
chr5	114531644	114531928	-	chr5	114531828	114531928	-	chr5	114531644	114531744	+	Id_19	1	1	0	no_gene	L1PBa/L1PBa	182/1392	+/+	FALSE		
chr6	4836356	4836547	+	chr6	4836356	4836456	+	chr6	4836447	4836547	-	Id_22	2	2	0	CDYL	L1PA3	5059	-	TRUE	Tongue tumor	191
chr7	*116364091*	*116371821*	*+*	*chr7*	*116364091*	*116364191*	*+*	*chr7*	*116371721*	*116371821*	-	*Id_27*	*0*	***1***	***1***	*MET*	*L1PA2*	*5993*	-	*TRUE*	***Tongue tumor, bladder carcinoma, placenta***	*7730*
chr9	115566286	115567142	+	chr9	115566286	115566386	+	chr9	115567045	115567142	-	Id_30	1	1	0	SNX30	L1HS	6033	-	TRUE	FALSE	**856**
chr14	31159850	31160082	+	chr14	31159850	31159950	+	chr14	31159982	31160082	-	Id_34	2	2	0	SCFD1	L1HS	6029	-	TRUE	bladder carcinoma	232
chr18	7966256	7966571	-	chr18	7966456	7966571	-	chr18	7966256	7966472	+	Id_37	4	4	1	PTPRM	L1PA2	6033	+	TRUE	FALSE	315
chr21	35303722	35304040	+	chr21	35303722	35303854	+	chr21	35303857	35304040	-	Id_39	3	3	0	LINC00649	L1PA2	5985	-	TRUE	liver, spleen, embryonic stem cells, placenta, lung tumor, nasopharynx	318

Legend Table S2B

This table shows data extracted from Table S1A for the 12 representative chimeras cited in the main text. Each line corresponds to a unique chimera.

3. Conclusion Partie 1.

L'analyse du jeu de données de RNA-seq de métastase ovarienne par CLIFinder a identifié 36 chimères. Parmi elles,

- 33/36 impliquent un L1 possédant une extrémité 5'UTR.
- 31/33 (94%) sont associées à un L1 récent (L1PA1 à 7) pour lesquels un ASP fonctionnel a été décrit (Macia *et al.*, 2011; Speek, 2001).
- 5/33 (15%) correspondent à des LCT déjà décrits dans la littérature dans d'autres tissus (Criscione *et al.*, 2016; Mätlik *et al.*, 2006).
- enfin, 4/31 impliquent une distance entre les deux lectures paired-end > 500 pb (allant jusqu'à 7 et 30 Kb), ce qui démontre que CLIFinder peut identifier des chimères épissées à l'image de ce qui a été décrit dans la littérature (Criscione *et al.*, 2016).

Dans un deuxième temps, l'analyse de 2 réplicas de RNA-seq sur des ARN de la lignée MCF7 a identifié 125 chimères parmi lesquelles 94,5% impliquent des L1 récents, 16% correspondent à des LCT déjà décrits et 22% sont communes entre les deux répliquas. De plus, si on prend en compte les chimères identifiées par au moins 2 ou 3 lectures dans au moins 1 réplica, les taux de chimères communes atteignent respectivement 52% et 82%. Ceci démontre une reproductibilité de l'approche pour les chimères fortement exprimées et suggère que l'analyse combinée de plusieurs échantillons doit permettre l'identification exhaustive pangénomique de LCT.

Enfin, l'analyse comparative des chimères obtenues dans la métastase ovarienne et les réplicas MCF7 identifie 8 chimères communes parmi lesquelles 4 correspondent à des LCT déjà décrits. Cette récurrence de LCT dans différents types tumoraux suggère que ceux-ci pourraient jouer un rôle dans les processus de carcinogénèse.

L'ensemble de ces résultats valide que les séquences chimères transcrites identifiées par CLIFinder à partir de données de RNA-seq paired-end orientées, correspondent à des LCT potentiels. Ainsi, CLIFinder se positionne comme un nouvel outil qui devrait permettre de mieux caractériser au niveau pangénomique l'impact de la transcription des LCT dans différents tissus, en conditions normales et tumorales.

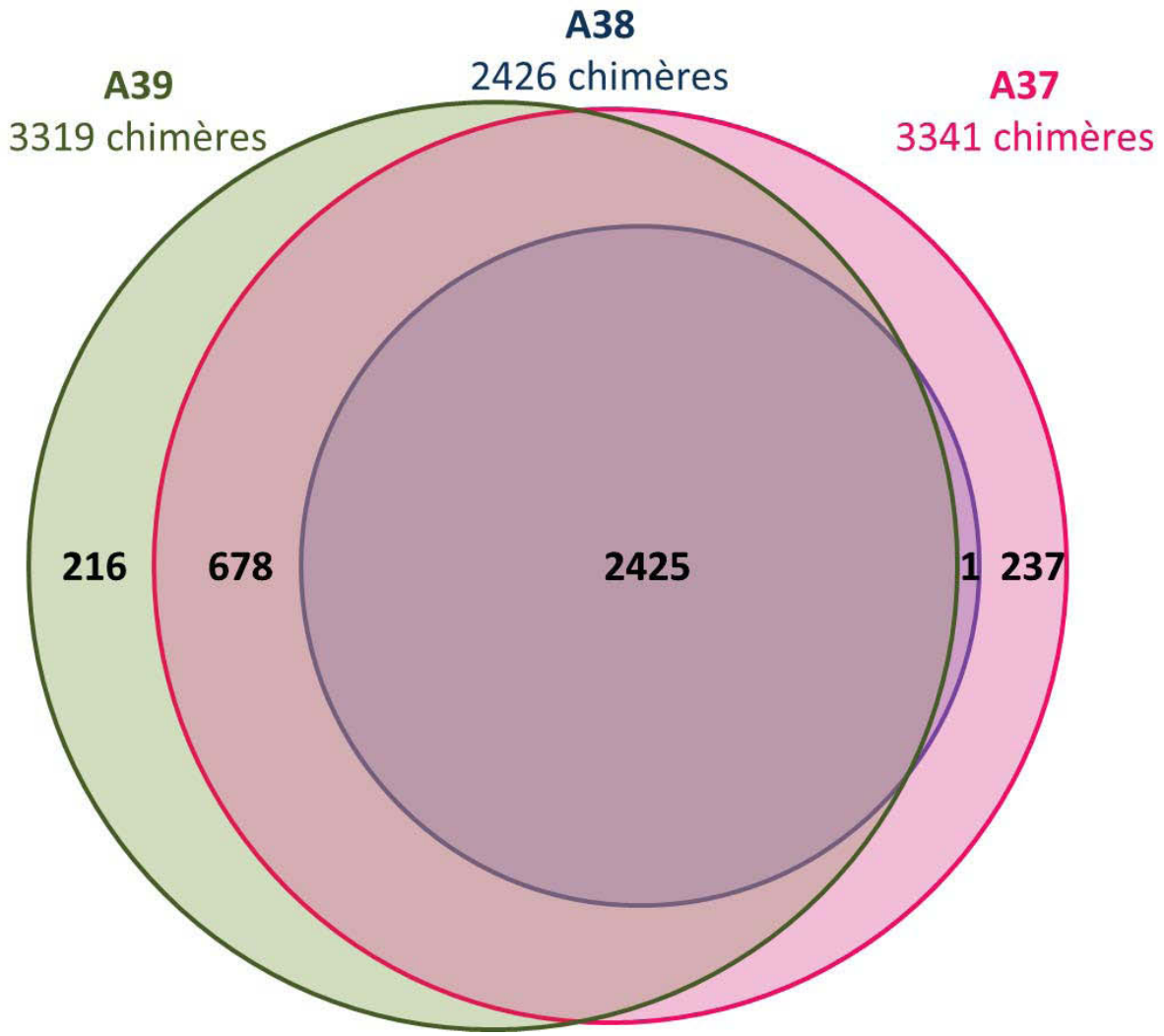


Figure 35 : La majorité des chimères sont communes aux 3 analyses A37, A38 et A39.

L'analyse A37 où le R1 doit posséder au moins 30 pb consécutives correspondant aux séquences consensus de 27 sous-familles avec 6 mésappariements tolérés et le R2 au moins 30 pb consécutives correspondant à une séquence unique du génome non annotée par RepeatMasker. L'alignement des séquences R1 et le R2 sur le génome humain (Hg19) à une distance de 50 Kb. Cette analyse A37 détecte 3624 chimères.

L'analyse A38 où le R1 doit posséder au moins 50 pb consécutives correspondant aux séquences consensus de 27 sous-familles avec 6 mésappariements tolérés et le R2 au moins 30 pb consécutives correspondant à une séquence unique du génome non annotée par RepeatMasker. L'alignement des séquences R1 et le R2 sur le génome humain (Hg19) à une distance de 50 Kb. Cette analyse A38 détecte 2426 chimères.

L'analyse A39 où le R1 doit posséder au moins 50 pb consécutives correspondant aux séquences consensus de 27 sous-familles avec 10 mésappariements tolérés et le R2 au moins 30 pb consécutives correspondant à une séquence unique du génome non annotée par RepeatMasker. L'alignement des séquences R1 et le R2 sur le génome humain (Hg19) à une distance de 50 Kb. Cette analyse A39 détecte 3319 chimères.

L'analyse A38 est totalement englobée dans les analyses A37 et A39. 93% des chimères sont communes entre les analyses A37 et A39.

Partie 2 : Etendue pangénomique de la dérégulation transcriptionnelle liée aux LCT dans les gliomes.

1. Objet de l'étude partie 2.

L'outil CLIFinder étant validé, le deuxième objectif de ma thèse qui consiste en l'étude à l'échelle pangénomique de l'implication des LCT dans les tumeurs a pu être entrepris.

Pour ce faire, les gliomes ont été choisis comme modèle d'étude. Les gliomes correspondent aux tumeurs du cerveau les plus fréquents chez l'adulte et ils se comportent à l'image d'un grand nombre d'autres cancers. En effet, parmi les gliomes, une hypométhylation globale de l'ADN est retrouvée spécifiquement dans 80% des GBM qui correspondent aux gliomes les plus sévères. Celle-ci affecte alors notamment les séquences promotrices de L1 et semble liée à l'agressivité tumorale (Cadieux *et al.*, 2006; Ohka *et al.*, 2011).

Une collaboration avec l'équipe CREaT EA 7282 dirigée par le Pr P. Verrelle (CRLCC Jean Perrin), nous a donné accès à la « Tumorothèque Auvergne Gliomes » (n°DC-2012-1584). Celle-ci inclut 87 échantillons de gliomes pour lesquels les données clinico-biologiques des patients (survie, traitement, âge au diagnostic, …) sont disponibles (Annexe 22).

Dans ce contexte, 13 gliomes ont été sélectionnés pour la réalisation des RNA-seq. Afin d'évaluer si des LCT peuvent être spécifiquement impliqués dans un sous-groupe tumoral, ceux-ci incluent 5 gliomes de bas grades (*IDH1* muté, LGG) et 8 gliomes de haut grade (*IDH1* WT, HGG). Par ailleurs, afin d'étudier également l'implication des LCT dans les tissus normaux, 3 tissus cérébraux contrôles provenant de la « Brain and Tissue Bank » du Maryland ont été sélectionnés. Ceux-ci correspondent à du corps calleux ou de la substance blanche corticale, deux zones du cerveau particulièrement riches en cellules gliales (oligodendrocytes et astrocytes) dont sont issus les gliomes.

Enfin, comme nous ne savons pas si les LCT sont polyadénylés ou non, nous avons choisi de réaliser un séquençage sans *a priori* de tous les transcrits en utilisant les ARN totaux après déplétion en ARN ribosomaux.

2. Résultats de la partie 2.

2.1. Optimisation des conditions d'analyse par CLIFinder.

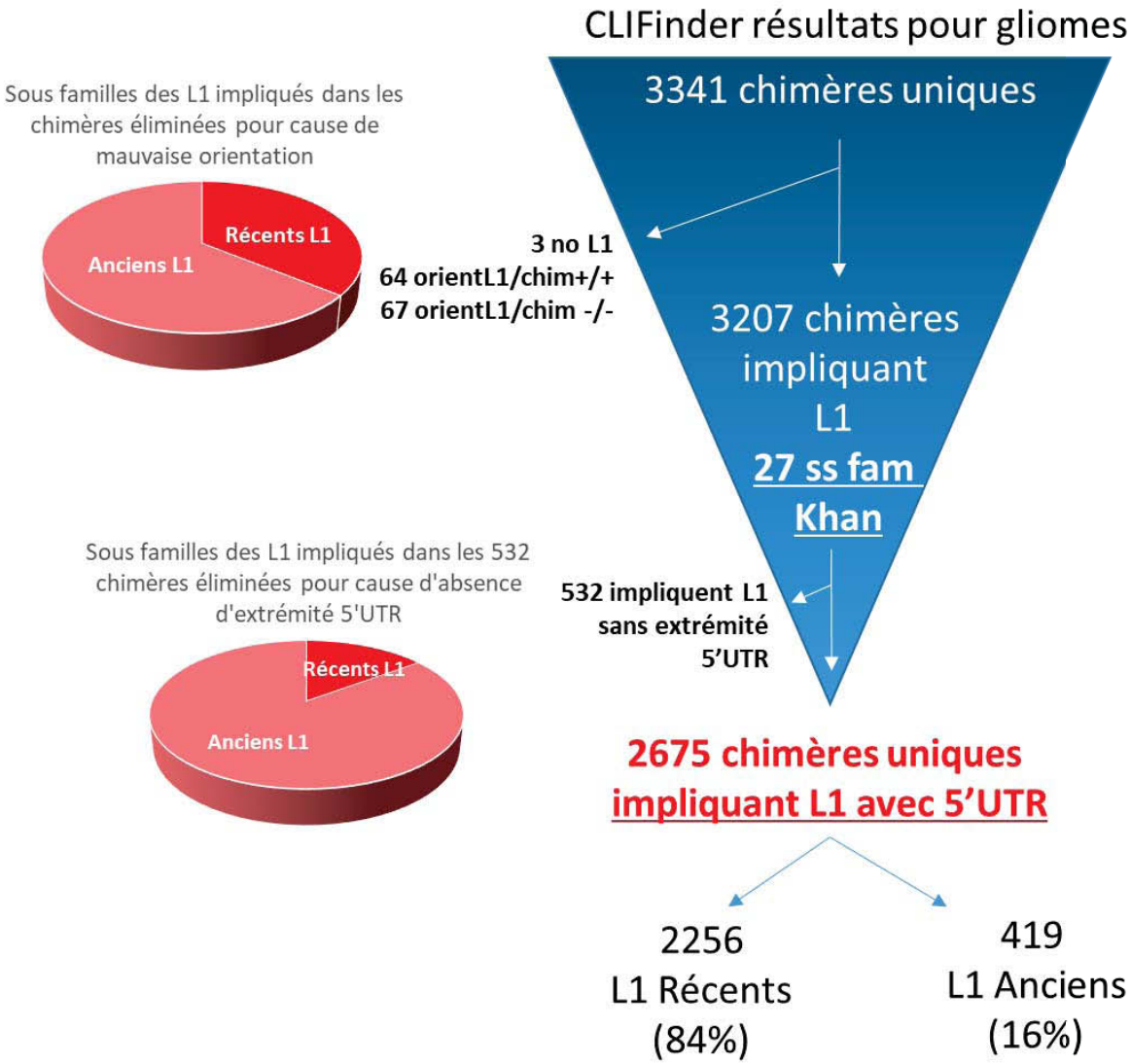


Figure 36: Elimination des chimères artéfactuelles identifiées par CLIFinder (A37)

3341 chimères uniques sont détectées initialement dans les données de RNA-seq avec CLIFinder. Parmi elles, 3 ne possèdent pas de L1 associé. 64 ont une orientation du L1 et de la chimère dans le sens + et 67 dans le sens -. 532 impliquent des L1 sans la région +400 et +600 de l'extrémité 5'UTR. Au final, 2675 chimères uniques impliquant des L1 avec une extrémité 5'UTR sont identifiées. La majorité des chimères artéfactuelles éliminées impliquent des L1 anciens (antérieurs à L1PA7).

Le logiciel CLIFinder permet d'identifier dans des données de RNA-seq des transcrits chimériques initiés à l'ASP. Ce RNA-seq a été réalisé à partir des ARN totaux déplétés en ARNr provenant de 13 échantillons de gliomes et 3 tissus contrôles. Ces RNA-seq sont paired-end orientés de 100 pb à chaque extrémité. Sur l'ensemble des lectures de RNA-seq, un premier filtre est appliqué pour sélectionner des paires dont l'extrémité en 5' (R1) correspond à une séquence de L1 en antisens. Les paramètres de ce premier filtre peuvent être adaptés à savoir : les séquences de référence utilisées pour la détection des L1, la longueur de la séquence L1 détectée au niveau du R1, et si des mésappariements sont tolérés entre la séquence de référence du L1 et la séquence R1. Le second filtre est appliqué sur les paires retenues afin d'identifier celles associées à une extrémité en 3' (R2) correspondant à une séquence unique du génome. Pour ce filtre la longueur de la séquence unique identifiée au niveau du R2 peut être choisie par l'utilisateur. Ensuite ces lectures filtrées sont alignées sur le génome humain avec un espacement toléré entre le R1 et le R2 de 50 Kb au maximum. Là encore, la distance d'alignement entre le R1 et le R2 peut être modifiée. Enfin, si plusieurs lectures correspondent au même locus, celles-ci sont concaténées afin d'établir une liste de chimères avec pour chaque échantillon de RNA-seq analysé le nombre de lectures définissant chaque chimère.

La liste des séquences consensus de 27 sous-familles de L1 récents décrit par Khan a été retenue pour notre analyse avec CLIFinder (Khan, 2005). Sachant que l'ASP est situé dans la région +400 à +600 pb dans la région 5'UTR de L1, les 540 premières paires de bases des séquences consensus des 27 sous-familles ont été utilisées pour le filtre 1 afin de détecter des chimères initiées à l'ASP (Figure 6).

Dans la première analyse (A1) réalisée pendant mon master 2, le filtre 1 était paramétré pour détecter des lectures possédant une extrémité R1 correspondant à 100% aux séquences consensus des 27 sous-familles de L1. Pour cette analyse A1, le filtre 2 devait sélectionner des extrémités R2 possédant au moins 70 pb consécutives correspondant à une séquence unique. Cette analyse A1 a identifié 535 chimères dans les données de RNA-seq. Ces conditions ont rapidement été identifiées comme trop restrictives sur le filtre 1 car :

1) en imposant 100% d'identité pour R1 avec les séquences consensus, elles ne tolèrent pas que la zone de jonction entre le L1 et la séquence unique puisse se situer dans R1, ce qui élimine des lectures.

2) elles ne tolèrent pas de mésappariements dans la séquence du R1 vis-à-vis des séquences consensus de 27 sous-familles qui ont été utilisées comme référence ; alors que des variations existent entre les différents éléments L1.

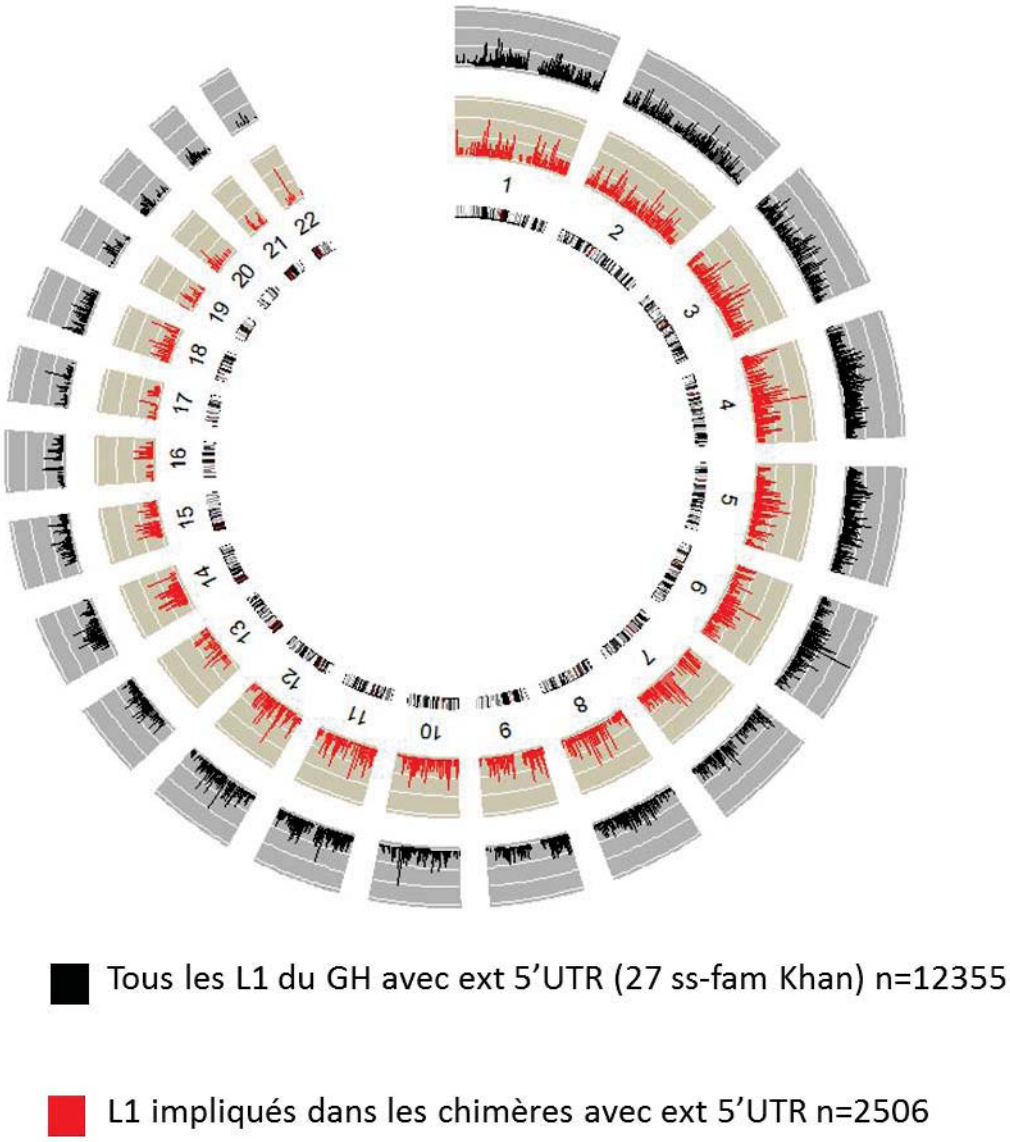


Figure 37 : Analyse de la distribution génomique des L1 impliqués dans les chimères.

Le nombre de L1 par Mb sur les 22 autosomes du génome humain est rapporté en noir pour tous les L1 du génome humain des 27 sous-familles de Khan possédant une extrémité 5'UTR (n=12355), en rouge pour les L1 avec une extrémité 5'UTR impliqués dans les chimères (n=2506).

Aussi cette analyse A1 sous-estimait le nombre de chimères pouvant être détectées dans les données de RNA-seq. Ces résultats nous ont indiqué qu'il fallait modifier le logiciel afin de capturer toutes les configurations pouvant avoir lieu. Aussi au début de ma thèse, le logiciel CLIFinder a été implémenté au niveau du filtre 1 permettant de choisir le nombre de pb correspondant à une séquence de L1 ainsi que le nombre de mésappariements tolérés entre R1 et les séquences consensus de référence. Différentes analyses ont été ensuite réalisées sur nos RNA-seq afin de définir les paramètres les plus pertinents en termes de sensibilité et spécificité. Trois analyses faisant varier les paramètres du filtre 1 ont été réalisées :

1) Analyse A37 où le R1 doit posséder au moins 30 pb consécutives correspondant aux séquences consensus de 27 sous-familles avec 6 mésappariements tolérés. Le filtre 2 sélectionne le R2 s'il possède au moins 30 pb consécutives correspondant à une séquence unique du génome non annotée par RepeatMasker. L'alignement des séquences sur le génome humain (hg19) tolère une distance de 50 Kb entre le R1 et le R2. Cette analyse A37 a permis de détecter 3341 chimères uniques dans les données de RNA-seq de gliomes.

2) Analyse A38 où le R1 doit posséder au moins 50 pb consécutives correspondant aux séquences consensus de 27 sous-familles avec 6 mésappariements tolérés. Le critère de sélection pour le filtre 2 et la distance d'alignement n'ont pas été modifiés. Cette analyse A38 détecte 2426 chimères uniques dans les données de RNA-seq.

3) Analyse A39 où le R1 doit posséder au moins 50 pb consécutives correspondant aux séquences consensus de 27 sous-familles avec 10 mésappariements tolérés. Le critère de sélection pour le filtre 2 et la distance d'alignement entre le R1 et le R2 restent inchangés. Cette analyse A39 détecte 3319 chimères.

La comparaison des chimères détectées dans ces 3 analyses (Figure 35) montre que les chimères détectées par l'analyse 38 sont totalement incluses dans les analyses A37 et A39. De plus, 93% des chimères sont communes entre les analyses A37 et A39. La condition A37 est la moins stringente car elle identifie des chimères supplémentaires présentes dans plusieurs échantillons alors que les chimères supplémentaires identifiées par l'analyse A39 sont détectées majoritairement que dans 1 échantillon parmi 13. Il a donc été décidé d'exploiter les résultats de l'analyse A37.

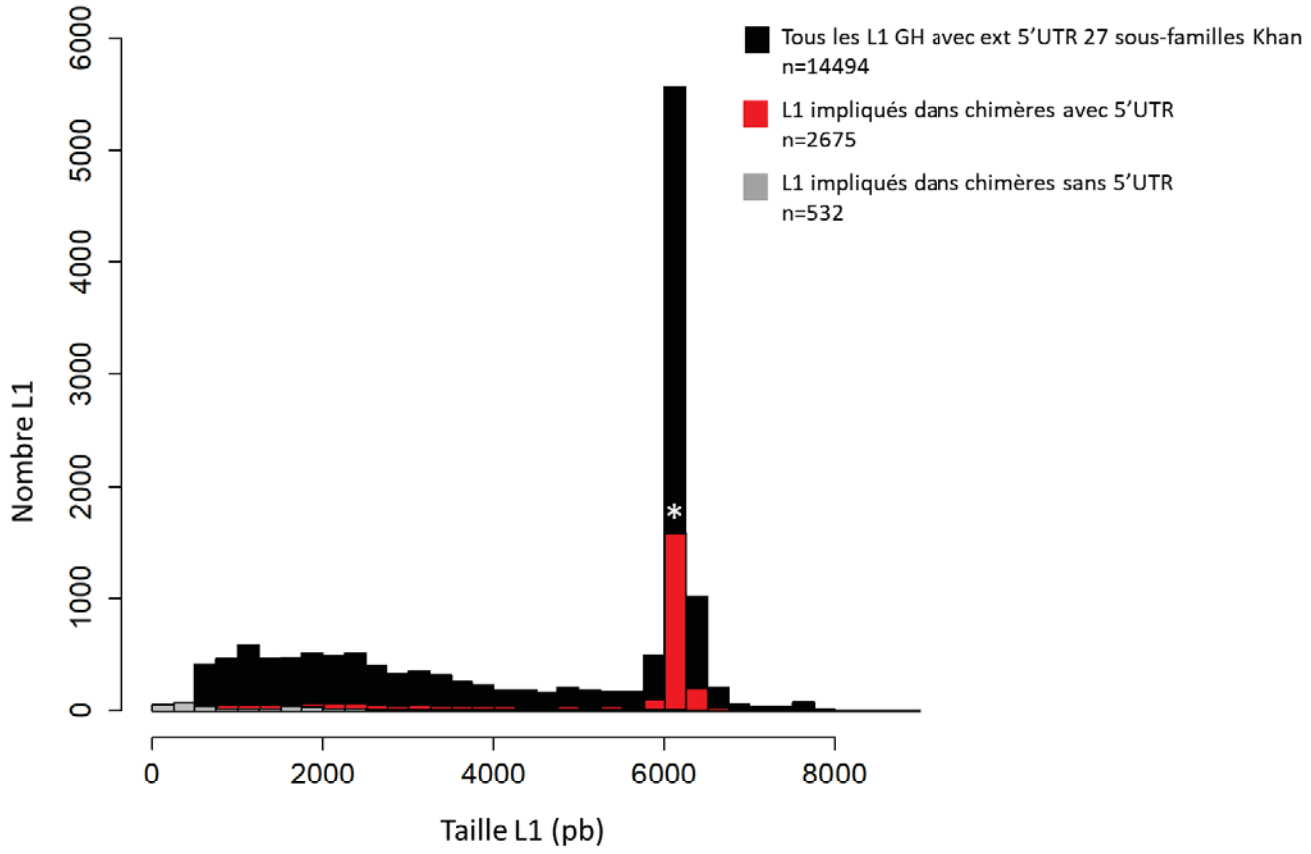


Figure 38 : Les L1 impliqués dans les chimères sont significativement enrichis en L1 pleine taille.

Distribution de la taille de l'ensemble des L1 (27 sous familles Kahn) du génome humain contre celle des L1 impliqués dans les chimères. La taille des L1 (en pb) en abscisses est rapportée au le nombre de L1 en ordonnées soit dans le génome humain en noir (n=14494), soit le nombre de L1 impliqués dans la formation des chimères possédant une extrémité 5'UTR en rouge (n=2675), soit le nombre de L1 impliqués dans la formation de chimères ne possédant pas d'extrémité 5'UTR en gris (n=532).

2.2. Identification de nombreuses chimères dans les gliomes et les tissus contrôles.

CLIFinder détecte 3341 chimères dans les données de RNA-seq de 13 gliomes et de 3 tissus contrôles. Lors de la validation du logiciel (Partie 1), il est apparu que certaines chimères identifiées par CLIFinder peuvent ne pas répondre aux critères d'un LCT potentiel, à savoir :

- Un sens de transcription de la chimère identique au sens de transcription du L1, ce qui est incompatible avec une initiation de la transcription de la chimère à l'ASP du L1.
- Le L1 associé ne possède pas la région 5'UTR +400 à +600 et ne contient donc pas l'ASP, au moins dans les sous-familles les plus récentes. Ceci peut être dû au mécanisme de rétrotransposition des L1, la majorité des copies présentes dans le génome humain étant tronquées en 5' mais aussi à des réarrangements comme l'insertion d'autres ET, tel que les éléments *Alu*, ou bien à des délétions au niveau de cette région.

Une première étape de contrôle des caractéristiques des chimères a donc été réalisée pour éliminer ces chimères artéfactuelles (Figure 36). Ainsi, ont été éliminées de la liste initiale :

- 3 chimères ne présentant pas de L1 associé,
- 131 chimères présentant une transcription orientée dans le même sens que celle du L1 à savoir 64 en orientation +/+ et 67 en orientation -/-,
- 535 chimères impliquant des L1 ne possédant pas la région critique +400 à +600 dans laquelle se trouve l'ASP.

Ceci correspond à 19% de la liste initiale. Toutefois, il est à noter que ces chimères éliminées impliquent majoritairement des L1 anciens, c'est-à-dire des sous-familles antérieures à L1PA7 (soit 64% pour le problème d'orientation, et 85% pour les L1 sans extrémité 5'UTR).

Au final, 2675 chimères uniques sont détectées dans les données de RNA-seq. Parmi celles-ci, 2256 impliquent des L1 récents (L1PA1 à 7) et 419 impliquent des L1 anciens (antérieurs à L1PA7). Ceci suggère que CLIFinder identifie majoritairement des chimères pouvant correspondre à un LCT puisque 84% impliquent des L1 récents pour lesquels un ASP fonctionnel a été décrit.

Sachant que les génomes tumoraux présentent de nombreux réarrangements dont des duplications et des délétions, la question s'est posée de savoir si les L1 impliqués dans la formation des chimères avaient une distribution chromosomique particulière comparée à la distribution de tous les L1 présents dans le génome humain. Quand la distribution des L1 impliqués dans la formation des chimères est comparée à celle des L1 présents dans le génome humain sur les autosomes (Figure 37), aucune différence de distribution n'est observée.

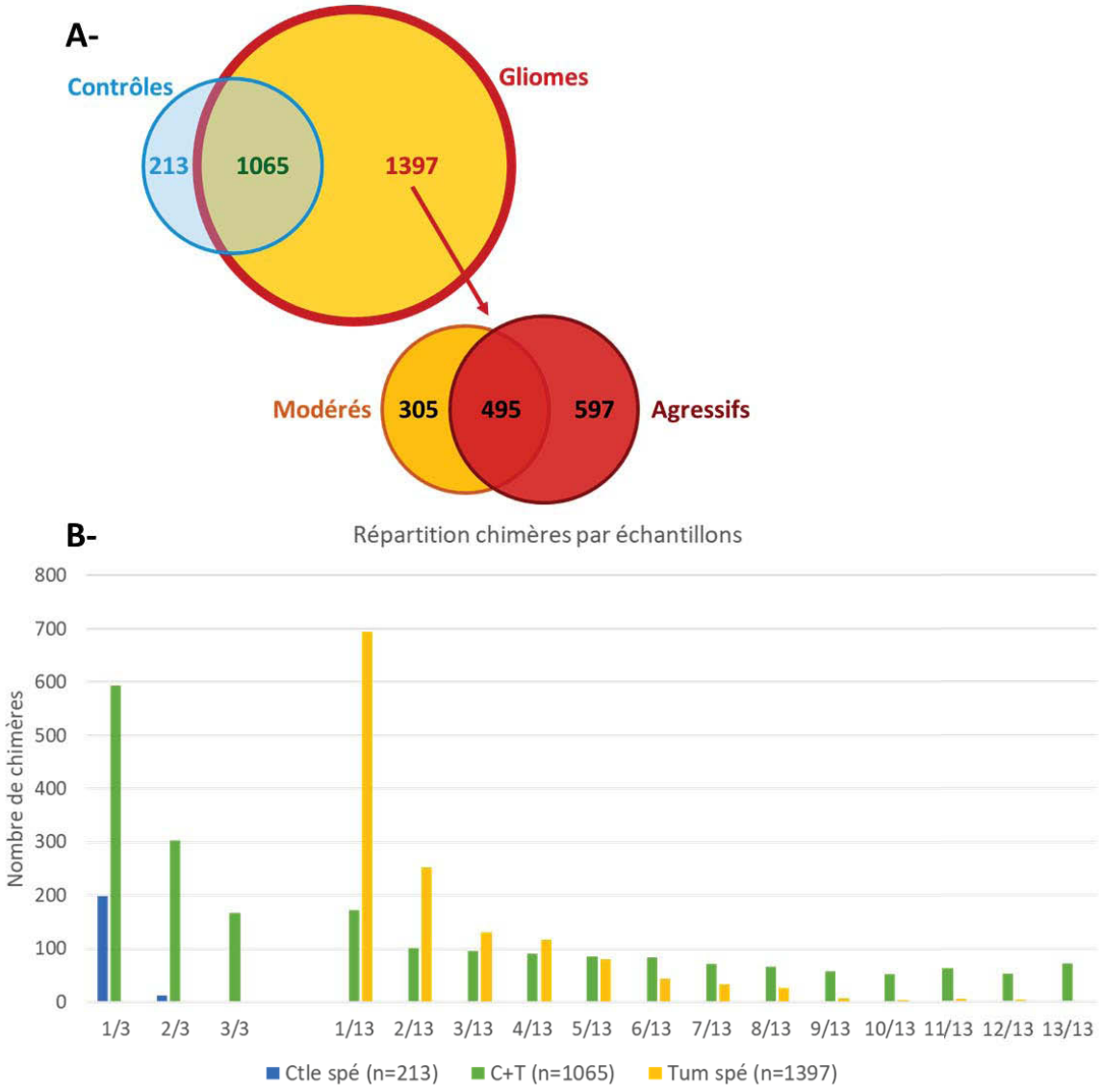


Figure 39 : La majorité des chimères détectées avec CLIFinder sont tumeurs spécifiques et récurrentes dans les échantillons tumoraux.

A- 2675 chimères impliquant un L1 possédant une extrémité 5'UTR sont identifiées par l'analyse A37. 213 chimères (soit 8%) sont uniquement détectées dans les échantillons contrôles (« Ctrl spé »), 1065 chimères (soit 40%) sont détectées dans les contrôles et les gliomes (« C+T ») et 1397 chimères (soit 52%) sont détectées uniquement dans les gliomes (« Tum spé »). Pour ces dernières, la répartition selon les deux groupes de sévérité de gliomes est reportée. 305 chimères (soit 22%) sont uniquement détectées dans les gliomes modérés, 597 chimères (soit 43%) sont uniquement détectées dans les gliomes agressifs et 495 chimères (soit 35%) sont détectées dans les deux groupes tumoraux.

B- Répartition des chimères par échantillon pour les chimères « Ctrle spé » (n=213), les chimères « Tum spé » (n=1397) et les chimères « C+T » (n=1065). Le nombre de chimère est reporté sur l'axe des ordonnées et en abscisse la fréquence d'apparition par échantillon soit dans les contrôles à gauche (1 contrôle parmi 3 = 1/3), soit dans les gliomes (2 tumeurs parmi 13 = 2/13).

Afin de mieux caractériser les L1 impliqués dans les chimères, l'analyse de leur taille a été réalisée. 67% des L1 impliqués dans la formation des 2675 chimères correspondent à des L1 pleine taille mesurant 6Kb. Par rapport à la distribution de la taille de tous les L1 possédant une extrémité 5'UTR, ceci représente un enrichissement significatif en L1 pleine taille. En revanche, les L1 impliqués dans la formation des 532 chimères qui ne possèdent pas d'extrémités 5'UTR sont majoritairement tronquées et ont une taille inférieure à 2 Kb (Figure 38). En conclusion, les chimères identifiées par CLIFinder impliquent en majorité des L1 récents pleine taille, supposés posséder un ASP fonctionnel, et n'ayant pas de localisation préférentielle dans le génome humain.

Qu'en est-il de la distribution tissulaire des chimères identifiées ?

L'analyse du nombre de lectures détectées par CLIFinder montre que le nombre moyen de chimères par échantillon est de 723,74 +/- 94,60 et qu'il n'y a pas de différence notable entre les échantillons contrôles et les échantillons tumoraux. De plus, 92% des chimères (2462/2672) sont détectées dans au moins 1 échantillon tumoral (Figure 39-A). Parmi celles-ci :

- 52% sont détectées uniquement dans les gliomes et semblent donc être tumeurs spécifiques (« Tum spé »),
- 40% sont détectées non seulement dans les gliomes mais aussi dans au moins 1 échantillon contrôle (« C+T »).

Par ailleurs, quelques chimères sont détectées uniquement dans les échantillons contrôles « Ctrl spé » (n=213). Parmi les 1397 chimères « Tum spé», 22% sont uniquement détectées dans les gliomes modérés possédant la mutation dans le gène *IDH1* (LGG), 43% sont uniquement détectées dans les gliomes agressifs ne possédant pas la mutation du gène *IDH1* (HGG). Aussi il semblerait que certaines de ces chimères pourraient être spécifiques des gliomes modérés ou agressifs.

Comment ces chimères se répartissent au sein des échantillons ? L'étude de la répartition des chimères par échantillon (Figure 39-B), montre que la majorité des chimères « Ctrl spé » (93%) n'est détectées que dans un échantillon alors que pour les chimères « Tum spé » seulement 50% ne sont détectées que dans un échantillon tumoral. Pour les chimères « C+T », 55% et 16% des chimères sont détectées respectivement dans 1 contrôle et dans 1 tumeur. Ces résultats montrent donc que la moitié des chimères « Tum spé » et 84% des chimères « C+T » sont détectées dans au moins 2 échantillons tumoraux différents. Cette observation conforte l'existence de ces chimères et implique une certaine récurrence. Cette dernière suppose que ces chimères pourraient jouer un rôle fonctionnel dans le développement tumoral que ce soit dans l'initiation,

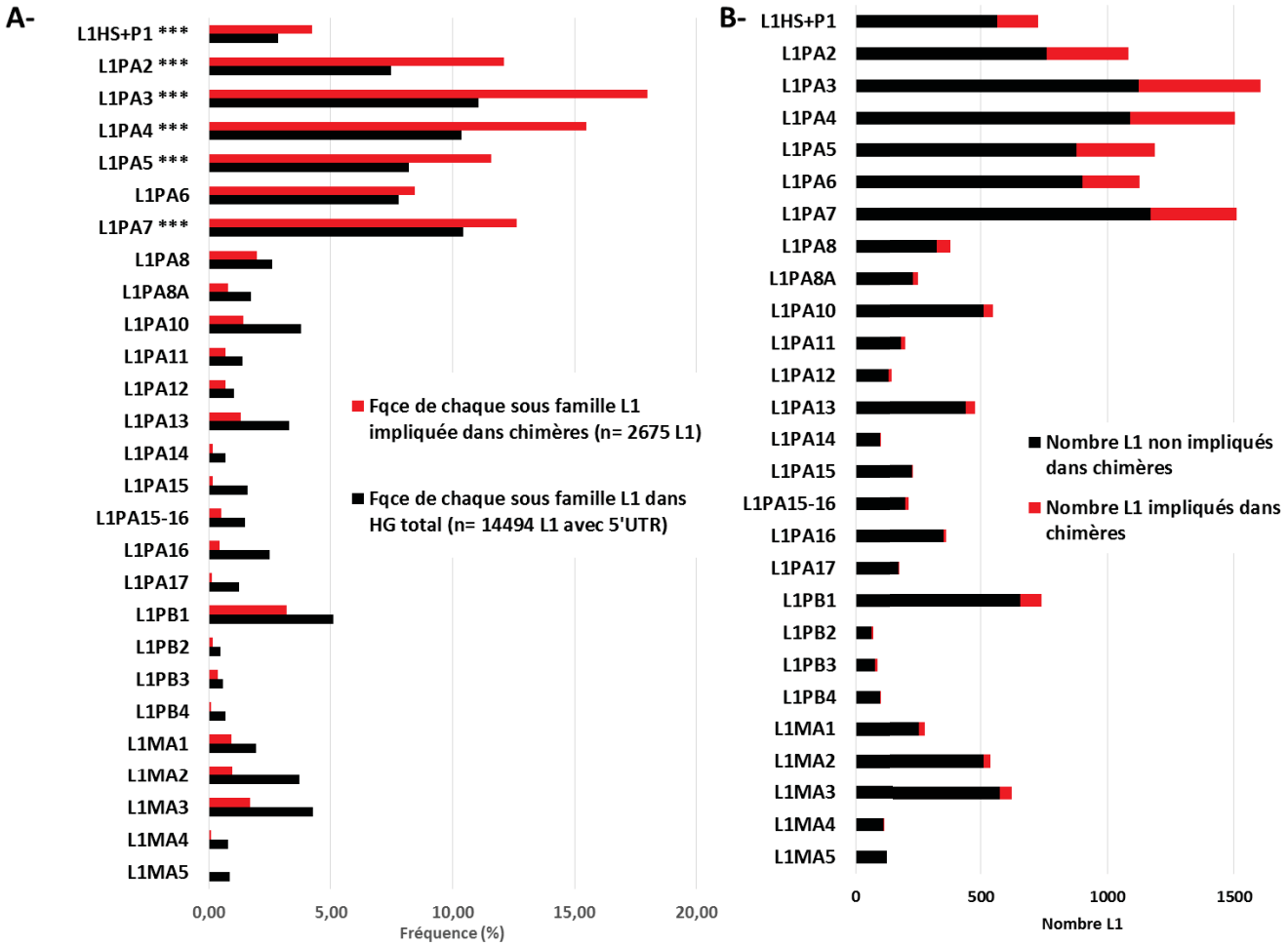


Figure 40 : Implication majoritaire des sous-familles de L1 récentes (L1PA1 à 7) dans les chimères identifiées par A37.

A- Comparaison de la fréquence de chaque sous-famille de L1 des 27 sous-familles de Khan présente dans le génome humain en noir et celle des L1 impliqués dans les chimères en rouge. Les étoiles représentent un enrichissement significatif pour les sous-familles les plus récentes (test binomial).

B- Nombres de L1 présents dans le génome humain non impliqués dans les chimères en noir et le nombre de L1 impliqués dans les chimères en rouge pour les 27 sous-familles de Khan. Au final, 25% les L1 récents du génome humain sont impliqués dans la formation des chimères.

la progression et/ou l'agressivité. Ces chimères pourraient également être des biomarqueurs moléculaires d'un grade tumoral.

Par ailleurs, parmi les 2675 chimères, 1169 (44%) sont en position intergénique et 1506 (56%) sont en position intragénique. Quand ces pourcentages sont comparés à ceux des 14 494 L1 du génome humain possédant une extrémité 5'UTR, cette tendance s'inverse puisque près de 60% des L1 sont intergéniques et que 40% sont en position intragénique. Ceci renforce l'idée selon laquelle les LCT pourraient jouer un rôle fonctionnel dans la tumorigenèse.

Enfin, la majorité des chimères identifiées semblent correspondre à des évènements de transcription « continus » entre le L1 et la séquence unique adjacente. En effet, la distance entre l'extrémité 3' de R1 avec l'extrémité 5' de R2 est :

- Dans 50% des cas (1323 chimères) négative ou nulle, ce qui démontre un chevauchement ou la juxtaposition des séquences R1 et R2. Cette situation implique alors que la chimère a été identifiée par plusieurs lectures (provenant d'un même échantillon ou de plusieurs échantillons différents)
- Dans 41 % des cas (1101 chimères) comprise entre 1 et 200 pb, ce qui implique que les fragments d'ADNc correspondant à R1 et R2 ont une taille correspondant à la sélection faite en amont de la réaction de séquençage.

Pour autant, 70 chimères (2,6%) présentent une distance entre R1 et R2 > 400 pb ce qui suggère qu'un évènement d'épissage a eu lieu entre les deux séquences. La distance maximale observée dans nos chimères est de 49 268 pb, concordante avec le critère de 50 kb maximum appliqué lors de l'analyse par CLIFinder.

2.3. La majorité des chimères impliquant des L1 récents semble correspondre à des LCT initiés à l'ASP.

Grâce au logiciel CLIFinder, 2675 chimères ont été détectées dans les données de RNA-seq de gliomes. La question se pose donc de savoir si ces chimères correspondent à des LCT initiés à l'ASP de L1 ?

2.3.1. Les chimères identifiées impliquent des L1 récents supposées posséder un ASP.

Nous avons vu que 84% des chimères identifiées dans l'analyse A37 impliquent des L1 récents (L1PA1 à 7). Les différentes sous-familles de L1 impliquées dans la formation des chimères

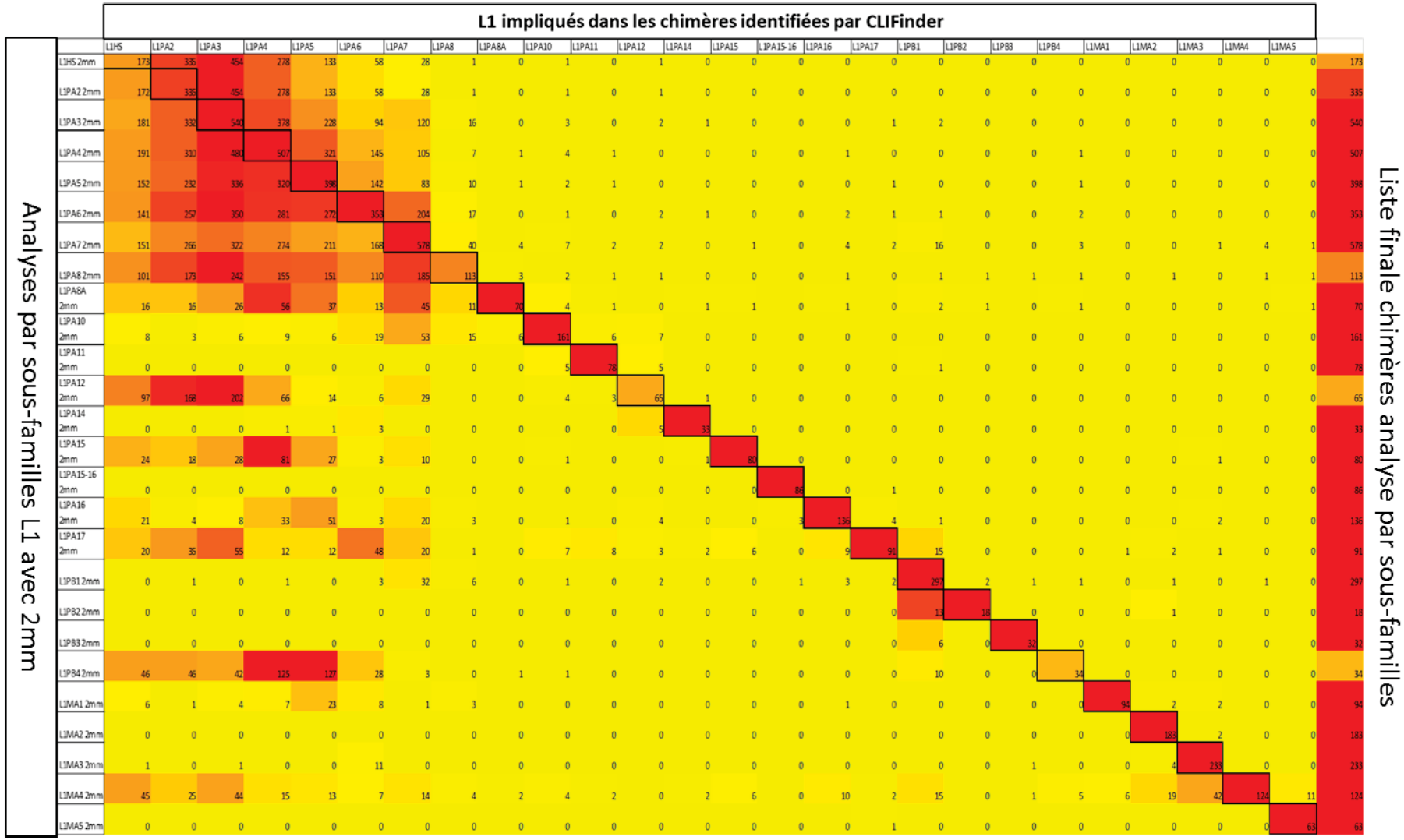


Figure 41 : **Ensemble des résultats obtenus pour l'analyse par sous-famille avec 2 mésappariements tolérés.**

Suite à l'analyse par sous-famille avec CLIFinder sur les données de RNA-seq avec les séquences de tous les L1 de chaque sous-famille présente dans le génome humain, les chiffres sont rapportés dans ce tableau. Plus la couleur est rouge plus le nombre de chimères détectées est élevée. Pour chaque sous-famille, le chiffre correspondant au nombre de chimères identifiées impliquant la sous-famille considérée (encadré en noir) est conservé (dans la dernière colonne). Ceci a permis d'établir les listes des chimères identifiées pour chaque analyse (0, 1 et 2mm) par sous-famille qui sont utilisé pour les analyses suivantes.

ont été analysées afin d'identifier leur caractéristique. Quand la fréquence de chaque sous-famille de L1 présente dans le génome humain est comparée à celle des sous-familles de L1 impliquées dans la formation des chimères, un enrichissement significatif pour les sous-familles de L1 les plus récentes (L1PA1 à 5 et L1PA7) est observé dans les chimères (Figure 40-A). Ceci est en accord avec l'hypothèse de Speek et les validations de Macia, qu'un ASP fonctionnel serait présent dans les sous-familles les plus récentes (L1PA1 à 7). Par ailleurs, quand le nombre de L1 impliqués dans la formation des chimères est rapporté au nombre total de L1 possédant une extrémité 5'UTR dans le génome humain, ceci montre que 25% des L1 récents du génome humain sont impliqués dans la formation des chimères (Figure 40-B). Ces résultats confortent notre hypothèse que les chimères identifiées par CLIFinder peuvent correspondre à des LCT initiés à l'ASP d'un L1 récent.

Pour autant, la question s'est posée de savoir si l'utilisation des séquences consensus des 27 sous-familles de L1 en autorisant « seulement » 6 mésappariements sur au moins 30 pb de séquence L1 ne pouvait pas induire un biais dans notre analyse. En effet, les éléments L1 les plus anciens ont pu accumuler au cours du temps des mutations ponctuelles dans leur séquence 5'UTR. Les 6 mésappariements tolérés peuvent donc être une condition trop stringente, empêchant l'identification de séquences transcrites chimères impliquant ces éléments. Aussi, les séquences de tous les L1 de chacune des 27 sous-familles ont été extraites du génome humain à partir de RepeatMasker dans UCSC. Puis le logiciel CLIFinder a été utilisé avec pour séquences de référence, non pas les séquences consensus des 27 sous-familles mais toutes les séquences promotrices des L1 du génome humain, par sous-famille. En théorie, l'utilisation des séquences promotrices de tous les L1 d'une sous-famille pourrait être réalisée en tolérant 0 mésappariement. Toutefois, l'existence de variations interindividuelles (SNP) pouvant aussi affecter les séquences L1, les analyses ont également été réalisées en tolérant 1 ou 2 mésappariements. L'observation des 81 résultats (analyse de 23 sous-familles x 3 conditions de mésappariements) montre que, malgré le fait que chaque analyse ait été réalisée contre les séquences promotrices des éléments d'une sous-famille donnée, il existe suffisamment d'homologie de séquence pour identifier des chimères impliquant des L1 d'autres sous-familles. Ceci est plus particulièrement noté pour les sous-familles les plus récentes (L1PA1 à 8A), les plus proches d'un point de vue évolutif (Figure 41). Toutefois, pour une sous-famille donnée, le nombre de chimère le plus important est toujours obtenu avec l'analyse de la sous-famille considérée. L'établissement des 3 listes globales de chimères obtenues pour les analyses par sous-familles pour 0, 1 et 2 mésappariements (mm) a donc consisté à retenir pour chaque sous-famille le nombre de chimères obtenu avec l'analyse de la sous-famille considérée. Ainsi

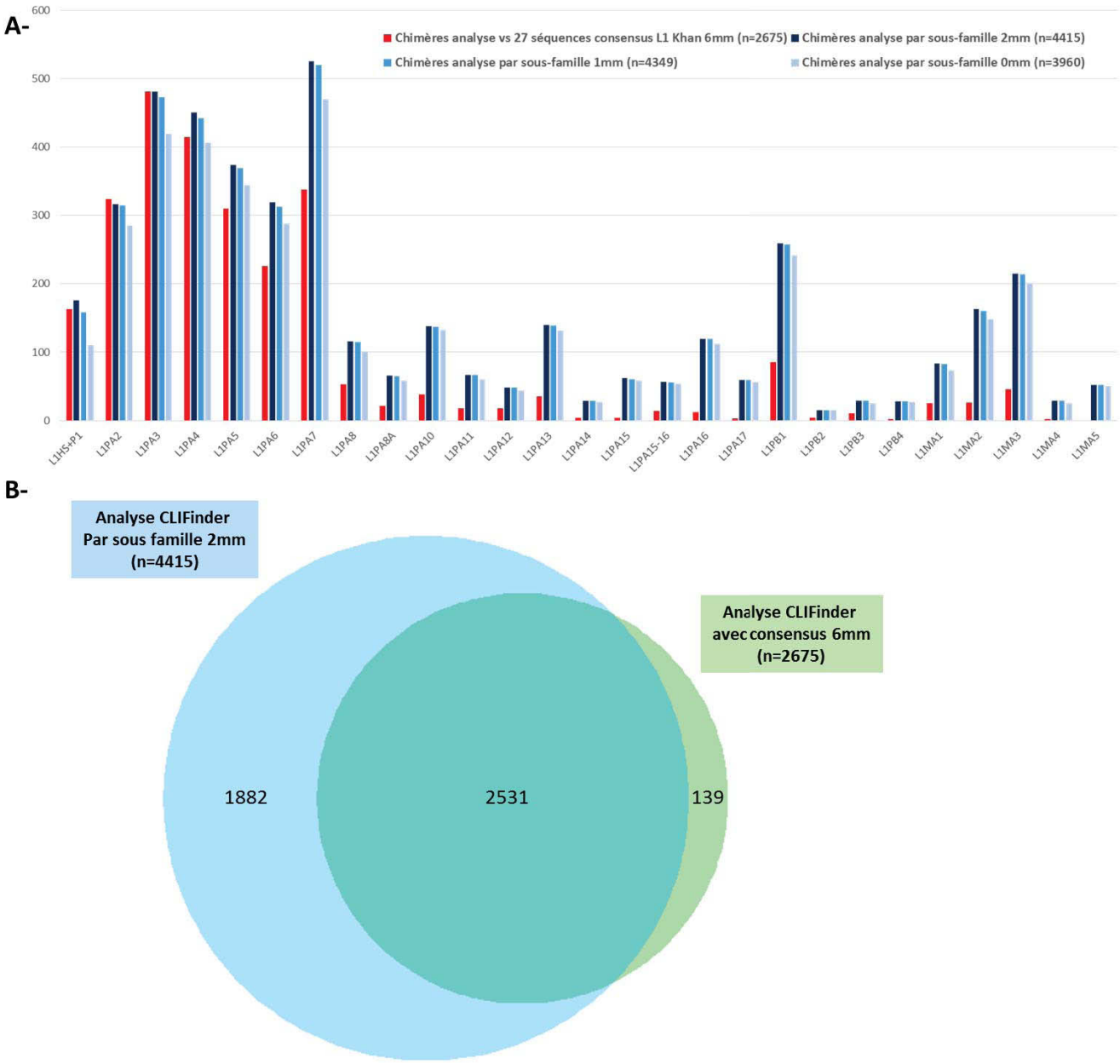


Figure 42 : Comparaison du nombre de chimères par sous-familles de L1 en fonction des conditions d'analyses.

A- Le nombre de chimères détectées pour chaque analyses par sous-familles est représenté, avec en rouge le nombre de chimère détectées par l'analyse A37 réalisée avec les séquences consensus (6 mésappariements tolérés), en bleu foncé le nombre de chimères détectées avec les séquences des L1 du génome humain par sous-familles avec 2 mésappariement tolérés, en bleu ciel le nombre de chimères détectées avec les séquences des L1 du génome humain par sous-familles avec 1 mésappariement toléré, en bleu clair le nombre de chimères détectées avec les séquences des L1 du génome humain par sous-familles avec 0 mésappariement toléré.

B- L'ensemble des chimères identifiées par l'analyse A37 utilisant les séquences consensus des 27 sous-familles avec 6 mésappariements tolérés (en vert) avec l'ensemble des chimères identifiées par l'approche utilisant les séquences des L1 des différentes sous-familles avec 2 mésappariements tolérés sont représentés. La majorité des chimères A37 sont détectées dans l'analyse par sous-famille avec 2 mésappariements.

les analyses 0, 1 et 2 mm ont identifiées respectivement 3 960, 4 349 et 4 415 chimères, ce qui est plus important que les 2 675 chimères de l'analyse A37. La comparaison du nombre de chimères détectées pour chaque analyse par sous-familles (0mm, 1mm et 2mm) avec celles de l'analyse A37 (Figure 42-A) montre globalement que les chimères supplémentaires identifiées par les analyses par sous-familles impliquent majoritairement des L1 anciens. En effet un nombre de chimères plus important est obtenu avec les 3 nouvelles analyses, pour toutes les sous-familles antérieures à L1PA6 inclus. Pour autant, concernant les chimères impliquant les L1 récents possédant un ASP, le nombre de chimères est relativement identique voire parfois plus faible notamment avec les analyses tolérant 0 et 1 mm. En termes de nombre de chimère pour les sous-familles récentes (L1PA1 à 7), c'est donc l'analyse avec 2mm qui se rapproche le plus, voire complète, l'analyse A37.

La comparaison des positions génomiques des chimères de l'analyse A37 et de l'analyse par sous-familles avec 2 mm (Figure 42-B) montre que 95% des chimères détectées par l'analyse A37 (2531/2675) sont également détectées par l'analyse par sous-familles avec 2mm. Cette dernière retrouve donc les chimères identifiées par l'analyse A37 et enrichi les résultats principalement en chimères impliquant des L1 anciens.

L'ensemble de ces résultats démontre que l'analyse A37, utilisant les séquences consensus des 27 sous-familles et tolérant 6 mm sur 30 pb a donc introduit un biais défavorable à l'identification de chimères impliquant des L1 anciens (où la variabilité de séquence est plus importante). L'enrichissement significatif en sous-famille de L1 dans l'analyse A37 correspond donc à un artéfact lié aux conditions expérimentales. Toutefois, l'analyse A37 a permis l'identification de 2256 chimères impliquant des L1 récents (L1PA1 à 7) possédant l'ASP.

2.3.2. Certaines chimères identifiées correspondent à des LCT déjà décrits.

Afin d'identifier si les chimères détectées avec CLIFinder peuvent correspondre à des LCT, les positions chromosomiques des chimères obtenues avec CLIFinder ont été comparées à celles de LCT déjà décrits dans la littérature. Jusqu'en 2016, 59 LCT étaient décrits dans la littérature, tous impliquant des L1 récents avec une initiation de la transcription à l'ASP et donc potentiellement identifiables par CLIFinder si ces LCT sont présents dans nos RNA-seq (Cruickshanks & Tufarelli, 2009; Mätlik *et al.*, 2006; Nigumann *et al.*, 2002; Speek, 2001; Wolff *et al.*, 2010). En 2016, les travaux de Criscione ont identifié des transcrits LCT-EST dont la transcription est initiée en antisens d'un L1. Toutefois, l'origine des L1 associés aux LCT-EST identifiés est beaucoup plus large que les 27 sous-familles de Khan considérées dans cette

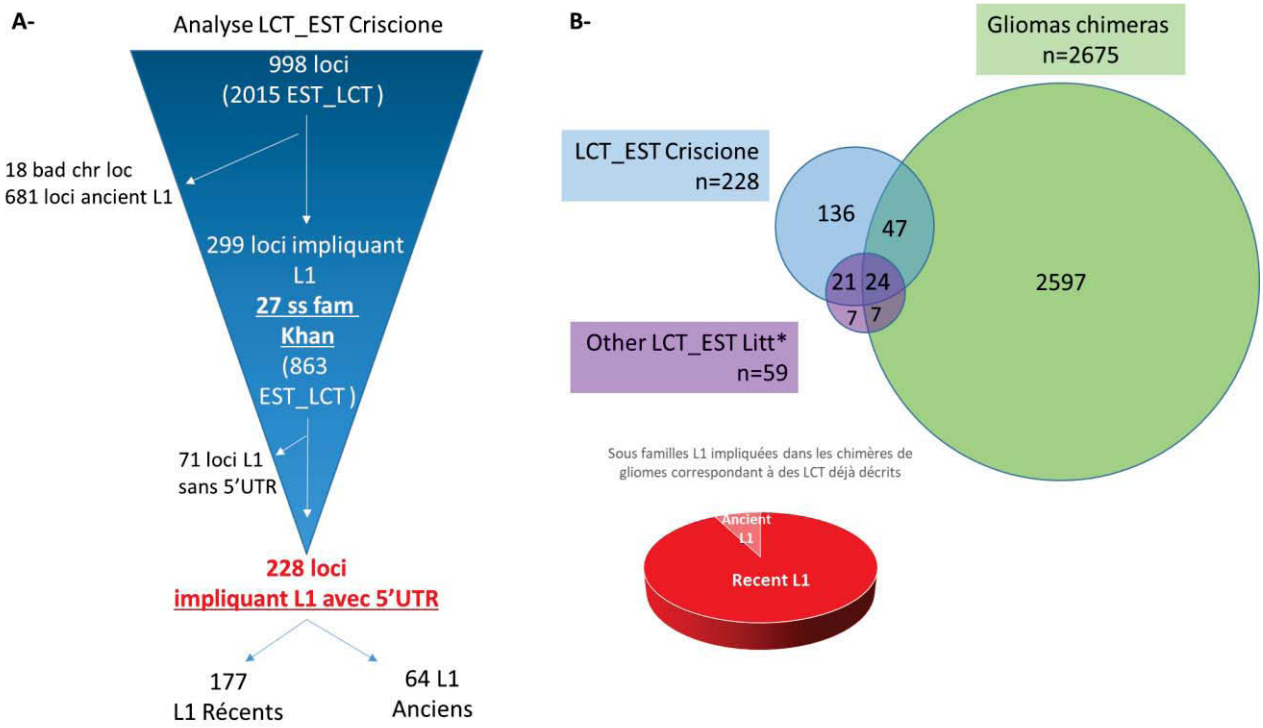


Figure 43 : 78 chimères identifiées dans les gliomes correspondent à des LCT déjà décrits dans d'autres types tissulaires.

A- Protocole de curation des chimères de la littérature au travers de l'exemple des chimères identifiées par Criscione (Criscione et al., 2016). Parmi les 2316 LCT identifiés dans les banques d'EST avec le protocole de Criscione, 1452 impliquent des L1 anciens non présents dans les sous-familles de Khan et 18 ont une mauvaise localisation. Au final, il y a 863 LCT impliquant des L1 des sous-familles de Khan dont 228 loci uniques possédant une extrémité 5'UTR avec 177 impliquant des L1 récents et 64 des L1 anciens.

B- Diagramme de Venn comparant la position des chimères de l'analyse A37 (en vert) avec celles des 228 LCT_EST de Criscione (en bleu) celles des LCT autres de la littérature (en violet). 78 chimères identifiées dans les gliomes correspondent à des LCT déjà décrits dans d'autres types tissulaires (normaux et tumoraux). Sur ces 78 LCT, 72 impliquent des L1 récents (L1PA1 à 7).

étude. Aussi, une première étape de tri des LCT-EST identifiées par Criscione a été réalisée afin de sélectionner les LCT-EST identifiables par CLIFinder pour les comparer aux chimères détectées dans les gliomes. Les 2015 LCT-EST identifiés par Criscione correspondent à près de 1000 loci génomiques différents. Parmi ceux-ci, seuls 299 loci uniques impliquent des L1 des 27 sous-familles considérées dans notre étude. Enfin, seuls 228 loci correspondent à des L1 possédant une extrémité 5'UTR, compatible avec la présence d'une région ASP. Ces 228 loci impliquent alors 177 L1 récents (L1PA1 à 7) et 64 L1 anciens (Figure 43-A).

La comparaison des chimères identifiées dans nos gliomes avec les LCT de la littérature montre que 78 chimères correspondent à des LCT avérés, majoritairement décrits dans d'autres types tissulaires (normaux et tumoraux) que les gliomes. Ainsi, 24 chimères des gliomes correspondent à des LCT décrits à la fois dans la littérature et par Criscione, 7 sont communes entre l'analyse A37 et les données de la littérature, et 47 sont communes entre l'analyse A37 et celle de Criscione (Figure 43-B). Ces résultats montrent que les chimères identifiées par CLIFinder peuvent correspondre à des LCT déjà décrits dans d'autres types cellulaires. De plus, les L1 impliqués dans la formation de ces 78 LCT impliquent à 92% des L1 récents.

L'ensemble de ces résultats suggèrent donc que les chimères identifiées peuvent correspondre à des LCT.

2.3.3. Les chimères impliquant des L1 récents montrent une initiation de leur transcription dans la région de l'ASP.

Afin de valider si les chimères détectées avec CLIFinder peuvent correspondre à des LCT initiés à l'ASP de L1, un groupe a été sélectionné parmi les 2675 chimères qui semblent initiés au niveau d'un L1 possédant une extrémité 5'UTR. Les critères de sélection des chimères sont les suivants :

- Implication d'un L1 des sous-familles L1PA1 à 6 pour identifier des LCT, et des L1 des sous-familles L1PA7 et L1PA8 afin d'identifier si ces chimères peuvent correspondre à des LCT,
- Détection dans un des 3 groupes « C spé », « Tum spé », « C+T »,
- Expression récurrente dans au moins 2 échantillons,
- Position intragénique ou intergénique.

Cette sélection permet d'avoir une vue d'ensemble de tous les scénarios possibles.

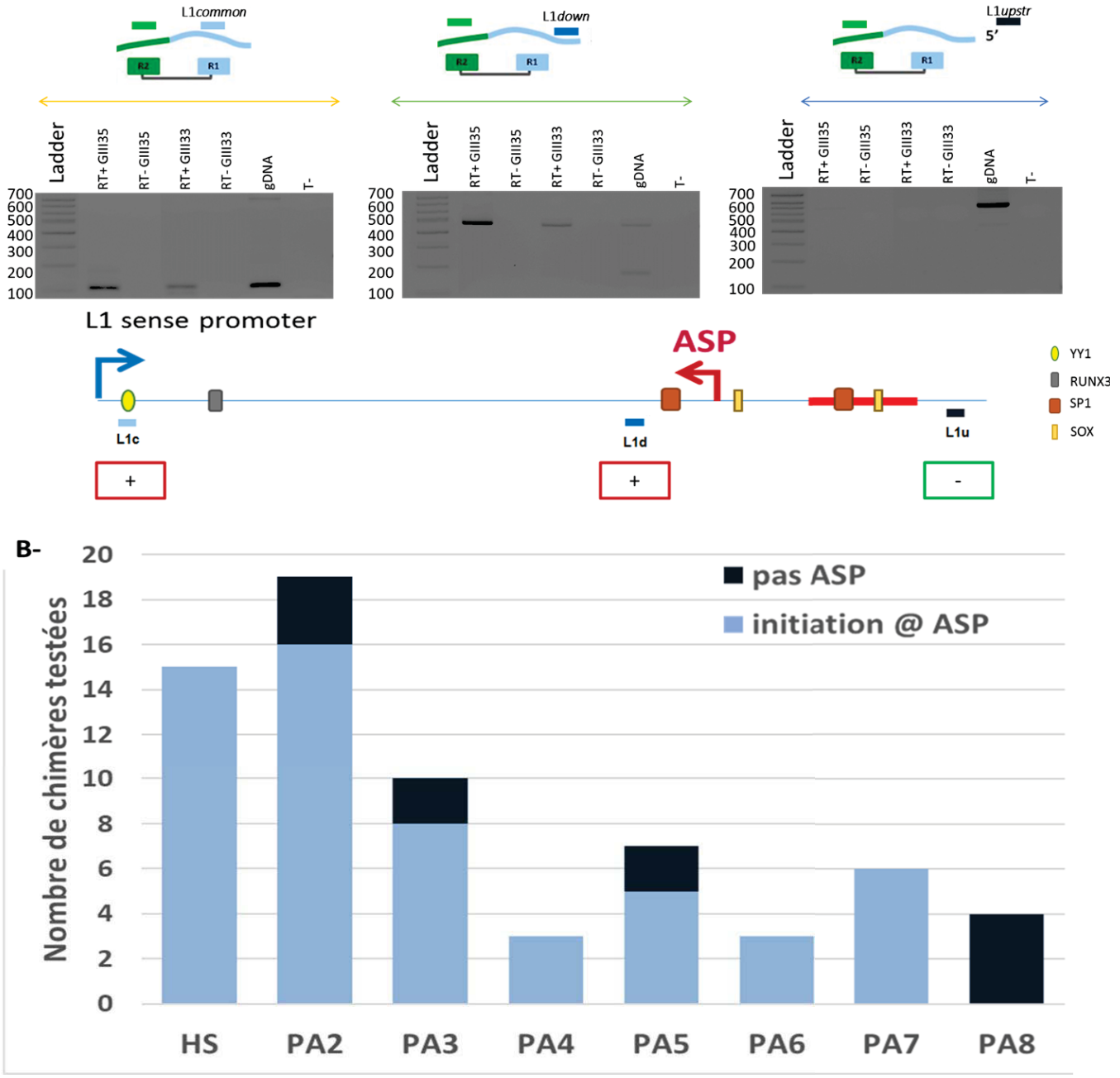


Figure 44 : La transcription de la majorité des chimères impliquant des L1 récents (PA1 à PA7) est initiée dans la région de l'ASP.

A- Résultats de marche en 5' pour le LCT 2705. Après amplification PCR avec les 3 couples d'amorces (SU-L1_common à gauche, SU-L1_downstream au milieu et SU-L1_upstream à droite) sur l'ADNg servant de témoin positif, avec l'H2O servant de témoin de contamination de la réaction de PCR, et sur les RT+ de 2 gliomes GIII-35 et GIII-33. Les RT- servent de témoins de contamination par l'ADNg. Pour le couple SU-L1_common, une amplification est observée à la taille attendue pour l'ADNg et pour les 2 RT+. Pour le couple SU-L1_downstream, une amplification est observée à la taille attendue pour l'ADNg et pour les 2 RT+. Pour le couple SU-L1_upstream, une amplification est observée à la taille attendue pour l'ADNg seulement. Aussi ce LCT2705 semble être initié à l'ASP. En dessous est représenté schématiquement l'extrémité 5'UTR du L1 avec le positionnement des amorces dans le L1 et l'amplification attendue pour un LCT, + en rouge pour le L1_common et L1_downstream et – en vert pour le L1_upstream.

B- Cette expérience de marche en 5'a été réalisée pour 63 chimères L1PA1 à 7 et 4 chimères L1PA8. Les résultats obtenus montrent que 56 chimères/63 impliquant les L1PA1 à 7 correspondant à des LCT, alors que 7 semblent être incluses dans un transcrit plus long. Par contre, pour toutes les chimères L1PA8 testées, toutes semblent être incluses dans un transcrit plus long incluant la séquence en amont de la région de l'ASP.

Ainsi 279 loci ont été sélectionnés afin de dessiner un couple d'amorces, avec une amorce localisée dans le R1 correspondant au L1, nommée L1_common, et une localisée dans le R2 correspondant à la séquence unique, nommée SU. À cause des contraintes dues au positionnement des amorces (taille et spécificité des amorces, taille de l'amplicon supérieure à 80 pb), 202 loci ont pu être testés. 84 chimères ont été validées par RT-PCR (sur les ADNc des échantillons dans lesquels les chimères ont été détectées par CLIFinder) et séquençage, dont 2 sont « Ctrl spé », 37 sont « Tum spé » et 45 sont « C+T ».

Suite à cette validation, la question se pose donc de savoir si ces chimères sont initiées à l'ASP.

L'approche expérimentale de choix pour répondre à cette question correspond à la 5'RACE. Toutefois, compte tenu de la difficulté de mise en œuvre pour étudier un grand nombre de transcrits, j'ai mis en place une approche de marche en 5' par RT-PCR. Celle-ci ne permet pas une définition précise du site d'initiation de la transcription des transcrits mais elle présente l'avantage de permettre l'analyse d'un grand nombre de transcrits.

Pour ce faire, en plus des amorces utilisées pour la validation, 2 autres amorces ont été dessinées dans la région 5'UTR du L1 : une localisée juste avant la position supposée de l'ASP nommée L1_downstream et une juste après, nommée L1_upstream. Ces amorces sont communes aux sous-familles L1PA1 à 6 et un dessin spécifique a été réalisé pour les sous-familles L1PA7 et L1PA8 de par la présence de 200 pb supplémentaires positionnant l'ASP aux positions +600 à +800 dans ces deux sous-familles (Annexe 1). Si la chimère correspond à un LCT initié à l'ASP de L1, les résultats attendus après amplification sont : une amplification positive pour le couple SU-L1_downstream et négative pour le couple SU-L1_upstream. Cette expérience a été réalisée sur les RT de 2 gliomes différents (Figure 44-A). L'exemple de la chimère 2705, montre une amplification avec l'ADNg pour les 3 couples d'amorces, ce qui signifie que les 3 conditions de PCR sont optimisées. Une amplification n'est obtenue avec les 2 RT+ que pour le couple d'amorces de validation et le couple SU-L1_downstream. L'absence d'amplification dans les RT+ de gliomes, pour le couple d'amorce SU-L1_upstream, suggère donc que le transcrit de la chimère 2705 est initiée dans la région génomique comprise entre les amorces L1_downstream et L1_upstream. Ceci indique que cette chimère correspond à un LCT dont la transcription est initiée dans la région de l'ASP (Figure 44-A). Cette expérience de marche en 5' a pu être réalisée pour 63 chimères L1PA1 à 7 et 4 chimères L1PA8. Au final, 56 chimères impliquant des L1 des sous-familles L1PA1 à 7 (soit 89%) montrent un profil d'amplification correspondant à des LCT initiés à l'ASP. Alors que pour les 4 chimères L1PA8 testées, celles-ci semblent toutes être incluses dans un transcrit plus long (Figure 44-B). Extrapolés à

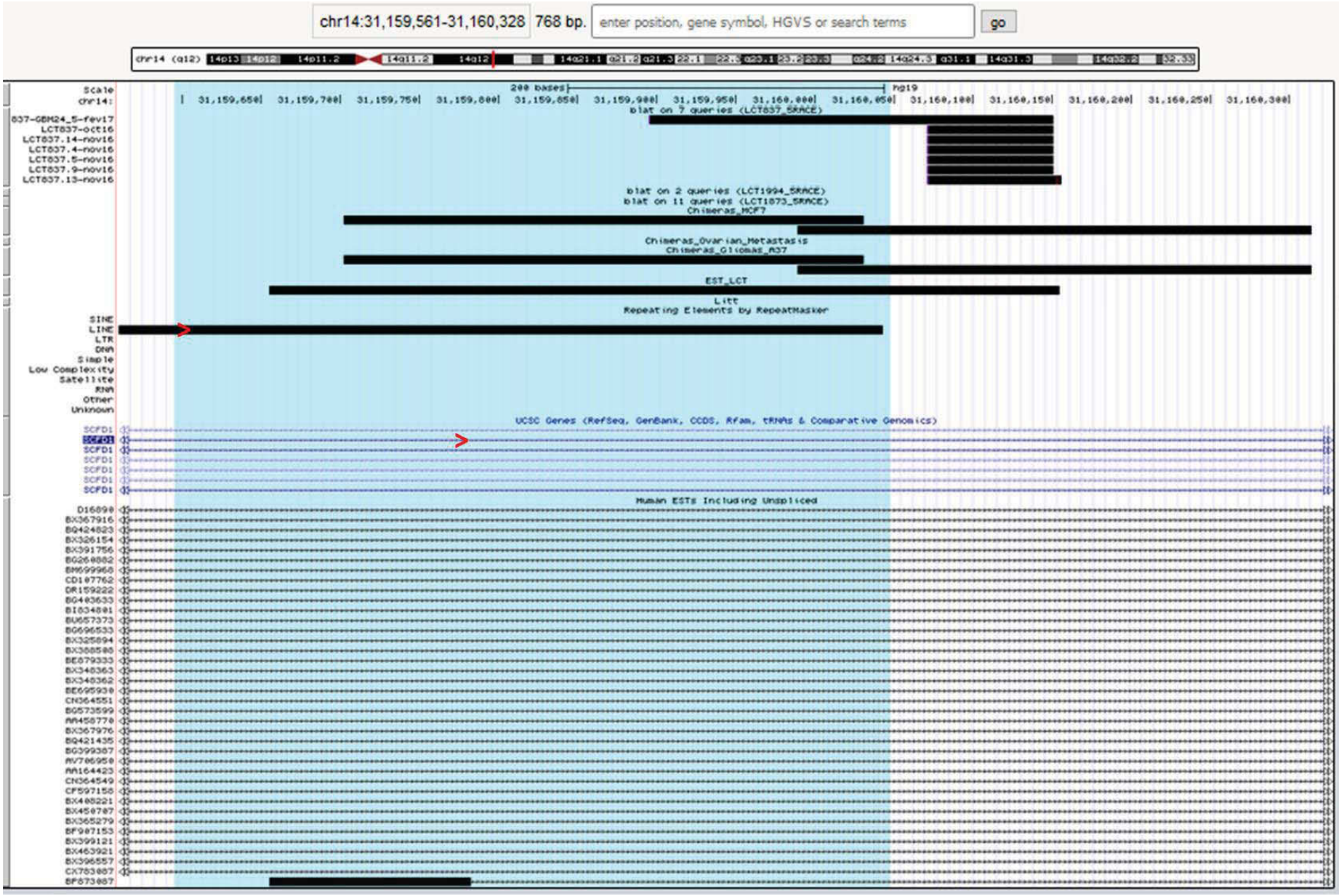


Figure 45 : Détermination du site d'initiation de la transcription du LCT837 au L1HS par 5'RACE.

Les séquences des 7 clones obtenus par 5'RACE sont localisées dans la séquence R1 et R2 détectées par CLIFinder dans les données de RNA-seq de MCF7 et des gliomes. Ce LCT a également été décrit par Criscione (EST_LCT). Ce LCT est localisé dans l'intron 14/24 du gène *SCFD1*, en sens. Le sens de transcription du gène et de l'ASP est indiqué par les flèches rouges. La région 5'UTR surlignée en bleu indique les 450 premières paires de bases du L1 dans lesquelles l'ASP est situé. Un clone remonte de +120 pb dans le L1 mais pas aussi loin que ce qui est détecté dans les RNA-seq.

l'ensemble des chimères détectées avec CLIFinder, ces résultats suggèrent que la majorité des 2256 chimères impliquant des L1 récents (L1PA1 à 7) pourraient correspondre à des LCT initiés à l'ASP.

Afin de valider précisément le site d'initiation de la transcription de ces LCT, des analyses de 5'RACE ont été réalisées pour 5 LCT : les LCT837 et LCT78 impliquant tous deux un L1HS, le LCT1873 impliquant un L1PA2, le LCT1994 impliquant un L1PA6 et le LCT2691 impliquant un L1PA7. Suite au dessin des amorces spécifiques pour l'amplification par 5'RACE (cf Annexe Matériels et Méthodes) et aux amplifications par PCR et séquençage, aucune séquence correspondante n'a été obtenue pour les clones testés pour les LCT2691 (n=3) et LCT78 (n=8).

Pour le LCT837 impliquant un L1HS localisé dans l'intron 14/24 du gène *SCFD1*, 8 clones /13 analysés correspondent à ce locus. 7 d'entre eux ne contiennent que 80 pb de la séquence unique et seulement 1 clone possède une séquence qui remonte dans le L1 aux environs de la position +120 pb. Comparé à la position 5' du R1 (en +350 du L1) qui a permis l'identification de la chimère dans nos RNA-seq, au résultat de marche en 5' qui a montré une amplification avec l'amorce L1_downstream située en +400 et à l'extrémité 5' du L1-EST décrit par Criscione en +400, ce résultat de 5'RACE ne semble pas identifier l'initiation de la transcription de ce LCT (Figure 45).

Pour le LCT1873 impliquant un L1PA2 localisé dans l'intron 14/15 du gène *GOLIM4*, 9 clones /9 correspondent à ce locus. Parmi eux, 1 ne contient que 89 pb, un autre 124 pb et 4 autres 142 pb correspondant à la séquence unique, mais ne remontent pas dans la séquence du L1. De plus, 1 transcrit épissé est identifié, celui-ci est transcrit à partir de l'exon 11 dans le sens du gène *GOLIM4* et se poursuit jusqu'à l'exon 14 (Annexe 3). Ce type de transcrit ne correspond pas à un LCT. Seulement 2 clones qui contiennent 363 pb de la séquence unique remontent de 51 pb dans l'extrémité 5'UTR du L1PA2. Comparé aux résultats de l'analyse A37, ce résultat de 5'RACE ne semble pas identifier l'initiation de la transcription de ce LCT, la position 5' du R1 étant aux environs de la position +400 (Annexe 2).

Enfin pour le LCT1994 impliquant un L1PA6 localisé dans l'intron 1 du gène *CCDC109B*, 4 clones/5 de 32 pb sont localisés au niveau de la séquence unique de la chimère détectée avec CLIFinder mais cette séquence correspond à celle de l'amorce séquence unique. En revanche, 1 clone de 216 pb correspond au locus et remonte de 82 pb dans le début du L1PA6. Comparé aux résultats de l'analyse A37, ce résultat de 5'RACE ne semble pas identifier l'initiation de la

transcription de ce LCT, la position 5' du R1 étant aux environs de la position +148 (Annexe 4).

Malgré l'utilisation du kit GeneRacer censé cibler les ARNm pleine taille coiffés pour effectuer notre étude de 5'RACE, les séquences amplifiées ne remontent jamais aussi loin dans le L1 que ce qui a été identifié par CLIFinder ou au niveau de l'extrémité EST-LCT. De plus, j'ai été confrontée à un problème récurrent de la présence de multiples adaptateurs liés en 5' ce qui surestimait la taille des inserts des clones qui ont été analysés. Ces résultats restent donc insatisfaisants et sont à compléter pour démontrer l'initiation à l'ASP.

L'ensemble des résultats obtenus par différentes analyses convergent donc vers la conclusion que la majorité des chimères impliquant des L1 récents correspondraient à des LCT dont la transcription serait initiée à l'ASP de L1. En effet, bien que j'ai démontré que l'enrichissement en chimères impliquant des L1 récents observé dans l'analyse A37 soit le résultat d'un biais lié aux conditions de paramétrage de CLIFinder (utilisation d'une séquence consensus pour chacune des 27 sous familles de L1 avec seulement 6 mésappariements tolérés), j'ai pu par ailleurs établir que :

- Une centaine de chimères correspondent à des LCT déjà décrits dans la littérature,

- Près de 90% des chimères impliquant des L1 récents (PA1 à PA7) ont une initiation de leur transcription dans la région de l'ASP,

- 100% des chimères analysées impliquant des L1PA8 sont incluses dans des transcrits plus longs incluant en 5' la séquence en amont de la région de l'ASP.

Les expériences de 5' RACE indispensables pour démontrer formellement une initiation à l'ASP restent à poursuivre. Compte tenu de la difficulté de mise en œuvre de cette approche, elles ne seront réalisées que pour quelques LCT candidats d'intérêt potentiel.

Au final, la majorité des 2256 chimères impliquant des L1 récents identifiées dans nos gliomes *via* l'analyse A37, semble donc correspondre à des LCT.

2.4. La dérégulation transcriptionnelle liée aux LCT dans les gliomes passe par une surexpression en contexte tumoral.

Les résultats du nombre de lectures détectées par CLIFinder pour chaque chimère suggèrent que la moitié d'entre elles (52%) serait « Tum spé » avec 50% détectées dans au moins 2 échantillons tumoraux parmi 13, mais dans aucun des 3 contrôles.

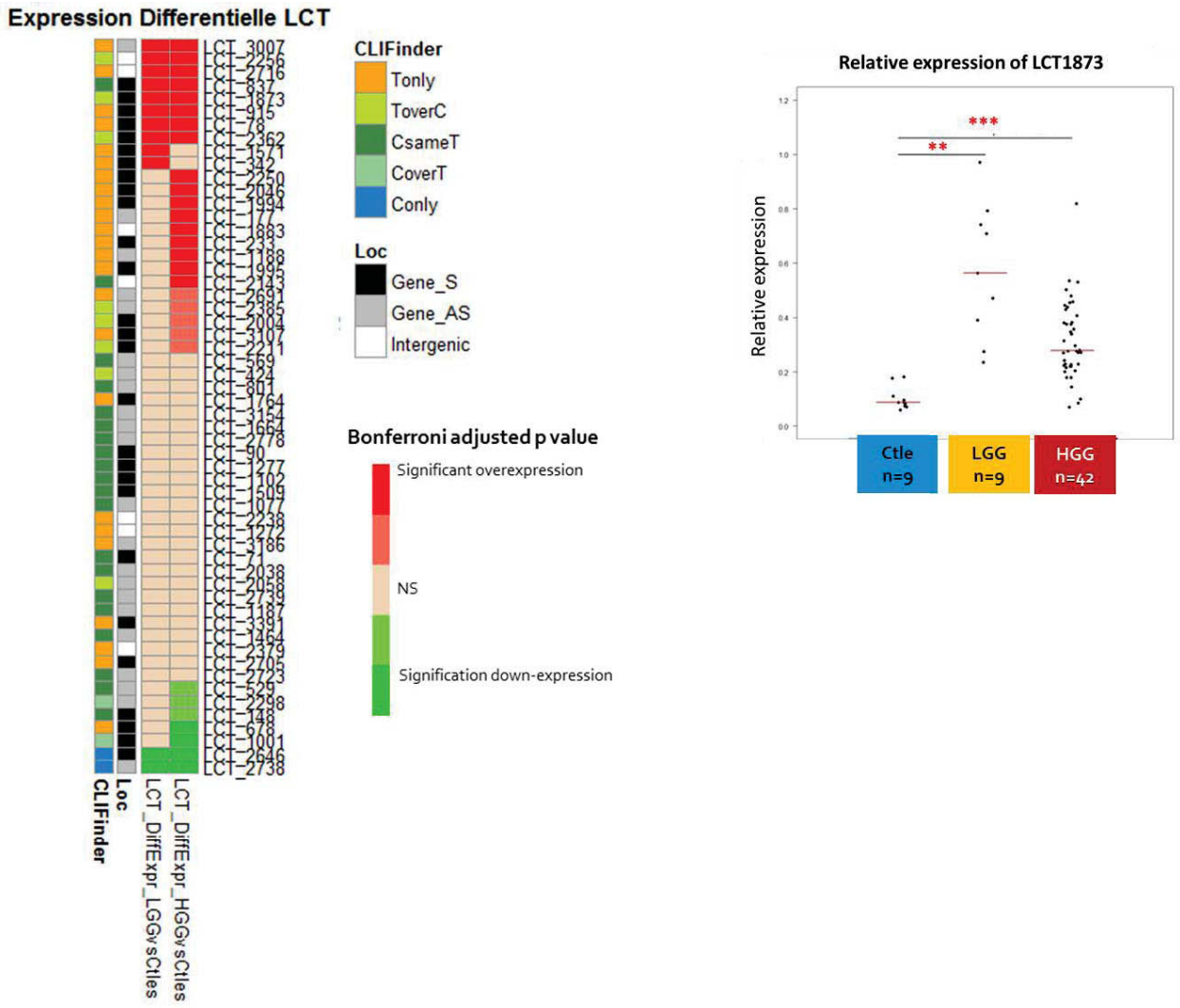


Figure 46 : L'analyse de l'expression des LCT validés par RT-qPCR sur une cohorte plus large montre une expression de tous les LCT dans les tissus contrôles et une expression différentielle en contexte tumoral pour certains d'entre eux.

A. L'expression des 56 LCT validés a été étudiée par RT-qPCR sur une cohorte plus large d'échantillons incluant 10 tissus contrôles, 9 LGG et 42 HGG. La heatmap représente la significativité des p values (test Mann Whitney) corrigées avec le coefficient de Bonferroni pour l'expression différentielle des LCT entre LGG (LCT_DiffExpr_LGGvsCtles) ou HGG (LCT_DiffExpr_HGGvsCtles) *versus* contrôles. Les valeurs de p-value représentées en rouge correspondent à une sur-expression significative, en rose à une tendance à la sur-expression proche de la significativité, en vert à une sous-expression significative, en vert clair à une tendance à la sous-expression proche de la significativité. Les p values représentées en beigne sont non significatives. Les deux colonnes à gauche reportent pour la colonne annotée CLIFinder la classification des chimères selon les résultats de l'analyse A37, à savoir si la chimère est « Ctrl spé », « Tum spé », ou « C+T » avec les sous-classifications « CoverT » si FC<0.5, « CoverT » si FC>2, « CsameT » si 0.5<FC<2. La colonne Loc indique si la chimère est en position intergénique ou intragénique en sens ou en antisens par rapport au gène dans lequel elle se trouve.

B. Exemple de résultat d'analyse d'expression différentielle pour le LCT1873. Le niveau d'expression relative mesuré pour chaque échantillon est reporté par groupe avec la médiane présentée en rouge. L'analyse statistique de Mann Whitney montre une surexpression significative de ce LCT dans les deux groupes tumoraux. Ce résultat confirme la classification « ToverC » établie à partir des nombres de lecture rapportés par CLIFinder pour ce LCT.

Sachant que les RNA-seq n'ont été réalisés que sur 16 échantillons (3 contrôles et 13 gliomes), une analyse de RT-qPCR, méthode de référence pour étudier l'expression des transcrits, a été entreprise sur une cohorte plus large incluant 10 contrôles et 51 gliomes (9 LGG et 42 HGG).

Suite à la validation des efficacités (comprises entre 1,8-2) des couples amorces L1_common-SU en qPCR (Lightcycler), l'expression de 56 LCT confirmés par marche en 5' a été mesurée sur la cohorte plus large grâce à une puce 96.96 Dynamic Array (Fluidigm). Celle-ci permet un multiplexage de 96 RT avec 96 couples d'amorces soit au total 9216 points de quantification obtenus en une réaction.

L'expression relative de chaque LCT a été mesurée pour chaque échantillon en prenant en compte l'efficacité des amorces et en normalisant vis-à-vis de l'expression de 3 gènes de ménage validés pour l'étude de l'expression génique dans les gliomes (Kreth *et al.*, 2010).

De façon inattendue, les résultats montrent que tous les LCT, dont ceux détectés uniquement dans les tumeurs par CLIFinder, sont retrouvés exprimés dans les échantillons contrôles. Cependant, la comparaison des niveaux d'expression relative des 56 LCT analysés dans les 10 tissus contrôles en fonction du groupe identifié par CLIFinder (« Ctrl spé » n= 2 ; « C+T » n= 30 ; « Tum spé » n=24) montre que l'expression dans les tissus contrôles des LCT « Tum spé » est significativement plus faible que celles des LCT détectés par CLIFinder dans les tissus contrôles (Figure 46). Cette observation, combinée au fait que les RNA-seq de 3 tissus contrôles seulement ont été analysés par CLIFinder explique pourquoi des LCT ont pu être détectés par CLIFinder uniquement dans les tumeurs alors qu'ils sont aussi exprimés dans les contrôles.

Aussi l'expression détectée dans les échantillons contrôles suggère qu'il existe une transcription basale à partir de l'ASP indépendamment de la méthylation ADN de la région promotrice.

L'expression différentielle des 56 LCT entre les 2 groupes tumoraux (LGG et HGG) et le groupe contrôle a ensuite été étudiée. Celle-ci montre une surexpression ou une sous-expression significative dans au moins 1 groupe tumoral par rapport aux contrôles pour respectivement 24 et 7 LCT sur les 56 étudiés (Figure 4). La confrontation des résultats d'expression des 56 LCT en RT-qPCR sur la cohorte plus large avec la classification établie à partir du nombre de lectures rapporté par CLIFinder montre :

- Pour les 24 chimères classées « Tum spé » que 16 (66%) sont significativement surexprimées. De plus, parmi les chimères « Tum spé », 7 semblaient HGG spécifique (0 lectures dans le 5 LGG). Parmi elles, 4 sont retrouvées surexprimées spécifiquement dans les HGG (LCT233, LCT1188, LCT1994 et LCT2691), 1 est retrouvée surexprimée

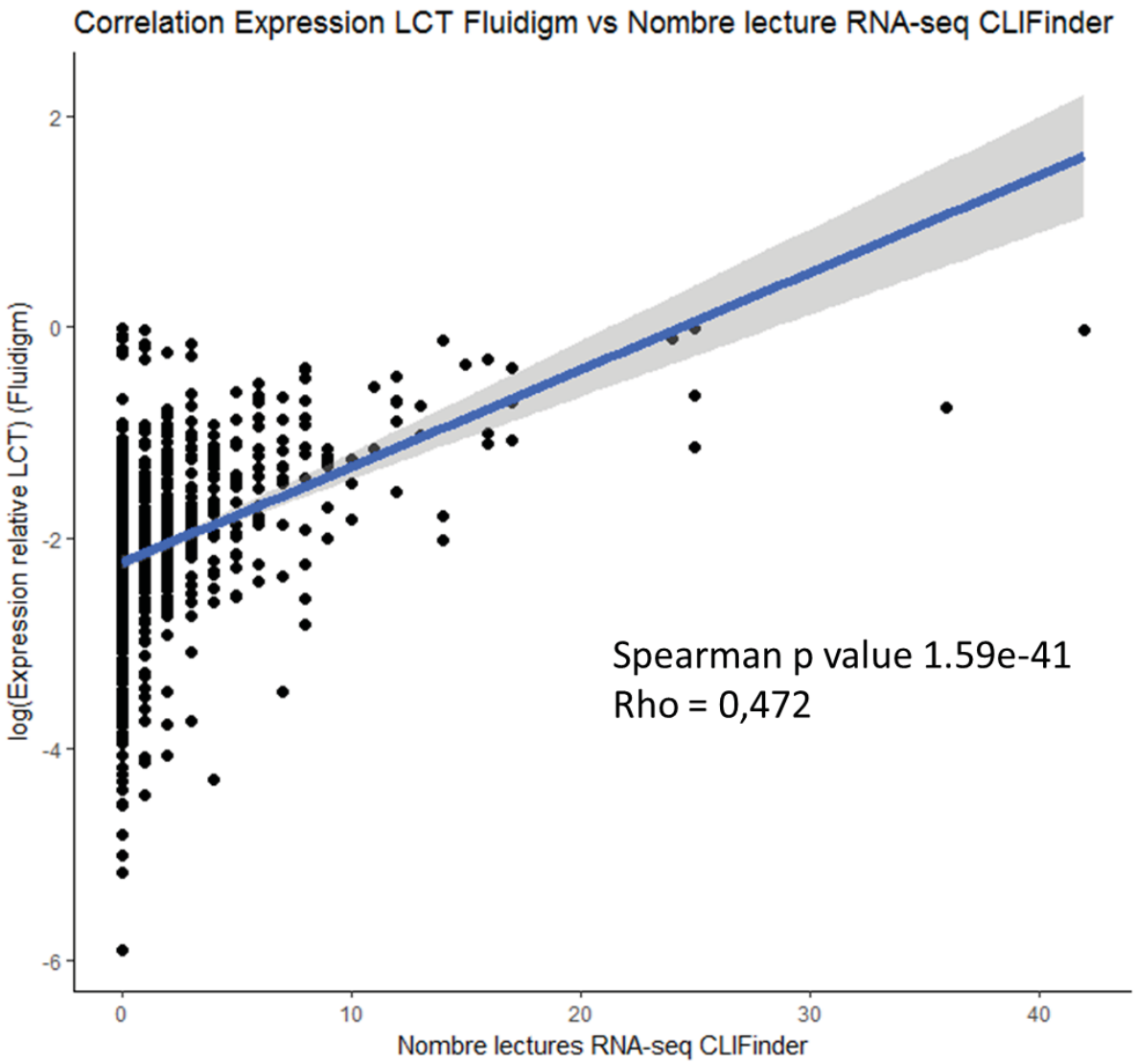


Figure 47 : Corrélation entre le nombre de lectures en RNA-seq (A37) et l'expression en RT-qPCT de 56 LCT.

Le nombre de lecture de RNA-seq issu de l'analyse A37, en abscisse, corrèle avec l'expression relative des 56 LCT validés en échelle logarithmique, selon la corrélation de Spearman. La valeur Rho est de 0,472. Cette observation démontre que les résultats de RNA-seq analysés avec CLIFinder sont au moins semi-quantitatif.

dans les HGG et les LGG (LCT915) et 2 ne sont pas différentiellement exprimées (LCT2705 et LCT3186).

- Pour les 2 chimères classées « Ctrl spé » les 2 (100%) sont significativement sous-exprimées dans les 2 groupes tumoraux.

Pour les 30 LCT classés « C+T », une sous-classification a été rajoutée à partir des données de CLIFinder. En effet, pour certaines chimères détectées dans les contrôles et les tumeurs, une différence importante du nombre de lectures était notée entre les groupes contrôles, HGG et LGG. La sous-classification s'est basée sur le FC (Fold-Change) entre le nombre moyen de lecture par groupe (Contrôle, HGG, LGG). Ainsi, un FC dans au moins 1 groupe tumoral par rapport aux contrôles inférieur à 0,5 a été annoté « CoverT » (n=2), un FC supérieur à 2 a été annoté « ToverC » (n=8) et un FC compris entre 0,5 et 2 a été annoté « CsameT » (n=20). La comparaison des résultats d'expression en RT-qPCR pour 30 LCT « C+T » montre alors :

- Pour les LCT « CoverT » : les 2 (100%) sont significativement sous-exprimés dans les HGG par rapport aux contrôles (LCT1001 et LCT2298),
- Pour les 8 LCT « ToverC » : 6 (75%) sont significativement surexprimés dans les 2 groupes tumoraux (LCT2256, LCT1873, LCT2362) ou seulement dans les HGG (LCT2385, LCT2004, LCT2211), conformément à ce qui était prédit par les données de CLIFinder,
- Enfin, pour les 20 LCT « CsameT » : 2 sont significativement surexprimés (LCT837 et LCT2143), 2 sont significativement sous-exprimés (LCT529 et LCT148) et 16 (80%) présentent comme attendu une expression similaire entre les tumeurs et les contrôles.

Au final, une surexpression significative est retrouvée dans la cohorte plus large pour 70% (22/32) des LCT supposés, d'après les données de CLIFinder, être « Tum spé » ou « ToverC ». Une sous-expression significative est quant à elle retrouvée pour 100% (4/4) des LCT classifiés « Ctrl spé » ou « CoverT ».

L'ensemble de ces résultats suggère donc :

- Qu'il existe une transcription basale de LCT à partir de l'ASP dans les tissus normaux,
- Que la dérégulation transcriptionnelle liée aux LCT, ne passe non pas par une induction d'expression en contexte tumoral mais par une surexpression.

Par ailleurs, la confrontation des résultats de quantification obtenus pour l'ensemble des 56 LCT par RT-qPCR avec le nombre de lectures de RNA-seq issu de l'analyse A37 a mis en évidence une corrélation positive significative entre les deux types de données (Figure 47). Cette observation démontre que le nombre de lectures de RNA-seq sont au moins semi-

quantitatives et devrait permettre d'appréhender de façon pangénomique le pourcentage de LCT surexprimés en contexte tumoral.

3. Conclusion partie 2.

Grâce au logiciel CLIFinder, 2675 chimères uniques ont été identifiées. Parmi elles, 2256 impliquent des L1 récents (L1PA1 à 7) pleines tailles supposées posséder un ASP fonctionnel et indiquent que 25% des L1 récents du génome humain seraient engagés dans la transcription de LCT.

L'identification de 78 chimères correspondant à des LCT déjà décrits et la démonstration que 89% des chimères impliquant des L1 récents ont une initiation de leur transcription dans la région de l'ASP conforte l'idée que la majorité des chimères impliquant un L1 récent identifiées par CLIFinder puisse correspondre à des LCT.

Malgré le fait que 50% des chimères aient été détectées uniquement dans des échantillons tumoraux par CLIFinder, les études d'expression quantitative sur une cohorte plus large m'ont permis d'établir que la dérégulation transcriptionnelle liée aux LCT dans les gliomes passe par une surexpression en contexte tumoral. En effet, les résultats obtenus montrent :

- Une transcription basale à partir de l'ASP pour tous les LCT testés dans les tissus normaux,
- Les LCT sont tous exprimés dans tous les tissus (contrôles et tumoraux) quand bien même ils n'ont été détectés que dans certains d'entre eux par CLIFinder,
- La détection « Tum spec » ou « ToverC » par CLIFinder implique dans 70% des cas une surexpression tumorale significative par rapport aux tissus contrôles. La surexpression peut alors être spécifique d'un sous-groupe tumoral.

Cette dernière observation suggère que certains LCT pourraient jouer un rôle fonctionnel dans l'initiation, la progression et/ou l'agressivité tumorale.

Partie 3 : Bases mécanistiques de la surexpression des LCT en contexte tumoral.

1. Objet de l'étude Partie 3.

Grâce au logiciel CLIFinder, 2256 chimères impliquant des L1 récents possédant un ASP ont été détectées dans les données de RNA-seq de gliomes. L'analyse de l'expression de 56 LCT validés sur la cohorte plus large par RT-qPCR, montre que la dérégulation transcriptionnelle liée aux LCT dans les gliomes implique une surexpression dans les tumeurs par rapport aux contrôles (partie 2). La question se pose donc de savoir quel est (sont) le(s) mécanisme(s) moléculaire(s) sous-jacent(s) pouvant expliquer cette surexpression.

Différentes hypothèses émergent des données de la littérature. Ainsi, pour le LCT *L1-MET* une corrélation inverse a été observée entre le pourcentage de méthylation du promoteur du L1 concerné et l'expression du LCT *L1-MET* (Figure 32). En effet, dans les lignées cellulaires contrôles, le promoteur du L1 fortement méthylé est associé à une faible expression du transcrit *L1-MET* alors que dans les lignées cellulaires tumorales de vessie, la diminution de la méthylation du promoteur du L1 est associée à une augmentation de l'expression du transcrit *L1-MET* (Wolff *et al.*, 2010). De cette observation découle une première hypothèse qui propose que la diminution de la méthylation des régions promotrices de L1 dans les tumeurs pourrait être à l'origine de la surexpression d'un LCT. Celle-ci implique alors que la méthylation du promoteur des L1, décrite pour être impliquée dans la répression de l'activité du promoteur sens, régulerait également l'activité de l'ASP.

Par ailleurs, un autre mécanisme a été proposé comme pouvant être à l'origine de l'expression d'éléments L1 dans le génome. En effet, plusieurs études ont montré que l'activation de la transcription de L1 dans les cellules somatiques serait influencée par l'activité transcriptionnelle du locus et ce, de façon type cellulaire spécifique. Ceci a été mis en évidence initialement pour quelques éléments L1 (Lavie *et al.*, 2004) et confirmé récemment plus largement pour les éléments de la sous-famille des L1HS (Philippe *et al.*, 2016). Ces observations suggèrent donc une deuxième hypothèse qui propose qu'une configuration permissive de la chromatine à un locus peut permettre l'activation du promoteur d'éléments L1. Dans ce contexte, on peut se demander si cette activation du promoteur sens pourrait concerner également l'ASP des L1 possédant un promoteur bidirectionnel et ainsi expliquer la surexpression des LCT que nous avons observée en contexte tumoral.

Afin de valider ces hypothèses, 2 approches expérimentales ont été mises en place et seront détaillées dans cette partie.

2. Résultats partie 3.

2.1. Surexpression et hypométhylation du promoteur de L1.

La première hypothèse est que l'hypométhylation globale observée dans les tumeurs, qui affecte notamment les promoteurs de L1, serait responsable de l'augmentation de la transcription au niveau de l'ASP ; induisant la surexpression de LCT comme décrit dans la littérature pour le transcrit *L1-MET* (Wolff *et al.*, 2010). Des puces de méthylation HM450K réalisées sur nos gliomes montrent une diminution globale de la méthylation ADN significative dans les HGG par rapport aux contrôles avec un taux méthylation médian de 55% dans les tumeurs HGG versus 60% dans les tissus contrôles (Figure 60-B, Matériels et Méthodes). Une diminution de la méthylation significative de même ordre est également observée lorsque les sondes localisées dans les L1 sont spécifiquement analysées. Aussi, nos gliomes se comportent à l'image de ce qui a été décrit dans la littérature, à savoir, qu'une hypométhylation globale est observée dans les tumeurs agressives affectant notamment les éléments L1.

Dans le but de quantifier le taux de méthylation des promoteurs de L1 spécifiquement impliqués dans les 24 LCT validés et surexprimés, une approche expérimentale permettant la mesure de la méthylation ADN par qPCR a été adaptée à notre étude. Basée sur la digestion de l'ADN en utilisant soit des enzymes de restriction sensibles à la méthylation (ici, HpaII et HhaI qui coupent uniquement quand leur site de reconnaissance n'est pas méthylé) soit une enzyme dépendante de la méthylation (McrBC qui ne coupe que si son site de reconnaissance est méthylé), cette technique nommée qAMP, permet l'analyse combinée de plusieurs sites CpG d'une région génomique qui sera ensuite analysée par qPCR (Oakes *et al.*, 2006). Après digestion de l'ADNg par chacune de ces enzymes indépendamment, une amplification locus spécifique est réalisée par qPCR sur les différents produits de digestion ainsi que sur un ADN non digéré. Le taux de méthylation pour un locus dans un échantillon correspond alors à la moyenne des taux obtenus pour chacune des 3 enzymes utilisées.

Initialement, les séquences L1 des 56 LCT validés ont été analysées pour définir pour chacune une amorce dans le promoteur de LINE permettant d'amplifier une région contenant au moins 1 site de restriction pour chaque enzyme et qui, combinée avec l'amorce séquence unique correspondante, aboutit à un produit d'amplification de taille inférieurs à 500 pb compatible avec une quantification par qPCR (Annexe 5). Les couples d'amorce séquence unique – amorce L1 ont ensuite été testés sur une gamme d'ADNg et les conditions de qPCR modifiées afin d'optimiser leur efficacité d'amplification. 32 couples ayant une efficacité > à 1,8 ont été

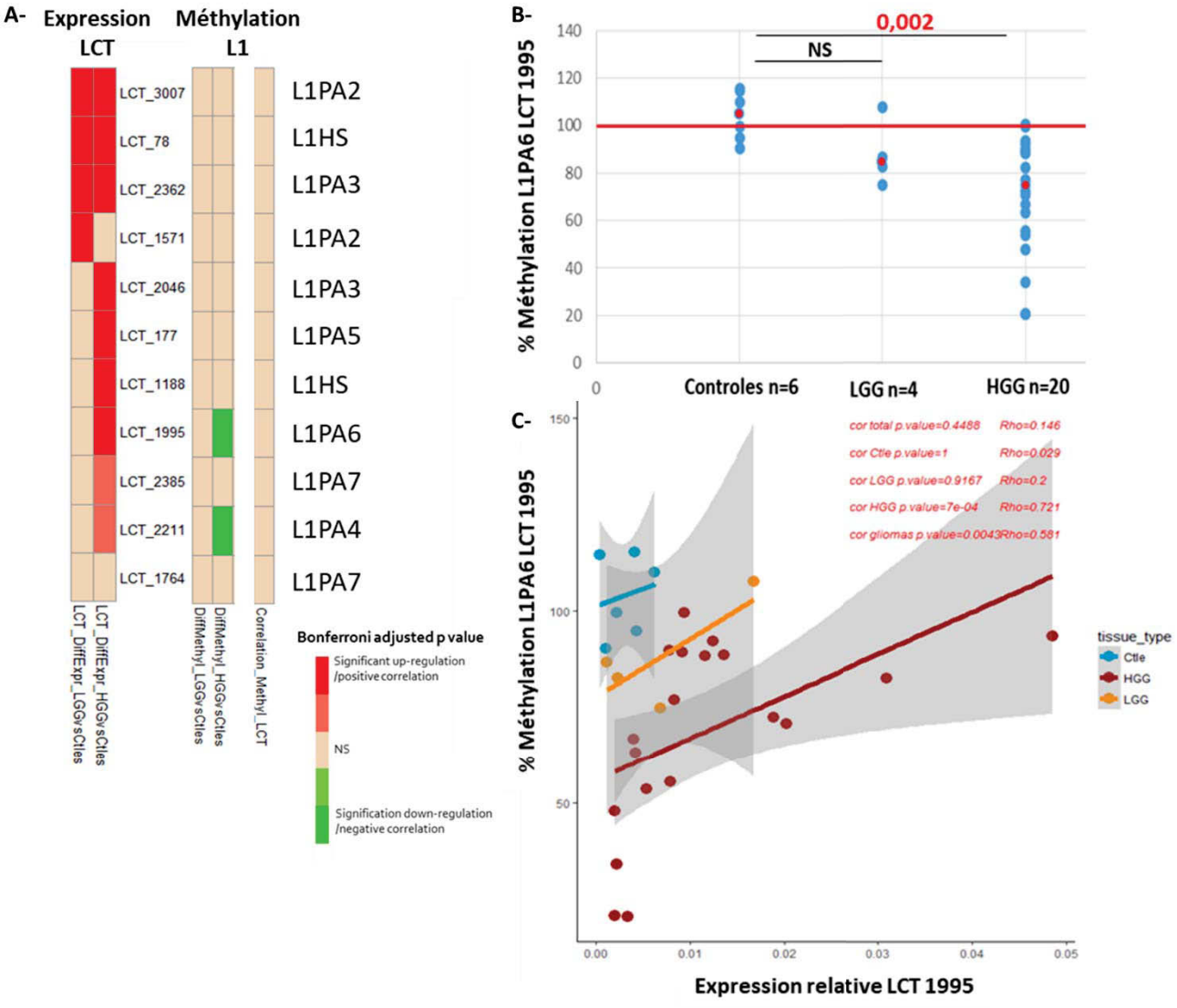


Figure 48 : L'hypométhylation des promoteurs de L1 n'est pas impliquée dans la surexpression des LCT.

A- La méthylation ADN des promoteurs de L1 associés à un LCT a été déterminée par qAMP pour 10 LCT significativement surexprimés dans au moins 1 groupe tumoral (LCT_DiffExpr) et 1 LCT dont l'expression ne varie pas dans les tumeurs (LCT_1764). 6 tissus contrôles, 4 LGG et 21 HGG ont été étudiés. Les résultats de méthylation différentielle (test Mann Whitney) dans les LGG (DiFFMethyl_LGGvsCtles) et les HGG (DiffMethyl_HGGvsCtles) sont représentés sous la forme d'une heatmap rapportant la significativité des p-values corrigées avec le coefficient de Bonferroni (rouge augmentation significative, rose tendance proche de la significativité, beige non significatif, vert clair tendance à la significativité, vert diminution significative) Les résultats montrent une diminution significative de la méthylation dans les HGG par rapport aux contrôles pour les LCT1995 et LCT2211, corrrespondant au groupe tumoral où la surexpression du LCT est observée. Les données d'expression des LCT et de taux de méthylation de L1 dans les échantillons ont ensuite été mises en corrélation. Les résultats de la significativité des p-values des corrélation de Spearman sont rapportées sur la heatmap Correlation_Methyl_LCT selon le code couleur décrit précédemment. Aucune corrélation n'est retrouvée.

B- Méthylation différentielle de la région promotrice du L1PA6 impliqué dans la formation du LCT1995 surexprimé dans les HGG. Une diminution significative de la méthylation est observée dans les HGG.

C- Mise en relation des données de pourcentage de méthylation au niveau du L1 et l'expression relative du LCT1995. Les résultats montrent une relation positive significative (Rho=0,721, p_value= 7^{e}-04) restreinte aux HGG. Toutefois, celle-ci ne répond pas à l'hypothèse testée qui implique l'identification d'une corrélation négative, l'augmentation de l'expression étant supposée être due à une diminution de la méthylation du promoteur de L1.

validés. Pour autant, les quantités d'ADN disponibles étant limitées, l'analyse n'a pu être réalisée que pour 11 loci incluant 10 LCT surexprimés pour lesquels les conditions de qAMP étaient validées et 1 LCT ne présentant pas de différence significative d'expression entre les contrôles et les tumeurs. Cette technique a été appliquée à une cohorte de 32 échantillons, incluant 7 tissus cérébraux sains et 25 gliomes (4 LGG et 21 HGG), plus 4 contrôles artificiels de méthylation (0, 30, 70, 100% de méthylation).

Dans un premier temps, une analyse de la méthylation différentielle a été effectuée, comparant les niveaux de méthylation obtenus entre les 3 groupes d'échantillons. Celle-ci montre que les L1 de 2 LCT présentent une diminution significative de leur méthylation et ce, spécifiquement dans le groupe des tumeurs sévères HGG (-28% LCT2211, p=0,001 et -30% LCT1995 ; p=0,002) (Figure 48-A et 48-B, Annexe 6). Cette observation concorde alors avec le fait que les LCT correspondants sont eux-mêmes surexprimées spécifiquement dans ce groupe tumoral.

Pour autant, est-ce que cette diminution de méthylation est retrouvée de façon concomitante avec la surexpression des LCT dans les tumeurs ? Afin de répondre à cette question, une analyse de corrélation a été réalisée confrontant, pour l'ensemble de la cohorte, l'expression relative de chaque LCT avec le taux de méthylation du L1 correspondant. Celle-ci montre une absence de corrélation pour tous les loci testés (Figure 48-A et 48-C, Annexes 7 et 8), y compris pour les 2 LCT présentant une diminution de méthylation dans les HGG. Une analyse complémentaire étudiant la corrélation par groupe montre pour le LCT 1995 une corrélation positive significative (Rho = 0,721 ; p=0,0007) dans les HGG (Figure 48-C). Toutefois, celle-ci ne répond pas l'hypothèse testée puisque les taux de méthylation les plus faibles sont retrouvées dans les tumeurs exprimant le moins LCT1995 et inversement.

Ceci signifie qu'un mécanisme, autre que l'hypométhylation du promoteur de L1, doit être impliqué dans la surexpression significative des LCT dans les tumeurs.

2.2. Surexpression et régulation transcriptionnelle au locus.

Une seconde hypothèse se dégage de mes résultats à savoir que la surexpression des LCT dans les tumeurs pourrait être due à une dérégulation transcriptionnelle au locus. En effet, parmi les 24 LCT que j'ai identifiés comme étant surexprimés en contexte tumoral, 20 sont localisés en position intragénique. A partir des données de RNA-seq, une analyse de l'expression génique a également été réalisée et nous permet de disposer de l'expression des gènes (en fpkm). Pour les 20 LCT intragéniques, j'ai donc pu confronter les résultats de l'expression du LCT à celle du gène associé dans les 15 échantillons analysés à la fois en RNA-seq et en Fluidigm. Malgré

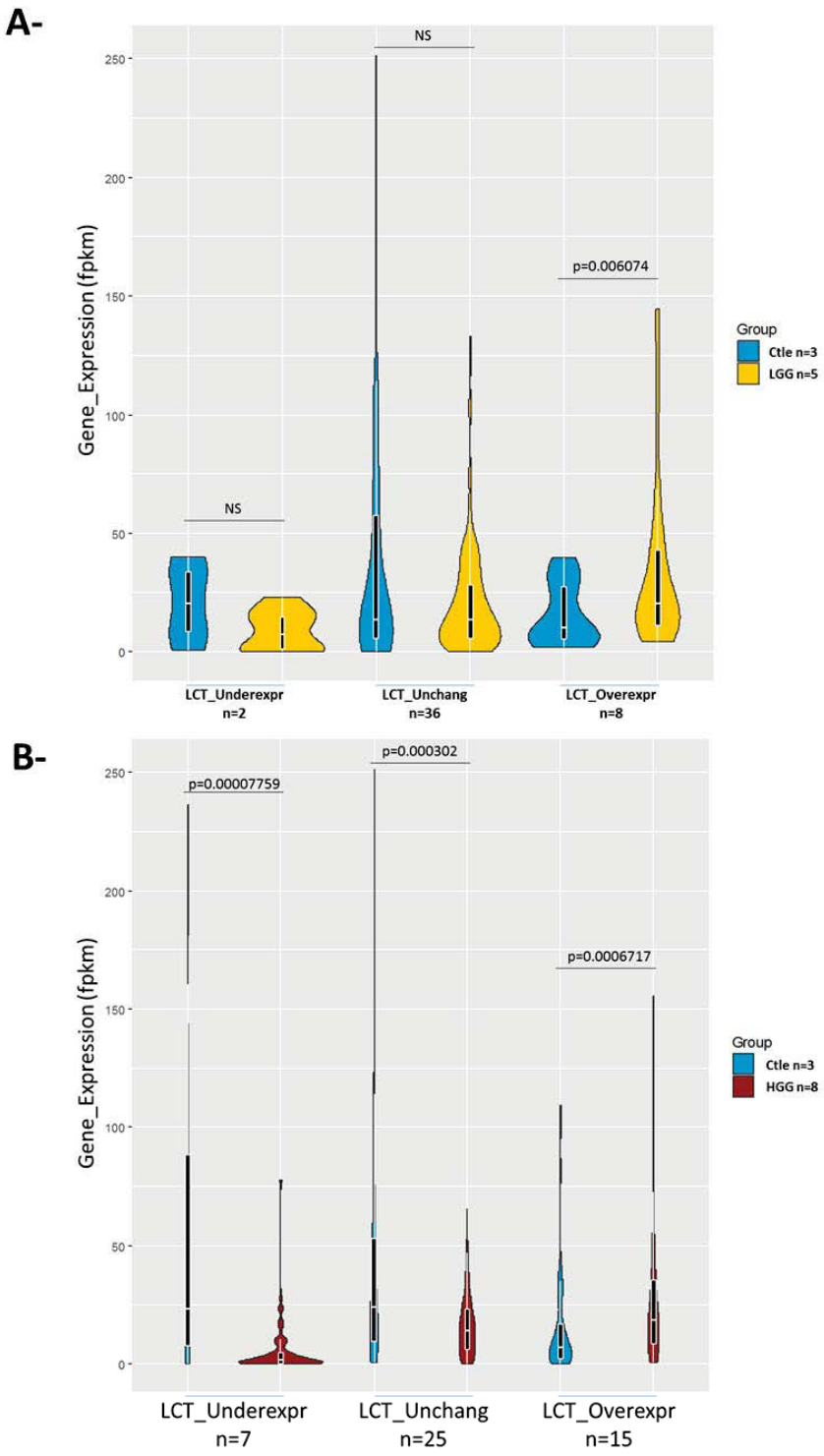


Figure 49 : Les gènes associés aux LCT montrent des variations d'expression similaires à celles des LCT.

A partir des données des 16 RNA-seq (3 contrôles, 5 LGG et 8 HGG), l'expression (fpkm) de 46 gènes associés à des LCT validés a été récupérée. Celle-ci est ensuite comparée entre tissus contrôles et LGG (A) ou HGG (B). Dans chaque cas, les gènes ont été répartis en 3 catégories selon la variation d'expression identifiée pour le LCT qui leur est associé, à savoir : LCT_Underexpr (LCT significativement sous exprimé dans les tumeurs), LCT_Unchang (LCT sans variation d'expression dans les tumeurs) et LCT_overexpr (LCT significativement surexprimé dans les tumeurs). La médiane du niveau d'expression ainsi que les limites des 25ème et 75ème percentiles sont rapportées par les box plots. Les diagrammes en violon représentent la densité de distribution des valeurs. L'existence d'une différence entre le groupe contrôle et le groupe tumoral est évaluée par un test de Mann-Whitney. Les gènes associés à des LCT sur- ou sous-exprimés montrent respectivement une diminution ou une augmentation d'expression tumorale significative qui va donc dans le même sens que celle des LCT.

le petit effectif d'échantillons, ceci a mis en évidence une corrélation positive significative dans 12 cas / 20 et ce, que la transcription du LCT se fasse dans le même sens que celle du gène (n=8) ou en antisens (n=4). L'observation que l'augmentation de l'expression des LCT dans les tumeurs se produit de façon concomitante à l'augmentation de l'expression du gène dans lequel il se trouve, suggère donc que l'activité transcriptionnelle du locus pourrait influer sur l'activité de l'ASP des L1.

L'existence d'une telle relation a été décrite dans la littérature pour l'expression en sens des L1 (Philippe *et al.*, 2016). Les auteurs décrivent que les éléments L1 exprimés dans une cellule somatique (normale ou tumorale) sont préférentiellement localisés dans un contexte génomique transcriptionnellement actif. Ainsi, l'activation du promoteur sens d'un L1 serait dépendante de l'activité transcriptionnelle de la région génomique avoisinante.

Est-ce qu'un mécanisme similaire peut avoir lieu à l'ASP, et être responsable de la surexpression tumorale de certains LCT ?

Afin de répondre à cette question, j'ai analysé dans un premier temps l'expression des gènes associés aux 49 LCT intragéniques que j'ai validés (Figure 46). Les données d'expression de 46 gènes (en fpkm) ont été récupérées des analyses transcriptomiques des 16 RNA-seq (2 LCT différents sont présents dans le même gène et 2 gènes référencés dans UCSC ne sont pas renseignés en termes d'expression dans l'analyse transcriptomique). Deux analyses comparatives ont été menées confrontant l'expression des gènes dans les échantillons contrôles (n=3) soit à celle dans les LGG (n=5) (Figure 49-A) soit à celle dans les HGG (n=8) (Figure 49-B). Dans chaque cas, les 46 gènes ont été répartis en 3 catégories selon l'expression différentielle du LCT associé, dans le groupe tumoral considéré, à savoir :

1) Des gènes associés à des LCT présentant une diminution de l'expression dans les tumeurs par rapport aux contrôles,
2) Des gènes associés à des LCT ne présentant pas de variation de l'expression entre les contrôles et les tumeurs,
3) Des gènes associés à des LCT présentant une surexpression dans les tumeurs par rapport aux contrôles.

Les représentations graphiques obtenues montrent que globalement l'expression des gènes suit le même sens de variation que celui noté pour les LCT auxquels ils sont associés. En effet, pour les groupes de gènes associés à des LCT sous-exprimés dans les LGG (n=2) et les HGG (n=7) par rapport aux contrôles, une expression plus faible de ces gènes dans les tumeurs par rapport aux contrôles est observée (tendance pour les LGG et diminution significative pour les HGG).

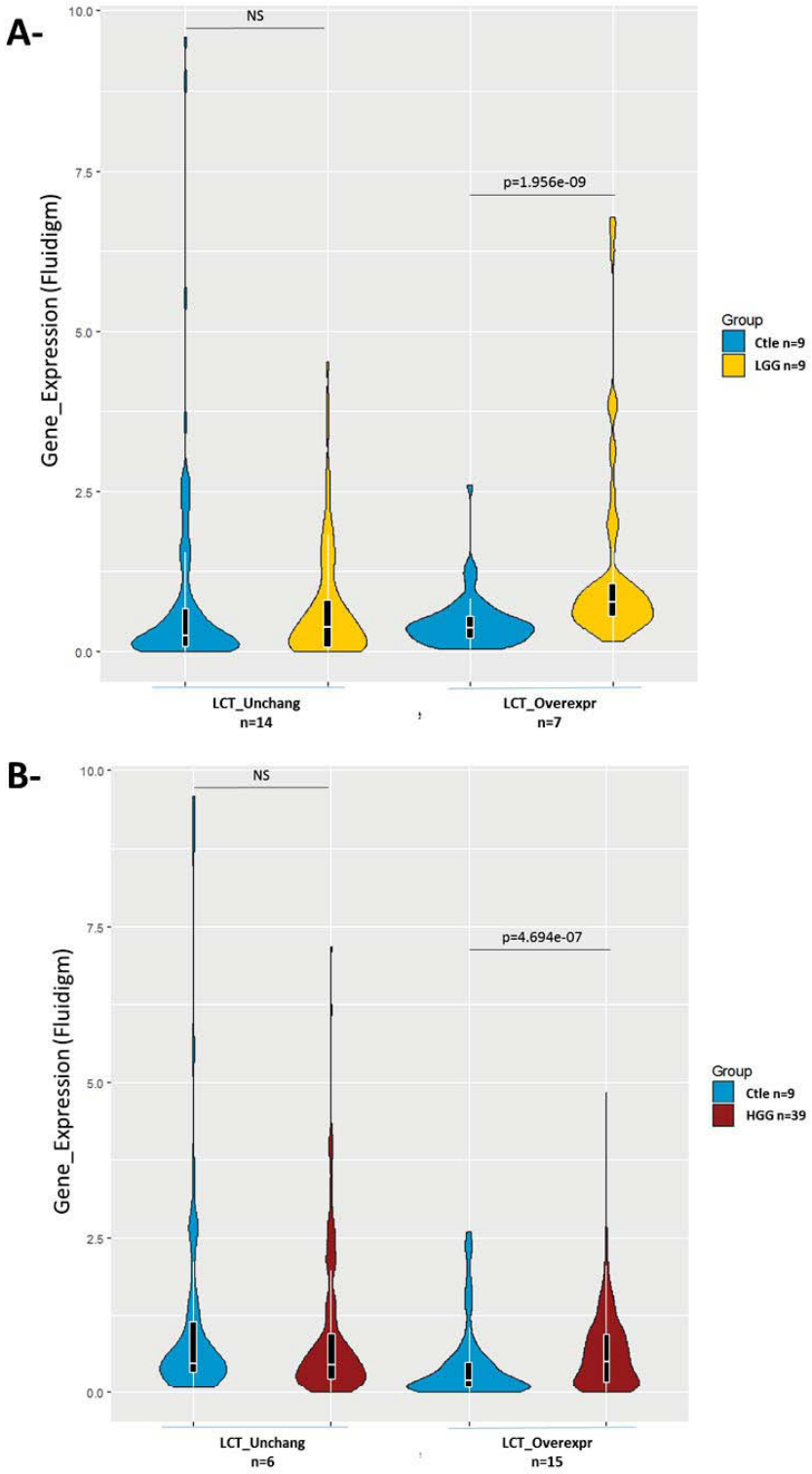


Figure 50 : Les gènes associés aux LCT montrent des variations d'expression similaires à celles des LCT.

A partir des données de RT-qPCR obtenus sur la cohorte plus large d'échantillons (9 contrôles, 9 LGG et 39 HGG), l'expression relative de 21 gènes associés à des LCT validés est comparée entre tissus contrôles et LGG (A) ou HGG (B). Dans chaque cas, les gènes ont été répartis en 2 catégories selon la variation d'expression identifiée pour le LCT qui leur est associé, à savoir : LCT_Unchang (LCT sans variation d'expression dans les tumeurs) et LCT_overexpr (LCT significativement surexprimé dans les tumeurs). La médiane du niveau d'expression ainsi que les limites des 25ème et 75ème percentiles sont rapportées par les box plots. Les diagrammes en violon représentent la densité de distribution des valeurs. L'existence d'une différence entre le groupe contrôle et le groupe tumoral est évaluée par un test de Mann-Whitney. Les gènes associés à des LCT surexprimés montrent une augmentation d'expression tumorale significative qui va donc dans le même sens que celle des LCT alors que les gènes associés à des LCT sans variation d'expression montrent des niveaux similaires d'expression.

De la même façon, pour les groupes de gènes associés à des LCT surexprimés dans les LGG (n=8) et HGG (n=15), une expression plus importante des gènes dans les tumeurs par rapport aux contrôles est observée (significative pour les LGG et les HGG). Enfin, pour les groupes de gènes associés à des LCT dont l'expression ne varie pas, il n'y a pas de différence significative d'expression entre les LGG et les contrôles et une diminution modérée mais significative de l'expression des gènes est observée dans les HGG par rapport aux contrôles. Ces résultats préliminaires montrent que les gènes contenant des LCT sur- ou sous-exprimés dans les tumeurs présentent eux aussi respectivement une sur- ou sous-expression. Ainsi, ils suggèrent que l'environnement génomique pourrait influer sur l'activité transcriptionnelle de l'ASP.

Sachant que les RNA-seq n'ont été réalisés que sur 16 échantillons, l'analyse de l'expression des gènes associés aux LCT intragéniques a été réalisée sur la cohorte plus large de gliomes et de contrôles par RT-qPCR Fluidigm. Suite aux contraintes du nombre de couples d'amorces analysables sur une puce Dynamic array, l'expression des gènes associés à seulement 21 LCT a été mesurée (17 LCT intragéniques surexprimés / 20 et 4 LCT intragéniques sans expression tumorale modifiée). Pour plusieurs LCT intragéniques, transcrits dans le même sens que le gène, j'ai pu observer qu'ils sont issus d'éléments L1 positionnés dans les premiers introns. Aussi, afin de quantifier spécifiquement l'expression des gènes associés aux LCT, sans qu'il y ait d'interférence avec l'expression du LCT pouvant lui aussi contenir les exons en aval, les couples d'amorces utilisés pour étudier l'expression génique ont été positionnés systématiquement en amont de l'initiation de la transcription du LCT.

Une analyse comparative, similaire à la précédente, a été réalisée confrontant les taux d'expression des gènes obtenus en RT-qPCR entre les échantillons contrôles (n=9) et tumeurs LGG (n=9) ou HGG (n=39) (Figure 50-A et -B). Dans cette deuxième analyse, la catégorie des gènes associés à des LCT sous-exprimés n'est plus présente, car les gènes associés aux LCT sous-exprimés n'ont pas été étudiés. Les résultats obtenus (Figure 48-B et C) montrent que :

- L'expression des gènes associés à des LCT surexprimés est significativement augmentée dans les tumeurs (LGG et HGG) par rapport aux contrôles.
- L'expression des gènes associés à des LCT dont l'expression ne varie pas ne présentent pas de différence significative entre les contrôles et les tumeurs (LGG et HGG).

Ainsi, l'analyse de données d'expression génique plus précises, obtenues sur une cohorte plus large, confirme les observations préliminaires.

Une analyse d'expression différentielle des gènes entre les groupes contrôle, LGG et HGG a ensuite été faite (Figure 51, Annexe 9). Dans 10 cas / 21 gènes étudiés, une surexpression

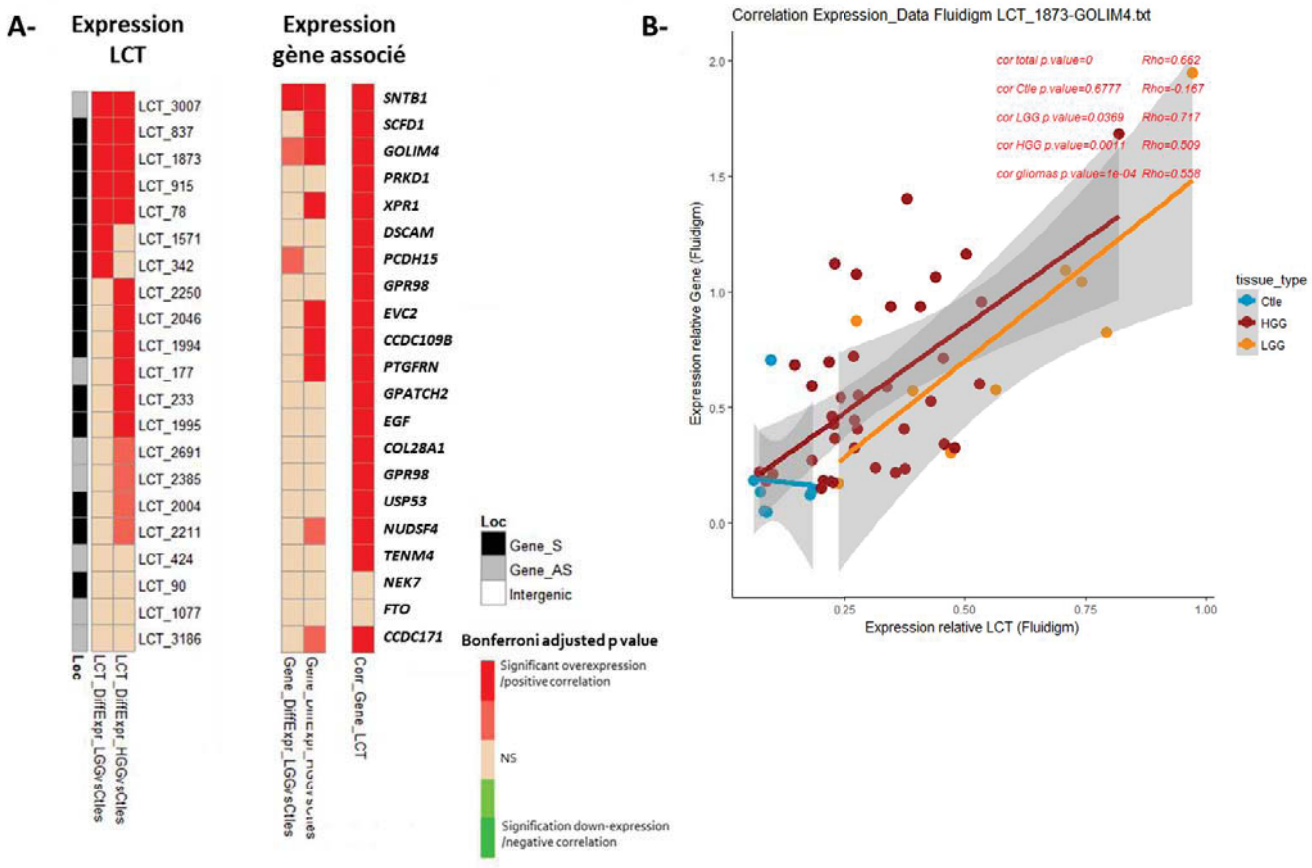


Figure 51 : Une corrélation est observée entre les niveaux d'expression du LCT et ceux du gène qui lui est associé pour tous les LCT surexprimés.

A. L'expression des gènes associés à 21 LCT intragéniques a été analysée par RT-qPCR Fluidigm sur la cohorte plus large d'échantillons (contrôles n= 9, LGG n=9, HGG n= 39). Parmi les 21 LCT considérés, 14 sont transcrits en sens du gène et 7 sont transcrits en antisens (Loc). De plus, 17 sont surexprimés dans au moins un groupe tumoral et 4 ne présentent pas de variation d'expresion dans les tumeurs (LCT_DiffExpr). Dans un premier temps, l'expression différentielle de chaque gène entre tumeur et contrôles a été évaluée pour chacun des groupes tumoraux LGG et HGG (Gene-DiffExpr) par un test de Mann-Whitney. Dans un deuxième temps, les données d'expression de LCT et d'expression du gène associées ont été corrélées entre elles (corrélation Spearman). Dans tous les cas, les p-values ont été corrigées par le coefficient de Bonferroni. Les heatmap représentent la significativité des p-values ajustées (en rouge sur-expression ou corrélation positive significative, en rose tendance à la sur-expression proche de la significativité, en beige p values non significatives). Une corrélation positive significative est observée pour tous les LCT surexprimés et ce, que le LCT soit transcrit en sens ou en antisens du gène dans lequel il se trouve.

B. Résultats des corrélations de Spearman entre l'expression relative du LCT1873 et l'expression relative du gène *GOLIM4* qui lui est associé qui montrent une corrélation positive significative.

significative des gènes dans un ou deux groupes tumoraux est observée. Celle-ci est alors le plus souvent identique à celle observée pour les LCT correspondants. Ainsi, comme pour l'expression des LCT 3007 et 1873, une surexpression de leurs gènes associés (respectivement *SNTB1* et *GOLIM4*) est observée dans les deux groupes tumoraux. Pour les LCT342, 2046, 1994, 177 et 2211 une surexpression des gènes associés n'est retrouvée que dans le groupe tumoral où chaque LCT est lui aussi significativement surexprimé. Enfin deux gènes, *SCFD1* et *XPR1* montrent une surexpression significative uniquement dans le groupe des HGG alors que leurs LCT associés (respectivement LCT 837 et 78) ont été identifiés comme surexprimés dans les deux groupes tumoraux. Seul le gène CCDC171 présente une tendance à la surexpression dans les HGG alors que le LCT3186 associé ne montre pas de surexpression. Dans les 6 autres cas de gènes associés à des LCT surexprimés, aucune expression différentielle significative des gènes associés n'est obtenue.

Afin de vérifier s'il existe une relation entre le niveau d'expression du LCT et le niveau d'expression du gène dans lequel il se trouve dans l'ensemble de la cohorte, une analyse de corrélation de Spearman a été réalisée pour les 17 LCT surexprimés et les 4 LCT sans variation d'expression (Figure 5-B, Annexes 10 et 11). Les résultats obtenus montrent une corrélation positive significative pour les 17 LCT surexprimés et pour 2 LCT sur les 4 sans variation d'expression (LCT 424 et 3186). Une analyse de régression linéaire univariée confirme ces corrélations pour 15 LCT surexprimés (tous sauf LCT 915 et 2211) et pour 1 LCT sans variation d'expression (LCT 424). Afin de savoir si les relations mises en évidence sur l'ensemble de la cohorte pourraient être due à un effet de groupe (contrôles, LGG ou HGG), une régression linéaire multivariée prenant en compte la variable groupe a été réalisée. Celle-ci montre que pour le LCT2046 la relation positive trouvée repose sur le groupe HGG (n=49). Par contre, pour tous les LCT surexprimés et pour le LCT424, la régression linéaire multivariée confirme que la relation positive forte entre l'expression du LCT et l'expression du gène dans lequel il se trouve existe indépendamment du groupe. L'analyse de la localisation du gène montre de plus que cette relation existe quand bien même le sens de transcription du LCT est antisens à celui du gène (LCT3007, 177, 2691, 2385 et 424).

L'ensemble de ces résultats montre que :

- Les gènes contenant des LCT sur- ou sous-exprimés dans les tumeurs présentent eux aussi respectivement une sur- ou sous-expression,

- Il existe dans la plupart des cas une relation positive forte entre l'expression du LCT et l'expression du gène et ce, dans les tumeurs comme dans les échantillons contrôles,
- La relation entre expression du LCT et expression du gène est retrouvée que les LCT soient transcrits en sens ou en antisens de la transcription du gène.

Ceci conforte l'hypothèse que l'environnement génomique d'éléments L1 producteurs de LCT influerait sur l'activité transcriptionnelle de l'ASP.

3. Conclusion partie 3.

La question posée dans cette partie concernait le(s) mécanisme(s) sous-jacents pouvant expliquer l'augmentation de l'activité transcriptionnelle, en contexte tumoral, des ASP de certains L1 impliqués dans la production de LCT ; conduisant à la surexpression de ces derniers.

Dans ce contexte, deux hypothèses, étayées par des données la littérature, ont été étudiées.

Mes résultats tendent à réfuter la première puisqu'ils montrent qu'une diminution de méthylation ADN n'est pas retrouvée au niveau de la région promotrice des L1 associés à des LCT surexprimés. Aussi, l'hypométhylation du promoteur des L1 dans les tumeurs ne semble pas expliquer l'augmentation de l'activité transcriptionnelle à partir de l'ASP.

Par contre, les gènes associés à des LCT dont l'expression est dérégulée en contexte tumoral présentent eux aussi une dérégulation dans le même sens et ce, quel que soit le sens de transcription du LCT par rapport à celui du gène associé. Cette observation suggère qu'il existerait une dérégulation transcriptionnelle aux loci des LCT sur- et sous-exprimés. Ainsi, dans les tumeurs, la sur- ou sous-expression d'un LCT pourrait être associée à une modification de la chromatine au locus. En contexte tumoral, l'augmentation, observée de façon concomitante, de la transcription à partir de l'ASP et du promoteur du gène associé pourrait être alors le reflet de l'acquisition d'une chromatine plus permissive, facilitant l'accès de l'ADN du locus à la machinerie de transcription. Ceci conforte la deuxième hypothèse et permet de proposer que l'activité transcriptionnelle de l'ASP d'un L1 pourrait être influencée par celle du locus.

Partie 4 : Fonctionnalité des LCT dans les processus tumoraux ?

1. Objet de l'étude partie 4.

Différentes études ont mis en évidence environ 250 LCT (228 décrits par Criscione et 14 autres de la littérature) dans différents tissus tumoraux et normaux à partir de bases de données d'EST ou grâce à une approche moléculaire dédiée, dont certains à l'image de *L1-MET* et du LCT13 semblent jouer un rôle fonctionnel dans les processus de tumorigenèse (Criscione *et al.*, 2016; Cruickshanks *et al.*, 2013; Cruickshanks & Tufarelli, 2009; Mätlik *et al.*, 2006; Nigumann *et al.*, 2002; Wolff *et al.*, 2010).

Plus de 2000 chimères ont été identifiées dans les données de RNA-seq de gliomes. Parmi celles-ci, 78 correspondent à des LCT déjà décrits dans des types tumoraux autres que les gliomes (Partie 2). Des données publiques de RNA-seq provenant de métastases ovariennes et de lignées de cancer du sein (MCF-7) ont été analysées et des chimères ont été identifiées avec CLIFinder dans ces tissus (Partie 1). La question se pose donc de savoir si certains de ces LCT pouvaient jouer un rôle fonctionnel dans les processus tumoraux. Dans un premier temps, la comparaison des résultats de CLIFinder obtenus sur les différents jeux de données a été réalisée afin d'identifier des chimères communes, pour lesquelles on pose l'hypothèse que leur détection dans différents types tumoraux pourrait indiquer qu'elles jouent un rôle fonctionnel. Dans un second temps, afin d'évaluer si parmi les LCT validés dans nos gliomes certains d'entre eux peuvent correspondre à de nouveaux biomarqueurs, des analyses de bio-statistiques confrontant les niveaux d'expression des LCT (obtenus par RT-qPCR sur la cohorte plus large) à la survie des patients ont été réalisées. Enfin, l'expression des gènes en aval de la transcription des LCT biomarqueurs et/ou présentant une surexpression tumorale a été étudiée afin d'essayer d'identifier des LCT candidats pour un rôle fonctionnel.

2. Résultats partie 4.

2.1. De nombreuses chimères de gliomes sont retrouvées dans d'autres types tumoraux.

Les chimères obtenues à partir de nos gliomes ont été comparées à celles obtenues avec les données publiques pour la publication de CLIFinder. Les données de RNA-seq analysées proviennent de séquençage paired-end orientés d'ARN messagers (polyA) extraits à partir de

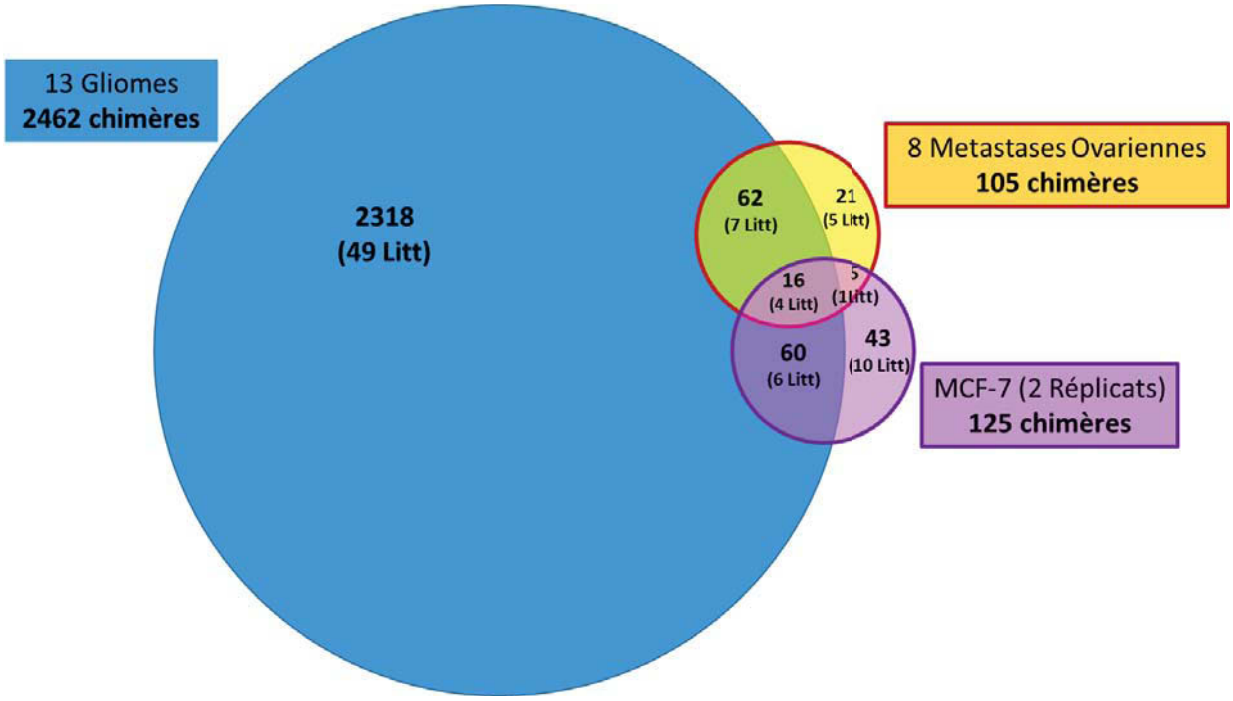


Figure 52 : De nombreuses chimères retrouvées dans des métastases ovariennes et une lignéesde cancer mammaire sont communes avec les chimères identifiées dans nos gliomes.

Le logiciel CLIFinder a identifié 2462 chimères uniques dans 13 gliomes, en bleu, 105 chimères uniques dans 8 métastases ovariennes (Böhm et al., 2016), en jaune, et 125 chimères uniques dans 2 réplicas de RNA-seq pour la lignée MCF-7 (Philippe et al., 2016), en violet. La comparaison des positions génomiques des chimères obtenues montre que 74% des chimères identifiées dans les métastases ovariennes et 61 % des chimères identifiées dans la lignée MCF-7 ont été également identifiées dans nos gliomes. L'indication « Litt » signifie que des chimères correspondant à des LCT déjà décrits dans la littérature sont identifiées. Au final, 16 chimères sont communes aux 3 analyses et parmi elles, 4 sont déjà décrites dans la littérature.

métastases ovariennes et de la lignée cellulaire MCF-7. Sur ces données de RNA-seq, les mêmes paramètres de l'analyse A37 ont été appliqués, à savoir :

- L'utilisation des séquences consensus des 27 sous-familles de L1 (Khan, 2005), avec au moins 30 pb du R1 correspondant à ces séquences avec 6 mésappariements tolérés,
- Pour le filtre 2, l'extrémité R2 doit contenir au moins 30 pb correspondant à une séquence unique.
- Réalisation de l'alignement de ces lectures filtrées sur le génome humain (hg19) avec un espacement maximal toléré de 50 Kb.

Les résultats obtenus sont triés selon les critères vus précédemment (Partie 2) :

- Elimination des chimères dont le sens de la transcription n'est pas compatible avec une initiation de la transcription à l'ASP du L1,
- Elimination des chimères dont le L1 ne possède pas dans sa région 5'UTR la région critique de l'ASP (+400 à +600).

L'analyse A37 identifie 2462 chimères uniques à partir des 13 gliomes. Pour l'analyse des 8 métastases ovariennes, 105 chimères uniques sont obtenues et pour l'analyse avec MCF-7, 125 chimères uniques sont détectées dans les données de RNA-seq. La comparaison des positions génomiques des chimères obtenues dans les 3 types de tissus tumoraux montre que 61% des chimères MCF7 et 74% des chimères des métastases ovariennes sont retrouvées dans les gliomes (Figure 52). Parmi celles-ci, 16 sont communes aux trois groupes, dont 4 ont déjà été décrites dans la littérature et 2 ont été validées dans notre étude, le LCT1277 qui fait partie des 4 décrits dans la littérature et le LCT801. Les captures d'écran relatives à ces 16 chimères sont détaillées en annexe. Parmi ces 16 chimères communes, 2 sont intergéniques et 14 sont intragéniques (Annexe 12). Différents types de configuration génique caractérisent les 14 chimères intragéniques (Figures 53 à 56, Annexes 13 à 21) :

- *Chimères transcrites dans le sens de gènes, dont le seul TSS annoté correspond à l'ASP du L1 des chimères (n=3) :*

Deux chimères sont initiées au niveau de l'ASP d'un L1PA2 localisé respectivement au niveau des TSS des gènes *LINC00649* et *ALG1L*. Une autre chimère semble initiée à l'ASP d'un L1PA3 localisé au niveau du TSS du gène *ARHGAP25*. Pour ce dernier LCT, un évènement d'épissage a lieu entre la séquence L1 en 5' et la séquence unique en 3' puisque la distance entre le R1 et le R2 est supérieure à 30 Kb (Figure supplémentaire S2 article partie 1).

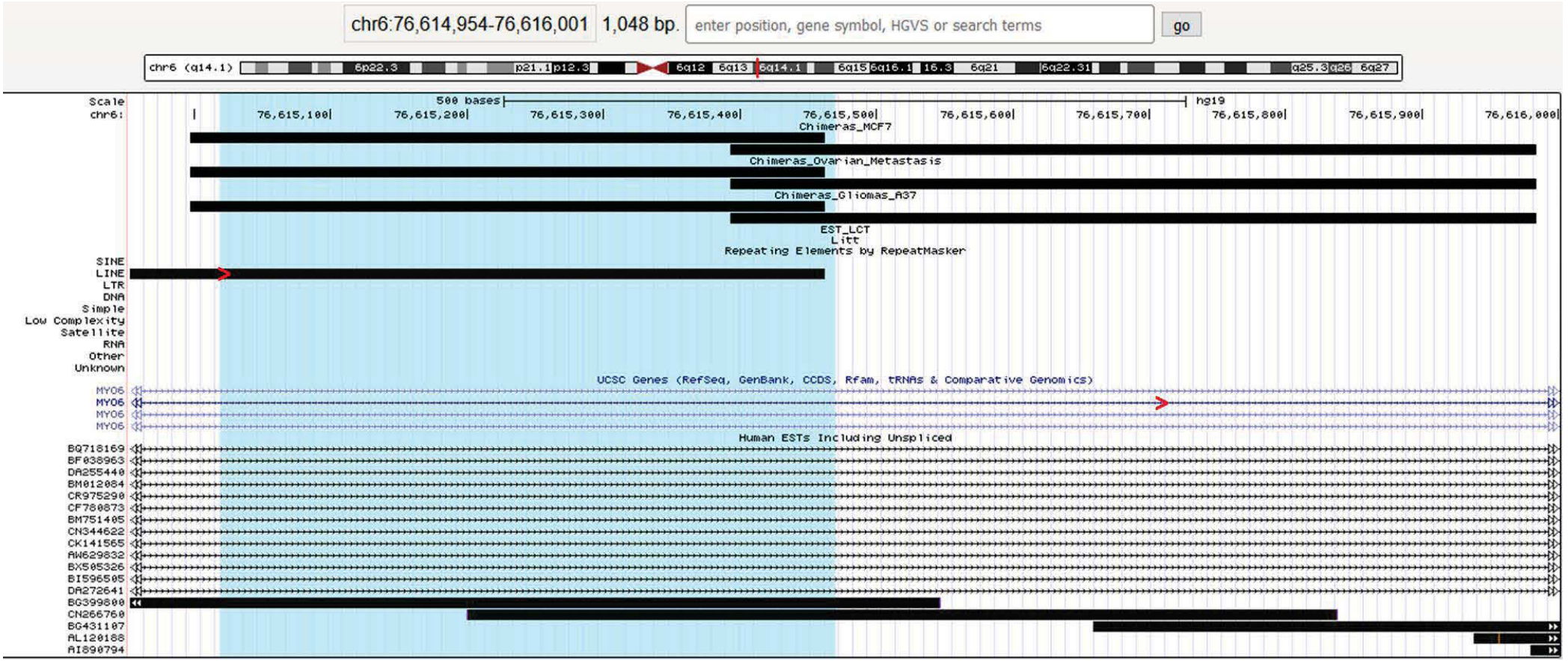


Figure 53 : Chimère 1552 initiée à l'ASP d'un L1PA3 localisé dans un intron du gène MYO6 en sens.

Région de la chimère 1552 qui semble initiée à partir de l'ASP du L1PA3 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquent le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 30/34 du gène *MYO6*. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes. 1 EST est initié à 200 pb du début du L1 et se poursuit dans la région adjacente.

- *Chimères transcrites dans le sens de gènes, localisées dans un intron, dont l'ASP semble se comporter alors comme un TSS alternatif (n=2) :*

La chimère 1280 est localisée dans l'intron 21/25 du gène *ERBB2IP*, qui serait initiée à l'ASP d'un L1PA4. 1 EST est décrit et débute au site +1 du L1.

La chimère 1552 est localisée dans l'intron 30/34 du gène *MYO6* et serait initiée à l'ASP d'un L1PA3. 1 EST est décrit et est initié à 200 pb du début du L1 (Figure 53).

- *Chimères transcrites dans le sens de gènes, impliquant une jonction intron-exon, dont l'ASP semble se comporter alors comme un TSS alternatif (n=3) :*

Le LCT1277 est initié à l'ASP d'un L1HS localisé dans l'intron 20/27 du gène *ZNF638* et a été validé au cours de mes travaux de thèse (Figure 54). Ce LCT semble être épissé puisque la distance entre le R1 et le R2 est de 5,3 Kb. De plus, le R2 est localisé dans l'exon 22/28 de ce gène. Par ailleurs, 1 EST débute au niveau de l'ASP et est épissé avec les exons 21 et 22 du gène *ZNF638*. Ce LCT semble être un candidat de choix pour la validation d'évènement d'épissage ayant lieu entre le L1 et les gènes adjacents. Les expériences sont donc à poursuivre grâce au dessin d'amorces dans les exons, afin de valider par RT-PCR l'amplification entre le L1_common et les amorces SU-exon21 et SU-exon22.

La chimère 2333 semble initiée à l'ASP d'un L1PA5 localisé dans l'intron 11/15 du gène *ITGB1* et dont le R2 est localisé dans l'exon 12/16 du gène. Il n'y a pas d'évidence d'EST à ce locus.

La chimère 1977 semble initiée à l'ASP d'un L1PA7 localisé dans l'intron 11/11 du gène *VPS37A* et dont le R2 est localisé dans l'exon 12/12. 3 EST sont décrits comme débutant dans le L1 et se poursuivant dans la région génomique adjacente.

- *Chimères transcrites en antisens de gènes, localisées dans un intron ou à une jonction intron-exon (n=6) :*

La chimère 1157 semble initiée au niveau de l'ASP d'un L1PA2 localisé dans l'intron 2/8 du gène *NR3C2* en antisens de celui-ci. 1 EST décrit remonte de 200 pb dans le L1 (Figure 55).

La chimère 1393 est localisée dans l'intron 5/6 du gène *COMMD10* et semble initiée dans un L1PB1. Sachant que la sous-famille L1PB1 est ancienne, je me questionne sur la réalité biologique qu'une telle chimère puisse correspondre à un LCT initié à l'ASP, malgré la présence d'1 EST décrit, initié dans ce L1.

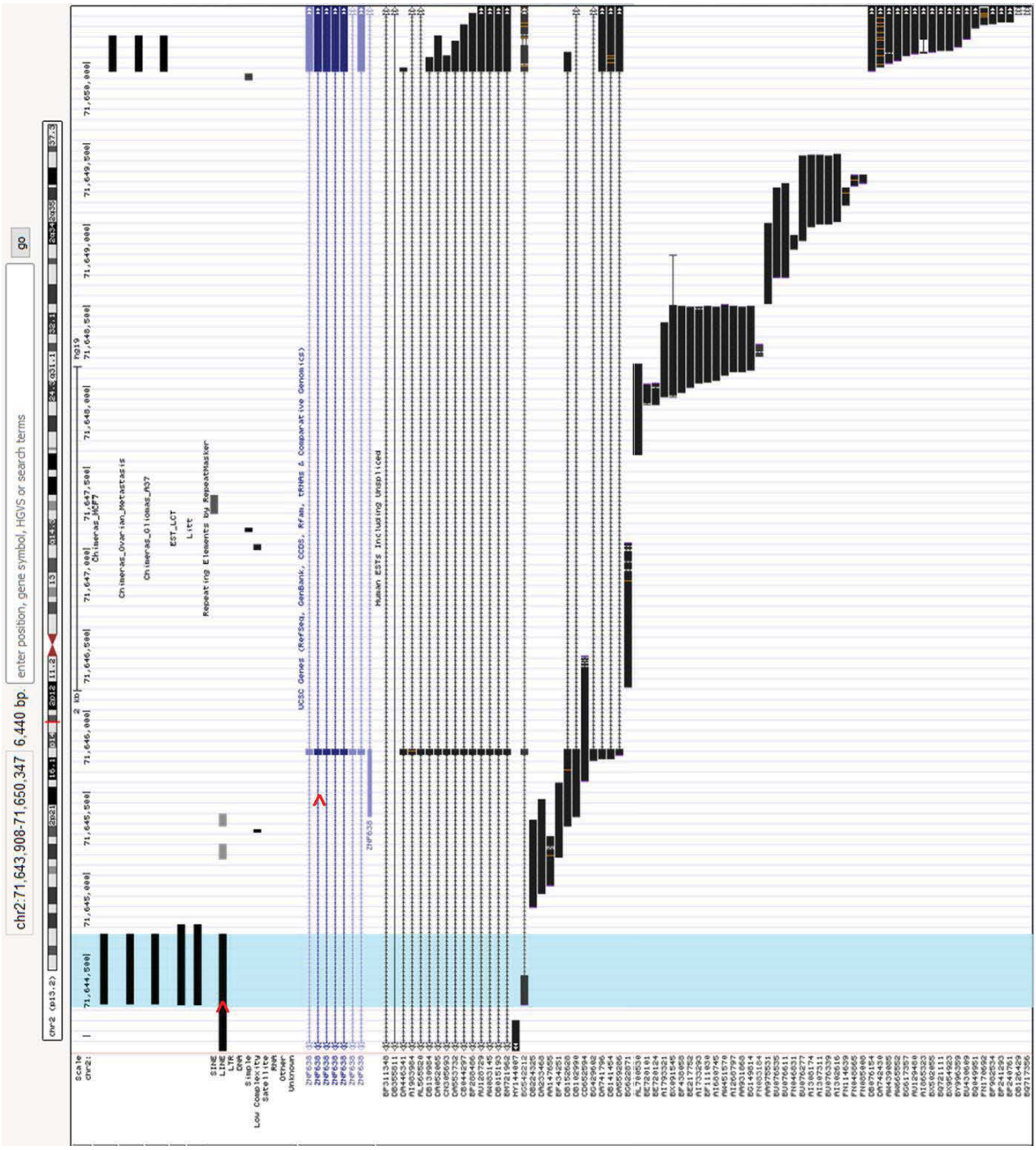


Figure 54: LCT1277 initiée à l'ASP d'un L1HS, localisée à la jonction intron/exon du gène ZNF638 en sens.

Région du LCT1277, initiée à partir de l'ASP du L1HS dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquent le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 20/27 du gène *ZNF638* et le R2 est localisé dans l'exon 22/28. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes. 1 EST est initié à 450 pb du début du L1 et se poursuit dans la région adjacente avec un évènement d'épissage, en effet, cet EST inclus les exons 21 et 22.

La chimère 3062 semble initiée à l'ASP d'un L1P1 (sous-famille qui possède une forte homologie de séquence avec la sous-famille L1PA1 ou L1HS) qui est localisé dans l'intron 11/13 du gène *TMEM62*. Aucune évidence d'EST n'est observée dans cette région.

De plus, 3 chimères 2468, 1984 et le LCT801 (Figure 56) sont localisées dans les gènes *PDE3B*, *BNIP3L* et *RB1* dont le R1 débute à l'ASP d'un L1PA3, L1PA2 ou L1HS respectivement, localisés dans un intron. Le R2 est quant à lui localisé dans un exon. Pour autant, aucune évidence d'EST n'est observée pour ces 3 chimères. Ces chimères sont transcrites de manière linéaire et le fait qu'elles contiennent une séquence exonique (en antisens), n'impliquent pas un épissage.

L'analyse par Gene Ontologie des gènes associés à ces chimères a été réalisée et aucun rôle biologique particulier n'est mis en évidence. Mais certaines chimères pourraient avoir un rôle fonctionnel, c'est pourquoi elles ont été sélectionnées dans le groupe de chimères à valider. En effet, la chimère 801 est localisée à la jonction intron-exon en antisens du gène *RB1*, qui est un régulateur négatif du cycle cellulaire et a été décrit comme ayant un rôle dans l'étiologie des cancers. La chimère 1277 est quant à elle localisée à la jonction intron-exon dans le sens du gène *ZNF638*, qui joue un rôle dans la maturation des transcrits dans le nucléoplasme. Les résultats de marche en 5' valident que ces chimères correspondent à des LCT mais celles-ci ne présentent pas de différence significative d'expression entre les tumeurs et les contrôles.

Les études sont à poursuivre quant aux rôles de ces LCT dans les processus tumoraux et seront discutées dans la partie suivante.

2.2. Des candidats fonctionnels parmi les LCT validés ?

Afin d'évaluer si parmi les LCT validés certains d'entre eux peuvent correspondre à de nouveaux biomarqueurs pour les gliomes, des analyses biostatistiques confrontant les niveaux d'expression des LCT (obtenus par RT-qPCR sur la cohorte plus large) à la survie des patients ont été réalisées. En effet, pour chaque gliome de la tumorothèque les données clinico-biologiques des patients sont disponibles. Ainsi les analyses biostatistiques ont permis d'identifier 10 LCT pouvant jouer un rôle de biomarqueur vis-à-vis de la survie. Pour autant aucun n'est plus fort que le biomarqueur actuellement utilisé qui est le statut de la mutation du gène *IDH1*. Ces analyses biostatistiques se basent sur l'expression des LCT associée à la survie des patients pour lesquels les échantillons (LGG et HGG) ont été analysés par RT-qPCR. Le premier et le dernier quartile ont été extraits afin de déterminer la relation entre la survie du patient et l'expression du LCT. Ainsi, 10 LCT ont été identifiés comme pouvant avoir un rôle

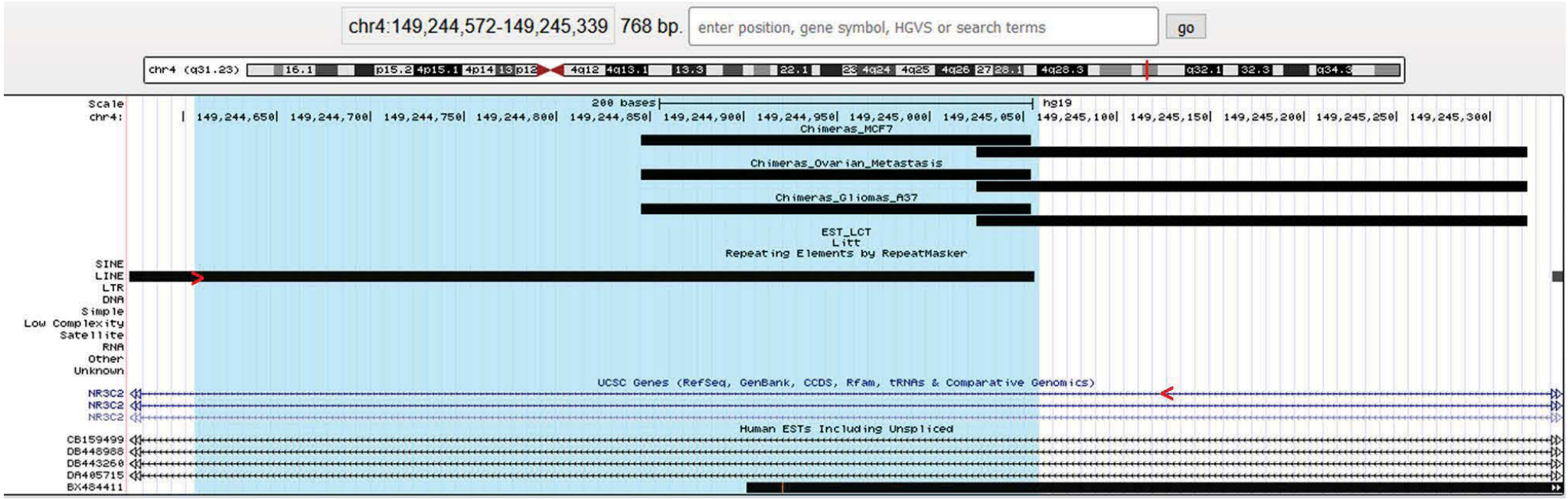


Figure 55 : Chimère 1157 initiée à l'ASP d'un L1PA2 localisé dans un intron du gène NR3C2 en antisens.

Région de la chimère 1157 qui semble initiée à partir de l'ASP du L1PA2 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquent le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 2/8 du gène *NR3C2*. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes. 1 EST est initié à 200 pb du début du L1 et se poursuit dans la région adjacente.

de biomarqueur prédictif vis-à-vis de la survie. Le « hazard ratio » (HR) est une valeur calculée au cours des analyses de survie qui nous est donnée pour chaque LCT ayant un potentiel de biomarqueur. Si le HR est inférieur à 1, cela signifie que l'expression du LCT dans les tumeurs est associée à une meilleure survie, alors que si celui-ci est supérieur à 1 cela signifie que l'expression du LCT est un facteur de mauvais pronostic. 5 LCT possèdent un HR inférieur à 1 ce qui signifie que leur expression dans les tumeurs est associée à un facteur de bon pronostic, dont 3 LCT (LCT1883, 1994 et 2046) sont significativement surexprimés dans les HGG. Les 2 autres LCT (LCT90 et 1077) ne présentent pas de différences significatives d'expression dans les tumeurs. 5 LCT possèdent un HR supérieur à 1 c'est-à-dire que leur expression dans les tumeurs est associée à un facteur de mauvais pronostic pour la survie des patients, dont 4 LCT (LCT837, 1873, 2004 et 2385) sont associés à une surexpression significative dans les gliomes par rapport aux contrôles. Le LCT424 quant à lui est exprimé dans les tumeurs et les contrôles.

Afin d'identifier si les LCT surexprimés et/ou biomarqueur peuvent avoir un rôle fonctionnel dans les processus de tumorigenèse, l'expression des gènes localisés dans une fenêtre de 500 Kb en aval du LCT a d'abord été analysée en utilisant les données d'expression relative mesurées dans les 16 RNA-seq (en fpkm). Ainsi, l'expression relative de 71 gènes localisés en aval des 24 LCT surexprimés dans les tumeurs par rapport aux contrôles (dont 7 sont également biomarqueurs) et des 3 LCT ayant seulement un rôle de biomarqueur a été extraite des données de RNA-seq. Les analyses de corrélation entre l'expression relative du gène en aval et l'expression du LCT correspondant ne montrent pas de corrélation significative. Toutefois, pour 6 LCT, une tendance à une corrélation (positive ou négative) entre l'expression du LCT et l'expression du gène en aval est observée et ces LCT pourraient jouer un rôle fonctionnel sur les variations d'expression de ces gènes.

Pour 5 LCT une tendance à une corrélation inverse est observée. Ainsi, un ou deux gènes en aval des LCT837 (*STRN3* et *HECTD1*), 78 (*STX6*), 1873 (*SERPINI1* et *PDCD10*), 1995 (*ELOVL6*) et 1994 (*PLA2G12A*) orientés en antisens (sauf *PDCD10*) du LCT ont une expression diminuée dans les tumeurs par rapport aux contrôles, de manière concomitante à une augmentation de l'expression de ces LCT. Pour 4 de ces candidats, l'hypothèse d'un mécanisme d'ARN interférence impliqué dans la diminution de l'expression d'un gène en aval orienté en antisens est envisagée. Pour le LCT1873, à l'image de ce qui a été décrit dans la littérature pour le LCT13 de Cruickshanks (Cruickshanks *et al.*, 2013), l'hypothèse d'un long transcrit non codant permettant la mise en place de marques d'histones répressives induisant la mise sous-silence de 2 gènes en aval est envisagée. En effet, ce LCT1873 est localisé à proximité de 2

Figure 56 : LCT801 initiée à l'ASP d'un L1HS, localisée à la jonction intron/exon du gène RB1 en antisens.

Région du LCT801, initiée à partir de l'ASP du L1HS dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquent le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 23/26 du gène *RB1* et le R2 est localisé dans l'exon 23/27. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.

gènes, *PDCD10* et *SERPINI1*, positionnés en orientation opposée sous la dépendance d'un seul promoteur bidirectionnel.

Pour le 6ème candidat, le LCT2691, la situation est inverse avec une corrélation positive entre l'expression du LCT et celle d'un gène en aval, *MIOS*, orienté dans le même sens de la transcription que le LCT.

Enfin, pour le LCT1994, il est à noter qu'une deuxième hypothèse fonctionnelle (différente de celle de régulation négative du gène en aval *PLA2G12A*) est envisagée. En effet, l'élément L1 à l'origine du LCT1994 est localisé dans l'intron 1 et le LCT est transcrit en sens du gène *CCDC109B* dans lequel il se trouve. L'hypothèse a été émise que le LCT1994 pourrait alors aboutir à la production d'un transcrit alternatif plus court pouvant conduire à l'expression d'une protéine dépourvue de son extrémité N-terminale.

Sachant que les RNA-seq n'ont été réalisées que pour 16 échantillons, afin de confirmer ces résultats et hypothèses, l'expression de ces gènes a été analysée par RT-qPCR sur la cohorte plus large (9 contrôles, 9 LGG et 39 HGG).

L'analyse de l'expression de 17 gènes situés en aval de 6 LCT candidats a été réalisée (incluant non seulement les gènes cibles identifiés mais aussi les autres gènes de la région de 500kb, Annexe Matériels et Méthodes Table 4).

Les analyses de corrélation entre l'expression du LCT et l'expression du gène cible potentiel en aval ne montrent pas de corrélation significative sauf pour le LCT1873 et le gène *SERPINI1 (Rho=-0,367 ; p=0,005)*. A l'image de ce qui a été décrit dans la littérature pour *PDCD10* et *SERPINI1* localisés en aval du LCT1873, (Chang *et al.*, 2000; Lambertz *et al.*, 2015), l'expression de ces 2 gènes est très fortement diminuée dans nos gliomes (en RNA-seq et confirmé par RT-qPCR). En effet, pour le gène *PDCD10,* dont la transcription va dans le sens de celui du LCT, une diminution significative de l'expression est observée dans les LGG et non dans les HGG. Pour ce qui est du gène *SERPINI1,* qui est orienté en antisens du LCT, l'expression différentielle de ce gène par RT-qPCR confirme une diminution significative dans les LGG et les HGG par rapport aux contrôles (Figure 57). Le gain de méthylation d'un promoteur étant classiquement décrit dans la littérature comme responsable d'une répression de la transcription en contexte tumoral, les données de puces de méthylation HM450K (qui ont été réalisées par ailleurs sur nos échantillons) ont été analysées. Celles-ci ne montrent aucune différence du profil de méthylation au niveau de ce promoteur entre les tumeurs et les contrôles, ce qui suggère l'implication d'un autre mécanisme dans la répression de la transcription de ces 2 gènes. Suite à l'observation d'une diminution d'expression de *SERPINI1* concomitante à une

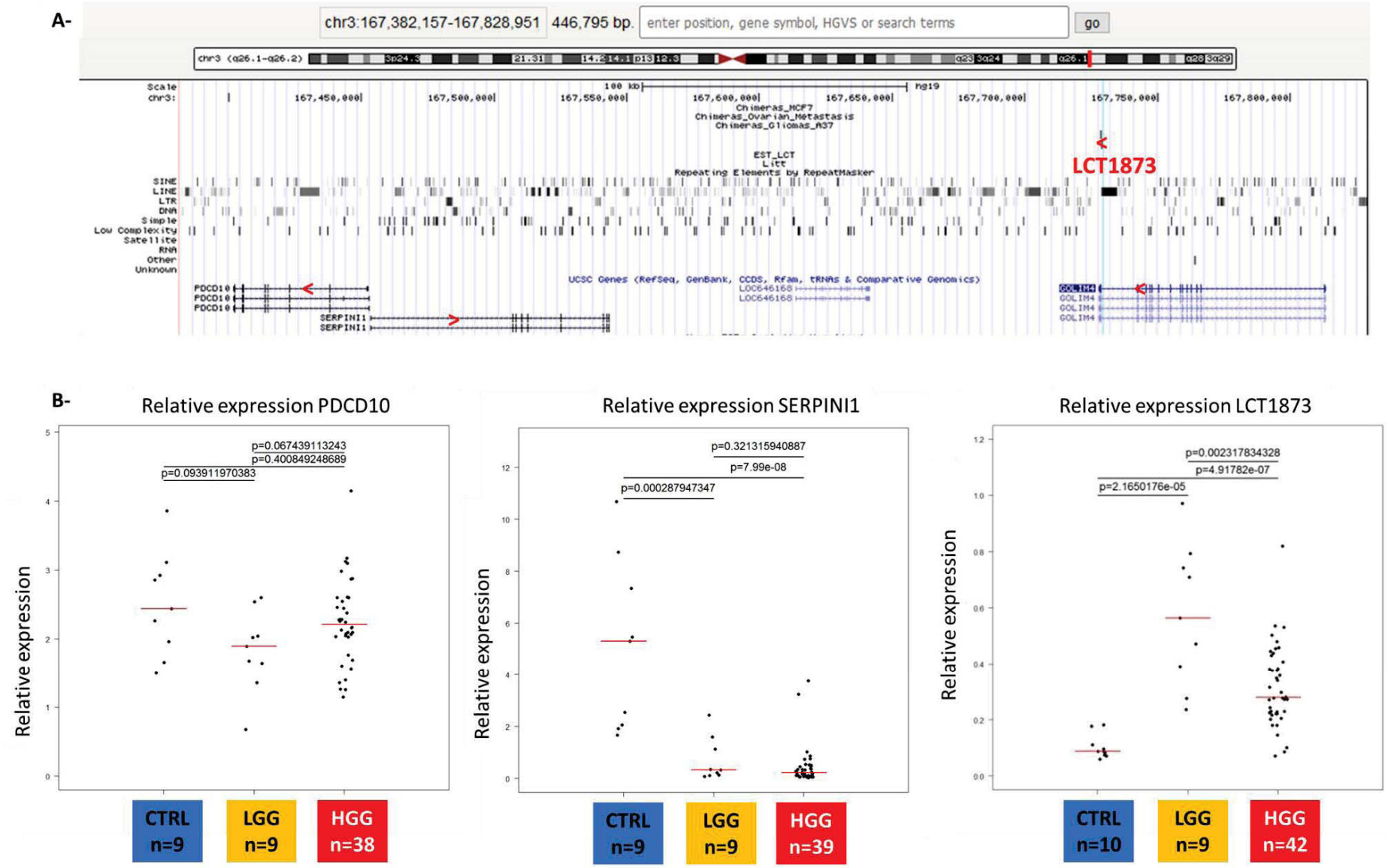


Figure 57 : Rôle fonctionnel du LCT1873 dans la mise sous silence des gènes PDCD10 et SERPINI1 ?

A- Impression d'écran de la fenêtre de 500 Kb en aval du LCT1873 dans UCSC. Le LCT1873 est intragénique au gène *GOLIM4* et est orienté dans le même sens que celui du gène (le sens de la transcription est indiqué par les flèches rouges). A 180 Kb de ce LCT se trouvent les gènes *PDCD10* et *SERPINI1* en orientation opposée l'un de l'autre sous la dépendance d'un promoteur bidirectionnel. Le gène *PDCD10* est orienté dans le même sens que celui du LCT1873 et le gène *SERPINI1* en antisens.

B- Expression différentielle des gènes *PDCD10*, *SERPINI1* et du LCT18736 sur la cohorte plus large de gliomes (RT-qPCR). Les p-values correspondent à un test de Mann-Whitney et indiquent si la différence d'expression par rapport aux contrôles ou entre les groupes tumoraux est significative.

augmentation d'expression du LCT1873, je me pose donc la question de l'implication de la surexpression du LCT1873 dans la diminution de ces gènes, notamment *SERPINI1*.

Enfin pour le LCT1994, les résultats d'expression, sur la cohorte plus large, ne confirment pas une corrélation inverse entre l'augmentation d'expression du LCT et une diminution du gène en aval *PLA2G12A*. Par contre, ce LCT, transcrit à partir de l'intron 1 de *CCDC109B*, pourrait inclure dans sa région 3' les exons 2 à 8 du gène, à l'image de ce qui a été décrit pour le LCT *L1-MET* dans les cancers de la vessie (Wolff *et al.*, 2010). Le gène *CCDC109B* code pour une protéine régulant négativement un canal calcique mitochondrial. De par la présence d'un codon ATG en début de l'exon 3, le LCT1994 pourrait alors coder une protéine tronquée ayant perdu son signal d'adressage à la mitochondrie et modifier l'absorption calcique de la cellule.

Au cours de mes travaux, 2 LCT candidats pour un rôle fonctionnel se dégagent de mes résultats à savoir le LCT1873 et le LCT1994. Les validations restent à poursuivre afin d'identifier les mécanismes par lesquels ils peuvent induire ou favoriser le développement tumoral.

3. Conclusion partie 4.

Au cours de mes travaux de thèse, des chimères ont été identifiées dans différents types tumoraux grâce au logiciel CLIFinder. En effet, 74% des chimères identifiées par CLIFinder dans les données de RNA-seq de métastases ovariennes et 61% des chimères détectées dans les données de RNA-seq de MCF7, sont retrouvées dans nos gliomes. 16 chimères sont communes aux 3 analyses et parmi elles, 2 ont été validées au cours de mes travaux. Le LCT1277 semble être un candidat à étudier pour la validation d'évènements d'épissage pouvant avoir lieu entre les sites donneurs d'épissage décrit dans la région 5'UTR de L1 au niveau de l'ASP et un site accepteur de la séquence unique adjacente.

Parmi les LCT intragéniques, 10 LCT semblent se comporter comme des biomarqueurs prédictifs de la survie des patients, bien qu'aucun ne soit plus fort que le biomarqueur actuellement utilisé qui est la mutation du gène *IDH1*.

Au final, 2 candidats potentiels, les LCT1873 et 1994, pourraient jouer un rôle fonctionnel dans les processus de tumorigenèse. Des analyses complémentaires sont à réaliser afin de valider ce rôle fonctionnel et seront détaillées dans la discussion.

En termes d'impact à plus long terme, il est important de noter ici que, des LCT ont été identifiées dans différents types tumoraux, que ce soit dans la littérature ou au cours de mes travaux de thèse avec CLIFinder. L'hypothèse que nous faisons est que certains LCT peuvent être une caractéristique récurrente des tumeurs agressives et que certains peuvent jouer un rôle

fonctionnel dans les processus de tumorigenèse à cause d'une dérégulation transcriptionnelle induite au cours de la transformation tumorale.

DISCUSSION

1. CLIFinder : un nouvel outil pertinent pour l'identification pangénomique de LCT dans les tumeurs ?

La majorité des chimères identifiées par CLIFinder impliquant des L1 récents semblent correspondre à des LCT dont l'initiation de la transcription se fait à l'ASP.

Au cours de ma thèse, 2675 chimères uniques ont été identifiées grâce au logiciel CLIFinder dans les données de RNA-seq de 13 gliomes et de 3 tissus cérébraux contrôles. Parmi elles, 2256 impliquent des L1 récents (L1PA1 à 7) supposés posséder un ASP. Un groupe de 84 chimères a été validé par des approches de biologie moléculaire (RT-PCR et séquençage).

Avec CLIFinder, une région chimérique est identifiée mais le logiciel ne nous permet pas de conclure précisément si la chimère est initiée à l'ASP du L1. Aussi, j'ai tout d'abord cherché à identifier la région d'initiation de la transcription pour les 84 chimères validées. La technique de choix pour identifier les sites d'initiation de transcription d'ARN est la technique de 5'RACE. Toutefois, la 5'RACE est coûteuse, longue et difficile à mettre en place pour étudier un grand nombre de transcrits. Ainsi, une approche de marche en 5' a été utilisée. Celle-ci permet l'identification d'une région promotrice mais de façon moins précise. Elle a été choisie comme alternative à la 5'RACE car elle présente les avantages de permettre l'étude d'un grand nombre de transcrits, rapidement et pour un coût raisonnable. 61 chimères ont été testées et cela m'a permis de valider que 89% d'entre elles, impliquant des L1 récents de L1PA1 à 7, sont initiées dans la région de l'ASP et correspondent donc à des LCT. Par contre, pour les 4 chimères testées impliquant un L1PA8, toutes ont montré une amplification en amont de la région. Ces résultats suggèrent donc qu'un ASP fonctionnel serait plus particulièrement retrouvé dans les sous familles L1PA7 à L1PA1 et que la majorité des chimères impliquant des L1 récents correspondent à des LCT.

À cause de ses contraintes, la technique de 5'RACE a été mise en œuvre uniquement dans un deuxième temps pour identifier précisément le site d'initiation de la transcription de 5 LCT d'intérêt. Ces LCT présentent une surexpression significative dans au moins un groupe tumoral et ont un potentiel de biomarqueur vis-à-vis de la survie des patients. Il est donc important de valider que leur initiation de transcription a lieu au niveau de l'ASP du L1 qui leur est associé. Des résultats préliminaires ont été obtenus pour seulement 2 LCT, le LCT1873 et le LCT1994

et restent insatisfaisants de par le fait notamment que l'extrémité 5' identifiée est plus précoce que celle annotée dans les résultats de RNA-seq ayant permis l'identification de la chimère.

Par ailleurs, ce qui me conforte également dans l'idée que les chimères identifiées par CLIFinder correspondent à des LCT est que parmi les 2675 chimères uniques détectées, 78 ont déjà été décrites dans la littérature (Criscione *et al.*, 2016; Cruickshanks & Tufarelli, 2009; Mätlik *et al.*, 2006; Nigumann *et al.*, 2002; Wolff *et al.*, 2010). Ces LCT décrits dans la littérature proviennent de différents types tumoraux : de lignées de cancer de la vessie, du sein, colorectal, de tumeurs issues de patients (vessie, sein), mais également de banques de données d'EST de tissus tumoraux mais aussi normaux. De plus, les analyses réalisées avec CLIFinder sur des métastases ovariennes et sur la lignée de cancer du sein (MCF7) identifient : 16 chimères communes aux 3 analyses, 62 communes aux gliomes et aux métastases ovariennes, 60 communes aux gliomes et à MCF7.

Pour les LCT identifiés par Cruickshanks des analyses de 5'RACE ont été réalisées et ils ont validés que ces LCT sont initiés à l'ASP (Cruickshanks *et al.*, 2013; Cruickshanks & Tufarelli, 2009). Parmi les chimères identifiées avec CLIFinder, certaines correspondent aux LCT décrits par Cruickshanks tels que les LCT8 et LCT9.

Parmi les LCT identifiés avec CLIFinder, certains pourraient correspondre à des LCT ayant subi un évènement d'épissage. En effet, les fragments d'ADNc sélectionnés pour le RNA-seq ont une taille comprise entre 280 et 340 pb. Sachant que le RNA-seq réalisé est paired-end orienté de 100 pb à chaque extrémité, si la distance entre R1 (extrémité 3') et R2 (extrémité 5') est supérieure à 400 pb, un mécanisme d'épissage peut être supposé. Pour les gliomes, 70 chimères (2,6%) possèdent une distance entre R1 et R2 supérieure à 400 pb et parmi elles, 11 ont été décrites dans la littérature comme des LCT épissés (Criscione *et al.*, 2016). Dans les RNA-seq polyA des 8 métastases ovariennes et de la lignée MCF-7, respectivement 39 (37,10%) et 37 (31,20%) chimères potentiellement épissées sont identifiées. Le nombre de chimères épissées obtenu est donc relativement proche entre les analyses des RNA-seq polyA et Ribozero. Pour autant, leur représentativité est drastiquement augmentée dans les résultats des chimères issues de RNA-seq polyA, ceci compte tenu d'un nombre de chimères identifiées plus faible. Enfin, 9 chimères épissées sont communes entre gliomes et métastases ovariennes et/ou MCF-7 et 3 supplémentaires sont retrouvées communes aux métastases ovariennes et à la lignée MCF7.

L'ensemble de mes résultats suggèrent que CLIFinder identifie des chimères correspondant à la définition d'un LCT, à savoir un transcrit dont l'initiation de la transcription se fait à l'ASP

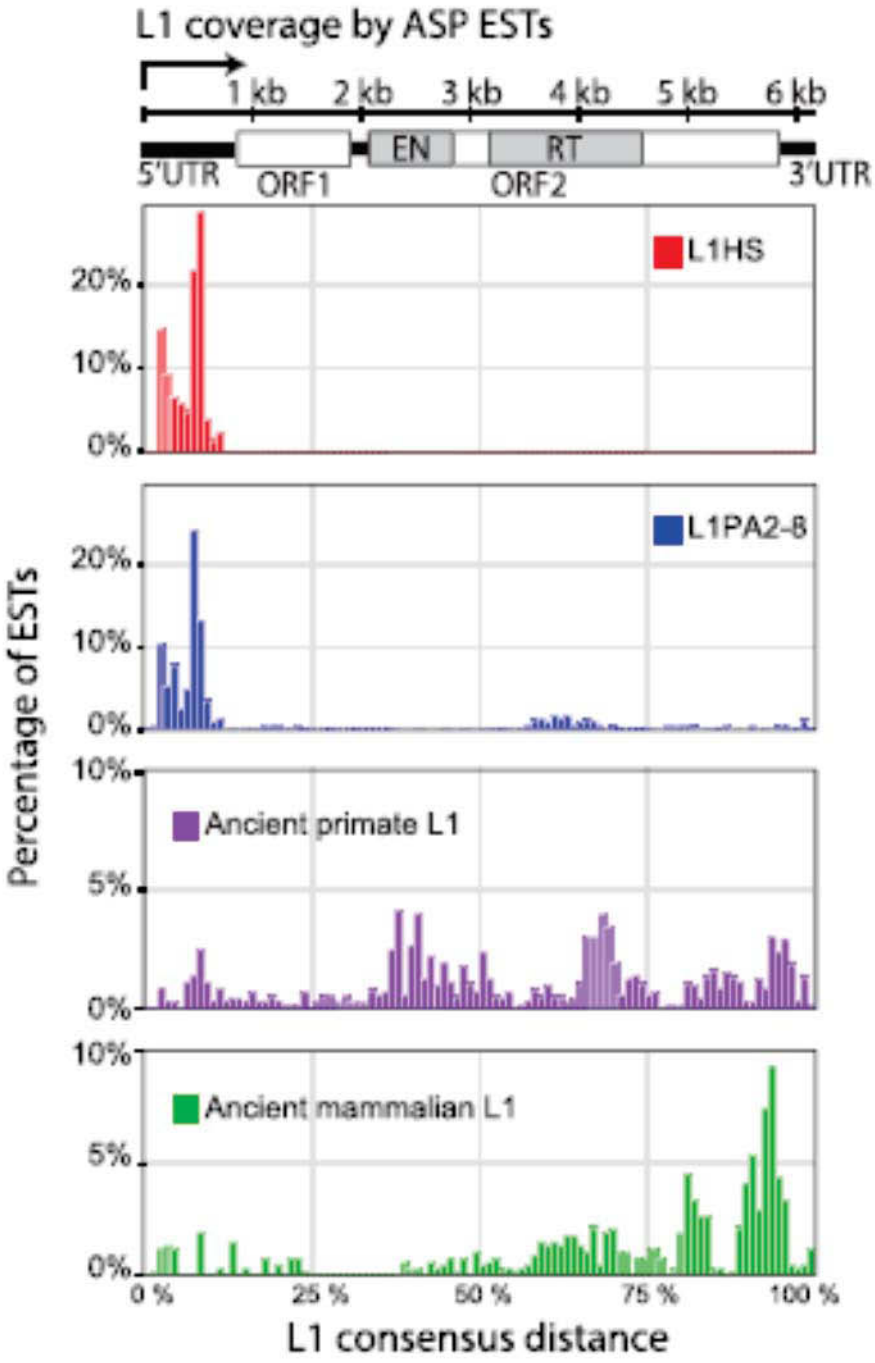


Figure 58 : L'ASP un promoteur alternatif pour les sous-familles L1PA1 à PA8 ?

Le % d'EST débutant dans chaque région du L1 a été rapportée selon l'origine du L1 associé. Aussi pour les L1HS, le début des EST (TSS supposé) est localisé dans la région 5'UTR de L1. Il en est de même pour les sous-familles L1PA2 à 8 avec une légère détection au niveau de la région ORF2. En revanche pour les sous-familles L1PA anciennes et L1MA anciennes, il y a très peu d'EST détectées dans la région 5'UTR, mais plutôt au niveau de l'ORF2 et de la région 3'UTR.

Tumor type	RNA-seq	nbre lectures	nbre chim uniq	nbre moy chim / échant	nbre lecture max ds 1 ech pour 1 chim	nbre moy lectures/ech +
Ov Metas	polyA	40m	105	16,75	5	1,28
MCF7	polyA	135m	125	76,5	15	1,74
Gliomes	Total RNA	90m	2675	723,75	122	1,67

Table 2 : Comparaison des données des 3 RNA-seq de gliomes, de métastases ovariennes et de la lignée MCF7.

Le nombre de chimères unique détectées avec le séquençage Ribozéro (2675) est plus important par rapport aux données de séquençage polyA. Ces résultats obtenus avec CLIFinder dépendent de la profondeur du séquençage qui est de 40 millions de lectures dans les métastases ovariennes (Ov meta), de 135 millions dans la lignée MCF7 (MCF7) et de 90 millions pour les gliomes (Gliomes). Ces résultats dépendent également du type d'ARN séquençé. Le séquençage Ribozéro semble plus exhaustif, mais le séquençage polyA apportent des informations plus pertinentes pour l'identifiaction de chimères avec un rôle fonctionnel.

d'un L1. Certaines de ces chimères correspondent à des LCT ayant subi un mécanisme d'épissage. Il serait intéressant de valider ces chimères et de voir si celles-ci peuvent jouer un rôle au cours du processus de tumorigenèse.

CLIFinder : une identification pangénomique des LCT ? RNA-seq Ribozero versus polyA ?

Le logiciel CLIFinder a été développé dans le but d'apporter une vision pangénomique exhaustive de la transcription initiée à l'ASP de L1 en conditions normale et tumorale.

L'analyse des RNA-seq de gliomes selon les paramètres de l'analyse A37 (30pb correspondant à la région 5'UTR de L1 en R1 avec 6mm) en utilisant les séquences consensus de 27 sous-familles de L1 identifie 2675 chimères parmi lesquelles 2256 impliquent des L1 récents. Ce chiffre est beaucoup plus important que les 250 LCT décrits dans la littérature. La réalisation d'analyses par sous-famille avec 0, 1, 2 mm détecte respectivement 3960, 4349 et 4415 chimères uniques. Ceci montre une différence importante du nombre de chimères identifiées par rapport à l'analyse A37. Pour autant, cette augmentation est due à l'identification de nouvelles chimères impliquant majoritairement des sous-familles antérieures à L1PA8. Or les analyses de Criscione ont montré que les LCT-EST impliquant ces sous-familles (antérieures à L1PA8) ont une initiation de leur transcription en antisens non pas à partir de la région 5'UTR mais à partir des régions ORF2 et 3'UTR (Figure 58). Ainsi, les résultats des analyses A37 et par sous-famille, démontrant un nombre équivalent de chimères impliquant les L1 récents, les chimères identifiées dans nos RNA-seq Ribozero par l'analyse A37 semblent être exhaustifs.

Comme nous ne savions pas si les LCT étaient polyadénylés ou non, les RNA-seq des gliomes ont été réalisés sur les ARN totaux déplétés en ARN ribosomaux pour un séquençage sans *a priori* de tous les transcrits. Le nombre de lectures totales obtenues dans les RNA-seq de gliomes est de 90 millions par échantillon. Par ailleurs, des données publiques de RNA-seq ont été analysées. Ces RNA-seq paired-end orientés ont été réalisés sur les ARN polyadénylés extraits de métastases ovariennes (8 échantillons) et de la lignée cellulaire d'adénocarcinome mammaire MCF7 (2 réplicas) dont le nombre moyen de lectures obtenues par échantillon est de 40 et 135 millions respectivement. Après utilisation du logiciel CLIFinder sur ces différents jeux de données (en utilisant les conditions d'analyse A37), 2675 chimères uniques sont identifiées dans nos 13 gliomes, 105 dans les métastases ovariennes et 125 dans la lignée MCF7. Le nombre moyen de chimères détectées par échantillon est de 723,75 pour les gliomes, 16,75 pour les métastases ovariennes et 76,5 pour MCF7. Le nombre de lectures maximal dans un échantillon pour 1 chimère est de 122 pour les gliomes, 5 pour les métastases ovariennes et 15

pour MCF7. Les plus grands nombres de chimères et de lectures détectées pour 1 chimère sont donc obtenus dans les gliomes provenant de séquençage des ARN totaux (Table 2). Pour autant, le nombre de lectures moyen par échantillon positif pour une chimère est de 1,67 pour les gliomes, 1,28 pour les métastases ovariennes et 1,74 pour MCF7. Ceci signifie que CLIFinder détecte des chimères de manière identique dans les 3 jeux de données. Ce qui différencie ces analyses (gliomes, métastases ovariennes et MCF7) sont :

- La profondeur de séquençage. En effet, le séquençage réalisé sur les ARNm extraits de la lignée MCF7 est 3 fois plus profond que celui réalisé sur les ARNm extraits des métastases ovariennes. Que ce soit le nombre maximal de lecture dans un échantillon (15 vs 5) ou le nombre moyen de chimères détectées par échantillon (76,5 vs 16,75), ces valeurs sont entre 3 et 4 fois plus élevées dans les données de MCF7 (135 millions de lectures) que dans les données de métastases ovariennes (40 millions de lectures). De même, le nombre de chimères détectées est supérieure (125 vs 105) dans les 2 réplicas MCF7 *vs* l'analyse des 8 métastases ovariennes, ceci semble être dû à la profondeur de séquençage plus importante.
- Le type d'ARN séquencés. L'utilisation des ARN polyadénylés, permet le séquençage seulement des transcrits messagers. Le séquençage des ARN totaux après déplétion des ARNr (évite de polluer le séquençage avec ces séquences fortement transcrites) permet également de séquencer les ARN messagers polyadénylés mais aussi les ARN non codants qui ne sont pas polyadénylés, voire les ARN pré-messagers. Ceci explique sans doute la grande différence des nombres de chimères détectées dans les données de RNA-seq des gliomes *versus* ceux détectés dans les MCF7 et les métastases ovariennes. Il serait intéressant de comparer les données de séquençage Ribozéro avec les données de séquençage polyA réalisé sur les mêmes échantillons. Une étude en cours par ailleurs dans l'équipe vient de séquencer les ARNm de cellules initiatrices de gliomes d'un patient en RNA-seq paired-end orientés de 100 pb à chaque extrémité. Or, les ARN totaux des mêmes cellules ont été analysés par séquençage RNA-seq Ribozéro paired-end orienté de 100 pb à chaque extrémité. La comparaison des listes de chimères obtenues par les deux types de séquençage sur le même échantillon devrait nous permettre de mieux caractériser le différentiel de détection de CLIFinder entre les deux approches.

L'exhaustivité des chimères identifiées par CLIFinder semble donc reposer largement sur le type d'ARN séquencé ainsi que sur le nombre d'échantillons analysés pour un même type tumoral et sur la profondeur de séquençage.

Enfin, il est à noter que CLIFinder identifie des transcrits chimères à partir du génome humain de référence. Or, dans les tumeurs des réarrangements génomiques (duplication, délétion, translocation…) ainsi que des néo-transpositions somatiques sont observés. Ceux-ci peuvent donc être à l'origine de l'expression de transcrits chimères spécifiques du génome tumoral. Dans ce cas-là, les lectures correspondantes seront retenues par les filtres 1 et 2 de CLIFinder mais le logiciel les éliminera ensuite car il ne pourra pas les aligner correctement en un locus du génome normal de référence. Pour identifier de telles chimères, il faudrait donc pouvoir disposer, pour l'étape d'alignement, du génome de chaque échantillon tumoral analysé en RNA-seq. Cependant, bien que des réarrangements et des transpositions *de novo* puissent être récurrents et caractéristiques de certains types tumoraux, la majorité d'entre eux vont être spécifique d'une tumeur. Aussi, l'impossibilité de CLIFinder à détecter des LCT spécifiques d'une tumeur semble correspondre à une limite mineure par rapport à l'objectif de caractériser des LCT récurrents (exprimés dans plusieurs tumeurs, voire dans différents types tumoraux) pouvant jouer un rôle dans les processus tumoraux.

CLIFinder est donc un logiciel qui permet l'identification de nombreux transcrits chimériques au sein d'un échantillon tumoral. L'utilisation de séquençage ARN Ribozéro tend à identifier beaucoup de chimères et est plus exhaustif. La détection de chimères dans les données de séquençage Ribozéro semble plus pangénomique mais des chimères faux-positifs peuvent être identifiées, comme des chimères provenant d'ARN pré messager où le L1 sera éliminé de la séquence du transcrit épissé et ne correspondront pas à la définition d'un LCT.

De plus, ces chimères ne sont pas toutes pertinentes pour un rôle fonctionnel au cours de la tumorigenèse. En effet, de nombreuses difficultés ont été rencontrés au cours de mes travaux afin de valider des candidats fonctionnels pleine taille. Peut-être que l'utilisation d'un séquençage polyA aurait permis de détecter plus rapidement des transcrits récurrents d'intérêt fonctionnel.

Aussi, un séquençage ARN polyA serait plus judicieux afin de répondre à cette question. En effet, le séquençage polyA permet d'identifier des ARNm pouvant être traduit en protéine, qui peuvent ainsi perturber l'homéostasie cellulaire et ainsi participer au développement tumoral.

CLIFinder : un outil bio-informatique adaptable.

Comme nous venons de le voir, CLIFinder est un outil pertinent permettant d'identifier des transcrits chimères initiées à partir de l'ASP de L1 dans des lectures de RNA-seq et ce, de façon plus exhaustive et plus précise (par type tumoral) que les approches précédemment utilisées. De plus, cet outil se présente comme un outil adaptable.

Premièrement, cet outil a été développé pour analyser préférentiellement des données de RNA-seq paired-end orientés. Ce type de séquençage permet d'obtenir une indication sur le sens de la transcription. L'outil CLIFinder prend en compte alors :

- Si le R1 est séquencé dans le sens du transcrit, alors R1 est considéré comme la partie contenant la séquence L1.
- Si le R1 est séquencé en antisens du transcrit, alors ce sera la séquence R2 qui sera assignée comme devant contenir la séquence L1.

Toutefois, la possibilité est laissée à l'utilisateur d'utiliser le logiciel pour analyser des données paires-end non orientées. Dans ce cas, les deux hypothèses énoncées ci-dessus seront prises en compte. Le risque d'une telle analyse sera alors d'identifier des faux positifs correspondant notamment à des évènements de transcription (en sens du L1) qui incluent les séquences promotrices de L1.

Un autre niveau d'adaptabilité concerne les séquences utilisées pour le filtre 1. Ce premier filtre permet de sélectionner les lectures dont le R1 contient une séquence de L1. Dans une première intention, les séquences consensus de 27 sous-familles de L1 (Khan, 2005) ont été utilisées en tolérant 6 mésappariements dans l'alignement d'au moins 30 pb consécutives. Dans une deuxième intention, une autre approche a été de modifier les séquences de référence du filtre 1 afin de valider si l'enrichissement observé en sous-familles récentes (L1PA1 à 7) ne provenait pas d'un biais du à l'utilisation des séquences consensus avec 6 mésappariements. Aussi les séquences des L1 présents dans tout le génome humain ont été extraites pour chaque sous-famille et les RNA-seq de gliomes ont été analysée avec 0, 1 et 2 mésappariements. Les résultats montrent que les analyses par sous-famille augmentent le nombre de chimères détectées dans les sous-familles les plus anciennes (antérieures à L1PA8). En revanche, 92% des chimères impliquant des L1 récents (L1PA1 à 7) sont retrouvées dans l'analyse par sous-famille par rapport à l'analyse A37 utilisant les séquences consensus. La transcription identifiée dans les sous-familles les plus anciennes provient de régions dans l'ORF2 ou le 3'UTR (Criscione *et al.*, 2016), et ne semble pas provenir d'un ASP. Ainsi, les sous-familles les plus récentes censées posséder l'ASP sont majoritairement détectées quel que soit le type d'analyse réalisée.

Enfin, CLIFinder peut s'adapter à l'identification de transcrits chimères issus d'autre type d'ET. Les séquences de L1 peuvent être remplacées par des séquences d'éléments LTR. En effet, il a été décrit un promoteur antisens dans la région 3'LTR à partir de laquelle la région génomique adjacente peut être transcrite, à l'image de ce qui a été décrit pour les LCT. Une transcription en antisens a été identifiée dans la région 3'LTR. Grâce à la transfection de cellules HEK293T transfectées avec un plasmide permettant la transcription en antisens à partir du promoteur 3'LTR de l'EGFP et permettant la transcription en sens à partir du promoteur 5'LTR de DsRed. La fluorescence des cellules est quantifiée par FACS et celle-ci indique une transcription à partir des 2 promoteurs dans ces cellules (Laverdure *et al.*, 2016). Il peut ainsi être envisagé que CLIFinder soit utilisé afin d'identifier des transcrits chimères initiées à partir du promoteur antisens en 3'LTR dans les données de RNA-seq.

2. La dérégulation transcriptionnelle liée aux LCT dans les tumeurs implique une surexpression liée au contexte du locus.

Une surexpression en contexte tumoral pour nombre de LCT identifiés par CLIFinder

L'analyse du nombre de lectures des chimères identifiées montrent que celles-ci sont majoritairement détectées dans les tumeurs (92%). De plus, 40% d'entre elles sont détectées dans les tumeurs mais aussi dans les contrôles. Ceci est en accord avec ce qui a été décrit dans la littérature à savoir que des LCT sont détectées dans les échantillons tumoraux mais aussi normaux, à l'image du LCT8 et LCT9 décrit par Cruickshanks (Cruickshanks & Tufarelli, 2009).

J'ai pu valider par des analyses de biologie moléculaire 56 LCT dont l'expression a été quantifiée par RT-qPCR sur une cohorte plus large. Tous les LCT, y compris ceux identifiés par CLIFinder comme « Tum spé », sont retrouvés exprimés dans tous les tissus, non seulement dans les tumeurs mais aussi dans les échantillons contrôles. Il existe donc une transcription basale provenant de l'ASP dans les tissus contrôles. Ceci indique que la méthylation de la région promotrice des L1 ne semble pas réprimer l'activité transcriptionnelle de l'ASP. En contexte tumoral, une surexpression de certains LCT est observée et ce, plus particulièrement pour les LCT initialement détectés « Tum spé » (66%) et pour les LCT « C+T » présentant un nombre de lectures/échantillon au moins 2 fois supérieur à celui dans les contrôles (75%). Ainsi, aucun des LCT testés ne présente une expression tumeur spécifique. Cette notion, initialement suggérée par les résultats de CLIFinder, est en partie liée au fait que dans notre étude seulement 3 échantillons contrôles ont été analysés contre 13 gliomes. De plus, CLIFinder identifie des

transcrits chimères dans les données de RNA-seq, uniquement sur la base de la lecture d'un ou plusieurs fragments de la banque localisés précisément au niveau de la zone de jonction L1/séquence unique du transcrit LCT. Il était donc prévisible que CLIFinder ne permette pas de quantifier l'expression d'un transcrit LCT de façon aussi précise qu'une analyse transcriptomique classique, basée sur le nombre de lectures réparties sur l'ensemble de la séquence d'un gène. Dans tous les cas, mes résultats de RT-qPCR, sur une cohorte plus large de gliomes et tissus contrôles, démontrent l'existence d'une augmentation de l'activité transcriptionnelle de l'ASP en contexte tumoral.

Pour 56 LCT, nous disposions à la fois de l'expression détectée par CLIFinder (nombre de lectures) et en RT-qPCR dans 15 échantillons. La mise en relation de ces deux types de données montre que 0 lecture détectée par CLIFinder ne signifie pas une absence d'expression (Figure 47). Comme discuté précédemment, ceci s'explique de par le principe de fonctionnement de CLIFinder. Toutefois, la mise en relation des données de CLIFinder et RT-qPCR montre également une corrélation positive significative avec un coefficient Rho de 0,47. Ce coefficient de corrélation modéré suggère que, bien que ne permettant pas d'appréhender de façon aussi précise l'expression des LCT que la RT-qPCR, l'expression détectée avec CLIFinder est au moins semi-quantitative. Aussi, afin de mieux décrire à l'échelle pangénomique le pourcentage de chimères présentant une surexpression tumorale, un modèle de prédiction est en cours d'établissement. Celui-ci se base sur la mesure des tailles d'effet. Celles-ci ont été calculées pour les 56 LCT étudiés à la fois en Fluidigm et avec CLIFinder dans 15 échantillons. Des analyses sont en cours afin de déterminer une valeur de taille d'effet seuil, présentant le meilleur compromis entre sensibilité et spécificité. L'analyse se focalisera ensuite sur les chimères « Tum spé » dont on a montré qu'un fort pourcentage présente une surexpression significative. Elle consistera à calculer les tailles d'effet pour chacune de ces chimères (n=1397) et la valeur seuil sera appliquée afin de prédire le pourcentage de chimères « Tum spé » surexprimées dans les gliomes.

La surexpression des LCT en contexte tumoral semble impliquer une augmentation d'activité transcriptionnelle de la région génomique avoisinante

L'analyse de l'expression des LCT sur une cohorte plus large de gliomes montre que ces LCT sont exprimés dans tous les échantillons, non seulement dans les contrôles mais aussi dans les tumeurs. Les LCT détectées comme « Tum spé » avec CLIFinder ou « ToverC » sont

majoritairement significativement surexprimées dans les tumeurs par rapport aux contrôles. Aussi la question s'est posée du mécanisme impliqué dans cette surexpression ?

La première hypothèse qui a été décrite dans la littérature est que la diminution de la méthylation observée dans les tumeurs induit l'activation de la transcription à l'ASP du L1.

Mes résultats montrent qu'une diminution de méthylation ADN n'est pas retrouvée au niveau de la région promotrice des L1 associée à des LCT surexprimés. Aussi, l'hypométhylation du promoteur des L1 dans les tumeurs ne semble pas expliquer l'augmentation de l'activité transcriptionnelle à partir de l'ASP.

La deuxième hypothèse émise est que cette surexpression peut être due à une dérégulation transcriptionnelle au locus.

Ainsi, l'expression des gènes dans lesquels se trouvent les LCT est corrélée positivement à la surexpression des LCT dans les tumeurs, quel que soit le sens de transcription du LCT par rapport à celui du gène associé. Cette observation suggère qu'il existerait une dérégulation transcriptionnelle aux loci des LCT sur- et sous-exprimés. Ainsi, dans les tumeurs, la sur- ou sous-expression d'un LCT pourrait être associée à une modification de la chromatine au locus. En contexte tumoral, l'augmentation, observée de façon concomitante, de la transcription à partir de l'ASP et du promoteur du gène associé pourrait être alors le reflet de l'acquisition d'une chromatine plus permissive, facilitant l'accès de l'ADN du locus à la machinerie de transcription. Ceci permet de proposer que l'activité transcriptionnelle de l'ASP d'un L1 pourrait être influencée par celle du locus.

Comment cette dérégulation transcriptionnelle au locus peut-elle avoir lieu ?

Afin de déterminer les bases mécanistiques de cette dérégulation, l'analyse des bases de données par une approche de datamining, des marques d'histones spécifiques soit de la chromatine transcrite comme H3K36me3 soit des marques d'histones activatrices comme H3K4me3 est envisagée au niveau des région ASP et dans la séquence unique adjacente des 56 LCT (qu'ils soient intra- ou intergéniques et qu'ils soient surexprimés ou non). Des données de ChIP-seq obtenues dans des gliomes et des tissus cérébraux sains sont disponibles (Rheinbay *et al.*, 2013). La comparaison des taux d'enrichissement entre les tissus cérébraux sains et les gliomes pour les 56 LCT permettra de voir si ces régions sont anormalement enrichies dans les gliomes, spécifiquement au loci des LCT surexprimés, attestant d'un état transcriptionnel permissif de la chromatine.

De plus, cette surexpression en contexte tumoral pourrait être due à des évènements de duplication du génome tumoral. Grâce à l'analyse de puces CNV qui ont été réalisées par

ailleurs sur nos échantillons de gliomes et qui sont en cours d'analyse, celles-ci pourraient indiquer si un évènement de duplication a eu lieu dans la région des loci surexprimés.

Dans une étude s'intéressant à la signature moléculaire des L1HS transcrits, Philippe et ses collaborateurs ont déterminer l'activité transcriptionnelle des L1HS dans des données de RNA-seq en utilisant une propriété de ceux-ci. En effet, pour une fraction de L1 transcrits une faiblesse au niveau du signal de polyadénylation en 3'UTR permet la transcription de la séquence unique adjacente. Aussi les auteurs ont quantifié dans leurs données de RNA-seq, ces transcrits L1 ayant un R1 localisé en 3' du L1 (en sens) et un R2 dans la séquence unique adjacente. Via cette mesure de l'activité transcriptionnelle des L1 et de la mesure de lectures provenant de l'ASP, ils démontrent que les activités promotrices du promoteur sens et de l'ASP d'un L1HS pleine taille ne sont pas nécessairement couplées à leur environnement chromosomique (Philippe *et al.*, 2016). Cette méthode pourrait être adaptée à nos données de RNA-seq afin de déterminer l'activité transcriptionnelle en sens des L1 impliqués dans la formation de nos chimères.

3. Rôle fonctionnel des LCT.

Au final, 2 LCT candidats pour un rôle fonctionnel ont été retenus : les LCT1873 et LCT1994 qui sont tous deux surexprimés dans les gliomes (HGG+LGG et HGG, respectivement) et possèdent un rôle de biomarqueur vis-à-vis de la survie des patients. Aussi il semble important d'identifier la séquence pleine taille de ces LCT grâce à des expériences de 5' et de 3' RACE. La connaissance de la séquence complète de chaque LCT permettra d'appréhender les gènes cibles pouvant être dérégulés par la surexpression de ces LCT.

Le LCT1994 est localisé en sens dans l'intron 1 du gène *CCCD109B*. D'après la littérature (Wolff *et al.*, 2010) et de par la présence d'un codon ATG dans l'exon 3, j'ai émis l'hypothèse que celui-ci pourrait coder une protéine tronquée (279AA au lieu de 336AA) pouvant aboutir à un défaut de localisation cellulaire suite à la perte du signal d'adressage à la mitochondrie et ainsi perturber l'homéostasie calcique mitochondriale. Pour vérifier cette hypothèse, une approche de western blot est envisagée (grâce à 2 anticorps dirigés contre l'extrémité C-terminale). Celle-ci sera réalisée sur des protéines extraites des échantillons de gliomes surexprimant le plus le LCT1994, dans le cas où la 3'RACE aura confirmé la présence des exons 3 à 8 dans la séquence pleine taille du LCT.

Le LCT1873 est localisé en sens dans l'intron 14/15 du gène *GOLIM4*. A l'image de ce qui a été décrit pour le LCT13 (Cruickshanks *et al.*, 2013), l'hypothèse d'un long transcrit non codant

permettant la mise en place de marques d'histones répressives induisant la mise sous-silence de 2 gènes localisés en aval est envisagée. En effet, le LCT1873 est localisé à 188 Kb de 2 gènes en orientation opposée sous la dépendance d'un seul promoteur bidirectionnel. A l'image de ce qui a été décrit dans la littérature pour ces 2 gènes *PDCD10* et *SERPINI1* (Chang *et al.*, 2000; Lambertz *et al.*, 2015), leur expression est diminuée dans nos gliomes. De plus, les résultats sur les puces de méthylation HM450K montrent que la méthylation de la région promotrice des gènes *SERPINI1* et *PDCD10* n'est pas modifiée entre les tumeurs et les contrôles. Aussi la diminution d'expression du gène *SERPINI1* observée dans les tumeurs n'est pas due à un gain de méthylation au niveau de sa région promotrice. Le gène *SERPINI1*, antisens au LCT, montre notamment une corrélation inverse significative entre la surexpression du LCT et la diminution d'expression de ce gène dans les gliomes (HGG+LGG). Ce gène a une expression spécifique au cerveau et code pour une protéase inhibant les serpines. Aussi afin de déterminer le rôle fonctionnel du LCT1873 des expériences de transfections cellulaires pourraient être réalisées. Après s'être assuré que les cellules à transfecter n'expriment pas le LCT1873, comme les cellules MO3-13 correspondant à des oligodendrocytes normaux où *SERPINI1* est exprimé, un plasmide surexprimant le LCT1873 pleine taille pourrait être transfecté afin de voir si celui-ci induit :

- La diminution d'expression de *SERPINI1*, quantifiée par RT-qPCR.
- La déposition de marques d'histones répressives, comme les marques répressives H3K9me3 et H3K27me3, pouvant être impliquée dans la diminution de l'expression de *SERPINI1*, grâce à la réalisation de ChIP.

Au cours de mes travaux de thèse, l'analyse par CLIFinder des données de RNA-seq de métastases ovariennes et de MCF7 issus de bases de données a détecté des chimères communes à différents types tumoraux. Il serait intéressant de continuer de rechercher dans les bases de données des RNA-seq paired-end orientés d'autres types tumoraux afin d'identifier des LCT récurrents qui pourraient jouer un rôle fonctionnel au cours de la transformation tumorale.

4. Le séquençage PacBio ciblé : Une stratégie alternative ?

Comme nous l'avons vu, de nombreuses difficultés ont été rencontrées pour la validation de candidats fonctionnels. En effet, il est impossible à partir des données de RNA-seq paired-end Illumina de reconstituer *de novo* la séquence pleine taille des transcrits LCT. Pour autant, la connaissance de la séquence pleine taille des LCT serait un élément facilitant pour mieux

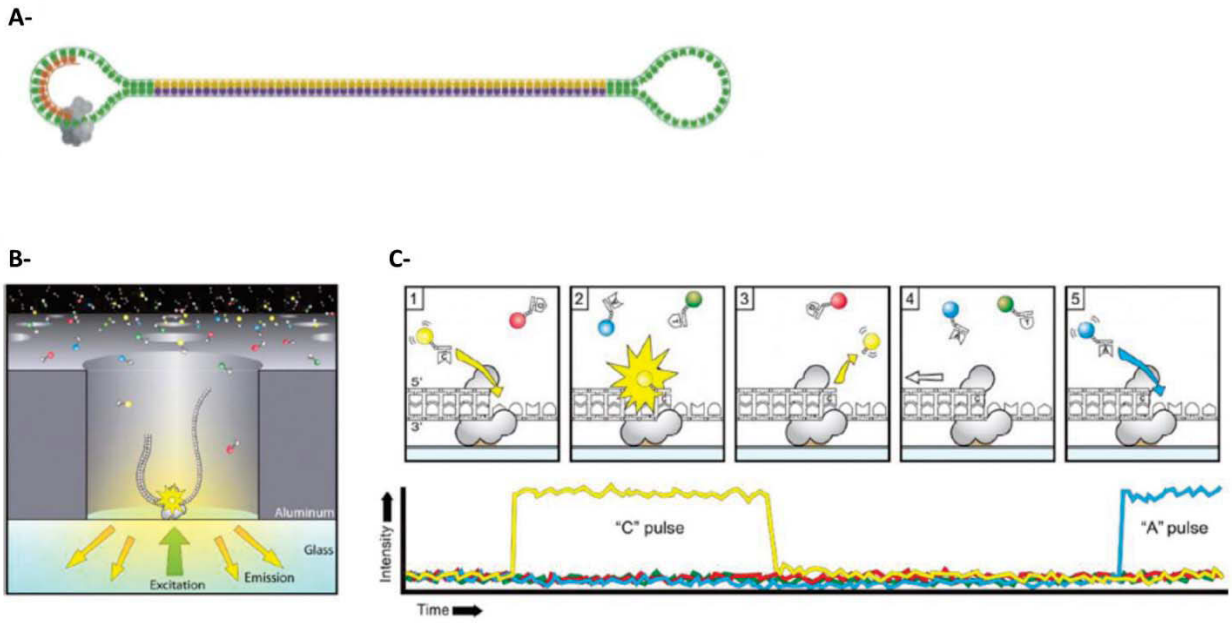


Figure 59 : *Fonctionnement du PacBio*

A- SMRTbell : les adaptateurs en épingle à cheveux (en vert) sont liés aux extrémités de l'ADN double brin (jaune et violet) et forment un cercle fermé. La polymérase (en gris) est ancrée à l'unité de séquençage ZMW (Zero-Mode Waveguide) et incorpore les nouvelles bases (en orange).

B- La cible SMRTbell diffuse dans l'unité de séquençage ZMW et les adaptateurs se lient à la polymérase.

C- Incorporation des 4 nucléotides fluorescents (représentés en rouge, en jeune, en vert et en bleu, respectivement pour G, C, T, A) qui ont des spectres d'émission de fluorescence. Le nucléotide fluorescent est associé au brin cible au niveau du site actif de la polymérase (1). La fluorescence alors émise correspond à celle de la base incorporée (2). Le fluorochrome lié au nucléotide par le pyrophosphate est clivé (3). La polymérase continue la réplication en passant au nucléotide suite (4). Incorporation du nucléotide suivant (5).

appréhender un rôle fonctionnel potentiel. Aussi l'approche de séquençage PacBio semble constituer une alternative qui nous permettrait de répondre à différentes questions.

Le séquençage PacBio est une nouvelle technologie permettant de séquencer des fragments de grande taille. Cette technologie se base sur du séquençage en temps réel de molécule simple (SMRT : single-molecule real-time). Ce séquençage fait l'acquisition de la séquence pendant le processus de réplication de la molécule d'ADN cible. Cette cible est appelée SMRTbell et correspond à une molécule d'ADN circulaire qui a été créée par ligation d'adaptateurs en épingle à cheveux aux 2 extrémités de la molécule d'ADN cible double brin (Figure 59-A). Après la ligation, l'échantillon SMRTbell est déposé sur la cellule de séquençage, nommée SMRTcell. Le séquençage va être assuré par une polymérase immobilisée au fond du puit qui peut se lier aux adaptateurs en épingle à cheveux du SMRTbell et commencer la réplication. Cette unité de séquençage est appelée ZMW (Zero-Mode Waveguide). Grâce aux 4 nucléotides fluorescents (A, T, G, C) présents dans le milieu réactionnel, un spectre d'émission de fluorescence sera généré à chaque incorporation de nucléotides. Cette acquisition permet d'obtenir des « films » de pulsations lumineuses qui peuvent durer entre 0,5 et 4 heures, dont le traitement permet d'obtenir la séquence nucléotidique de la cible. Comme la SMRTbell a une forme circulaire, après la réplication du 1er brin de l'ADN cible double brin, la polymérase continue à incorporer les bases correspondant à l'adaptateur puis réplique le second brin (Figure 59-B et -C). La molécule d'ADN pourra être lue plusieurs fois, ce qui génèrera des lectures multiples. Un des avantages du PacBio est que celui-ci peut séquencer des fragments dont la taille peut aller jusqu'à 10 Kb (Rhoads & Au, 2015). Aussi cette technologie devrait nous permettre de séquencer des transcrits LCT pleine taille.

Afin d'enrichir nos échantillons de gliomes en LCT, il est envisagé de coupler ce séquençage PacBio à une étape préliminaire de capture. A partir des séquences consensus de L1, des sondes vont être dessinées grâce au logiciel KASpOD (Gasc & Peyret, 2017). Celui-ci appliqué aux séquences promotrices des L1 des sous familles récentes a permis de dessiner 4 sondes (80mer) couvrant la région promotrice des L1 (450 pb). L'étape de capture réalisée avec ces sondes permettra d'enrichir nos échantillons de gliomes en transcrits contenant la séquence promotrice de L1. Ensuite une banque d'ADNc SMRTbell pleine taille sera construite par échantillon pour être séquencée. Ces expériences nous permettront d'identifier des LCT pleine taille, pouvant jouer un rôle fonctionnel dans la tumorigenèse. La limite de cette approche est qu'elle ne permettra l'étude que des LCT polyadénylés, la construction de la banque ADNc pleine taille

utilisant un amorçage avec une amorce oligodT ciblant les ARNm polyA. Les avantages de cette expérience pour notre étude sont :

1) D'identifier des transcrits pleine taille et ainsi préciser le site d'initiation de la transcription des LCT qui est supposée à l'ASP. Ceci nous affranchira des expériences de 5'RACE qui sont difficiles à mettre en œuvre sur nos échantillons et de valider un plus grand nombre de LCT.
2) D'identifier précisément la composition de la séquence des LCT. Ainsi, des LCT ayant subi des phénomènes d'épissage pourront être détectés dans nos échantillons.
3) De nouvelles cibles pourront être mises en lumière.

Ces LCT ainsi identifiés seront comparés avec les LCT obtenus avec CLIFinder sur les données de RNA-seq RiboZero Illumina. Nous disposons de quantités suffisantes d'ARN pour pouvoir étudier en PacBio 8 gliomes sur les 13 analysés précédemment.

Cela validera, par exemple pour le LCT1873, si celui-ci code un long transcrit pouvant induire la mise sous silence du gène *SERPINI1*, ou pour le LCT1994, si celui-ci possède dans sa séquence les exons 3 à 8 du gène *CCDC109B*.

De nouvelles cibles pourront être détectées. Certaines pourront avoir subi un évènement d'épissage induisant la caractérisation de nouveaux variants d'épissage dont l'initiation a lieu à l'ASP d'un L1. Ces transcrits pourront être traduits en protéines pouvant aboutir à des effets gain de fonction, à l'image de ce qui est supposé pour le transcrit *L1-MET* (Wolff *et al.*, 2010). Des LCT correspondant à des longs transcrits pouvant aboutir à la mise sous silence de gènes cibles pourront être détectés à l'image de ce qui a été décrit pour le LCT13 (Cruickshanks *et al.*, 2013).

En conclusion, mes travaux de thèse ont permis de valider un nouvel outil bio-informatique permettant de détecter au niveau pangénomique des LCT dans des données de RNA-seq. J'ai montré que 25% des éléments L1 des sous familles récentes du génome humain sont associés à la production de LCT dans les gliomes et les tissus cérébraux contrôles. Sur les 56 LCT analysés plus en avant, tous sont retrouvés exprimés à un niveau basal dans les tissus normaux et en contexte tumoral la dérégulation transcriptionnelle observée passe principalement par une surexpression. Cette surexpression semble alors dépendre, non pas de la diminution de méthylation ADN de la région promotrice des L1, mais plutôt d'une augmentation de l'activité transcriptionnelle de la région génomique avoisinante de l'ASP. Des études complémentaires restent à réaliser pour mieux étayer cette hypothèse. Enfin, parmi ces LCT validés, 2 ont été

retenus pour un rôle fonctionnel possible au cours du développement tumoral. Les analyses sont à poursuivre pour ces deux candidats mais l'identification de LCT candidats fonctionnels pertinents reste très difficile à partir des seules données de RNA-seq Illumina. Aussi, une approche alternative combinant une étape de capture des transcrits LCT suivie d'un séquençage PacBio est envisagée. Ceci permettrait d'obtenir des séquences de transcrits pleine taille permettant d'assigner pour chacun le site d'initiation de la transcription (ASP ou pas) et la composition de la séquence. La connaissance de cette dernière sera utile pour identifier plus facilement des LCT candidats pertinents et pouvant jouer une fonction dans les processus de transformation tumorale.

ANNEXES

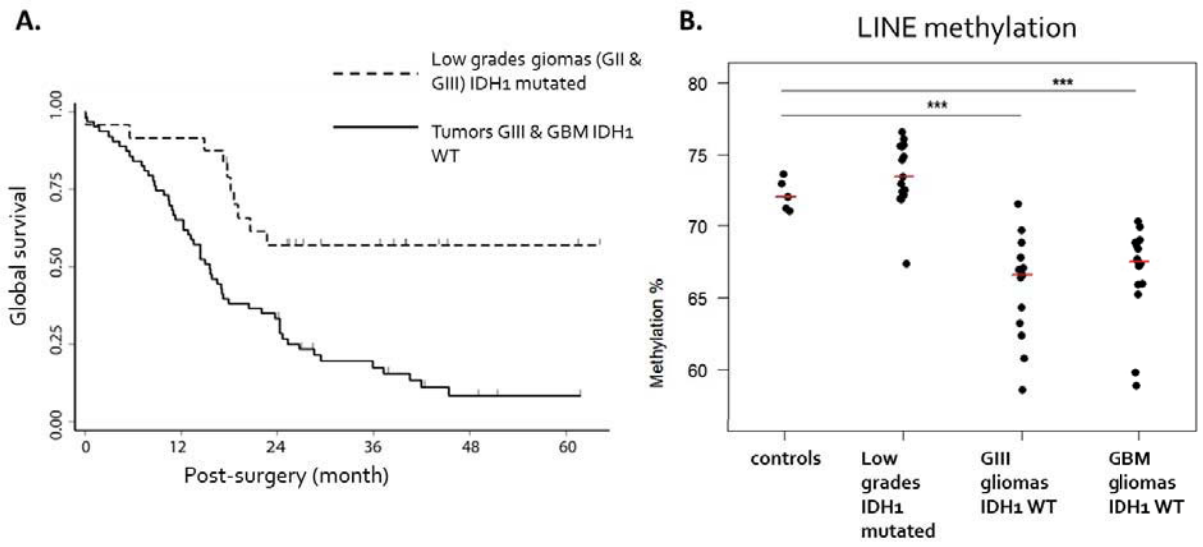


Figure 60 : Caractéristiques des tumeurs de bas grades et de haut grade selon le statut de mutation IDH1.

A- Deux populations de patients bien distinctes sont définies en termes de survie selon le statut de mutation du gène *IDH1* dans les gliomes. La courbe de Kaplan-Meier met en évidence que la survie des patients présentant un gliome *IDH1* muté est meilleure (Courbe en pointillé) comparé à celle des patients présentant une tumeur *IDH1* sauvage (trait plein).

B- Pourcentage de méthylation obtenus au niveau des sondes localisées dans les régions de L1 à partir des données extraites des puces de méthylation HM450K. Une diminution de la méthylation est présente dans les HGG (GIII et GBM *IDH1* WT) par rapport aux LGG (*IDH1* mutés) et aux contrôles.

1. Matériels et méthodes.

1.1. Tumorothèque de gliomes et échantillons contrôles.

Grâce à une collaboration avec l'équipe du Pr. P. Verrelle (Centre Jean Perrin – UCA CREaT EA7283), ce projet s'appuie sur la Tumorothèque Auvergne Gliomes (n°DC-2012-1584) qui inclue actuellement 87 gliomes. Après signature d'un consentement éclairé par le patient, les prélèvements sont anonymisés. Une analyse par l'anatomopathologiste va permettre d'établir le grade tumoral et ne conserver pour la recherche que les échantillons tumoraux présentant un pourcentage de cellules tumorales supérieur à 50%. Pour ces échantillons, en plus du grade histologique (selon les critères de l'OMS), les données clinico- biologiques des patients sont rapportées incluant la survie des patients, le statut de mutation de l'*isocitrate déshydrogénase 1* (*IDH1*) et pour 53 tumeurs du profil de méthylation de l'ADN tumoral, obtenu par analyses sur puce Illumina HM450K (Annexe 22). En accord avec les données de la littérature, les analyses biostatistiques de cette population de gliomes ont montré que, bien plus que le grade, le statut muté ou non d'*IDH1* est un facteur pronostic puissant. Ainsi, lorsque la survie des patients présentant des tumeurs *IDH1* mutés (LGG) est comparée à celle des patients présentant des tumeurs *IDH1* sauvages (ou Wild-Type, WT) (HGG), une survie beaucoup plus longue est observée pour les patients avec une tumeur *IDH1* mutée. Cette dernière sera alors caractérisée de tumeur de bas grade (Figure 60-A). De plus, l'analyse de sondes particulières de la puce HM450K localisées dans les séquences de LINE du génome humain montre que ces tumeurs se comportent comme décrit dans la littérature (Ohka *et al.*, 2011). Ainsi, une méthylation normale est maintenue dans les tumeurs de bas grade et une hypométhylation est observée dans les tumeurs de haut grade (Figure 60-B). Parmi ces échantillons, 50 tumeurs ont été sélectionnées pour le projet, dont les caractéristiques sont détaillées dans l'annexe 22.

Les échantillons contrôles, quant à eux, proviennent de la « Brain and Tissue Bank » de l'université du Maryland. Ces échantillons correspondent à du corps calleux et de la substance blanche corticale, deux régions cérébrales riches en cellules gliales (Annexe 22).

Au début de l'étude, les tumeurs et tissus sains ont été cryobroyés. Pour chaque échantillon, la poudre obtenue a été répartie dans 3 tubes : 1 pour l'extraction d'ADN, 1 pour l'ARN total et 1 pour une utilisation ultérieure selon les besoins (ADN, ARN ou protéines). Cette procédure permet de disposer d'ADN et d'ARN provenant d'une composition cellulaire identique, notion importante notamment pour l'étude de corrélation entre expression des LCT et % de méthylation des L1 associés (Résultats Partie 3). La qualité des ARN a été évaluée sur BioAnalyseur (Agilent).

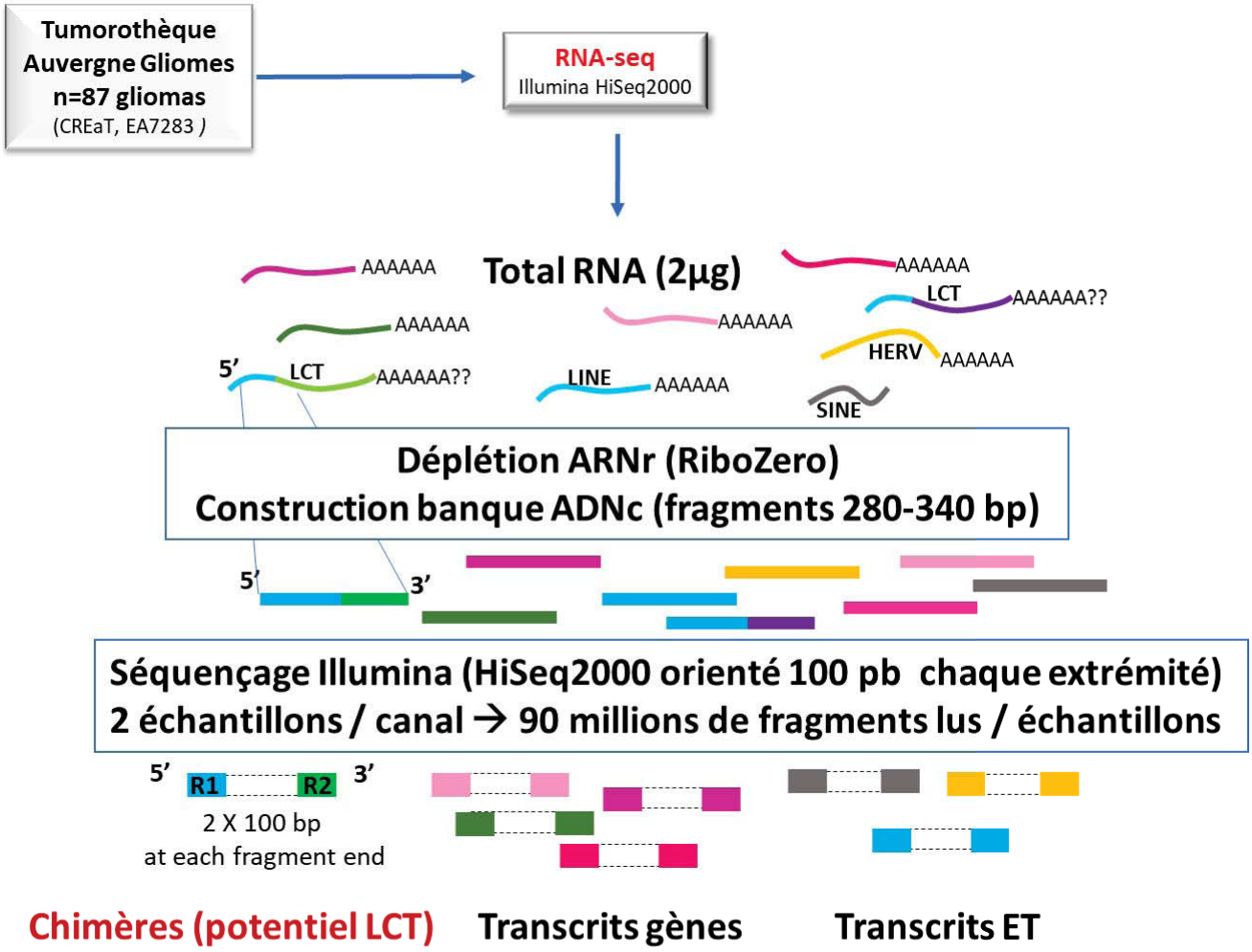


Figure 61 : Protocole utilisé pour le séquençage ARN de 13 gliomes et de 3 tissus contrôles.

A partir des échantillons disponibles dans la tumorothèque, les ARN totaux ont été extraits et déplétés en ARN ribosomaux (RiboZero). Parmi les échantillons ayant un RIN supérieur à 7 sur puce Agilent (c'est-à-dire des ARN de bonne qualité), 13 gliomes ainsi que 3 échantillons contrôles ont été sélectionnés. Pour le séquençage ARN des banques d'ADNc ont été obtenues en sélectionnant les fragments de 280 à 340 pb. Puis un séquençage Illumina Hiseq2000 en lecture orientées de 100pb à chaque extrémité a été réalisé. 2 échantillons sont assemblés par canal, ce qui permet d'obtenir 90 millions de fragments lus par échantillon.

1.2. Séquençage ARN (RNA-seq)

Les ARN totaux sont extraits de ces tumeurs et la qualité de ces ARN a été testée grâce à une puce Agilent. Parmi ceux ayant un RIN supérieur à 7, qui est le coefficient reflétant la qualité des ARN, 13 gliomes ainsi que 3 tissus cérébraux contrôles ont été sélectionnés. Au final, 16 RNA-seq ont été réalisés par la société *Integragen* sur 3 échantillons contrôle, 5 gliomes de bas grade *IDH1* muté (LGG) et 8 gliomes de haut grade *IDH1* WT (HGG). Comme nous ne savons pas si les transcrits chimères sont polyadénylés ou non, les ARN totaux ont été déplétés en ARN ribosomiques pour permettre un séquençage sans *a priori* de tous les transcrits. Un séquençage Illumina HiSeq2000 en lecture orientée de 100pb à chaque extrémité a été réalisé sur une banque de fragments d'ADNc de 280 à 340 pb. Deux banques d'ADNc ont été séquencées en même temps dans un canal, ce nous a permis d'obtenir en moyenne 90 millions de fragments lus par échantillon (Figure 61).

1.3. PCR quantitative sur puce Fluidigm

La RT-qPCR sur Biomark HD System (Fluidigm) avec la puce 96.96 Dynamic Array est une approche moyen débit qui permet la réalisation de 9216 réactions individuelles de PCR quantitative en temps réel en une réaction grâce à une technologie de micro fluidique. Les RT ont été réalisées sur 10 échantillons contrôles ainsi que 51 gliomes grâce à la *SuperScript III* selon les instructions du fournisseur *Invitrogen*, à partir de 750 ng d'ARN. La procédure expérimentale a été réalisée sur la plateforme Gentyane de l'INRA de Clermont-Ferrand. Cette RT-qPCR débute par une étape de pré amplification des ADNc en utilisant tous les couples d'amorces qui vont être utilisés pour la qPCR proprement dite. Puis la puce est réalisée avec d'un côté 96 puits dans lesquels sont déposés les ADNc des RT1 et RT2 préamplifiés, et de l'autre 96 puits contenant chacun un couple d'amorce (sens et antisens à 20µM chaque). Les amorces sont celles qui ont été dessinées pour la validation des chimères avec une amorce L1_common et une amorce dans la séquence unique (SU). Les couples d'amorces possédant une efficacité en RT-qPCR supérieure à 1,8 (SYBR green sur Lightcycler) ont été retenues (Table 3) et diluées à 10µM chaque pour la RT-qPCR. La cellule centrale de la puce Fluidigm permet le multiplexage des analyses avec 9216 RT-qPCR qui ont lieu en même temps. Au final, 61 échantillons sont analysés (10 contrôles, 42 gliomes de haut grade, 9 gliomes de bas grade) pour l'expression de 56 chimères et de 3 gènes de ménage validés pour les gliomes (*TBP*,

ID	séquence (5'-NNN-3')	Taille amplification L1_common-SU (pb)	Taille amplification L1_downstream1-SU (pb)	Taille amplification L1_downstream2-SU (pb)	Taille amplification L1_downstream3-SU (pb)	Taille amplification L1_upstream-SU (pb)
Chim1001_R	GTGAGGTACCACGTTTAACAAGA	200	423	531	586	801
Chim1509_F	TCAACCATTGGATAAACACATTACA	216	439	547	602	817
Chim2298_R	AGGAAATCTTGATTCTGACAGTTTT	150	373	481	536	751
Chim0148_F	TGGTGGATCAACATCTTTCTCC	225	456	569	624	839
Chim0071_R	AAAAGAGCGTGTTATTTTAGGCA	200	423	531	586	801
Chim0090_R	ACGGCTTTGCTTTCAGTTGT	160	383	493	549	764
Chim0529_F	AGAGCTCTTGATTTTCTGGGGA	103	326	433	488	703
Chim0569_R	GGGACTCCTCTGAAGCCTTT	316	540	649	704	918
Chim0801_F	GTCTGCCAACACCAACAAAA	247	470	578	633	848
Chim0837_R	TCTACAACATGGGCTCAAAAGAG	180	403	511	566	781
Chim1077_F	TCCTATTGTCGTGGATCCCTG	181	404	512	567	782
Chim1102_R	GTGAAACCTGAGCGCTCCAC	170	393	501	556	771
Chim1187_F	TCTTGAGATTACACAGGCTTAGC	141	364	472	527	742
Chim1277_R	GTGTTTACTAGCAGCTTTCCATT	250	473	581	636	850
Chim1464_F	TGCACAGCTGTCCATAGTGA	131	354	462	517	732
Chim1664_R	GGGAGATGGGTTAAAGCAAGG	235	458	566	621	836
Chim2143_F	GGCCCTACCCAAAATTTATGAAA	212	435	543	598	813
Chim2038_R	GGTAACAAAAGAATCATGCACTAAGA	70	292	400	455	670
Chim2723_R	AGCGTGATTATCCCCTTCGT	150	374	483	538	753
Chim2739_R	ACTTCCTTCTTCCAAAGCTAGTC	311	534	642	697	912
Chim2778_R	AATATGACTTTCTCTTCTAAATGG	111	332	441	496	711
Chim3154_R	CATGGGTGGTCTATCTCGGT	231	455	564	619	834
Chim0424_R	GAGATGACATTTTAGCAGAGCCT	246	469	577	632	847
Chim2058_F	ATTCCCCATGAGTCACCTTG	156	379	487	542	757
Chim1873_F	TCCTGAACCTAAAACAAACCACA	249	471	580	635	850
Chim2004_R	GCAATGTGTTCAAGGGTCCC	198	421	529	584	799
Chim2211_R	ACTATGTTACAGTTCACAAAGAGAGT	210	434	543	598	810
Chim2362_F	GCATAGTTCCCATGGTTATTTGC	134	357	466	521	740
Chim2256_R	TGGCACACGTTAAGCATCAT	227	451	560	615	832
Chim3107_R	TGTCACAGCAGCCAAATCCATA	228	438	547	601	815
Chim0342_F	CCATCAGGGTCCAAGGTGTTAA	291	515	624	679	899
Chim3007_R	GTTTTGTCAGATCAGCCCAACA	221	444	553	608	823
Chim0177_F	GGTCACTTCCAGGCCTATTTCA	204	428	537	592	806
Chim1883_F	AATCATCCGGGTTTTCTAGCACT	248	471	579	634	849
Chim2046_F	AAAGGTGCCACAGATTCGTACA	164	388	497	552	767
Chim1995_R	GGGTTAACATCCTTTTACTCAGG	126	337	446	501	733
Chim2238_R	TTAGGGCAAGTTGTAAGGCTT	165	388	497	552	769
Chim2379_F	ACTGGTACACTGACTCCACA	241	465	574	629	843
Chim0078_R	TGGCCACCTTCTCTTTCCAT	143	366	474	529	744
Chim2716_R	GCCTGAATTCCAACCTAAGTATTAA	118	340	449	504	715
Chim1571_F	GCATGTTTACAGCCTCCAACTAA	260	482	591	646	865
Chim2250_R	ACAACTTAAAGGTAGTCAATGCAGT	199	422	530	585	800
Chim1272_R	GGCACAGTCAGACCTCTCAT	181	405	514	569	782
Chim0915_F	AGCAATCAGGTATAGCAGAAGTCA	211	435	543	598	812
Chim1188_F	GCCTCCTCATCATCCAAGGT	121	343	451	506	720
Chim1994_R	TTGATCTCAATACTTATGCAACACTT	193	404	513	568	781
Chim2705_R	AATGATTTCAGTGCTCTTTACAAGT	111	330	439	494	709
Chim3186_F	GGTTAGACCCTAAGAATCATGGG	110	334	443	498	713
Chim0678_F	GGTCTTCTGCAAATTATTCATCCT	129	352	460	515	730
Chim3391_F	AGTCTCCCATGTAGTGTTCAAAC	228	452	561	616	830
Chim1163_F	GCTGATCCTCCATCCATGTT	198	421	529	584	799
Chim166_F	AATTGGCCACGTCGTTCTAC	227	450	559	614	825
Chim1766_F	TGTTCATATTGGCTTGCAAGAG	98	322	430	485	700
Chim752_R	GGTGACAGTCATAGGTTGTAAGC	298	522	610	665	880
Chim2996_R	GTGATTTAAGGAGATAATTGGAGGC	142	366	475	530	745
Chim3142_R	TCTGTTCTTTGGACCACACACA	186	396	505	560	774
Chim2773_R	CGTGGGCCATGTGATGTAG	175	398	525	580	794
Chim3351_R	AGCAAAATGTATGCCGAAGA	113	336	444	499	714
Chim2691_R	TGAACTGACCCTCCATAGCT	187	715	1009		
Chim233_F	TGTTACCCTTTCAGCCGACT	102	633	928		
Chim526_F	TCAATGTTACCACTTCAGAGCAC	250	788	1082		
Chim1764_F	GCCCAAAGTCTCAAGTACCTGA	200	731	970		
Chim172_F	TCCTAGGATAAAACGTACCCCT	244	775	1053		
Chim2385_F	GAGAAGCTCAGCAGATACGAAGA	216	764	1056		
Chim2738_R	AGGAAGAATGTGGCAATAACAGC	191	554	858		
Chim2646_F	CCTAAGGCTAAACAGAGGAATACT	245	781	1049		
Chim2850_F	AAGCCAGCACCACTTGTATG	314	775	957		
Chim2549_R	GCTCAGATCATTTGTGAAAGCT	389	825	1102		
Chim1502_F	TGATCTCAGCCATGGTTCTAA	253	700	981		
Chim3343_R	TGCTTGAGTATTATAGGATTGGAAGT	278	730	1014		
L1_common (PA1-6)	GCTGTTCCTATTCGGCCATC					
L1_downstream-1 (PA1-6	CGCCTTGCAGTTTGATCTCA					
L1_downstream-2 (PA1-6	CGTAGGACCCTCCGAGC					
L1_downstream-3 (PA1-6	CGTCCGTCACCCCTTTCTTT					
L1_upstream (PA1-6)	GCTCTCTTCAAAGCTGTCAGA					
L1_common (PA7)	GATGGCCGAATAGGAACAGC					
L1_downstream-1 (PA7)	AGATCCACTGGCTTGAAATTCT					
L1_upstream (PA7)	AVCAGAGCACCTGGGGGAAG					
L1_common (PA8)	AACAGCTCTGGTCTGCAG					
L1_downstream-1 (PA8)	CTGAAGCCAGGGAGCCAAGT					
L1_upstream (PA8)	CTGGGCAGGGCATCTCTG					

Table 3 : Listes des amorces utilisées pour les puces Fluidigm (LCT) et la marche en 5'.

RPL13A et *PPIA*). Pour chaque échantillon, les quantifications sont réalisées en dupliqua sur 2 RT indépendantes. A la fin de la réaction de qPCR, les résultats sont visualisés avec le logiciel « Fluidigm Real-Time PCR Analysis ». Celui-ci donne accès pour chaque réaction individuelle de qPCR à une valeur de Ct et à la courbe de fusion. Les amplifications ne présentant pas un Tm unique sur leur courbe de fusion sont éliminées. Les résultats sont ensuite extraits sous forme d'un tableau Excel pour calculer l'expression relative de chaque chimère. Celle-ci est exprimée sous la forme du ratio $R = (E_{chim}{}^{- Ctchim}) / (\text{moy geom } E_{HK}{}^{- CtHK})$ où E = efficacité d'amplification des amorces, chim = chimère quantifiée par les amorces et HK = gènes de ménage. Puis grâce aux logiciels R et XLSTAT, le test statistique de Mann-Whitney étudie l'expression différentielle entre les groupes et permet de générer les résultats sous forme graphique.

La même approche a été utilisée pour quantifier l'expression des gènes se trouvant à proximité des LCT dont l'expression pouvait être dérégulée dans les tumeurs par rapport aux contrôles, ou bien pour mesurer l'expression des gènes dans lesquels le LCT se trouve. En effet, la quantification de l'expression des gènes grâce aux données de RNA-seq se fait non seulement sur un nombre limité d'échantillons, mais aussi comptabilise le nombre de lectures sur l'ensemble du gène, non seulement celles provenant de l'expression propre du gène mais aussi potentiellement, celle provenant de l'expression du LCT inclus en son sein. C'est la raison pour laquelle, les amorces pour les gènes possédant un LCT en position intragénique ont été positionnées en amont du site d'initiation de la transcription du LCT afin de n'obtenir que l'expression du gène associé. Au final, l'expression de 42 gènes a pu être quantifiée par RT-qPCR selon la technologie Fluidigm, sur 9 échantillons contrôles, 40 HGG et 9 LGG (Table 4).

1.4. La technique de quantification de la méthylation par digestion d'enzyme de restriction (qAMP, quantitative Analysis of DNA methylation using qPCR) (Oakes *et al.*, 2006)

Cette technique permet de quantifier la méthylation ADN d'un locus grâce à l'amplification par qPCR de l'ADN suite à la digestion de celui-ci par des enzymes de restriction sensibles ou non à la méthylation. Les enzymes de restriction choisies sont HhaI et HpaII dont les sites de reconnaissance sont respectivement 5'-GCG/C-3' et 5'-C/CGG-3'. Ces deux enzymes sont sensibles à la méthylation (MSRE). Elles coupent lorsque les cytosines du locus ne sont pas méthylées. Une autre enzyme qui elle est dépendante de la méthylation, McrBC est utilisée.

LCT	Gène	Amorce F	Amorce R
LCT1873	GOLIM4	gaacagttggaacagcagagact	aacgatatcattatccatagcatca
	SERPINI1	tgcaaatttgacaggcctctct	aaaatggatggtcgacaataactt
	PDCD10	cagagccagaattccaagacct	cctgcggttctggtattgatat
	WDR49	ttgattatcactctcggctcaa	aggacacccactggtttagaga
LCT1994	CCDC109B	ccccaggttttgcgtgtgaa	tggctgcagttttgataccct
	CASP6	aattcagacattaactggcttgttc	ctactacatccaaaggaatgactgg
	PLA2G12A	ccgagatgtacagaaaacactagga	ggcatctcctttaaagatcagtttt
	CFI	acagaatgcacattcgaaggaa	cagaatccaacagccaccaata
	GAR1	ggtgccttatttcaatgctcct	gaagccttcatgttttctgaca
LCT78	XPR1	caggcaccttctgtggaagtta	tggcaagttctttttcacaggt
	STX6	ggctggacaatgtgatgaagaa	aagaggatgagcacaaccaaca
	MR1	ttaatggtgaagcacagcgatt	cagttccaccttgaacatctgc
	IER5	ccacctttctctcccaatgaaga	tccaagcttgaactcttctgtga
LCT2691	COL28A1	aacaaggaccacaaggcttcc	agatccaggctctccaatttctc
	RPA3	catgctagctcaattcatcgac	atcaactcgatggttccatttt
	MIOS	actggagatgttcaaacagcaa	gggatccaacttactcctgtga
LCT837	SCFD1	ccagaggcccttattagtcctt	ccaaaatttatcaaccggagtt
	STRN3	tggtaacctctttcaacactggta	taacaggaagtgtgggatgactta
	HECTD1	tgtgattctgatgttggtcaca	tctggctagctgatccaaaaat
	AP4S1	gagctgtctctctcgatccaat	ccacaatgaagagagctgcata
	COCH	tggcatccagtctcaaatgct	gccagttttcttctcgggtgt
LCT1995	EGF	tgctgctcactcttatcattctgt	attggttccttctgtgtcaatcct
	ELOVL6	tcctgttttctgctctgtatgct	tttggtcatcaaaatgtacacca
LCT3007	SNTB1	tgtgctacgtcactcggagtat	ggtcattaacgttggaatggat
LCT90	NEK7	gctaatgtgttcattacagccact	tctggagacatgtaataaggcgta
LCT3186	CCDC171	gacagcattggaggagtttagatt	gcatattttgcacagcagtatctt
LCT2046	EVC2	aagaagatggaaaaccagtaccag	agacttcctcaagtgctcctgtt
LCT2250/LCT2385	GPR98	tttccttctgagttggggacta	tgaacaaagcagcatagacgtt
LCT424	TENM4	agaccctgaaggcctacgac	aggccaatgtctgtccggta
LCT2004	USP53	acagtcacacaggtgtagggaaa	ttcaatagctttcagagcacattc
LCT233	GPATCH2	ttgtttaccaatgatgagggaag	gacataaggggaaaagaaccagt
LCT1571	DSCAM	acaatgggaattacacctgcata	gaatttccacacgatggtaggta
LCT342	PCDH15	cagctttgagtgaactttgacaga	atcatactggcccaagcagatt
LCT915	PRKD1	ggggcatctcgttccatct	gacaatggagcaagccatctc
LCT177	PTGFRN	cctgttaacatattttgggcatt	agccatgatgagaacagagtagc
LCT2211	NUDSF4	gagtttgataccagagagcgatg	tctgcaaaggaaactgcatct
LCT1077	FTO	gtttcaaggcaatcgatacagaa	caagatttcattcctttgttcca

Table 4 : Listes des amorces utilisées pour la quantification de l'expression des gènes en RT-qPCT (Fluidigm).

Les amorces utilisées pour la quantification de l'expression du gène dans lequel se situe le LCT sont indiquées en vert foncé. Pour 6 LCT, les amorces utilisées pour la quantification des gènes en aval de ce LCT sont indiquées en vert clair.

Cette enzyme coupe lorsqu'une cytosine du locus est méthylée, dont le site de reconnaissance est complexe. L'enzyme McrBC reconnait 2 sites 5'-G/A^{m}C-3' séparés par une distance optimale de 55 à 103 pb. Chez les mammifères, les cytosines méthylées sont dans un contexte de dinucléotides CpG. Aussi les séquences de reconnaissance non palindromiques de McrBC seront : 5'-G^{m}CG-3' et 5'-A^{m}CG-3', les séquences du brins complémentaires étant 5'-^{m}CGC-3' et 5'-^{m}GCT-3' respectivement. Cette enzyme va alors cliver l'ADN à environ 30 pb de chaque site. Des échantillons contrôles présentant une méthylation artificielle de 0, 30, 70 et 90% de méthylation sont utilisés pour valider la quantification de méthylation à chaque locus étudié. L'ADN 0% de méthylation est obtenu après amplification PCR du génome entier avec le kit Repli-G Mini kit (ou Whole Genome Amplification) de *Qiagen* réalisé sur de 50ng d'ADNg extrait de fibroblastes humains. L'ADN contrôle 90% de méthylation est obtenu par méthylation artificielle avec l'enzyme SssI à partir de 6µg d'ADNg 0% obtenu avec l'amplification WGA selon les instructions du fournisseur (NEB). Puis les contrôles 30 et 70% sont obtenus par mélange entre le 0 et le 90% de méthylation. Quatre échantillons contrôles de tissus cérébraux sains riches en cellules gliales sont analysés ainsi que 25 gliomes (4 LGG et 21 HGG). Suite à l'extraction des ADNg à partir des échantillons cryobroyés (tumeurs et contrôles) selon les instructions du fournisseur avec le kit *Qiagen*, chaque ADN est réparti dans 4 tubes à raison de 250 ng/tube. 3 tubes sont consacrés à la digestion avec une des enzymes de restriction HhaI, HpaII ou McrBC à 37°C pendant 2 heures. Les ADN sont ensuite traités avec 1µl de protéinase K (PK) à 4mg/µl pendant 30 minutes à 40°C pour éliminer les enzymes et la PK est inactivée pendant 10 minutes à 95°C. Le 4ème tube correspondant à l'ADN témoin non digéré subit les mêmes incubations et traitements (Figure 62). La concentration obtenue est de 3.125ng/µl, et 2µl ont été utilisés pour les qPCR. Pour certains 4µl ont été utilisés, ce qui a limité le nombre de loci pouvant être testés par digestion. Celles-ci sont d'abord réalisées sur 3 régions génomiques contrôles :

- 1 région sans site de coupure pour les enzymes de restriction utilisées (contrôle 1),

- 1 région hémi-méthylée correspondant à l'ICR du locus *PEG10*

-1 région non méthylée qui est la région promotrice du gène *GAPDH*.

À ces régions contrôles s'ajoutent 10 loci présentant une expression différentielle entre les contrôles et les tumeurs et 1 locus pour lequel il n'y a pas de différence d'expression (Table 5). La localisation des amorces permet d'analyser au moins 1 site de coupure analysé pour chaque enzyme HhaI et HpaII et plusieurs sites McrBC, c'est-à-dire au moins 3 cytosines qui peuvent

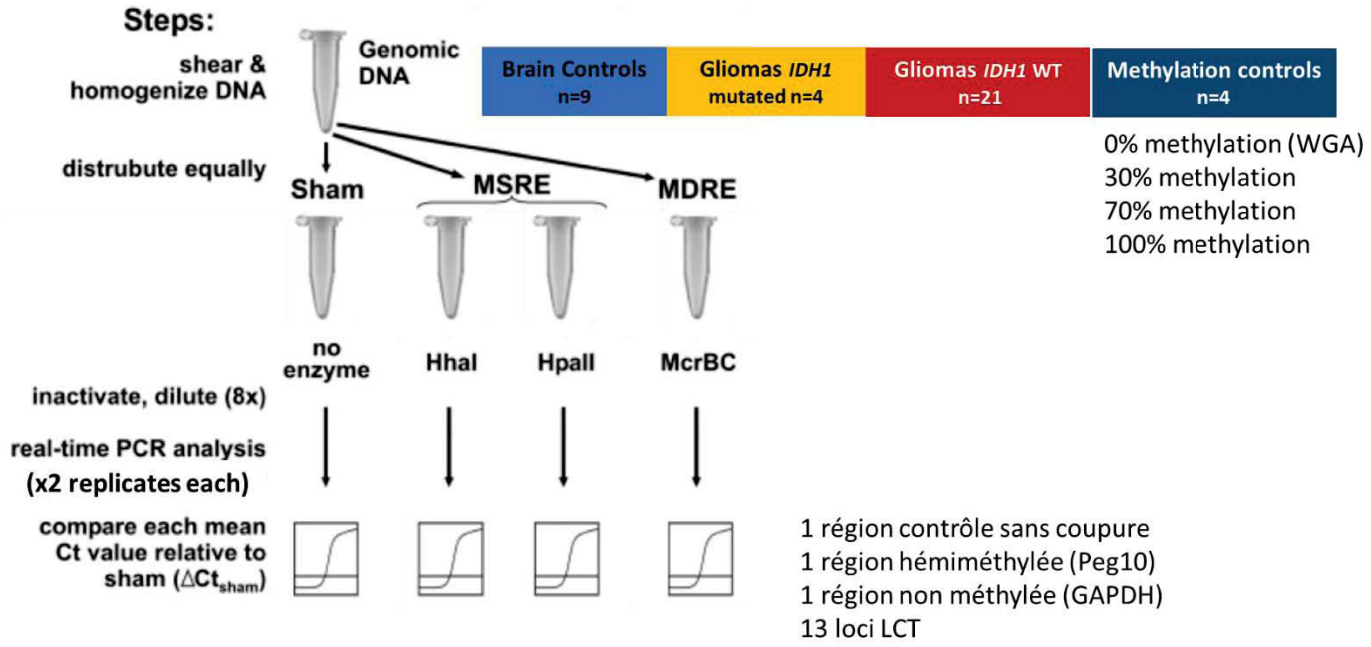


Figure 62 : Protocole de la qAMP

Suite à l'extraction de l'ADNg des échantillons de 8 tissus cérébraux contrôles, de 4 LGG et 21 HGG, l'ensemble est réparti en 4 tubes équivalents en quantité d'ADN. Un tube ne contient pas d'enzyme (Sham), ou des enzymes sensibles à la méthylation (MSRE) sont ajoutées comme HhaI et HpaII, ou bien une enzyme dépendante de la méthylation (MDRE) comme McrBC. Suite à la digestion de l'ADNg selon les instructions du fournisseur (NEB), les échantillons sont dilués. Les quantifications par qPCR de 3 régions contrôles ainsi que de 13 loci LCT sont réalisées. (D'après Oakes *et al*., 2006)

ID	séquence (5'-NNN-3')	Combinaison amorces
L1-qAMP#1.1	GGAACTCCCTGACCCCTTG	
L1-qAMP#1.2	GGAATTCCCTGACCCCTTG	
L1-qAMP#1.3_1571	GAACTCCCTGACCCCTGTG	
L1-qAMP#1.6_1764	GAGTTCCCTGACCCCTAGC	
L1-qAMP#1.9_2385	AAATTCCCCAACCCCTTGTG	
L1-qAMP#2.3-1995	GCAAGGCTCCATGGGTATG	
qAMP-controle1-F	TGAACTCTGAAGGACTGCTTGA	
qAMP-controle1-R	AGACTTTCAGGTTTCAGAGCAA	
Peg10_F	CTATAGTCCCACCTCCCTCAGA	
Peg10_R	CAACACAGCAGAGAAACCGC	
GAPDH_CGI_F	AAAGTAGGGCCCGGCTACTA	
GAPDH_CGI_R	TCGAACAGGAGGAGCAGAGA	
LCT2385_SU2	ATTCCTGGTGCCATGTTGCT	L1-qAMP#1.9_2385
LCT2211_SU2	AGGAATTGCACTTAAGAACCTCT	L1-qAMP#1.2
LCT2046_SU2	GGCTAAAGAAACGCTGCTGA	L1-qAMP#1.1
LCT0177_SU3	AGCTCCTTCTCTTGTGTCTTTCA	L1-qAMP#1.2
LCT2362_F	GCATAGTTCCCATGGTTATTTGC	L1-qAMP#1.1
LCT1188_F	GCCTCCTCATCATCCAAGGT	L1-qAMP#1.1
LCT1995_R	GGGTTAACATCCTTTTACTCAGG	L1-qAMP#2.3-1995
LCT1571_F	GCATGTTTACAGCCTCCAACTAA	L1-qAMP#1.3_1571
LCT1764_F	GCCCAAAGTCTCAAGTACCTGA	L1-qAMP#1.6_1764
LCT3007_R	GTTTTGTCAGATCAGCCCAACA	L1-qAMP#1.1
LCT0078_R	TGGCCACCTTCTCTTTCCAT	L1-qAMP#1.1

Table 5 : Liste des amorces utilisées pour la qAMP.

potentiellement être méthylées, sont testées dans la région amplifiée de la région promotrice du L1 (Annexe 5).

1.5. 5'RACE (Rapid Amplification cDNA End)

Afin de valider si l'initiation de la transcription a bien lieu à l'ASP du L1, l'approche de 5'RACE (kit GeneRacer d'*Invitrogen*) est réalisée. Ce kit a la particularité d'analyser spécifiquement les ARNm pleine taille (capés et polyadénylés), ainsi que de pouvoir réaliser des approches de 5' et de 3' RACE sur une même banque. Ainsi, 2µg d'ARN totaux du GBM24 et du GIII33 ont été utilisés pour réaliser 2 banques de RACE.

La première étape de cette expérience est de déphosphoryler les ARNm dégradés en 5' (qui ne sont plus coiffés) et les ARN non polyA grâce à la CIP (Phosphatase alkaline). La coiffe des ARNm pleine taille est ensuite enlevée grâce à la TAP (Tobacco acid pyrophosphatase) et un ARN oligo de 44 bases servant d'adaptateur est ligué à l'extrémité 5' phosphate libérée des ARNm (5'-CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3'). Puis une reverse transcription est réalisée avec la SuperScript IV (*Invitrogen*) selon les instructions du fournisseur avec une amorce oligo dT (5'-GCTGTCAACGATACGCTACGTAACGGCATGACAGTG(T)24-3′) de 60 bases. Les banques de RACE peuvent ensuite être amplifiées par PCR (Figure 63).

Afin de réaliser les amplifications de 5' ou 3' RACE, des conditions particulières de dessin des amorces spécifiques des séquences cibles doivent être respectées. Les amorces doivent être longues (de 23 à 28 nucléotides recommandés), présenter un fort pourcentage de GC (50 à 70%) et une température d'hybridation supérieure à 72°C. La liste des amorces utilisées est rapportée dans l'annexe 23.

Les conditions de PCR optimisées avec la Taq polymérase Advantage HD (Clontech) pour les LCT837, LCT1873 et LCT1994 sont :

<u>PCR1 5'RACE</u>

98°C 2min

94°C 30sec	72°C 1min		X5
94°C 30sec	70°C 1min		X5
94°C 30sec	68°C 30sec	72°C 1min	X25

72°C 5min

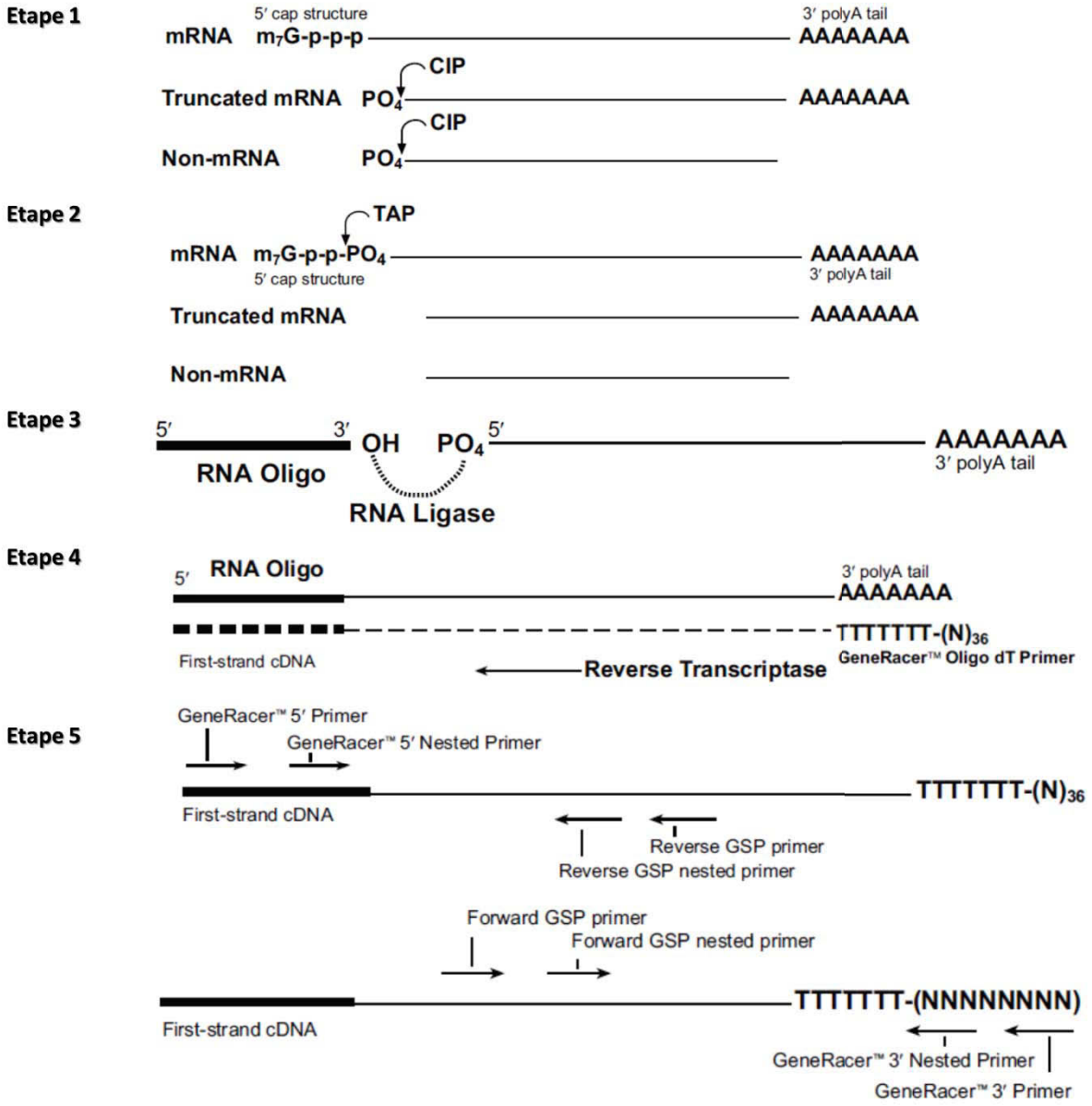


Figure 63 : Protocole de l'approche GeneRacer pour la réalisation de 5'RACE.

L'étape 1 consiste à déphosphoryler les ARNm dégradés en 5' (qui ne sont plus coiffés) et les ARN non polyA (non coiffés) de l'échantillon. Au cours de l'étape 2, les ARNm pleine taille sont décoiffés grâce à la pyrophosphatase acide de tabac (TAP). Ceci libère une extrémité 5' phosphate spécifiquement sur les ARNm qui étaient coiffés. L'étape 3 va permettre la ligation de l'adaptateur (RNA Oligo) en 5' des ARNm pleine taille décoiffés. Au cours de l'étape 4, une transcription inverse est réalisée en amorçant la réaction au niveau de la queue polyA grâce à une amorce oligodT couplée à une séquence adaptatrice (GeneRacer OligodT primer). Enfin les réactions de PCR pourront avoir lieu, soit en combinant les amorces dans l'adaptateur en 5' et les amorces de la séquence unique pour la 5'RACE pour l'identification du site d'initiation de la transcription, soit en combinant les amorces dans l'adaptateur en 3' et les amorces de la séquence unique pour identifier la séquence 3' des ARNm testés.

PCR2 5'RACE

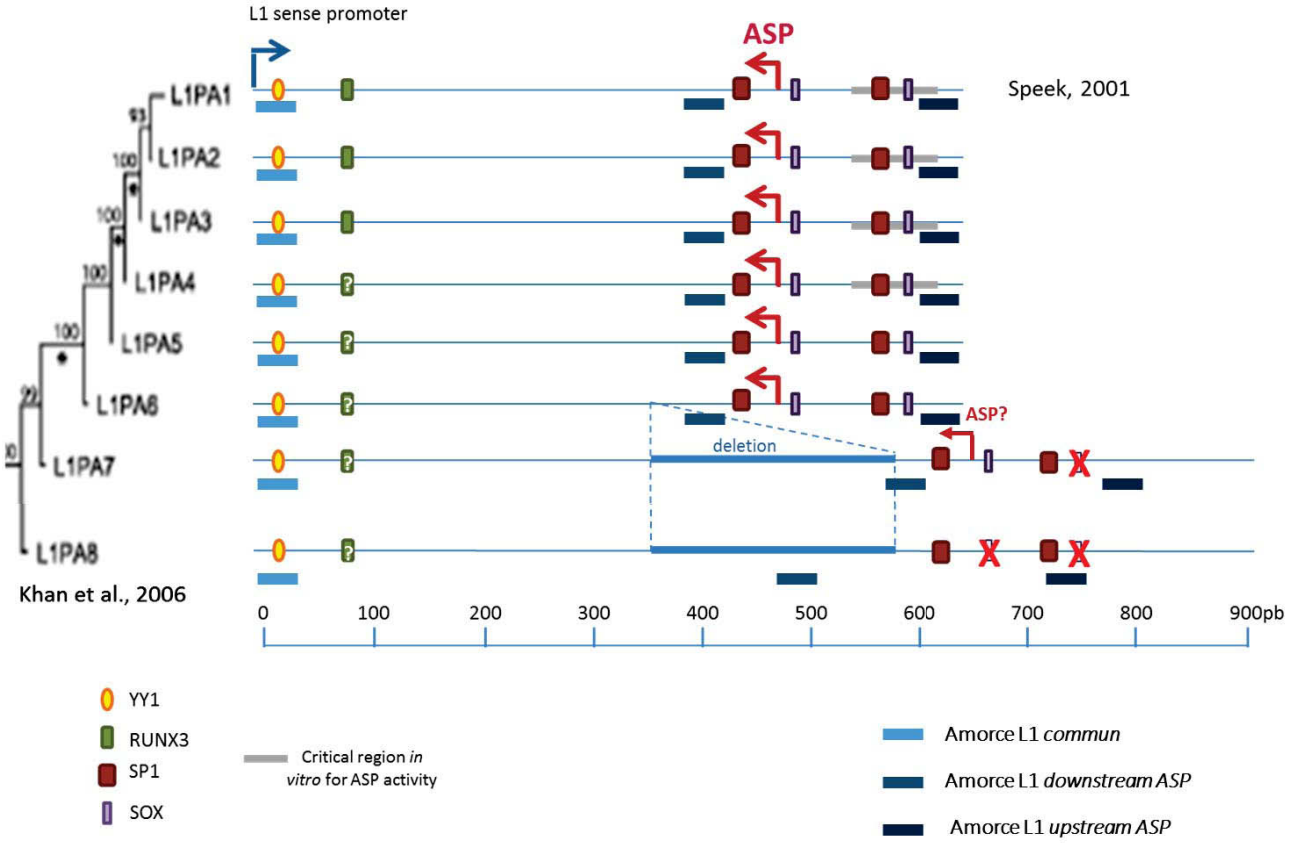
98°C 2min

94°C 20sec 65°C 20sec 72°C 1min X25

72°C 10min

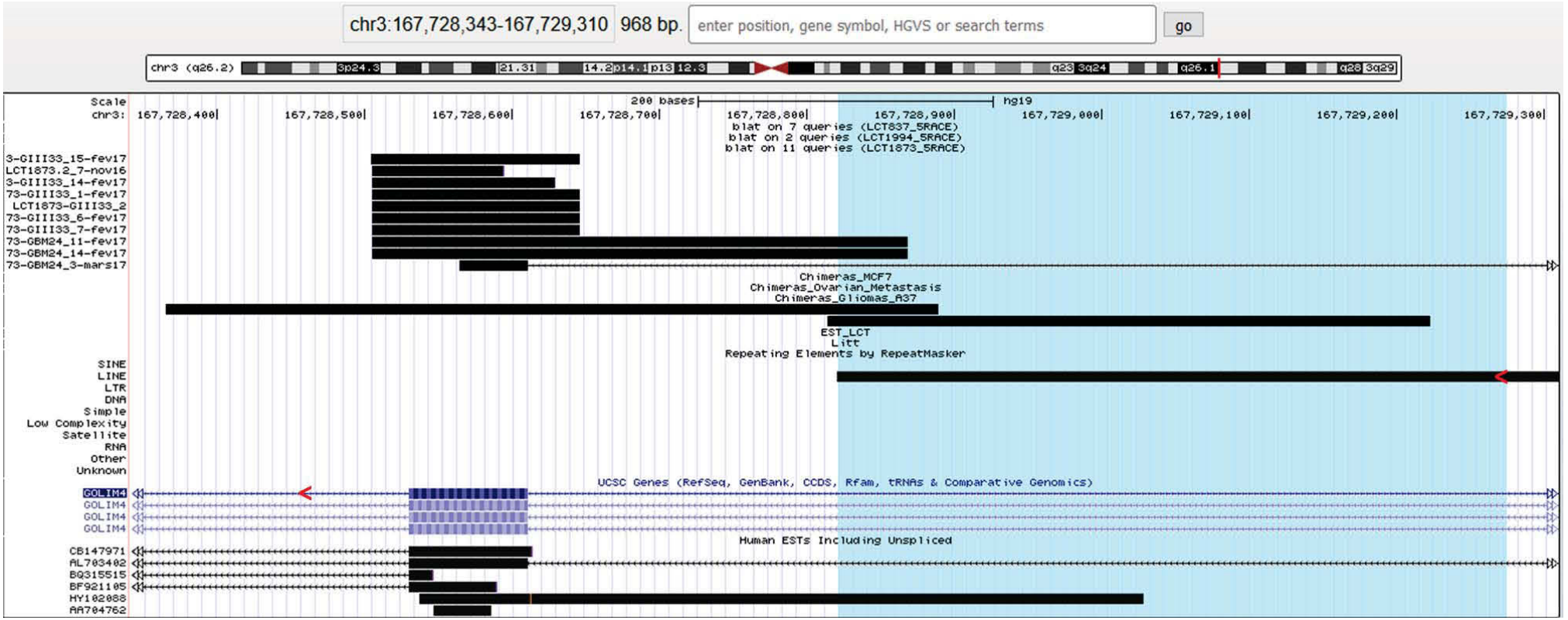
De nombreux temps d'élongation ont été testés pour la PCR1 de 2 minutes et pour la PCR2 de 1 minute à 3 minutes, sachant que la taq polymérase synthétise 1 Kb par minute, et que les tailles des amplicons attendues sont de 532 pb pour le LCT837, 735 pb pour le LCT1873 et 601pb pour le LCT1994. Malgré cela, les tailles d'amplificon ne remontent pas jusqu'à l'extrémité 5' du LCT, supposée à l'ASP.

2. Annexes Figures



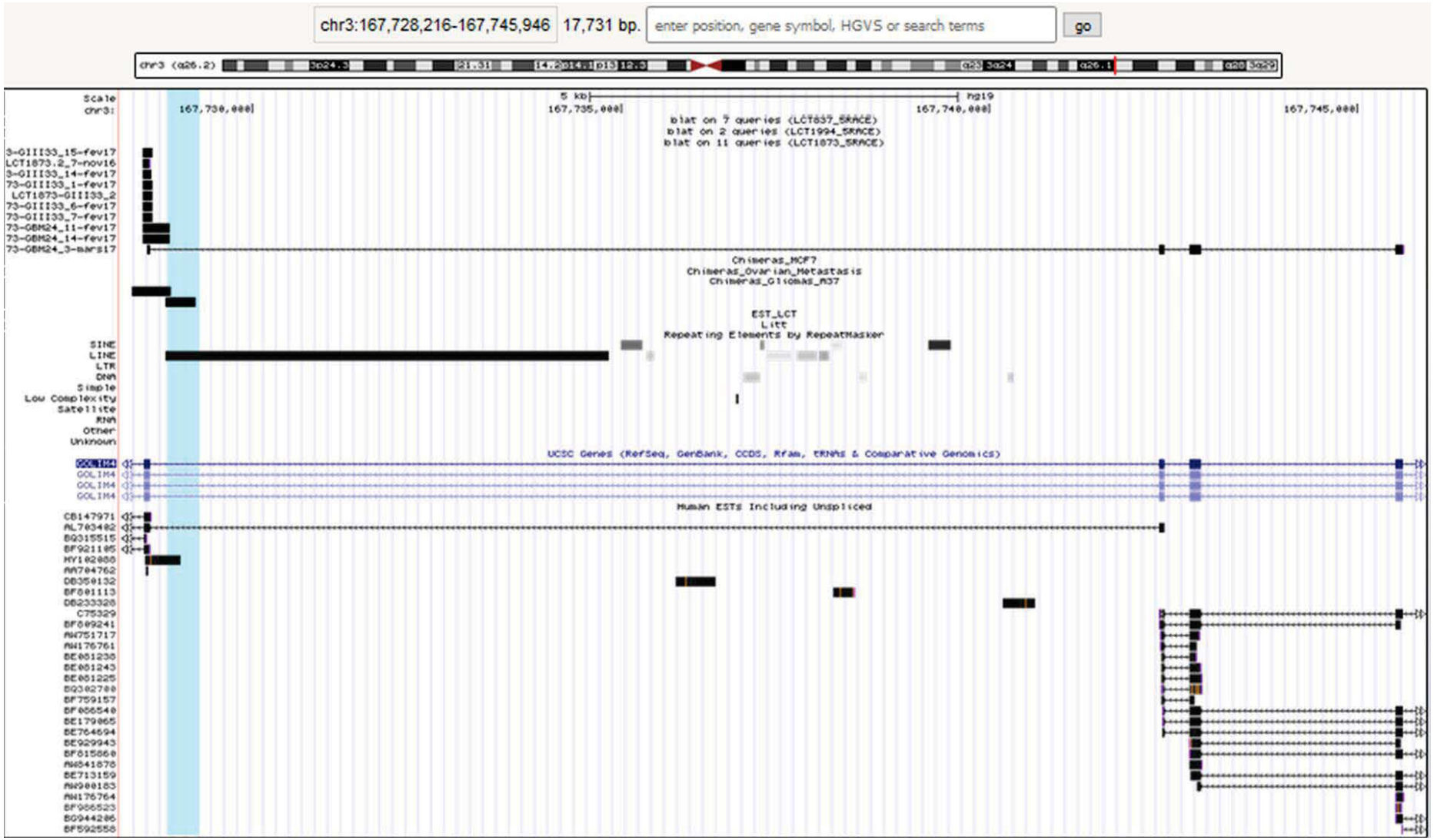
Annexe 1 : Position des amorces utilisées pour la marche en 5'.

Alignement des séquences des extrémités 5'UTR des sous-familles L1PA1 à 8, avec le positionnement des sites de fixation pour les facteurs de transcription présents dans cette région 5'UTR. Pour les amorces L1_downstream et L1_upstream, les mêmes amorces ont pu être utilisées pour les séquences L1PA1 à 6. En revanche, pour les sous-familles L1PA7 et PA8, de par la présence d'une séquence d'ADN de 200 pb qui a été délétée au cours de l'évolution, des amorces L1_downstream et L1_upstream spécifiques des sous-familles L1PA7 et L1PA8 ont été dessinées.



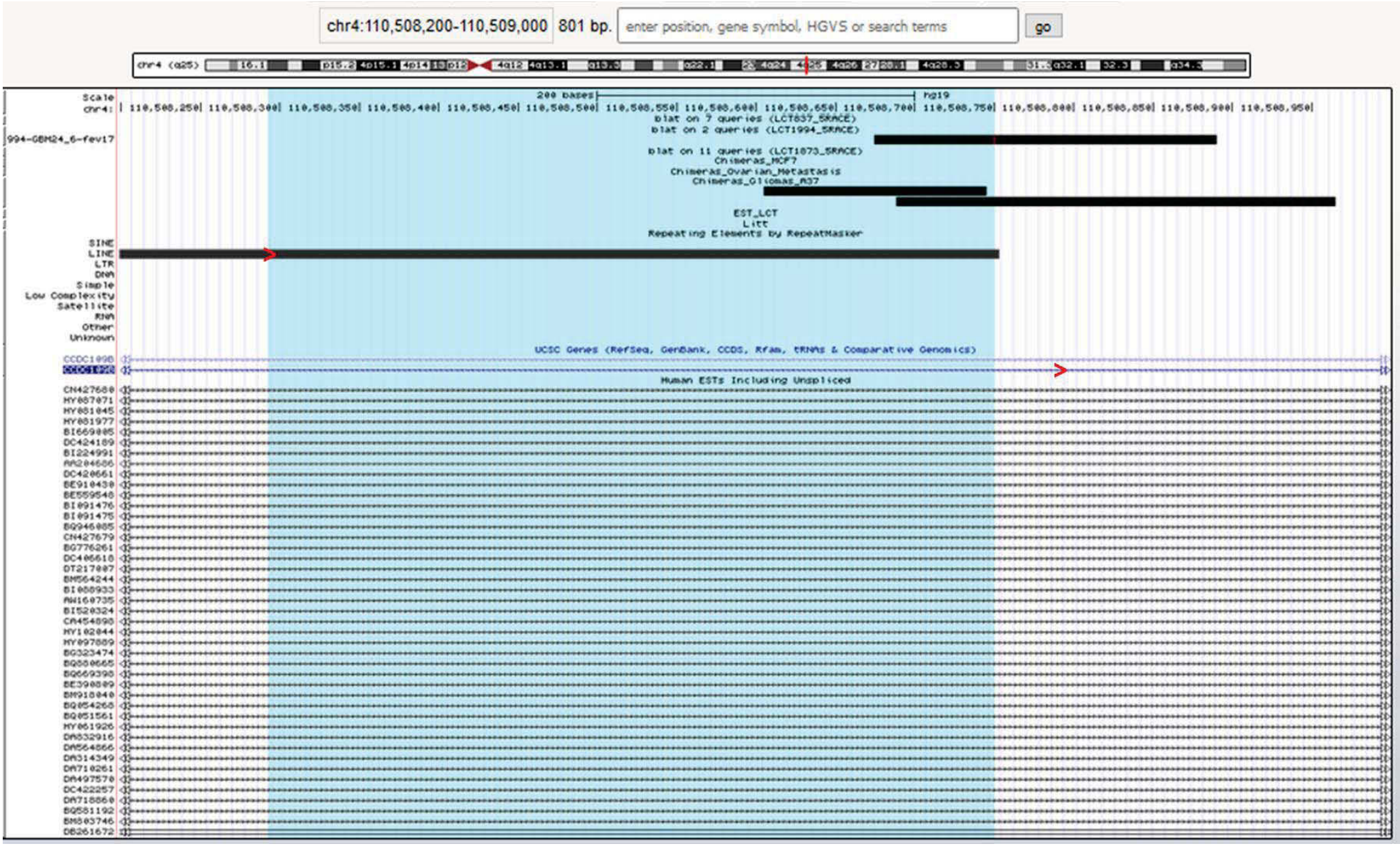
Annexe 2 : Détermination du site d'initiation de la transcription du LCT1873 au L1PA2 par 5'RACE.

Les séquences des 10 clones obtenus par 5'RACE sont localisées dans la séquence R1 et R2 détectées par CLIFinder dans les données de RNA-seq des gliomes. Ce LCT est localisé dans l'intron 14/15 du gène *GOLIM4*, en sens. Le sens de transcription du gène et de l'ASP est indiqué par les flèches rouges. La région 5'UTR surlignée en bleu indique les 450 premières paires de bases du L1 dans lesquelles l'ASP est situé. Une évidence d'EST est identifiée. Cet EST débute dans les 150 premières paires de bases du L1.



Annexe 3 : Exons du gène GOLIM4 capturés par 5'RACE

Au cours des expériences de 5'RACE pour identifier le TSS du LCT1873, un clone est amplifié et correspond aux exons 11 à 14 du gène *GOLIM4* en sens.



Annexe 4 : Détermination du site d'initiation de la transcription du LCT1994 au L1PA6 par 5'RACE.

Les séquences de 1 clone obtenu par 5'RACE est localisé dans la séquence R1 et R2 détectées par CLIFinder dans les données de RNA-seq des gliomes. Ce LCT est localisé dans l'intron 1/7 du gène *CCDC109B*, en sens. Le sens de transcription du gène et de l'ASP est indiqué par les flèches rouges. La région 5'UTR surlignée en bleu indique les 450 premières paires de bases du L1 dans lesquelles l'ASP est situé.

Annexe 5 : Localisation des amorces utilisées pour la qAMP ainsi que le nombre de site C-G interrogé.

	Nbre de sites CpG interrogés			Amplif qPCR (bp)
	HpaII	HhaI	McrBC	
LCT_3007	1	2	multiple	412
LCT_0078	2	4	multiple	332
LCT_2362	2	2	multiple	325
LCT_1571	1	2	multiple	450
LCT_2046	2	1	multiple	356
LCT_0177	1	1	multiple	396
LCT_1188	2	2	multiple	312
LCT_1995	1	1	multiple	461
LCT_2385	1	0	multiple	393
LCT_2211	1	1	multiple	402
LCT_1764	1	1	multiple	378

Annexe 6 : Tableau des résultats de la méthylation différentielle (selon le test de Mann-Whitney) pour le LCT testés en qAMP

Methyl Differentielle Mann Whitney		
	LGG vs Ctles	HGG vs Ctles
2046%CH3	1,000	0,265
2362%CH3	0,637	0,071
1188%CH3	0,925	0,064
2385%CH3	0,777	0,243
1764%CH3	0,219	0,222
2211%CH3	0,299	**0,001**
177 %CH3	0,508	0,599
1995 %CH3	0,070	**0,002**
1571 %CH3	0,241	0,191
0078 %CH3	0,219	0,029
3007 %CH3	0,112	0,027

p ajusté Bonferroni p< 0,00454

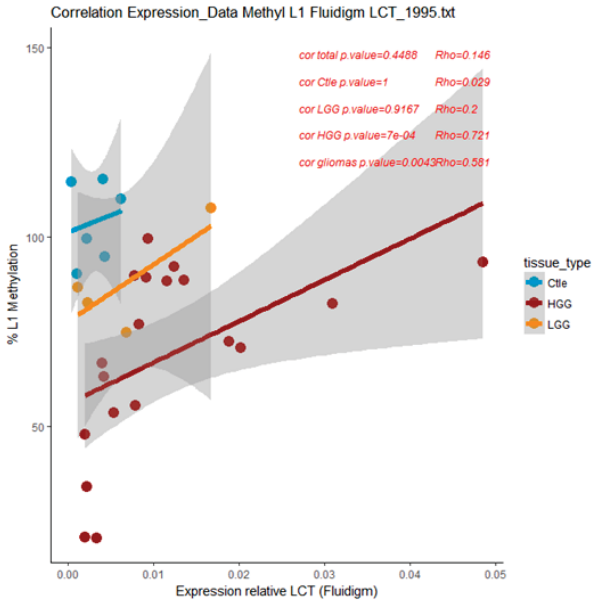
Annexe 7 : Tableau des résultats des corrélations de Spearman entre la méthylation de la région promotrice du L1 et l'expression des LCT.

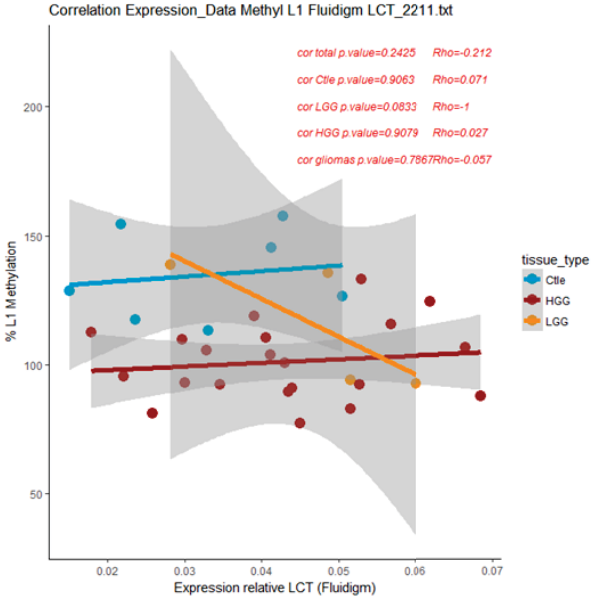
			Ctles n=7		LGG n=4		HGG n=21	
	cor total p.value	Rho	cor Ctle p.value	Rho	cor LGG p.value	Rho	cor HGG p.value	Rho
Data Methyl L1 Fluidigm LCT_0078.	0,166720924	-0,254435484	0,661507937	-0,214285714	0,75	-0,4	0,926549607	0,022556391
Data Methyl L1 Fluidigm LCT_0177.	0,83604639	0,038709677	0,497619048	0,321428571	0,083333333	1	0,621381283	-0,117293233
Data Methyl L1 Fluidigm LCT_1188.	0,018279033	-0,423790323	0,555952381	-0,285714286	0,416666667	0,6	0,015706579	-0,538345865
Data Methyl L1 Fluidigm LCT_1571.	0,328684332	-0,191023536	1	-0,028571429	0,916666667	0,2	0,196815079	-0,318885449
Data Methyl L1 Fluidigm LCT_1764.	0,635464672	-0,088306452	0,353571429	-0,428571429	0,083333333	1	0,95184079	-0,015037594
Data Methyl L1 Fluidigm LCT_1995.	0,448751753	0,145812808	1	0,028571429	0,916666667	0,2	0,000718586	0,721052632
Data Methyl L1 Fluidigm LCT_2046.	0,542257554	-0,11143695	0,138888889	0,642857143	0,333333333	-0,8	0,74538947	0,075324675
Data Methyl L1 Fluidigm LCT_2211.	0,242496505	-0,212243402	0,906349206	0,071428571	0,083333333	-1	0,907914494	0,027272727
Data Methyl L1 Fluidigm LCT_2362.	0,626602038	0,089076246	0,83968254	-0,107142857	0,75	-0,4	0,122459551	0,348051948
Data Methyl L1 Fluidigm LCT_2385.	0,540921624	0,111803519	0,266666667	0,5	0,333333333	0,8	0,766895706	-0,068831169
Data Methyl L1 Fluidigm LCT_3007.	0,021558965	-0,435139573	0,333333333	0,8	0,333333333	-0,8	0,414490572	-0,192481203

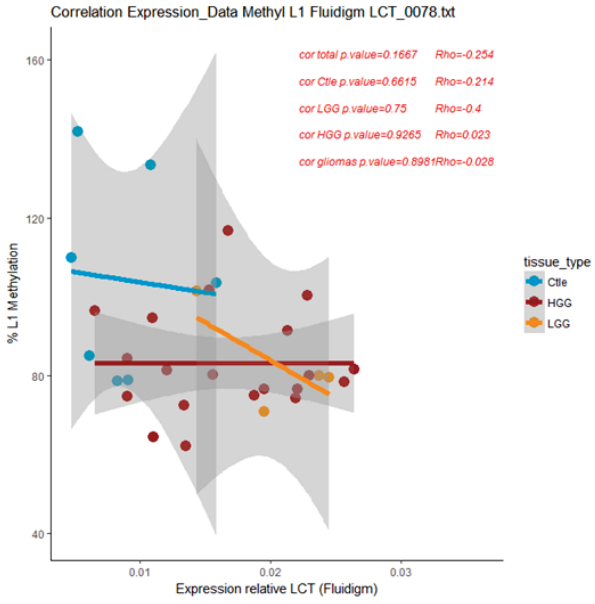
0,004545455 p adjust Bonferroni

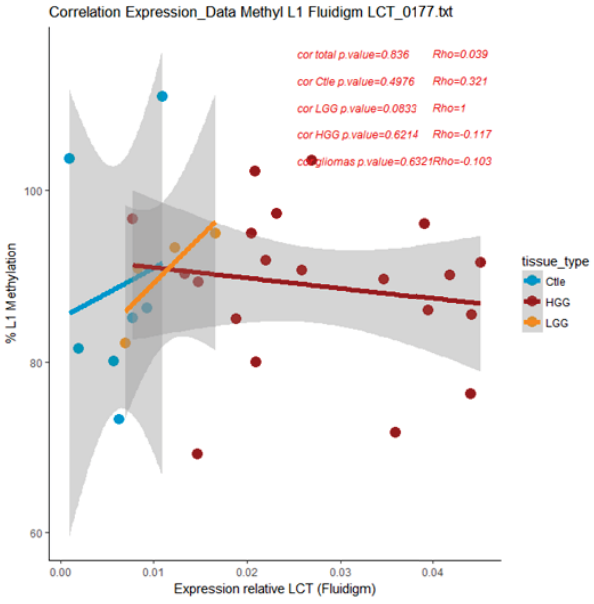
Annexe 8 : Mise en relation des données de pourcentage de méthylation au niveau du L1 et l'expression relative du LCT.

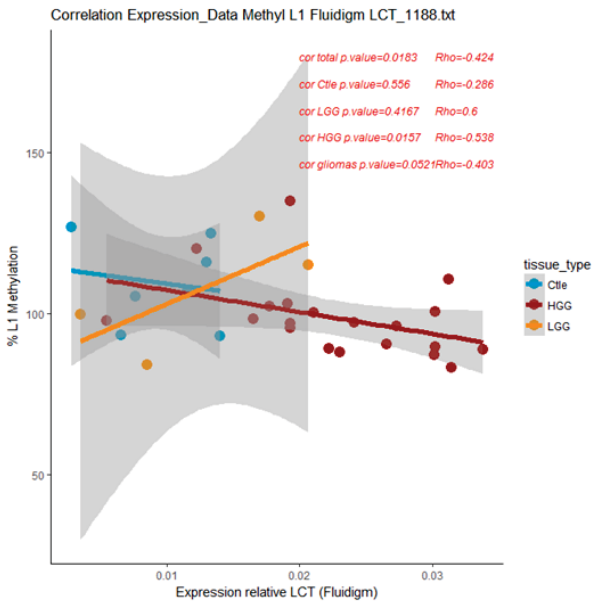
Graphiques pour les LCT 1995, 2211, 78, 177, 1188, 1571, 2362, 2385 et 3007.

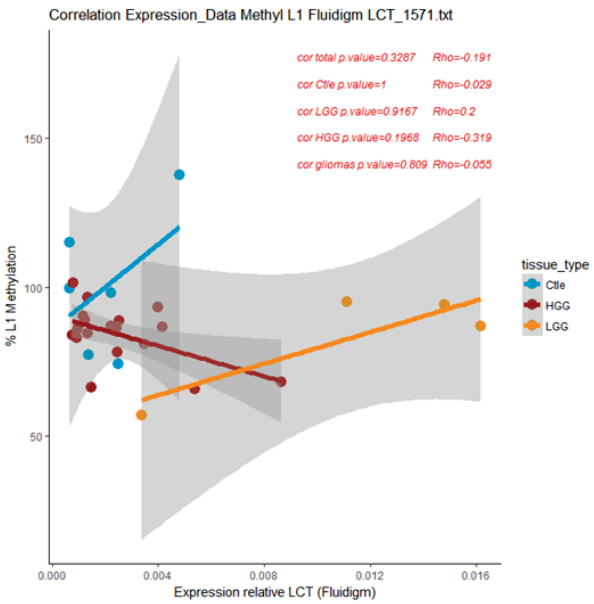


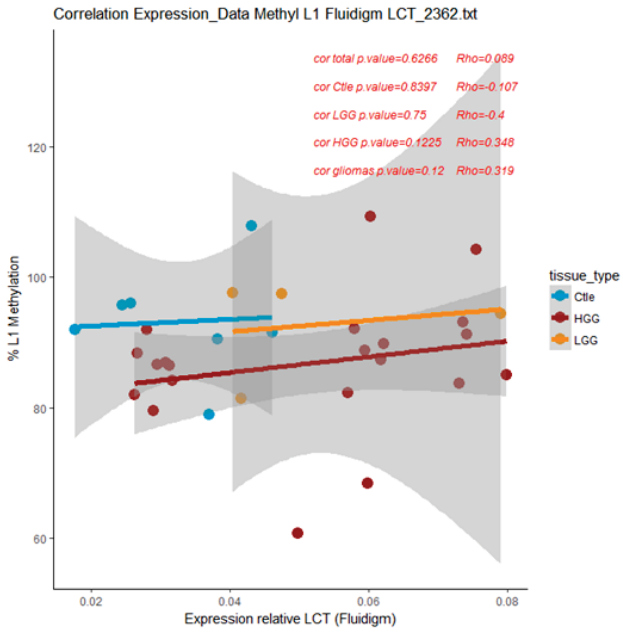


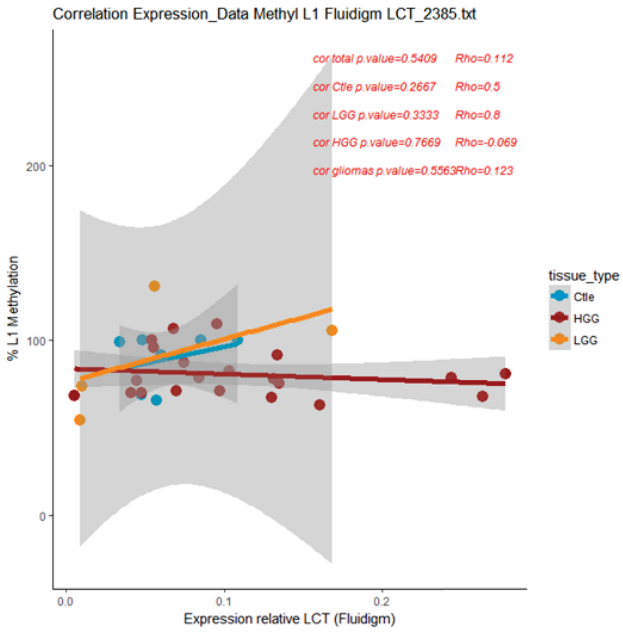


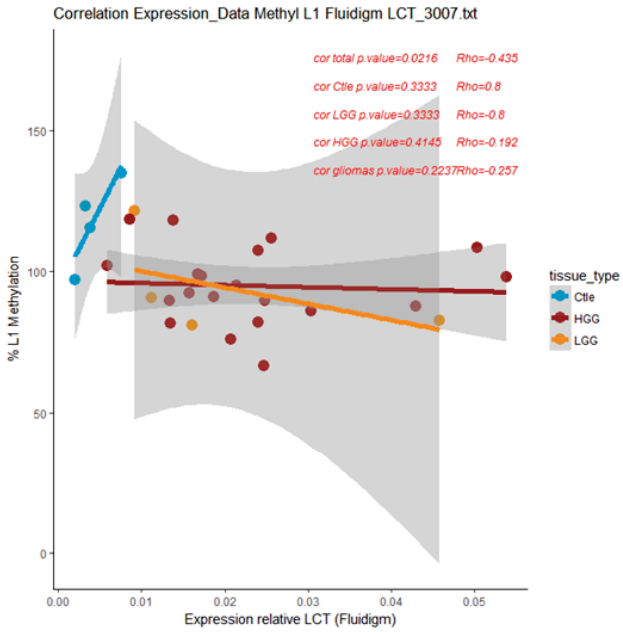







Correlation Expression_Data Methyl L1 Fluidigm LCT_2362.txt
cor total p.value=0.6266 Rho=0.089
cor Ctle p.value=0.8397 Rho=-0.107
cor LGG p.value=0.75 Rho=-0.4
cor HGG p.value=0.1225 Rho=0.348
cor gliomas p.value=0.12 Rho=0.319
% L1 Methylation
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG


Correlation Expression_Data Methyl L1 Fluidigm LCT_2385.txt
cor total p.value=0.5409 Rho=0.112
cor Ctle p.value=0.2667 Rho=0.5
cor LGG p.value=0.3333 Rho=0.8
cor HGG p.value=0.7669 Rho=-0.069
cor gliomas p.value=0.5563 Rho=0.123
% L1 Methylation
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG


Correlation Expression_Data Methyl L1 Fluidigm LCT_3007.txt
cor total p.value=0.0216 Rho=-0.435
cor Ctle p.value=0.3333 Rho=0.8
cor LGG p.value=0.3333 Rho=-0.8
cor HGG p.value=0.4145 Rho=-0.192
cor gliomas p.value=0.2237 Rho=-0.257
% L1 Methylation
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG

Annexe 9 : Résultats de l'expression différentielle (selon le test de Mann-Whitney) des gènes dans lesquels le LCT se situe dans les groupes contrôles VS les LGG ou HGG

		LGG vs Ctles	HGG vs Ctles
LCT	Variable\Test	Mann-Whitney	Mann-Whitney
LCT 3007	SNTB1	**0,001**	**< 0,0001**
LCT 837	SCFD1	**0,022**	**0,001**
LCT 1873	GOLIM4	**0,003**	**0,000**
LCT 915	PRKD1	0,427	0,526
LCT 78	XPR1	0,077	**0,001**
LCT 1571	DSCAM	**0,013**	0,070
LCT 342	PCDH15	**0,004**	0,831
LCT 2250	GPR98	0,163	**0,037**
LCT 2046	EVC2	0,391	**< 0,0001**
LCT 1994	CCDC109B	0,312	**< 0,0001**
LCT 177	PTGFRN	0,077	**< 0,0001**
LCT 233	GPATCH2	0,916	0,783
LCT 1995	EGF	0,561	0,944
LCT 2691	COL28A1	0,862	0,067
LCT 2385	GPR98	0,163	**0,037**
LCT 2004	USP53	0,158	**0,023**
LCT 2211	NUDSF4	0,185	**0,005**
LCT 424	TENM4	0,077	**0,032**
LCT 90	NEK7	1,000	**0,009**
LCT 1077	FTO	0,331	0,051
LCT 3186	CCDC171	0,331	**0,003**

Annexe 10 : Résultats de la corrélation (Corrélation de Spearman) entre l'expression du LCT et l'expression du gène dans lequel il se trouve.

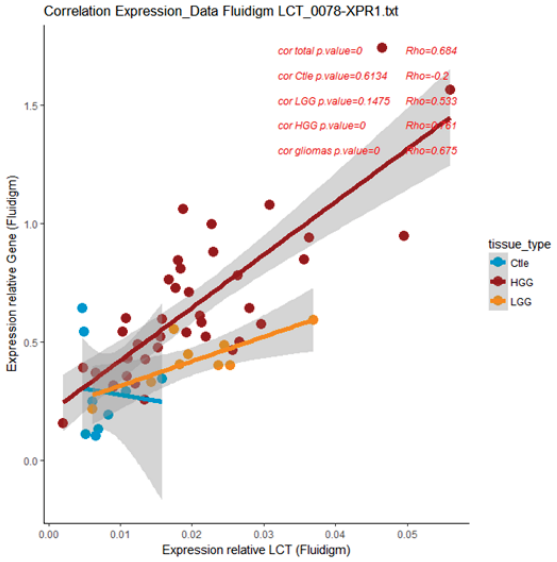
					Corrélation Spearman							
	Overexpr LCT		Regression Lineaire		Ctles+LGG+HGG		Ctle n=9		LGG n=9		HGG n=39	
	LGG	HGG	Univariée	Multivariée	cor total p.val	Rho	cor Ctle p.va	Rho	cor LGG p.va	Rho	cor HGG p.va	Rho
Data Fluidigm LCT_3007-SNTB1.txt			1.89e-10	1.75e-07	0,0000	0,6993	0,6436	-0,1833	0,0172	0,7833	0,0007	0,5279
Data Fluidigm LCT_0837-SCFD1.txt			1.5e-09	9.58e-08	0,0000	0,7697	0,0833	1,0000	0,2992	0,4286	0,0000	0,7681
Data Fluidigm LCT_1873-GOLIM4.txt			8.23e-10	2.38e-07	0,0000	0,6617	0,6777	-0,1667	0,0369	0,7167	0,0011	0,5089
Data Fluidigm LCT_0915-PRKD1.txt			0.015	0.000606	0,0001	0,5151	0,0138	0,8000	0,0760	0,6333	0,0002	0,5737
Data Fluidigm LCT_0078-XPR1.txt			2.93e-12	1.85e-11	0,0000	0,6841	0,6134	-0,2000	0,1475	0,5333	0,0000	0,7611
Data Fluidigm LCT_1571-DSCAM.txt			2.75e-13	1.45e-08	0,0004	0,5149	0,7520	0,1429	0,0108	0,8167	0,0078	0,5061
Data Fluidigm LCT_0342-PCDH15.txt			<2e-16	<2e-16	0,0000	0,8960	0,2125	0,4667	0,0007	0,9333	0,0000	0,8775
Data Fluidigm LCT_2250-GPR98.txt			< 2e-16	<2e-16	0,0000	0,8941	0,0154	0,8333	0,0000	0,9833	0,0000	0,9052
Data Fluidigm LCT_2046-EVC2.txt			0.000219	0.0127	0,0000	0,7703	0,0833	1,0000	0,0833	0,9000	0,0004	0,5709
Data Fluidigm LCT_1994-CCDC109B.txt			5.21e-14	5.48e-09	0,0000	0,8702	0,1323	0,5952	0,0083	0,8333	0,0000	0,6960
Data Fluidigm LCT_0177-PTGFRN.txt			8.15e-10	5.08e-06	0,0000	0,8785	0,0433	0,7000	0,0589	0,6667	0,0000	0,7405
Data Fluidigm LCT_0233-GPATCH2.txt			0.000524	0.000389	0,0001	0,4982	0,2000	0,5714	0,8801	0,0667	0,0002	0,5743
Data Fluidigm LCT_1995-EGF.txt			6.25e-08	5.64e-07	0,0002	0,7223	1,0000	-1,0000	0,2333	0,7000	0,0003	0,8000
Data Fluidigm LCT_2691-COL28A1.txt			< 2e-16	<2e-16	0,0000	0,7815	0,6615	0,2143	0,0011	0,9524	0,0000	0,7878
Data Fluidigm LCT_2385-GPR98.txt			< 2e-16	<2e-16	0,0000	0,8880	0,0072	0,8810	0,0007	0,9333	0,0000	0,8706
Data Fluidigm LCT_2004-USP53.txt			<2e-16	<2e-16	0,0000	0,9153	1,0000	0,0000	0,0046	0,9048	0,0000	0,9173
Data Fluidigm LCT_2211-NUDSF4.txt			0.00393	9.18e-06	0,0022	0,4002	0,8432	0,0833	0,0369	0,7167	0,0001	0,6014
Data Fluidigm LCT_0424-TENM4.txt			6.88e-08	1.86e-07	0,0000	0,8017	0,7500	-0,4000	0,0107	0,8571	0,0000	0,8246
Data Fluidigm LCT_0090-NEK7.txt			0.033473	0.07275	0,0123	0,3804	0,9167	0,2000	0,5008	0,2857	0,0865	0,3133
Data Fluidigm LCT_1077-FTO.txt			0.000211	0.00175	0,0026	0,4509	0,7500	0,4000	0,0458	0,7381	0,0182	0,4242
Data Fluidigm LCT_3186-CCDC171.txt			0.003117	0.00734	0,0005	0,4489	0,2696	-0,4167	0,0760	0,6333	0,0023	0,4785

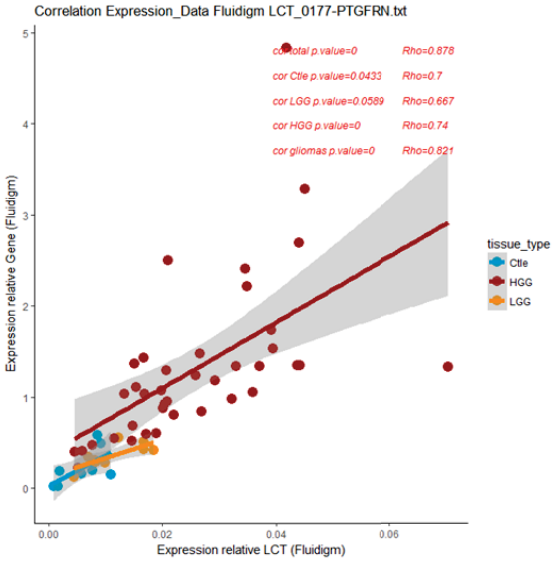
0,0024 padjust bonferroni

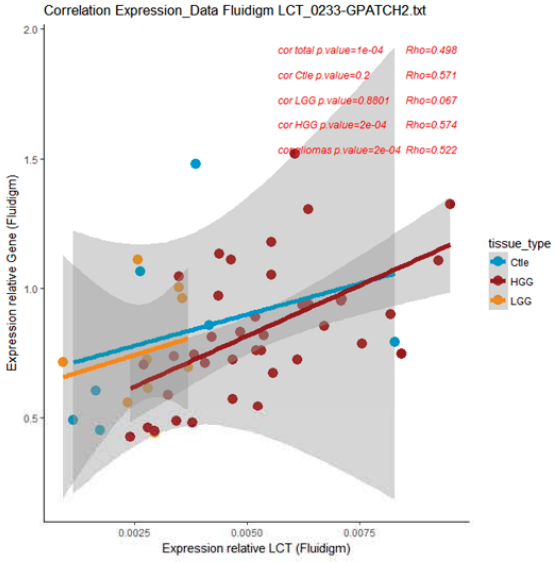
LCT antisens gène
LCT unchanged in T vs Ctles

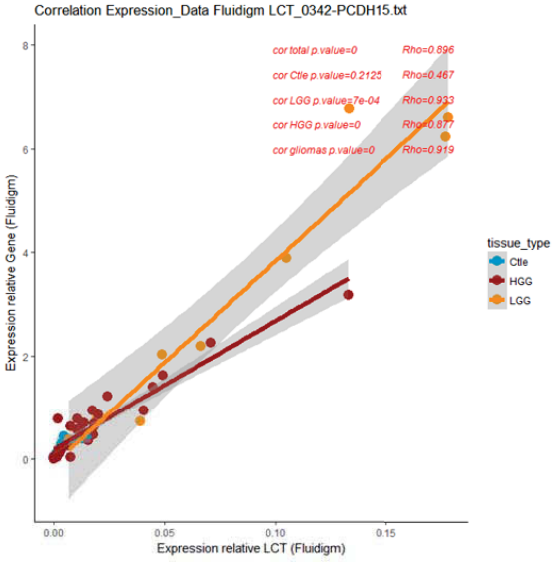
Annexe 11 : Résultats des corrélations de Spearman entre l'expression relative du LCT et l'expression du gène qui lui est associé.

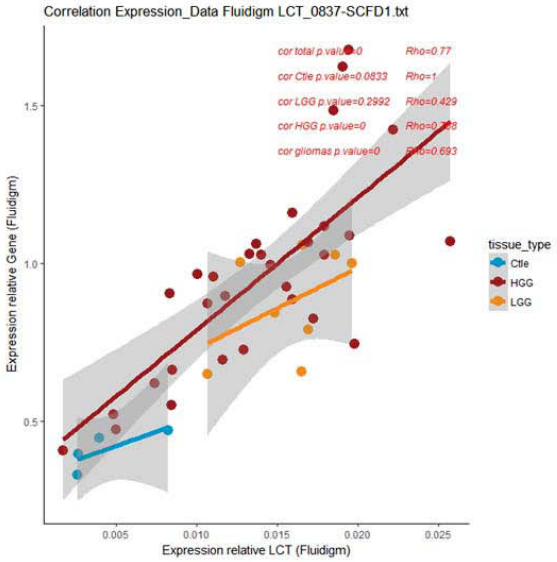
Graphiques pour les LCT 78, 177, 233, 342, 837, 915, 1571, 1873, 1994, 1995, 2004, 2046, 2211, 2250, 2385, 2691 et 3007.

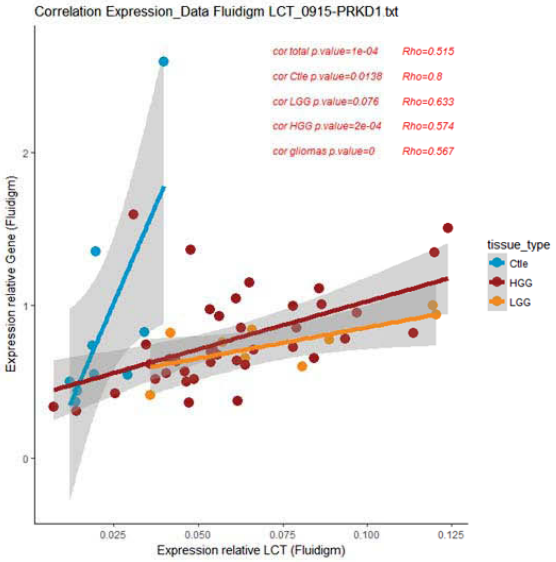


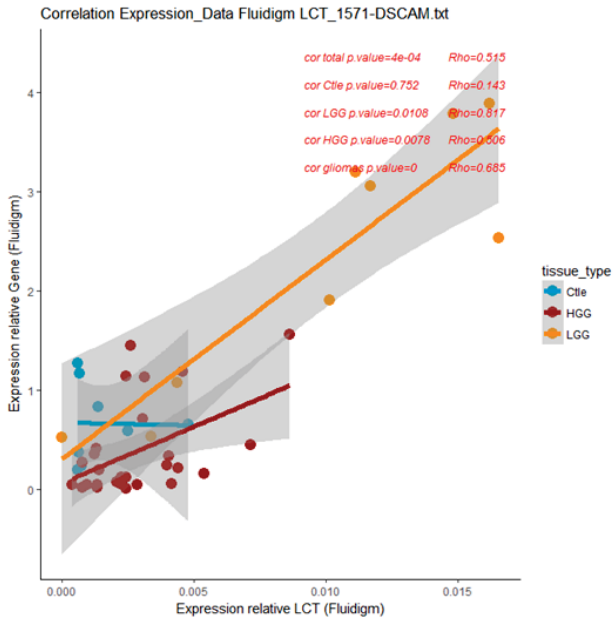


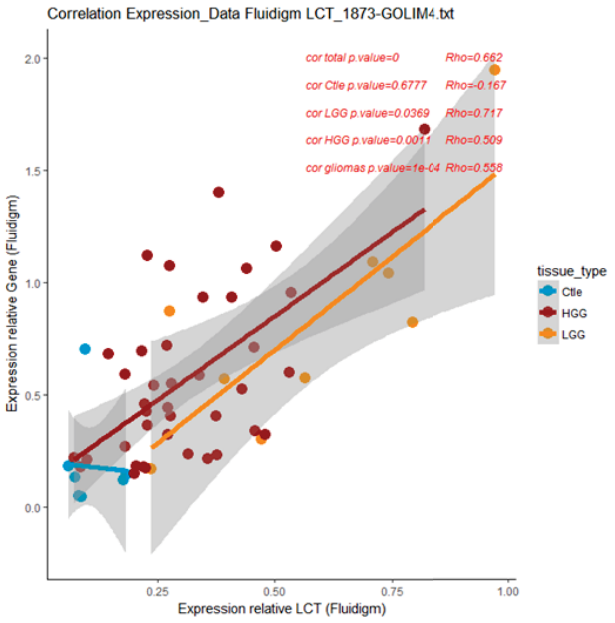


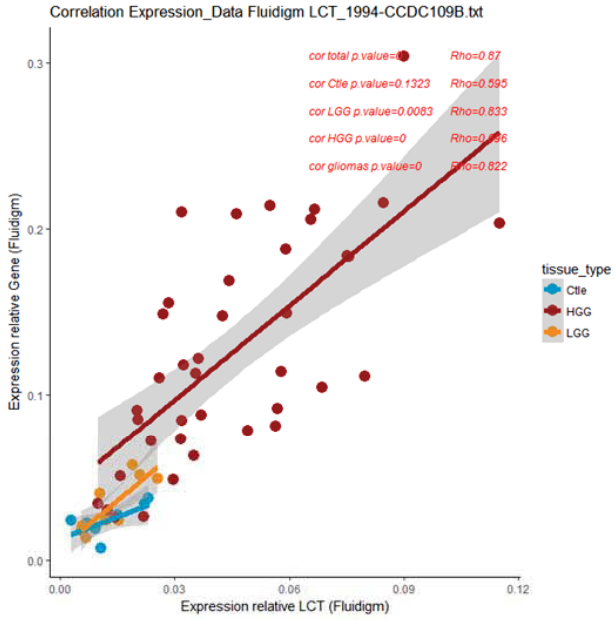


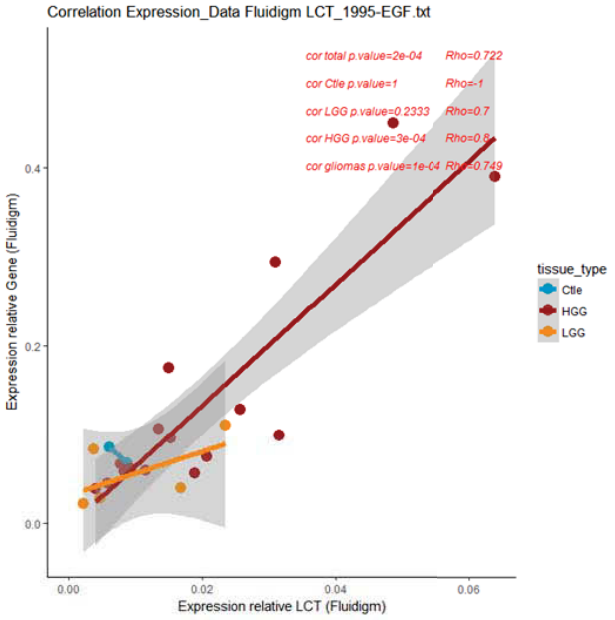


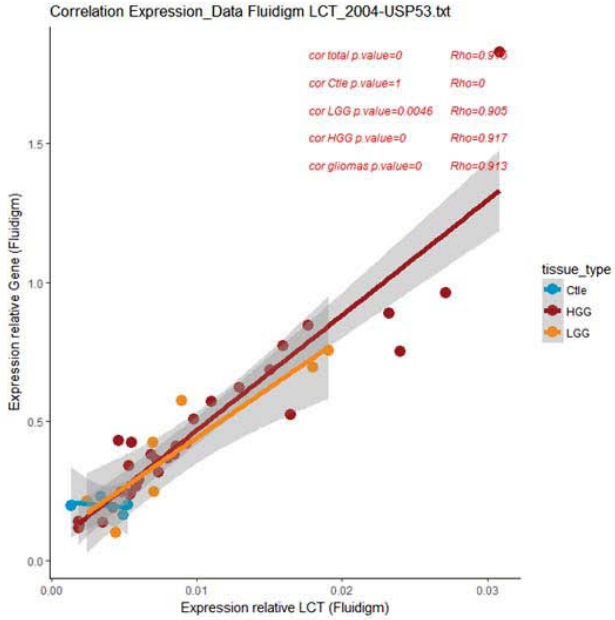


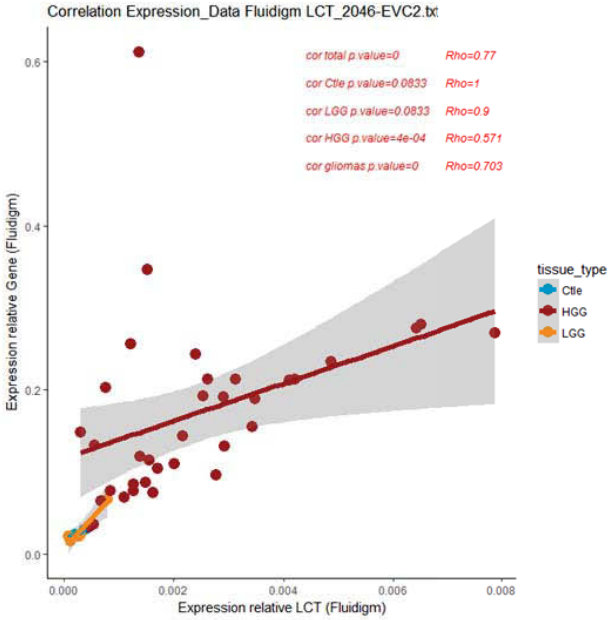

Correlation Expression_Data Fluidigm LCT_1571-DSCAM.txt
cor total p.value=4e-04 Rho=0.515
cor Ctle p.value=0.752 Rho=0.143
cor LGG p.value=0.0108 Rho=0.817
cor HGG p.value=0.0078 Rho=0.506
cor gliomas p.value=0 Rho=0.685
tissue_type
Ctle
HGG
LGG
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)

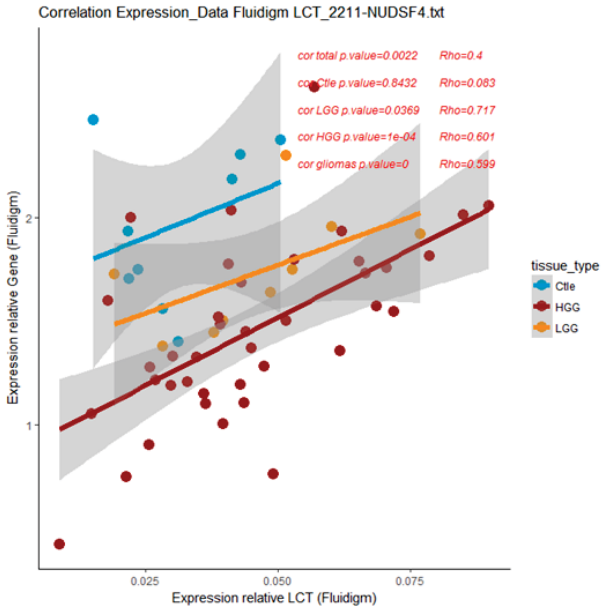

Correlation Expression_Data Fluidigm LCT_1873-GOLIM4.txt
cor total p.value=0 Rho=0.662
cor Ctle p.value=0.6777 Rho=-0.167
cor LGG p.value=0.0369 Rho=0.717
cor HGG p.value=0.0011 Rho=0.509
cor gliomas p.value=1e-04 Rho=0.558
tissue_type
Ctle
HGG
LGG
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)

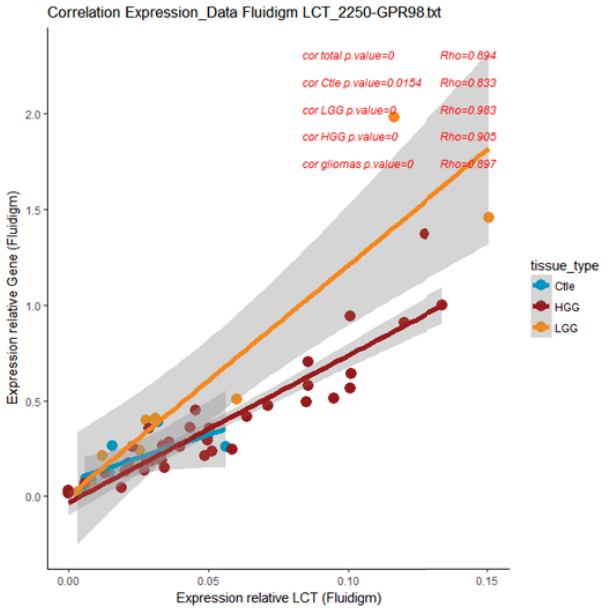

Correlation Expression_Data Fluidigm LCT_1994-CCDC109B.txt
cor Ctle p.value=0.1323 Rho=0.595
cor LGG p.value=0.0083 Rho=0.833
cor gliomas p.value=0 Rho=0.822
tissue_type
Ctle
HGG
LGG
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)

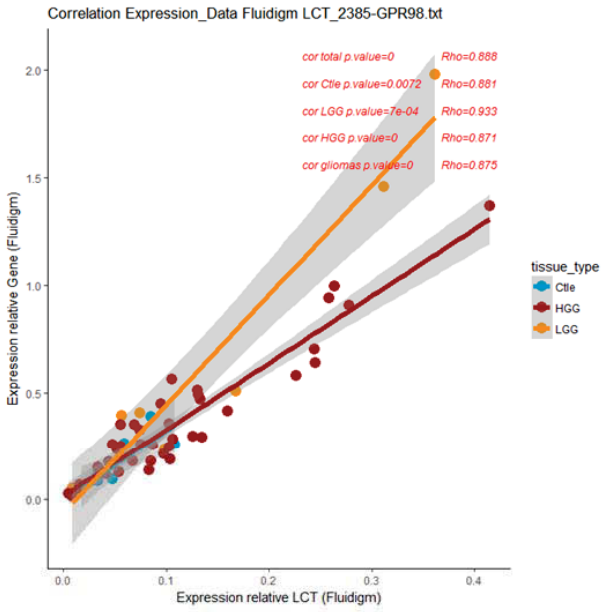

Correlation Expression_Data Fluidigm LCT_1995-EGF.txt
cor total p.value=2e-04 Rho=0.722
cor Ctle p.value=1 Rho=-1
cor LGG p.value=0.2333 Rho=0.7
cor HGG p.value=3e-04
cor gliomas p.value=1e-04 Rho=0.749
tissue_type
Ctle
HGG
LGG
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)

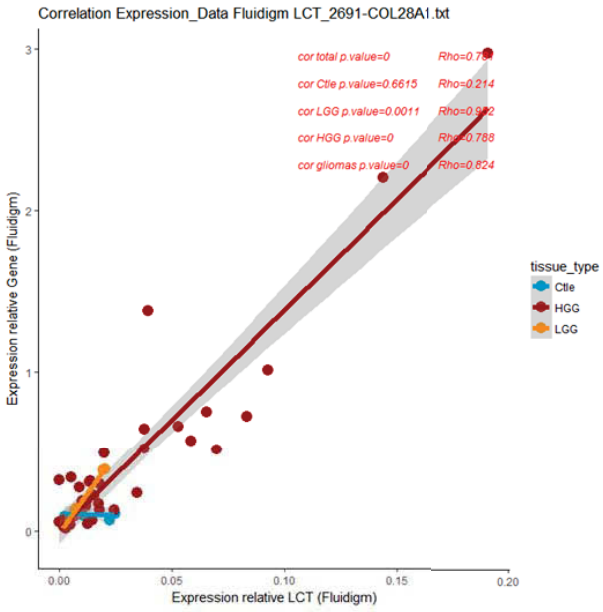

Correlation Expression_Data Fluidigm LCT_2004-USP53.txt
cor Ctle p.value=1 Rho=0
cor LGG p.value=0.0046 Rho=0.905
cor HGG p.value=0 Rho=0.917
cor gliomas p.value=0 Rho=0.913
tissue_type
Ctle
HGG
LGG
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)

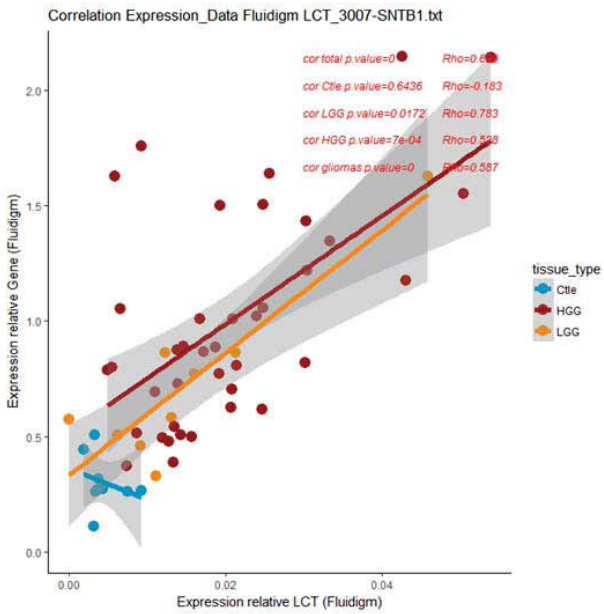

Correlation Expression_Data Fluidigm LCT_2046-EVC2.txt
cor total p.value=0 Rho=0.77
cor Ctle p.value=0.0833 Rho=1
cor LGG p.value=0.0833 Rho=0.9
cor HGG p.value=4e-04 Rho=0.571
cor gliomas p.value=0 Rho=0.703
tissue_type
Ctle
HGG
LGG
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)


Correlation Expression_Data Fluidigm LCT_2211-NUDSF4.txt
cor total p.value=0.0022 Rho=0.4
cor Ctle p.value=0.8432 Rho=0.083
cor LGG p.value=0.0369 Rho=0.717
cor HGG p.value=1e-04 Rho=0.601
cor gliomas p.value=0 Rho=0.599
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG


Correlation Expression_Data Fluidigm LCT_2250-GPR98.txt
cor total p.value=0 Rho=0.894
cor Ctle p.value=0.0154 Rho=0.833
cor LGG p.value=0 Rho=0.983
cor HGG p.value=0 Rho=0.905
cor gliomas p.value=0 Rho=0.897
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG


Correlation Expression_Data Fluidigm LCT_2385-GPR98.txt
cor total p.value=0 Rho=0.888
cor Ctle p.value=0.0072 Rho=0.881
cor LGG p.value=7e-04 Rho=0.933
cor HGG p.value=0 Rho=0.871
cor gliomas p.value=0 Rho=0.875
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG


Correlation Expression_Data Fluidigm LCT_2691-COL28A1.txt
cor total p.value=0 Rho=0.7
cor Ctle p.value=0.6615 Rho=0.214
cor LGG p.value=0.0011 Rho=0.942
cor HGG p.value=0 Rho=0.788
cor gliomas p.value=0 Rho=0.824
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG

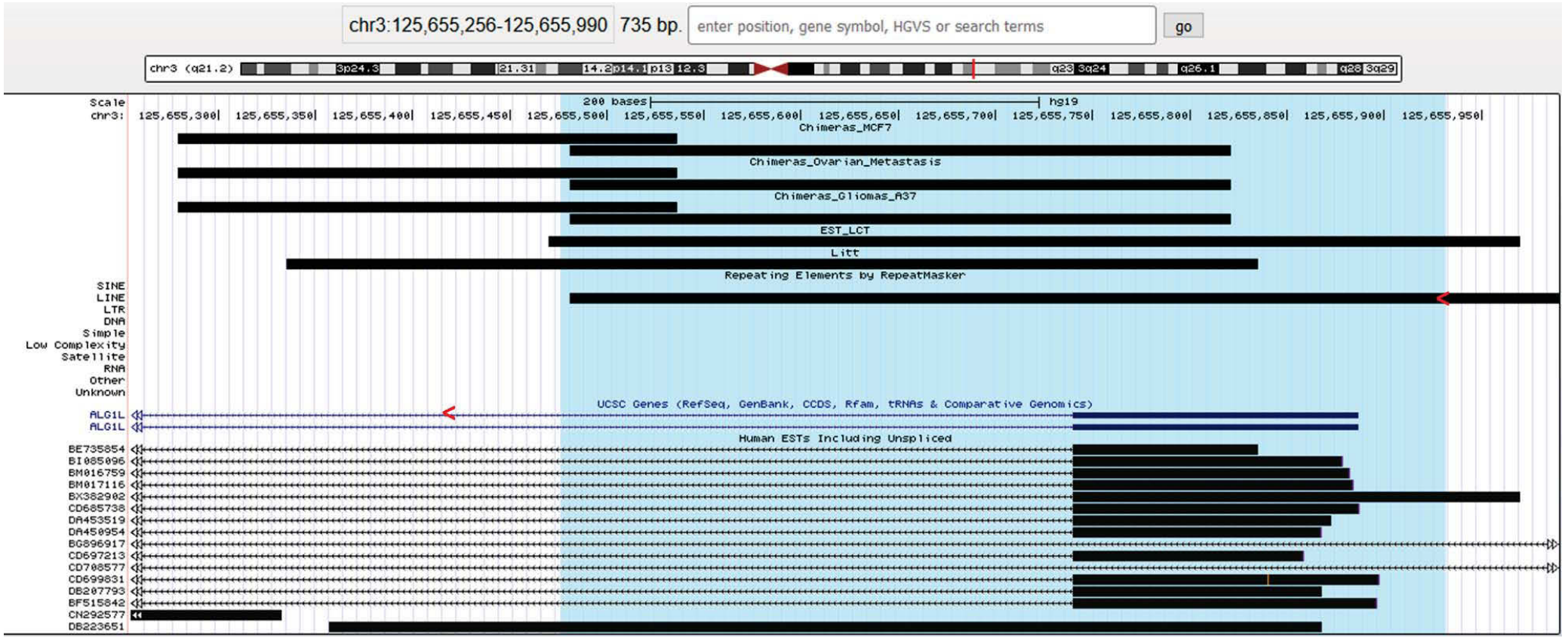

Correlation Expression_Data Fluidigm LCT_3007-SNTB1.txt
cor Ctle p.value=0.6436 Rho=-0.183
cor LGG p.value=0.0172 Rho=0.783
cor gliomas p.value=0 Rho=0.587
Expression relative Gene (Fluidigm)
Expression relative LCT (Fluidigm)
tissue_type
Ctle
HGG
LGG

Annexe 12 : Caractéristiques des 16 chimères communes aux 3 analyses A37, Métastases ovariennes et MCF7.

Caractéristiques	Chimera.Chr	Chimera.Start	Chimera.End	Chimera.Strand	ID_final	GeneName	repName	repStrand
TSS L1	chr21	35303560	35304120	+	ID_3336	LINC00649	L1PA2	-
	chr3	125655279	125655821	-	ID_775	ALG1L	L1PA2	+
	chr2	68907102	68937793	+	ID_342	ARHGAP25	L1PA3	-
Intron sens	chr5	65359074	65359896	+	ID_1280	ERBB2IP	L1PA4	-
	chr6	76614997	76615984	+	ID_1552	MYO6	L1PA3	-
Jct sens	chr2	71644194	71650165	+	LCT_1277	ZNF638	L1HS	-
	chr10	33200816	33201447	-	ID_2333	ITGB1	L1PA5	+
	chr8	17151959	17152747	+	ID_1977	VPS37A	L1PA7	-
Intron AS	chr4	149244845	149245320	+	ID_1157	NR3C2	L1PA2	-
	chr5	115526502	115527004	-	ID_1393	COMMD10	L1PB1	+
	chr15	43467693	43468263	-	ID_3062	TMEM62	L1P1	+
Jct AS	chr11	14865561	14866264	-	ID_2468	PDE3B	L1PA3	+
	chr13	49039298	49040051	-	LCT_801	RB1	L1HS	+
	chr8	26252727	26253486	-	ID_1984	BNIP3L	L1PA2	+
Intergénique	chrX	53173868	53174670	-	ID_3449	no_gene	L1PA4	+
	chr1	119562957	119563586	-	ID_120	no_gene	L1PA3	+

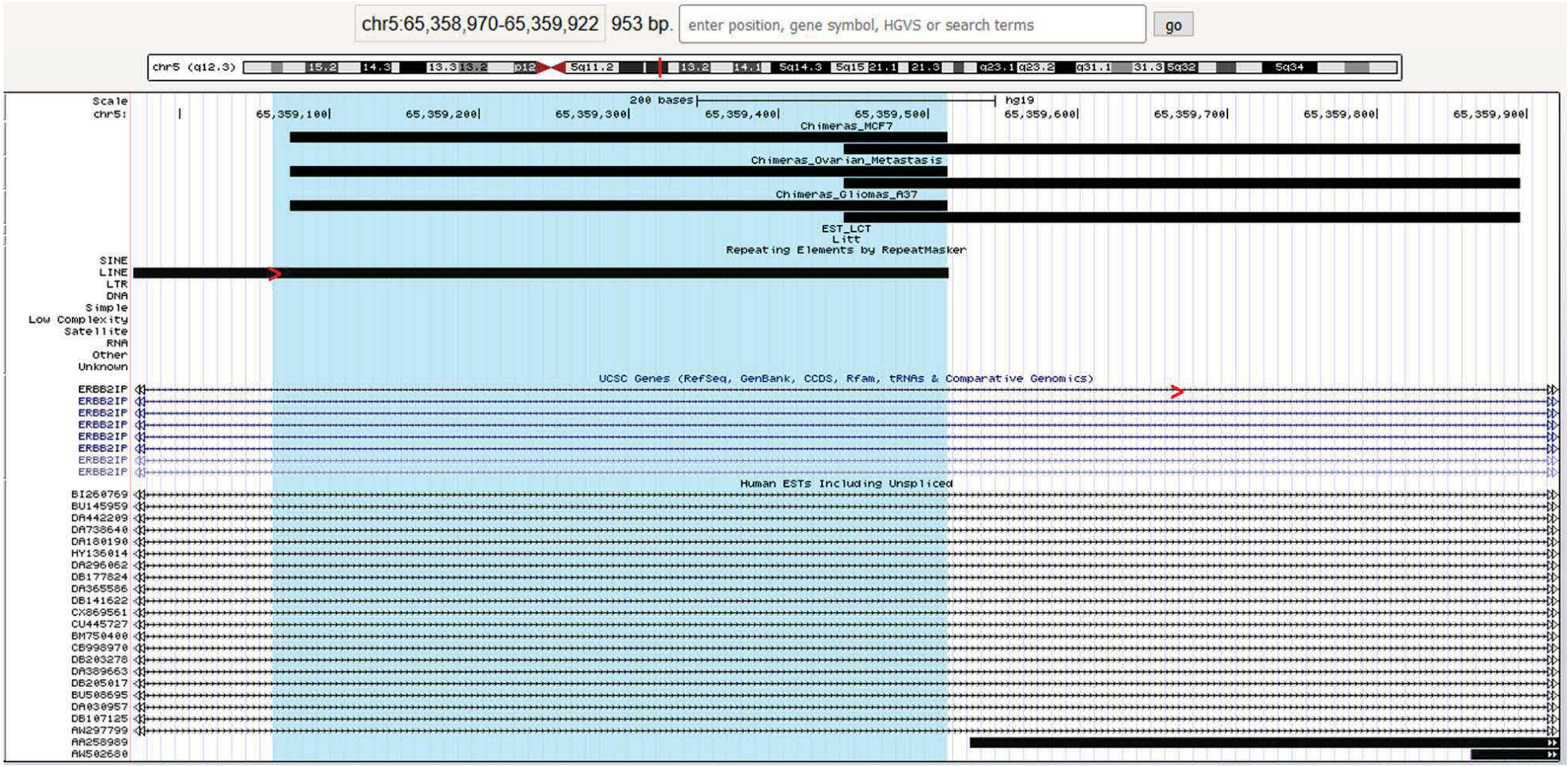
*Annexe 13 : **Chimère 3336** dont l'ASP sert de promoteur au gène LINC00649.*

Région de la chimère 3336, initiée à partir de l'ASP du L1PA2 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Cet ASP sert de TSS au gène *LINC00649*. Des évidences d'EST sont présentes de par la présence du TSS. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.



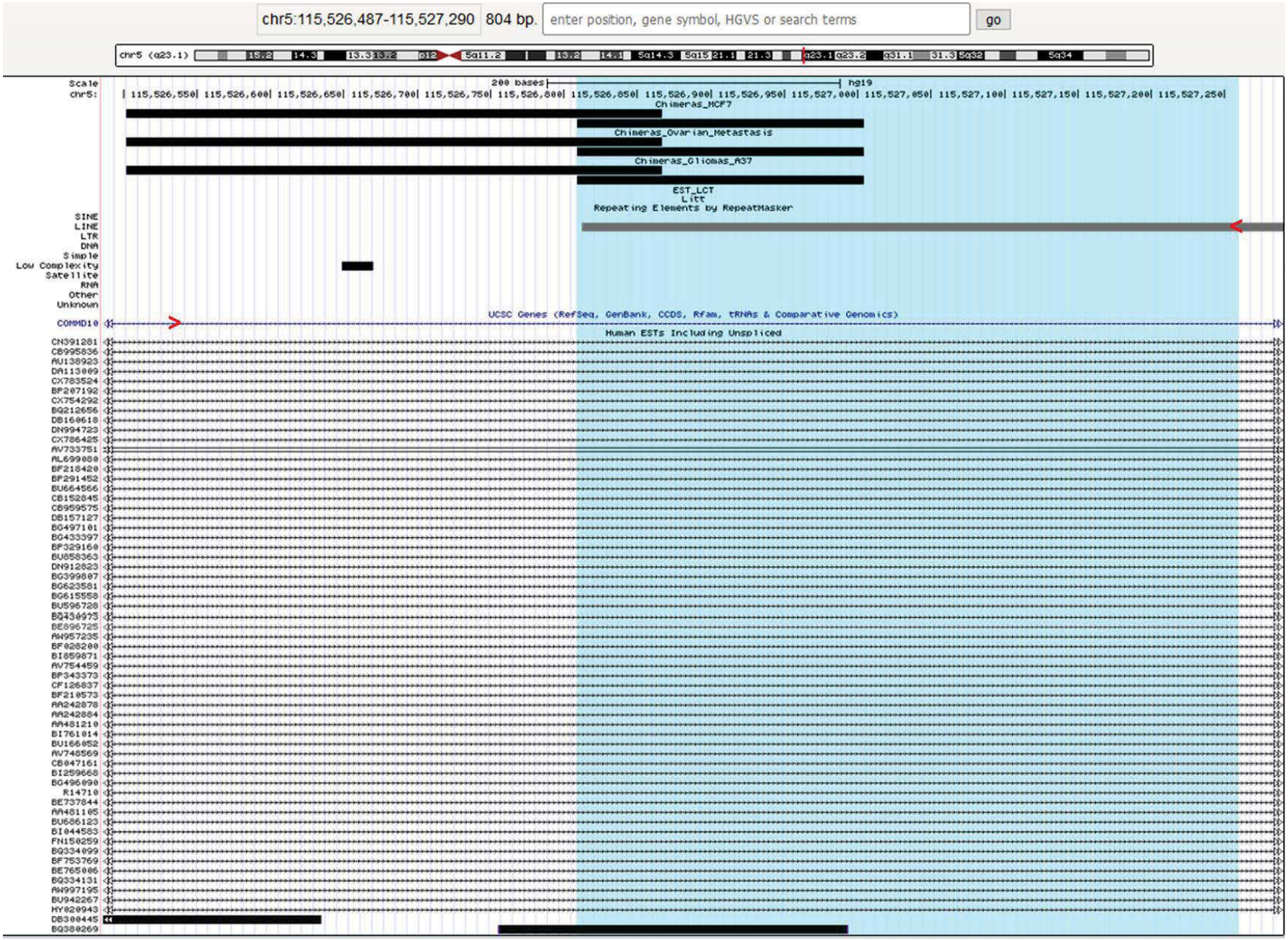
Annexe 14 : Chimère 775 dont l'ASP sert de promoteur au gène ALG1L.

Région de la chimère 775, initiée à partir de l'ASP du L1PA2 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Cet ASP sert de TSS au gène *ALG1L*. Des évidences d'EST sont présentes de par la présence du TSS. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.



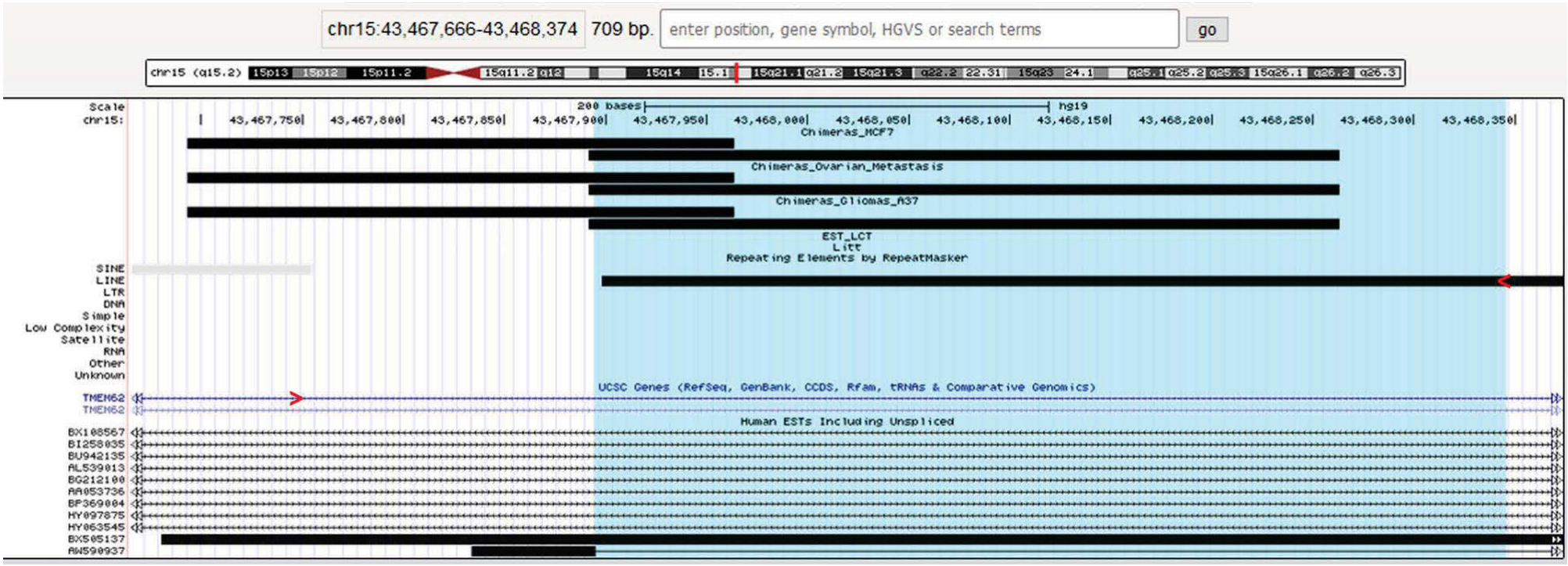
Annexe 15 : Chimère 1280 initiée à l'ASP d'un L1PA4 localisé dans un intron du gène ERBB2IP en sens.

Région de la chimère 1280 qui semble initiée à partir de l'ASP du L1PA4 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 21/25 du gène *ERBB2IP*. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes. 1 EST initié au site +1 du début du L1.



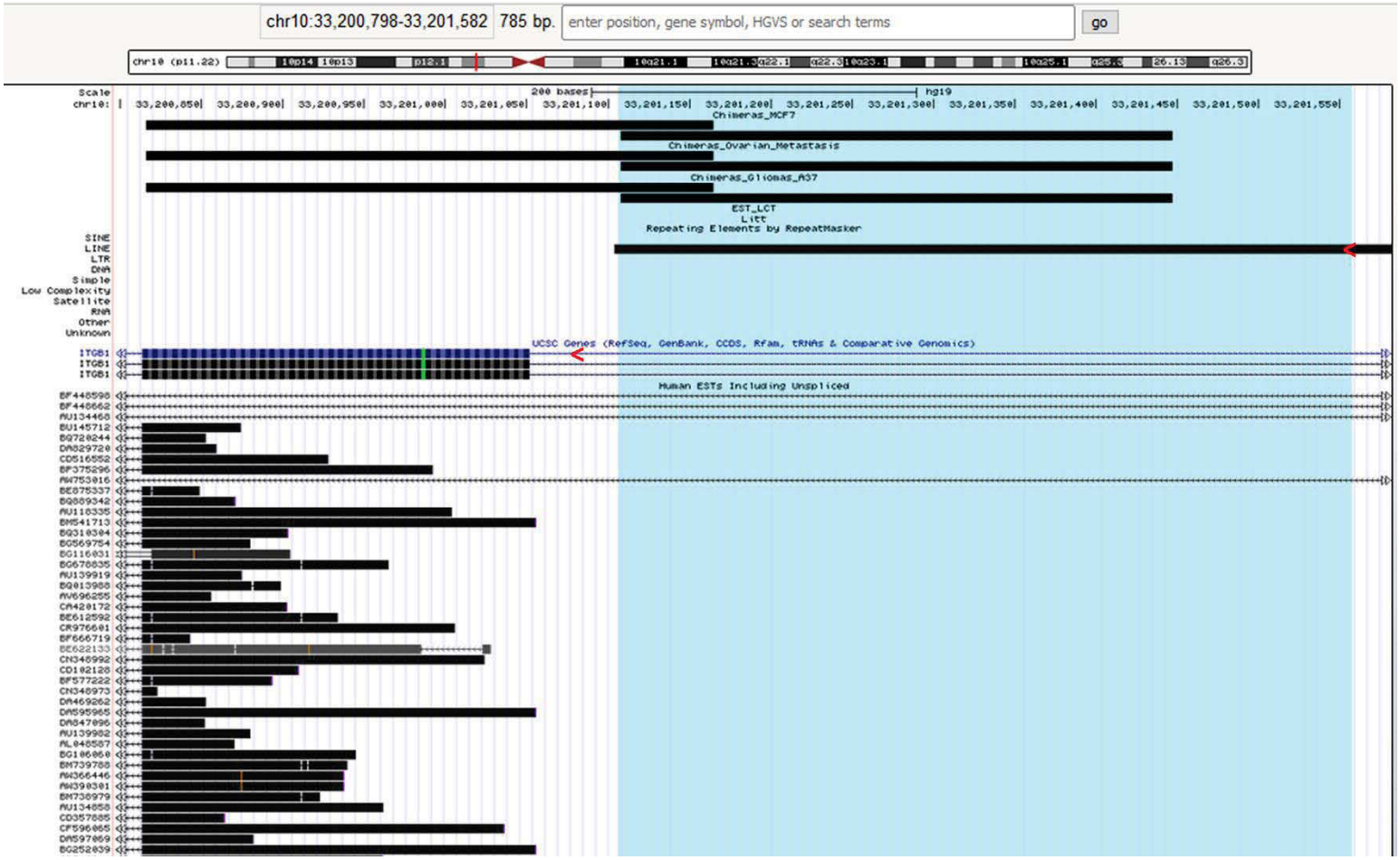
Annexe 16 : Chimère 1393 initiée à l'ASP d'un L1PB1 localisé dans un intron du gène COMMD10 en antisens.

Région de la chimère 1393 qui semble initiée à partir d'un L1PB1 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 5/6 du gène *COMMD10*. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes. 1 EST est détecté dans la région du L1 et se poursuit dans la région génomique adjacente.



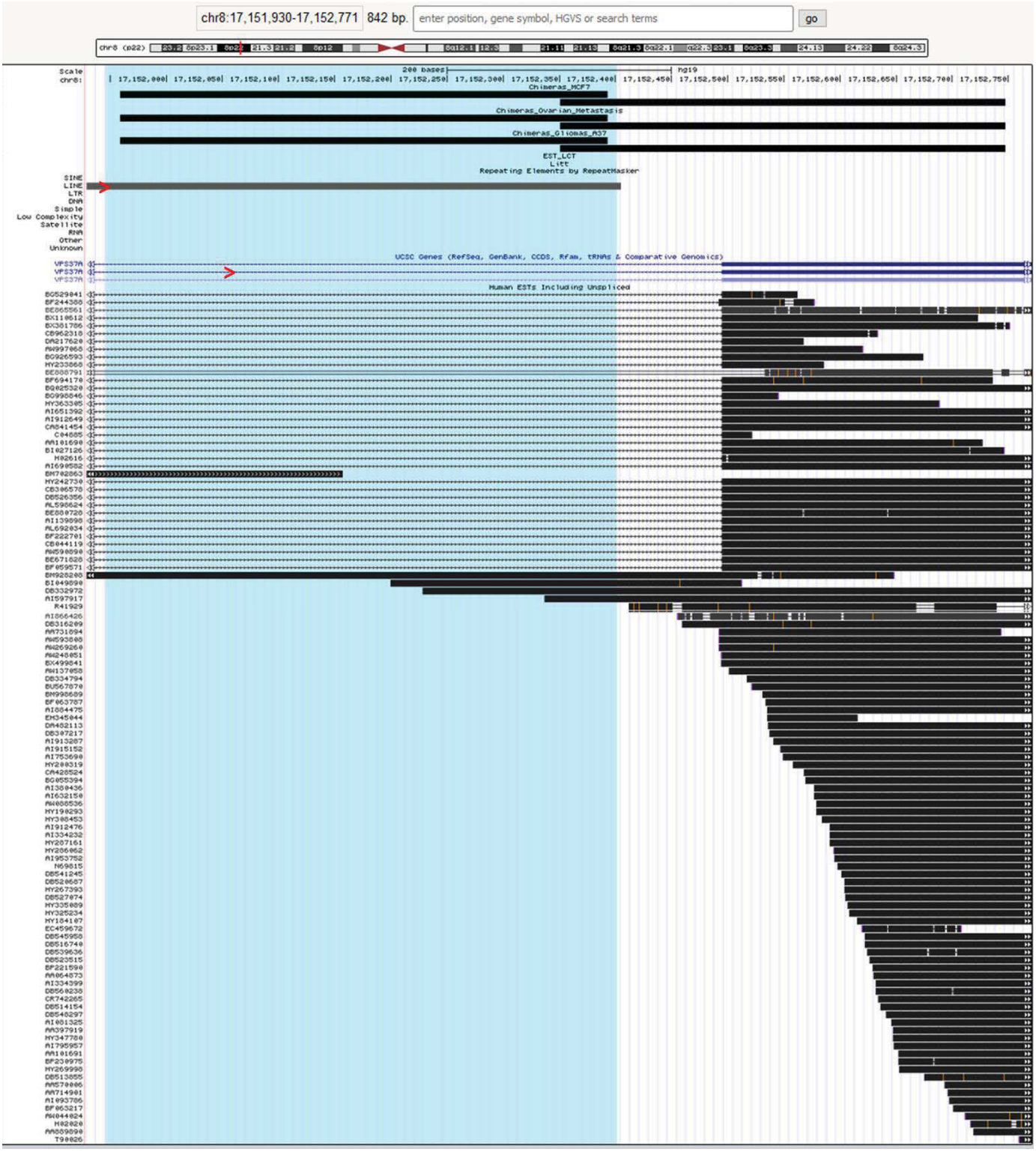
Annexe 17 : Chimère 3062 initiée à l'ASP d'un L1P1 localisé dans un intron du gène TMEM62 en antisens.

Région de la chimère 3062 qui semble initiée à partir de l'ASP du L1P1 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 11/13 du gène *TMEM62*. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.



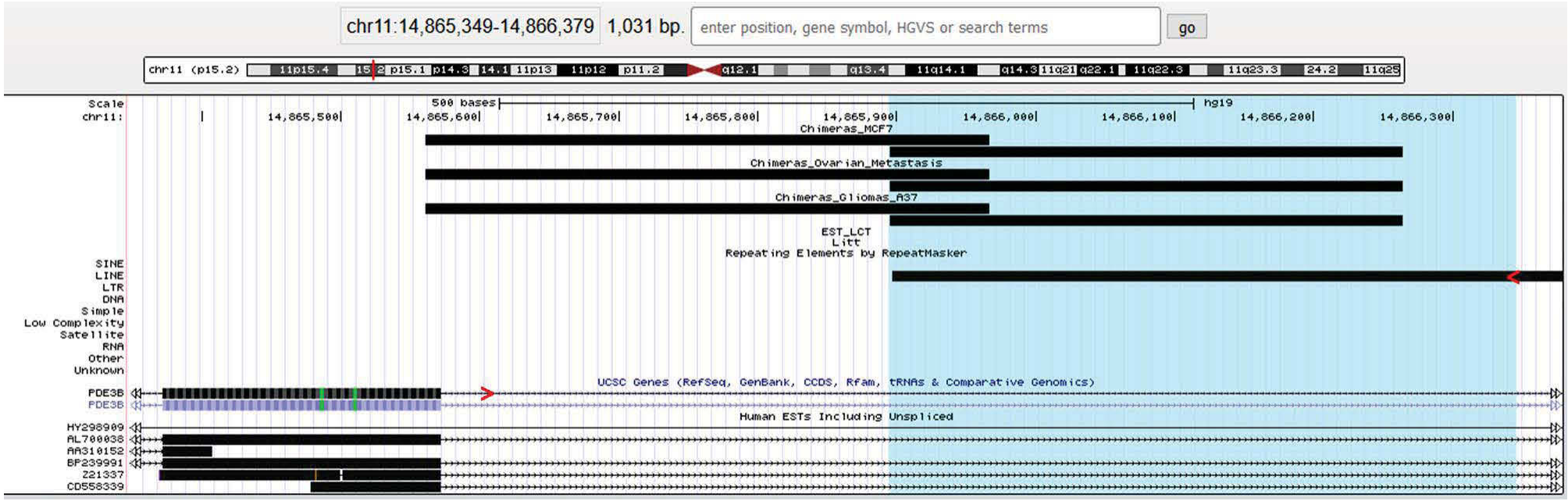
Annexe 18 : Chimère 2333 initiée à l'ASP d'un L1PA5, localisée à la jonction intron/exon du gène ITGB1 en sens.

Région de la chimère 2333, initiée à partir de l'ASP du L1PA5 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 11/15 du gène *ITGB1* et le R2 est localisé dans l'exon 12/16. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.



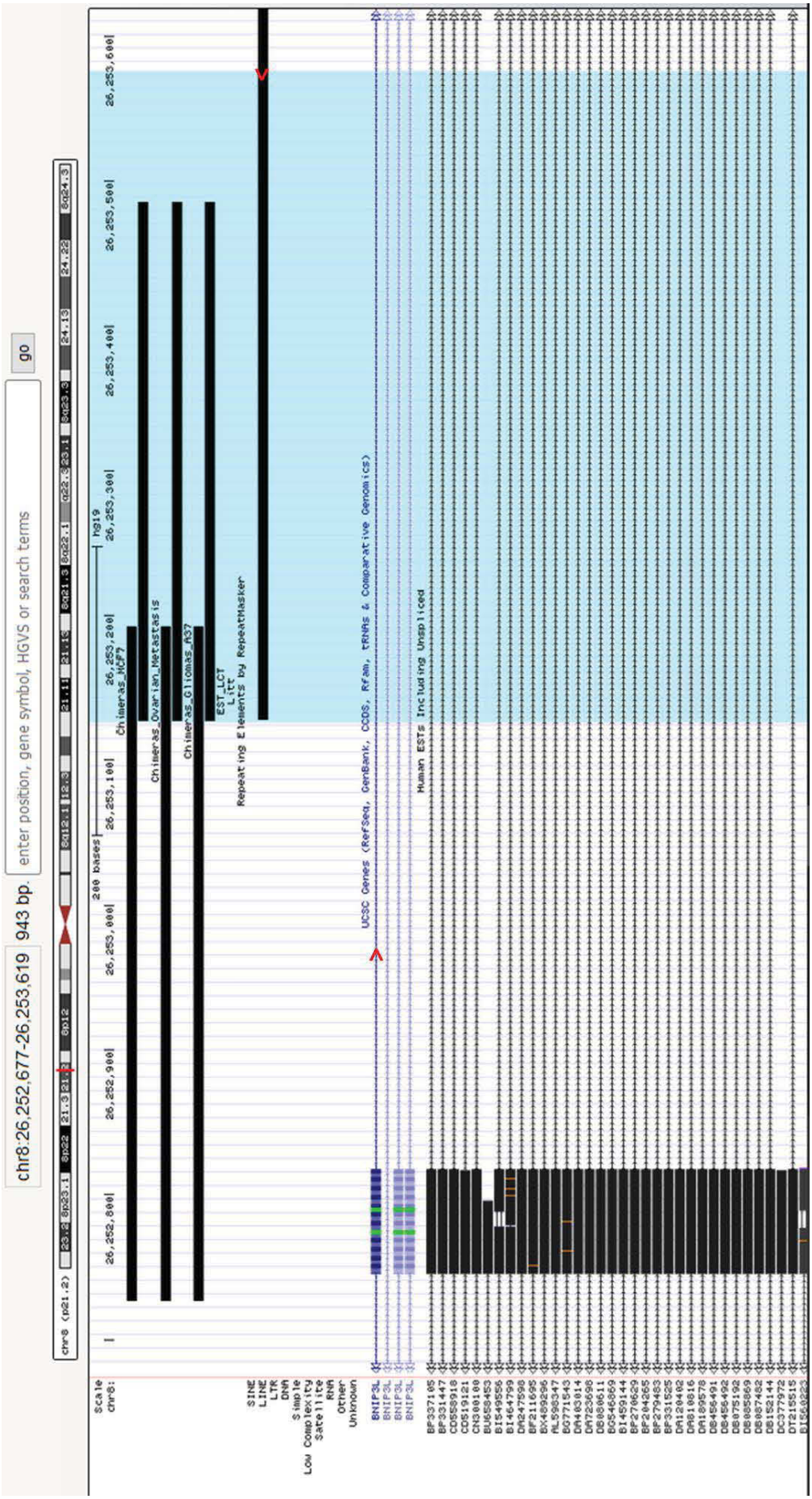
Annexe 19: Chimère 1977 initiée à l'ASP d'un L1PA7, localisée à la jonction intron/exon du gène VPS37A en sens.

Région de la chimère 1977, initiée à partir de l'ASP du L1PA7 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 11/11 du gène *VPS37A* et le R2 est localisé dans l'exon 12/12. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes. 3 EST sont décrites comme débutant dans le L1 et se poursuit dans la région adjacente.



Annexe 20: Chimère 2468 initiée à l'ASP d'un L1PA3, localisée à la jonction intron/exon du gène PDE3B en antisens.

Région de la chimère 2468, initiée à partir de l'ASP du L1PA3 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquant le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 12/15 du gène *PDE3B* et le R2 est localisé dans l'exon 12/16. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.



Annexe 21 : Chimère 1984 initiée à l'ASP d'un L1PA2, localisée à la jonction intron/exon du gène BNIP3L en antisens.

Région de la chimère 1984, initiée à partir de l'ASP du L1PA2 dont la région 5'UTR est surlignée en bleu. Les flèches rouges indiquent le sens de la transcription du gène et de l'ASP. Ce L1 est localisé dans l'intron 3/5 du gène *BNIP3L* et le R2 est localisé dans l'exon 3/6. Cette chimère est détectée par CLIFinder dans les données de RNA-seq de MCF7, de métastases ovariennes et des gliomes.

sujet	groupe_corr OMS	Groupe tumoral	Composante	Survie 1/0 (1=mort)	SURVIE (a)	IDH (1=mut)	Méthylation HM450K	RNA-seq	Fluidigm
CC1981	CTRL	CTRL	-	-	-	0		X	X
CC1011	CTRL	CTRL	-	-	-	0		X	
CC1465	CTRL	CTRL	-	-	-	0			X
CC3608	CTRL	CTRL	-	-	-	0			X
CC3809	CTRL	CTRL	-	-	-	0			X
CF1409	CTRL	CTRL	-	-	-	0			X
CF1465	CTRL	CTRL	-	-	-	0			X
CF3702	CTRL	CTRL	-	-	-	0			X
CF1	CTRL	CTRL	-	-	-	0			X
CF3	CTRL	CTRL	-	-	-	0			X
CF4	CTRL	CTRL	-	-	-	0		X	X
GII 1	GII	LGG	oligo	0	2,71047	1			X
GII 2	GIII	LGG	oligo	0	6,36277	1	X	X	X
GII 3	GII	LGG	oligo	0	6,59274	1	X		X
GII 4	GIII	LGG	oligo	1	0,00548	1	X	X	X
GII 5	GII	LGG	oligo	0	5,00479	1	X	X	X
GII 6	GIII	LGG	oligo	1	1,48118	1	X		
GII 7	GII	LGG	oligo	0	4,58042	1	X		
GII 8	GIII	LGG	oligo	0	4,91444	1			
GII 9	GII	LGG	oligo	0	3,68789	1	X		
GII 10	GII	LGG	oligo	0	3,43326	1	X		
GII 11	GIII	HGG	oligo	1	0,90623	0			
GII 12	GII	LGG	oligo	0	3,50719	1	X		
GII 13	GII	LGG	oligo	0	3,33744	1	X		X
GII 14	GII	LGG	oligo	0	3,36208	1	X		
GIII 1	GIII	LGG	oligo	0	4,56674	1	X		X
GIII 2	GIII	HGG	oligo astro	1	0,46543	0	X		
GIII 3	GIII	HGG	oligo astro	1	0,91992	0	X		
GIII 4	GIII	HGG	oligo astro	1	1,43463	0	X	X	X
GIII 5	GIII	HGG	oligo	0	5,3306	0			X
GIII 6	GIII	HGG	oligo	1	1,89185	0			
GIII 7	GIII	HGG	oligo	1	2,02601	0			X
GIII 8	GIII	HGG	oligo astro	1	1,2512	0	X		
GIII 9	GIII	HGG	oligo	1	0,66256	0			X
GIII 10	GIII	HGG	oligo astro	1	1,24298	0	X		
GIII 11	GIII	HGG	oligo astro	1	1,02669	0	X	X	X
GIII 12	GIII	HGG	oligo astro	0	4,38877	0	X	X	X
GIII 13	GIII	HGG	oligo	1	1,51403	0			
GIII 14	GIII	HGG	oligo	1	1,71937	0			
GIII 15	GIII	HGG	oligo	1	1,20192	0			X
GIII 16	GIII	HGG	oligo	1	0,87611	0			X
GIII 17	GIII	HGG	oligo astro	1	1,54962	0	X	X	X
GIII 18	GIII	HGG	oligo	1	1,02396	0			X
GIII 19	GIII	HGG	oligo	1	0,3614	0			X
GIII 20	GIII	HGG	oligo astro	1	0,72005	0	X		
GIII 21	GIII	HGG	oligo astro	1	0,81862	0			
GIII 22	GIII	LGG	oligo	1	1,59069	1	X	X	X
GIII 23	GIII	HGG	oligo astro	1	2,45038	0			X
GIII 24	GIII	LGG	oligo	0	4,44627	1	X	X	X

Annexe 22 : Récapitulatif des caractéristiques relatives aux gliomes et aux échantillons contrôles utilisés dans l'étude. (1/2)

Sujet : nom de l'échantillon après anonymisation du patient. Group_corr_OMS : grade anatomo-pathologique selon la classification de l'OMS. Groupe Tumoral : classification utilisée dans le manuscrit selon le statut de la mutation du gène *IDH1*. Composante : composante tissulaire correspondant à l'origine du type tumoral. Survie des patients : 1 mort, 0 en vie et durée de survie après diagnostic en années. La dernière mise à jour des informations relatives aux patients date de novembre 2015. IDH : statut de mutation IDH 0 non muté, 1 muté. Les suivantes reportent les échantillons pour lesquels les données de méthylation sont disponibles à partir des puces HM450K, des données de RNA-seq et de Fluidigm.

sujet	groupe_corr OMS	Groupe tumoral	Composante	Survie 1/0 (1=mort)	SURVIE (a)	IDH (1=mut)	Méthylation HM450K	RNA-seq	Fluidigm
GIII 25	GIII	HGG	oligo	0	5,52498	0			
GIII 26	GIII	HGG	oligo astro	1	0,4627	0	X		X
GIII 27	GIII	HGG	oligo astro	1	1,8371	0	X		
GIII 28	GIII	HGG	oligo astro	1	0,0219	0			
GIII 29	GIII	HGG	oligo	1	0,87064	0			X
GIII 30	GIII	HGG	oligo astro	1	2,02875	0			X
GIII 31	GIII	HGG	oligo	1	1,97399	0			X
GIII 32	GIII	LGG	oligo	0	4,3039	1	X		X
GIII 33	GIII	HGG	oligo astro	1	0,71184	0	X		X
GIII 34	GIII	HGG	oligo	1	1,20192	0			X
GIII 35	GIII	HGG	oligo astro	1	2,38467	0	X		X
GIII 36	GIII	HGG	oligo	1	3,37851	0			X
GIII 37	GIII	HGG	oligo	1	0,42984	0			X
GIII 38	GIII	HGG	oligo astro	1	0,5859	0			X
GBM 1	GIV	HGG		0	3,48802	0			X
GBM 2	GIV	HGG		0	3,99179	0			
GBM 3	GIV	HGG		0	3,60575	0			
GBM 4	GIV	HGG		1	0,61602	0	X	X	X
GBM 5	GIV	HGG		1	0,28474	0			
GBM 6	GIV	HGG		1	0,73648	0	X		
GBM 7	GIV	HGG		1	2,11088	0	X	X	X
GBM 8	GIV	HGG		1	2,05339	0	X		
GBM 9	GIV	HGG		1	1,10335	0			X
GBM 10	GIV	HGG		1	0,93908	0			
GBM 11	GIV	HGG		1	3,78371	0	X		
GBM 12	GIV	HGG		1	1,41821	0	X		
GBM 13	GIV	HGG		1	1,29774	0			
GBM 14	GIV	HGG		1	0,15058	0	X		X
GBM 15	GIV	HGG		0	6,38467	0	X		
GBM 16	GIV	HGG		1	2,23135	0	X		
GBM 17	GIV	HGG		0	3,24983	0	X		X
GBM 18	GIV	HGG		1	1,30595	0			X
GBM 19	GIV	HGG		1	1,49213	0	X		X
GBM 20	GIV	HGG		1	0,24641	0	X		X
GBM 21	GIV	HGG		1	1,19918	0			X
GBM 22	GIV	HGG		1	2,02875	0	X		
GBM 23	GIV	HGG		1	3,10472	0			
GBM 24	GIV	HGG		1	1,42368	0			X
GBM 25	GIV	HGG		1	3,49897	0	X		X
GBM 26	GIV	HGG		1	2,99521	0			
GBM 27	GIV	HGG		1	1,12799	0	X	X	X
GBM 28	GIV	HGG		1	1,43737	0			
GBM 29	GIV	HGG		1	0,09035	0			X
GBM 30	GIV	HGG		1	0,01095	0			X
GBM 31	GIV	HGG		1	1,70568	0			X
GBM 32	GIV	HGG		1	1,07324	0	X		X
GBM 33	GIV	HGG		1	0,49555	0			X
GBM 34	GIV	HGG		1	1,31691	0	X	X	X
GBM 35	GIV	HGG		1	1,3744	0	X		X

Annexe 22 : Récapitulatif des caractéristiques relatives aux gliomes et aux échantillons contrôles utilisés dans l'étude. (2/2)

Sujet : nom de l'échantillon après anonymisation du patient. Group_corr_OMS : grade anatomo-pathologique selon la classification de l'OMS. Groupe Tumoral : classification utilisée dans le manuscrit selon le statut de la mutation du gène *IDH1*. Composante : composante tissulaire correspondant à l'origine du type tumoral. Survie des patients : 1 mort, 0 en vie et durée de survie après diagnostic en années. La dernière mise à jour des informations relatives aux patients date de novembre 2015. IDH : statut de mutation IDH 0 non muté, 1 muté. Les suivantes reportent les échantillons pour lesquels les données de méthylation sont disponibles à partir des puces HM450K, des données de RNA-seq et de Fluidigm.

ID	séquence (5'-NNN-3')	taille (pb)
LCT1873-RACE1	CTTTTCCTAATTTCAGGTGGGCCTGTT	851
LCT1873-RACE2	GCACACCCACAAACACTCAAGTGCTCA	735
LCT1994_RACE1	AAGGTTTTGATCTCAATACTTATGCAACACTT	601
LCT837-RACE1	TGGCTTCCTTCTATAGGCCTTTTCTCA	646
LCT837-RACE2	TGGGCTCAAAAGAGAATACAGTTTGGGTA	532
LCT78-RACE1	TGAATGTGAAATTTTTGGGAAGGGGAAC	648
LCT78-RACE2	GCCACCTTCTCTTTCCATCCCTAGAGACC	502
LCT2691_RACE1	CAGACCTACAAACTGGCTGTGTCCAA	794
LCT2691_RACE2	GCCAACCTCAGCAACTGCAGCAGAA	678
GeneRacer 5' Primer	CGACTGGAGCACGAGGACACTGA	
GeneRacer 5' Nested Primer	GGACACTGACATGGACTGAAGGAGTA	

Annexe 23 : Liste des amorces utilisées pour la validation de l'initiation de la transcription par 5'RACE.

La taille des amplifications correspond aux tailles attendues entre la position supposée de l'ASP et le positionnement de l'amorce dans la séquence unique.

BIBLIOGRAPHIE

ATHANIKAR, J. N., BADGE, R. M. & MORAN, J. V., (2004), A YY1-binding site is required for accurate human LINE-1 transcription initiation, *Nucleic Acids Research*, vol. 32, n°13, p. 3846- 3855.

BABA, Y., WATANABE, M., MURATA, A., SHIGAKI, H., MIYAKE, K., ISHIMOTO, T., … BABA, H., (2014), LINE-1 Hypomethylation, DNA Copy Number Alterations, and CDK6 Amplification in Esophageal Squamous Cell Carcinoma, *Clinical Cancer Research*, vol. 20, n°5, p. 1114- 1124.

BAILLIE, J. K., BARNETT, M. W., UPTON, K. R., GERHARDT, D. J., RICHMOND, T. A., DE SAPIO, F., … FAULKNER, G. J., (2011), Somatic retrotransposition alters the genetic landscape of the human brain, *Nature*, vol. 479, n°7374, p. 534- 537.

BECK, C. R., GARCIA-PEREZ, J. L., BADGE, R. M. & MORAN, J. V., (2011), LINE-1 Elements in Structural Variation and Disease, *Annual Review of Genomics and Human Genetics*, vol. 12, n°1, p. 187- 215.

BELANCIO, V. P., (2006), LINE-1 RNA splicing and influences on mammalian gene expression, *Nucleic Acids Research*, vol. 34, n°5, p. 1512- 1521.

BELANCIO, V. P., ROY-ENGEL, A. M., POCHAMPALLY, R. R. & DEININGER, P., (2010), Somatic expression of LINE-1 elements in human tissues, *Nucleic Acids Research*, vol. 38, n°12, p. 3909- 3922.

BERGHOFF, A. S., HAINFELLNER, J. A., MAROSI, C. & PREUSSER, M., (2015), Assessing MGMT methylation status and its current impact on treatment in glioblastoma, *CNS Oncology*, vol. 4, n°1, p. 47- 52.

BROUHA, B., SCHUSTAK, J., BADGE, R. M., LUTZ-PRIGGE, S., FARLEY, A. H., MORAN, J. V. & KAZAZIAN, H. H., (2003), Hot L1s account for the bulk of retrotransposition in the human population, *Proceedings of the National Academy of Sciences*, vol. 100, n°9, p. 5280–5285.

BURNS, K. H., (2017), Transposable elements in cancer, *Nature Reviews Cancer*, vol. 17, n°7, p. 415- 424.

BURWINKEL, B. & KILIMANN, M. W., (1998), Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease11Edited by J. Karn, *Journal of Molecular Biology*, vol. 277, n°3, p. 513- 517.

CADIEUX, B., CHING, T.-T., VANDENBERG, S. R. & COSTELLO, J. F., (2006), Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation, *Cancer research*, vol. 66, n°17, p. 8469–8476.

CASTRO-DIAZ, N., ECCO, G., COLUCCIO, A., KAPOPOULOU, A., YAZDANPANAH, B., FRIEDLI, M., … TRONO, D., (2014), Evolutionally dynamic L1 regulation in embryonic stem cells, *Genes & Development*, vol. 28, n°13, p. 1397- 1409.

CHANG, W.-S. W., CHANG, N.-T., LIN, S.-C., WU, C.-W. & WU, F. Y.-H., (2000), Tissue-specific cancer-related serpin gene cluster at human chromosome band 3q26, *Genes, Chromosomes and Cancer*, vol. 29, n°3, p. 240–255.

CHUONG, E. B., ELDE, N. C. & FESCHOTTE, C., (2016), Regulatory evolution of innate immunity through co-option of endogenous retroviruses, *Science*, vol. 351, n°6277, p. 1083- 1087.

COHEN, C. J., LOCK, W. M. & MAGER, D. L., (2009), Endogenous retroviral LTRs as promoters for human genes: a critical assessment, *Gene*, vol. 448, n°2, p. 105- 114.

COUFAL, N. G., GARCIA-PEREZ, J. L., PENG, G. E., YEO, G. W., MU, Y., LOVCI, M. T., … GAGE, F. H., (2009), L1 retrotransposition in human neural progenitor cells, *Nature*, vol. 460, n°7259, p. 1127- 1131.

CRISCIONE, S. W., THEODOSAKIS, N., MICEVIC, G., CORNISH, T. C., BURNS, K. H., NERETTI, N. & RODIĆ, N., (2016), Genome-wide characterization of human L1 antisense promoter-driven transcripts, *BMC Genomics*, vol. 17, n°1, .

CRUICKSHANKS, H. A. & TUFARELLI, C., (2009), Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter, *Genomics*, vol. 94, n°6, p. 397- 406.

CRUICKSHANKS, H. A., VAFADAR-ISFAHANI, N., DUNICAN, D. S., LEE, A., SPROUL, D., LUND, J. N., … TUFARELLI, C., (2013), Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer, *Nucleic Acids Research*, vol. 41, n°14, p. 6857- 6869.

DEININGER, P., MORALES, M. E., WHITE, T. B., BADDOO, M., HEDGES, D. J., SERVANT, G., … BELANCIO, V. P., (2017), A comprehensive approach to expression of L1 loci, *Nucleic Acids Research*, vol. 45, n°5, p. e31.

DENLI, A. M., NARVAIZA, I., KERMAN, B. E., PENA, M., BENNER, C., MARCHETTO, M. C. N., … GAGE, F. H., (2015), Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity, *Cell*, vol. 163, n°3, p. 583- 593.

DOTZLAW, H., ALKHALAF, M. & MURPHY, L. C., (1992), Characterization of estrogen receptor variant mRNAs from human breast cancers., *Molecular Endocrinology*, vol. 6, n°5, p. 773- 785.

ELBARBARY, R. A., LUCAS, B. A. & MAQUAT, L. E., (2016), Retrotransposons as regulators of gene expression, *Science*, vol. 351, n°6274, p. aac7247-aac7247.

ERGÜN, S., BUSCHMANN, C., HEUKESHOVEN, J., DAMMANN, K., SCHNIEDERS, F., LAUKE, H., … SCHUMANN, G. G., (2004), Cell Type-specific Expression of LINE-1 Open Reading Frames 1 and 2 in Fetal and Adult Human Tissues, *Journal of Biological Chemistry*, vol. 279, n°26, p. 27753- 27763.

FAULKNER, G. J., KIMURA, Y., DAUB, C. O., WANI, S., PLESSY, C., IRVINE, K. M., … CARNINCI, P., (2009), The regulated retrotransposon transcriptome of mammalian cells, *Nature Genetics*, vol. 41, n°5, p. 563- 571.

FEINBERG, A. P. & VOGELSTEIN, B., (1983), Hypomethylation distinguishes genes of some human cancers from their normal counterparts, *Nature*, vol. 301, n°5895, p. 89- 92.

GAINETDINOV, I., SKVORTSOVA, Y., KONDRATIEVA, S., FUNIKOV, S. & AZHIKINA, T., (2017), Two modes of targeting transposable elements by piRNA pathway in human testis, *RNA (New York, N.Y.)*.

GARCIA-PEREZ, J. L., MARCHETTO, M. C. N., MUOTRI, A. R., COUFAL, N. G., GAGE, F. H., O'SHEA, K. S. & MORAN, J. V., (2007), LINE-1 retrotransposition in human embryonic stem cells, *Human Molecular Genetics*, vol. 16, n°13, p. 1569- 1577.

GASC, C. & PEYRET, P., (2017), Revealing large metagenomic regions through long DNA fragment hybridization capture, *Microbiome*, vol. 5, n°1, .

GASIOR, S. L., WAKEMAN, T. P., XU, B. & DEININGER, P. L., (2006), The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks, *Journal of Molecular Biology*, vol. 357, n°5, p. 1383- 1393.

GASTON, K. & FRIED, M., (1994), YY1 is involved in the regulation of the bi-directional promoter of the Surf-1 and Surf-2 genes, *FEBS Letters*, vol. 347, n°2- 3, p. 289- 294.

GOODIER, J. L. & KAZAZIAN, H. H., (2008), Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites, *Cell*, vol. 135, n°1, p. 23- 35.

HA, H., SONG, J., WANG, S., KAPUSTA, A., FESCHOTTE, C., CHEN, K. C. & XING, J., (2014), A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements, *BMC Genomics*, vol. 15, n°1, .

HANCKS, D. C. & KAZAZIAN, H. H., (2012), Active human retrotransposons: variation and disease, *Current Opinion in Genetics & Development*, vol. 22, n°3, p. 191- 203.

HATA, K. & SAKAKI, Y., (1997), Identification of critical CpG sites for repression of L1 transcription by DNA methylation, *Gene*, vol. 189, n°2, p. 227–234.

HULME, A. E., KULPA, D. A., PEREZ, J. L. G. & MORAN, J. V., (2006), The Impact of LINE-1 Retro transposition on the Human Genome, In : J. R. LUPSKI et P. STANKIEWICZ (éd.), *Genomic Disorders*, Humana Press, Totowa, NJ, p. 35- 55.

KAER, K., BRANOVETS, J., HALLIKMA, A., NIGUMANN, P. & SPEEK, M., (2011), Intronic L1 Retrotransposons and Nested Genes Cause Transcriptional Interference by Inducing Intron Retention, Exonization and Cryptic Polyadenylation, *PLoS ONE*, vol. 6, n°10, p. e26099.

KAESSMANN, H., VINCKENBOSCH, N. & LONG, M., (2009), RNA-based gene duplication: mechanistic and evolutionary insights, *Nature reviews. Genetics*, vol. 10, n°1, p. 19- 31.

KAPITONOV, V. V. & JURKA, J., (2005), RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons, *PLoS Biology*, vol. 3, n°6, p. e181.

KAZAZIAN, H. H., WONG, C., YOUSSOUFIAN, H., SCOTT, A. F., PHILLIPS, D. G. & ANTONARAKIS, S. E., (1988), Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man, *Nature*, vol. 332, n°6160, p. 164- 166.

KHAN, H., (2005), Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates, *Genome Research*, vol. 16, n°1, p. 78- 87.

KINOMOTO, M., KANNO, T., SHIMURA, M., ISHIZAKA, Y., KOJIMA, A., KURATA, T., … TOKUNAGA, K., (2007), All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition, *Nucleic Acids Research*, vol. 35, n°9, p. 2955- 2964.

KRETH, S., HEYN, J., GRAU, S., KRETZSCHMAR, H. A., EGENSPERGER, R. & KRETH, F. W., (2010), Identification of valid endogenous control genes for determining gene expression in human glioma, *Neuro-Oncology*, vol. 12, n°6, p. 570- 579.

LAI, C. B., ZHANG, Y., ROGERS, S. L. & MAGER, D. L., (2009), Creation of the two isoforms of rodent NKG2D was driven by a B1 retrotransposon insertion, *Nucleic Acids Research*, vol. 37, n°9, p. 3032- 3043.

LAMBERTZ, N., EL HINDY, N., KREITSCHMANN-ANDERMAHR, I., STEIN, K. P., DAMMANN, P., OEZKAN, N., … ZHU, Y., (2015), Downregulation of programmed cell death 10 is associated with tumor cell proliferation, hyperangiogenesis and peritumoral edema in human glioblastoma, *BMC Cancer*, vol. 15, n°1, .

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., … OTHERS, (2001), Initial sequencing and analysis of the human genome.

LAVERDURE, S., POLAKOWSKI, N., HOANG, K. & LEMASSON, I., (2016), Permissive Sense and Antisense Transcription from the 5′ and 3′ Long Terminal Repeats of Human T-Cell Leukemia Virus Type 1, *Journal of Virology*, vol. 90, n°7, p. 3600- 3610.

LAVIE, L., MALDENER, E., BROUHA, B., MEESE, E. U. & MAYER, J., (2004), The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity, *Genome Research*, vol. 14, n°11, p. 2253- 2260.

LEE, E., ISKOW, R., YANG, L., GOKCUMEN, O., HASELEY, P., LUQUETTE, L. J., … THE CANCER GENOME ATLAS RESEARCH NETWORK, (2012), Landscape of Somatic Retrotransposition in Human Cancers, *Science*, vol. 337, n°6097, p. 967- 971.

LIANG, W., XU, J., YUAN, W., SONG, X., ZHANG, J., WEI, W., … YANG, Y., (2016), APOBEC3DE inhibits LINE-1 retrotransposition by interacting with ORF1p and influencing LINE reverse transcriptase activity, *PloS one*, vol. 11, n°7, p. e0157220.

LINDTNER, S., FELBER, B. K. & KJEMS, J., (2002), An element in the 3' untranslated region of human LINE-1 retrotransposon mRNA binds NXF1(TAP) and can function as a nuclear export element., *RNA*, vol. 8, n°3, p. 345- 356.

LYON, M. F., (2006), Do LINEs Have a Role in X-Chromosome Inactivation?, *Journal of Biomedicine and Biotechnology*, vol. 2006, .

MACIA, A., MUNOZ-LOPEZ, M., CORTES, J. L., HASTINGS, R. K., MORELL, S., LUCENA-AGUILAR, G., … GARCIA-PEREZ, J. L., (2011), Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells, *Molecular and Cellular Biology*, vol. 31, n°2, p. 300- 316.

MACIA, A., WIDMANN, T. J., HERAS, S. R., AYLLON, V., SANCHEZ, L., BENKADDOUR-BOUMZAOUAD, M., … GARCIA-PEREZ, J. L., (2017), Engineered LINE-1 retrotransposition in nondividing human neurons, *Genome Research*, vol. 27, n°3, p. 335- 348.

MARTÍNEZ-GARAY, I., BALLESTA, M., OLTRA, S., ORELLANA, C., PALOMEQUE, A., MOLTÓ, M., … MARTÍNEZ, F., (2003), Intronic L1 insertion and F268S, novel mutations in RPS6KA3 (RSK2) causing Coffin–Lowry syndrome, *Clinical Genetics*, vol. 64, n°6, p. 491- 496.

MÄTLIK, K., REDIK, K. & SPEEK, M., (2006), L1 Antisense Promoter Drives Tissue-Specific Transcription of Human Genes, *Journal of Biomedicine and Biotechnology*, vol. 2006, p. 1- 16.

MIKI, Y., NISHISHO, I., HORII, A., MIYOSHI, Y., UTSUNOMIYA, J., KINZLER, K. W., … NAKAMURA, Y., (1992), Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer, *Cancer research*, vol. 52, n°3, p. 643–645.

MINE, M., CHEN, J.-M., BRIVET, M., DESGUERRE, I., MARCHANT, D., DE LONLAY, P., … MARSAC, C., (2007), A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element, *Human Mutation*, vol. 28, n°2, p. 137- 142.

MIOUSSE, I. R. & KOTURBASH, I., (2015), The Fine LINE: Methylation Drawing the Cancer Landscape, *BioMed Research International*, vol. 2015, p. 1- 8.

MONTOYA-DURANGO, D. E., LIU, Y., TENENG, I., KALBFLEISCH, T., LACY, M. E., STEFFEN, M. C. & RAMOS, K. S., (2009), Epigenetic control of mammalian LINE-1 retrotransposon by retinoblastoma proteins, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 665, n°1- 2, p. 20- 28.

MORRISH, T. A., GARCIA-PEREZ, J. L., STAMATO, T. D., TACCIOLI, G. E., SEKIGUCHI, J. & MORAN, J. V., (2007), Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres, *Nature*, vol. 446, n°7132, p. 208- 212.

MORRISH, T. A., GILBERT, N., MYERS, J. S., VINCENT, B. J., STAMATO, T. D., TACCIOLI, G. E., … MORAN, J. V., (2002), DNA repair mediated by endonuclease-independent LINE-1 retrotransposition, *Nature Genetics*, vol. 31, n°2, p. 159- 165.

MUKHERJEE, S., MUKHOPADHYAY, A., BANERJEE, D., CHANDAK, G. R. & RAY, K., (2004), Molecular pathology of haemophilia B: identification of five novel mutations including a LINE 1 insertion in Indian patients, *Haemophilia*, vol. 10, n°3, p. 259- 263.

NIGUMANN, P., REDIK, K., MÄTLIK, K. & SPEEK, M., (2002), Many Human Genes Are Transcribed from the Antisense Promoter of L1 Retrotransposon, *Genomics*, vol. 79, n°5, p. 628- 634.

NOUSHMEHR, H., WEISENBERGER, D. J., DIEFES, K., PHILLIPS, H. S., PUJARA, K., BERMAN, B. P., … ALDAPE, K., (2010), Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma, *Cancer Cell*, vol. 17, n°5, p. 510- 522.

OAKES, C. C., LA SALLE, S., ROBAIRE, B. & TRASLER, J. M., (2006), Evaluation of a Quantitative DNA Methylation Analysis Technique using Methylation-Sensitive/Dependent Restriction Enzymes and Real-Time PCR, *Epigenetics*, vol. 1, n°3, p. 146- 152.

OHKA, F., NATSUME, A., MOTOMURA, K., KISHIDA, Y., KONDO, Y., ABE, T., … WAKABAYASHI, T., (2011), The Global DNA Methylation Surrogate LINE-1 Methylation Is Correlated with MGMT Promoter Methylation and Is a Better Prognostic Factor for Glioma, *PLoS ONE*, vol. 6, n°8, p. e23332.

PALKA, K. T., MOBLEY, B., PERKINS, S., COOPER, M., SILLS, A. K. & MOOTS, P. L., (2012), Glioma and Other Neuroepithelial Neoplasms, In : D. R. M. PRESIDENT,ESSOR/MEDICINE , FACP, FRACP, FASCO, C. D. B. M. CHIEF FACP, FRCPC, FASCO Vice-President,essor and, D. H. J. MD, P. L. M. M. A. of N. and M. CHIEF, G. H. R. Md. of P. and MEDICINE, P. G. R. M. HEAD et M. A. S. M. DIRECTOR MS (éd.), *Textbook of Uncommon Cancer*, John Wiley & Sons, Inc., p. 767- 794.

PHILIPPE, C., VARGAS-LANDIN, D. B., DOUCET, A. J., VAN ESSEN, D., VERA-OTAROLA, J., KUCIAK, M., … CRISTOFARI, G., (2016), Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci, *Elife*, vol. 5, p. e13926.

PISKAREVA, O., LACKINGTON, W., LEMASS, D., HENDRICK, C., DOOLAN, P. & BARRON, N., (2011), The human L1 element: a potential biomarker in cancer prognosis, current status and future directions, *Current Molecular Medicine*, vol. 11, n°4, p. 286- 303.

RANGWALA, S. H., ZHANG, L. & KAZAZIAN, H. H., (2009), Many LINE1 elements contribute to the transcriptome of human somatic cells, *Genome biology*, vol. 10, n°9, p. R100.

REBOLLO, R., ROMANISH, M. T. & MAGER, D. L., (2012), Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes, *Annual Review of Genetics*, vol. 46, n°1, p. 21- 42.

RHEINBAY, E., SUVÀ, M. L., GILLESPIE, S. M., WAKIMOTO, H., PATEL, A. P., SHAHID, M., … BERNSTEIN, B. E., (2013), An aberrant transcription factor network essential for Wnt signaling and stem cell maintenance in glioblastoma, *Cell reports*, vol. 3, n°5, p. 1567- 1579.

RHOADS, A. & AU, K. F., (2015), PacBio Sequencing and Its Applications, *Genomics, Proteomics & Bioinformatics*, vol. 13, n°5, p. 278- 289.

RODIĆ, N., SHARMA, R., SHARMA, R., ZAMPELLA, J., DAI, L., TAYLOR, M. S., … BURNS, K. H., (2014), Long Interspersed Element-1 Protein Expression Is a Hallmark of Many Human Cancers, *The American Journal of Pathology*, vol. 184, n°5, p. 1280- 1286.

SAMUELOV, L., FUCHS-TELEM, D., SARIG, O. & SPRECHER, E., (2011), An exceptional mutational event leading to Chanarin–Dorfman syndrome in a large consanguineous family, *British Journal of Dermatology*, vol. 164, n°6, p. 1390- 1392.

SCARFÒ, I., PELLEGRINO, E., MEREU, E., KWEE, I., AGNELLI, L., BERGAGGIO, E., … OTHERS, (2016), Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts, *Blood*, vol. 127, n°2, p. 221–232.

SMIT, A. F., TÓTH, G., RIGGS, A. D. & JURKA, J., (1995), Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences, *Journal of molecular biology*, vol. 246, n°3, p. 401–417.

SOLYOM, S., EWING, A. D., HANCKS, D. C., TAKESHIMA, Y., AWANO, H., MATSUO, M. & KAZAZIAN, H. H., (2012), Pathogenic orphan transduction created by a non-reference LINE-1 retrotransposon, *Human Mutation*, vol. 33, n°2, p. 369- 371.

SPEEK, M., (2001), Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes, *Molecular and Cellular Biology*, vol. 21, n°6, p. 1973- 1985.

STUPP, R. & WEBER, D. C., (2005), The Role of Radio- and Chemotherapy in Glioblastoma, *Oncology Research and Treatment*, vol. 28, p. 315- 317.

SWERGOLD, G. D., (1990), Identification, characterization, and cell specificity of a human LINE-1 promoter., *Molecular and cellular biology*, vol. 10, n°12, p. 6718–6729.

TUBIO, J. M. C., LI, Y., JU, Y. S., MARTINCORENA, I., COOKE, S. L., TOJO, M., … CAMPBELL, P. J., (2014), Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes, *Science (New York, N.Y.)*, vol. 345, n°6196, p. 1251343.

TURCAN, S., ROHLE, D., GOENKA, A., WALSH, L. A., FANG, F., YILMAZ, E., … CHAN, T. A., (2012), IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype, *Nature*, vol. 483, n°7390, p. 479- 483.

TURCHIANO, G., LATELLA, M. C., GOGOL-DÖRING, A., CATTOGLIO, C., MAVILIO, F., IZSVÁK, Z., … RECCHIA, A., (2014), Genomic Analysis of Sleeping Beauty Transposon Integration in Human Somatic Cells, *PLoS ONE*, vol. 9, n°11, .

VAN DEN HURK, J. A. J. ., MEIJ, I. C., SELEME, M. DEL C., KANO, H., NIKOPOULOS, K., HOEFSLOOT, L. H., … CREMERS, F. P. M., (2007), L1 retrotransposition can occur early in human embryonic development, *Human Molecular Genetics*, vol. 16, n°13, p. 1587- 1592.

VASSETZKY, N. S. & KRAMEROV, D. A., (2013), SINEBase: a database and tool for SINE analysis, *Nucleic Acids Research*, vol. 41, n°Database issue, p. D83- D89.

VIOLLET, S., MONOT, C. & CRISTOFARI, G., (2014), L1 retrotransposition: The snap-velcro model and its consequences, *Mobile Genetic Elements*, vol. 4, n°2, p. e28907.

WANG, J., SU, H., ZHAO, H., CHEN, Z. & TO, S. T., (2015), Progress in the application of molecular biomarkers in gliomas, *Biochemical and Biophysical Research Communications*, vol. 465, n°1, p. 1- 4.

WICK, W., HARTMANN, C., ENGEL, C., STOFFELS, M., FELSBERG, J., STOCKHAMMER, F., … OTHERS, (2009), NOA-04 randomized phase III trial of sequential radiochemotherapy of anaplastic glioma with procarbazine, lomustine, and vincristine or temozolomide, *Journal of clinical oncology*, vol. 27, n°35, p. 5874–5880.

WICKER, T., SABOT, F., HUA-VAN, A., BENNETZEN, J. L., CAPY, P., CHALHOUB, B., … OTHERS, (2007), A unified classification system for eukaryotic transposable elements, *Nature Reviews Genetics*, vol. 8, n°12, p. 973–982.

WILLIAMS, Z., MOROZOV, P., MIHAILOVIC, A., LIN, C., PUVVULA, P. K., JURANEK, S., … TUSCHL, T., (2015), Discovery and Characterization of piRNAs in the Human Fetal Ovary, *Cell Reports*, vol. 13, n°4, p. 854- 863.

WOLFF, E. M., BYUN, H.-M., HAN, H. F., SHARMA, S., NICHOLS, P. W., SIEGMUND, K. D., … LIANG, G., (2010), Hypomethylation of a LINE-1 Promoter Activates an Alternate Transcript of the MET Oncogene in Bladders with Cancer, *PLoS Genetics*, vol. 6, n°4, p. e1000917.

WOODCOCK, D. M., LAWLER, C. B., LINSENMEYER, M. E., DOHERTY, J. P. & WARREN, W. D., (1997), Asymmetric Methylation in the Hypermethylated CpG Promoter Region of the Human L1 Retrotransposon, *Journal of Biological Chemistry*, vol. 272, n°12, p. 7810- 7816.

XING, J., WANG, H., BELANCIO, V. P., CORDAUX, R., DEININGER, P. L. & BATZER, M. A., (2006), Emergence of primate genes by retrotransposon-mediated sequence transduction, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, n°47, p. 17608- 17613.

YANG, N., (2003), An important role for RUNX3 in human L1 transcription and retrotransposition, *Nucleic Acids Research*, vol. 31, n°16, p. 4929- 4940.

YANG, N. & KAZAZIAN, H. H., (2006), L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells, *Nature Structural & Molecular Biology*, vol. 13, n°9, p. 763- 771.

Abstract :

Impact of transcriptional deregulation linked to the production of chimeric transcripts intiated from LINE-1 repeat elements in gliomas.

LINE-1 (L1) is the most abundant class of retrotransposons which represents 17% of the human genome. The 5' region of the youngest L1 sub-families (L1PA1 to 6) contains a bidirectional promoter consisting, in addition to the internal sense promoter, of an antisense promoter, called ASP. In normal cells, the main defense mechanism, developed to counteract the deleterious effect of L1 activity, consists in L1 promoter DNA methylation.
A hallmark of cancer genomes consists in a global DNA hypomethylation which affects especially L1 promoters. In tumors, evidences suggest that this hypomethylation could result in transcription from ASP of aberrant L1-Chimeric Transcripts (LCTs) composed of L1 5'end and its adjacent sequence. To investigate the pangenomic extent of this transcriptional deregulation and its impact in tumoral processes, a dedicated bioinformatic tool, CLIFinder, was designed to select putative LCTs among RNA-seq oriented paired-end reads. RNA-seq analyses of 13 gliomas, which are the most common brain cancer in adults, and 3 control brains were performed.

CLIFinder identifies 2675 chimeras in gliomas, among which 84% involves recent L1 (PA1 to 7) full size, supposed to possess a functional ASP, and 50% are detected specifically in tumors samples. 78 chimeras correspond to LCT already described in literature. In addition, study of additional RNA-seq data from other tumor types (MCF7 and ovarian metastasis) by CLIFinder identifies common chimeras suggesting that some of them can be recurrent. The analysis of a group of chimeras by 5'walk RT-PCR validate that 89% (56/63) of chimeras implying recent L1 (L1PA1 to 7) are initiated at the ASP region and therefore correspond to LCT; whereas all tested chimeras implying an L1PA8 element are transcribed from an upstream region. RT-qPCR studies on a larger cohort of 51 gliomas show that all 56 tested LCT, even identified by CLIFinder as "tumor specific", are not only expressed in tumors but also in controls. Nevertheless, 70% of the "tumor specific" LCTs are significantly overexpressed in tumors. My results suggest that, even L1 5'UTR methylation, some ASP are active in normal tissues and lead to a basal LCT expression in normal tissues. Moreover, a transcriptional deregulation linked to LCTs in tumors exists and implies a LCTs' overexpression.

In order to determine the underlying mechanisms involved in the increase of transcriptional activity of ASP, two hypothesis were tested. The first one implies L1 promoter hypomethylation. My results tend to refute this hypothesis because no decrease of the DNA methylation is found at the promoter region of L1 linked to overexpressed LCTs. On the other hand, the genes associated to LCT presenting an expression deregulation in tumors demonstrate a deregulation in the same way. Moreover, gene expression variations correlates systematically with the one corresponding LCTs. This suggests that an increase of transcriptional activity at the LCTs loci would be responsible of their overexpression.

Finally, 2 candidate LCT overexpressed and presenting as potential predictive biomarkers for patient's survival, could play a functional role in initiation, progression and/or the tumoral aggressiveness.

In conclusion, my work has validated CLIFinder as a useful tool to identify, at pangenomic level, LCTs expressed in different tumor types from paired-end stranded RNA-seq data. The observation of the recurrence and tumoral overexpression for some LCTs suggests that they may play a functional role in tumoral processes.

Keywords : LINE-1, LINE-1 Chimeric Transcripts, Gliomas, CLIFinder, RNA-seq

Résumé :

Etude de l'impact de la dérégulation transcriptionnelle liée à des transcrits chimères initiés à partir d'éléments répétés de type LINE-1 dans la tumorigenèse gliale.

Les éléments LINE-1 (L1) sont une classe abondante de rétrotransposons représentant 17% du génome humain. La région 5'UTR des sous-familles les plus récentes (L1PA1 à 6) contient un promoteur bidirectionnel contenant non seulement un promoteur sens interne mais aussi un promoteur antisens, nommé ASP. Dans les cellules normales, l'un des mécanismes impliqués dans la régulation du promoteur de L1 est la méthylation ADN. Dans les tumeurs, une hypométhylation globale affectant notamment les L1 est observée. Il a été mis en évidence que cette hypométhylation pouvait induire la transcription, à partir de l'ASP, de transcrits chimères ou LCT (L1 Chimeric Transcript). Ces LCT sont composés en 5' de la séquence L1 et se poursuivent dans la région génomique adjacente. Afin d'étudier l'impact pangénomique de cette dérégulation et son implication dans les processus tumoraux, un outil bio-informatique dédié, nommé CLIFinder, a été développé pour identifier dans des données de RNA-seq paired-end orientés des LCT putatifs. Les RNA-seq de 13 gliomes, qui sont les cancers du cerveau les plus fréquents chez l'adulte, et de 3 tissus cérébraux contrôles ont été étudiés.

CLIFinder identifie 2675 chimères dans les gliomes dont 84% impliquent des L1 récents (PA1 à 7) pleine taille supposés posséder un ASP fonctionnel et 50% sont détectées spécifiquement dans les échantillons tumoraux. 78 chimères correspondent à des LCT déjà décrits dans la littérature. De même, l'étude de RNA-seq d'autres types tumoraux (lignée MCF7 et métastases ovariennes) par CLIFinder identifie des chimères en commun suggérant une récurrence de certaines d'entre elles. L'étude d'un groupe de chimères par marche en 5' par RT-PCR valide que 89% (56/63) des chimères impliquant des L1 récents (L1PA1 à PA7) sont initiées dans la région de l'ASP et correspondent à des LCT alors que toutes les chimères testées impliquant des L1PA8 sont initiées en amont de cette région. Des études de RT-qPCR sur une cohorte plus large de 51 gliomes montrent que les 56 LCT testés, incluant des LCT tumeurs spécifiques, sont exprimés non seulement dans les tumeurs mais aussi dans les contrôles. Par contre, 70% des LCT tumeur spécifique montrent alors une surexpression tumorale significative. Ces résultats suggèrent donc une transcription basale provenant de l'ASP dans les tissus normaux et que la dérégulation transcriptionnelle liées aux LCT dans les gliomes passe par une surexpression.

Par ailleurs, afin de déterminer le ou les mécanismes sous-jacents impliqués dans l'augmentation de l'activité transcriptionnelle de l'ASP, deux hypothèses ont été testées. La première implique l'hypométhylation du promoteur de L1. Toutefois mes résultats tendent à réfuter cette hypothèse puisqu'aucune diminution de la méthylation ADN n'est retrouvée au niveau de la région promotrice des L1 impliqués dans la transcription de LCT surexprimés. Par contre, les gènes associés à des LCT dont l'expression est dérégulée en contexte tumoral présentent une dérégulation dans le même sens que celle du LCT. De plus, les variations d'expression de gènes corrèlent systématiquement avec celle des LCT correspondants. Ceci suggère qu'une augmentation d'activité transcriptionnelle aux loci des LCT serait responsable de la surexpression de ceux-ci.

Enfin 2 LCT candidats surexprimés et ayant un potentiel de biomarqueur prédictif de la survie des patients, pourraient jouer un rôle fonctionnel dans l'initiation, la progression et/ou l'agressivité tumorale.

En conclusion, mes travaux ont validé que CLIFinder se positionne comme un outil pertinent permettant d'identifier, de façon pangénomique, les LCT exprimés dans différents types tumoraux à partir de données de RNA-seq paired-end orientées. L'observation d'une récurrence ainsi que d'une surexpression tumorale de certains LCT suggère qu'ils pourraient jouer un rôle fonctionnel dans les processus de tumorigenèse.

Mots-clés : LINE-1, Transcrits chimères (LCT), CLIFinder, Gliomes, RNA-seq.