

PROCEDE DE CARACTERISATION D'UNE TUMEUR

Domaine de l'invention

- La présente invention a pour objet un procédé *in vitro* de caractérisation d'une tumeur, fondé sur l'analyse quantitative de nucléosides modifiés et non modifiés isolés d'un échantillon biologique. Plus particulièrement, l'invention a pour objet un procédé de prédiction du grade d'une tumeur gliale. Selon un autre aspect particulier, l'invention a pour objet un procédé de détection de la présence d'une tumeur.
- La présente invention se situe donc dans les domaines de la cancérologie et de la biologie moléculaire plus particulièrement appliquée au diagnostic médical.

Etat de la technique

- La caractérisation d'une tumeur est une condition préalable essentielle au choix du traitement le plus approprié pour le patient. Par « caractérisation d'une tumeur », on entend la caractérisation du statut ou degré d'évolution d'une tumeur donnée, il peut s'agir notamment par exemple de l'évaluation du degré d'évolution d'une tumeur d'un tissu connu, de l'attribution à une tumeur d'un grade préalablement défini, ou de tout autre caractérisation telle que notamment la détermination du caractère initial ou métastatique d'une tumeur.

- Les gliomes, ou tumeurs gliales, sont les tumeurs les plus courantes du système nerveux central, elles sont caractérisées par une variabilité significative de l'âge d'apparition, du classement, des caractéristiques histologiques et de la capacité à progresser et éventuellement à métastaser.

- Les gliomes sont classés en fonction de leur morphologie et de leur degré de malignité. La classification consensuelle de l'Organisation mondiale de la santé (OMS) attribue un degré de malignité de I à IV aux gliomes, les glioblastomes, ou tumeurs de grade IV, étant la forme la plus agressive et la plus mortelle.

- L'une des principales limites de la prise en charge des gliomes et des glioblastomes est liée au manque actuel de stratégies de diagnostic efficaces. La sélection d'un traitement personnalisé nécessite une classification précise des tumeurs. Actuellement, les principales méthodes de diagnostic utilisées cliniquement pour la détection des gliomes reposent sur des tests neurologiques et des méthodes de neuro-imagerie, réalisés lorsque la maladie est déjà à un stade avancé.

Le diagnostic de la tumeur nécessite une analyse des tissus du patient provenant d'une biopsie ou d'une résection chirurgicale. À partir de cet échantillon, plusieurs analyses moléculaires sont effectuées : test d'expression de gènes candidats, comptage du nombre de copies d'ADN, profil de méthylation, profilage de la voie phospho-protéique et séquençage génétique. Pourtant, les diagnostics basés sur la biopsie ont des limites concernant la détermination du grade des tumeurs et la stratification des patients. En effet, en ce qui concerne par exemple plus particulièrement les tumeurs gliales, les grades du gliome sont difficiles à distinguer, et plus particulièrement les grades II et III. L'établissement du grade requiert une analyse anatomo-pathologique délicate, souvent réalisée indépendamment par deux spécialistes. Le grade II désigne une tumeur bénigne tandis que le grade III représente une transition vers le glioblastome multiforme, qui est l'état le plus agressif.

Les classifications purement histologiques sont difficilement reproductibles, elles sont basées sur une expertise visuelle, et exigent l'intervention de deux spécialistes. Le couplage anatomo-pathologique avec analyse des images par imagerie par résonance magnétique (IRM) est coûteux et long, il dépend notamment du délai d'accès à l'IRM. A l'heure actuelle, aucun biomarqueur n'est suffisant à lui seul pour orienter les décisions thérapeutiques anti-cancéreuses.

Il existe donc un besoin général d'un procédé *in vitro* de caractérisation d'une tumeur, ledit procédé étant objectif, précis, reproductible, aisé et réalisable à un stade si possible précoce de la maladie. Ledit procédé permettrait de renforcer le diagnostic et faciliter la stratification des patients.

La publication de Janzer (« *Neuropathologie et pathologie moléculaire des gliomes.* » R-C Janzer, *Rev. Med. Suisse*, 5, 1501-4, 2009) décrit la classification des gliomes selon l'OMS, basée sur des critères histologiques et immuno-histochimiques, ainsi que sur des profils génétiques mettant en évidence l'altération de l'ADN des cellules : la détermination de l'hyperméthylation du promoteur du gène MGMT (pour les glioblastomes) et la détection de pertes des chromosomes 1p et 19q (pour les tumeurs oligodendrogiales).

La publication de Relier *et al* (« *FTO-mediated cytoplasmic m⁶A_m demethylation adjusts stem-like properties in colorectal cancer cell.* », *Nat. Commun* 12, 1716, 2021) décrit la régulation dans le cytoplasme du niveau de la méthylation m⁶A_m par la FTO (en anglais *Fat mass and obesity associated protein*) dans des lignées de cellules souches cancéreuses. Les auteurs mettent en évidence la fonction biologique de la modification m⁶A_m et ses effets secondaires potentiels pour le suivi du cancer colorectal. Ce document mentionne une

étape d'analyse par spectrométrie de masse (LC-MS/MS) d'ARNm fragmenté. Seuls les nucléosides m^6A , A , m^6A_m et A_m sont détectés et quantifiés.

La demande internationale WO 2007/008647 « *Diagnostic et classement de gliomes à l'aide d'une approche protéomique* » a pour objet un procédé de diagnostic et de classification des gliomes utilisant une approche protéomique. Dans ce procédé, un tissu tumoral est analysé par spectrométrie de masse et un profil des protéines exprimées est obtenu.

10 Il existe donc un besoin particulier d'un procédé *in vitro* d'évaluation du degré de malignité d'une tumeur gliale, en particulier sa classification. Il existe en particulier un besoin d'un procédé objectif permettant distinguer entre le grade II et le grade III des tumeurs gliales, pour renforcer le diagnostic et faciliter la stratification des patients.

15 Par ailleurs, il existe un besoin particulier d'un procédé de détection de la présence d'une tumeur dès les stades les plus précoces. En effet, la prise en charge précoce de la plupart des cancers accroît considérablement la survie d'un patient ou permet même sa guérison. Il existe donc un besoin d'un procédé objectif permettant de caractériser précocement la présence d'une tumeur.

20

Exposé de l'invention

Les inventeurs ont maintenant mis au point un procédé de caractérisation d'une tumeur qui exploite les données quantitatives de l'épitranscriptome.

25 L'épitranscriptome englobe l'ensemble des modifications chimiques portées par les bases des acides ribonucléiques (ARN), ensemble que l'on dénomme aussi par les termes « épigénétique de l'ARN ». Un procédé selon l'invention comprend la fourniture d'un échantillon biologique d'un sujet atteint d'une tumeur et l'obtention des quantités de nucléosides modifiés et non modifiés issus dudit échantillon, lesdites quantités sont regroupées dans un vecteur (au sens mathématique du terme). Selon un aspect particulier, un procédé selon l'invention comporte l'analyse informatique subséquente dudit vecteur pour la caractérisation d'une tumeur. Ladite caractérisation d'une tumeur permet de prédire des informations cliniques et médicales sur la tumeur à partir de l'échantillon sujet à l'analyse. Plus particulièrement, un procédé selon l'invention comporte l'analyse
35 informatique dudit vecteur pour la prédiction du grade de ladite tumeur.

Par simplicité, pour un échantillon donné, on nommera « profil épitranscriptomique » ou tout simplement « profil », le vecteur qui regroupe les quantités de chaque nucléoside, modifié ou non modifié.

- 5 Les nucléosides modifiés et non modifiés sont issus : i) de l'ARN total extrait de cellules d'un échantillon biologique d'un patient, ii) de l'ARN extracellulaire provenant d'un échantillon biologique d'un patient, et/ou iii) d'un extrait de métabolites d'un échantillon biologique isolé d'un patient.
- 10 Les nucléosides issus de l'ARN total extrait de cellules d'un échantillon biologique d'un patient et/ou de l'ARN extracellulaire provenant d'un échantillon biologique d'un patient sont obtenus par la fragmentation de l'ARN en nucléotides puis leur déphosphorylation. Les nucléosides issus d'un extrait de métabolites d'un échantillon biologique isolé d'un patient sont obtenus par une extraction des métabolites d'un échantillon biologique puis la
- 15 déphosphorylation desdits métabolites, selon des procédés appropriés bien connus de l'homme du métier. Lesdits métabolites sont notamment issus du catabolisme de l'ARN, les nucléosides présents sous forme monomérique peuvent également être désignés par l'expression de nucléosides dits « libres ».
- 20 Plus particulièrement, les nucléosides modifiés et non modifiés sont les nucléosides présents dans l'ARN total extrait de cellules d'une biopsie de ladite tumeur.

Par « nucléosides » on désigne les glycosamines constituées d'une base nucléique liée à l'atome de carbone anomérique d'un résidu de pentose par une liaison glycosidique depuis

- 25 l'atome d'azote N1 d'une pyrimidine ou l'atome N9 d'une purine. Selon un aspect particulier d'un procédé selon l'invention, lorsque ledit pentose est le ribose, le terme « nucléosides » désigne dans ce cas les ribonucléosides.

- On désigne aussi les nucléosides modifiés des ARN par les termes de « marques
- 30 épitranscriptomiques » ou de « modifications épitranscriptomiques ». Outre les nucléosides usuels des ARN (Tableau 1) utilisables pour la caractérisation des tumeurs, des nucléosides modifiés utilisables dans un procédé selon l'invention, en particulier dans l'analyse de gliomes, sont listés (Tableau 2).

Tableau 1

Nucléoside	Dénomination chimique
A	adénosine
C	cytidine
G	guanosine
U	uridine

5

10

15

20

25

30

Tableau 2

Nucléoside	Dénomination chimique	Nucléoside d'origine
Am	2'-O-methyladenosine	A
m1A	1-methyladenosine	A
m66A	N6,N6-dimethyladenosine	A
m66Am	N6,N6,2'-O-trimethyladenosine	A
m6A	N6-methyladenosine	A
m6Am	N6,2'-O-dimethyladenosine	A
ac4C	N4-acetylcytidine	C
Cm	2'-O-methylcytidine	C
hm5C	5-hydroxymethylcytidine	C
m3C	3-methylcytidine	C
m5C	5-methylcytidine	C
Gm	2'-O-methylguanosine	G
m1G	1-methylguanosine	G
m227G	N2,N2,7-trimethylguanosine	G
m27G	N2,7-dimethylguanosine	G
m7G	7-methylguanosine	G
oxo8G	8-hydroxyguanosine	G
I	inosine	I
Psi	pseudouridine	P
Q	Queuosine	Q
m3Um	3,2'-O-dimethyluridine	U
mcm5s2U	5-methoxycarbonylmethyl-2-	U

Selon les modes de réalisation d'un procédé selon l'invention, un profil épitranscriptomique peut inclure, selon les besoins de l'application, un nombre plus important de nucléosides modifiés, à déterminer parmi les nucléosides connus (Jonkhout *et al*, « *The RNA modification landscape in human disease* », RNA, Dec ;23 (12) :1754-1769, 2017). Une
5 liste complète de tous les nucléosides modifiés, qui pourront, selon les besoins de l'analyse, être inclus dans les profils transcriptomiques est accessible publiquement.

Le profil épitranscriptomique d'un échantillon caractérise ledit échantillon. Ledit profil
10 épitranscriptomique peut être obtenu par toute technique connue de l'état de l'art, et notamment par spectrométrie de masse, notamment par spectrométrie de masse couplée à la chromatographie.

Pour désigner les informations médicales à prédire, on parlera aussi de
15 « variables cliniques » ou « caractéristiques cliniques ».

Dans un procédé selon l'invention, l'étape d'analyse du profil épitranscriptomique à des fins de prédiction clinique est basé sur une méthode d'apprentissage automatique supervisé. L'apprentissage est effectué sur les profils issus d'une cohorte, c'est-à-dire des
20 échantillons cellulaires pour lesquels on connaît au préalable la variable caractéristique clinique à prédire. Le « modèle informatique » ainsi créé par l'apprentissage peut être ensuite utilisé (en mode prédiction) afin de prédire la variable clinique pour tout nouvel échantillon.

25 Les inventeurs ont également développé un procédé de normalisation des données quantitatives brutes permettant d'obtenir un profil épitranscriptomique contenant des quantités relatives comparables.

Préalablement au procédé d'apprentissage, l'analyse exploratoire des profils des
30 échantillons de la cohorte a révélé des variations des profils, lesdites variations étant corrélées au grade des tumeurs des sujets dont ont été extraits les échantillons. A partir des profils de la cohorte signature et au moyen d'un outil de prédiction par apprentissage automatique, un procédé selon l'invention permet de prédire le grade d'un gliome à partir d'un échantillon biologique d'un patient atteint de tumeur, en particulier à partir d'un
35 échantillon comprenant des cellules tumorales. Plus particulièrement, un procédé selon l'invention permet de distinguer les grades II et III d'un gliome à partir d'un échantillon comprenant des cellules tumorales.

Enfin, en combinant les profils épitranscriptomiques normalisés et les données de survie de patients, et au moyen d'un outil de prédiction par apprentissage automatique, un procédé selon l'invention permet de prédire la survie d'un patient à partir d'un échantillon biologique isolé dudit patient, en particulier à partir d'un échantillon tumoral.

5

Les inventeurs ont également développé un procédé de détection de la présence d'une tumeur chez un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :

- 10 a) isolement des nucléosides dudit échantillon biologique, par l'extraction : i) de l'ARN cellulaire total et sa fragmentation en nucléosides, ii) de l'ARN extracellulaire et sa fragmentation en nucléosides, et/ou iii) des nucléosides issus des catabolites monomériques,
- 15 b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents obtenus lors de l'étape a), et
- c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de la présence de ladite tumeur.

- 20 Les étapes techniques d'un procédé de détection de la présence d'une tumeur chez un individu répondent aux mêmes caractéristiques que les étapes techniques d'un procédé de caractérisation d'une tumeur.

- 25 Les inventeurs ont donc maintenant mis au point un procédé qui exploite les données quantitatives de l'épitranscriptome pour la caractérisation d'une tumeur, d'une part, et pour la détection de la présence d'une tumeur, d'autre part. Selon un mode de réalisation, un procédé selon l'invention est donc avantageusement utilisé pour caractériser une tumeur. Selon un autre mode de réalisation, un procédé selon l'invention est avantageusement utilisé pour détecter la présence d'une tumeur.

30

Description détaillée de l'invention

Selon un premier aspect, l'invention a pour objet un procédé *in vitro* de caractérisation d'une tumeur d'un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :

- 35 a) isolement des nucléosides dudit échantillon biologique, par l'extraction : i) de l'ARN cellulaire total et sa fragmentation en nucléosides, ii) de l'ARN extracellulaire et sa fragmentation en nucléosides, et/ou iii) des nucléosides issus des catabolites monomériques,

- b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents obtenus lors de l'étape a), et
- c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur.

Selon un premier mode de réalisation, un procédé selon l'invention est fondé sur l'analyse simultanée de la quantité de différents nucléosides issus de l'ARN cellulaire total d'un échantillon biologique, et/ou issus de l'ARN extracellulaire et sa fragmentation en nucléosides et/ou issus de nucléosides obtenus à partir des catabolites monomériques présents dans ledit échantillon, un procédé selon l'invention comprend donc l'analyse simultanée de multiples variables, et non sur la détection quantitative d'un marqueur unique.

Par « profil » ou « profil de nucléosides » on entend un vecteur de quantités de nucléosides.

Par « ARN cellulaire total » on entend la totalité de l'ARN cellulaire extrait selon les méthodes bien connues et accessibles. L'ARN cellulaire total inclut l'ARN de transfert (ARNt), l'ARN messenger (ARNm), l'ARN ribosomique (ARNr) et d'autres ARN non codants. Ledit ARN cellulaire total est donc ici présent sous une forme polymérique.

Par « ARN extracellulaire » on entend la totalité de l'ARN extracellulaire présent sous forme polymérique, extrait selon les méthodes bien connues et accessibles. Cette forme polymérique de l'ARN extracellulaire est notamment également désignée par l'expression « ARN circulant ». Ledit ARN extracellulaire est issu de la dégradation enzymatique *in vivo* de l'ARN de transport (ARNt), l'ARN messenger (ARNm) et/ou l'ARN ribosomal (ARNr) et des autres types d'ARN, notamment les ARN non codants.

Par « nucléosides issus des catabolites monomériques » on entend les nucléosides obtenus, selon les méthodes bien connues et accessibles, à partir des catabolites présents sous une forme monomérique dans l'échantillon. Ces catabolites monomériques sont issus de la dégradation enzymatique *in vivo* de l'ARN de transport (ARNt), l'ARN messenger (ARNm) et/ou l'ARN ribosomal (ARNr) et des autres types d'ARN, notamment les ARN non codants.

Par « isolement et détermination d'une quantité respective d'au moins 3 nucléosides différents » on entend l'isolement et la détermination d'une quantité de chacun des « au moins 3 » nucléosides pris individuellement.

- 5 Selon ce premier mode de réalisation, l'invention a donc pour objet un procédé *in vitro* de caractérisation d'une tumeur d'un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :
- a) isolement des nucléosides dudit échantillon biologique par l'extraction de l'ARN cellulaire total et sa fragmentation en nucléosides,
 - 10 b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents obtenus lors de l'étape a), et
 - c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b),
 - 15 ledit profil étant caractéristique de ladite tumeur.

Les nucléosides peuvent également être présents dans l'échantillon biologique sous une forme polymérique extracellulaire, notamment également désignée par l'expression « ARN circulant ». Les nucléosides peuvent également être présents sous une forme

20 monomérique (métabolites) dans l'échantillon biologique. Lesdits ARN extracellulaires et nucléosides monomériques sont issus de la dégradation enzymatique *in vivo* de l'ARN de transport (ARNt), l'ARN messager (ARNm) et/ou l'ARN ribosomal (ARNr) et des autres types d'ARN, notamment les ARN non codants.

- 25 Selon un deuxième mode de réalisation, un procédé selon l'invention est fondé sur l'analyse simultanée de la quantité de différents nucléosides issus de l'ARN extracellulaire d'un échantillon biologique.

- Selon ce deuxième mode de réalisation, l'invention a pour objet un procédé *in vitro* de
- 30 caractérisation d'une tumeur d'un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :
- a) isolement des nucléosides dudit échantillon biologique, par l'extraction de l'ARN extracellulaire et sa fragmentation en nucléosides,
 - b) isolement et détermination d'une quantité respective d'au moins 3, de préférence
 - 35 au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents obtenus lors de l'étape a), et

- c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur.

- 5 Selon un troisième mode de réalisation, un procédé selon l'invention est fondé sur l'analyse simultanée de la quantité de différents nucléosides issus des catabolites monomériques présents dans un échantillon biologique.

10 Selon ce troisième mode de réalisation, l'invention a pour objet un procédé *in vitro* de caractérisation d'une tumeur d'un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :

- a) isolement des nucléosides dudit échantillon biologique, par l'extraction des catabolites monomériques présents dans ledit échantillon,
b) isolement et détermination d'une quantité respective d'au moins 3, de préférence
15 au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents obtenus lors de l'étape a), et
c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur.

20

Dans un procédé *in vitro* de caractérisation d'une tumeur d'un individu, à partir d'un échantillon biologique isolé dudit individu, ledit échantillon biologique est notamment être choisi parmi :

- un échantillon biologique solide, particulièrement une biopsie, et plus
25 particulièrement une biopsie de ladite tumeur, et
- un échantillon biologique liquide, notamment un prélèvement d'un fluide corporel dudit individu, plus particulièrement un échantillon de sang, de plasma, de sérum ou d'urine.

- 30 Par « biopsie » on entend le prélèvement d'une très petite partie d'un organe ou d'un tissu. Lorsque l'échantillon biologique est une biopsie, le premier mode de réalisation du procédé selon l'invention, dans lequel l'ARN cellulaire total est extrait puis fragmenté, est préféré.

- Lorsque l'échantillon biologique est un échantillon biologique liquide, le deuxième et le
35 troisième mode de réalisation du procédé selon l'invention, dans lesquels respectivement l'ARN extracellulaire est extrait puis fragmenté, ou dans lequel l'ARN sous forme de nucléosides isolés est extrait, sont préférés.

Dans un procédé selon l'invention, ledit échantillon biologique est en volume suffisant, ou comporte un nombre de cellules suffisant, pour permettre une détermination quantitative fiable d'au moins 3 nucléosides issus de la fragmentation d'un extrait de l'ARN cellulaire total dudit échantillon.

5

Dans le cas d'une biopsie, l'ARN cellulaire est total est extrait selon un procédé choisi parmi les procédés accessibles à l'homme du métier, notamment un procédé tel que décrit dans le présent exemple. Dans le cas d'un échantillon liquide, tel que le sang ou l'urine, ledit échantillon est préalablement traité si nécessaire, afin notamment d'éliminer d'éventuels composés interférents, de concentrer ledit échantillon et/ou de déterminer une valeur standard de concentration d'un élément de référence, tel que la créatinine dans l'urine, cette valeur standard servant à étalonner la concentration de l'échantillon à partir duquel le profil de nucléosides est établi.

15 Dans un procédé selon l'invention, l'ARN cellulaire total, l'ARN extracellulaire et les nucléosides isolés sont obtenus d'un échantillon biologique par tout procédé connu d'un homme du métier, ledit procédé comprend notamment une étape d'extraction, éventuellement une étape de fragmentation, et une étape de déphosphorylation.

20 Selon un aspect particulier, l'invention a donc pour objet un procédé *in vitro* de caractérisation d'une tumeur d'un individu à partir d'une biopsie de ladite tumeur, ledit procédé comprenant la préparation, à partir de ladite biopsie, d'un extrait de l'ARN cellulaire total et fragmentation dudit ARN en nucléosides.

25 Selon un mode de réalisation, dans un procédé *in vitro* de caractérisation d'une tumeur d'un individu selon l'invention, au moins 3 nucléosides isolés issus de l'échantillon biologique, obtenus par i) la préparation d'un extrait de l'ARN cellulaire total et de sa fragmentation en nucléosides, ii) par la préparation d'un extrait de l'ARN extracellulaire et de sa fragmentation en nucléosides, et/ou iii) par l'extraction des nucléosides isolés, sont isolés et leur quantité respective déterminée, lesdits au moins 3 nucléosides sont choisis parmi :

- les nucléosides non modifiés : adénosine (A), cytidine (C), guanosine (G), uridine (U), et
- les nucléosides modifiés (voir Tableau 2).

35

Les nucléosides modifiés résultent de l'action d'un grand nombre d'enzymes hautement spécifiques, les nucléosides subissent notamment méthylation et réarrangement de liaisons carbone-azote. Lesdits nucléosides modifiés sont tous les nucléosides modifiés

connus à la date de la présente demande, ces nucléosides sont notamment cités dans la publication de Jonkhout *et al* (« *The RNA modification landscape in human disease* », RNA, Dec ;23 (12) :1754-1769, 2017), et dans le Tableau 2 de la présente demande.

5 Selon différents modes de réalisation d'un procédé objet de l'invention, lesdits au moins 3 nucléosides sont choisis dans les groupes constitués par :

- les nucléosides non modifiés : adénosine (A), cytidine (C), guanosine (G), uridine (U),
- 2'-O-méthyladénosine (Am), 1-méthyladénosine (m1A), N6,N6-diméthyladénosine (m66A), N6,N6,2'-O-triméthyladénosine (m66Am), N6-méthyladénosine (m6A),
10 N6,2'-O-diméthyladénosine (m6Am), N4-acetylcytidine (ac4C), 2'-O-méthylcytidine (Cm), 5-hydroxyméthylcytidine (hm5C), 3-méthylcytidine (m3C), 5-méthylcytidine (m5C), 2'-O-méthylguanosine (Gm), 1-méthylguanosine (m1G), N2,N2,7-triméthylguanosine (m227G), N2,7-diméthylguanosine (m27G), 7-méthylguanosine (m7G), 8-hydroxyguanosine (oxo8G), inosine (I), pseudouridine (Psi), queuosine (Q), 3,2'-O-diméthyluridine (m3Um), 5-méthoxycarbonylméthyl-
15 2-thiouridine (mcm5s2U), 5- méthoxycarbonylméthyluridine (mcm5U), 5-carbamoylméthyluridine (ncm5U), 2'-O-méthyluridine (Um), et/ou
- 3-(3-amino-3-carboxypropyl)uridine (acp3U), 2'-O-ribosyladénosine (phosphat) (Ar(p)), 5-carboxyméthylaminométhyl-2-thiouridine (cmnm5s2U), 5-carboxyméthylaminométhyluridine (cmnm5U), 5-carboxyméthylaminométhyl-2'-
20 O-méthyluridine (cmnm5Um), dihydrouridine (D), 5-formylcytidin (f5C), galactosyl-queuosine (galQ), 2'-O-méthyl-5-hydroxyméthylcytidine (hm5Cm), 5-hydroxyuridine (ho5U), 5-hydroxyadénosine (ho8A), 8-hydroxyguanosine (ho8G), N6-isopentenyladénosine (i6A), N6-(cis-hydroxyisopentenyl)adénosine (io6A), 1-méthylinosine (m1I), 1-méthylpseudouridine (m1psi), N2,N2-diméthylguanosine (m22G), 2-méthyladénosine (m2A), N2-méthylguanosine (m2G), 5-méthyluridine (m5U),
25 5, 2'-O-diméthyluridine (m5Um), N6-méthyl-N6-threonylcarbamoyladénosine (m6t6A), mannosyl-queuosine (manQ), 5-(carboxyhydroxyméthyl)uridineméthyl ester (mchm5U), 5-méthylaminométhyl-2-thiouridine (mnm5s2U), 2-méthylthio-N6-isopentenyladénosine (ms2i6A), 2-méthylthio-N6-threonylcarbamoyladénosine (ms2t6A), peroxywybutosine (o2yW), 2'-O-méthylpseudouridine (psi m), 2-thiouridine (s2U), N6-threonylcarbamoyladénosine (t6A), wybutosine (yW).

35

Selon un mode de réalisation d'un procédé objet de l'invention, lesdits au moins 3 nucléosides sont choisis dans les groupes constitués par :

- les nucléosides non modifiés : adénosine (A), cytidine (C), guanosine (G), uridine (U), et
- 2'-O-méthyladénosine (Am), 1-méthyladénosine (m1A), N6,N6-diméthyladénosine (m66A), N6,N6,2'-O-triméthyladénosine (m66Am), N6-méthyladénosine (m6A),
5 N6,2'-O-diméthyladénosine (m6Am), N4-acetylcytidine (ac4C), 2'-O-méthylcytidine (Cm), 5-hydroxyméthylcytidine (hm5C), 3-méthylcytidine (m3C), 5-méthylcytidine (m5C), 2'-O-méthylguanosine (Gm), 1-méthylguanosine (m1G),
N2,N2,7-triméthylguanosine (m227G), N2,7-diméthylguanosine (m27G), 7-méthylguanosine (m7G), 8-hydroxyguanosine (oxo8G), inosine (I), pseudouridine
10 (Psi), queuosine (Q), 3,2'-O-diméthyluridine (m3Um), 5-méthoxycarbonylméthyluridine (mcm5s2U), 5-méthoxycarbonylméthyluridine (mcm5U), 5-carbamoylméthyluridine (ncm5U), 2'-O-méthyluridine (Um).

Selon un mode de réalisation plus particulier, un procédé objet de l'invention comprend
15 l'isolement et la détermination quantitative d'au moins 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 ou 29 nucléosides différents issus de la fragmentation de l'ARN total dudit échantillon biologique.

Selon un autre mode de réalisation plus particulier, un procédé objet de l'invention
20 comprend l'isolement et la détermination quantitative d'au moins 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 ou 29 nucléosides différents issus de la fragmentation de l'ARN extracellulaire dudit échantillon biologique.

Selon un autre mode de réalisation plus particulier, un procédé objet de l'invention
25 comprend l'isolement et la détermination quantitative d'au moins 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 ou 29 nucléosides différents issus de l'extraction de nucléosides dudit échantillon biologique.

Selon un mode de réalisation plus particulier, un procédé objet de l'invention comprend
30 l'isolement et la détermination quantitative d'au moins 3 nucléosides différents issus de la fragmentation de l'ARN total dudit échantillon biologique et/ou de la fragmentation de l'ARN extracellulaire et/ou de l'extraction des nucléosides isolés, lesdits nucléosides étant choisis parmi les suivants : adénosine (A), cytidine (C), guanosine (G), uridine (U), 2'-O-méthyladénosine (Am), 1-méthyladénosine (m1A), N6,N6-diméthyladénosine (m66A),
35 N6,N6,2'-O-triméthyladénosine (m66Am), N6-méthyladénosine (m6A), N6,2'-O-diméthyladénosine (m6Am), N4-acetylcytidine (ac4C), 2'-O-méthylcytidine (Cm), 5-hydroxyméthylcytidine (hm5C), 3-méthylcytidine (m3C), 5-méthylcytidine (m5C), 2'-O-méthylguanosine (Gm), 1-méthylguanosine (m1G), N2,N2,7-triméthylguanosine

(m227G), N2,7-diméthylguanosine (m27G), 7-méthylguanosine (m7G), 8-hydroxyguanosine (oxo8G), inosine (I), pseudouridine (Psi), queuosine (Q), 3,2'-O-diméthyluridine (m3Um), 5-méthoxycarbonylméthyl-2-thiouridine (mcm5s2U), 5-méthoxycarbonylméthyluridine (mcm5U), 5-carbamoylméthyluridine (ncm5U), 2'-O-méthyluridine (Um).

Dans un procédé selon l'invention, l'isolement et la détermination d'une quantité respective d'au moins 3 nucléosides sont mis en œuvre par tout moyen d'analyse connu de l'homme du métier. Ces moyens comprennent notamment la chromatographie, en particulier la chromatographie liquide en phase inverse de haute performance (RP-HPLC) ou l'électrophorèse capillaire (CE).

Ces moyens comprennent également des moyens de spectrométrie, en particulier la spectrométrie de masse. Plus particulièrement, ces moyens comprennent la spectrométrie de masse en tandem couplée à la Chromatographie en phase liquide (LC-MS/MS), une technique analytique qui associe le pouvoir de séparation de la chromatographie en phase liquide à la capacité d'analyse de masse hautement sensible et sélective de la spectrométrie de masse à triple quadripôles. Le point fort de cette technique réside dans le pouvoir de séparation de la chromatographie en phase liquide pour une large gamme de composés, combiné à la capacité de la spectrométrie de masse à quantifier les composés avec un degré élevé de sensibilité et de sélectivité, en fonction des transitions uniques masse / charge (m/z) de chaque composé d'intérêt.

Selon un aspect particulier, dans un procédé selon l'invention, le mélange de nucléosides obtenu par fragmentation est analysé en utilisant une chromatographie liquide haute performance couplée à une spectrométrie de masse en tandem (LC-MS / MS) de type triple quadripôles dans le mode Multiple Reaction Monitoring (MRM). Le mode MRM est une technique hautement sensible et spécifique qui permet la quantification de molécules par spectrométrie de masse. Ce mode de scan est dépendant de la spectrométrie de masse en tandem et plus particulièrement de triple quadripôles ou de systèmes spectrométrie de masse à trappes hybrides. Le MRM scan mode est basée sur la sélection d'ions de masse et charge spécifiques d'une molécule, ions appelés ions précurseurs ou ions parents, ainsi que sur les ions fragments correspondants après fragmentation dans la cellule de collision. Le premier quadripôle va permettre la sélection précise des ions précurseurs spécifiques des molécules d'intérêts qui vont alors être fragmentés dans le second quadripôle. Les ions fragments résultants sont alors sélectionnés dans le troisième quadripôle. Les deux ions (masse/charge) correspondent alors à une transition hautement spécifique de la molécule d'intérêt.

Selon un mode de réalisation plus particulier, l'invention a pour objet un procédé *in vitro* de caractérisation d'une tumeur d'un individu, à partir d'une biopsie de cet individu, comprenant les étapes de :

- 5 a) préparation, à partir dudit échantillon biologique, d'un extrait de l'ARN cellulaire total et fragmentation de l'ARN polymérique en nucléosides,
- b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents issus l'étape a),
- 10 c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur.

Selon un mode de réalisation encore plus particulier, l'invention a pour objet un procédé

15 *in vitro* de caractérisation d'une tumeur d'un individu, ladite tumeur étant une tumeur située dans l'un des organes suivants : rectum, colon, sein, pancréas, rein, poumon, ou une tumeur hématologique, notamment une leucémie.

Selon un mode de réalisation plus particulier, l'invention a pour objet un procédé *in vitro*

20 de caractérisation d'une tumeur gliale d'un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :

- a) préparation, à partir dudit échantillon biologique, d'un extrait de l'ARN cellulaire total et fragmentation dudit ARN en nucléosides,
- b) isolement et détermination d'une quantité respective d'au moins 3, de préférence
- 25 au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents issus l'étape a),
- c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur, et
- 30 d) prédiction d'un grade de ladite tumeur gliale par un premier modèle de classification préalablement entraîné, à partir du profil établi lors de l'étape c).

Les termes de « tumeur gliale » ou « gliome » regroupent diverses tumeurs du cerveau qui se développent à partir des cellules gliales normales du cerveau. Le grade d'une tumeur

35 gliale représente le plus important déterminant pour la survie d'un individu portant une telle tumeur. Le tissu cérébral non tumoral est caractérisée par de nombreuses cellules ayant des caractéristiques normales et quelques caractéristiques mitotiques, sans prolifération endothéliale. Les tumeurs de grade II, aussi appelées « astroblastomes »

comprennent un nombre plus important de cellules comprenant des noyaux polymorphiques en mitose. Les tumeurs de grade III, sont aussi appelées « astroblastomes anaplastiques ». Les tumeurs de grade IV correspondent au glioblastome multiforme.

- 5 On entend par « modèle de classification » un algorithme d'apprentissage automatique préalablement entraîné, en particulier lors d'un apprentissage supervisé, ainsi qu'un jeu de données d'apprentissage permettant l'entraînement de l'algorithme susmentionné, et un jeu de données d'évaluation.
- 10 Suivant des modes de réalisation, ledit premier modèle de classification peut comprendre :
- un algorithme d'apprentissage automatique,
 - plus particulièrement un réseau neuronal à apprentissage supervisé, ou
 - un algorithme de classification probabiliste multi-classes,
- préalablement entraîné avec un jeu de données d'apprentissage.

15

Les données d'apprentissage, d'une part, sont spécifiques de la question posée et, d'autre part, spécifiques du type de cancer ciblé. Ainsi, la phase d'entraînement de l'algorithme d'apprentissage utilise les données d'apprentissage afin de produire un modèle de classification, qui est lui-même spécifique de la question posée et spécifique du type de cancer ciblé. La phase d'apprentissage infère les paramètres du modèle en fonction de ces données et de la question. Par exemple, pour une question de détermination du grade le modèle de classification renvoie une réponse parmi quatre réponses possibles si on distingue quatre grades. En revanche, pour la question de la détection de la présence d'une tumeur, le modèle de classification répond par "tumeur" ou "sain", soit un choix parmi deux réponses possibles.

25

Le modèle de classification produit par la phase d'apprentissage est un programme mis en oeuvre sur un ordinateur afin d'obtenir une prédiction sur la question considérée à partir de la donnée d'un profil épitranscriptomique issu d'un échantillon. Ce programme est téléchargeable et installable, permettant ainsi une installation sur un autre système que celui sur lequel il a été produit.

30

Le jeu de données d'apprentissage peut comprendre une multitude de couples de données, chacun des couples de données comprenant une première donnée représentant un profil de nucléosides et une deuxième donnée représentant le grade de tumeur pour ce profil.

35

Le jeu données d'apprentissage peut comprendre un jeu d'entraînement et un jeu de test, également désigné par « jeu d'évaluation [CF1] », du modèle. Le modèle peut ainsi être

testé sur le jeu d'entraînement et le jeu de test peut être utilisé pour déterminer si l'apprentissage du modèle est satisfaisant ou non.

Le jeu d'entraînement et le jeu de test peuvent être différents. Alternativement, le jeu de
5 test peut correspondre à une partie du jeu d'entraînement.

Le jeu de données d'apprentissage peut être préalablement constitué à partir de données
obtenues en laboratoire par analyse d'échantillons obtenus à partir d'individus atteints de
cancer et dont le grade de la tumeur a préalablement été déterminé.

10

On estime que le modèle de classification a atteint un niveau d'apprentissage satisfaisant
sur l'ensemble des profils du jeu de test si la classification atteint par exemple 85 % de
précision ; autrement dit on estime que le modèle de classification a atteint un niveau
d'apprentissage satisfaisant sur l'ensemble des profils du jeu de test si la classification
15 atteint par exemple au plus 15 % d'erreur.

Le modèle de classification peut consister en un programme d'ordinateur. Selon un mode
de réalisation préféré de l'invention, un modèle de classification mis en œuvre dans un
procédé selon l'invention consiste en un programme d'ordinateur qui exécute en puissance
20 une fonction technique consistant en des étapes du procédé de classification. L'exécution
dudit programme par un ordinateur produit un objet numérique, qui est un objet technique.

Ledit programme d'ordinateur peut être écrit en tout langage informatique tel que par
exemple en C, C++, JAVA, Python, etc.

25

Suivant des exemples de réalisation, le modèle de classification peut comprendre une
machine à vecteur support, une forêt aléatoire, une analyse discriminante linéaire ; ces
méthodes se nomment en anglais respectivement : "Support Vector Machines", "Random
Forests" et "Linear Discriminant Analysis" (LDA).

30

Plus particulièrement, ledit algorithme d'apprentissage est notamment choisi parmi:

- une Machine à Vecteurs Supports munie soit d'un noyau linéaire, soit d'un noyau RBF
(Radial Basis Function) avec une valeur faible de paramètre de coût,
- un algorithme de LDA utilisant les solutions de moindres carrés avec détermination
35 automatique du paramètre de réduction de dimension selon la procédure de Ledoit-Wolf.

Ces trois familles d'algorithmes d'apprentissage automatique sont décrites
conceptuellement dans la littérature (Cornuejols et Miclet, « *Apprentissage Artificiel* :

Concepts et Algorithmes » Eyrolles, 2012 ; Hastie *et al.* "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2nd Edition. Springer Series in Statistics, Springer 2009, ISBN 9780387848570) et sont parfaitement adaptées à la classification multi-classes.

5

Suivant des modes de réalisation d'un procédé selon l'invention, la prédiction d'un grade d'une tumeur gliale peut comprendre :

- la prédiction d'une tumeur gliale de grade II,
- la prédiction d'une tumeur gliale de grade III ou
- 10 - la prédiction d'une tumeur gliale de grade IV.

L'invention a plus particulièrement pour objet un procédé *in vitro* de prédiction d'un grade d'une tumeur gliale d'un individu, à partir d'un échantillon biologique dudit individu, et notamment une biopsie de ladite tumeur gliale, dans laquelle la prédiction d'un grade de

15 ladite tumeur gliale par un modèle de classification préalablement entraîné, comprend : la prédiction d'une tumeur gliale de grade II, la prédiction d'une tumeur gliale de grade III et la prédiction d'une tumeur gliale de grade IV. Plus particulièrement, un procédé selon l'invention de prédiction d'un grade d'une tumeur gliale d'un individu comprend la distinction entre une tumeur gliale de grade II et une tumeur gliale de grade III ou IV ; la

20 distinction entre une tumeur gliale de grade III et une tumeur gliale de grade II ou IV ; la distinction entre une tumeur gliale de grade IV et une tumeur gliale de grade II ou III.

L'invention a encore plus particulièrement pour objet un procédé *in vitro* de caractérisation d'une tumeur gliale d'un individu, à partir d'une biopsie de ladite tumeur, comprenant :

25 a) la préparation, à partir de ladite biopsie, d'un extrait de l'ARN cellulaire total et fragmentation dudit ARN en nucléosides,

b) l'isolement et détermination quantitative d'au moins 3 nucléosides issus de ladite fragmentation, choisis parmi : adénosine (A), cytidine (C), guanosine (G), uridine (U), 2'-O-méthyladénosine (Am), 1-méthyladénosine (m1A), N6,N6-

30 diméthyladénosine (m66A), N6,N6,2'-O-triméthyladénosine (m66Am), N6-méthyladénosine (m6A), N6,2'-O-diméthyladénosine (m6Am), N4-acetylcytidine (ac4C), 2'-O-méthylcytidine (Cm), 5-hydroxyméthylcytidine (hm5C), 3-méthylcytidine (m3C), 5-méthylcytidine (m5C), 2'-O-méthylguanosine (Gm), 1-méthylguanosine (m1G), N2,N2,7-triméthylguanosine (m227G), N2,7-

35 diméthylguanosine (m27G), 7-méthylguanosine (m7G), 8-hydroxyguanosine (oxo8G), inosine (I), pseudouridine (Psi), queuosine (Q), 3,2'-O-diméthyluridine (m3Um), 5-méthoxycarbonylméthyl-2-thiouridine (mcm5s2U), 5-

méthoxycarbonylméthyluridine (mcm5U), 5-carbamoylméthyluridine (ncm5U), 2'-O-méthyluridine (Um).

- 5 c) l'établissement, pour ladite tumeur, d'un profil à partir des valeurs quantitatives respectives des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur, et
- 10 d) la prédiction d'un grade de ladite tumeur gliale par un modèle de classification préalablement entraîné, à partir du profil établi lors de l'étape c), dans lequel la prédiction d'un grade d'une tumeur gliale est choisie parmi : la prédiction d'une tumeur gliale de grade II, la prédiction d'une tumeur gliale de grade III et la

Selon un autre aspect, l'invention a pour objet un procédé *in vitro* de caractérisation d'une tumeur gliale d'un individu, ledit procédé comprenant les étapes de :

- 15 a) préparation, à partir dudit échantillon biologique, d'un extrait de l'ARN cellulaire total et fragmentation dudit ARN en nucléosides,
- b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10, de préférence au moins 20, nucléosides différents issus de l'étape a),
- 20 c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur, et
- d) prédiction d'un état de survie dudit individu, par un deuxième modèle de classification préalablement entraîné, à partir du profil établi lors de l'étape c).

- 25 Suivant des modes de réalisation, ledit deuxième modèle de classification peut comprendre :

- un algorithme d'apprentissage automatique,
 - plus particulièrement un réseau neuronal à apprentissage supervisé, ou
 - un algorithme de classification probabiliste,
- 30 préalablement entraîné avec un deuxième jeu de données d'apprentissage.

Ledit deuxième jeu de données d'apprentissage peut comprendre une multitude de couples de données, chacun des couples de données comprenant une première donnée représentant un profil de nucléosides et une deuxième donnée représentant l'état de survie

35 pour ce profil.

Ce jeu de données d'apprentissage peut comprendre un jeu d'entraînement et un jeu d'évaluation du modèle. Le modèle peut ainsi être testé sur le jeu d'entraînement et le jeu

d'évaluation peut être utilisé pour déterminer si l'apprentissage du modèle est satisfaisant ou non. Le jeu d'entraînement et le jeu d'évaluation peuvent être différents. Alternativement, le jeu d'évaluation peut correspondre à une partie du jeu d'entraînement. Le jeu de données d'apprentissage peut être préalablement constitué à partir de données
5 obtenues en laboratoire par analyse d'échantillons obtenus à partir d'individus atteints de cancer et dont le statut de survie a préalablement été déterminé.

On estime que le modèle de classification a atteint un niveau d'apprentissage satisfaisant sur l'ensemble des profils du jeu d'évaluation si la classification atteint 85 % de précision ;
10 autrement dit on estime que le modèle de classification a atteint un niveau d'apprentissage satisfaisant sur l'ensemble des profils du jeu d'évaluation si la classification atteint au plus 15 % d'erreur.

Tout comme le premier modèle de classification, le deuxième modèle de classification peut
15 consister en un programme d'ordinateur. Le programme d'ordinateur peut être écrit en tout langage informatique tel que par exemple en C, C++, JAVA, Python, etc.

Suivant des exemples de réalisation, le deuxième modèle de classification peut comprendre une machine à vecteur support, une forêt aléatoire, une analyse discriminante linéaire ;
20 ces méthodes se nomment en anglais respectivement : "Support Vector Machines", "Random Forests" et "Linear Discriminant Analysis".

Selon un autre aspect, l'invention a pour objet un modèle de classification, préalablement entraîné sur un jeu de données d'apprentissage, pour prédire un grade d'une tumeur gliale
25 d'un individu atteint d'une tumeur, à partir d'un profil de nucléosides obtenu par la mise en œuvre d'un procédé selon l'invention.

Ledit modèle de classification pour prédire le grade d'une tumeur gliale comprend un algorithme d'apprentissage automatique préalablement entraîné et évalué, en particulier
30 lors d'un apprentissage supervisé, avec un jeu de données d'apprentissage relatives à la prédiction d'un grade d'une tumeur gliale, ledit jeu d'apprentissage comprenant un jeu d'entraînement et un jeu d'évaluation, tous deux relatifs à la prédiction d'un grade d'une tumeur gliale.

35 L'invention a également pour objet un procédé de construction d'un modèle de classification pour prédire le grade d'une tumeur gliale, comprenant au moins :

- la sélection d'un algorithme d'apprentissage automatique pour une tâche de classification,

- la fourniture d'un jeu de données d'apprentissage relatives à la prédiction d'un grade d'une tumeur gliale, comprenant un jeu d'entraînement et un jeu de test,
- une étape d'apprentissage de la prédiction d'un grade d'une tumeur gliale par ledit algorithme, à l'aide dudit jeu de données d'apprentissage.

5

Selon un autre aspect particulier, l'invention a pour objet un deuxième modèle de classification, préalablement entraîné sur un jeu de données d'apprentissage, pour prédire un statut de survie d'un individu atteint d'une tumeur, à partir d'un profil de nucléosides obtenu par la mise en œuvre d'un procédé selon l'invention.

10

Ledit modèle de classification pour prédire un statut de survie d'un individu atteint d'une tumeur comprend un algorithme d'apprentissage automatique préalablement entraîné et évalué, en particulier lors d'un apprentissage supervisé, avec un jeu de données d'apprentissage relatives à la prédiction d'un statut de survie d'un individu, ledit jeu d'apprentissage comprend un jeu d'entraînement et un jeu de test, tous deux relatifs à la

15

L'invention a également pour objet un procédé de construction d'un modèle de classification pour prédire un statut de survie d'un individu, comprenant au moins :

20

- la sélection d'un algorithme d'apprentissage automatique pour une tâche de classification,
- la fourniture d'un jeu de données d'apprentissage relatives à la prédiction d'un statut de survie d'un individu, comprenant un jeu d'entraînement et un jeu de test,
- une étape d'apprentissage de la prédiction d'un statut de survie d'un individu, par ledit algorithme, à l'aide dudit jeu de données d'apprentissage.

25

Selon un autre aspect, la présente invention a pour objet l'utilisation d'un modèle de classification selon l'invention pour la prédiction d'un grade d'une tumeur gliale.

30

Selon un aspect, la présente invention a pour objet l'utilisation d'un modèle de classification selon l'invention pour la stratification d'un patient atteint d'une tumeur gliale, en combinaison avec au moins un autre marqueur biologique caractéristique dudit patient.

35

Selon un autre aspect, la présente invention a pour objet l'utilisation d'un modèle de classification selon l'invention pour la prédiction d'un état de survie d'un individu.

Selon un autre mode de réalisation particulier, l'invention a pour objet un procédé *in vitro* de détection de la présence d'une tumeur chez un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :

- 5 a) isolement des nucléosides dudit échantillon biologique, par l'extraction : i) de l'ARN cellulaire total et sa fragmentation en nucléosides, ii) de l'ARN extracellulaire et sa fragmentation en nucléosides, et/ou iii) des nucléosides issus des catabolites monomériques, et de préférence des nucléosides issus des catabolites monomériques,
- 10 b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents obtenus lors de l'étape a), et
- c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de la présence de ladite tumeur.

15

Selon un mode de réalisation encore plus particulier, l'invention a pour objet un procédé *in vitro* de détection de la présence d'une tumeur chez un individu, à partir d'un échantillon sanguin isolé de cet individu, comprenant les étapes de :

- 20 a) isolement des nucléosides dudit échantillon biologique, par l'extraction des nucléosides issus des catabolites monomériques,
- b) isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents issus l'étape a),
- 25 c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur, et
- d) prédiction de la présence de ladite tumeur par un modèle de classification préalablement entraîné, à partir du profil établi lors de l'étape c).

- 30 Selon un mode de réalisation encore plus particulier, l'invention a pour objet un procédé *in vitro* de détection de la présence d'une tumeur colorectale chez un individu, à partir d'un échantillon sanguin isolé de cet individu, comprenant les étapes de :

- a. isolement des nucléosides dudit échantillon biologique, par l'extraction des nucléosides issus des catabolites monomériques,
- 35 b. isolement et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents issus l'étape a),

- c. établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur, et
- d. prédiction de la présence de ladite tumeur colorectale par un modèle de classification
- 5 préalablement entraîné, à partir du profil établi lors de l'étape c).

L'invention a également pour objet l'utilisation d'un procédé selon l'invention pour la détection de la présence d'une tumeur, ladite tumeur étant une tumeur située dans l'un des organes suivants : rectum, colon, sein, pancréas, rein, poumon, ou une tumeur

10 hématologique, notamment une leucémie.

L'invention a également pour objet l'utilisation d'un procédé selon l'invention pour la détection de la présence d'une tumeur du tube digestif, notamment une tumeur colorectale.

15 Selon un autre aspect, l'invention a pour objet un modèle de classification, préalablement entraîné sur un jeu de données d'apprentissage, pour détecter la présence d'une tumeur chez un individu, à partir d'un profil de nucléosides obtenu par la mise en œuvre d'un procédé selon l'invention. Ce modèle de classification comprend un algorithme

20 d'apprentissage automatique préalablement entraîné et évalué, en particulier lors d'un apprentissage supervisé, avec un jeu de données d'apprentissage relatives à la détection de la présence d'une tumeur chez un individu, ledit jeu d'apprentissage comprend un jeu d'entraînement et un jeu de test, tous deux relatifs à la détection de la présence d'une tumeur chez un individu.

25 Plus particulièrement, l'invention a pour objet un modèle de classification, préalablement entraîné sur un jeu de données d'apprentissage, pour détecter la présence d'une tumeur colorectale chez un individu, à partir d'un profil de nucléosides obtenu par la mise en œuvre d'un procédé selon l'invention. Ledit modèle de classification comprend un algorithme

30 d'apprentissage automatique préalablement entraîné et évalué, en particulier lors d'un apprentissage supervisé, avec un jeu de données d'apprentissage relatives à la détection de la présence d'une tumeur colorectale chez un individu.

L'invention a également pour objet un procédé de construction dudit modèle de classification pour la détection de la présence d'une tumeur, comprenant au moins :

35 -

- la sélection d'un algorithme d'apprentissage automatique pour une tâche de classification,

- la fourniture d'un jeu de données d'apprentissage relatives à la détection de la présence d'une tumeur chez un individu, comprenant un jeu d'entraînement et un jeu de test,

5 - une étape d'apprentissage de la prédiction de la présence d'une tumeur chez un individu, par ledit algorithme, à l'aide dudit jeu de données d'apprentissage.

Selon un mode de réalisation particulier, l'invention a également pour objet un procédé de construction d'un modèle de classification pour la détection de la présence d'une tumeur colorectale, comprenant au moins : la sélection d'un algorithme d'apprentissage
10 automatique pour une tâche de classification, la fourniture d'un jeu de données d'apprentissage relatives à la détection de la présence d'une tumeur colorectale chez un individu, comprenant un jeu d'entraînement et un jeu de test, et une étape d'apprentissage de la prédiction de la présence d'une tumeur colorectale chez un individu, par ledit algorithme, à l'aide dudit jeu de données d'apprentissage.

15

Selon un autre aspect, la présente invention a pour objet l'utilisation d'un modèle de classification selon l'invention pour la détection d'une tumeur, notamment une tumeur colorectale.

20 Selon un aspect, la présente invention a pour objet l'utilisation d'un modèle de classification selon l'invention pour la détection d'une tumeur, notamment une tumeur colorectale, en combinaison avec au moins un autre marqueur biologique caractéristique dudit patient.

Selon un autre aspect particulier, la présente invention a enfin pour objet une méthode de
25 diagnostic comprenant la mise en œuvre d'un procédé selon l'invention de caractérisation d'une tumeur. La présente invention a aussi pour objet une méthode de diagnostic comprenant la mise en œuvre d'un procédé selon l'invention de prédiction d'un grade d'une tumeur gliale. La présente invention a aussi pour objet une méthode de diagnostic comprenant la mise en œuvre d'un procédé selon l'invention de prédiction l'état de survie
30 d'un patient. Ladite méthode de diagnostic peut en outre comprendre une analyse histologique des tissus.

Selon un autre aspect particulier, la présente invention a enfin pour objet une méthode de diagnostic comprenant la mise en œuvre d'un procédé selon l'invention de détection d'une
35 tumeur. La présente invention a aussi pour objet une méthode de diagnostic comprenant la mise en œuvre d'un procédé selon l'invention de détection d'une tumeur colorectale. Ladite méthode de diagnostic peut comprendre une analyse histologique des tissus.

Description des figures et modes de réalisation

D'autres avantages et caractéristiques apparaîtront à l'examen de la description détaillée d'un mode de réalisation nullement limitatif, et des dessins annexés, dans lesquels :

- 5 La figure 1 représente le schéma global de l'expérimentation, où LC-MS/MS désigne la chromatographie liquide associée à la spectrométrie de masse et les données brutes (data) sont les profils épitranscriptomiques obtenus par LC-MS/MS.

- La figure 2 représente le schéma global du process bioinformatique, les données brutes
10 sont les profils épitranscriptomiques obtenus par LC-MS/MS, les données normalisées sont les profils épitranscriptomiques après normalisation, MS désigne la spectrométrie de masse (combinée à la chromatographie liquide).

- Les figures 3A, 3B et 3C représentent, sous forme de boîte à moustache, six graphiques
15 représentant respectivement la quantité relative (en pourcentage) de six nucléosides modifiés selon le grade de tumeur gliale. Pour chacun des graphes, ledit grade est désigné en abscisses par : « Normal », « Grade-II », « Grade-III » ou « Grade IV » indiquant respectivement un échantillon de tissu glial non tumoral ou un échantillon de tumeur gliale de grade II, III ou IV. La figure 3A montre deux exemples de nucléosides dont la quantité
20 diminue avec l'augmentation du grade de la tumeur gliale : (de gauche à droite) oxo8G et m1G. La figure 3B représente deux exemples de nucléosides dont la quantité augmente avec l'augmentation du grade de la tumeur gliale : (de gauche à droite) m6Am et Gm. La figure 3C représente deux exemples de nucléosides dont la quantité varie faiblement avec l'augmentation du grade de la tumeur gliale : (de gauche à droite) m1A et m7G. Les
25 échelles sont différentes selon les graphiques.

- La figure 4 représente le pourcentage de variance expliquée des premières composantes de l'Analyse en Composantes Principales (ACP) des profils épitranscriptomiques de la cohorte. En abscisse, les composantes sont numérotées à partir de 0 jusqu'à 9. En
30 ordonnée, les pourcentages de variance expliquée par ces composantes.

- La figure 5 représente la visualisation en trois dimensions des profils de la cohorte selon lesdites trois premières composantes de l'Analyse en Composantes Principales (ACP), c'est-à-dire les trois composantes qui détiennent $39,2 + 23,3 + 8,6 = 71,1\%$ de la variance des
35 profils épitranscriptomiques de la cohorte. Chacun des axes représente, respectivement, la composante principale 0 (39,24 %), la composante principale 1 (23,27 %) et la composante principale 2 (8,58%). Les symboles « étoile » représentent le garde

« normal », « triangle » le grade II, « carré » le grade III et « croix » le grade IV, respectivement.

- Il est bien entendu que les modes de réalisation qui seront décrits par la suite ne sont nullement limitatifs. On pourra notamment imaginer des variantes de l'invention ne comprenant qu'une sélection de caractéristiques décrites par la suite isolées des autres caractéristiques décrites, si cette sélection de caractéristiques est suffisante pour conférer un avantage technique ou pour différencier l'invention par rapport à de l'état de la technique antérieur. La présente invention se comprendra mieux à la lecture de l'exemple suivant, qui est donné pour illustrer l'invention et non pour en limiter la portée.

EXEMPLE 1 : Analyse des données transcriptomiques d'échantillons de cellules gliales

- Cette section présente la cohorte utilisée, la préparation des échantillons, la méthode d'obtention des profils épitranscriptomiques et le programme d'analyse informatique. Cette section présente ensuite les résultats d'analyse exploratoire des profils de la cohorte, de prédiction des grades des tumeurs et de prédiction de survie.
- La préparation des échantillons et l'obtention des profils par spectrométrie de masse sont réalisés comme suit : Cinquante-huit échantillons issus de tumeurs réséquées chirurgicalement chez des patients adultes diagnostiqués avec un gliome, aucun des patients n'ayant reçu de traitement chimique ou de radiothérapie avant la chirurgie, ont été utilisés conformément aux lois françaises de bioéthique concernant l'information et le consentement des patients. Au moment de la résection, pour chaque tumeur, une aliquote a été immédiatement congelée et conservée à -80°C et le tissu restant a été fixé dans du formol à 4%, incorporé dans de la paraffine et des sections de 3 microns ont été coupées et colorées avec de l'hématoxyline éosine. Le type histopathologique de la tumeur a été déterminé, selon la classification révisée de l'Organisation mondiale de la santé (Wesseling & Capper, « *WHO 2016 Classification of gliomas* ». *Neuropathol Appl Neurobiol.* 44, 139-150, 2018). Le groupe des tumeurs est constitué de gliomes de grade II (n = 20), de grade III (n = 20) et de glioblastomes de grade IV (n = 18). En outre, 19 échantillons « contrôle » de cellules gliales non tumorales (n = 19) ont été préparés selon le même protocole (décrit ci-après) que les échantillons tumoraux.
- L'ARN total a été extrait d'échantillons de tumeurs en utilisant la méthode acide-phénol guanidium. La qualité des échantillons d'ARN a été déterminée par électrophorèse sur gels d'agarose et coloration au bromure d'éthidium, et les bandes d'ARN 18S et 28S ont été

visualisées sous lumière UV. Le traitement de l'échantillon biologique débute par l'extraction de l'ARN par séparation de phase afin d'obtenir un échantillon d'ARN d'au moins 100 ng. Le traitement de poursuit par l'hydrolyse enzymatique de l'ARN polymérique et la déphosphorylation des nucléosides.

5

La digestion enzymatique de l'ARN est réalisée comme suit : une quantité de 400 ng d'ARN est diluée dans un volume total de 20 µL d'eau milliQ, auquel on ajoute 3 µl d'acétate d'ammonium (0,1 M pH 5,3) et 0,001 unité enzymatique (U) de Nuclease P1 (Sigma, N8630). Une incubation à 42°C est réalisée pendant 2 heures. Puis, 3 µl d'acétate d'ammonium 1 M et 0,001 U de phosphatase alcaline (Sigma, P4252) sont ajoutés. Le mélange est ensuite incubé à 37°C pendant 2 heures. Enfin, la solution de nucléosides est diluée deux fois et filtrée avec des filtres de 0,22 µm (Millex®-GV, Millipore, SLGVR04NL). Finalement, 5 µL de chaque échantillon est injecté et tous les échantillons sont analysés en triplicats par LC-MSMS.

10

15

La chromatographie liquide (LC) est réalisée comme suit : les nucléosides sont séparés par des systèmes Nexera LC-40 (Shimadzu) en utilisant une colonne Synergi™ Fusion-RP C18 (taille de particule 4 µm, 250 mm x 2 mm, 80 Å) (Phenomenex, 00G-4424-B0). La phase mobile est constituée d'acétate d'ammonium 5 mM ajusté à pH 5,3 avec de l'acide acétique (solvant A) et de l'acétonitrile pur (solvant B). Le gradient d'élution de 30 minutes commence avec 100 % de phase A suivi d'un gradient linéaire jusqu'à 8 % de solvant B à 13 minutes. Le solvant B est encore augmenté à 40 % en 10 minutes. Après 2 minutes, le solvant B est ramené à 0 % à 25,5 minutes. Les conditions initiales sont régénérées par rinçage avec 100 % de solvant A pendant 4,5 minutes supplémentaires. Le débit est de 0,4 ml/min et la température de la colonne de 35 °C.

20

25

La spectrométrie de masse en mode « Multiple Reaction Monitoring » (MRM) est réalisée comme suit : la détection est réalisée par Shimadzu TripleQuad 8060 en mode ions positifs. La spectrométrie de masse fonctionne en mode MRM dynamique avec une fenêtre de temps de rétention de 3 min et un temps de cycle maximum fixé à 258 ms. Les aires des pics sont déterminées à l'aide du logiciel Skyline 4.1 (Pino LK et al, « The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. » *Mass Spectrom Rev.* 2020 May;39(3):229-244. 2020).

30

35

Le spectromètre de masse a été calibré pour identifier et quantifier avec précision 25 nucléosides modifiés (Tableau 2) et 4 nucléosides non modifiés (A, U, G, T) (Tableau 1). L'appareil de spectrométrie de masse utilisé est un Shimadzu TripleQuad 8060 en mode suivi de réaction multiple (« Multiple Reaction Monitoring »).

Chaque échantillon a été injecté trois fois, fournissant ainsi trois répliques techniques. Pour chaque nucléoside, l'homogénéité du temps de rétention donné par le spectromètre de masse est vérifiée. Les mesures montrant une divergence de plus de 6% ont été écartées.

- 5 Il en résulte un tableau de données contenant les mesures de quantité de chaque nucléoside, dans chaque réplicat, pour tous les échantillons. Ce tableau est ensuite analysé grâce à nos programmes informatiques.

- 10 Toutes les analyses bio-informatiques sont effectuées avec des programmes python développés en interne. Pour cela les auteurs ont utilisé des modules (open source) bien connus : « Pandas » pour la gestion des données tabulaires (Reback *et al*, Pandas-dev/pandas : Pandas 1.0.3 (Version v1.0.3). Zenodo March 18, 2020), « scikit-learn » pour les analyses statistiques exploratoires des données et pour l'apprentissage automatique (Pedregosa *et al*, Journal of Machine Learning Research, vol. 12, pp. 2825-
15 2830, 2011), « Matplotlib » pour la visualisation (JD Hunter, Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007).

- Les caractéristiques nécessaires des programmes sont les suivantes : a) ils prennent en entrée des données de spectrométrie de masse dans un fichier au format tabulaire (format
20 CSV) ; b) les quantifications d'aire et de temps de rétention issues du spectromètre doivent être données sous forme de valeurs réelles avec une précision d'au moins 10⁻¹ ; c) ils implantent un algorithme d'apprentissage automatique supervisé multi-classe parmi ceux susmentionnés ; d) ils implantent la phase d'apprentissage, la phase d'évaluation, et le mode de prédiction ; e) ils utilisent le modèle de classification en mode prédiction pour
25 classifier le profil épitranscriptome d'un échantillon de patient afin de prédire le grade de la tumeur.

- Pour le prétraitement informatique et la normalisation, le tableau de quantités brutes est chargé en mémoire et son format est vérifié. Ensuite, la quantité moyenne de chaque
30 nucléoside est calculée, et le tableau est reformaté pour obtenir toutes les mesures sur une ligne pour chaque échantillon biologique. La spectroscopie de masse ne produit pas des décomptes absolus de molécules mais des mesures relatives. Les inventeurs proposent une formule nouvelle de normalisation, dans laquelle sont sommées les quantités des nucléosides non modifiés A, C, G et U. Cette somme sert de référence. Ensuite toutes les
35 mesures de référence sont divisées par cette somme. Ainsi, on obtient des mesures relatives, toutes comprises dans l'intervalle [0, 1]. A titre d'exemple, un extrait d'un tel tableau de données est proposé dans les tableaux 3, 4 et 5.

Tableau 3

Grade	A	Am	C	Cm	G	Gm	I	Psi	Queuosine	U
Grade-II	1,455 E-01	5,507 E-02	5,185 E-01	1,219 E-02	3,258 E-01	7,194 E-03	1,774 E-04	4,391 E-03	5,045E- 05	1,021 E-02
Grade-III	7,084 E-02	7,919 E-02	4,303 E-01	1,670 E-02	4,915 E-01	1,006 E-02	1,632 E-04	4,090 E-03	3,430E- 05	7,414 E-03
Grade-IV	1,360 E-01	7,150 E-02	3,681 E-01	1,776 E-02	4,894 E-01	9,499 E-03	2,988 E-04	3,922 E-03	2,500E- 05	6,571 E-03
Normal	1,599 E-01	6,030 E-02	5,014 E-01	1,406 E-02	3,286 E-01	7,311 E-03	1,992 E-04	2,928 E-03	9,334E- 05	1,007 E-02

Tableau 4

Grade	Um	ac4C	hm5C	m1A	m1G	m227G	m27G	m3C	m3Um	m5C
Grade-II	9,863E- 05	1,255E- 04	2,075E- 06	3,333E- 02	1,073E- 03	1,300E- 04	3,744E- 06	3,017E- 03	2,522E- 06	2,728E- 02
Grade-III	1,489E- 04	8,378E- 05	3,249E- 07	2,740E- 02	3,793E- 04	2,149E- 04	3,739E- 06	2,587E- 03	4,434E- 06	2,273E- 02
Grade-IV	1,760E- 04	1,112E- 04	1,660E- 06	5,166E- 02	6,559E- 04	1,593E- 04	6,544E- 06	5,398E- 03	5,382E- 06	4,159E- 02
Normal	1,246E- 04	1,780E- 04	3,208E- 06	4,849E- 02	1,914E- 03	1,553E- 04	9,019E- 06	4,470E- 03	1,799E- 06	3,654E- 02

5

Tableau 5

Grade	m66A	m66A m	m6A	m6Am	m7G	mcm5 U	mcm5s2 U	ncm5U	oxo8G
Grade-II	3,035E-03	1,746E-06	2,785E-03	1,303E-04	6,834E-03	1,599E-06	2,199E-05	3,996E-05	2,454E-06
Grade-III	4,077E-03	8,600E-07	4,115E-03	2,538E-04	6,673E-03	9,258E-07	3,282E-05	1,757E-05	2,357E-07
Grade-IV	3,899E-03	9,457E-07	4,110E-03	3,981E-04	1,185E-02	2,162E-06	7,748E-05	7,354E-05	1,060E-06
Norma I	3,368E-03	6,777E-07	3,183E-03	1,283E-04	9,999E-03	8,795E-06	1,118E-05	5,544E-05	5,862E-06

Les tableaux 3, 4 et 5 indiquent, pour chacun des nucléosides analysés, la valeur de donnée normalisée pour chacun des grades II, III et IV de gliome, et pour les tissus sains (« normal »).

L'analyse conjointe des profils épitranscriptomiques et de variables cliniques d'intérêt (Fig. 1) est réalisée, en particulier concernant le grade dans le cas des gliomes. Ce processus peut être adapté à tout type de variable clinique. Dans cet exemple, nous cherchons à distinguer les grades de cancer, qui peuvent être difficiles à établir au moyen d'un examen anatomopathologique.

Le prétraitement des profils de la cohorte a abouti à un tableau de 77 lignes, avec une ligne par échantillon, et 29 colonnes, avec une colonne par mesure. Pour chacun des échantillons, la mention des grades des tumeurs ou la mention « normal » pour les échantillons sains, a été ajoutée. Une analyse statistique exploratoire de ce tableau a été réalisée pour évaluer la pertinence du signal contenu dans les profils sur l'information de grade.

Premièrement, les variations des quantités de chaque nucléoside sont étudiées dans les échantillons d'un même grade, et comparons ces variations entre les grades. Comme présenté dans les graphiques en boîtes à moustaches des Figures 3A, 3B et 3C, les résultats expérimentaux suggèrent un regroupement des nucléosides en quatre groupes : i) ceux dont la quantité augmente avec le grade, c'est-à-dire entre le tissu cérébral non tumoral (désigné par simplification comme « normal » en ordonnées des graphiques) et les grades, II, III et IV, notamment les nucléosides oxo8G, m1G, queuosine et Ac4C (comme montré

par exemple dans Fig. 3A) ; ii) ceux dont la quantité diminue avec le grade (comme montré par exemple dans Fig. 3B), iii) ceux qui varient faiblement avec les grades (comme montré par exemple dans Fig. 3C) et iv) les nucléosides restants, qui ne satisfont pas les conditions d'appartenance aux trois premiers groupes.

5

À première vue, aucun de ces groupes n'est lié à une caractéristique spécifique connue de ses constituants (par exemple, bord modifié du nucléoside). Néanmoins, il convient de noter que les 2'-O-méthylations (Am, Um, Cm, Gm), principalement trouvées dans l'ARN ribosomal (ARNr) et le petit ARN nucléaire (snRNA), se comportent de manière similaire dans un cluster central contenant m6Am, une modification spécifique de l'ARNr.

10

Ensuite, une Analyse en Composante Principale (ACP) de ces données a été réalisée, afin d'effectuer une réduction de dimension, à ne pas confondre avec une sélection des "caractéristiques", autrement dit des nucléosides, pour voir si les variations de quantité pourraient être combinées en un petit nombre de composantes. Fig. 4 montre le pourcentage de variance expliquée pour les 10 premières composantes de l'ACP : clairement les trois premières composantes regroupent une large majorité des variations des profils. On constate en effet que les trois premières composantes regroupent à elles seules : $39,2 + 23,3 + 8,6 = 71,1\%$ de la variance des profils épitranscriptomiques de la cohorte.

15

20

Chaque profil épitranscriptomique comprenant les mesures pour x nucléosides est vu mathématiquement parlant comme un point dans un espace à x dimensions. L'ACP est une méthode d'analyse multivariée exploratoire qui permet de réduire les dimensions des données tout en capturant leur variabilité. Les composantes sont de nouvelles variables qui combinent les données des observations initiales afin de capturer au mieux leur variabilité tout en réduisant le nombre de variables à analyser. Les composantes résultent de la projection des données initiales sur d'autres axes de l'espace multidimensionnel. Les composantes sont ordonnées par ordre décroissant de pourcentage de variance expliquée. Ce pourcentage associé à chaque composante indique son importance pour décrire les données initiales. La Fig. 4 présente le graphique du pourcentage de variance expliquée pour les 10 premières composantes. L'ACP est une technique classique de l'analyse de données.

25

30

35

La visualisation en 3 dimensions des profils projetés sur les trois premières composantes est montrée en Fig. 5. Premièrement, les échantillons de tissu non tumoral et de grade II se séparent clairement de ceux de grades III et IV. Par ailleurs, les échantillons de grade III occupent un volume relativement séparé de ceux de grade IV. Ces résultats

exploratoires suggèrent que des algorithmes d'apprentissage automatique supervisé devraient pouvoir apprendre une frontière entre les groupes d'échantillons de différents grades.

5 Méthode d'apprentissage automatique permettant la prédiction avec précision du grade des tumeurs et des échantillons sains

Une méthode d'apprentissage automatique a été testée pour déterminer si le grade des échantillons pouvait être prédit à partir des seuls profils épitranscriptomiques, c'est-à-dire sans l'aide d'aucune autre information que les quantités de nucléosides (Fig. 1). Pour ce faire, les profils ont été partitionnés en deux sous-ensembles distincts : le premier a été utilisé uniquement pour entraîner le modèle d'apprentissage automatique (n=60, soit 78%), le second a servi à évaluer le modèle (n=17, soit 22%).

Comme la variable à prédire (ici le grade) est une donnée catégorielle, la méthode d'apprentissage doit appartenir à la catégorie de classification. Parmi les grands types d'algorithmes d'apprentissage, un algorithme de classification de Machine à Vecteurs Support (SVM pour Support Vector Machine, en anglais) a été choisi pour la possibilité qu'il offre d'adapter les formules des frontières en changeant de type de noyau, comme c'est la norme en apprentissage. La précision de prédiction de l'algorithme SVM équipé d'un noyau linéaire sur les profils du sous ensemble de testé est de 0,90, sur un maximum de 1, ce qui est remarquable. Le niveau de précision de prédiction se maintient lorsque l'on réitère l'apprentissage puis les tests avec de nouveaux partitionnements aléatoires du jeu de données, ce qui montre la robustesse de l'outil d'apprentissage développé.

En outre, les résultats de l'évaluation permettent de comparer notre méthode de normalisation (désignée par SUM, pour somme) avec les formules utilisées dans la littérature. En effet, la normalisation classique qui consiste à diviser la mesure d'un nucléoside modifié, par exemple m1A, par celle du nucléoside correspondant non modifié, ici la mesure de A. Dans le tableau 6, la précision selon l'utilisation de différentes formules est comprise entre 0,8 et 0,9, et est donc toujours inférieure ou égale (mais jamais supérieure) à la précision de la formule de normalisation SUM.

Tableau 6

Normalisation	A	C	G	U	SUM
Précision de la prédiction	0.85	0.90	0.85	0.80	0.90

En outre, la prédiction des grades est robuste au changement d'algorithme de classification. Au lieu d'un algorithme SVM, si un algorithme basé sur une approche d'Analyse Discriminante Linéaire est utilisé, une précision de 92 % est obtenue, avec un
5 rappel (ou sensibilité) de 90% et un F1-score de 90%. Le détail des prédictions pour chaque grade est donné dans le tableau 7.

Tableau 7

Grade	précision	rappel	score-f1
Normal	1.00	0.80	0.89
Grade-II	0.67	1.00	0.80
Grade-III	1.00	0.83	0.91
Grade-IV	0.88	1.00	0.93
Moyenne pondérée	0.92	0.90	0.90

10

En conclusion, la qualité de prédiction des grades n'est pas particulièrement liée à l'optimisation d'une méthode d'apprentissage sur une cohorte donnée, puisque deux méthodes d'apprentissage très différentes obtiennent des résultats similaires. La qualité de la prédiction est donc liée à la puissance du signal contenu dans les profils
15 transcriptomiques.

20

En outre, les modèles d'apprentissage dont les résultats sont rapportés ici n'ont volontairement pas été optimisés vis-à-vis de leurs paramètres, afin d'éviter un risque de sur-apprentissage qui entacherait la capacité de généralisation des modèles.

Prédiction du statut de survie des patients

La même démarche par apprentissage supervisé a été utilisée pour prédire la variable clinique indiquant le statut de survie, c'est-à-dire le statut « vivant » ou « décédé » à la fin du suivi de la cohorte, c'est-à-dire en 2020. Ici, la classification est binaire : « vivant »
25 ou « décédé ». L'algorithme d'apprentissage SVM donne une prédiction correcte à 80%, ce qui est convaincant au regard de la taille de la cohorte considérée (Tableau 8).

Tableau 8

Classe	précision	rappel	score-f1
Faux (vivant)	0.75	0.86	0.80
Vrai (décédé)	0.89	0.80	0.84
Moyenne pondérée	0.83	0.82	0.82

Conclusion

- 5 Des différences de quantités relatives de certaines modifications épigénétiques des ARN ont été mises en évidence selon différents échantillons, qu'ils soient sains ou tumoraux. Ces différences permettent notamment de séparer les différents grades tumoraux. Un algorithme d'apprentissage automatique supervisé appliqué aux vecteurs de quantités de nucléosides permet de distinguer efficacement les grades des gliomes, et permet en particulier de distinguer les grades II et III, avec une précision remarquable compte tenu de la taille relativement restreinte de la cohorte. Par ailleurs, ce procédé permet également, à partir des mêmes données, d'estimer la survie des patients à l'aide d'une méthode d'apprentissage automatique supervisé.

15 **EXEMPLE 2 : Analyse des données transcriptomiques d'échantillons de sérum sanguin de sujets**

- Quarante-sept échantillons sanguins de patients adultes diagnostiqués avec un cancer colorectal ou de sujets contrôle (Etablissement Français du Sang, n=20) aucun des patients n'ayant reçu de traitement chimique ou de radiothérapie avant la chirurgie, ont été utilisés conformément aux lois françaises de bioéthique concernant l'information et le consentement des patients. Un comité d'éthique local (CORT) a évalué et autorisé l'utilisation de ces échantillons.

- L'ARN circulant est extrait à partir de plasma en utilisant un kit (miRNeasy Serum/Plasma). Les ARNs sont digérés avec de la nucléase P1 et traités avec de la phosphatase alcaline afin d'obtenir un mélange de nucléosides. Les nucléosides libres circulants sont extraits des mêmes échantillons de plasma en utilisant une procédure d'extraction par le méthanol. Ils ne nécessitent pas de traitement enzymatique avant passage en spectrométrie de masse.

- La chromatographie liquide (LC) et la spectrométrie de masse et les analyses bioinformatiques sont réalisées comme indiqué dans l'exemple 1. En particulier, chaque échantillon est analysé trois fois indépendamment, permettant ainsi d'obtenir trois

répliques techniques pour chacun. Les données brutes sont traitées afin d'être normalisées comme dans l'exemple 1. Ainsi, ces étapes produisent un profil épitranscriptomique par échantillon. Une méthode d'apprentissage automatique a été développée et testée pour déterminer la présence ou non d'une tumeur à partir des seuls profils épitranscriptomiques, c'est-à-dire sans l'aide d'aucune autre information que les quantités de nucléosides. Comme dans l'exemple 1, un algorithme de classification de Machine à Vecteurs Support (SVM pour Support Vector Machine, en anglais) muni d'un noyau linéaire a été choisi pour cette tâche dite de « classification » binaire. L'algorithme a d'abord été entraîné et ensuite testé afin d'évaluer sa capacité de prédiction de la présence ou non d'une tumeur dans l'échantillon. Avec les profils épitranscriptomiques contenant les mesures des nucléosides libres, la méthode d'apprentissage automatique donne une prédiction avec une précision de 100 % et une sensibilité de 100 %.

REVENDEICATIONS

1. Procédé *in vitro* de caractérisation d'une tumeur d'un individu, à partir d'un échantillon biologique isolé de cet individu, comprenant les étapes de :
 - 5 a) isolement des nucléosides dudit échantillon biologique, par l'extraction : i) de l'ARN cellulaire total et sa fragmentation en nucléosides, ii) de l'ARN extracellulaire et sa fragmentation en nucléosides et/ou iii) des nucléosides issus des catabolites monomériques,
 - 10 b) isolement par chromatographie et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents issus de l'étape a), et
 - c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides, obtenues lors de l'étape b), ledit profil étant caractéristique de ladite tumeur.
- 15 2. Procédé selon la revendication précédente, dans lequel ledit échantillon biologique est une biopsie ou un échantillon biologique liquide, ledit échantillon biologique liquide étant choisi parmi le sang, le sérum, le plasma et l'urine.
- 20 3. Procédé selon l'une quelconque des revendications 1 ou 2, dans lequel lesdits nucléosides sont choisis parmi les suivants :
 - les nucléosides non modifiés : adénosine (A), cytidine (C), guanosine (G), uridine (U), et
 - les nucléosides modifiés : 2'-O-méthyladénosine (Am), 1-méthyladénosine (m1A), N6,N6-diméthyladénosine (m66A), N6,N6,2'-O-triméthyladénosine (m66Am), N6-méthyladénosine (m6A), N6,2'-O-diméthyladénosine (m6Am), N4-acetylcytidine (ac4C), 2'-O-méthylcytidine (Cm), 5-hydroxyméthylcytidine (hm5C), 3-méthylcytidine (m3C), 5-méthylcytidine (m5C), 2'-O-méthylguanosine (Gm), 1-méthylguanosine (m1G), N2,N2,7-triméthylguanosine (m227G), N2,7-diméthylguanosine (m27G), 7-méthylguanosine (m7G), 8-hydroxyguanosine (oxo8G), inosine (I), pseudouridine (Psi), queuosine (Q), 3,2'-O-diméthyluridine (m3Um), 5-méthoxycarbonylméthyl-2-thiouridine (mcm5s2U), 5-méthoxycarbonylméthyluridine (mcm5U), 5-carbamoylméthyluridine (ncm5U), 2'-O-méthyluridine (Um).
- 35 4. Procédé selon l'une quelconque des revendications précédentes, caractérisé en ce que ladite tumeur est une tumeur gliale et en ce qu'il comprend une étape de

prédiction d'un grade de ladite tumeur gliale par un premier modèle de classification préalablement entraîné, à partir du profil établi lors de l'étape c.

5. Procédé selon la revendication 4, caractérisé en ce que le premier modèle de classification comprend :
- un algorithme d'apprentissage automatique,
 - un réseau neuronal à apprentissage supervisé, ou
 - un algorithme de classification probabiliste multi-classes, préalablement entraîné avec un jeu de données d'apprentissage.
6. Procédé selon l'une quelconque des revendications 4 ou 5, dans lequel la prédiction d'un grade d'une tumeur gliale est choisie parmi : la prédiction d'une tumeur gliale de grade II, la prédiction d'une tumeur gliale de grade III et la prédiction d'une tumeur gliale de grade IV.
7. Procédé selon l'une quelconque des revendications précédentes, caractérisé en ce qu'il comprend une étape de prédiction d'un état de survie dudit individu, par un deuxième modèle de classification préalablement entraîné, à partir du profil établi lors de l'étape c).
8. Procédé selon la revendication 7, caractérisé en ce que le deuxième modèle de classification comprend :
- un algorithme d'apprentissage automatique,
 - un réseau neuronal à apprentissage supervisé, ou
 - un algorithme de classification probabiliste, préalablement entraîné avec un jeu de données d'apprentissage.
9. Procédé *in vitro* de détection de la présence d'une tumeur chez un individu, à partir d'un échantillon biologique isolé dudit individu, comprenant les étapes de :
- a) isolement des nucléosides dudit échantillon biologique, par l'extraction : i) de l'ARN cellulaire total et sa fragmentation en nucléosides, ii) de l'ARN extracellulaire et sa fragmentation en nucléosides et/ou iii) des nucléosides issus des catabolites monomériques,
 - b) isolement par chromatographie et détermination d'une quantité respective d'au moins 3, de préférence au moins 5, de préférence au moins 10 de préférence au moins 20, nucléosides différents issus de l'étape a), et

c) établissement, pour ledit échantillon biologique, d'un profil de nucléosides à partir des quantités respectives de chacun des nucléosides, obtenues lors de l'étape b), ledit profil étant caractéristique de la présence de ladite tumeur.

- 5 10. Modèle de classification, préalablement entraîné sur un jeu de données d'apprentissage, pour prédire, dans un procédé selon l'une quelconque des revendications 4 à 8
- un grade d'une tumeur, et/ou
 - un état de survie d'un individu,
- 10 à partir du profil établi lors de l'étape c).
11. Utilisation d'un modèle de classification selon la revendication 9, pour la prédiction d'un grade d'une tumeur et/ou pour la prédiction d'un état de survie d'un individu.
- 15 12. Utilisation d'un procédé selon l'une des revendications 1 à 3 pour la détection d'une tumeur.
13. Utilisation selon la revendication 12 pour la détection d'une tumeur colorectale.