

Exploration de l'âge de méthylation de l'adénocarcinome pulmonaire humain (LUAD)

De plus en plus d'études suggèrent que certaines modifications dues au vieillissement se manifestent via l'épigénétique. Dans cette idée Steve Horvath a réalisé une étude en 2013 afin de prédire l'âge d'un individu à partir de la méthylation de son ADN (Horvath, 2013). Il a ensuite étendu son étude aux tissus cancéreux. Il a ainsi démontré que dans 20 types de cancers l'accélération de l'âge était significativement altérée. Une faible accélération est associée à certaines mutations somatiques et une forte accélération est plutôt associée à des mutations dans les récepteurs stéroïdiens.

Afin d'étudier l'âge méthylé et ses conséquences dans les cancers humains nous utilisons les données provenant de la base de données *The Cancer Genome Atlas*.

Analyse de l'âge méthylé par rapport à l'âge chronologique

1. L'âge de méthylation tumoral semble globalement plus bas que l'âge réel du patient

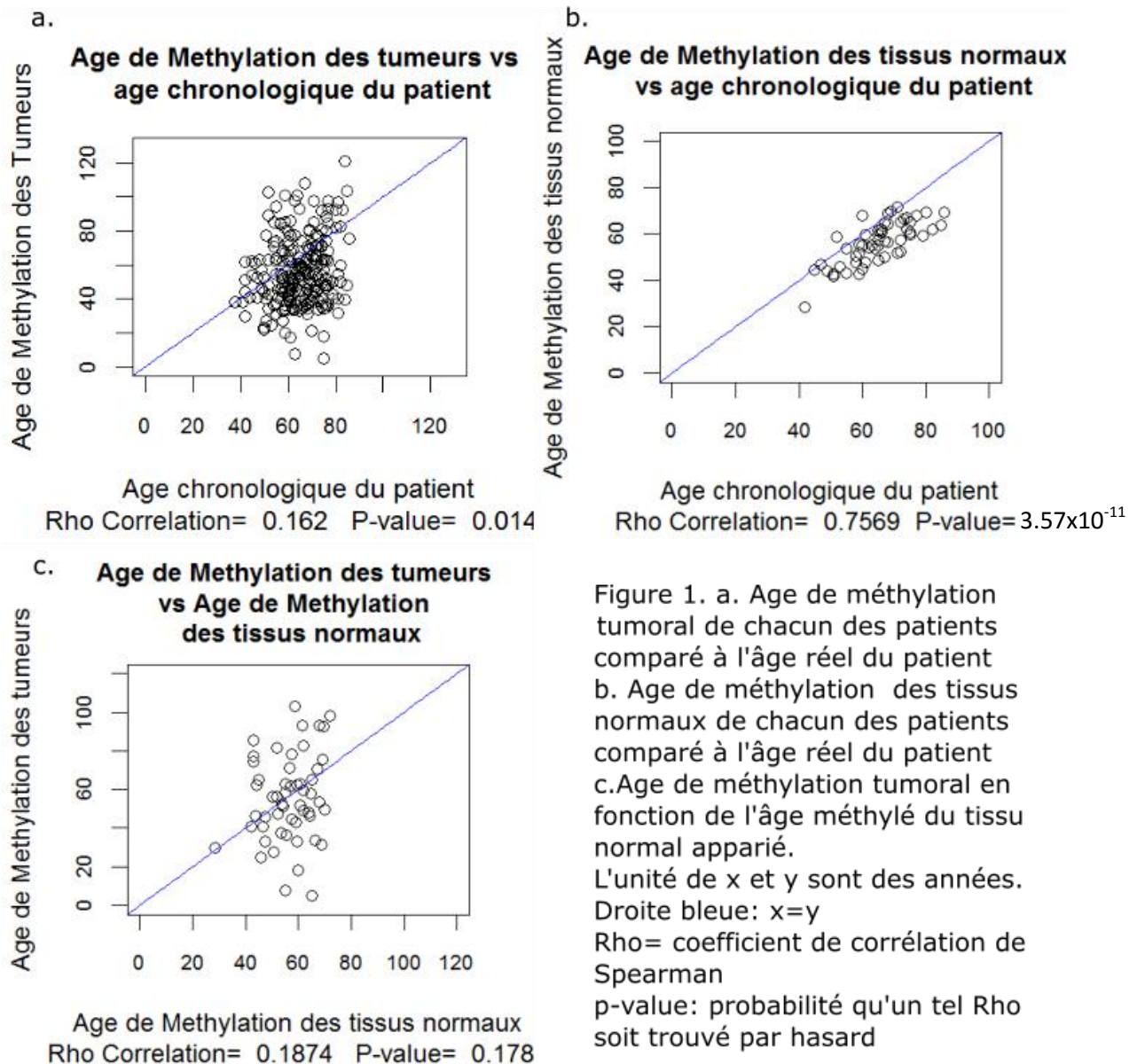
En comparant l'âge de méthylation tumoral à l'âge réel du patient correspondant on observe que l'âge de méthylation tumoral a une légère tendance globale à être plus bas que l'âge réel. En effet on observe une plus grande masse de points rassemblée sous la diagonale ($x=y$) (Fig.1a.). Ceci suggère que les tumeurs seraient plus jeunes. Cependant, on observe également que l'âge de méthylation est assez dispersé allant d'une quinzaine d'années à plus de 100 ans, tandis que l'âge réel est situé dans une fenêtre plus étroite entre une quarantaine d'années et moins de 90 ans. La corrélation de Spearman entre ces deux variables n'est pas très élevée ($p=0.16$) et tout juste significative ($p\text{-valeur}=0.014$). Cette tendance des tumeurs à être plus jeunes que l'âge chronologique est peut-être simplement due à l'imprécision de la mesure de l'âge de méthylation par rapport à l'âge réel. En conséquence, il est intéressant de comparer l'âge de méthylation calculé pour les tissus normaux par rapport à l'âge réel du patient correspondant.

2. L'âge de méthylation du tissu normal est positivement corrélé à l'âge chronologique du patient correspondant

En comparant l'âge de méthylation du tissu sain avec l'âge réel du patient correspondant on observe une forte corrélation de Spearman de 0.76 très significative avec une p-valeur de 3.57×10^{-11} . De manière intéressante, on observe un décalage des points sous la diagonale ($x=y$) (Fig.1b). Ceci suggère que l'âge méthylé mesuré est globalement plus petit que l'âge réel malgré que ces ceux-ci soient fortement corrélés. Il semble même que ce décalage de quelques années soit assez systématique quel que soit l'âge réel. Contrairement à l'âge de méthylation tumoral pour lequel le décalage par rapport à l'âge réel semblait moins régulier.

3. Les âges méthylés tumoraux et normaux ne sont pas corrélés de manière significative

En comparant l'âge méthylé tumoral avec l'âge méthylé du tissu sain correspondant on n'observe pas de corrélation de Spearman significative (p-valeur de 0.18 est > 0.05). Malgré le fait que certains points sont situés sur la diagonale ($x=y$) on ne peut observer de tendance globale (Fig.1c). Puisque l'âge chronologique, principalement dans les tissus normaux, semble avoir une influence sur l'âge méthylé mesurée il serait intéressant d'essayer de corriger cette influence dans les tumeurs en définissant l'accélération de l'âge méthylé par rapport à l'âge chronologique et également d'évaluer s'il existe une différence si cette correction se fait via l'âge méthylé du tissu normal apparié.



4. Plusieurs définitions de l'accélération de l'âge méthylé sont envisageables

En comparant le ratio de l'âge méthylé dans les tumeurs sur l'âge méthylé dans les tissus normaux (âge mét. tumeur/âge mét. normal), au ratio de l'âge méthylé dans les tumeurs sur l'âge chronologique du patient correspondant (âge mét. tumeur/âge chron.) nous obtenons une corrélation très significative ($< 2e-16$) et très forte de 0.94 (données non montrées). Ceci n'est pas étonnant puisque l'âge de méthylation des tissus normaux et l'âge chronologique sont fortement corrélés (0.76). Nous en déduisons que choisir l'une ou l'autre définition de l'accélération n'impliquera pas de conséquences majeures. Nous choisissons de manière

arbitraire le ratio de l'âge méthylé tumoral sur l'âge chronologique. Ceci nous permet de traiter plus d'échantillons et d'augmenter ainsi notre puissance statistique. En effet nous n'avons que 56 âges méthylés pour les échantillons sains contre 353 âges méthylés pour les échantillons tumoraux. Dans les données cliniques nous retrouvons 228 âges chronologiques correspondants aux patients pour lesquels nous avons l'âge tumoral méthylé.

Une autre définition de l'accélération peut être envisagée et est retrouvée dans l'article d'origine. Celle-ci se mesure comme la différence entre l'âge méthylé et l'âge chronologique.

Nous analysons les distributions des accélérations mesurées soit comme une différence soit comme un ratio via un QQ plot.

Nous observons qu'en considérant l'accélération de l'âge méthylé comme un ratio nous obtenons une distribution relativement normale de l'accélération de l'âge méthylé pour les tissus normaux et les tissus tumoraux (Fig2.ab). De même, en considérant que l'accélération de l'âge méthylé se mesure comme une différence, la distribution des valeurs suit bien le modèle théorique et est donc relativement normale (Fig2.cd).

Cependant lorsqu'on détaille un peu plus cette distribution via un histogramme couplé à un graphe de densité on observe que la distribution de l'accélération mesurée par la différence semble plus normale (Fig.3abcd) mais sans différence majeure.

Dans notre étude nous avons décidé de considérer l'accélération comme la différence entre l'âge méthylé et l'âge chronologique, afin de garder un ordre de grandeur en années et non un ratio, même si les deux définitions conviendraient probablement.

En comparant l'accélération de l'âge méthylé des tissus normaux à l'accélération de l'âge des tissus tumoraux on remarque qu'il n'y a pas de lien significatif entre ces valeurs, $\rho = 0.168$ et une p-valeur de 0.23 (données non montrées). L'accélération de l'âge du tissu normal ne semble donc pas avoir d'influence significative sur l'accélération de l'âge de la tumeur.

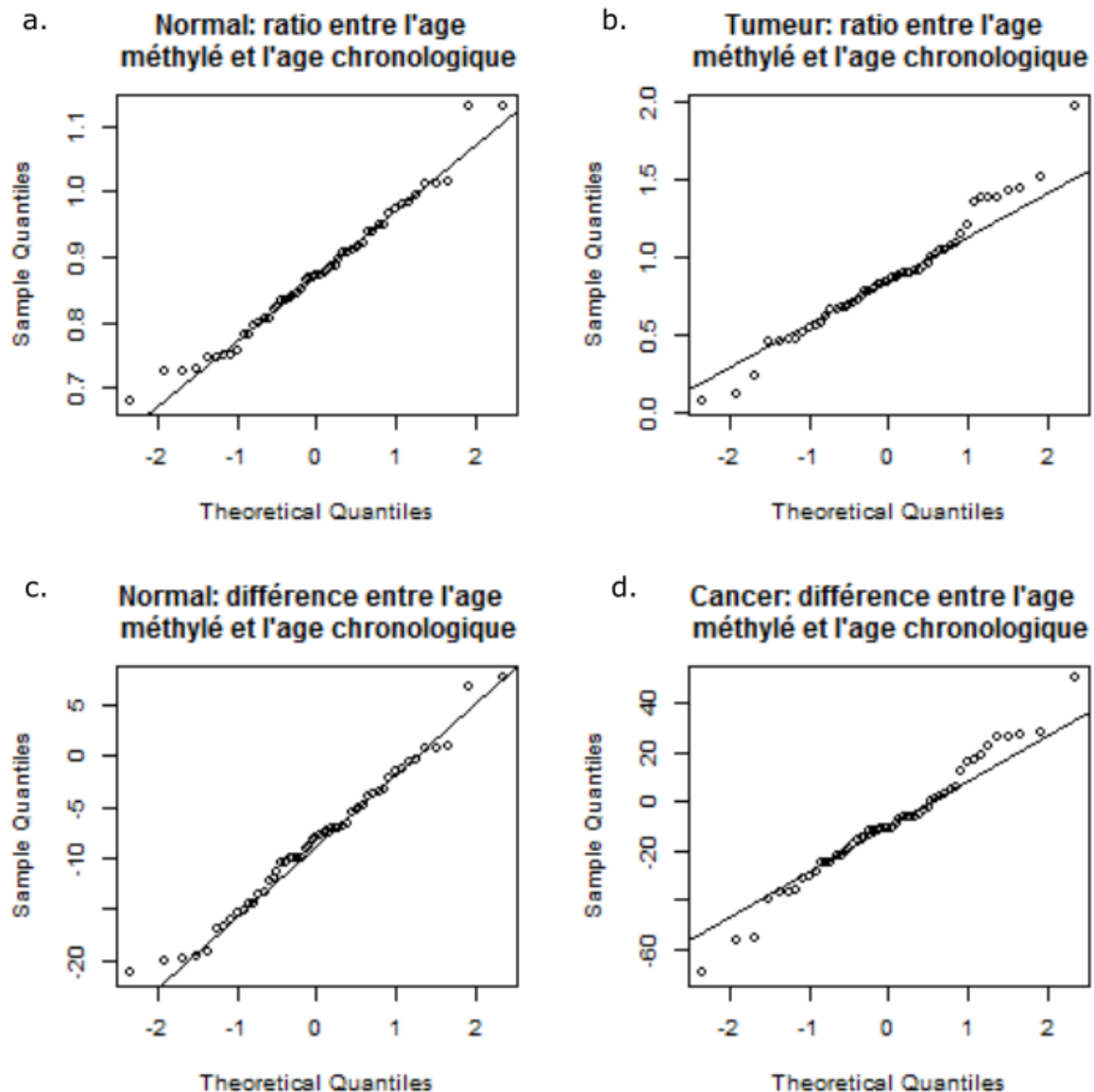


Figure2: QQplot distribution des quantiles des accélérations de l'âge méthylé en fonction des quantiles théoriques suivant une distribution normale. Haut: Accélération mesurée comme le ratio entre l'âge méthylé et l'âge chronologique du patient correspondant. a. distribution de l'accélération des tissus normaux. b. distribution de l'accélération des tissus tumoraux. Bas: Accélération mesurée comme la différence entre l'âge méthylé et l'âge chronologique. c. distribution de l'accélération des tissus normaux. d. distribution de l'accélération des tissus tumoraux

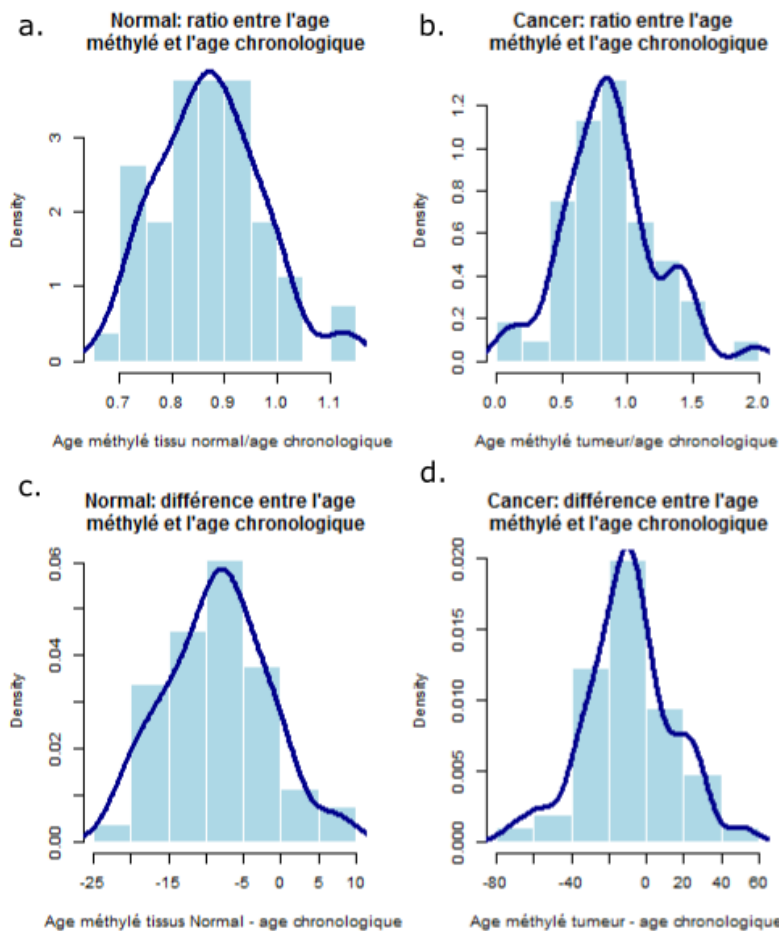


Figure.3. histogramme couplé à un graph de densité des accélérations mesurées soit par le ratio entre l'âge méthylé et l'âge normal (ab) soit par la différence entre l'âge méthylé et l'âge chronologique (cd). Pour les tissus normaux (ac) et les tissus tumoraux (bd)

Analyse de l'âge méthylé par rapport aux variables cliniques

1. Il ne semble pas exister de lien entre l'âge méthylé et une variable clinique

En comparant l'âge de méthylation à différentes variables cliniques aucune ne semble spécifiquement liée à une quelconque variation de l'âge de méthylation tumoral observé.

Lorsqu'on compare les âges méthylés tumoraux retrouvés dans les différents groupes décrivant les stades tumoraux on n'observe pas de différence. Que ce soit concernant la catégorie N donnant une idée de la propagation tumorale dans les ganglions lymphatiques (Annexe Fig1a.), la catégorie T donnant la catégorie de la tumeur primaire (Annexe Fig1b.), ou la catégorie M donnant une idée de l'apparition de métastases (Annexe Fig1c.). Les médianes sont similaires et les intervalles de confiance à 95% autour de la médiane (symbolisés par la partie plus étroite de la boîte) se recoupent (visible par la droite bleue qui

traverse tous les intervalles). Lorsqu'il n'y a pas de chevauchement entre ces intervalles on peut considérer que les médianes diffèrent (Chambers, 1983). Un test statistique n'est donc pas nécessaire pour confirmer que les médianes sont semblables. Dans les stades tumoraux concernant les métastases la catégorie m1a a une médiane qui semble différer à première vue des autres, cependant ce groupe ne comprend qu'un seul échantillon. À noter que dans les catégories t4 et tx (Annexe Fig1b.) l'intervalle de confiance à 95% dépasse le range interquartile (IQR) donnant cette forme atypique au boxplot.

Lorsqu'on compare les âges méthylés retrouvés dans les différents groupes histologiques il ne semble pas y avoir de différence significative. Les médianes sont semblables et les intervalles de confiance se recoupent pour la grande majorité des groupes (Annexe Fig2a.). La majorité des échantillons (161) sont retrouvés dans la catégorie non spécifiée (not otherwise specified (nos)), une grande partie (50) se retrouve dans la catégorie mixte (mixed subtype) tandis que les autres catégories comprennent 10 échantillons ou moins. Le nombre d'échantillons exacte peut être retrouvé sur le graphique indiqué par "n=" au dessus de chaque boxplot. La catégorie "lung acinar adenocarcinoma" (en rouge) est la seule ayant un intervalle de confiance autour de la médiane ne se recoupant pas avec tous les autres groupes. Afin de tester la différence des valeurs retrouvées dans ce groupe par rapport à celles retrouvées dans les autres, nous réalisons un test des rangs de Wilcoxon entre ce groupe et le groupe "not otherwise specified (nos)". Le groupe "nos" a été choisi pour cette comparaison car il comprend le plus d'échantillons, l'intervalle de confiance le plus étroit et est l'intervalle qui ne se recoupe pas avec "lung acinar adenocarcinoma". Le test nous rend une p-valeur de 0.057. Nous ne rejetons pas l'hypothèse nulle selon laquelle les deux distributions sont égales. On peut donc conclure que ces distributions sont similaires. Le groupe "lung acinar adenocarcinoma" contient seulement 6 échantillons ce qui est peu pour faire une bonne analyse statistique. À noter qu'il n'y a pas de chevauchement également avec les catégories "micropapillary adenocarcinoma" et "clear cell adenocarcinoma" mais ces catégories comprennent respectivement 3 et 1 échantillons, ce qui est trop peu pour réaliser un test statistique.

Il ne semble pas y avoir de différence significative entre les médianes d'âges de méthylation retrouvées dans les différents stades de la maladie (Annexe Fig2b.). En effet tous les intervalles de confiances se recoupent excepté le stade I qui ne comprend qu'un seul échantillon.

Lorsqu'on compare le nombre de paquets de cigarettes fumés par an à l'âge de méthylation on n'observe pas de corrélation significative entre les deux (p -valeur=0.3) (Annexe Fig2c.).

Lorsqu'on compare l'âge méthylé à l'année de début de tabagisme on n'observe pas de subdivision particulière des groupes. Cependant on calcule une corrélation négative de -0.2 et une p -valeur juste significative de 0.02 (données non montrées).

Lorsqu'on compare la tendance de l'âge de méthylation à la prise de thérapie ou non, à la classe de la tumeur résiduelle, à l'origine ethnique du patient, à son sexe, à son origine, ou encore à l'indice de Karnofsky (allant de 100% pas de maladie, à 10% processus fatal progressant rapidement) on n'observe aucune différence apparente entre les groupes. Les médianes sont très proches et la droite bleue nous indique que les intervalles de confiance se recourent (Annexe Fig3abcdeg).

Finalement lorsqu'on compare les médianes des âges de méthylations retrouvées chez les patients en vie ou décédés on n'observe pas non plus de différence significative (Annexe Fig3f). Cependant, ce test n'est pas approprié pour ce genre d'analyse de survie. L'idéal est de faire un test statistique de Cox pour l'âge de méthylation tumoral en fonction de la survie. Ce test nous donne un coefficient bêta ("hazard ratio") quasiment égal à 1 (0.993) et une probabilité élevée de trouver par hasard une valeur égale ou plus élevée en valeur absolue ($\Pr(>|z|) = 0.24$). Ceci veut dire que la survie des patients atteints de LUAD qui peut être représentée par la courbe Fig4 ne change pas significativement en fonction de l'âge de méthylation tumoral.

Mesurer directement une association entre l'âge de méthylation et une variable clinique n'est peut-être pas approprié et il serait intéressant de prendre en compte les variations de l'âge chronologique, c'est pourquoi nous allons tester ces associations avec l'accélération de l'âge de méthylation selon la définition choisie plus tôt.

**Graphe de Kaplan-Meier pour les patients atteints d'un
adénocarcinome du poumon**

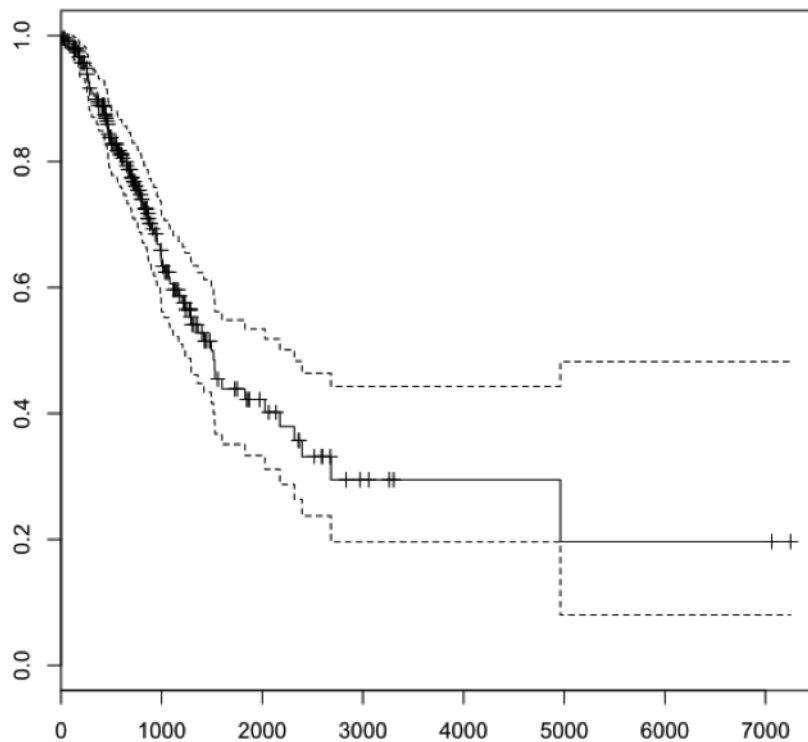


Figure4: Graphe de Kaplan-Meier représentant la survie des patients en fonction du temps pour les patients atteints d'un adénocarcinome du poumon. Chaque barre verticale représente un évènement censuré. L'axe des y représente le ratio de décès, tandis que l'axe des x représente l'évolution du temps en jour.

2. Il ne semble pas exister de lien entre l' accélération de l'âge méthylé et une variable clinique

Lorsqu'on compare l'accélération des âges méthylés tumoraux retrouvés dans les différents groupes décrivant les stades tumoraux (N,T,M) (Annexe Fig4.abc) on n'observe pas de différence. Les médianes sont similaires et les intervalles de confiance à 95% se recoupent. Le stade n3 (Annexe Fig4.a) a une médiane décalée cependant il ne comprend qu'un seul échantillon, tout comme le stade m1a ayant une médiane légèrement décalée.

L'accélération de l'âge méthylé ne semble pas significativement différente d'un groupe histologique à l'autre. Pour la grande majorité des échantillons, les médianes sont semblables et les intervalles de confiance se recoupent (Annexe Fig5a.). Comme précédemment, la catégorie "lung acinar adenocarcinoma" (rouge) est la seule ayant un intervalle de confiance autour de la médiane ne se recoupant pas avec une partie des autres groupes. Le test des rangs de Wilcoxon évalue la différence des valeurs retrouvées dans ce groupe par rapport à celles retrouvées dans le groupe "not otherwise specified (nos)". Comme évoqué précédemment, ce groupe contient le plus d'échantillons, l'intervalle de confiance le plus étroit et est l'intervalle qui ne se recoupe pas avec "lung acinar adenocarcinoma". Le test nous rend une p-valeur significative de 0.014. Nous rejetons l'hypothèse nulle selon laquelle les deux distributions sont égales. On peut donc conclure que ces distributions sont différentes. L'appartenance au groupe "lung acinar adenocarcinoma" semblerait significativement associée à une accélération de l'âge de méthylation plus basse que celle trouvée dans le groupe histologique "nos". De la même manière l'intervalle de "lung acinar adenocarcinoma" ne croise pas celui de "lung bronchioloalveolar carcinoma nonmucinous" (n=9). Le test des rangs de Wilcoxon nous retourne une p-valeur tout juste significative de 0.0496. Suggérant également l'accélération de l'âge de méthylation retrouvé dans le groupe "lung acinar adenocarcinoma" est significativement plus bas. À noter qu'il n'y a pas de chevauchement avec les catégories "micropapillary adenocarcinoma" et "lung bronchioloalveolar carcinoma mucinous" comprenant respectivement 2 et 1 échantillons, ce qui est trop peu pour réaliser un test statistique.

Aucun stade pathologique ne semble associé à une variation de l'accélération de l'âge méthylé (Annexe Fig5b.). À nouveau le stade I est décalé par rapport aux autres médianes mais ne comprend qu'un seul échantillon.

Lorsqu'on compare le nombre de paquets de cigarettes fumés par an à l'accélération de l'âge de méthylation on n'observe pas de corrélation significative entre les deux (Annexe Fig5c.).

Lorsqu'on compare l'accélération de l'âge méthylé à l'année de début de tabagisme on n'observe pas de subdivision particulière des groupes. On calcule une corrélation non significative de 0.1, p-valeur=0.26 (données non montrées).

Lorsqu'on compare la tendance de l'âge de méthylation à la prise de thérapie ou non, à la classe de la tumeur résiduelle, à l'origine ethnique du patient, à son sexe ou à son origine, à la classe de la tumeur résiduelle, ou à l'indice de Karnofsky, on n'observe aucune différence apparente entre les groupes. Les médianes sont très proches et les intervalles de confiance se recoupent (Annexe Fig6abcdef). Seul le score de Karnofsky à 40 semble avoir une médiane différente des autres, cependant ce groupe ne contient que 2 échantillons.

En réalisant l'analyse de Cox en fonction de l'accélération nous obtenons un coefficient bêta légèrement moins proche de 1 (0,987) qui est significatif avec $\Pr(>|z|) = 0,022$.

Une manière plus appropriée pour corriger l'âge de méthylation en fonction de l'âge chronologique est d'utiliser l'âge chronologique comme une covariable de l'âge de méthylation lors de l'analyse de Cox. Dans ce cas nous obtenons un coefficient bêta quasiment le même que précédemment (0,988) qui est également significatif avec $\Pr(>|z|) = 0,040$. Cette valeur étant à la limite de la significativité définie comme <0.05 et le coefficient est très proche de 1. En conséquence, on ne peut tirer de conclusion concernant l'influence de l'âge méthylé tumoral sur la survie et donc la progression du cancer.

Analyse de l'expression différentielle de gènes en fonction de l'âge de méthylation

1. Aucun geneset n'est significativement lié à l'âge de méthylation tumoral

Cette analyse a été réalisée avec la fonction `camera` du package `limma`.

Lorsque nous évaluons les genesets différenciellement exprimés avec l'âge de méthylation tumoral nous observons qu'aucun geneset des voies canoniques ne ressort significativement (False Discovery Rate (FDR) < 0.05). Néanmoins en ne s'intéressant qu'aux voies provenant des bases de données KEGG, Reactome et Biocarta, on observe que la voie des lymphocytes B a le FDR le plus bas (0.32).

Lorsque nous corrigeons l'âge de méthylation en définissant l'âge chronologique comme une covariable, nous éliminons la voie des lymphocytes B. Ceci suggère que celle-ci était liée à une augmentation de l'âge chronologique du patient plutôt que réellement à l'âge méthylé. En revanche, nous observons que la voie des récepteurs purinergique a un meilleur FDR

qu'avant (0.24). Celle-ci est activée avec l'augmentation de l'âge de méthylation. La voie des récepteurs purinergiques implique plusieurs fonctions cellulaires telles les réactions vasculaires, l'apoptose et la sécrétion de cytokines. Un FDR de 0.24 n'est pas significatif mais ne signifie pas forcément que cette voie doit être exclue. En effet, un tel FDR signifie qu'il y a 24% de chance que cette voie ait été trouvée par hasard suite à l'analyse Camera et est donc un faux positif.

La correction ne permet donc pas de faire sortir des voies métaboliques avec un FDR < 0.05. Ceci signifie que toutes les voies annoncées comme différentiellement exprimées avec l'âge de méthylation ont une probabilité supérieure à 5 % d'avoir été trouvées par hasard (Table 2).

2. Aucun gène ne semble fortement surexprimé ou réprimé avec l'âge de méthylation

L'analyse par Limma permet de trouver des gènes différentiellement exprimés avec l'âge méthylé. En ne s'intéressant qu'aux gènes différentiellement exprimés avec un FDR < 0.05 on trouve 6031 gènes. Néanmoins en y regardant de plus près la surexpression maximale significative est pour IL1F7 avec un log Fold Change (LFC) de 0.08 et un FDR de 1.83×10^{-16} . Le LFC est le logarithme en base 2 du Fold Change qui est la différence d'expression d'un gène mesurée entre les échantillons ayant un âge de méthylation plus et moins élevé. Un LFC de 0.08 équivaut donc à une surexpression d'environ 1.05 fois plus, ce qui est peu. Le gène le plus sous exprimé est le gène PRAME avec un LFC de -0.11 et un FDR de 7.72×10^{-21} . Ceci équivaut à une sous expression de 0.93, très peu également (Table 3).

Lorsqu'on corrige l'influence de l'âge chronologique sur les résultats obtenus pour l'âge de méthylation, on observe que le gène PRAME reste le gène avec le meilleur FDR et l'expression différentielle la plus élevée. Néanmoins son LFC reste très bas -0.12. On observe que l'expression différentielle de certains gènes est moins significative suite à la correction tandis qu'elle l'est plus pour d'autres (Table 4).

PRAME est un antigène préférentiellement exprimé dans le mélanome reconnu par les lymphocytes T. Il agit comme un répresseur au récepteur de l'acide rétinoïque et confère un avantage de croissance au cancer via cette fonction. Ce gène semble donc légèrement moins

exprimé lorsque l'âge de méthylation tumoral augmente. Suggérant qu'une tumeur "plus jeune" pourrait avoir un léger avantage de croissance comparé à une tumeur "plus âgée".

Table1 Analyse Camera: voies métaboliques différentiellement activées en fonction de l'âge de méthylation, 20 voies métaboliques avec le FDR le plus petit. La direction Up signifie que la voie est activée.

Voies métaboliques	Direction	FDR
BIOCARTA_BLYMPHOCYTE_PATHWAY	Up	0,316504343
BIOCARTA_ASBCCELL_PATHWAY	Up	0,316504343
KEGG_ALLOGRAFT_REJECTION	Up	0,316504343
KEGG_ASTHMA	Up	0,316504343
KEGG_TYPE_I_DIABETES_MELLITUS	Up	0,316504343
REACTOME_INTERFERON_GAMMA_SIGNALING	Up	0,316504343
KEGG_LEISHMANIA_INFECTION	Up	0,316504343
REACTOME_P2Y_RECEPTORS	Up	0,316504343
REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS	Up	0,316504343
REACTOME_INFLAMMASOMES	Up	0,316504343
REACTOME_THE_NLRP3_INFLAMMASOME	Up	0,316504343
REACTOME_DOWNSTREAM_TCR_SIGNALING	Up	0,316504343
KEGG_VIRAL_MYOCARDITIS	Up	0,316504343
REACTOME_ACTIVATION_OF_IRF3_IRF7_MEDIATED_BY_TBK1_IKK_EPSILON	Up	0,316504343
REACTOME_EXTRINSIC_PATHWAY_FOR_APOPTOSIS	Up	0,316504343
REACTOME_SYNTHESIS_OF_PIP2_AT_THE_PLASMA_MEMBRANE	Up	0,316504343
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	Up	0,316504343
REACTOME_IKK_COMPLEX_RECRUITMENT_MEDIATED_BY_RIP1	Up	0,316504343
KEGG_FC_EPSILON_R1_SIGNALING_PATHWAY	Up	0,316504343
REACTOME_TOLL_RECEPTOR_CASCADES	Up	0,316504343

Table2: Analyse Camera: voies métaboliques différentiellement exprimées en fonction de l'âge de méthylation ajusté pour l'âge chronologique, 20 voies métaboliques avec le FDR le plus petit. La direction Up signifie que la voie est activée.

Voies métaboliques	Direction	FDR
REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS	Up	0,23962988
KEGG_APOPTOSIS	Up	0,23962988
REACTOME_TOLL_RECEPTOR_CASCADES	Up	0,23962988
REACTOME_ACTIVATED_TLR4_SIGNALING	Up	0,23962988
REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE	Up	0,23962988
REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	Up	0,23962988
REACTOME_IKK_COMPLEX_RECRUITMENT_MEDIATED_BY_RIP1	Up	0,249399817
REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	Up	0,249399817
REACTOME_TRAF6_MEDIATED_INDUCED_OF_NFKB_AND_MAP_KINASES_UPON_TLR7_8_OR_9_ACTIVATION	Up	0,249399817
REACTOME_ERK_MAPK_TARGETS	Up	0,300966056
KEGG_ASTHMA	Up	0,331034577
REACTOME_P2Y_RECEPTORS	Up	0,331034577
KEGG_LEISHMANIA_INFECTION	Up	0,331034577
REACTOME_EXTRINSIC_PATHWAY_FOR_APOPTOSIS	Up	0,331034577
REACTOME_INFLAMMASOMES	Up	0,331034577
REACTOME_ACTIVATION_OF_IRF3_IRF7_MEDIATED_BY_TBK1_IKK_EPSILON	Up	0,331034577
REACTOME_DOWNSTREAM_TCR_SIGNALING	Up	0,331034577
BIOCARTA_CBL_PATHWAY	Up	0,331034577
REACTOME_SYNTHESIS_OF_PIP2_AT_THE_PLASMA_MEMBRANE	Up	0,331034577
BIOCARTA_TCR_PATHWAY	Up	0,331034577

Table3 Analyse Limma: gènes différentiellement exprimés en fonction de l'âge de méthylation (20 gènes les plus significatifs)

Gènes	logFold Change	Expression moyenne	FDR
PRAME	-0,11	0,94	7,72E-21
DRGX	0,07	-4,13	6,73E-18
STK32A	0,05	1,74	2,65E-17
IL1F7	0,08	-1,19	1,83E-16
IGF2BP1	-0,08	-1,46	1,70E-15
MMP28	0,05	3,44	2,61E-15
MSI1	-0,06	-0,23	1,56E-14
TLR2	0,03	5,25	1,67E-13
DPYD	0,03	5,42	1,72E-13
ECEL1	-0,06	-0,71	1,72E-13
CD97	0,02	6,82	3,24E-13
SLC15A1	-0,06	-0,87	3,24E-13
PFN2	-0,03	6,20	5,62E-13
ICAM1	0,03	8,62	5,80E-13
LOC100130776	0,02	2,74	5,85E-13
SPERT	-0,06	-2,93	6,40E-13
UBA7	0,02	5,58	1,10E-12
ANKRD24	0,03	0,18	1,25E-12
ATP10A	0,03	3,98	1,64E-12
FAS	0,02	4,39	1,64E-12

Table 4 Analyse Limma: gènes différentiellement exprimés en fonction de l'âge de méthylation ajusté pour l'âge chronologique (20 gènes les plus significatifs)

Gènes	logFold Change	Expression moyenne	FDR
PRAME	-0,12	0,96	1,25E-16
DRGX	0,07	-4,29	3,89E-14
IGF2BP1	-0,09	-1,50	9,74E-13
STK32A	0,05	1,67	1,01E-12
MSI1	-0,07	-0,25	1,01E-12
HOXB13	-0,09	-2,32	2,62E-12
MMP28	0,05	3,29	1,70E-10
ATP10A	0,03	3,96	2,45E-10
LOC100130776	0,03	2,78	2,78E-10
ECEL1	-0,06	-0,77	3,41E-10
CASKIN1	-0,05	-1,29	4,25E-10
STEAP3	0,03	6,24	5,64E-10
CD97	0,02	6,76	1,37E-09
WNK2	-0,05	2,54	1,37E-09
KIF1A	-0,09	0,29	2,00E-09
SLC15A1	-0,06	-0,90	2,24E-09
ETV1	0,03	4,99	4,72E-09
DVWA	-0,04	-3,59	4,72E-09
EPHA10	0,04	2,73	6,40E-09
TLR2	0,03	5,25	6,48E-09

Conclusion

En conclusion mesurer l'âge de méthylation d'une tumeur est assez aisé et pourrait donc être révolutionnaire si l'on arrivait à démontrer un quelconque lien entre cet âge et certaines variables cliniques pronostiques, certains gènes ou certaines voies métaboliques impliqués

dans la tumorigénèse. L'étude réalisée jusqu'à présent suggère cependant qu'il est difficile de tirer de telles conclusions en se basant simplement sur l'âge méthylé, même si celui-ci est ajusté pour l'âge chronologique. Néanmoins, l'âge de méthylation pourrait être utilisé en combinaison avec une autre variable afin d'identifier des associations plus significatives.

Matériel et Méthodes

Matériel

Les données RNAseqv2_RSEM_genes et Clinical_Pick provenant du *The Cancer Genome Atlas* ont été téléchargées à partir du serveur The Genome Data Analysis Centers (GDACs) firehose version 2016_01_28 (Broad Institute TCGA Genome Data Analysis Center (2016): Firehose stddata__2016_01_28 run. Broad Institute of MIT and Harvard. doi:10.7908/C11G0KM9). Les données concernant l'âge de méthylation ont été calculées par le Pr. Vincent Detours.

Les 6 premiers chiffres des noms d'échantillons reprennent l'identifiant du patient (p.ex. 05-4244), les deux chiffres suivant (p.ex. 01) indiquent la nature du prélèvement, les chiffres suivants identifient les aliquots et ne seront pas utilisés.

Les données provenant d'échantillons sanguins (n° "10") de tumeurs récurrentes (n° "02") ou de métastases (n° "06") ont été éliminées. En conséquence, nous avons utilisé uniquement les échantillons provenant de tumeurs solides (identifiés par "01") et de tissus sains solides adjacents à la tumeur (identifié par "11"). Comparer nos données tumorales aux échantillons sanguins considérés comme tissu sain ne nous semblait pas pertinent car la transcription des gènes est spécifique à chaque tissu.

Dans les données concernant l'âge méthylé, nous avons quelques échantillons (3) pour lesquelles deux aliquots étaient disponibles pour le prélèvement tumoral du même patient p.ex: "TCGA.44.2656.01A.02D.A276.05" âge méthylé: 86.05057

"TCGA.44.2656.01A.02D.0945.05" âge méthylé: 84.27625

Les âges méthylés étaient semblables mais pas identiques. Nous avons fusionné ces échantillons en leur attribuant la valeur correspondant à la moyenne de ces deux âges méthylés.

Les valeurs de compte de RNA-seq non normalisées ont été récupérées dans la colonne "raw_count" pour chaque échantillon. Les données de RNA-seq présentaient 29 sondes de gènes non identifiés ("?",) qui ont été systématiquement éliminés. Deux sondes étaient identifiées par le même symbole: "SLC35E2". Nous avons renommé la deuxième "SLC35E2bis".

Les âges chronologiques ont été récupérés dans les données cliniques sous la catégorie "years_to_birth". Les patients pour lesquels l'âge chronologique n'était pas disponible (NA) ont été éliminés de l'étude.

Méthodes

Toutes les analyses ont été réalisées avec le logiciel R version 3.2.4 Revised (2016-03-16 r70336). La majorité des packages utilisés proviennent du projet Bioconductor.

Analyse d'association avec l'âge chronologique et l'âge méthylé normal

La fonction `intersect` nous a permis de sélectionner uniquement les échantillons communs aux données utilisées pour chaque analyse.

Les graphiques *scatter plot* ont été tracés via la fonction `plot`.

Les corrélations de Spearman ont été calculées via la fonction R `cor.test`. Le coefficient de Spearman et sa p-valeur (représentant la probabilité qu'une telle corrélation soit trouvée par hasard alors qu'il n'y en pas en réalité) ont été arrondis au 10 000^e via la fonction `round`. Concernant les points 1,2 et 3, l'idée était, entre autres, de comparer la différence entre les valeurs chronologiques et méthylées. Nous avons donc tracé une droite $x=y$ plutôt qu'une droite de régression linéaire. La droite $x=y$ indique l'endroit où devraient se situer les points si les âges étaient exactement les mêmes. Le calcul de corrélation de Spearman nous indique la force et la direction de l'association entre ces deux variables.

Afin d'évaluer la normalité des distributions en ce qui concerne les différents calculs d'accélération nous avons tout d'abord réalisé un "Q-Q plot". Le Q-Q plot ou diagramme quantile-quantile permet d'évaluer la pertinence de l'ajustement d'une distribution à un modèle théorique. Le modèle théorique est ici une distribution normale. Le graphique de la distribution des points d'accélération sera réalisé via la fonction `qqnorm` et le modèle théorique sera ajouté dans le même graphique via la fonction `qqline`.

La représentation des histogrammes couplée au graphique de densité a été réalisée via la fonction `hist` à laquelle on ajoute la ligne de densité obtenue par la fonction `density`.

Analyse d'association avec les variables cliniques

La comparaison de l'âge méthylé des cancers ou de son accélération avec les variables cliniques a été réalisée par des représentations graphiques en boxplot via la fonction `boxplot`. Un boxplot représente des groupes de valeurs numériques à travers leurs quartiles (valeurs divisant les données en quatre groupes reprenant chacun un quart des données). Les boxplots possèdent également des "moustaches" représentant la variabilité hors des deux quartiles les plus extrêmes. Les points individuels représentent les outliers. L'avantage des représentations boxplots est qu'elles n'émettent aucune hypothèse sur la distribution des données qui ne doivent donc pas être distribuées normalement. Le paramètre `notch` de ces boxplot a été fixé à `TRUE`. Ceci ajoute un rétrécissement autour de la médiane qui représente l'intervalle de confiance à 95%. La largeur de ces *notches* est donc proportionnelle à l'intervalle interquartile de l'échantillon et inversement proportionnel à la racine carrée de la taille de l'échantillon. D'après *Graphical methods for data analysis* (Chambers, 1983) deux *notches* ne se chevauchant pas peut être considéré comme une "preuve solide" que les deux médianes diffèrent. Au contraire, si ces deux *notches* se chevauchent on peut affirmer que les deux médianes sont similaires (ne diffèrent pas). Nous utiliserons cette propriété pour évaluer s'il existe une différence entre les distributions des âges méthylés en fonction d'une variable clinique. La droite bleue est tracée de manière à traverser un maximum d'intervalles de confiances appartenant aux différents groupes. Si celle-ci traverse tous les intervalles cela signifie que les *notches* se chevauchent et nous pouvons affirmer que les médianes d'âge méthylé ou d'accélération des différents groupes sont similaires.

Dans le cas où deux intervalles ne se chevauchent pas, la différence entre les deux groupes peut être évaluée via un test des rangs de Wilcoxon. Ce test vérifie si la distribution des valeurs appartenant à deux groupes est la même. Une p-valeur < 0.05 nous permet de rejeter l'hypothèse nulle selon laquelle les deux distributions sont les mêmes, et donc d'affirmer que les distributions sont différentes. Ce test ne nécessite pas que les distributions soient normales et est réalisé via la fonction R `wilcox.test`.

La variable clinique "tumor_tissue_site" n'a pas été testée puisqu'elle ne contient qu'une seule catégorie ("lung").

L'analyse de survie a été réalisée via le package *Survival*. L'analyse de Cox pour évaluer une association entre la survie et l'âge de méthylation ou l'accélération de l'âge méthylée a été réalisée via la fonction `coxph`. Les colonnes "days_to_death" et "days_to_last followup" mutuellement exclusives ont été fusionnées afin de prendre en compte les évènements censurés.

Analyse d'association avec une expression différentielle

À l'aide des packages `limma` et `edgeR` de Bioconductor, la matrice de comptes de RNA-seq non normalisée a été transformée en objet liste par la fonction `DGEList`. Ensuite, la variance des données de RNA-seq a été stabilisée via la fonction `voom`. Suite à ces transformations nos données sont traitables par des algorithmes de détection de l'expression différentielle classiques conçus au départ pour les microarrays.

L'analyse des voies différentiellement activées en fonction de l'âge méthylé tumoral a été réalisée avec la fonction `Camera` disponible à partir du package `Limma`. Le design de notre analyse a été précisé via la fonction `model.matrix` en utilisant la variable continue "Âges méthylés tumoraux". Les groupes de gènes utilisés pour cette étude proviennent de "C2: curated gene sets" de la base de données MSigDB (<http://bioinf.wehi.edu.au/software/MSigDB/>). Les données ont été téléchargées directement sous forme d'objet R. Nous ne nous sommes intéressés qu'aux groupes de gènes impliqués dans les voies métaboliques canoniques répertoriées dans les bases de données Biocarta, Reactome et Kegg que nous avons récupérée à l'aide de la fonction `grep`.

L'analyse de l'expression différentielle des gènes en fonction de l'âge méthylé a été réalisée avec le package `limma`. Le design de l'analyse a été déterminé de la même manière que pour l'analyse `camera`. La fonction `lmFit` permet l'utilisation d'un modèle linéaire afin d'analyser l'entièreté des données comme un ensemble. L'utilisation du modèle empirique de Bayes via la fonction `eBayes` permet le partage d'informations entre les échantillons.

Nous avons réalisé ces deux analyses (`limma` et `camera`) une seconde fois en tenant compte des variations de l'âge chronologique pouvant influencer l'âge de méthylation. Pour ce faire, nous avons défini un design multivarié " ~ Âge méthylé tumoral + âge chronologique" via la fonction `model.matrix`.

3. Références bibliographiques

Chambers, J.M. (1983). Graphical methods for data analysis (Wadsworth International Group).

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* *14*, R115.

Annexes

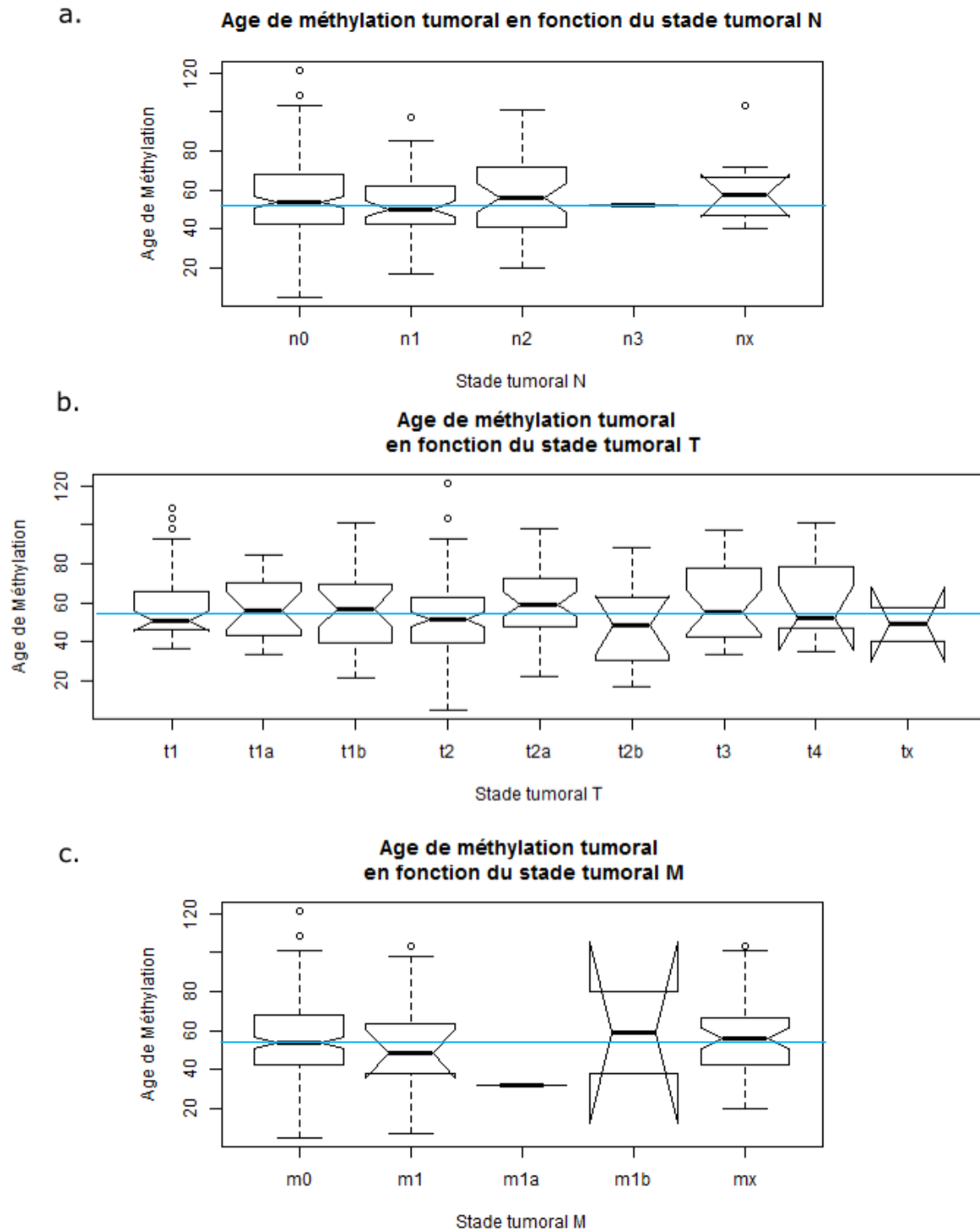
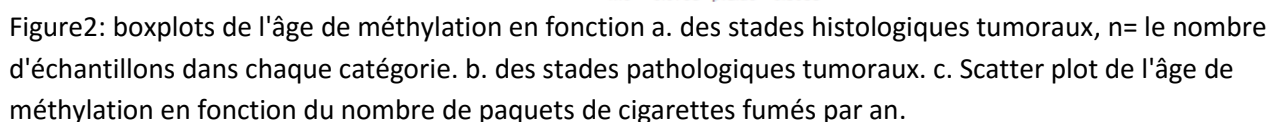
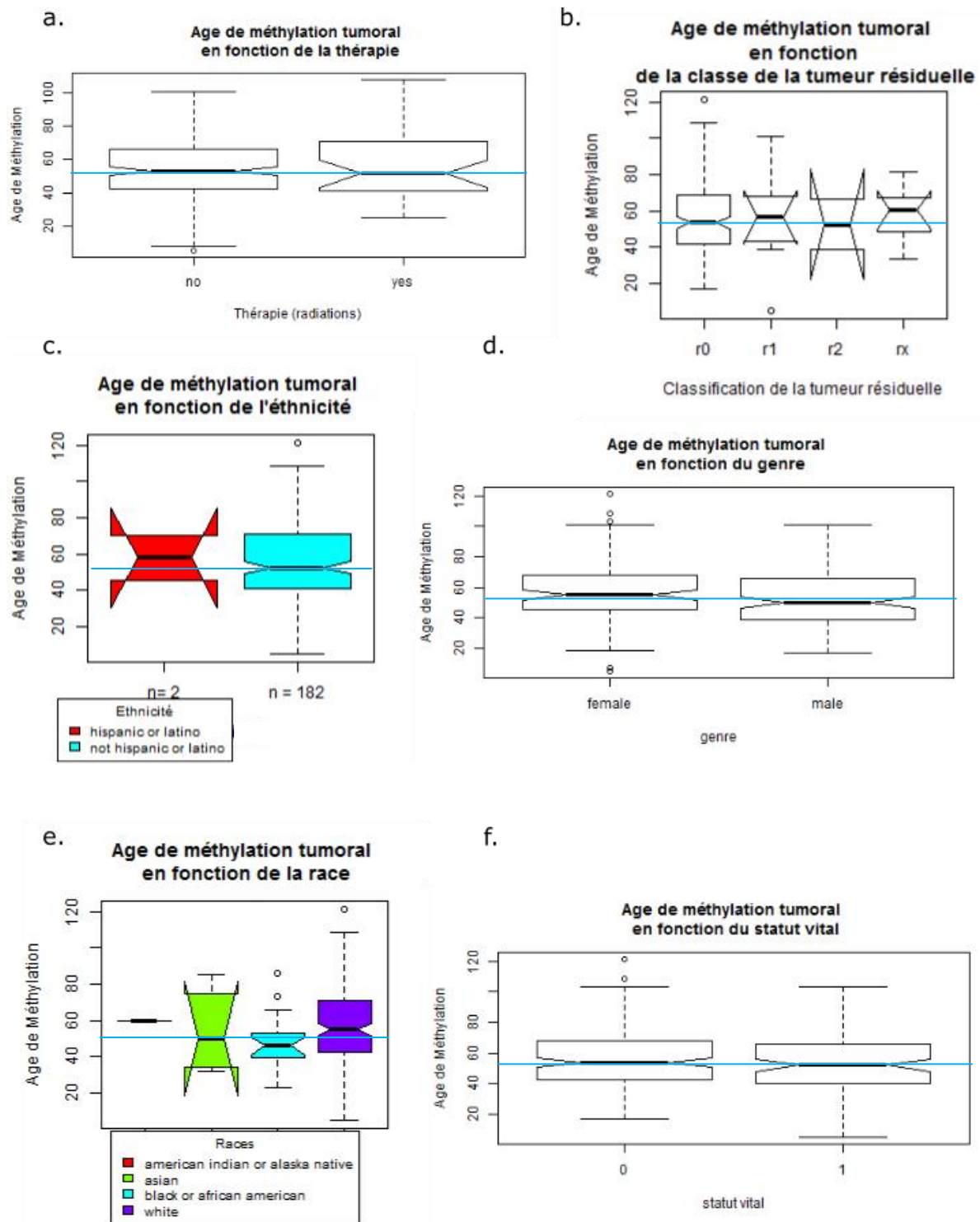


Figure1: boxplots de l'âge de méthylation en fonction a. de l'extension tumorale dans les nœuds lymphatiques. b. du stade tumoral de la tumeur primitive. c. de l'extension métastatique. Les droites bleues sont tracées de manière à maximiser le nombre d'intervalles (*notches*) qu'elles traversent.





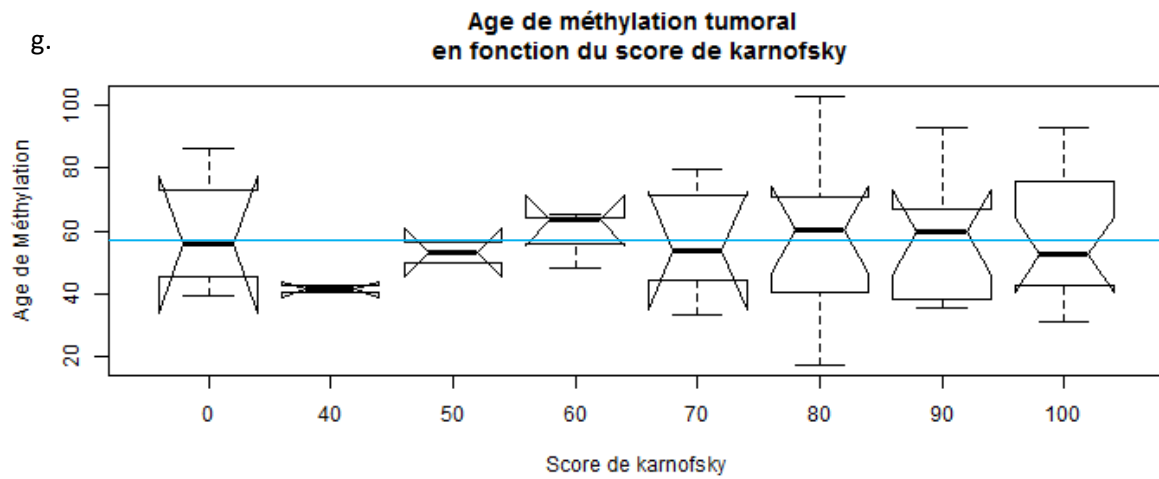


Figure 3. Boxplots des âges de méthylation en fonction a. du suivi d'une thérapie, b. de la classe de la tumeur résiduelle, c. de l'ethnicité, d. du sexe, e. de l'origine, f. du statut vital (0 en vie ,1 décédé) g. du score de Karnofsky

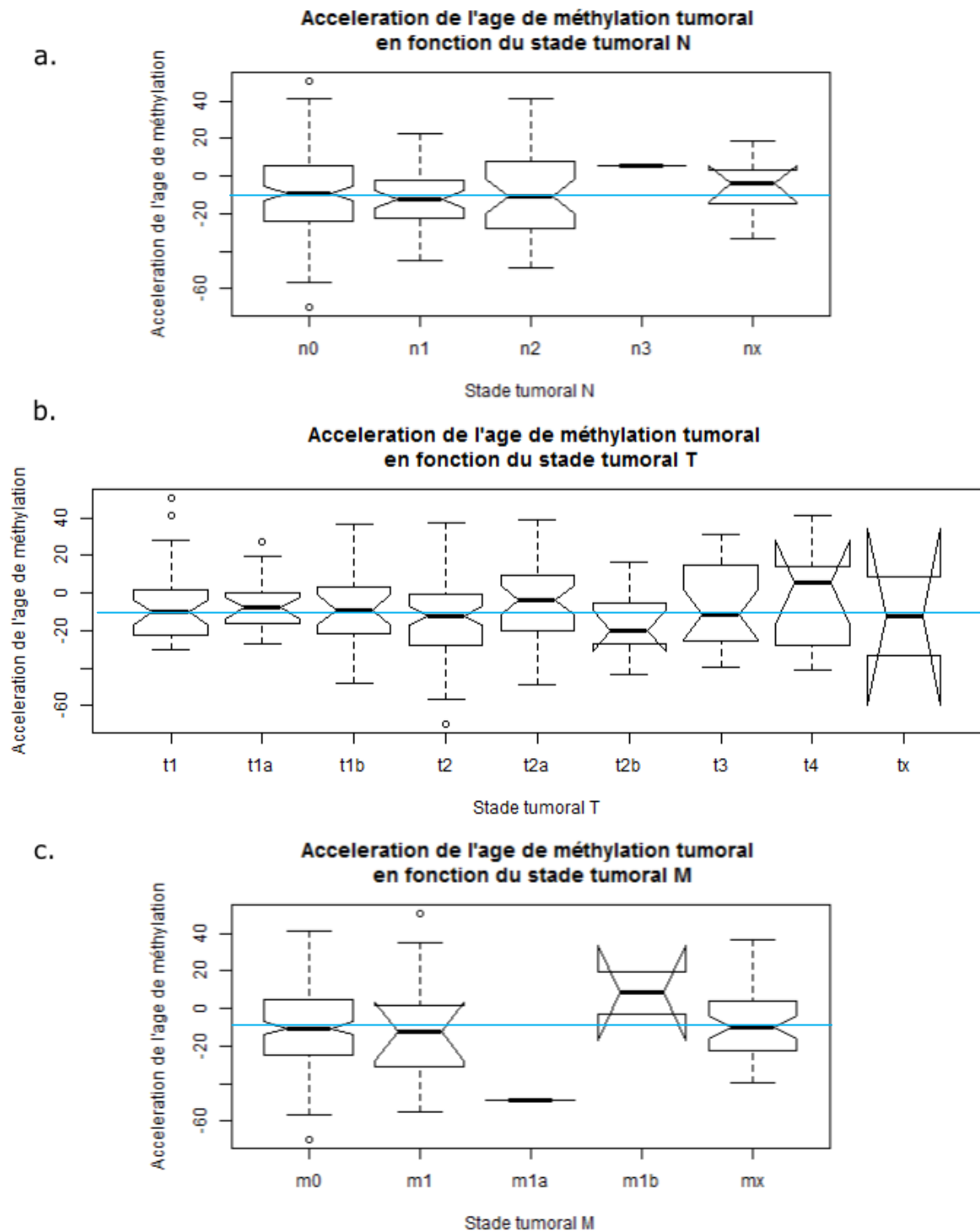


Figure4: boxplots de l'accélération de l'âge de méthylation en fonction a. de l'extension tumorale dans les nœuds lymphatiques. b. du stade tumoral de la tumeur primitive. c. de l'extension métastatique.

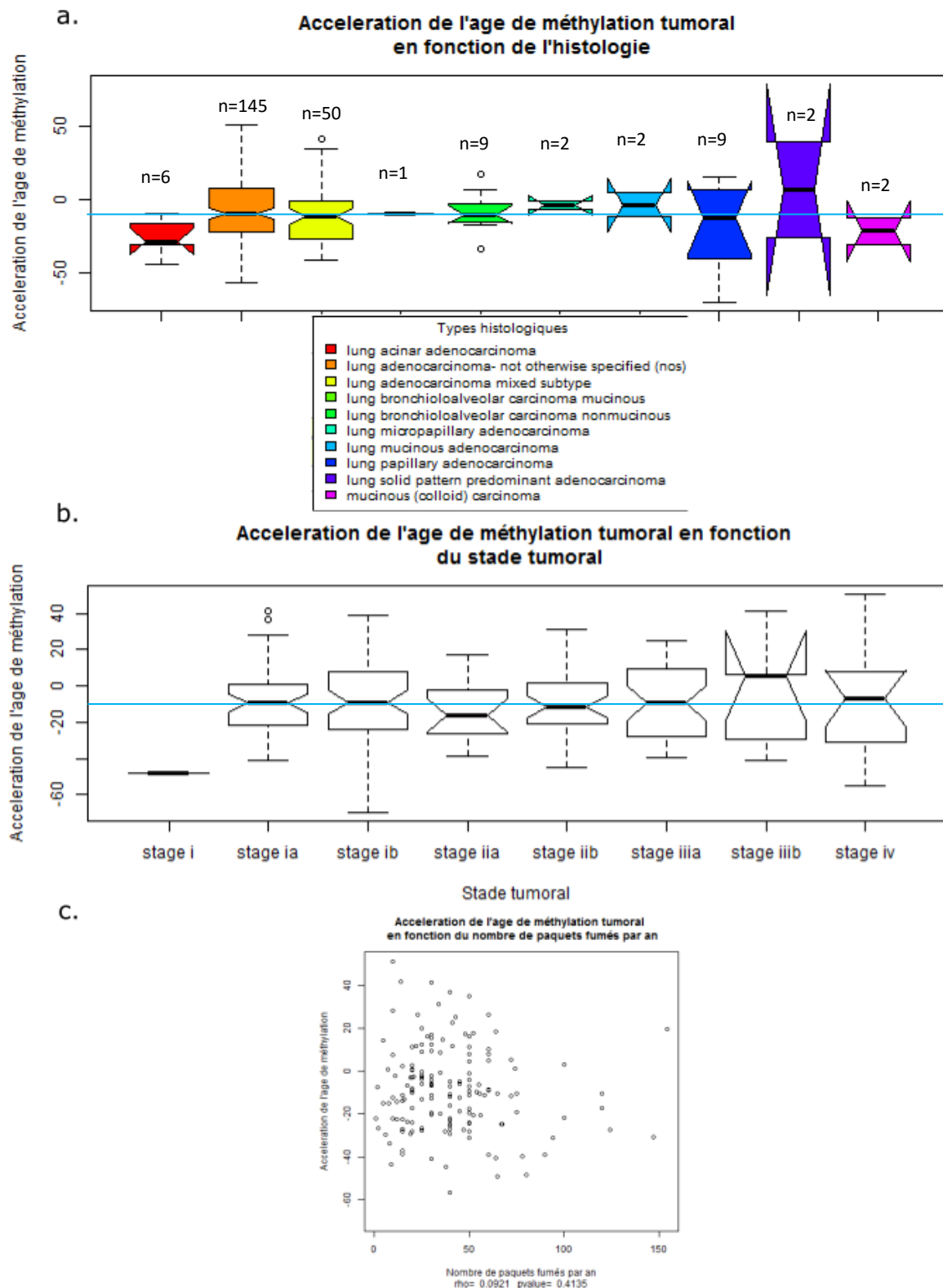
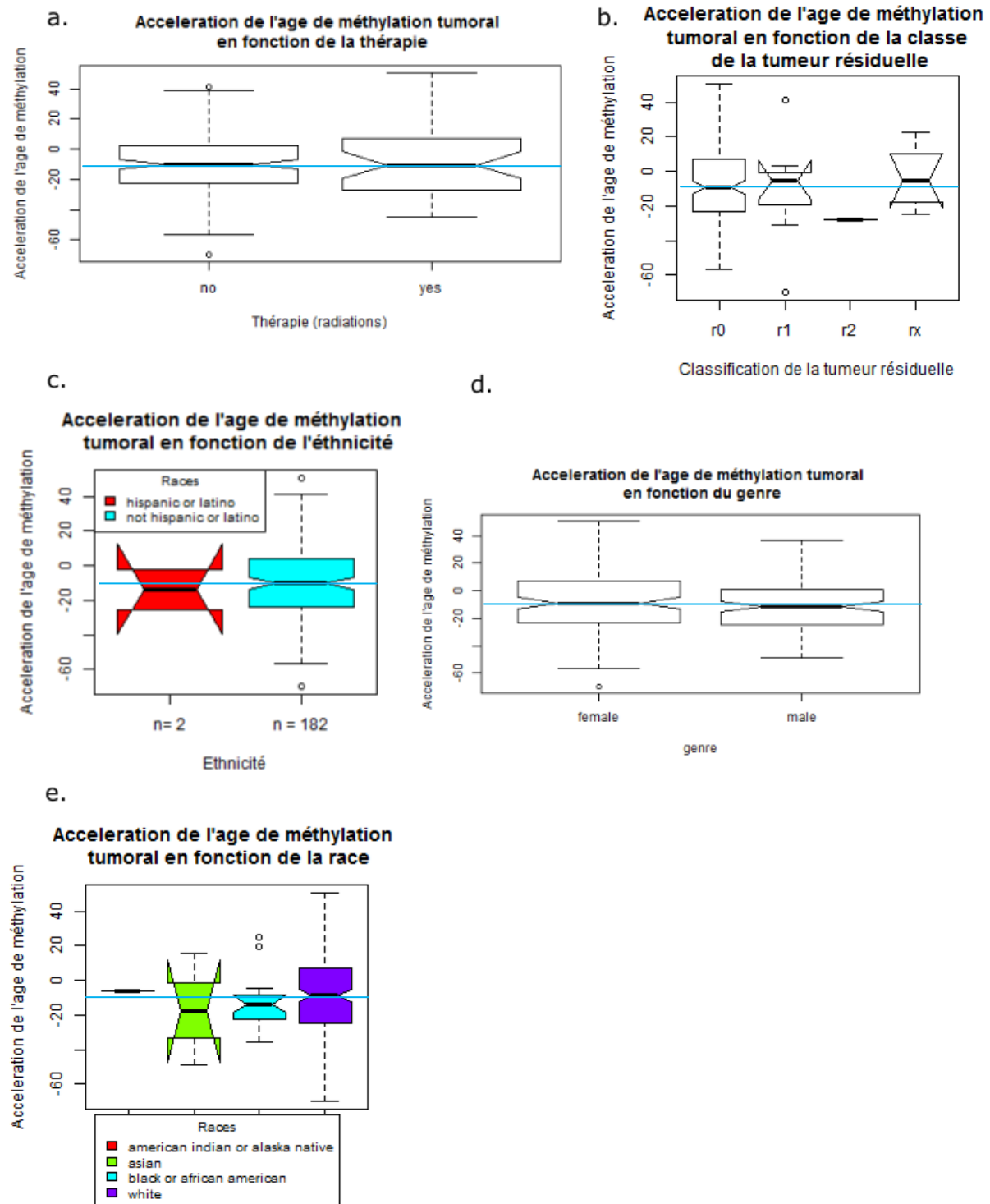


Figure5: boxplots de l'accélération de l'âge de méthylation en fonction a. des stades histologiques tumoraux, n= le nombre d'échantillons dans chaque catégorie. b. des stades pathologiques tumoraux.c. Scatter plot de l'âge de méthylation en fonction du nombre de paquets de cigarettes fumés par an



f.

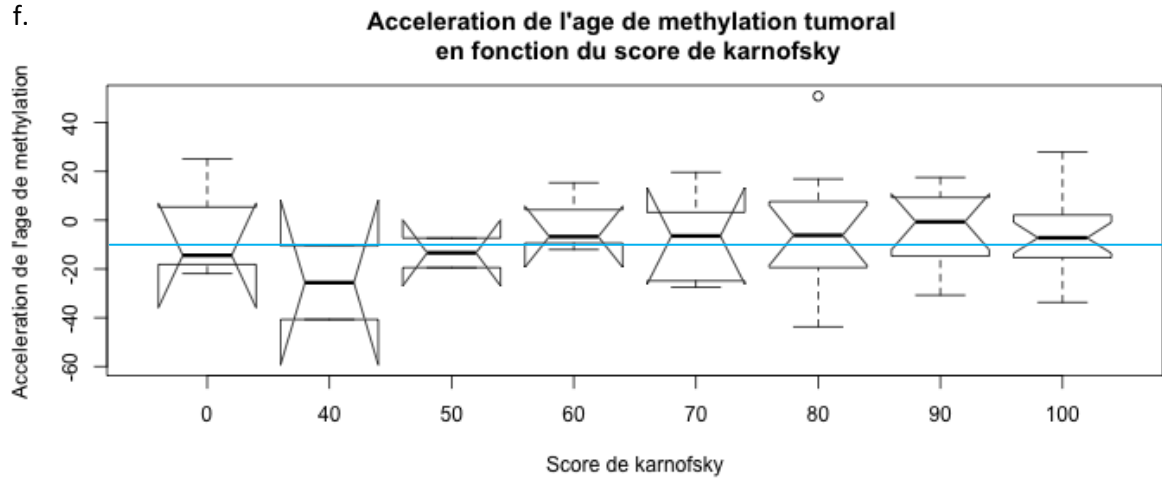


Figure 6. Boxplots de l'accélération de l'âge de méthylation en fonction a. du suivi d'une thérapie. b. de la classe de la tumeur résiduelle. c. de l'ethnicité. d. du sexe. e. de l'origine. f. du score de karnofsky